LEARNING SELF-SUPERVISED STYLE REPRESENTATIONS FOR DETECTING AI GENERATED FACES

Tharun Anand, Sanjeev Manivannan, Siva Sankar, Kaushik Mitra
Indian Institute of Technology Madras
{ed20b068, be21b034, ch20b103}@smail.iitm.ac.in, kmitra@ee.iitm.ac.in

ABSTRACT

The proliferation of AI-generated photorealistic faces—from GANs to diffusion models have become indistinguishable from authentic images. This poses significant privacy and security risks, enabling misinformation and identity fraud at scale on social media and other platforms. To detect these AI-Generated faces effectively, we propose a fundamentally new approach inspired by the intrinsic stylistic discrepancies between authentic and synthetic images. Our key insight is that even highly realistic AI-generated faces exhibit persistent differences in style representations, which manifest as distinguishable patterns in the $W+$ Style Space. We introduce a self-supervised style representation learning approach that captures intrinsic differences between actual and synthetic faces. By first learning the style distribution of authentic images, our method identifies deviations indicative of AI generation without relying on explicit generative watermarks. This enables strong generalization across unseen generators, including diffusion-based models. Experiments show high detection accuracy (93%+) across multiple generative datasets and significant improvements in cross-domain settings.

1 INTRODUCTION

AI-generated faces are increasingly pervasive across the internet, appearing in social media as fake profiles or manipulated depictions of real individuals, contributing to misinformation and identity fraud. Detecting such synthetic faces is an escalating challenge as generative models continually improve, reducing the presence of artifacts that traditional detection methods rely on.

Prior approaches have primarily focused on identifying transient watermarks in facial attributes, such as inconsistencies in eyes, teeth, or lighting (Guo et al. (2021); Hu et al. (2020); Mundra et al. (2023); Yang et al. (2018); Zhang et al. (2019)), or its pixel/spectral features (Chai et al. (2020); Corvi et al. (2022); Gragnaniello et al. (2021); Liu et al. (2022)). However, these strategies often target surface-level artifacts that diminish as generators improve . Other methods adopt data-driven strategies (Wang et al. (2019); Tan et al. (2023); Ju et al. (2022); Porcile et al. (2023)) to classify AI-generated faces, which often struggle with generalization to images from out-of-distribution (OOD) generators.

In this work, we introduce a novel approach that leverages the highly disentangled $W+$ Style Space (Karras et al. (2019)) to distinguish authentic and AI-generated faces based on style inconsistencies in facial attributes. Our core insight is that while generators may perfect low-level artifacts, they cannot fully replicate the compositional style patterns inherent to authentic facial images encoded in the $W+$ space's axis-aligned disentanglement. To enhance robustness and generalizability, we employ self-supervised learning (SSL) to pre-train a model that faithfully learns the distribution of real image styles. This enables our method to generalize effectively to OOD scenarios, addressing key limitations of prior works.

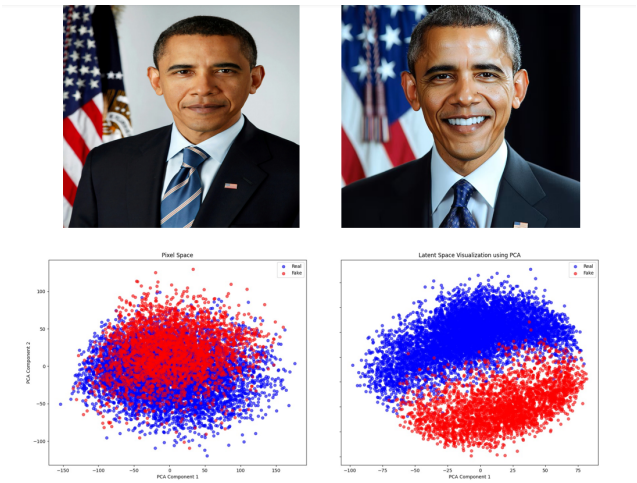

Figure 1: Comparison of AI-generated and real face distributions. (Top) GAN-generated (left) and diffusion-generated (right) faces, illustrating different synthesis methods. (Bottom) PCA visualization of real (blue) vs. synthetic (red) faces in pixel space (left) and StyleGAN's $W+$ space (right), showing improved separability in the style space.

2 PRELIMINARIES

2.1 STYLEGAN AND PSP ENCODER

StyleGANs (Karras et al. (2019; 2018; 2021)) are generative adversarial networks (GAN) that synthesizes high-quality images by leveraging an intermediate latent space, denoted as $W+$. Unlike traditional GAN latent spaces, $W+$ consists of *layer-wise style codes*, allowing precise control over different aspects of an image's appearance. The $W+$ space provides a disentangled and expressive representation, making it particularly useful for encoding fine-grained facial details. The **pSp encoder** (Richardson et al. (2020)) is a pretrained style encoder that maps real images into the $W+$ space, enabling real-image inversion into the latent style space.

2.2 SELF-SUPERVISED LEARNING FOR LATENT REPRESENTATIONS

Self-supervised learning (SSL) has emerged as an effective paradigm for learning robust representations without labeled data. Approaches such as BEiT (Bao et al. (2021)) and MAE (He et al. (2021)) demonstrate the effectiveness of vision transformer based encoder-decoder (Dosovitskiy et al. (2020)) to reconstruct the original image through self-supervised learning, thereby learning powerful representations. Inspired by these methods, we extend the idea of self supervised learning to StyleGAN's latent space rather than raw pixels. By ensuring the encoder learns to structure real-style distributions without supervision, our method captures the stylistic fingerprint of authentic faces, which serves as a strong prior for distinguishing real and synthetic images.

3 MOTIVATION

The $W+$ latent space of StyleGAN is widely used to encode facial attributes—such as pose, age, and texture—along semantically meaningful axes, enabling precise feature modifications without unintended changes to other aspects of the image (Härkönen et al. (2020)). Given that $W+$ is designed to disentangle facial properties, we hypothesize that its structured representation amplifies discrepancies between real and AI-generated faces, making it a strong candidate for detecting synthetic faces.

To validate this, we analyze the separability of real and synthetic images in $W+$ space. We used 50000 real images and 50000 synthetic images generated from both GAN and diffusion generators. We use the PSP encoder (Richardson et al. (2020)) to extract $W+$ style codes, apply PCA, and plot

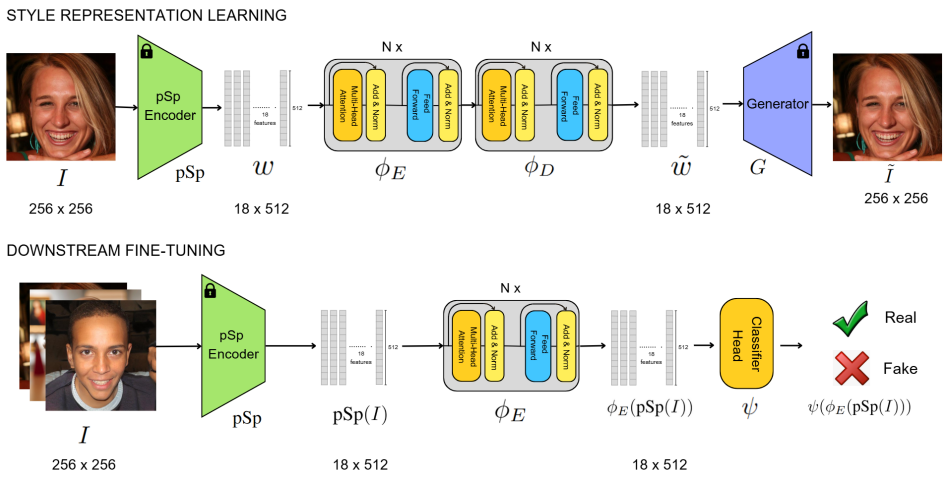

Figure 2: Two-stage framework for detecting AI-generated faces. **Stage 1 (Top):** A transformer-based encoder-decoder learns self-supervised style representations by reconstructing StyleGAN's $\mathbf{W}^+$ latent codes from real faces, optimizing style consistency (L_{style}) and image fidelity (L_{img}). **Stage 2 (Bottom):** A classifier head is fine-tuned on pretrained style embeddings to detect synthetic faces by identifying deviations from authentic style distributions.

the first two principal components. As shown in Figure 1, these PCA projections of style codes form distinct clusters for real and synthetic faces. In contrast, when we perform PCA on pixel-space, the projections exhibit significant overlap. This shows that stylistic deviations introduced by AI-generated content are inherently encoded in $W+$ style representations.

4 METHOD

Our framework follows a two-stage approach: (I) self-supervised pretraining to learn robust style representations from real faces and (II) supervised fine-tuning for real/fake classification. This decoupled learning strategy allows the model to first internalize the stylistic "fingerprint" of authentic images before adapting to discriminative features for synthetic image detection.

4.1 STYLE REPRESENTATION LEARNING

The objective of this stage is to learn a generalizable prior over the style distribution of real faces using only unlabeled data. Given a dataset of real facial images:

$$D_{\text{real}} = \{I_i\}_{i=1}^{N}, \quad I \in \mathbb{R}^{H \times W \times 3} \tag{1}$$

we use the pretrained pSp encoder (Richardson et al. (2020)) to project each image into the W+ style space:

$$w = \text{pSp}(I), \quad w \in \mathbb{R}^{18 \times 512} \tag{2}$$

Here w represents the latent projection, where 18 corresponds to the number of style vectors across different layers, and 512 is the latent space dimension. To learn meaningful style representations from real images, we train a transformer-based encoder-decoder network, ϕ_E and ϕ_D, to reconstruct the latent style codes w. Let $\tilde{w}$ be the reconstructed style codes:

$$\tilde{w} = \phi_D(\phi_E(w)) \tag{3}$$

The decoder output $\tilde{w}$ is then passed through a frozen StyleGAN generator G to synthesize a reconstructed image:

$$\tilde{I} = G(\tilde{w}) \tag{4}$$

The model is trained using the following objectives:

Style Consistency Loss: This loss ensures that the reconstructed style codes $\tilde{w} = \phi_D(\phi_E(w))$ remain close to the original style representation. It is formulated as:

$$\mathcal{L}_{\text{style}} = \|\tilde{w} - w\|_1 \tag{5}$$

Minimizing this loss encourages the encoder to capture fine-grained details of the style space and ensures a faithful reconstruction.

Image Fidelity Loss: To further enforce the preservation of image-specific attributes, we pass the reconstructed style codes through a frozen StyleGAN generator G to obtain the reconstructed image $\tilde{I} = G(\tilde{w})$. The reconstruction quality is enforced using:

$$\mathcal{L}_{\text{img}} = \|\tilde{I} - I\|_1 \tag{6}$$

The total loss function for pretraining is:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{style}} + \mathcal{L}_{\text{img}} \tag{7}$$

To ensure robust learning, both the pSp encoder and the StyleGAN generator remain frozen, allowing the network to focus solely on capturing the structure of authentic styles.

4.2 Downstream Fine-tuning

In this stage, we leverage the pre-trained encoder ϕ_E for real/fake classification. A classifier head ψ is added atop ϕ_E, and the model is fine-tuned on a mixed dataset containing real and fake images:

$$D_{\text{mix}} = \{(I_i, y_i)\}, \quad y_i \in \{\text{real}, \text{fake}\} \tag{8}$$

where fake samples are sourced from both GAN- and diffusion-based generators. The model predicts:

$$\hat{y} = \psi(\phi_E(\text{pSp}(I))) \tag{9}$$

Optimization is performed using cross-entropy loss:

$$L_{\text{CE}} = -\sum y \log \hat{y} \tag{10}$$

By leveraging the pre-trained encoder's ability to encode real image styles, the model effectively detects stylistic inconsistencies in synthetic images.

5 Experiments

5.1 Datasets and Implementation

For pretraining, we use the FFHQ dataset (Matuzevi$\boldsymbol{c}ius$ (2024)), which consists of 70,000 high-quality real face images, enabling the model to learn a robust style prior of authentic human faces. In the fine-tuning stage, we construct a balanced dataset of 3,000 real faces (from CelebA-HQ (Liu et al. (2015)) and 3,000 AI-generated faces. The synthetic images used in fine-tuning are sourced from StyleGAN 1 (Karras et al. (2018)) and Stable Diffusion 1 (Rombach et al. (2021b)), ensuring controlled learning. Also, our model is computationally lightweight and trained using a single NVIDIA 4090 GPU. We trained the model for 800 epochs with a batch size of 8, followed by fine-tuning for 100 epochs using the same batch size, using the Adam optimizer.

To evaluate out-of-distribution (OOD) generalization, we test the model on synthetic images generated by unseen models, which differ from those used during training:

- **GAN-generated test set**: 5,000 synthetic face images from StyleGAN 2 (Karras et al. (2019)), StyleGAN 3 (Karras et al. (2021)), and EG3D (Lan et al. (2023)), which were not used in the training phase.
- **Diffusion-generated test set**: 5,000 synthetic face images from Stable Diffusion 2 (Rombach et al. (2021b)), DALL·E (Rombach et al. (2021a)), Stable Diffusion XL (Podell et al. (2023)), and MidJourney (Borji (2022)), ensuring exposure to unseen synthesis techniques.

This evaluation strategy ensures that the model's performance is assessed on distribution shifts rather than memorization of specific generators used during training.

Table 1: Performance comparison on validation (seen generators) and test (unseen generators) datasets. We report AUC (%) and F1-scores (%).

Method	**Validation (GAN Generators)**		**Test (Diffusion Generators)**	
	AUC% ↑	**F1** ↑	**AUC%** ↑	**F1** ↑
Corvi et al. (2022)	82.81	74.2	66.20	60.5
Mundra et al. (2023)	77.8	70.1	60.28	55.8
Porcile et al. (2023)	85.74	78.5	84.50	79.2
Ours	**95.27**	**93.4**	**93.76**	**91.70**

Table 2: Ablation study on key components of our method. We report AUC (%) and F1-score (%) on test datasets (unseen generators).

Method	**Validation (GAN Generators)**		**Test (Diffusion Generators)**	
	AUC% ↑	**F1** ↑	**AUC%** ↑	**F1** ↑
Baseline (No Self-Supervision)	77.46	72.1	70.4	67.3
Only Self-Supervised Encoder	66.1	61.8	61.3	59.4
Full Model (Ours)	**95.27**	**93.4**	**93.76**	**91.70**

5.2 Results and Benchmark Comparisons

Table 1 compares our method with existing benchmarks on both test datasets. Our approach achieves the highest AUC and F1-score, demonstrating superior performance.

Specifically, our model attains an AUC of 95.2% on the GAN test set and 93.7% on the more challenging diffusion test set, significantly outperforming prior methods while maintaining strong generalization. These results highlight the effectiveness of our style-based representation learning for AI-generated face detection.

5.3 Ablation Study

To understand the impact of each component in our method, we conduct an ablation study by systematically removing different modules. Table 2 presents the performance of our model under different settings on the test dataset (unseen generators).

Baseline (No Self-Supervision). We remove self-supervised learning and train a simple classifier using the StyleGAN encoder's style codes. The performance drops significantly, confirming the importance of self-supervised learning.

Only Self-Supervised Encoder. We exclude the pSp encoder and the StyleGAN decoder during training and use only the self-supervised encoder-decoder trained with image reconstruction. During fine-tuning we use only the trained self-supervised encoder for classification. This configuration lacks the ability to leverage high-quality style embeddings from StyleGAN fully.

Full Model (Ours). The complete model, incorporating both the StyleGAN encoder-decoder and self-supervised learning, achieves the best performance across all metrics, demonstrating that the synergy of both components leads to stronger generalization to unseen generators.

6 Conclusion

In this work, we propose a novel deepfake detection framework that leverages self-supervised style representation learning in the W+ latent space of StyleGAN. Moving forward, we aim to explore adversarial robustness and cross-domain generalization, extending our framework to detect AI-generated manipulations in real images. We hope our findings inspire future research in leveraging self-supervised learning for more generalizable and interpretable detection of AI-generated images.

REFERENCES

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254, 2021. URL https://api.semanticscholar.org/CorpusID:235436185.

Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *ArXiv*, abs/2210.00586, 2022. URL https://api.semanticscholar.org/CorpusID:252683252.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, 2020. URL https://api.semanticscholar.org/CorpusID:221266121.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. URL https://api.semanticscholar.org/CorpusID:253254809.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. URL https://api.semanticscholar.org/CorpusID:233033768.

Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2904–2908, 2021. URL https://api.semanticscholar.org/CorpusID:237372454.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *ArXiv*, abs/2004.02546, 2020. URL https://api.semanticscholar.org/CorpusID:214802845.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021. URL https://api.semanticscholar.org/CorpusID:243985980.

Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504, 2020. URL https://api.semanticscholar.org/CorpusID:221949373.

Yan Ju, Shan Jia, Jia Cai, Haiying Guan, and Siwei Lyu. Glff: Global and local feature fusion for ai-synthesized image detection. *IEEE Transactions on Multimedia*, 26:4073–4085, 2022. URL https://api.semanticscholar.org/CorpusID:253553484.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018. URL https://api.semanticscholar.org/CorpusID:54482423.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2019. URL https://api.semanticscholar.org/CorpusID:209202273.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235606261.

Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. E3dge: Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:253121047.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Dalius Matuzevi**c***ius*. $Diversedatasetforeyeglassesdetection : Extendingtheflickr - faces - hq(ffhq)dataset.$ *Sensors (Basel, Switzerland)*, 24, 2024. URL.

Shivansh Mundra, Gonzalo J. Aniano Porcile, Smit Marvaniya, James R. Verbus, and Hany Farid. Exposing gan-generated profile photos from compact embeddings. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 884–892, 2023. URL https://api.semanticscholar.org/CorpusID:259257906.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Gonzalo J. Aniano Porcile, Jack Gindi, Shivansh Mundra, James R. Verbus, and Hany Farid. Finding ai-generated faces in the wild. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4297–4305, 2023. URL https://api.semanticscholar.org/CorpusID:265213082.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2287–2296, 2020. URL https://api.semanticscholar.org/CorpusID:220936362.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021a. URL https://api.semanticscholar.org/CorpusID:245335280.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021b.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114, 2023. URL https://api.semanticscholar.org/CorpusID:259226993.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701, 2019. URL https://api.semanticscholar.org/CorpusID:209444798.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, 2018. URL https://api.semanticscholar.org/CorpusID:53295714.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2019. URL https://api.semanticscholar.org/CorpusID:196622700.