Robust CLIP-Based Detector for Exposing Diffusion Model-Generated Images

Santosh[1], Li Lin[1], Irene Amerini[2], Xin Wang[3], Shu Hu[1*]

[1]Purdue University, {santosh2, lin1785, hu968}@purdue.edu
[2]Sapienza University of Rome amerini@diag.uniroma1.it
[3]University at Albany, State University of New York xwang56@albany.edu

Abstract—Diffusion models (DMs) have revolutionized image generation, producing high-quality images with applications spanning various fields. However, their ability to create hyper-realistic images poses significant challenges in distinguishing between real and synthetic content, raising concerns about digital authenticity and potential misuse in creating deepfakes. This work introduces a robust detection framework that integrates image and text features extracted by CLIP model with a Multilayer Perceptron (MLP) classifier. We propose a novel loss that can improve the detector's robustness and handle imbalanced datasets. Additionally, we flatten the loss landscape during the model training to improve the detector's generalization capabilities. The effectiveness of our method, which outperforms traditional detection techniques, is demonstrated through extensive experiments, underscoring its potential to set a new state-of-the-art approach in DM-generated image detection. The code is available at https://github.com/Purdue-M2/Robust_DM_Generated_Image_Detection.

Index Terms—Diffusion models, CLIP, Robust, AI images

I. INTRODUCTION

Diffusion models (DMs) [1, 2] are a class of generative models that simulate a stochastic process to transform noise into structured data, effectively producing high-quality images. Unlike traditional generative approaches (*e.g.*, GAN), DMs gradually construct images through an iterative process of denoising, which allows for the generation of images with a high degree of fidelity and detail [3, 4]. This unique method of image creation results in DM-generated images that are not only visually compelling but also versatile in their application across various fields such as health, business, and the creative industries [5, 6]. The advantages of DMs and their generated images lie in their unparalleled ability to produce customized content tailored to specific needs, their flexibility in handling diverse data types, and their effectiveness in tasks that demand high-quality visual representations. This flexibility and high precision offer substantial benefits for enhancing synthetic datasets for machine learning and spearheading innovations in personalized content creation, establishing DMs as a groundbreaking technology in the image generation domain [7–9].

Despite the significant advantages of diffusion models (DMs) in generating high-quality, detailed images, their capability poses considerable challenges, particularly in authenticity and trust in digital media [12]. The core disadvantage
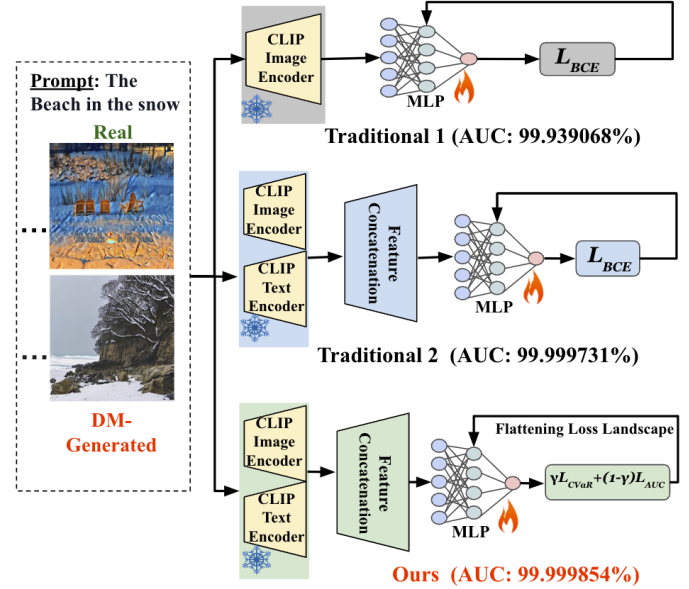
*Corresponding author

Fig. 1: *Comparison of our method with traditional methods.* **First row:** *Traditional Method 1 [10] utilizes CLIP image features combined with a Multilayer Perceptron (MLP) classifier and Binary Cross-Entropy (BCE) loss L_{BCE}.* **Second Row:** *Traditional Method 2 [11] incorporates both CLIP image and text features with an MLP classifier and BCE loss L_{BCE}.* **Third row:** *Our model enhances DM-generated image detection by using CLIP image and text features, a lightweight MLP classifier, and a combination of Conditional Value at Risk (CVaR) and Area Under the Curve (AUC) losses, across a flattened loss landscape in order to train a robust detector.*

of DM-generated images lies in their potential misuse for creating highly realistic yet entirely synthetic images, which can be nearly indistinguishable from genuine photographs. This capability raises serious concerns about misinformation, identity theft, and the creation of deepfakes, which can have far-reaching implications in politics, society, and personal security [13, 14]. The convincing nature of these images can be exploited to fabricate evidence, impersonate individuals in sensitive positions, or spread disinformation, undermining trust in digital content and potentially influencing public opinion and personal reputations [15].

Recent research has emphasized the critical need to tackle this issue, promoting sophisticated analysis methods that cap-

italize on the unique characteristics of content produced by diffusion models. These include minor irregularities or digital artifacts that do not occur in authentic images [16, 17]. These detection capabilities are crucial to mitigate the risks associated with the proliferation of synthetic images, ensuring a more secure and trustworthy digital environment [18, 19]. Notably, a range of approaches have been proposed to enhance the detection accuracy and counter the misuse of synthetic imagery. Physical and physiological methods focus on inconsistencies in images, such as aberrant reflections or lighting, and physiological anomalies in human depictions [20–23]. Techniques such as analyzing diffuser fingerprints detect unique artifacts inherent to specific image synthesis processes, pointing to their synthetic origins. Additionally, spatial-based and frequency-based methods focus on the textual and spectral anomalies that distinguish fake images from authentic ones [24–30].

Despite advancements, the robustness of these detection techniques faces significant challenges as generative models improve at evading detection. The reasons can be summarized into four key points: 1) **Limited Dataset Size:** Detectors are often trained on relatively small datasets, which may not adequately represent the broad diversity of images needed for comprehensive learning. 2) **Difficulty with Hard Images:** Some images pose greater challenges in learning. Standard training losses, such as cross-entropy, treat all images with equal importance, which can result in inadequate detection of these complex examples. 3) **Imbalanced Training Data:** Typically, neural network-based detectors are trained on datasets where DM-generated images outnumber real images, leading to training strategies that might cause the detector to exhibit biased performance by favoring the detection of the majority class. 4) **Suboptimal Optimization Techniques:** Traditional stochastic gradient descent optimization methods can lead to suboptimal training, where the model settles into local optima, thus reducing its generalization ability across diverse scenarios. Several recent studies [10, 11] have employed Contrastive Language–Image Pre-training (CLIP) [31] ViT as feature extractor, either focusing solely on image representations or integrating both image and text features with Multilayer Perceptrons (MLP) and Binary Cross-Entropy (BCE) loss for the DM-generated images detection. However, these approaches primarily tackle the issue of limited datasets for training.

To simultaneously address the above four challenges, in this work, we introduce a robust classifier designed to detect DM-generated images. Specifically, we utilize a pre-trained CLIP model to extract both image-level and corresponding text (prompt)-level features for each example. These features are then concatenated to represent the full spectrum of characteristics of the input example. For classification, we employ a trainable MLP model to differentiate between real and DM-generated images. To improve focus on hard examples, we implement a distributionally robust optimization technique known as Conditional Value at Risk (CVaR) loss. To manage the imbalanced training data, we incorporate an

Area under the ROC curve (AUC) loss. Additionally, to boost the model's generalization capability, we use Sharpness-Aware Minimization (SAM) to optimize model parameters. A comparative overview of our method against traditional approaches is illustrated in Figure 1. The main contributions of this work can be summarized as follows:

- We propose a lightweight, robust detector to detect images generated by diffusion models effectively.
- Our method enhances robustness comprehensively by establishing a framework that operates across three critical levels: features, loss, and optimization.
- Our method outperforms the state-of-the-art approaches, as highlighted in the extensive experiments and results.

II. RELATED WORK

A. DM-generated Image

Advancements in diffusion models (DMs) have led to their application in a wide array of generative tasks, demonstrating unparalleled ability in synthesizing realistic images and content. Following foundational work, Stable Diffusion (SD) [32] emerged as a versatile framework, enabling detailed image generation from textual prompts with various iterations including SD v1.4 and SD v2.0. Latent Diffusion Models (LDMs) represent a significant leap, reducing computational demands by operating in a latent space, thus facilitating higher resolution outputs. Additionally, commercial platforms have begun integrating these technologies, with DALL·E 2 by OpenAI [33] and Imagen by Google [34] leading the charge in text-to-image generation, pushing the boundaries of creative potential and application. Models such as Guided Diffusion [35] and Cascaded Diffusion [36] employ nuanced approaches to refine the synthesis process, thereby reducing artifacts and enhancing the naturalness of generated images. The introduction of Midjourney [37] and the expansion of Stable Diffusion into user-friendly interfaces like DreamStudio highlight the transition of DMs from research to mainstream use, indicating a growing interest and investment in generative AI technologies. These models, from their academic roots to commercial implementations, illustrate the dynamic evolution of diffusion models, setting new standards for artificial intelligence in creative domains.

B. DM-generated Image Detection

In the field of synthetic image detection, especially for images generated by diffusion models (DMs), methodologies can be distinctly categorized into two primary approaches: methods from the first category analyze solely the image data, and methods from the second category incorporate both images and textual information for classifying the input images. Specifically, the first category of methods primarily focuses on analyzing image data, employing a range of sophisticated techniques specifically tailored for diffusion models (DMs), and in some instances, methodologies originally developed for generative adversarial networks (GANs) have been effectively adapted to address the distinct challenges posed by DM-generated images. For example, [38] focuses
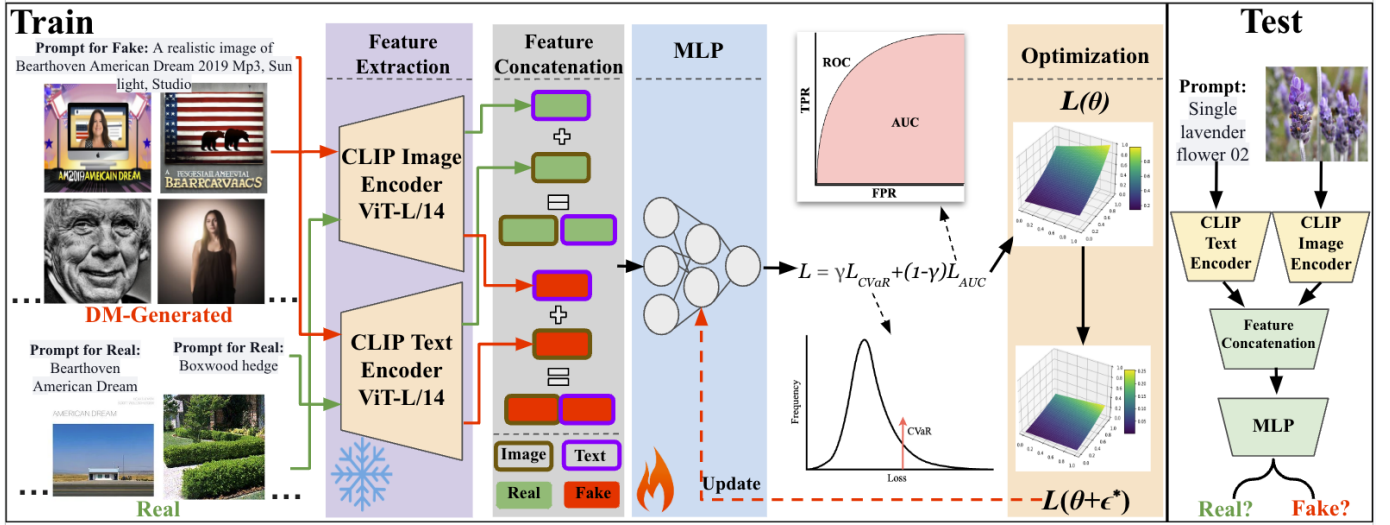
Fig. 2: *Overview of our proposed model using CLIP for encoding the input images and text, concatenating the image and text features, an MLP module with robust CVaR + AUC loss, and an optimization step involving a flattened loss landscape for detecting DM-generated images apart from the real images. The snowflake represents the module is fixed. The fire means the module will be trained.*

on physical/physiological features for classifying the real vs. fake content by categorizing qualitative failures of image generation models into five main types (*e.g.*, human parts, geometry, physics) and identifying common failure points such as unnatural textures or incorrect geometries. Similarly, [39] assesses if the 3D structures, shadows, and reflections in generated images conform to the physical and geometric rules that typically govern real-world photography and [20] applies a spherical harmonic-based approach to analyze the lighting in images generated by DALL-E-2 from text prompts. The specified focus on the datasets in these methodologies limit their generalization capabilities. [40] introduces a low-resource method for synthetic image detection and model lineage analysis that uses deep image fingerprints derived from convolutional neural network (CNN) architectures. Similarly, [41] integrates stepwise error analysis at multiple stages of the diffusion process. Deep neural network-based and transformer-based classifiers [10, 42–44] have proven effective in detecting images manipulated by image-to-image translation networks. However, their application to DM-generated images is often limited by challenges in generalizing to content developed by unseen models.

In the second category, methods utilize both visual and textual data, thereby broadening the scope of analysis. This dual-input strategy significantly enhances the detection capabilities by combining the strengths of image and text-based features. A notable implementation of this approach is [11], which employs CLIP's dual encoders to process images alongside their corresponding textual prompts. The visual and textual features are then concatenated to form a composite input. This method exemplifies the enhanced sensitivity and accuracy that can be achieved in classifying DM-generated content by addressing its inherent multimodal nature. However, this method does not consider robustness problems in real-world scenarios, which is the focus of our proposed work.

III. METHOD

Figure 2 provides an overview of our proposed framework.

A. Feature Space Modeling

We design our classification algorithm based on the image and text features encoded by the CLIP (Contrastive Language-Image Pre-training) ViT L/14 model, which has been trained on a large dataset containing 400 million image-text pairs [31]. The significance of this module lies in its ability to enable the model to identify both high-level semantic and detailed visual features, which are crucial for the classification task without requiring a large dataset for training. Consequently, we can ensure that our method is both lightweight and highly effective, facilitating easier deployment.

For a Dataset $D = (X_i, T_i, Y_i)_{i=1}^n$, each element comprises an image X_i, its associated text T_i, and a binary label $Y_i \in \{0, 1\}$, where 0 represents an authentic image and 1 a diffusion-generated image. By processing each image-text pair through the CLIP model, we obtain a joint feature representation, which incorporates the attributes essential for distinguishing between the two types of images. This approach generates a 1536-dimensional feature vector for each dataset entry, equally divided between image-derived and text-derived features. These vectors are collected into a feature bank $C = (F_i, Y_i)_{i=1}^n$, where F_i is the combined feature vector of the i-th dataset entry, and Y_i is its label.

B. MLP Classifier Architecture

Upon establishing the feature bank, we introduce a 3-layer Multilayer Perceptron (MLP) as our classifier. To ensure

stability and generalization, batch normalization and ReLU activation follow each linear layer, with dropout included to prevent overfitting. The classifier's output layer adapts to our binary classification task using a single neuron with sigmoid activation.

C. Objective Function

To effectively differentiate real images from diffusion-generated ones, we propose a dual-objective loss framework that synergizes Conditional Value-at-Risk (CVaR) and an AUC loss. This framework is designed to tackle the challenges posed by hard examples and imbalanced training sets.

Conditional Value-at-Risk (CVaR) Loss. The CVaR loss [45–54] component is formulated to make the model focus on the hardest examples in the dataset, thereby improving robustness. It is defined as:

$$L_{CVaR}(\theta) = \min_{\lambda \in \mathbb{R}} \left\{ \lambda + \frac{1}{\alpha n} \sum_{i=1}^{n} [\ell(\theta; F_i, Y_i) - \lambda]_+ \right\}, \quad (1)$$

where ℓ is the classification loss function (*e.g.*, binary cross-entropy loss), θ symbolizes the model parameters, and (F_i, Y_i) are the feature-label pairs from the dataset. $[a]_+ = \max\{0, a\}$ is the hinge function. The parameter $\alpha \in (0, 1)$ facilitates tuning the focus from the most challenging to more average examples.

AUC loss. For enhancing the model's discrimination ability, especially under class imbalance, an area under the curve (AUC) loss [55, 56] component is utilized. This component is strategically developed to optimize the AUC metric directly, ensuring the model effectively ranks pairs of positive and negative examples. Suppose $P = \{i | y_i = +1\}$ and $N = \{i | y_i = 0\}$ are sets of indices for positive (diffusion-generated) and negative (real) examples, respectively. Then, the AUC loss can be written as follows:

$$L_{AUC}(\theta) = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} R(s(\theta; F_i), s(\theta; F_j)), \quad (2)$$

where

$$\begin{aligned} &R(s(\theta; F_i), s(\theta; F_j)) \\ &= \begin{cases} (-(s(\theta; F_i) - s(\theta; F_j) - \eta))^p, & s(\theta; F_i) - s(\theta; F_j) < \eta, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (3)$$

$\eta \in (0, 1]$ and $p > 1$ are two hyperparameters. $s(\theta; F)$ indicates the score function of the model (*i.e.*, MLP) parameterized by θ for feature F. This formulation is aimed at increasing the margin between scores of positive and negative examples, thereby enhancing the AUC metric's contribution directly.

Total Loss Function. The overarching training objective harmonizes the CVaR and AUC losses into a unified framework:

$$L(\theta) = \gamma L_{CVaR}(\theta) + (1 - \gamma) L_{AUC}(\theta), \quad (4)$$

where γ serves as a balancing parameter to modulate the contributions from both loss components. This compound loss

---

Algorithm 1: End-to-end Training

Require: A training dataset C with size n, γ, α, η, p, max_iterations, num_batch, learning rate β
Ensure: Optimized parameters
1: **Initialization:** θ_0, $l = 0$
2: **for** $e = 1$ to *max_iterations* **do**
3:    **for** $b = 1$ to *num_batch* **do**
4:       Sample a mini-batch C_b from C
5:       Compute $\ell(\theta_l; F_i, Y_i)$, $\forall (F_i, Y_i) \in C_b$
6:       Use binary search to find λ that minimizes L (Eq. 4) on C_b
7:       Compute ϵ^* based on Eq.(5)
8:       Compute gradient approximation for the total loss
9:       Update θ: $\theta_{l+1} \leftarrow \theta_l - \beta \nabla_\theta L \big|_{\theta_l + \epsilon^*}$
10:       $l \leftarrow l + 1$
11:    **end for**
12: **end for**
13: **return** θ_l {Return the optimized model parameters}

---

function enables targeted learning from the most challenging instances, as prioritized by the CVaR component, while also promoting an improved classification performance that is not affected by the imbalanced data, as facilitated by the AUC-related component. The integrated approach provides a comprehensive solution to model optimization, tailored to the nuanced requirements of differentiating real from diffusion-generated images.

D. Optimization Method

Enhancing the model's generalization capabilities is achieved through Sharpness-Aware Minimization (SAM) [57], which seeks a flat loss landscape conducive to robust classification across diverse scenarios. This is realized by optimizing for a perturbation ϵ^* that maximizes the loss $L(\theta)$ within a defined perturbation magnitude δ:

$$\epsilon^* = \arg \max_{\|\epsilon\|_2 \leq \delta} L(\theta) \approx \arg \max_{\|\epsilon\|_2 \leq \delta} \epsilon^\top \nabla_\theta L = \delta \, \texttt{sign}(\nabla_\theta L), \quad (5)$$

The approximation is obtained using first-order Taylor expansion with the assumption that ϵ is small. The final equation is obtained by solving a dual norm problem, where `sign` represents a sign function and $\nabla_\theta L$ is the gradient of L with respect to θ. As a result, the model parameters are updated by solving the following problem:

$$\min_\theta L(\theta + \epsilon^*). \quad (6)$$

End-to-end Training. The training begins with the initialization of the model parameters θ. For each training iteration, a mini-batch C_b is randomly selected from the comprehensive feature bank C. The steps involved in each iteration are laid out as follows (also see Algorithm 1):

- Fix θ and use binary search to find the global optimum of λ since (4) is convex with respect to λ.
- Fix λ, compute ϵ^* based on Eq. (5).

- Update θ based on the gradient approximation for (6): $\theta \leftarrow \theta - \beta \nabla_\theta L(\theta + \epsilon^*)$, where β is the learning rate.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. Our experiments are based on the Diffusion-generated Deepfake Detection dataset (D3) [58] which includes a comprehensive collection of approximately 2.3 million records and 11.5 million images in the training set. This dataset comprises real images sourced from the LAION-400M [59] dataset and four synthetic images for each entry. The synthetic images are produced from a single textual prompt using a diverse array of text-to-image generation models, namely Stable Diffusion 1.4 (SD-1.4) [32], Stable Diffusion 2.1 (SD-2.1) [32], Stable Diffusion XL (SD-XL) [60], and DeepFloyd IF (DF-IF) [61]. This selection provides a broad representation of current synthetic image generation capabilities, including both variations of the Stable Diffusion technique and the distinct approach inspired by Imagen, represented by Deep-Floyd IF. The dataset showcases a variety of aspect ratios, specifically 256×256, 512×512, 640×480, and 640×360. However, we resize all images to 224×224. It also utilizes a broad spectrum of encoding and compression formats to mirror the prevalent image distribution found, with a focus on JPEG encoding to match the common distribution within the LAION dataset. The test set follows the similar format and includes a smaller set of 4,800 records, consisting of 4,800 real images and 19,200 synthetic images from the aforementioned generative models.

Evaluation Metrics. The model's effectiveness is measured by the Area Under the ROC Curve (AUC). This metric is utilized to measure the model's discriminative performance, serving as the primary benchmark for comparison.

Baseline Methods. We benchmark our approaches against two baseline models that utilize the CLIP architecture for feature extraction in the detection of AI-generated images. 1) The first baseline [10] employs a combination of CLIP image features with a Multilayer Perceptron (MLP) and Binary Cross-Entropy (BCE) loss for binary classification of images. 2) The second baseline [11] extends this approach by integrating both the visual and textual features provided by CLIP. It processes images alongside their corresponding textual prompts through CLIP's dual encoders, concatenating these features to form a composite input for the MLP classifier, also utilizing BCE loss. We implement and conduct extensive testing on both the baseline models for a thorough assessment of their performance in the same dataset.

Implementation Details. Our method and the second baseline employ a hybrid approach, integrating visual and textual features within a singular model framework. This integration results in a feature space of 1536 dimensions (768 for visual features, 768 for textual features) per data instance. All model architectures comprise three Multi-Layer Perceptron (MLP) layers, each equipped with 1536 units to navigate the complexities of the enhanced feature space. We conduct the training on a batch size of 32, optimizing data throughput for each

TABLE I: Comparison with the baseline methods in terms of AUC score. The best results are shown in **Bold**.

Metric	Traditional 1 [10]	Traditional 2 [11]	Ours
AUC(%)	99.939068	99.999731	**99.999854**

TABLE II: An ablation study of key components in our framework.

Method	CVaR Loss	AUC Loss	SAM	AUC (%)
V1	✓			99.999764
V2		✓		99.999467
V3	✓	✓		99.999747
V4	✓		✓	99.999849
Ours	✓	✓	✓	**99.999854**

iteration. An Adam optimizer, with an initial learning rate of $\beta = 1e - 3$, is utilized, complemented by a Cosine Annealing Learning Rate Scheduler to adaptively refine the learning rate throughout the training process. δ is 0.05 as specified in (5). Tuning the hyperparameters, we experiment the α values for CVaR Loss from 0.1 to 0.9. Once the best α hyperparameter is found, we experiment with trade-off hyperparameter γ from 0.1 to 0.9 with η and p set to 0.6 and 2, respectively. The implementation leverages the PyTorch framework, with computational support from an NVIDIA RTX A6000 GPU.

B. Results

We report the performance of our method and the compared methods in Table I. It is clear that our method outperforms the baselines, which achieves an outstanding AUC of 99.999854%, demonstrating near-perfect discrimination between real and DM-generated images.

C. Ablation Study

Effects of CVaR, AUC, and SAM. In this ablation study, we systematically examine and compare the performance of variants (V1, V2, V3, and V4) of our method against our final method to demonstrate the effectiveness of each module. As detailed in Table II, the performance comparison between V4 and ours shows an improvement of 0.0005%, indicating the effectiveness of AUC loss. Comparing V3 with ours, the AUC increases by 0.0107%, highlighting the impact of SAM. Additionally, the transition from V2 to V3, which incorporates the CVaR component, results in a 0.028% improvement in AUC. Therefore, the impact of each component in our method is: CVaR loss > SAM > AUC loss.

Visualization of Loss Landscape. Figure 3 clearly demonstrates the benefits of incorporating SAM optimization into our methodology. The absence of SAM results in a complex and uneven loss landscape, complicating the optimization efforts and potentially causing erratic generalization. On the other hand, the figure on the right illustrates a significantly smoother loss landscape when SAM is integrated. This smoother landscape implies a more stable model that is better at generalizing to unfamiliar data.
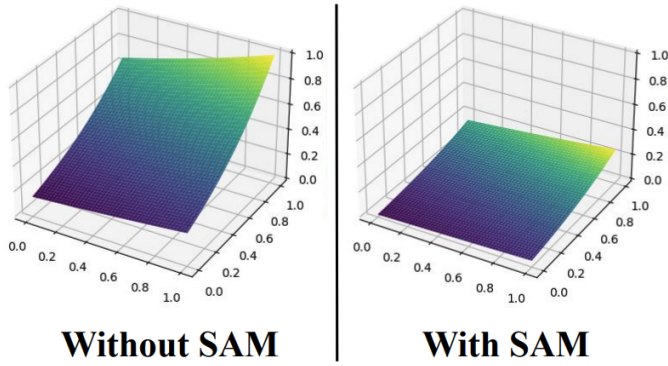
Fig. 3: The loss landscape visualization of our proposed method without (left) and with (right) using the Sharpness Aware minimization (SAM) method.
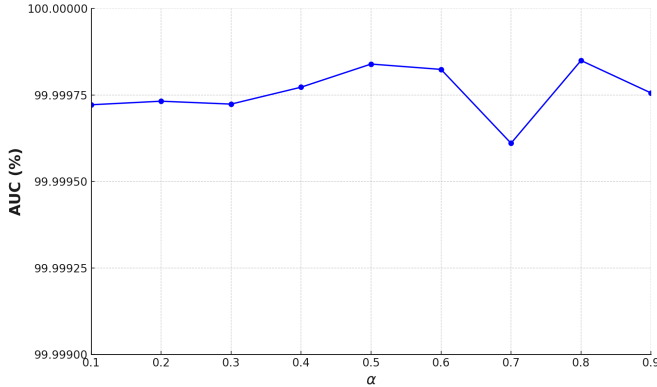

Fig. 4: AUC score with respect to different α values.

D. Sensitivity Analysis

In the sensitivity analysis, the influence of α in the CVaR loss is examined. The range of α values from 0.1 to 0.9 was explored to identify the optimal setting for CVaR, crucial for maximizing detection capabilities. Figure 4 illustrates the AUC as a function of various CVaR α values. This analysis pinpointed an optimal CVaR α value of 0.8, correlating with the highest AUC.

Following the determination of the optimal CVaR α at 0.8, a further analysis was conducted to evaluate the trade-off between CVaR loss and AUC loss in Eq. (4). The range of γ values from 0.1 to 0.9 was revisited for this purpose. The results, presented in Figure 5, indicated that an γ of 0.5 for the CVaR and AUC trade-off facilitated the most substantial improvement in model performance.

V. CONCLUSION

The challenge of accurately detecting images generated by diffusion models (DMs) is compounded by the sophistication of the generative processes and the realism of the output images. Addressing these issues, we introduce a robust detection framework that employs CLIP image and text features combined with a Multilayer Perceptron (MLP) classifier. We utilize a combination of CVaR and AUC losses for model training. The model is then trained with a loss landscape flattening
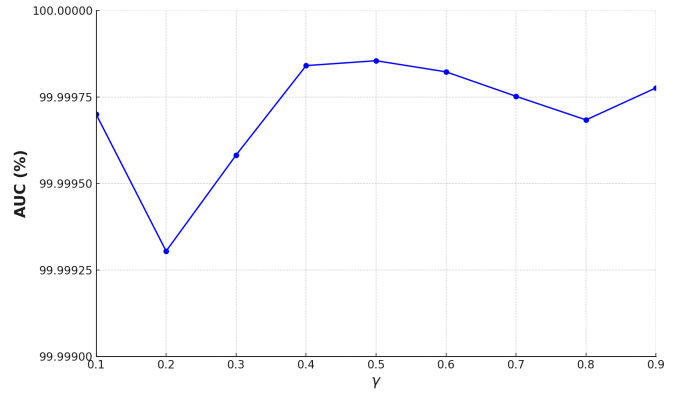

Fig. 5: AUC score with respect to different γ values by fixing $\alpha = 0.8$.

strategy. Our approach not only enhances the model's robustness but also improves the model's generalization capability, which has demonstrated superior capabilities in distinguishing between real and DM-generated images, as confirmed by extensive experiments that yielded nearly perfect discrimination capabilities.

Limitations. One limitation is that our method can only test images that have corresponding text information.

Future Work. We plan to broaden the scope of our methodology by incorporating additional modalities and datasets to demonstrate the robustness and scalability of our detection techniques. We also aim to extend our approach to address newer generative models beyond DMs, such as advanced versions of GANs.

REFERENCES

[1] Jonathan Ho et al. Denoising diffusion probabilistic models. *neurips*, 33, 2020.
[2] Yang Song et al. Score-based generative modeling through stochastic differential equations. *arXiv*, 2020.
[3] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. 2024.
[4] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. *arXiv*, 2024.
[5] Siyuan Mei et al. Segmentation-guided knee radiograph generation using conditional diffusion models. *arXiv*, 2024.
[6] Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Future-proofing class incremental learning. *arXiv*, 2024.
[7] Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. Would deep generative models amplify bias in future models? *arXiv*, 2024.
[8] Dongzhi Jiang et al. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv*, 2024.
[9] Rinon Gal et al. Lcm-lookahead for encoder-based text-to-image personalization. *arXiv*, 2024.

[10] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. *arXiv*, 2023.

[11] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC CCS*, 2023.

[12] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images. *ACM TOMM*, 2024.

[13] Siwei Lyu. Deepfake detection: Current challenges and next steps. *arXiv*, 2020.

[14] Andreas Rossler et al. Faceforensics++: Learning to detect manipulated facial images. In *CVPR*, 2019.

[15] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *MIPR*, 2018.

[16] Mamadou Keita, Chloé Clavel, and Thomas Pellegrini. Bilora: A vision-language approach for synthetic image detection. *arXiv*, 2024.

[17] Wei Shen and Li Liu. Lare2: Latent reconstruction error based method for diffusion-generated image detection. *arXiv*.

[18] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. *arXiv*, 2024.

[19] Yuhang Lu and Touradj Ebrahimi. Towards the detection of ai-synthesized human face images. *arXiv*, 2024.

[20] Hany Farid. Lighting (in)consistency of paint by text. *arXiv*, 2022.

[21] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Open-eye: An open platform to study human performance on identifying ai-synthesized faces. In *MIPR*, 2022.

[22] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP*, 2022.

[23] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP*, 2021.

[24] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *arXiv*, 2024.

[25] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv*, 2024.

[26] Bing Fan, Shu Hu, and Feng Ding. Synthesizing black-box anti-forensics deepfakes with high visual quality. In *ICASSP*. IEEE, 2024.

[27] Lei Zhang, Hao Chen, Shu Hu, Bin Zhu, Xi Wu, Jinrong Hu, and Xin Wang. X-transfer: A transfer learning-based framework for robust gan-generated fake image detection. *arXiv*, 2023.

[28] Shanmin Yang, Shu Hu, Bin Zhu, Ying Fu, Siwei Lyu, Xi Wu, and Xin Wang. Improving cross-dataset deepfake detection with deep information decomposition. *arXiv*, 2023.

[29] Bing Fan, Zihan Jiang, Shu Hu, and Feng Ding. Attacking identity semantics in deepfakes via deep feature fusion. In *MIPR*, 2023.

[30] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu. Gan-generated faces detection: A survey and new perspectives. *ECAI*, 2022.

[31] Alec Radford et al. Learning transferable visual models from natural language supervision, 2021.

[32] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022.

[34] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. *neurips*, 2022.

[35] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *neurips*, 2021.

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. 2021.

[37] Midjourney. https://mid-journey.ai/.

[38] Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vis. Comput.*, 2023.

[39] Hany Farid. Perspective (in) consistency of paint by text. *arXiv*, 2022.

[40] Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection. *arXiv*, 2023.

[41] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv*, 2023.

[42] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[44] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[45] Li Lin, Yamini Sri Krubha, Zhenhuan Yang, Cheng Ren, Xin Wang, and Shu Hu. Robust covid-19 detection in ct images with clip. *arXiv*, 2024.

[46] Li Lin et al. Robust light-weight facial affective behavior recognition with clip. *arXiv*, 2024.

[47] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. *CVPR*, 2024.

[48] Yan Ju, Shu Hu, Shan Jia, George H Chen, and Siwei Lyu. Improving fairness in deepfake detection. In *WACV*, 2024.

[49] Shu Hu, Zhenhuan Yang, Xin Wang, Yiming Ying, and Siwei Lyu. Outlier robust adversarial training. In *ACML*, 2024.

[50] Shu Hu, Xin Wang, and Siwei Lyu. Rank-based decomposable losses in machine learning: A survey. *TPAMI*, 2023.

[51] Shu Hu and George H Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, 2022.

[52] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. In *ICCV*, 2021.

[53] Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu. Sum of ranked range loss for supervised learning. *Journal of Machine Learning Research*, 2022.

[54] Shu Hu, Yiming Ying, Siwei Lyu, et al. Learning by minimizing the sum of ranked range. *neurips*, 2020.

[55] Wenbo Pu et al. Learning a deep dual-level network for robust deepfake detection. *Pattern Recognition*, 2022.

[56] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Robust attentive deep neural network for detecting gan-generated faces. *IEEE Access*, 2022.

[57] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020.

[58] Diffusion-generated deepfake detection datase. In *https://huggingface.co/datasets/elsaEU/ELSA_D3*.

[59] Christoph Schuhmann et al. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv*, 2021.

[60] S. Chen and J. Lee. Stable diffusion xl: Enhancing the resolution and quality of ai-generated images. *Journal of Artificial Intelligence Research*, 2023.

[61] R. Floyd et al. Deepfloyd-if: A novel image forensics method for identifying deepfake images. *Journal of Digital Forensics, Security and Law*, 2023.