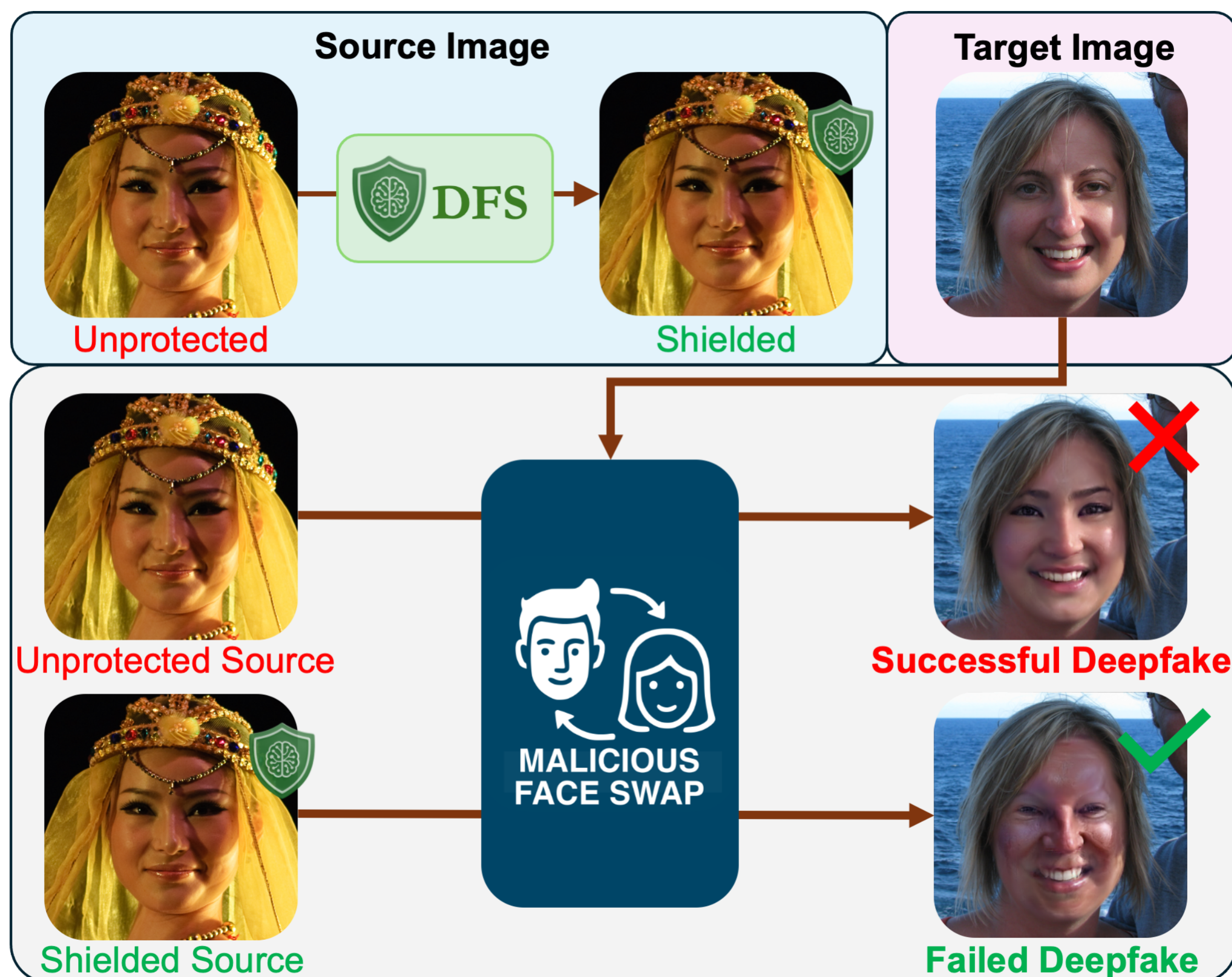


THE DEEFAKE THREAT

Malicious actors can use face-swappers for non-consensual explicit content, misinformation, identity theft, defamation, and social engineering.



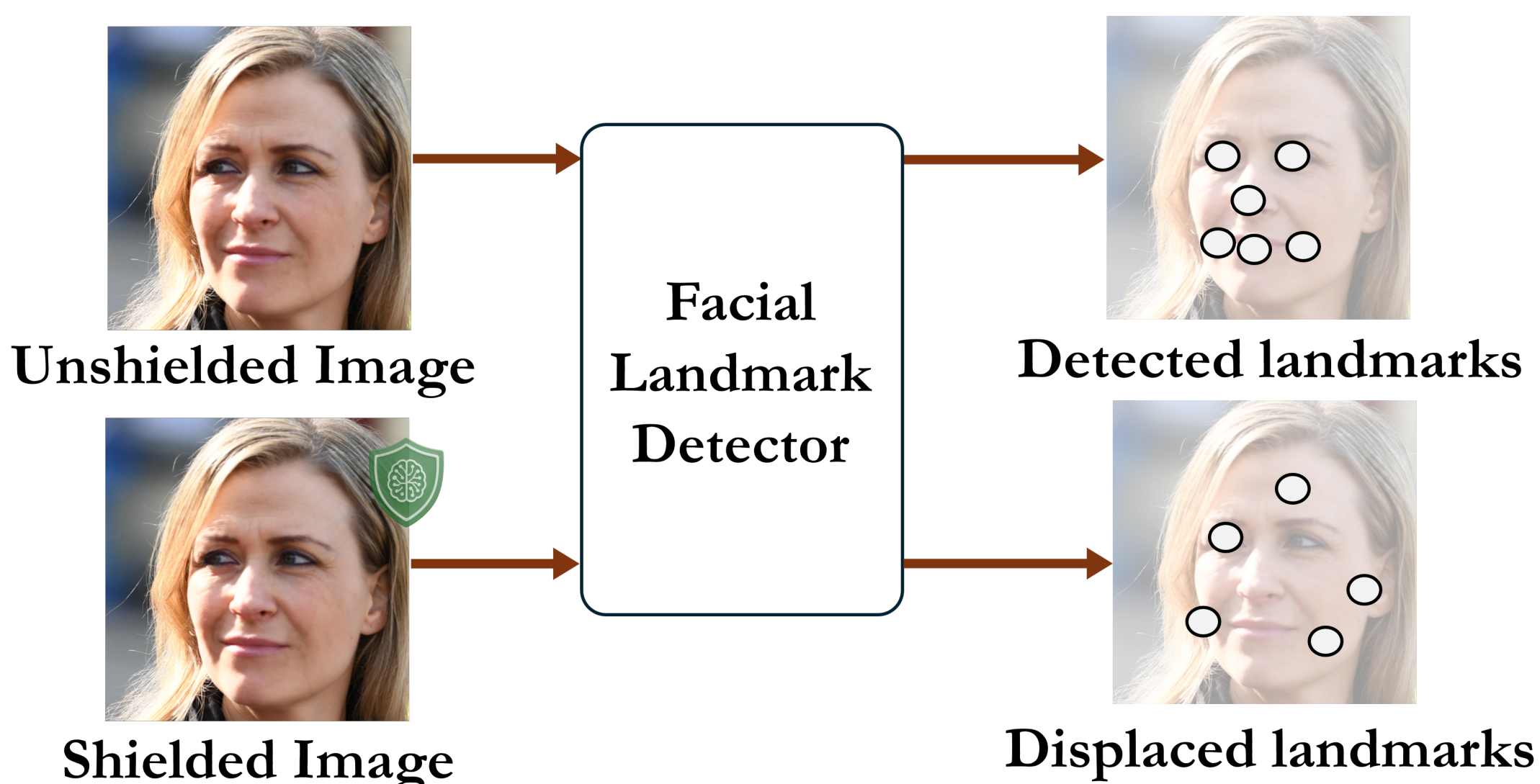
Shortcomings of current solutions:

- **Passive Detectors:** flag fakes after they are shared – harm already done.
- **Budget-Constrained:** overlook the texture, lighting, and semantics.
- **Fragile Shield:** sensitive to changes such as compression and blurring.

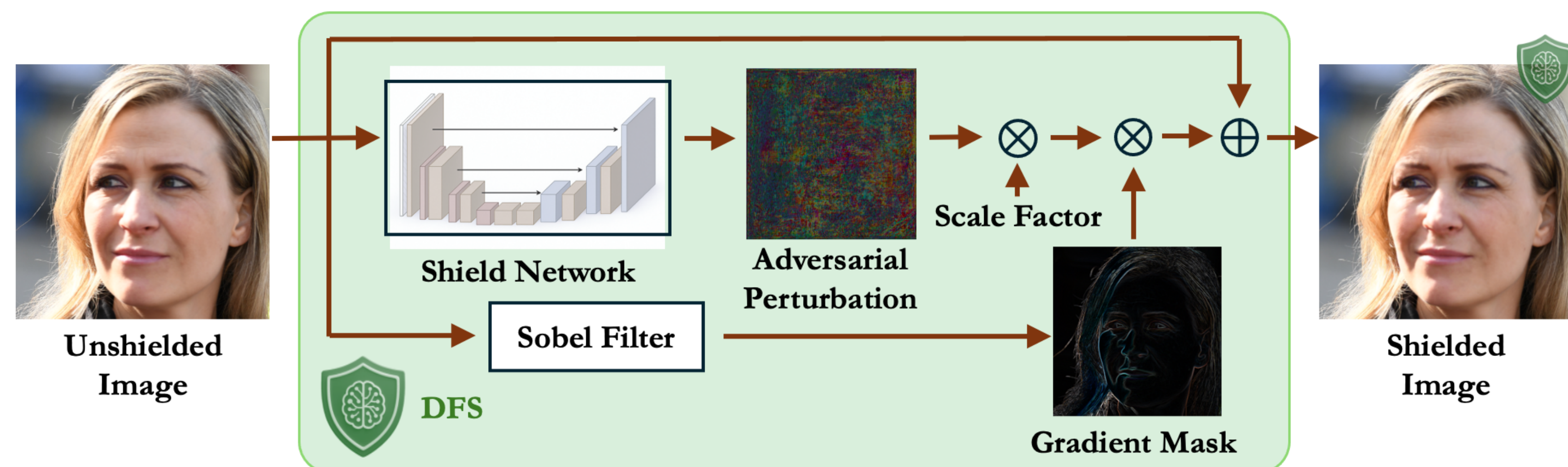
OUR SOLUTION

Key Insight: Many face-swapping methods rely on facial landmark detectors for precise **alignment** of the source and target faces.

Idea: Inject imperceptible adversarial noise to mislocalize facial landmarks.

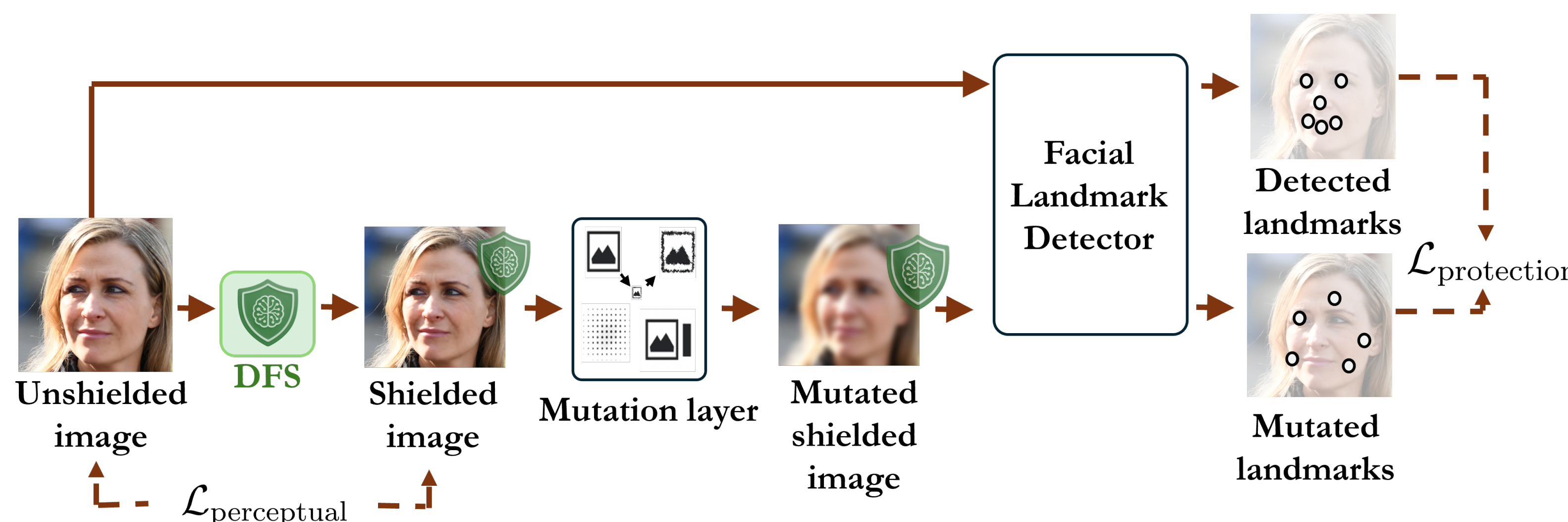


SHIELD ARCHITECTURE



- The shield is applied before sharing, rendering face-swap attempts unusable.
- A Sobel-guided gradient mask confines noise to edges, enhancing the shield's imperceptibility.

TRAINING PIPELINE

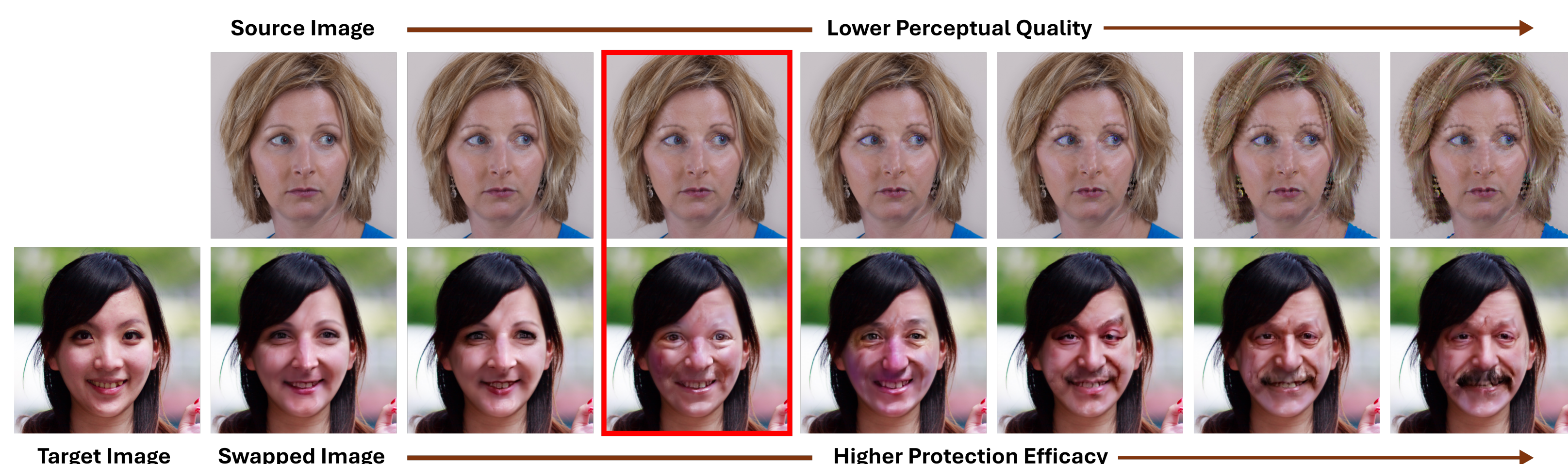


Objective function:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{perceptual}} \cdot \mathcal{L}_{\text{perceptual}} + \mathcal{L}_{\text{protection}}$$

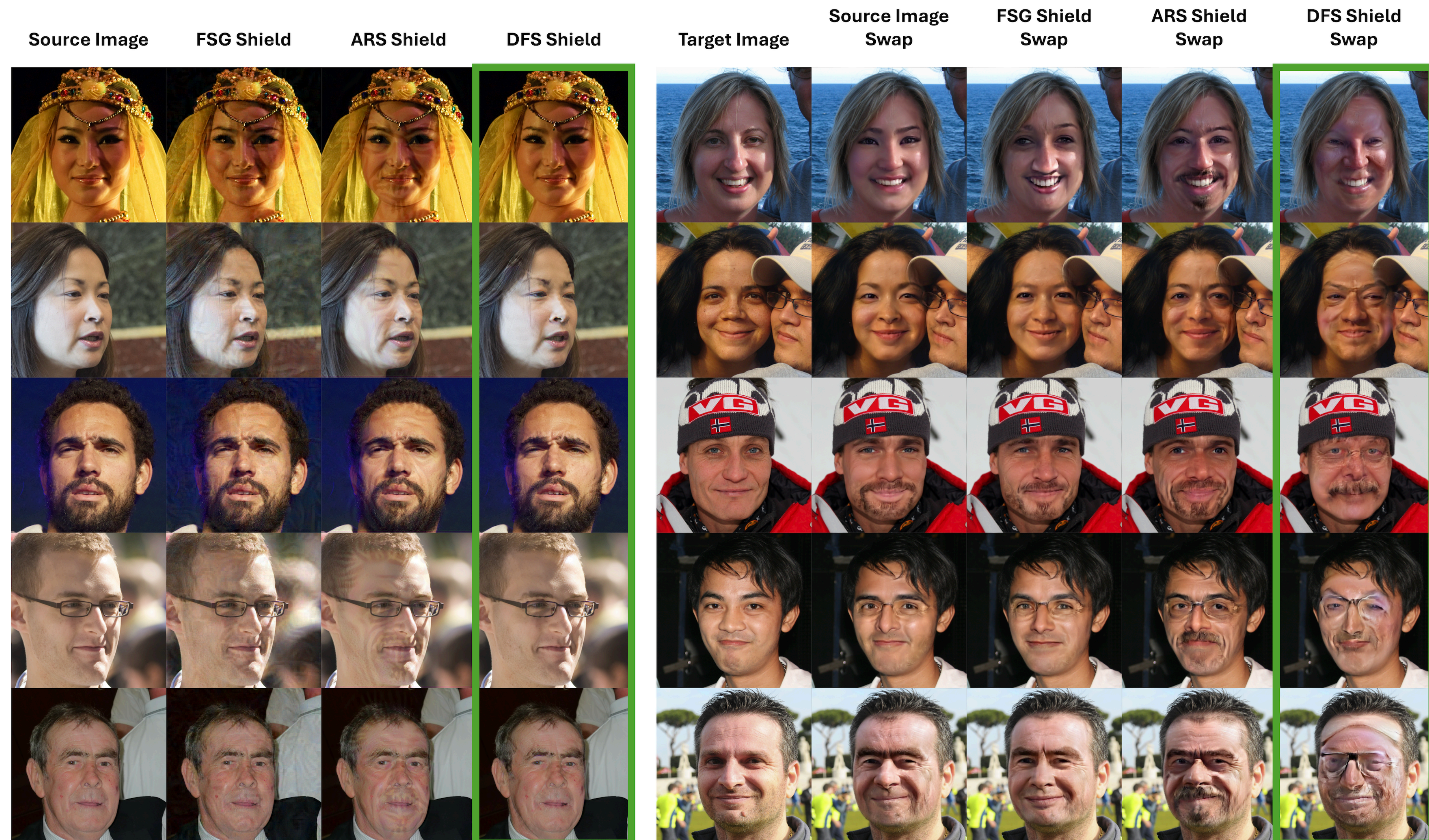
- A **stage-wise training** scheme with progressively increasing λ stabilizes the training.
- The **mutation layer** ensures that the shield remains robust to image transformations.

TRADEOFF BETWEEN PERCEPTUAL QUALITY AND PROTECTION EFFICACY

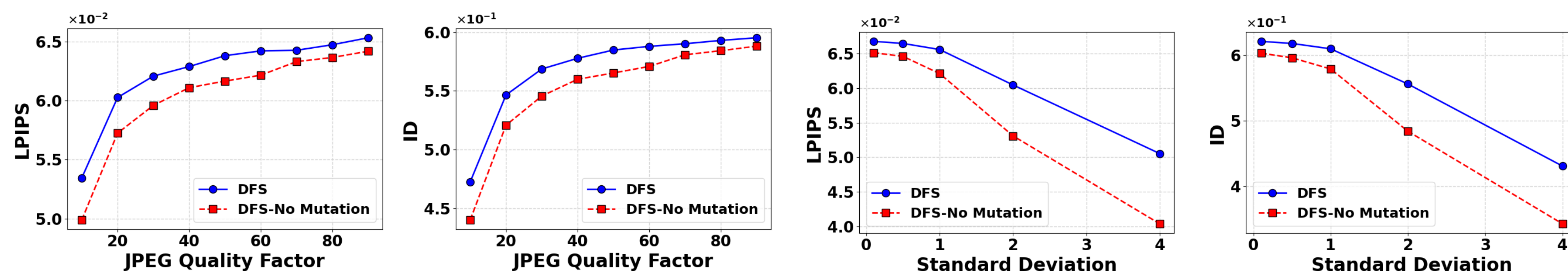


Higher protection efficacy comes at the cost of lower visual quality and vice versa.

SUPERIOR VISUAL QUALITY, INCREASED PROTECTION, AND ENHANCED ROBUSTNESS



Our shield (DFS) qualitatively outperforms other methods in terms of perceptual quality and protection efficacy.



The mutation layer enhances the shield's robustness against its induced transformations.

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	0.0253	0.0309	0.0499	2.10e-4	0.9789	36.739

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

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	0.0253	0.0309	0.0499	2.10e-4	0.9789	36.739

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

Our shield (DFS) quantitatively outperforms both other methods and our budget-based version.