2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

IVORY: Adversarial Purification of Obfuscated Faces to Extract Soft-Biometrics using Diffusion Transformers

Shaikh Akib Shahriyar[1], Matthew Wright[1] and Armon Barton[2]
[1] Department of Cybersecurity, Rochester Institute of Technology, Rochester, NY, USA
[2] Computer Science Department, Naval Postgraduate School, Monterey, CA, USA

***Abstract*— The proliferation of online face images has heightened privacy concerns, as adversaries can exploit facial features for nefarious purposes. While adversarial perturbations have been proposed to safeguard these images, their effectiveness remains questionable. This paper introduces IVORY, a novel adversarial purification method leveraging Diffusion Transformer-based Stable Diffusion 3 model to purify perturbed images and improve facial feature extraction. Evaluated across gender recognition, ethnicity recognition and age group classification tasks with CNNs like VGG16, SENet and MobileNetV3 and vision transformers like SwinFace, IVORY consistently restores classifier performance to near-clean levels in white-box settings, outperforming traditional defenses such as Adversarial Training, DiffPure and IMPRESS. For example, it improved gender recognition accuracy from 37.8% to 96% under the PGD attack for VGG16 and age group classification accuracy from 2.1% to 52.4% under AutoAttack for MobileNetV3. In black-box scenarios, IVORY achieves a 22.8% average accuracy gain. IVORY also reduces SSIM noise by over 50% at 1x resolution and up to 80% at 2x resolution compared to DiffPure. Our analysis further reveals that adversarial perturbations alone do not fully protect against soft-biometric extraction, highlighting the need for comprehensive evaluation frameworks and robust defenses.**

I. INTRODUCTION

The sharing of personal images has become a daily routine for billions of users worldwide. Platforms such as Facebook, Instagram and X see the constant uploading of pictures, many of which prominently feature human faces. These faces contain a wealth of biometric data, including not only identity, but also more nuanced soft-biometrics such as gender, age and ethnicity [8], [39]. Researchers have raised concerns regarding the misuse of this information [7], [36], such as surveillance, unauthorized profiling and targeting in social engineering attacks. The lack of awareness among users about the extent of this data mining further exacerbates the issue, as they inadvertently expose highly personal information with every shared photo [24], [48], [52]. As a result, there is a growing need for privacy-preserving techniques that can allow users to maintain control over their biometric data while continuing to share images online.

One of the most promising methods for safeguarding facial data involves using adversarial perturbations to obfuscate soft-biometrics [11], [42], [43], [57], [62]. These perturbations manipulate the images to confuse models trained to identify facial soft-biometrics while maintaining high fidelity for human viewers.

In addition to its use in protecting facial soft-biometrics, adversarial perturbation has been explored as a means to shield artists' unique styles from being mimicked by text-to-image generative models [40], [56], [64]. Style mimicry, wherein a model can replicate an artist's distinctive aesthetic, poses a significant threat to contemporary artists who rely on their creative identity to sustain their careers. In response, adversarial techniques have been applied to published artworks, introducing subtle perturbations that disrupt the fine-tuning process of generative models, making it difficult for them to learn and replicate artists' styles accurately.

Unfortunately for both privacy-concerned users and artists, researchers have demonstrated that these uses of adversarial examples are not secure. Recent advancements in *adversarial purification*, such as DiffPure [47] and IMPRESS [9], can reverse the effects of style protection perturbations, rendering them ineffective. These methods operate by purifying images of the adversarial noise, thus enabling generative models to bypass the embedded protections and once again accurately mimic an artist's style.

Given the success of adversarial purification, we sought to explore how emerging diffusion-transformer models could further enhance the purification process, specifically in the context of adversarially perturbed facial images. The Multi-Modal Diffusion Transformer (MMDiT) [23] found in Stable Diffusion 3 has demonstrated strong performance across a variety of image restoration tasks, making it a promising candidate for adversarial purification.

In this paper, we introduce IVORY, a novel purification method that leverages the MMDiT architecture. By integrating MMDiT's diffusion capabilities, IVORY effectively purifies adversarially perturbed images, restoring them to a state where facial feature extraction is significantly enhanced as illustrated in Fig. 1.

The key contributions of this papers are:

- We propose IVORY, a novel diffusion-based purification method leveraging diffusion transformers to effectively counter adversarial perturbations.
- IVORY demonstrates significant improvements in adversarial defense, restoring classification accuracy across tasks like gender, ethnicity and age group recognition in both white-box and black-box attack scenarios.
- Our results show that IVORY consistently outperforms traditional adversarial training, DiffPure and IMPRESS, achieving superior purification across various CNN and Vision Transformer (ViT) architectures and attack methods, gaining 22.8% average accuracy in the black-box

979-8-3315-5341-8/25/$31.00

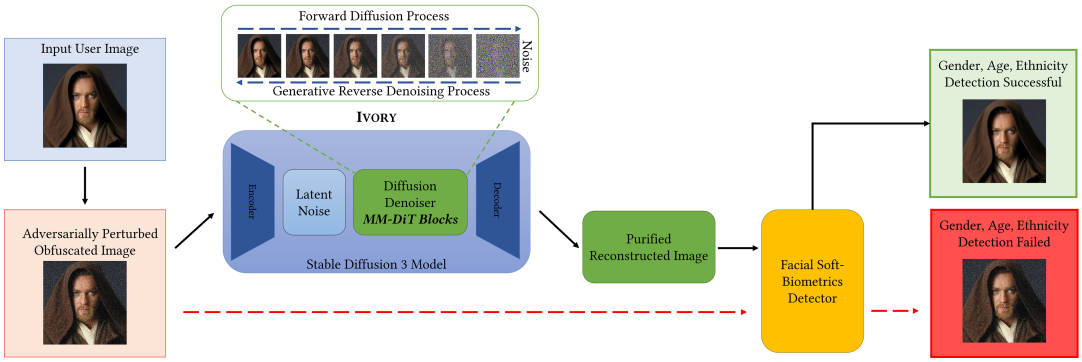

Fig. 1. An overview of IVORY, which integrates a MMDiT (Stable Diffusion 3) model to purify adversarially perturbed images. The process begins with the input of user image and the generation of adversarially perturbed obfuscated images. The MMDiT adds Gaussian noise to the obfuscated image in the forward diffusion process which is then denoised using the Img2Img pipeline with a guidance scale to obtain the purified images, ready for soft-biometric extraction.

setting.

- We show that IVORY also reduces SSIM noise by over 50% at 1x resolution and up to 80% at 2x resolution compared to DiffPure.

This study highlights the limitations of adversarial perturbations in soft-biometric privacy protection and emphasizes the need for more robust privacy-preserving techniques.

II. Related Work

Adversarial Training. Adversarial training has emerged as one of the most prominent and effective defense strategies against adversarial examples [4], [25]–[27], [41], [44], [51], [67], [68]. Despite extensive research, adversarial training still struggles with an inherent trade-off: more robust models lose performance on clean data [6].

Adversarial Purification. Adversarial purification has gained significant attention as an alternative to adversarial training, focusing on removing adversarial noise from perturbed images before classification [14], [28], [31], [32], [53], [61]. More recent developments, such as DiffPure [47], highlight the capacity of diffusion models to reduce the KL divergence between clean and adversarial images, leading to more aligned and purified outputs. DiffPure, however, is time-consuming and impractical for real-time applications [65]. IMPRESS [9] marked another leap forward by targeting adversarial style mimicry and offering robust purification specifically against artistic style extraction, but it also suffers from a longer purification time. Diffusion models like those proposed by Ankile et al. [3] and Shi et al. [59] continue to enhance purification techniques with novel architectures and improved noise-scheduling methods.

Diffusion Models. Diffusion models have shown remarkable capabilities for image-editing tasks, including image-to-image translation [18], [45] and text-guided image manipulation [37], [46]. Adversarial purification shares similarities with these tasks, as seen in DiffPure [47] and SDEdit [45]. Few works, however, have explored the most recent diffusion models, such as SDXL [49], SDXL-Turbo [54], Pixart-α [17] and Stable Diffusion 3 [23], for improving model robustness. The only work we know of is by Honig et al. [33], who showed the value of SDXL as a purifier. No work has yet evaluated Stable Diffusion 3's MMDiT architecture for adversarial purification.

III. Threat Model

In this work, we aim to evaluate the purification capabilities of the MMDiT in reducing adversarial noise for facial soft-biometric obfuscation tasks.

Our assumptions are grounded in the real-world challenge of adversarial purification, where an adversary seeks to reverse the obfuscation applied by a user to preserve their privacy. Specifically, we consider the following assumptions:

- *Attacker's Goal*: The attacker aims to purify images that have been obfuscated by the user through methods such as AutoAttack [19] or Carlini & Wagner (C&W) [16]. This noise is designed to prevent soft-biometrics (e.g., gender, age, ethnicity) from being extracted.
- *Soft-Biometrics Extraction*: Once the adversary purifies the image, they attempt to extract soft-biometrics from it using a CNN-based or ViT-based soft-biometric classifier. Thus, the purification process should not degrade or modify the image and ideally would restore the image visually as close as possible to the original.
- *Unknown Adversarial Noise*: The adversary is unaware of the specific type of adversarial noise added to the image.

a) White-Box Scenario: In the white-box scenario, we assume that the user and the adversary share the same CNN or ViT models (including weights) for soft-biometric extraction. The user uses the model to create adversarial perturbations for their images, which improves their effectiveness in fooling the same model used by the adversary. Although the adversary may know which model the user has, it provides no advantage to the purification algorithms.

b) Black-Box Scenario: The black-box scenario reflects a more realistic setting where the user and adversary are not aware of what the other one is using. We model this by having each party use a different model architecture: the user generates their adversarial samples with one model, while the

adversary evaluates the sample after purification on another model. This will generally reduce the effectiveness of the user's adversarial perturbations. We note that our black-box experiments are in what may more properly be called a *gray-box* scenario, as both parties train on a single training dataset and share knowledge of the task (gender, ethnicity, or age group classification).

IV. IVORY: METHODOLOGY

A. Soft Biometric Extraction

Over the years, various methods have been proposed for extracting soft-biometrics from facial images [12], [20], [29], [30]. The tasks we use are: (i) gender recognition is typically a binary classification task (male/female) [5]; (ii) age estimation classifies individuals into predefined age groups [12]; and (iii) ethnicity recognition involves classifying individuals into groups such as Caucasian, African American, East Asian and Asian Indian [29]. These classification tasks serve as the foundation for evaluating soft-biometrics extraction in the context of adversarial purification.

In our work, we selected three CNN models that have been employed in one of the most recent studies [11] on soft-biometrics extraction and consistently demonstrated strong performance in facial analysis tasks: VGG-16 [60], SENet [35] and MobileNetV3 [34]. These models have been pre-trained on large-scale datasets such as VGGFace2 [10] and ImageNet [21], making them ideal for transfer learning and soft-biometrics classification tasks.

To examine Transformer-based approaches, we also included SwinFace [50], a ViT-based face perception model that uses the Swin Transformer as its backbone. Unlike CNN-based models, SwinFace uses a hierarchical feature representation that captures multi-scale facial features, making it particularly robust to pose variations and occlusions. Additionally, it incorporates the Multi-Level Channel Attention (MLCA) module to resolve conflicts in multi-task learning, allowing it to jointly learn facial attributes more effectively. SwinFace achieves state-of-the-art performance while efficiently handling over 40 facial perception tasks, including face recognition, facial expression analysis, age estimation and facial attribute classification.

B. Datasets

We use the same datasets as Carletti et al. in the most recent studies on soft-biometrics extraction: VGGFace2, VMER, VMAGE and Adience [11].

VGGFace2. VGGFace2 [10] is one of the largest publicly available face datasets, consisting of over 3 million images of more than 9,000 individuals. It includes a wide range of variations in pose, age, gender and ethnicity.

VMER and VMAGE. The VGGFace2 MIVIA Ethnicity Recognition (VMER) [29] and VGGFace2 MIVIA Age (VMAGE) [30] datasets extend the original VGGFace2 dataset by providing additional labels for ethnicity and age, respectively. VMER categorizes individuals into four ethnic groups: African American, East Asian, Caucasian Latin and Asian Indian. VMAGE assigns each individual to an age group (0-15, 16-25, 26-35, 36-45, 46-60, 61+).

Adience. This dataset [38] is specifically designed for gender and age classification with 26,580 face images in a wide range of variations in appearance, lighting and image quality, making it suitable for evaluating the robustness of soft-biometrics extraction models under real-world conditions. Its age categories range from 0-2 years to 60+ years.

C. Generating Adversarial Samples

We employ three well-known adversarial attacks: AutoAttack [19], PGD [44] and C&W [16]. These methods are widely regarded for their robustness and efficacy in generating adversarial examples [55].

D. Purification of Adversarial Samples with MMDiT

IVORY leverages the power of the MMDiT to purify adversarial samples. In this section, we describe how IVORY handles adversarial input images, how it performs forward diffusion and how the reverse process restores the clean image for soft-biometrics extraction.

a) Processing the Input Adversarial Image: The purification process in IVORY's application of MMDiT begins by processing the adversarial image $x_{\text{adv}} = x + \delta$, where x represents the original image and δ is the adversarial noise. The image is first tokenized into a sequence of patches, with each patch linearly embedded into a vector of dimension d. This is similar to the patch embedding mechanism used in the ViT architecture [22]. Given an image x_{adv} of size $I \times I \times C$, MMDiT divides it into patches of size $p \times p$, resulting in a sequence of tokens of length $T = (I/p)^2$. Each patch is then embedded as a token, forming the input sequence for the MMDiT transformer blocks. In addition to image tokens, MMDiT incorporates auxiliary inputs such as noise timesteps t and potentially textual descriptions or class labels c, which are appended as extra tokens to the input sequence. This multi-modal approach enables MMDiT to process both image data and additional context.

b) Forward Diffusion Process: Once the input sequence has been prepared, the forward diffusion process is applied. Each forward diffusion step adds Gaussian noise to the image up to a predefined timestep t^*, reducing the impact of the adversarial noise.This process can be described by Eqn. 1, where $\bar{\alpha}_{t^*}$ controls the noise intensity at timestep t^* and $\epsilon \sim \mathcal{N}(0, I)$ represents samples from a normal distribution.

$$(x_{\text{adv}})_{t^*} = \left(\sqrt{\bar{\alpha}_{t^*}}\right) x_{\text{adv}} + \left(\sqrt{1 - \bar{\alpha}_{t^*}}\right) \epsilon \tag{1}$$

If we set t^* too high, the noisy image will lose key image details, making accurate recovery difficult. So we must select t^* to balance the amount of purification against the risk of losing image features.

c) Reverse Diffusion and Purification: Starting from the noisy image x_{t^*}, the reverse diffusion process gradually removes the noise by approximating the reverse transitions using a neural network $p_\theta(x_{t-1}|x_t)$. This reverse process is

mathematically defined in Eqn. 2, where p_θ is the network parameterized by θ, which predicts the denoised image at each timestep by learning the reverse Markov process.

$$\hat{x}_0^{\text{adv}} = p_\theta(x_{t-1}|x_t) \quad \text{for} \quad t = t^*, t^* - 1, \ldots, 0 \qquad (2)$$

MMDiT improves the effectiveness of this reverse process by utilizing multi-head attention mechanisms that integrate both image and auxiliary token information (such as noise timesteps t and class labels c). This attention mechanism allows the model to consider both the visual details of the image and any contextual information that can guide the purification process. More details on diffusion purification and MMDiT are provided in Appendix B.

By iteratively applying the reverse denoising steps, IVORY outputs a purified image $\hat{x}_0^{\text{adv}}$ that closely resembles the original image x while being stripped of adversarial perturbations. This purified image can then be used by the classifier C to accurately extract soft-biometrics, free from the distortions caused by adversarial noise.

V. EXPERIMENTAL SETUP

Our experiments were conducted on a machine equipped with an NVIDIA A100 GPU with 80 GB of VRAM and 125 GB of system RAM running on Ubuntu 22.04.4 LTS.

a) Models for Soft Biometrics Recognition: The CNNs and ViT used for facial soft biometrics recognition were trained on VGGFace2, along with its extensions VMER (for ethnicity recognition) and VMAGE (for age estimation). For evaluating performance, gender and ethnicity recognition tasks were tested on the VGGFace2 test set, which contains 170,000 images from 500 identities. Age estimation was tested using the Adience dataset. While some models used in the experiments were finetuned from publicly available pre-trained weights, others were trained specifically for this study. Pre-trained models for age classification [30] and gender recognition (VGG-16 and SENet) were publicly available from [12], [29]. These pre-trained networks were chosen due to their state-of-the-art performance and robustness against common real-world corruption. For this work, we trained MobileNetV3 [34] for gender recognition and ethnicity classification, following similar training procedures as those used for the pre-trained models.

b) Face Cropping and Data Preprocessing: We followed the preprocessing procedure described by Carletti et al. [11]. We used a Single Shot Detector (SSD) based on the ResNet-10 architecture to crop the faces from the images in the VGGFace2 dataset. As the cropped faces may vary in shape, padding was applied to make the final input size 224×224 pixels, ensuring that the face was centered and occupied approximately 80% of the input image [10]. This preprocessing step was crucial for ensuring that the CNNs received consistent input during training, eliminating potential errors due to face detection. The images were normalized by subtracting the mean value of each color channel, calculated over the entire dataset.

c) Training Procedure: The CNNs were trained using SGD with a batch size of 128 for MobileNetV3 and 32 for VGG-16 and SENet. The training process was initialized with a learning rate of 0.005, which decayed by a factor of 0.2 every 20 epochs and a weight decay of 0.05 for regularization. We fine-tuned SwinFace according to its publicly available codebase, carefully adjusting hyperparameters to optimize performance.

d) Adversarial Attack Parameters: The user's objective is to confuse soft biometric classifiers without modifying the image too much when posting about themselves online. In our experimental setup, we thus configured the adversarial attack parameters to balance the effectiveness of the perturbations with the preservation of image quality. Carletti et al. [11] prepared samples with varying noise intensities for each attack type and had five human observers assess the maximum level of noise that remained imperceptible. Based on their findings, they constrained the ℓ_∞ and ℓ_2 norms to 15 and 900, respectively. We use the same values.

For generating adversarial examples, we used the Adversarial Robustness Toolbox (ART) to implement the attacks and applied the following attack parameters:

- *PGD*: The attack was iterated for a maximum of 40 steps with a step size of $\alpha = 0.01$ and a perturbation size of $\epsilon = 0.005$ for ethnicity and gender recognition tasks. For age estimation, we used $\alpha = 0.007$ and $\epsilon = 0.007$.
- *Carlini & Wagner (C&W)* ℓ_2: We used a learning rate of 0.02 and a maximum of 50 iterations.
- *AutoAttack*: The parameters included a maximum iteration count of 40, a step size of 0.005 and $\alpha = 0.01$.

The rest of the attack parameters were kept at default values, following the ART Library instructions.

e) Adversarial Training: As a baseline, we implemented adversarial training for both white-box and black-box attack scenarios. In a real-world white-box setting, the attacker would adversarially train the best-performing model using the most transferable adversarial samples. As shown in Table I, we found that MobileNetV3 had the best performance among the three base CNN models, so we focused on it for our computationally expensive adversarial training tests for both white-box and black-box attack scenarios.

We performed adversarial training by finetuning MobileNetV3 rather than training it from scratch, following the training setup in [11]. We extracted 750,000 random samples from the original training set and created adversarial examples using PGD and AutoAttack. Note that PGD and AutoAttack are designed to be more transferable compared to C&W [19], [44]. This resulted in a total of 2.2 million images per task across clean, AutoAttack and PGD samples, culminating in a dataset of 6.75 million images.

In the black-box setting, we evaluated the adversarially trained MobileNetV3 on PGD and AutoAttack samples generated using VGG16 and SENet. In this scenario, the adversarially trained model had no direct knowledge of the adversarial samples generated with the other models.

The training was conducted using SGD with an initial learning rate of 0.001 and we applied a learning rate decay factor of 0.5 every five epochs.

f) DiffPure: Among recent diffusion-based purification techniques, we chose to benchmark the CNN and ViT models against DiffPure [47]. Hönig et al.'s recent report found that DiffPure performed about as well as the more complex IMPRESS++ [33]. We implemented DiffPure with Stable Diffusion XL 1.0 (SDXL) [49] using the HuggingFace `AutoPipelineForImage2Image` pipeline, with a guidance scale of 7.5, prompt P = *"High Quality,"* and strength parameter of 0.2. The guidance scale and strength parameters were based on empirical best practices [33]. As the prompt is subject-agnostic, it guides the reverse diffusion process to reconstruct the original image, rather than modifying it.

g) IMPRESS: IMPRESS [9] is one of the most recent diffusion-based adversarial purification methods and it optimizes images in the latent space using the Variational Autoencoder (VAE) of a latent diffusion model (LDM). Given the significant computational cost and extended runtime of IMPRESS, we limit its evaluation to white-box experiments. We used the IMPRESS implementation provided by the authors. To maintain consistency, we have configured IMPRESS using the same SDXL model, pipeline, parameters and prompt as used for DiffPure.

h) IVORY: Instead of a regular diffusion model, IVORY uses the best-performing publicly available diffusion-transformer model, namely Stable Diffusion 3 (SD3). We used SwarmUI 0.9.1 Beta to interface with the model. Based on our empirical testing, we used the Img2Img pipeline with a guidance scale of 7, a prompt P = *"High Quality,"* and a strength parameter of 0.2 for the number of diffusion timesteps. We used `FlowMatchEulerDiscreteScheduler` as the scheduler, which leverages the flow-matching sampling technique introduced in SD3 [23] to denoise the encoded image latents.

i) Evaluation Metrics: We primarily gauge the effectiveness of IVORY by the accuracy we get with standard classifiers on our three downstream tasks – gender, ethnicity and age group classification. We also measure the amount of adversarial noise before and after purification by calculating the Structural Similarity Index Measure (SSIM) [66], which quantifies the perceptual similarity between the original image and its perturbed or purified counterpart.

VI. RESULTS

Table I shows the baseline accuracy of our unprotected models on unmodified samples. As expected, SwinFace outperforms the CNN models [58]. We also use MobileNetV3 for our key findings, with results for VGG16 and SENet in the appendices (available from the authors upon request).

TABLE I
CLEAN ACCURACY OF UNPROTECTED MODELS

Task	MobileNetV3	VGG16	SENet	SwinFace
Gender	97.24%	96.85%	96.50%	97.51%
Ethnicity	93.67%	91.25%	92.30%	95.22%
Age	54.62%	57.51%	63.04%	66.88%

A. White-Box Purification

In the white-box setting, the obfuscated adversarial samples are generated and evaluated, after the purification, by the same model architecture. We find that IVORY effectively purifies adversarial noise and restores classifier performance across various tasks and architectures.

Table II shows the main results of our whitebox experiments. On the MobileNetV3 model, IVORY performs as well as or better than DiffPure and IMPRESS in all settings. It restores model accuracy to near-clean levels in many cases. On age group classification, for example, IVORY improves the accuracy from 2.14% against AutoAttack samples to 52.44% versus 51.61% for DiffPure, 51.60% for IMPRESS, and 31.94% for adversarial training.

Using the more powerful SwinFace model, IVORY is even more effective. The model suffered significant accuracy degradation under adversarial attacks. For example, AutoAttack reduced SwinFace's accuracy on Age Group Classification to just 3.2%, highlighting the effectiveness of the perturbations. IVORY successfully restored SwinFace's accuracy to 63.92%, compared to DiffPure's 59.35%, demonstrating its superior purification capabilities. Across most task and attack settings, IVORY consistently outperformed DiffPure by **2-3%** for SwinFace.

B. Black-Box Purification

In the black-box setting, the adversary attempts to extract soft-biometrics from the purified samples using a different model than the one the user generates adversarial perturbations with. IVORY demonstrates strong resilience across various tasks and perturbation methods in this more realistic setting as well.

On both the MobileNetV3 and SwinFace models, IVORY outperforms DiffPure by 1% or more on nearly every task. It produces an average accuracy gain across all tasks of 22.80%, compared to 19.48% for DiffPure. For instance, in the gender recognition task under the C&W attack, MobileNetV3's unprotected accuracy dropped to 51.2% under attack, whereas it maintained a significantly higher accuracy of 94.12% with IVORY (90.56% for DiffPure).

Overall, the performance gains across all these experiments demonstrate that IVORY consistently provides superior purification, regardless of the underlying model architecture.

C. Versus Adversarial Training & IMPRESS

In addition to comparisons with DiffPure, we compare IVORY to adversarial training and IMPRESS using MobileNetV3. Given IMPRESS's substantial computational cost, we limited its evaluation to a subset of 250 samples per task-attack setting, while adversarial training, DiffPure and IVORY were tested on the full dataset. Hönig et al. and Cao et al. used 180 and 80 samples, respectively, to evaluate IMPRESS [9], [33]. Each image took approximately 5-7

TABLE II

WHITE-BOX: COMPARISON OF IVORY AND DIFFPURE ON CNN (MOBILENETV3) AND VIT (SWINFACE)

Task	Model / Attack	MobileNetV3 Undef.	MobileNetV3 Adv. Train	MobileNetV3 DiffPure	MobileNetV3 IMPRESS (250)	MobileNetV3 IVORY	SwinFace Undef.	SwinFace DiffPure	SwinFace Ivory	Ave. Accuracy Gain DiffPure	Ave. Accuracy Gain Ivory
Gender	Clean	**97.24%**	96.11%	**97.24%**	**96.80%**	**97.24%**	**97.51%**	**97.51%**	**97.51%**	-	-
	PGD	54.96%	94.39%	**96.22%**	**95.60%**	**96.38%**	61.28%	95.25%	**97.09%**	37.62%	**38.62%**
	AA	53.15%	**91.87%**	**91.97%**	**92.00%**	**92.51%**	56.68%	92.46%	**93.14%**	**37.30%**	**37.91%**
	C&W	49.19%	90.26%	94.55%	94.80%	**96.52%**	43.97%	94.03%	**95.42%**	47.71%	**49.39%**
Ethnicity	Clean	**93.67%**	81.40%	**93.67%**	**93.20%**	**93.67%**	**95.22%**	**95.22%**	**95.22%**	-	-
	PGD	31.57%	83.81%	**90.72%**	88.00%	**91.68%**	42.08%	91.28%	**92.38%**	54.18%	**55.21%**
	AA	28.33%	84.07%	90.88%	**92.00%**	**92.05%**	30.73%	90.89%	**93.63%**	61.36%	**63.31%**
	C&W	26.76%	87.64%	**90.58%**	**90.40%**	**90.24%**	21.84%	87.26%	**90.71%**	64.62%	**66.18%**
Age	Clean	**54.62%**	47.29%	**54.62%**	52.80%	**54.62%**	**66.88%**	**66.88%**	**66.88%**	-	-
	PGD	3.74%	41.03%	49.80%	**51.20%**	**51.07%**	7.21%	61.82%	**64.53%**	50.34%	**52.33%**
	AA	2.14%	34.59%	**51.61%**	**51.60%**	**52.44%**	3.20%	59.35%	**63.92%**	52.81%	**55.51%**
	C&W	46.06%	31.94%	**52.02%**	**52.00%**	**52.65%**	29.16%	62.09%	**65.70%**	19.45%	**21.57%**
										(Ave.) 47.26%	**(Ave.) 48.89%**

TABLE III

BLACK-BOX: COMPARISON OF IVORY AND DIFFPURE ON CNN (MOVILENETV3) AND VIT(SWINFACE)

Task	Model / Attack	MobileNetV3 Undef.	MobileNetV3 DiffPure	MobileNetV3 IVORY	SwinFace Undef.	SwinFace DiffPure	SwinFace Ivory	Ave. Accuracy Gain DiffPure	Ave. Accuracy Gain Ivory
Gender	PGD	58.07%	93.87%	**95.21%**	91.09%	92.35%	**96.78%**	18.53%	**21.42%**
	AA	59.64%	91.64%	**92.39%**	89.67%	90.42%	**92.25%**	16.38%	**17.67%**
	C&W	51.22%	90.56%	**94.12%**	88.24%	91.74%	**94.83%**	21.42%	**24.75%**
Ethnicity	PGD	45.32%	88.73%	**91.24%**	86.47%	89.36%	**91.49%**	23.15%	**25.47%**
	AA	43.95%	88.60%	**91.32%**	85.89%	90.65%	**92.04%**	24.71%	**26.76%**
	C&W	38.09%	84.29%	**89.82%**	82.61%	85.20%	**90.15%**	24.40%	**29.64%**
Age	PGD	21.08%	46.03%	**49.12%**	51.98%	58.29%	**62.39%**	15.63%	**19.23%**
	AA	27.58%	47.89%	**51.74%**	49.68%	57.34%	**62.08%**	13.99%	**18.28%**
	C&W	31.93%	50.26%	**52.18%**	40.43%	56.43%	**64.19%**	17.17%	**22.01%**
								(Ave.) 19.48%	**(Ave.) 22.80%**

minutes for purification with our simplified implementation of IMPRESS, whereas IVORY completed purification within seconds (follow Sec VI-E for more details).

Across most settings, IVORY consistently restored model accuracy more effectively than adversarial training, DiffPure and IMPRESS. While IMPRESS maintained image fidelity well, it marginally reduced clean accuracy by approximately 1%, a behavior also observed by its original authors [9]. Among the nine task-attack settings, DiffPure outperformed IVORY in only one setting (Ethnicity Recognition, C&W attack), while IMPRESS surpassed IVORY in two settings (Ethnicity Recognition, C&W attack and Age Group Classification, PGD attack) as shown in Table II. However, the performance gains of these methods over IVORY were minimal, ranging from only 0.13% to 0.34%. In all remaining settings, IVORY achieved the highest accuracy, demonstrating its superior purification effectiveness.

For black-box testing, we evaluated MobileNetV3 and SwinFace using adversarial samples generated from VGG16 and SENet. While DiffPure and IVORY were evaluated on these black-box samples, running IMPRESS on this larger dataset was computationally infeasible due to its iterative nature. Thus, IMPRESS was excluded from black-box experiments.

As shown in Table III, adversarial perturbations are generally less effective in this setting, as the user does not know what model the attacker is using. Still, the user's tactic does meaningfully degrade accuracy, particularly on age group classification, where it reaches a low of 21.08% against PGD for MobileNetV3 and 40.43% against C&W for SwinFace. However, IVORY consistently outperformed DiffPure across all tasks and models with an average accuracy gain of 22.80%. Notably, in the Ethnicity Recognition task under C&W attack, IVORY achieved 90.15% accuracy, surpassing DiffPure's 85.2%. Overall, IVORY demonstrated substantial accuracy gains in both white-box and black-box settings, reaffirming its effectiveness across various adversarial defense scenarios.

D. SSIM Noise Analysis

SSIM provides a quantitative assessment of the perceptual similarity between an original clean image and a modified image and we can use it to measure both the level of adversarial perturbation added by the user and the noise remaining after the image purification process in IVORY and DiffPure. Lower SSIM noise values indicate higher similarity

and therefore less perceptible degradation in image quality.

Fig. 2 shows the amount of noise introduced by IVORY and DiffPure compared to adversarial examples. During our experiments we noticed that IVORY consistently introduces less noise compared to DiffPure, demonstrating its superior ability to restore image quality. For instance, in the case of AutoAttack, DiffPure (SDXL | 224) introduces an average noise of 0.071, whereas IVORY (SD3 | 224) introduces 0.0341. Although IVORY introduces slightly higher SSIM noise compared to adversarial samples, the values remain imperceptible to the human eye.

Additionally, we explored the "Just Resize" technique, available within SwarmUI, where the purified image is upscaled using a latent upscaler, to upscale the purified image by 2x. We observed further reduction of SSIM noise. For example, with PGD adversarial samples, the SSIM noise introduced by IVORY (SD3 | 224) reduces from 0.0331 to 0.0191 for the IVORY (SD3 + Just Resize | 448). These results suggest that upscaling the purified images serves as a viable strategy to enhance image quality without sacrificing purification effectiveness. Overall, IVORY reduces SSIM noise by over 50% at 1x resolution and up to 80% at 2x resolution compared to DiffPure.

Fig. 3 showcases some purified samples by IVORY and DiffPure. The adversarial images were generated using AutoAttack on MobileNetV3 model, which has the lowest SSIM. We observe that the adversarial images purified by DiffPure (SDXL) lack some fine details that IVORY (SD3) retains – faces in rows 2-4 are all smoother with DiffPure. Perhaps surprisingly, both IVORY and DiffPure keep odd patterns added by the adversarial perturbation (the background of the first and second rows, forehead of the third row, hat in the fourth row). It may be that these visible patterns cannot be treated as noise by the diffusion models, as they are visible and could be real image features, but the less visible perturbations may have been removed. In that sense, purification does not fully restore the original image. The fourth column in Fig. 3 illustrates the output from the combination of IVORY and Just Resize technique which further improves the quality of IVORY outputs. Overall, IVORY provides minimal image degradation.

E. Purification Time Analysis

To evaluate the performance of IVORY's purification time against DiffPure, we randomly selected 1000 samples from three different adversarial attack techniques: PGD, AutoAttack and C&W, all generated using MobileNetV3. The purification process was conducted on a machine equipped with an NVIDIA A100 GPU.

Table IV reports single-image purification times for IVORY and DiffPure across adversarial methods and diffusion timesteps t^*. As expected, inference time increases with t^*. At $t^* = 0.2$, IVORY takes 2.54s (PGD), 2.21s (AutoAttack) and 2.93s (C&W), slightly higher than DiffPure's 1.79s, 2.26s and 2.57s, respectively. This marginal difference ($\leq$ 0.5s) reflects SD3's larger model size (8B vs. 3.5B in SDXL) and more complex architecture, which contributes to

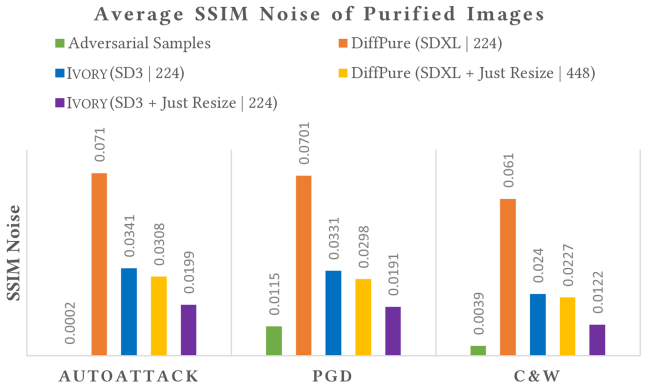

Fig. 2. SSIM noise introduced by DiffPure and IVORY across different adversarial samples and resolutions.

Fig. 3. Adversarial examples (second column) generated from clean images are purified by *DiffPure (SDXL)*,IVORY *(SD3)* and IVORY *(SD3) + Just Resize*. The examples purified by IVORY (fourth column) are relatively sharper and closer to the clean images compared to DiffPure (third column).

better purification. Compared to the 10–12s times originally reported in [47] using DDPM on a V100 GPU, our setup with SDXL on an A100 — leveraging Diffusers and PEFT (Parameter Efficient Fine-Tuning) libraries — achieves a 10× speedup (1.1–1.5s at $t^* = 0.1$), due to hardware improvements and diffuser codebase optimizations.

In contrast, IMPRESS required approximately 300–400 seconds to purify a single image, significantly longer than both IVORY and DiffPure. The original authors of IMPRESS reported an average purification time of 141 seconds per image [9]; however, in our experiments, we observed nearly double that time. We hypothesize that this discrepancy arises from the larger size of the underlying latent diffusion model's VAE (SDXL in our case), which may introduce additional computational overhead. Regardless, IVORY remains dramatically faster than IMPRESS, making it far more viable for real-time deployment.

While purification times can still be optimized, particu-

TABLE IV

INFERENCE TIME (IN SECONDS) FOR IVORY AND DIFFPURE ACROSS DIFFERENT ADVERSARIAL SAMPLES AND DIFFUSION TIMESTEPS t^*

Adv. Method	IVORY $t^*=0$	DiffPure $t^*=0$	IVORY $t^*=0.1$	DiffPure $t^*=0.1$	IVORY $t^*=0.2$	DiffPure $t^*=0.2$
PGD	0.34	0.32	1.26	1.18	2.54	1.79
AutoAttack	0.51	0.52	1.34	1.23	2.21	2.26
C&W	0.87	0.79	1.78	1.59	2.93	2.57

larly by exploring techniques like LoRAs with LCM, we have focused on demonstrating the practical performance of IVORY. Our results show that, despite being slightly slower than DiffPure, IVORY's inference times remain competitive, especially given the more computationally complex attacks like C&W. This solidifies IVORY 's feasibility for real-world deployment, where slight increases in computation time are offset by its superior purification capabilities. Further improvement and optimization of the purification time is left for future work.

VII. DISCUSSION AND BROADER IMPACT

IVORY *outperforms SDXL-based methods.* From our experimental results, it is evident that IVORY consistently surpasses traditional adversarial defenses such as adversarial training, as well as recent purification methods like IMPRESS and DiffPure. The primary distinction between these methods lies in the underlying diffusion model architecture. SD3 differs significantly from SDXL, leveraging its transformer-based architecture for image generation, which employs separate weight sets for text and image modalities. This enables a bidirectional flow of information between image and text tokens, improving overall image generation [23]. In contrast, SDXL relies on a simpler U-Net-based architecture. Another crucial difference between the two models is their Variational AutoEncoder (VAE). SD3-VAE operates with a 16-channel latent space, significantly improving image detail retention compared to the 4-channel latent space of SDXL-VAE [1], [2]. Essentially, SD3's VAE is designed to capture a richer latent representation, leading to higher-quality reconstructions. In the context of adversarial purification, we hypothesize that SD3-VAE's higher latent capacity enables more effective reconstruction of obfuscated images, which could explain the observed performance improvements over SDXL-based purification methods. Unlike SDXL's VAE, which applies 48x compression, SD3-VAE uses a more modest 12x compression, ensuring better preservation of critical details. These architectural advancements contribute to IVORY 's superior purification performance.

Limitations of Adversarial Perturbations in Protecting Soft-Biometrics. Our experiments reveal that advanced adversarial purification techniques, such as IVORY, using off the shelf Diffusion Transformer models like SD3, could easily remove adversarial perturbations from obfuscated images. Our results further prove that these perturbation methods do not offer effective protection against soft-biometric extraction. The fundamental issue lies in the fact that while adversarial perturbations can obscure soft-biometric features to some extent, they do not entirely shield them from extraction methods designed to work even with perturbed data. This limitation emphasizes that adversarial defenses should not be relied upon for privacy preservation in soft-biometric extraction systems.

The Need for Rigorous Evaluation of Protection Methods. Our findings highlight the need for rigorous and adaptive evaluations of protection methods. Similar to how adversarial defenses in machine learning have often been found vulnerable to adaptive attacks [15], [63], protections against soft-biometric extraction are similarly susceptible to breaches if not thoroughly tested. This observation underscores the importance of continuously evolving evaluation strategies to keep pace with advancements in both attack and defense methodologies. This issue is exacerbated when protections are widely publicized without sufficient validation, potentially leading to a false sense of security among users.

Limitations and Future Work. Our study has several limitations that warrant further investigation. First, the evaluation was conducted with a limited number of adversarial samples and defense techniques. Expanding the study to include a broader range of adversarial attacks and defensive methods could provide a more comprehensive understanding of their effectiveness and limitations. Additionally, future work should explore the application of faster sampling techniques, such as Low-Rank Adaptation (LoRA) with Latent Consistency Models (LCM), to enhance the efficiency of diffusion-based purification methods. Moreover, while IVORY shows promise, there is room for improvement in reducing purification time and SSIM noise. Alternative architectures and optimization strategies could lead to more effective and efficient methods. It is also important to investigate the potential of countering IVORY with better privacy-preserving techniques to achieve more comprehensive protection against soft-biometric extraction.

VIII. CONCLUSIONS

We presented IVORY, an adversarial purification technique using diffusion transformers to enhance robustness against adversarial perturbations. IVORY effectively mitigates adversarial noise and restores classifier performance across tasks like gender, ethnicity and age group classification, consistently outperforming adversarial training, DiffPure and IMPRESS. For example, IVORY restored an average of 55%-66% accuracy on ethnicity recognition in the white-box setting. In the more realistic black-box setting, it restored an average of 17%-29% in age group classification accuracy. Our findings highlight that adversarial perturbations alone are insufficient for protecting soft-biometric data, indicating the need for better privacy-protecting techniques to help users.

IX. ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation under Awards No. 2040209, 2422241 and 2429835.

ETHICAL IMPACT STATEMENT

IVORY is an attack against users who would attempt to protect their privacy by adding adversarial perturbations to their images. It might also be adapted to remove perturbations from artwork, undoing efforts meant to protect artists from style copying. Some may argue that work on attacks is unethical, but we strongly disagree. If adversarial examples are found to be ineffective at protecting users, then more effective approaches must be developed. Furthermore, if anyone is relying on these insecure approaches for protection, they need to be made aware of the vulnerability as soon as possible so they can take down the images [13].

REFERENCES

[1] S. AI. Sd3/vae/config.json · stabilityai/stable-diffusion-3 at main. 8

[2] S. AI. Sdxl/vae/config.json · stabilityai/stable-diffusion-xl-base-1.0 at main. 8

[3] L. L. Ankile, A. Midgley, and S. Weisshaar. Denoising diffusion probabilistic models as a defense against adversarial attacks. *arXiv preprint arXiv:2301.06871*, 2023. 2

[4] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018. 2

[5] G. Azzopardi, A. Greco, A. Saggese, and M. Vento. Fusion of domain-specific and trainable features for gender recognition from face images. *IEEE Access*, 6:24171–24183, 2018. 3

[6] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 2

[7] F. Becerra-Riera, A. Morales-González, and H. Méndez-Vázquez. A survey on facial soft biometrics for video surveillance and forensic applications. *Artificial Intelligence Review*, 52(2):1155–1187, 2019. 1

[8] B. T. Bell. "you take fifty photos, delete forty nine and use one": A qualitative study of adolescent image-sharing practices on social media. *International Journal of Child-Computer Interaction*, 20:64–71, 2019. 1

[9] B. Cao, C. Li, T. Wang, J. Jia, B. Li, and J. Chen. IMPRESS: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative AI. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5, 6, 7

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 3, 4

[11] V. Carletti, P. Foggia, A. Greco, A. Saggese, and M. Vento. Facial soft-biometrics obfuscation through adversarial attacks. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 1, 3, 4

[12] V. Carletti, A. Greco, G. Percannella, and M. Vento. Age from faces in the deep learning revolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2113–2132, 2019. 3, 4

[13] N. Carlini. Why i attack. https://nicholas.carlini.com/writing/2024/why-i-attack.html, 2024. Accessed: 2024-09-27. 9

[14] N. Carlini, F. Tramer, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter. (certified!!) adversarial robustness for free! *ICLR*, 2023. 2

[15] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 8

[16] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 2, 3

[17] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2

[18] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 2

[19] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 13–18 Jul 2020. 2, 3, 4

[20] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2015. 3

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 3

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[23] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 5, 8

[24] K. Ghazinour and J. Ponchak. Hidden privacy risks in sharing pictures on social media. *Procedia computer science*, 113:267–272, 2017. 1

[25] Y. Gong, L. Huang, and L. Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4322, 2022. 2

[26] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 2

[27] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021. 2

[28] W. Grathwohl, K.-C. J. Wang, J.-H. Jacobsen, D. K. Duvenaud, M. Norouzi, and K. Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ICLR*, abs/1912.03263, 2019. 2

[29] A. Greco, G. Percannella, M. Vento, and V. Vigilante. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, 31:1–13, 2020. 3, 4

[30] A. Greco, A. Saggese, M. Vento, and V. Vigilante. Effective training of convolutional neural networks for age estimation based on knowledge distillation. *Neural Computing and Applications*, pages 1–16, 2022. 3, 4

[31] M. Hill, J. Mitchell, and S.-C. Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. *ICLR*, 2020. 2

[32] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[33] R. Hönig, J. Rando, N. Carlini, and F. Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024. 2, 5

[34] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 3, 4

[35] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[36] A. K. Jain, S. C. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? In *Biometric technology for human identification*, volume 5404, pages 561–572. Spie, 2004. 1

[37] G. Kim, T. Kwon, and J. C. Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[38] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 3

[39] Y. Li and Y. Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of marketing research*, 57(1):1–19, 2020. 1

[40] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023. 1

[41] W.-A. Lin, C. P. Lau, A. Levine, R. Chellappa, and S. Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *Advances in Neural Information Processing Systems*, 33:3487–3498, 2020. 2

[42] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021. 1

[43] Y. Liu, W. Zhang, and N. Yu. Protecting privacy in shared photos via adversarial examples based stealth. *Security and Communication Networks*, 2017(1):1897438, 2017. 1

[44] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017. 2, 3, 4

[45] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[46] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. 2

[47] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022. 1, 2, 5, 7

[48] M. S. Nixon, P. L. Correia, K. Nasrollahi, T. B. Moeslund, A. Hadid, and M. Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 68:218–230, 2015. 1

[49] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The 12th International Conference on Learning Representations*, 2024. 2, 5

[50] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[51] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021. 2

[52] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. *Handbook of statistics*, 31:327–352, 2013. 1

[53] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ArXiv*, abs/1805.06605, 2018. 2

[54] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial diffusion distillation. *ArXiv*, abs/2311.17042, 2023. 2

[55] S. A. Shahriyar and M. Wright. Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*, WDC '22 @ ASIA CCS, page 13–18, New York, NY, USA, 2022. Association for Computing Machinery. 3

[56] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 1

[57] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020. 1

[58] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh. On the adversarial robustness of vision transformers, 2022. 5

[59] Y. Shi, M. Du, X. Wu, Z. Guan, J. Sun, and N. Liu. Black-box backdoor defense via zero-shot image purification. *Advances in Neural Information Processing Systems*, 36:57336–57366, 2023. 2

[60] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[61] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ArXiv*, abs/1710.10766, 2017. 2

[62] S. Thys, W. Van Ranst, and T. Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[63] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020. 8

[64] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 1

[65] J. Wang, Z. Lyu, D. Lin, B. Dai, and H. Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 2

[66] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[67] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263. PMLR, 2023. 2

[68] C. Yu, B. Han, M. Gong, L. Shen, S. Ge, B. Du, and T. Liu. Robust weight perturbation for adversarial training. *arXiv preprint arXiv:2205.14826*, 2022. 2