ProtoFair: Fair Self-Supervised Contrastive Learning via Pseudo-Counterfactual Pairs

Marah Halawa and Olaf Hellwich

Technische Universität Berlin
halawa@campus.tu-berlin.de, olaf.hellwich@tu-berlin.de

Abstract. Self-supervised learning methods learn high-quality visual representations, yet recent studies show that these representations often capture demographic biases present in the training data. Existing fairness-aware methods address this by redesigning the self-supervised objective itself, limiting portability across the rapidly evolving landscape of self-supervised learning (SSL) frameworks. We propose ProtoFair, a fairness-aware contrastive loss designed to work alongside existing SSL objectives without modifying them. ProtoFair leverages unsupervised prototype clustering to identify pseudo-counterfactual pairs: samples sharing the same cluster assignment but belonging to different sensitive groups. By pulling these content-matched, cross-group samples together in the embedding space, ProtoFair encourages the encoder to learn representations that are invariant to the sensitive attribute. The method requires only sensitive attribute annotations, no target labels, and integrates seamlessly with both SimCLR and SupCon. Experiments on CelebA and UTKFace demonstrate consistent fairness improvements while maintaining competitive accuracy.

1 Introduction

Self-supervised contrastive learning has emerged as a leading approach for learning visual representations without requiring manual annotation. Methods such as SimCLR [4], MoCo [9], and SupCon [14] have demonstrated that contrastive objectives can produce representations that compete with their supervised counterparts across a range of downstream tasks. However, recent work has revealed that self-supervised representations are not immune to societal biases. Models trained without explicit labels still encode demographic information present in the training data [23], [24] leading to unfair outcomes in downstream applications such as facial attribute classification.

Addressing fairness in representation learning has attracted growing attention. Two prominent lines of work have emerged. The first focuses on adversarial debiasing methods [15], which trains an auxiliary discriminator to remove sensitive information from representations. The second develops modified contrastive objectives [21,27], which redesign the loss function to reduce bias. While effective, both strategies share a common limitation: they require replacing or fundamentally altering the base training objective. Adversarial methods introduce minimax optimization known for its instability, while fairness-aware contrastive losses

such as FSCL [21] modify the negative sampling strategy and require target task labels. As self-supervised learning continues to evolve rapidly, with new methods regularly achieving state-of-the-art performance, fairness approaches tightly coupled to a specific loss function must be redesigned for each new method.

In this paper, we ask a different question: rather than redesigning SSL objectives for fairness, can we introduce a lightweight regularizer that makes *other* existing SSL methods fairer? We propose **ProtoFair Loss**, a fairness-aware contrastive loss added to the base SSL loss as an auxiliary term, leaving the original objective entirely unchanged. ProtoFair identifies pseudo-counterfactual pairs, which are samples that share similar semantic content but belong to different demographic groups, and encourages their representations to be similar. The key conceptual motivation comes from counterfactual fairness [16]: a representation is fair if it would remain unchanged were an individual's sensitive attribute different. By pulling such pairs together, ProtoFair actively encourages the learned representation to reflect content rather than demographic membership.

The main contributions of this paper are summarized as follows

1. A **plug-in fairness-aware contrastive loss** that can be added to other SSL objective without modifying them, requiring only sensitive attribute annotations, not target task labels.
2. A **pseudo-counterfactual pair construction strategy**, where samples from different sensitive groups assigned to the same cluster are treated as positives in a contrastive objective, actively encouraging group-invariant representations. To the best of our knowledge, *this is the first work to* (i) employ unsupervised clustering as a mechanism for fairness in contrastive learning and (ii) construct pseudo-counterfactual pairs through cluster assignments.
3. **Empirical validation** on CelebA [18], UTKFace [28], and NIH Chest X-rays [25] showing improved equalized odds [8] across multiple base SSL methods while maintaining competitive accuracy.

2 Related Work

2.1 Self-Supervised Learning

Self-supervised learning (SSL) has appeared as a powerful paradigm for learning visual representations without requiring annotation. Among the notable methods in this domain is contrastive learning, which learns representations by pulling similar (positive) pairs together while pushing dissimilar (negative) pairs apart in the embedding space. SimCLR [4] established a framework for self-supervised contrastive learning by treating two augmented views of the same image as positives and all other samples in the batch as negatives. Leveraging this concept, Khosla et al. [14] generalized the contrastive loss to the supervised setting with SupCon, which leverages label information to treat all samples sharing the same class as positives, achieving superior performance over the standard cross-entropy loss on several benchmarks. One significant challenge in contrastive

learning is obtaining sufficient negative samples for effective training. He et al. [9] addressed this through Momentum Contrast (MoCo), which maintains a dynamic queue of representations encoded by a momentum-updated encoder, decoupling dictionary size from mini-batch size and enabling a large set of negatives, without needing enormous batches. The momentum encoder and queue mechanism introduced by MoCo have become influential design patterns in SSL. In our work, we draw inspiration from MoCo's cross-batch queue to construct our feature queue, which stores representations, cluster assignments, and sensitive attribute labels from past batches, enabling discovery of pseudo-counterfactual pairs beyond the current mini-batch.

While contrastive methods focus on distinguishing individual instances, clustering based approaches aim to capture higher-level semantic structure in self-supervised representation learning. DeepCluster [1] pioneered this direction by alternating between clustering features with K-Means and using the resulting cluster assignments as pseudo-labels for discriminative training. However, this offline alternation can be computationally expensive. SwAV [2] addressed this by using an online clustering approach that enforces consistency between cluster assignments of different augmented views of the same image. This eliminates the need for explicit pairwise comparisons and enabling scalable training. Prototypical Contrastive Learning (PCL) [17] further advanced clustering-based SSL by introducing momentum-updated prototypes as latent cluster centroids, using an expectation-maximization framework where prototypes concentrate representations around semantically meaningful modes. Our ProtoFair method builds directly on the momentum prototype mechanism of PCL: we maintain a set of K prototypes updated via exponential moving average (EMA) that define cluster assignments over feature representations. However, while PCL and previously mentioned methods use clustering exclusively for improving representation quality, we repurpose the cluster structure for a fundamentally different goal, namely fairness. While clustering-based pseudo-labels have been widely adopted in SSL [2, 17], their use for constructing fairness-aware contrastive objectives remains largely unexplored.

Despite the success of self-supervised methods in representation learning for multiple downstream tasks, recent studies have shown that these methods can encode societal biases present in the training data. Steed and Caliskan [24] demonstrated that the examined self-supervised models encode demographic biases. Although the representations are learned without explicit labels, they continue to capture misleading correlations between visual features and sensitive attributes. Additionally, Sirotkin et al. [23] examined three types of self-supervised learning models, including contrastive, geometric, and clustering-based, and found that contrastive models tend to inherit more biases than other SSL approaches. This motivates the need for fairness-aware self-supervised methods.

2.2 Fairness in Representation Learning

Fairness in machine learning has been extensively investigated, particularly within supervised learning settings. Hardt et al. [8] formalized the notion of equalized

odds, requiring that a classifier's true positive and false positive rates be equal across demographic groups. This approach has emerged as a widely accepted fairness standard. Accordingly, we use equalized odds as our main evaluation metric to assess fairness in downstream classification tasks.

Early approaches to fair representation learning relied on adversarial debiasing. Kim et al. [15] proposed "Learning Not to Learn" which employs an adversarial training objective that penalizes the encoder for producing representations from which sensitive attributes can be predicted. Despite their effectiveness, training adversarial methods often faces instability, and require careful hyperparameter tuning to balance the minimax objectives [15]. Moreover, adversarial approaches typically operate as modifications to the training procedure itself, replacing or fundamentally altering the base loss rather than complementing it.

In the context of face recognition and facial attribute classification, several works have addressed fairness from different angles. Karkkainen and Joo [13] introduced the FairFace dataset and highlighted the racial biases present in existing face recognition systems. Park et al. [21] proposed FSCL (Fair Supervised Contrastive Learning), which achieves fairness by modifying the selection of negative samples in the supervised contrastive loss so that sensitive attribute information is not exploited. While FSCL represents an important step toward fairness-aware contrastive learning, it operates by constraining what the model learns *not* to distinguish, rather than actively encouraging group-invariant representations for similar content. Furthermore, FSCL requires target task labels to define positive pairs within the supervised contrastive framework. Our ProtoFair loss is complementary to FSCL and can be combined with it. FSCL provides fairness through restriction to the denominator of supervised contrastive, while our loss provides fairness by explicitly pulling cross-group pairs together.

Several other works have tackled fairness in facial attribute classification. Chiu et al. [5] proposed a fair multi-exit framework which uses intermediate classifiers to produce accurate yet less biased predictions. Park et al. [20] introduced a disentangled representation approach for fair facial attribute classification via fairness-aware information alignment, which separates sensitive attribute information from task-relevant features. In other work [6] they explored vision-language driven image augmentation to improve fairness by generating augmented training samples that balance demographic representation. In the broader fairness-aware contrastive learning literature, Chai and Wang [3] investigated self-supervised fair representation learning and demonstrated that standard self-supervised objectives do not inherently provide fairness guarantees, when trained alone without access to demographic labels. Zhang et al. [27] proposed fairness-aware contrastive learning methods that modify the contrastive objective to reduce bias, especially when sensitive attribute labels are only partially available. Primarily through re-weighting and augmentation strategies tied to sensitive attributes. A critical distinction of our approach is its role as a *plug-in regularizer*. Existing fairness methods for contrastive learning typically replace the base loss (e.g., adversarial debiasing [15]), require target labels (e.g., FSCL [21]). In contrast, `ProtoFair` is an additive regularization term that can be ap-

pended to any self-supervised objective without modifying it. This plug-and-play design preserves representation quality while improving fairness, requiring only sensitive attribute annotations rather than target task labels, and is thus suitable for fully self-supervised settings.

Our method draws conceptual motivation from the counterfactual fairness framework of Kusner et al. [16], which defines a decision as fair if it would remain unchanged were an individual's sensitive attribute different, relying on structural causal models to reason about interventions while holding causally upstream variables fixed. We implement this intuition by constructing *pseudo-counterfactual pairs*: samples from different sensitive groups assigned to the same unsupervised cluster, which serves as a proxy for shared semantic content. Unlike true counterfactuals, our approach does not require knowledge of the causal graph; instead, it approximates the counterfactual condition under the assumption that cluster assignments capture task-relevant, non-sensitive factors. While this assumption is weaker than full causal identification, our empirical results demonstrate it is sufficient to yield substantial fairness improvements in practice.

3 Methodology

We propose **ProtoFair Loss**, a fairness-aware contrastive regularizer that can be incorporated into existing self-supervised learning frameworks to promote representations invariant to sensitive attributes. This ensures samples with similar semantic content are embedded close to one another, regardless of their demographic group membership. The core idea is to leverage unsupervised cluster assignments as pseudo-labels for semantic content, enabling the construction of *pseudo-counterfactual pairs*, i.e., samples that share similar content but differ in sensitive group membership, without requiring task-specific labels. Only sensitive attribute annotations are required. Concretely, ProtoFair Loss pulls together representations of samples that are semantically similar (assigned to the same cluster) but belong to different demographic groups, encouraging the model to learn features invariant to the sensitive attribute within each content cluster.

3.1 Problem Formulation

Let f_θ denote an encoder network parameterized by θ, and let g_ϕ denote a projection head that maps encoder outputs to a normalized embedding space. For an input sample x_i, we obtain the L2-normalized representation $z_i = g_\phi(f_\theta(x_i)) \in \mathbb{R}^d$. Each sample is associated with a sensitive attribute $s_i \in \{0, 1\}$ (e.g., gender), which is assumed to be known during training.

Existing self-supervised methods such as SimCLR [4], SupCon [14] have demonstrated strong representation learning capabilities. However, incorporating fairness typically requires redesigning the contrastive objective itself, limiting portability across SSL frameworks and often necessitating costly retraining from scratch. We take a different approach: rather than replacing the SSL objective,

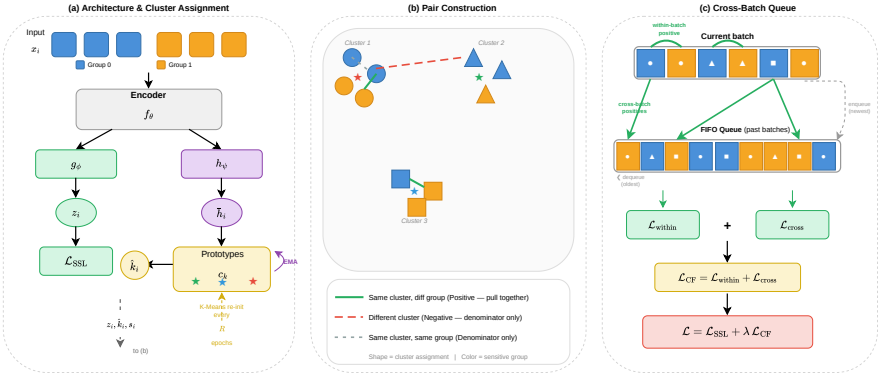


Fig. 1: Illustration of the key steps in the ProtoFair loss. **(a)** A shared encoder f_θ produces representations projected by two heads: a contrastive head g_ϕ (*green*) for the base SSL loss and a cluster head h_ψ (*purple*) for computing cluster assignments via momentum-updated prototypes (*stars*). Prototypes are initialized with K-Means and tracked between re-initializations using exponential moving average updates. **(b)** Pseudo-counterfactual pair construction in the embedding space. Samples are colored by sensitive group and shaped by cluster assignment. Pairs sharing the same cluster but from different groups (green solid lines) are positives, whereas different-cluster pairs (red dashed lines) and same-cluster same-group pairs (gray dotted lines) are non-positives, appear only in the denominator. **(c)** Within-batch positives are identified in the current mini-batch, while cross-batch positives are discovered by matching against a FIFO queue of past batch representations. The two components form the ProtoFair regularizer $\mathcal{L}_{\text{CF}}$, which is added to the base SSL loss $\mathcal{L}_{\text{SSL}}$.

we introduce ProtoFair Loss as an auxiliary regularizer $\mathcal{L}_{\text{CF}}$ that is added to an existing SSL loss $\mathcal{L}_{\text{SSL}}$:

$$\mathcal{L} = \mathcal{L}_{\text{SSL}} + \lambda \mathcal{L}_{\text{CF}} \tag{1}$$

where $\lambda > 0$ controls the strength of the fairness regularization. The base loss $\mathcal{L}_{\text{SSL}}$ can be any self-supervised objective (e.g., SimCLR, SupCon) and remains *unchanged*. This plug-in design offers two practical advantages: (i) it avoids modifying existing SSL methods that already produce high-quality representations, and (ii) it enables fairness regularization to benefit from ongoing advances in SSL without requiring method-specific adaptations. Beyond the projection head g_ϕ used by the base SSL loss, our method introduces a separate *cluster projection head* $h_\psi : \mathbb{R}^m \to \mathbb{R}^d$ that maps encoder representations to a clustering embedding space. This head operates on the same encoder output as g_ϕ but is updated only through the ProtoFair loss, decoupling the base SSL and the fairness objectives. The cluster head output is used exclusively by the ProtoFair regularizer for computing cluster assignments, as described in the following subsection.

3.2 ProtoFair: Cluster-Fair Contrastive Loss

Momentum-Updated Cluster Prototypes To construct pseudo counterfactual pairs without target labels, we require a notion of semantic content similarity. We obtain this through a set of K prototype vectors $\{c_k\}_{k=1}^{K} \subset \mathbb{R}^d$, maintained as non-learnable running estimates in the cluster embedding space produced by h_ψ. The prototypes are not updated via backpropagation; instead, they are maintained through K-Means initialization and momentum-based tracking, as described below.

Initialization. After a warmup period of several epochs, during which the encoder is trained using only the base SSL loss, the prototypes are initialized by performing K-means clustering over the full set of training representations in cosine space:

$$\{c_k\}_{k=1}^{K} \leftarrow \texttt{K-Means}\Big(\big\{\bar{h}_i = \tfrac{h_\psi(f_\theta(x_i))}{\|h_\psi(f_\theta(x_i))\|}\big\}_{i=1}^{N}\Big). \tag{2}$$

To prevent prototype drift as the representation space evolves during training, K-Means re-initialization is performed periodically every R epochs over the full training set.

Momentum update. Between re-initializations, prototypes are updated every training iteration via exponential moving average (EMA) to smoothly track the evolving feature space. Given a mini-batch of cluster-head features $\{\bar{h}_i\}_{i=1}^{B}$ with hard assignments $\hat{k}_i = \arg\max_k \bar{h}_i^\top c_k$, the update for prototype c_k is:

$$c_k \leftarrow \text{normalize}\Big(m \cdot c_k + (1-m) \cdot \frac{\sum_{i:\hat{k}_i=k} \bar{h}_i}{|\{i : \hat{k}_i = k\}|}\Big), \tag{3}$$

where $m \in [0, 1)$ is the momentum coefficient. This continuous tracking ensures that cluster assignments remain meaningful as the encoder improves, without the cost of running full-dataset K-Means at every iteration.

Cluster assignments. Each sample is assigned to its nearest prototype in the cluster embedding space based on cosine similarity:

$$\hat{k}_i = \arg\max_k \ \bar{h}_i^\top c_k \tag{4}$$

These hard assignments serve as pseudo-content labels for constructing counterfactual pairs, as described next.

Gradient flow. The cluster assignments are detached from the computational graph when passed to the ProtoFair loss. This design follows an EM-style alternating optimisation: the cluster assignments act as a fixed E-step, providing pseudo-labels that define which pairs serve as positives, while the fairness contrastive update acts as the M-step, optimising the encoder and projection head,

mirroring PCL [17]. Detachment is essential because, without it, gradients from $\mathcal{L}$CF could manipulate the cluster assignments themselves rather than improving the representations. In the degenerate case, the model could collapse all samples into a single cluster, trivially satisfying the same-cluster criterion without producing fairer representations. By treating cluster assignments as fixed pseudo-labels, the clustering module reflects the underlying content similarity and is neither rewarded nor penalised by the fairness objective. The fairness loss $\mathcal{L}$CF therefore updates only the encoder f_θ and the contrastive projection head g_ϕ through the feature similarity computation in its contrastive objective. The cluster head h_ψ is not directly updated by $\mathcal{L}$CF; it evolves indirectly as the shared encoder $f\theta$ is updated by both $\mathcal{L}$SSL and $\mathcal{L}$CF, as defined in Eq. (1).

Pseudo-Counterfactual Pair Construction The central idea of ProtoFair is to identify pairs of samples that share semantic content but differ in their sensitive attribute, approximating counterfactual pairs without requiring target labels. Given a mini-batch of B samples with features $\{z_i\}_{i=1}^B$, hard cluster assignments $\{\hat{k}_i\}_{i=1}^B$, and sensitive attributes $\{s_i\}_{i=1}^B$, we define the *pseudo-counterfactual positive set* for sample i as:

$$\mathcal{P}_i = \left\{ j \neq i \;\middle|\; \hat{k}_j = \hat{k}_i \;\text{ and }\; s_j \neq s_i \right\} \tag{5}$$

That is, sample j is a positive for sample i if and only if they belong to the same cluster (proxy for shared content) and to different sensitive groups. This yields three categories of pairs:

- **Same cluster, different sensitive group** ($\hat{k}_j = \hat{k}_i,\ s_j \neq s_i$): pseudo-counterfactual pairs (*pulled together*).
- **Different cluster** ($\hat{k}_j \neq \hat{k}_i$): semantically dissimilar pairs (*pushed apart*).
- **Same cluster, same sensitive group** ($\hat{k}_j = \hat{k}_i,\ s_j = s_i$): same-content, same-group pairs (*pushed apart*).

The third category is key: by excluding same-group pairs from the numerator even when they share the same cluster, the loss specifically targets cross-group alignment rather than general within-cluster similarity.

Within-Batch Contrastive Fairness Loss Given the positive set $\mathcal{P}_i$ defined above, we formulate the within-batch fairness loss following the contrastive learning framework. For each sample i with $|\mathcal{P}_i| > 0$, the loss is:

$$\mathcal{L}_{\text{within}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log \frac{\exp\big(\text{sim}(z_i, z_j)/\tau\big)}{\sum_{\substack{k=1 \\ k \neq i}}^{B} \exp\big(\text{sim}(z_i, z_k)/\tau\big)} \tag{6}$$

where $\text{sim}(z_i, z_j) = z_i^\top z_j$ is the cosine similarity between L2-normalized features, τ is the temperature, and $\mathcal{V} = \{i : |\mathcal{P}_i| > 0\}$ is the set of samples that have at least one pseudo-counterfactual partner in the batch. The denominator sums over *all* non-self pairs in the batch, following the standard contrastive denominator.

Cross-Batch Queue for Expanded Pair Discovery A practical limitation of the within-batch loss is that pseudo-counterfactual pairs require samples from the same cluster but different sensitive groups to co-occur in a mini-batch. When batches are small or sensitive groups are imbalanced, such pairs may be scarce, weakening the fairness signal. To address this, we maintain a first-in-first-out (FIFO) queue $\mathcal{Q}$ that stores representations from the M most recent batches. For each past sample, the queue retains a tuple of its feature vector, cluster assignment, and sensitive attribute: $\mathcal{Q} = \{(z_j^q,\, \hat{k}_j^q,\, s_j^q)\}_{j=1}^{Q}$, where $Q = M \times B$ is the total queue capacity. Queue entries are stored as detached tensors and are not involved in gradient computation.

At each training step, current-batch samples are matched against the queue to find additional cross-group pairs. The positive set for sample i with respect to the queue is:

$$\mathcal{P}_i^q = \left\{\, j \in \mathcal{Q} \;\middle|\; \hat{k}_j^q = \hat{k}_i \;\text{ and }\; s_j^q \neq s_i \,\right\} \tag{7}$$

and the cross-batch loss follows the same contrastive formulation:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{|\mathcal{V}^q|} \sum_{i \in \mathcal{V}^q} \frac{1}{|\mathcal{P}_i^q|} \sum_{j \in \mathcal{P}_i^q} \log \frac{\exp\big(\text{sim}(z_i, z_j^q)/\tau\big)}{\sum_{k \in \mathcal{Q}} \exp\big(\text{sim}(z_i, z_k^q)/\tau\big)} \tag{8}$$

where $\mathcal{V}^q = \{i : |\mathcal{P}_i^q| > 0\}$. Note that the denominator sums over all queue entries, providing a large and diverse set of negatives. After each forward pass, the queue is updated by enqueuing the current batch and dequeuing the oldest entries.

Full Loss Formulation The complete ProtoFair regularizer combines the within-batch and cross-batch components:

$$\mathcal{L}_{\text{CF}} = \mathcal{L}_{\text{within}} + \mathcal{L}_{\text{cross}} \tag{9}$$

This is combined with the base SSL objective as defined in Eq. (1):

$$\mathcal{L} = \mathcal{L}_{\text{SSL}} + \lambda\, \mathcal{L}_{\text{CF}}$$

where $\lambda > 0$ determines the magnitude of the fairness regularization effect. The ProtoFair regularizer is not applied from the beginning of training. The model first trains for several warmup epochs using only the base SSL loss $\mathcal{L}_{\text{SSL}}$, allowing the encoder to learn meaningful representations before prototypes are initialized and the fairness loss is activated. The queue is initially empty and populates naturally over time. The cross-batch loss contributes zero until the queue contains entries.

4 Experiments

4.1 Datasets

We evaluate ProtoFair on two widely-used facial-attribute benchmarks (CelebA, UTKFace) and one medical-imaging benchmark (NIH Chest X-rays).

CelebA [18] contains over 200K face images annotated with 40 binary attributes. We use the standard train/val/test partition and follow [21] for target and sensitive attribute notation in all table results.

UTKFace [28] consists of over 20K face images labeled with age, gender, and ethnicity. Following [21], we binarize ethnicity (Caucasian vs. non-Caucasian). We construct training sets with varying imbalance ratios $\alpha \in \{2, 3, 4\}$, where α controls the skew between demographic groups. Val/test sets are kept balanced.

NIH Chest X-rays [25] is a large-scale medical imaging benchmark of 112,120 chest radiographs from 30,805 patients, each annotated with up to 14 thoracic pathologies.

4.2 Implementation Details

Encoder training, and downstream evaluation. We use ResNet-18 [10] as the backbone encoder with two separate MLP projection heads: a contrastive head for the base SSL loss, and a cluster head for prototype computation. Training uses SGD with momentum 0.9, weight decay 10^{-4}, and an initial learning rate of 0.1 with cosine annealing. Following the linear evaluation protocol of [21], we freeze the trained encoder and train a linear classifier on top of the encoder features using cross-entropy loss.

Evaluation metrics. Following [21], we report classification accuracy (ACC) on the target attribute and measure the degree of equalized odds (EO) [8] across sensitive groups, where lower EO corresponds to fairer predictive outcomes.

4.3 Evaluation Results

Table 1 presents classification results on CelebA across multiple target and sensitive attribute scenarios. To enable direct comparison with existing approaches, we adopt the evaluation protocol and baseline results in [21] and apply ProtoFair on top of SupCon under the same setup. Among the baselines, CE [11] shows the highest EO violations, confirming that unconstrained training encodes sensitive attribute biases. Prior fairness methods such as FSCL [21] and MFD [12] improve EO, with varying fairness-accuracy trade-offs.

For all SupCon + ProtoFair experiments in Table 1, we use 10 clusters whose prototypes are initialized via K-Means, with λ=0.3. For the *Bags Under Eyes* and *Big Nose* targets, we apply a 100-epoch warmup followed by 5 additional epochs of ProtoFair training. For the *Attractiveness* target with *Male* as the sensitive attribute, we keep the 100 epoch warmup but train with ProtoFair for 10 epochs instead of 5. For the *Attractiveness* target with *Young* as the sensitive attribute, we start SupCon + ProtoFair optimization at epoch 10 and continue for 90 epochs. We evaluate the model by freezing the encoder and training a linear classifier on top. ProtoFair achieves substantial fairness improvements. On (T:b, S:y), EO drops from 16.9 to 6.9 with under one point of accuracy loss. Similarly, On (T:a, S:m), EO drops from 30.5 to 14.2 with accuracy nearly unchanged. On (T:e, S:y), ProtoFair achieves an EO of 1.9, matching the best result of the

Table 1: Classification results on CelebA. We measure classification accuracy (ACC) and equalized odds (EO) in various target (T)–sensitive (S) attribute scenarios. Lower EO is better. Attributes: a=attractiveness, b=big nose, e=bags under eyes, m=male, y=young.

	T:a, S:m		T:a, S:y		T:b, S:m		T:b, S:y		T:e, S:m		T:e, S:y	
Method	ACC	EO	ACC	EO	ACC	EO	ACC	EO	ACC	EO	ACC	EO
CE [11]	79.6	27.8	79.8	16.8	84.0	17.6	84.5	14.7	83.9	15.0	83.8	12.7
GRL [22]	77.2	24.9	74.6	14.7	82.5	14.0	83.3	10.0	81.9	6.7	82.3	5.9
LNL [15]	79.9	21.8	74.3	13.7	82.3	10.7	82.3	6.8	81.6	5.0	80.3	3.3
FD-VAE [20]	76.9	15.1	77.5	14.8	81.6	11.2	81.7	6.7	82.6	5.7	84.0	6.2
MFD [12]	78.0	7.4	80.0	14.9	78.0	7.3	78.0	5.4	79.0	8.7	78.0	5.2
FSCL [21]	79.1	11.5	79.1	13.0	82.1	7.0	83.8	6.4	82.7	3.8	82.0	1.8
SupCon [14]	80.5	30.5	80.1	21.7	84.6	20.7	84.4	16.9	84.3	20.8	84.0	10.8
SupCon + ProtoFair (Ours)	80.3	14.2	81.5	17.7	84.2	15.8	83.5	6.9	84.0	10.1	83.3	1.9

Table 2: CelebA results without target labels during representation learning, using SimCLR as the base loss. Only sensitive attributes are used. ACC: classification accuracy, EO: equalized odds (lower is fairer). a=attractiveness, m=male.

	T:a, S:m	
Method	ACC ↑	EO ↓
FSCL* [21]	74.6	14.8
SimCLR [4] + GRL [22]	72.3	21.9
SimCLR [4]	75.7	29.4
SimCLR + ProtoFair (Ours)	73.3	21.9

dedicated fairness method FSCL (1.8) while retaining 1.3 points higher accuracy. These gains require no adversarial training or supervision beyond the sensitive attribute labels already required by all compared fairness methods.

Notably, SupCon + ProtoFair achieves competitive or superior fairness compared to dedicated fairness methods such as GRL [22], LNL [15], and FD-VAE [20], while consistently maintaining higher classification accuracy.

We further evaluate ProtoFair in a self-supervised setting where target labels are unavailable during training. Since ProtoFair requires no target labels by design, it naturally extends to this setting. We first combine SimCLR [4] as the base loss with ProtoFair as a fairness regularizer, following the training and evaluation protocol of Table 1 but replacing SupCon with SimCLR. Prototypes are initialized via K-Means with 10 clusters after a 100-epoch warmup, followed by 30 epochs of training with the ProtoFair loss at λ=0.3. The encoder is then frozen and a linear classifier is trained for 2 epochs. We compare against the baselines in [21]: SimCLR [4], SimCLR with gradient reversal (SimCLR+GRL [22]), and FSCL* [21], a variant of FSCL that uses only the augmentation pair as the positive sample, analogous to SimCLR.

Table 3: CelebA results without target labels during representation learning, applying ProtoFair to BarlowTwins and BYOL. Only sensitive attributes are used. ACC: classification accuracy, EO: equalized odds (lower is fairer). Attributes: b=big nose, e=bags under eyes, m=male, y=young.

	T:e, S:y		T:e, S:m		T:b, S:y		T:b, S:m	
Method	ACC ↑	EO ↓	ACC ↑	EO ↓	ACC ↑	EO ↓	ACC ↑	EO ↓
BarlowTwins [26]	79.87	1.23	80.03	1.64	81.71	6.72	81.72	12.07
BarlowTwins [26]+ ProtoFair	80.40	1.16	79.99	1.36	80.0	1.15	80.41	5.98
BYOL [7]	82.66	11.21	82.66	17.04	81.77	8.9	81.77	13.54
BYOL [7] + ProtoFair	81.75	8.98	81.11	7.11	80.42	4.59	80.16	6.83

Table 4: Classification results on UTKFace (T: gender, S: ethnicity) under varying data imbalance ratios α. ACC: classification accuracy, EO: equalized odds.

	$\alpha = 4$		$\alpha = 3$		$\alpha = 2$	
Method	ACC ↑	EO ↓	ACC ↑	EO ↓	ACC ↑	EO ↓
FSCL [21]	90.1	2.7	92.3	1.7	91.6	1.0
SupCon [14]	89.8	10.6	91.6	8.4	92.0	4.5
SupCon + ProtoFair (Ours)	89.8	6.6	90.3	4.8	90.7	2.8

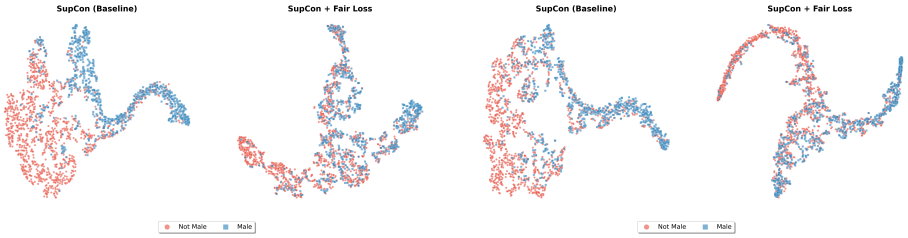
As shown in Table 2, vanilla SimCLR achieves the highest accuracy (75.7) but the worst EO (29.4). Adding ProtoFair reduces EO to 21.9, matching SimCLR+GRL while maintaining higher accuracy (73.3 vs. 72.3). FSCL* achieves the best fairness (EO of 14.8) owing to its specialized contrastive sampling strategy. Nevertheless, ProtoFair operates as a simple auxiliary loss requiring no architectural modifications, demonstrating its flexibility across supervised and self-supervised regimes.

To further assess this flexibility, we apply ProtoFair as a regularizer to two additional self-supervised methods, BarlowTwins [26] and BYOL [7], using the same setup (10 clusters, 100 epoch warmup). Both converge faster than SimCLR, requiring only 5 to 15 epochs of ProtoFair training. We use λ=0.3 for BarlowTwins [26] (except λ=0.2 for T:e/S:m) and λ=0.2 for all BYOL [7] experiments. As shown in Table 3, ProtoFair consistently improves EO across both methods and all target/sensitive-attribute pairs, at only a minor accuracy cost. This confirms ProtoFair's compatibility with diverse self-supervised objectives.

To assess the robustness of ProtoFair under varying degrees of dataset bias, we evaluate on UTKFace with gender as the target attribute and ethnicity as the sensitive attribute. The imbalance ratio α controls the skew between demographic groups, with higher values indicating stronger bias. We adopt the baseline results reported in [21] and apply ProtoFair on top of SupCon. We initialize prototypes via K-Means after a 10-epoch warmup and train with the combined SupCon and ProtoFair loss for the remaining 90 epochs. As described in Sec-

Table 5: Results on NIH Chest X-rays [25] with pneumothorax as the target and patient sex as the sensitive attribute, using SupCon as the base loss.

Method	ACC ↑	AUROC ↑	AUROC gap ↓	EO ↓
SupCon [14]	89.34	0.70	0.03	1.87
SupCon + ProtoFair	89.28	**0.72**	**0.01**	**1.79**



(a) Target: *Big Nose* **(b)** Target: *Bags Under Eyes*

Fig. 2: t-SNE visualizations colored by the sensitive attribute (*Male* vs. *Not Male*). In each subfigure, baseline SupCon is shown on (*left*) and SupCon + Fair Loss on (*right*).

tion 3, prototypes are updated via momentum after each batch and re-initialized with K-Means every 5 epochs to prevent drift from the evolving feature space. As shown in Table 4, ProtoFair consistently reduces EO compared to the SupCon baseline across all imbalance levels. EO decreases from 10.6 to 6.6 at $\alpha = 4$, from 8.4 to 4.8 at $\alpha = 3$, and from 4.5 to 2.8 at $\alpha = 2$, with minimal loss in accuracy. The relative improvement remains stable as bias increases, suggesting that ProtoFair scales gracefully with the degree of imbalance rather than being effective only in low-bias regimes.

To verify that ProtoFair generalizes beyond facial images, we evaluate it in the medical-imaging domain on NIH Chest X-rays [25], using SupCon as the base loss and pneumothorax as the target. As shown in Table 5, ProtoFair reduces the AUROC gap between sex groups (from 0.03 to 0.01) and lowers EO (1.87 to 1.79), while slightly improving overall AUROC (0.70 to 0.72) at negligible accuracy cost. This shows that ProtoFair improves both fairness and AUROC.

4.4 Qualitative Analysis of Representation Fairness

To qualitatively assess whether ProtoFair mitigates the encoding of sensitive attribute information, we visualize the CelebA test-set embeddings using t-SNE [19]. We compare the baseline SupCon [14] against our model (SupCon + ProtoFair) from Table 1 on two target attributes, Big Nose and Bags Under Eyes, both with Male as the sensitive attribute, coloring points by the sensitive attribute (Male vs. Not Male). In the baseline embedding (Figure 2(a), *left*), Male and Not Male samples form largely disjoint clusters, indicating that the encoder has entangled gender with the target features. After applying ProtoFair

Table 6: Sensitive attribute predictability from frozen features on CelebA. Lower linear-probe accuracy indicates less sensitive information is retained. Sensitive attribute: Male.

	Big Nose		Bags Under Eyes	
Method	Probe-ACC ↓	EO ↓	Probe-ACC ↓	EO ↓
SupCon [14]	87.76	20.7	82.41	20.8
SupCon + ProtoFair	81.02	15.8	76.07	10.1

for only 5 epochs (Figure 2(a), *right*), the two groups become substantially more intermixed, suggesting the regularizer reduces reliance on gender as a distinguishing feature. The same pattern holds for Bags Under Eyes (Figure 2(b)). To corroborate these observations, we train a linear classifier on the frozen features to predict the sensitive attribute, where lower accuracy indicates less retained sensitive information. As shown in Table 6, ProtoFair reduces sensitive-attribute predictability on both tasks, consistent with the t-SNE intermixing and the corresponding EO reductions.

4.5 Ablation Study

We study the sensitivity of ProtoFair to its two main hyperparameters: the number of prototypes K and the fairness weight λ. Figure 3 reports EO disparity and accuracy on four CelebA (T, S) pairs as we vary K at λ=0.3 (left) and λ at K=10 (right). Accuracy remains stable across K (within 83.0–84.5), indicating that ProtoFair is robust to the cluster count. Varying λ reveals a sweet spot in $[0.1, 0.3]$, where EO is reduced with little accuracy cost. Larger λ values overconstrain the learned representation, leading to reduced accuracy when $\lambda \geq 0.7$ or higher. These trends support our default choice of K=10 and $\lambda \in [0.1, 0.3]$ used throughout our experiments.

5 Discussion and Conclusions

We introduced ProtoFair, a fairness-aware contrastive loss that complements existing self-supervised learning objectives without modifying them. The central idea is to leverage unsupervised clustering, which is typically used in self-supervised learning to enhance representation quality, in our fairness mechanism. Momentum-updated prototypes, initialized with K-Means and refined via exponential moving average, assign each sample to a cluster that serves as a proxy for its semantic content. Samples assigned to the same cluster but belonging to different sensitive groups form *pseudo-counterfactual pairs*: they share content but differ in the sensitive attribute. Pulling these pairs together in the embedding space encourages the encoder to produce representations that are invariant to the sensitive attribute while preserving semantic content, without requiring

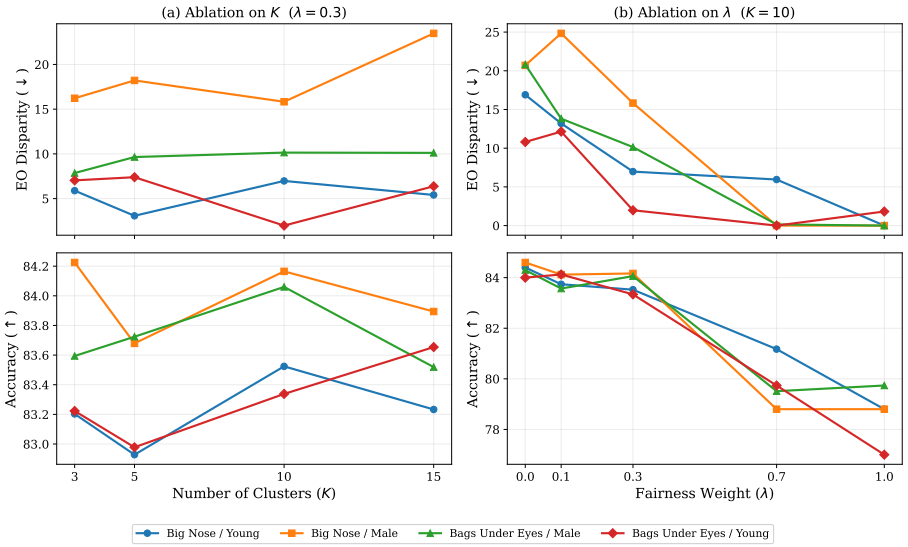


Fig. 3: Ablation on K and λ. EO (top) and accuracy (bottom) on four CelebA (T, S) pairs, varying the number of clusters K at λ=0.3 (left) and the fairness weight λ at K=10 (right).

target task labels. To the best of our knowledge, leveraging clustering based pseudo labels to construct fairness-aware contrastive objectives, in particular, using them to identify cross-group counterfactual pairs has not been previously explored. A cross-batch queue further expands pair discovery beyond the current mini-batch, strengthening the fairness signal under group imbalance.

Experiments on CelebA, UTKFace, and NIH Chest X-rays show that ProtoFair consistently reduces equalized odds across supervised (SupCon) and self-supervised (SimCLR, BarlowTwins, BYOL) base methods, achieving fairness competitive with dedicated methods while maintaining higher accuracy. Results on UTKFace further confirms robustness across varying levels of data imbalance. While dedicated fairness architectures may reach stronger performance, ProtoFair offers a practical and modular alternative that meaningfully closes the fairness gap without requiring target label supervision.

ProtoFair depends on the quality of cluster assignments, which approximate content similarity. If clusters fail to capture meaningful semantic structure, the resulting pairs may yield unreliable counterfactual approximations. We mitigate this by deriving clusters from *learned* representations and scheduling the ProtoFair loss only after a warmup period, once the encoder has matured.

Future work includes extending ProtoFair to safety-critical perception tasks such as autonomous driving, where pedestrian detection should not vary across demographic groups, and human-robot interaction, where social and assistive robots must perceive people equitably.

References

1. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (ECCV) (2018)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
3. Chai, J., Wang, X.: Self-supervised fair representation learning without demographics. Advances in Neural Information Processing Systems **35**, 27100–27113 (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
5. Chiu, C.H., Chung, H.W., Chen, Y.J., Shi, Y., Ho, T.Y.: Fair multi-exit framework for facial attribute classification. CoRR **abs/2301.02989** (2023), https://arxiv.org/abs/2301.02989
6. D'Incà, M., Tzelepis, C., Patras, I., Sebe, N.: Improving fairness using vision-language driven image augmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4695–4704 (2024)
7. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
8. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. vol. 29, pp. 3315–3323 (2016)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (2019), http://arxiv.org/abs/1911.05722, cite arxiv:1911.05722Comment: CVPR 2020 camera-ready. Code: https://github.com/facebookresearch/moco
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), https://api.semanticscholar.org/CorpusID:206594692
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (2016). https://doi.org/10.1109/CVPR.2016.90
12. Jung, S., Lee, D., Park, T., Moon, T.: Fair Feature Distillation for Visual Recognition . In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12110–12119. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01194, https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01194
13. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

15. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning Not to Learn: Training Deep Neural Networks With Biased Data . In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9004–9012. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.00922, https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00922
16. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
17. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: International Conference on Learning Representations (ICLR) (2021)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
19. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (2008)
20. Park, S., Hwang, S., Kim, D., Byun, H.: Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. Proceedings of the AAAI Conference on Artificial Intelligence **35**(3), 2403–2411 (May 2021). https://doi.org/10.1609/aaai.v35i3.16341, https://ojs.aaai.org/index.php/AAAI/article/view/16341
21. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10379–10388 (2022). https://doi.org/10.1109/CVPR52688.2022.01014
22. Raff, E., Sylvester, J.: Gradient reversal against discrimination: A fair neural network learning approach. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 189–198. IEEE (2018). https://doi.org/10.1109/DSAA.2018.00029
23. Sirotkin, K., Carballeira, P., Escudero-Viñolo, M.: A study on the distribution of social biases in self-supervised learning visual models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10442–10451 (2022). https://doi.org/10.1109/CVPR52688.2022.01019
24. Steed, R., Caliskan, A.: Image representations learned with unsupervised pretraining contain human-like biases. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 701–713. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445932, https://doi.org/10.1145/3442188.3445932
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3462–3471 (2017). https://doi.org/10.1109/CVPR.2017.369
26. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 12310–12320. PMLR (2021), http://proceedings.mlr.press/v139/zbontar21a.html
27. Zhang, F., Kuang, K., Chen, L., Liu, Y., Wu, C., Xiao, J.: Fairness-aware contrastive learning with partially annotated sensitive attributes. In: Proceedings of

the 11th International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=woa783QMul
28. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4352–4360 (2017)