

# DeepFakeShield: A Proactive Defense Against Malicious Face Swapping

## Supplementary Material

This supplementary material is organized as follows:

To further demonstrate the applicability of our framework to other facial landmark detectors, Section 6 evaluates DFS when HRNet [2] is employed in place of SynergyNet [3]. In Section 7, we extend our initial ablation study comparing the performance of DFS and DFS-NoMutation, presenting the metrics that were not included in Fig. 7. An additional qualitative comparison between FSG, ARS, and our (DFS) method is provided in Section 8.

### 6. Extension to HRNet Landmark Detector

To demonstrate that DFS can be applied to any differentiable facial landmark detector, we evaluate its performance when HRNet [2] is used in place of SynergyNet.

#### 6.1. Experimental Setup

Following the same setup for the dataset, baselines, and metrics described in Section 4.1, we conduct experiments on the FFHQ dataset and compare DFS against FSG and ARS in terms of LPIPS, DISTs, Identity loss (ID), MSE, SSIM, and PSNR.

**Training Procedure:** Training follows a two-stage protocol. In the first stage, we set  $\lambda_{\text{perceptual}} = 0$  and optimize  $\mathcal{L}_{\text{protection}}$  for 60 epochs. At the end of this stage, the shielded images show strong protection efficacy, although with low perceptual quality. In the second stage,  $\lambda_{\text{perceptual}}$  is gradually increased by 0.25 every 1,000 iterations with a training batch size of 8, and training continues for an additional 16 epochs, at which point both LPIPS and DISTs metrics fall below 1.5%.

**Implementation Details:** Given an image  $x$ , HRNet generates a set of 98 heatmaps from which the coordinates of 98 facial landmarks are extracted. We select  $K = 31$  out of the 98 heatmaps corresponding to landmarks around the eyes, the tip of the nose, and the corners of the mouth, disregarding the remaining landmarks, such as those around the jawline. We sum all  $K = 31$  heatmaps to obtain a single heatmap  $h$ , which is then normalized to sum to 1. The heatmap  $h$  is treated as a distribution that visually resembles a Gaussian mixture with  $K = 31$  components. If  $h_s$  and  $\tilde{h}_p$  are the corresponding heatmaps 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 a weighted combination of MSE, KL-Divergence, and Wasserstein Distortion<sup>2</sup> (WD) [1] between

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	36.768
DFS	<b>0.0163</b>	<b>0.0165</b>	<b>0.0632</b>	<b>1.67e-4</b>	<b>0.9815</b>	<b>38.025</b>

Method	Protection Efficacy (SimSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
FSG	0.0435	0.0281	<b>0.6687</b>	12.70e-4	0.9476	30.845
ARS	0.0571	0.0386	0.5357	20.12e-4	0.9339	<b>28.895</b>
DFS	<b>0.0675</b>	<b>0.0397</b>	0.6177	<b>25.78e-4</b>	<b>0.9278</b>	28.948

Method	Protection Efficacy (MobileFaceSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
FSG	0.0106	0.0139	0.3481	2.38e-4	0.9878	36.812
ARS	0.0147	0.0181	0.3780	3.63e-4	0.9841	35.119
DFS	<b>0.0198</b>	<b>0.0195</b>	<b>0.5006</b>	<b>5.95e-4</b>	<b>0.9778</b>	<b>33.966</b>

Method	Protection Efficacy (Ghost)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
FSG	0.0166	0.0179	0.2853	4.33e-4	0.9827	34.482
ARS	0.0222	0.0230	0.3367	6.09e-4	0.9770	33.075
DFS	<b>0.0315</b>	<b>0.0253</b>	<b>0.4843</b>	<b>11.84e-4</b>	<b>0.9668</b>	<b>31.355</b>

Table 4. DFS’s advantages in perceptual quality and protection efficacy over FSG and ARS hold when HRNet is used in place of SynergyNet for facial landmark detection and image alignment. ↓ and ↑ denote lower is better and higher is better, respectively.

$h_s$  and  $\tilde{h}_p$ :

$$\begin{aligned} \mathcal{L}_{\text{protection}} = & -\lambda_{\text{protection-MSE}} \cdot \mathcal{L}_{\text{MSE}}(h_s, \tilde{h}_p) \\ & -\lambda_{\text{protection-KL}} \cdot \mathcal{L}_{\text{KL}}(h_s, \tilde{h}_p) \\ & -\lambda_{\text{protection-WD}} \cdot \mathcal{L}_{\text{WD}}(h_s, \tilde{h}_p). \end{aligned} \quad (5)$$

The design of the mutation layer remains identical to that described in Section 4.1. Models are trained using an NVIDIA L4 GPU. The training batch size and the scaling factor are set to 8 and  $\delta = 0.5$ , respectively. Adam optimizer with the learning rate  $5e-4$  is used to fine-tune the weights of the shield network. We set  $\lambda_{\text{MSE}} = \lambda_{\text{protection-MSE}} = 10$ ,  $\lambda_{\text{LPIPS}} = \lambda_{\text{DISTS}} = \lambda_{\text{protection-KL}} = 1$ , and  $\lambda_{\text{protection-WD}} = 0.1$ . Inference is carried out on an NVIDIA GeForce RTX 4070 Laptop GPU and takes 32.3 ms per image, including data transfer and I/O.

#### 6.2. Comparison Against Baselines

In Table 1, we presented the improved perceptual quality and protection efficacy of DFS relative to FSG and ARS when SynergyNet is used as the facial landmark detector. Table 4 demonstrates that these improvements persist when HRNet is employed in place of SynergyNet. Specifically, DFS attains higher perceptual quality and surpasses both FSG and ARS across all six metrics, often by clear margins. Moreover, DFS exhibits superior protection efficacy, achieving the highest scores in 16 out of 18 metrics. These results indicate that, regardless of the facial landmark detector utilized, DFS offers a more favorable balance between

<sup>2</sup><https://github.com/balle-lab/wasserstein-distortion>

Method	Perceptual Quality					
	LPIPS↓	DISTS↓	ID↓	MSE↓	SSIM↑	PSNR↑
S-DFS	0.0253	0.0309	<b>0.0499</b>	2.10e-4	0.9789	36.739
H-DFS	<b>0.0163</b>	<b>0.0165</b>	0.0632	<b>1.67e-4</b>	<b>0.9815</b>	<b>38.025</b>

Method	Protection Efficacy (HRNet + SimSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0106	0.0109	0.0778	2.06e-4	0.9865	39.216
H-DFS	<b>0.0675</b>	<b>0.0397</b>	<b>0.6177</b>	<b>25.78e-4</b>	<b>0.9278</b>	<b>28.948</b>

Method	Protection Efficacy (HRNet + MobileFaceSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0025	0.0050	0.0479	0.44e-4	0.9967	44.398
H-DFS	<b>0.0198</b>	<b>0.0195</b>	<b>0.5006</b>	<b>5.95e-4</b>	<b>0.9778</b>	<b>33.966</b>

Method	Protection Efficacy (HRNet + Ghost)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0054	0.0082	0.0546	1.14e-4	0.9944	40.743
H-DFS	<b>0.0315</b>	<b>0.0253</b>	<b>0.4843</b>	<b>11.84e-4</b>	<b>0.9668</b>	<b>31.355</b>

Table 5. The protection efficacy of S-DFS (trained with SynergyNet) is lower than that of H-DFS (trained with HRNet) at worse perceptual quality, showing S-DFS generalizes poorly to HRNet-based face swapping algorithms. ↓ and ↑ denote lower is better and higher is better, respectively.

perceptual quality and protection efficacy, outperforming the other methods on both fronts.

We complement these quantitative results with a video that qualitatively illustrates how DFS can prevent deepfake swaps. This video, provided in the supplementary material as demo.mp4, has the following three main parts:

- The first part presents the target face (left), the source face<sup>3</sup> (bottom-right), and the source face swapped onto the target video’s frames (right). SimSwap with HRNet as its facial landmark detector is used to generate the swapped frames. The swapped video on the right demonstrates the vulnerability of unshielded faces to deepfake swaps.
- The second part shows the original source face (left) and the shielded face (right) for qualitative assessment of the shield’s perceptual quality.
- The third part displays the swapped source video (left) and the swapped shielded video (right), demonstrating the protection efficacy provided by DFS.

### 6.3. Ablation Studies

**Generalization across different landmark detectors.** During training, the shield network learns to generate adversarial perturbations tailored to the specific facial landmark detector under attack. In this section, we examine whether the perturbations learned in this manner remain effective when applied to other facial landmark detectors.

We refer to the DFS models presented in Tables 1 and 4, which are trained to attack SynergyNet and HRNet, as S-DFS and H-DFS, respectively. The question is whether the shielded images generated by the S-DFS model offer sufficient protection when face-swapping models employ HRNet for facial landmark detection. Table 5 summarizes the

<sup>3</sup>The target face video and the source face image are part of the CelebV-HQ and CelebA-HQ datasets, respectively.

Method	Perceptual Quality					
	LPIPS↓	DISTS↓	ID↓	MSE↓	SSIM↑	PSNR↑
S-DFS	0.0253	0.0309	0.0499	2.10e-4	0.9789	36.739
H-DFS	0.0163	0.0165	0.0632	1.67e-4	0.9815	38.025
SH-DFS	0.0400	0.0403	0.1237	4.12e-4	0.9603	34.076

Method	Protection Efficacy (SynergyNet + SimSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0659	0.0391	0.6013	24.42e-4	0.9311	29.725
SH-DFS	0.0705	0.0418	0.6325	26.167e-4	0.9262	28.987

Method	Protection Efficacy (SynergyNet + MobileFaceSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0243	0.0223	0.5713	8.12e-4	0.9744	33.534
SH-DFS	0.0261	0.0240	0.6076	8.61e-4	0.9727	32.638

Method	Protection Efficacy (SynergyNet + Ghost)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
S-DFS	0.0369	0.0284	0.5530	15.35e-4	0.9615	30.929
SH-DFS	0.0390	0.0300	0.5859	16.07e-4	0.9595	29.938

Method	Protection Efficacy (HRNet + SimSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
H-DFS	0.0675	0.0397	0.6177	25.78e-4	0.9278	28.948
SH-DFS	0.0631	0.0375	0.5859	23.37e-4	0.9314	29.307

Method	Protection Efficacy (HRNet + MobileFaceSwap)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
H-DFS	0.0198	0.0195	0.5006	5.95e-4	0.9778	33.966
SH-DFS	0.0193	0.0194	0.4755	5.68e-4	0.9785	34.096

Method	Protection Efficacy (HRNet + Ghost)					
	LPIPS↑	DISTS↑	ID↑	MSE↑	SSIM↓	PSNR↓
H-DFS	0.0315	0.0253	0.4843	11.84e-4	0.9668	31.355
SH-DFS	0.0301	0.0250	0.4497	10.99e-4	0.9684	31.544

Table 6. SH-DFS achieves protection efficacy comparable to S-DFS and H-DFS when the face-swapping models utilize SynergyNet and HRNet, respectively. This indicates that, rather than interfering with one another, aggregating adversarial perturbations from different models preserves the protection efficacy of each individual model. However, this enhanced generalizability comes at the expense of reduced perceptual quality. ↓ and ↑ denote lower is better and higher is better, respectively.

performance of S-DFS when SimSwap, MobileFaceSwap, and Ghost utilize HRNet for facial landmark detection, and compares these results with those of H-DFS reported in Table 4. H-DFS obtains higher protection efficacy than S-DFS without compromising perceptual quality, indicating adversarial perturbations learned by S-DFS fail to effectively protect when face-swapping models rely on HRNet for facial landmark detection and image alignment.

This lack of generalization arises from the fact that different facial landmark detectors employ distinct techniques and rely on different feature representations; consequently, perturbations that produce a strong adversarial effect on one set of features may not have the same effect on another. In the remainder of this section, we showcase one potential pathway to alleviate this issue, although a comprehensive investigation of this open-ended problem lies beyond the scope of the current work.

Based on the aforementioned S-DFS and H-DFS models, we create an ensemble model—hereafter referred to as SH-DFS—as follows: Given a source image  $x_s$ , SH-DFS calculates the adversarial perturbation as the sum of the perturbations produced by S-DFS and H-DFS when  $x_s$  is used as input. This combined adversarial perturbation is then

added to the source image, and the result is clipped to the range  $[0, 1]$  to ensure that the shielded image remains within a valid pixel domain.

As shown in Table 6, SH-DFS achieves protection efficacy comparable to that of S-DFS when face-swapping models employ SynergyNet for facial landmark detection. Likewise, SH-DFS provides protection efficacy similar to that of H-DFS when the models use HRNet for landmark detection. These findings indicate that ensembling is a viable approach for ensuring that shielded images remain robust against multiple facial landmark detectors. However, this improved generalizability comes at the cost of reduced perceptual quality, as also reported in Table 6. This degradation arises because the ensembling method—constructed by aggregating adversarial perturbations from each model—tends to amplify the overall perturbation magnitude, resulting in more noticeable artifacts. Future research could explore more sophisticated ensembling strategies or the joint optimization of adversarial perturbations for multiple landmark detectors.

## 7. Additional Ablation Results

In Fig. 7 we showed that DFS exhibits greater robustness to JPEG compression and Gaussian blurring compared to DFS-NoMutation, as measured by LPIPS and ID. In Fig. 8, we also report results for DISTs, MSE, SSIM, and PSNR. These results follow the same trend, indicating DFS consistently outperforms DFS-NoMutation across all six metrics and all compression and blurring values, underscoring the effectiveness of the mutation layer against adversarial attacks.

## 8. Additional Qualitative Comparisons

In Fig. 6 we qualitatively compared FSG, ARS, and DFS on a variety of source and target image pairs, both in terms of perceptual quality and protection efficacy. We extend this qualitative comparison in Fig. 9 to 10 additional image pairs. As before, source images shielded by DFS remain qualitatively more similar to the original source images than those processed by ARS and DFS; when artifacts appear in DFS-shielded images, they are substantially less perceptible than those produced by the other two methods. Likewise, DFS-shielded images that undergo face swaps are noticeably more distorted than those shielded by FSG or ARS. Similar to ARS, DFS often distorts images by inserting or modifying facial hair (rows 1, 2, 7, and 10); however, DFS also distorts eye shape, skin color, and texture to a greater extent than either FSG or ARS. Notably, DFS-shielded swap images (column 9) exhibit the lowest texture similarity and the least resemblance to a human face; however, the same is not true for the images produced by the other two methods (columns 7 and 8). Overall, these results suggest that DFS is more effective at balancing impercep-

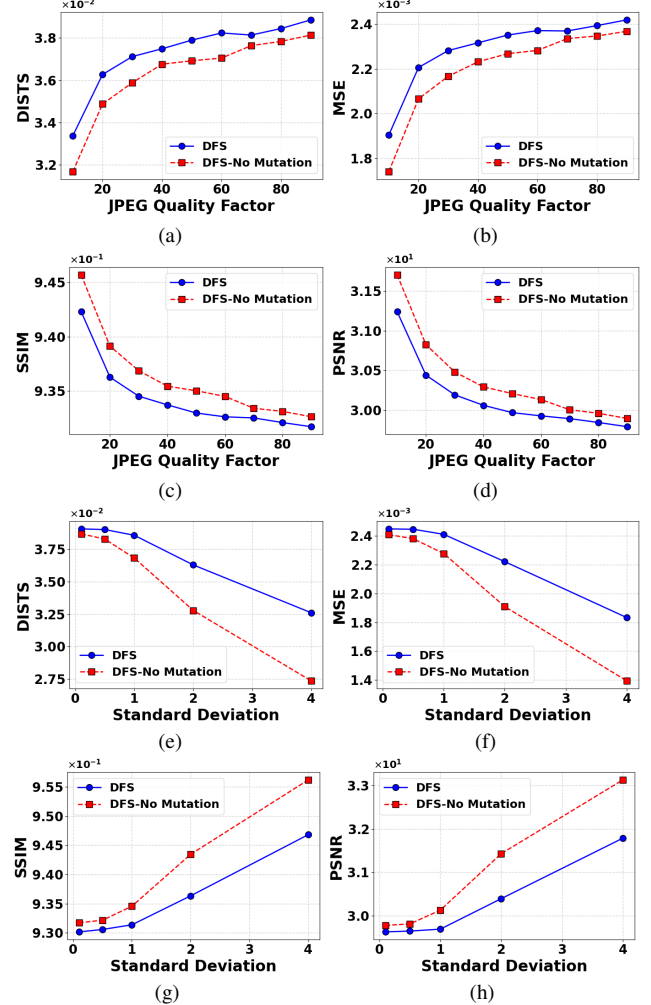


Figure 8. As measured by DISTs, MSE, SSIM and PSNR between source swap and shielded swap images, DFS offers higher protection efficacy against SimSwap than DFS-NoMutation for JPEG compression (plots (a)–(d)) and blurring (plots (e)–(h)), further justifying the need for the mutation layer.

tibility and protection efficacy, making it a viable approach for practical applications.



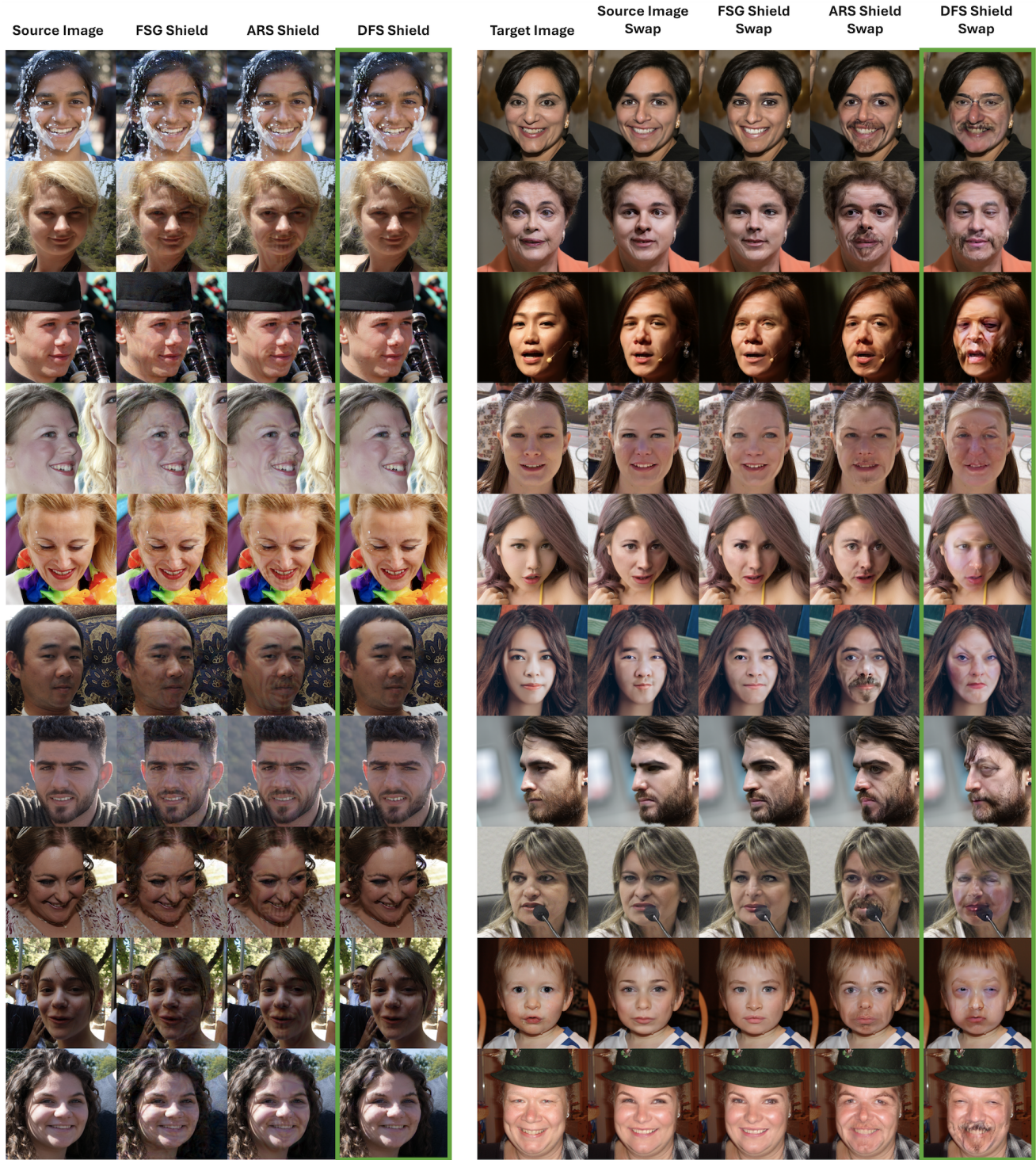


Figure 9. Additional qualitative comparison between FSG, ARS, and our (DFS) method. Left: source images and shielded images by FSG, ARS, and DFS (ours, highlighted in green). Right: from left to right, the target image, the target image face-swapped with the source, and each of the shielded versions face-swapped with the target (DFS highlighted in green). As in Fig. 6, 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).



## References

- [1] Yang Qiu, Aaron B. Wagner, Johannes Ballé, and Lucas Theis. Wasserstein distortion: Unifying fidelity and realism, 2024. [1](#)
- [2] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. [1](#)
- [3] 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. [1](#)