Neurocomputing 655 (2025) 131356

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Degradation-adaptive attack-robust self-supervised facial representation learning

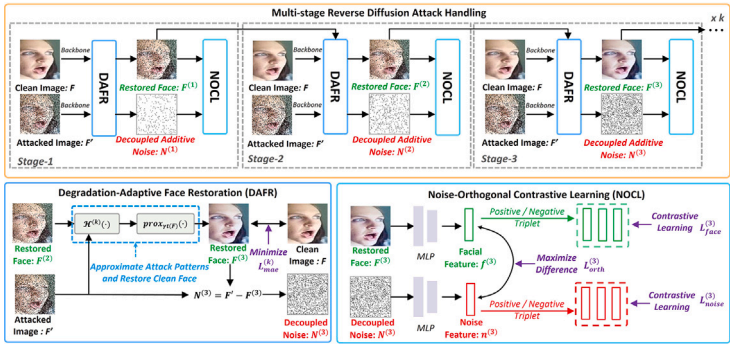
Ke Wang [a], Yuanyuan Liu [a,*], Chang Tang [a], Kun Sun [a], Yibing Zhan [b], Zhe Chen [c]

[a] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, 430074, China
[b] School of Computer Science, Wuhan University, Wuhan, China
[c] Cisco–La Trobe Centre for Artificial Intelligence and Internet of Things, School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, Australia

HIGHLIGHTS

- We propose DAR-SFRL, a novel degradation-adaptive SFRL method via reverse diffusion.
- DAFR models and inverts structured degradation using Bayesian theory and Taylor expansion.
- NOCL disentangles unstructured noise with three tailored contrastive learning losses.
- DAR-SFRL integrates semantic recovery and robustness learning in a unified framework.
- DAR-SFRL improves adversarial robustness by up to 96.98 % across multiple face tasks.

GRAPHICAL ABSTRACT



ARTICLE INFO

Communicated by J. Gui

Keywords:
Self-supervised learning
Face representation
Adversarial robustness
Facial semantic degradation
Contrastive learning

ABSTRACT

Self-supervised face representation learning (SFRL) shows strong potential for scalable face-related applications, yet remains vulnerable to adversarial attacks that cause dual facial semantic degradations, namely (1) structured distortions in key facial regions (*e.g.*, subtle inter-ocular distance shifts) that disrupt identity-related features, and (2) unstructured additive noise (*e.g.*, illumination artifacts) that entangles with face-related features in latent space. Existing defense methods struggle to deal with both facial semantic degradations in SFRL, resulting in limited robustness. To address this, inspired by existing reverse Diffusion approaches that effectively tackle the image denoising, we propose **DAR-SFRL**, a novel Degradation-adaptive Attack-Robust Self-supervised Face Representation Learning framework. DAR-SFRL models adversarial attacks as a degradation-based function composed of geometric distortions and additive noise, applying a multi-stage reverse Diffusion iterative process to recover facial semantics. At each stage of the process, DAR-SFRL employs: (1) an adaptive degraded-face restoration method that progressively reverses the degradation function and recovers fine-grained details from structured distortions, and (2) a noise-orthogonal contrastive learning mechanism to mitigate the impact of unstructured additive noise by maximizing the dissimilarity between noisy and clean image features in the latent space. Extensive experiments across tasks—including face recognition, facial expression recognition, and facial action unit detection—demonstrate that DAR-SFRL significantly outperforms state-of-the-art defenses under various adversarial attacks, highlighting its robustness and generalization in real-world face-aware applications. Our evaluation code is available at https://github.com/23wk/DAR-SFRL

* Corresponding author.
Email addresses: liuyy@cug.edu.cn (Y. Liu), zybjy@mail.ustc.edu.cn (Y. Zhan).

https://doi.org/10.1016/j.neucom.2025.131356
Received 5 June 2025; Received in revised form 19 July 2025; Accepted 20 August 2025
Available online 26 August 2025

1. Introduction

Learning facial representations is an important task in computer vision, which is widely applied in various face-related tasks, such as Face Recognition (FR), Face Emotion Recognition (FER), Human-Computer Interaction (HCI), Financial Security (FS), and Medical Diagnosis (MD). Although supervised learning has helped deep neural networks achieve promising facial understanding results, it heavily relies on large-scale annotations which require substantial labor costs. Recently, self-supervised facial representation learning (SFRL) has emerged as a promising alternative without overly relying on large-scale manual annotations. By learning from self-generated labels, SFRL enables effective utilization of vast unlabeled data, generating effective face-related models that are scalable for large-scale applications [1,2]. However, existing SFRL methods are often vulnerable to the threat of adversarial attacks in real-world scenarios, resulting in limited robustness in real-world face-related applications. Therefore, developing a robust SFRL method against various adversarial attacks remains a key and pressing research challenge.

To deal with adversarial attacks, current research has explored two primary defense methods for general image representation: adversarial training and adversarial purification [3,4]. Adversarial training methods improve robustness by incorporating adversarial disturbances directly during training. For example, Kim et al. [5] utilized unlabeled data for adversarial training and tried to defend against attacks by maximizing the similarity between randomly augmented samples and their adversarially perturbed counterparts at the instance level. Jiang et al. [6] combined self-supervised contrastive learning with adversarial training thereby improving the robustness of the model against introduced adversarial attacks during training. Adversarial purification-based methods aim to restore attacked images before inference, typically using generative priors or scoring functions. Nie et al. [7] relied on finding the optimal time step in the forward process of the diffusion model to uncover data from adversarial disturbances. Yong et al. [8] proposed a scoring function to distinguish attacks from clean data.

Despite the progress in general image representation tasks, we find that both defense paradigms face inherent limitations in dealing with SFRL [9,10]. Adversarial training methods primarily optimize for global feature robustness by augmenting the training set with adversarial examples and do not explicitly model or correct localized geometric distortions at key facial landmarks. As a result, the learned representations remain vulnerable to subtle structural degradations, especially in regions critical for identity and expression modeling. Adversarial purification methods typically perform denoising in the pixel domain and lack mechanisms to disentangle noise from meaningful semantics within the latent space. This limitation is particularly pronounced when noise is entangled with legitimate facial features such as texture or illumination, making it difficult to recover fine-grained, structurally relevant identity information. Compared to generalized image representation learning, SFRL heavily relies on fine-grained facial structures and subtle semantic cues to support downstream applications like identity recognition and emotion perception. However, under various adversarial attacks (*e.g.*, FGSM [11], PGD [12], and MIFGSM [13]), these delicate facial semantics are easily disrupted, resulting in significant degradation in discriminative performance.

As shown in Fig. 1, our analysis reveals two primary degradation patterns in SFRL attacks: (1) ***Structural Semantic Degradation***, where adversarial perturbations disrupt key facial regions, such as the interocular distance, nose bridge, and mouth corners, thereby impairing the extraction of identity-relevant structural cues; and (2) ***Unstructured Additive Noise***, where high-frequency noise entangles with genuine semantic features (*e.g.*, illumination artifacts) that entangle with face-related features in latent space. Existing adversarial training and adversarial purification methods typically focus on either global feature robustness or pixel-level denoising, making it challenging to simultaneously address these intertwined, fine-grained degradations. This gap fundamentally limits their effectiveness in defending against semantic degradation in facial representation learning.

To overcome these limitations, recent work has explored reverse diffusion for denoising via iterative refinement, leveraging diffusion models to reconstruct high-fidelity images from adversarial inputs [7]. Building on this, we propose Degradation-Adaptive Attack-Robust Self-supervised Face Learning, *i.e.*, **DAR-SFRL**, a multi-stage reverse diffusion framework that models both structural semantic degradation and unstructured noise as a unified degradation function, enabling targeted and robust defense against face adversarial attacks in SFRL. Fig. 1 shows a brief motivation for our approach, and Fig. 2 presents a training pipeline of our approach for attack-robust SFRL. Specifically, we introduce two key modules at each stage of DAR-SFRL: Degradation-Adaptive Face Recovery (DAFR) and Noise-Orthogonal Contrastive Learning (NOCL). First, DAFR employs a maximum a posteriori (MAP) strategy to progressively reverse the degradation function and recover fine-grained details from structured distortions. Then, NOCL comprises a noise orthogonal disentangling loss, a facial-robust contrastive loss, and a noise-sensitive contrastive loss, to mitigate the impact of unstructured additive noise by maximizing the dissimilarity between noisy and clean image features in the latent space. Through the multi-stage collaborative training of DAFR and NOCL in DAR-SFRL, we effectively capture the facial semantic degradation caused by face adversarial attacks, enabling a more precise understanding of adversarial attack patterns in SFRL. This enhances the robustness of DAR-SFRL in several face-related downstream tasks.

In summary, the main contributions of this paper are as follows:

(1) We propose a novel Degradation-adaptive Attack-Robust SFRL method in a multi-stage reverse diffusion learning manner, termed DAR-SFRL, which aims to effectively address the facial semantic degradation caused by adversarial attacks for obtaining attack-robust SFRL. To achieve this, we introduce two key modules in each stage of DAR-SFRL: Degradation-Adaptive Face Recovery (DAFR) and Noise Orthogonal Contrastive Learning (NOCL), to formulate semantic degradations caused by face adversarial attacks.
(2) We propose the DAFR component, which adaptively simulates and reverses facial semantic degradation to recover fine-grained details from structured distortions. In DAFR, we introduce Bayesian theory and Taylor expansion to iteratively approximate the optimal degradation process, capturing the relationship between degraded and clean images to effectively disentangle different structured degradation patterns.
(3) We devise a novel NOCL to further mitigate the impact of unstructured additive noise using three loss functions: noise-orthogonal disentangling loss, face-robust contrastive loss, and noise-sensitive contrastive loss. By learning in appropriate feature spaces, NOCL effectively decouples unstructured additive noise from face adversarial attacks by maximizing the dissimilarity between noisy and clean image features in the latent space.
(4) We evaluated the performance of DAR-SFRL on several face-related downstream tasks, including facial expression recognition (FER), face recognition (FR), and facial action unit detection (FAU). Extensive results show that DAR-SFRL outperforms existing methods in defending against seven types of face adversarial attacks during inference. For example, under the UPGD attack in the FER task, the baseline's performance dropped by 17.15 %, while DAR-SFRL's dropped by only 0.52 %, offering a 96.98 % relative improvement. These results demonstrate DAR-SFRL's effectiveness in enhancing robustness for face-related tasks in real-world scenarios.

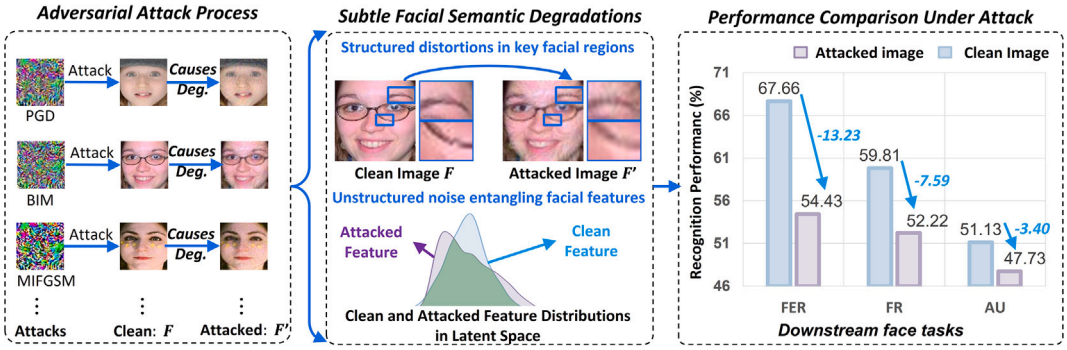


Fig. 1. The intuitive motivation of our method. Various adversarial attacks introduce subtle semantic degradations (Deg.) in facial data, which contains: (1) structured distortion of key facial regions, and (2) unstructured additive noise that disrupts face-related features in latent space. Existing SFRL methods lack a targeted and holistic learning strategy to defend against such adversarial perturbations, resulting in significantly degraded performance in several downstream face tasks.

2. Related work

2.1. Self-supervised facial representation learning

Self-supervised learning has shown a wide range of application prospects in the field of facial representation learning [14–18]. For example, He et al. [19] enhance face recognition by self-supervised 3D reconstruction. MAE [20] introduces a non-trivial and meaningful self-supervised task by masking large portions of random blocks in the input image and reconstructing the missing pixels. MCF [21] utilizes image-level contrastive learning and masked image modeling, as well as facial representation learning knowledge extracted from pre-trained models of external image networks. PCL [22] decouples the facial and pose features, and then conducts comparative learning on these features, achieving strong performance on both pose and facial analysis tasks. FRA [23] proposes a new self-supervised face representation learning framework to learn consistent global and local face representations. Although some progress has been made, most of the existing work is based on learning facial representations in pre-trained "black box" networks. This opaque working mechanism makes them vulnerable to attacked samples, that is, attackers can use attacked samples to specifically corrupt highly personalized and sensitive facial features. Misleading the network to learn incorrect features to fool the face recognition system leads to face information leakage, leaving a huge security risk.

To provide a more comprehensive context, recent surveys offer valuable overviews of self-supervised learning. Liu et al. [24] categorize SSL methods into generative, contrastive, and hybrid types, outlining their theoretical principles and practical applications. Balestriero et al. [25] provide practical training recipes and conceptual insights, calling SSL the "dark matter of intelligence" for its hidden complexity and power. Gui et al. [26] highlight current challenges, such as robustness and scalability, which are critical for advancing SSL. These reviews collectively establish a foundational understanding that informs the design of more robust and interpretable self-supervised facial representation models.

2.2. Adversarial defense

Adversarial attacks involve adding small, carefully designed disturbances to the original input data, which are difficult for humans to detect, leading to attacked samples [4,27,28]. These samples cause deep neural networks to make incorrect predictions in tasks like facial recognition and image classification. To counter this, researchers have developed various defense methods [5–7] to enhance model robustness and accuracy.For example, Chen et al. [29] used the improved FGSM to generate attacked samples for adversarial search, and employed a reconstructor to help the classifier learn key features under disturbances. Wang et al. [30] mapped attacked samples back to clean sample manifolds through an image-to-image generator, enhanced sample complexity, and integrated adversarial training into the GAN process to eliminate the problem of confusing gradients and improve defense effectiveness.Mao et al. [31] drew on NLP-style adversarial training, converted images into discrete visual words through VQGAN, and used symbolic adversarial disturbances to minimize risks, significantly improving the performance of visual representations. Yoon et al. [8] proposed an EBM adversarial purification method based on denoising score matching (DSM) training, which can quickly purify attacked images in a small number of steps.Yang et al. [3] proposed a defense method based on matrix estimation, which destroys the adversarial noise structure by randomly deleting pixels and reconstructing the image, strengthening the global structure of the original image, and making the network more consistent with human classification perception. Although adversarial defense methods have made progress, most existing adversarial defense methods do not explicitly account for the dual nature of adversarial degradation in SFRL, namely, structural semantic degradation and unstructured additive noise.Together, these two types of degradation pose a compounded challenge that existing adversarial defense methods are not designed to handle effectively. Most adversarial training techniques focus on improving instance-level robustness but fail to capture the subtle structural misalignments caused by perturbations in critical facial regions. Conversely, purification-based defenses typically operate in the image domain and lack mechanisms to disentangle high-frequency noise entangled in the latent feature space. This motivates the need for defense strategies that jointly model both structural and unstructured degradations to robustly enhance the resilience of self-supervised facial representation learning.

3. The proposed method

3.1. Problem definition and overview

The overview of the proposed DAR-SFRL is shown in Fig. 2. Our primary goal is to comprehensively defend SFRL models against semantic degradations caused by adversarial attacks, which include: (1) structured distortions in key facial regions (*e.g.*, subtle inter-ocular distance shifts), and (2) imperceptible unstructured additive noise. These two degradation forms jointly undermine the model's ability to capture fine-grained facial details and perform accurate discrimination. To provide a theoretical foundation for this decomposition, we start with the mathematical definition of adversarial attacks. Mathematically, let F denote a clean facial image, and F' denote the attacked image:

$$F' = F + \delta, \tag{1}$$

where δ represents the semantic degradation introduced by the adversarial attack. Although adversarial attacks are crafted to mislead models, their effect on images can be interpreted as a form of degradation, which

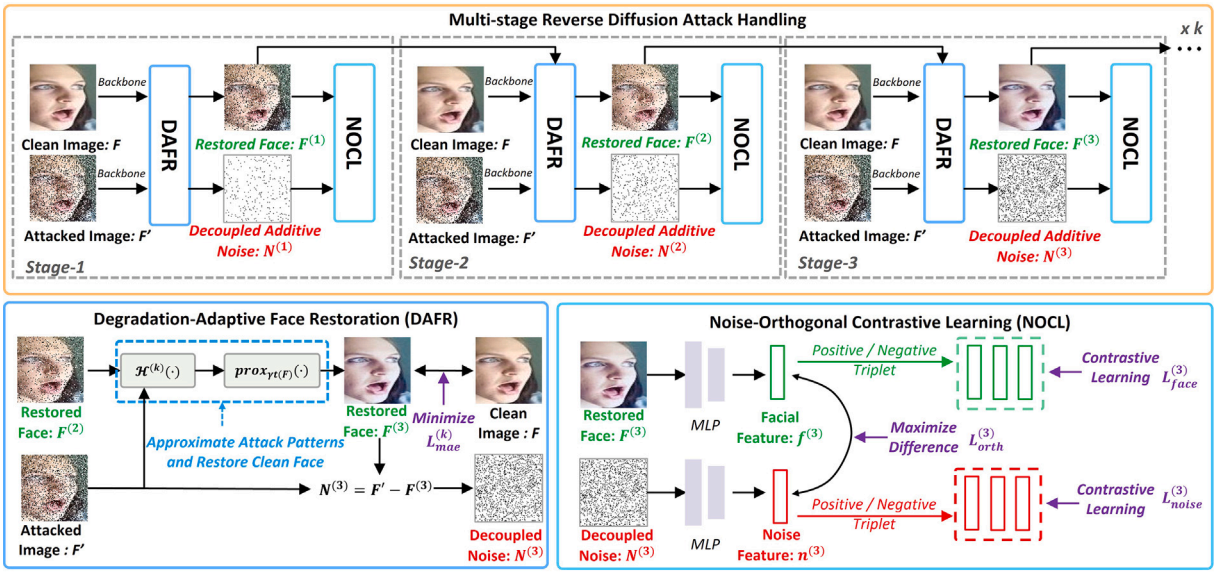


Fig. 2. The training pipeline of our proposed DAR-SFRL. DAR-SFRL first summarizes the two types of subtle semantic degradation caused by adversarial attacks into a unified degradation formula ($\mathbf{F}' = \mathbf{F} \cdot \mathbf{M} + \epsilon$) that includes structured geometric distortion and unstructured additive noise, and then learns complex and diverse degradation patterns through $\boldsymbol{k}$ stages of iterative refinement, gradually reducing the impact of degradation attacks. Each stage consists of Degradation-Adaptive Face Restoration (**DAFR**) and Noise-Orthogonal Contrastive Learning (**NOCL**), where DAFR addresses the degradation matrix M from structured distortions and NOCL further alleviate the influence of the additive noise ϵ from unstructured artifacts.

motivates the application of classical image degradation modeling [32] to analyze and decompose adversarially attacked images. Accordingly, we further decompose F' as:

$$F' = F \cdot M + \epsilon, \tag{2}$$

where M is a degradation matrix modeling structured distortions in key facial regions, and ϵ is an unstructured additive noise component. From the above, the semantic degradation introduced by the adversarial attack can be equivalently expressed as:

$$\delta = F' - F = (M - I) \cdot F + \epsilon, \tag{3}$$

where I is the identity matrix. This decomposition shows that δ consists of a structured distortion term $(M - I) \cdot F$ that affects essential facial structures, and an unstructured additive noise term ϵ.

Based explicitly on this unified degradation model Eq. (2), a straightforward goal of our DAR-SFRL is to recover the fine-grained details of F' and make high-dimensional features of F' and F as similar as possible through a multi-stage reverse diffusion iterative optimization process, thus enabling better resilience against the subtle semantic degradation caused by adversarial attacks during inference. As mentioned previously, to reverse the effects of the adversarial attack and make the attacked image F' close to the clean image F, we introduce two key strategies in each stage of DAR-SFRL. Firstly, we attempt to progressively reverse the degradation function and restore fine-grained image details. This primarily corresponds to handling structural distortions in key facial regions, which are controlled by the degradation matrix M in Eq. (2). We term the related technique Degradation-Adaptive Face Restoration (DAFR). Secondly, we devise a Noise-Orthogonal Contrastive Learning (NOCL) scheme to further deal with the additive noise ϵ in Eq. (2). After the restoration of clean facial image, we also obtain additive noise representations decomposed from the features of the original attacked image. We maximize the difference between this decomposed additive noise representation and the restored facial features so that the final facial features are better distinguished from noise representations. As a result, through the multi-stage joint training of DAFR and NOCL, our approach effectively alleviates the facial semantic degradation modeled by the unified degradation function in Eq. (2). We will discuss more details in the following sections.

3.2. Degradation-adaptive face restoration (DAFR)

SFRL generally shows poor robustness when faced with adversarial attacks. A fundamental reason is their difficulty in handling structural semantic distortions in key facial regions, which can be formalized using a degradation matrix M in Eq. (2), as it directly impacts the integrity of facial structural information. Therefore, we focus on mitigating the impact of M, as defined in Eq. (2), which accounts for these structural distortions. In our assumptions, adversarial attacks often cause varying degrees of structural distortions to key facial regions in unpredictable ways. These non-fixed, unpredictable degradation patterns make it difficult for us to construct comprehensive and appropriate M that covers all potential cases. Unlike previous studies that only deal with fixed degradation types, we treat M as a potential random variable and perform inference in a data-driven manner. To this end, we introduce Bayes' theorem [33] and try to model M pairs without strong prior assumptions, gradually reversing the degradation function and recovering fine-grained image details.

In particular, we infer the most likely clean image F from the observed data (the attacked image F') using the maximum a posteriori (MAP) principle, aiming to maximize the posterior probability $P(F|F')$. This results in our DAFR component, which models and reverses the degradation function based on Bayes' theorem. The DAFR treats the attacked image F' with structural distortion as an observable variable and reverses the degradation process by recovering the clean image F. More specifically, based on the MAP principle and the formulation of Eq. (2), we can prove that maximizing the negative logarithmic transformation of $P(F|F')$ is equivalent to solving the following non-convex optimization problem:

$$\arg\min_{F} \log P(F|F') = \arg\min_{F} \|F' - F \cdot M\|_2^2 + \gamma t(F), \tag{4}$$

where $\|F' - F \cdot M\|_2^2$ represents the data fidelity term, which ensures that the solution conforms to the degradation process, $t(F)$ is a regularization term approximating the prior distribution $P(F)$ over clean

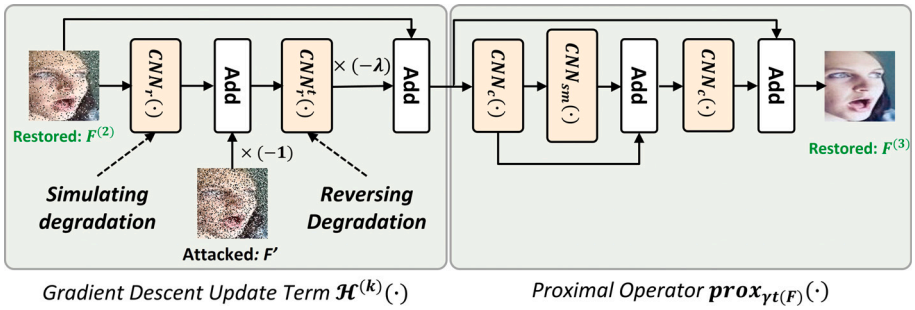


Fig. 3. The structure of $\mathcal{H}^{(k)}(\cdot)$ and $prox_{\gamma t(F)}(\cdot)$ in DAFR. $\mathcal{H}^{(k)}(\cdot)$ uses two independent residual blocks $CNN_r(\cdot)$ and $CNN_r^t(\cdot)$, to simulate the degradation matrix M and its transpose M^T, respectively, thus modeling the degradation pattern. In contrast, the proximal operator $prox_{\gamma t(F)}(\cdot)$ is approximated by a deep neural network comprising 3×3 convolutional layers $CNN_c(\cdot)$ and a sequential module $CNN_{sm}(\cdot)$. The latter integrates 3×3 convolutional layers, PReLU activation, and a channel attention mechanism to capture and model the structural and semantic degradation induced by adversarial attacks.

facial representations. We adopt the L_1 norm for its edge-preserving and sparsity-inducing properties, and γ is a hyperparameter that weights the regularization term $t(F)$.

Directly solving Eq. (4) for the global optimal solution is non-trivial because we assume that the degradation matrix M is unknown and the data fidelity term in Eq. (4) is non-convex. To solve this problem, we attempt to apply Taylor's expansion formula [34] with the gradient descent algorithm to gradually approximate the local optimal solution of Eq. (4). This approach reformulates the data fidelity term of Eq. (4) as a solvable multi-stage iterative refinement process comprising k stages. Accordingly, the update formula for the k-th stage of the above optimization can be described as follows:

$$F^{(k)} = \arg\min_F g(F^{(k-1)}) + \frac{1}{2\lambda}\|F - F^{(k-1)}\|_2^2 + \left\langle F - F^{(k-1)}, \nabla g(F^{(k-1)})\right\rangle + \gamma t(F), \tag{5}$$

where $F^{(k)}$ is the facial image restored at the k-th stage, $g(F^{(k-1)}) = \|F' - F^{(k-1)} \cdot M\|_2^2$, ∇ represents the gradient operator, weighted by the step size λ. Next, to make Eq. (5) even easier to solve, we follow the gradient descent update rule to combine the quadratic and gradient terms to simplify the expression and use the gradient information to update the local optimal solution $F^{(k)}$. Specifically, we simplify Eq. (5) as follows:

$$F^{(k)} = \arg\min_F \frac{1}{2}\|F - (F^{(k-1)} - \lambda\nabla g(F^{(k-1)}))\|_2^2 + \gamma t(F). \tag{6}$$

Then, to handle the non-differentiable regularization term $t(F)$, we apply the proximal operator [35] to Eq. (6) to adjust the solution during each iteration to ensure that it stays within a reasonable range and satisfies the regularization constraint. The updated form of Eq. (6) is as follows:

$$\mathcal{H}^{(k)}(F^{(k-1)}, F') = F^{(k-1)} - \lambda M^T(F^{(k-1)} \cdot M - F'), \tag{7}$$

$$F^{(k)} = prox_{\gamma t(F)}(\mathcal{H}^{(k)}(F^{(k-1)}, F')), \tag{8}$$

where $\mathcal{H}^{(k)}(F^{(k-1)}, F')$ is the conventional gradient descent update term that adjusts the current solution to approximate the ideal restored image. $prox_{\gamma t(F)}(\cdot)$ represents the proximal operator corresponding to the regularization term $t(F)$. Through this proximal operation, the updated solution is supposed to not only fit the data as closely as possible but also ensure that the image restoration process adheres to the assumption about prior probability $P(F)$.

In practice, we implement this process as a learnable DNN. As shown in Fig. 2, we combine Eqs. (7) and (8) with a DNN to construct a degradation-adaptive face restoration module (DAFR) at each stage of DAR-SFRL, which learns to understand non-fixed, unpredictable degradation patterns and simulates their impact on the clean image.

Based on the above derivations, our DAFR adopts a data-driven strategy to approximate M and M^T in Eq. (7), using two independent residual blocks $CNN_r(\cdot)$ and its transpose $CNN_r^t(\cdot)$ to model degradation and restoration operators. As shown in Fig. 3, $CNN_r(\cdot)$ and $CNN_r^t(\cdot)$ replace M and M^T in Eq. (7), enabling a data-driven implementation of Eq. (7). The updated form is as follows:

$$\mathcal{H}^{(k)}(F^{(k-1)}, F') = F^{(k-1)} - \lambda CNN_r^t(CNN_r(F^{(k-1)}) - F'). \tag{9}$$

Next, directly deriving an explicit $prox_{\gamma t(F)}(\cdot)$ of Eq. (8) is difficult because the regularization term $t(F)$ is usually nonlinear and non-differentiable. Therefore, we employ numerical approximation methods to approximate the proximal operator, utilizing DNNs to simulate the operation and model the impact of structural distortions on clean data. As shown in Fig. 3, by combining traditional optimization models with data-driven strategies, the updated form of Eq. (8) is as follows, forming the final update rule:

$$F^{(k)} = CNN_c(CNN_{sm}(CNN_c(\mathcal{H}^{(k)}(F^{(k-1)}, F'))) + CNN_c(\mathcal{H}^{(k)}(F^{(k-1)}, F'))) + \mathcal{H}^{(k)}(F^{(k-1)}, F'), \tag{10}$$

$$N^{(k)} = F' - F^{(k)}. \tag{11}$$

where $F^{(k)}$ and $N^{(k)}$ represent the reconstructed image and the decoupled structured distortion produced by DAFR at the k-th stage of DAR-SFRL, respectively. $CNN_{sm}(\cdot)$ represents a serialization module consisting of a 3×3 convolutional layer, a PReLU activation function, and a channel attention layer. $CNN_c(\cdot)$ represents another 3×3 convolutional layer. Subsequently, to constrain the progressive refinement process of DAFR, we employ the L_1 loss [36] to measure the consistency between the recovered facial image and the original clean image. Mathematically, the consistency loss is as follows:

$$L_{mae}^{(k)} = \|F^{(k)} - \mathbf{F}\|_1, \tag{12}$$

where $\|\cdot\|_1$ represents the L_1 loss.

Through the above iterative optimization, DAFR gradually refines the restored image by learning non-fixed, unpredictable degradation patterns and applying consistency loss constraints, enabling the DAFR network to robustly handle various structural distortions.

3.3. Noise-orthogonal contrastive learning (NOCL)

When facing semantic degradations caused by adversarial attacks, SFRL must handle not only structural distortions but also unstructured additive noise ϵ (as formulated in Eq. (2)). This noise, often entangled with fine-grained facial semantics such as micro-expression features, leads to latent-space interference that corrupts the learned identity

representations and degrades robustness. To further address this, we introduce Noise Orthogonal Contrastive Learning (NOCL) after DAFR at each stage of DAR-SFRL.

As shown in Fig. 2, NOCL consists of three loss functions: noise-orthogonal disentangling loss, face-robust contrastive loss, and noise-sensitive contrastive loss. In our formulation, the additional noise introduces subtle perturbations, making it difficult to directly remove it from the corrupted image. We thus introduce the noise orthogonal disentangling loss to separate unstructured additive noise from the reconstructed images in the feature space, so that perturbations from ϵ can be better exposed and eliminated after training. After separating the unstructured additive noise, we further use the face-robust and noise-sensitive contrastive losses to learn the disentangled facial and noise features in their respective feature spaces, separately. This decoupled learning strategy enables each component to operate independently within its own feature space, minimizing the unstructured semantic interference from various potential unstructured additive noise patterns on the restored facial features.

Noise-orthogonal Disentangling Loss: We first use a multilayer perceptron to project the decoupled reconstruction image $F^{(k)}$ and degraded pattern $N^{(k)}$ into high-dimensional feature spaces as $f^{(k)} = MLP(F^{(k)})$, $n^{(k)} = MLP(N^{(k)})$, where $MLP()$ represents a multilayer perceptron consisting of two fully connected layers. Then, to further separate unstructured additive noise, we aim to maximize the divergence $d^k(f^{(k)}, n^{(k)})$ between noise and reconstructed samples in the high-dimensional feature space. Formally, $d^k(f^{(k)}, n^{(k)})$ can be described as follows:

$$d^k(f^{(k)}, n^{(k)}) = \|f^{(k)} - n^{(k)}\|^2. \tag{13}$$

We can prove that maximizing the divergence $d^k(f^{(k)}, n^{(k)})$ is approximately equivalent to minimizing the orthogonality loss between $f^{(k)}$ and $n^{(k)}$. To this end, we introduce a noise-orthogonal disentangling loss, defined as follows:

$$L_{orth}^{(k)} = \frac{1}{R}\sum_{r=1}^{R} \|f^{(k)} \cdot n^{(k)}\|_2^2. \tag{14}$$

After separating the unstructured additive noise, we introduce face-robust contrastive loss and noise-sensitive contrastive loss to ensure that the model achieves high robustness in reconstructing facial features from images while maintaining sensitivity to various types of unstructured additive noise. The face-robust contrastive loss is designed to guide the model in learning more robust facial features, minimizing the influence of irrelevant information or disturbances. In contrast, the noise-sensitive contrastive loss focuses on capturing the diversity of unstructured additive noise, further enhancing the model's ability to distinguish between unstructured additive noise and genuine facial features. Specifically, we first apply stochastic data augmentation to transform any image F' within the same batch, resulting in two correlated views of the same face F_i' and F_j', and then select another image F_z' from the same batch. Next, we feed F_i', F_j' and F_z' into the DAFR, the multilayer perceptron, and the noise-orthogonal disentangling loss, obtaining the restored facial features $f_i^{(k)}$, $f_j^{(k)}$, and $f_z^{(k)}$, along with the unstructured additive noise features $n_i^{(k)}$, $n_j^{(k)}$, and $n_z^{(k)}$ at the k-th stage. Subsequently, we construct positive and negative sample pairs based on these features to calculate the face-robust contrastive loss and the noise-sensitive contrastive loss, respectively.

Face-robust Contrastive Loss: We perform face-robust contrastive learning on these face-only features $f_i^{(k)}$, $f_j^{(k)}$, and $f_z^{(k)}$ via the designed face-robust contrastive loss. Specifically, we take the restored facial features $f_i^{(k)}$ and $f_j^{(k)}$ of two related views of the same face image as positive samples and maximize the similarity between them. Meanwhile, we treat $f_z^{(k)}$ as a negative sample and minimize its similarity to the positive samples. Formally, the k-th stage of the face-robust contrastive loss can be written as:

$$L_{face}^{(k)}(f_i^{(k)}, f_j^{(k)}, f_z^{(k)}) = l_f(f_i^{(k)}, f_j^{(k)}) + l_f(f_j^{(k)}, f_i^{(k)}), \tag{15}$$

$$l_f(f_i^{(k)}, f_j^{(k)}) = -\log \frac{exp(\frac{sim(f_i^{(k)}, f_j^{(k)})}{\tau})}{\sum_{z=1,[z\neq i]} exp\left(\frac{sim(f_i^{(k)}, f_z^{(k)})}{\tau}\right)}, \tag{16}$$

where $sim(\cdot)$ is the cosine similarity of pairs. τ is the temperature parameter.

Noise-sensitive Contrastive Loss: We use the noise-sensitive contrastive loss designed for these unstructured additive noise features $n_i^{(k)}$, $n_j^{(k)}$, and $n_z^{(k)}$ to perform noise-sensitive contrastive learning. In detail, we treat the unstructured additive noise features $n_i^{(k)}$ and $n_j^{(k)}$ with the same degradation pattern as positive samples and minimize the distance between them in the feature space. Meanwhile, we treat the noise feature $n_z^{(k)}$ with different degradation pattern as a negative sample and maximize the distance between it and the positive samples in the feature space. Formally, the noise-sensitive contrastive loss at the k-th stage can be written as:

$$L_{noise}^{(k)}(n_i^{(k)}, n_j^{(k)}, n_z^{(k)}) = l_n(n_i^{(k)}, n_j^{(k)}) + l_n(n_j^{(k)}, n_i^{(k)}), \tag{17}$$

$$l_n(n_i^{(k)}, n_j^{(k)}) = -\log \frac{exp(\frac{sim(n_i^{(k)}, n_j^{(k)})}{\tau})}{\sum_{z=1,[z\neq i]} exp\left(\frac{sim(n_i^{(k)}, n_z^{(k)})}{\tau}\right)}, \tag{18}$$

where $sim(\cdot)$ is the cosine similarity of pairs. τ is the temperature parameter.

The combination of these loss functions provides a comprehensive noise decoupling strategy that explicitly separates unstructured additive noise from meaningful facial semantics in the latent space. By mitigating noise-induced interference, this strategy enables the model to focus on learning more robust and identity-preserving facial representations.

3.4. Overall learning objective

Overall, the proposed DAR-SFRL incorporates four types of objective functions: consistency loss, noise orthogonal disentangling loss, face-robust contrastive loss, and noise-sensitive contrastive loss. Therefore, the total loss function L is the weighted sum of $L_{mae}^{(k)}$, $L_{orth}^{(k)}$, $L_{noise}^{(k)}$, and $L_{face}^{(k)}$, which can be given by:

$$L = \sum_{k=1}^{K}\left(L_{mae}^{(k)} + L_{orth}^{(k)} + \alpha_n L_{noise}^{(k)} + \alpha_f L_{face}^{(k)}\right), \tag{19}$$

where K represents the total number of stages in DAR-SFRL. The parameters α_n and α_f are dynamic weights that adaptively balance the learning objectives based on the contributions of noise sensitivity and face sensitivity to face representation. Following prior work [22,37], we use dynamic weight averaging to obtain α_n and α_f during training.

4. Experiment and analysis

To evaluate the robustness and generalizability of the proposed model DAR-SFRL, we conduct experiments on face-related downstream tasks, including Facial Expression Recognition (FER), Face Recognition (FR), and AU detection. Compared to other self-supervised face representation and defense methods, DAR-SFRL demonstrates impressive robustness and generalizability to various adversarial attacks during the inference phase. Finally, we perform ablation experiments to verify the effectiveness of the proposed key modules.

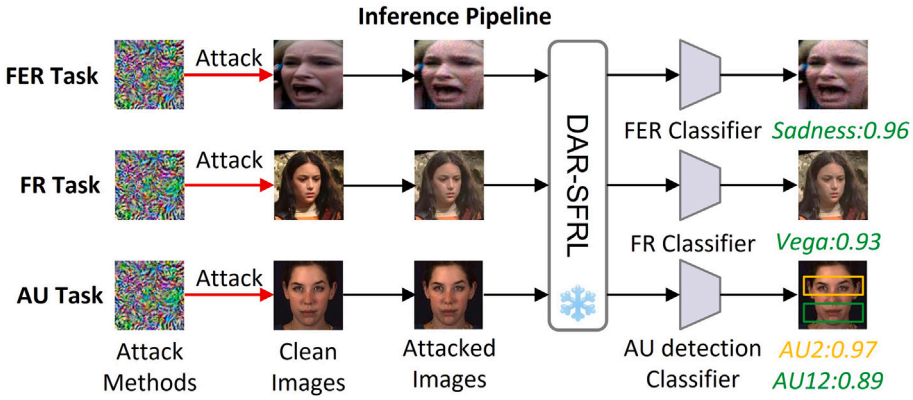


Fig. 4. The inference pipeline of our proposed DAR-SFRL. With the generate attacked samples as test data, we first use the pre-trained DAR-SFRL model to extract the facial representation, then employ a simple liner classifier to perform three face-related downstream tasks (FER, FR, AU detection), thereby verifying the robustness and generalization ability of DAR-SFRL.

4.1. Implementation details

Pre-training Phase: Our proposed model is implemented using the PyTorch framework and trained for 1000 epochs with the Adam optimizer ($\beta 1 = 0.9$, and $\beta 2 = 0.999$). The batch size and initial learning rate are set to 256 and 0.0001, respectively. We utilize cosine annealing to decrease the learning rate, with the temperature parameter τ set to 0.07. All models are trained and tested on an NVIDIA GTX 3090 GPU. The training process follows the data augmentation techniques and negative interpolation used in the baseline PCL [22].

Inference Phase: As shown in Fig. 4, to evaluate the robustness and generalizability of DAR-SFRL, we employ seven different adversarial attacks, including PGD [12], UPGD [27], BIM [11], MIFGSM [13], EOTPGD [38], DIFGSM [4], and NIFGSM[28]. Each of these attack methods is used to generate a set of attacked samples, which serve as the test datasets for evaluating our model's performance under different attack scenarios. Next, we use the pre-trained DAR-SFRL model to extract facial representations of the attacked samples, and then train three simple linear classifiers to perform three face-related downstream tasks (FER, FR, AU detection) respectively to verify the robustness and generalization ability of DAR-SFRL.

4.2. Datasets

Pre-training Datasets: During the pre-training phase, the VoxCeleb1 [39] and VoxCeleb2 [40] datasets introduce different levels of noise disturbances to represent the diverse adversarial attack patterns and are preprocessed using conventional data augmentation methods. Additionally, the aforementioned datasets lack annotations. They include a total of 299,085 video clips from approximately 7,000 speakers. Frames are extracted from the videos at a rate of 6 frames per second, cropped to center the face, and resized to a resolution of 64×64.

Downstream Task Datasets: For FER evaluation, we used the RAF-DB [41] dataset, which contains 12,271 training images and 3,068 test images. For FR evaluation, we used the CPLFW [42] dataset, which includes 3,000 pairs of frontal images with pose differences. For facial Action Unit (AU) detection, we used the BP4D [43] dataset, a spontaneous Facial Action Coding System (FACS) dataset containing 328 videos from 41 subjects (18 males and 23 females).

4.3. Overall performance on face-related downstream tasks

4.3.1. Performance on FER

Comparison of SFRL Methods: Table 1 presents the performance of different SFRL methods on the FER task under both clean (w/o attack) and adversarial attack conditions. The results showed that DAR-SFRL not only achieved competitive accuracy on clean samples but also demonstrated strong robustness against subtle semantic degradations introduced by various adversarial attacks. Specifically, on clean samples, DAR-SFRL achieved an accuracy of 67.14 %, ranking second only to PCL [22], which indicates its competitive feature learning capability in the absence of attacks. Furthermore, DAR-SFRL outperformed other methods when subjected to subtle semantic degradation caused by seven different adversarial attacks. For instance, under the NIFGSM [28] attack, DAR-SFRL maintained an accuracy of 66.33 %, reflecting only a 0.81 % drop compared to the clean condition. In comparison, the accuracy of PCL dropped from 67.66 % to 59.32 %, corresponding to an 8.34 % decline. This stark difference highlighted the superior capability of DAR-SFRL in simultaneously addressing both types of degradation, whereas other methods such as SimCLR[44] and MoCo [45] showed much greater performance drops (32.49 % and 15.12 %, respectively) under similar conditions. Overall,DAR-SFRL exhibited stronger robustness against semantic degradations than other SFRL methods.

Comparison of Defense Methods: Table 1 also reports the performance of DAR-SFRL and existing adversarial defense methods on the FER task, including RoCL [5], ACL [6], and the supervised method TRADES [46], all of which enhance model robustness through PGD-based adversarial training. In addition, DiffPure [7] separates the interference of adversarial attacks through adversarial purification. The results demonstrate that DAR-SFRL outperforms these methods on clean samples (w/o attack), achieving an accuracy of 67.14 %, significantly higher than RoCL, ACL, and TRADES. Under adversarial attacks, DAR-SFRL consistently achieves higher accuracy than all comparison methods. For instance, under the UPGD [27] attack, DAR-SFRL maintains an accuracy of 66.62 %, reflecting only a 0.52 % drop. Meanwhile, DiffPure, RoCL, ACL, and TRADES achieve accuracies of 47.04 %, 43.48 %, 15.48 %, and 37.32 %, corresponding to declines of 4.30 %, 11.61 %, 32.73 %, and 5.41 %, respectively, from their performance under clean conditions.

In summary, these results highlight the superior capability of DAR-SFRL in defending against semantic degradations introduced by adversarial attacks. Unlike existing methods that handle only one type of degradation when faced with such compound perturbations, DAR-SFRL offers a unified solution that addresses both, providing better protection in real-world scenarios.

4.3.2. Performance on FR

Comparison of SFRL Methods: Table 2 presents the performance of various SFRL methods on the FR task under both clean (w/o attack) and

Table 1
Evaluation for facial expression recognition (FER) on RAF-DB using accuracy (Acc) and drop accuracy (Drop Acc). The ↑ represents that larger values are better, while the ↓ represents that smaller values are better. The best results are in bold.

Methods		w/o Attacks	Adversarial Attacks													
			PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]		DIFGSM [4]		NIFGSM[28]	
		Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
SFRL	MAE [20]	46.38	23.60	22.60	24.93	21.45	20.70	25.68	24.54	21.84	23.60	22.60	30.54	15.84	34.55	11.83
	FRA [23]	52.24	38.59	13.65	38.59	13.65	36.05	16.19	36.57	15.67	33.51	18.73	32.69	19.55	33.60	18.64
	FaceCycle [14]	62.13	52.53	9.60	52.56	9.57	61.76	0.37	52.56	9.57	52.44	9.69	54.23	7.90	52.54	9.69
	MoCo [45]	63.36	48.53	14.83	48.66	14.70	62.03	1.33	48.66	14.70	48.53	14.83	48.50	14.86	48.24	15.12
	SimCLR [44]	65.52	33.56	31.96	33.24	32.28	32.10	33.42	32.28	33.24	33.98	31.54	33.95	31.57	33.03	32.49
	PCL [22]	**67.66**	55.37	12.29	54.66	13.00	54.43	13.23	54.66	13.00	55.38	12.28	57.27	10.39	59.32	8.34
	DAR-SFRL	67.14	**66.23**	**0.91**	**66.62**	**0.52**	**66.87**	**0.27**	**66.62**	**0.52**	**66.23**	**0.91**	**66.30**	**0.84**	**66.33**	**0.81**
Defense	TRADES [46]	42.73	40.13	2.60	37.32	5.41	37.81	4.92	36.83	5.90	37.87	4.86	37.68	5.05	38.33	4.40
	ACL [6]	48.21	30.40	17.81	15.48	32.73	14.64	33.57	15.48	32.73	14.99	33.22	14.64	33.57	24.71	23.50
	RoCL [5]	55.09	51.30	3.79	43.48	11.61	43.68	11.41	43.68	11.41	43.74	11.35	43.64	11.45	45.83	9.26
	DiffPure [7]	51.34	45.16	6.18	47.04	4.30	45.93	5.41	48.41	2.93	44.11	7.23	46.01	5.33	47.92	3.42
	DAR-SFRL	67.14	**66.23**	**0.91**	**66.62**	**0.52**	**66.87**	**0.27**	**66.62**	**0.52**	**66.23**	**0.91**	**66.30**	**0.84**	**66.33**	**0.81**

Table 2
Evaluation for facial recognition (FR) on CPLFW dataset using accuracy (acc) and drop accuracy (Drop Acc).

Methods		w/o Attacks	Adversarial Attacks													
			PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]		DIFGSM [4]		NIFGSM[28]	
		Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
SFRL	FRA [23]	**63.92**	57.59	6.33	54.11	9.81	60.13	3.79	42.41	21.51	58.23	5.69	48.73	15.19	47.78	16.14
	FaceCycle [14]	58.81	52.44	6.37	52.69	6.12	56.00	2.81	52.69	6.12	52.44	6.37	52.44	6.37	52.19	6.62
	MoCo [45]	54.43	46.52	7.91	45.89	8.54	46.52	7.91	45.89	8.54	46.52	7.91	45.25	9.18	45.57	8.86
	SimCLR [44]	55.38	43.99	11.39	38.61	16.77	41.14	14.24	38.61	16.77	43.98	11.40	50.95	4.43	46.20	9.18
	PCL [22]	59.81	51.27	8.54	51.58	8.23	52.22	7.59	51.58	8.23	53.16	6.21	57.28	2.53	56.33	3.48
	DAR-SFRL	63.61	**63.61**	**0.00**	**62.66**	**0.95**	**62.97**	**0.64**	**62.66**	**0.95**	**63.61**	**0.00**	**63.29**	**0.32**	**63.29**	**0.32**
Defense	TRADES [46]	59.87	58.60	1.27	55.76	4.11	56.71	3.16	56.39	3.48	57.97	1.90	55.76	4.11	56.96	2.91
	ACL [6]	52.30	51.30	1.00	49.37	2.93	47.78	4.52	51.27	1.03	49.45	2.85	48.70	3.60	49.77	2.53
	RoCL [5]	62.34	61.39	0.95	58.23	4.11	58.86	3.48	59.18	3.16	60.13	2.21	59.49	2.85	58.54	3.80
	DiffPure [7]	59.41	58.19	1.22	56.25	3.16	56.97	2.44	57.53	1.88	57.96	1.45	56.99	2.42	58.33	1.08
	DAR-SFRL	63.61	**63.61**	**0.00**	**62.66**	**0.95**	**62.97**	**0.64**	**62.66**	**0.95**	**63.61**	**0.00**	**63.29**	**0.32**	**63.29**	**0.32**

adversarial attack conditions. As a continuation of the FER evaluation, this task serves as a cross-task validation of DAR-SFRL's generalizability in face-related scenarios. In contrast to FER, which requires fine-grained sensitivity to facial expression variations, FR focuses on preserving identity-related features under semantic degradations. From the results, DAR-SFRL not only exhibits strong robustness in preserving identity consistency under clean conditions but also demonstrates high resistance to subtle semantic degradations caused by adversarial attacks.

In detail, DAR-SFRL achieved an accuracy of 63.61 % under clean conditions, which demonstrates its effective learning of identity representations. Under the BIM [11] attack, its performance slightly decreased to 62.97 %, with a minimal drop of only 0.64 %. Similarly, DAR-SFRL maintained stable performance when subjected to PGD [12], EOTPGD [38], and NIFGSM [28] attacks, with accuracy consistently close to 63.61 %. In contrast, the performance of other SFRL models showed more substantial declines. For instance, PCL [22], SimCLR [44], and FRA [23] experienced accuracy drops of 7.59 %, 14.24 %, and 3.79 %, respectively. MoCo [45] and FaceCycle [14] also saw performance reductions of 7.78 % and 2.81 %, respectively.

Comparison of Defense Methods: Next, we also reported the performance of DAR-SFRL and existing adversarial defense methods on the FR task in Table 2. The results indicated that existing defense methods such as DiffPure [7], RoCL [5], ACL [6], and the supervised method TRADES [46] struggled to effectively handle the semantic degradations caused by adversarial attacks, which limited their overall robustness. In contrast, DAR-SFRL was explicitly designed to simultaneously address semantic degradations, allowing it to preserve identity consistency even under adversarial attack conditions. For example, under the UPGD [27] attack, DAR-SFRL experienced only a 0.95 % drop in accuracy, demonstrating superior capability in preserving identity-relevant features compared to DiffPure, RoCL, ACL, and TRADES. By contrast, these methods suffered significantly larger drops in accuracy, ranging from 2.93 % to 4.11 %.

Overall, these performance advantages underscore DAR-SFRL's stronger adaptability to semantic degradations caused by adversarial attacks, making it a more robust and reliable solution for face recognition and significantly enhancing the overall security and stability of face recognition systems.

4.3.3. Performance on facial AU detection

Comparison of SFRL Methods: Table 3 compares the performance of different SFRL methods on the facial AU task under both clean and adversarial conditions. Unlike FER and FR tasks, AU detection requires finer recognition of facial actions, making it more vulnerable to small structural distortions and noise, which can lead to missed or incorrect detections. This high precision demand makes AU detection particularly sensitive to semantic degradations caused by adversarial attacks. The results in Table 3 show that DAR-SFRL exhibited remarkable robustness, maintaining strong performance even under the subtle semantic degradations induced by adversarial attacks. Specifically, under the UPGD [27] attack, DAR-SFRL achieved an F1 score of 51.20, reflecting only a 1.47 drop compared to the clean condition. When subjected to other attacks such as PGD [12], BIM [11], and MIFGSM [13], the declines in F1 score remained within the narrow range of 0.97 to 1.17, indicating a high level of overall performance stability. In contrast, the performance degradation of other SFRL methods was significantly more pronounced. Under the UPGD attack, the F1 scores of PCL [22], SimCLR [44], FRA [23], MoCo [45], and FaceCycle [14] dropped by 5.50, 7.12, 1.95, 7.33, and 5.78 points, respectively—substantially higher than the 1.47 drop observed for DAR-SFRL. Similar trends were observed under other adversarial attacks, further confirming that DAR-SFRL demonstrates superior robustness against semantic degradations compared to existing SFRL methods.

Comparison of Defense Methods: Next, we also compared the performance of DAR-SFRL with existing adversarial defense methods (such as RoCL [5], ACL [6], and TRADES [46]) on the facial AU detection task in Table 3. The results indicate that these defense methods, which relied on supervised adversarial training to enhance model stability under attacks, typically fail to effectively handle the semantic degradations caused by adversarial attacks. However, in AU detection, structural distortions and localized noise often occur simultaneously, forming compound attacks that limit the robustness of these methods. In contrast, DAR-SFRL achieved stronger generalization robustness by jointly modeling the semantic degradations introduced by adversarial attacks. For example, under the UPGD [27] attack, DAR-SFRL achieved an F1 score of 51.20, representing only a 1.47 drop from the clean condition. This was significantly better than the performance losses observed for RoCL (2.12), ACL (1.57), and TRADES (2.43). In summary, DAR-SFRL demonstrated exceptional robustness against the semantic degradations introduced by various adversarial attacks in AU detection tasks. It effectively identified and resisted adversarial interference, ensuring the accuracy and stability of AU detection under various challenging conditions.

4.4. Ablation study

4.4.1. Effect of DAFR and NOCL

Table 4 examines the effects of each component of DAR-SFRL on FER task performance, with a focus on robustness under both clean and adversarial conditions. The results indicated that, for the FER task, DAFR and NOCL modules played a crucial role in detecting subtle facial expression variations under adversarial attacks and ensuring the model's robustness. When DAFR was removed, the robustness of DAR-SFRL(C) significantly decreased under attacks, especially under PGD [12], BIM [11], and EOTPGD [38], with accuracy drops of 7.27 %, 9.65 %, and 9.42 %, respectively, compared to the full model DAR-SFRL. This highlighted the critical role of DAFR in reversing semantic degradation and preserving fine-grained facial semantics. In contrast, when NOCL was removed, DAR-SFRL(B) showed significant performance degradation both under clean and adversarial conditions, with accuracy dropping by over 1 %, which was greater than the accuracy drop of the full model DAR-SFRL. This emphasized the importance of NOCL in mitigating the interference of unstructured noise and maintaining discriminative features. These results suggest that DAFR and NOCL are key components in the robustness and performance of the DAR-SFRL model.

4.4.2. Effect of various M architectures in DAFR

Table 5 explores the impact of different designs of degradation matrix M in DAFR on the model performance. Given that adversarial attacks introduce unpredictable structural distortions, conventional fixed-form degradation matrices are insufficient to capture such variability. To address this, we designed M as a learnable, lightweight neural module that adaptively models the structural degradation induced by attacks. We evaluated several representative lightweight DNN architectures: (i) a **CNN** consisting of two 3×3 convolutional layers, each followed by ReLU and BatchNorm; (ii) a **ResNet** composed of two basic residual blocks with skip connections, each block containing two convolutional layers; (iii) a **LSTM** with two layers and a hidden dimension of 256; and (iv) a **Transformer** with two standard blocks, each equipped with 4 attention heads and a hidden dimension of 256. The results revealed substantial differences in these architectures' abilities to simulate structural semantic degradation and recover fine-grained details under adversarial attacks. Specifically, while CNNs are effective at modeling local spatial features, they are limited in capturing global context, which resulted in a 2.28 % accuracy drop in clean conditions compared to ResNet. The LSTM, though suitable for sequential modeling, is less capable of preserving spatial structure in images, which led to reduced robustness and accuracy drops ranging from 0.82 % to 2.32 % under attacks. Transformers achieved the highest accuracy (68.34 %) in clean conditions, with accuracy drops ranging from 2.01 % to 2.45 %. In contrast, ResNet struck a better balance between clean performance and robustness. Benefiting from its skip connection mechanism, it facilitated

Table 3
Evaluation for facial AU detection on BP4D dataset using F1 score (F1) and drop F1 score (drop F1).

Methods		w/o Attacks	Adversarial Attacks													
			PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]		DIFGSM [4]		NIFGSM[28]	
		F1(↑)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)	F1(↑)	Drop F1(↓)
SFRL	FRA [23]	42.90	41.48	1.42	40.95	1.95	41.33	1.57	41.16	1.74	39.36	3.54	40.74	2.16	38.95	3.95
	FaceCycle [14]	49.98	44.26	5.72	44.20	5.78	44.26	5.72	44.26	5.72	44.26	5.72	44.26	5.72	44.26	5.72
	MoCo [45]	51.55	49.90	1.65	44.22	7.33	49.64	1.91	49.19	2.36	49.90	1.65	49.61	1.94	48.64	2.71
	SimCLR [44]	51.35	47.56	3.79	44.23	7.12	47.58	3.77	47.35	4.00	47.58	3.77	47.59	3.76	47.14	4.21
	PCL [22]	51.13	47.92	3.21	45.63	5.50	47.73	3.40	45.69	5.44	48.15	2.98	47.69	3.44	46.75	4.38
	DAR-SFRL	**52.67**	**51.56**	**1.11**	**51.20**	**1.47**	**51.69**	**0.98**	**51.56**	**1.11**	**51.66**	**1.01**	**51.70**	**0.97**	**51.50**	**1.17**
Defense	TRADES [46]	52.23	50.51	1.72	49.80	2.43	51.02	1.21	51.06	1.17	50.14	2.09	50.11	2.12	50.95	1.28
	ACL [6]	52.00	49.97	2.03	50.43	1.57	50.71	1.29	49.84	2.16	50.19	1.81	50.71	1.29	50.15	1.85
	RoCL [5]	51.80	50.17	1.63	49.68	2.12	50.44	1.36	50.52	1.28	49.68	2.12	50.45	1.35	50.61	1.19
	DAR-SFRL	**52.67**	**51.56**	**1.11**	**51.20**	**1.47**	**51.69**	**0.98**	**51.56**	**1.11**	**51.66**	**1.01**	**51.70**	**0.97**	**51.50**	**1.17**

Table 4
Module ablation of DAR-SFRL with and without DAFR and NOCL for the FER task on the RAF-DB dataset.

Methods	DAFR	NOCL	w/o Attacks	Adversarial Attacks									
				PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]	
			Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
DAR-SFRL(A)	×	×	53.58	50.42	3.16	50.65	2.93	50.42	3.16	50.59	2.99	50.39	3.19
DAR-SFRL(B)	✓	×	50.88	49.64	1.24	49.77	1.11	49.61	1.27	49.77	1.11	49.64	1.24
DAR-SFRL(C)	×	✓	60.56	53.29	7.27	50.98	9.58	50.91	9.65	51.17	9.39	51.14	9.42
DAR-SFRL	✓	✓	**67.14**	**66.23**	**0.91**	**66.62**	**0.52**	**66.87**	**0.27**	**66.62**	**0.52**	**66.23**	**0.91**

Table 5
Performance of different degradation matrices M in DAFR for the FER task on the RAF-DB dataset.

Methods	w/o Attacks	Adversarial Attacks											
		PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]		DIFGSM [4]		NIFGSM[28]	
	Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
ResNet	67.14	**66.23**	**0.91**	**66.62**	**0.52**	**66.87**	**0.27**	**66.62**	**0.52**	**66.23**	**0.91**	**66.30**	**0.84**	**66.33**	**0.81**
CNN	64.86	64.54	0.32	64.66	0.20	64.44	0.42	64.60	0.26	64.50	0.36	64.41	0.45	64.53	0.33
LSTM	63.43	61.21	2.22	61.11	2.32	61.28	2.15	61.28	2.15	61.21	2.22	61.77	1.66	62.61	0.82
Transformer	**68.34**	66.07	2.27	66.21	2.13	66.14	2.20	66.33	2.01	65.89	2.45	65.99	2.35	66.19	2.15

effective information flow and retention throughout the network, enabling more stable recovery of fine-grained semantic details and more robust performance under diverse attack scenarios.

4.4.3. *Effect of various losses in NOCL*

Table 6 presents the impact of various loss functions in NOCL on model performance. NOCL consists of the noise-orthogonal disentangling loss (L_{orth}), the noise-sensitive contrastive loss (L_{noise}), and the facial-robust contrastive loss (L_{face}), and it is designed to mitigate semantic degradation caused by adversarial attacks, particularly non-structural additive noise. We conducted ablation experiments by progressively removing L_{orth}, L_{face}, and L_{noise}. The results demonstrated that these loss functions significantly affected the accuracy and robustness of DAR-SFRL under both clean and adversarial conditions, indicating that these losses play critical roles in enhancing the model's overall defense capability. Specifically, DAR-SFRL(a), which removed L_{orth}, resulted in a 2.6 % accuracy drop under clean conditions compared to the complete DAR-SFRL. Under adversarial attacks, its accuracy further decreased by 2.19 % to 3.49 % relative to its own accuracy in clean conditions. Building upon this, DAR-SFRL(b), which further removed L_{face}, exhibited a substantial 11.02 % accuracy decrease in clean conditions compared to DAR-SFRL(a), while the drop under adversarial attacks was relatively small—only 0.58 % to 1.05 %. Finally, DAR-SFRL(c), which removed all three losses, showed an additional 2.64 % accuracy drop in clean conditions compared to DAR-SFRL(b). Under adversarial conditions, its accuracy declined by 1.11 % to 1.27 %, which was greater than the drop observed in DAR-SFRL(b). In summary, the performance comparison between DAR-SFRL and its three ablated variants under both clean and adversarial conditions demonstrates that each loss component in NOCL plays a distinct role in robustness modeling. Their joint design and complementary effects enable the model to achieve a well-balanced trade-off between robustness and accuracy in challenging environments.

4.4.4. *Effect of iteration count k*

Table 7 presents the effect of varying the number of reverse diffusion iterations k on model performance. Without adversarial attacks, setting $k = 3$ achieved the highest accuracy (67.14 %), which was 5.15 %, 2.24 %, and 7.00 % higher than those obtained with $k = 2$, 4, and 5, respectively. These results suggest that a moderate iteration depth provides the best trade-off between semantic recovery and noise amplification. Under adversarial attacks, the influence of k is less pronounced. Once key semantic cues are sufficiently restored, further refinement yields diminishing returns in robustness. Based on these findings, we adopt $k = 3$ as the default setting throughout our experiments. This choice balances efficiency and representational fidelity, while promoting stable training and generalization across tasks.

4.4.5. *Complexity analysis*

Table 8 provides a computational complexity comparison of our DAR-SFRL and current defense attack methods. From the table, DAR-SFRL significantly outperforms the supervised TRADES [46] and self-supervised ACL [6] and RoCL [5] in inference time (5.97 s), FLOPs (0.04 G), and parameter size (0.03 M). Specifically, ACL and RoCL rely on adversarial training, which requires the continual generation of adversarial samples during training to enhance robustness. This mechanism drives the model architecture toward larger capacity to accommodate increasingly complex adversarial perturbations. Although adversarial sample generation is confined to the training phase, the resulting increase in model complexity still imposes a substantial computational burden during inference, with ACL and RoCL incurring inference times of 14.57 and 13.10 s and FLOPs of 1.34 G and 0.56 G, respectively. Moreover, the parameter counts of TRADES, ACL, and RoCL all approach 11 M, further increasing storage and memory demands. In contrast, DAR-SFRL eliminates the need for adversarial training and adopts a lightweight self-supervised framework

Table 6
Performance of different losses in NOCL for the FER task on the RAF-DB dataset.

Methods	L_{orth}	L_{face}	L_{noise}	w/o Attacks	Adversarial Attacks									
					PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]	
				Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
DAR-SFRL	✓	✓	✓	**67.14**	**66.23**	**0.91**	**66.62**	**0.52**	**66.87**	**0.27**	**66.62**	**0.52**	**66.23**	**0.91**
DAR-SFRL(a)	×	✓	✓	64.54	62.35	2.19	61.05	3.49	61.47	3.07	61.05	3.49	62.35	2.19
DAR-SFRL(b)	×	×	✓	53.52	52.27	1.05	52.67	0.65	52.27	1.05	52.27	1.05	52.74	0.58
DAR-SFRL(c)	×	×	×	50.88	49.64	1.24	49.77	1.11	49.61	1.27	49.77	1.11	49.64	1.24

Table 7
Performance of different iteration amounts k in DAR-SFRL for the FER task on the RAF-DB dataset.

Methods	w/o Attacks	Adversarial Attacks													
		PGD [12]		UPGD [27]		BIM [11]		MIFGSM [13]		EOTPGD [38]		DIFGSM [4]		NIFGSM[28]	
	Acc(↑)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)	Acc(↑)	Drop Acc(↓)
$k = 2$	61.99	61.83	0.16	61.93	0.06	61.86	0.13	61.93	0.06	61.83	0.16	61.80	0.19	61.73	0.26
$k = 3$	**67.14**	**66.23**	0.91	**66.62**	0.52	**66.87**	0.27	**66.62**	0.52	**66.23**	0.91	**66.30**	0.84	**66.33**	0.81
$k = 4$	64.90	64.15	0.75	64.22	0.68	64.19	0.71	64.22	0.68	64.16	0.74	64.19	0.71	64.29	0.61
$k = 5$	60.14	59.94	0.20	59.88	0.26	59.97	0.17	59.88	0.26	59.94	0.20	59.78	0.36	59.88	0.26

Table 8
Computational complexity comparison of our DAR-SFRL and **current defense attack methods**.

Methods	TRADES [46]	ACL [6]	RoCL [5]	DAR-SFRL
Inference time	10.02 s	14.57 s	13.10 s	**5.97 s**
FLoPs	0.74 G	1.34 G	0.56 G	**0.04 G**
Params.	11.30 M	11.30 M	11.17 M	**0.03 M**

enhanced by degradation-adaptive modeling, which ensures robustness while significantly reducing inference cost and model size. These advantages demonstrate DAR-SFRL's superior computational and resource efficiency.

4.5. Visualization

4.5.1. Visualization of DAFR and NOCL impact on feature distribution

Fig. 5 presents T-SNE visualizations of the feature space on the FER task, offering a qualitative perspective on how DAFR and NOCL contribute to the robustness of DAR-SFRL. Compared to the quantitative ablation analysis in Section 4.4.1, this analysis intuitively demonstrates the impact of each component on the distribution of emotion categories under adversarial conditions. As shown in Fig. 5, under the PGD attack, the sequential integration of NOCL and DAFR progressively improves the separation of emotion categories in the feature space. Specifically, Fig. 5(a) shows the baseline configuration without either module, where substantial category overlap and scattered feature points are evident. Fig. 5(b) and (c) demonstrate that introducing either NOCL or DAFR individually helps partially restore feature compactness and class boundaries. Notably, Fig. 5(d) reveals that the combined use of both modules leads to clearer separation among categories such as "Surprise," "Neutral," and "Happiness," effectively minimizing overlap. These visual patterns serve as qualitative evidence that DAFR and NOCL collaboratively enhance adversarial robustness by mitigating semantic degradation and suppressing noise-induced feature distortion.

4.5.2. Visualization of per category performance using confusion matrix

Fig. 6 visualizes the classification results of various SFRL methods under PGD [12] attacks using confusion matrices. As shown, DAR-SFRL maintains relatively stable and balanced performance across all emotion categories. In contrast, methods such as PCL [22], SimCLR [44], and MoCo [45] exhibit more severe misclassification patterns under the same adversarial conditions. Specifically, DAR-SFRL surpasses PCL by an average of approximately 15 % in six emotion categories, excluding "Surprise." While SimCLR performs slightly better than DAR-SFRL on the "Disgust" category, it suffers substantial performance drops in others, such as "Fear," where the accuracy gap reaches up to 47 %. MoCo is particularly vulnerable, trailing behind DAR-SFRL in most categories except "Neutral." These confusion matrix visualizations provide intuitive evidence of classification biases across different emotion categories, offering insights into each model's prediction weaknesses and potential areas for improvement under adversarial conditions.

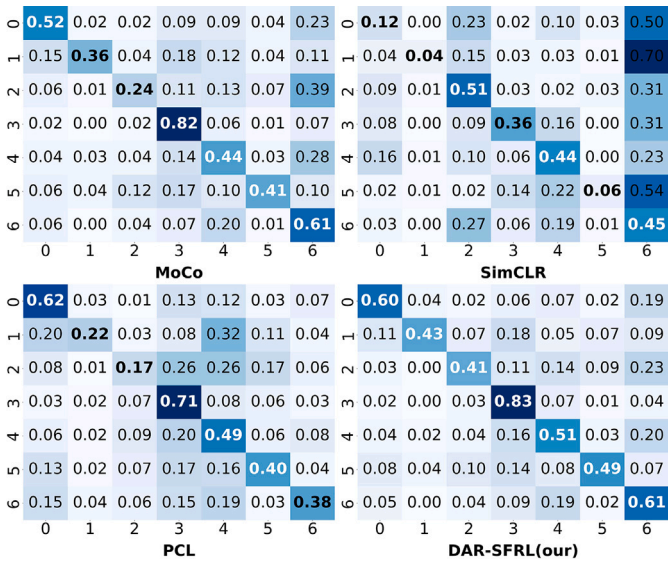


Fig. 6. The confusion matrix of DAR-SFRL and other SFRL methods on FER task under the PGD attack. The horizontal and vertical axes represent the predicted and true labels, respectively. The axis values correspond to the following emotions: 0: Surprise, 1: Fear, 2: Disgust, 3: Happiness, 4: Sadness, 5: Anger, 6: Neutral.

4.5.3. Visualization of feature distribution under structured and unstructured noise

Fig. 7 visualizes the distributions of original feature without attacks and the attacked feature using Kernel Density Estimation (KDE), aiming to assess the specific impact of structured distortion and unstructured noise introduced by adversarial attacks on the high-dimensional feature space. By comparing different variants of the DAR-SFRL model (with or without DAFR and NOCL), we intuitively analyzed the effect of these two types of semantic degradation on feature consistency. As shown in Fig. 7(a), the Baseline model did not incorporate either DAFR or NOCL, making it difficult to handle the structured distortion and unstructured noise introduced by adversarial attacks, resulting in a low overlap score (0.5357) between original feature without attacks and the attacked feature, indicating a significant feature shift. Considering that DAFR and NOCL are specifically designed to address structured distortion and unstructured noise respectively, Fig. 7(b) and (c) further present the performance of models incorporating only NOCL and only DAFR. The results showed that both types of semantic degradation introduced by adversarial attacks severely affect feature consistency: when only NOCL was used, only unstructured noise was mitigated with an overlap area of 0.7204; when only DAFR was applied, only structured distortion was resolved with an overlap area of 0.9476. In contrast, the complete DAR-SFRL model in Fig. 7(d), which integrated both DAFR and NOCL, achieved the highest feature overlap (0.9849), effectively aligning the original feature without attacks and the attacked feature. In summary,

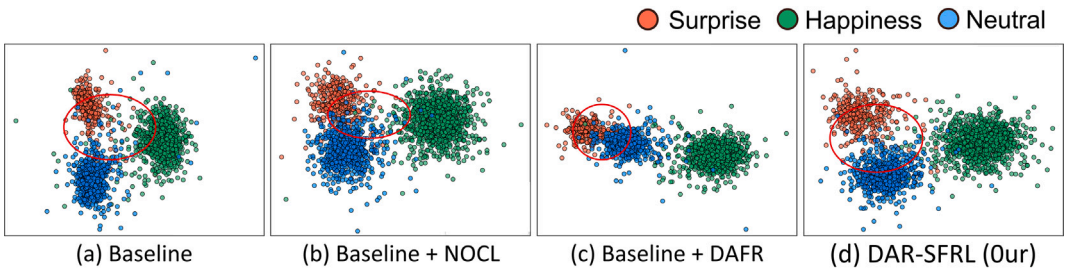


Fig. 5. Feature distribution visualization for the FER task using DAR-SFRL with and without DAFR and NOCL on the RAF-DB dataset.

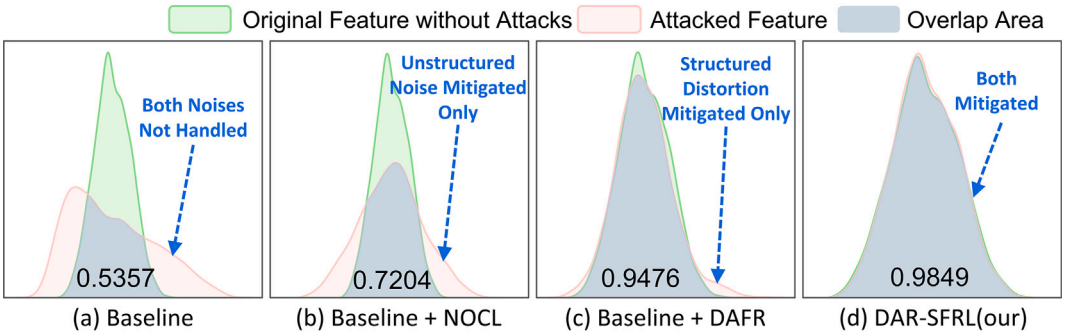


Fig. 7. Visualization of the data distribution relationships among the attacked feature, original feature without attacks, and their overlapping regions using kernel density estimation. Each subfigure corresponds to a variant of DAR-SFRL: (a) Baseline (both noises not handled), (b) +NOCL only (handles unstructured noise), (c) +DAFR only (handles structured distortion), and (d) full DAR-SFRL (handles both). Overlap scores are provided to quantify distribution similarity.

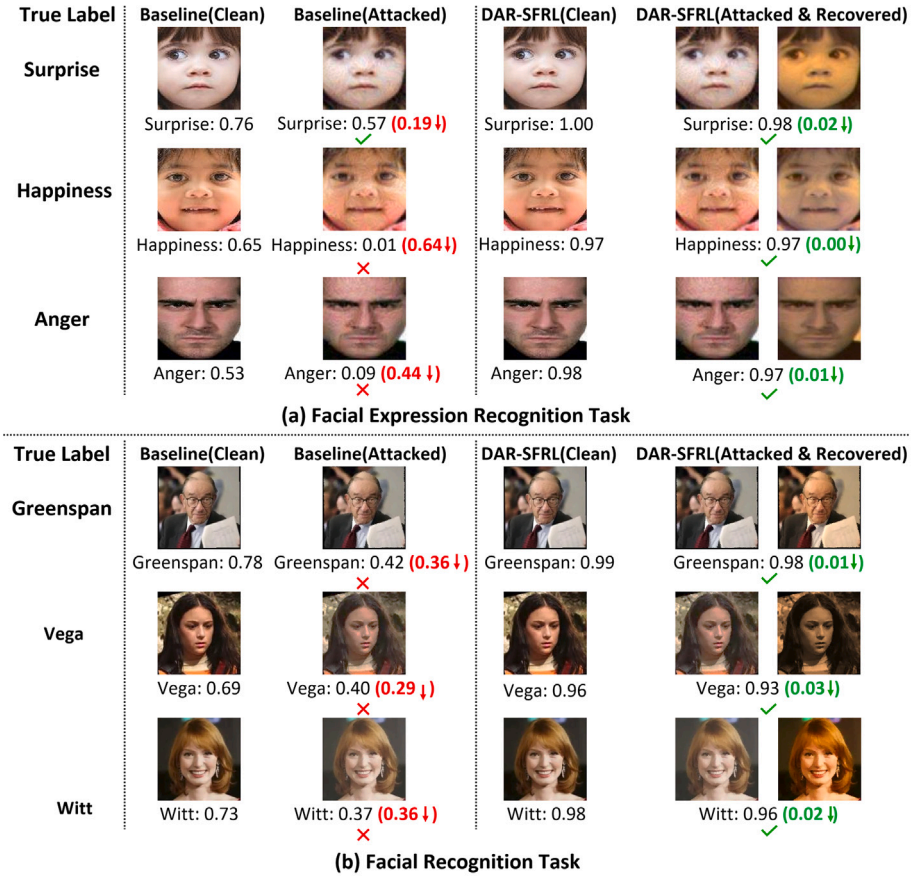


Fig. 8. Comparison of prediction probability scores on the baseline and our method under attacks. The values under the face images represent the prediction probability to the truth category. The ✓ and × symbols indicate the correct and wrong identification for the ground truth when under attacks, respectively.

these KDE visualizations clearly verify the effectiveness of DAR-SFRL in mitigating both structured and unstructured semantic degradation, and further reveal the independent and synergistic contributions of its components to enhancing feature robustness.

4.5.4. Visualization of prediction probability scores under adversarial attacks

Fig. 8 illustrates the model's robustness to adversarial attacks from the perspective of output probability scores in face-related downstream tasks. We fed both unattacked and adversarially perturbed samples into the Baseline and DAR-SFRL models for the FER and FR tasks, respectively. The visualization results show that the Baseline model exhibits significant fluctuations in predicted probability scores when exposed to adversarial attacks, often resulting in incorrect predictions or low confidence. In contrast, the DAR-SFRL model consistently maintains high confidence in the correct labels for both clean and attacked samples, demonstrating strong robustness to adversarial perturbations. Together with the findings in Section 4.5.3, these results further confirm that DAR-SFRL not only stabilizes feature-level representations but also improves the reliability of task-level decisions under adversarial attacks.

5. Conclusion

In this paper, we propose a novel framework, Degradation-based Attack-Robust Self-supervised Face Representation Learning (DAR-SFRL), which is designed to address the challenges posed by adversarial attacks on clean facial data. This method interprets adversarial attacks from the perspective of facial semantic degradation, and models the resulting subtle degradations as a composite degradation function that incorporates both structured geometric distortions and unstructured additive noise. Theoretically, this formulation enables a comprehensive and targeted defense against adversarial attacks. To systematically address these two components, we introduce two key modules in DAR-SFRL: Degradation-Adaptive Face Recovery (DAFR) and Noise-Orthogonal Contrastive Learning (NOCL). DAFR utilizes maximum a posteriori (MAP) estimation to progressively reverse the degradation function and recover fine-grained image details. It accurately models the relationship between degradation patterns and clean data and learns different degradation patterns during the gradual disentangling process.To further enhance robustness, NOCL incorporates a noise-orthogonal disentangling loss, a facial-robust contrastive loss, and a noise-sensitive contrastive loss. This ensures that the model not only discriminates between the additive noise of adversarial attacks and clean images but also generalizes well to various adversarial attacks. Through the synergistic training of DAFR and NOCL, DAR-SFRL effectively captures the perturbations caused by adversarial attacks, enabling a more precise understanding of adversarial attack patterns. This enhances the robustness of DAR-SFRL in face-related tasks, providing greater resilience against various adversarial attacks during inference. Despite the effectiveness of our approach, we find that DAR-SFRL still has some room for improvement. At present, the model lacks learning of the degradation process, which introduces uncertainty in the progressive recovery process. In future work, we plan to build a degradation-recovery framework based on physical information to provide richer and more comprehensive feature information for self-supervised face representation tasks in open environments.

CRediT authorship contribution statement

Ke Wang: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation. **Yuanyuan Liu:** Writing – review & editing, Validation, Supervision, Resources. **Chang Tang:** Validation, Supervision. **Kun Sun:** Validation, Supervision. **Yibing Zhan:** Validation, Supervision. **Zhe Chen:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

Data availability

I have shared the code in the abstract section of the manuscript .

References

[1] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, Z. Chen, Open-set video-based facial expression recognition with human expression-sensitive prompting, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5722–5731.
[2] Y. Liu, H. Zhang, Y. Zhan, Z. Chen, G. Yin, L. Wei, Z. Chen, Noise-resistant multimodal transformer for emotion recognition, Int. J. Comput. Vis. (2024) 1–21.
[3] Y. Yang, G. Zhang, D. Katabi, Z. Xu, Me-Net: towards effective adversarial robustness with matrix estimation, in: International Conference on Machine Learning, PMLR, 2019, pp. 7025–7034.
[4] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A.L. Yuille, Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.
[5] M. Kim, J. Tack, S.J. Hwang, Adversarial self-supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 2983–2994.
[6] Z. Jiang, T. Chen, T. Chen, Z. Wang, Robust pre-training by adversarial contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 16199–16210.
[7] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, A. Anandkumar, Diffusion models for adversarial purification, in: International Conference on Machine Learning, PMLR, 2022, pp. 16805–16827.
[8] J. Yoon, S.J. Hwang, J. Lee, Adversarial purification with score-based generative models, in: International Conference on Machine Learning, PMLR, 2021, pp. 12062–12072.
[9] M. Naseer, S. Khan, M. Hayat, F.S. Khan, F. Porikli, A self-supervised approach for adversarial robustness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 262–271.
[10] D. Deb, X. Liu, A.K. Jain, Faceguard: a self-supervised defense against adversarial face images, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–8.
[11] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, et al., Adversarial attacks and defences competition, in: The NIPS'17, Building Intelligent Systems, Springer, Competition, 2018.
[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.
[13] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.
[14] J.-R. Chang, Y.-S. Chen, W.-C. Chiu, Learning facial representations from the cycle-consistency of face, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9680–9689.
[15] H. Wang, M. Li, Y. Song, Y. Zhang, L. Chi, UCOL: unsupervised learning of discriminative facial representations via uncertainty-aware contrast, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 2510–2518.
[16] Y. Wang, Y. Liu, S. Zhou, Y. Huang, C. Tang, W. Zhou, Z. Chen, Emotion-oriented cross-modal prompting and alignment for human-centric emotional video captioning, in: IEEE Transactions on Multimedia, 2025.
[17] Y. Liu, N. Zhou, Y. Huang, S. Liu, L. Liu, W. Zhou, C. Tang, K. Wang, Beyond boundaries: hierarchical-contrast unsupervised temporal action localization with high-coupling feature learning, Pattern Recognit. (2025) 111421.
[18] Y. Liu, S. Feng, S. Liu, Y. Zhan, D. Tao, Z. Chen, Z. Chen, Sample-cohesive pose-aware contrastive facial representation learning, Int. J. Comput. Vis. (2025) 1–19.
[19] M. He, J. Zhang, S. Shan, X. Chen, Enhancing face recognition with self-supervised 3D reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4062–4071.
[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[21] Y. Wang, J. Peng, J. Zhang, R. Yi, L. Liu, Y. Wang, C. Wang, Toward high quality facial representation learning, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5048–5058.
[22] Y. Liu, W. Wang, Y. Zhan, S. Feng, K. Liu, Z. Chen, Pose-disentangled contrastive learning for self-supervised facial representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9717–9728.
[23] Z. Gao, I. Patras, Self-supervised facial representation learning with facial region awareness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2081–2092.
[24] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: generative or contrastive, IEEE Trans. Knowl. Data Eng. 35 (1) (2021) 857–876.
[25] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, et al., A cookbook of self-supervised learning, arXiv preprint arXiv:2304.12210, (2023).
[26] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, D. Tao, A survey on self-supervised learning: algorithms, applications, and future trends, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[27] M. Elsayed, A.R. Mahmood, Utility-based perturbed gradient descent: an optimizer for continual learning, in: OPT 2023: Optimization For Machine Learning, 2023.
[28] J. Lin, C. Song, K. He, L. Wang, J.E. Hopcroft, Nesterov accelerated gradient and scale invariance for adversarial attacks, in: International Conference on Learning Representations, 2020.
[29] E.-C. Chen, C.-R. Lee, Towards fast and robust adversarial training for image classification, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[30] D. Wang, W. Jin, Y. Wu, A. Khan, Atgan: adversarial training-based GAN for improving adversarial robustness generalization on image classification, Appl. Intell. 53 (20) (2023) 24492–24508.
[31] X. Mao, Y. Chen, R. Duan, Y. Zhu, G. Qi, X. Li, R. Zhang, H. Xue, et al., Enhance the visual representation via discrete adversarial training, Adv. Neural Inf. Process. Syst. 35 (2022) 7520–7533.
[32] B. Zha, S. Yang, J. Lei, Z. Xu, N. Ye, B. Feng, A dual branch attention network based on practical degradation model for face super resolution, Sci. Rep. 14 (1) (2024) 28064.
[33] R. Swinburne, Bayes' theorem, revue philosophique de La France et de l 194 (2) (2004).
[34] A. El-Ajou, O.A. Arqub, M. Al-Smadi, A general form of the generalized taylor's formula with some applications, Appl. Math. Comput. 256 (2015) 851–859.

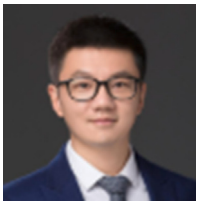
[35] R.T. Rockafellar, Monotone operators and the PROXIMAL POINT algorithm, SIAM J. Control Optim. 14 (5) (1976) 877–898.
[36] M.A. Error, Mean absolute error, Retrieved September 19 (2016) 2016.
[37] S. Liu, E. Johns, A.J. Davison, End-to-end multi-task learning with attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[38] X. Liu, Y. Li, C. Wu, C.-J. Hsieh, Adv-BNN: improved adversarial defense through robust Bayesian neural network, in: International Conference on Learning Representations, 2019.
[39] A. Nagrani, J.S. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, Interspeech (2017).
[40] J.S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, Interspeech, in, 2018.
[41] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2852–2861.
[42] T. Zheng, W. Deng, Cross-pose LFW: a database for studying cross-pose face recognition in unconstrained environments, Beijing University of Posts and Telecommunications, Tech. Rep. 5 (7) (2018) 5.
[43] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, A high-resolution spontaneous 3D dynamic facial expression database, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.
[44] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for Contrastive LEARNING of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[45] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[46] H. Zhang, Y. Yu, J. Jiao, E. Xing, L.E. Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International Conference on Machine Learning, PMLR, 2019, pp. 7472–7482.

Author biography

Ke Wang received the master's degree from Taiyuan University of Science and Technology in 2023. He is currently pursuing a Ph.D degree at China University of Geosciences (Wuhan). His research interests include computer vision and affective computing.

Yuanyuan Liu received the Ph.D. degree from Central China Normal University, Wuhan, in 2015. She is currently an Associate Professor at the School of Computer Science, China University of Geosciences (Wuhan). She has pub lished more than 40 peer-reviewed papers, including those in highly regarded journals and conferences such as CVPR, ACM MM, PR, INS, IEEE TGRS, FGRandICIP, etc. Her research interests include im age processing, computer vision, pattern recognition, affective computing, and multimodal interaction, etc.

Chang Tang received the PhD degree from Tianjin University, Tianjin, China, in 2016. He joined the AMRL Lab of the Uni versity of Wollongong between September 2014 and September 2015. He is currently an Associate Professor at the School of Computer Science, China University of Geosciences, Wuhan, China. He has published more than 50 peer-reviewed papers, in cluding those in highly regarded journals and conferences such as IEEE TPAMI, TKDE, TMM, THMS, SPL, AAAI, IJCAI, ICCV, CVPR, ACMM, ICME, etc. His current research interests include machine learning and computer vision.



Kun Sun received his PhD in June, 2017 from Institute of Image Recognition & Artificial Intelligence (IPRAI), School of Artificial Intelligence and Automation, Huazhong University of Science & Technology (HUST), Wuhan, China. He has published several articles on CVPR, ACPR, PRCV, IEEE TIP, IEEE TMM, Information Sciences, IEEE GRSL, IEEE SPL and so on. His research focuses on machine learning and computer vision algorithms, such as multi-view image matching, large scale Structure from Motion (SfM), 3D point cloud processing and medical image understanding.

Yibing Zhan received the bachelor's and doctor's degrees from the School of Information Science and Technology, University of Science and Technology of China, Hefei, China, in 2012 and 2018, respectively. From 2018 to 2020, he was an Associate Researcher with the School of Computer Science, Hangzhou Dianzi University, Hangzhou, China. Served as an algorithm scientist at JD Explore Academy from 2021 to 2025. He is currently employed at the School of Computer Science of Wuhan University. He has authored or coauthored many scientific papers in top conferences and journals such as NeurIPS, CVPR, ACM MM, ICCV. His research interests include graph generation, foundation model, and graph neural networks.

Zhe Chen received his PhD from the University of Sydney in 2019. He is now a lecturer at La Trobe University and is affiliated with the Cisco-La Trobe Centre for Artificial In telligence and Internet of Things. He is a highly cited researcher with regular publications in top conferences and journals such as CVPR, ECCV, ICCV, IJCV, and TIP. He has also participated in and won championship international computer vision competitions like ImageNet 2017 Detection from Video. His research focuses on visual understanding and deep learning applications on healthcare, robotics, space exploration, and so on.