

# DeepFakeShield: A Proactive Defense Against Malicious Face Swapping

Saeed Karimi-Bidhendi, Joseph DeGol, Eric Wengrowski, Dominic Roberts, Kristin Dana  
Steg.AI Corporation

{saeed.karimi, joe, eric, dominic.roberts, kristin}@steg.ai

## Abstract

Face-swapping algorithms enable malicious actors to spread misinformation and invade personal privacy and security by inserting victims' faces into images and videos without consent. Existing countermeasures are easily defeated and unfavorably trade perceptual quality with protection efficacy. We present DeepFakeShield (DFS), a proactive method that thwarts face-swapping algorithms by adding adversarial noise that disrupts facial landmark detection. DFS's primary contributions are (1) jointly optimizing both perceptual quality and protection efficacy, and (2) introducing a mutation layer during training to make the shield robust against common removal attacks such as blurring and JPEG compression. Using both qualitative and quantitative results, we show that DFS optimally trades perceptual quality with protection efficacy, remaining robust to image removal attacks and outperforming prior approaches. With ablation studies, we also show that DFS generalizes to out-of-domain face images and that both the joint optimization and mutation layer contribute significantly to our state-of-the-art results.

## 1. Introduction

In this paper, we present DeepFakeShield (DFS), a proactive method that protects human face images from being deepfaked with face-swapping (Figure 1). Cutting edge face-swapping algorithms enable malicious actors to generate high-fidelity, convincing deepfake photos and videos of victims without their consent [3]; with negative outcomes including personal privacy violations [23, 44], identity theft [27], misinformation [35], and security risks in facial recognition systems [28, 42]. Like other proactive approaches, DFS thwarts deepfakes by adding imperceptible adversarial noise to the input images. However, DFS differs in that it disrupts the alignment step in face-swappers by fooling the facial landmark detectors into mislocalizing key facial features such as the mouth, nose, and eyes. Compared to other proactive approaches, DFS represents a major leap towards practical adoption because DFS produces shields

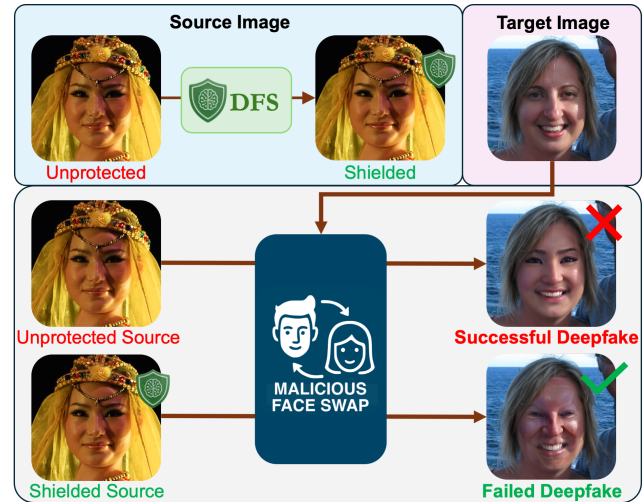


Figure 1. Top-left: DFS consumes source images and outputs perceptually-similar shielded images. Top-right: The target image. Bottom: When malicious actors attempt to generate deepfakes by swapping shielded source faces onto target faces, the output face appears distorted, resulting in a failed deepfake.

with fewer perceptible artifacts with greater robustness to mutations like compression and blur.

DFS achieves more potent shields with fewer perceptible artifacts because of a novel multi-objective loss that is jointly optimized to balance perceptual quality of the shield with efficacy of landmark displacement. This is important because noticeable artifacts will dissuade people from using proactive deepfake protection measures on their photos. Previous approaches (e.g., [16, 47]) use a budget to constrain the perceptibility of artifacts in their shields. The problem with budget-based approaches is that the optimal budget differs for every image based on the textures and colors in that image, and a fixed budget artificially constrains the network from finding the optimal balance. With our approach, we leverage state-of-the-art models of human perception to penalize DFS when artifacts become too obvious; thereby, allowing DFS to find the best balance of landmark displacement efficacy and imperceptibility.

DFS shields survive image mutations like compression

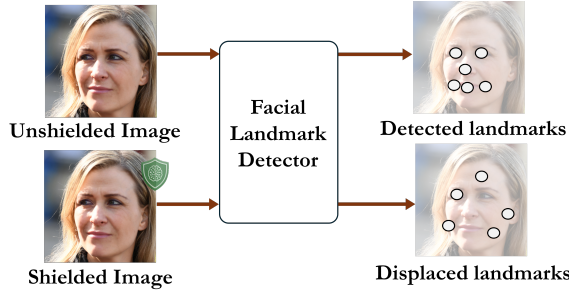


Figure 2. DFS breaks face-swapping algorithms by causing their landmark detectors to mislocalize the detected landmarks.

and blur because we introduce a mutation layer into the training pipeline. This is important because the shield loses effectiveness if an attacker can circumvent the protection using simple image manipulations or filtering. We train DFS by adding facial landmark detector(s) into the training loop and encouraging DFS to produce adversarial noise that moves detected facial landmarks away from their ground truth positions. During training, shielded images are mutated before being processed by facial landmark detectors. Thus, DFS learns to produce shields that are capable of disrupting facial landmark detectors while remaining robust to image mutations.

To demonstrate the value of our approach, we evaluate both the perceptual quality and shield efficacy with qualitative and quantitative experiments. First, we evaluate on 5000 face pairs from the test set of the commonly used FFHQ dataset and show that our shield is less perceptible and more effective at breaking deepfakes than two cutting edge baseline approaches. We then also perform two ablation studies to demonstrate the importance of the multi-objective loss and mutation layers. Finally, we test on the out-of-domain CelebA-HQ dataset to show that the method generalizes beyond the FFHQ dataset.

In summary, the **contributions** of this paper are (1) DFS, a proactive protection method for shielding images against face-swapping that uses a novel multi-objective loss for jointly optimizing the perceptual quality of the shield and landmark displacement; (2) a mutation layer in the training pipeline that enables DFS to generate structured noise that is robust to image mutations; and (3) comprehensive qualitative and quantitative experiments that demonstrate cutting edge efficacy of DFS at breaking face-swappers without sacrificing the quality of the shielded image.

## 2. Related Work

**Proactive Protection Against Deepfakes:** Proactive methods modify images such that deepfake algorithms will produce poor results if using these protected images as input at inference time. Protection against deepfakes is achieved by crafting and adding imperceptible, targeted

perturbations to source images [4, 16, 37, 38]. Gradient-based disruptions in white-box settings have been explored in [30, 37, 38], while finding appropriate surrogate models for face-swapping is studied in [12, 21] for the black-box scenarios. Model-agnostic methods, such as CMUA-watermark [11, 20, 30] and Anti-Forgery [8, 41], are developed to maintain effectiveness against various models. Improving generalization remains an open problem, with techniques such as model ensembling [4, 10], layer-specific attacks [49], and frequency-domain perturbations [31, 57] proposed to this end. By constraining perturbations to important facial regions, the work in [29] and [36] aims to balance protection effectiveness with imperceptibility. A recent method, FaceShield [22], focuses on preventing deepfakes generated from prompting diffusion models. We choose to focus on face-swappers and show in Section 4 that DFS yields superior performance against face-swappers because it specifically crafts perturbations that target the landmark detector component, while also introducing a novel approach to handle image mutations and jointly optimize for both perceptual quality and protection efficacy.

Data poisoning (e.g., Nightshade, [39]; Glaze, [40]) is a related proactive protection approach, but rather than targeting inference operations, Nightshade and Glaze modify images with the intent of foiling methods during fine-tuning/training. While similar, these methods are distinct from the deepfake-oriented methods detailed in this section.

**Reactive Defenses Against Deepfakes:** Reactive deepfake detection often involves classification to identify fake content after creation [34]. Early detection methods exploit spatial artifacts such as pixel-level inconsistencies [1], texture differences [36], and blending boundaries [26, 57]. Frequency-domain analysis can reveal the innate characteristics of GAN-synthesized images [4, 32, 43]. Some methods exploit biological signals, such as inconsistent eye blinking [30], discrepancies in head poses [52], and lip movement [18, 51], for deepfake detection. Temporal inconsistencies have been exploited in [14, 15, 33] for video deepfake detection.

Despite recent advances, passive methods often learn to detect artifacts specific to the deepfake generation method they are trained on instead of universal forgery patterns [9]. Therefore, they do not generalize well against unseen deepfake generation methods [9, 41]. Contrastive learning [56], dynamic augmentation [46], and continual learning [32, 49] are proposed to improve generalization; however, all are vulnerable to adversarial attacks [1, 12] and their performance on datasets such as DFDC [8, 41] remains subpar. There is an important fundamental downside of passive approaches: they can not prevent deepfakes nor undo the damage they have caused after creation. That is why proactive approaches like DFS are crucial.

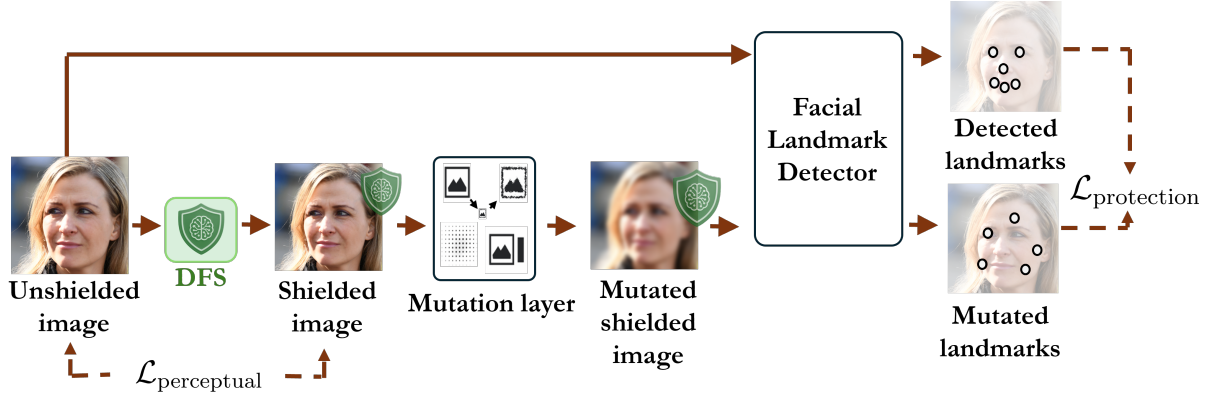


Figure 3. DFS Training pipeline. The forward pass begins with an unshielded training image. This image is shielded via DFS. Then the mutation layer optionally applies mutations. Then landmarks are detected on both the mutated shielded image and the unshielded image. Lastly, joint optimization computes one step towards minimizing the perceptual and protection losses.

### 3. Methodology

Figure 3 outlines the overall framework that was used to train DFS. The core idea is that DFS is trained in a feedback loop with facial landmark detectors, and two losses incentivize (1) spoiling the outputs of the facial landmark detectors and (2) making the perceptual quality of the shielded image high. A mutation layer is placed between DFS and the facial landmark detector which encourages DFS to produce shields that still spoil facial landmark detectors despite image mutations like compression and blur. The remainder of this section details the facial landmark detector, DFS architecture, losses, and mutation layer.

#### 3.1. Facial Landmark Detector

**Face-Swapper Components:** Face-swappers are typically built as three-stage pipelines [45]: face detection, alignment, and identity fusion. (1) The face detection stage outputs bounding boxes that tightly localize faces in an image. Since downstream modules operate in the selected regions, accurate face detection is crucial to maintain consistency and efficiency in the face-swapping pipeline. Modern face detectors, such as MTCNN [53] and RetinaFace [6], detect with high precision while exhibiting robustness to pose, illumination, and occlusion. (2) The alignment stage extracts information on the geometry and expression of a face using a landmark detector that identifies key facial points around the eyes, nose, mouth, and jawline [17, 48]. This information is used to align the source and target images, which is critical for the source face identity to fit naturally to the target face. Without proper alignment, the face-swapper output is distorted due to geometric mismatches between the source and target faces. (3) The identity fusion stage performs the “swap” by transferring the source’s identity features to the target while maintaining realistic expressions, lighting, and orientation. Recent methods use encoder–decoder architectures, GANs,

or diffusion-based models to seamlessly blend the identity of the source with the target face [2]. In principle, adversarial attacks can be carried out against one or more of the aforementioned three key stages of face-swappers to hinder deepfake generation. Previous work focuses mostly on identity fusion; whereas, in this work, we focus specifically on the landmark detector, as described below.

**Attacking the Facial Landmark Detectors:** Of the various components of face-swapper algorithms, we found that facial landmark detectors were an ideal weak point to attack because (1) alignment using facial landmark detectors is a critical and standard component of face-swappers, so this approach will work for many different existing face-swappers; and (2) the alignment process is highly sensitive to errors such that even small errors in the alignment matrix will manifest as noticeable artifacts in the face swap.

Figure 2 provides a depiction of the shield disrupting a facial landmark detector. The facial landmark detector consumes the mutated, shielded image as input and generates an estimate of the 2D coordinates of its  $K$  facial landmarks. Here,  $K$  represents the number of landmark points that we aim to attack. During training, the shield network is optimized to produce a perturbation that—even after mutation—misleads the facial landmark detector about the true locations of key facial points.

#### 3.2. DeepFakeShield (DFS) Architecture

DFS protects images from face-swapping by attacking the alignment component of face-swappers as described in Section 3.1. The core idea is to generate imperceptible additive perturbations that disrupt the alignment by preventing seamless integration of facial features from the source face to the target face. Since the alignment heavily depends on the correct prediction of key facial landmarks, DFS adversarially attacks the facial landmark detector.

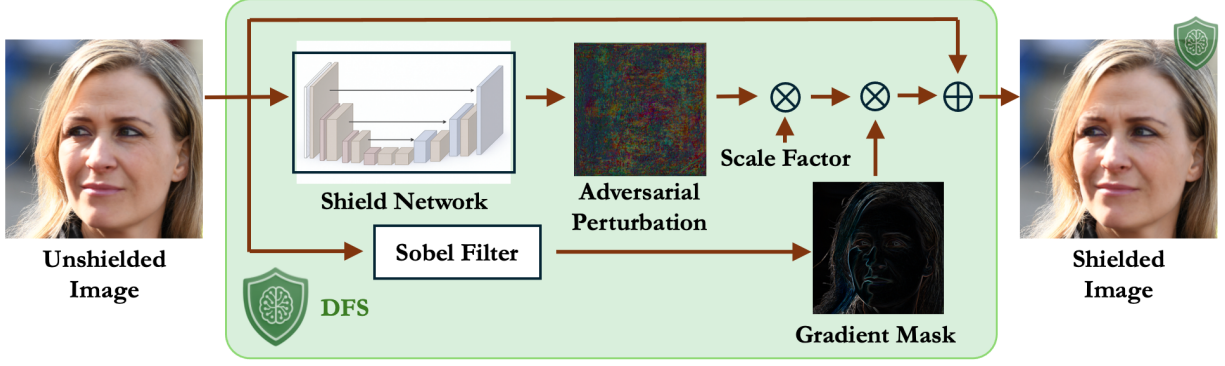


Figure 4. Shield Architecture. The shield network uses a U-Net architecture with a pretrained ResNet-34 backbone as the encoder. The gradient mask enhances structured noise at edges to increase its imperceptibility when added to the source image.

Figure 4 shows the design of DFS, which consists of three components: shield network, scaling factor, and Sobel filter.

**Shield Network:** The shield network is based on the U-Net architecture with the ResNet-34 backbone as the encoder, pre-trained on the ImageNet dataset to provide strong initial feature representations. The network processes standard three-channel RGB inputs and uses five encoding blocks to capture hierarchical features on multiple scales. The decoder mirrors the encoder structure and uses decreasing channel dimensions of 256, 128, 64, 32, and 16 to gradually reconstruct the spatial resolution and generate a three-channel output with the same dimensions as the input. Skip connections are made between the corresponding encoder and decoder layers to improve gradient flow and mitigate vanishing gradient problems. Batch normalization is used to encourage training stability and convergence. Finally, tanh is used as the final activation function, ensuring the output remains within  $[-1, 1]$ .

**Scaling Factor:** The output of the shield network is multiplied by a scalar hyperparameter  $0 < \delta < 1$  to constrain the perturbation value at each pixel to  $[-\delta, \delta]$ . Although this procedure resembles imposing a hard budget on the magnitude of additive perturbations, as is commonly done in previous studies, it represents only one of several mechanisms that we use to enhance the imperceptibility of the perturbation. The value of  $\delta$  is a design factor and depends on the robustness of the particular facial-landmark detector that the network tries to attack, with more robust detectors often requiring larger values.

**Sobel Filter:** We further improve the perceptual quality of shielded images via the Sobel filter, an image processing operation that identifies areas of high-frequency content, such as edges and highly textured regions. Since distortions are more perceptually discernible in flatter regions [19], we multiply the additive perturbations from the previous step

by the gradient mask derived from Sobel-filtering and stack the source image’s individual RGB channels separately. This suppresses perturbations in flat areas, such as the forehead and cheeks, enhancing the perceptual quality of the shielded image.

**Shielding an Image:** Mathematically, we define the shielding of an image as

$$x_p = \text{clip}(x_s + SN(x_s) \cdot \delta \cdot SF(x_s)), \quad (1)$$

where the source and shielded images are denoted by  $x_s$  and  $x_p$  respectively, the shield network and Sobel filter are denoted by  $SN(\cdot)$  and  $SF(\cdot)$  respectively, and  $\text{clip}(\cdot)$  ensures a valid pixel range  $[0, 1]$  for the shielded image.

### 3.3. Mutation Layer

The shielded image, after being distributed, may be subjected to edits such as compression and blurring; it is crucial for it to maintain its protection efficacy after these edits. Thus, we introduce the mutation layer — an operation that alters the shielded image via a random set of image transformations, simulating real-world alterations and improving robustness to such changes. Within the mutation layer, any differentiable mutation can be applied; making it practical and adaptable to many application-specific mutations. We denote the mutated, shielded image by  $\tilde{x}_p = M(x_p)$  where  $M(\cdot)$  represents the mutation layer.

### 3.4. Loss Functions

The shielded image generated by DFS should (1) be as perceptually faithful as possible to the source image; and (2) disrupt the locations of facial landmarks even after mutation. We introduce the following loss components to achieve these two properties.

**Perceptual Quality:** Our perceptual loss consists of a weighted combination of MSE, LPIPS [55], and DISTS [7]:



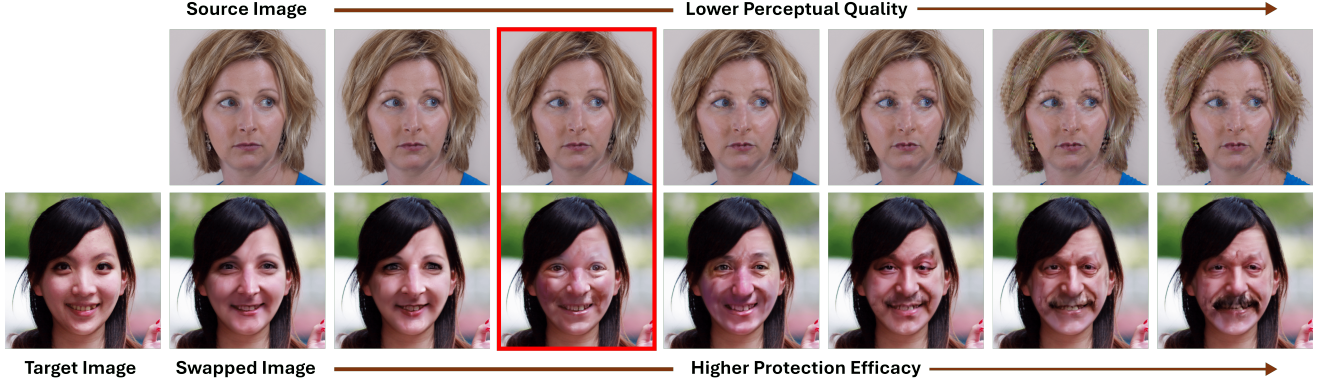


Figure 5. Illustration of the tradeoff between perceptual quality and protection efficacy. Shielded and swapped images at various iterations in the second stage of DFS training are shown. Moving from left to right, the shield perceptual quality worsens, but the protection efficacy improves. The pair highlighted in red is a qualitative example of the model chosen by our stopping criteria.

$$\begin{aligned} \mathcal{L}_{\text{perceptual}}(x_s, x_p) = & \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}}(x_s, x_p) \\ & + \lambda_{\text{LPIPS}} \cdot \mathcal{L}_{\text{LPIPS}}(x_s, x_p) \\ & + \lambda_{\text{DISTS}} \cdot \mathcal{L}_{\text{DISTS}}(x_s, x_p), \end{aligned} \quad (2)$$

where  $\lambda_{\text{MSE}}$ ,  $\lambda_{\text{LPIPS}}$ , and  $\lambda_{\text{DISTS}}$  are scalar hyperparameters.

**Protection Efficacy:** We additionally optimize for a protection loss, denoted by  $\mathcal{L}_{\text{protection}}(x_s, \tilde{x}_p)$ , to ensure that the estimated facial landmarks of the mutated shielded image and those of the source image are misaligned. The specific form of  $\mathcal{L}_{\text{protection}}(x_s, \tilde{x}_p)$  can be tailored to the facial landmark detector being targeted, as shown in Section 4.

The total loss function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{perceptual}} \cdot \mathcal{L}_{\text{perceptual}} + \mathcal{L}_{\text{protection}}, \quad (3)$$

where  $\lambda_{\text{perceptual}}$  is a scalar hyperparameter.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets:** We conduct our experiments on the FFHQ dataset [25], which consists of 70K high-quality human face images at a resolution of  $1024 \times 1024$ . We bilinearly interpolate each image to a resolution of  $512 \times 512$  and randomly split the dataset into training, validation, and test sets containing 55K, 5K, and 10K images, respectively. The test set is split into 5K pairs of source and target images.

**Baselines:** As discussed in Section 2, several proactive protection methods have been proposed. However, these methods are either not designed to address the same problem as DeepFakeShield or lack publicly available code and implementations. To enable a meaningful comparison, we implement (1) Adversarial Robust Safeguard (ARS)

[16] and (2) FaceSwapGuard (FSG) [47]. ARS protects images from a manipulation model by training its outputs to resemble a monochromatic white image. As simulated in [16], we set the manipulation model to SimSwap [2] and implement ARS according to their experimental setup. FSG protects images by adding imperceptible adversarial perturbations that disrupt the identity fusion stage of deepfake swapping systems. Although different  $\epsilon$ -budgets are considered in ARS [16] and FSG [47], we set  $\epsilon = 0.03$  for both methods to ensure a fair comparison.

**Training Procedure:** Training follows a two-stage protocol. In the first stage, we set  $\lambda_{\text{perceptual}} = 0$  and optimize only  $\mathcal{L}_{\text{protection}}$ . We train the first stage for 75 epochs. At the end of the first stage, the shield network generates the strongest adversarial perturbations. During the second stage, the network gradually shifts toward more moderate perturbations to enhance the perceptual quality of the shielded image, at the cost of reduced protection effectiveness. The choice of the optimal shield network during the second stage is application-specific and depends on the desired trade-off between the perceptual quality and protection efficacy. For this paper, we gradually increase  $\lambda_{\text{perceptual}}$  by 0.02 every 1000 iterations until both LPIPS and DISTS are below 0.03. We choose these stopping criteria to strike a balance between protection efficacy and perceptual quality. We illustrate this trade-off in Figure 5 and highlight a qualitative example corresponding to LPIPS and DISTS meeting the stopping criteria of 0.03.

**Metrics:** To quantitatively evaluate perceptual quality and protection efficacy, we use standard metrics for comparing two images: LPIPS, DISTS, Identity loss (ID)<sup>1</sup>, mean

<sup>1</sup>Identity loss between two images  $I_1$  and  $I_2$  is defined as  $\text{ID}(I_1, I_2) = 1 - \cos(\text{id}_1, \text{id}_2)$  where  $\text{id}_1$  and  $\text{id}_2$  are identity embeddings of  $I_1$  and  $I_2$ , respectively, extracted via ArcFace [5].

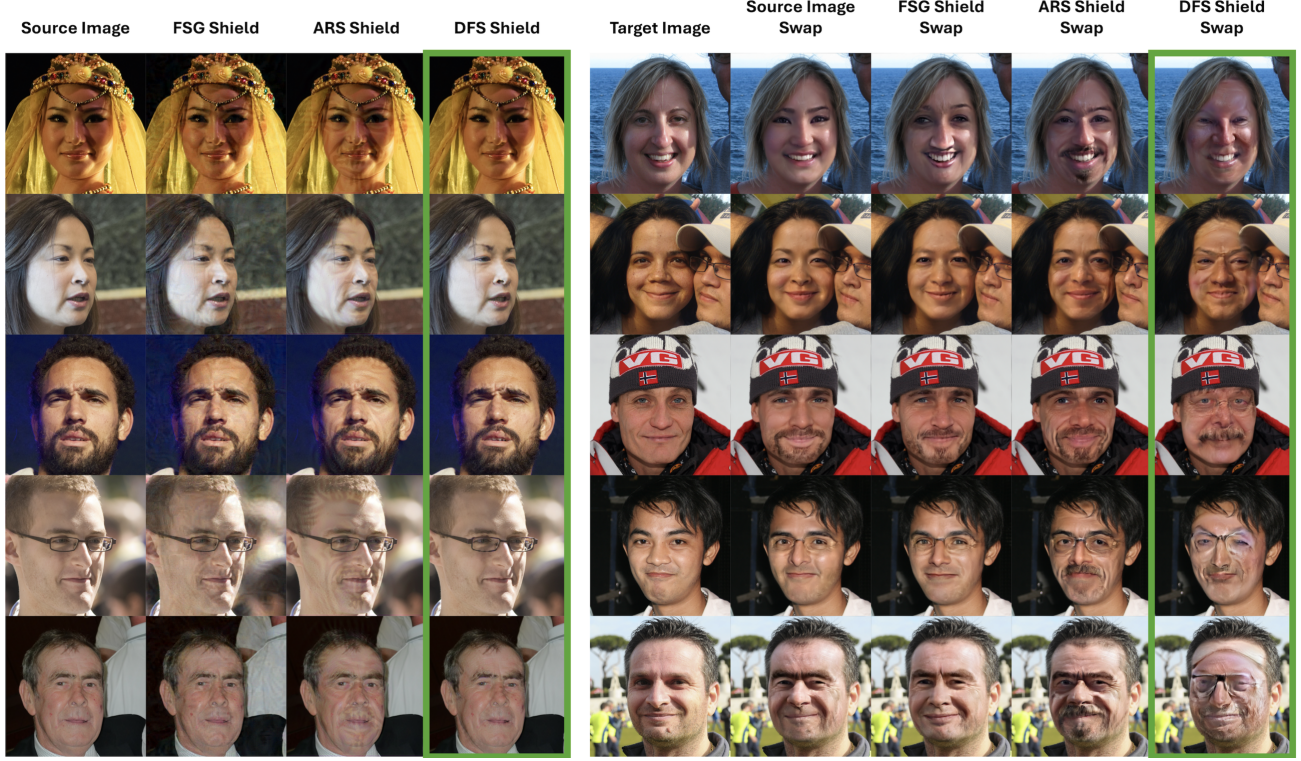


Figure 6. Qualitative comparison of FSG, ARS, and our (DFS) method. Left: source images and shielded source images by FSG, ARS and DFS (ours, highlighted in green). Right: from left to right, the target image, the target image face-swapped with the source, and each of the shielded versions face-swapped with the target (DFS highlighted in green). The columns highlighted in green show that DFS achieves the best perceptual quality and most drastic distortions to the deepfake outputs (zoom in for best viewing).

square error (MSE), SSIM, and PSNR. In the case of perceptual quality, scores are computed between the source image and each method’s shielded version. Ideal scores indicate no changes between the source image and the shielded version (i.e., lower LPIPS, DISTS, ID, and MSE scores, and higher SSIM and PSNR scores). In the case of protection efficacy, scores are computed between the source image swap and each method’s swap. Ideal scores indicate large changes between the source image swap and the shielded swap (i.e., higher LPIPS, DISTS, ID, and MSE scores, and lower SSIM and PSNR scores).

**Implementation Details:** Our framework can accommodate any facial landmark detector in its training pipeline (see Figure 3). In this work, we adopt SynergyNet [48] as the facial landmark detector and pair it with three face-swapping systems: SimSwap [2], MobileFaceSwap [50], and Ghost [13]. Due to space limitations, analyses of additional facial landmark detector choices are provided in the supplementary material. Given an image  $x$ , SynergyNet generates a latent vector that encodes pose, shape, and expression information, from which 68 facial landmarks are extracted. We focus our adversarial attack on a subset of  $K = 25$  landmark points around eyes, tip of the nose, and corners

of the mouth. With  $\alpha_s$  and  $\tilde{\alpha}_p$  denoting the latent vectors for the source image  $x_s$  and the mutated shielded image  $\tilde{x}_p$ , respectively, we define  $\mathcal{L}_{\text{protection}}$  in Eq. (3) as

$$\mathcal{L}_{\text{protection}} = -\lambda_{\text{protection-MSE}} \cdot \mathcal{L}_{\text{MSE}}(\alpha_s, \tilde{\alpha}_p). \quad (4)$$

Adversarial perturbations are susceptible to low-pass filtering [54, 58]. Thus, we design the mutation layer in Figure 3 to include JPEG compression with quality factors  $\in [50, 95]$ , as well as Gaussian blurring with kernel sizes  $\in [9, 21]$  and standard deviations  $\in [1, 4]$ . During each training iteration, each mutation is applied with  $p = 0.5$  probability. All models are trained using an NVIDIA L4 GPU. The training batch size is set to 8, the scaling factor is set to  $\delta = 0.25$ , and the Adam optimizer with learning rate  $5e-4$  is used to fine-tune the weights of the shield network. We set  $\lambda_{\text{MSE}} = 10$  and  $\lambda_{\text{LPIPS}} = \lambda_{\text{DISTS}} = \lambda_{\text{protection-MSE}} = 1$ . On an NVIDIA GeForce RTX 4070 Laptop GPU, inference takes 32.3 ms per image, including data transfer and I/O.

## 4.2. Qualitative Comparison Against Baselines

Figure 6 shows qualitative results of both perceptual quality and protection efficacy of FSG, ARS, and our (DFS) approach. The left four columns show the source images and



| Method | Perceptual Quality |               |               |                |               |               |
|--------|--------------------|---------------|---------------|----------------|---------------|---------------|
|        | LPIPS↓             | DISTS↓        | ID↓           | MSE↓           | SSIM↑         | PSNR↑         |
| FSG    | 0.1213             | 0.1335        | 1.8846        | 5.67e-4        | 0.8864        | 32.493        |
| ARS    | 0.0329             | 0.0517        | 0.5492        | 2.13e-4        | 0.9778        | <b>36.768</b> |
| DFS    | <b>0.0253</b>      | <b>0.0309</b> | <b>0.0499</b> | <b>2.10e-4</b> | <b>0.9789</b> | 36.739        |

| Method | Protection Efficacy (SimSwap) |               |               |                 |               |               |
|--------|-------------------------------|---------------|---------------|-----------------|---------------|---------------|
|        | LPIPS↑                        | DISTS↑        | ID↑           | MSE↑            | SSIM↓         | PSNR↓         |
| FSG    | 0.0450                        | 0.0288        | <b>0.7088</b> | 13.35e-4        | 0.9459        | 31.076        |
| ARS    | 0.0577                        | 0.0389        | 0.5401        | 20.57e-4        | 0.9334        | <b>29.322</b> |
| DFS    | <b>0.0659</b>                 | <b>0.0391</b> | 0.6013        | <b>24.42e-4</b> | <b>0.9311</b> | 29.725        |

| Method | Protection Efficacy (MobileFaceSwap) |               |               |                |               |               |
|--------|--------------------------------------|---------------|---------------|----------------|---------------|---------------|
|        | LPIPS↑                               | DISTS↑        | ID↑           | MSE↑           | SSIM↓         | PSNR↓         |
| FSG    | 0.0102                               | 0.0134        | 0.3341        | 2.27e-4        | 0.9883        | 37.075        |
| ARS    | 0.0141                               | 0.0174        | 0.3689        | 3.41e-4        | 0.9846        | 35.428        |
| DFS    | <b>0.0243</b>                        | <b>0.0223</b> | <b>0.5713</b> | <b>8.12e-4</b> | <b>0.9744</b> | <b>33.534</b> |

| Method | Protection Efficacy (Ghost) |               |               |                 |               |               |
|--------|-----------------------------|---------------|---------------|-----------------|---------------|---------------|
|        | LPIPS↑                      | DISTS↑        | ID↑           | MSE↑            | SSIM↓         | PSNR↓         |
| FSG    | 0.0157                      | 0.0171        | 0.2762        | 4.12e-4         | 0.9833        | 34.756        |
| ARS    | 0.0207                      | 0.0218        | 0.3242        | 5.65e-4         | 0.9782        | 33.437        |
| DFS    | <b>0.0369</b>               | <b>0.0284</b> | <b>0.5530</b> | <b>15.35e-4</b> | <b>0.9615</b> | <b>30.929</b> |

Table 1. DFS (ours) achieves better perceptual quality for the shielded source images and better protection efficacy on the swapped images when compared to FSG [47] and ARS [16]. ↓ and ↑ denote lower is better and higher is better, respectively.

shielded versions of the source images by each approach. The right five columns show the target image, the swap of the source image (i.e., an unprotected face-swap) onto the target image (denoted *Source Image Swap*), and the FSG, ARS, and our (DFS) shielded source images swapped onto the target image (denoted *FSG Shield Swap*, *ARS Shield Swap*, and *DFS Shield Swap*), respectively.

**DFS shields are the highest perceptual quality.** Using the source image as a reference, we compare the artifacts in each shielded image to the source image. We find that in all cases, the artifacts are more clearly visible in the FSG and ARS shielded images when compared to the DFS shielded images; but, DFS producing shields with the fewest perceptible artifacts is only notable because DFS is also the most effective at breaking the face-swaps, as we will discuss next.

**DFS shields most effectively break the attempted face-swaps.** The goal of the shield is to produce face-swaps that look convincingly dissimilar from the intended swap (*Source Image Swap*). FSG performs the worst, with some results (e.g., row 3, 4, 5) still looking close to the intended face swap. On the other hand, both ARS and DFS successfully thwart the face-swap in all five cases. However, DFS results are the most convincing because the face-swap results do not look similar to the source face; thereby, completely thwarting any attempt to face-swap with that image. Conversely, some of the ARS results have some resemblance to the source face despite the clear artifacts. This is especially notable because DFS’s perceptual quality is better than ARS. See the supplementary material for more examples showing the high visual quality and efficacy of DFS.

### 4.3. Quantitative Comparison Against Baselines

Table 1 provides quantitative perceptual quality and protection efficacy comparisons between FSG, ARS, and ours

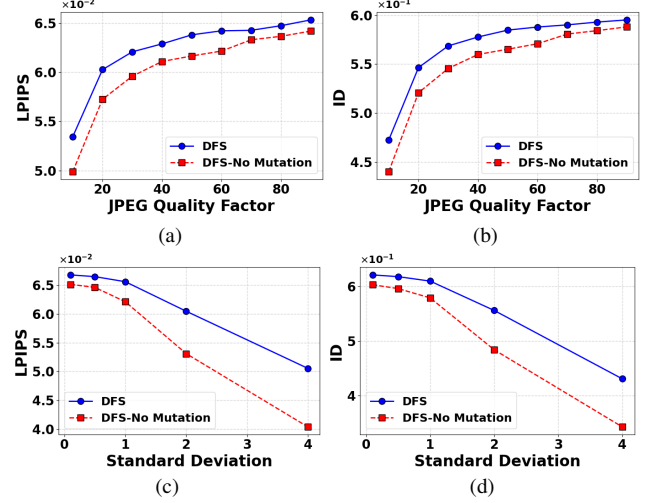


Figure 7. DFS has higher protection efficacy compared to DFS-NoMutation for JPEG compression and blurring on shielded images, indicating that the mutation layer is effective.

(DFS). For each method, we create shielded versions of each source image from the source-target pair test set. These are used for perceptual quality testing. We then create face swaps to evaluate protection efficacy by inputting the 5K source-target pairs of the test set into three face swappers: SimSwap, MobileFaceSwap, and Ghost. This results in 15K face-swapped images per method.

**DFS achieves the best perceptual quality scores overall.** DFS achieves the best perceptual quality scores for five metrics (LPIPS, DISTS, ID, MSE, and SSIM), and nearly the top score for PSNR. Moreover, DFS is the clear winner for the cutting-edge deep learning based perceptual metrics (LPIPS, DISTS, and ID). These metrics are the most important because they are trained specifically to model the human vision system for judging perceptual quality.

**DFS achieves the best protection efficacy overall.** DFS achieves the best scores in 16 out of 18 metrics for protection efficacy (often by clear, obvious margins). Moreover, when focusing on LPIPS, DISTS, and ID metrics, DFS is the clear winner in 8 out of 9 cases. In the case where FSG scores higher than DFS for the ID metric, this is likely because FSG is trained specifically to maximize the ID loss.

**Most importantly, DFS is the clear winner for both perceptual quality and protection efficacy.** This distinction is important because a trade-off exists between perceptual quality and protection efficacy. By winning for both, it shows that DFS is out-performing the other methods, not just trading off quality for efficacy or vice-versa.

### 4.4. Ablation Studies

**The mutation layer improves robustness of the shield significantly.** To ablate the impact of the mutation layer, we train an additional DFS variant referred to as DFS-

| Method | Perceptual Quality |               |               |                |               |               |
|--------|--------------------|---------------|---------------|----------------|---------------|---------------|
|        | LPIPS↓             | DISTS↓        | ID↓           | MSE↓           | SSIM↑         | PSNR↑         |
| B-DFS  | 0.1318             | 0.1516        | 0.1536        | 5.80e-4        | 0.8904        | 32.432        |
| DFS    | <b>0.0253</b>      | <b>0.0309</b> | <b>0.0499</b> | <b>2.10e-4</b> | <b>0.9789</b> | <b>36.739</b> |

| Method | Protection Efficacy (SimSwap) |               |               |                 |               |               |
|--------|-------------------------------|---------------|---------------|-----------------|---------------|---------------|
|        | LPIPS↑                        | DISTS↑        | ID↑           | MSE↑            | SSIM↓         | PSNR↓         |
| B-DFS  | <b>0.1005</b>                 | <b>0.0567</b> | <b>0.8290</b> | <b>39.51e-4</b> | <b>0.8995</b> | <b>26.451</b> |
| DFS    | 0.0659                        | 0.0391        | 0.6013        | 24.42e-4        | 0.9311        | 29.725        |

| Method | Protection Efficacy (MobileFaceSwap) |               |               |                 |               |               |
|--------|--------------------------------------|---------------|---------------|-----------------|---------------|---------------|
|        | LPIPS↑                               | DISTS↑        | ID↑           | MSE↑            | SSIM↓         | PSNR↓         |
| B-DFS  | <b>0.0383</b>                        | <b>0.0320</b> | <b>0.8129</b> | <b>13.05e-4</b> | <b>0.9624</b> | <b>29.371</b> |
| DFS    | 0.0243                               | 0.0223        | 0.5713        | 8.12e-4         | 0.9744        | 33.534        |

| Method | Protection Efficacy (Ghost) |               |               |                 |               |               |
|--------|-----------------------------|---------------|---------------|-----------------|---------------|---------------|
|        | LPIPS↑                      | DISTS↑        | ID↑           | MSE↑            | SSIM↓         | PSNR↓         |
| B-DFS  | <b>0.0520</b>               | <b>0.0375</b> | <b>0.7822</b> | <b>22.46e-4</b> | <b>0.9468</b> | <b>27.148</b> |
| DFS    | 0.0369                      | 0.0284        | 0.5530        | 15.35e-4        | 0.9615        | 30.929        |

Table 2. B-DFS, trained with an  $\epsilon$ -budget instead of  $\mathcal{L}_{\text{perceptual}}$ , offers superior protection efficacy at the cost of perceptual quality across the three face-swapping models. ↓ and ↑ denote lower is better and higher is better, respectively.

NoMutation. DFS-NoMutation uses the same configuration as DFS, but with the mutation layer removed. For a fair comparison, we select the DFS-NoMutation checkpoint where shielded images achieve perceptual quality metrics equivalent to those of DFS and use SimSwap to generate swapped images for both models. Figure 7 shows greater robustness achieved by DFS with the addition of the mutation layer across a wide range of JPEG compression values (plots (a) and (b)) and gaussian blurring sigmas (plots (c) and (d)). These results emphasize the importance of the mutation layer in the training pipeline and its role in preserving deepfake protection against adversarial attacks. Depending on the application, the mutation layer can be adapted and optimized to defend against other image manipulations. Plots for the omitted metrics show the same trends and can be found in the supplemental material.

$\mathcal{L}_{\text{perceptual}}$  improves both perceptual quality and protection efficacy together. To ablate the impact of the  $\mathcal{L}_{\text{perceptual}}$  loss, we train an additional DFS variant referred to as Budget-DFS (B-DFS). B-DFS uses the same configuration as DFS, but replaces optimization of  $\mathcal{L}_{\text{perceptual}}$  with an  $\epsilon = 0.03$  constraint on the perturbation magnitude as is done in FSG and ARS. Table 2 shows that B-DFS achieves better protection efficacy than DFS at the cost of lower perceptual quality. Notably, the perceptual quality metrics for B-DFS are even lower than those of FSG and ARS in Table 1, despite using the same  $\epsilon$ -budget. This result underscores that enforcing only a hard constraint on perturbation magnitude does not provide adequate control over the perceptual quality of shielded images, emphasizing the need for explicit optimization over perceptual quality metrics.

DFS generalizes to out-of-domain examples. To evaluate the generalization of the DFS model to images outside the FFHQ dataset, we randomly select 10K images from the CelebA-HQ dataset [24] and form 5K source-target image pairs. We then bilinearly resized the images to  $512 \times 512$

| Data | Perceptual Quality |        |        |         |        |        |
|------|--------------------|--------|--------|---------|--------|--------|
|      | LPIPS↓             | DISTS↓ | ID↓    | MSE↓    | SSIM↑  | PSNR↑  |
| FFHQ | 0.0253             | 0.0309 | 0.0499 | 2.10e-4 | 0.9789 | 36.739 |
| C-HQ | 0.0254             | 0.0361 | 0.0533 | 2.19e-4 | 0.9795 | 37.001 |

| Data | Protection Efficacy (SimSwap) |        |        |          |        |        |
|------|-------------------------------|--------|--------|----------|--------|--------|
|      | LPIPS↑                        | DISTS↑ | ID↑    | MSE↑     | SSIM↓  | PSNR↓  |
| FFHQ | 0.0659                        | 0.0391 | 0.6013 | 24.42e-4 | 0.9311 | 29.725 |
| C-HQ | 0.0610                        | 0.0407 | 0.5938 | 24.66e-4 | 0.9323 | 29.283 |

| Data | Protection Efficacy (MobileFaceSwap) |        |        |         |        |        |
|------|--------------------------------------|--------|--------|---------|--------|--------|
|      | LPIPS↑                               | DISTS↑ | ID↑    | MSE↑    | SSIM↓  | PSNR↓  |
| FFHQ | 0.0243                               | 0.0223 | 0.5713 | 8.12e-4 | 0.9744 | 33.534 |
| C-HQ | 0.0229                               | 0.0233 | 0.5461 | 8.08e-4 | 0.9752 | 33.974 |

| Data | Protection Efficacy (Ghost) |        |        |          |        |        |
|------|-----------------------------|--------|--------|----------|--------|--------|
|      | LPIPS↑                      | DISTS↑ | ID↑    | MSE↑     | SSIM↓  | PSNR↓  |
| FFHQ | 0.0369                      | 0.0284 | 0.5530 | 15.35e-4 | 0.9615 | 30.929 |
| C-HQ | 0.0329                      | 0.0283 | 0.5109 | 13.30e-4 | 0.9652 | 32.048 |

Table 3. DFS performs similarly on both FFHQ and CelebA-HQ (C-HQ) across multiple face-swapping algorithms despite DFS not being trained on CelebA-HQ. This shows DFS generalizes effectively beyond its training distribution. ↓ and ↑ denote lower is better and higher is better, respectively.

and input them to the DFS model trained on the FFHQ dataset. Table 3 presents the performance on the CelebA-HQ dataset and compares it with the results on FFHQ reported in Table 1. The DFS model achieves comparable performance on CelebA-HQ and FFHQ for perceptual quality and protection efficacy, demonstrating our model’s generalizability to images beyond the training domain.

## 5. Conclusion

We present DeepFakeShield (DFS), a method for thwarting face-swapping algorithms by adding adversarial perturbations to a source image. We train DFS to foil landmark detectors used in face-swapping algorithms, while modifying images imperceptibly and remaining robust to common image transformations. We show the qualitative and quantitative improvement of DFS over prior methods and justify the need for the mutation layer, and perceptual loss through ablation studies. Our work contributes to the broader literature around protection against nefarious uses of generative modeling methods. Beyond our academic contributions, our work demonstrates the feasibility and practicality of protecting facial imagery against current face-swapping algorithms, and we hope our findings incentivize social media platforms and camera manufacturers to adopt similar solutions to protect facial imagery. DFS, as presented in this work, is limited by how well the perceptual losses (in our case, LPIPS and DISTS) reflect true human perception, though it is straightforward to switch to any differentiable perceptual loss. DFS also requires training once per landmark detector, which limits its scalability and applicability. Follow-up work could consist of ensembling or jointly optimizing over adversarial perturbations corresponding to multiple different landmark detectors.



## References

- [1] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 658–659, 2020. 2
- [2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 3, 5, 6
- [3] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019. 1
- [4] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15108–15117, 2021. 2
- [5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5
- [6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotisa, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 3
- [7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 4
- [8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2
- [9] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European conference on computer vision*, pages 18–35. Springer, 2022. 2
- [10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2
- [11] Jiri Fridrich. Image watermarking for tamper detection. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, pages 404–408. IEEE, 1998. 2
- [12] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 2
- [13] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022. 6
- [14] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 744–752, 2022. 2
- [15] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. Delving into sequential patches for deepfake detection. *Advances in Neural Information Processing Systems*, 35:4517–4530, 2022. 2
- [16] Jiazhi Guan, Yi Zhao, Zhuoer Xu, Changhua Meng, Ke Xu, and Youjian Zhao. Adversarial robust safeguard for evading deep facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 118–126, 2024. 1, 2, 5, 7
- [17] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfd: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019. 3
- [18] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don’t lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 2
- [19] Tingyu Hu, Haibing Yin, Hongkui Wang, Ning Sheng, and Yafen Xing. Pixel-domain just noticeable difference modeling with heterogeneous color features. *Sensors*, 23(4):1788, 2023. 4
- [20] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmu-a-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI conference on artificial intelligence*, pages 989–997, 2022. 2
- [21] Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1619–1627, 2021. 2
- [22] Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. Faceshield: Defending facial image against deepfake threats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10364–10374, 2025. 2
- [23] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision*, 130(7):1678–1734, 2022. 1
- [24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 8
- [25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5
- [26] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Difface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025. 2
- [27] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1

- [28] Changjiang Li, Li Wang, Shouling Ji, Xuhong Zhang, Zhao-han Xi, Shanqing Guo, and Ting Wang. Seeing is living? rethinking the security of facial liveness verification in the deepfake era. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2673–2690, 2022. 1
- [29] Qilei Li, Mingliang Gao, Guisheng Zhang, and Wenzhe Zhai. Defending deepfakes by saliency-aware attack. *IEEE Transactions on Computational Social Systems*, 11(4):5060–5067, 2023. 2
- [30] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In icu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018. 2
- [31] Jie Ling, Jinhui Chen, and Honglei Li. Fdt: Improving the transferability of adversarial examples with frequency domain transformation. *Computers & Security*, 144:103942, 2024. 2
- [32] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019. 2
- [33] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision*, pages 667–684. Springer, 2020. 2
- [34] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023. 2
- [35] Xiangtao Meng, Li Wang, Shanqing Guo, Lei Ju, and Qingchuan Zhao. Ava: Inconspicuous attribute variation-based adversarial attack bypassing deepfake detection. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 74–90. IEEE, 2024. 1
- [36] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 2
- [37] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European conference on computer vision*, pages 236–251. Springer, 2020. 2
- [38] Nataniel Ruiz, Sarah Adel Bargal, Cihang Xie, and Stan Sclaroff. Practical disruption of image translation deepfake networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14478–14486, 2023. 2
- [39] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models, 2024. 2
- [40] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models, 2025. 2
- [41] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022. 2
- [42] Divyang Teotia, Agata Lapedriza, and Sarah Ostadabbas. Interpreting face inference models using hierarchical network dissection. *International Journal of Computer Vision*, 130(5):1277–1292, 2022. 1
- [43] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh 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. 2
- [44] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020. 1
- [45] Tomasz Walczynna and Zbigniew Piotrowski. Quick overview of face swap deep fakes. *Applied Sciences*, 13(11):6711, 2023. 3
- [46] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021. 2
- [47] Li Wang, Zheng Li, Xuhong Zhang, Shouling Ji, and Shanqing Guo. Faceswapguard: Safeguarding facial privacy from deepfake threats through identity obfuscation. *arXiv preprint arXiv:2502.10801*, 2025. 1, 5, 7
- [48] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 international conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021. 3, 6
- [49] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 2
- [50] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. Mobilefaceswap: A lightweight framework for video face swapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2973–2981, 2022. 6
- [51] Chen-Zhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung Liew. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, 16:1841–1854, 2020. 2
- [52] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2
- [53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 3
- [54] Liangqi Zhang, Yihao Luo, Haibo Shen, Tianjiang Wang, and Kate Larson. A fourier perspective of feature extraction and adversarial robustness. In *International Joint Conferences on Artificial Intelligence Organization*, pages 1715–1723, 2024. 6

- [55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [4](#)
- [56] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deep-fake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. [2](#)
- [57] Hegui Zhu, Yuchen Ren, Chong Liu, Xiaoyan Sui, and Libo Zhang. Frequency-based methods for improving the imperceptibility and transferability of adversarial examples. *Applied Soft Computing*, 150:111088, 2024. [2](#)
- [58] Vadim Ziyadinov and Maxim Tereshonok. Low-pass image filtering to achieve adversarial robustness. *Sensors*, 23(22): 9032, 2023. [6](#)