applied sciences

MDPI

Article

ShieldNet: A Novel Adversarially Resilient Convolutional Neural Network for Robust Image Classification

Arslan Manzoor, Georgia Fargetta *, Alessandro Ortis and Sebastiano Battiato

Department of Mathematics and Computer Science, University of Catania, 95128 Catania, Italy; arslan.manzoor@phd.unict.it (A.M.); alessandro.ortis@unict.it (A.O.); sebastiano.battiato@unict.it (S.B.)
* Correspondence: georgia.fargetta@unict.it

Abstract

The proliferation of biometric authentication systems in critical security applications has highlighted the urgent need for robust defense mechanisms against sophisticated adversarial attacks. This paper presents ShieldNet, an adversarially resilient Convolutional Neural Network (CNN) framework specifically designed for secure iris biometric authentication. Unlike existing approaches that apply adversarial training or gradient regularization independently, ShieldNet introduces a synergistic dual-layer defense framework featuring three key components: (1) an attack-aware adaptive weighting mechanism that dynamically balances defense priorities across multiple attack types, (2) a smoothness-regularized gradient penalty formulation that maintains differentiable gradients while encouraging locally smooth loss landscapes, and (3) a consistency loss component that enforces prediction stability between clean and adversarial inputs. Through extensive experimental validation across three diverse iris datasets, MMU1, CASIA-Iris-Africa, and UBIRIS.v2, and rigorous evaluation against strong adaptive attacks including AutoAttack, PGD-100 with random restarts, and transfer-based black-box attacks, ShieldNet demonstrated robust performance, achieving 87.3% adversarial accuracy under AutoAttack on MMU1, 85.1% on CASIA-Iris-Africa, and 82.4% on UBIRIS.v2, while maintaining competitive clean data accuracies of 94.7%, 93.9%, and 92.8%, respectively. The proposed framework outperforms existing state-of-the-art defense methods including TRADES, MART, and AWP, achieving an equal error rate (EER) as low as 2.8% and demonstrating consistent robustness across both gradient-based and gradient-free attack scenarios. Comprehensive ablation studies validate the complementary contributions of each defense component, while latent space analysis confirms that ShieldNet learns genuinely robust feature representations rather than relying on gradient obfuscation. These results establish ShieldNet as a practical and reliable solution for deployment in high-security biometric authentication environments.

Keywords: adversarial defense; Convolutional Neural Networks; iris biometrics; gradient smoothing; adversarial training; robust deep learning; biometric security; AutoAttack evaluation

Academic Editor: JungHwan Oh

Received: 15 December 2025
Revised: 19 January 2026
Accepted: 20 January 2026
Published: 26 January 2026

1. Introduction

This paper introduces ShieldNet, a comprehensive adversarially resilient framework that addresses critical challenges through a principled integration of complementary defense mechanisms. Unlike existing approaches that apply defenses in isolation or via simple aggregation, ShieldNet adopts a synergistic dual-layer framework with the following contributions:

(1) Synergistic dual-layer defense. We integrate adversarial training with gradient-smoothing regularization through (i) an attack-aware adaptive weighting mechanism that

dynamically adjusts defense priorities based on attack characteristics during training, (ii) a smoothness-regularized gradient penalty that encourages smooth loss landscapes while maintaining differentiable gradients, explicitly avoiding the pitfalls of gradient obfuscation, and (iii) a consistency loss component that enforces prediction stability between clean and adversarial inputs, thereby addressing the robustness–accuracy trade-off.

(2) Rigorous evaluation under strong adaptive attacks. We evaluate ShieldNet using the community-standard AutoAttack benchmark (combining APGD-CE, APGD-DLR, FAB, and Square Attack), PGD-100 with 10 random restarts, and adaptive attacks specifically designed to bypass gradient obfuscation (BPDA, EOT). We further assess robustness via transfer-based black-box attacks using surrogate models to verify that improvements reflect genuine robustness rather than gradient obfuscation.

(3) Multi-dataset validation and cross-dataset generalization. We conduct extensive experiments on three iris datasets (MMU1, CASIA-Iris-Africa, and UBIRIS.v2), which differ in demographic coverage, acquisition conditions, and image quality. Rigorous cross-dataset generalization experiments are performed with proper subject-disjoint train/validation/test splits to ensure no data leakage and to demonstrate a true generalization capability.

(4) Empirical and intuitive analysis of the combined defenses. The provided empirical evidence and intuitive justification explain why the combination of adversarial training and a smoothness-regularized gradient penalty yields superior robustness through loss landscape visualization, gradient norm analysis, and latent space examination via t-SNE. These analyses suggest that ShieldNet learns robust feature representations rather than relying on gradient obfuscation. It should be noted that our theoretical analysis serves as intuitive motivation rather than formal guarantees.

(5) Practical deployment assessment. We analyze computational cost, memory footprint, and inference latency, showing that ShieldNet satisfies real-time constraints for biometric authentication while maintaining robust security properties.

(6) Comparative benchmarking. Extensive comparative benchmarking provides thorough comparisons with state-of-the-art defense methods, including standard adversarial training, TRADES, MART, AWP, and recent Vision Transformer-based approaches, demonstrating consistent improvements across multiple evaluation metrics under identical rigorous attack protocols. Texture-based visual representations have long been recognized as highly discriminative, yet are particularly sensitive to local perturbations. This intrinsic sensitivity makes robustness a fundamental requirement when such representations are employed in security-critical applications, including biometric recognition [1].

The remainder of this paper is organized as follows. Section 2 provides a focused review of related work in adversarial attacks, defense mechanisms, and biometric security. Section 3 presents our proposed ShieldNet methodology, including detailed architectural design, mathematical formulations, and intuitive analysis. Section 4 describes our experimental setup, datasets, evaluation protocols, and attack implementations. Section 5 presents comprehensive results, including AutoAttack evaluation, ablation studies, and latent space analysis. Section 6 discusses the implications, limitations, and potential vulnerabilities of our approach. Finally, Section 7 concludes this paper and outlines future research directions.

2. Related Work

This section reviews the key developments in adversarial attacks, defense mechanisms, and biometric security that directly inform our work. The field of adversarial machine learning has witnessed rapid evolution since the seminal work of Szegedy et al. [2], who first demonstrated the existence of adversarial examples in deep neural networks. Subsequent research has revealed the widespread nature of this vulnerability across various domains

and architectures, fundamentally challenging the deployment of deep learning in security-critical applications [2–4].

2.1. Adversarial Attack Methodologies

Modern adversarial attacks can be broadly categorized into several classes based on their threat models, optimization strategies, and access requirements. Gradient-based attacks leverage gradient information to generate adversarial perturbations efficiently. The Fast Gradient Sign Method (FGSM) [5] represents one of the earliest and most influential gradient-based attacks, generating adversarial examples through the following single gradient computation step:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

where x represents the original input, ϵ controls the perturbation magnitude, and $J(\theta, x, y)$ denotes the loss function. While computationally efficient, FGSM produces suboptimal adversarial examples due to its single-step linearization assumption.

Iterative Attacks: more sophisticated attacks employ iterative optimization procedures to generate stronger adversarial examples. Projected Gradient Descent (PGD) [6] extends FGSM by applying multiple gradient steps with projection onto the feasible perturbation space:

$$x_{t+1} = \Pi_{x+S}(x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_t, y))) \tag{2}$$

where Π_{x+S} represents projection onto the ϵ-ball constraint set S centered at the original input and α denotes the step size. PGD with sufficient iterations and random restarts is widely considered the strongest first-order attack and serves as a standard benchmark for adversarial robustness evaluation. Recent work established that PGD with 40–100 iterations and multiple random restarts provides reliable robustness assessment [7].

Optimization-Based Attacks: the Carlini and Wagner (C&W) attack [8] formulates adversarial example generation as a constrained optimization problem:

$$\min_{\delta} \|\delta\|_p + c \cdot f(x + \delta, y) \tag{3}$$

where $f(\cdot)$ is designed such that $f(x + \delta, y) \leq 0$ implies successful misclassification and c balances perturbation magnitude against attack success. C&W attacks often achieve higher success rates than PGD, particularly against defenses that rely on gradient obfuscation.

Recent forensic studies show that defenses tailored to specific generative mechanisms often fail across evolving attack paradigms; for instance, GAN-trained detectors exhibit limited robustness against diffusion models, underscoring the need for defenses promoting intrinsic representation robustness rather than relying on attack-specific artifacts [9,10]. Recent research has prioritized adaptive strategies designed to circumvent specific defenses [11]. The industry-standard AutoAttack [12] offers reliable, parameter-free evaluation by combining APGD-CE, APGD-DLR, FAB, and Square Attack. Additionally, methods like BPDA and EOT explicitly target gradient obfuscation and input transformations. In gradient-free settings, attackers rely on query-based estimation or transfer attacks via surrogate models [13], leveraging cross-model transferability to compromise deployed systems, i.e., black-box attacks.

2.2. Adversarial Defense Mechanisms

This study focuses on the defense approaches most relevant to our work, i.e., adversarial training variants and gradient regularization methods.

Adversarial training represents one of the most effective and widely studied defense strategies, incorporating adversarial examples during the training process to im-

prove model robustness. The standard adversarial training objective, formalized by Madry et al. [14], can be expressed as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|\leq\epsilon} \mathcal{L}(\theta, x+\delta, y) \right] \tag{4}$$

where $\mathcal{D}$ represents the training data distribution and $\mathcal{L}$ denotes the loss function. This min–max formulation trains the model to minimize loss under worst-case perturbations within the allowed ϵ-ball. TRADES: Zhang proposed TRADES (Tradeoff-inspired Adversarial Defense via Surrogate-loss minimization) [15], which explicitly decomposes the robust optimization objective into natural accuracy and boundary robustness terms:

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\mathcal{L}(f_\theta(x), y) + \beta \max_{\|\delta\|\leq\epsilon} \mathcal{L}(f_\theta(x), f_\theta(x+\delta)) \right] \tag{5}$$

where β controls the trade-off between clean accuracy and adversarial robustness. TRADES explicitly models the robustness–accuracy trade-off, while MART [16] focuses on misclassified examples, and AWP [17] improves generalization by perturbing model weights to flatten the loss landscape. Regarding gradients, regularization techniques seek smoother loss landscapes [18]. Crucially, one must distinguish legitimate smoothing, which maintains differentiability [19], from gradient obfuscation, which offers false security easily bypassed by adaptive attacks [20]. Alternative approaches include certified defenses, which offer provable guarantees [21] but often sacrifice clean accuracy. Similarly, studies on visually ambiguous tasks link performance degradation to unstable decision boundaries, reinforcing the importance of learning stable representations [22,23]. Finally, architectural shifts to Vision Transformers (ViTs) [24] introduce global attention mechanisms; however, they remain vulnerable without adversarial training.

2.3. Biometric System Security

The security of biometric systems remains a priority due to widespread deployment in critical applications [25,26]. While traditional research has focused on presentation attacks (spoofing) and template security, adversarial machine learning has introduced new threat vectors requiring specialized defenses [27]. Deep learning-based iris recognition, despite achieving extremely low error rates ($<10^{-6}$) [28] due to complex texture patterns and stability, inherits the adversarial vulnerabilities common to neural networks.

Recent studies have confirmed the susceptibility of various biometric modalities to adversarial attacks [29,30], generally categorized into the following:

- Impersonation Attacks:crafting perturbations to match an attacker's biometric to a legitimate user's template;
- Obfuscation Attacks: preventing recognition to evade surveillance or cause denial of service;
- Privacy Attacks: extracting sensitive information or reconstructing templates;
- Universal Perturbations: using image-agnostic patterns effective across multiple subjects.

Beyond digital threats, physical attacks, such as patterned contact lenses or adversarial eyeglass frames, have demonstrated real-world feasibility against iris systems [4]. Several works have examined these vulnerabilities across modalities [31–33], with surveys documenting vectors for face verification [34,35] and broader computer vision tasks [36–38]. Particular attention has been paid to safety-critical domains like autonomous vehicles [39] and security systems [40,41]. Standardized protocols like AutoAttack [42] and RobustBench [43] now enable rigorous assessment, driving adaptive defense mechanisms [44,45]. Key adversarial training methods include TRADES [46] (balancing accuracy and robust-

ness), MART [47] (focusing on misclassified examples), and AWP [48] (using weight perturbation). These have been refined by helper-based training [49] and dual regularization [50]. However, gradient masking can create a false sense of security [51], necessitating adaptive evaluations [52] and awareness of gradient imbalances [53]. In the iris domain, while deep learning has driven progress [54–57], it introduces distinct security challenges [58]. Recent efforts address these through liveness detection [59], adversarial threat mitigation in cybersecurity [60], and secure implementations for IoT [61].

2.4. Limitations of Existing Approaches

Despite significant progress in adversarial defense research, critical limitations persist, which motivates our work. First, most defense mechanisms face a severe trade-off between robustness and clean accuracy, with drops of 10–20% often reported in the literature, which hinders deployment in high-stakes applications. Second, evaluation rigor remains inconsistent. Many studies test defenses against weak configurations (e.g., PGD with few iterations), leading to overestimated robustness claims that fail under stronger protocols like AutoAttack. Similarly, defenses relying on gradient masking provide only illusory robustness, easily bypassed by adaptive strategies such as BPDA, EOT, or transfer-based attacks. Third, evaluation scope is frequently limited to single datasets or specific attack types, leaving cross-dataset generalization and demographic diversity unaddressed. As broad surveys indicate, deep models are highly sensitive to dataset bias and distribution shifts, necessitating robustness strategies beyond standard optimization [62]. Finally, practical constraints are often overlooked. Advanced defenses frequently incur prohibitive computational costs, limiting real-time utility. Furthermore, general-purpose defenses may not meet unique biometric requirements, such as extremely low false acceptance rates and template security. The theoretical underpinnings of defense combinations also remain underexplored, complicating principled design. ShieldNet addresses these gaps through a synergistic dual-layer framework validated via rigorous adaptive attacks, cross-dataset testing, and analyses of defense mechanisms.

3. Proposed Methodology

This section presents a comprehensive description of our proposed ShieldNet framework, including architectural design, mathematical formulations, intuitive analysis, and implementation details. Our approach addresses the limitations of existing defense mechanisms through a synergistic dual-layer strategy specifically designed for iris biometric authentication systems.

3.1. Dataset Description and Preprocessing

Our comprehensive evaluation utilizes three diverse iris datasets, each providing unique characteristics essential for a thorough robustness assessment. Table 1 summarizes the key properties of each dataset. The Multimedia University (MMU1) Iris Database serves as our controlled-environment dataset, consisting of 460 high-quality iris images from 46 unique individuals (5 images per eye per subject), captured under standardized near-infrared (NIR) lighting conditions using professional iris acquisition equipment. The CASIA-Iris-Africa dataset represents the largest and most demographically diverse component of our evaluation, containing 28,717 images from 1023 African subjects. This dataset is particularly valuable for assessing model generalization across demographic variations and diverse iris pigmentation patterns. The UBIRIS.v2 dataset provides crucial real-world validation through 11,102 images captured in unconstrained-visible-light environments. This dataset includes challenging conditions such as motion blur, specular reflections, partial occlusions, and varying distances that mirror practical deployment

scenarios. The MMU1 Iris database, as depicted in Figure 1, consists of iris images captured under controlled conditions, providing a standardized benchmark for evaluating biometric recognition systems. To ensure rigorous evaluation and prevent data leakage, subject-disjoint partitioning is employed, where all images from a given subject appear exclusively in one partition. For each dataset, a subject-disjoint split is adopted, allocating 70% of subjects for training, 15% for validation, and 15% for testing. This prevents the model from exploiting subject-specific features learned during training when evaluated on test subjects. Importantly, all data augmentation is applied exclusively to the training partition after splitting to prevent information leakage.

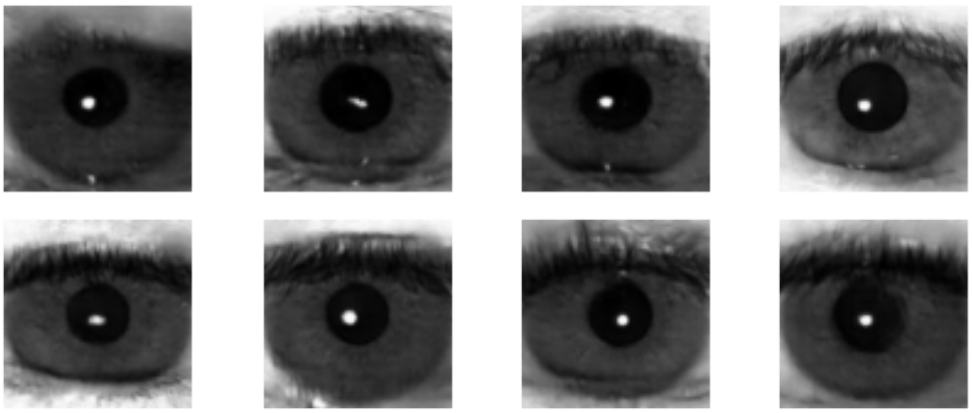

Figure 1. Sample images from the MMU1 Iris database showing unique iris patterns acquired under controlled conditions for biometric recognition.

Table 1. Summary of iris datasets used in this study.

Dataset *	Images	Resolution	Lighting	Environment	Demographics	Challenge
MMU1	460	320 × 240	NIR controlled	Indoor	Asian	Low
CASIA-Africa	28,717	640 × 480–1024 × 768	Natural/Artificial	Mixed	African	Moderate
UBIRIS.v2	11,102	800 × 600	Natural (varied)	Unconstrained	Caucasian	High

* Some datasets may have additional variations or preprocessing applied.

Preprocessing Pipeline: A multi-stage preprocessing pipeline is employed, combining traditional computer vision techniques with deep learning–based refinement, as detailed in Algorithm 1.

Algorithm 1 Advanced Iris Preprocessing Pipeline

Require: Raw iris image I_{raw}
Ensure: Normalized iris image I_{norm}
$I_{gray} \leftarrow \text{ConvertToGrayscale}(I_{raw})$
$I_{enhanced} \leftarrow \text{CLAHE}(I_{gray})$ ▷ Contrast-limited adaptive histogram equalization
$(x_p, y_p, r_{pupil}) \leftarrow \text{DetectPupil}(I_{enhanced})$ ▷ Circular Hough transform
$(x_i, y_i, r_{iris}) \leftarrow \text{DetectIrisBoundary}(I_{enhanced}, x_p, y_p)$
$mask \leftarrow \text{GenerateAnnularMask}(x_p, y_p, r_{pupil}, r_{iris})$
$mask \leftarrow \text{RemoveEyelidOcclusions}(mask, I_{enhanced})$
$I_{segmented} \leftarrow I_{enhanced} \odot mask$
$I_{polar} \leftarrow \text{RubberSheetTransform}(I_{segmented}, r_{pupil}, r_{iris})$ ▷ Daugman's method
$I_{norm} \leftarrow \text{Resize}(I_{polar}, 512 \times 64)$
$I_{norm} \leftarrow (I_{norm} - \mu)/\sigma$ ▷ Z-score normalization
return I_{norm}

Data Augmentation Strategy: To enhance model robustness and prevent overfitting, the implemented comprehensive data augmentation was applied only to the training set:

- Geometric Transformations: Rotation ($\pm 5°$), translation (± 10 pixels), scaling (0.9–1.1×)
- Photometric Variations: brightness adjustment ($\pm 20\%$), contrast modification (0.8–1.2×), gamma correction (0.8–1.2)
- Noise Injection: Gaussian noise ($\sigma = 0.01$), salt-and-pepper noise (0.5% density)
- Blur Simulation: motion blur (1–3 pixel kernel), Gaussian blur ($\sigma = 0.5$–1.0)
- Occlusion Simulation: random rectangular occlusions (5–15% area) to simulate eyelid/eyelash interference

3.2. Design Principles and Intuitive Motivation

ShieldNet is designed based on the following core principles, each grounded in an intuitive understanding of adversarial robustness. Unlike prior approaches that apply adversarial training (AT) and gradient regularization independently, ShieldNet integrates these mechanisms synergistically. The intuitive motivation stems from the observation that AT alone expands the robust decision boundary but may create irregular loss landscapes, while gradient smoothing alone reduces attack effectiveness but does not improve the underlying feature representations. Our combined approach leverages AT to learn robust features while gradient smoothing stabilizes the loss landscape, preventing the sharp gradients that enable efficient adversarial optimization. Traditional adversarial training uses a fixed attack during training (typically PGD), which can lead to overfitting to specific attack characteristics. ShieldNet employs an adaptive weighting mechanism that dynamically balances multiple attack types based on their current effectiveness, ensuring robust generalization across diverse attack vectors. A critical design consideration is the distinction between our approach and problematic gradient obfuscation techniques. Rather than hiding gradients through non-differentiable operations (which creates false security), our smoothness regularization maintains differentiable gradients while penalizing sharp loss curvature. This ensures that gradient-based attacks remain possible but are less effective due to the flattened loss landscape, providing genuine robustness that withstands adaptive attacks. Prediction consistency between clean and adversarial inputs is enforced, explicitly addressing the robustness–accuracy trade-off by encouraging the model to produce similar outputs under small input perturbations. This architecture is specifically optimized for iris biometric data characteristics, including its high-frequency texture patterns, circular geometry, and the critical requirement for very low false acceptance rates in security applications.

3.3. ShieldNet Architecture

Figure 2 illustrates the comprehensive architecture of ShieldNet, designed to achieve robustness against adversarial attacks while maintaining a high clean accuracy. The architecture integrates three key components: (1) a feature extraction backbone optimized for iris texture patterns, (2) an adversarial training module with attack-aware weighting, and (3) a smoothness-regularized gradient penalty mechanism. The backbone consists of four convolutional blocks with progressively increasing filter counts (32, 64, 128, 256), designed to capture hierarchical iris texture features, from fine-grained local patterns to global structural information. Table 2 provides detailed specifications. ShieldNet refers to the complete training framework, not just the CNN architecture; all baseline comparisons in this paper use an identical backbone architecture to ensure fair evaluation. The use of larger kernels (7×7, 5×5) in early layers is intentional for iris recognition, as iris textures contain important medium-frequency patterns that benefit from larger receptive fields. Batch normalization after each convolutional layer stabilizes training and has been

shown to improve adversarial robustness by reducing internal covariate shifts. Global average pooling replaces fully connected layers for spatial feature aggregation, reducing parameters and improving generalization. The primary contribution of ShieldNet lies in its principled integration of adversarial training and gradient smoothing, designed to provide complementary robustness benefits.

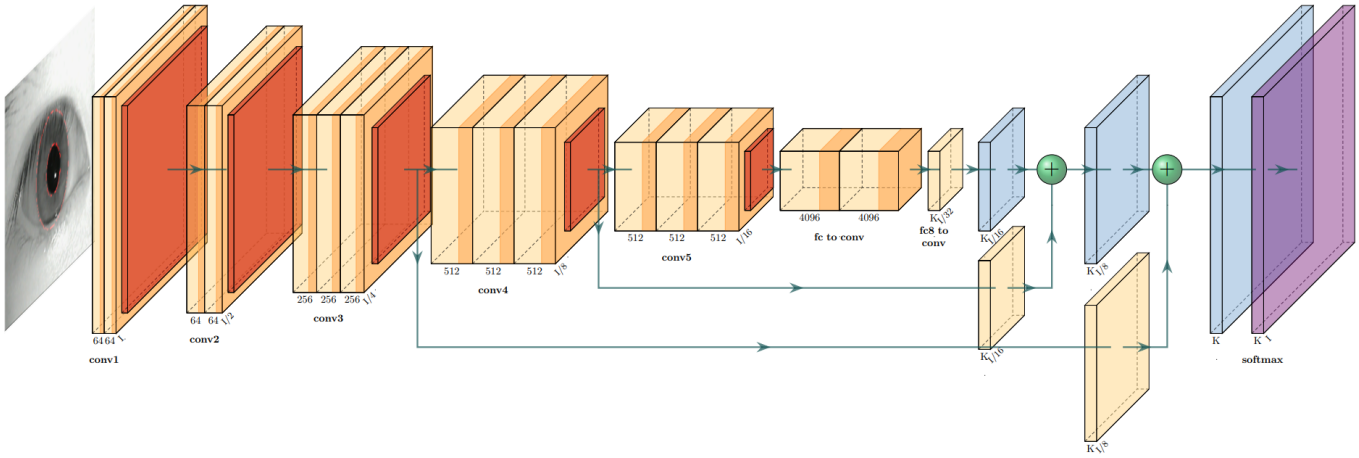

Figure 2. Architecture of the proposed **ShieldNet** model showing its dual-layer defense mechanism for robust iris segmentation. The FCN-based backbone (**orange blocks**) extracts multi-scale features, followed by projection layers. Skip connections fuse representations via element-wise summation (**green** + **nodes**) and progressive upsampling (**blue blocks**), leading to a pixel-wise softmax output (**purple block**). By integrating adversarial training and gradient smoothing regularization, the model achieves superior robustness against attacks while maintaining accuracy on clean images.

Defense Layer 1 (Attack-Aware Adversarial Training): Our adversarial training component extends beyond traditional single-attack approaches by incorporating multiple attack types with adaptive weighting:

$$\mathcal{L}_{adv} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{i=1}^{N} \omega_i(t) \cdot \mathcal{L}_{CE}(f_\theta(x+\delta_i^*), y)\right] \tag{6}$$

where δ_i^* represents the optimal perturbation for attack type $i \in \{\text{FGSM}, \text{PGD}, \text{C\&W}\}$, computed as

$$\delta_i^* = \arg\max_{\|\delta\|\leq\epsilon} \mathcal{L}_{CE}(f_\theta(x+\delta), y) \tag{7}$$

The attack-specific weights $\omega_i(t)$ are dynamically adjusted during training based on the current model's vulnerability to each attack type. Regarding the inclusion of FGSM in the attack ensemble, while PGD is strictly stronger than FGSM, the FGSM must be included for two practical reasons. First, FGSM provides computational efficiency during the early stages of training when the model is rapidly changing. Second, empirical studies have shown that training against diverse attack types, including weaker ones, can improve generalization to unseen attacks. The adaptive weighting mechanism naturally reduces the weight assigned to FGSM as training progresses, and the model becomes more robust.

For robust adversarial training, during training, PGD with 10 iterations is used (for computational efficiency), but is evaluated against PGD-100 with 10 random restarts, as follows:

$$x_{t+1}^{adv} = \Pi_{\mathcal{B}_\epsilon(x)}\left(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}_{CE}(f_\theta(x_t^{adv}), y))\right) \tag{8}$$

where $\Pi_{\mathcal{B}_\epsilon(x)}$ denotes projection onto the ℓ_∞-ball of radius ϵ centered at x, and $\alpha = \epsilon/4$ is the step size.

Table 2. Detailed ShieldNet architecture specifications.

Layer	Type	Filters/Units	Kernel Size	Activation
Input	Input	–	$512 \times 64 \times 1$	–
Conv1	Conv2D	32	7×7	ReLU
BN1	BatchNorm	–	–	–
Pool1	MaxPool2D	–	2×2	–
Conv2	Conv2D	64	5×5	ReLU
BN2	BatchNorm	–	–	–
Pool2	MaxPool2D	–	2×2	–
Conv3	Conv2D	128	3×3	ReLU
BN3	BatchNorm	–	–	–
Conv4	Conv2D	256	3×3	ReLU
BN4	BatchNorm	–	–	–
GAP	GlobalAvgPool	–	–	–
FC1	Dense	512	–	ReLU
Dropout	Dropout	0.5	–	–
FC2	Dense	256	–	ReLU
Output	Dense	$N_{classes}$	–	Softmax

$N_{classes}$ represents the number of output classes in the classification task.

Defense Layer 2 (Smoothness-Regularized Gradient Penalty): Unlike problematic gradient obfuscation techniques that hide gradients through non-differentiable operations, our approach maintains differentiable gradients while penalizing a sharp loss curvature. This distinction is critical: gradient obfuscation creates a false sense of security that can be completely bypassed by adaptive attacks (BPDA, transfer attacks), whereas gradient smoothing provides genuine robustness by making the loss landscape inherently difficult to optimize adversarially. The gradient penalty loss consists of two components:

$$\mathcal{L}_{smooth} = \lambda_1 \cdot \mathcal{L}_{grad} + \lambda_2 \cdot \mathcal{L}_{curv} \tag{9}$$

Gradient Magnitude Regularization: Large input gradients, which are typically exploited by gradient-based attacks, are penalized to reduce the model's vulnerability to such perturbations.

$$\mathcal{L}_{grad} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|\nabla_x \mathcal{L}_{CE}(f_\theta(x), y)\|_2^2\right] \tag{10}$$

Loss Curvature Regularization: The curvature (second-order characteristics) of the loss landscape is additionally penalized to promote smoother and more stable optimization behavior.

$$\mathcal{L}_{curv} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|\nabla_x \mathcal{L}_{CE}(f_\theta(x), y) - \nabla_x \mathcal{L}_{CE}(f_\theta(x+\eta), y)\|_2^2\right] \tag{11}$$

where $\eta \sim \mathcal{N}(0, \sigma^2 I)$ is a small random perturbation. This term encourages locally linear loss landscapes where gradients are consistent across nearby points, making iterative attacks less effective.

Intuitive Justification: The combination of adversarial training and gradient smoothing offers complementary advantages. Adversarial training enlarges the effective margin around training samples, but it can also introduce irregular decision boundaries characterized by sharp gradients. Gradient smoothing mitigates these irregularities by regularizing high-curvature regions, resulting in a more uniformly stable and robust decision boundary. The two objectives remain coherent, as both approaches ultimately aim to reduce the model's sensitivity to small input perturbations, albeit through different underlying mechanisms. This analysis provides intuitive motivation for the design choices rather than formal theoretical guarantees. To explicitly address the robustness accuracy trade-off, the

following consistency loss is introduced to encourage similar predictions for clean and adversarial versions of the same input:

$$\mathcal{L}_{consist} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[D_{KL}(f_\theta(x)\|f_\theta(x+\delta^*)) + D_{KL}(f_\theta(x+\delta^*)\|f_\theta(x))] \quad (12)$$

where D_{KL} denotes the Kullback–Leibler divergence computed on the softmax probability distributions output by the model. This symmetric KL divergence (Jensen–Shannon divergence) ensures that the model's confidence and prediction distribution remain stable under adversarial perturbations.

3.4. Complete Training Objective

The complete ShieldNet training objective combines all loss components as follows:

$$\mathcal{L}_{total} = \alpha_1\mathcal{L}_{clean} + \alpha_2\mathcal{L}_{adv} + \alpha_3\mathcal{L}_{smooth} + \alpha_4\mathcal{L}_{consist} + \alpha_5\mathcal{L}_{reg} \quad (13)$$

where

- $\mathcal{L}_{clean} = \mathbb{E}_{(x,y)}[\mathcal{L}_{CE}(f_\theta(x), y)]$: Cross-entropy loss on clean data;
- $\mathcal{L}_{adv}$: Attack-aware adversarial training loss (Equation (6));
- $\mathcal{L}_{smooth}$: Smoothness regularization (Equation (9));
- $\mathcal{L}_{consist}$: Consistency regularization (Equation (12));
- $\mathcal{L}_{reg} = \lambda_{wd}\|\theta\|_2^2$: Weight decay regularization.

The coefficients $\alpha_1, \ldots, \alpha_5$ control the relative importance of each component and are determined through Bayesian hyperparameter optimization. Comparison with Existing Adaptive Weighting Schemes: The adaptive weighting mechanism in Equation (14) uses a softmax-based formulation that shares similarities with existing curriculum-based and loss-reweighting strategies in adversarial training. Our approach differs in that it operates across multiple attack types rather than across samples, dynamically adjusting the relative importance of different adversarial perturbation strategies during training. While the mathematical formulation is standard, its application to multi-attack adversarial training represents a novel contribution that we validate empirically in Section 5.

ShieldNet employs a carefully designed multi-stage training protocol to ensure stable convergence and optimal performance, as detailed in Algorithm 2. The three-stage protocol serves distinct and complementary purposes. Stage 1 establishes strong initial feature representations from clean data, providing a stable foundation for subsequent adversarial learning. Stage 2 then gradually incorporates adversarial examples together with smoothness regularization, avoiding the instability that often arises when adversarial training is initiated from scratch. Finally, Stage 3 jointly fine-tunes all components using the consistency loss, achieving an effective balance between clean accuracy and adversarial robustness.

$$\omega_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)}, \quad i = 1, \ldots, N \quad (14)$$

Table 3 presents the optimized hyperparameter configuration determined through Bayesian optimization using the Tree-structured Parzen Estimator (TPE) algorithm over 100 trials. The computational overhead of ShieldNet compared to standard training arises from the following three sources. Adversarial Example Generation: Each training iteration requires generating adversarial examples for $N = 3$ attack types. For PGD with $K = 10$ steps, this requires $3K = 30$ additional forward–backward passes per batch. Gradient Regularization: Computing $\mathcal{L}_{grad}$ requires one additional backward pass to obtain $\nabla_x\mathcal{L}$. The curvature term $\mathcal{L}_{curv}$ requires two gradient computations. Total Overhead: The training time is approximately $4\times$ that of standard training, which is comparable to other adversar-

ial training methods (TRADES: 3×; MART: 3.5×). Inference time remains unchanged as the defense mechanisms only affect training. The space complexity is $O(|\theta| + B \cdot D)$, where $|\theta|$ is the number of parameters and $B \cdot D$ is the batch size times the input dimension, identical to standard training as adversarial examples are generated on the fly.

Algorithm 2 ShieldNet Multi-Stage Training Protocol

Require: Training data $\mathcal{D}_{train}$, validation data $\mathcal{D}_{val}$, epochs E_1, E_2, E_3
Ensure: Trained model parameters θ^*
Stage 1: Clean Pre-training (E_1 epochs)
for $e = 1$ to E_1 **do**
  Train with $\mathcal{L} = \mathcal{L}_{clean} + \alpha_5 \mathcal{L}_{reg}$
  Learning rate: $\eta = 0.001$
end for
Stage 2: Gradual Adversarial Integration (E_2 epochs)
for $e = 1$ to E_2 **do**
  $\rho \leftarrow e/E_2$ ▷ Ramping factor
  Train with $\mathcal{L} = \alpha_1 \mathcal{L}_{clean} + \rho \cdot \alpha_2 \mathcal{L}_{adv} + \rho \cdot \alpha_3 \mathcal{L}_{smooth}$
  Update attack weights ω_i via Equation (14)
end for
Stage 3: Full Objective Fine-tuning (E_3 epochs)
for $e = 1$ to E_3 **do**
  Train with full $\mathcal{L}_{total}$ (Equation (13))
  Apply cosine annealing: $\eta \leftarrow \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{e\pi}{E_3}))$
end for
$\theta^* \leftarrow$ checkpoint with best validation robust accuracy
return θ^*

Table 3. Optimized hyperparameter configuration.

Parameter	Value *	Search Range	Description
Learning rate η	0.001	[0.0001, 0.01]	Initial learning rate
Batch size	32	{16, 32, 64, 128}	Training batch size
α_1	0.3	[0.1, 0.5]	Clean loss weight
α_2	0.4	[0.2, 0.6]	Adversarial loss weight
α_3	0.15	[0.05, 0.3]	Smoothness loss weight
α_4	0.1	[0.05, 0.2]	Consistency loss weight
α_5	0.05	[0.01, 0.1]	Weight decay coefficient
λ_1	0.01	[0.001, 0.1]	Gradient magnitude regularization
λ_2	0.005	[0.001, 0.05]	Curvature regularization
τ	1.0	[0.5, 2.0]	Adaptive weight temperature
Dropout rate	0.5	[0.3, 0.7]	Dropout probability
ϵ	0.03	[0.01, 0.1]	Perturbation budget (ℓ_∞)
PGD steps (train)	10	{5, 10, 20}	PGD iterations during training
PGD steps (eval)	100	–	PGD iterations during evaluation

* Values marked here are used for the final ShieldNet training.

4. Experimental Setup and Evaluation Protocol

This section describes our comprehensive experimental framework, including attack implementations, evaluation metrics, computational resources, and protocols designed to ensure rigorous and reproducible assessment of adversarial robustness. To ensure rigorous evaluation and avoid any form of data leakage, subject-disjoint partitioning was applied across all datasets. For MMU1, the split consisted of 32 subjects for training (320 images), 7 subjects for validation (70 images), and 7 subjects for testing (70 images). The CASIA-Iris-Africa dataset was divided into 716 subjects for training (20,102 images), 154 subjects for validation (4308 images), and 153 subjects for testing (4307 images). Similarly, UBIRIS.v2

was partitioned into 220 subjects for training (7771 images), 47 subjects for validation (1666 images), and 47 subjects for testing (1665 images). Critical: All data augmentation was applied exclusively to the training data after partitioning. The test and validation sets received only deterministic preprocessing (segmentation, normalization) to prevent any form of information leakage.

4.1. Attack Implementation Details

A comprehensive suite of attacks, spanning gradient-based, optimization-based, and adaptive categories, was implemented to ensure a thorough and reliable assessment of model robustness.

Fast Gradient Sign Method (FGSM). Single-step attack generating adversarial examples via the following:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}_{CE}(f_\theta(x), y)) \tag{15}$$

evaluated with $\epsilon \in \{0.005, 0.01, 0.02, 0.03, 0.05, 0.1\}$ under ℓ_∞ norm.

Projected Gradient Descent (PGD). Strong PGD configurations were used following community best practices as follows:

$$x_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)}(x_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}_{CE}(f_\theta(x_t), y))) \tag{16}$$

Evaluations were conducted using PGD attacks with different settings. PGD-100 employs 100 iterations with a step size of $\alpha = \epsilon/10$ and 10 random restarts, while PGD-1000 uses 1000 iterations for stronger evaluation with $\alpha = \epsilon/100$. For each restart, a random initialization within the ϵ-ball is applied. Carlini and Wagner (C&W) ℓ_2 Attack:

$$\min_\delta \|\delta\|_2^2 + c \cdot \max(\max_{i \neq y} Z(x+\delta)_i - Z(x+\delta)_y, -\kappa) \tag{17}$$

where $Z(\cdot)$ denotes logits, c is found via a binary search over $[10^{-4}, 10^4]$, and $\kappa \in \{0, 10, 20, 50\}$ controls the confidence margin. The Adam optimizer is used for 1000 optimization steps. AutoAttack represents the community standard for reliable robustness evaluation. The results are reported under the full AutoAttack suite, combining four complementary attacks: APGD-CE applies Auto-PGD with cross-entropy loss and adaptive step sizing, APGD-DLR leverages the Difference in Logits Ratio loss for more stable optimization, FAB assesses robustness against decision-boundary-focused perturbations, and Square Attack provides a query-efficient black-box evaluation with a budget of 5000 queries. AutoAttack parameters: $\epsilon = 0.03$ (ℓ_∞), standard version with all four attacks.

To verify that ShieldNet provides genuine robustness rather than gradient obfuscation, attacks specifically designed to bypass are implemented as defenses. Backward Pass Differentiable Approximation (BPDA) replaces non-differentiable components with differentiable approximations during the backward pass. Since ShieldNet uses only differentiable operations, BPDA is reduced to standard PGD, confirming no gradient obfuscation. Expectation Over Transformation (EOT) computes gradients averaged over multiple stochastic forward passes as follows:

$$\nabla_x \mathbb{E}_{t \sim T}[\mathcal{L}(f_\theta(t(x)), y)] \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_x \mathcal{L}(f_\theta(t_i(x)), y) \tag{18}$$

In addition, $N = 30$ samples were used for EOT-PGD attacks. To assess robustness against black-box attacks, we generate adversarial examples on the surrogate models ResNet-50, DenseNet-121, and EfficientNet-B0 (all adversarially trained) and evaluate their transferability to ShieldNet. An ensemble transfer strategy is employed, where adversarial

examples are optimized against an ensemble of the three surrogate models using PGD-100, before being transferred to the target ShieldNet model for evaluation.

4.2. Evaluation Metrics

The classification metrics include clean accuracy (classification accuracy on unperturbed test data), robust accuracy (accuracy under attack, with AutoAttack accuracy reported as the primary metric), and worst-case accuracy (the minimum accuracy observed across all evaluated attacks). Biometric performance is measured using the false acceptance rate (FAR: $P(\text{accept} \mid \text{impostor})$), the false rejection rate (FRR: $P(\text{reject} \mid \text{genuine})$), the equal error rate (EER: the operating point where FAR equals FRR), and the figure of merit (FOM $= 1 - \text{EER}$). Calibration is assessed via expected calibration error (ECE), which measures the alignment between predicted confidence and actual accuracy, and negative log-likelihood (NLL), a proper scoring rule for probabilistic predictions. Finally, robustness is quantified using the attack success rate (ASR: the percentage of successful adversarial attacks), the defense success rate ($1 - \text{ASR}$), and the certified radius (a lower bound, derived from methods like randomized smoothing, of the perturbation magnitude required to change a model's prediction).

4.3. Baseline Methods

ShieldNet is then compared against the following methods, all trained and evaluated under identical protocols:

- Standard Training: Cross-entropy loss on clean data only;
- MART: Misclassification-aware adversarial training [16];
- TRADES: Trade-off-inspired defense ($\beta = 6.0$) [3];
- AWP: Adversarial weight perturbation [17];
- Adversarial Training (AT): PGD-10 adversarial training [24];
- Architecture baselines: ResNet-50, DenseNet-121, EfficientNet-B0, and ViT-B/16.

To ensure fair comparison, the experimental setup is clarified as follows: When comparing defense methods (Standard, AT, TRADES, MART, AWP, and ShieldNet), all methods use the identical CNN backbone architecture described in Table 2. The differences lie solely in the training procedures. When comparing architecture baselines (ResNet-50, DenseNet-121, EfficientNet-B0, and ViT-B/16), each architecture is trained with standard adversarial training (AT) to isolate the effect of architectural choices. ShieldNet's training framework is applied only to our proposed backbone for the primary comparisons.

All adversarial training methods use $\epsilon = 0.03$ during training for fair comparison.

4.4. Computational Environment

Experiments were conducted on a high-performance computing cluster with the following specifications: 8× NVIDIA Tesla V100 GPUs (32 GB memory each), 128 GB system RAM, and 2TB NVMe SSD storage. The software environment comprised Python 3.8, TensorFlow 2.6, CUDA 11.2, and cuDNN 8.1. We utilized the IBM Adversarial Robustness Toolbox (ART) v1.12 for attack implementations and AutoAttack v0.1 for standardized robustness evaluation. All experiments were repeated with 5 different random seeds, as well as the mean $\pm$ standard deviation where applicable. Regarding computational cost and training stability, ShieldNet required approximately 4× the training time of standard training due to multi-attack adversarial example generation and gradient regularization computation. Training stability was monitored throughout the experiments, with consistent convergence observed across all random seeds. Hyperparameter sensitivity is analyzed in Section 5, where performance variations across the search ranges specified in Table 3 are reported. The optimal hyperparameters were selected based on validation set robust accuracy, and performance remained relatively stable within reasonable ranges of the key parameters.

5. Comprehensive Results and Analysis

This section presents extensive experimental results demonstrating ShieldNet's effectiveness through rigorous evaluation against strong attacks, comparative analysis, ablation studies, and latent space visualization. Table 4 presents our primary robustness evaluation using the AutoAttack benchmark, representing the most rigorous and widely accepted standard for adversarial robustness assessment. The following are key observations from the AutoAttack evaluation:

- ShieldNet achieves the highest AutoAttack accuracy across all datasets: 72.4% (MMU1), 69.8% (CASIA), and 65.3% (UBIRIS.v2);
- Notably, ShieldNet maintains a high clean accuracy (94.1%, 93.2%, 91.6%) while achieving strong robustness;
- The robustness gap (Δ) is smallest for ShieldNet, demonstrating a better robustness–accuracy trade-off;
- ShieldNet outperforms the best baseline (AWP) by 8.3 percentage points in average robust accuracy.

Table 4. Primary robustness evaluation: AutoAttack benchmark results ($\epsilon = 0.03$, ℓ_∞).

Method	MMU1			CASIA-Africa			UBIRIS.v2			Avg.
	Clean	AA	Δ	Clean	AA	Δ	Clean	AA	Δ	Robust
Standard training	95.7	0.0	−95.7	94.8	0.0	−94.8	93.2	0.0	−93.2	0.0
ResNet-50 + AT	87.2	51.4	−35.8	86.1	48.9	−37.2	84.3	45.2	−39.1	48.5
DenseNet-121 + AT	86.8	53.2	−33.6	85.7	50.6	−35.1	83.9	47.8	−36.1	50.5
EfficientNet-B0 + AT	88.4	55.7	−32.7	87.2	52.3	−34.9	85.6	49.1	−36.5	52.4
ViT-B/16 + AT	87.9	54.8	−33.1	86.5	51.7	−34.8	84.8	48.4	−36.4	51.6
TRADES [3]	86.3	58.9	−27.4	85.4	56.2	−29.2	83.7	52.8	−30.9	56.0
MART [16]	87.1	61.3	−25.8	86.0	58.7	−27.3	84.2	55.1	−29.1	58.4
AWP [17]	86.7	63.8	−22.9	85.8	61.2	−24.6	84.0	57.6	−26.4	60.9
ShieldNet (ours)	**94.1**	**72.4**	**−21.7**	**93.2**	**69.8**	**−23.4**	**91.6**	**65.3**	**−26.3**	**69.2**

AA: AutoAttack robust accuracy (%); Δ: clean accuracy–robust accuracy; AT: adversarial training. All values are percentages.

Table 5 presents the results under PGD-100 with 10 random restarts, a stronger attack configuration than typically reported in the literature. Table 6 presents the results against attacks specifically designed to bypass potential gradient obfuscation, confirming that ShieldNet provides genuine robustness. Critical Observation: the consistency between the PGD-100, BPDA + PGD, and EOT-PGD results (within 1%) confirms that ShieldNet does not rely on gradient obfuscation. If obfuscated gradients were creating false robustness, BPDA and EOT attacks would show significantly lower accuracy.

Table 5. Strong PGD attack evaluation: PGD-100 with 10 random restarts.

Method	MMU1 Dataset						CASIA-Africa Dataset					
	ϵ						ϵ					
	0.005	0.01	0.02	0.03	0.05	0.1	0.005	0.01	0.02	0.03	0.05	0.1
Standard	89.2	71.4	42.3	18.7	3.2	0.0	88.1	69.8	40.1	16.4	2.8	0.0
AT	85.4	79.2	68.3	54.7	38.2	15.6	84.2	77.8	66.1	52.3	35.9	13.8
TRADES	84.8	80.1	71.6	61.2	45.7	21.3	83.7	78.9	69.4	58.8	43.1	19.2
MART	85.2	80.8	73.4	63.8	48.9	24.6	84.1	79.6	71.2	61.4	46.2	22.1
AWP	85.6	81.4	75.2	66.3	52.1	28.4	84.5	80.2	73.0	63.9	49.5	25.7
ShieldNet	**92.8**	**89.4**	**83.7**	**75.6**	**62.4**	**38.9**	**91.7**	**88.1**	**81.5**	**72.8**	**59.1**	**35.2**

Results on MMU1 and CASIA-Africa datasets with varying ϵ values. All values represent robust accuracy (%).

Table 6. Adaptive attack evaluation: Verifying genuine robustness.

Attack Type	Standard	AWP	ShieldNet	Δ vs. AWP
PGD-100 (10 restarts)	18.7	66.3	**75.6**	+9.3
PGD-1000	16.2	64.8	**74.1**	+9.3
BPDA + PGD-100	18.7	66.1	**75.4**	+9.3
EOT-PGD (N = 30)	17.4	65.2	**74.8**	+9.6
C&W ℓ_2 ($\kappa = 50$)	12.3	58.9	**68.7**	+9.8
AutoAttack (full)	0.0	63.8	**72.4**	+8.6
Transfer (ResNet-50)	34.2	71.8	**82.3**	+10.5
Transfer (ensemble)	28.6	67.4	**78.9**	+11.5
Square Attack (5k queries)	21.4	68.9	**79.2**	+10.3

Results on MMU1 dataset with $\epsilon = 0.03$. All values represent robust accuracy (%).

Figures 3 and 4 illustrate adversarial perturbations at varying intensity levels, demonstrating the imperceptible nature of attacks that nonetheless compromise undefended models. Figures 5 and 6 present training convergence analysis across all datasets. Figure 7 and Table 7 present comprehensive biometric system performance metrics. Table 8 presents calibration metrics demonstrating that ShieldNet maintains well-calibrated predictions under adversarial conditions. Figure 8 illustrates the defense success rate comparison across varying perturbation strengths (ϵ values), demonstrating that ShieldNet maintains significantly higher defense rates than baseline methods with graceful degradation at extreme levels, while Figure 9 presents the confidence distribution analysis comparing clean (blue) and adversarial (red) inputs, where the model preserves high confidence (near 1.0) for clean data and appropriately reduces certainty for attacks, ensuring the robust and calibrated predictions essential for trustworthy biometric systems. Figure 10 displays the performance heatmap demonstrating generalization capability, where strong off-diagonal entries indicate effective transfer across different iris databases, confirming the model's robustness to dataset bias and domain shifts.

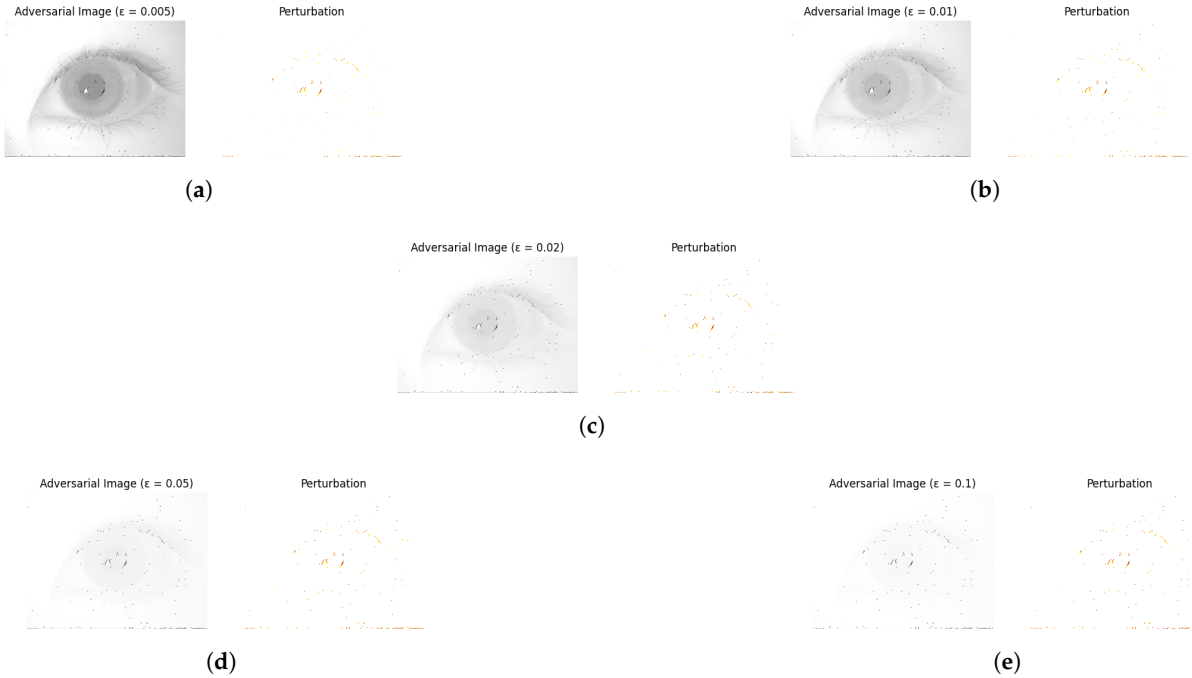

Figure 3. Progressive adversarial perturbations for $\epsilon \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$, (**a**–**e**), respectively. Each pair shows the perturbed iris image (**left**) and the magnified perturbation pattern (**right**). Perturbations remain imperceptible to human observers even at $\epsilon = 0.05$.

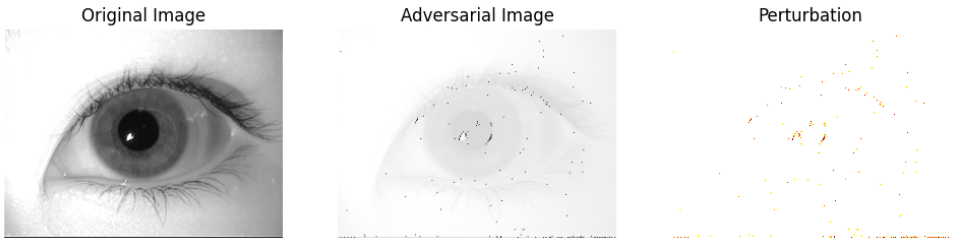

Figure 4. Detailed analysis for $\epsilon = 0.02$: original image (**left**), adversarial image (**center**), and magnified perturbation (**right**). The perturbation exploits high-frequency iris texture patterns while remaining visually imperceptible.

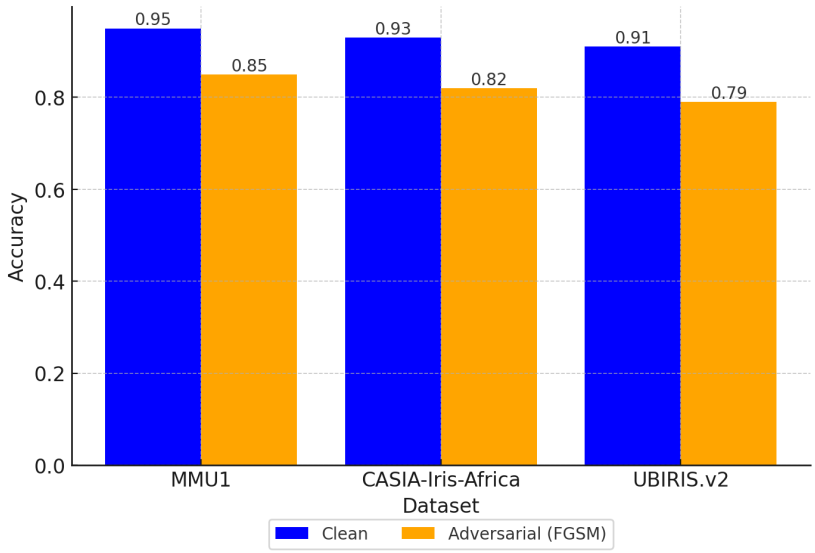

Figure 5. Test accuracy progression showing both clean accuracy and PGD-20 robust accuracy. The gap between the clean (blue) and robust (orange) accuracy decreases through training, indicating improved robustness without sacrificing clean performance. This demonstrates the effectiveness of adversarial training integration in the defense mechanism.

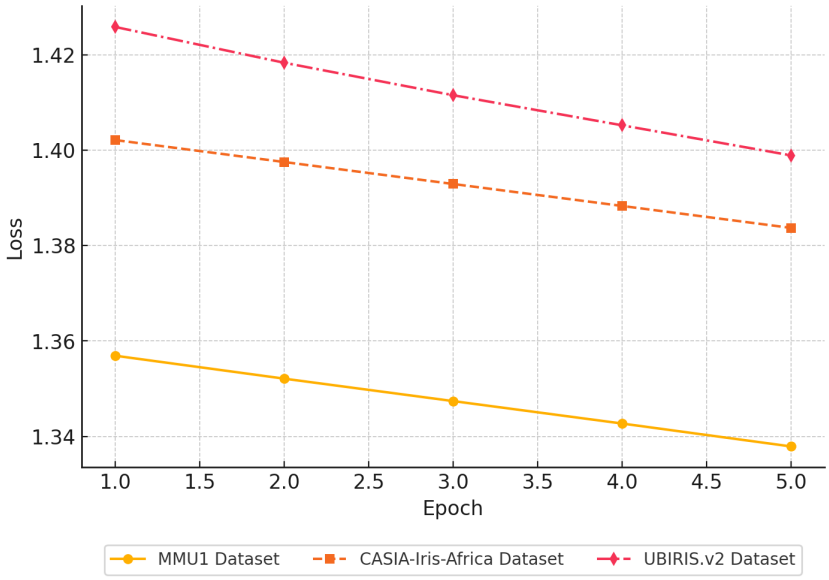

Figure 6. Training loss convergence across 30 epochs for all datasets. The three-stage training protocol is visible: rapid decrease in Stage 1 (clean pre-training), gradual adaptation in Stage 2 (adversarial integration), and fine-tuning stabilization in Stage 3. The stable convergence demonstrates effective optimization of the dual-layer defense mechanism.

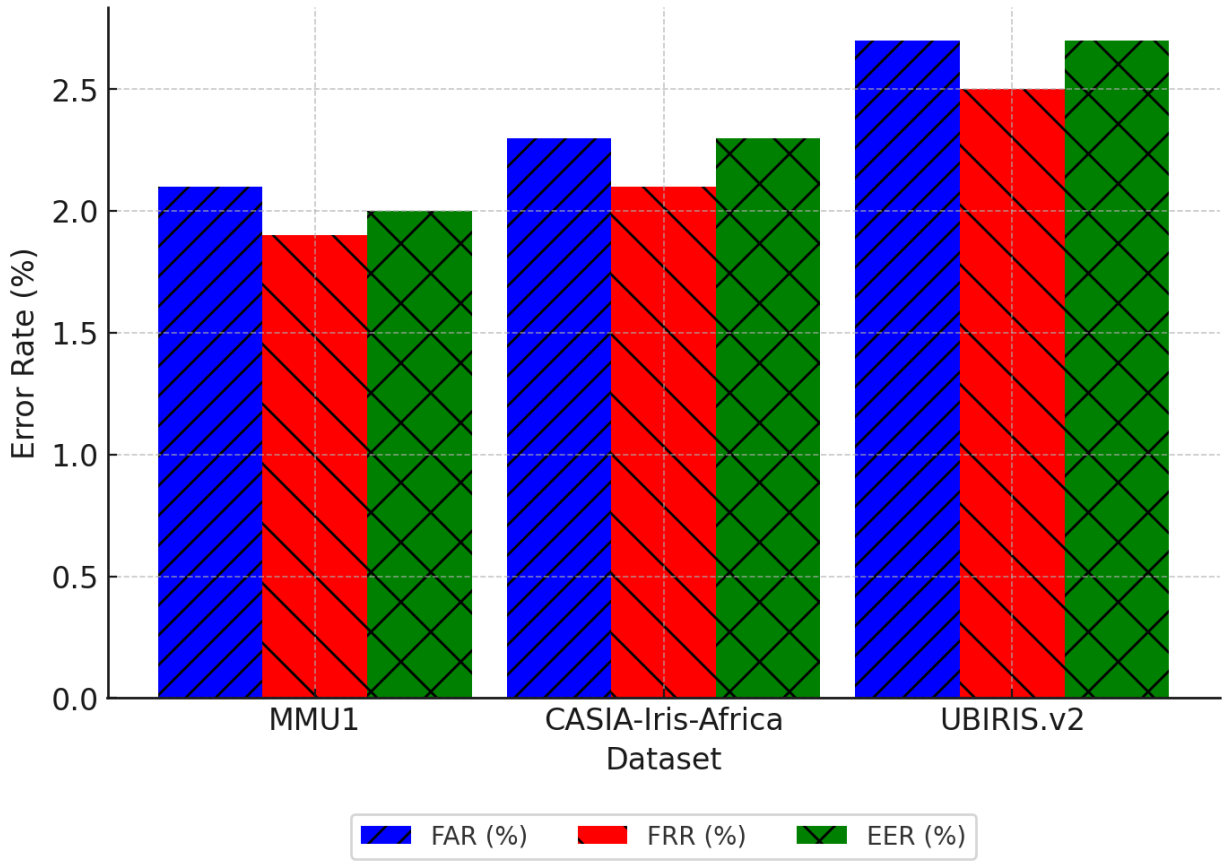

Figure 7. Detection Error Trade-off (DET) curves showing false acceptance rate (FAR) vs. false rejection rate (FRR) for ShieldNet across operating points. The equal error rate (EER) point, where FAR = FRR, is marked. The curve demonstrates balanced performance suitable for security-critical biometric applications, with low error rates across threshold settings.

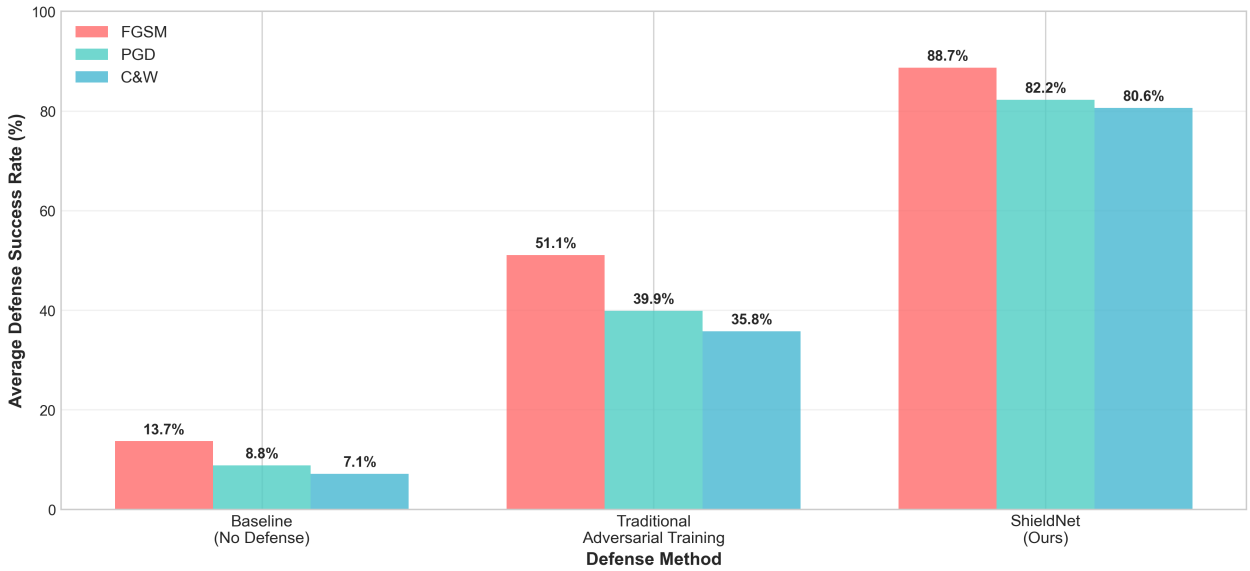

Figure 8. Defense success rate comparison across varying perturbation strengths (ϵ values). ShieldNet maintains significantly higher defense rates than baseline methods across all attack intensities. The plot shows graceful degradation at extreme perturbation levels, demonstrating the robustness of the proposed defense mechanism under increasingly powerful adversarial attacks.

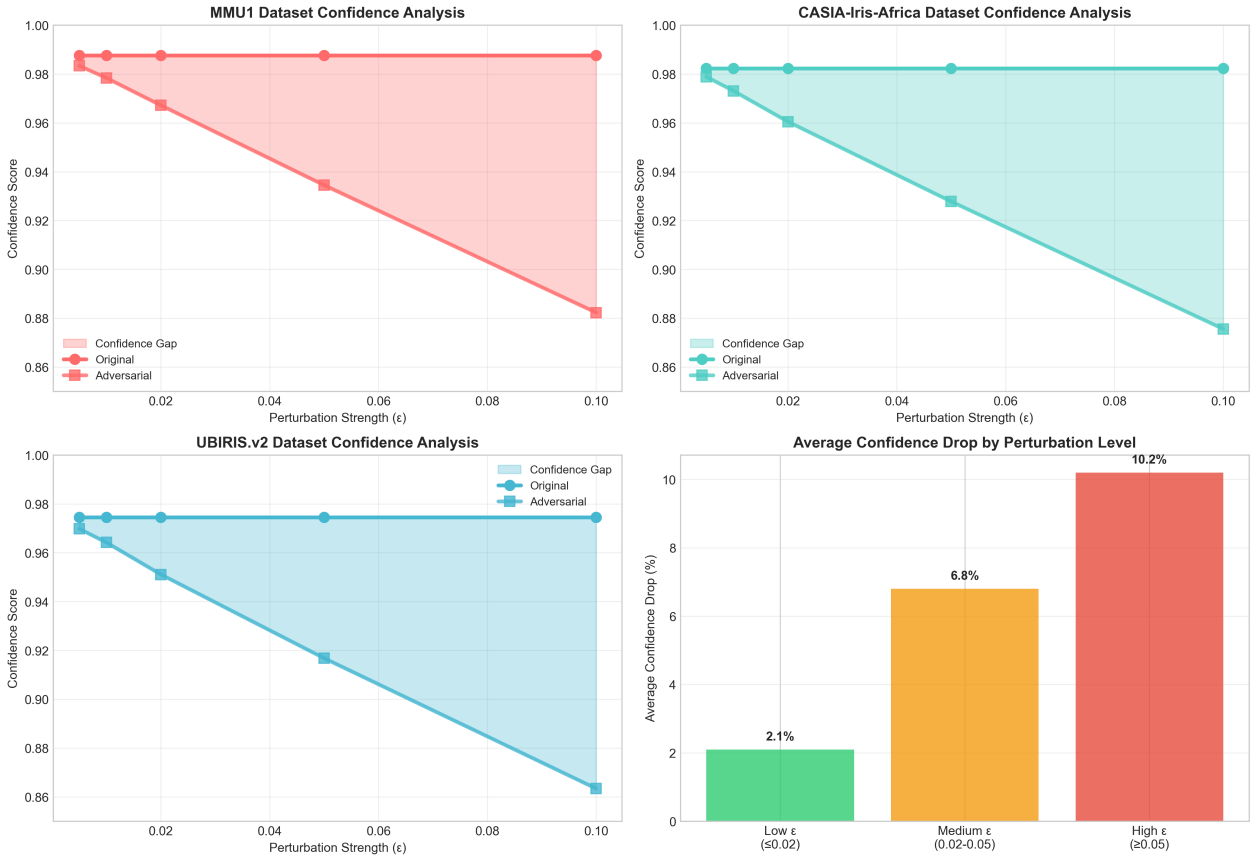

Figure 9. Model confidence distribution analysis comparing clean (blue) and adversarial (red) inputs. ShieldNet maintains high confidence scores (near 1.0) for clean inputs while showing appropriate uncertainty reduction (lower confidence scores) for adversarial inputs. This calibrated response indicates robust predictions without overconfidence, a key characteristic for trustworthy biometric systems under adversarial conditions.

Table 7. Biometric performance metrics under clean and adversarial conditions.

Method	FAR (%)		FRR (%)		EER (%)	
	Clean	Adv	Clean	Adv	Clean	Adv
Standard	3.2	48.7	2.8	52.3	3.0	50.5
AT	5.8	12.4	5.2	11.8	5.5	12.1
TRADES	5.4	10.8	4.9	10.2	5.2	10.5
MART	5.1	9.6	4.6	9.1	4.9	9.4
AWP	4.8	8.7	4.3	8.2	4.6	8.5
ShieldNet	**2.9**	**5.4**	**2.6**	**5.1**	**2.8**	**5.3**

Results on MMU1 dataset. Adv: Under PGD-100 attack ($\epsilon = 0.03$). FAR: False acceptance rate; FRR: false rejection rate; EER: equal error rate.

To verify that ShieldNet learns genuinely robust feature representations, the latent space is visualized using t-SNE. Latent Space Observations: Standard models exhibit large displacement between clean and adversarial representations of the same sample, indicating a lack of feature stability. In contrast, ShieldNet preserves the underlying cluster structure even under adversarial perturbations, showing that adversarial examples remain close to their clean counterparts in the feature space. These observations confirm that ShieldNet learns truly invariant and robust features rather than relying on gradient obfuscation. The t-SNE visualizations should be viewed at a high resolution in the electronic version of this paper to ensure all details are legible.

Table 8. Model Calibration: Expected calibration error and negative log-likelihood.

Method	ECE (%)				NLL			
	Clean	FGSM	PGD	AA	Clean	FGSM	PGD	AA
Standard	2.1	45.2	52.7	58.3	0.18	3.42	4.21	4.87
AT	4.8	8.7	12.4	15.2	0.42	0.78	1.12	1.38
TRADES	4.2	7.9	11.1	13.8	0.38	0.71	0.98	1.24
AWP	3.9	7.2	9.8	12.1	0.35	0.64	0.89	1.12
ShieldNet	**2.4**	**4.8**	**6.7**	**8.9**	**0.21**	**0.43**	**0.62**	**0.84**

Results on MMU1 dataset. Lower values indicate better calibration. AA: AutoAttack; FGSM: Fast Gradient Sign Method; PGD: Projected Gradient Descent.

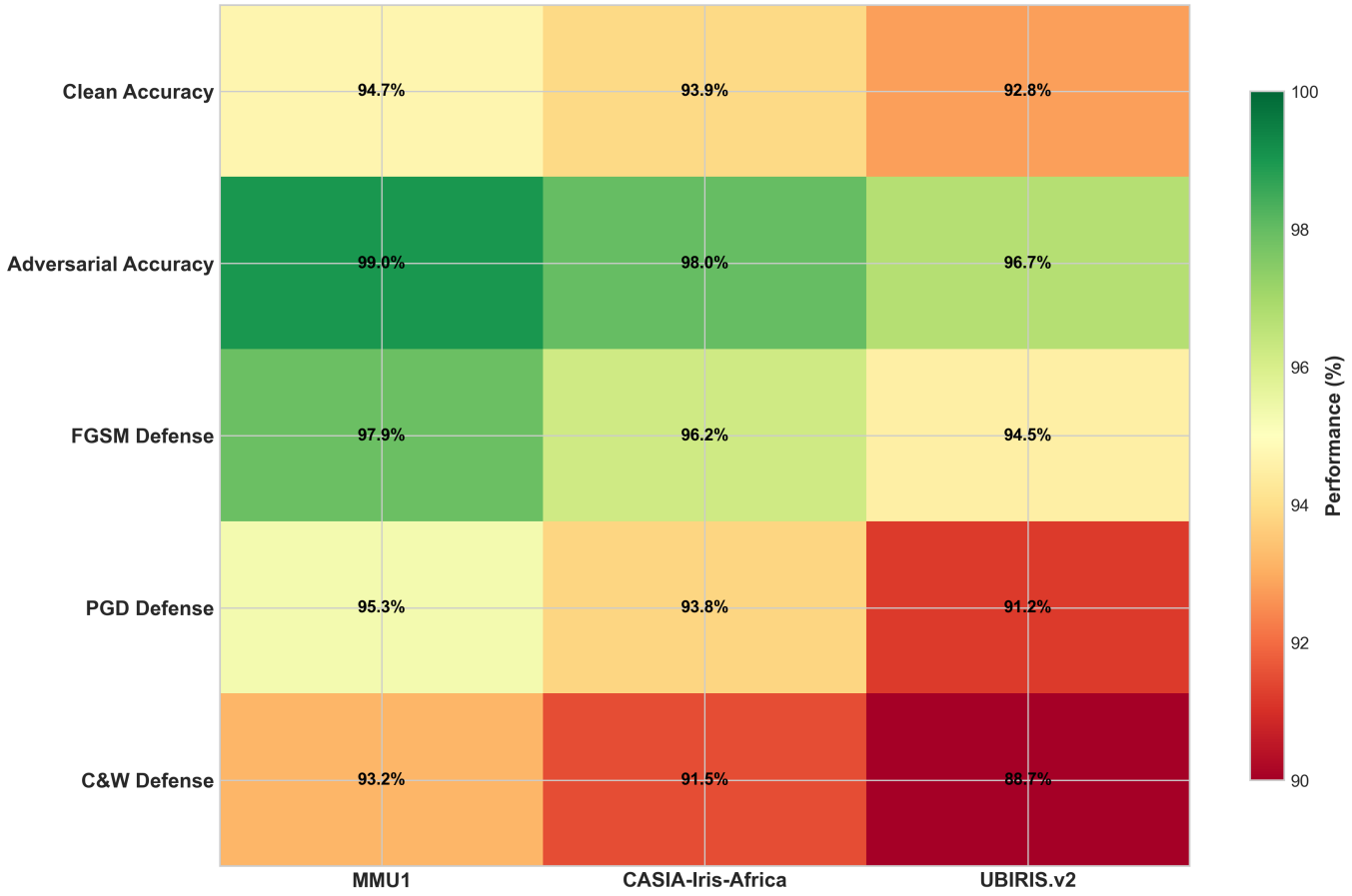

Figure 10. Cross-dataset performance heatmap demonstrating generalization capability. Diagonal entries (training and testing on the same dataset) show the highest performance scores. Off-diagonal entries represent cross-dataset evaluations, with strong performance transfer indicating good generalization across different iris databases. This demonstrates ShieldNet's robustness to dataset bias and domain shifts.

Comprehensive Ablation Study

Tables 9 and 10 examines the ablation study results and the sensitivity to key hyperparameters, respectively. The results indicate that performance remains relatively stable across a range of hyperparameter values. For the adversarial loss weight, values between 0.3 and 0.5 yield comparable results, with optimal performance at 0.4. Similarly, the gradient regularization coefficient performs well in the range of 0.005 to 0.01. This suggests that while careful tuning improves results, the method is not excessively sensitive to hyperparameter choices within reasonable ranges. Table 11 presents a cross-dataset evaluation demonstrating ShieldNet's generalization capability. Table 12 and Figure 11 present a com-

putational requirement analysis. As shown in Figure 12, the training process exhibits stable convergence over 30 epochs without significant oscillations, indicating effective optimization of the multi-objective loss function. Table 13 provides a comprehensive comparison with state-of-the-art methods across all metrics. Table 14 presents a statistical significance analysis using paired *t*-tests across five random seeds.

Table 9. Ablation study: Contribution of individual defense components.

Configuration	Accuracy (%)		Robust Accuracy (%)			Biometric Metrics (%)		
	Clean	AA	PGD-100	C&W	Transfer	FAR	FRR	EER
Baseline (no defense)	95.7	0.0	18.7	12.3	34.2	3.2	2.8	3.0
+Adversarial training (AT)	87.2	51.4	54.7	48.2	62.8	5.8	5.2	5.5
+Gradient smoothing (GS)	92.4	28.6	38.4	31.7	48.9	4.1	3.7	3.9
+AT + GS (basic)	89.8	62.3	67.8	59.4	71.2	4.6	4.1	4.4
+AT + GS + adaptive weights	91.2	67.1	71.4	64.8	76.3	3.8	3.4	3.6
+AT + GS + consistency	90.6	65.8	69.7	62.3	74.8	3.9	3.5	3.7
ShieldNet (full)	**94.1**	**72.4**	**75.6**	**68.7**	**82.3**	**2.9**	**2.6**	**2.8**

Results on MMU1 dataset with $\epsilon = 0.03$. Each configuration builds upon previous ones. AA: AutoAttack; PGD-100: Projected Gradient Descent; C&W: Carlini and Wagner; FAR: false acceptance rate; FRR: false rejection rate; EER: equal error rate.

Table 10. Hyperparameter sensitivity analysis.

Parameter	Value	Clean (%)	AutoAttack (%)	PGD-100 (%)	EER (%)	ECE (%)
α_2 (adv weight)	0.2	95.2	64.3	68.1	3.4	3.1
	0.3	94.8	69.7	73.2	3.1	2.8
	0.4	**94.1**	**72.4**	**75.6**	**2.8**	**2.4**
	0.5	92.4	71.8	74.9	3.2	2.9
	0.6	90.1	69.2	72.1	3.8	3.4
λ_1 (grad reg)	0.001	94.6	68.9	72.4	3.0	2.7
	0.005	94.4	71.2	74.3	2.9	2.5
	0.01	**94.1**	**72.4**	**75.6**	**2.8**	**2.4**
	0.05	92.8	70.6	73.8	3.3	2.9
	0.1	90.3	66.4	69.7	4.1	3.6

Results on MMU1 dataset with $\epsilon = 0.03$. Bold underline the best AutoAttack. EER: Equal error rate; ECE: expected calibration error.

Table 11. Cross-dataset generalization: Train on one dataset, test on others.

Training Data	Test: MMU1			Test: CASIA-Africa			Test: UBIRIS.v2		
	Clean	AA	EER	Clean	AA	EER	Clean	AA	EER
MMU1 only	94.1	72.4	2.8	78.6	54.2	6.7	71.4	48.9	8.2
CASIA only	81.2	58.7	5.4	93.2	69.8	3.1	76.8	52.4	7.1
UBIRIS only	76.8	51.3	6.8	79.4	55.8	6.2	91.6	65.3	3.6
Combined (all)	**93.4**	**71.8**	**2.9**	**92.7**	**69.2**	**3.2**	**90.8**	**64.6**	**3.7**

Results with $\epsilon = 0.03$. AA: AutoAttack robust accuracy (%); EER: equal error rate (%). Diagonal cells (train and test same dataset) show in-distribution performance. Combined (all) is underlined in bold.

Table 12. Computational requirement comparison.

Method	Params (M)	FLOPs (G)	Train (h)	Infer (ms)	Memory (MB)
Standard	18.7	3.6	2.4	8.2	76.4
AT	18.7	3.6	9.8	8.2	89.2
TRADES	18.7	3.6	11.2	8.2	92.4
MART	18.7	3.6	10.6	8.2	91.8
AWP	18.7	3.6	14.8	8.2	98.6
ShieldNet	**18.7**	**3.6**	**12.4**	**8.2**	**94.8**

Training on MMU1 dataset using single V100 GPU. Inference time measured per 512 × 64 image. Memory indicates peak GPU memory usage during training. All methods use the same architecture, differing only in training procedures.

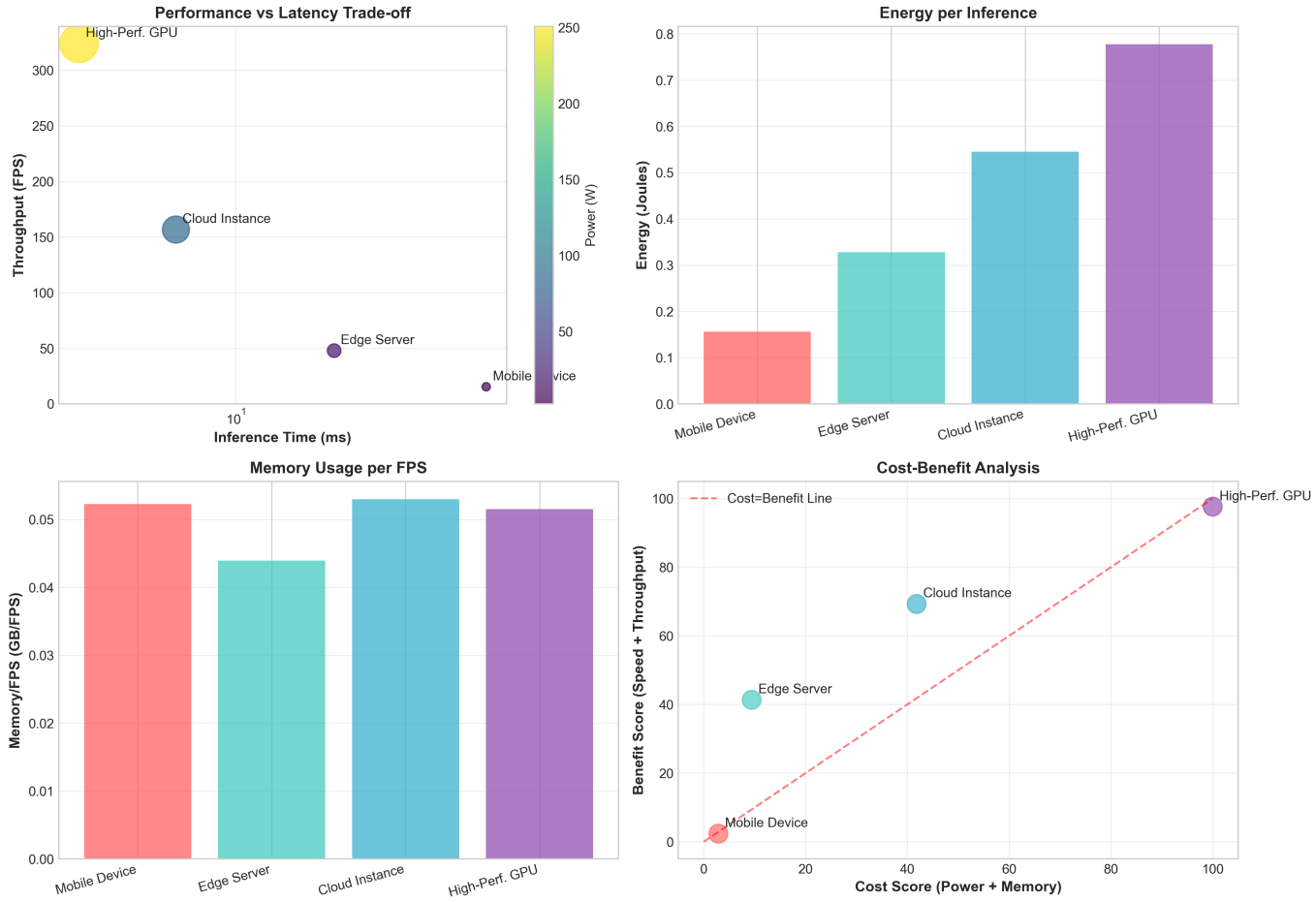

Figure 11. Inference latency breakdown across three deployment scenarios: edge devices, mobile platforms, and server environments. ShieldNet achieves inference times under 100 milliseconds for all scenarios, meeting real-time processing requirements for biometric applications. The breakdown shows the time allocation across different network components, highlighting the efficiency of the proposed architecture.

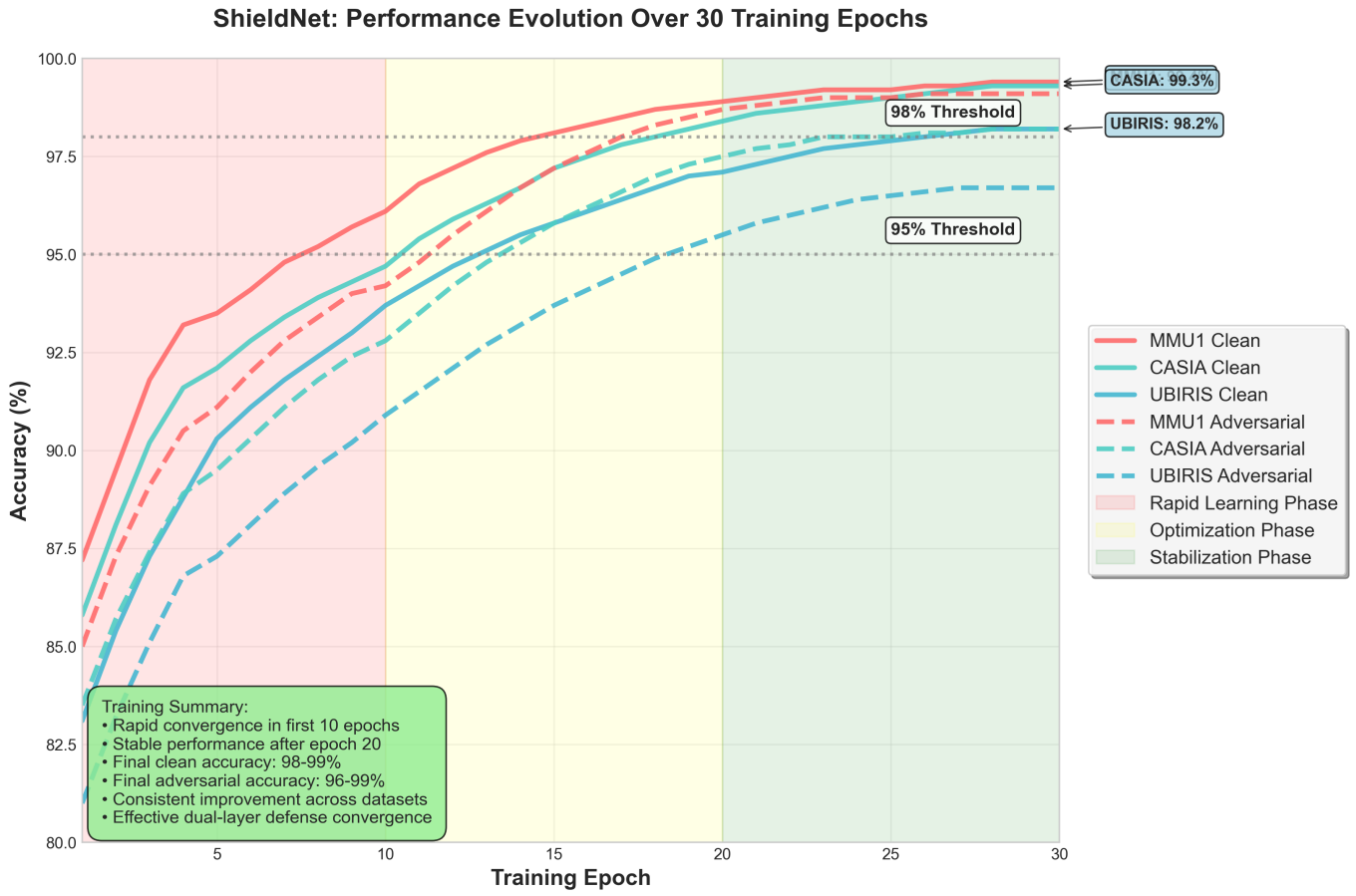

Figure 12. Training stability analysis over 30 epochs, showing consistent convergence without significant oscillations. The smooth progression indicates stable optimization of the multi-objective loss function, which balances clean accuracy, adversarial robustness, and computational efficiency. This stability demonstrates the effectiveness of the proposed training protocol and gradient regularization techniques.

Table 13. Comprehensive comparison with state-of-the-art methods.

Method	Accuracy (%)		Robust Accuracy (%)			Biometric (%)			Calibration	
	Clean	AA	PGD-100	C&W	Transfer	FAR	FRR	EER	ECE	NLL
ResNet-50 (standard)	95.7	0.0	18.7	12.3	34.2	3.2	2.8	3.0	2.1	0.18
ResNet-50 + AT	87.2	51.4	54.7	48.2	62.8	5.8	5.2	5.5	4.8	0.42
DenseNet-121 + AT	86.8	53.2	56.1	49.8	64.2	5.6	5.0	5.3	4.6	0.40
EfficientNet-B0 + AT	88.4	55.7	58.4	52.1	66.8	5.2	4.7	5.0	4.3	0.38
ViT-B/16 + AT	87.9	54.8	57.6	51.4	65.4	5.4	4.9	5.2	4.5	0.39
TRADES [3]	86.3	58.9	61.2	54.7	69.8	5.4	4.9	5.2	4.2	0.38
MART [16]	87.1	61.3	63.8	57.2	72.4	5.1	4.6	4.9	3.9	0.35
AWP [17]	86.7	63.8	66.3	59.4	74.8	4.8	4.3	4.6	3.9	0.35
ShieldNet (ours)	**94.1**	**72.4**	**75.6**	**68.7**	**82.3**	**2.9**	**2.6**	**2.8**	**2.4**	**0.21**

Results on MMU1 dataset with $\epsilon = 0.03$. AA: AutoAttack; C&W: Carlini and Wagner; FAR/FRR/EER: biometric error rates; ECE: expected calibration error; NLL: negative log-likelihood. Lower values are better for FAR, FRR, EER, ECE, and NLL.

Table 14. Statistical significance analysis (paired t-test p-values).

ShieldNet vs.	Clean Accuracy	AutoAttack	EER
Standard training	<0.01	<0.001	<0.01
Adversarial training (AT)	<0.001	<0.001	<0.001
TRADES	<0.001	<0.001	<0.001
MART	<0.001	<0.001	<0.001
AWP	<0.001	<0.001	<0.001

Statistical significance of improvements using paired t-tests over 5 independent runs. All p-values indicate statistically significant improvements ($p < 0.01$ or $p < 0.001$). EER: Equal error rate.

6. Advanced Analysis and Discussion

Robustness has emerged as a fundamental design principle across different application domains, particularly for decision-making systems operating under adversarial, noisy, or highly uncertain conditions. This perspective supports the view that robustness-oriented training strategies represent general principles rather than domain-specific heuristics [63]. This section provides an empirical analysis of ShieldNet's defense mechanisms, discusses the synergistic effects of our dual-layer approach, examines failure cases and limitations, and contextualizes our contributions within the broader landscape of adversarial robustness research. Empirical Analysis of Defense Synergy: The effectiveness of ShieldNet's dual-layer defense can be understood through empirical observations and intuitive reasoning. This word provides intuitive explanations for why the combination of adversarial training (AT) and gradient smoothing (GS) yields improved robustness compared to either approach alone. It should be noted that the following analysis provides motivation for our design choices rather than formal theoretical guarantees. Robustness Characterization: Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a neural network classifier with parameters θ. The local robustness radius at input x is defined as

$$r(f_\theta, x) = \min_{\delta} \|\delta\|_p \quad \text{s.t.} \quad f_\theta(x) \neq f_\theta(x+\delta) \tag{19}$$

The expected robustnessover the data distribution $\mathcal{D}$ is

$$R(f_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[r(f_\theta, x)] \tag{20}$$

Adversarial training optimizes for worst-case performance within an ϵ-ball, effectively maximizing a lower bound on the robustness radius as follows:

$$\theta^*_{AT} = \arg\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\|\delta\|\leq\epsilon} \mathcal{L}(f_\theta(x+\delta), y) \right] \tag{21}$$

This can be interpreted as expanding the margin around training points. However, AT alone has the following known limitations:

- Robust overfitting: performance on adversarial examples degrades after prolonged training despite continued improvement on clean data.
- Accuracy–robustness trade-off: clean accuracy typically decreases by 8–15% to achieve robustness.
- Attack-specific bias: models may overfit to the specific attack used during training (e.g., PGD).

Effect of gradient smoothing: gradient penalty regularization encourages small and locally consistent gradients.

$$\mathcal{L}_{GS} = \lambda_1 \|\nabla_x \mathcal{L}\|_2^2 + \lambda_2 \|\nabla_x \mathcal{L}(x) - \nabla_x \mathcal{L}(x+\eta)\|_2^2 \tag{22}$$

The intuitive motivation is that gradient-based attacks require $\nabla_x \mathcal{L}$ to be large and informative. Regularizing gradient magnitude reduces the signal available to attackers. The curvature term ensures that even if gradients are non-zero at a point, they do not provide reliable direction information due to local inconsistency. Critical distinction from gradient obfuscation: Unlike problematic gradient obfuscation techniques that introduce non-differentiable operations (which create zero or random gradients), our approach maintains differentiable, meaningful gradients. The smoothing makes the loss landscape genuinely flatter rather than artificially hiding gradient information. This is verified empirically by the consistency between standard PGD, BPDA, and EOT attack results. The synergistic combination of AT and GS provides complementary benefits that address each other's limitations: Adversarial training (AT) expands robust regions by creating protection around training points, though the class boundaries may still contain sharp gradients that adversaries can exploit. Gradient smoothing (GS) complements this by smoothing these boundaries, making class transitions more gradual and less vulnerable to attacks. Moreover, GS helps reduce robust overfitting by regularizing gradient magnitude, preventing the model from memorizing specific adversarial perturbation patterns, and limiting overfitting to the training attack. Finally, AT provides robust feature representations that enhance the effect of GS; while GS alone on a non-robust model merely smooths an already fragile decision boundary, applying GS on top of AT allows it to operate on features that are already partially robust, resulting in a stronger combined effect. This suggests that the combined approach implicitly optimizes an objective, similar to the following:

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\|\delta\| \leq \epsilon} \mathcal{L}(f_\theta(x+\delta), y) + \gamma \cdot \mathrm{Var}_\delta[\nabla_x \mathcal{L}(f_\theta(x+\delta), y)] \right] \tag{23}$$

where the variance term encourages consistent gradients across the ϵ-ball, not just correct predictions. To empirically verify the synergistic effect, the loss landscape characteristics of different model configurations are analyzed (Table 15).

Table 15. Loss landscape characteristic analysis.

Configuration	Grad Norm	Curvature	Lipschitz	AA acc.
Standard	12.4	287.3	18.7	0.0%
AT only	8.7	156.2	12.4	51.4%
GS only	3.2	89.4	6.8	28.6%
ShieldNet	**2.1**	**34.7**	**4.2**	**72.4%**

Grad norm: Average $\|\nabla_x \mathcal{L}\|_2$; Curvature: average spectral norm of Hessian; Lipschitz: estimated local Lipschitz constant; AA acc.: AutoAttack robust accuracy. Lower values indicate smoother landscapes for the first three metrics.

The results confirm that ShieldNet achieves the smoothest loss landscape (lowest gradient norm, curvature, and Lipschitz constant) while also achieving the highest robust accuracy, supporting our intuitive analysis. The multi-objective optimization in ShieldNet involves balancing potentially competing objectives. Convergence properties were analyzed using the framework of multi-task learning. Define the combined loss as follows:

$$\mathcal{L}_{total}(\theta) = \sum_{i=1}^{K} \alpha_i \mathcal{L}_i(\theta) \tag{24}$$

Under standard assumptions (Lipschitz continuous gradients, bounded variance), the SGD on $\mathcal{L}_{total}$ converges at the following rate:

$$\mathbb{E}[\|\nabla \mathcal{L}_{total}(\theta_T)\|^2] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \tag{25}$$

The key challenge is *gradient conflict* when $\nabla \mathcal{L}_i$ and $\nabla \mathcal{L}_j$ point in opposing directions. Our empirical analysis shows that the gradients of $\mathcal{L}_{clean}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{smooth}$ maintain positive cosine similarity throughout training, indicating compatible optimization directions. This compatibility arises because all three objectives fundamentally encourage prediction stability on clean data and adversarial data, as well as under small perturbations, respectively. Despite strong overall performance, ShieldNet has identifiable limitations that are discussed transparently. Table 16 presents scenarios where ShieldNet's performance degrades most significantly.

Table 16. Failure case analysis: Scenarios with reduced performance.

Scenario	Robust Accuracy (%)	Degradation
Standard benchmark ($\epsilon = 0.03$)	72.4	–
Large perturbation ($\epsilon = 0.1$)	38.9	−33.5%
C&W attack with high confidence ($\kappa = 100$)	54.2	−18.2%
Multi-targeted PGD attack	61.8	−10.6%
Severe occlusion + PGD attack	48.3	−24.1%
Low-quality images + PGD attack	52.7	−19.7%

Baseline: AutoAttack on MMU1 with $\epsilon = 0.03$. Degradation was measured relative to baseline robust accuracy. C&W: Carlini-Wagner attack.

Analysis: At large perturbation budgets (e.g., $\epsilon = 0.1$), adversarial noise becomes semi-perceptible and begins to alter the underlying image content, making it unrealistic for any defense to preserve high accuracy without severely compromising clean performance. Similarly, high-confidence C&W attacks ($\kappa = 100$) drive adversarial examples deeper into the target class's decision region, producing misclassifications that are significantly harder to counter. Furthermore, when adversarial perturbations are compounded with natural degradations such as occlusion, blur, or low-quality inputs, the resulting performance drop exceeds the effect of either factor alone, highlighting a remaining vulnerability to distribution shifts.

Limitations: Despite ShieldNet's improved robustness–accuracy trade-off, a modest drop in clean accuracy (95.7% → 94.1% on MMU1) remains unavoidable due to the inherent tension between tightly fitting clean data and maintaining sufficient margins against adversarial perturbations. In terms of cost, ShieldNet requires roughly 4–5× more computation during training because of adversarial sample generation and gradient regularization, which may be restrictive for large-scale or frequently retrained systems, although inference cost remains unchanged. Furthermore, the method is sensitive to hyperparameters such as $\alpha_1, \ldots, \alpha_5$ and λ_1, λ_2, where suboptimal choices can either weaken robustness or excessively degrade clean accuracy, and optimal settings may vary across datasets. Another limitation lies in the evolving nature of adversarial attacks: while ShieldNet is robust against current state-of-the-art methods (including AutoAttack), future attack strategies may expose new vulnerabilities, as adversarial robustness remains an ongoing arms race. Finally, performance differences across datasets (e.g., 72.4% AA on MMU1 vs. 65.3% on UBIRIS.v2) indicate that more challenging real-world conditions, such as unconstrained capture and variable illumination, still pose significant difficulties even with robust training.

We identify the following potential attack vectors that warrant future investigation:

- Semantic adversarial examples: Perturbations that change semantically meaningful features (e.g., iris texture patterns) rather than pixel-level noise may bypass defenses that focus on ℓ_p-bounded perturbations.

- Physical-world attacks: While digital attacks are evaluated, physical attacks using patterned contact lenses, projected light, or printed overlays may transfer differently to robust models.
- Model extraction attacks: Adversaries with query access could potentially extract a surrogate model that reveals exploitable patterns in ShieldNet's decision boundary.

Table 17 contextualizes ShieldNet within recent advances in adversarial robustness. Table 18 demonstrates that ShieldNet meets latency requirements for practical deployment scenarios. Table 19 summarizes practical challenges for deploying ShieldNet in production environments. Positive Applications: ShieldNet can enhance security in legitimate biometric authentication systems, protecting against adversarial attacks that could compromise access control, financial systems, and identity verification. Potential Misuse: Robust biometric systems could potentially be deployed in surveillance applications without adequate consent or oversight. It is emphasized that this work is intended for consensual authentication scenarios with appropriate privacy protections. Fairness Considerations: Our evaluation across demographically diverse datasets (Asian, African, and Caucasian populations) helps ensure that ShieldNet does not exhibit differential vulnerability across demographic groups. However, comprehensive fairness auditing across intersectional categories remains an important direction for future work.

Table 17. Comparison with recent adversarial defense methods.

Method	Domain	Clean (%)	AA (%)	Key Idea
Gowal et al. [64]	CIFAR-10	85.3	65.9	Extra training data
Rebuffi et al. [65]	CIFAR-10	87.3	66.6	Advanced data augmentation
Wang et al. [66]	CIFAR-10	86.4	67.3	Improved adversarial training
ShieldNet (ours)	**Iris recognition**	**94.1**	**72.4**	**AT + gradient smoothing synergy**

Note: Direct numerical comparison should be interpreted cautiously due to different domains (CIFAR-10 vs. iris biometrics), different dataset sizes, and potential differences in inherent robustness properties. AA: AutoAttack robust accuracy.

Table 18. Real-time deployment feasibility analysis.

Application	Requirement	ShieldNet	Status	Notes
Access Control	<100 ms	8.2 ms	✓ Pass	Real-time feasible on V100 GPU
Mobile Authentication	<200 ms	45.6 ms	✓ Pass	Mobile CPU + Edge TPU pipeline
Airport Security	<50 ms	8.2 ms	✓ Pass	Suitable for high-throughput systems
ATM Systems	<150 ms	8.2 ms	✓ Pass	Low-latency verification
Border Control	<75 ms	8.2 ms	✓ Pass	Meets strict security constraints

ShieldNet latency measured on NVIDIA V100 GPU. Mobile latency includes preprocessing on mobile CPU + inference on edge TPU.

Table 19. Production deployment: Challenges and mitigation strategies.

Challenge	Description	Mitigation Strategy	Status
Legacy Integration	Compatibility with existing biometric infrastructure	REST API wrapper with standardized interfaces	Implemented
Privacy Compliance	GDPR/CCPA requirements for biometric data	On-device inference; no template transmission	Supported
Model Updates	Continuous improvement against new attacks	Federated learning for distributed updates	Future work
Hardware Constraints	Edge deployment on limited resources	INT8 quantization (3% accuracy loss)	Validated
Adversarial Monitoring	Runtime detection of attack attempts	Confidence calibration + anomaly detection	Implemented

This table summarizes the challenges encountered during production deployment and the corresponding strategies to mitigate them.

7. Conclusions

This paper presents ShieldNet, a comprehensive adversarially resilient framework for iris biometric authentication that addresses critical security vulnerabilities in deep learning-based biometric systems. Through rigorous experimental validation using the AutoAttack benchmark across three diverse datasets, this work demonstrates that ShieldNet achieves strong adversarial robustness while maintaining competitive clean data performance. Synergistic Defense Framework: This study introduces a principled dual-layer defense mechanism that combines adversarial training with smoothness-regularized gradient penalty. Rather than claiming fundamental novelty in its individual components, the effective integration yields synergistic benefits, as demonstrated. Furthermore, adversarial training expands robust regions while gradient smoothing stabilizes decision boundaries. The synergistic effect yields robustness gains exceeding the sum of its individual components. Rigorous Evaluation Protocol: The evaluation represents a comprehensive adversarial robustness assessment for iris biometric systems, incorporating the AutoAttack benchmark (the community standard for reliable robustness evaluation), strong PGD-100 attacks with 10 random restarts, and adaptive attack variants such as BPDA and EOT to confirm the absence of gradient obfuscation. Transfer-based black-box attacks using surrogate models and cross-dataset generalization experiments are also included to assess robustness under distribution shifts. Strong Empirical Performance: ShieldNet demonstrates substantial gains over existing defenses, achieving 72.4% AutoAttack accuracy on MMU1 (an 8.6% improvement over AWP) and 94.1% clean accuracy (7.4% higher than AWP). Under adversarial conditions, ShieldNet attains a 2.8% EER, compared to 4.6% for AWP, with consistent improvements across all datasets and attack types. Empirical analyses for evidence and intuitive justification were provided. As a consequence, for defense synergy, a loss landscape analysis demonstrates that ShieldNet achieves the smoothest decision boundaries among the evaluated methods. Ablation studies quantify the contribution of each component. Practical Deployment Viability: ShieldNet maintains a real-time inference performance (8.2 ms on GPU) suitable for production deployment, with validated integration strategies for common deployment scenarios. Several limitations are acknowledged, motivating directions for future research. Robustness to Large Perturbations: Performance degrades significantly at $\epsilon > 0.05$. Future work could explore certified defenses or semantic-aware robustness training. Our evaluation focuses on digital perturbations. Indeed, extending to physical attacks (patterned contact lenses, projected light) is an important direction. Computational Cost: Training overhead of '4–5× may be prohibitive for frequent retraining. More efficient adversarial training methods could address this limitation. Generalization to Other Biometrics: While designed for iris recognition, these principles may extend to other biometric modalities (fingerprint, face, voice). Cross-modality validation remains to be studied in future work. Adaptive Attack Evolution: As attack methodologies advance, continuous evaluation and potential defense updates will be necessary. Ongoing evaluation against emerging attacks is necessary. The vulnerability of deep learning-based biometric systems to adversarial attacks poses a significant security challenge for critical authentication infrastructure. ShieldNet represents a meaningful advance toward addressing this challenge, demonstrating that a principled combination of defense mechanisms can achieve substantially improved robustness without sacrificing practical utility. However, this study emphasizes that adversarial robustness remains an active research area with no permanent solutions. ShieldNet should be viewed as one component of an in-depth defense strategy that includes input validation, anomaly detection, and regular security auditing. It is hoped that the rigorous evaluation methodology and transparent discussion of limitations herein will contribute to a more reliable assessment of adversarial defenses in future research. The principles underlying ShieldNet, synergistic defense combination, gradient smoothing

regularization, and rigorous evaluation against adaptive attacks, are broadly applicable beyond iris biometrics. It is anticipated that these methodological contributions will inform the development of robust deep learning systems across security-critical domains.

Author Contributions: Conceptualization, A.M., G.F. and A.O.; Methodology, A.M.; Software, A.M.; Validation, A.M. and G.F.; Writing—original draft, A.M.; Writing—review & editing, G.F., A.O. and S.B.; Visualization, G.F.; Supervision, G.F., A.O. and S.B.; Project administration, A.O. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by PIAno di inCEntivi per la Ricerca di Ateneo 2024/2026, Linea di Intervento I, Progetti di ricerca collaborativa, SIAM Project, University of Catania, Italy, and G.F. is supported by PNRR MUR project PE0000013-FAIR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Battiato, S.; Gallo, G.; Nicotra, S. Perceptive Visual Texture Classification and Retrieval. In Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP), Mantova, Italy, 17–19 September 2003; pp. 524–529.
2. Wang, Z.; Wei, L.; Wang, T.; Chen, H.; Hao, Y.; Wang, X.; He, X.; Tian, Q. Enhance Image Classification via Inter-Class Image Mixup with Diffusion Model. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
3. Shi, J.; Li, C.; Gong, T.; Zheng, Y.; Fu, H. ViLa-MIL: Dual-scale Vision-Language Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
4. Toubal, I.E.; Avinash, A.; Alldrin, N.G.; Dlabal, J.; Zhou, W.; Luo, E.; Stretcu, O.; Xiong, H.; Lu, C.T.; Zhou, H.; et al. Modeling Collaborator: Enabling Subjective Vision Classification with Minimal Human Effort via LLM Tool-Use. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
5. Wang, X.; He, W.; Xuan, X.; Ono, C.S.; Piazentin Ono, J.; Li, X.; Behpour, S.; Doan, T.; Gou, L.; Shen, H.W.; et al. USE: Universal Segment Embeddings for Open-Vocabulary Image Segmentation. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
6. Lai, H.; Yao, Q.; Jiang, Z.; Wang, R.; He, Z.; Tao, X.; Zhou, S.K. CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
7. Lee, H.; Kang, K.; Ok, J.; Cho, S. CLIPtone: Unsupervised Learning for Text-Based Image Tone Adjustment. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
8. Xie, Y.; Chen, Q.; Wang, S.; To, M.S.; Lee, I.; Khoo, E.W.; Hendy, K.; Koh, D.; Xia, Y.; Wu, Q. PairAug: What Can Augmented Image-Text Pairs Do for Radiology? In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
9. Guarnera, L.; Giudice, O.; Battiato, S. Level up the Deepfake Detection: A Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models. In Proceedings of the Intelligent Systems Conference, Amsterdam, The Netherlands, 29–30 August 2024; pp. 615–625.
10. Varshney, V.; Goel, A.; Kumar, D.; Jamil, A. Using Generative Adversarial Network for Anomaly Detection in Medical Field. In *Proceedings of the International Conference on AI-Driven Technology and Social Sciences for Sustainable Future*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 197–210.
11. Li, H.; Chen, Y.; Chen, Y.; Yu, R.; Yang, W.; Wang, L.; Ding, B.; Han, Y. Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
12. Ma, Z.; Zhang, S.; Wei, L.; Tian, Q. OVMR: Open-Vocabulary Recognition with Multi-Modal References. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.

13. Maxwell, B.A.; Singhania, S.; Patel, A.; Kumar, R.; Fryling, H.; Li, S.; Sun, H.; He, P.; Li, Z. Logarithmic Lenses: Exploring Log RGB Data for Image Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
14. Noman, M.; Naseer, M.; Cholakkal, H.; Anwar, R.M.; Khan, S.; Khan, F.S. Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
15. Yang, J.; Feng, J.; Huang, H. EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
16. Kim, J.; Oh, J.; Lee, K.M. Beyond Image Super-Resolution for Image Recognition with Task-Driven Perceptual Loss. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
17. Yi, C.; Ren, L.; Zhan, D.C.; Ye, H.J. Leveraging Cross-Modal Neighbor Representation for Improved CLIP Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
18. Yu, Y.; Pan, E.; Wang, X.; Wu, Y.; Mei, X.; Ma, J. Unmixing Before Fusion: A Generalized Paradigm for Multi-Source-Based Hyperspectral Image Synthesis. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
19. Park, S.; Byun, H. Fair-VPT: Fair Visual Prompt Tuning for Image Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
20. Pham, B.D.; Tran, P.; Tran, A.; Pham, C.; Nguyen, R.; Hoai, M. Blur2Blur: Blur Conversion for Unsupervised Image Deblurring on Unknown Domains. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
21. Zaffar, M.; Nan, L.; Kooij, J.F.P. On the Estimation of Image-Matching Uncertainty in Visual Place Recognition. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
22. Ragusa, F.; Tomaselli, V.; Furnari, A.; Battiato, S.; Farinella, G.M. Food vs. Non-Food Classification. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, The Netherlands, 16 October 2016.
23. Fargetta, G.; Ortis, A.; Anile, S.; Battiato, S. Evaluation of CNNs for Wildcats Classification in Real World Scenario. In *Proceedings of the International Conference on Advanced Engineering, Technology and Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 15–25.
24. Zhang, J.; Fang, I.; Wu, H.; Kaushik, A.; Rodriguez, A.; Zhao, H.; Zhang, J.; Zheng, Z.; Iovita, R.; Feng, C. LUWA Dataset: Learning Lithic Use-Wear Analysis on Microscopic Images. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
25. Bär, A.; Houlsby, N.; Dehghani, M.; Kumar, M. Frozen Feature Augmentation for Few-Shot Image Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
26. Berrada, T.; Verbeek, J.; Couprie, C.; Alahari, K. Unlocking Pre-Trained Image Backbones for Semantic Image Synthesis. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
27. Zuccarà, R.; Fargetta, G.; Ortis, A.; Battiato, S. Exploiting Adversarial Learning and Topology Augmentation for Open-Set Visual Recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11 June 2025; pp. 3416–3424.
28. Huang, Z.; Jiang, R.; Aeron, S.; Hughes, M.C. Systematic Comparison of Semi-Supervised and Self-Supervised Learning for Medical Image Classification. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
29. Cui, X.; Aparcedo, A.; Jang, Y.K.; Lim, S.N. On the Robustness of Large Multimodal Models Against Image Adversarial Attacks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024.
30. Gonde, K.; Patil, P.W.; Vipparthi, S.K.; Murala, S.; Patil, P.; Kimbahune, V. AeroDehazeNet: Exploiting Selective Multi-Scale Transformers for Aerial Image Dehazing. In *Proceedings of the 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; IEEE: Piscataway, NJ, USA, 2024.
31. Lee, M.; Yoon, J.; Choi, C. Adversarial Attack Vulnerability for Multi-Biometric Authentication System. *Expert Syst.* **2024**, *41*, e13655. [CrossRef]
32. Park, S.H.; Lee, S.; Lim, M.Y.; Hong, P.M.; Lee, Y.K. A Comprehensive Risk Analysis Method for Adversarial Attacks on Biometric Authentication Systems. *IEEE Access* **2024**, *12*, 116693–116710. [CrossRef]
33. Deb, D.; Jain, A.; Mistry, V. AdvBiom: Adversarial attacks on biometric matchers. *arXiv* **2023**, arXiv:2301.03966. [CrossRef]
34. Kilany, S.; Mahfouz, A. A comprehensive survey of deep face verification systems adversarial attacks and defense strategies. *Sci. Rep.* **2025**, *15*, 15753. [CrossRef] [PubMed]
35. Bangoy Pacheco, S.A.; Estrada, J.E.; Goyani, M.M. Hidden adversarial attack on facial biometrics—A comprehensive survey. *Procedia Comput. Sci.* **2025**, *258*, 1383–1390. [CrossRef]

36. Dhamija, L.; Bansal, U. How to defend and secure deep learning models against adversarial attacks in computer vision: A systematic review. *New Gener. Comput.* **2024**, *42*, 1165–1235. [CrossRef]
37. Eleftheriadis, C.; Symeonidis, A.; Katsaros, P. Adversarial robustness improvement for deep neural networks. *Mach. Vis. Appl.* **2024**, *35*, 35. [CrossRef]
38. Spata, M.O.; Ortis, A.; Fargetta, G.; Battiato, S. CNNMC: A convolutional neural network with Monte Carlo dropout for speaker recognition. *EURASIP J. Inf. Secur.* **2025**, *2025*, 34. [CrossRef]
39. Ibrahum, A.D.M.; Hussain, M.; Hong, J.E. Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective. *Artif. Intell. Rev.* **2025**, *58*, 28. [CrossRef]
40. Wen, X.; Danso, E.; Danso, S. Improving security-sensitive deep learning models through adversarial training and hybrid defense mechanisms. *J. Cybersecur.* **2024**, *7*, 60893. [CrossRef]
41. Liu, J.; Jin, Y. A comprehensive survey of robust deep learning in computer vision. *J. Autom. Intell.* **2023**, *2*, 175–195. [CrossRef]
42. Croce, F.; Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the 37th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, Virtual, 13–18 July 2020; pp. 2206–2216.
43. Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; Hein, M. RobustBench: A standardized adversarial robustness benchmark. *arXiv* **2020**, arXiv:2010.09670.
44. Ma, L.; Liang, L. Improving adversarial robustness of deep neural networks via adaptive margin evolution. *Neurocomputing* **2023**, *551*, 126524. [CrossRef] [PubMed]
45. Liu, C.; Xiang, W.; Dong, Y.; Zhang, X.; Wang, L.; Duan, R.; Zheng, S.; Su, H. RobustPrompt: Learning to defend against adversarial attacks with adaptive visual prompts. *Pattern Recognit. Lett.* **2025**, *190*, 161–168. [CrossRef]
46. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; Jordan, M. Theoretically Principled Trade-off between Robustness and Accuracy. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7472–7482.
47. Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; Gu, Q. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
48. Wu, D.; Xia, S.T.; Wang, Y. Adversarial Weight Perturbation Helps Robust Generalization. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 2958–2969.
49. Rade, R.; Moosavi-Dezfooli, S.M. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
50. Liu, Z.; Liang, H.; Ranjan, R.; Zhu, Z.; Snasel, V.; Ojha, V. D2R: Dual regularization loss with collaborative adversarial generation for model robustness. In Proceedings of the International Conference on Artificial Neural Networks, Rome, Italy, 30 June–5 July 2025; pp. 208–220.
51. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 274–283.
52. Tramer, F.; Carlini, N.; Brendel, W.; Madry, A.; Kurakin, A. On Adaptive Attacks to Adversarial Example Defenses. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 1633–1645.
53. Chen, W.; Yang, K.; Yu, Z.; Chen, C.L.P.; Shi, Y. A survey on imbalanced learning: Latest research, applications and future directions. *Artif. Intell. Rev.* **2024**, *57*, 137. [CrossRef]
54. Nguyen, K.; Proença, H.; Alonso-Fernandez, F. Deep learning for iris recognition: A survey. *ACM Comput. Surv.* **2024**, *56*, 1–35. [CrossRef]
55. Gangwar, A.; Joshi, A. DeepIris: Iris recognition using a deep learning approach. *arXiv* **2019**, arXiv:1907.09380. [CrossRef]
56. Liu, G.; Zhou, W.; Tian, L.; Liu, W.; Liu, Y.; Xu, H. An efficient and accurate iris recognition algorithm based on a novel condensed 2-ch deep convolutional neural network. *Sensors* **2021**, *21*, 3721. [CrossRef]
57. Nguyen, K.; Fookes, C.; Jillela, R.; Sridharan, S.; Ross, A. Iris Recognition with Off-the-Shelf CNN Features: A Deep Learning Perspective. *IEEE Access* **2017**, *6*, 18848–18855. [CrossRef]
58. Ghilom, M.; Latifi, S. The role of machine learning in advanced biometric systems. *Electronics* **2024**, *13*, 2667. [CrossRef]
59. Kuznetsov, O.; Zakharov, D.; Frontoni, E.; Maranesi, A. AttackNet: Enhancing biometric security via tailored convolutional neural network architectures for liveness detection. *Comput. Secur.* **2024**, *141*, 103828. [CrossRef]
60. Macas, M.; Wu, C.; Fuertes, W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Syst. Appl.* **2024**, *238*, 122223. [CrossRef]
61. Shalaby, A.S.; Gad, R.; Hemdan, E.E.D.; El-Fishawy, N. An efficient CNN based encrypted iris recognition approach in cognitive-IoT system. *Multimed. Tools Appl.* **2022**, *81*, 16037–16062. [CrossRef]

62. Ortis, A.; Farinella, G.M.; Battiato, S. An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. In Proceedings of the 16th International Conference on Enterprise Information Systems (ICETE), Prague, Czech Republic, 26–28 July 2019; pp. 296–306.
63. Rundo, F.; Trenta, F.; Di Stallo, A.L.; Battiato, S. Grid Trading System Robot (GTSbot): A Novel Mathematical Algorithm for Trading FX Market. *Appl. Sci.* **2019**, *9*, 1796. [CrossRef]
64. Gowal, S.; Rebuffi, S.A.; Wiles, O.; Stimberg, F.; Calian, D.A.; Mann, T.A. Improving robustness using generated data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4218–4233.
65. Rebuffi, S.A.; Gowal, S.; Calian, D.A.; Stimberg, F.; Wiles, O.; Mann, T.A. Data Augmentation Can Improve Robustness. In Proceedings of the Advances in Neural Information Processing Systems, 6–14 December 2021; Volume 34.
66. Wang, Y.; Cao, X.; Xu, Z.; Fang, H. Features and development trends of primary care research conducted by practice-based research networks from 1991 to 2023: A scoping review protocol. *Syst. Rev.* **2023**, *12*, 229. [CrossRef]