ARTIFACT EVALUATED USENIX AVAILABLE

Sy-FAR: Symmetry-based Fair Adversarial Robustness

Haneen Najjar
Tel Aviv University

Eyal Ronen
Tel Aviv University

Mahmood Sharif
Tel Aviv University

Abstract

Security-critical machine-learning (ML) systems, such as face-recognition systems, are susceptible to adversarial examples, including real-world physically realizable attacks. Various means to boost ML's adversarial robustness have been proposed; however, they typically induce unfair robustness: It is often easier to attack from certain classes (e.g., individuals) or groups (e.g., genders) than from others. Several techniques have been developed to improve adversarial robustness while seeking perfect fairness between classes. Yet, prior work has focused on settings where security and fairness are less critical (e.g., classifying objects such as cars and ships).

Our insight is that achieving perfect parity in realistic fairness-critical tasks, such as face recognition, is often infeasible—some classes (e.g., siblings) may be highly similar, leading to more misclassifications between them. Instead, we suggest that seeking symmetry—i.e., attacks from class i to j would be as successful as from j to i—is more tractable. Intuitively, symmetry is desirable because class resemblance is a symmetric relation in most domains. Additionally, as we prove theoretically, symmetry between individuals induces symmetry between any set of sub-groups, in contrast to other fairness notions where group-fairness is often elusive.

We develop Sy-FAR, a technique to encourage symmetry while also optimizing adversarial robustness and extensively evaluate it using five datasets, with three model architectures, including against targeted and untargeted realistic attacks. The results show Sy-FAR significantly improves fair adversarial robustness compared to state-of-the-art methods. Moreover, we find that Sy-FAR is faster and more consistent across runs. Notably, Sy-FAR also ameliorates another type of unfairness we discover in this work—target classes that adversarial examples are likely to be classified into become significantly less vulnerable after inducing symmetry.[1]

[1]An extended version of this paper is available at: https://arxiv.org/abs/2509.12939.

1 Introduction

Adversarial examples—perturbed inputs that lead machine-learning (ML) models to misclassify at deployment time—pose a profound challenge to ML [3, 51]. Notably, adversarial examples are not a hypothetical risk to ML, but can be implemented under real-world constraints, thus harming ML systems' integrity in practical settings [14, 44]. For instance, adversaries may physically realize attacks against face-recognition systems via eyeglasses they can don to evade surveillance or impersonate others [44, 45].

Realizing the imminent risk to ML systems that are becoming increasingly ubiquitous, researchers have explored various directions to enhance models' adversarial robustness (e.g., [9, 15, 21, 32, 59]). Different classes of defenses have been proposed, ranging from inference-time countermeasures that may refine classification boundaries or derive security guarantees (e.g., [9, 21]) to adversarial training, which augments training with attacks to improve robustness (e.g., [15, 32, 53]). From these, adversarial training is particularly appealing due to significant improvements in robust accuracy (i.e., accuracy under attacks) without increasing inference-time latency.

Fairness Challenges in Robust Classification. While defenses substantially improve adversarial robustness, they suffer from fairness issues, due to disparities between classes (e.g., individuals in the case of face recognition) or groups (e.g., genders or ethnicities) [36, 39, 58]. In particular, while defenses enhance adversarial robustness, they often do so unevenly across classes. Said differently, it is often significantly easier to produce adversarial examples for inputs from certain (source) classes than inputs from other classes. Besides ethical implications, these disparities in robust accuracy across classes—a phenomenon termed *unfair source-class adversarial robustness*—have direct security implications. For example, if such disparities exist in a face-recognition system in surveillance, then certain subjects may still easily evade detection although the overall robustness seems high. As Fig. 1 shows, source-class unfairness is a significant issue in

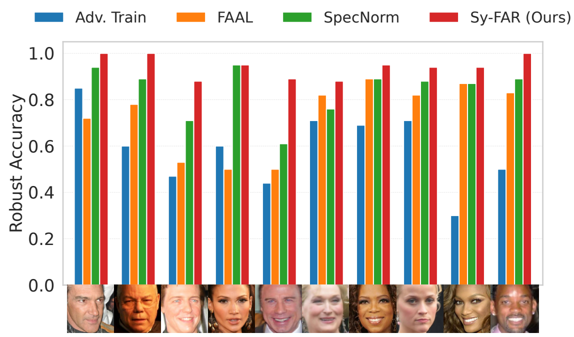


Figure 1: Per-class robust accuracy of face-recognition models trained using different defensive methods on a subset of PubFig dataset [23] using the VGG-16 architecture [48]. The models were trained to recognize a set of ten celebrities with an equal number of males and females (§5). We report results obtained with four methods: adversarial training [53], two leading approaches for enhancing fair source-class adversarial robustness (FAAL [62] and SpecNorm [18]), and our proposed method, Sy-FAR. Adversarial examples were created with (untargeted) eyeglass attacks [44].

defended face-recognition models: certain subjects receive significantly lower robust accuracy than others (e.g., Antonio Banderas, first from the left, and Beyoncé, second from the right, for the adversarially trained model).

Yet another fairness issue facing defenses that, to our knowledge, we characterize for the first time (§6.3), is that of *unfair target-class robustness*. Specifically, we find that certain classes are more likely to be the "sinks" of misclassified adversarial examples. Put simply, adversarial examples have higher probability of being misclassified *into* certain classes than others. This form of unfairness directly affects security: For example, in a face-recognition system that exhibits target-class unfairness, specific individuals would be at a higher risk of impersonation than others. Hence, in addition to countering unfair source-class adversarial robustness, as has been the focus of prior work (e.g., [18, 62]), it is critical to address unfair target-class adversarial robustness.

Last, but not least, besides achieving fairness for individuals, we should also seek to enhance fairness for different groups, such as different genders or ethnicities, so no group remains disadvantaged. Prior work has demonstrated that unfair source-class robust accuracy is exhibited not only for individual classes, but also for groups [36, 39]. A core challenge arising when seeking fairness for groups is that their number may be excessively large (exponential in the number of classes), rendering it prohibitive to improve fairness across *all* groups [40]. Accordingly, to our knowledge, no existing technique is able to promote fair adversarial robustness for an arbitrary group that may be unknown at training time.

Established Solutions. Several methods have been proposed to ameliorate unfair source-class robustness [18, 27, 31, 50, 56, 58, 62]. These roughly share the same principle: While optimizing for overall benign and robust accuracy (so these are preserved or improved compared to standard training), they also seek to directly decrease the gap in robust accuracy between classes (e.g., by assigning higher weights for classes with lower robustness) or to increase the robust accuracy of the class with minimal robustness. In doing so, leading approaches such as FAAL [62] and SpecNorm [18] substantially improve source-class fairness. However, as shown in Fig. 1, specific classes still remain highly vulnerable even under these defenses (e.g., John Travolta, fifth from the left). Notably, previous work evaluated setups where fairness and security have little implications, as they focus on artificial object classification tasks (mainly, the CIFAR datasets [22]) and impractical attacks where adversaries introduce imperceptible perturbations (bounded in ℓ_p-norm) to induce misclassifications. As a result, prior work's ability to improve fairness while preserving robustness and benign accuracy (i.e., accuracy on clean inputs) in realistic fairness- and security-critical settings remains unknown. We address this gap by porting previous approaches to realistic face-recognition settings (while also evaluating them on the original setups).

Our Solution. In this work, we argue that achieving perfect source-class fairness, as sought by prior work [18, 27, 31, 50, 56, 58, 62], may often be infeasible. For instance, assume a setting where face recognition is trained to recognize unrelated individuals as well as two siblings (or even twins). Naturally, misclassifications between the siblings would be more challenging to prevent than misclassifications between other individuals. Achieving perfect parity in this setting may even be counterproductive, as it may require reducing the robust accuracy of classes other than the siblings.

As an alternative, we propose to tackle fairness issues in adversarially robust classification via a novel technique—by *encouraging symmetry between classes*, as part of a framework for *S*ymmetry-based *F*air *A*dversarial *R*obustness (*Sy-FAR*). Besides optimizing for benign and robust accuracy, Sy-FAR trains models that would be nearly as likely to misclassify adversarial examples from class i to j as they are to misclassify in the other direction (§4). Intuitively, as class resemblance is typically a symmetric relation, it is natural to seek symmetry, such that misclassifications (of adversarial examples) between class pairs would be equally likely. Consequently, no class would be significantly more disadvantaged against the other. Crucially, we find that promoting symmetry boosts fair adversarial robustness from both source- and target-class perspectives (§6). Furthermore, as we formally prove, ensuring symmetry for individual classes also leads to symmetry between arbitrary groups (§7). Consequently, Sy-FAR, in a way completely agnostic to group definitions, encourages symmetry between arbitrary groups and, as a by-product, enforces fair source- and target-group robustness for any arbitrary partition of classes into groups.

Our Contributions. We make several key contributions:

- We propose the Sy-FAR (§4), a defense encouraging symmetry in misclassification patterns of adversarial examples, while optimizing benign and robust accuracy.
- We suggest a tractable, computationally efficient method to optimize symmetry (through a differentiable loss function) as part of Sy-FAR, lending itself to efficient integration in the training of neural networks (§4.3).
- We extensively evaluate Sy-FAR on two vision tasks (face and object recognition), using three model architectures, and two attack types (§5). Importantly, in contrast to pervious work, we consider a realistic adversary producing adversarial eyeglasses that would lead face-recognition to misclassify [44]. Thus, our work considers a real-world setting where adversarial robustness and fairness are critical. Moreover, we compare Sy-FAR empirically to FAAL [62] and SpecNorm [18], two state-of-the-art techniques.
- We find that, compared to past methods, by improving symmetry, Sy-FAR more substantially improves fair source-class adversarial robustness (§6.2). For instance, on challenging face-recognition tasks with siblings where previous methods are particularly asymmetric, Sy-FAR reduces the gap between the most and least robust classes by $\geq$41%. compared to baselines. Importantly, Sy-FAR also preserves and sometimes even improves benign and robust accuracy.
- We identify, to our knowledge for the first time, that ML models exhibit unfair target-class adversarial robustness (§6.3). Additionally, we find that Sy-FAR helps ameliorate this form of unfairness by encouraging symmetry. For example, on face recognition, Sy-FAR makes the most vulnerable class $\geq$56% less likely to be erroneously selected as the output of adversarial examples compared to baseline methods.
- We theoretically prove that encouraging symmetry for classes leads to symmetry for *any* partitioning of classes into mutually exclusive groups (§7). As a result, Sy-FAR encourages symmetry between any pair of arbitrary groups, thus promoting fair source- and target-group adversarial robustness.
- Interestingly, we find that Sy-FAR is more computationally efficient (§6.4) and consistent at optimizing models for fairness and robustness (§6.5) than alternative approaches.

The paper proceeds as follows. We next discuss related work (§2), lay out the threat model (§3), and present the technical approach behind Sy-FAR (§4). Subsequently, we describe our experimental setup (§5) before we present our results for fair (class-level) source- and target-class adversarial robustness (§6) and group fairness (§7). We wrap the paper with discussion of limitations and future work (§8) and concluding remarks (§9).

2 Related Work

2.1 Adversarial Examples

The proliferation of ML has raised researchers' interest in ML's implications on system security. Researchers have shown various attacks on ML privacy (e.g., [46]), integrity (e.g., [4, 15, 51]), and availability (e.g., [47]), among others. Specifically, adversarial examples have received special attention due to the potential risk they pose to deployed safety- and security-critical systems during deployment: Numerous attacks targeting models of varied modalities under different assumptions and with various attack objectives were proposed since the initial discovery of adversarial examples (e.g., [14, 30, 37, 42, 44, 63]).

While one class of adversarial examples leverages imperceptible adversarial perturbations, bounded in ℓ_p-norm, to mislead ML models [6, 12, 15, 35, 52], another class targets more realistic settings where adversaries may be more bound by practical constraints [14, 30, 42, 44, 45, 60]. For instance, in the *eyeglass attack* against face recognition systems [44], adversaries may need to physically realize artifacts that are robust to printing and camera noise to mislead systems under real-world constraints. Different types of attacks may have untargeted or targeted variants [38]—i.e., aiming to achieve arbitrary misclassification or misclassification into specific classes, respectively—as well as other variants suitable for other assumptions (e.g., white- and black-box [5, 28, 37]). In this work, we study the fairness properties of ML models subject to different attack types, with a particular focus on the realistic eyeglass attack.

Defenses. Different forms of defenses against adversarial examples have been proposed, including, but not limited to, defenses that detect attacks (e.g., [34]); filter out adversarial perturbations (e.g., [59]); smoothen the classification boundaries to reduce model vulnerability (e.g., [7, 9, 41]); verify robustness against specific adversaries (e.g., [21]); and adversarially train models to inherently increase their robustness (e.g., [24, 32]). Due to its intuitive nature, its ability to improve adversarial robustness in a practical manner against different attack types, and absence of impact on model's inference time, adversarial training is particularly appealing and was widely studied (e.g., [15, 25, 27, 29, 32, 43, 53, 57]). Accordingly, in this work, we focus on improving the fairness properties of adversarially trained models, especially against more realistic attacks. As adversarial training is often combined with other defenses (e.g., randomized smoothing [41]), we expect that our approach and findings would also apply to defenses other than adversarial training.

2.2 Fairness and Adversarial Robustness

Algorithmic fairness, in general, and fairness in ML, specifically, have been extensively studied in recent years [54].

Researchers have defined different fairness notions between individuals and groups, ranging from precision parity to recall parity [54], studied when notions can or cannot be satisfied simultaneously (e.g., [8]), and proposed means to promote different notions of fairness (e.g., [10,49]). In addition to ethical considerations that often motivate fairness, another key motivation behind fairness is improving the general performance of the system (e.g., benign accuracy), which sometimes stems as a side effect of improving fairness [20].

Prior work has mainly focused on studying fairness in the context of benign inputs (e.g., [1, 17, 20]). However, a line of work has studied fairness in the context of adversarial robustness—particularly, disparities in adversarial robustness between different (source) classes and groups [36, 58]. While defenses against adversarial examples such as adversarial training significantly improve robust accuracy, they do not guarantee fairness across classes or demographic groups. In high-stakes applications such as face recognition, this limitation is particularly concerning: a system may appear robust overall, yet still perform disproportionately worse on certain classes or subpopulations, exposing vulnerable groups to higher risks of misidentification or security breaches. In concurrent work, Nanda et al. [36] and Xu et al. [58] noticed that certain classes or groups may have significantly lower robust accuracy than others. These efforts highlighted the shortcomings of evaluating fairness only under benign conditions, and reinforced the need for defenses that explicitly address both robustness and fairness under attack. As noted by Rosenberg et al. [39], similar disparities are also exhibited by face-recognition models where face-obfuscation with ℓ_p-norm bounded perturbations to attain privacy are less effective for minority groups.

Improving Fair Adversarial Robustness. Several methods have emerged to tackle fairness in adversarial settings [18, 27, 31, 50, 56, 58, 62]. In particular, *Fair Robust Learning (FRL)* [58] aims to improve worst-class robust accuracy through a min–max framework that reweights adversarial samples and adjusts attack strengths during training. *Balance Adversarial Training (BAT)* [50] tackles fairness by encouraging uniformity in predicted classes and equalizing attack difficulty across source classes. *Fairness-Aware Adversarial Learning (FAAL)* [62] improves worst-class robustness by dynamically reweighting samples and using a form of so-called Distributionally Robust Optimization (DRO) to optimize worst-case performance. Last, *Confusional Spectral Regularization (SpecNorm)* [18] penalizes the spectral norm of the confusion matrix to bound the worst-case class-level robust error. Similar to SpecNorm, Sy-FAR optimizes particular properties of the confusion matrix on adversarial examples (§4). However, Sy-FAR departs from norm-based constraints and instead focuses on directional imbalances in misclassification rates. SpecNorm and FAAL are considered the leading approaches for enhancing fair source-class adversarial robustness, thus we include them in our experiments (§5).

In contrast to prior work, Sy-FAR aims to improve fair adversarial robustness through novel means, by optimizing symmetry—while this property sometime naturally arises, misclassification patterns of adversarial examples between most class pairs are often asymmetric [33]. Moreover, in this work we identify disparities in target-class adversarial robustness as a key limitation of existing techniques and show that Sy-FAR helps make it less severe. Last, previous work has mainly focused on predefined groups (e.g., genders or races), while we show how to improve fair adversarial robustness for any arbitrary group—addressing an issue that is known to be computationally challenging for certain notions of fairness.

We also note that previous defenses aiming to promote fairness and robustness have mainly conducted evaluations on object-classification benchmarks (e.g., CIFAR-10 and CIFAR-100) where robustness and fairness have little-to-no security or ethical implications. Moreover, they considered unrealistic adversaries that may fail to harm integrity under real-world constraints. In contrast, in this work, we mainly focus on high-stakes settings by adapting prior techniques and evaluating them along with Sy-FAR on face-recognition models against realistic eyeglass attacks.

3 Threat Model

We consider settings where adversaries produce adversarial examples to mislead ML-based image classifiers, particularly, ones based on neural networks. The adversaries we consider have white-box access to models, enabling them to efficiently produce worst-case perturbations using their knowledge of model weights, in contrast to black-box counterparts, which often slower and achieve lower attack success [12]. Moreover, we consider adversaries launching either untargeted or targeted attacks. Furthermore, we focus on adversaries that use realistic attacks against face-recognition (specifically, using the eyeglass attack [44]). However, for better compatibility with prior work on fair adversarial robustness [18,62], we also study adversaries that produce imperceptible adversarial perturbations bounded in ℓ_p-norm against object recognition. In both cases, the adversary's goal is to maximize their success rate—i.e., achieving any misclassification in the case of untargeted attacks, or obtaining misclassification to target classes in the case of targeted attacks—thus maximally harming the integrity of the ML-based system.

Contrastively, the defender aims to prevent attempts to mislead the ML model to the extent possible (i.e., *maximizing robust accuracy*) while *maximizing benign accuracy* on clean inputs. Specifically, the defender induces adversarial robustness at training time, through a form of adversarial training. Crucially, in this work, we place strong emphasis on fair adversarial robustness. Specifically, additionally to deterring attacks and correctly classifying benign samples, our defender seeks to *maximize fair source and target adversarial robustness, both at the individual class and group levels*.

4 Sy-FAR's Technical Approach

Our proposed symmetry-based regularization builds upon confusion-matrix analysis, which has been recognized as a foundation for fairness evaluation [16]. Here, however, we focus on confusion matrices over adversarial examples. We follow a standard adversarial training framework, where the objective balances benign accuracy and adversarial robustness, and extend it with a fairness regularizer. Unlike earlier fairness-oriented regularizers that rely on specialized optimization schemes [18, 50, 58], Sy-FAR is conceptually straightforward yet effective: it introduces a differentiable regularization term that explicitly enforces symmetry between class pairs, thereby counteracting asymmetry in adversarial settings. The intuition is that if examples of class i can be perturbed to resemble class j, then the reverse should hold as well—class similarity is fundamentally a symmetric relation. Framing fairness in this way provides a natural and tractable objective even when considering unbalanced or biased datasets. Moreover, the approach is computationally lightweight, as evaluating the regularizer requires only a single pass over the confusion matrix, making it well-suited for scalable adversarial training.

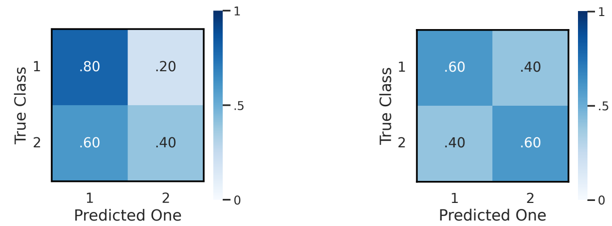
Next, we provide intuition for why symmetry helps advance fair adversarial robustness (§4.1), describe the symmetry regularizer (§4.2), and present Sy-FAR's overall objective and algorithm (§4.3).

4.1 Intuition: Why Symmetry?

Fig. 2 illustrates two confusion matrices over adversarial examples. While one matrix is asymmetric and another is symmetric, both represent models with an overall 60% robust accuracy. Note, however, that the model with the asymmetric confusion matrix exhibits unfair source- and target-class adversarial robustness: class 2 has robust accuracy lower by 40% than class 1, and adversarial examples have 40% higher likelihood of being classified into class 1 than into class 2. In contrast, the model with the symmetric confusion matrix achieves fair source- and target-class adversarial robustness (both classes have identical robust accuracy and they are equally likely of being selected as output of successful adversarial examples). Thus, overall, this illustrative example shows that if we are able to maintain (or increase) robust accuracy while improving symmetry, it is possible to promote fair adversarial robustness. In our experiments (§6–7), we find that symmetry regularization indeed preserves or boosts robust accuracy, and, as expected, also improved fair adversarial robustness from source- and target-class perspectives.

4.2 Symmetry Regularization

Confusion matrices provide a natural way to visualize how predictions are distributed across classes. In a standard



(a) Asymmetric confusion (b) Symmetric confusion

Figure 2: An illustration of asymmetric vs. symmetric confusion matrices.

("hard") confusion matrix, each entry C_{ij} counts how many examples from class i are predicted as class j. The diagonal entries represent correct classifications, while the off-diagonal entries correspond to misclassifications. A symmetric confusion matrix, C, is one where $C_{ij} = C_{ji}$, meaning that if samples of class i are often confused with class j, the reverse confusion occurs at a similar rate. By contrast, asymmetry arises when one direction dominates, e.g., class i is frequently misclassified as j but not vice versa. Such directional imbalances reflect fairness issues: one class disproportionately "loses" to another, making it more vulnerable under adversarial perturbations (as illustrated in §4.1).

Producing standard confusion matrices requires computing predicted class labels. However, since we cannot use hard predictions in gradient-based optimization due to non-differentiability, analogous to how the benign loss relies on softmax probabilities in lieu of discrete labels, we adopt a similar approach here. Specifically, we define a *soft confusion matrix*, where each entry C_{ij} represents the expected probability of samples from class i being predicted as class j. This construction leverages the softmax outputs of the model, making it directly compatible with gradient-based optimization.

Soft Confusion Matrices From Probabilities. Let K be the number of classes. For a minibatch of B adversarial examples $\{(x_b, y_b)\}_{b=1}^{B}$, where x_b denotes the input image and y_b its ground-truth label, with model logits $z_b \in \mathbb{R}^K$, we first compute the softmax probabilities $p_b = \text{softmax}(z_b)$ for each sample. To estimate the confusion matrix in a differentiable way, we form $\widehat{C} \in \mathbb{R}^{K \times K}$ by accumulating these probabilities according to ground-truth labels: for each class i, we sum (or equivalently average) the probability vectors p_b over all samples with $y_b = i$, and assign the result to row i of $\widehat{C}$. In this construction, each row corresponds to a true class, while each column reflects the (soft) predicted distribution for that class. We also track the number of samples per class using a count vector $n \in \mathbb{R}^K$. Starting with $\widehat{C}$ initialized with zeros, after accumulating the soft predictions, we normalize each row by the corresponding count to obtain a final soft confusion matrix $C \in \mathbb{R}^{K \times K}$. In other words, for each prediction p_b, we accumulate $\widehat{C}_{y_b,:} \leftarrow \widehat{C}_{y_b,:} + p_b^\top$. Subsequently, we normalize each row i by n_i, the number of samples in class i (i.e., we set $C_{i,:} \leftarrow \frac{\widehat{C}_{i,:}}{\max(n_i,1)}$). Here, $\max(n_i, 1)$ prevents division by zero in case a class is missing from the minibatch. Eventually, each cell C_{ij} estimates the conditional class distribution $\mathbb{P}(\widehat{y} = j \mid y = i)$ using soft predictions, where $\widehat{y}$ denotes the predicted class of the adversarial example.

Pairwise Asymmetry Penalty. Sy-FAR's goal is to reduce directional bias between pairs of classes: if samples from class i are often misclassified as class j, then misclassifications in the reverse direction should occur at a similar rate. To capture this, we define a penalty for each class pair, as follows. For a class pair (i, j) with $i < j$, denote $a = C_{ij}$ and $b = C_{ji}$. We then penalize *relative* asymmetry, scaled by the *total misclassification mass*:

$$\text{penalty}_{i,j} \;=\; \frac{|a-b|}{a+b+\varepsilon} \cdot (a+b), \tag{1}$$

where ε is a small constant for numerical stability, set in proportion to the symmetry-loss weight hyperparameter (default $\varepsilon = \frac{1}{K}$). The first factor $\frac{|a-b|}{a+b+\varepsilon}$ measures the degree of directional imbalance (zero when $a \approx b$), while $(a+b)$ scales the penalty more strongly for class pairs that the model frequently confuses (i.e., large off-diagonal entries). This scaling is important: without it, even negligible cases—such as one misclassification in one direction and none in the other—would contribute a large relative imbalance despite being practically insignificant. By weighting with $(a+b)$, the loss emphasizes meaningful asymmetries from frequent confusions while downplaying spurious imbalances from rare or single-sample errors.

The total symmetry loss then aggregates penalties across all unordered class pairs:

$$\mathcal{L}_{\text{sym}}(C) \;=\; \sum_{1 \le i < j \le K} \text{penalty}_{i,j}. \tag{2}$$

Interpretation and Differentiability. The symmetry regularizer is minimal when $C_{ij} = C_{ji}$ for all $i \neq j$, and it increases smoothly with both (i) the magnitude of asymmetry and (ii) the amount of mutual confusion. This design directly targets directional bias (e.g., $A \to B \gg B \to A$) while de-emphasizing rare, noisy interactions. Moreover, all steps from logits $\to$ softmax $\to$ row-normalized $C \to$ Eq. 2 involve mostly continuous and differentiable operations (addition, multiplication, absolute value, etc.), making the loss fully differentiable.

4.3 Training With Symmetry Regularization

Loss Composition. We follow the adversarial training framework, where the objective balances benign accuracy and adversarial robustness, and extend it with our symmetry regularizer. Specifically, the total loss combines clean loss, adversarial loss, and symmetry loss:

$$\mathcal{L} \;=\; \lambda_{\text{clean}}\, \mathcal{L}_{\text{CE}}(x,y) \;+\; \lambda_{\text{adv}}\, \mathcal{L}_{\text{CE}}(x^{\text{adv}},y) \;+\; \lambda_{\text{sym}}\, \mathcal{L}_{\text{sym}}(C),$$

Algorithm 1 Sy-FAR's Algorithm

Require: dataset $\mathcal{D}$, model f_θ, attack $\mathcal{A}$, weights $(\lambda_{\text{clean}}, \lambda_{\text{adv}}, \lambda_{\text{sym}})$

for each minibatch $(x,y) = \{(x_b, y_b)\}_{b=1}^{B}$ **do**
- **Clean pass:** $z \leftarrow f_\theta(x)$, $\quad \mathcal{L}_{\text{clean}} \leftarrow \mathcal{L}_{\text{CE}}(z,y)$
- **Adversarial examples:** $x^{\text{adv}} \leftarrow \mathcal{A}(f_\theta, x, y)$
- **Adversarial pass:**
  - $z^{\text{adv}} \leftarrow f_\theta(x^{\text{adv}})$, $\quad p \leftarrow \text{softmax}(z^{\text{adv}})$
- **Soft confusion:** build C per §4.2
- **Symmetry loss:** $\mathcal{L}_{\text{sym}}(C) \leftarrow \sum_{i<j} \frac{|C_{ij} - C_{ji}|}{C_{ij} + C_{ji} + \varepsilon}(C_{ij} + C_{ji})$
- **Adversarial loss:** $\mathcal{L}_{\text{adv}} \leftarrow \mathcal{L}_{\text{CE}}(z^{\text{adv}}, y)$
- **Total loss:** $\mathcal{L} \leftarrow \lambda_{\text{clean}} \mathcal{L}_{\text{clean}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{sym}} \mathcal{L}_{\text{sym}}(C)$
- **Update:** backpropagate $\nabla_\theta \mathcal{L}$ and step optimizer

end for

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss on clean inputs (x, y) and adversarial inputs (x^{adv}, y); C is the soft confusion matrix of the adversarial batch; and $\lambda_{\text{clean}}, \lambda_{\text{adv}}, \lambda_{\text{sym}} > 0$ are scalar hyperparameters balancing the three losses.

Alg. 1 summarizes the training procedure, where each minibatch involves a clean forward pass, adversarial example generation, construction of the soft confusion matrix, and the computation of the symmetry loss. The resulting objective integrates these components and updates the model parameters accordingly via gradient descent.

Complexity. Computing $\mathcal{L}_{\text{sym}}(C)$ involves comparing each unordered class pair (i, j) exactly once, resulting in $O(K^2)$ operations per batch for K classes. The computation consists only of simple element-wise differences and sums over the confusion matrix, which are fully vectorizable on a GPU. In typical regimes, the added cost is negligible compared to the standard forward and backward passes, making Sy-FAR highly efficient even for tasks with many classes.

5 Experimental Setup

This section outlines the baseline methods, datasets, models, attack settings, and evaluation metrics used throughout our experiments. A high-level summary of the experimental configurations is provided in Table 1. We present the results in §6, where we highlight consistent gains in robustness and fairness over all baselines.

Datasets and Models. To evaluate fair adversarial robustness under realistic settings, we conduct experiments on three variations of the PubFig face-recognition dataset [23]. First, we use a gender-balanced subset of PubFig featuring ten celebrity identities (five female, five male), with approximately 300 images per identity.[2] This selection was motivated by the

[2]Males: Antonio Banderas; Colin Powell; Hugh Grant; John Travolta; Will Smith. Females: Jennifer Lopez; Meryl Streep; Oprah Winfrey; Reese Witherspoon; Tyra Banks.

Data	Model	Def.	Eval	Train
PubFig	VGG-16	ROA	Eyeglass U+T	10e FT
PubFig + Sibs	VGG-16	ROA	Eyeglass U+T	5e FT
PubFig	ViT	ROA	Eyeglass U+T	5e FT
CIFAR-10	PreAct-ResNet18	PGD-ℓ_∞ ($\varepsilon = 8/255$, $\alpha = 2/255$, 10 iters)	AutoAttack	200e Scratch
CIFAR-100	PreAct-ResNet18	PGD-ℓ_∞ ($\varepsilon = 8/255$, $\alpha = 2/255$, 10 iters)	AutoAttack	200e Scratch

Table 1: Summary of experimental setups. U and T stand for untargeted and targeted attacks, respectively. FT stands for fine-tuning epochs, and Scratch denotes training from scratch.

need to control for demographic imbalance: by ensuring equal male–female representation, differences in misclassification rates can be attributed to model behavior rather than dataset bias. As our main model, we use the VGG-16 architecture [48], a well-established convolutional neural network (CNN). Second, we modify the dataset to PubFig$_{\text{SIB}}$ by replacing two female identities in PubFig with a pair of visually similar female siblings.[3] Doing so preserves the gender balance (five male, five female identities) and the dataset size ($\sim$300 images per identity, as in PubFig), with the two replaced identities selected at random to maintain diversity. PubFig$_{\text{SIB}}$ setup stress-tests the system's ability to ensure fairness not only across gender but also among individuals with high visual similarity—a critical scenario for symmetry-aware fairness evaluation. Finally, we test generalization beyond CNNs by applying Sy-FAR to a Vision Transformer (ViT) trained on PubFig$_{\text{ViT}}$.

We pre-process all face images using the PyTorch-based FaceX-Zoo toolkit [55], which provides standard pipelines for face alignment and cropping. We adopt an 80/10/10 split for training, validation, and testing, a standard ratio that provides sufficient training diversity while keeping held-out data for reliable evaluation. Across all experiments, we perform ten independent repetitions per setup and report averaged results, where each training method is evaluated under its optimized hyperparameters (see the extended version for details on the training method).

All methods are fine-tuned on top of adversarially trained models, where we empirically find that five epochs suffice for the PubFig$_{\text{SIB}}$ and PubFig$_{\text{ViT}}$ settings, while ten epochs yielded optimal results for PubFig.

To enable direct comparison with prior work that focused primarily on CIFAR-10 and CIFAR-100 [22], and to demonstrate that our results generalize beyond face recognition, we also evaluate Sy-FAR on these benchmarks [22] with Preact-ResNet18 model, as commonly used in the adversarial robustness and fairness literature [18,50,58,62]. Since these datasets do not capture the real-world fairness challenges central to our work (e.g., demographic balance or sibling similarity), we defer the detailed results in App. A. For both CIFAR-10 and CIFAR-100 with PreAct-ResNet18, all models are adversarially trained from scratch for 200 epochs.

Defenses and Baselines. All face recognition experiments adopt the Rectangular Occlusion Attack (ROA) defense, following the FACESEC framework [53]. ROA simulates physical-world patch-based occlusions by performing an exhaustive search over fixed-size rectangular regions to find the most damaging location, then runs constrained attack within that region. We use FACESEC's default parameters: rectangle size 70×70, stride (10,10), step size of α=20, and 100 attack iterations, with occlusion color initialized to mid-gray (255/2 in pixel space). These settings give the best results in our experiments, offering strong robustness while maintaining competitive clean accuracy, and we adopt them consistently across all face-recognition setups.

On CIFAR-10 and CIFAR-100, we adversarially train the PreAct-ResNet18 model using PGD-ℓ_∞ [32], with perturbation budget ε=8/255, step size α= 2/255, and ten iterations per step. This configuration is common in recent fair adversarial robustness studies, including FAAL [62] and SpecNorm [18], and thus provides a standard benchmark for comparison.

Fine-tuning for Enhancing Robust Fairness. Since our focus is specifically on the challenge of fair adversarial robustness, we build on adversarially trained models that already achieve a reasonable level of average robustness. This strategy is consistent with prior work, such as FAAL [62], which fine-tunes on top of adversarial training to further improve fairness-oriented objectives. In this context, we investigate whether fairness can be effectively improved by fine-tuning robust models, rather than retraining entirely from scratch. Our experiments indicate that adversarial fine-tuning is a more efficient approach: it enhances both fairness and robustness while avoiding the overhead of training from scratch.

Accordingly, we adopt fine-tuning as our primary setting for face-recognition tasks. For completeness, in the object recognition benchmarks (CIFAR-10/100), we explored both training from scratch and fine-tuning, and found their performance to be nearly identical; reported numbers (App. A) correspond to the best-performing configuration.

Attacks. We evaluate the face-recognition models under the realistic eyeglass attack [44]. The attack overlays a fixed eyeglass-frame mask onto the face image and perturbs pixel values only within the masked region. The perturbation is initialized to a fixed color maximizing the (mis)classification loss. Adversarial examples are then refined through iterative gradient-based updates constrained to the eyeglass region, with momentum to stabilize optimization. We test a wide range of iterations capturing both weak and strong attack regimes; results are reported for the strongest attack configuration (with 300 iterations). Moreover, we use both the

[3] We replace Meryl Streep and Oprah Winfrey with Dakota and Elle Fanning, while retaining the remaining identities from PubFig.

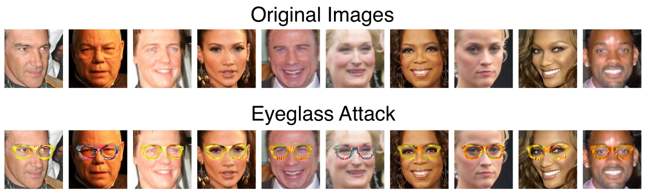


Figure 3: Illustration of the eyeglass attack on PubFig images [23]. Original face images are at the top, and images perturbed with adversarial eyeglasses are at the bottom [44]. The attack only modifies pixels within the eyeglass region, producing perturbations that mislead face recognition.

untargeted (primarily) and targeted variants of the attack. This setting reflects realistic threats where adversarial glasses can be realized by adversaries to mislead face recognition, making it a natural testbed for fairness evaluation. Fig. 3 depicts attack examples with images from PubFig, demonstrating how visually plausible perturbations can drastically impact recognition outcomes.

On CIFAR-10 and CIFAR-100, evaluation is performed using AutoAttack [12], a parameter-free ensemble of four strong adversarial attacks: (1) APGD-CE, (2) APGD-DLR [12], (3) FAB [11], and (4) Square Attack [2]. Consistent with prior fairness-aware adversarial training approaches, including FAAL [62] and SpecNorm [18], we adopt this evaluation setup as it provides a reliable benchmark for comparison.

Evaluation Metrics. We assess accuracy, robustness, and fairness with multiple metrics:

- **Benign Accuracy:** The standard classification accuracy under clean (unperturbed) inputs. This serves as a sanity check for overall model performance and ensures that robustness or fairness improvements do not come at the expense of degraded clean accuracy.
- **Robust Accuracy:** The average accuracy under adversarial attack (untargeted or targeted, depending on the setup). This is the primary measure of adversarial robustness and is widely used in the literature (e.g., [13, 19, 26, 32, 61]).
- **Source-class Fairness:** We also evaluate fair adversarial robustness with respect to the source class. We adopt two standard metrics from the literature to quantify whether robustness is distributed evenly across classes [18, 50, 58, 62]: *(1) Min (Worst-Class) Accuracy* reports the lowest diagonal entry of the row-normalized confusion matrix, capturing the most vulnerable class under attack; and *(2) Accuracy Gap* measures the difference between the maximum and minimum values on the confusion matrix's diagonal, measuring disparity across classes.
- **Symmetry-based Metrics:** These metrics quantify directional bias in misclassifications using the hard confusion matrix (C): *(1) Max Asymmetry Gap* measures the largest imbalance between two classes, $\max_{i,j}|C_{ij} - C_{ji}|$; and *(2) Symmetry Loss* is our regularization loss, $\mathcal{L}_{\text{sym}}(C)$, from Eq. (2). These metrics reveal whether class confusions are bidirectionally balanced (i.e., misclassification of adversarial examples from $i \to j$ occurs as often as from $j \to i$). Lower values indicate more symmetric and fair behavior.
- **Target-class Fairness:** Here, metrics evaluate fairness from the target-class perspective. For each target class j, we compute the fraction of total misclassifications that are directed into j:

$$T_j = \frac{\sum_{i \neq j} C_{ij}}{\sum_{i \neq j} \sum_{k \neq i} C_{ik}}.$$

  This *target confusion share* indicates how often class j becomes the destination of misclassifications. Based on the fraction of total misclassifications, we define three quantities: *(1)* TgtMax $= \max_j T_j$: here, lower is better, since this quantity reveals whether one class receives a disproportionately large share of errors (bias); and *(2)* TgtMin $= \min_j T_j$: here, higher is better, since very low values imply that some classes are almost never predicted, rendering them more advantaged than others. *(3)* TgtStd $= \text{Std}_j(T_j)$: the standard deviation across all $\{T_j\}$ values, capturing how evenly misclassifications are distributed among target classes (lower indicates greater fairness).

6 Class-Level Fairness: Experiment Results

This section presents our experimental results, demonstrating that, by improving symmetry between individual classes (§6.1), Sy-FAR improves robust fairness from source- and target-class perspectives while mostly preserving or improving benign and robust accuracy compared to baseline approaches (§6.2–6.3). Our experiments also demonstrate faster run time compared to baselines (§6.4), and higher stability across runs (§6.5). Overall, these results provide a comprehensive view of how symmetry-based regularization advances robustness and fairness simultaneously across diverse settings (datasets, model architectures, etc.).

In this section, we focus on untargeted eyeglass attacks against the face-recognition models. We note that the results on CIFAR-10 and CIFAR-100, reported in App. A follow consistent trends with those obtained on the face-recognition models. Similarly, results for the *targeted* eyeglass attack (App. B) and powerful *face-mask attacks* [53] (App. E) also corroborate our main findings.

6.1 Symmetry Loss

We begin by evaluating whether explicitly optimizing for symmetry in parallel to other objectives (as described in §5) indeed improves symmetry on the adversarial examples produced from the test samples of the dataset. Table 2 summarizes the results for the face-recognition datasets and models, using the untargeted eyeglass attack to produce adversarial examples. It can be seen that, compared to vanilla adversarial training, FAAL and SpecNorm, both yield better symmetry (i.e., lower asymmetry gap and symmetry loss), showcasing the connection between fairness and symmetry—optimizing for fair source-class robustness directly improves symmetry even without explicitly optimizing for symmetry. In comparison, Sy-FAR has the largest impact on symmetry, achieving the lowest asymmetry gap and symmetry loss across all methods evaluated. Sy-FAR's gains are particularly pronounced in the $\text{PubFig}_{\text{SIB}}$ setting, where visually similar identities exacerbate directional misclassifications.

A deeper analysis shows that asymmetry in FAAL and SpecNorm stems from misclassifications between particular class pairs, where adversarial examples from one class are markedly more likely to be classified as the other class than vice versa. For instance, in the case of the $\text{PubFig}_{\text{SIB}}$ dataset (Fig. 4), FAAL and SpecNorm are particularly asymmetric for classes 1 and 2 (Dakota and Elle Fanning, two sisters), 6 and 8 (Jennifer Lopez and Reese Witherspoon), and 6 and 9 (Jennifer Lopez and Beyoncé). By contrast, Sy-FAR yields higher symmetry, including for challenging classes where FAAL and SpecNorm struggle. Said differently, with Sy-FAR, pairwise confusions are more symmetric and no class disproportionately "loses" to another. Similar trends are observed for the CIFAR datasets (Table 5 in App. A) and targeted attacks (for full results see App. B in the extended version).

6.2 Source-Class Robustness and Fairness

We next explore the effect of optimizing symmetry as part of Sy-FAR on fair source-class adversarial robustness as well as models' overall accuracy and robustness. Specifically, we focus on benign accuracy, robust accuracy, worst-class robust accuracy, and robustness disparities between classes, as these metrics are the standard metrics in the literature on fair adversarial robustness [18, 50, 58, 62].

Table 3 presents the results for the face-recognition datasets and models. It can be seen that Sy-FAR preserves or even slightly improves benign accuracy compared to baseline approaches. In terms of robust accuracy, Sy-FAR consistently surpasses prior methods, with the most striking improvement on the challenging $\text{PubFig}_{\text{SIB}}$ setting: robustness improves by 6.5% over the strongest baseline (SpecNorm) and 10.7% over FAAL. This demonstrates that symmetry regularization remains effective even when visually similar identities exacerbate misclassifications.

For worst-class accuracy and the fairness gap, the gains are even more dramatic. On $\text{PubFig}_{\text{SIB}}$, compared to FAAL and SpecNorm, Sy-FAR improves the minimum robust accuracy by a factor of 62.6$\times$ and 3.5$\times$, respectively, while narrowing the disparity gap by 51.3% and 40.9%.

Overall, these results show that explicitly enforcing symmetry yields both robustness and fairness gains even in the most challenging scenarios, while preserving high clean accuracy—a combination unmatched by established baselines. As shown in Fig. 5, Sy-FAR yields the strongest diagonals, reflecting the highest source-class robust accuracy among all methods. Comparable patterns are evident for the CIFAR models (Table 6 in App. A) and for targeted attacks (for full results see App. B in the extended version).

6.3 Target-Class Fairness

If a target class is more likely to be selected as the output of misclassified adversarial examples, this class could be significantly disadvantaged compared to others. Particularly, in the context of face recognition, such a class (i.e., individual) may be at a disproportionally higher risk of impersonation than other classes. Indeed, we find, to our knowledge, for the first time, that certain classes are markedly more likely to be selected as the output of adversarial examples than others. Fig. 5 depicts this form of unfair target-class robustness. Specifically, it can be seen that, for models trained via FAAL or SpecNorm on PubFig, certain classes are disproportionally more likely to be selected as the output of misclassified adversarial examples (especially, Hugh Grant, Colin Powell, and Will Smith). This issue is less pronounced for Sy-FAR.

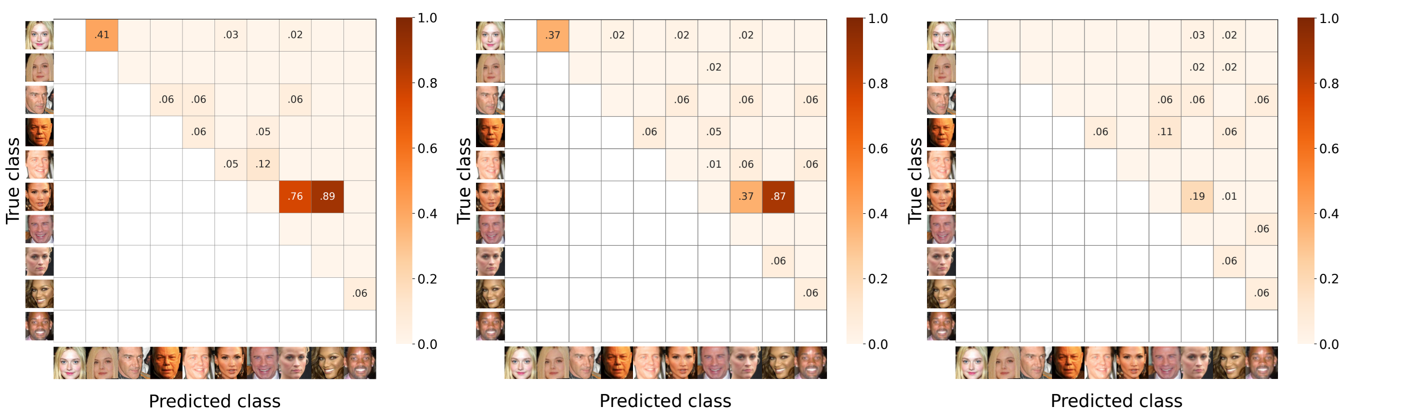
Table 4 reports Target-Class Fairness metrics, which evaluate how misclassifications are distributed across target classes. We consider the maximum target confusion share (MaxTgt$\downarrow$), the minimum target confusion share (MinTgt$\uparrow$), and the standard deviation across classes (Std$\downarrow$).

In terms of MaxTgt, Sy-FAR consistently achieves the lowest concentration of errors across all datasets. On the challenging $\text{PubFig}_{\text{SIB}}$ setup, Sy-FAR significantly lowers the worst-case error share to 0.2008, compared to 0.4543 for SpecNorm and 0.5254 for FAAL, reducing the dominant error sink by 55.8% and 61.8%, respectively. For MinTgt, Sy-FAR raises the smallest target share across all setups, ensuring that no identity is systematically neglected as a predicted class and the gap between the most vulnerable and least vulnerable target classes is small. Finally, for Std, which measures dispersion of target shares, Sy-FAR consistently yields the lowest values. On $\text{PubFig}_{\text{SIB}}$, for example, it reduces dispersion to 0.0505 compared to 0.1226 for SpecNorm and 0.1480 for FAAL, cutting skew by 58.8% and 65.9%, respectively. Overall, the results confirm that enforcing symmetry improves not only robustness and source-class fairness, but also target-class fairness—ensuring that no single class disproportionately carries the burden of errors, even in the most challenging setups.

In Fig. 5, Sy-FAR also produces the faintest off-diagonals across columns, indicating fewer concentrated misclassifications and improved target-class fairness. Consistent trends are also confirmed for CIFAR (Table 7 in App. A) and targeted attacks (for full results see App. B in the extended version).

Method	PubFig		$\text{PubFig}_{\text{SIB}}$		$\text{PubFig}_{\text{ViT}}$	
	Asym. Gap ↓	**Sym. Loss ↓**	**Asym. Gap ↓**	**Sym. Loss ↓**	**Asym. Gap ↓**	**Sym. Loss ↓**
Adv. Train	0.5491	2.1812	0.3781	1.2967	0.3052	1.4260
FAAL	0.2728	1.0803	0.8159	1.8245	0.4210	1.7487
SpecNorm	0.1864	0.4854	0.6641	1.2933	0.2571	1.4029
Sy-FAR	**0.1704**	**0.4162**	**0.2320**	**0.7214**	**0.2386**	**1.3159**

Table 2: The impact of different training approaches on symmetry, as captured by the Max Asymmetry Gap and the Symmetry Loss. The results are reported for three setups (PubFig, $\text{PubFig}_{\text{SIB}}$, and $\text{PubFig}_{\text{ViT}}$), using the untargeted eyeglass attack to generate adversarial examples, and are averaged across ten runs.



(a) FAAL (b) SpecNorm (c) Sy-FAR

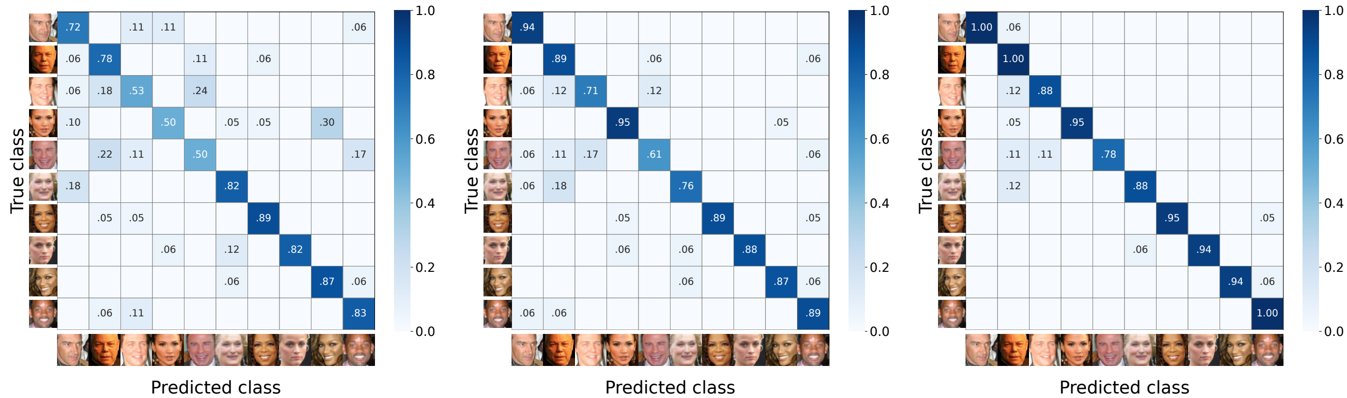
Figure 4: Asymmetry heatmaps on $\text{PubFig}_{\text{SIB}}$, using the untargeted eyeglass attack to create adversarial examples. Each cell (i, j) in the upper triangle reports the *Asymmetry Gap*, i.e., $|C_{ij} - C_{ji}|$ (see §5). Darker regions indicate stronger directional bias, where adversarial examples from one class are more likely to be classified into the other class than vice versa. We show representative heatmaps from individual randomly selected runs out of the ten repetitions.

Method	PubFig				$\text{PubFig}_{\text{SIB}}$				$\text{PubFig}_{\text{ViT}}$			
	Benign ↑	**Robust ↑**	**Min ↑**	**Gap ↓**	**Benign ↑**	**Robust ↑**	**Min ↑**	**Gap ↓**	**Benign ↑**	**Robust ↑**	**Min ↑**	**Gap ↓**
Adv. Train	96.91	58.34	2.50	92.62	95.35	68.62	28.65	67.39	82.02	63.22	32.25	61.66
FAAL	96.91	71.27	40.16	53.02	88.19	70.65	0.78	99.22	81.41	60.06	21.97	70.80
SpecNorm	98.26	86.41	66.25	33.75	93.04	73.46	14.06	81.84	83.39	65.68	30.86	62.86
Sy-FAR	**98.53**	**89.44**	**70.72**	**29.28**	**95.47**	**78.23**	**48.85**	**48.37**	**83.87**	**67.82**	**33.66**	**58.57**

Table 3: The impact of different training approaches on accuracy and source-class fairness. The results are reported for three setups (PubFig, $\text{PubFig}_{\text{SIB}}$, and $\text{PubFig}_{\text{ViT}}$), using the untargeted eyeglass attack to generate adversarial examples, and are averaged across ten runs. We report benign accuracy (Benign↑), robust accuracy (Robust↑), worst-class accuracy (Min↑), and the class-level robust accuracy gap (Gap↓).

6.4 Run-Time Comparison

We compare the computational cost of Sy-FAR against the baseline methods on the PubFig dataset using a VGG16 backbone. Experiments were conducted on a cluster node equipped with a single NVIDIA TITAN Xp GPU (12 GB), and the reported times are averaged across ten independent runs. Fig. 6 shows that Sy-FAR is substantially more efficient than FAAL and SpecNorm. By contrast, FAAL is nearly five times slower than Sy-FAR because of its bi-level optimization procedure that repeatedly solves inner adversarial updates for fairness-aware weighting. SpecNorm also incurs a significant slowdown—more than a factor of two—stemming from the need to compute the spectral norm of confusion matrix via singular value decomposition once per epoch. By comparison, standard adversarial training serves as the baseline cost (30



(a) FAAL (b) SpecNorm (c) Sy-FAR

Figure 5: Confusion matrices for different methods on adversarial examples produced with untargeted eyeglass attack against the PubFig setup. Diagonals indicate source-class robust accuracy; off-diagonals are misclassifications. We show representative heatmaps from individual randomly selected runs out of the ten repetitions.

Method	**PubFig**			**PubFig$_{\text{SIB}}$**			**PubFig$_{\text{ViT}}$**		
	MinTgt $\uparrow$	**MaxTgt** $\downarrow$	**Std** $\downarrow$	**MinTgt** $\uparrow$	**MaxTgt** $\downarrow$	**Std** $\downarrow$	**MinTgt** $\uparrow$	**MaxTgt** $\downarrow$	**Std** $\downarrow$
Adv. Training	0.0000	0.3174	0.0983	0.0276	0.2353	0.0631	0.0162	0.2291	0.0700
FAAL	0.0048	0.3145	0.0880	0.0048	0.5254	0.1480	0.0001	0.2141	0.0725
SpecNorm	0.0099	0.3120	0.0886	0.0141	0.4543	0.1226	0.0146	0.1805	0.0601
Sy-FAR	**0.0103**	**0.2563**	**0.0796**	**0.0339**	**0.2008**	**0.0505**	**0.0224**	**0.1709**	**0.0591**

Table 4: The impact of different training approaches on target-class fairness. The results are reported for three setups (PubFig, PubFig$_{\text{SIB}}$, and PubFig$_{\text{ViT}}$), using the untargeted eyeglass attack to generate adversarial examples, and are averaged across ten runs. We report the minimum and maximum normalized misclassification into each target class (MinTgt$\uparrow$ and MaxTgt$\downarrow$, respectively) and the standard deviation (Std) across target classes.

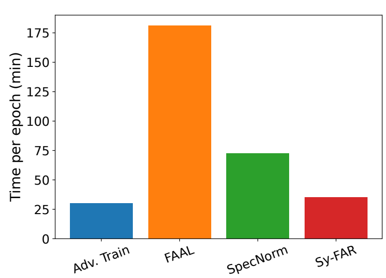


Figure 6: Training time per epoch across different methods.

minutes/epoch), and Sy-FAR adds only a negligible overhead on top of it, since the $O(K^2)$ operations from the confusion matrix (§4.3) are minor compared to the far larger cost of weight updates and adversarial example generation. Additional run-time analysis is included in the extended version.

6.5 Stability Across Runs

Beyond average performance, a key requirement for defensive methods is *stability*: producing consistently strong results across different training runs. An unstable training method may require multiple attempts or fail to converge reliably under practical time and computational constraints. Indeed, we find that certain approaches are markedly more stable than others. We measure stability by examining the variance in robust accuracy, worst class accuracy and accuracy gap across ten independent runs. The results (shown in Figs. 8–10 in App. C) demonstrate that Sy-FAR achieves superior stability (i.e., lower variance) compared to alternative approaches not only in terms of robust accuracy, but also in terms of fair adversarial robustness. One possible explanation is that FAAL relies on bilevel optimization for fairness-aware weighting, which introduces additional non-convexity and sensitivity to initialization, while SpecNorm applies global spectral constraints that can interact unpredictably with adversarial updates. By comparison, Sy-FAR introduces a symmetry regularization term that, we posit, acts as a smoother objective, reducing run-to-run variance without adding significant optimization instability.

7 Fairness for Subgroup

Ensuring fairness across all subgroups is a central challenge in ML. As Rothblum and Yona [40] point out, the number of potential subgroups in a classification task grows exponentially with the number of classes K. Specifically, the number of possible partitions of K classes is given by the Bell number, which already exceeds 100,000 for K=10. Even if we restrict attention to fairness between pairs of disjoint subgroups, the combinatorial explosion remains substantial—there are over 28,000 such pairs when K=10—rendering exhaustive subgroup fairness enforcement computationally infeasible.

To address this, Sy-FAR sidesteps the need to explicitly define or enforce fairness for all possible subgroups. Instead, Sy-FAR leverages a simple yet effective regularizer that promotes pairwise symmetry in the confusion matrix: it directly penalizes asymmetries in the likelihood of misclassifying class i as j vs. j as i. Consequently, as we now show, by minimizing directional imbalances across all class pairs, Sy-FAR promotes symmetry not only at the individual class level but also across any subgroup composed of these classes. While symmetry does not explicitly enforce fairness, our experiments demonstrate that encouraging such symmetry consistently leads to improved fairness outcomes at the subgroup level. This property is powerful: when symmetry holds between all class pairs, it induces symmetry over arbitrary partitions and combinations of these classes. As a result, and as a by product of the regularization, Sy-FAR's symmetry-based regularization enhances *subgroup fairness*, without requiring explicit group labels, domain knowledge, or combinatorial enumeration.

7.1 Defining (Sub)group Symmetry

To define subgroup symmetry in classification, we build on the soft confusion matrix defined in §4, where each entry C_{ij} denotes the expected probability of predicting class j given true class i. Recall that the matrix is constructed from softmax probabilities over adversarial examples and each row sums to one. Let $P = \{G_1, \ldots, G_m\}$ be a partition of the K classes into m disjoint subgroups, with $G_i \subseteq \{1, \ldots, K\}$ and $G_i \cap G_j = \emptyset$ for $i \neq j$. Each subgroup G_i may represent a demographic attribute (e.g., gender), a shared feature (e.g., similar appearance), or any meaningful grouping not necessarily known *a priori* during training. Using this notation, we now define the notion of symmetry at the (sub)group-level.

Definition 1 (Subgroup Misclassification Rate)**.** *Given a partition $P = \{G_1, \ldots, G_m\}$ of the K classes, the* subgroup misclassification rate *from G_a to G_b is defined as*

$$\widehat{C}_{G_aG_b} := \frac{1}{n_a n_b} \sum_{i \in G_a} \sum_{j \in G_b} C_{ij}.$$

where $n_a = |G_a|$ and $n_b = |G_b|$ are the subgroup sizes.

This quantity captures the average probability of misclassifying a sample from subgroup G_a into any class in subgroup G_b, adjusted for subgroup sizes.

Definition 2 (Subgroup Symmetry)**.** *Two subgroups G_a and G_b are said to be* symmetric *if their misclassification rates are approximately equal:*

$$\widehat{C}_{G_aG_b} \approx \widehat{C}_{G_bG_a}.$$

Subgroup symmetry ensures that errors between G_a and G_b are not directionally biased.

7.2 From Individual to Subgroup Symmetry

Sy-FAR promotes a novel form of fairness by enforcing symmetry in the confusion matrix, requiring that the probability of misclassifying class i as j matches the reverse direction, i.e., $C_{ij} = C_{ji}$ for all $i \neq j$. This criterion, which we refer to as *class-level symmetry*, is both intuitive and efficient to compute. To ensure fairness at the *group level*, however, we must show that the same principle extends to partitions of the label space. Formally, given any partition $P = \{G_1, \ldots, G_m\}$ of the K classes, we require that any two subgroups satisfy subgroup symmetry:

$$\widehat{C}_{G_aG_b} \approx \widehat{C}_{G_bG_a}, \quad \forall G_a, G_b \in P,\ a \neq b.$$

At first glance, enforcing such subgroup symmetry appears to demand evaluating an exponential number of subgroup pairs, which is computationally prohibitive. Remarkably, we prove that our class-level symmetry regularizer is sufficient: enforcing $C_{ij} = C_{ji}$ at the class level *implies subgroup symmetry fairness* across all partitions and vice versa. The key insight is that both class-level and subgroup-level symmetry arise from the same confusion matrix. Symmetry in individual off-diagonal entries naturally aggregates into symmetric relationships between groups. We now formally state and prove the equivalence between class- and subgroup-level symmetry.

Theorem 1. *Let $C \in \mathbb{R}^{K \times K}_{\geq 0}$ be a normalized confusion matrix. Let $P = \{G_1, G_2, \ldots, G_m\}$ be a partition of $\{1, \ldots, K\}$ into disjoint subgroups, then:*

1. *If C is symmetric, i.e., ($C_{ij} = C_{ji}$ for all $i \neq j$), then subgroup symmetry holds: $\widehat{C}_{G_aG_b} = \widehat{C}_{G_bG_a}$ for all $G_a \neq G_b$ in any partition P.*
2. *Conversely, if subgroup symmetry holds for* all *partitions P, then C must be symmetric.*

Said concisely, C is symmetric if and only if $\widehat{C}$ is symmetric for all subgroup partitions.

Proof. **(1) Class symmetry ⇒ Symmetry for subgroups.** Assume $C_{ij} = C_{ji}$ for all $i \neq j$, and consider any two distinct subgroups G_a, G_b. By definition,

$$\widehat{C}_{G_aG_b} = \frac{1}{n_an_b}\sum_{i\in G_a}\sum_{j\in G_b} C_{ij} = \frac{1}{n_an_b}\sum_{i\in G_a}\sum_{j\in G_b} C_{ji}.$$

Switching the order of summation (finite sums) gives

$$\widehat{C}_{G_aG_b} = \frac{1}{n_an_b}\sum_{j\in G_b}\sum_{i\in G_a} C_{ji} = \widehat{C}_{G_bG_a}.$$

Thus $\widehat{C}_{G_aG_b} = \widehat{C}_{G_bG_a}$ for all $a \neq b$.

(2) Symmetry for all subgroup partitions ⇒ Class Symmetry. Suppose $\widehat{C}_{G_aG_b} = \widehat{C}_{G_bG_a}$ holds for all partitions. For two classes $i \neq j$, examine the following three-part partition:

$$G = \{G_1 = \{i\},\ G_2 = \{j\},\ G_3 = \{1,\ldots,K\} \setminus \{i,j\}\}.$$

We get that

$$\widehat{C}_{G_1G_2} = \frac{1}{1\cdot 1}\sum_{p\in G_1}\sum_{q\in G_2} C_{pq} = C_{ij},$$

and

$$\widehat{C}_{G_2G_1} = \frac{1}{1\cdot 1}\sum_{p\in G_2}\sum_{q\in G_1} C_{pq} = C_{ji}.$$

(Notice that $n_1 = n_2 = 1$.) By subgroup symmetry, $\widehat{C}_{G_1G_2} = \widehat{C}_{G_2G_1}$, hence $C_{ij} = C_{ji}$. As this equality holds for any pair of classes $i \neq j$, C is class symmetric. □

7.3 Empirically Assessing Subgroup Fairness

In §6, we empirically showed that Sy-FAR promotes fairness for individuals under evasion attacks; in particular, Sy-FAR increases symmetry as well as source- and target-class fair adversarial robustness. Complementing the theoretical result from §7.2, we now empirically demonstrate that optimizing for symmetry at the level of individual classes also improves symmetry, and as a by-product, source- and target-group fair adversarial robustness, for subgroups. Notably, these improvements are attained in a manner completely agnostic to the subgroup definitions, as Sy-FAR does not receive any subgroup information as part of its input.

Datasets, Models, and Groups. We evaluated all three face-recognition setups presented in §5. To study subgroup-level fairness, we focus on two demographic attributes provided in the PubFig dataset [23]: *gender* (male vs. female) and *ethnicity* (white vs. non-white). Class-to-subgroup assignments follow directly from the PubFig annotations.

Evaluation Metrics. To evaluate robustness and fairness at the level of subgroups, we report robust accuracy for each subgroup separately. The differences between subgroups' robust accuracies inform us about the degree of unfair source-group robust accuracy. Note that, because we consider two subgroups in each of the partitions, differences in unfair

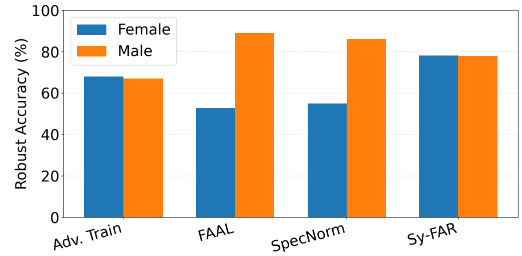


Figure 7: Subgroup robust accuracy by gender (male vs. female) on $\text{PubFig}_{\text{SIB}}$.

target-group robust accuracies are similar to the differences in source-group robust accuracy.

Evaluation Results. Fig. 7 reports subgroup robust accuracy for gender on the $\text{PubFig}_{\text{SIB}}$ setup between males and females. It can be observed that adversarial training exhibits a small gap between males and females (0.08), but the overall robust accuracy is low. FAAL and SpecNorm achieve higher overall robust accuracy than adversarial training; however, males attain significantly higher robust accuracy, with the gender gap reaching 0.36 for FAAL and 0.32 for SpecNorm. By contrast, Sy-FAR not only achieves the highest overall robust accuracy, but the gap between males and females is the lowest (0.04), with females receiving higher robust accuracy than males. Additional dataset and ethnicity-partition results are available in the extended version. Note that on all settings and between all subgroups, the gaps in benign accuracy between subgroups is small ($\leq$0.04). Overall, these results confirm that symmetry-based regularization not only encourages class-level fair adversarial robustness, but also propagates to subgroup-level fair adversarial robustness, substantially reducing demographic disparities under attacks.

8 Limitations and Future Work

While Sy-FAR achieves strong and consistent improvements in fair adversarial robustness, several limitations remain, opening promising directions for future research. First, our experiments were conducted under controlled and balanced conditions to isolate the core fairness–robustness challenge. We show that this problem is already difficult in balanced datasets and becomes considerably harder when bias is introduced, as seen in the $\text{PubFig}_{\text{SIB}}$ setup. Extending Sy-FAR to highly imbalanced or noisy datasets, where labels may be uncertain or ambiguous, is an important next step toward deployment in real-world systems.

Second, the theoretical proof that class-level symmetry implies subgroup-level fairness assumes a finite and correctly labeled class space. In open-world or continuous-class settings—where new or unseen classes may appear—these guarantees may not strictly hold. Future work should explore how to generalize the concept of symmetry to continuous or open-set domains, possibly through embedding-based or probabilistic formulations.

Third, similar to prior work, this work focuses on classification and vision tasks, further extending them to security-critical applications such as face recognition. We believe that the underlying idea of symmetry can also extend to other modalities (e.g., natural language or audio). However, these domains may require domain-specific definitions of symmetric behavior, as similarity definitions and misclassification patterns may differ fundamentally from those in vision.

Finally, extending Sy-FAR to generative and multimodal models will likely require new formalizations of symmetry and fairness that account for continuous output spaces. Investigating these directions constitutes an interesting avenue for future work.

9 Conclusion

This work introduced Sy-FAR, an approach for enhancing fair source- and target-class adversarial robustness through symmetry. Sy-FAR is based on the intuitive insight that class resemblance is usually a symmetric relationship, i.e., misclassifications of adversarial examples between classes should be symmetric. We extensively evaluate Sy-FAR, comparing it leading approaches [18, 62], on fairness- and security-critical face-recognition tasks against realistic eyeglass attacks as well as standard object-classification benchmarks commonly used in the literature. Our results, among others, demonstrate substantial improvements in fair source-class adversarial robustness while benign and robust accuracy are preserved or improved; identify unfair target-class robustness as a risk that Sy-FAR helps ameliorate; extend to group-based fairness where we show theoretically and empirically how symmetry advances fairness for arbitrary sub-groups in a computationally efficient manner. Beyond fairness, Sy-FAR offers strong practical advantages: It incurs only negligible overhead relative to standard adversarial training, in sharp contrast to established methods, and it exhibits lower variance across independent runs, highlighting its stability and reproducibility.

Taken together, our work shows that enforcing symmetry is an effective and principled way to mitigate biased error concentration while preserving robustness and scalability. We view Sy-FAR as a step toward fair and reliable adversarial training for safety-critical ML-based systems, and anticipate that the idea of symmetry regularization can be extended to broader domains and threat models in future work.

Ethical Considerations

Our work proposes Sy-FAR, a training method aimed at improving fairness and adversarial robustness of ML models, with a particular focus on face-recognition tasks. Below, we discuss the relevant stakeholders, the ethical implications of our research process and publication, and the measures we took to mitigate potential negative impacts.

Stakeholders. The stakeholders relevant to this work include:

- individuals whose images appear in publicly available datasets used in our experiments;
- end users of ML-based security systems that may adopt fairness- and robustness-enhancing methods;
- organizations and practitioners developing, deploying, or auditing such systems;
- members of society at large, particularly populations that may be disproportionately affected by face recognition or automated decision-making; and
- the research community, including researchers building on our methods or findings.

Research Process and Experimental Scope. We evaluate face- and object-recognition models using only existing, publicly available datasets. Importantly, no human subjects or deployed systems were involved in our experiments. We did not collect new biometric data, scrape online platforms, or attempt to identify or track individuals beyond the labels already present in the datasets. We used each public dataset strictly in accordance with its published usage guidelines and terms, without augmenting it with any additional personal information.

All experiments are conducted offline, and we apply only known adversarial perturbation techniques introduced in prior work. We do not introduce new attack vectors that could increase risk to existing systems. As a result, our work does not expose new vulnerabilities or create additional attack capabilities beyond what is already documented in the adversarial ML literature. We do not experiment with live systems and do not violate any terms of service.

Positive and Negative Potential Impacts. On the positive side, Sy-FAR has the potential to improve fairness and promote more equal treatment across demographic groups in ML-based security systems. By encouraging more symmetric misclassification behavior and reducing worst-class performance gaps, our work can help identify and mitigate biases that disproportionately harm certain individuals or groups. Our analysis also contributes to academic understanding of fairness–robustness trade-offs, supporting the design and evaluation of fairer and more trustworthy ML systems.

At the same time, we acknowledge that improvements in robustness and fairness may have dual-use implications. More effective face-recognition systems could be misused in surveillance or monitoring applications that threaten privacy or civil liberties, particularly when deployed without transparency or regulation. More generally, technical improvements alone do not address broader social or political concerns regarding the appropriateness of such systems. Our goal is not to promote or legitimize such deployments. Rather, we aim to highlight (un)fairness and robustness issues in existing models and to inform the research community about potential trade-offs and mitigation strategies. We do not endorse the deployment of face-recognition or surveillance systems, particularly in high-risk or sensitive contexts.

Mitigations and Ethical Choices. To mitigate potential negative impacts, we make several deliberate choices in how we design and conduct our study. We restricted experiments to offline evaluations on public datasets, avoided live systems and real users, and used only known adversarial methods. We did not scrape data, violate terms of service, or introduce new attack techniques. Throughout the paper, we present Sy-FAR primarily as a mitigation approach for these issues during training, rather than as a justification for deploying face-recognition systems. These limitations and caveats are clearly stated to ensure responsible interpretation and use of our results.

Deception and Researcher Wellbeing. There were no user studies, no deception, and no involvement of outside participants. Consequently, there was no risk of harm to participants. The research did not involve exposure to disturbing content or conditions affecting researcher wellbeing, and standard academic safeguards were sufficient.

Reflection. We believe it is important to publish this work despite potential dual-use risks, as failing to study and report fairness, robustness issues, and possible mitigations may allow biased or unevenly robust systems to persist unexamined. However, we emphasize that technical solutions alone cannot resolve all ethical challenges surrounding biometric technologies. We hope this work informs both technical research and broader discussions on governance, regulation, and responsible use.

Open Science

Our code repository[4] includes all necessary artifacts to fully reproduce the results of the paper. A detailed `README` file provides step-by-step instructions on how to set up the environment, run training, and reproduce both our baseline adversarial training and the proposed fairness-aware methods (FAAL [62], SpecNorm [18], and Sy-FAR). The instructions cover both clean training and adversarial training pipelines, as well as evaluation against benign and adversarial inputs (e.g., created with the eyeglass attack). To facilitate reproduction of results without training models from scratch, in addition to the source code, we provide pre-trained model checkpoints under reported settings trained with different robustness–fairness objectives. All artifacts are documented to allow end-to-end reproduction of experiments, ensuring transparency and verifiability of our contributions.

[4] https://zenodo.org/records/17901662 or https://github.com/haneenn24/Sy-FAR-Symmetry-based-Fair-Adversarial-Robustness.

Acknowledgments

This work has been supported in part by grant No. 2023641 from the United States-Israel Binational Science Foundation (BSF); by ISF grant no. 1807/23; by Len Blavatnik and the Blavatnik Family foundation; by a Maof prize for outstanding young scientists; by the Ministry of Energy, Israel (grant number 224-11-026); by the Ministry of Innovation, Science & Technology, Israel (grant number 1001702114); by the Stellar Development Foundation; and by a grant from the Tel Aviv University Center for AI and Data Science (TAD).

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *ECCV*, 2020.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *ECML/PKDD*, 2013.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *ICML*, 2012.

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[6] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv*, 2019.

[7] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and Zico Kolter. (Certified!!) Adversarial robustness for free! In *ICLR*, 2023.

[8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[9] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.

[11] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020.

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[13] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv*, 2018.

[14] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. In *CVPR*, 2018.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[16] Furkan Gursoy and Ioannis A Kakadiaris. Equal confusion fairness: Measuring group-based disparities in automated decision systems. In *ICDMW*, 2022.

[17] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy S Liang. Fairness without demographics in repeated loss minimization. In *ICML*, 2018.

[18] G Jin, S Wu, J Liu, T Huang, and R Mu. Enhancing robust fairness via confusional spectral regularization. In *ICLR*, 2025.

[19] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint*, 2018.

[20] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021.

[21] Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. The Marabou framework for verification and analysis of deep neural networks. In *CAV*, 2019.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[23] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLRW*, 2017.

[25] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

[26] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *CVPR*, 2020.

[27] Boqi Li and Weiwei Liu. Wat: Improve the worst-class robustness in adversarial training. In *AAAI*, 2023.

[28] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

[29] Keane Lucas, Samruddhi Pai, Weiran Lin, Lujo Bauer, Michael K. Reiter, and Mahmood Sharif. Adversarial training for raw-binary malware classifiers. In *USENIX Security*, 2023.

[30] Keane Lucas, Sharif, Mahmood, Bauer, Lujo, Michael K. Reiter, and Saurabh Shintre. Malware makeover: Breaking ML-based static analysis by modifying executable bytes. In *AsiaCCS*, 2021.

[31] Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. In *NeurIPS*, 2022.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[33] Shibin Mei, Chenglong Zhao, Bingbing Ni, and Shengchao Yuan. Towards interpreting and utilizing symmetry property in adversarial examples. In *AAAI*, 2023.

[34] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017.

[35] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[36] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *FAccT*, 2021.

[37] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *AsiaCCS*, 2017.

[38] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2015.

[39] Harrison Rosenberg, Brian Tang, Kassem Fawaz, and Somesh Jha. Fairness properties of face recognition and obfuscation systems. In *USENIX Security*, 2023.

[40] Guy N. Rothblum and Gal Yona. Probably approximately metric-fair learning. In *ICML*, 2018.

[41] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019.

[42] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In *NDSS*, 2019.

[43] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[44] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, 2016.

[45] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. A general framework for adversarial examples with objectives. *ACM TOPS*, 22(3):16:1–16:30, 2019.

[46] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *S&P*, 2017.

[47] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. In *EuroS&P*, 2021.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[49] Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. *arXiv*, 2024.

[50] Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fariness via balance adversarial training. In *AAAI*, 2023.

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[53] Liang Tong, Zhengzhang Chen, Jingchao Ni, Wei Cheng, Dongjin Song, Haifeng Chen, and Yevgeniy Vorobeychik. FaceSec: A fine-grained robustness evaluation framework for face recognition systems. In *CVPR*, 2021.

[54] Sahil Verma and Julia Rubin. Fairness definitions explained. In *IWSF*, 2018.

[55] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. FaceX-Zoo: A pytorch toolbox for face recognition. In *MM*, 2021.

[56] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. CFA: Class-wise calibrated fair adversarial training. In *CVPR*, 2023.

[57] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

[58] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*, 2021.

[59] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *NDSS*, 2018.

[60] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *NDSS*, 2016.

[61] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv*, 2020.

[62] Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *CVPR*, 2024.

[63] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv*, 2023.

A Evaluation on CIFAR-10 and CIFAR-100

While our primary focus is on face recognition, where fairness disparities pose immediate security and social risks, we also evaluate Sy-FAR on the widely used CIFAR-10 and CIFAR-100 benchmarks with the PreAct-ResNet18 model. These datasets and model are standard in the adversarial robustness and fairness literature [18, 62], enabling direct comparison with prior methods. Since they lack real-world fairness challenges such as demographic balance or sibling similarity, we present the detailed results here in the Appendix, keeping the main text centered on security-critical face recognition scenarios.

Results Summary. Across both CIFAR-10 and CIFAR-100, Sy-FAR achieves the strongest balance of robustness and fairness. Table 6 shows that Sy-FAR improves robust accuracy over adversarial training and FAAL, while matching or slightly exceeding SpecNorm, particularly on worst-class accuracy. In terms of asymmetry (Table 5), Sy-FAR consistently reduces directional bias, achieving the lowest asymmetry gap and symmetry loss on both benchmarks. Finally, target-class fairness results in Table 7 confirm that Sy-FAR distributes misclassifications more evenly. Together, these results demonstrate that even in standard object classification benchmarks, Sy-FAR delivers robust improvements across all fairness perspectives while preserving competitive robustness.

B Evaluation Under Targeted Eyeglass Attack

In addition to the untargeted setting, we also evaluate all methods under the *targeted eyeglass attack*, where adversarial perturbations are optimized to force misclassification specifically into a chosen (incorrect) target identity. We run the same experimental protocol across all three datasets (PubFig, $\text{PubFig}_{\text{SIB}}$, and $\text{PubFig}_{\text{ViT}}$). Consistent with the untargeted results, Sy-FAR achieves the strongest improvements across all metrics, showing the lowest asymmetry and fairness disparities while maintaining high robustness. These findings confirm that symmetry regularization is effective not only against arbitrary misclassifications but also in resisting targeted adversarial attempts that deliberately push predictions toward specific identities. Complete tables and figures are available in the extended version.

C Stability

Figs. 8–10 demonstrate Sy-FAR consistently optimize adversarial robustness and fairness across different runs.

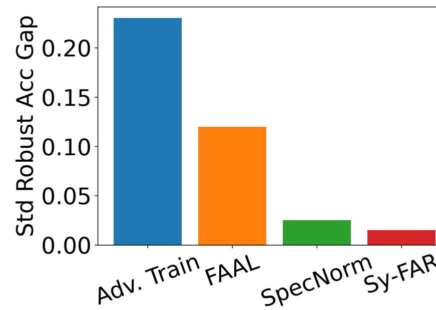


Figure 8: Standard deviation in robust accuracy across ten runs in PubFig under untargeted glass attack.

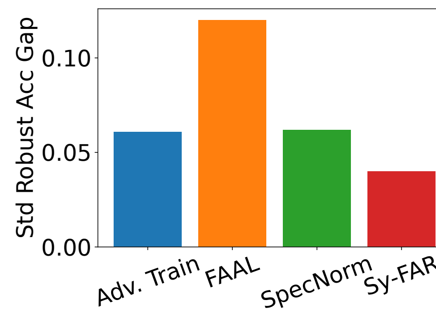


Figure 9: Standard deviation in worst-class accuracy across ten runs in PubFig under untargeted glass attack.

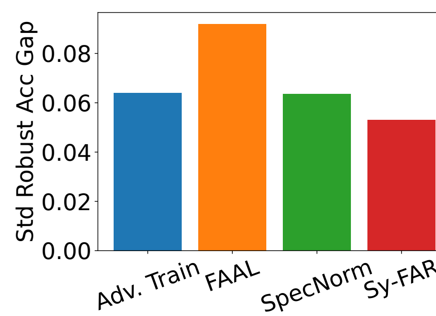


Figure 10: Standard deviation in accuracy gap (min-max) across ten runs in PubFig under untargeted glass attack.

D Scalability and Run-Time Analysis

To evaluate the computational scalability of Sy-FAR, we analyze the complexity and runtime contribution of the proposed symmetry regularization term across datasets with different numbers of classes. During training, nearly all computation time is spent on the forward and backward propagation through the deep network and on adversarial example generation–these steps involve billions of floating-point operations. In contrast, the symmetry penalty performs only lightweight arithmetic operations over class-pair statistics derived from the model's existing softmax outputs. The penalty does not require additional forward or backward passes, nor does it alter the adversarial optimization process. Although the number of class pairs grows quadratically with the number of classes ($C(C-1)/2$), each symmetry computation is an extremely small GPU tensor operation over already available data. Empirically, we find that:

- For 10 classes (e.g., PubFig), the symmetry term accounts for $<$0.001% of the total training time per batch.
- For 100 classes (e.g., CIFAR-100), the cost increases slightly to 0.01–0.05%.
- Even for 1,000 classes (e.g., ImageNet-scale settings), the overhead remains $<$0.5%.

Method	CIFAR-10		CIFAR-100	
	Asym. Gap ↓	Sym. Loss ↓	Asym. Gap ↓	Sym. Loss ↓
Adv. Train	11.2	0.81	17.6	53.46
FAAL	9.8	1.17	15.9	38.80
SpecNorm	8.5	0.71	14.4	39.84
Sy-FAR	**6.7**	**0.63**	**12.2**	**37.66**

Table 5: Symmetry on CIFAR-10/100 (PRN18): The impact of different training approaches on symmetry, as captured by the Max Asymmetry Gap and the Symmetry Loss. Models are adversarially trained with PGD-ℓ_∞ and evaluated with AutoAttack [12]. Results are averaged across ten runs. We report the Max Asymmetry Gap (Gap↓) and the Symmetry Loss (Loss↓).

Method	CIFAR-10				CIFAR-100			
	Benign ↑	Robust ↑	Min ↑	Gap ↓	Benign ↑	Robust ↑	Min ↑	Gap ↓
Adv. Train	82.07	47.12	12.90	34.22	54.70	22.15	1.00	21.15
FAAL	82.38	49.80	33.50	16.30	56.82	21.96	3.00	18.96
SpecNorm	**83.51**	49.94	34.94	15.00	57.36	25.17	**4.00**	21.17
Sy-FAR	82.02	**52.60**	**35.31**	**17.29**	**57.60**	**25.77**	**4.00**	**21.77**

Table 6: Accuracy and source-class fairness on CIFAR-10/100 (PRN18): The impact of different training approaches on accuracy and fairness across true classes. Models are adversarially trained with PGD-ℓ_∞ and evaluated with AutoAttack [12]. Results are averaged across ten runs. We report benign accuracy (Benign↑), robust accuracy (Robust↑), worst-class accuracy (Min↑), and the class-level robust accuracy gap (Gap↓).

In other words, the computation remains negligible compared to the dominant convolutional and adversarial training operations. As the number of classes increases, the total epoch time rises modestly due to the linear growth of the final fully connected (FC) layer and softmax computations, while the symmetry term—despite its theoretical $O(C^2)$ complexity, still contributes an almost imperceptible fraction of the total runtime.

E Extended Attack Evaluation: Face Masks

To evaluate Sy-FAR under more realistic physical threats, we introduce a *face-mask attack* inspired by FACESEC [53]. Unlike eyeglass attacks, which apply *pixel-level*, small, localized, high-frequency perturbations around the eyes [44], the face-mask attack operates at the *grid level*, occluding the lower face (nose and mouth) and inducing a large, low-frequency, spatially coherent perturbation that removes critical global shape and landmark information. Despite this stronger and fundamentally different threat model, Sy-FAR consistently achieves the best *symmetry loss*, *source- and target-class robustness*, and *fairness* (Tables 8–10), demonstrating that its symmetry-based regularization remains effective even under severe facial occlusion.

Method	CIFAR-10			CIFAR-100		
	MinTgt ↑	**MaxTgt ↓**	**Std ↓**	**MinTgt ↑**	**MaxTgt ↓**	**Std ↓**
Adv. Train	0.03	0.29	0.085	0.004	0.15	0.041
FAAL	0.04	0.24	0.071	0.006	0.13	0.036
SpecNorm	0.05	0.22	0.066	0.007	0.11	0.033
Sy-FAR	**0.07**	**0.18**	**0.052**	**0.010**	**0.09**	**0.028**

Table 7: Target-class fairness on CIFAR-10/100 (PRN18): The impact of different training approaches on the distribution of misclassifications across predicted classes. Models are adversarially trained with PGD-ℓ_∞ and evaluated with AutoAttack [12]. Results are averaged across ten runs. We report the minimum normalized misclassification into a target class (MinTgt↑), the maximum normalized misclassification into a target class (MaxTgt↓), and the standard deviation across classes (Std↓).

Method	PubFig		PubFig$_{\text{SIB}}$		PubFig$_{\text{ViT}}$	
	Asym. Gap ↓	**Sym. Loss ↓**	**Asym. Gap ↓**	**Sym. Loss ↓**	**Asym. Gap ↓**	**Sym. Loss ↓**
Adv. Train	0.6183	2.0856	0.4048	1.1772	0.3602	1.4638
FAAL	0.2117	0.7686	0.3333	1.0288	0.3962	1.8115
SpecNorm	0.2475	0.8215	0.3418	1.0426	0.3462	1.5138
Sy-FAR	**0.1748**	**0.4946**	**0.3117**	**0.9276**	**0.2974**	**1.0920**

Table 8: Impact of different training methods on symmetry under the untargeted *face-mask attack*. Reported are the Max Asymmetry Gap and Symmetry Loss on PubFig, PubFig$_{\text{SIB}}$, and PubFig$_{\text{ViT}}$ using VGG16 and ViT backbones, averaged over ten runs.

Method	PubFig				PubFig$_{\text{SIB}}$				PubFig$_{\text{ViT}}$			
	Clean ↑	**Robust ↑**	**Min ↑**	**Gap ↓**	**Clean ↑**	**Robust ↑**	**Min ↑**	**Gap ↓**	**Clean ↑**	**Robust ↑**	**Min ↑**	**Gap ↓**
Adv. Train	97.05	56.35	1.48	88.01	95.60	64.18	25.33	60.45	82.26	54.46	17.65	65.69
FAAL	98.60	79.36	57.78	39.44	93.04	68.54	26.67	58.33	81.06	57.97	20.82	69.82
SpecNorm	98.10	77.02	55.85	38.42	95.75	72.04	28.33	60.84	82.19	56.61	26.16	63.63
Sy-FAR	**99.09**	**84.76**	**72.88**	**27.12**	**95.76**	**77.05**	**44.12**	**55.88**	**83.87**	**65.06**	**33.89**	**58.96**

Table 9: Performance comparison across training methods under the untargeted *face-mask attack*. Reported are Clean, Robust, and Minimum (Worst-Class) Accuracies, and Accuracy Gap (Min–Max diagonal difference) on PubFig, PubFig$_{\text{SIB}}$, and PubFig$_{\text{ViT}}$, averaged over ten runs.

Method	PubFig			PubFig$_{\text{SIB}}$			PubFig$_{\text{ViT}}$		
	MinTgt ↑	**MaxTgt ↓**	**Std ↓**	**MinTgt ↑**	**MaxTgt ↓**	**Std ↓**	**MinTgt ↑**	**MaxTgt ↓**	**Std ↓**
Adv. Train	0.0049	0.2736	0.0831	0.0000	0.2485	0.0882	0.0088	0.3719	0.0900
FAAL	0.0038	0.2913	0.0904	0.0000	0.3432	0.0950	0.0016	0.3683	0.0872
SpecNorm	0.0078	0.2435	0.0801	0.0044	0.2595	0.0899	0.0055	0.2461	0.0933
Sy-FAR	**0.0105**	**0.1991**	**0.0623**	**0.0062**	**0.2146**	**0.0669**	**0.0115**	**0.1907**	**0.0582**

Table 10: Target-class fairness under the untargeted *face-mask attack*. Reported are the minimum and maximum misclassification shares (MinTgt and MaxTgt) and their standard deviation across classes, for PubFig, PubFig$_{\text{SIB}}$, and PubFig$_{\text{ViT}}$ setups, averaged over ten runs.