Fairness Without Labels: Pseudo-Balancing for Bias Mitigation in Face Gender Classification

Haohua Dong[1,2] Ana Manzano Rodríguez[1,3] Camille Guinaudeau[4] Shin'ichi Satoh[1]
[1]National Institute of Informatics, Japan
[2]INESC-ID, Instituto Superior Técnico, University of Lisbon, Portugal
[3]University of Amsterdam, The Netherlands
[4]LIMSI, CNRS / Université Paris-Saclay, France
haohua@tecnico.ulisboa.pt

Abstract

Face gender classification models often reflect and amplify demographic biases present in their training data, leading to uneven performance across gender and racial subgroups. We introduce pseudo-balancing, a simple and effective strategy for mitigating such biases in semi-supervised learning. Our method enforces demographic balance during pseudo-label selection, using only unlabeled images from a race-balanced dataset without requiring access to ground-truth annotations.

We evaluate pseudo-balancing under two conditions: (1) fine-tuning a biased gender classifier using unlabeled images from the FairFace dataset, and (2) stress-testing the method with intentionally imbalanced training data to simulate controlled bias scenarios. In both cases, models are evaluated on the All-Age-Faces (AAF) benchmark, which contains a predominantly East Asian population. Our results show that pseudo-balancing consistently improves fairness while preserving or enhancing accuracy. The method achieves 79.81% overall accuracy—a 6.53% improvement over the baseline—and reduces the gender accuracy gap by 44.17%. In the East Asian subgroup, where baseline disparities exceeded 49%, the gap is narrowed to just 5.01%. These findings suggest that even in the absence of label supervision, access to a demographically balanced or moderately skewed unlabeled dataset can serve as a powerful resource for debiasing existing computer vision models.

1. Introduction

Automated gender classification systems are widely deployed in modern computer vision applications, yet their performance disparities across demographic groups remain a persistent challenge. While these systems achieve high accuracy on Western populations, their reliability degrades significantly when applied to non-Western population such as Black or East Asian faces. This performance disparity often stems from two compounding factors: biased training datasets that under-represent non-Western populations [18, 27], and algorithmic approaches that amplify these biases during learning [1, 22].

Balanced training datasets like FairFace (FF) [20] demonstrate that careful curation can improve fairness, but recollecting such data is often prohibitively expensive. Semi-supervised learning offers a cost-effective alternative, but standard pseudo-labeling approaches like FixMatch [30] often reinforce existing biases rather than mitigate them, as the model is trained on pseudo-labels it generates itself—thus amplifying its initial bias.

In this work, we explore how unlabeled images from a race-balanced dataset like FairFace can be strategically used to reduce bias—without relying on FairFace's ground truth labels. We introduce pseudo-balancing, a lightweight yet effective enhancement to self-training pipelines, which dynamically adjusts pseudo-label selection to enforce gender balance. Our method (1) modifies only the sample selection step and prevents majority-class domination during self-training by manually re-balancing pseudo-labels, (2) preserves model accuracy through confidence-based sample selection, and (3) avoids the need for adversarial training or distributional assumptions common in domain adaptation techniques [11]. This makes our approach highly compatible with existing semi-supervised methods such as FixMatch [30] and FlexMatch [35], which leverage confident pseudo-labeling and adaptive thresholding strategies to improve learning from unlabeled data. We structure our study around two experimental scenarios designed to evaluate both the benefits and the limits of this approach:

- **Scenario 1**: We start with a standard case where a classifier is trained on a publicly available but demo-

graphically skewed dataset Kaggle Gender Classification Dataset [19]. We then apply pseudo-balancing using unlabeled FF images to rebalance the training distribution. Evaluation is conducted on the All-Age-Faces (AAF) dataset [6], which contains predominantly East Asian images. This scenario tests whether incorporating balanced unlabeled data during training can improve fairness over a biased baseline;

- **Scenario 2**: We simulate controlled bias conditions by training the initial classifier on data from only varying demographic group imbalances (e.g., gender and race subsets from FF). We then apply pseudo-balancing using unlabeled FF images. This more challenging setup evaluates the robustness of our method when the labeled training data lacks diversity entirely.

In both scenarios, we evaluate on the AAF benchmark using overall accuracy and selection rate (accuracy gap across genders). Notably, even without FairFace labels, leveraging its balanced composition reduces gender disparities. In Scenario 1, our method achieves a more uniform accuracy than a Kaggle-trained baseline. In Scenario 2, pseudo-balancing remains effective despite initial bias in training data, demonstrating its robustness.

Our main contributions are as follows:

- We propose pseudo-balancing, a simple bias mitigation method that enforces pseudo-gender balance during self-training without requiring labeled source data;
- We introduce a framework to evaluate compounding dataset and model biases via controlled experiments using FairFace and All-Age-Faces;
- We empirically show that pseudo-balancing reduces gender disparities while maintaining or improving accuracy on the AAF benchmark.

Our method achieves 79.81% accuracy—a 6.53% gain over the baseline—while cutting the gender gap by 44.17%. Pseudo-balancing particularly benefits underrepresented groups in the training data, offering a fairer gender classification without costly labeled data collection.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 details our approach. Section 4 presents experiments. Section 5 discusses implications, with limitations and conclusions in Sections 6 and 7.

2. Related Work

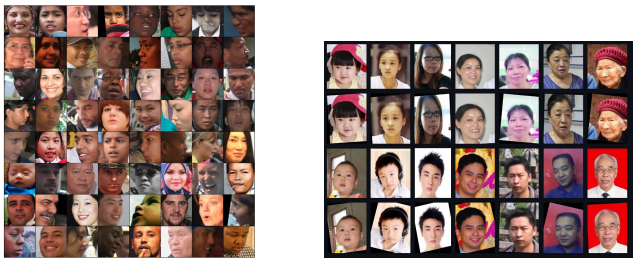
2.1. Gender Classification Algorithms: Its Applications and Bias

Automated gender classification systems is widely used across multiple domains, including video surveillance [7], demographic analysis [13], targeted advertising [29], and human-computer interaction [21]. However, extensive research has revealed significant performance disparities in commercial systems, particularly for non-Western populations such as East Asians or Black people. For example, in a 2018 study, a dermatologist approved Fitzpatrick skin type classification system was shown to be have error rates as high as 34.7% for dark-skinned females compared to light-skinned males [1].

These technical shortcomings have real-world consequences, especially in high-stakes applications. Law enforcement systems using biased facial recognition have been shown to subject African-American individuals to searches at disproportionate rates [12], while commercial systems frequently reinforce harmful stereotypes by associating women with domestic activities [36] and being misgendered by automatic gender recognition systems (AGR) could lead to a greater insult due to the perceived objectivity of computer systems [23]. The root causes of these biases can often be traced to unrepresentative training datasets.

For example, Color FERET is a 8.5GB database benchmark for facial recognition under controlled environments [27]. Color FERET contains only 10% as many African male identities and 17% as many African female identities compared to their white counterparts. Labeled Faces in The Wild (LFW), a database composed of celebrity facial images designed for facial verification [18], is estimated to be composed of 77.5% male and 83.5% white faces [13]. Recent facial recognition systems reported a 97.35% accuracy on the LFW database, however it is not clear how the algorithm is performing on the underrepresented subgroups inside the benchmark [1, 31].

Recent initiatives have attempted to address these representation gaps through improved dataset design. The Diversity in Faces (DiF) benchmark [26] provides comprehensive anthropometric measurements across one million images, while FairFace (FF) [20] explicitly balances representation across seven racial groups (White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino) with near-equal gender distribution. These datasets demonstrate that balanced training data can improve model fairness. However, even carefully curated datasets struggle with intersectional biases (e.g., dark-skinned women) [22] and require complementary algorithmic solutions.



a). FairFace (FF) dataset b). All-Ages-Face (AAF) dataset

Figure 1. FairFace (FF) and All-Ages-Face (AAF) dataset examples.

2.2. Traditional Approaches for Fairness Improvement

While our previous discussion highlighted how skewed training data distributions propagate bias, the machine learning architectures themselves can introduce and amplify these disparities. Traditional approaches to improving fairness involve fine-tuning models with ground truth labels, which, while effective, require extensive manual annotation [17]. Another strategy involves retraining models on more balanced datasets, though this is resource-intensive and may still be affected by the biases in the new dataset [11, 22]. This motivates our investigation of alternative solutions, particularly for applications like surveillance and targeted advertising where annotation budgets are limited.

To address these challenges, semi-supervised and unsupervised learning techniques have been proposed. Self-supervised pre-training methods like MoCo [5, 15] and SimCLR [4] offer another pathway by learning robust representations from uncurated data. However, such approaches still risk encoding societal biases present in the pre-training corpus [36]. Domain-Adversarial Neural Networks (DANN) [11] leverage adversarial learning to mitigate domain shifts and reduce bias, making them promising candidates for fairer gender classification, but require careful tuning to handle demographic imbalances. Pseudo-labeling techniques have emerged as effective approaches for fine-tuning pre-trained models, generating synthetic labels from unlabeled data to improve feature alignment across demographic groups. Recent advances in curriculum-based methods like FlexMatch [35] demonstrate how adaptive thresholding - dynamically adjusting confidence requirements based on model performance - can outperform fixed-threshold approaches like FixMatch [30] in biased learning scenarios. Building on these foundations, contemporary research highlights the importance of balanced pseudo-label distributions [16], suggesting that careful management of demographic representation during training can further enhance model fairness.

This critical gap in existing approaches motivates our comprehensive study of bias mitigation techniques using FairFace's carefully balanced demographic structure as both an experimental framework and source domain. FairFace's controlled composition (equal representation across seven racial groups and near-balanced gender distribution) provides an interesting testbed for evaluating algorithmic fairness interventions. Our experimental design isolates and analyzes specific bias factors - particularly the intersection of racial and gender biases that affect East Asian populations in the All-Ages-Faces dataset [6].

3. Methods

3.1. Model and Datasets

We employ a readily available ResNet18-based convolutional neural network [14] pre-trained on the Kaggle gender classification dataset as our baseline model[1]. This choice is motivated by its demonstrated effectiveness when fine-tuned on FairFace and mixed datasets, achieving a significant reduction in the gender accuracy gap for Japanese data [28], despite its known limitations in performance on minority subgroups (See Table 1). The model is fine-tuned using FairFace [20], All-Ages-Face (AAF) datasets [6], and Kaggle Gender Classification Dataset [19]. FairFace contains 108k training images balanced across seven racial groups (with a rate of 53% male, 47% female), AAF comprises 13.3k images primarily featuring individuals of Asian descent (45% male, 55% female), while Kaggle contains 47k training images (49.5% female, 50.5% male).

We measure bias using the **Selection Rate (SR)**, defined as the ratio of the minimum to maximum gender accuracy:

$$\text{SR} = \frac{\text{min accuracy}}{\text{max accuracy}}.$$

Following prior work [9, 33], SR $\geq 80\%$ indicates mitigated bias. Bias reduction is considered effective when SR increases along with overall model accuracy.

3.2. Supervised Learning

We conducted fine-tuning experiments in a supervised learning setting [28], using a gender-balanced test set of 2,504 samples from the All-Ages-Face (AAF) dataset to ensure fair evaluation across male and female subgroups. Fine-tuning the baseline model on the FairFace dataset improved overall accuracy on AAF but left a significant gender accuracy gap of 26%—a reduction from the original 49%. The most equitable performance was achieved by sequentially fine-tuning on FairFace and then a mixed dataset (AAF and Kaggle Gender Classification), reducing the gender disparity to under 1%, as shown in Table 1.

3.3. Semi-Supervised Learning

Since ground-truth labels are often unavailable or expensive to obtain, we adopt pseudo-labeling to improve model performance using unlabeled data. A model initially trained on labeled data is used to generate high-confidence predictions (pseudo-labels), which are then treated as ground truth for further training. We apply this approach using the FairFace dataset and generate pseudo-labels via the FixMatch framework [30], testing confidence thresholds $\epsilon \in \{0.1,\ 0.3,\ 0.6,\ 0.9\}$.

[1] https://github.com/ndb796/Face-Gender-Classification-PyTorch

Model	Accuracy	Selection Rate
Baseline	73.28% **97.94%** / *48.61%*	49.63%
Fine-tuned FF	87.54% **95.36%** / *79.71%*	83.59%
Fine-tuned Mixed	95.17% **93.94%** / *96.39%*	97.46%
Fine-tuned FF + Mixed	96.22% **96.20%** / *96.24%*	99.95%

Table 1. Impact of fine-tuning with ground-truth labels of FairFace and mixed datasets. The first line corresponds to the general accuracy. Accuracy values for male and female are in **bold** and in *italic*, respectively.

Model	Accuracy	Selection Rate
Baseline	73.28% **97.94%** / *48.61%*	49.63%
$\epsilon = 0.1$	77.26% **97.39%** / *57.13%*	58.66%
$\epsilon = 0.3$	79.04% **97.62%** / *60.46%*	61.93%
$\epsilon = 0.6$	78.80% **96.32%** / *61.29%*	63.58%
$\epsilon = 0.9$	79.42% **93.46%** / *65.37%*	69.94%

Table 2. Performance of one iteration of pseudo-labeling with no pseudo-balancing (no PB) under varying confidence thresholds on the AAF dataset.

3.3.1. Pseudo-Balancing

We define **pseudo-balancing** as the enforcement of equal class representation during pseudo-label selection in self-training. Unlike standard pseudo-labeling, which accepts all high-confidence predictions from an initial model, pseudo-balancing selectively subsamples these predictions to ensure demographic parity—e.g., an equal number of pseudo-male and pseudo-female samples. This prevents the model from reinforcing majority-class biases during retraining. In our implementation, we apply pseudo-balancing after each pseudo-labeling step by partitioning the confident predictions into gender groups and sampling an equal number from each before feeding them into the retraining phase. This strategy complements confidence-based selection by adding a fairness constraint to the sample distribution, ensuring balanced learning signals during iterative self-training.

3.3.2. Iterative Self-Training

One iteration of pseudo-labeling using a single-confidence threshold improves the performance of the baseline model (Table 2). We extend single-step pseudo-labeling into an iterative self-training framework, where the model alternates between generating pseudo-labels for unlabeled data and retraining on these labels (Figure 2). This approach has shown strong performance in both classification and segmentation tasks [3, 10, 34, 37, 38].

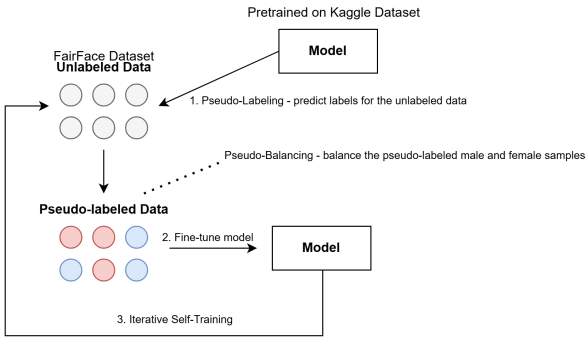


Figure 2. Iterative self-training using pseudo-labeled data.

We implement two variants of this strategy. FixMatch [30] combines weak and strong augmentations with confidence-based pseudo-label selection. Weak augmentations generate reliable pseudo-labels, while strong augmentations provide more robust training signals. After each step, we enforce demographic parity via **pseudo-balancing**, subsampling the pseudo-labeled set to maintain equal representation of male and female faces. FlexMatch [35] builds on FixMatch by introducing a curriculum-based thresholding scheme that adapts confidence thresholds per class based on learning progress. This dynamic adjustment may improve learning for underrepresented classes, particularly female samples, and reduces bias amplification common in imbalanced datasets.

4. Experiments

To assess how pseudo-labeling and pseudo-balancing strategies behave under skewed demographic conditions, we conducted a set of controlled experiments using the FairFace dataset. The experiments are structured around two key scenarios: 1) mitigating bias when using a balanced unlabeled dataset with a demographically biased initial model; and 2) analyzing the effects of demographic skew within the unlabeled training data.

We split the FF dataset into 86,744 samples for training (40,758 female and 45,986 males) and 1,250 East-Asian samples (777 males, 773 females) for validation. For evaluation, a gender-balanced test set of 2,504 samples was curated from AAF. To investigate bias patterns, we created specialized subsets of FF including a 43,372-sample gender-balanced variant (GB-FF), race-specific collections (12,287 East-Asian, 12,233 Black, and 6,141 East-

Asian samples), and gender-skewed versions (41,803 samples with a 80% female distribution and 44,941 samples with 80% male distribution).

4.1. Scenario 1: Bias Mitigation with Balanced Unlabeled Data

In the first scenario, we simulate a realistic setting where an off-the-shelf model exhibits inherent biases due to its training history. We use the pre-trained Kaggle CNN PyTorch model, which is known to perform poorly on minority subgroups, as our initial model. For fine-tuning, we employ unlabeled data from FairFace—specifically its race-balanced (FF) and gender-balanced (GB-FF) subsets. Although FairFace includes gender labels, these are used only for evaluation, not training.

Evaluation is conducted on the AAF benchmark, which predominantly features East Asian individuals and allows for detailed demographic bias analysis. We compare the baseline Kaggle model against variants fine-tuned on FF and GB-FF using FixMatch and FlexMatch, with and without pseudo-balancing (PB).

4.2. Scenario 2: Learning with Biased Unlabeled Data

In this scenario, we investigate how pseudo-balancing performs when the unlabeled training data itself is biased. We design the following controlled bias conditions:

1. *Gender Bias*: Datasets with 80%-20% or 50%-50% male-female distributions.
2. *Severe Racial Bias*: Datasets containing only one race (e.g., East Asian or Black).
3. *Gender and Racial Bias*: An East Asian female FairFace subset to simulate overlapping gender and racial bias, reflecting AAF test set characteristics.

4.3. Results

FixMatch was evaluated using two confidence thresholds (ϵ = 0.6 and ϵ = 0.9) to analyze the impact of pseudo-labeling confidence on model performance. FlexMatch used a nonlinear threshold adjustment based on class-specific learning status, with a base threshold of 0.95, in which the adjustment function is given by $\frac{x}{2-x}$. Training was conducted for 10 epochs per iteration using SGD optimization (learning rate = 0.001, momentum = 0.9) and a batch size of 16, with early stopping applied after 2 epochs of no improvement. These parameters are based on previous ablation studies of FlexMatch performance [35]. The best models can be seen in Table 3 and Table 4, highlighted in yellow.

Pseudo-balancing (PB) improves performance in most cases, particularly for datasets that are initially balanced or diverse like FF and GB-FF (Fig. 3). For instance, FairFace with PB achieves 83.36% accuracy with FlexMatch, compared to 79.18% without PB; FixMatch (ϵ = 0.9, PB) achieves a accuracy of 79.81% and 75.73% (ϵ = 0.9, No PB). On severely biased datasets like Severe East-Asian Female and Severe Male datasets, pseudo-balancing may de-

Model	ϵ	PB	Accuracy	SR
Baseline	-	-	73.28% **97.94** / *48.61*	49.63%
FairFace	0.6	yes	80.05% **91.12** / *69.98*	76.80%
FairFace	0.6	no	76.92% **96.20** / *57.65*	59.93%
FairFace	0.9	yes	79.81% **82.37** / *77.26*	93.80%
FairFace	0.9	no	75.73% **97.35** / *54.12*	55.59%
Gender-Balanced FF	0.9	yes	78.65% **95.96** / *61.33*	63.91%
Severe East-Asian Female FF	0.9	yes	71.63% **90.33** / *44.10*	48.82%
Severe East-Asian Female FF	0.9	no	81.30% **87.99** / *74.60*	84.78%
Severe Black FF	0.9	yes	74.41% **96.72** / *52.10*	53.87%
Severe Black FF	0.9	no	71.35% **99.05** / *43.66*	44.08%
Severe Male FF	0.6	yes	78.11 % **86.33** / *69.89*	80.96%
Severe East-Asian FF	0.6	yes	79.33% **90.65** / *69.22*	76.36%
Severe East-Asian FF	0.9	yes	80.59% **95.56** / *65.61*	68.66%
Severe Female FF	0.9	no	78.86% **95.84** / *61.89*	64.58%

Table 3. Performance of the fine-tuned model with the pseudo-labeled data from FairFace with different confidence scores of FixMatch with and without pseudo-balancing, evaluated on the AAF dataset. Initial model is the pre-trained Kaggle CNN PyTorch model. Accuracy values for male and female are in **bold** and in *italic*, respectively.

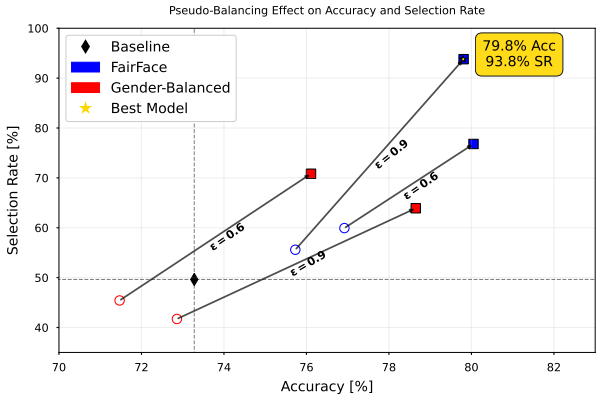


Figure 3. Scenario 1. Accuracy vs. selection rate for FF variants using FixMatch. Squares denote pseudo-balanced (PB) results and circles non-PB counterparts. Arrows indicate performance shifts. The best model achieves 82.4% accuracy at 93.8% selection rate (highlighted).

Model	PB	Accuracy	SR
Baseline	-	73.28% **97.94** / *48.61*	49.63%
FairFace	yes	83.36% **89.62** / *77.10*	86.03%
FairFace	no	79.18% **67.75** / *90.61*	74.77%
Gender-Balanced FF	yes	67.49% **94.93** / *40.05*	42.19%
Severe East-Asian Female FF	yes	71.87% **83.44** / *60.30*	72.27%
Severe East-Asian Female FF	no	56.04% **76.07** / *36.01*	47.37%
Severe Male FF	yes	62.92% **52.61** / *73.21*	71.86%
Severe Male FF	no	54.16% **19.02** / *89.30*	21.28%

Table 4. Performance of the fine-tuned model with the pseudo-labeled data from FairFace with FlexMatch, evaluated on the AAF dataset. Initial model is the pre-trained Kaggle CNN PyTorch model. Accuracy values for male and female are in **bold** and in *italic*, respectively.

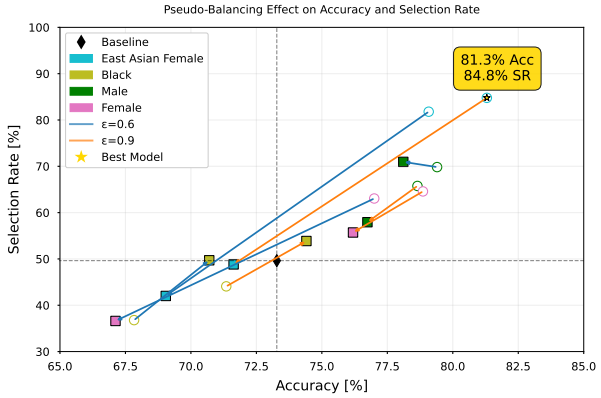


Figure 4. Scenario 2. Bias-specific performance analysis showing East Asian Female, Black and Male subsets using FixMatch. Squares denote pseudo-balanced results and circles show non-PB results ($\epsilon = 0.6$: blue arrows; $\epsilon = 0.9$: orange arrows). The model trained on the East Asian Female subset achieves 81.3% accuracy at 84.8% selection rate (highlighted).

grade model performance (Fig. 4). For example, on the Severe East-Asian Female dataset, FixMatch (ϵ = 0.9, PB) achieves 81.30% accuracy, compared to 71.63% with PB. This suggests that pseudo-balancing is most effective when applied to training datasets that are initially diverse or balanced, as it may help maintain class equilibrium during training and prevent the model from overfitting to dominant classes. However, in cases where the training dataset exhibits severe biases, pseudo-balancing can inadvertently amplify existing imbalances, leading to accelerated bias. This occurs because pseudo-balancing may reinforce the

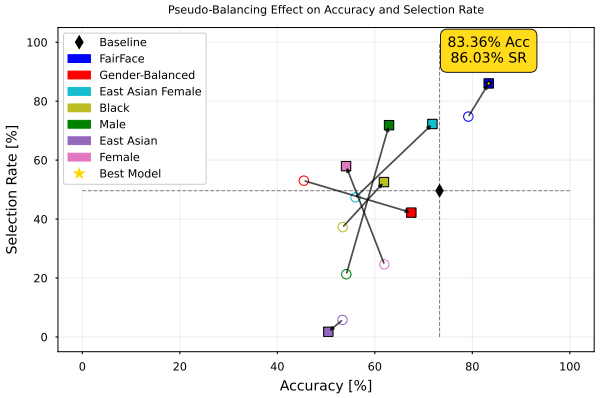


Figure 5. Scenarios 1 and 2. FlexMatch adaptation results across all bias conditions. Squares denote PB results and circles non-PB results. The best model achieves 83.36% accuracy at 86.03% selection rate using FairFace dataset with PB.

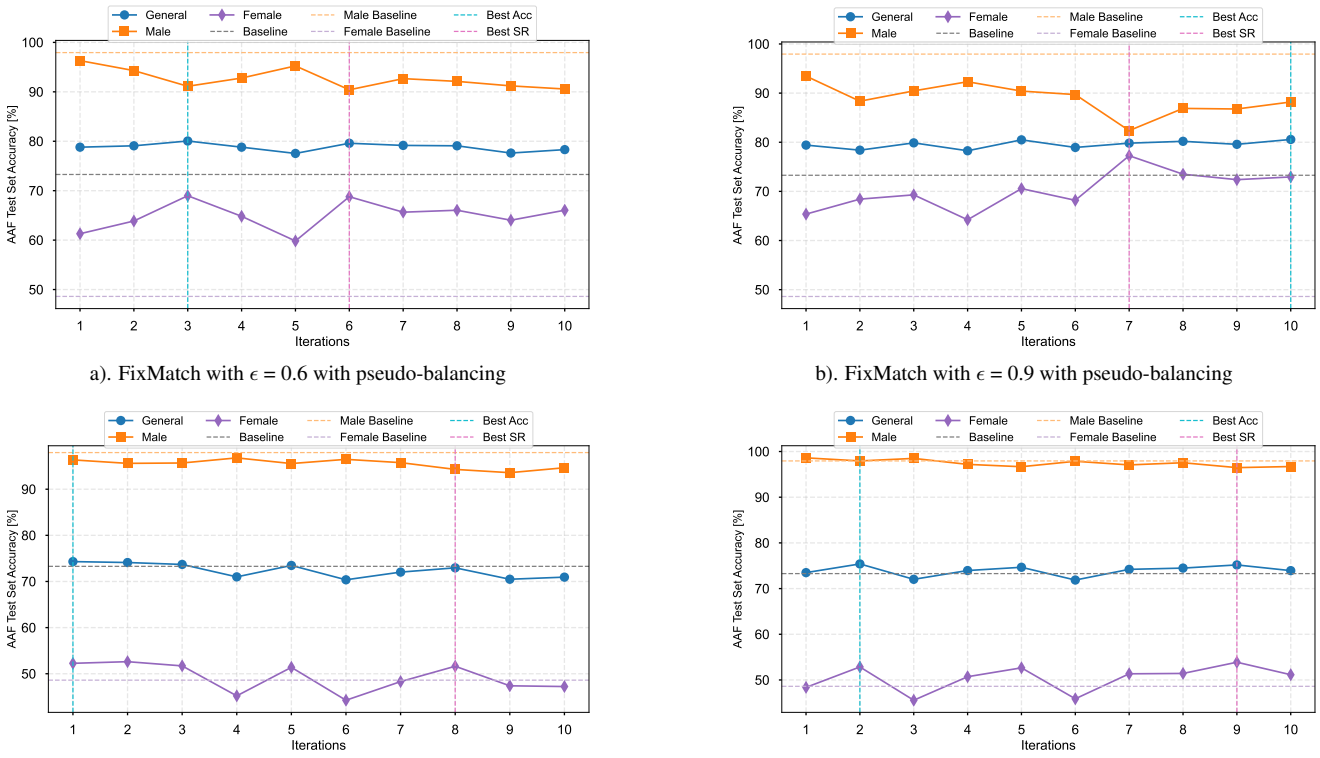
skewed distribution of pseudo-labels, further disadvantaging underrepresented classes. This observation is supported by [24], which emphasizes the importance of maintenance of dataset composition in pseudo-labeling strategies.

FlexMatch achieves its best performance of 83.36% accuracy (with a SR of 86.03%) when trained with FF and pseudo-balancing. However, its effectiveness diminishes significantly when dealing with severely biased datasets, where it underperforms compared to both FixMatch and the baseline model. This suggests that adaptive thresholding methods, while effective in balanced or moderately biased scenarios, are insufficient to address extreme gender or racial biases without additional balancing mechanisms such as pseudo-balancing.

Furthermore, the presence of severe bias in the training data significantly impacts model performance on the AAF dataset. When fine-tuning the model with FixMatch using severely biased datasets such as the East-Asian FairFace or East-Asian Female FairFace, the model achieves accuracies of 79.33% (with a SR of 76.36%) and 81.30% (with a SR of 84.78%), respectively. These results demonstrate that models fine-tuned on datasets with specific racial or gender biases can perform on par with those fine-tuned on the entire FF or FF-GB dataset, provided the training data aligns with the target domain's demographic characteristics. Moreover, the Severe Male FF dataset performs well under FixMatch ($\epsilon = 0.6$, PB). In this setting, pseudo-balancing effectively compensates for extreme gender bias by redistributing pseudo-labels, while the reduced threshold helps preserve gender diversity and enhances recall for the underrepresented female class.

4.4. Self-Training Performance Analysis

The iterative self-training results of FixMatch and FlexMatch highlight critical differences in bias mitigation effi-



a). FixMatch with $\epsilon = 0.6$ with pseudo-balancing

b). FixMatch with $\epsilon = 0.9$ with pseudo-balancing

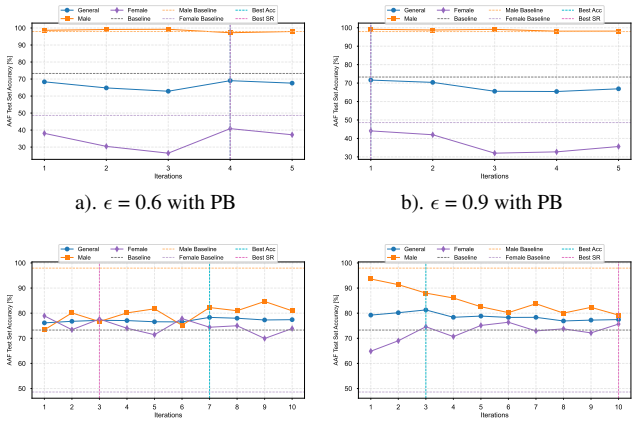
c). FixMatch with $\epsilon = 0.6$ without pseudo-balancing

d). FixMatch with $\epsilon = 0.9$ without pseudo-balancing

Figure 6. Scenario 1. Self-training per-iteration evolution of fine-tuned models using FixMatch, trained with unlabeled FF and tested on AAF dataset.

cacy and training stability for the current classification task (Fig. 6 and 7). While both methods achieve comparable overall accuracy (around 80%), FixMatch ($\epsilon = 0.9$, PB) reduces the gender accuracy gap to 5.01% by iteration 7, surpassing FlexMatch by 7.51%. This aligns with [30], demonstrating that higher confidence thresholds enhance pseudo-label quality by filtering uncertain predictions, thereby mitigating bias propagation during training. The synergy of pseudo-balancing and elevated confidence thresholds prevents overfitting to dominant classes, allowing fairer gender performance.

FlexMatch achieves the best overall accuracy (83.36%) with FF dataset but exhibits instability when fine-tuning on severely biased distributions. The adaptive thresholding mechanism [35] fails to stabilize training under extreme racial or gender imbalances (e.g., Severe East-Asian Female), resulting in erratic accuracy fluctuations (See Fig. 8 and Table 4). Pseudo-balancing improves FlexMatch's performance by (1) compensating for FlexMatch's tendency to ignore underrepresented classes early in training, and (2) stabilizing threshold adjustments when class-wise learning status is unreliable (See Fig. 5). This explains why PB improves FlexMatch's accuracy on GB-FF (67.49% $\rightarrow$ 83.36%) but fails on Severe East-Asian Female (71.87% vs. no PB 56.04%)—the baseline model's poor initial performance on East Asian faces skews pseudo-label quality irrecoverably.



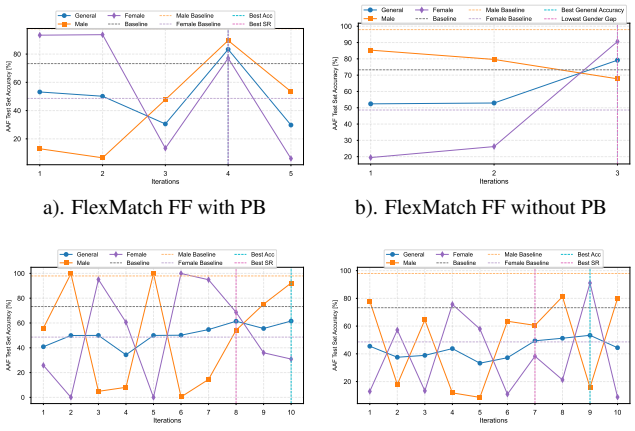
a). $\epsilon = 0.6$ with PB

b). $\epsilon = 0.9$ with PB

c). $\epsilon = 0.6$ without PB

d). $\epsilon = 0.9$ without PB

Figure 7. Self-training per-iteration evolution of fine-tuned models using FixMatch, trained with unlabeled East Asian Female biased FF and tested on AAF dataset.



a). FlexMatch FF with PB

b). FlexMatch FF without PB

c). FlexMatch East Asian Female with PB

d). FlexMatch East Asian Female without PB

Figure 8. Self-training per-iteration evolution of fine-tuned models using FlexMatch, trained with unlabeled FF and East Asian Female biased FF and tested on AAF dataset.

5. Discussion

Our experiments reveal that pseudo-balancing effectively mitigates bias in semi-supervised learning when applied to balanced or moderately biased datasets, improving both accuracy and fairness. However, its efficacy diminishes when severe dataset biases align with the baseline model's intrinsic biases—particularly for East Asian subgroups, where compounding biases degrade performance. The Black subset benefits from partial alignment with the Western-centric pre-training data, highlighting how pseudo-balancing's success depends on the interplay between dataset composition, model bias, and target domain distribution. FlexMatch excels on balanced data but struggles with severe biases, underscoring the need for adaptive thresholding that accounts for these demographic imbalances.

For race-balanced datasets like FairFace, pseudo-balancing mitigates the "Matthew effect"—where the model reinforces existing biases by generating more confident pseudo-labels for already well-performing groups—by enforcing proportional sampling during self-training [2]. This is particularly effective at higher confidence thresholds (ϵ=0.9), where the 93.80% selection rate indicates near-complete utilization of reliable pseudo-labels while maintaining a accuracy of 79.81%. However, the East Asian performance reveals a critical limitation: when the baseline model's architectural bias (from Kaggle dataset pre-training) compounds with dataset bias, pseudo-balancing inadvertently reinforces incorrect feature representations. This manifests most severely in the East Asian Female subset, where pseudo-balancing amplifies the baseline model's accuracy deficit between male and female classes. The relative success of the Black FF subset (SR: 53.87% with PB, 44.08% without PB) indicates that the Kaggle model's Western-centric training provides better initial representations for African facial features than Asian ones. This difference likely arises because Western-centric datasets used for pre-training contain more diverse representations of African-descended faces than East Asian ones [1, 25], leading to better feature extraction for Black faces.

6. Limitations and Future Work

The current study focuses primarily on gender and racial biases in facial analysis, though we recognize that real-world applications face additional challenges in age, expression, and cultural context variations [32]. The effectiveness of pseudo-balancing depends on the initial model's capability to generate reasonable pseudo-labels – a limitation when dealing with completely novel demographic groups not represented in the pre-training data. Future work should investigate hybrid approaches combining our pseudo-balancing technique with demographic-aware augmentation strategies to handle more complex intersectional biases [8]. The promising results on East Asian populations implies that our method could be extended to other underrepresented groups, though this requires validation through larger-scale multicultural studies.

7. Conclusion

We presented a simple yet effective method for mitigating gender bias in face gender classification through semi-supervised learning. Our approach, *pseudo-balancing*, enforces demographic parity during pseudo-label selection without relying on any ground-truth demographic annotations. By leveraging demographically balanced pools of unlabeled data—such as the FairFace dataset—we guide biased models toward fairer outcomes.

Through controlled experiments under both realistic and synthetic bias conditions, we demonstrate that pseudo-balancing improves classification accuracy while significantly reducing gender disparities, particularly on the All-Age-Faces benchmark, which predominantly features East Asian individuals. Although its effectiveness diminishes under severe data imbalance, our results reveal a key insight: access to a balanced or moderately skewed unlabeled dataset, combined with simple balancing constraints, can serve as a practical and scalable strategy for debiasing existing computer vision models.

These findings highlight pseudo-balancing as a simple and practical fairness intervention, especially in settings where demographic labels are unavailable. It offers a general framework for improving fairness in semi-supervised learning across broader computer vision tasks.

Acknowledgment

This work was developed under the NII International Internship Program at National Institute of Informatics (NII) in Tokyo, Japan.

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. 1, 2, 8

[2] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning, 2022. 8

[3] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation, 2020. 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 3

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[6] Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019. 2, 3

[7] Meltem Demirkus, Kshitiz Garg, and Sadiye Guler. Automated person categorization for video surveillance using soft biometrics. In *Defense + Commercial Sensing*, 2010. 2

[8] Samuel Dooley, Rhea Sanjay Sukthanker, John P. Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition, 2023. 8

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery. 3

[10] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *ArXiv*, abs/2004.08514, 2020. 4

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016. 1, 3

[12] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up: Unregulated police face recognition in america, 2016. 2

[13] Hu Han and Anil Jain. Age, gender and race estimation from unconstrained face images, 2014. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 3

[16] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation, 2021. 3

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017. 3

[18] Gary Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Tech. rep.*, 2008. 1, 2

[19] Kaggle. Kaggle Gender Classification Dataset. https://www.kaggle.com/datasets/uciml/gender-classification-dataset, 2020. Accessed: 2025-06-09. 2, 3

[20] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 1, 2, 3

[21] S.A. Khan, Munir Ahmad, Muhammad Nazir, and N. Riaz. A comparative analysis of gender classification techniques. *Middle - East Journal of Scientific Research*, 20:1–13, 2014. 2

[22] Anoop Krishnan and Ajita Rattani. A novel approach for bias mitigation of gender classification algorithms using consistency regularization. *Image and Vision Computing*, 137: 104793, 2023. 1, 2, 3

[23] Vijay Kumar, Raghavendra Ramachandra, Anoop Namboodiri, and Christoph Busch. Robust transgender face recognition: Approach based on appearance and therapy factors. pages 1–7, 2016. 2

[24] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013. 6

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. 8

[26] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019. 2

[27] P. Jonathon Phillips, Harry Wechsler, Jeffrey Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.*, 16:295–306, 1998. 1, 2

[28] Ana Manzano Rodríguez, Camille Guinaudeau, and Shin'ichi Satoh. Uncovering gender biases in gender identification models for japanese data analysis, 2025. CVPR 2025 DemoDiv Workshop, available online: https://cvpr2025.thecvf.com/workshops/demodiv. 3

[29] Namrata Sandhu. Impact of gender cues in advertisements on perceived gender identity meanings of the advertised product. *FIIB Business Review*, 7:231971451880582, 2018. 2

[30] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020. 1, 3, 4, 7

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. 2014. 2

[32] Yitong Wang, Shuang Song, Jing Wang, and Ying Huang. Mitigating gender bias in face recognition using skewness-aware regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12320–12327, 2020. 8

[33] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. Gender classification and bias mitigation in facial images. In *Proceedings of the 12th ACM Conference on Web Science*, page 106–114, New York, NY, USA, 2020. Association for Computing Machinery. 3

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. 4

[35] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling, 2022. 1, 3, 4, 5, 7

[36] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, 2017. 2, 3

[37] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training, 2020. 4

[38] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training, 2018. 4