IDProtector: An Adversarial Noise Encoder to Protect Against ID-Preserving Image Generation

Supplementary Material

Figure 6. Visualization of face localization prior. The prior channel is overlayed as the difference in brightness levels onto the original RGB images. Bright regions indicate 1s in the face localization prior channel, which would be pixels visible to ArcFace. Dark regions indicate 0s in the prior channel, which would be discarded before reaching ArcFace.

A. Contribution Statement

- **Yiren Song** proposed the topic and conducted literature review.
- **Yiren Song**, **Pei Yang**, and **Hai Ci** jointly validated the feasibility and proposed the method.
- **Pei Yang** trained the models.
- **Yiren Song** and **Pei Yang** jointly conducted the baseline comparison experiments. **Pei Yang** conducted other experiments.
- **Yiren Song** wrote Section 1-3 of the paper. **Pei Yang** wrote Section 4-5.
- **Hai Ci** and **Mike Zheng Shou** oversaw and advised the project.

B. Details of Prior Channels

As introduced in Sec. 4.2, IDProtector incorporates additional prior channels alongside the RGB channels to reduce the network's learning burden. Here, we elaborate on the details of these two prior channels and explain how they facilitate the network's learning process.

B.1. Face Localization Prior

The first prior channel is a face localization prior, implemented as a binary mask channel with values of 0 or 1, which indicates the facial region within the image (bright regions as visualized in Fig. 6). This prior channel explicitly marks the regions that InstantID's preprocessing step will preserve, thereby eliminating the need for the network to learn face localization capabilities independently, thus reducing the training burden.

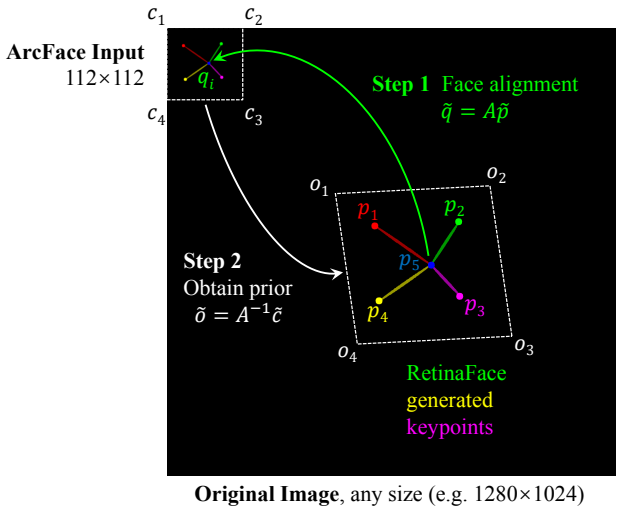

Figure 7. Visualization of ArcFace alignment and how the face localization prior is obtained.

The face localization region is defined by a quadrilateral whose vertices are computed through the following process. Initially, the three-channel 224×224 RGB image is processed by RetinaFace, which outputs five facial landmark points, denoted as $\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_5$. These landmarks correspond to the eyes, nose tip, and mouth corners. These points are used to align the facial features with ArcFace's required input format, which defines five predefined landmark locations $\tilde{\mathbf{q}}_1, ..., \tilde{\mathbf{q}}_5$. The alignment is achieved through an affine transformation, as illustrated in Figure X, resulting in a 112×112 aligned facial image. The affine transformation is given by

$$\tilde{\mathbf{q}} = A\tilde{\mathbf{p}}, \tag{6}$$

where A represents the affine transformation matrix. This matrix is computed by minimizing the transformation fitting error:

$$A = \arg\min_{A} \sum_{\text{landmark}i} \|\tilde{\mathbf{q}}_i - A\tilde{\mathbf{p}}_i\|_2^2, \tag{7}$$

which is solved using least squares fitting. Subsequently, the 112×112 region in the upper-left corner is cropped and fed into ArcFace as the aligned face. To mark this cropped region in the original image (which constitutes our face localization prior), we perform an inverse affine transformation. Denoting the coordinates of the square region's vertices as $\tilde{\mathbf{c}}_1, ..., \tilde{\mathbf{c}}_4$, their corresponding positions in the orig-

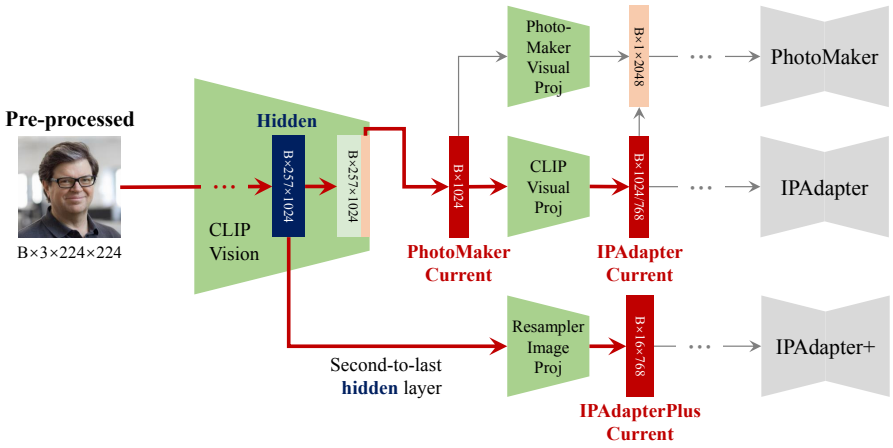

Figure 8. Alternative victim embedding that could be attacked during ID protection.

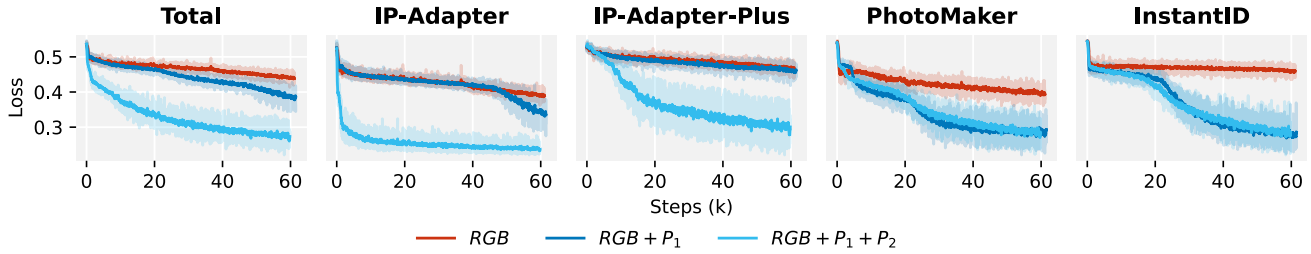

Figure 9. Loss convergence of the first 60k training steps when training with different prior channels.

inal image are computed as:

$$\tilde{\mathbf{o}}_i = A^{-1}\tilde{\mathbf{c}}_i. \tag{8}$$

The face localization prior is then obtained by shading the region bounded by $\tilde{\mathbf{o}}_i$ with value 1, leaving other pixels with value 0.

Our experiments demonstrate that this prior channel is crucial for effective training. As shown by the light and dark blue curves in Fig. 9 (corresponding to RGB + P_1 in the legend), it significantly reduces both PhotoMaker and InstantID losses. Without this prior, InstantID's cosine similarity plateaus around 0.5 and fails to achieve effective ID protection.

B.2. Aspect Ratio Prior

Similar to the Face Localization Prior, we provide an additional Center Crop Prior channel to the adversarial noise encoder model, serving as its fifth channel and second prior channel input. This channel also consists of binary values (0/1) and indicates the image regions that will be preserved after CLIP preprocessing. Providing this prior effectively communicates CLIP's preprocessing approach to the network through the prior format, eliminating the need for the network to learn this information independently and thereby reducing the training burden. As shown by the light blue curves in Fig. 9 (corresponding to RGB + P_1 + P_2 in the legend), this prior effectively helps IP-Adapter and IP-Adapter-Plus converge more rapidly during the early stages of training.

C. Ablation on Alternative Target Embeddings

In Sec. 4.3, we introduced three principles for selecting the target embedding for attacks, including blocking all potential pathways of information flow and ensuring the embedding possesses a certain level of semantic abstraction to facilitate the attack. However, our selection is not the only viable choice.

Beyond the current option, the penultimate hidden state of the CLIP Vision model also exhibits these two properties to some extent. This embedding is shown in blue in Fig. 8, while the current attack target is marked in red. We investigated replacing the current optimization target with this alternative embedding. The results, shown in Table 7, indicate that while the drop in facial similarity (measured using ISM) is less pronounced compared to the current target, the identity protection performance on IP-Adapter and IP-Adapter-Plus remains nearly comparable. This demonstrates that, although the current embedding achieves superior results, the choice of attack embedding is not uniquely constrained.

Table 7. Protection performance when choosing different target embeddings for attack. Each branch is evaluated using either the second-to-last hidden state (hidden) or the embedding obtained after the CLIP vision output (current).

Victim Embedding	ISM ↓		
	IP-Adapter	IP-Adapter-Plus	PhotoMaker
Hidden	0.047	0.075	0.134
Current	0.042	0.074	0.005

D. More Visualizations on Protecting against Unseen Models

Fig. 10 provides more visualizations of ID protection performance against customization models unseen during training.

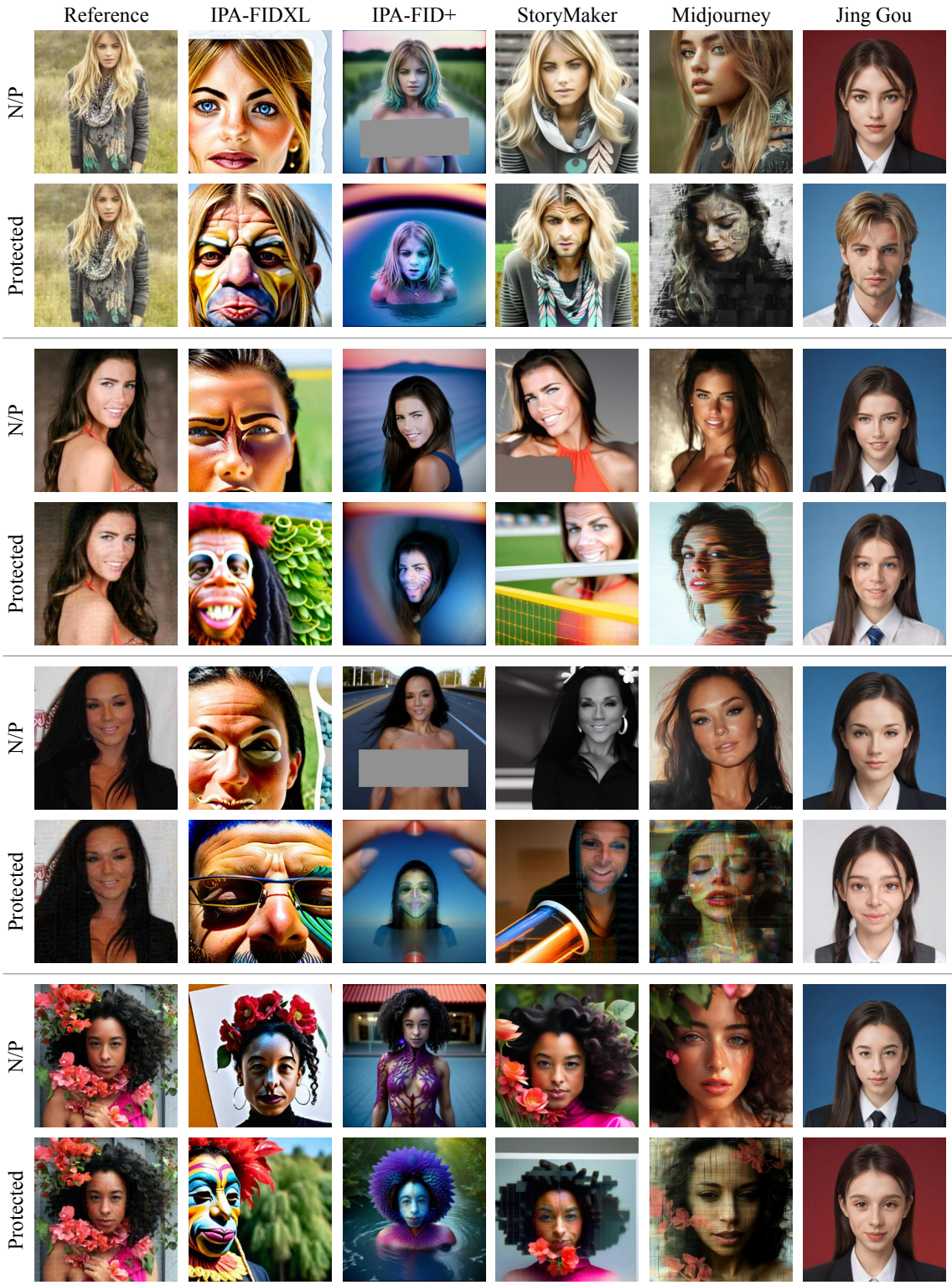

Figure 10. More visualizations of ID protection performance on unseen generators.