Sunglasses-Region Adversarial Perturbations for Feature-Space Face Recognition Attacks

Sheng-Pin Chang[†], Chun-Yao Chen[†], I-Syuan Lee, and Chia-Po Wei (✉)

Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[†] These authors contributed equally.
cpwei@mail.nsysu.edu.tw

Abstract. Deep face recognition systems remain vulnerable to adversarial accessory attacks, where localized perturbations can induce both dodging and impersonation failures. Prior work has largely focused on thin eyeglass-frame regions and often formulates attacks at the classifier level, limiting both attack strength and relevance to verification settings. In this work, we study adversarial attacks in the face embedding space and enlarge the perturbation region from eyeglass frames to a sunglasses-shaped region covering both frames and lenses. Using a generative optimization framework, we synthesize adversarial accessories and evaluate them on standard verification benchmarks, LFW and CALFW, under a white-box setting. We also assess cross-model transfer by generating adversarial overlays on AdaFace and testing them on ArcFace. Experiments show that sunglasses-region attacks consistently outperform eyeglass-based counterparts in both dodging and impersonation, producing larger drops in verification accuracy, greater shifts in cosine similarity, and stronger black-box transfer. Qualitatively, eyeglass-frame attacks tend to produce noise-like perturbations, whereas sunglasses-region attacks often induce structured eye-like patterns in the lens area, suggesting stronger interaction with identity-discriminative facial cues. These results highlight the importance of perturbation-region design in embedding-based face verification attacks and offer insights into adversarial robustness and security evaluation for modern pattern recognition systems.

Keywords: Adversarial Attack · Face Verification · Face Recognition.

1 Introduction

Face recognition (FR) has become a core component in authentication, access control, and media applications. Recent embedding-based models trained with large-margin metric learning have achieved very high accuracy on unconstrained benchmarks such as LFW and the cross-age variant CALFW [5, 20]. Representative embedding backbones include FaceNet, ArcFace, and AdaFace [12, 3, 7], where identity decisions are typically made by comparing cosine similarity between feature vectors rather than by relying on a closed-set classifier. As these

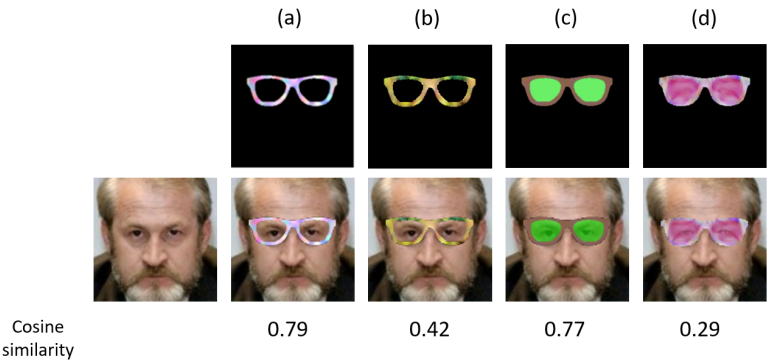

Fig. 1. Qualitative comparison of gallery-sampled (non-adversarial) and generated (adversarial) eyeglasses/sunglasses overlays and their effects on face embeddings. The leftmost image shows the reference face without glasses. Each column (a–d) presents the overlay patch (top) and the resulting composite image (bottom). (a) Eyeglasses randomly sampled from the eyeglasses dataset. (b) Eyeglasses generated by our eyeglasses attack. (c) Sunglasses randomly sampled from the sunglasses dataset. (d) Sunglasses generated by the proposed method. The number beneath each column is the cosine similarity between the reference and composite embeddings, computed using the fixed face feature extractor.

systems are increasingly deployed in real applications, understanding their robustness under adversarial manipulation has become an important problem for the evaluation and security of modern pattern recognition systems.

Despite this progress, FR inherits the broader vulnerability of deep networks to adversarial examples [16, 4, 10]. Perturbations can be optimized to induce large changes in network outputs while remaining visually subtle, and robustness depends strongly on the threat model and the optimization objective [9, 14]. In the verification setting, attacks are naturally expressed in feature space, for example by decreasing similarity to the true identity (dodging) or increasing similarity to a target identity (impersonation), without requiring access to a classifier head. Fig. 1 previews the motivation of this work by contrasting eyeglass- and sunglasses-region overlays and their induced changes in embedding similarity.

As shown in Fig. 1, extending the perturbation region from thin eyeglass frames to a sunglasses-shaped region that includes the lenses can yield substantially larger changes in cosine similarity under the same backbone. Moreover, sunglasses-region optimization often produces structured, eye-like textures inside the lens area, whereas eyeglass-frame perturbations tend to be more noise-like. This qualitative difference suggests that the lens region may interact more directly with identity-discriminative cues, motivating a systematic study of sunglasses-region perturbations under embedding-space objectives.

Patch-based attacks are especially relevant to FR because perturbations can be localized to an accessory-shaped region that resembles natural occlusions. General adversarial patches demonstrate that a localized pattern can reliably

induce misbehavior under varying backgrounds and viewpoints [2]. Robust optimization via expectation over transformations further improves tolerance to nuisance factors [1]. In the FR domain, eyeglass-style accessories have been shown to enable both dodging and impersonation against strong recognition models [13]. Recent work also explores learning-based generators to synthesize adversarial patches for FR and improve robustness and transferability [17, 6]. Beyond small accessory patches, several lines of work study more structured and semantically aligned adversarial changes, including transfer-oriented optimization, embedding-level losses, semantic image editing, makeup-based perturbations, and generalized manifold attacks [18, 21, 11, 19, 15, 8]. However, most accessory-based FR attacks still focus on relatively thin eyeglass-frame regions. This design limits the available perturbation area and may also restrict interaction with identity-discriminative facial cues around the eyes. By contrast, a sunglasses-shaped region covers both the frame and lens areas, providing a larger and more semantically aligned perturbation region. This motivates us to study whether perturbation-region design itself can significantly affect attack strength, transferability, and robustness evaluation in embedding-based face verification.

This paper studies sunglasses-region adversarial perturbations for feature-space face-recognition (FR) attacks in a digital white-box setting. We adopt the white-box setting to isolate the effects of region size and embedding-space objectives under full gradient access, which provides a clear upper bound and enables controlled comparisons to prior accessory-based attacks. Within a generative accessory optimization pipeline, we expand the perturbation region of interest from a thin eyeglass frame to a sunglasses-shaped region that covers both the frame and lens areas, and show that this enlarged region consistently yields stronger dodging and impersonation on LFW and CALFW under standard verification protocols, providing a security-oriented perspective on the robustness of modern recognition systems. Our contributions are threefold: we study adversarial accessory generation in a sunglasses region that extends beyond eyeglass frames and quantify its impact on attack strength; we formulate attacks directly in the FR embedding space, enabling both dodging and impersonation under verification protocols without relying on a classifier head; and we provide an empirical analysis on LFW and CALFW that highlights both quantitative gains and the emergence of structured eye-like patterns within the sunglasses lens region.

2 Related Work

Adversarial accessories and patches for face recognition. Localized adversarial patches provide an effective way to induce recognition errors while remaining compatible with natural occlusions and accessories [2]. In face recognition, the most representative early example is the eyeglass-based attack of Sharif et al., which demonstrated both dodging and impersonation using accessory-shaped perturbations [13]. Robust patch optimization under transformations has also been widely adopted to improve tolerance to geometric and photometric varia-

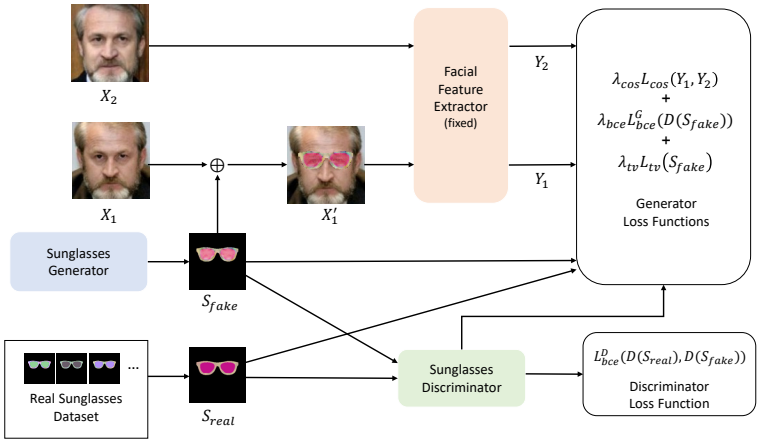

Fig. 2. Proposed pipeline for adversarial sunglasses generation. The generator synthesizes an RGBA sunglasses sample S_{fake}, which is aligned and alpha-blended onto the source face X_1 to form the perturbed image X_1'. A fixed face feature extractor $F(\cdot)$ produces embeddings $Y_1 = F(X_1')$ and $Y_2 = F(X_2)$, where X_2 is the reference/target image. The generator is optimized to manipulate the embedding similarity between Y_1 and Y_2 while maintaining visual plausibility through adversarial training and total-variation regularization. The sunglasses discriminator is trained with real samples S_{real} from the sunglasses dataset to distinguish real from generated sunglasses, thereby encouraging realistic outputs.

tions [1]. More recent FR-specific studies further explored generator-based adversarial patches to improve attack strength and transferability [6, 17].

Embedding-space objectives for verification attacks. For modern face recognition systems, attacks are naturally formulated in embedding space because identity decisions are typically made by comparing feature similarity rather than by using a closed-set classifier. Prior work has shown that embedding-level objectives improve transferability and are well aligned with verification-style evaluation [21]. In contrast to prior studies that mainly focus on relatively thin eyeglass-frame regions, our work studies a larger sunglasses-shaped ROI that includes the lens area, and evaluates its effect directly on feature-space dodging and impersonation under standard verification benchmarks.

3 Main Method

3.1 Overall Pipeline

Fig. 2 summarizes our framework for generating adversarial-yet-realistic sunglasses. The system comprises three components: (i) a sunglasses generator G, (ii) a sunglasses discriminator D, and (iii) a fixed face feature extractor F (the attacked FR model).

Forward computation. Given a source face image X_1 to be modified and a reference/target image X_2, we first compute the reference embedding $Y_2 = F(X_2)$. The generator samples a latent vector z and synthesizes an RGBA sunglasses pattern $S_{\text{fake}} = G(z)$. Using facial landmarks L, we align S_{fake} to the glasses region of X_1 and composite it by alpha blending (Eq. (2)) to obtain the attacked image X_1'. We then extract the attacked embedding $Y_1 = F(X_1')$ and optimize G to achieve a desired embedding-space effect under cosine distance.

Realism via adversarial learning. To ensure the synthesized sunglasses remain visually plausible, we train a discriminator D to distinguish generated samples S_{fake} from real sunglasses $S_{\text{real}} \sim \mathcal{D}_g$. The generator is updated to fool D while simultaneously optimizing the embedding-space objective (dodging or impersonation), optionally augmented by additional smoothness/perceptual regularizers.

3.2 Problem Setup

Let $F(\cdot)$ be a pretrained face-recognition feature extractor that is kept fixed, mapping an RGB face image to a d-dimensional embedding, $F : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$. Given a source image X_1 and a target (reference) image X_2, we compute $Y_2 = F(X_2)$ and $Y_1 = F(X_1')$, where X_1' is the attacked version of X_1 after compositing the generated sunglasses. We measure embedding discrepancy using the cosine distance: $d_{\cos}(Y_1, Y_2) = 1 - \cos\_\text{sim}(Y_1, Y_2)$, where $\cos\_\text{sim}(Y_1, Y_2)$ denotes the cosine similarity, defined as

$$\cos\_\text{sim}(Y_1, Y_2) = \frac{Y_1^\top Y_2}{\|Y_1\|_2 \, \|Y_2\|_2}, \qquad Y_1, Y_2 \in \mathbb{R}^d.$$

Sunglasses generator. We introduce a generator G that synthesizes an *RGBA* sunglasses pattern from a latent vector $z \sim \mathcal{N}(0, I)$:

$$G : \mathbb{R}^m \to \mathbb{R}^{h\times w\times 4}, \qquad G(z) = S_{\text{fake}}, \tag{1}$$

where the first three channels of S_{fake} are RGB values and the fourth channel is an alpha matte. Let $S_{\text{rgb}} \in \mathbb{R}^{h\times w\times 3}$ denote the RGB channels and $\alpha \in [0,1]^{h\times w}$ the normalized alpha channel.

Compositing (alpha blending). Let L denote facial landmarks used to place the generated sunglasses on X_1. We denote the geometric alignment operator by $T(\cdot\,; L)$, which scales/translates S_{fake} to the glasses region on X_1. The attacked image is obtained by a compositing operator $X_1' = \mathcal{C}(X_1, S_{\text{rgb}}, \alpha; L)$ that replaces the region-of-interest (ROI) pixels of X_1 with

$$I_{\text{out}} = \alpha \odot I_{\text{glass}} + (1 - \alpha) \odot I_{\text{face}}, \tag{2}$$

where I_{face} are ROI pixels from X_1, I_{glass} are ROI-aligned sunglasses pixels, and $\odot$ denotes element-wise multiplication. By construction, $X_1' = X_1$ outside the ROI.

Attack objectives (dodging and impersonation). We consider two verification attack goals:

- **Dodging (evasion, untargeted).** X_1 and X_2 correspond to the *same* identity. The goal is to induce a false reject by optimizing the generator so that the cosine distance, $d_{\cos}(Y_1, Y_2)$, is maximized.
- **Impersonation (targeted).** X_1 corresponds to an attacker identity and X_2 to a *target* identity. The goal is to induce a false accept by optimizing the generator so that the cosine distance, $d_{\cos}(Y_1, Y_2)$, is minimized.

Realism constraint. In both cases, the synthesized sunglasses should remain visually plausible. We therefore introduce a discriminator D and a real sunglasses dataset $\mathcal{D}_g$, where D is trained to distinguish S_{fake} from $S_{\text{real}} \sim \mathcal{D}_g$, and G is optimized under a combined objective that balances (i) embedding manipulation and (ii) realism enforced by adversarial learning (and optional regularizers).

3.3 Training Sunglasses Dataset and Preprocessing

To encourage visually plausible generations, we construct a sunglasses training set $\mathcal{D}_g$ with diverse appearances under a fixed geometric template. All samples share the same binary glasses mask M, which defines the canonical frame-and-lens region and facilitates consistent alignment during face compositing. Given a generated RGBA sample $S = (S_{\text{rgb}}, \alpha)$, we apply the mask as $S_{\text{rgb}} \leftarrow S_{\text{rgb}} \odot M$ and $\alpha \leftarrow \alpha \odot M$, so that pixels outside the template are removed before overlay.

To increase appearance diversity, we procedurally synthesize frame textures by combining a random linear gradient with multi-scale noise, followed by random HSV jittering. For the lens region, we use semi-transparent colors with random transparency $\alpha_\ell \in [0.2, 0.8]$. To keep the lens perceptually distinguishable from the frame after blending with skin tone, we select the lens color in CIELAB space by maximizing its perceptual distance from the frame color:

$$C_{\text{final}} = \arg\max_{C \in \mathcal{C}} \left\| F_{\text{Lab}}(C_{\text{frame}}) - F_{\text{Lab}}(C_{\text{mixed}}(C)) \right\|_2, \tag{3}$$

where $C_{\text{mixed}}(C) = \alpha_\ell C + (1 - \alpha_\ell) C_{\text{skin}}$ denotes the expected lens appearance after blending with a representative skin-tone background.

Each synthesized sample is stored as a four-channel RGBA image. The frame is set to be fully opaque, while the lens transparency is determined by α_ℓ. During training and attack-time overlay, we use the same alpha normalization and blending rule as in Eq. (2), ensuring consistency between dataset generation and deployment.

3.4 Network Architecture

Our framework consists of three components: (i) a generator G that synthesizes an RGBA glasses pattern, (ii) a discriminator D that encourages visual realism in the glasses domain, and (iii) a fixed face feature extractor F (AdaFace) that provides identity features for attack optimization.

We adopt a DCGAN-style architecture for both G and D. The generator maps a latent code $z \in \mathbb{R}^{100}$ to an RGBA output $G(z) \in \mathbb{R}^{4\times160\times160}$ through a

sequence of transposed-convolution upsampling layers with batch normalization and ReLU activations, followed by a final Tanh output layer. The discriminator takes an RGBA glasses sample $S \in \mathbb{R}^{4\times160\times160}$ and outputs a scalar realism score using stacked convolutional layers with LeakyReLU activations and batch normalization, followed by a final binary real/fake prediction.

This architecture is sufficient for our goal because the attack is driven primarily by the feature-space objective and the perturbation region design, while the adversarially trained generator-discriminator pair serves to maintain visual plausibility of the synthesized accessories.

3.5 Training Objective

We train the generator G to synthesize an RGBA glasses pattern that (i) achieves the desired embedding-space effect after alpha blending and (ii) remains visually plausible as a realistic glasses sample. Our objective combines a feature-space attack loss, an adversarial realism loss, and a lightweight smoothness regularizer based on total variation (TV).

Adversarial (realism) loss. We adopt a GAN formulation in the glasses domain. Let $G(z)$ denote the generated RGBA glasses sample and S_{real} denote a real RGBA glasses sample drawn from the glasses dataset $\mathcal{D}_g$. The discriminator D outputs a logit indicating whether its input is real. We use binary cross-entropy with logits:

$$\mathcal{L}^D_{\text{bce}} = \mathbb{E}_{S_{\text{real}}\sim\mathcal{D}_g}\big[\text{BCE}(D(S_{\text{real}}), 1)\big] + \mathbb{E}_z\big[\text{BCE}(D(G(z)), 0)\big], \tag{4}$$

$$\mathcal{L}^G_{\text{bce}} = \mathbb{E}_z\big[\text{BCE}(D(G(z)), 1)\big]. \tag{5}$$

The discriminator is optimized to minimize $\mathcal{L}^D_{\text{bce}}$, while the generator is optimized to minimize $\mathcal{L}^G_{\text{bce}}$.

Feature-space attack loss. Given a source face image X_1 and a target/reference image X_2, we composite the generated glasses onto X_1 via alpha blending to obtain X_1', and extract embeddings $Y_1 = F(X_1')$ and $Y_2 = F(X_2)$ using a frozen face recognition network F. We define the attack loss by the cosine distance $d_{\cos}(\cdot,\cdot)$ in the embedding space:

$$\mathcal{L}_{\cos} = s \cdot d_{\cos}(Y_1, Y_2), \qquad s = \begin{cases} +1, & \text{impersonation (reduce distance)}, \\ -1, & \text{dodging (increase distance)}. \end{cases} \tag{6}$$

With this sign convention, minimizing $\mathcal{L}_{\cos}$ supports both objectives: impersonation pulls Y_1 toward Y_2, while dodging pushes Y_1 away from Y_2.

Smoothness regularization (total variation). To suppress high-frequency artifacts and encourage visually smooth textures, we regularize the RGB channels of the generated glasses using a masked total-variation (TV) penalty. To match the evaluation/constraints reported in Sec. 4.2 and the tables, we compute TV separately on the *rim* and *lens* regions.

Let M_{rim} and M_{lens} denote binary masks for the rim and lens regions, respectively. For a generic binary mask M, we define the masked TV as

$$\begin{aligned}\text{TV}(G(z); M) = \frac{1}{|M|} \sum_{(u,v)\in M} \Big(&\big\|G(z)_{\text{rgb}}(u+1,v) - G(z)_{\text{rgb}}(u,v)\big\|_2 \\ &+ \big\|G(z)_{\text{rgb}}(u,v+1) - G(z)_{\text{rgb}}(u,v)\big\|_2 \Big),\end{aligned} \tag{7}$$

where $|M|$ is the number of pixels in the masked region and $\|\cdot\|_2$ denotes the Euclidean norm over RGB channels. We then define

$$\text{TV}_{\text{rim}}(G(z)) \triangleq \text{TV}(G(z); M_{\text{rim}}), \quad \text{TV}_{\text{lens}}(G(z)) \triangleq \text{TV}(G(z); M_{\text{lens}}). \tag{8}$$

We apply hinge-style activation with separate thresholds:

$$\mathcal{L}_{\text{tv,rim}} = \max\big(0,\, \text{TV}_{\text{rim}}(G(z)) - \tau_{\text{rim}}\big), \quad \mathcal{L}_{\text{tv,lens}} = \max\big(0,\, \text{TV}_{\text{lens}}(G(z)) - \tau_{\text{lens}}\big). \tag{9}$$

In the eyeglasses setting, we only enforce rim smoothness (i.e., we omit $\mathcal{L}_{\text{tv,lens}}$); in the sunglasses setting, both $\mathcal{L}_{\text{tv,rim}}$ and $\mathcal{L}_{\text{tv,lens}}$ can be enabled with their respective thresholds.

Overall generator objective. The final generator loss is

$$\mathcal{L}_G = \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\text{bce}}\mathcal{L}^G_{\text{bce}} + \lambda_{\text{rim}}\mathcal{L}_{\text{tv,rim}} + \lambda_{\text{lens}}\mathcal{L}_{\text{tv,lens}}. \tag{10}$$

The discriminator is trained by minimizing $\mathcal{L}^D_{\text{bce}}$ in Eq. (4).

3.6 Optimization Procedure

We optimize the generator–discriminator pair (G, D) in two stages. Stage I pretrains (G, D) on the standalone glasses dataset $\mathcal{D}_g$ using only the adversarial realism loss, so that the generator captures the appearance distribution of realistic eyeglasses or sunglasses. Stage II performs pairwise attack optimization on a face pair (X_1, X_2) under a fixed face recognizer.

In Stage II, we first compute the reference embedding $Y_2 = F(X_2)$ and keep it fixed. At each iteration, we sample real glasses from $\mathcal{D}_g$ and latent codes $z \sim \mathcal{N}(0, I)$, update D using the discriminator loss, generate $S_{\text{fake}} = G(z)$, alpha-blend it onto X_1 to obtain the attacked image X_1', compute the attacked embedding $Y_1 = F(X_1')$, and update G using the combined objective in Eq. (10). This objective balances the feature-space attack term, the adversarial realism term, and the masked TV regularizer.

The latent code z is re-sampled rather than optimized. For each face pair, we retain the generated glasses and composite image with the lowest generator objective value within the optimization budget.

4 Experimental Results

4.1 Datasets and Verification Protocol

Labeled Faces in the Wild (LFW) is a widely-used benchmark for unconstrained face verification, containing 13,233 face images from 5,749 identities collected

from the web. Following the standard *View 2* evaluation setting, LFW provides 6,000 verification pairs (3,000 mated and 3,000 non-mated) partitioned into 10 folds, each with 300 mated and 300 non-mated pairs [5].

Cross-Age LFW (CALFW) is a "renovation" of LFW designed to emphasize age variation by selecting positive pairs with large age gaps, while making negative pairs more confusable by matching demographic attributes. CALFW contains 12,174 images from 4,025 identities and adopts the same 6,000-pair, 10-fold protocol as LFW (3,000 mated / 3,000 non-mated; 300/300 per fold) [20].

We use AdaFace [7] as the pretrained face recognizer and denote its embedding extractor by $F(\cdot)$ (with an IR-50 backbone and 112×112 input). For a pair of aligned face images (X_1, X_2), we compute the cosine similarity $\mathrm{CSM}(X_1, X_2) = \frac{\langle F(X_1), F(X_2) \rangle}{\|F(X_1)\|_2 \, \|F(X_2)\|_2}$. To report verification accuracy, we follow the AdaFace evaluation protocol and make accept/reject decisions by thresholding the ℓ_2 distance between embeddings, i.e., $d(X_1, X_2) = \|F(X_1) - F(X_2)\|_2$, where a pair is accepted as the same identity if $d(X_1, X_2) \leq \tau$ and rejected otherwise. The threshold τ is calibrated on clean (unattacked) verification pairs using AdaFace and then kept fixed for all attack evaluations. In our experiments, the AdaFace-based thresholds are $\tau = 1.5070$ on LFW and $\tau = 1.6200$ on CALFW. Note that AdaFace embeddings are normalized, so ℓ_2 distance and cosine similarity are monotonic transforms; thus, optimizing cosine distance is consistent with an ℓ_2-threshold verification system.

We evaluate two attack objectives using the predefined pair lists: (i) *dodging*, where we perturb X_1 in each **mated** pair to reduce its similarity to X_2 (aiming for false rejects); and (ii) *impersonation*, where we perturb X_1 in each **non-mated** pair to increase its similarity to X_2 (aiming for false accepts). In both cases, X_2 remains unmodified and serves as the reference image. We report the verification accuracy under the fixed threshold (lower is better for the attacker) and the mean cosine similarity (CSM) over the attacked subset (mated pairs for dodging; non-mated pairs for impersonation).

4.2 Compared Methods

We compare seven evaluation settings under the same frozen face recognizer AdaFace (IR-50 backbone) [7]: (1) **Clean (AdaFace)**, i.e., original aligned faces without any overlay. (2) **Gallery eyeglasses**, where an eyeglass RGBA sample is randomly drawn from the eyeglasses set and alpha-blended onto the face (non-adversarial); (3) **Eyeglasses attack (TV-constrained)**, where we optimize the generator to produce an adversarial rim-region perturbation while enforcing a smoothness bound $TV_{\mathrm{rim}} \leq 0.4$; and (4) **Eyeglasses attack (unconstrained)**, where the same optimization is performed but without the TV constraint. (5) **Gallery sunglasses**, where a sunglasses RGBA sample is randomly drawn from the sunglasses set and alpha-blended (non-adversarial); (6) **Sunglasses attack (TV-constrained)**, where we optimize a sunglasses-region adversarial perturbation while enforcing $TV_{\mathrm{rim}} \leq 0.4$ and $TV_{\mathrm{lens}} \leq 0.1$; and (7) **Sunglasses attack (unconstrained)**, where the optimization is performed without these TV

Method	Acc↓	CSM↓	TV_{rim} ↓	TV_{lens} ↓
Clean (no overlay)	0.9967	0.6698	–	–
Eyeglasses				
Gallery-sampled overlay	0.9947	0.5713	**0.3289**	–
Ours (TV-constrained)	0.9633	0.4639	0.3970	–
Ours (unconstrained)	0.9583	0.4596	0.4620	–
Sunglasses				
Gallery-sampled overlay	0.9933	0.5247	0.3392	**0.0364**
Ours (TV-constrained)	0.3302	0.1953	0.3653	0.0953
Ours (unconstrained)	**0.2217**	**0.1576**	0.4877	0.3933

Table 1. Dodging attack results on LFW (AdaFace threshold $\tau = 1.5070$).

Method	Acc↓	CSM↓	TV_{rim} ↓	TV_{lens} ↓
Clean (no overlay)	0.9250	0.5191	–	–
Eyeglasses				
Gallery-sampled overlay	0.9197	0.4501	**0.3289**	–
Ours (TV-constrained)	0.8087	0.3206	0.3984	–
Ours (unconstrained)	0.8033	0.3161	0.4708	–
Sunglasses				
Gallery-sampled overlay	0.9097	0.4149	0.3392	**0.0364**
Ours (TV-constrained)	0.2103	0.0694	0.3651	0.0952
Ours (unconstrained)	**0.1480**	**0.0362**	0.4822	0.4016

Table 2. Dodging attack results on CALFW (AdaFace threshold $\tau = 1.6200$).

bounds. For all adversarial settings, the perturbation is produced by a generator and optimized directly in the AdaFace embedding space (dodging or impersonation, depending on the protocol) while the overlay/compositing procedure is kept identical to the non-adversarial gallery baselines. We use a generator learning rate of $LR_G = 3 \times 10^{-4}$ for eyeglasses and $LR_G = 9 \times 10^{-4}$ for sunglasses, reflecting the different ROI sizes and optimization difficulty.

4.3 Results on Dodging Attacks

Tables 1 and 2 summarize dodging results on LFW and CALFW. We attack *mated* verification pairs by perturbing only X_1 to reduce its similarity to the reference image X_2, thereby increasing false rejects under the fixed AdaFace threshold. We report verification accuracy and mean cosine similarity (CSM), where lower values indicate a stronger attack.

Gallery overlays have limited impact. Random gallery-sampled overlays cause only minor degradation from the clean baseline on both datasets, indicating that

Method	Acc↓	CSM↑	TV_{rim} ↓	TV_{lens} ↓
Clean (no overlay)	1.0000	0.0011	–	–
Eyeglasses				
Gallery-sampled overlay	1.0000	0.0008	**0.3289**	–
Ours (TV-constrained)	0.9910	0.0800	0.3832	–
Ours (unconstrained)	0.9897	0.0813	0.4305	–
Sunglasses				
Gallery-sampled overlay	1.0000	0.0005	0.3392	**0.0364**
Ours (TV-constrained)	0.7227	0.2021	0.3601	0.0939
Ours (unconstrained)	**0.6283**	**0.2226**	0.4585	0.3901

Table 3. Impersonation attack results on LFW (AdaFace threshold $\tau = 1.5070$).

Method	Acc↓	CSM↑	TV_{rim} ↓	TV_{lens} ↓
Clean (no overlay)	0.9963	0.0057	–	–
Eyeglasses				
Gallery-sampled overlay	0.9943	0.0052	**0.3289**	–
Ours (TV-constrained)	0.8777	0.1085	0.3900	–
Ours (unconstrained)	0.8750	0.1107	0.4485	–
Sunglasses				
Gallery-sampled overlay	0.9967	0.0047	0.3392	**0.0364**
Ours (TV-constrained)	0.2573	0.2527	0.3628	0.0945
Ours (unconstrained)	**0.1780**	**0.2776**	0.4675	0.4147

Table 4. Impersonation attack results on CALFW (AdaFace threshold $\tau = 1.6200$).

non-adversarial accessories alone are generally insufficient to induce consistent dodging failures.

Adversarial sunglasses are substantially stronger. Both eyeglasses and sunglasses attacks reduce verification performance, but sunglasses consistently produce much larger drops in both accuracy and CSM. This result shows that enlarging the perturbation region from the rim alone to the rim+lens region yields a much stronger feature-space dodging attack.

Effect of TV constraints. TV constraints improve smoothness, especially in the lens region, while slightly weakening attack strength. A similar but milder trend is observed for eyeglasses.

4.4 Results on Impersonation Attacks

Tables 3 and 4 report impersonation results on LFW and CALFW. Here we attack *non-mated* pairs by perturbing X_1 to increase its similarity to a target identity image X_2. Lower verification accuracy and higher CSM indicate a stronger impersonation attack.

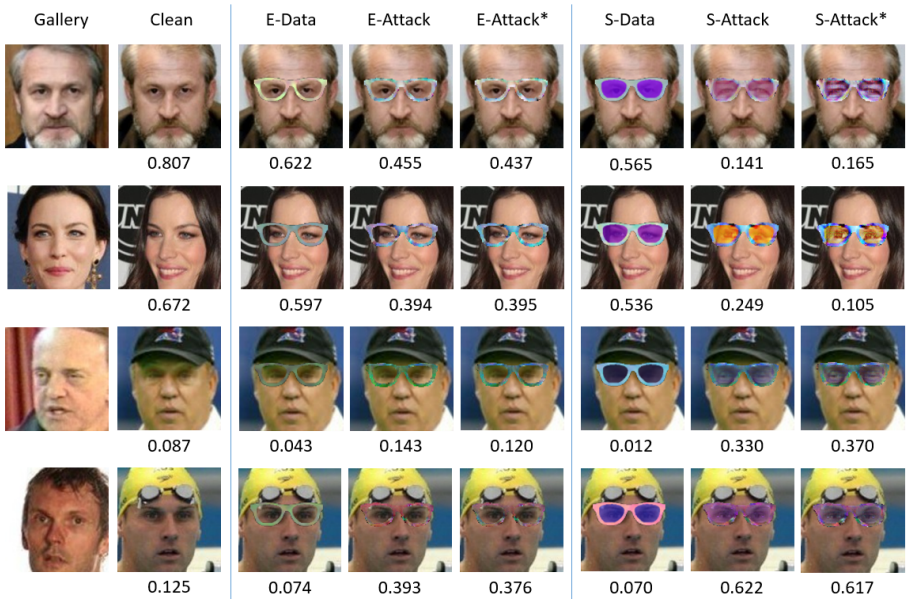

Fig. 3. Qualitative results on LFW/CALFW for dodging and impersonation. Each row shows the gallery image, the clean probe, and composites with eyeglasses or sunglasses overlays. E-Data and S-Data denote non-adversarial gallery-sampled overlays. E-Attack / E-Attack* and S-Attack / S-Attack* denote the TV-constrained and unconstrained adversarial variants, respectively. Numbers indicate cosine similarity between AdaFace embeddings of the probe and gallery.

Gallery overlays are near-neutral. Gallery-sampled eyeglasses and sunglasses have negligible impact on impersonation, with accuracy remaining close to the clean baseline and CSM staying near zero.

Adversarial perturbations induce strong impersonation, especially with sunglasses. Eyeglasses attacks already increase impostor similarity, but sunglasses are consistently more effective, yielding the highest CSM and the lowest verification accuracy across both datasets. This again confirms the importance of the lens-inclusive region for feature-space attacks.

Effect of TV constraints. TV constraints again improve visual smoothness while slightly reducing attack success, offering a practical trade-off between attack strength and visual regularity.

4.5 Qualitative Results on LFW/CALFW

Fig. 3 shows representative examples for dodging and impersonation on LFW and CALFW. In both tasks, gallery-sampled overlays usually cause only minor similarity changes, whereas adversarial overlays produce much larger shifts, especially for sunglasses that include the lens region. This observation is consistent with the quantitative results.

Method	Dodging		Impersonation	
	Acc↓	CSM↓	Acc↓	CSM↑
Clean (no overlay)	0.9960	0.6528	1.0000	0.0012
Eyeglasses				
Gallery-sampled overlay	0.9923	0.5610	1.0000	0.0000
Ours (TV-constrained)	0.9823	0.4950	0.9990	0.0366
Ours (unconstrained)	0.9820	0.4932	0.9987	0.0369
Sunglasses				
Gallery-sampled overlay	0.9890	0.5171	1.0000	−0.0001
Ours (TV-constrained)	0.7153	0.3043	0.9717	0.1019
Ours (unconstrained)	**0.6440**	**0.2834**	**0.9637**	**0.1114**

Table 5. Black-box LFW dodging and impersonation results on ArcFace (threshold $\tau = 1.5140$). Adversarial images are generated by attacking AdaFace and then evaluated on ArcFace without further optimization. For dodging, lower Acc/CSM indicates a stronger attack; for impersonation, lower Acc and higher CSM indicates a stronger attack.

A notable visual difference is that sunglasses attacks often induce structured eye-like patterns in the lens region, while eyeglasses attacks mainly introduce high-frequency perturbations around the rim. These examples support our main claim that enlarging the perturbation region from eyeglasses to sunglasses yields stronger and more semantically effective feature-space attacks.

4.6 Black-box Transfer to ArcFace on LFW

Table 5 reports black-box transfer to ArcFace [3], where adversarial overlays are generated on AdaFace and directly evaluated on ArcFace without further optimization. This setting measures cross-backbone transferability. For ArcFace, we follow the same evaluation protocol as in Sec. 4.1 and calibrate the verification threshold on clean LFW pairs, yielding $\tau = 1.5140$.

The main trend is clear: eyeglasses show only limited transfer, whereas sunglasses transfer substantially better in both dodging and impersonation. In particular, sunglasses reduce dodging accuracy much more strongly and also produce larger impersonation CSM increases, indicating that the enlarged lens-inclusive region yields more transferable perturbations. As in the white-box results, gallery-sampled overlays remain close to the clean baseline, confirming that the observed performance drop comes from adversarial optimization rather than generic occlusion.

Overall, the black-box results support our central claim that extending the perturbation region from eyeglasses to sunglasses improves not only white-box attack strength but also cross-model transferability.

5 Conclusion

We studied adversarial accessory attacks on embedding-based face verification and presented a generative framework for synthesizing adversarial RGBA sunglasses, optimized directly in feature space for both dodging and impersonation. Experiments on LFW and CALFW show that expanding the perturbation region from eyeglasses to sunglasses consistently yields stronger attacks, causing larger drops in verification performance and greater shifts in cosine similarity. Black-box evaluation further shows that sunglasses-based perturbations transfer more effectively across backbones than eyeglass-based ones. We also find that TV regularization improves visual smoothness while remaining competitive with, and in some cases outperforming, the unconstrained setting. Qualitatively, sunglasses attacks often produce structured eye-like patterns in the lens region, suggesting stronger interaction with identity-discriminative facial cues. Overall, these results highlight the importance of perturbation-region design for adversarial robustness and security evaluation in embedding-based pattern recognition systems.

References

1. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the International Conference on Machine Learning (ICML) (2018)
2. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
4. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015)
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2008)
6. Hwang, R.H., Lin, J.Y., Hsieh, S.Y., Lin, H.Y., Lin, C.L.: Adversarial patch attacks on deep-learning-based face recognition systems using generative adversarial networks. Sensors **23**(2), 853 (2023)
7. Kim, M., Jain, A.K., Liu, X.: AdaFace: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18729–18738 (2022)
8. Li, Q., Hu, Y., Liu, Y., Zhang, D., Jin, X., Chen, Y.: Discrete point-wise attack is not enough: Generalized manifold adversarial attack for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20575–20584 (2023)
9. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
10. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2574–2582 (2016)

11. Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H., Li, B.: SemanticAdv: Generating adversarial examples via attribute-conditioned image editing. In: European Conference on Computer Vision (ECCV). pp. 19–37 (2020)
12. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)
13. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS) (2016)
14. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security **22**(3), 1–30 (2019)
15. Sun, Y., Yu, L., Xie, H., Li, J., Zhang, Y.: DiffAM: Diffusion-based adversarial makeup transfer for facial privacy protection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24584–24594 (2024)
16. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014)
17. Xiao, Z., Gao, X., Fu, C., Dong, Y., Gao, W., Zhang, X., Zhou, J., Zhu, J.: Improving transferability of adversarial patches on face recognition with generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11845–11854 (2021)
18. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2730–2739 (2019)
19. Yin, B., Wang, W., Yao, T., Guo, J., Kong, Z., Ding, S., Li, J., Liu, C.: Adv-Makeup: A new imperceptible and transferable attack on face recognition. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 1252–1258 (2021)
20. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
21. Zhong, Y., Deng, W.: Towards transferable adversarial attack against deep face recognition. IEEE Transactions on Information Forensics and Security **16**, 1452–1466 (2021)