FORENSICWCT: WHITENING AND COLORING TRANSFORM IN FORENSIC LATENT SPACE FOR ROBUST LOCALIZATION OF AI-EDITED IMAGE REGIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The spread of generative image editing tools poses a growing threat to digital safety. Developing forensic models with strong generalization is therefore critical, yet remains challenging due to two key issues: (1) semantic bias, where models rely on spurious correlations between image content and generator artifacts, and (2) the limited diversity of localization data, which restricts generalization to unseen forgery techniques. To address these challenges, we propose to **reformulate forgery localization as a forensic style classification problem**. We conceptualize generator-specific artifacts as a distinct "style," enabling the disentanglement of forensic traces from semantic content. Building on this formulation, we present ForensicWCT, a novel training framework that instantiates this concept. During training, ForensicWCT applies a whitening transform to decompose image features into semantic content and forensic style components. It then performs a style-swapping augmentation by exchanging styles across samples, while dynamically updating the corresponding ground-truth labels. This training strategy, which does not introduce additional computational cost at inference time, forces the model to learn content-agnostic artifact representations. Moreover, by elegantly injecting styles from fully-generated images into arbitrary regions, our approach enables the effective use of large-scale, diverse, fully-generated image datasets to improve localization generalization. Extensive experiments on multiple benchmarks demonstrate the effectiveness of our ForensicWCT.

1 INTRODUCTION

The rapid advancement in generative AI empowers users to produce highly realistic image forgeries effortlessly. This capability blurs the line between authentic and manipulated content, posing significant risks to information integrity and public trust Ma et al. (2024). Malicious actors can subtly alter images for fraudulent purposes, such as faking evidence or creating disinformation. Consequently, it is critical to develop robust methods for localizing AI-generated forgeries.

However, training a generalized forgery localization model remains a major challenge due to two fundamental obstacles: **First, semantic bias**. The subtle artifacts left by a generative model are inherently interwoven with the visual content of the image. During training, a forensic model can easily learn to associate specific semantic content (e.g., object categories, particular appearances) with forgery, rather than learning the underlying manipulation artifacts themselves Sun et al. (2023). This "shortcut" results in overfitting to the content of training data, leading to limited generalization when facing images with different semantics. **Second, data scarcity and task mismatch.** Effective generalization hinges on training with diverse data, yet current pixel-level forgery datasets suffer from **a critical lack of generator diversity** (< 50 generators). While large, diverse datasets of fully-generated images are available (≈ 5000 generators), they cannot be used directly due to a fundamental task mismatch: image classification models learn by aggregating global features, whereas localization must analyze local artifacts on a granular, region-by-region basis.

To address these challenges, we reformulate forgery localization as a problem of image style classification. Our core insight is that manipulation artifacts share intrinsic properties with image style: both strongly relate to texture and pattern, independent of the semantic content or shape they inhabit. Distinct from generic visual style, we conceptualize generator artifacts as a unique "forensic style." Within this framework, authentic regions possess an "authentic style," while AI-forged regions exhibit a "generated style."

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

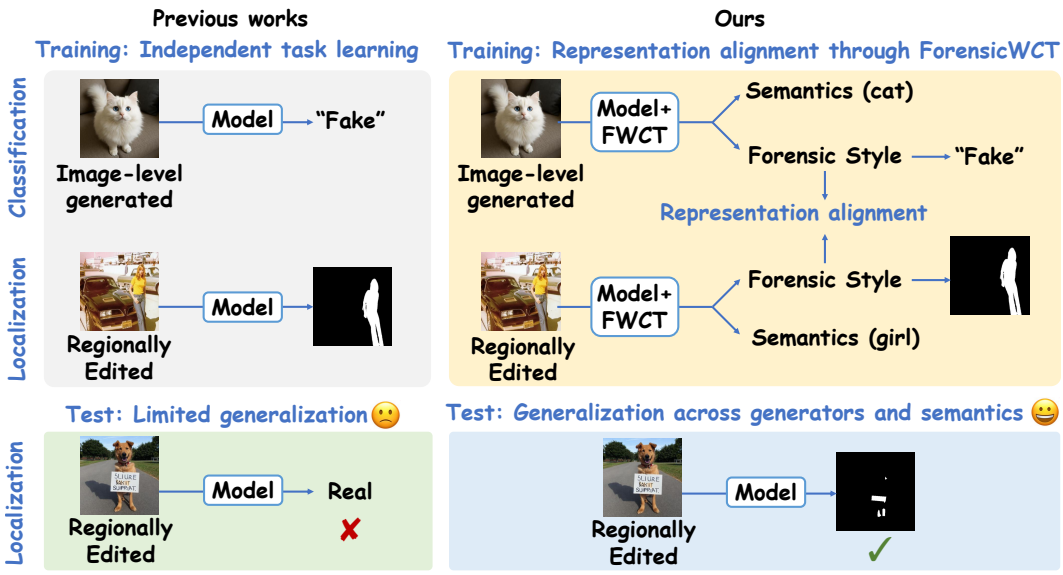


Figure 1: Previous methods train on isolated tasks and yield poor generalization. By contrast, our approach treats forgery artifacts as a unique image style, disentangled from semantic content. Unifying image- and pixel-level forensic tasks via style transfer, our method delivers better generalization.

This reframing is powerful. By forcing the model to classify a style that is naturally independent of content, it directly mitigates semantic bias. Furthermore, this provides a theoretical bridge to unify image-level and pixel-level data. By treating the artifacts of a fully-generated image as a transferable style, we can synthetically apply this style to arbitrary regions in other images, effectively creating diverse, pixel-annotated training samples as if they were generated by that image generator. This allows us to leverage vast, image-level datasets to improve a forgery localization model, elegantly addressing the challenge.

Building on this insight, we introduce ForensicWCT, a novel training framework. During training, ForensicWCT uses a forensic whitening transform to decouple feature representations into content and forensic style. It then creates new hybrid training examples by swapping the styles between different feature maps and recombining them with a forensic coloring transform. Crucially, the ground-truth segmentation masks are dynamically updated to reflect the newly applied style. For example, if a "generated style" is applied to the content of an authentic region, the ground-truth for that region is temporarily changed to "generated." The same semantic features, when combined with different forensic styles, can result in completely opposite ground-truths. This style swapping forces the decoder to be semantically unbiased. In addition, by swapping with forensic styles from the large-scale diverse image-level generated data, the forgery localization data is significantly augmented, leading to superior model generalization. Notably, our framework is active only in training and does not add any storage or computational burden in inference, providing free improvement.

We conducted extensive experiments on multiple representative benchmarks. The superior performance of our method validates its effectiveness and demonstrates a significant step forward in robust AI-generated content localization. For example, our model achieves 7.7 points (18%) higher average IoU than MaskCLIP (CVPR'25), while being much simpler (122M versus 535M).

In summary, our key contributions are:

- We introduce a new paradigm for forgery localization by reformulating it as a forensic style classification problem, which naturally decouples manipulation artifacts from semantic content and enables the effective use of diverse image-level generated data.
- ForensicWCT, a novel effective training framework that instantiates this idea using forensic whitening/coloring transforms and a dynamic style-swapping augmentation strategy.
- Through extensive experiments, we demonstrate that our method establishes a new state of the art, significantly outperforming previous approaches in generalizing to unseen generators and semantic content.

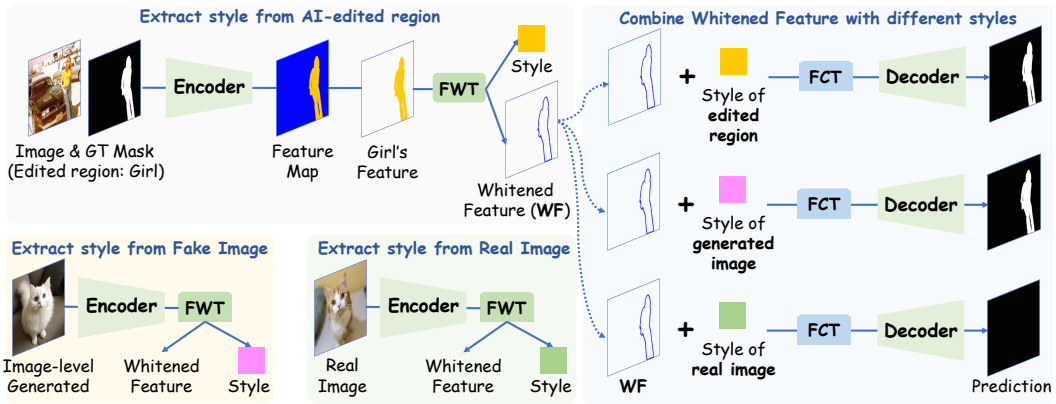


Figure 2: The core idea of ForensicWCT: training semantically unbiased forensic models and making full use of diverse fully-generated images through dynamic style swapping.

2 METHOD

2.1 CORE IDEA

To address semantic bias and the lack of data diversity, we propose to reframe forgery localization as a style classification task by conceptualizing manipulation traces as a distinct "forensic style." This reframing is motivated by the observation that manipulation artifacts, much like image style, are characterized by texture and pattern, independent of semantic content or shape. This allows us to leverage established style transfer theories Huang & Belongie (2017); Li et al. (2017; 2019) to separate style from content.

The core mechanism of our method is style swapping, which elegantly solves both fundamental challenges. First, to mitigate semantic bias, we inject different forensic styles (e.g., real vs. AI-generated) into the same content features and dynamically update the ground-truth labels accordingly. This forces the model to base its predictions solely on the transferred style, breaking any spurious correlation with content. Second, to bridge the task gap, we leverage the fact that, since the style descriptor matrix shape depends only on feature depth, style transfer can be flexibly applied across features of varying spatial sizes and contours. We can extract styles from large-scale, diverse fully-generated images and inject them flexibly into arbitrary regions, treating the resulting style-transferred region as if it were edited by the generator of the style donor. This directly bridges the gap between image-level generated data and region-level edited images, greatly alleviating the data diversity limitation in forgery localization.

We realize this concept through our ForensicWCT (FWCT) framework, as illustrated in Figure 2. The process begins by using a Forensic Whitening Transform (FWT) to decouple features into content and style components. The framework then swaps these style components between different samples and recombines them with the original content using a Forensic Coloring Transform (FCT). Crucially, pixel-level ground-truth masks are dynamically updated to match the newly injected style. For example, if an AI-generated style is transferred to a region, its ground-truth mask is updated to "manipulated." Our full implementation encompasses four key swapping operations, as detailed in Figure 12: swapping styles (I) between partially-edited and fully-generated images, (II) between different partially-edited images, (III) between real and fully-generated images, and (IV) between real and partially-edited images. As demonstrated in our ablation studies (Section 3.3), each of these operations contributes to the model's overall generalization capability.

2.2 THEORETICAL BASIS OF LEARNING STYLE TRANSFER

We begin by establishing the theoretical foundation for our style transfer approach. Let $F(I) \in \mathcal{R}^{HW \times C}$ represent the feature map produced by the encoder for an input image I, where H, W, and C denote the height, width, and channel dimensions, respectively. Accordingly, we denote the feature maps for a content image and a style donor image as F_c and F_s. The problem of style transfer can be formulated as learning a linear transformation to produce a stylized feature map, F_d. Let ϕ_s

represent an ideal feature map embodying the target style, and let $\bar{F}$ denote the mean-centered version of a feature map F. The objective is to learn a transformation matrix T that minimizes the discrepancy between the covariance of the transformed features and that of the target style features Li et al. (2017). This can be expressed as minimizing the style loss L_{style}:

$$\begin{aligned} \mathcal{L}_{style} = \quad & \tfrac{1}{NC}\bar{F}_d\bar{F}_d{}^\top - \bar{\phi}_s\bar{\phi}_s{}^\top{}_F^2 \\ s.t. \quad & \bar{F}_d = T\bar{F}_c \end{aligned} \tag{1}$$

By substituting the linear constraint into $\mathcal{L}_{style}$, the minimum is achieved under the condition:

$$T\bar{F}_c\bar{F}_c{}^\top T^\top = \bar{\phi}_s\bar{\phi}_s{}^\top \tag{2}$$

Let the Singular Value Decomposition (SVD) of the centered covariance matrices for the content and target style features be $cov\,(F_c) = \bar{F}_c\bar{F}_c{}^\top = V_cD_cV_c^\top$ and $cov\,(\phi_s) = \bar{\phi}_s\bar{\phi}_s{}^\top = V_sD_sV_s^\top$. Following this decomposition, a set of solutions for T that satisfies Equation 2 can be derived as:

$$T = \left(V_sD_s^{\frac{1}{2}}V_s^\top\right)U\left(V_cD_c^{-\frac{1}{2}}V_c^\top\right) \tag{3}$$

where $U \in \mathcal{R}^{\mathcal{C}\times\mathcal{C}}$ is an orthogonal matrix. This result demonstrates that the transformation matrix T is determined entirely by the covariance statistics of the content and style features. The complete style transfer process involves applying the transformation to the mean-centered content features and then adding back the mean of the style features: $F_d = T\bar{F}_c + mean(F_s)$. This procedure, known as a Whitening and Coloring Transform (WCT) Li et al. (2017), aligns both the mean and covariance of the content features with those of the style donor.

In theory, one could directly implement this style transfer by performing SVD on the feature covariances, as prescribed by the original WCT method Li et al. (2017). However, our objective differs from conventional style transfer like WCT. We aim to transfer a binary "forensic style"—distinguishing between real and AI-generated content—rather than a common artistic one. This requires specific constraints, such as supervising the output with updated ground-truth masks and classifying the resulting style matrix. A direct SVD-based implementation presents significant practical challenges in this context. First, the computational cost of SVD is overly high, leading to prohibitive latency during training. Second, and more critically, the SVD operation has a complex and often unstable gradient flow, which frequently causes training convergence failures, especially when integrated with our specialized forensic loss functions.

Therefore, we opt to learn the transformation matrix T implicitly using a neural network. Since T is conditioned only on the second-order statistics of the content and style features, a feasible strategy is to design a network that takes these features as input and directly predicts the transformation matrix T, bypassing the need for explicit SVD computation.

2.3 Detailed Framework Implementation

Figure 3 illustrates the ForensicWCT framework, which comprises three primary steps: (1) calculating a style transfer matrix T, (2) performing the style transfer using T, and (3) optimizing the process with four specialized loss functions to facilitate the learning of forensic-specific styles.

1. Obtain the Transfer Matrix T. As established in Eq. 3, the transfer matrix T is determined solely by the second-order statistics of the content and style image features. We leverage this property to compute T. Let $F_c \in \mathcal{R}^{H\times W\times C}$ be the feature map extracted by the encoder from the content target image I_c and $F_s \in \mathcal{R}^{H\times W\times C}$ be the corresponding feature map from the style donor image I_s. We first mean-center these features along their spatial dimensions:

$$\bar{F}_c = F_c - mean(F_c), \bar{F}_s = F_s - mean(F_s)$$

These mean-centered features are then processed by two small convolutional networks, CNet and SNet (each comprising three convolutional layers with ReLU activations), to extract compact style representations and reduce channels:

$$F_{c'} = CNet(\bar{F}_c), F_{s'} = SNet(\bar{F}_s)$$

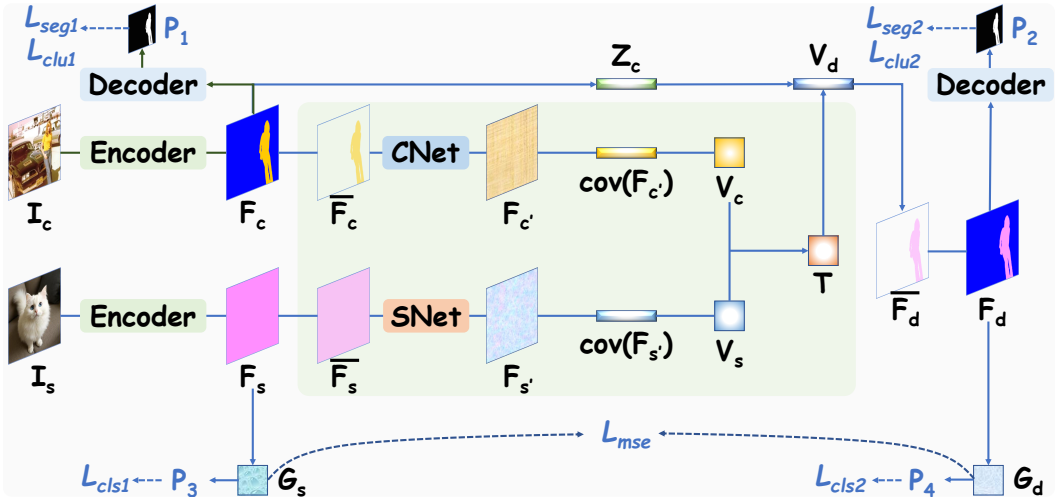


Figure 3: The detailed workflow of the proposed ForensicWCT framework. Green lines (top-left) exist in both training and inference. Blue lines exist only in training.

where $F_{c'}, F_{s'} \in \mathcal{R}^{H\times W\times c}$. Next, for the target region of interest, we flatten the corresponding features and compute their self-covariance:

$$cov(F_{c'}) = flatten(F_{c'})^\top flatten(F_{c'})$$
$$cov(F_{s'}) = flatten(F_{s'})^\top flatten(F_{s'})$$

The resulting covariance matrices of size $c \times c$ are passed through separate linear layers (FC_1 and FC_2) to produce intermediate matrices:

$$V_c = FC_1(cov(F_{c'})), V_s = FC_2(cov(F_{s'}))$$

The final transfer matrix T is computed as their matrix product: $T = V_c V_s$.

2. Perform Style Transfer with T. The style transfer is applied to the original content features F_c. First, the channel dimension of F_c is reduced to c, and the feature map is flattened spatially to produce a one-dimensional feature map $Z_c \in \mathcal{R}^{HW\times c}$. This matrix is then multiplied by the transfer matrix T to yield the stylized features: $V_d = Z_c T$, where $V_d \in \mathcal{R}^{HW\times c}$. Next, V_d is reshaped back to its spatial dimensions and its channel depth is expanded back to C using a 1×1 convolutional layer, resulting in $\bar{F}_d \in \mathcal{R}^{H\times W\times C}$. Finally, the new style is fully incorporated by adding the mean of the style donor features, completing the transformation: $F_d = \bar{F}_d + mean(F_s)$. According to the theoretical analysis in Section 2.2, both the forensic whitening transform (FWT) and forensic coloring transform (FCT) are achieved simultaneously in this step.

3. Optimization. Four distinct loss functions guide the training process, ensuring that the model learns a meaningful style transfer targeted specifically at forensic analysis, rather than an arbitrary mapping or a generic artistic style.

(1) L_{seg}**:** A binary cross-entropy (BCE) segmentation loss, which provides the primary supervision for the forgery localization task. We compute two segmentation losses. The first, L_{seg1}, is calculated between the prediction P_1 (from decoding the original feature map F_c) and the original ground-truth mask. The second, L_{seg2}, is calculated between the prediction P_2 (from decoding the style-transferred feature map F_d) and a correspondingly updated ground-truth mask. The total segmentation loss is $L_{seg} = L_{seg1} + L_{seg2}$.

(2) L_{mse}**:** A mean squared error (MSE) loss that drives the learning of the style transfer mechanism. It ensures that the style of the target region in the transferred features F_d matches the style of the donor features F_s. Following standard practice in style transfer Li et al. (2017; 2019), we use the Gram matrix as a style descriptor. The loss is thus defined as $L_{mse} = MSE(G_s, G_d)$, where G_s and G_d are the Gram matrices of F_s and the target region of F_d, respectively. This loss optimizes the

auxiliary networks to perform the style transformation correctly. Notably, minimizing the distance between Gram matrices inherently suppresses the transfer of semantic information. This is because Gram matrices are semantic-agnostic by nature; any feature map shares the same Gram matrix as a spatially shuffled version of itself (with broken semantic structure).

(3) L_{cls}: A binary cross-entropy (BCE) classification loss applied to the style representations. We feed the Gram matrices G_s and G_d into a linear classifier to predict whether their corresponding style is real or AI-generated. The losses, L_{cls1} and L_{cls2}, are computed from these predictions. This loss, $L_{cls} = L_{cls1} + L_{cls2}$, serves a dual purpose: it enforces a clear distinction between real and fake forensic styles and ensures that the transferred style G_d is classified in the same category as G_s.

(4) L_{clu}: An L2 clustering loss applied to the decoder's input features, F_c and F_d:

$$\mathcal{L}_{clu} = \frac{1}{|\mathcal{P}_{\text{in}}|} \sum_{(p,q)\in\mathcal{P}_{\text{in}}} \|F_p - F_q\|_2^2 - \frac{1}{|\mathcal{P}_{\text{out}}|} \sum_{(p,q)\in\mathcal{P}_{\text{out}}} \|F_p - F_q\|_2^2$$

where p and q are spatial indices on the input feature map $F \in \mathbb{R}^{H\times W\times C}$, and $F_p \in \mathbb{R}^C$ represents the feature vector at spatial position p. $M \in \{0,1\}^{H\times W}$ is the binary ground-truth mask. $\mathcal{P}_{\text{in}} = \{(p,q) \mid M_p = M_q\}$ and $\mathcal{P}_{\text{out}} = \{(p,q) \mid M_p \neq M_q\}$ correspond to the sets of intra-class and inter-class pixel pairs. $\|\cdot\|_2$ denotes the Euclidean norm, and $|\cdot|$ denotes the cardinality of a set.

This loss encourages feature compactness by minimizing the intra-class variance for each forensic style (real vs. AI-generated). It effectively pulls features of the same style closer together in the feature space. The loss is computed on both F_c and F_d, yielding $L_{clu} = L_{clu1} + L_{clu2}$.

The total training loss L_{FWCT} is a weighted sum:

$$L_{FWCT} = L_{seg} + L_{mse} + L_{cls} + \frac{L_{clu}}{16}$$

The ForensicWCT (FWCT) is active only in the training phase. During inference, the FWCT is detached, and the base model operates via its standard pipeline (i.e., $I_c \rightarrow Encoder \rightarrow F_c \rightarrow Decoder \rightarrow P_1$, as shown in the top-left of Figure 3), incurring no additional computation.

Rationale for the Loss Design. The synergy between these loss functions is critical for learning the desired binary "forensic style". While L_{mse} ensures that a style transfer occurs by matching Gram matrices, it is the combination of L_{seg} and L_{cls} that constrains this transfer to be forensically meaningful. L_{seg2} forces the decoder to produce a correct segmentation mask from the style-transferred features F_d. If the model were to transfer a generic artistic style, the resulting features would not contain the discriminative information needed for the decoder to distinguish forged from real regions, leading to a high L_{seg2}. Similarly, L_{cls} penalizes the model if the original style or transferred style cannot be correctly classified as either real or AI-generated. To jointly minimize these losses, the model is compelled to learn and transfer only the essential, underlying artifact features that characterize the two forensic categories. Finally, L_{clu} supports this process by promoting a more separable feature space, making the segmentation tasks more effective.

Prediction with a Lightweight Decoder. A potential challenge arises from applying style transfer exclusively to the target region, which could introduce boundary artifacts in the feature map F_d. A decoder with a larger receptive field might erroneously learn to detect these boundaries rather than the intrinsic forensic style. To mitigate this, we employ a lightweight MLP decoder with a receptive field of 1. This design ensures that each output prediction is based solely on the features at its corresponding spatial location, forcing the decoder to rely on the transferred style itself. This architectural choice is well-supported in the literature, with similar lightweight decoders being successfully used in numerous segmentation and forgery localization models Chen et al. (2021); Kwon et al. (2022); Guillaro et al. (2023); Xie et al. (2021) without compromising performance. Some models' decoders originally have a receptive field of 3, and we modify it to 1. This does not lead to obvious performance degradation since the encoder already provides a sufficient global receptive field. In models featuring a feature pyramid neck, such as ConvNeXt-UPerHead Liu et al. (2022b), we consider the neck to be an extension of the encoder.

Distinction from WCT: While inspired by the WCT Li et al. (2017), our framework represents a fundamental departure in its objective and mechanism. **First, Style Domain.** Whereas WCT transfers conventional visual styles like color and texture, we define a binary forensic style that isolates the imperceptible traces of manipulation. **Second, Supervision**. WCT is an unguided process,

whereas ForensicWCT is a supervised training regularizer that uses dynamic label updates to force the model to learn forensic artifacts, not just swap arbitrary features. **Third, Objective**. WCT's goal is creative stylization; our goal is to mitigate semantic bias and bridge the data-task gap—objectives the original WCT is neither designed for nor capable of achieving.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Benchmarks. We conduct a comprehensive evaluation of our ForensicWCT framework across six diverse and challenging AI-generated forgery localization benchmarks: OpenSDI Wang et al. (2025b), SID-Set Huang et al. (2025b), So-Fake-Set Huang et al. (2025c), So-Fake-OOD Huang et al. (2025c), AutoSplice Jia et al. (2023), and GIT10k Hao et al. (2025). These benchmarks collectively span a wide spectrum of forgery types, including those generated by GANs, Diffusion Models, and Large Multimodal Models (e.g., Qwen-Image-Edit Wu et al. (2025)). They also feature varied image sources and cross-domain evaluation settings. For fair comparison, we adhere to the official training and testing splits defined for each benchmark. Detailed statistics for each dataset are provided in the Appendix. Following common practice, we report pixel-level F1-score and IoU.

Baselines. We compare our method against a comprehensive suite of state-of-the-art forgery localization models. This includes methods based on multi-scale features Chen et al. (2021); Liu et al. (2022a), frequency-domain guidance Kwon et al. (2022); Guillaro et al. (2023), large multimodal models Wang et al. (2025b); Huang et al. (2025b;c), and reconstruction-based models Hao et al. (2025). All results for competing models are taken directly from the official papers of their respective benchmarks, ensuring that they represent officially fine-tuned and reported performance.

Implementation Details. We train our models for up to 160,000 iterations for each dataset with an input resolution of 512 and a batch size of 32. We use the AdamW Loshchilov & Hutter (2017) optimizer with a learning rate initialized at 1e-4 and linearly decayed to 1e-6. Our default base model is the ConvNeXt-Base-UPerHead architecture Liu et al. (2022b) with its official ADE-20k Zhou et al. (2017) pretrained weights for initialization. When incorporating fully-generated images into the training process, we sample them from the CommunityForensics dataset Park & Owens (2025), which contains approximately 2.7 million images from nearly 5,000 distinct generators. Corresponding authentic images are randomly sampled from large-scale datasets such as COCO2017 Lin et al. (2014), LAION Schuhmann et al. (2022), and SA-1B Kirillov et al. (2023). During training, we employ the same data augmentations from previous works Ma et al. (2024); Wang et al. (2025b), including random flips and compression. In each iteration, we sample a batch of images along with a corresponding batch for style swapping. One of the four swapping operations is then chosen uniformly at random and applied to the batch.

3.2 STATE-OF-THE-ART COMPARISON

As shown in Tables 1, 2, and 3, our model consistently and significantly outperforms state-of-the-art methods across all evaluated benchmarks. For instance, on the OpenSDID dataset Wang et al. (2025b), our method achieves a 7.7-point higher average IoU than MaskCLIP Wang et al. (2025b) (CVPR'25), despite having a substantially smaller and faster model (122M vs. 535M parameters). This superior performance is attributable to our novel design. By effectively decoupling image semantics from forensic artifacts, ForensicWCT mitigates the semantic bias that plagues existing models. Furthermore, our style-swapping paradigm enables the model to effectively leverage diverse, fully-generated images, fostering a more robust, generator-agnostic predictive capability. The superior generalization of our model is further evidenced by the qualitative comparisons in Figure 4. More experiments, including image-level evaluation, cross-dataset evaluation, further results and analyses are provided in the Supplement.

3.3 ABLATION STUDY

Ablation on Swapping Operations. The left portion of Table 4 presents an ablation study on ForensicWCT swapping operations detailed in Figure 12. **(a)**: The baseline is a standard ConvNeXt-UPerHead model trained without ForensicWCT. **(b)**: In this setting, we augment training with fully-

Table 1: Pixel-level (localization) performance on OpenSDID dataset Wang et al. (2025b)

Methods	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
MVSS-Net Chen et al. (2021)	57.8	65.3	44.9	51.7	14.6	18.5	26.9	32.7	4.7	6.3	29.8	34.9
CAT-Net Kwon et al. (2022)	66.3	74.8	54.5	62.3	25.5	30.7	35.5	42.0	4.9	6.5	37.3	43.3
PSCC-Net Liu et al. (2022a)	54.7	64.2	36.6	44.7	19.7	26.0	29.2	37.2	8.1	11.5	29.7	36.7
ObjectFormer Wang et al. (2022)	51.1	65.6	41.4	47.3	7.4	9.8	9.4	12.5	5.2	7.3	24.1	27.3
TruFor Guillaro et al. (2023)	63.4	71.0	54.6	61.8	26.5	31.8	32.2	38.5	7.6	9.7	36.9	42.5
DeCLIP Smeu et al. (2025)	37.1	43.4	35.6	41.8	14.5	18.2	27.3	33.4	11.2	14.2	25.2	30.2
IML-ViT Ma et al. (2024)	66.5	73.6	44.7	50.6	21.4	25.9	23.6	28.3	6.1	7.9	32.5	37.3
MaskCLIP Wang et al. (2025b)	67.1	75.6	55.5	62.8	30.9	37.0	43.7	51.2	16.2	20.3	42.7	49.4
Ours	**75.8**	**82.8**	**63.6**	**71.6**	**44.3**	**52.6**	**52.5**	**61.8**	**16.3**	**22.0**	**50.4**	**58.2**

Table 2: Pixel-level performance on SID-Set Huang et al. (2025b) and So-Fake Huang et al. (2025c).

Methods	SID-Set		Methods	So-Fake-Set		So-Fake-OOD	
	IoU	F1		IoU	F1	IoU	F1
MVSS Chen et al. (2021)	23.7	31.6	PSCC Liu et al. (2022a)	46.3	54.8	41.1	48.7
HIFI Guo et al. (2023)	21.1	45.9	TruFor Guillaro et al. (2023)	47.5	57.6	32.3	41.1
PSCC Liu et al. (2022a)	35.7	71.3	Qwen2.5-VL Wang et al. (2024)	42.7	50.1	42.2	49.9
LISA Lai et al. (2024)	32.5	69.1	SIDA Huang et al. (2025b)	44.1	58.9	40.1	49.3
SIDA Huang et al. (2025b)	43.8	73.9	So-Fake Huang et al. (2025c)	48.6	63.9	47.8	59.1
Ours	**80.9**	**88.1**	Ours	**51.5**	**66.4**	**49.6**	**61.0**

Table 3: Pixel-level F1 (localization) performance on AutoSplice Jia et al. (2023) and GIT10k Hao et al. (2025) datasets, where our model is compared with MVSS-Net Chen et al. (2021), MFI-Net Ren et al. (2023), TA-Net Shi et al. (2023), CFL-Net Niloy et al. (2023), EMF-Net Ren et al. (2024) and DcDsDiff Hao et al. (2025).

Benchmark	MVSS-Net	MFI-Net	TA-Net	CFL-Net	EMF-Net	DcDsDiff	Ours
AutoSplice Jia et al. (2023)	91.3	92.2	93.4	91.7	92.9	94.9	**95.2**
GIT10k Hao et al. (2025)	51.7	72.7	65.5	55.0	81.2	86.5	**95.0**

Table 4: Ablation study. Pixel-level average IoU (localization) performance on OpenSDI Wang et al. (2025b) dataset. "J.T." (Joint Training) denotes directly training with the same extra fully-generated images and real images while setting their ground-truth mask to all-white or all-black. Operations I, II, III, IV are those in Figure 12. SegFormer-B5 and ConvNeXt-Base are trained with their official UPerHead Xiao et al. (2018) decoder.

Swapping operations			Loss functions			Base models			
Ablation		IoU	Ablation		IoU	Ablation (w.o. or w/ ours)		IoU	
								w.o.	w/
(a)	None	38.2	(h)	L_{seg} (baseline)	39.2				
(b)	J.T.	38.8	(i)	L_{seg}+L_{cls}	45.0	(o)	MVSS-Net Chen et al. (2021)	29.8	32.5
(c)	I	48.7	(j)	L_{seg}+L_{mse}	47.1	(p)	CAT-Net Kwon et al. (2022)	37.3	42.2
(d)	II	41.5	(k)	L_{seg}+L_{cls}+L_{mse}	49.2	(q)	IML-ViT Ma et al. (2024)	32.5	43.7
(e)	I+II	49.2	(l)	L_{seg}+L_{cls}+L_{clu}	46.6	(r)	TruFor Guillaro et al. (2023)	36.9	45.1
(f)	III+IV	43.8	(m)	L_{seg}+L_{mse}+L_{clu}	48.8	(s)	SegFormer Xie et al. (2021)	39.0	48.6
(g)	All (Ours)	**50.4**	(n)	All (Ours)	**50.4**	(t)	ConvNeXt Liu et al. (2022b)	38.2	**50.4**

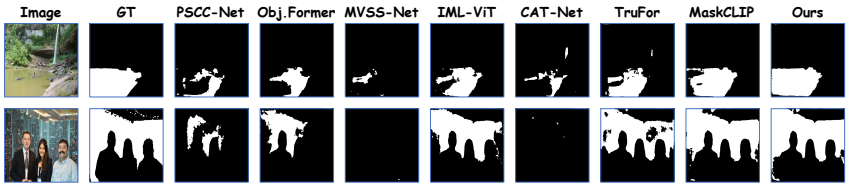


Figure 4: Qualitative comparison on forgery localization.

generated and real images, labeling them with all-black or all-white masks, respectively. The performance gain over (a) is negligible, highlighting the task gap between image-level classification and pixel-level localization, which prevents effective knowledge transfer. **(c)**: By incorporating style-

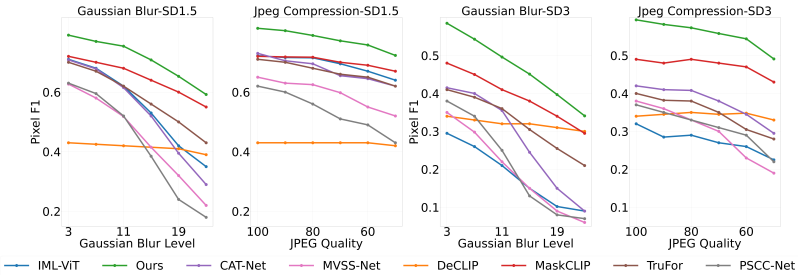


Figure 5: Robustness evaluation on OpenSDID (pixel-F1).

swapping for the fully-generated data, this setting outperforms (b) by a remarkable 9.9 points in IoU. This demonstrates that our method successfully bridges the task gap, enabling the model to leverage diverse, image-level data. **(d)**: Applying style-swapping to partially-edited images yields a 3.3-point IoU improvement over (a), confirming that this operation effectively reduces semantic bias and promotes a more generator-agnostic decoder. **(e)**: The combination of operations I and II is better than each separate component, demonstrating their synergy. **(f)**: This setting augments training with large-scale authentic object segmentation data, improving IoU by 5.0 points over (b) and showcasing the strong scalability of our approach. **(g)**: Our full model, which integrates all four swapping operations, achieves the best performance, validating the necessity of each component.

Ablation on Loss Functions. The center of Table 4 details the contribution of each loss function. (h) serves as the loss baseline, where all swapping operations are used but only the standard BCE segmentation loss (L_{seg}) is active. Its performance is comparable to the base model (a), indicating that the auxiliary losses are essential for guiding the model to separate forensic style from content. The classification loss (L_{cls}) trains a classifier on the Gram matrices and significantly improves performance (e.g., (i) vs. (h) and (n) vs. (m)). This loss forces the forensic styles of real and fake regions to become more distinct, which is crucial for accurate localization. The MSE style loss (L_{mse}), computed between the Gram matrices of the style donor and the transformed region, also proves effective (e.g., (j) vs. (h) and (n) vs. (l)). It constrains the auxiliary networks to learn the style transfer transformation correctly, aligning with our theoretical analysis. The clustering loss (L_{clu}), an L2 loss on the decoder's input features, consistently improves performance (e.g., (l) vs. (i), (m) vs. (j), and (n) vs. (k)). By encouraging feature compactness, this loss helps better align the forensic styles of regional and global forgeries, enhancing the leverage of diverse training data.

Ablation on Base Models. The right part of Table 4 demonstrates the versatility of ForensicWCT. When applied to various existing forgery localization and segmentation models, our method yields substantial performance gains (e.g., +8.2 IoU for TruFor, +12.2 IoU for ConvNeXt-Base) **without** introducing any additional parameters or computational overhead during inference. This highlights the strong flexibility and broad applicability of our approach.

Robustness Evaluation in Figure 5 shows that our approach maintains high performance and consistently outperforms SOTA methods under distortions.

4 Conclusion

We have introduced a new paradigm for forgery localization by reformulating the task as image style classification. This conceptual shift directly confronts the critical, dual challenges of semantic bias and scarcity of data that have long hindered progress in the field. Our proposed framework, ForensicWCT, realizes this paradigm through a novel style-swapping approach, enabled by a forensic whitening and coloring transform. This technique disentangles generator-specific artifacts from semantic content, compelling the model to learn content-agnostic representations to mitigate semantic bias. Furthermore, it elegantly bridges the task gap between localization and large-scale image-level datasets, vastly enriching training data diversity. The result is a significant leap in generalization across unseen generators and content, achieved without introducing any computational overhead at inference. Ultimately, ForensicWCT establishes a powerful and efficient approach for training highly generalizable forensic models, offering a critical tool to safeguard digital trust in an era of proliferating generative AI.

REFERENCES

Xiuli Bi, Zhipeng Zhang, and Bin Xiao. Reality transform adversarial generators for image splicing forgery detection and localization. In *proceedings of the IEEE/CVF international conference on computer vision*, pp. 14294–14303, 2021.

Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C Cosman. Detecting generated images by real images only. *arXiv preprint arXiv:2311.00962*, 2023.

Lvpan Cai, Haowei Wang, Jiayi Ji, Yanshu Zhoumen, Shen Chen, Taiping Yao, and Xiaoshuai Sun. Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2534–2542, 2026.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Beijing Chen, Yuting Hong, Ziqiang Li, and Zhangjie Fu. Dfpd: Dual-forgery proactive defense against both deepfakes and traditional image manipulations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11697–11705, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025b.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14185–14193, 2021.

Alex Costanzino, Woody Bayliss, Juil Sock, Marc Gorriz Blanch, Danijela Horak, Ivan Laptev, Philip Torr, and Fabio Pizzati. Towards reliable identification of diffusion-based image manipulations. *arXiv preprint arXiv:2506.05466*, 2025.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217. IEEE, 2025.

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*, pp. 422–426, 2013. doi: 10.1109/ChinaSIP.2013.6625374.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3040–3048, 2025.

Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, 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)*, pp. 63–72. IEEE, 2019.

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*, pp. 20606–20615, 2023.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694. IEEE, 2025.

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*, pp. 3155–3165, 2023.

Suhyeon Ha, Woo Jae Kim, Joonsung Jeon, Sooel Son, and Sung-eui Yoon. Apt: Anchor-aligned perturbations for tamper localization in fully regenerated images. *arXiv preprint arXiv:2608.30656*, 2026.

Qixian Hao, Shaozhang Niu, Jiwei Zhang, and Kai Wang. Dcdsdiff: Dual-conditional and dual-stream diffusion model for generative image tampering localization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1071–1079, 2025.

Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14862–14870, 2021.

Yongkang Hu, Yu Cheng, Yushuo Zhang, Yuan Xie, and Zhaoxia Yin. Saido: Generalizable detection of ai-generated images via scene-aware and importance-guided dynamic optimization in continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3876–3886, 2026.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Yingsong Huang, Hui Guo, Jing Huang, Bing Bai, and Qi Xiong. Diffusion epistemic uncertainty with asymmetric learning for diffusion-generated image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17097–17107. IEEE, 2025a.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025b.

Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, et al. So-fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.18660*, 2025c.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 68549–68570, 2026.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 893–903, 2023.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947. IEEE, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Shashank Kotyan, Makoto Shing, Yuki Imajuku, Rujikorn Charakorn, and Tarin Clanuwat. Prior-conditioned gaussian discriminants for generalizable ai-generated image detection. *arXiv preprint arXiv:2608.18523*, 2026.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

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.

Myung-Joon Kwon, Wonjun Lee, Seung-Hun Nam, Minji Son, and Changick Kim. Safire: Segment any forged image region. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4437–4445, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9579–9589, 2024.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20379–20389. IEEE, 2025.

Songlin Li, Guofeng Yu, Zhiqing Guo, Yunfeng Diao, Dan Ma, and Gaobo Yang. Beyond fully supervised pixel annotations: Scribble-driven weakly-supervised framework for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 677–685, 2026.

Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022a.

Yaqi Liu, Shuhuan Chen, Haichao Shi, Xiao-Yu Zhang, Song Xiao, and Qiang Cai. Mun: image forgery localization based on m^3 encoder and un decoder. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5685–5693, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution ai-generated image detection. In *International Conference on Learning Representations*, volume 2026, pp. 80332–80355, 2026.

Ju-Hyeon Nam, Dong-Hyun Moon, and Sang-Chul Lee. M2sformer: Multi-spectral and multi-scale attention with edge-aware difficulty guidance for image forgery localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15927–15938. IEEE, 2025.

Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S Woo. Cfl-net: Image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 4642–4651, 2023.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257, June 2025.

Rongxuan Peng, Shunquan Tan, Chenqi Kong, Anwei Luo, Alex C Kot, and Jiwu Huang. Forensic-ssam: Toward robust and unified image forgery detection and localization resisting to adversarial attack. *arXiv preprint arXiv:2508.07402*, 2025.

Chenfan Qu, Yiwu Zhong, Xuekang Zhu, Junchi Li, Changjiang Jiang, Lianwen Jin, et al. Detect any ai-counterfeited text image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35437–35450, 2026.

Ruyong Ren, Qixian Hao, Shaozhang Niu, Keyang Xiong, Jiwei Zhang, and Maosen Wang. Mfi-net: Multi-feature fusion identification networks for artificial intelligence manipulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):1266–1280, 2023.

Ruyong Ren, Qixian Hao, Feng Gu, Shaozhang Niu, Jiwei Zhang, and Maosen Wang. Emf-net: An edge-guided multi-feature fusion network for text manipulation detection. *Expert Systems with Applications*, 249:123548, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Xiangfei Sheng, Weidong Zou, Tianjiao Gu, Zhichao Yang, Pengfei Chen, and Leida Li. Agidefect-4k: A richly annotated dataset for ai-generated image defect detection, localization and explanation. *arXiv preprint arXiv:2608.20713*, 2026.

Senyuan Shi, Hao Tan, Zichang Tan, Shuhan Feng, Ajian Liu, Sergio Escalera, and Jun Wan. Hydraprompt: An adaptive and asymmetric framework of vision-language models for synthetic image detection. *arXiv preprint arXiv:2605.26421*, 2026.

Zenan Shi, Haipeng Chen, and Dong Zhang. Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4907–4920, 2023.

Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22424–22433, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in ai-generated images. In *International Conference on Learning Representations*, volume 2026, pp. 111907–111942, 2026a.

Hao Tan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 66420–66477, 2026b.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Hanyi Wang, Jun Lan, Yaoyu Kang, Huijia Zhu, Weiqiang Wang, Zhuosheng Zhang, and Shilin Wang. Generalizable and adaptive continual learning framework for ai-generated image detection. *IEEE Transactions on Multimedia*, 2026a.

Jiaan Wang, Sirui Liu, Yu Li, Kaiyuan Yang, Juan Cao, and Sheng Tang. Fleet: Few shots lead effective ai-generated image detection. *arXiv preprint arXiv:2606.31082*, 2026b.

Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373, 2022.

Mengxi Wang, Shaozhang Niu, and Jiwei Zhang. Forgdiffuser: General image forgery localization with diffusion models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1954–1962, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Opensdi: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4291–4301, 2025b.

Youqi Wang, Shen Chen, Haowei Wang, Rongxuan Peng, Taiping Yao, Shunquan Tan, Changsheng Chen, Bin Li, and Shouhong Ding. Forgeryvcr: Visual-centric reasoning via efficient forensic tools in mllms for image forgery detection and localization. *arXiv preprint arXiv:2602.14098*, 2026c.

Bihan 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)*, pp. 161–165, 2016. doi: 10.1109/ICIP.2016.7532339.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13440–13449, 2022.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Enze Xie, Wenhai 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.

Zhipei Xu, Xuanyu Zhang, Youmin Xu, Qing Huang, Shen Chen, Taiping Yao, Shouhong Ding, and Jian Zhang. Genshield: Unified detection and artifact correction for ai-generated images. *arXiv preprint arXiv:2605.16122*, 2026.

Haozhen Yan, Ruoxin Chen, Jiahui Zhan, Bo Wang, Youchang Xiao, Shouhong Ding, Liqing Zhang, Taiping Yao, and Jianfu Zhang. Lorc: Detecting ai-generated images via low-rank collapse in semantic residuals. In *European Conference on Computer Vision*, pp. 207–224. Springer, 2026.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, et al. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, volume 2026, pp. 121742–121761, 2026.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 70440–70522, 2025.

Tianyi Zhang, Qinglong Lin, Yang Hu, Pengming Feng, and Rubo Zhang. Edge-aware affinity enhancement for image manipulation localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 324–332, 2025a.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. D 3 qe: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16292–16301. IEEE, 2025b.

Yanran Zhang, Wenzhao Zheng, Yifei Li, Bingyao Yu, Yu Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. Unigendet: A unified generative-discriminative framework for co-evolutionary image generation and generated image detection. *arXiv preprint arXiv:2604.21904*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. Pgc: Peak-guided calibration for generalizable ai-generated image detection. *arXiv preprint arXiv:2605.21207*, 2026.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18746–18758. IEEE, 2025.

A Appendix

A.1 Related Works

A.1.1 AI-Edited Image Region Localization

Some works focus on localizing fake image regions created by traditional splicing operation Ha et al. (2026); Liu et al. (2025); Peng et al. (2025); Zhang et al. (2025a); Chen et al. (2025a); Li et al. (2026); Nam et al. (2025); Kwon et al. (2025). However, AI-editing leaves much less edge artifacts compared to splicing, making them less effective for this task. The localization of AI-generated edits is hindered by two fundamental challenges: semantic bias, where models incorrectly learn to associate edits with object appearance, and a lack of generator diversity in pixel-level training data. Prior methods have struggled to overcome these hurdles. Early approaches using noise-domain modeling Chen et al. (2021); Guillaro et al. (2023) require complex feature alignment and incur significant computational overhead. While powerful, more recent solutions leveraging large pre-trained models Wang et al. (2025b); Costanzino et al. (2025); Huang et al. (2025b); Hao et al. (2025) are computationally expensive and still fail to resolve the underlying generator and semantic biases. A core unresolved issue is the inability to leverage large-scale, fully-synthetic datasets like CommunityForensics Park & Owens (2025). The reason is a fundamental task mismatch: image-level classification relies on global feature aggregation, whereas localization demands fine-grained, patch-to-patch comparison. This disparity not only renders joint training ineffective but also causes naive style augmentations Bi et al. (2021); Hong et al. (2021) to corrupt authentic regions with confusing artifacts. In response, we propose a paradigm shift: reframing forgery localization as a style classification problem. This approach elegantly sidesteps these longstanding issues, simultaneously addressing both semantic bias and data scarcity.

A.1.2 Image Style Transfer

Image style transfer seeks to transfer the style from a reference image to a content image by aligning their deep feature statistics. Seminal works progressed from matching first-order statistics (mean, variance) with AdaIN Huang & Belongie (2017) to aligning second-order statistics (covariance) with the Whitening and Coloring Transform Li et al. (2017) for more expressive results. While effective, its reliance on Singular Value Decomposition (SVD) is computationally intensive; this was later mitigated by methods like LST Li et al. (2019) that learn an efficient approximation of the transformation. **Unlike prior research on generic visual styles**, we introduce and target a binary **forensic style** that distinguishes between authentic and AI-generated content. We achieve the extraction of forensic style with novel losses.

A.1.3 AI-Generated Image Detection

Several works detect AI-generated image in image-level Sheng et al. (2026); Li et al. (2025); Zhou et al. (2025) PCGD Kotyan et al. (2026) utilizes closed-form heads built from first- and second-order statistics of frozen encoder features, serve as strong analytical baselines for AIGI detection transfer. Fleet Wang et al. (2026b) replaces static generalization with continuous few-shot adaptation, using constrained avoidance routing to redirect novel AI samples away from real-dominated routes in decoupled subspaces. AIDE Yan et al. (2025) combines CLIP semantic embeddings with high- and low-frequency patch-wise low-level features. DRCT Chen et al. (2024) generates hard negatives through high-quality diffusion reconstruction. PGC Zhou et al. (2026) aggregates the most discriminative local cues through a peak-focusing mechanism. LoRC Yan et al. (2026) decouples semantic dominance to capture the low-rank collapse arising in the semantic-residual orthogonal subspace. B-Free Guillaro et al. (2025) generates fakes from real images via the stable-diffusion conditioning procedure so that the only difference stems from generative artifacts, removing content-level dataset bias. DDA Chen et al. (2025b) aligns real and synthetic images in both the pixel and frequency domains to eliminate spurious high-frequency correlations. NFA-ViT Cai et al. (2026) mines heterogeneous regions via noise fingerprints. UniGenDet Zhang et al. (2026) unifies generation and detection in a co-evolutionary loop. HiDA-Net Mu et al. (2026) fuses features from multiple full-resolution tiles with a down-sampled global view. Recently, some works introduce solutions that based on large multimodal models Wen et al. (2026); Ji et al. (2026); Tan et al. (2026a); Zhou et al. (2025); Kang et al. (2025); Tan et al. (2026b); Xu et al. (2026); Wang et al. (2026c); Ye et al. (2025). Despite the effectiveness, the detection is constrained in image-level Shi et al. (2026); Yang et al.

(2026); Yan et al. (2024); Cui et al. (2025); Fu et al. (2025); Zhang et al. (2025b); Qu et al. (2026); Hu et al. (2026); Wang et al. (2026a); Huang et al. (2025a).

B Detailed Dataset Statistics

Tables 5 and 6 present detailed statistics for the datasets used in our experiments. We adopt the official train/test split for each dataset. Pixel-level performance evaluation is conducted on the edited data. During training, we introduce additional image-level generated data from CommunityForensics Park & Owens (2025) and ensure its generator types do not overlap with those in the test set.

Table 5: Dataset statistics of the OpenSDID dataset Wang et al. (2025b).

Model	Training Set			Test Set			Total
	Real	Edited	Fake	Real	Edited	Fake	Images
SD1.5	100K	50K	50K	10K	5K	5K	220K
SD2.1	-	-	-	10K	5K	5K	20K
SDXL	-	-	-	10K	5K	5K	20K
SD3	-	-	-	10K	5K	5K	20K
Flux.1	-	-	-	10K	5K	5K	20K
Total	100K	50K	50K	50K	25K	25K	300K

B.1 Additional Experiments

B.1.1 Experiments on Image-level Classification

The results of our image-level forgery classification experiments are presented in Table 7. The separate scores are presented in Table 8. Although our method is developed to leverage extra image-level generated data for improving regional forgery localization, its introduction clearly enhances the model's performance on the image-level classification task as well, leading to state-of-the-art performance on representative benchmarks. The robustness evaluation in Figure 6 confirms our model's robustness against distortions.

B.1.2 Cross Dataset Evaluation

Table 9 shows the performance where models are trained on one dataset and evaluated on another dataset. The superior performance of our model validates that it learns more generalized features under the proposed ForensicWCT framework.

B.1.3 Semantic Suppression

Our paper focuses on localizing AI-edited regions, which mostly do not have obvious boundary artifacts. Consequently, the advantage of reducing the bias from appearance or boundary notably overweights the drawback. OpenSDID's training-set has 547 generated cups and 52 real ones, 45 generated hats and 934 real ones (based on its meta-info). This authenticity-category imbalance drives model to identify cup as generated and hat as real. We evaluate cosine similarity $cos(\cdot)$ between different models' features on its test-set: R_{cup}, R_{hat}: features of real cup and hat, G_{cup}, G_{hat}: features of generated cup and hat. Higher $\cos(R_{cup}, G_{cup})$, $\cos(R_{hat}, G_{hat})$ or lower $\cos(R_{cup}, R_{hat})$, $\cos(G_{hat}, G_{cup})$ indicate that model is biased to semantics. The results in Table 10 confirm the reduced semantic correlation of our model.

B.1.4 Experiments on non-AIGC Benchmarks

We evaluated the base models on traditional forgery localization benchmarks, using their official implementations and augmenting them with our approach. As shown in Table 11, our method—primarily designed for AI-generated forgeries—does not degrade models' ability to localize traditional forgeries. It improves performance on samples with texture anomalies.

B.2 Semantic- and Shape-Agnostic Nature

The forensic style we propose is defined by the feature covariance matrix, a representation that is inherently both semantic- and shape-agnostic. The semantic-agnostic property arises because the

Table 6: Dataset statistics of the SID-Set Huang et al. (2025b), So-Fake-Set Huang et al. (2025c), So-Fake-OOD Huang et al. (2025c), AutoSplice Jia et al. (2023) and GIT10K Hao et al. (2025).

Dataset	Training Set			Test Set			Total
	Real	Edited	Fake	Real	Edited	Fake	Images
SID-Set	76819	76732	76740	20000	20000	20000	290291
So-Fake-Set	517262	509292	752804	129316	127324	188200	2224198
So-Fake-OOD	-	-	-	33207	34046	32469	99722
AutoSplice	1364	2173	-	909	1448	-	5894
GIT10K	-	9000	-	-	1000	-	10000

Table 7: Image-level classification performance. 'ConvNeXt*' denotes ConvNeXt-Base Liu et al. (2022b) trained with our ForensicWCT framework. The compared methods include UnivFD Ojha et al. (2023), RINE Koutlis & Papadopoulos (2024), IML-ViT Ma et al. (2024), MaskCLIP Wang et al. (2025b), LNP Bi et al. (2023), LGrad Tan et al. (2023), SIDA Huang et al. (2025b), Qwen2.5-VL Wang et al. (2024) (7B version), So-Fake-R1 Huang et al. (2025c), IF-OSN Wu et al. (2022), MVSS-Net Chen et al. (2021), CAT-Net Kwon et al. (2022), TruFor Guillaro et al. (2023).

OpenSDID Wang et al. (2025b)						
Method	UnivFD	RINE	IML-ViT	MaskCLIP	ConvNeXt	ConvNeXt*
F1	72.2	76.0	53.7	77.8	67.8	**83.4**
SID-Set Huang et al. (2025b)						
Method	UnivFD	LNP	LGrad	SIDA	ConvNeXt	ConvNeXt*
F1	85.3	58.2	55.5	93.5	81.6	**95.2**
So-Fake-Set Huang et al. (2025c)						
Method	UnivFD	Qwen2.5-VL	SIDA	So-Fake-R1	ConvNeXt	ConvNeXt*
F1	63.8	90.8	91.5	92.9	83.5	**95.0**
So-Fake-OOD Huang et al. (2025c)						
Method	UnivFD	Qwen2.5-VL	SIDA	So-Fake-R1	ConvNeXt	ConvNeXt*
F1	40.2	72.5	72.2	75.3	54.4	**78.9**
NIST16 Guan et al. (2019)						
Method	IF-OSN	MVSS-Net	CAT-Net	TruFor	ConvNeXt	ConvNeXt*
Acc	55.3	53.8	59.7	66.2	64.9	**66.5**

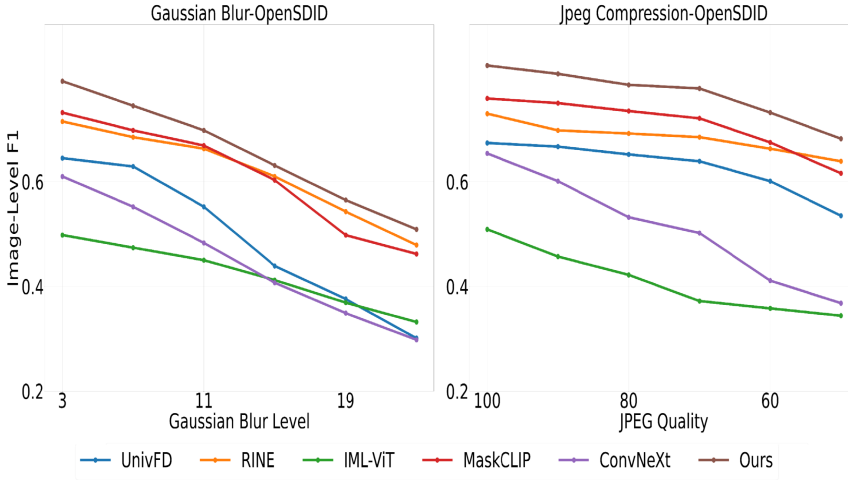


Figure 6: Robustness evaluation on image-level forgery classification.

covariance calculation is invariant to the spatial arrangement of features. As illustrated in Figure 9, spatially shuffling a feature map disrupts its semantic content, yet the resulting covariance matrix

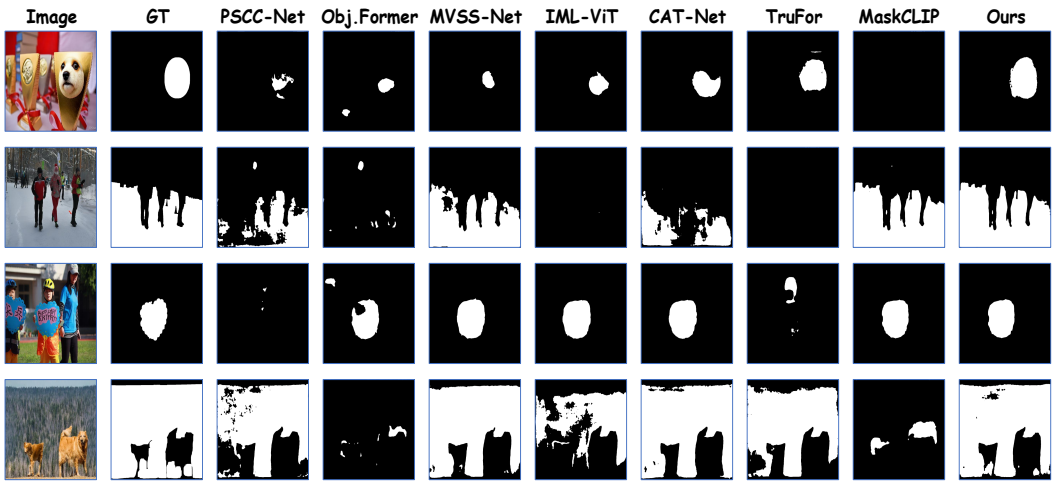


Figure 7: Qualitative comparison with state-of-the-art models on forgery localization.

remains unchanged. This demonstrates that our style representation is independent of the high-level semantic information encoded in the image. Furthermore, the style is shape-agnostic. Before covariance is computed, a 2-dimensional feature map from an image region of any shape is flattened into a 1-dimensional feature vector. The dimension of the resulting covariance matrix depends only on the vector's channel depth, not its spatial dimensions (i.e., length). This key property allows the forensic style to be flexibly extracted from and transferred between image regions of arbitrary shapes, sizes.

B.3 Additional Visualizations

Prediction visualization. The qualitative comparison in Figure 7 shows that our model excels in detecting forgeries from unseen generators, arbitrary objects and shapes. Figure 8 provides a qualitative ablation study, with settings corresponding to Table 4 of the main paper. The results visually validate the contribution of each design component in our ForensicWCT framework.

Feature visualization. In the left panel of Figure 11, the Gram matrices of the style and transferred features differ only negligibly, indicating successful style transfer. In the right panel, our method yields a more unified and broader feature distribution between the edited region and the fully-generated image, demonstrating improved exploitation of diverse image-classification-task data.

B.4 Limitations

The ForensicWCT framework is designed for common forgery localization architectures that use a trainable encoder and a lightweight decoder. However, it is not directly applicable to some models with frozen backbones Wang et al. (2025b), pure-diffusion structures Wang et al. (2025a), or decoders based on large language models Huang et al. (2025b). In addition, although the ForensicWCT framework does not add any additional computational burden in inference, it almost doubles the training time (51h→97h in 8*A100) and training memory (27GB→44GB) requirement. Failure Cases in Figure 10 show that there is still room for improvement in detecting tiny forgery regions, poor image conditions and abrupt texture change in real images.

Method	UnivFD	RINE	IML-ViT	MaskCLIP	ConvNeXt	ConvNeXt+Ours
Fake	73.5	76.7	51.7	78.9	65.3	**82.9**
Real	70.6	75.3	55.6	76.8	70.3	**84.0**

Table 8: Separated F1 scores for real and fake samples on OpenSDID (detailed version of Table 4).

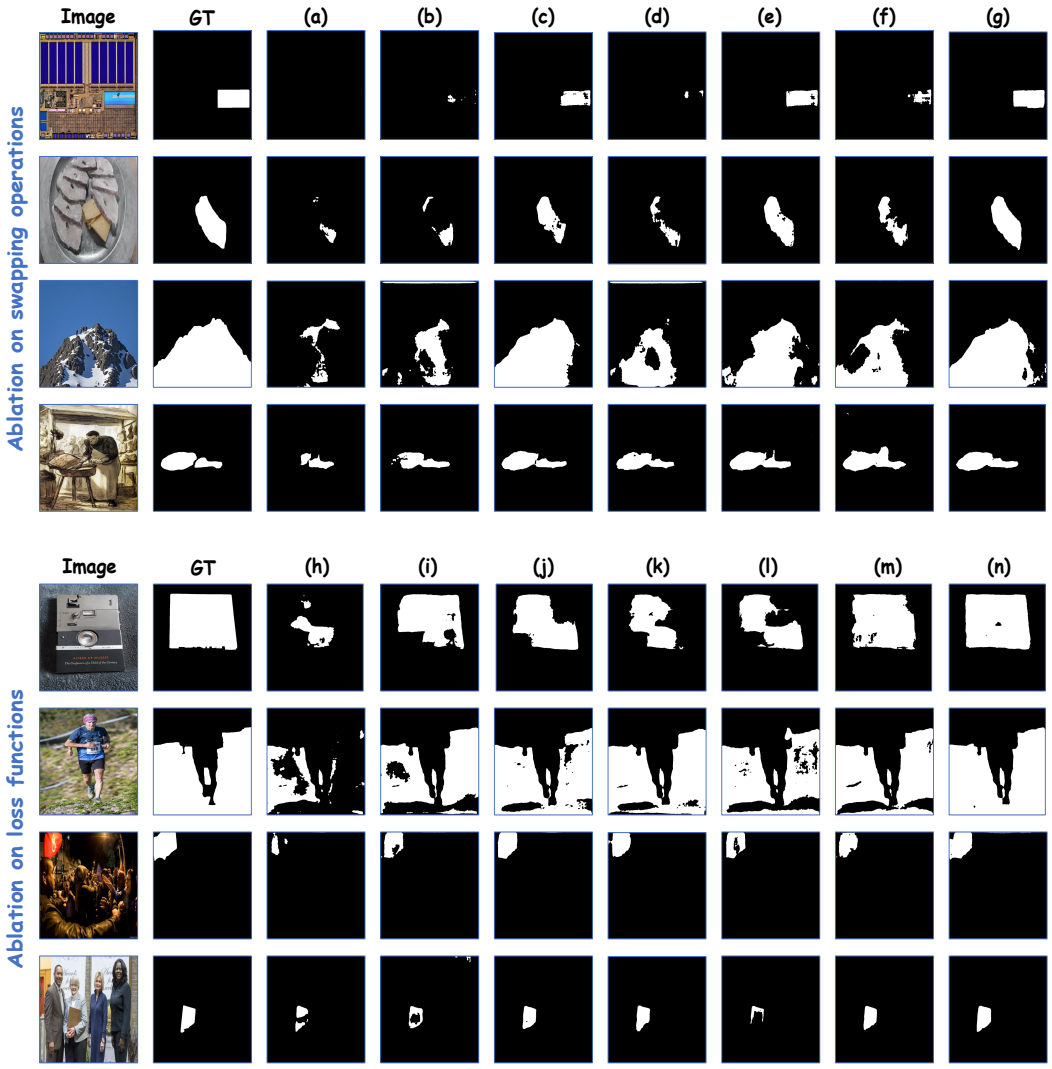


Figure 8: Qualitative comparison for ForensicWCT ablation. The ablation setting index is from Table 4 of the main paper.

Train: So-Fake Test: OpenSDID					
Method	UnivFD	PSCC-Net	TruFor	MaskCLIP	Ours
Pixel-F1	-	23.3	28.4	36.1	**39.0**
Image-F1	65.2	57.5	63.9	72.2	**73.8**
Train: SID-Set Test: So-Fake-OOD					
Method	UnivFD	PSCC-Net	TruFor	SIDA-7B	Ours
Pixel-F1	-	25.4	25.6	29.5	**33.6**
Image-F1	47.7	52.4	57.1	59.8	**64.9**

Table 9: Cross-dataset evaluation with both pixel-level and image-level F1.

Model	$cos(R_{cup}, R_{hat}) \uparrow$	$cos(G_{cup}, G_{hat}) \uparrow$	$cos(R_{cup}, G_{cup}) \downarrow$	$cos(G_{hat}, R_{hat}) \downarrow$
MVSS-Net	0.593	0.708	0.619	0.531
TruFor	0.662	0.787	0.454	0.368
IML-ViT	0.541	0.752	0.553	0.457
MaskCLIP	0.710	0.826	0.414	0.409
ConvNeXt	0.512	0.728	0.491	0.516
Ours	**0.794**	**0.855**	**0.330**	**0.279**

Table 10: Evaluating semantic bias via cosine similarity.

Dataset	CAT-Net		IML-ViT		TruFor		ConvNeXt	
	w.o. ours	w/ ours	w.o. ours	w/ ours	w.o. ours	w/ ours	w.o. ours	w/ ours
CASIAv1	75.2	75.4	72.1	71.5	73.7	73.2	74.4	74.0
COVER.	38.1	36.8	41.0	40.8	60.0	58.9	45.3	44.7
NIST16	30.8	32.0	33.1	34.6	39.9	42.3	40.5	42.0

Table 11: Pixel-level localization performance with F1-score metric on CASIAv1 Dong et al. (2013), COVER. Wen et al. (2016) and NIST16 Guan et al. (2019) datasets. 'w.o.' denotes 'without' and 'w/' denotes 'with'. Compared models include: CAT-Net Kwon et al. (2022), IML-ViT Ma et al. (2024), TruFor Guillaro et al. (2023), ConvNeXt Liu et al. (2022b)

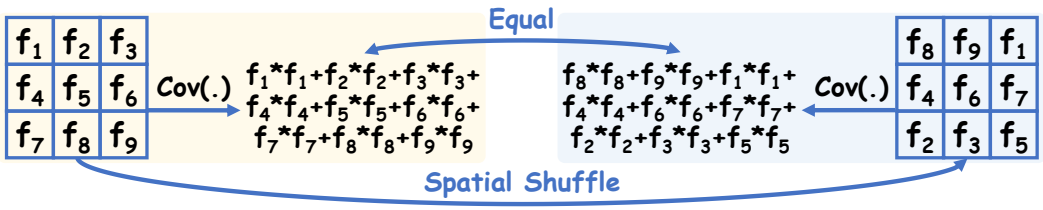


Figure 9: The forensic style, calculated as the feature covariance, is invariant to spatial shuffling (breaks semantic). A feature map (left) and its spatially shuffled version (right) produce an identical result, demonstrating the style's semantic-agnostic nature.

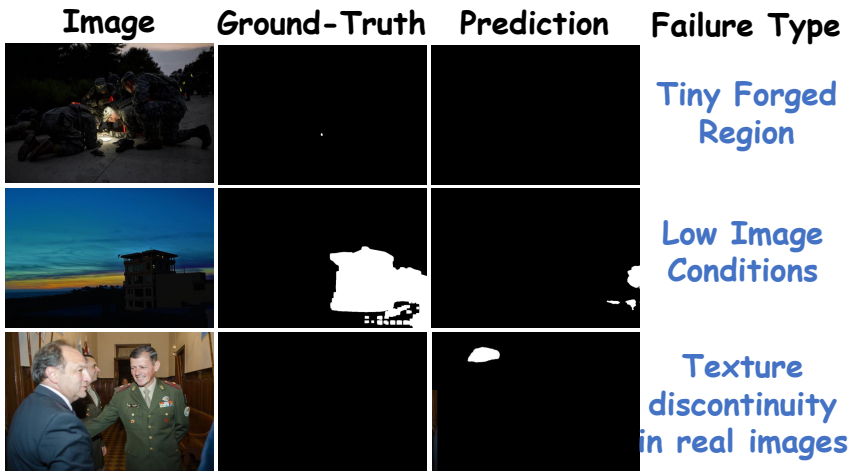


Figure 10: Failure cases. There is still room for improvement in detecting tiny forgery regions, poor image conditions and abrupt texture change in real images.

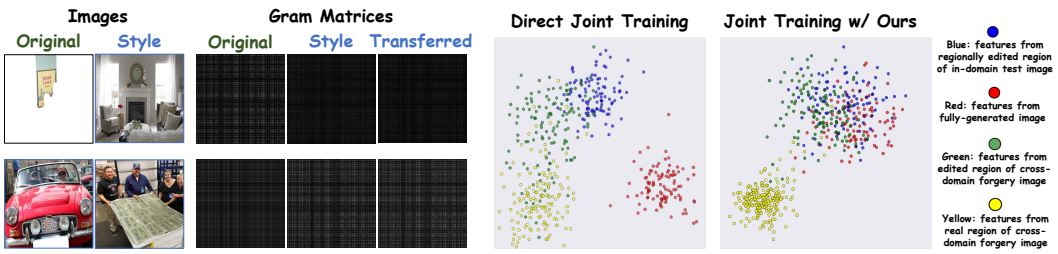


Figure 11: (Left) Images for forensic style transfer and their corresponding feature Gram matrices. (Right) A t-SNE Van der Maaten & Hinton (2008) feature visualization.

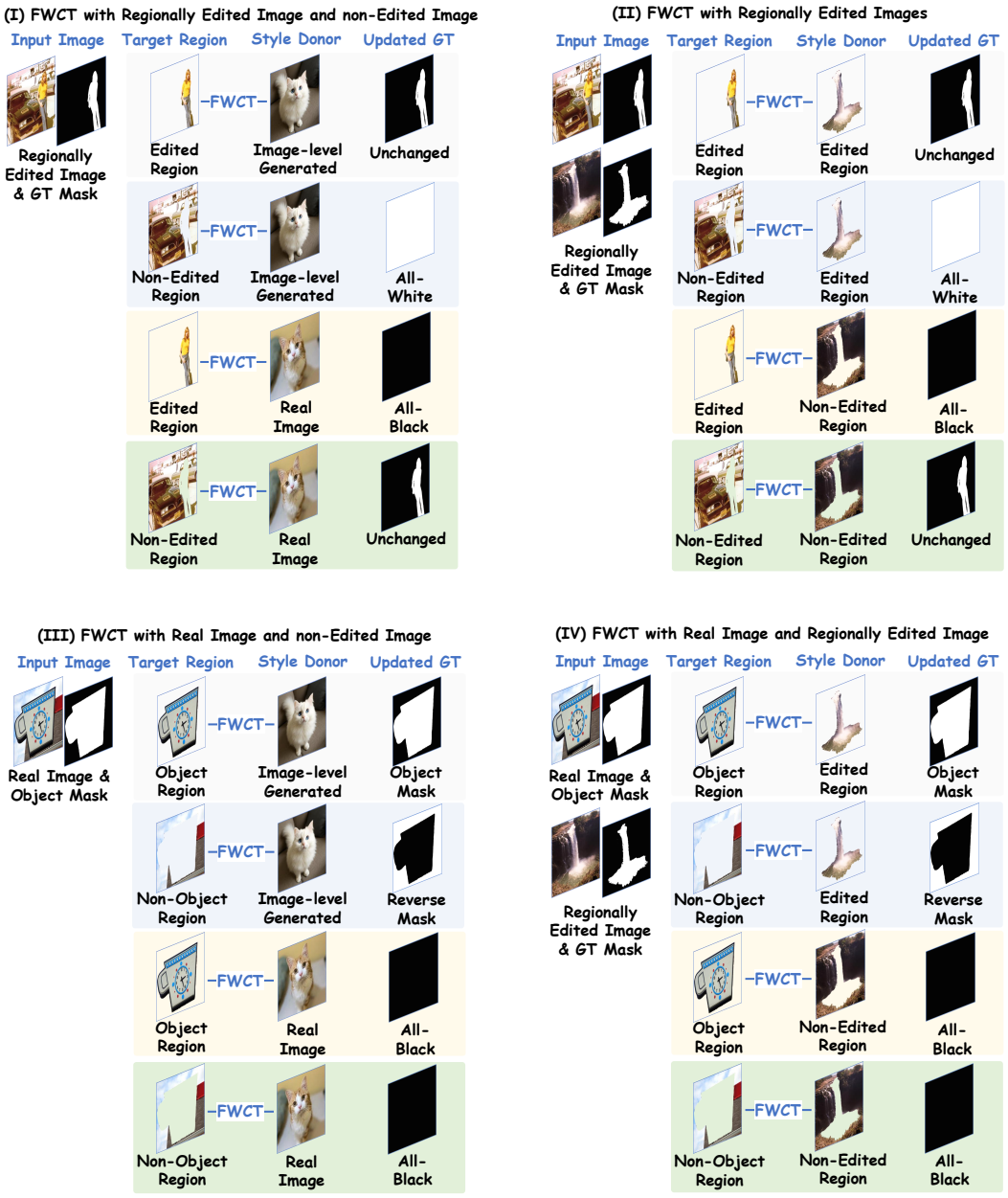


Figure 12: ForensicWCT (FWCT) swapping operations and the corresponding ground-truth (GT) update rules. We combine the features from real or AI-generated regions with forensic styles from other images, and dynamically update the ground-truth mask to reflect the newly applied style.