StyleSentinel: Reliable Artistic Copyright Verification via Stylistic Fingerprints

Lingxiao Chen[1], **Liqin Wang**[1], **Wei Lu**[1*]

[1]MoE Key Laboratory of Information Technology, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

chenlx67@mail2.sysu.edu.cn, wanglq37@mail2.sysu.edu.cn, luwei3@mail.sysu.edu.cn

Abstract

The versatility of diffusion models in generating customized images has led to unauthorized usage of personal artwork, which poses a significant threat to the intellectual property of artists. Existing approaches relying on embedding additional information, such as perturbations, watermarks, and backdoors, suffer from limited defensive capabilities and fail to protect artwork published online. In this paper, we propose StyleSentinel, an approach for copyright protection of artwork by verifying an inherent stylistic fingerprint in the artist's artwork. Specifically, we employ a semantic self-reconstruction process to enhance stylistic expressiveness within the artwork, which establishes a dense and style-consistent manifold foundation for feature learning. Subsequently, we adaptively fuse multi-layer image features to encode abstract artistic style into a compact stylistic fingerprint. Finally, we model the target artist's style as a minimal enclosing hypersphere boundary in the feature space, transforming complex copyright verification into a robust one-class learning task. Extensive experiments demonstrate that compared with the state-of-the-art, StyleSentinel achieves superior performance on the one-sample verification task. We also demonstrate the effectiveness through online platforms.

Introduction

The versatility of diffusion models such as DALL-E (Ramesh et al. 2021), Midjourney, Kandinsky (Razzhigaev et al. 2023), and Stable Diffusion (Rombach et al. 2022) represents a revolutionary advance in generative AI. They can transform text into detailed and stylistically diverse images. However, simple prompts fail to meet specific requirements and full model retraining remains prohibitively expensive. Fine-tuning techniques like LoRA (Hu et al. 2021), Dreambooth (Ruiz et al. 2023), and Textual Inversion (Gal et al. 2022) democratize AI art by enabling large foundation models to learn specific concepts, objects, or styles from small image sets, significantly lowering technical barriers.

This powerful capability for visual understanding and style transfer allows users to create personalized artworks. However, it also creates a convenient way for unauthorized usage of artworks. Malicious attackers can easily scrape unauthorized artworks from public sources and rapidly fine-tune a model on consumer-grade GPUs to mimic the styles. It constitutes plagiarism of the artwork, devaluing the original art while directly threatening intellectual property rights and economic interests (Moayeri et al. 2024).

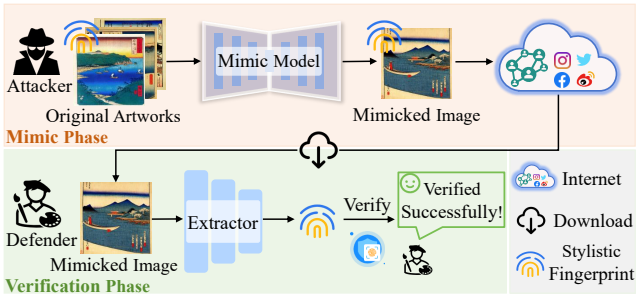

Figure 1: Illustration of our approach for copyright verification. The stylistic fingerprint in the artwork is preserved during the mimic phase, and it can be extracted as robust evidence for copyright verification.

Current research on artwork copyright protection can be primarily categorized into three types of methods. Perturbation-based methods aim to disrupt the output of malicious models (Chen et al. 2024; Shan et al. 2023; Van Le et al. 2023; Zhao et al. 2024b; Wang et al. 2025), typically achieved by injecting imperceptible perturbations into the images. These perturbations are designed to corrupt the training process, causing distorted and stylistically incongruous output. Watermark-based methods (Cui et al. 2023; Luo et al. 2023; Ma et al. 2023; Zhu et al. 2023) embed invisible watermarks into images. Models trained on such images will produce outputs with these watermarks, enabling copyright verification by detecting them. Backdoor-based methods (Wang et al. 2023a; Chou, Chen, and Ho 2023) introduce backdoors into datasets, which allow infringement detection by testing whether a model exhibits specific backdoor behaviors triggered by pre-defined inputs.

However, both perturbation-based, watermark-based, and backdoor-based methods are subject to critical limitations. First, they rely on embedding extrinsic signals into images, but the instability of the signals results in limited defensive capabilities against potential attacks. Second, these pre-

*Corresponding authors

processing dependent methods cannot retroactively protect artworks already circulating online without prior safeguards.

To address these, as illustrated in Figure 1, we propose StyleSentinel to verify the copyright by extracting a shared stylistic fingerprint from an artist's artworks. Unlike existing approaches that embed invisible signals into images, our solution neither compromises visual quality nor requires preprocessing. Our work is motivated by two key observations. First, the images generated by the diffusion models exhibit significant visual and stylistic similarities with the training data (Wang et al. 2023b; Somepalli et al. 2023), suggesting that style functions as a persistent fingerprint preserved during training. Second, the distinct styles possessed by most artists differentiate their artworks from others (Moayeri et al. 2024). This inherent distinctiveness makes the extraction of such stylistic fingerprints feasible.

Specifically, considering the limited number of artworks available for a single artist, we employ a semantics self-reconstruction process to enhance stylistic expressiveness within the artworks. This process leverages semantics and style of an image to synthesize new augmented samples, which establishes a style-consistent and data-rich manifold foundation for subsequent feature learning. Subsequently, due to the multi-level nature of artistic style, we design a multi-layer attention style extractor to generate a compact stylistic fingerprint from each image. This module extracts features from multiple layers of a backbone network and adaptively fuses them into a style feature vector, encoding a discriminative fingerprint for subsequent verification. Finally, we use a hypersphere-based verifier to model stylistic boundaries, enabling reliable verification of the extracted stylistic fingerprint. This component learns a minimal enclosing boundary for all style embeddings of the target artist, transforming copyright verification into a more robust one-class learning task. During verification, copyright attribution is determined simply by checking whether a suspect image's style vector falls within the prelearned hypersphere boundary. In summary, our contributions are as follows:

- We propose StyleSentinel for artistic copyright protection by verifying inherent artistic fingerprints from artworks. It requires no image preprocessing and enables verification for unprotected images published online.
- We introduce a semantic self-reconstruction process to strengthen the data foundation for stylistic feature learning. We further design a multi-layer attention style extractor with a hypersphere-based verifier to encode stylistic fingerprint and achieve reliable verification.
- Extensive experiments demonstrate that StyleSentinel achieves superior performance against state-of-the-art baselines in the one-sample verification task. The effectiveness is further validated on two real-world platforms.

Related Works

Style Mimicry with Diffusion Models

Large-scale pre-trained diffusion models, trained on massive image-text pair datasets like LAION (Schuhmann et al. 2022), demonstrate remarkable capabilities in generating high-quality and diverse images. However, the prohibitive cost of retraining these foundation models has limited the proliferation of their use for stylistic mimicry.

To enable low-cost personalized generation, the research community has developed efficient fine-tuning techniques, which have led to the rampant problem of style mimicry. Among these, DreamBooth and LoRA deliver the most potent style replication. DreamBooth, a representative high-fidelity method, achieves sophisticated style learning with minimal images (typically 3-5) by exposing the model to a unique identifier (e.g., a rare token [V]) and fine-tuning the entire U-Net. It further mitigates language drift and overfitting through class-specific prior preservation loss. In contrast to DreamBooth's extensive weight adjustments, LoRA freezes all pretrained parameters and operates via parallel low-rank injections into key modules. With trainable parameters representing merely 0.01%-0.1% of the original model, LoRA dramatically reduces computational and storage demands. While the lightweight and efficient nature of these techniques has profoundly democratized AI art, it paradoxically enables unauthorized and low-cost artistic style replication, posing significant threats to creative rights.

Dataset Copyright Protection

Current dataset copyright protection for generative models relies on embedding external information. Perturbation-based methods introduce adversarial noise, impair the semantic learning of the model and cause corrupted outputs. However, they are vulnerable to defenses like adversarial purification (Cao et al. 2023), where preprocessing can nullify the effect. Watermark-based methods embed covert signals designed to persist through training and appear in outputs, while backdoor-based methods implant triggers for detection. Crucially, both watermark-based and backdoor-based methods face the challenge of ensuring signal robustness against image transformations and the training process (Zhao et al. 2024a) while preserving host image quality.

Some approaches explore a distinct path requiring no image modification. However, approaches like membership inference (Chen et al. 2020; Shokri et al. 2017) are often impractical, as they require massive sample generation for statistical analysis. ArtistAuditor (Du et al. 2025) leverage style representation for copyright verification, but it fails to perform adequately in one-sample verification scenarios.

Problem Statement

Threat Model

Attacker's Goal and Capability. Attackers aim to generate new images that closely mimic or replicate the artist's distinctive style for low-cost appropriation and commercial exploitation. Attackers can easily aggregate an artist's publicly available artworks from online portfolios, galleries, and social media platforms to construct a dataset. They potentially preprocess the images with methods like secondary fine-tuning, prompt attacks, or data augmentation. With this dataset, attackers can fine-tune mimic models on consumer-grade GPUs. In most cases, attackers solely publish the

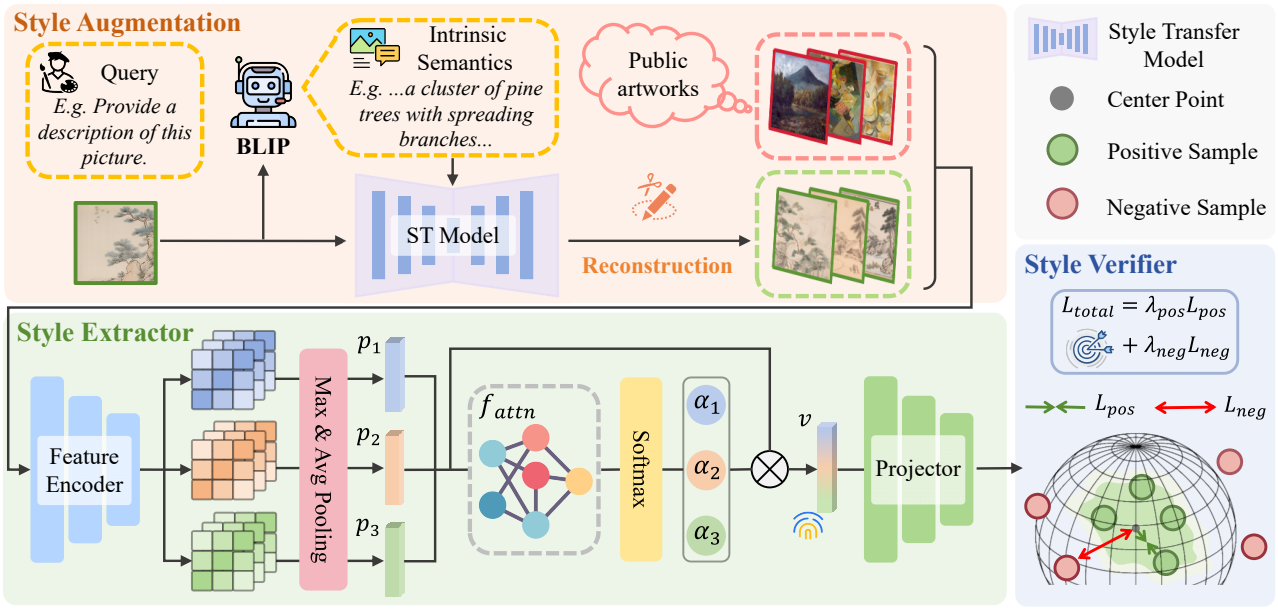

Figure 2: Overview of the proposed StyleSentinel. The pipeline comprises three main stages. (1) Style Augmentation: A semantic self-reconstruction process generates style-consistent augmented data from the original works. (2) Style Extraction: A multi-layer attention style extractor encodes each artwork into a compact style fingerprint. (3) Style Verifier: A hypersphere-based verifier performs one-class classification with the style fingerprint.

generated images while deliberately concealing the mimic model and its implementation details.

Defender's Goal and Capability. The defender's primary objective is to reliably verify the copyright of a suspect artwork based on minimal evidence, ideally a single instance. The verification is performed under a reasonable and challenging assumption: The defender acquires the suspect image accidentally, without any additional information about the mimic model or the text prompt used.

Design Challenges

To achieve reliable verification with stylistic fingerprints for copyright protection, we face two challenges. The first challenge is to accurately represent artistic style. An abstract style is composed of various elements, prompting each artist to develop a unique focus. For example, Claude Monet focuses on the overall atmosphere created by light and shadow, whereas Vincent van Gogh emphasizes local expression through brushwork and color. We must extract a robust fingerprint sensitive to these hierarchical distinctions.

The second challenge is the inherent data sparsity. An individual artist's portfolio is typically a limited corpus, often comprising several tens to hundreds of works. Consequently, our model must learn a generalizable stylistic signature from the limited dataset, avoiding overfitting to the specific content of the training images.

Methods

Overview

We aim to extract a robust stylistic fingerprint for reliable verification from the artworks. As illustrated in Figure 2, our pipeline comprises three sequential stages. Initially, we introduce a novel module of style augmentation. This module employs a semantic self-reconstruction process to synthesize augmented samples that are not only highly consistent in style with the original artwork but also rich in detailed variations. Subsequently, a multi-layer attention style extractor is used to extract hierarchical features from the images and adaptively fuse them into a single feature vector that holistically represents the style of the image. Finally, the extracted style vector is passed to a hypersphere-based style verifier, which performs verification by determining whether the style features of an unknown sample fall within the decision boundary defined by a hypersphere.

Semantic Self-Reconstruction Style Augmentation

Recalling design challenges, to address the inherent scarcity of data, an effective approach is data augmentation. However, traditional methods such as rotation, compression, and color jittering are insufficient because they operate at a superficial geometric or pixel level randomly. They introduce limited variations and probably disrupt the model's internal representation of artistic style.

To introduce diverse augmented samples without compromising the integrity of the original artistic style, we propose a novel augmentation approach motivated by the insight that artistic style is intrinsically linked to its semantics. An artist's style emerges from their sustained depiction of specific subjects, compositional choices, and atmospheric rendering. For example, Claude Monet's impressionist style of rapid and loose brushstrokes is a result of his semantic focus on capturing the fleeting effects of light on landscapes

Figure 3: Some original images (a) and reconstructed images (b) with their intrinsic semantics.

and water. Preserving this inherent style-semantic coupling is essential for creating meaningful augmentations. Based on this insight, we defined our augmentation approach as the artistic reinterpretation of an image's intrinsic semantics using its own stylistic representation. This can generate high-quality augmented samples that preserve strict style consistency while varying contextual details. Specifically, for an image X_{ori} to be protected, we perform the following two steps to achieve data augmentation.

Intrinsic Semantic Extraction. Prior work (Wang et al. 2024) has demonstrated that, in contrast to the ambiguous nature of stylistic attributes, semantics can often be effectively represented by natural language text. Building upon this insight, we propose using the textual description of an image as its intrinsic semantics. To achieve this, we use BLIP model (Li et al. 2022) to generate a textual description T_{ori} of the image X_{ori} serving as its intrinsic semantic information.

Self-Reconstruction. With intrinsic semantics, we then perform a reconstruction process to generate augmented images. Specifically, we use X_{ori} as the style reference and T_{ori} as the content guide, both of which are fed into a style transfer model (Wang et al. 2024) to synthesize the augmented images. This step leverages the inherent stochasticity of the generative model to introduce valuable generative variance, while strictly preserving the style-semantic coupling (Figure 3). Consequently, we create augmented images that are both stylistically faithful and diverse. They establish a dense and style-consistent manifold foundation, improving the robustness of subsequent feature learning.

Multi-Layer Attention Style Extractor

As analyzed in design challenges, artistic style is abstract and difficult to quantify. It constitutes a complex and cross-layer visual pattern, manifesting not only in low-level attributes such as textures and brushstrokes but also in mid-level compositional arrangements and object deformations, as well as in the high-level atmosphere. To learn a comprehensive style representation, we design a multi-layer attention style extractor to capture image features across multiple hierarchical levels, and fuse them into a final feature vector that accurately reflects the style. The style extractor is constructed of two fundamental blocks, a multi-layer feature extraction backbone and an attentional fusion module.

Multi-layer Feature Extraction Backbone. Inspired by (Zhang et al. 2022), we fine-tune the VGG-19 (Simonyan and Zisserman 2014) architecture, which is pre-trained on ImageNet (Deng et al. 2009). We select the output feature maps from three specific layers corresponding to low, mid, and high level features. For each selected feature map, we apply parallel average and max pooling operations to extract both global average and peak feature activations. The outputs of these operations are then concatenated and processed by a final convolutional layer to produce the level-specific feature encoding.

Attentional Fusion Module. After obtaining the feature encodings from different hierarchical levels, an effective approach is to fuse them into a single final style representation. Simple averaging or concatenation operations are suboptimal because they assign equal weight to all feature levels, ignoring the reality that their relative importance varies significantly with the specific artistic style. For instance, a style characterized by prominent brushstrokes might rely more heavily on low-level features, whereas a style defined by a unique atmosphere could be more correlated with high-level features (Gatys, Ecker, and Bethge 2015). To this end, we design an attentional fusion module that dynamically learns the weights for different hierarchical feature encodings based on the input.

Specifically, given the N feature encodings $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N$ each with a potentially different dimension, we first unify their dimensionality by projecting each $\boldsymbol{c}_i$ into a common m-dimensional space using a linear layer, yielding a set of projected vectors $\boldsymbol{p}_i$. These vectors are concatenated to form a single tensor $\boldsymbol{P} \in R^{N \times m}$ and passed through a multi-layer perceptron f_{attn} to generate attention scores. Then, these scores are normalized via Softmax function to compute the final attention weights $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ as:

$$\boldsymbol{\alpha} = \text{softmax}(f_{\text{attn}}(\boldsymbol{P})) \tag{1}$$

The final feature vector $\boldsymbol{v}$, as the stylistic fingerprint, is calculated as the weighted sum of the projected feature vectors, where each $\boldsymbol{p}_i$ is modulated by its corresponding attention weight α_i:

$$\boldsymbol{v} = \sum_{i=1}^{N} \alpha_i \boldsymbol{p}_i \tag{2}$$

Hypersphere-based Style Verifier

By far, we have obtained a feature vector $\boldsymbol{v}$ to represent the style. The subsequent step is to verify if this fingerprint belongs to the protected artist. A conventional approach is to construct a binary classifier and train it with a binary cross-entropy loss to learn a separating hyperplane. However, in our specific scenario, the negative class (i.e., non-target artist styles) is inherently diverse and unbounded. It encompasses an infinite spectrum ranging from the styles of all other artists to photographic works and even random images. It is challenging to define a distinct separating boundary for a class that lacks a fixed distribution. This could compel the

binary classifier to learn an excessively complex or poorly generalizing decision boundary.

Therefore, we learn a compact description of the positive class, transforming the conventional binary classification problem into a more suitable one-class problem (Ruff et al. 2018). Specifically, we project the feature vectors obtained in the previous step into a feature space where positive samples are clustered around a center point, forming a minimal-volume hypersphere. In contrast, all negative samples are repelled to the exterior of this hypersphere. Given N samples $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ of feature vectors, the loss function for positive samples is defined as:

$$L_{pos} = \frac{1}{N} \sum_{i=1}^{N} (\sqrt{d_i^2 + 1} - m) \tag{3}$$

where $d_i = \|\phi(\boldsymbol{v}_i) - o\|_2$ is the Euclidean distance, m defines a soft margin, ϕ represents the projection module and o is the learnable center point of the hypersphere. We aim to minimize the distance from the style feature vectors to the center o in the projected space. For negative samples, the loss is formulated as:

$$L_{neg} = -\frac{1}{N} \sum_{i=1}^{N} \log(1 - \exp(-\beta \cdot (\sqrt{d_i^2 + 1} - m)) + \epsilon) \tag{4}$$

where β is a hyperparameter that governs the repulsion intensity, and ϵ is a constant for the numerical stability. The goal is to penalize style vectors that are insufficiently distant from the center, effectively repelling them. The total training objective is defined as:

$$L_{total} = \lambda_{pos} L_{pos} + \lambda_{neg} L_{neg} \tag{5}$$

where λ_{pos} and λ_{neg} are corresponding hyperparameters. The minimization of overall compels the model to sculpt an optimal feature space. The space is organized so that positive samples are enclosed within the hypersphere, while negative samples are expelled to the outside, achieving a fine-grained characterization and separation of the target artistic style.

Training and Inference

Regarding the training dataset, for the negative samples, we recommend selecting from large-scale public artwork datasets. The complex and diverse artistic styles contained within such a dataset benefit the construction of a well-defined decision boundary. For the positive samples, we select the artworks of the target artist. If any of these works are present in the negative dataset, they are excluded from it. We then augment the positive samples by style augmentation based on intrinsic semantics, in conjunction with conventional techniques such as random flipping, JPEG compression, Gaussian noise, and color jittering. Furthermore, we employ a weighted random sampling strategy to address the class imbalance between positive and negative samples.

Upon completion of training, we perform a line search on the validation set to determine the optimal threshold for the radius R. During the inference phase, the center point o and the radius R are utilized to ascertain whether a given sample falls within the hypersphere.

Experiment

Experimental Setting

Datasets and Models. Our experiments are conducted on subsets of two large-scale art datasets: WikiArt (Tan et al. 2018) and ArtBench (Liao et al. 2022). To emulate unauthorized usage of artworks, we fine-tune a set of three surrogate text-to-image models with DreamBooth and LoRA, including Stable Diffusion v1.5 (Rombach et al. 2022), Stable Diffusion v2.1, and Kandinsky (Razzhigaev et al. 2023). Both Dreambooth and LoRA are applied to Stable Diffusion, while Kandinsky is fine-tuned exclusively with LoRA for memory efficiency.

Implementation Details. All experiments are conducted on an NVIDIA RTX 3090 GPU. Before training, we generate 1 to 3 augmented images for each positive sample. Negative samples are sourced from the WikiArt dataset. The verifier is trained with the AdamW optimizer (learning rate 5×10^{-4}) with key hyperparameters: $\lambda_{pos} = 1.0$, $\lambda_{neg} = 1.0$, $\beta = 0.3$, and $m = 1.0$. The test set is balanced, comprising 1,000 mimicked images and 1,000 unrelated style images. All attack models are fine-tuned using default parameters from their original implementations.

Evaluation Metrics. We use two primary metrics: the area under the curve (AUC) and the True Positive Rate at a 10^{-2} False Positive Rate (TPR@FPR= 10^{-2}), where AUC quantifies the overall discriminative ability across all thresholds and TPR@FPR=10^{-2} measures the verification sensitivity under strict constraints.

Baseline. We mainly compare our StyleSentinel with four state-of-the-art methods, including two watermarking-based methods SIREN (Li et al. 2025) and RoSteALS (Bui et al. 2023), a backdoor-based method DIAGNOSIS (Wang et al. 2023a), and a verification method ArtistAuditor (Du et al. 2025). To ensure a fair comparison in a unified way, we convert their native metrics to AUC and TPR@FPR= 10^{-2} inspired by (Li et al. 2025).

Performance Evaluation

As demonstrated in Table 1, our StyleSentinel consistently and significantly outperforms four baselines across all tested settings. On the WikiArt dataset, our approach achieves the highest AUC scores between 0.993 and 0.998 while maintaining the highest on TPR@FPR= 10^{-2} between 0.925 and 0.953. These results demonstrate the remarkable robustness of StyleSentinel, validating the efficacy of using a stylistic fingerprint for copyright verification. In contrast, the baselines either fail to effectively detect the unauthorized data usage or have very fluctuating performance across different settings. The suboptimal baseline ArtistAuditor achieves high AUC scores, but its performance on TPR@FPR= 10^{-2} represents a significant gap of over 25 percent compared to our method. The remaining baselines perform even worse. The backdoor-based method, DIAGNOSIS, shows unstable performance across different fine-tuning techniques, while the watermarking methods, SIREN and RoSteALS, are rendered almost entirely ineffective. Their extremely low AUC scores and TPR values confirm that their embedded signals are too fragile to survive the fine-tuning process and to be

Datasets	Methods	SD1.5-Db	SD1.5-LoRA	SD2.1-Db	SD2.1-LoRA	Kandinsky
		AUC / T@10^{-2}F	AUC / T@10^{-2}F	AUC / T@10^{-2}F	AUC / T@10^{-2}F	AUC / T@10^{-2}F
WikiArt	RoSteALS	0.578 / 0.010	0.589 / 0.013	0.571 / 0.008	0.576 / 0.014	0.583 / 0.012
	DIAGNOSIS	0.901 / 0.256	0.767 / 0.153	0.831 / 0.139	0.840 / 0.231	0.802 / 0.120
	SIREN	0.721 / 0.196	0.580 / 0.016	0.809 / 0.237	0.564 / 0.031	0.699 / 0.067
	ArtistAuditor	0.973 / 0.626	0.970 / 0.601	0.972 / 0.632	0.969 / 0.652	0.973 / 0.681
	Ours	**0.993 / 0.928**	**0.994 / 0.938**	**0.994 / 0.925**	**0.998 / 0.953**	**0.998 / 0.949**
ArtBench	RoSteALS	0.575 / 0.018	0.574 / 0.012	0.571 / 0.015	0.565 / 0.014	0.554 / 0.012
	DIAGNOSIS	0.785 / 0.155	0.743 / 0.138	0.785 / 0.157	0.744 / 0.134	0.864 / 0.193
	SIREN	0.613 / 0.019	0.621 / 0.010	0.712 / 0.129	0.612 / 0.012	0.688 / 0.063
	ArtistAuditor	0.918 / 0.470	0.916 / 0.506	0.921 / 0.576	0.912 / 0.505	0.942 / 0.634
	Ours	**0.990 / 0.901**	**0.993 / 0.922**	**0.992 / 0.882**	**0.989 / 0.862**	**0.995 / 0.930**

Table 1: Performance comparison across five fine-tuning settings on the WikiArt and ArtBench datasets, where Db represents Dreambooth and T@10^{-2}F represents TPR@FPR= 10^{-2}. The best result under each metric is marked with **bold**.

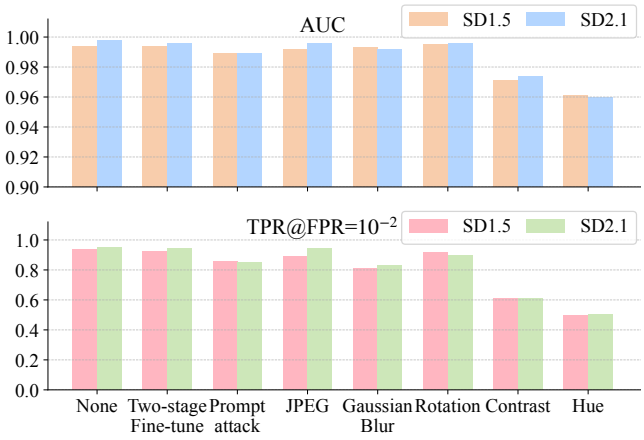

Figure 4: Robustness of StyleSentinel against common image transformations and adaptive attacks. We evaluate the performance on Stable Diffusion v1.5 and v2.1.

detected in one-sample scenarios. These results confirm that intrinsic stylistic fingerprints provide more effective and robust verification evidence than the extrinsic signals used by watermark-based or backdoor-based methods.

Notably, when migrating to the lower-resolution (256×256) ArtBench dataset, the performance of most baseline methods declines markedly due to inferior fine-tuning results. Although our method is also slightly affected, it consistently maintains the highest performance, demonstrating superior robustness to variations in image quality. Furthermore, StyleSentinel exhibits uniform verification capabilities across different fine-tuning settings on both datasets. This resilience stems from the ability of StyleSentinel to extract overall and intrinsic stylistic features, which are robust and transferable.

Robustness Study

We evaluate the robustness of StyleSentinel against two primary categories of adversarial manipulations: common image transformations encountered during online distribution and sophisticated adaptive attacks. The former category comprises five distinct transformations: rotation, JPEG compression (50% quality), Gaussian blurring (3x3 kernel), color jittering (hue factor of 0.2) and contrast adjustment (factor of 2.0). The adaptive attack strategies included two-stage fine-tuning on generated images and black-box prompt attacks employing different textual descriptions.

As shown in Figure 4, StyleSentinel demonstrates remarkable robustness against common transformations and adaptive attacks. It exhibits high resilience to geometric manipulations and compression artifacts. Although alterations in the color space influence style features and cause a decline in TPR@FPR= 10^{-2}, the consistently high AUC demonstrates retained robustness in verification performance. Moreover, the method effectively withstands adaptive strategies such as second-stage fine-tuning and maintains its efficacy against prompt attacks. This resilience against both superficial transformations and adaptive attacks suggests that the learned fingerprint captures the deep feature of an artist's style, rather than just low-level textures.

Configuration	SD1.5		SD2.1	
	AUC	T@10^{-2}F	AUC	T@10^{-2}F
Our Full Method	**0.994**	**0.938**	**0.998**	**0.953**
w/o Aug.	0.975	0.659	0.978	0.707
w/ Trad. Aug.	0.988	0.791	0.986	0.797
Concatenation	0.987	0.803	0.978	0.741
BCE Loss	0.992	0.838	0.989	0.851

Table 2: Ablation study on method components.

Generalization Study

We evaluate the generalization of StyleSentinel when the artworks used to train the verifier differ from those used to fine-tune the suspicious model. Specifically, we use a

Dataset	Completely Disjoint				Partially Overlapping		
	SD1.5-Db		SD2.1-Db		SD1.5-LoRA	SD2.1-LoRA	Kandinsky
	AUC	T@10^{-2}F	AUC	T@10^{-2}F	AUC / T@10^{-2}F	AUC / T@10^{-2}F	AUC / T@10^{-2}F
WikiArt	0.991	0.810	0.990	0.793	0.997 / 0.942	0.995 / 0.938	0.996 / 0.900
ArtBench	0.992	0.846	0.988	0.857	0.994 / 0.885	0.995 / 0.895	0.994 / 0.886

Table 3: Generalization study of StyleSentinel. We test against models fine-tuned with completely disjoint dataset for DreamBooth and partially overlapping dataset for LoRA. The excellent performance shows strong generalization of StyleSentinel.

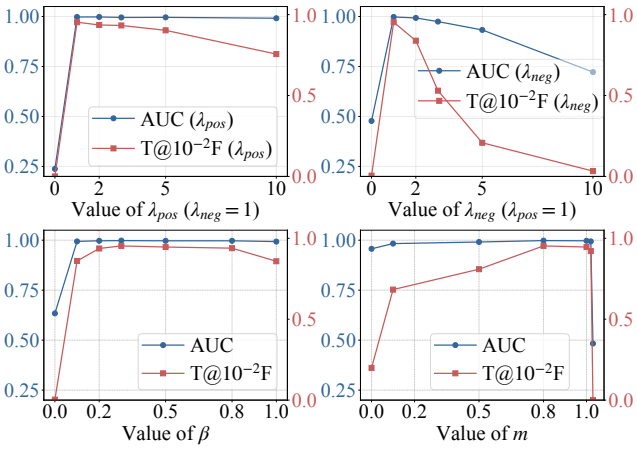

Figure 5: Ablation study on hyperparameters.

partially overlapping dataset for the suspicious model fine-tuned with LoRA, and a completely disjoint dataset for the one fine-tuned with Dreambooth. As demonstrated in Table 3, our method shows remarkable generalization, consistently achieving AUC scores around 0.99 and high TPR@FPR= 10^{-2} ranging from 0.793 to 0.942 in the WikiArt and ArtBench datasets. The consistent performance across all settings confirms that our approach learns a transferable stylistic fingerprint, representing the intrinsic rules of the style.

Ablation Study

Component Ablation. As shown in Table 2, we systematically validate the contribution of each component. First, our semantic self-reconstruction style augmentation demonstrates a significant performance improvement compared to both a no-augmentation baseline and traditional augmentation methods, confirming its effectiveness for subsequent style feature learning. Furthermore, replacing the attentional fusion module with simple feature concatenation leads to a performance decline, which highlights the necessity of adaptively fusing features from different hierarchical levels. Finally, substituting the hypersphere-based verifier with a standard BCE-loss binary classifier results in degraded performance, validating that a one-class learning approach is more suitable for stylistic fingerprint verification.

Hyperparameter Sensitivity. As shown in Figure 5, we reveal several key sensitivities of the hyperparameters. We found that balancing the loss weights at $\lambda_{pos} = \lambda_{neg} = 1.0$ is critical, as any imbalance leads to a performance degradation. Similarly, the repulsion intensity β has an optimal value at 0.3, which best separates the classes without creating an overly rigid boundary. The soft margin m also shows a clear trade-off with performance peaking between 0.8 and 1.0, but excessively high values (>1.0) introduce training instability, resulting in significant performance degradation.

Platform	WikiArt		ArtBench	
	AUC	T@10^{-2}F	AUC	T@10^{-2}F
Shakker	0.998	0.944	0.994	0.933
LibLibAI	0.994	0.928	0.993	0.922

Table 4: Real-world performance of StyleSentinel. We mainly test the effect on two common online platforms.

Real-World Performance

We demonstrate the effectiveness of StyleSentinel in two real-world online fine-tuning service, Shakker[1] and LibLibAI[2]. The two platforms allow users to fine-tune a model with their own uploaded images and provide an API endpoint for generating mimicked images. We conducted tests on WikiArt and ArtBench using the default models and fine-tuning methods offered by this service. As shown in Table 4, our method maintained an AUC of more than 0.99 and a TPR of more than 0.90 in an FPR of 10^{-2}, proving its reliability in practical applications.

Conclusion

In this paper, we introduced StyleSentinel, a novel approach for artistic copyright verification that operates by learning an intrinsic stylistic fingerprint directly from the artworks. It requires no preprocessing and enables protection for artworks already in circulation online. Our approach employs a semantic self-reconstruction process to overcome data sparsity and uses a multi-layer attention style extractor to encode the stylistic fingerprint. Moreover, it formulates the verification task as a robust one-class learning problem with a hypersphere-based style verifier. Extensive experiments demonstrated that StyleSentinel outperforms state-of-the-art baselines in challenging one-sample scenarios. The effectiveness of our method was further validated on two real-world online platforms. We believe StyleSentinel offers a reliable verification for artistic copyright protection.

[1]https://www.shakker.ai
[2]https://www.liblib.art

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62441237, No.62261160653).

References

Bui, T.; Agarwal, S.; Yu, N.; and Collomosse, J. 2023. Rosteals: Robust steganography using autoencoder latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 933–942.

Cao, B.; Li, C.; Wang, T.; Jia, J.; Li, B.; and Chen, J. 2023. IMPRESS: Evaluating the Resilience of Imperceptible Perturbations Against Unauthorized Data Usage in Diffusion-Based Generative AI. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 10657–10677. Curran Associates, Inc.

Chen, D.; Yu, N.; Zhang, Y.; and Fritz, M. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 343–362.

Chen, R.; Jin, H.; Liu, Y.; Chen, J.; Wang, H.; and Sun, L. 2024. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In *Proceedings of the European Conference on Computer Vision*, 126–142. Springer.

Chou, S.-Y.; Chen, P.-Y.; and Ho, T.-Y. 2023. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4015–4024.

Cui, Y.; Ren, J.; Xu, H.; He, P.; Liu, H.; Sun, L.; Xing, Y.; and Tang, J. 2023. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Du, L.; Zhu, Z.; Chen, M.; Su, Z.; Ji, S.; Cheng, P.; Chen, J.; and Zhang, Z. 2025. ArtistAuditor: Auditing Artist Style Pirate in Text-to-Image Generation Models. In *Proceedings of the ACM on Web Conference*, 2500–2513.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Li, B.; Wei, Y.; Fu, Y.; Wang, Z.; Li, Y.; Zhang, J.; Wang, R.; and Zhang, T. 2025. Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2564–2582.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900.

Liao, P.; Li, X.; Liu, X.; and Keutzer, K. 2022. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*.

Luo, G.; Huang, J.; Zhang, M.; Qian, Z.; Li, S.; and Zhang, X. 2023. Steal my artworks for fine-tuning? a watermarking framework for detecting art theft mimicry in text-to-image models. *arXiv preprint arXiv:2311.13619*.

Ma, Y.; Zhao, Z.; He, X.; Li, Z.; Backes, M.; and Zhang, Y. 2023. Generative watermarking against unauthorized subject-driven image synthesis. *arXiv preprint arXiv:2306.07754*.

Moayeri, M.; Basu, S.; Balasubramanian, S.; Kattakinda, P.; Chengini, A.; Brauneis, R.; and Feizi, S. 2024. Rethinking artistic copyright infringements in the era of text-to-image generative models. *arXiv preprint arXiv:2404.08030*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, 8821–8831.

Razzhigaev, A.; Shakhmatov, A.; Maltseva, A.; Arkhipkin, V.; Pavlov, I.; Ryabov, I.; Kuts, A.; Panchenko, A.; Kuznetsov, A.; and Dimitrov, D. 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, 4393–4402.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, 25278–25294.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *Proceedings of the 32nd USENIX Security Symposium*, 2187–2204.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning

models. In *Proceedings of the IEEE Symposium on Security and Privacy*, 3–18.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6048–6058.

Tan, W. R.; Chan, C. S.; Aguirre, H. E.; and Tanaka, K. 2018. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Wang, H.; Spinelli, M.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.

Wang, L.; Hu, Q.; Lu, W.; and Luo, X. 2025. Diffusion-based Adversarial Identity Manipulation for Facial Privacy Protection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11562–11571.

Wang, Z.; Chen, C.; Lyu, L.; Metaxas, D. N.; and Ma, S. 2023a. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. *arXiv preprint arXiv:2307.03108*.

Wang, Z.; Chen, C.; Zeng, Y.; Lyu, L.; and Ma, S. 2023b. Alteration-free and model-agnostic origin attribution of generated images. *arXiv preprint arXiv:2305.18439*.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *Proceedings of the ACM SIGGRAPH Conference Proceedings*, 1–8.

Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.-X.; and Li, L. 2024a. Invisible Image Watermarks Are Provably Removable Using Generative AI. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, 8643–8672.

Zhao, Z.; Duan, J.; Xu, K.; Wang, C.; Zhang, R.; Du, Z.; Guo, Q.; and Hu, X. 2024b. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24398–24407.

Zhu, H.; Liu, M.; Fang, C.; Deng, R.; and Cheng, P. 2023. Detection-performance tradeoff for watermarking in industrial control systems. *IEEE Transactions on Information Forensics and Security*, 18: 2780–2793.