BENCHMARKING FAIRNESS IN SPIKING NEURAL NETWORKS: DATA BIAS AND SPURIOUS FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating fairness in Spiking Neural Networks (SNNs) demands rigorous benchmarks that reflect real-world complexities, yet existing assessments remain limited by superficial dataset diversity. This work introduces the first systematic fairness benchmark for SNNs, addressing two critical dimensions of realism: (1) demographic coverage gaps in training data and (2) spurious feature leakage (e.g., skin tone as a proxy for class labels). Our framework integrates four cross-demographic datasets together with a chroma-aware cleaning protocol that removes 17.9% of RFW images as color anomalies and yields paired original and cleaned variants, enabling controlled comparison across data conditions. Standardized evaluations of state-of-the-art SNNs 12 reveal pervasive fairness deficiencies, including stark disparities between demographic groups and systematic shortcut learning driven by spurious low-level cues. Our code is available at: `https://anonymous.4open.science/r/SNN-Benchmarks-8017`.

1 INTRODUCTION

By mimicking the event-driven mechanism and binary spatiotemporal dynamics of biological neurons, Spiking Neural Networks (SNNs) show strong potential for low-power computing in edge devices and Internet of Things (IoT) applications (Zhang et al., 2025). Their hardware–algorithm co-design makes them efficient in resource-limited settings (M Ferreira et al., 2025). Recent advances, such as spike-driven Transformers, further improve performance and versatility across various visual tasks by leveraging sparse addition paradigms (Yao et al., 2024). These advances motivate a systematic evaluation of fairness in SNN-based facial image analysis, particularly with respect to differences in predictive performance across demographic groups.

However, the rapid progress of SNNs has hidden an important socio-technical risk: algorithmic fairness. While fairness has been widely studied in many machine learning systems, it is still largely overlooked in SNNs (Altamirano et al., 2025). Consequently, when applied in human-centered vision tasks, these models may produce biased outcomes across demographic groups (Li et al., 2025). For example, in facial recognition or identity classification, individuals from underrepresented racial groups may experience lower accuracy or higher misclassification rates (Pham et al., 2025).

In Artificial Neural Networks (ANNs), such bias is often caused by imbalanced datasets (Schaap et al., 2024), and mitigation typically relies on continuous representation disentanglement or selective fine-tuning (Zhao et al., 2025). Whether conventional debiasing approaches are effective for SNNs requires evaluation in the context of spike-based processing. Current neuromorphic research places considerable emphasis on predictive performance and computational efficiency, including activity pruning (Bu et al., 2025). How spike-based processing affects demographic disparities is therefore an important question for empirical investigation. Complementing existing fairness tools such as OxonFair (Delaney et al., 2024), our benchmark evaluates group disparities across SNN configurations and datasets and tracks their evolution during training.

To address this gap, we propose a fairness benchmark for SNNs. Our framework maps the aforementioned neuromorphic challenges to two dimensions of realism: (1) **demographic coverage gaps** to evaluate fundamental data-level bias; and (2) **spurious feature leakage** to examine possible shortcut learning associated with color-related image characteristics. Additionally, we design fine-grained evaluation protocols to better understand fairness in SNNs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Specifically, we evaluate gender classification on UTKFace, FairFace, and DemogPairs, using race as the sensitive attribute to assess disparities across demographic groups. We also conduct controlled experiments on image grayscale, since grayscale is critical in spike-based processing, to study its impact on fairness. Furthermore, we track per-class accuracy during training on RFW to examine how learning trajectories differ across its racial categories. Through comprehensive experiments across multiple datasets, our benchmark enables a systematic study of fairness–performance trade-offs. Our main contributions are summarized as follows:

- We design a standardized benchmark for evaluating fairness in SNNs on four facial image datasets. For gender classification on UTKFace, FairFace, and DemogPairs, we jointly report SP and EO using race as the sensitive attribute. For RFW, we evaluate 11 SNN configurations on race classification and use the maximum accuracy gap across racial groups as an alternative fairness diagnostic.
- We compare group disparities between the original RFW dataset and a cleaned subset. We construct a refined subset, *RFW-Cleaned*, by removing 7,280 black-and-white and abnormal-color images (17.9% of 40,607). The maximum inter-group accuracy gap decreases for some SNN configurations but increases for others, showing that the observed differences in group accuracy parity between the two dataset variants depend on the model configuration.
- We track group-wise race-classification accuracy over 150 training epochs on RFW and RFW-Cleaned. We observe uneven accuracy trajectories across racial groups, with pronounced disparities already evident in the early stages of training. These early disparities may reflect differences in representation learning across racial groups. By tracing how these disparities evolve, our analysis complements fairness evaluation based on final model performance.

2 METHODOLOGY

2.1 FAIRNESS METRICS

We evaluate the fairness of SNNs using two primary metrics: Statistical Parity (SP) (Dwork et al., 2012) and Equal Opportunity (EO) (Hardt et al., 2016). For both metrics, a value closer to 0 indicates a fairer SNN model. For the binary-group case, let $\mathbf{A} \in \{0, 1\}$ represent the sensitive attribute. Let $\mathbf{Y} \in \{0, 1\}$ denote the ground-truth gender label and $\hat{\mathbf{Y}} \in \{0, 1\}$ the predicted gender label. Statistical parity difference (Δ_{SP}) measures whether the model provides equal positive outcomes to all groups. It requires the prediction $\hat{\mathbf{Y}}$ to be independent of the sensitive attribute $\mathbf{A}$. We calculate the absolute difference in acceptance rates between groups:

$$\Delta_{\text{SP}} = |\mathbb{P}(\hat{\mathbf{Y}} = 1|\mathbf{A} = 1) - \mathbb{P}(\hat{\mathbf{Y}} = 1|\mathbf{A} = 0)|. \tag{1}$$

Equal opportunity difference (Δ_{EO}) focuses on the fairness of outcomes for qualified individuals. It requires the true positive rate to be equal across groups. We define the disparity as:

$$\Delta_{\text{EO}} = |\mathbb{P}(\hat{\mathbf{Y}} = 1|\mathbf{A} = 1, \mathbf{Y} = 1) - \mathbb{P}(\hat{\mathbf{Y}} = 1|\mathbf{A} = 0, \mathbf{Y} = 1)|. \tag{2}$$

In practice, our benchmark involves more than two racial groups (5 on UTKFace, 7 on FairFace, and 3 on DemogPairs), and the binary formulations above are therefore generalized to the multi-group setting. Let G_i denote the group of samples belonging to racial category i. For each group, we compute the conditional positive-prediction rate $P(\hat{\mathbf{Y}} = 1|G_i)$, and define Δ_{SP} as the maximum absolute pairwise difference across all groups:

$$\Delta_{\text{SP}} = \max_{i,j} \left| P(\hat{\mathbf{Y}} = 1|G_i) - P(\hat{\mathbf{Y}} = 1|G_j) \right|. \tag{3}$$

Analogously, Δ_{EO} compares the true positive rate $P(\hat{\mathbf{Y}} = 1|\mathbf{Y} = 1, G_i)$ across all group pairs:

$$\Delta_{\text{EO}} = \max_{i,j} \left| P(\hat{\mathbf{Y}} = 1|\mathbf{Y} = 1, G_i) - P(\hat{\mathbf{Y}} = 1|\mathbf{Y} = 1, G_j) \right|. \tag{4}$$

We adopt the maximum pairwise difference rather than a mean-difference statistic because it is sensitive to extreme deviations: mean-based metrics can be smoothed by the overall data distribution and conceal severe discrimination against specific group pairs, whereas the maximum-difference operator explicitly identifies the worst-affected pairs and directly pinpoints where bias is most pronounced.

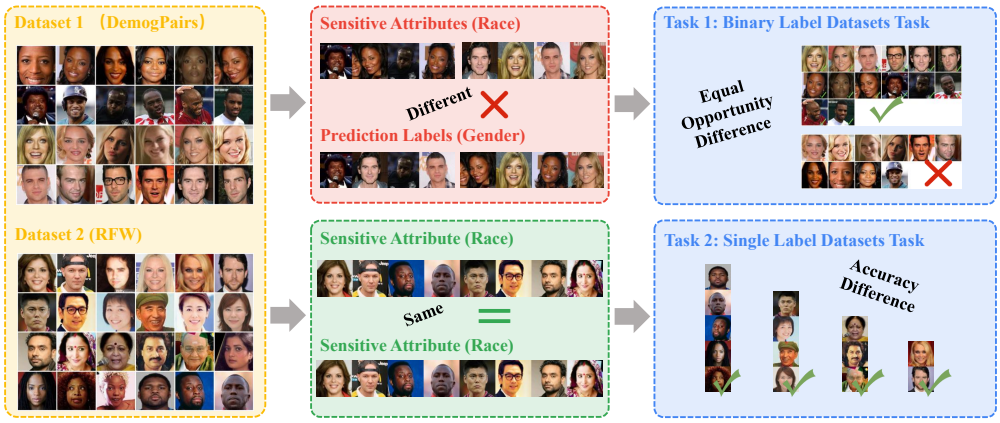


Figure 1: Fairness evaluation framework for the two task settings. The upper panel illustrates gender classification with race as the sensitive attribute, evaluated using SP and EO. The lower panel illustrates the alternative protocol for RFW: four-class race classification, with group disparities measured by the maximum accuracy gap across groups defined by true race labels.

2.2 Problem Formulation: The Fairness Gap as Optimization Constraints

Deep SNNs, including Spike-ResNet and QKFormer, achieve high performance using Surrogate Gradient (SG) (Zhou et al., 2024a) methods. Their event-driven nature ensures energy efficiency for edge deployment in visual tasks (Zhang et al., 2025). Standard SNN training focuses solely on minimizing task-specific loss $\mathcal{L}_{\text{task}}$, where θ represents the trainable parameters:

$$\min_{\theta} \mathcal{L}_{\text{task}}(f_\theta(X), Y). \tag{5}$$

Minimizing the task loss does not explicitly enforce SP or EO. For the gender-classification tasks considered here, requirements on these two metrics can be expressed conceptually as constraints on the training objective:

$$\min_{\theta} \mathcal{L}_{\text{task}}(f_\theta(X), Y) \quad \text{s.t.} \quad \Delta_{\text{SP}} \leq \tau_{\text{SP}}, \quad \Delta_{\text{EO}} \leq \tau_{\text{EO}}. \tag{6}$$

Here, τ_{SP} and τ_{EO} specify tolerances for the corresponding group disparities. This formulation illustrates how fairness requirements could enter model training. Our benchmark evaluates standard SNN training configurations and reports these disparities after training; it does not optimize models under these constraints.

LIF neurons integrate inputs over time, with membrane leakage and threshold-based spike generation. These dynamics motivate investigating how visual cues contribute to predictions across demographic groups. From the perspective of shortcut learning (Geirhos et al., 2020), one possible explanation is that readily learned color-related cues contribute more strongly to predictions than other facial features. Interactions between input-driven activation, membrane leakage, and firing thresholds could influence this process. However, the group-wise accuracy patterns alone do not establish this mechanism.

To quantify disparities in prediction outcomes, we treat race as the sensitive attribute $\mathbf{A}$ and gender as the ground-truth label $\mathbf{Y}$ for the classification task. We evaluate the trained SNNs using SP and EO to measure disparities across racial groups. The evaluation pipeline is illustrated in the upper panel of Figure 1.

2.3 Alternative Evaluation Protocol for RFW

Our primary fairness evaluation jointly uses statistical parity (SP) and equal opportunity (EO). On UTKFace, FairFace, and DemogPairs, gender is the prediction target and race is the sensitive attribute. The RFW annotations used in our experiments do not include gender labels, so this gender-classification-based evaluation protocol cannot be applied directly to RFW (Wang & Deng, 2021).

Table 1: Statistics and task settings of the datasets used in our SNN fairness benchmark. Race is the sensitive attribute considered across all four datasets. For UTKFace, FairFace, and DemogPairs, we evaluate SP and EO in gender classification. For RFW, race also serves as the prediction target, and we assess group disparities using the maximum accuracy gap across groups defined by true race labels.

Dataset	Images	Races	Sensitive attributes	Target
RFW	$40,607$	4	race	race
FairFace	$97,698$	7	race	gender
UTKFace	$20,856$	5	race	gender
DemogPairs	$10,800$	3	race	gender

We therefore adopt an engineering workaround based on disparities in race-classification accuracy across racial groups, as described below.

Specifically, we perform four-class race classification on RFW using the dataset-provided Caucasian, Asian, Indian, and African categories. Let $\mathcal{A}$ denote the set of these four categories. For each category $a \in \mathcal{A}$, Acc_a is the proportion of evaluation samples with true race label a that are correctly classified. The denominator is therefore the number of evaluation samples belonging to that true racial category, rather than the number predicted as that category. If a sample is misclassified into another racial category, it is still counted as an error for its true group. We use the maximum accuracy gap across these racial groups as an alternative measure of group performance disparity within our fairness evaluation:

$$\Delta_{\mathrm{Acc}} = \max_{a \in \mathcal{A}} \mathrm{Acc}_a - \min_{a \in \mathcal{A}} \mathrm{Acc}_a. \tag{7}$$

A smaller Δ_{Acc} indicates greater parity in race-classification accuracy across the four racial groups. The measure summarizes the largest pairwise difference among the four group accuracies. The groups attaining the highest and lowest accuracies are determined by the evaluated results, rather than designated in advance. This engineering workaround provides a performance-based fairness diagnostic for RFW.

As illustrated in the lower panel of Figure 1, we report the classification accuracy for each racial group and use Δ_{Acc} to summarize the largest inter-group difference. Reporting the four group accuracies alongside this gap shows which groups occupy the upper and lower ends of the observed performance range and where the remaining groups lie. Unlike SP and EO in our gender-classification tasks, this measure compares accuracies across categories that also serve as the prediction target. Its values should therefore be interpreted within the RFW task rather than treated as directly comparable to the SP and EO disparities reported for the other datasets.

3 Experiments

3.1 Datasets and Baselines

We employ four benchmark datasets (UTKFace (Zhang et al., 2017), FairFace (Karkkainen & Joo, 2021), RFW (Wang & Deng, 2021), and DemogPairs (Hupont & Fernández, 2019)) to evaluate SNN fairness. These datasets complement each other in scale, demographic balance, and evaluation focus. Detailed dataset descriptions and the summary in Table 1 are provided in Appendix B.

- RFW (Racial Faces in-the-Wild) : Designed specifically to quantify racial bias in face recognition. It overcomes the lack of racial diversity found in traditional benchmarks like LFW.
- UTKFace : A large-scale dataset featuring over $20,000$ images with a wide age range (0–116 years). It supports multi-attribute evaluation across age, gender, and race.
- DemogPairs : A balanced verification dataset was constructed to address biased data distributions. It ensures that model biases are not obscured by skewed test samples.
- FairFace : Aims to mitigate racial imbalance in public datasets (e.g., ImageNet), where certain groups exceed 80% of the data. It provides a representative distribution across seven racial groups.

Table 2: Comparison on accuracy and fairness (Δ_{SP} and Δ_{EO}) in percentage (%) with three real world datasets. ↑ denotes the larger, the better; ↓ denotes the opposite. Results are reported as mean$^{\pm\text{std}}$ over multiple runs with different random seeds.

Method	ACC(%) ↑	Δ_{SP}(%) ↓	Δ_{EO}(%) ↓	ACC(%) ↑	Δ_{SP}(%) ↓	Δ_{EO}(%) ↓	ACC(%) ↑	Δ_{SP}(%) ↓	Δ_{EO}(%) ↓
	UTKFace			FairFace			DemogPairs		
QKFormer	$66.22^{\pm0.08}$	$10.85^{\pm0.05}$	$11.85^{\pm0.09}$	$61.26^{\pm0.11}$	$21.74^{\pm0.05}$	$27.33^{\pm0.05}$	$57.14^{\pm0.02}$	$12.86^{\pm0.21}$	$17.89^{\pm0.05}$
SEW-ResNet	$85.14^{\pm0.18}$	$6.03^{\pm0.17}$	$0.37^{\pm0.17}$	$96.21^{\pm0.17}$	$19.00^{\pm0.06}$	$7.74^{\pm0.05}$	$64.94^{\pm0.17}$	$60.89^{\pm0.00}$	$73.83^{\pm0.10}$
SdT	$93.24^{\pm0.05}$	$14.27^{\pm0.14}$	$7.99^{\pm0.12}$	$91.10^{\pm0.09}$	$14.37^{\pm0.10}$	$11.78^{\pm0.15}$	$86.85^{\pm0.17}$	$1.59^{\pm0.24}$	$11.15^{\pm0.18}$
Spikformer	$88.68^{\pm0.11}$	$12.40^{\pm0.21}$	$5.42^{\pm0.10}$	$92.33^{\pm0.13}$	$6.05^{\pm0.06}$	$4.91^{\pm0.12}$	$57.24^{\pm0.16}$	$15.33^{\pm0.18}$	$16.42^{\pm0.01}$
SdT V3	$87.19^{\pm0.03}$	$15.79^{\pm0.17}$	$12.28^{\pm0.03}$	$91.02^{\pm0.02}$	$7.83^{\pm0.03}$	$9.08^{\pm0.10}$	$65.86^{\pm0.13}$	$7.04^{\pm0.04}$	$9.17^{\pm0.11}$
SRN18	$86.05^{\pm0.12}$	$6.58^{\pm0.16}$	$0.57^{\pm0.10}$	$93.07^{\pm0.18}$	$17.64^{\pm0.01}$	$7.61^{\pm0.01}$	$79.24^{\pm0.16}$	$54.08^{\pm0.09}$	$78.33^{\pm0.06}$
VGGSNN (T)	$85.04^{\pm0.30}$	$13.83^{\pm0.09}$	$11.03^{\pm0.13}$	$89.76^{\pm0.06}$	$17.70^{\pm0.08}$	$13.76^{\pm0.06}$	$80.67^{\pm0.18}$	$4.04^{\pm0.02}$	$9.75^{\pm0.22}$
VGGSNN (F)	$83.53^{\pm0.07}$	$14.56^{\pm0.11}$	$13.71^{\pm0.14}$	$89.02^{\pm0.04}$	$17.81^{\pm0.10}$	$12.26^{\pm0.04}$	$78.97^{\pm0.21}$	$3.49^{\pm0.01}$	$9.36^{\pm0.06}$
SRN19 (T)	$83.39^{\pm0.17}$	$15.58^{\pm0.12}$	$16.32^{\pm0.09}$	$89.61^{\pm0.05}$	$18.46^{\pm0.10}$	$9.73^{\pm0.11}$	$82.00^{\pm0.17}$	$3.23^{\pm0.16}$	$9.13^{\pm0.18}$
SRN19 (F)	$82.91^{\pm0.12}$	$15.83^{\pm0.30}$	$15.32^{\pm0.13}$	$89.66^{\pm0.24}$	$17.43^{\pm0.02}$	$10.88^{\pm0.07}$	$77.69^{\pm0.03}$	$1.56^{\pm0.19}$	$9.27^{\pm0.02}$
SdT V2	$83.70^{\pm0.07}$	$19.36^{\pm0.15}$	$17.74^{\pm0.03}$	$90.21^{\pm0.05}$	$8.24^{\pm0.04}$	$9.42^{\pm0.05}$	$78.32^{\pm0.13}$	$19.09^{\pm0.02}$	$12.06^{\pm0.02}$
STAtten	$90.19^{\pm0.22}$	$15.76^{\pm0.19}$	$9.37^{\pm0.20}$	$89.78^{\pm0.21}$	$8.23^{\pm0.05}$	$9.08^{\pm0.08}$	$92.37^{\pm0.22}$	$15.71^{\pm0.17}$	$7.37^{\pm0.13}$

Our framework jointly reports SP and EO for gender classification on UTKFace, FairFace, and DemogPairs, using race as the sensitive attribute. For RFW, we perform race classification and use the maximum accuracy gap across racial groups as an alternative measure of group performance disparity within our fairness evaluation. This multi-perspective design effectively identifies biases that are often hidden in single-dataset evaluations. To ensure a comprehensive benchmark, we select representative SNNs based on three criteria:

- Competitive accuracy on ImageNet-1K (Deng et al., 2009) and CIFAR (Doon et al., 2018).
- Covering convolutional and Transformer-based SNN architectures, including models with residual connections.
- Compatibility with unified direct-training and evaluation protocols.

We describe the selected methods in the following three parts:

- Residual and Direct-Training Baselines: These models represent the standard surrogate gradient paradigm. We include SNN-ResNet18 (SRN18 as for short), SEW-ResNet (identity mapping optimization) (Fang et al., 2021), and Spike-ResNet19, VGGSNN with Temporal Efficient Training (TET) (Deng et al., 2022), which we refer to as SRN19 (T/F) and VGGSNN (T/F) for short.
- Spiking Transformers: This group captures the evolution of attention mechanisms in SNNs. The benchmark covers early designs like Spikformer (Zhou et al., 2023) and Spike-driven Transformer (SdT) (Yao et al., 2023), to advanced architectures like Meta-SpikeFormer (SdT V2) (Yao et al., 2024), E-SpikeFormer (SdT V3) (Yao et al., 2025), and QKFormer (Zhou et al., 2024b).
- Attention Enhancement Mechanism: We introduce STAtten (Lee et al., 2025), which leverages block-wise spatial-temporal attention to disentangle and enhance spatial-temporal feature extraction while preserving the convolutional inductive bias, without increasing the asymptotic computational complexity.

The benchmark includes both different model families and alternative training configurations within a model family. In particular, the TET and non-TET versions of Spike-ResNet19 and VGGSNN are reported separately. These paired configurations make it possible to compare the reported group disparities under the two training settings within each model family. Comparisons across model families address a different question: how the evaluated SNN configurations differ under the benchmark's task settings. Distinguishing these comparisons helps relate the results to the specific configuration being evaluated.

By comparing these model configurations, we examine whether higher classification accuracy is consistently accompanied by smaller disparities across racial groups. The experimental protocol is detailed in Appendix C.

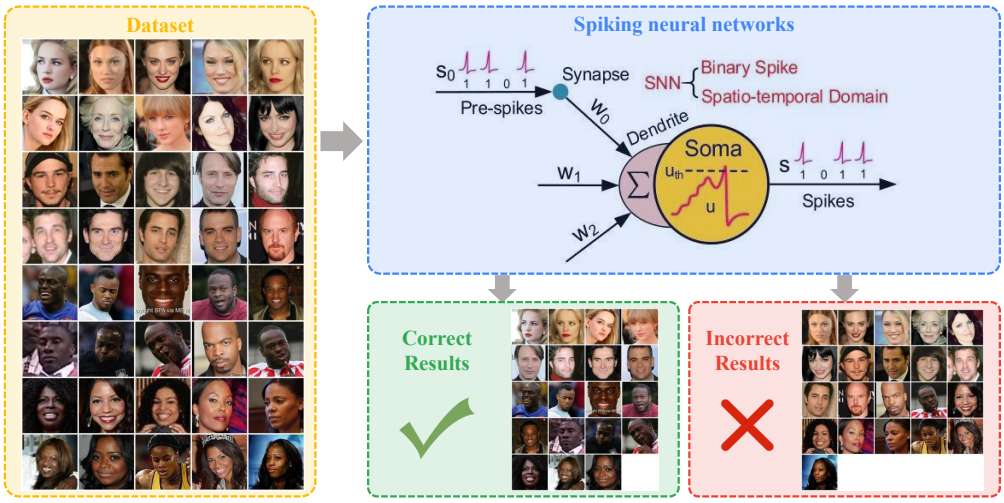


Figure 2: Qualitative results: Bias in SNN classification.

3.2 Empirical Results and Comparative Analysis

Table 2 compares classification accuracy and fairness metrics across the three datasets. Higher accuracy is not consistently accompanied by smaller group disparities. On UTKFace, SdT achieves the highest accuracy (93.24%) but has a larger Δ_{SP} (14.27%) than SEW-ResNet, which achieves 85.14% accuracy and a Δ_{SP} of 6.03%. On DemogPairs, STAtten achieves higher accuracy than SdT (92.37% versus 86.85%) but also has a larger Δ_{SP} (15.71% versus 1.59%). These comparisons highlight the need to evaluate accuracy and group disparities jointly when assessing SNN performance.

The relationship between accuracy and group disparities also depends on the fairness metric. On DemogPairs, STAtten achieves both the highest accuracy (92.37%) and the lowest Δ_{EO} (7.37%) among the evaluated configurations, despite having a larger Δ_{SP} than SdT. This result illustrates that a model can perform well in terms of accuracy and equal opportunity without achieving the smallest statistical parity disparity.

The reported disparities also vary across datasets. For example, SEW-ResNet has a mean Δ_{SP} of 6.03% on UTKFace, compared with 19.00% on FairFace and 60.89% on DemogPairs. These results show that its reported SP disparity is not uniform across the three evaluation datasets. They motivate evaluating models on multiple datasets and interpreting each disparity in the context of the corresponding task and demographic composition.

Within FairFace, comparing SEW-ResNet with Spikformer further illustrates the distinction between classification accuracy and group disparities. As reported in Table 2, SEW-ResNet achieves higher mean accuracy (96.21% versus 92.33%), but also larger mean disparities under both fairness criteria: 19.00% versus 6.05% for Δ_{SP}, and 7.74% versus 4.91% for Δ_{EO}. Thus, the model with higher accuracy in this comparison has larger group disparities under both SP and EO. Considering all three measures makes this difference visible, whereas selecting a model solely by accuracy would omit the fairness differences between these two configurations.

3.3 Microscopic Insights and Interpretability

To dissect the failure modes of high-performance SNNs, we examine QKFormer on the RFW dataset, revealing a severe performance imbalance across racial groups. The model exhibits a dramatic accuracy gradient: 83.68% for African samples versus a shockingly low 32.64% for Caucasian samples. Inspection of the misclassified samples reveals recurring color-related characteristics, including grayscale images and color shifts. Pronounced tone sensitivity bias exists: when Caucasian faces are converted to grayscale, QKFormer misclassifies them as “African" at a rate of 31.5%. These observations suggest a possible association between color-related image characteristics and QKFormer’s race-classification errors.

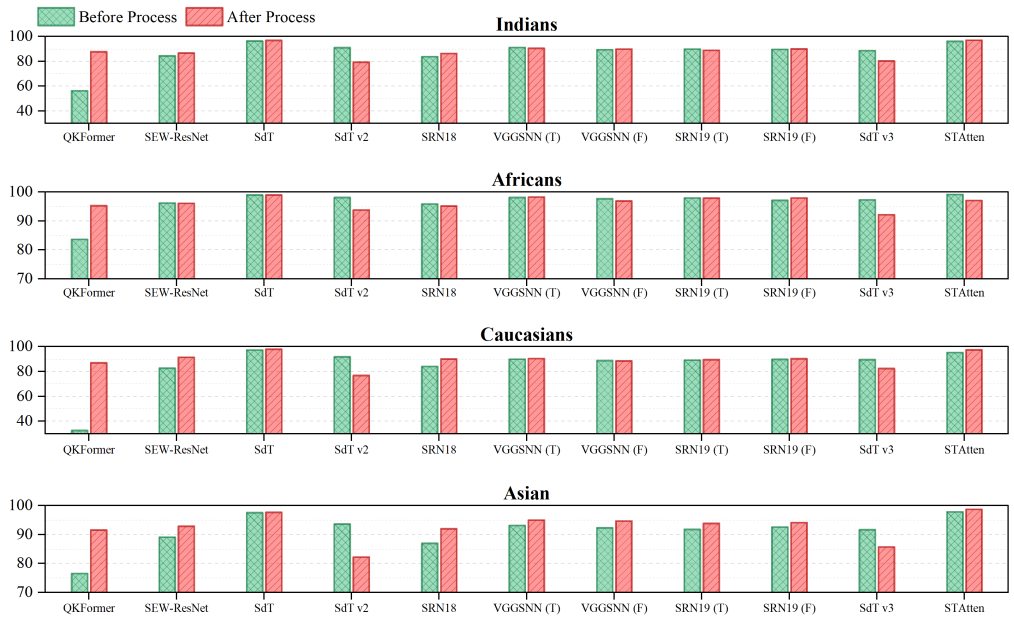


Figure 3: Group-wise race-classification accuracy (%) of the evaluated SNN models on RFW and RFW-Cleaned. Each panel corresponds to a group defined by its true race label. Green and red bars show results on RFW and RFW-Cleaned, respectively.

We also examine color-related patterns in QKFormer's race-classification errors. In color images, Caucasian samples with yellow-leaning tones or color temperature shifts are frequently misclassified as Asian (4.9%) or Indian (11.2%).

We also observe that some grayscale Indian samples are misclassified as African by QKFormer.

We also examine gender-classification performance across race–gender subgroups using SEW-ResNet on DemogPairs. In the examined set of 40 images, the model correctly classifies 8/10 African male samples and 3/10 African female samples. This difference illustrates unequal gender-classification accuracy between these two subgroups within the examined samples and motivates further evaluation of intersectional group disparities.

3.4 Asymmetric Learning Trajectories Across Racial Groups

Fairness evaluation based on final model performance does not capture how group disparities evolve during training. By tracking group-wise race-classification accuracy on RFW and RFW-Cleaned, we observe Asymmetric Learning Trajectories: racial groups differ in their accuracy levels and in how their accuracies change over training. In our reporting, epoch indices start at 0, and the results at Epoch 0 are obtained by evaluating the model after the first complete training epoch. Pronounced group disparities are already evident at this early checkpoint and continue to evolve during subsequent training. These asymmetric trajectories may reflect differences in representation learning across racial groups, highlighting the value of examining fairness throughout training.

- Case A (Early Group Disparity): On RFW-Cleaned, Spike-ResNet19 with TET achieves 92.89% accuracy on the African group and 24.34% on the Indian group at Epoch 0, a difference of 68.55 percentage points. This shows that a substantial group disparity is already present after the first complete training epoch.
- Case B (Early Disparity in VGGSNN): On RFW-Cleaned, VGGSNN without TET achieves 87.38% accuracy on the African group and 46.73% on the Caucasian group at Epoch 0. Thus, early group disparities are observed in both the Spike-ResNet19 and VGGSNN configurations considered here.

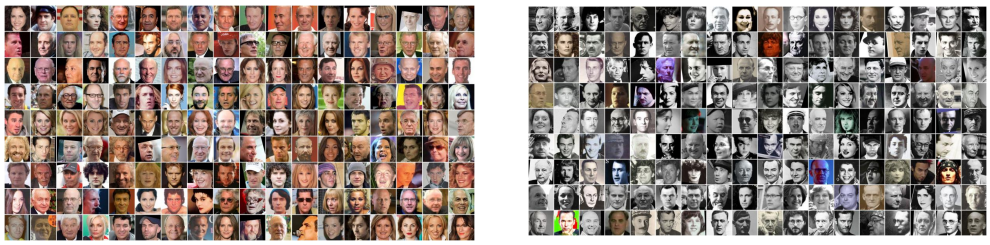


Figure 4: QKFormer's race-classification errors on Caucasian samples are primarily associated with grayscale and color-distorted images, whereas natural color images are mostly classified correctly.

- Case C (Early Disparity on Original RFW): On the original RFW dataset, Spike-ResNet19 with TET achieves 79.85% accuracy on the African group but only 4.50% on the Indian group at Epoch 0. The latter is below the 25% expected accuracy of uniform random guessing in a four-class task, highlighting the pronounced imbalance in early group performance.

Tracking the first ten reported checkpoints (E_0–E_9) reveals asymmetric and non-monotonic accuracy trajectories across racial groups.

- Case D (Divergent Accuracy Changes): On RFW-Cleaned, Spike-ResNet19 with TET shows an increase in African-group accuracy from 93.69% at E_1 to 98.54% at E_2, while Caucasian-group accuracy decreases from 52.73% to 23.69% over the same interval. Caucasian-group accuracy subsequently rises to 62.54% at E_3. This illustrates that improvements for one group can coincide with a temporary decline for another.
- Uneven Early Progress: For the same model configuration and dataset, Indian-group accuracy rises from 24.34% at E_0 to 87.90% at E_9, whereas African-group accuracy is already 92.89% at E_0. These trajectories illustrate differences in the timing and extent of accuracy improvements across racial groups.

Possible Gradient-Based Explanation: These asymmetric learning trajectories may reflect an imbalance in how different features contribute to parameter updates. One hypothesis is that gradients associated with readily learned features dominate early optimization, potentially limiting the learning of features useful for other racial groups. Such a mechanism could contribute to the observed differences in learning trajectories, but establishing it requires further analysis of group-wise gradients and learned representations.

- **Case G (Non-monotonic Group Accuracy):** On RFW-Cleaned, VGGSNN without TET achieves 83.94% accuracy on the Caucasian group at E_3, followed by a decrease to 69.18% at E_4 and a recovery to 81.02% at E_5. This fluctuation provides another example of non-monotonic group performance within the asymmetric learning trajectories.
- **Implications for Fairness Evaluation and Training:** The presence of early group disparities and subsequent fluctuations motivates evaluating fairness throughout training, alongside the final model performance. Future work could investigate whether training objectives that account for group disparities can improve group accuracy parity. Such approaches should be evaluated using fairness measures appropriate to each task setting.

4 Limitation

Our analysis considers grayscale and other color-related image characteristics, as well as training dynamics. Further work could more systematically examine how encoding strategies, architectural choices, and training objectives affect group disparities. Developing methods to mitigate these disparities is another direction for future work.

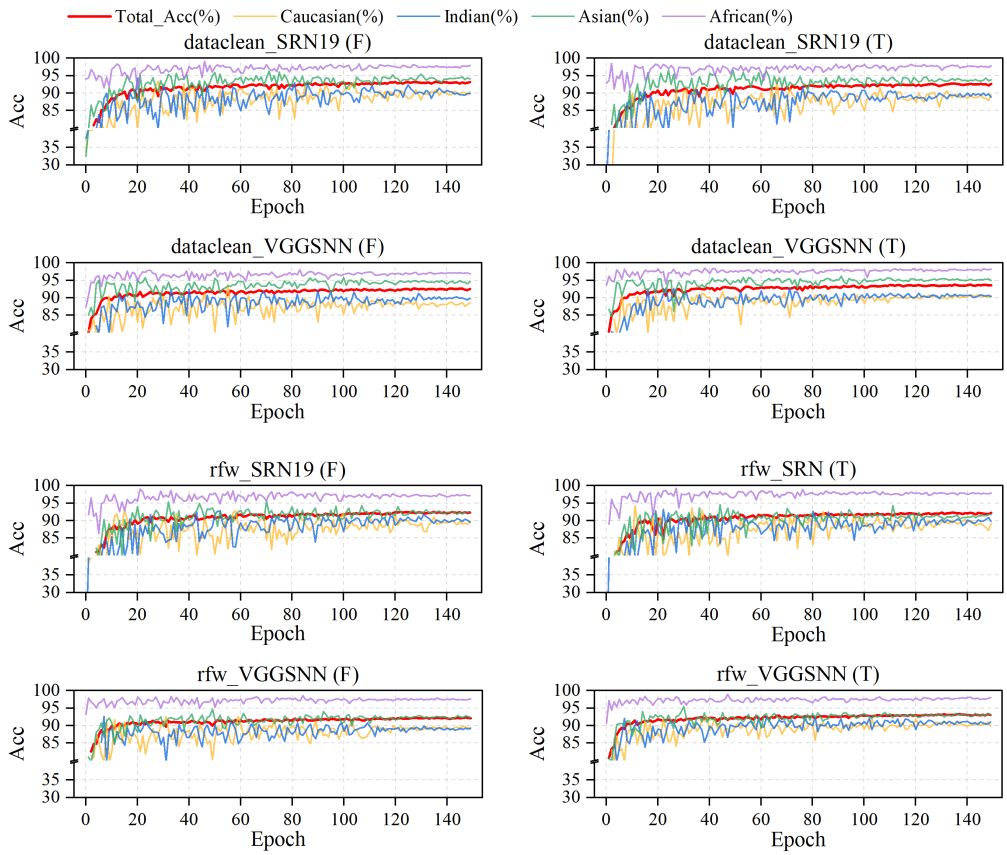


Figure 5: Group-wise and overall test accuracy (%) over 150 training epochs for Spike-ResNet19 and VGGSNN, with and without TET, on RFW-Cleaned and the original RFW dataset. The red curve represents overall accuracy, while the other curves represent accuracy for groups defined by true race labels. The group-wise curves illustrate asymmetric learning trajectories across racial groups. Epoch indices start at 0, with Epoch 0 reporting evaluation after the first complete training epoch.

5 Conclusion

In this work, we introduce the first systematic fairness benchmark for Spiking Neural Networks (SNNs), addressing a critical gap in equitable neuromorphic computing. By formalizing two dimensions of realism (demographic coverage and spurious feature leakage), we challenge the adequacy of existing SNN evaluations, which often overlook the interplay between algorithmic bias and spike-based processing. By combining four cross-demographic datasets, our benchmark supports the assessment of classification performance and demographic disparities across different data conditions.

Through standardized evaluations of 12 state-of-the-art SNNs, we expose non-trivial disparities in accuracy and fairness across demographic groups, along with phenomena such as texture-based shortcut learning and asymmetric convergence trajectories. These findings underscore the need for fairness-aware evaluation protocols and mitigation strategies specifically tailored to the unique learning dynamics of SNNs.

Future work will investigate how encoding strategies and training objectives affect group disparities and explore fairness-aware training methods for SNN-based facial image analysis. Our benchmark provides an empirical basis for assessing group disparities in SNNs on facial image datasets and supports further research on the roles of demographic coverage and spurious visual cues in SNN fairness.

AI USE STATEMENT

We used OpenAI Codex to assist with English-language editing,translation between Chinese and English, literature search, and checks of terminology, numerical consistency, and the alignment between reported results and their interpretation. We also used the tool to obtain feedback on the description of the evaluation methodology and to identify claims that required additional evidence or clarification. AI assistance was used to flag potential issues and suggest revisions; decisions about changes to the manuscript remained with the authors. The authors take full responsibility for the final content of this work, including AI-assisted text and interpretations.

REPRODUCIBILITY STATEMENT

All experiments in this work are designed to be fully reproducible. The four benchmark datasets employed (UTKFace, FairFace, RFW, and DemogPairs) are publicly available, and the complete experimental protocol, including training configurations, hyperparameters, and hardware specifications, is detailed in Appendix C. Our source code, model configurations, and evaluation scripts are released at `https://anonymous.4open.science/r/SNN-Benchmarks-8017`.

ETHICS STATEMENT

This study analyzes facial images annotated with demographic attributes, which constitutes sensitive personal information. We exclusively employ established, publicly released research benchmarks (UTKFace, FairFace, RFW, and DemogPairs) curated by their original authors under appropriate consent and privacy protocols; no new personal data are collected or annotated in this work. Our benchmark is intended solely for auditing and exposing fairness deficiencies in SNNs. The reported performance disparities across demographic groups should be interpreted as artifacts of datasets, training procedures, and architectures rather than attributes of the groups themselves. We further note that face analysis technologies carry inherent dual-use risks, and we advocate that any real-world deployment be accompanied by systematic fairness audits of the kind this benchmark enables.

REFERENCES

Sara Altamirano, Arjan Vreeken, and Sennay Ghebreab. Machine learning and public health: Identifying and mitigating algorithmic bias through a systematic review. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 138–151, 2025.

Tong Bu, Xinyu Shi, and Zhaofei Yu. Activity pruning for efficient spiking neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. Oxonfair: A flexible toolkit for algorithmic fairness. *Advances in Neural Information Processing Systems*, 37:94209–94245, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *The Tenth International Conference on Learning Representations, ICLR2022, Virtual Event, April 25-29, 2022*, 2022.

Raveen Doon, Tarun Kumar Rawat, and Shweta Gautam. Cifar-10 classification using deep convolutional neural network. In *2018 IEEE Punecon*, pp. 1–5, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. In *Advances in Neural Information Processing Systems*, pp. 21056–21069, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2), 2024.

Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7, 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1), 2012.

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*, pp. 1548–1558, 2021.

Donghyun Lee, Yuhang Li, Youngeun Kim, Shiting Xiao, and Priyadarshini Panda. Spiking transformer with spatial-temporal attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13948–13958, 2025.

Yang Li, Xiang He, Qingqun Kong, and Yi Zeng. Spiking neural networks with consistent mapping relations allow high-accuracy inference. *Information Sciences*, 677:120822, 2024.

Ziqi Li, Tao Gao, Yisheng An, Ting Chen, Jing Zhang, Yuanbo Wen, Mengkun Liu, and Qianxi Zhang. Brain-inspired spiking neural networks for energy-efficient object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3552–3562, 2025.

Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. *On the fairness of disentangled representations*. Curran Associates Inc., 2019.

Pietro M Ferreira, Siqi Wang, Yueyuan Gao, and Aziz Benlarbi-Delai. A comparative review of deep and spiking neural networks for edge ai neuromorphic circuits. *Frontiers in Neuroscience*, 19: 1676570, 2025.

Ayesha Manzoor and Ajita Rattani. Fineface: Fair facial attribute classification leveraging fine-grained features. In *Pattern Recognition*, pp. 81–97, 2025.

Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Otavio Parraga, Martin D. More, Christian M. Oliveira, Nathan S. Gavenski, Lucas S. Kupssinskü, Adilson Medronha, Luis V. Moura, Gabriel S. Simões, and Rodrigo C. Barros. Fairness in deep learning: A survey on vision and language research. *ACM Comput. Surv.*, 57(6), 2025.

Duc Thien Pham, Maryam Khoshkhooy Titkanlou, and Roman Mouček. A hybrid spiking neural network–transformer architecture for motor imagery and sleep apnea detection. *Frontiers in Neuroscience*, 19:1716204, 2025.

Fardin Ahsan Sakib, Ziwei Zhu, Karen Trister Grace, Meliha Yetisgen-Yildiz, and Ozlem Uzuner. Spurious correlations and beyond: Understanding and mitigating shortcut learning in sdoh extraction with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1097–1106, 2025.

Anthonie Schaap, Sofoklis Kitharidis, and Niki van Stein. Towards fairness in machine learning: Balancing racially imbalanced datasets through data augmentation and generative ai. In *IJCCI*, pp. 584–592, 2024.

Guobin Shen, Jindong Li, Tenglong Li, Dongcheng Zhao, and Yi Zeng. Spikepack: Enhanced information flow in spiking neural networks with high hardware compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23385–23395, 2025.

Pirzada Suhail, Vrinda Goel, and Amit Sethi. Shortcut learning susceptibility in vision classifiers (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 41390–41392, 2026.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8433–8448, 2021.

Man Yao, JiaKui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64043–64058, 2023.

Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *The Twelfth International Conference on Learning Representations , (ICLR)*, 2024.

Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2973–2990, 2025.

Brian Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 12 2018.

Guanlei Zhang, Lei Feng, Fanqin Zhou, Zhixiang Yang, Qiyang Zhang, Alaa Saleh, Praveen Kumar Donta, and Chinmaya Kumar Dehury. Spiking neural networks in intelligent edge computing. *IEEE Consumer Electronics Magazine*, 14(4):66–75, 2025.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4352–4360, 2017.

Zengqun Zhao, Ziquan Liu, Yu Cao, Shaogang Gong, and Ioannis Patras. Aim-fair: Advancing algorithmic fairness via selectively fine-tuning biased models with contextual synthetic data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28748–28758, 2025.

Chenlin Zhou, Han Zhang, Liutao Yu, Yumin Ye, Zhaokun Zhou, Liwei Huang, Zhengyu Ma, Xiaopeng Fan, Huihui Zhou, and Yonghong Tian. Direct training high-performance deep spiking neural networks: a review of theories and methods. *Frontiers in Neuroscience*, 18, 2024a.

Chenlin Zhou, Han Zhang, Zhaokun Zhou, Liutao Yu, Liwei Huang, Xiaopeng Fan, Li Yuan, Zhengyu Ma, Huihui Zhou, and Yonghong Tian. Qkformer: Hierarchical spiking transformer using q-k attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13074–13098, 2024b.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=frE4fUwz_h`.

A Related Work

A.1 Learning Methods and Architectural Evolution of SNNs

SNNs rely on discrete spikes and leaky integrate-and-fire dynamics, introducing temporal leakage and non-linear firing thresholds (Shen et al., 2025). This temporal dynamic acts as an asymmetric information filter. Strong, superficial visual cues (such as skin tone or global image contrast) can

rapidly drive membrane potentials to cross-firing thresholds in early timesteps, thereby hijacking the network's gradient updates. Consequently, SNNs exhibit a unique susceptibility to texture-based shortcut learning (Suhail et al., 2026; Sakib et al., 2025), while fine-grained geometric features, which are crucial for equitable recognition across diverse groups, often suffer from temporal starvation.

Currently, the optimization of high-performance SNNs is primarily dominated by two paradigms. The first is the ANN-to-SNN conversion approach (Li et al., 2024), which equivalently maps the continuous activation functions (e.g., ReLU) of pre-trained ANNs to the firing rates of spiking neurons. Despite partial success, this paradigm is typically bottlenecked by the requirement for extremely long inference timesteps to match the original activation precision, and exhibits limited adaptability to non-standard architectures. In contrast, the direct training paradigm has emerged as the mainstream methodology. By introducing SG techniques (Neftci et al., 2019), this approach effectively circumvents the non-differentiability of the Heaviside step function during spike generation, thereby enabling ultra-low-latency, end-to-end backpropagation, and highly flexible architectural exploration.

Catalyzed by the SG method, the architectural design of SNNs has undergone significant evolution. Early deep SNNs (such as SEW-ResNet (Fang et al., 2021)) mitigated gradient degradation in deep layers by devising spike-element-wise residual connections. Recently, inspired by the tremendous success of Vision Transformers, Spiking Transformers have profoundly reshaped the performance boundaries of SNNs. Pioneering architectures like Spikformer (Zhou et al., 2023) introduced the Spiking Self-Attention (SSA) mechanism, achieving fully spike-driven feature interaction by discarding the computationally expensive Softmax operation. Subsequent advancements (including Spike-driven Transformer (SdT) (Yao et al., 2023) and QKFormer (Zhou et al., 2024b)) further integrated mask operations and 1D Q-K attention mechanisms. These innovations achieved $\mathcal{O}(N)$ linear complexity and extreme energy efficiency, while pushing the Top-1 accuracy of directly trained SNNs on ImageNet-1K beyond 85% for the first time. However, despite these novel architectures relentlessly breaking accuracy records on standard benchmarks, the current evaluation system remains overwhelmingly performance-centric, thereby severely neglecting their representational reliability when processing sensitive demographic features. Consequently, the fairness of SNNs with respect to sensitive attributes has so far been largely overlooked in the existing literature. (Hort et al., 2024)

A.2 Algorithmic Fairness and Shortcut Learning in Computer Vision

As deep learning models are widely deployed in high-stakes, human-centric domains such as facial recognition and biometric analysis, algorithmic fairness has emerged as an urgent research agenda (Parraga et al., 2025). In conventional ANNs, extensive empirical studies have demonstrated that models frequently exhibit systematic performance disparities against specific minority groups. Such algorithmic bias originates not only from skewed training data distributions but, more fundamentally, from the deep models' inherent susceptibility to shortcut learning (Manzoor & Rattani, 2025).

Specifically, in facial analysis tasks, continuous ANNs often display a strong texture and color bias. They tend to capture and amplify spurious correlations, relying excessively on shallow, low-level visual statistics (such as global image contrast, ambient illumination, skin tone intensity, or highly saturated hair colors) rather than extracting robust, race-invariant intrinsic facial topological structures. To quantify and mitigate these biases, the academic community has established core fairness evaluation metrics (such as SP (Dwork et al., 2012) and EO (Hardt et al., 2016)) and actively explored algorithmic interventions, including data resampling (Kamiran & Calders, 2012), adversarial debiasing (Zhang et al., 2018), and feature disentanglement (Locatello et al., 2019).

Nevertheless, these fairness investigations have thus far been predominantly confined to conventional continuous ANNs, while the fairness of SNNs in vision tasks, despite their rapid architectural advances, remains almost entirely unexplored, leaving a critical gap in assessing whether their distinct spike-based processing exacerbates or mitigates representational biases.

B Datasets

To eliminate the confounding influence of chromatic artifacts, we introduce a color-anomaly cleaning procedure into the data preprocessing stage. The motivation is twofold. First, web-collected face datasets such as RFW inevitably contain grayscale photographs, faded scans, and images captured

under improperly white-balanced conditions; these artifacts constitute spurious low-level cues that are orthogonal to the sensitive attributes under investigation. Second, since spiking encoders transform pixel intensities directly into spike trains, SNNs are particularly susceptible to global chroma statistics; anomalous coloration may thus be exploited as a shortcut feature, contaminating both the learned representations and the subsequent fairness evaluation.

Concretely, each RGB image is converted to its HSV representation, where chromatic information is explicitly disentangled from luminance, and the saturation channel is summarized by the per-image mean saturation $\bar{S}$ (OpenCV 0–255 scale). Two complementary criteria are then applied. First, a high-saturation threshold τ_{high} is set to the 95th percentile of the empirical $\bar{S}$ distribution, giving $\tau_{\text{high}} = 148.8$ on RFW; images with $\bar{S} > \tau_{\text{high}}$ are flagged as *high-saturation images*, a pattern frequently induced by strong, unnatural color casts due to white-balance errors or color temperature shifts. This criterion flags $2{,}026$ images (5.0%). Second, *low-saturation images*, corresponding to grayscale, faded, or severely desaturated photographs, are identified by a chroma test on the mean absolute inter-channel difference (threshold 10), which flags $5{,}254$ images (12.9%). We note that a symmetric lower saturation percentile degenerates on RFW ($\tau_{\text{low}} = 0$, since over 5% of the images are near-grayscale with an exactly zero $\bar{S}$), which motivates this chroma-based criterion instead. All flagged samples are excluded from the benchmark: in total $7{,}280$ of $40{,}607$ images (17.9%) are removed, with consistent removal rates across the training and validation splits (17.98% and 17.81%), yielding the refined subset *RFW-Cleaned* with $33{,}327$ images. We emphasize that, as a criterion operating on global chroma statistics, this screening targets extreme chromatic anomalies only; mild color casts whose saturation remains within the normal range are not captured, and detecting such subtler shifts would require hue-domain statistics, which we leave to future work.

The necessity of this cleaning step is empirically corroborated by the qualitative analysis in Figure 4: the Caucasian samples misclassified by QKFormer are dominated by grayscale and color-distorted images—consistent with the two anomaly categories identified by the cleaning procedure—whereas normally saturated color images are mostly classified correctly. This correspondence between the observed failure modes and the per-image chroma statistics corroborates that chroma-based screening serves as an effective proxy for identifying such artifacts.

This cleaning step enhances the overall data quality in two respects. On the one hand, it removes non-representative samples whose chromatic anomalies would otherwise dominate early-stage feature learning and reinforce spurious correlations between color statistics and demographic labels. On the other hand, by ensuring that the retained images share a consistent color regime, it suppresses acquisition-induced variance, so that cross-group performance disparities become more attributable to genuine model behavior rather than data imperfections. Moreover, retaining both the original and the cleaned versions of the dataset enables controlled attribution studies, allowing the contribution of color cues to SNN fairness to be isolated and quantified.

Racial Faces in-the-Wild (RFW) (Wang et al., 2021) introduced by Mei Wang et al., is specifically designed to quantify and analyze racial bias in face recognition systems. Its primary goal is to overcome the lack of racial diversity in traditional test sets such as LFW.

1. **Subset Composition:** RFW consists of four independent test subsets: Caucasian, Indian, Asian, and African. Each subset includes approximately $10,000$ images from around $3,000$ individuals.

2. **Data Collection and Cleaning:** Images are mainly selected from MS-Celeb-1M. Initial race estimation is performed using Freebase nationality attributes and commercial APIs (e.g., Face++), followed by manual cleaning to ensure annotation accuracy. The dataset strictly removes identities overlapping with common training datasets such as CASIA-WebFace.

3. **Original and Adopted Evaluation Protocols:** RFW was developed to evaluate racial disparities in face verification. In our benchmark, we use its facial images and race annotations for four-class race classification, assessing group disparities through the maximum inter-group accuracy gap.

UTKFace (Zhang et al., 2017) is a large-scale facial attribute dataset designed to support research on age estimation, gender classification, and race recognition. Collected and organized by Zhang et al., it contains over $20,000$ facial images spanning a wide age range from 0 to 116 years.

1. **Data Source and Composition:** The images are primarily collected from the Internet (e.g., Flickr), exhibiting substantial variations in pose, illumination, and expression (PIE), as well as varying degrees of occlusion and blur. This makes the dataset representative of real-world, "in-the-wild" conditions.
2. **Annotations:** Each image is labeled with precise age, gender, and race information. The dataset is constructed by integrating multiple sources, including MORPH and CACD, combined with web crawling techniques, addressing the limitation of insufficient large-scale paired samples in earlier datasets.
3. **Applications:** Due to its broad age coverage and rich attribute annotations, UTKFace is widely used for evaluating model robustness in cross-age and cross-race classification tasks.

DemogPairs (Hupont & Fernández, 2019) is a carefully constructed, demographically balanced verification dataset aimed at addressing the limitations of traditional benchmarks (e.g., LFW), where biased data distributions may obscure model biases.

1. **Design Philosophy:** To counter severe gender and racial imbalances in datasets such as VGGFace and CASIA-WebFace, DemogPairs establishes a fully balanced evaluation environment through meticulous sample selection.
2. **Data Distribution:** The dataset contains $10,800$images distributed across six non-overlapping demographic groups: Asian female, Asian male, African female, African male, White female, and White male. Each group includes 100 identities, with 18 images per identity.
3. **Evaluation Capability:** DemogPairs can generate up to 58.3 million verification pairs, enabling fine-grained cross-demographic analysis. For instance, it supports targeted evaluation under conditions such as "same-race different-gender" or "cross-race cross-gender." Experimental results demonstrate that DemogPairs can reveal shortcut learning behaviors, where model accuracy drops significantly for specific groups (e.g., Asian females), phenomena that are often difficult to detect in imbalanced datasets.

FairFace (Karkkainen & Joo, 2021) was proposed by Karkkainen et al. to address the racial imbalance issue in existing public face datasets, such as ImageNet-1K (Deng et al., 2009) and MS-Celeb-1M, where white people account for approximately 80% of the data.

1. **Scale and Source:** FairFace contains $108,501$ images in total, primarily sampled from the YFCC-100M Flickr dataset, supplemented with images from Twitter and online news sources to ensure diversity. Of these, the $97,698$ training and validation images listed in Table 1 are employed in our benchmark.
2. **Race Taxonomy:** Unlike traditional coarse-grained classifications (e.g., White, Black, Asian), FairFace defines seven racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. It is the first large-scale in-the-wild dataset to distinguish between East Asian and Southeast Asian populations while also including Latino as a separate category.
3. **Balanced Design:** Through adaptive sampling across countries, FairFace achieves a high degree of balance in race, gender, and age distributions, effectively mitigating "White bias" and making it an ideal benchmark for fairness evaluation.

Together, these datasets support the evaluation of classification accuracy and group disparities under two task settings. For gender classification on UTKFace, FairFace, and DemogPairs, we jointly report SP and EO using race as the sensitive attribute. For race classification on RFW, we use the maximum accuracy gap across racial groups as an alternative measure of group performance disparity within our fairness evaluation.

C Experimental Protocol

Our experiments are conducted on a server equipped with an Intel Xeon Gold 6455B processor and three NVIDIA L40 GPUs (46 GB memory each). The system features 377 GB of RAM and a

hybrid storage configuration comprising an 894 GB NVMe SSD and a 14.6 TB HDD. The software environment is based on Ubuntu $22.04.5$ LTS.

To ensure meaningful comparisons, we evaluate all SNN models under a strictly unified protocol. All models are trained on identical race-sensitive dataset splits. For UTKFace, FairFace, and DemogPairs, gender is the prediction target, while race is the sensitive attribute used to compute SP and EO. The RFW annotations used in our experiments do not include gender labels, preventing direct application of the same gender-classification-based evaluation protocol. We therefore perform four-class race classification on RFW and use the maximum accuracy gap across its four racial groups as an alternative measure of group performance disparity within our fairness evaluation. We maintain consistent preprocessing, including input resolution, AdamW optimizer, and cosine annealing scheduler. Models are implemented either by adapting official code or through faithful re-implementation.

We fix the inference timestep (T) to 4 across all methods to ensure comparable spike statistics and latency. For methods requiring specific configurations (e.g., TET variants), we report results under their original settings and distinguish them from the standardized setup. This control is vital as timestep variations may impact prediction stability across demographic groups.

We do not introduce external debiasing techniques, such as adversarial objectives or post-processing. Instead, each model is evaluated in its standard predictive form. Under this protocol, we compare group disparities across the evaluated SNN configurations. These comparisons characterize model behavior under the specified datasets and training settings.

The SP and EO disparities for the gender-classification tasks on UTKFace, FairFace, and DemogPairs follow the maximum-pairwise-difference definitions given in Section 2.1. For RFW, we instead use the previously defined Δ_{Acc} to measure disparities in race-classification accuracy across racial groups.

D Results

D.1 Performance Disparity and Demographic Bias

We evaluate the selected SNN configurations on RFW and RFW-Cleaned to examine group disparities in race-classification accuracy. Table 4 and Table 3 report the accuracy for each racial group on the original and cleaned datasets, respectively. For each configuration and dataset variant, we use the previously defined Δ_{Acc} to summarize the maximum accuracy gap across the four racial groups.

Group-wise Accuracy Patterns: On the original RFW dataset, the African group has the highest race-classification accuracy for all reported model configurations. This pattern also holds for most configurations on RFW-Cleaned, but not for STAtten. For STAtten, the highest group accuracy is observed on the African group in RFW (99.13%) and on the Asian group in RFW-Cleaned (98.70%, compared with 97.08% for the African group). These results show that the group with the highest accuracy is not fixed across model configurations and dataset variants.

Group Disparities Across Dataset Variants: The differences between RFW and RFW-Cleaned vary across model configurations and racial groups. For QKFormer, accuracy on the Caucasian group is 32.64% on RFW and 87.04% on RFW-Cleaned, while Δ_{Acc} decreases from 51.04 to 8.22 percentage points. However, smaller group disparities are not observed for every configuration: Δ_{Acc} increases from 7.08 to 16.87 percentage points for Spike-driven Transformer V2 and from 8.65 to 11.83 percentage points for V3. Thus, the cleaned dataset is associated with smaller group disparities for some configurations, but this pattern does not hold uniformly across the evaluated models.

D.2 Convergence Dynamics and Fairness Evolution

To further understand how fairness emerges during training, we analyze the test accuracy trajectories over the initial 10 epochs (Table 5 and Table 6).

Asymmetric Learning Trajectories: We observe pronounced differences in group accuracy during early training. For Spike-ResNet19 with TET on the original RFW dataset, the African group already

Table 3: Group-wise race-classification accuracy (%) on RFW-Cleaned. Results are reported as mean$^{\pm \text{std}}$ over multiple runs with different random seeds.

Method	Asian	African	Caucasians	Indians
QKFormer	$91.57^{\pm 0.11}$	$95.26^{\pm 0.17}$	$87.04^{\pm 0.12}$	$87.65^{\pm 0.11}$
SEW-ResNet	$92.87^{\pm 0.21}$	$96.07^{\pm 0.25}$	$91.46^{\pm 0.18}$	$86.68^{\pm 0.15}$
Spike-driven Transformer	$97.64^{\pm 0.08}$	$98.99^{\pm 0.07}$	$97.95^{\pm 0.10}$	$96.78^{\pm 0.11}$
Spike-driven Transformer V2	$82.24^{\pm 0.14}$	$93.79^{\pm 0.17}$	$76.92^{\pm 0.14}$	$79.27^{\pm 0.02}$
SNN-ResNet18	$91.99^{\pm 0.06}$	$95.13^{\pm 0.10}$	$90.06^{\pm 0.03}$	$86.23^{\pm 0.06}$
VGGSNN(TET = True)	$94.99^{\pm 0.15}$	$98.28^{\pm 0.17}$	$90.49^{\pm 0.20}$	$90.45^{\pm 0.13}$
VGGSNN(TET = False)	$94.69^{\pm 0.33}$	$96.87^{\pm 0.20}$	$88.60^{\pm 0.07}$	$89.88^{\pm 0.15}$
Spike-ResNet19(TET = True)	$93.85^{\pm 0.10}$	$97.88^{\pm 0.11}$	$89.59^{\pm 0.05}$	$88.86^{\pm 0.07}$
Spike-ResNet19(TET = False)	$94.10^{\pm 0.20}$	$97.93^{\pm 0.05}$	$90.39^{\pm 0.06}$	$89.92^{\pm 0.06}$
Spike-driven Transformer V3	$85.72^{\pm 0.04}$	$92.16^{\pm 0.21}$	$82.46^{\pm 0.25}$	$80.33^{\pm 0.09}$
STAtten	$98.70^{\pm 0.10}$	$97.08^{\pm 0.11}$	$97.50^{\pm 0.06}$	$97.00^{\pm 0.10}$

Table 4: Group-wise race-classification accuracy (%) on RFW. Results are reported as mean$^{\pm \text{std}}$ over multiple runs with different random seeds.

Method	Asian	African	Caucasians	Indians
QKFormer	$76.60^{\pm 0.15}$	$83.68^{\pm 0.12}$	$32.64^{\pm 0.10}$	$56.27^{\pm 0.15}$
SEW-ResNet	$89.10^{\pm 0.15}$	$96.18^{\pm 0.09}$	$82.73^{\pm 0.10}$	$84.32^{\pm 0.15}$
Spike-driven Transformer	$97.55^{\pm 0.22}$	$98.96^{\pm 0.16}$	$97.39^{\pm 0.20}$	$96.31^{\pm 0.14}$
Spike-driven Transformer V2	$93.65^{\pm 0.15}$	$98.10^{\pm 0.12}$	$91.83^{\pm 0.17}$	$91.02^{\pm 0.11}$
SNN-ResNet18	$86.99^{\pm 0.22}$	$95.88^{\pm 0.11}$	$84.16^{\pm 0.15}$	$83.75^{\pm 0.14}$
VGGSNN(TET = True)	$93.13^{\pm 0.02}$	$98.07^{\pm 0.04}$	$89.98^{\pm 0.06}$	$91.08^{\pm 0.15}$
VGGSNN(TET = False)	$92.34^{\pm 0.06}$	$97.67^{\pm 0.06}$	$88.88^{\pm 0.11}$	$89.35^{\pm 0.19}$
Spike-ResNet19(TET = True)	$91.79^{\pm 0.18}$	$97.92^{\pm 0.07}$	$89.14^{\pm 0.09}$	$89.80^{\pm 0.15}$
Spike-ResNet19(TET = False)	$92.62^{\pm 0.13}$	$97.14^{\pm 0.25}$	$89.83^{\pm 0.16}$	$89.57^{\pm 0.01}$
Spike-driven Transformer V3	$91.69^{\pm 0.23}$	$97.27^{\pm 0.10}$	$89.63^{\pm 0.12}$	$88.62^{\pm 0.11}$
STAtten	$97.81^{\pm 0.16}$	$99.13^{\pm 0.14}$	$95.24^{\pm 0.10}$	$96.03^{\pm 0.20}$

achieves a relatively high accuracy of 79.85% after the first complete training epoch (E_0), whereas Indian-group accuracy is only 4.50% at the same stage (Table 6).

Stabilization via TET: The integration of Temporal Efficient Training (TET) serves as a regularizer that stabilizes the learning trajectory. Models trained with TET (e.g., VGGSNN with TET=True) demonstrate lower inter-epoch variance and a more synchronized accuracy climb across all four ethnic groups compared to their non-TET counterparts.

Early Asymmetry in Learning Trajectories: Here, E_0 denotes evaluation after the first complete training epoch, consistent with the main text. Pronounced differences in group-wise race-classification accuracy are already observed at this checkpoint. Together with the subsequent accuracy trajectories, these results illustrate the Asymmetric Learning Trajectories discussed in the main text and may reflect differences in representation learning across racial groups.

Table 5: Group-wise race-classification test accuracy (%) over the first ten training epochs on RFW-Cleaned. Epoch indices start at 0; E_0 reports evaluation after the first complete training epoch, and E_9 after the tenth.

Model & Setting	Demographic	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9
Spike-ResNet19 (TET=True)	**African**	92.89	93.69	98.54	91.17	94.84	94.22	96.47	90.56	95.94	91.44
	Caucasian	41.13	52.73	23.69	62.54	74.58	63.68	81.37	84.69	77.80	71.66
	Indian	24.34	42.78	69.67	78.52	80.99	84.95	74.08	77.82	87.19	87.90
	Asian	54.17	74.95	81.10	86.71	75.10	85.02	86.16	86.81	82.59	90.87
Spike-ResNet19 (noTET)	**African**	94.09	94.35	96.78	95.81	93.69	95.06	94.57	91.57	94.62	91.75
	Caucasian	33.99	47.77	39.84	76.66	70.81	58.57	81.22	73.89	75.87	83.50
	Indian	37.68	53.21	65.32	67.52	78.83	83.85	77.38	88.51	89.92	83.10
	Asian	32.59	76.39	86.61	83.28	86.01	87.75	87.50	84.87	85.42	91.52
VGGSNN (TET=True)	**African**	93.60	96.20	94.88	94.62	97.97	97.09	95.37	98.23	96.78	95.98
	Caucasian	52.08	69.72	79.09	83.85	75.57	83.70	91.18	83.89	87.86	88.40
	Indian	77.42	68.18	80.41	76.85	76.76	74.08	82.13	84.68	80.81	84.55
	Asian	66.12	86.76	84.72	89.14	94.49	94.00	90.58	93.06	94.99	94.89
VGGSNN (noTET)	**African**	87.38	90.78	94.53	94.31	96.07	96.29	93.20	96.47	95.32	96.65
	Caucasian	46.73	80.57	81.22	83.94	69.18	81.02	86.27	83.20	85.58	90.58
	Indian	78.21	64.61	73.11	76.85	82.04	76.89	87.28	87.72	84.60	78.48
	Asian	74.26	85.12	87.30	84.87	93.95	94.94	89.73	91.22	94.59	94.10

Table 6: Group-wise race-classification test accuracy (%) over the first ten training epochs on the original RFW dataset. Epoch indices start at 0; E_0 reports evaluation after the first complete training epoch, and E_9 after the tenth.

Model & Setting	Demographic	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9
Spike-ResNet19 (TET=True)	**African**	79.85	89.01	96.13	94.17	93.09	90.62	96.81	96.92	96.35	97.82
	Caucasian	47.16	49.45	67.86	81.34	76.64	85.28	81.05	75.84	79.37	84.58
	Indian	4.50	61.24	67.21	58.41	72.95	77.56	70.04	74.76	84.21	66.72
	Asian	66.15	74.60	72.42	80.12	81.43	71.39	86.11	88.97	85.08	91.15
Spike-ResNet19 (noTET)	**African**	78.74	91.73	96.53	91.23	92.13	86.36	94.88	95.31	97.39	95.67
	Caucasian	66.62	38.37	55.72	75.15	76.17	83.05	79.99	60.02	78.24	87.21
	Indian	10.80	70.65	73.14	74.20	73.18	76.65	72.91	84.13	84.10	77.94
	Asian	47.22	77.18	68.97	77.26	82.22	78.41	84.37	90.28	85.24	89.21
VGGSNN (TET=True)	**African**	90.59	97.21	94.45	96.85	95.38	97.57	95.78	96.56	97.32	96.35
	Caucasian	71.36	73.07	83.05	68.11	87.06	82.91	83.38	77.44	87.68	88.92
	Indian	67.09	71.29	76.88	86.06	76.28	83.45	87.57	92.07	85.80	84.85
	Asian	65.87	80.83	78.17	85.08	90.28	90.63	90.60	90.16	91.43	91.35
VGGSNN (noTET)	**African**	93.27	97.96	96.81	96.13	95.10	96.49	96.89	97.75	95.24	96.21
	Caucasian	66.51	61.30	81.45	80.79	87.57	85.35	81.71	73.91	91.87	77.84
	Indian	68.79	70.08	77.67	78.58	67.96	78.54	85.27	92.63	77.11	88.70
	Asian	66.55	81.03	73.29	82.98	88.21	88.93	90.83	86.75	91.31	92.18