Pedestrian Attribute Recognition as Label-balanced Multi-label Learning

Yibo Zhou [1 2] **Hai-Miao Hu** [1 2 3] **Yirong Xiang** [4] **Xiaokang Zhang** [1 2] **Haotian Wu** [1 2]

Abstract

Rooting in the scarcity of most attributes, realistic pedestrian attribute datasets exhibit unduly skewed data distribution, from which two types of model failures are delivered: (1) *label imbalance*: model predictions lean greatly towards the side of majority labels; (2) *semantics imbalance*: model is easily overfitted on the under-represented attributes due to their insufficient semantic diversity. To render perfect label balancing, we propose a novel framework that successfully decouples label-balanced data re-sampling from the curse of attributes co-occurrence, i.e., we equalize the sampling prior of an attribute while not biasing that of the co-occurred others. To diversify the attributes semantics and mitigate the feature noise, we propose a Bayesian feature augmentation method to introduce true in-distribution novelty. Handling both imbalances jointly, our work achieves best accuracy on various popular benchmarks, and importantly, with minimal computational budget.

1. Introduction

In visual tasks, human attribute is generally not a precisely defined concept, and can encompass a spectrum of disparate soft-biometrics that range from locatable body parts to comprehensive human descriptors (Wang et al., 2022; Liu et al., 2017). Thus, for the pedestrian attribute recognition (PAR), it is inviable to craft a universal framework that efficiently yields level performance among myriad attributes of distinct characteristics. Specifically, for accessary attribute like hat or boot, the task of PAR essentially mirrors weakly supervised object detection (Zhang et al., 2021b), as the model should infer on minimal area as relevant as possible for a discriminative recognition (Jia et al., 2022). While for the attribute of whole-body semantics like action or ages, any explicit mechanism to discard spatial information may result in insufficiency of information exchange, revealing that in this case, PAR is more akin to regular visual classification.

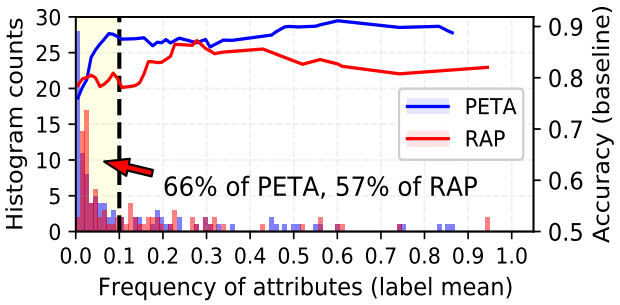


Figure 1. The dominance of negative labels in PAR datasets, and mean accuracy as a function of the label mean. In PETA (Deng et al., 2014), 66% attributes occur with a frequency under 0.1, while that for RAP (Li et al., 2016) is 57%. Also, label imbalance is the main performance bottleneck of contemporary PAR model, as it is significantly brittle to attributes with label mean ≤ 0.1.

As a result, the broadness of 'attribute' implies it a loose umbrella term, and motivates us not to approach PAR from the perspective of over-specialized architectures (Lu et al., 2023; Jia et al., 2021a). Thus, we question that, *is there a more pervasive problem existing in realistic pedestrian attribute tasks, by solving which the predictions on overall attributes are expected to be evenly boosted?* Equipped with this aspiration, we concisely distill PAR into a problem of multi-label classification under significant data imbalance.

This simplification makes sense as: (1) the ambiguity and variety within attributes require a much general PAR definition; (2) since most attributes occur with small empirical frequencies, PAR datasets are profoundly label-imbalanced. Worse, attribute label priors exhibit great unpredictability across various sceneries (Zhou et al., 2023), making it an immense data selection bias that hardly generalizes; (3) previous work only partly alleviates label imbalance by experimentally setting different penalizing weights for labels (Wang et al., 2022), or just abstains the overly infrequent attributes from benchmarks to display decent overall results

[1]Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China [2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China [3]Hangzhou Innovation Institute of Beihang University, Hangzhou 310051, China [4]The University of Manchester, UK. Correspondence to: Hai-Miao Hu <hu@buaa.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(Jia et al., 2021b). Thus, label imbalance is de facto the *grey rhino* that fundamentally bottlenecks the performance of PAR (Figure 1), and remains critically under-addressed.

While data re-sampling (over-sample the images from under-represented label or under-sample the others) can facilitate unbiased label prior for long-tailed recognition (Zhang et al., 2021d), it is infeasible to be directly adopted into PAR owing to the intricate entanglement of attributes in images. In specific, given the limited patterns of label co-occurrence in dataset, repeating/dropping images to equalize the sampling distribution of one attribute will inevitably skew the balance of others (Guo & Wang, 2021). In contrast to segmenting attributes in pixel space for independent sampling, we demonstrate that such a gap can be absolutely bridged if the re-sampling space is shifted from images to latent features. Consequentially, we develop a novel training pipeline of feature re-sampled learning to offer immunity to this curse of label co-occurrences, and thus ensure true **label balance** for PAR. Both theoretical insights and experimental evidence suggest that our method is an ideal drop-in instantiation for the intractable label-balanced image re-sampling of PAR.

However, since the features of under-represented attributes may not suffice to describe the intra-class expressivity, when they are over-repeatedly sampled for label balancing, severe overfitting can be posed. To palliate such incidental overfitting, we aim to enrich feature novelty to attain **semantics balance**. One principled solution for it is resorting to feature augmentation techniques (DeVries & Taylor, 2017a), and a prevalent recipe in this topic is built with an implicit assumption that the intra-class translating direction is homogeneous across the feature space, and samples synthetic points from identical gaussian clouds centering at different features (Wang et al., 2019; Li et al., 2021).

Unfortunately, we unveil that no novel variety is introduced by these homogeneous methods as they can be essentially reformulated as large-margin optimizers (Liu et al., 2016) with static margins. As a counter, we state the necessity of heterogeneous feature augmentation for genuine semantics diversification, and promote a Bayesian method for it. With our approach, feature of impoverished labels is augmented by non-trivial gradient-based stochasticity, in effect relieving the exacerbated overfitting. Also, we theoretically prove that our method is able to assuage the data noise from spurious feature during feature re-sampling.

Our method establishes state-of-the-art performance on various benchmarks, and the contribution is three-fold:

- To our best knowledge, this is the first work that develops true label-balanced learning for multi-label tasks.
- We elaborate on the whys and wherefores of the pitfall of existing feature augmentation methods, and propose a Bayesian approach to create true novel features.
- By mitigating two types of imbalance, our lightweight framework scores best w.r.t. mean accuracy on realistic PAR benchmarks. Extensive ablation and robustness studies also validate a suite of merits of our proposal.

2. Related Work

Pedestrian Attribute Recognition. Basically, there are two common paradigms in PAR. First class of studies has delved into enhancing attributes localization to reduce the accuracy drop from predicting on extraneous area. Various attention mechanisms (Liu et al., 2018; Jia et al., 2022; Liu et al., 2017), attributes partition strategies (Fabbri et al., 2017; Li et al., 2017) and body-attributes dependencies (Liu et al., 2018; Lu et al., 2023) were leveraged to better capture the spatial topological structure of attributes. Another active research stream regards attributes correlation as a concrete prior (Li et al., 2022; Fan et al., 2020; Wang et al., 2017), and attempts to exploit attributes interdependencies by graph models. However, both lines of work are questionable. (Jia et al., 2020) showed that attribute positioning may not be the core performance bottleneck of PAR. Also, (Zhou et al., 2023) discovered that attributes co-occurrence is more like a mutable data selection bias that impairs the PAR performance. Such paradoxical results make us rethink, what is indeed a fundamental factor for PAR to scale well?

Imbalance in Multi-label Tasks. Limited by the label co-occurrences, existing multi-label methods ease the label imbalance mainly by loss re-weighting (Jia et al., 2021b), such as using the inverse of label-wise sample size in loss function to up-weight minority data (Xu et al., 2022), or other alternative weighting functions (Li et al., 2015; Tan et al., 2020a). Differently, this work achieves label-balanced re-sampling for multi-label recognition. Moreover, not only the numerical asymmetry of labels distribution, we also milden the twined semantics imbalance.

3. Method

3.1. On the Label-balanced Re-sampling of PAR

Formally, let X be a distribution characterized by all of the pedestrian surveillance images. Some data points $\{\boldsymbol{x}_i\}_{i=1}^N$ are sampled from X, jointly with their corresponding labels $\{\boldsymbol{y}_i\}_{i=1}^N$ of certain attributes to form a dataset D, where N denotes the dataset cardinality $|D|$, $\boldsymbol{y}_i \in \{0,1\}^C$ and C is the number of total annotated attributes. Each element in $\boldsymbol{y}_i$ serves as the 0/1 indicator of the occurrence of an attribute in $\boldsymbol{x}_i$. Practically, such a dataset D is collected from X with small empirical attribute frequencies. It results in that $\frac{N^k}{N}, \forall k = 1, 2, ..., C$, can be far from 0.5, where N^k is the number of images in D with attribute label $\boldsymbol{y}^k$ being 1.

Consequentially, the separating hyperplane in the decision space will be heavily skewed to the label of relatively few number, from where poor PAR performance is delivered.

Label-balanced re-sampling is the most straightforward approach to facilitate recognition with such imbalanced labels.

Label-balanced Image Re-sampling (LIR): Adjust the sampling function of images, to let the attributes images fed into model perfectly balanced between binary labels.

LIR is achievable only if there exists $\{a_i\}_{i=1}^N$ satisfying

$$\begin{aligned} &\sum_{i=1}^{N} \boldsymbol{y}_i \cdot a_i + \sum_{i=1}^{N} (\boldsymbol{y}_i - \mathbf{1}) \cdot a_i = \mathbf{0}, \\ \text{s.t.} \ &\sum_{i=1}^{C} a_i = 1, \ \ a_i > 0, i = 1, 2, ..., N. \end{aligned} \tag{1}$$

Since patterns of attributes co-occurrence can be quite limited (Zhou et al., 2023), gathering a dataset meeting Eq.1 is difficult. It reveals that, re-adjusting the sampling function of a certain attribute to balance its label prior would yield another biased distribution for others. Also, as a_i represents the probability of $\boldsymbol{x}_i$ to be sampled, it is expected that all $a_i > 0$ and have a similar value such that data points can be sampled with comparable odds, making an acceptable sampling function much impracticable to get. Essentially, such curse of label co-occurrence roots from that all attributes are entangled in input images, implying that for independent balancing of each attribute, LIR would be preconditioned on some challenging methods to precisely segment attributes in pixel space. Instead of attributes segmenting, we attempt to label-balanced re-sample attributes in a label-disentanglable feature space to unconditionally bridge this gap.

3.2. Feature Re-sampled Decoupled Learning

For multi-class recognition, decoupling is one of the training schemes most successful on long-tailed datasets (Zhang et al., 2023). Its two-stage workflow is streamlined as

Decoupled Learning (DL): *Stage#1*: Do vanilla training with instance-balanced sampled images to learn a whole model. *Stage#2*: The images are label-balanced sampled, and only fine-tune the classifier with other modules fixed.

Compared to label-balanced image re-sampling, DL renders better accuracy on long-tailed dataset, since it not only gives same neutral decision boundaries in classifier, but also produces more discriminative latent representations thanks to that the feature extractor in DL is not overfitting on the over-sampled images of minority classes (Kang et al., 2019). Inspired by it, we conjecture that solving the impossibility of label-balanced sampling in the attributes-entangled pixel space might not be technically indispensable for true balanced PAR, as we actually do not need a label-balanced learned feature extractor. In other words, *all we need is a label-balanced classifier.* Importantly, this concept remedies the curse of attributes co-occurrence of LIR for PAR, *since unlike feature extractor, classifiers weight is not shared among attributes, meaning that the inferences of attributes are already structurally disentangled in the final classification step, and independent attributes re-sampling is thus viable for PAR classifier.* To this end, we devise the pipeline

Feature Re-sampled Decoupled Learning (FRDL): *Stage#1*: Do vanilla training with instance-balanced data sampling to learn a whole model. *Stage#2*: Input image is still instance-balanced sampled and fed into fixed feature extractor to produce representations. Differently, features are saved in memory banks according to their labels, and classifier is re-trained on label-balanced sampled features.

as an upper substitution of LIR. Concretely, we denote with $\boldsymbol{f}_i = \mathcal{H}_\theta(\boldsymbol{x}_i) \in \mathbb{R}^M$ the representation of $\boldsymbol{x}_i$, where $\mathcal{H}_\theta(\cdot)$ is a feature extractor parameterized by θ. Sequentially, $\boldsymbol{f}_i$ is decomposed into M-dimensional attribute-specific features $\{\boldsymbol{f}_i^k\}_{k=1}^C = \mathcal{T}_\psi(\boldsymbol{f}_i)$ by a fully-connected layer $T_\psi(\cdot)$. Attribute posterior is finally estimated with a linear classifier function $\tilde{\boldsymbol{y}_i^k} = \boldsymbol{w}^{k\top} \boldsymbol{f}_i^k + b^k$, where $\boldsymbol{w}^k \in \mathbb{R}^M$ represents the classifier weight, and $b^k \in \mathbb{R}$ the bias, $\forall k = 1, 2, ..., C$.

For *Stage#1*, we train whole model on the instance-balanced sampled images by plain binary cross-entropy (BCE) loss. When model converges, we feed the whole dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ into fixed $\mathcal{H}_\theta(\cdot)$ and $\mathcal{T}_\psi(\cdot)$, and collect the output representations $\{(\boldsymbol{f}_i^1, \boldsymbol{f}_i^2, ..., \boldsymbol{f}_i^C)\}_{i=1}^N$ into C pairs of attribute-specific feature banks $\{(Q_0^k, Q_1^k)\}_{k=1}^C$. Specifically, Q_0^k and Q_1^k save all $\boldsymbol{f}_i^k$ with label $\boldsymbol{y}_i^k$ being 0 and 1, respectively. Finally, the *Stage#2* of FRDL draws between Q_0^k vs. Q_1^k with an equal probability, and a feature from the selected bank is uniformly sampled with replacement to form a label-balanced training batch, atop which $(\boldsymbol{w}^k, b^k)$ is fine-tuned.

Seemingly, FRDL and DL make no difference in multi-class tasks. However, in the context of PAR, FRDL is non-trivial as: (1) it unconditionally achieves label-balanced classifier, by transferring the unsatisfiable label-balanced image re-sampling in the *Stage#2* of DL to a tractable label-balanced feature re-sampling; (2) even if Eq.1 is satisfied for DL, the over-sampled images to balance an attribute will be uncalled-for repeated in the classifiers learning of other attributes, propagating the overfitting issue coupled with balanced re-sampling of one attribute to all attributes. Differently, as the attributes inferences are already disentangled in classifier, FRDL enables not only label-balanced, but also independent

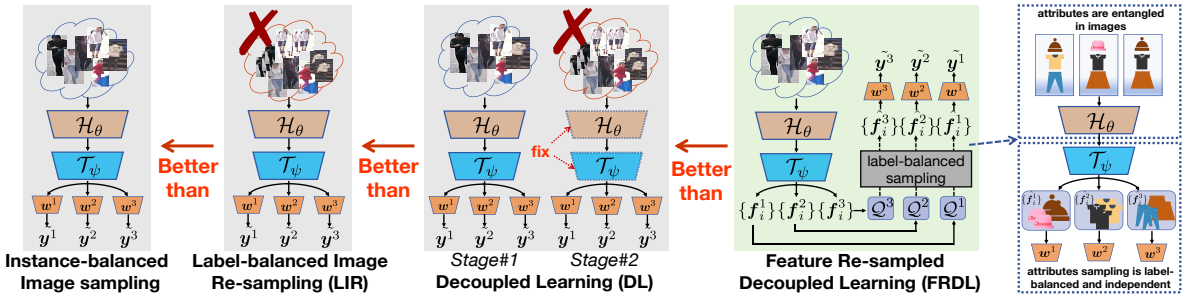


Figure 2. Schematic presentation of the main idea of FRDL. Although DL can not be naively implemented for PAR due to the unsatisfiable label-balanced image re-sampling (Eq.1), its better form of FRDL is workable and thus acts as a better drop-in substitution of LIR.

classifier learning for each attribute, and thus performs better than DL. Main concept of FRDL is illustrated in Figure 2.

3.3. Pitfall of Homogeneous Feature Augmentation

To obviate overfitting aggravated by the over-sampled features in FRDL, an intuitive solution is to diversify the limited statistics of minority attributes by strong data augmentation (Chawla et al., 2002). However, most image augmentation techniques can potentially obliterate the delicate signatures of small attributes within the pixel space, thus leading to subpar performance (see Appendix D.3). Hence, we resort to augment data in latent space. Postulating that certain directions in feature space are aligned with intra-class semantics variation, ISDA (Wang et al., 2019) and its follow-ups translate the features linearly in some latent directions to augment additional representations. For PAR, they can be expressed as a feature re-sampling process of $\tilde{\boldsymbol{f}}_i^k \sim \mathcal{N}(\boldsymbol{f}_i^k, \lambda^k \boldsymbol{\Sigma}^k)$, and just differ by the specific choice of $\{\boldsymbol{\Sigma}^k\}_{k=1}^N$.

Since the translating directions at different $\boldsymbol{f}_i^k$ are sampled from a same prior $\mathcal{N}(0, \lambda^k \boldsymbol{\Sigma}^k)$, they are actually presumed, by ISDA, homogeneous across the whole latent space. However, on one hand latent direction of intra-class variation is not as homogeneous as consistent gaussian clouds, since in practice features are distributed heterogeneously (Wan et al., 2018). On the other hand, to explore all directions in $\mathcal{N}(0, \lambda^k \boldsymbol{\Sigma}^k)$, one should minimize the expectation of the BCE loss of PAR, under all possible augmented features, as

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\boldsymbol{f}}_i^k}[\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\log(1+e^{-\mathbb{I}(\boldsymbol{y}_i^k)\cdot(\boldsymbol{w}^{k\top}\tilde{\boldsymbol{f}}_i^k+b^k)})] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\log \mathbb{E}_{\tilde{\boldsymbol{f}}_i^k}[1+e^{-\mathbb{I}(\boldsymbol{y}_i^k)\cdot(\boldsymbol{w}^{k\top}\tilde{\boldsymbol{f}}_i^k+b^k)}] \\
&= \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\log(1+e^{-\mathbb{I}(\boldsymbol{y}_i^k)\cdot(\boldsymbol{w}^{k\top}\boldsymbol{f}_i^k+b^k)+\frac{1}{2}\boldsymbol{w}^{k\top}\lambda^k\boldsymbol{\Sigma}^k\boldsymbol{w}^k})}_{\text{binary cross-entropy on } \boldsymbol{f}_i^k}.
\end{aligned}
\tag{2}
$$

In Eq.2, $\mathbb{I}(\boldsymbol{y}_i^k) = 1$ if $\boldsymbol{y}_i^k = 1$ and $\mathbb{I}(\boldsymbol{y}_i^k) = -1$ if $\boldsymbol{y}_i^k = 0$. The inequality follows from the Jensen inequality and the final step is obtained by the moment-generating function for the gaussian variable $\tilde{\boldsymbol{f}}_i^k$. It reveals that Eq.2, a closed-form upper bound of the homogeneous feature augmentation loss, is in essence a vanilla BCE loss with fixed inter-label margins since $\{\frac{1}{2}\boldsymbol{w}^{k\top}\lambda^k\boldsymbol{\Sigma}^k\boldsymbol{w}^k\}_{k=1}^C$ are just constants. Thus, *homogeneous methods are endogenously large-margin optimizers*, and carefully tuning $\{\lambda^k\}_{k=1}^C$ like their original paper is intrinsically enumerating the priori-unknown best inter-label margin and will finally smooth out any difference in the specific choices of $\{\boldsymbol{\Sigma}^k\}_{k=1}^N$. As a result, we argue that no novel diversity regarding distribution exploration can be inherently introduced by homogeneous methods.

3.4. Gradient-oriented Augment Translating

We are now in a position to overcome above issue. A desirable translating direction to augment features should comprise: (1) **in-distribution**, the augmented features still reside in the latent domain of same attribute identity; (2) **meaningful**, the translating directions co-linear with attribute semantics shifting, instead of some random noise; (3) **heterogeneous**, the translating direction of each feature is computed from its own neighborhood of the distribution. Hence, for any feature point $\boldsymbol{f}_i^k$ within a trained model, we translate it along its local gradient to augment new feature

$$
\tilde{\boldsymbol{f}}_i^k = \boldsymbol{f}_i^k - \eta\nabla_{\boldsymbol{f}^k=\boldsymbol{f}_i^k}|\mathcal{L}_{cls}(\boldsymbol{f}^k) - \mathbb{E}_{\boldsymbol{f}^k}[\mathcal{L}_{cls}(\boldsymbol{f}^k)]|, \tag{3}
$$

where $\mathcal{L}_{cls}(\cdot)$ computes the BCE loss of $\boldsymbol{f}$, and η is a positive step size. During this process, the classifier utilized for the gradient computation is well-trained and remains fixed. Conversely, a fresh classifier is independently trained from scratch with $\tilde{\boldsymbol{f}}_i^k$, and finally takes over for the test-time classification. The rationales behind applying Eq.3 for feature augmentation are: (1) the translating is high-density oriented as it always points to the distribution centroid $\mathbb{E}_{\boldsymbol{f}^k}[\mathcal{L}_{cls}(\boldsymbol{f}^k)]$. Therefore, the over-confident features

(small loss) would be pulled back to be less-confident, while the noisy features (large loss) would be relaxed into high-density zone. Consequentially, no outliers are created, leading to **in-distribution**; (2) the feature is transferred in the direction of loss gradient, which is most relevant to the attribute informativeness across the entire space. It enables that, instead of a quasi replication, the augmented feature is novel w.r.t. its initial representation in term of the embedded attributes semantics, i.e., the translating is **meaningful**; (3) with subsequent non-linear classifier, the gradient varies among different feature points, making Eq.3 form a **heterogeneous** sampling field of translating directions.

Practically, the proposed Gradient-Oriented Augment Translating (GOAT) in Eq.3 can be seamlessly implemented without further efforts. In specific, if we optimize the feature extractor $\mathcal{T}_{\psi_t}(\mathcal{H}_{\theta_t}(\cdot))$ at training step t by gradient descend w.r.t. a succinct loss $\mathcal{L}_{goat}$ of

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}|\mathcal{L}_{cls}(\boldsymbol{f}_{i,t}^k)-\mu_t^k|, \\ s.t.\quad &\{\boldsymbol{f}_{i,t}^k\}_{k=1}^C=\mathcal{T}_{\psi_t}(\mathcal{H}_{\theta_t}(\boldsymbol{x}_i)), \end{aligned} \tag{4}$$

where μ_t^k is $\mathbb{E}_{\boldsymbol{f}_t^k}[\mathcal{L}_{cls}(\boldsymbol{f}_t^k)]$, sequentially, $\tilde{\boldsymbol{f}_{i,t}^k}$ that translated from $\boldsymbol{f}_{i,t}^k$ by Eq.3 would be identical to $\boldsymbol{f}_{i,t+1}^k$ generated by $\mathcal{T}_{\psi_{t+1}}(\mathcal{H}_{\theta_{t+1}}(\cdot))$. The reason is that, to minimize Eq.4, the feature extractor would be updated to translate $\boldsymbol{f}_{i,t}^k$ along the same direction of $-\nabla_{\boldsymbol{f}_t^k=\boldsymbol{f}_{i,t}^k}|\mathcal{L}_{cls}(\boldsymbol{f}_t^k)-\mathbb{E}_{\boldsymbol{f}_t^k}[\mathcal{L}_{cls}(\boldsymbol{f}_t^k)]|$ in Eq.3. Importantly, *it reveals the inherent equivalence between gradient-oriented feature augmentation and the feature extractor gradient-descending.* Thus, to incorporate additional stochasticity, we optimize $(\theta_{t_0},\psi_{t_0})$, which is the optimum feature extractor pre-trained on D, w.r.t. Eq.4 and treat the features collected along a short stochastic gradient descent (SGD) trajectory of $\{\theta_{t_0+s},\psi_{t_0+s}\}_{s=0}^T$ as representations aptly augmented from the features at t_0, where $T\geq 1$. As such, GOAT approximates Bayesian feature sampling, as we can use (θ,ψ) at different steps to produce the probabilistic representations of a same input data. In this regard, GOAT essentially constructs a high-density-oriented *heterogeneous Bayesian sampling cloud* around $\boldsymbol{f}_i$, which is in contrast to the homogeneous sampling cloud of prior arts. Notably, throughout the entire process, the likely feature distortion towards out-of-distribution is mitigated, since the classifier for gradient computation is fixed, resulting in that the subsequent $\{\theta_{t_0+s},\psi_{t_0+s}\}_{s=1}^T$ would evolve within the vicinity of the initial classifier solution. Also, we set T as a small number, and reload the model with $(\theta_{t_0},\psi_{t_0})$ when the SGD trajectory reaches T (larger T produces stochasticity beyond Eq.3, but with more risk of off-distribution).

Eq.4 also can be reasoned and adopted within the setting of feature de-noising. Practically, no matter how well-trained a feature extractor is, it inevitably encounters failures in pinpointing certain attributes from hard-case images, deriving information from the background as discriminative attribute representations and mistakenly pairing them with the positive labels of their original images. When the mismatched feature-label pair of minority attributes are over-sampled by the *Stage#2* of FRDL, noise can be greatly overfitted by the classifier, to which we refer as the feature noise in FRDL. In Appendix C.1, we prove following proposition,

Proposition 3.1. *Eq.4 is upper bounded by the optimum feature de-noising BCE loss:*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}|\mathcal{L}_{cls}(\boldsymbol{f}_i^k)+\log(1-\sigma^k)|, \tag{5}$$

where σ^k represents the feature noise rate of attribute k.

Consequently, we also apply Eq.4 in the *Stage#2* of FRDL to train the label-balanced classifier, thereby rendering less overfitting on the spurious features. Finally, GOAT can be realized implicitly to enhance FRDL in term of semantics diversification and feature de-noising, and Algorithm 1

Algorithm 1 Pseudo-code of the GOAT-enhanced FRDL

Input: Training set D; ending step T_1, T_2 and T; initialized modules $\mathcal{H}_{\theta_0}(\cdot)$ and $\mathcal{T}_{\psi_0}(\cdot)$; initialized classifiers $\mathcal{G}_{W_0}^{cls}(\cdot)$ and $\mathcal{G}_{W_0}^{ft}(\cdot)$; empty feature banks $\{(Q_0^a,Q_1^a)\}_{a=1}^C$

Output: Instanced-balanced $\mathcal{H}_{\theta_{T_1}}(\cdot),\mathcal{T}_{\psi_{T_1}}(\cdot)$; Label-balanced $\mathcal{G}_{W_{T_2}}^{ft}(\cdot)$ # final model is $\mathcal{G}_{W_{T_2}}^{ft}(\mathcal{T}_{\psi_{T_1}}(\mathcal{H}_{\theta_{T_1}}(\cdot)))$

for $j=1$ **to** T_1 **do** #*Stage#1* for FRDL
  Instance-balanced draw a batch $B=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{|B|}$
  Calculate loss # can be any alternative of PAR loss
  $\mathcal{L}_B=\frac{1}{|B|}\sum_{i=1}^{|B|}L_{bce}(\mathcal{G}_{W_{j-1}}^{cls}(\mathcal{T}_{\psi_{j-1}}(\mathcal{H}_{\theta_{j-1}}(\boldsymbol{x}_i))),\boldsymbol{y}_i)$
  Update: $W_j\leftarrow W_{j-1}-\alpha\nabla_W\mathcal{L}_B$; # α: learning rate
  $\theta_j\leftarrow\theta_{j-1}-\alpha\nabla_\theta\mathcal{L}_B$; $\psi_j\leftarrow\psi_{j-1}-\alpha\nabla_\psi\mathcal{L}_B$
Compute the attribute-wise loss centroid μ^b in Eq.4 by $\mathcal{G}_{W_{T_1}}^{cls}(\cdot),\mathcal{H}_{\theta_{T_1}}(\cdot)$ and $\mathcal{T}_{\psi_{T_1}}(\cdot)$ on $D,\forall b=1,2,...,C$
for $j=T_1$ **to** T_2 **do** #*Stage#2* for FRDL + GOAT
  Instance-balanced draw a batch $B=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{|B|}$
  if $(j\,\%\,T=0$ **or** $j=T_1):k=T_1$ **else**: $k=k+1$
  Calculate $\mathcal{L}_{goat}$ on B w.r.t. $\mathcal{G}_{W_{T_1}}^{cls}(\mathcal{T}_{\psi_k}(\mathcal{H}_{\theta_k}(\cdot)))$ and $\{\mu^b\}_{b=1}^C$ by Eq.4
  Update: # optimize $\theta\&\psi$ around W_{T_1} solution
  $\theta_{k+1}\leftarrow\theta_k-\alpha\nabla_\theta\mathcal{L}_{goat}$; $\psi_{k+1}\leftarrow\psi_k-\alpha\nabla_\psi\mathcal{L}_{goat}$
  Save the produced features into $\{(Q_0^a,Q_1^a)\}_{a=1}^C$
  for $l=1$ **to** C **do** #label-balanced train $\mathcal{G}^{ft}(\cdot)$
    Form a batch $\hat{B}=\{(\tilde{\boldsymbol{f}_i^l},\boldsymbol{y}_i^l)\}_{i=1}^{|\hat{B}|}$ label-balanced drawn from (Q_0^l,Q_1^l) with replacement
    Calculate loss # W^l denotes l-th column of W
    $\mathcal{L}_{\hat{B}}=\frac{1}{|\hat{B}|}\sum_{i=1}^{|\hat{B}|}|\mathcal{L}_{cls}^{W_j^l}(\tilde{\boldsymbol{f}_i^l})-\mu^l|$
    Update: $W_{j+1}^l\leftarrow W_j^l-\alpha\nabla_{W^l}\mathcal{L}_{\hat{B}}$

Table 1. Comparisons with the state-of-the-art methods. We refer the strong baseline results on PA100k and PETA from (Specker et al., 2022), and the baseline results on RAPv1 are from our experiments. All values are percentages, and the best are highlighted in **bold** fonts.

Method	Network	PA100k		RAPv1		PETA	
		mA	F1	mA	F1	mA	F1
Baseline (Specker et al., 2022)	*ResNet-50*	81.6-	88.1-	80.18	79.32	84.0-	86.3-
	ConvNeXt-base	82.2-	88.5-	80.61	81.76	86.1-	88.1-
	Swin Transformer	83.2-	88.5-	82.12	**82.30**	86.6-	87.7-
DAFL (Jia et al., 2022)	*ResNet-50*	83.54	**88.90**	83.72	80.29	87.07	86.40
DRFormer (Tang & Huang, 2022)	*ViT-B/16*	82.47	88.04	81.81	81.42	**89.96**	88.30
IAA-Caps (Wu et al., 2022)	*OSNet*	81.94	87.80	81.72	80.37	85.27	85.64
EALC (Weng et al., 2023)	*EfficientNet-B4*	81.45	88.14	83.26	81.67	86.84	88.40
DFDT (Zheng et al., 2023)	*Swin Transformer*	83.63	88.74	82.34	82.15	87.44	88.19
FEMDAR (Cao et al., 2023)	*ResNet-50*	81.02	87.32	79.71	78.76	84.73	85.90
VAL-PAT (Bao et al., 2023)	*ResNet-50&Transformer*	82.3-	88.5-	80.8-	81.0-	83.1-	84.4-
PARFormer-B (Fan et al., 2023)	*Swin Transformer*	83.95	87.69	83.84	81.16	88.65	88.66
FRDL + GOAT	*ConvNeXt-base*	**89.44**	88.05	**87.72**	79.16	88.59	**89.03**

overviews the whole workflow, where we ignore the classifier bias for brevity, and $W = (\boldsymbol{w}^1, \boldsymbol{w}^2, ..., \boldsymbol{w}^C)$.

4. Experiments

4.1. Experimental Setup

Evaluation protocol. We perform experiments on popular large-scale PAR datasets of PA100k (Liu et al., 2017), PETA (Deng et al., 2014) and RAPv1 (Li et al., 2016). For the datasets configuration, we strictly follow (Jia et al., 2020) to make a wide and fair comparison with prior arts. It is noteworthy that, for this datasets protocol, there are total 60 annotated attributes in PETA, but 25 attributes are dismissed from evaluation due to their great label asymmetry. For RAPv1, 21 attributes are disregarded for the same reason. Considering that some dropped attributes only have a handful of samples, also to be consistent with the datasets configuration of prior arts, we do not use full attributes of PETA and RAP in our testing as well. We discern between methods by reporting their scores on the label-based metric mean Accuracy (mA), which computes the mean of all attributes recognition accuracy on the positive and negative data. Instance-based metric F1-score (F1) is also evaluated. Details are placed in Appendix B.1 and Appendix D.1.

Implementation details. We adopt ConvNeXt-base (Liu et al., 2022) as the backbone of $\mathcal{H}_\theta(\cdot)$, due to its desirable trade-off between performance and efficiency. Classifiers and $\mathcal{T}_\psi(\cdot)$ are instantiated by single fully-connected layer as the simplest form among possible variants. Image is spatially resized to 256×192 for input, and batch size is set as 64. Adam solver is applied with weight decay of 5e-4. Horizontal flip and random crop are the only image augmentation methods. The learning rate starts at 1e-4 and decays by a factor of 10 at certain steps. Unless otherwise stated, we on default set the T in Algorithm 1 as 20. Other details can be referred in our code at github.

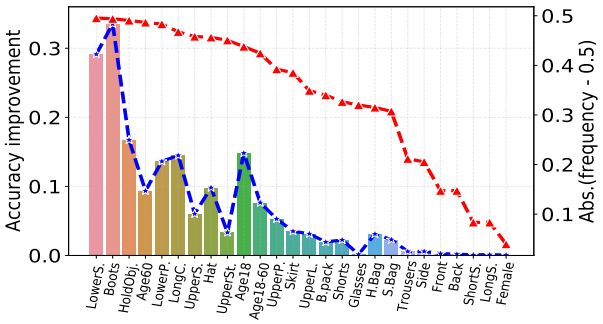


Figure 3. The boost of attributes mA of our method w.r.t the baseline for PA100k. Attributes are placed in a decreasing order of the absolute difference between their label mean and 0.5 (red-dashed line), which can quantitatively measure their label imbalance.

4.2. Benchmark Results

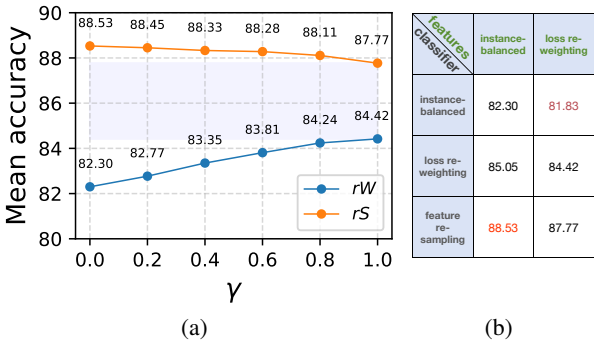
We conduct a thorough evaluation of our method, comparing with strong baselines and a range of recent approaches. The results are presented in Table 1. Basically, our method enjoys a wide range of meritorious superiorities with practical significance in the challenging real-world scenarios:

● **Strong performance.** FRDL and GOAT in tandem excel existing methods utterly in mA on PA100k and RAPv1. While for other settings, our method is at least on-par with others. Also, prior arts comparable scores on PETA might attribute to the data leakage in its training set about the test data (Jia et al., 2021b), and thus are likely overrated (Zhou et al., 2023). In Appendix B.3, when the data leakage on PETA is tackled, our method outperforms prior methods with considerable margins. Overall, the result not only highlights the effectiveness of our proposal, but also signifies that modern PARs do not fuel significantly performant models due to their ineffective treatment of the pivotal label imbalances, reinforcing the driving principle of this paper.

Table 2. Comparison of our methods with their existing alternatives, and the ablation results. All values are mA.

Method	PA100k	RAPv1	PETA
Baseline	82.30	80.61	86.29
re-weighting#1 (Li et al., 2015)	84.29	83.11	87.46
re-weighting#2 (Tan et al., 2020a)	84.42	83.47	87.67
re-weighting#3 (Zhang et al., 2021c)	84.77	83.55	**87.85**
FRDL(Ours)	**88.53**	**86.31**	87.55
FRDL+ISDA(Wang et al., 2019)	88.70	86.52	87.94
FRDL+ISDA ($\mathbf{\Sigma}^*$ is random noise)	88.79	86.46	87.93
FRDL+MetaSAug(Li et al., 2021)	88.63	86.58	87.80

FRDL+GOAT of Eq.3	Eq.4 (SGD)	feature de-noising	PA100k	RAPv1	PETA
✓			88.84	86.75	87.96
	✓		89.15	87.52	88.27
	✓	✓	**89.44**	**87.72**	**88.59**



(a) (b)

Figure 4. (a): We contrast the representation quality of backbones learned with various label-balancing ratio γ on PA100k. From γ of 0 to 1, feature extractor transitions from that of a plain baseline to that of the loss re-weighted model. The y-axis shows the mA of a classifier re-trained on each of these converged feature extractors, by which we discern between the feature quality of them. rS denotes that the classifier is re-trained by the *Stage#2* of FRDL, while rW represents the classifier is trained by loss re-weighting. (b): The mA of different feature extractor + classifier combinations on PA100k. For PAR, it manifests that: (1) label balancing obstructs a performant feature extractor; (2) label-balanced feature re-sampling produces fairly better classifier; (3) LIR is supposed to lie between 84.42 - 87.77% mA, falling behind FRDL (88.53%).

● **Good generalizability.** Our method emphasizes on a general problems of the asymmetry in label distribution, thereby functioning with less inductive biases. Figure 3 reports the attribute-wise accuracy increase of our method over the baseline, and we observe sizable improvements on all attributes: performance gain is larger for infrequent attributes, less for balanced attributes, but never negative.

● **Minimal computational burden.** Unlike prior arts, we do not facilitate PAR in a multi-modal or multi-task manner by involving related tasks, and nor do we pay a premium regarding parameters by stacking costly modules. During inference, our method exercises with the computational footprint as minimal as that of any baseline model, but still yields overall best accuracy, without the bells and whistles.

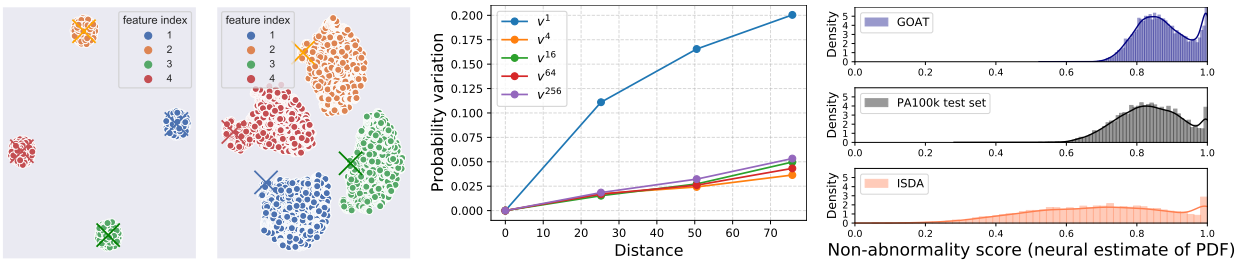
● **High compatibility.** Both FRDL and GOAT are macro learning pipelines that lean on no specific or customized network architectures. Thus, our work is of great applicability and can be employed as an effortless plug-and-play companion onto any existing methods.

4.3. FRDL Achieves Label Balance

Present works to dampen the label imbalance in multi-label tasks rely on loss re-weighting techniques. In Table 2, FRDL competes with some of the best performing re-weighting functions for PAR (Jia et al., 2021b), which scale attributes loss by their labels mean, and have been widely integrated into notable works (Lu et al., 2023; Zhou et al., 2023). In Table 2, FRDL outscores both the baseline and re-weightings with substantial margins. Also, it is noteworthy that re-weighting alone brings about 1.5-3% improvement of mA, which is not trivial as the total improvements of prior methods over baseline vary within about 1-4% mA. Consequentially, it double-verifies our point of view that label imbalance is the main performance bottleneck for PAR.

Moreover, we state that FRDL develops true label balancing for PAR, since it always delivers performance better than LIR. As is validated by (Kang et al., 2019): (1) learning the feature extractor with instance-balanced sampling produces more generalizable features; (2) learning the classifier with label-balanced sampling sets proper decision boundaries over the learned representations. Correspondingly, if the PAR model is decoupled into a feature extractor of (θ, ψ) and classifiers denoted by W, for FRDL, (θ, ψ) is trained by plain instance-balanced sampling, while W is updated with label-balanced features, meeting the expectations of both (1) and (2). Hence, by inferencing on a better feature extractor, FRDL is an empirical upper replacement of LIR.

We experimentally prove it in Figure 4(a), where we train a sequence of feature extractors on PA100k with different label-balancing ratio (detailed in Appendix D.2), and examine the feature quality of them by comparing the accuracy of classifiers re-trained atop their representations. Since a perfectly label-balanced feature extractor of PAR is practically impossible due to Eq.1, we apply loss re-weighting (Tan et al., 2020a) in this study to simulate (θ, ψ) learned from relatively balanced labels. In Figure 4(a), by decreasing the degree of label balancing in feature extractor training, the feature quality upgrades persistently, and the best feature is obtained at (θ, ψ) from instance-balanced sampling, which is exactly FRDL. Since the (θ, ψ) of LIR would be fully label-balanced trained, its accuracy corresponds to a point in the blue-shaded region of Figure 4(a), revealing an inferior performance of LIR when competed with FRDL. Thus, as



(a) (b) (c) (d)

Figure 5. Proof-of-concept experiments for GOAT. T-SNE map of the augmented features distribution for (a) ISDA and (b) GOAT ($\times$ denotes the original feature). (c): Posterior variation along the rays corresponding to different eigenvectors of GOAT translation covariance matrix. (d): For PA100k, the distributions of the PDF scores for the features from test data, ISDA and GOAT augmentations, respectively.

it is intractable and unnecessary to disentangle attributes in image, FRDL could serve as an optimum paradigm for label balancing in multi-label visual tasks.

4.4. GOAT Approximates Bayesian Feature Sampling

In Table 2, ISDA and MetaSAug do not give much boosting over FRDL. It is expectable, since the homogeneity assumption of them is too strong and invites their equivalence to constant-margin optimizer. We experimentally justify it by replacing the feature-sampling covariance $\boldsymbol{\Sigma}$ in ISDA with random gaussian noise $\boldsymbol{\Sigma}^*$, and fine-tune the λ to get best results. It reports that specific form of $\boldsymbol{\Sigma}$ does not make much difference in the final results, as its values would be anyway balanced out, by the optimized λ^*, to achieve the expected margin of $\boldsymbol{w}^\top \lambda \boldsymbol{\Sigma} \boldsymbol{w} = \boldsymbol{w}^\top \lambda^* \boldsymbol{\Sigma}^* \boldsymbol{w}$, i.e., what matters for prior arts is not the design of sampling distribution, but the carefully tuned final inter-label margin in Eq.2.

In contrast, when switching to GOAT, the performance of FRDL is fostered about 1% mA, indicating that the heterogeneously exploited semantics can milden its overfitting to some extent. For a pictorial grasp of it, we exemplify some sampling distributions of ISDA and GOAT in Figure 5(a)-5(b). It shows that the synthetic representations of GOAT are heterogeneously scattered, while those from ISDA simply form repeated gaussian clouds encircling original points. Moreover, GOAT enjoys additional stochasticity from the fact that we run its iterates in a SGD manner. To examine the quality of such stochastic representations, we study the posterior variation along the directions corresponding to the eigenvectors of the heterogeneous cloud. In detail, we use the randomized SVD (Halko et al., 2011) to compute the eigenvectors of the covariance matrix of 1024 translating directions sampled from the heterogeneous cloud of a feature $\boldsymbol{f}_i$. Then, we calculate the attributes posterior as a function of the distance t from $\boldsymbol{f}_i$ along its l-th eigenvector $\boldsymbol{v}_i^l$, and visualize the expectation of it under all features from the training set $\mathbb{E}_{\boldsymbol{f}_i}[|\mathcal{G}(\boldsymbol{f}_i + t \cdot \frac{\boldsymbol{v}_i^l}{\|\boldsymbol{v}_i^l\|}) - \mathcal{G}(\boldsymbol{f}_i)|]$ in Figure 5(c), where $\mathcal{G}(\cdot)$ represents the classifier function to give sigmoid probabilities. It discovers a strong correlation between the main variance of the GOAT features and the attributes informativeness. In other words, GOAT iterates co-linear with the local geometry of semantics transition, and thus ensure the augmented data semantically novel from its initials.

Although GOAT produces meaningful features that are distinct from their originals, the augmented points are still in-distribution. This is because that GOAT solely translates feature along its high-density direction, evident in Figure 5(b) that the initial features predominantly reside at the peripheries of their clouds. To confirm it, we apply (Zhou, 2022) to estimate the probability density function (PDF) of the features in PA100k training data, and utilize this PDF as an in-distribution metric to quantify the non-abnormality of augmented features. Our findings, presented in the Figure 5(d), demonstrate a significant overlap between the PDFs of GOAT features and the inlier features from PA100k test images, indicating that our method is safe in term of not generating outliers. In this regard, GOAT is endued with the Bayesian feature sampling power to give probabilistic representations. A further discussion is in Appendix D.3.

5. Conclusion

We show that label imbalance is the overlooked *grey rhino* that primarily hinders PAR on realistic datasets. We address this long-standing issue by proposing two complementary methods, FRDL and GOAT, to facilitate unprecedented label balancing and ameliorate the consequential semantics imbalance, in a highly unified framework. Comprehensive discussion and experiments underscore our proposals state-of-the-art outperforming and compelling applicabilities: it is generic, lightweight, simple, catering and orthogonal to previous architectural approaches. At a higher level, label imbalance is a thorny problem for numerous multi-label tasks, endorsing our work shedding light not only on PAR, but a wide array of real-world multi-label recognitions so.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62122011 and U2lA20514, in part by the Pioneer and Leading Goose R&D Program of Zhejiang under Grant 2023C01030, and in part by the Fundamental Research Funds for the Central Universities. (Corresponding author: Hai-Miao Hu, frank0139@163.com)

Impact Statement

Our work has positive academic and social impacts as: 1. not exclusive to PAR, our methods solve general limitations in multi-label paradigm; 2. it reduces information leakage of privacy data from being modeled: to achieve the Holy Grail of unbiased intelligence, AI models must minimize the amount of distinctive information they reflect from the underlying datasets, i.e., dataset bias. This is particularly crucial for tasks like PAR, which involve modeling of structured privacy data. Thus, it's essential to prioritize bias minimization to prevent any potential forms of data leakage, particularly concerning specific identity groups. With the proposed methods, we demonstrate in this paper that the pursuit of high-performance PAR actually co-lines with the protection of pedestrian privacy regarding attributes empirical frequency and co-occurring pattern, thus contributing to the acceptability of society to PAR.

References

Bao, L., Wei, L., Qiu, X., Zhou, W., Li, H., and Tian, Q. Learning transferable pedestrian representation from multimodal information supervision. *arXiv preprint arXiv:2304.05554*, 2023.

Cao, Y., Fang, Y., Zhang, Y., Hou, X., Zhang, K., and Huang, W. A novel self-boosting dual-branch model for pedestrian attribute recognition. *Signal Processing: Image Communication*, 115:116961, 2023.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, Q., Jiang, W., Li, K., and Wang, Y. Improving energy-based out-of-distribution detection by sparsity regularization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 539–551. Springer, 2022.

Cormier, M., Specker, A., Junior, J., Jacques, C., Florin, L., Metzler, J., Moeslund, T. B., Nasrollahi, K., Escalera, S., and Beyerer, J. Upar challenge: Pedestrian attribute recognition and attribute-based person retrieval–dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 166–175, 2023.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Deng, Y., Ping, L., Chen, C. L., and Tang, X. Pedestrian attribute recognition at far distance. *ACM*, 2014.

DeVries, T. and Taylor, G. W. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017a.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017b.

Fabbri, M., Calderara, S., and Cucchiara, R. Generative adversarial models for people attribute recognition in surveillance. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pp. 1–6. IEEE, 2017.

Fan, H., Hu, H. M., Liu, S., Lu, W., and Pu, S. Correlation graph convolutional network for pedestrian attribute recognition. *IEEE Transactions on Multimedia*, PP(99): 1–1, 2020.

Fan, X., Zhang, Y., Lu, Y., and Wang, H. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

Guo, H. and Wang, S. Long-tailed multi-label visual recognition by collaborative training on uniform and rebalanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15089–15098, 2021.

Guo, H., Zheng, K., Fan, X., Yu, H., and Wang, S. Visual attention consistency under image transforms for multi-label image classification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Jia, Chen, X., and Huang, K. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 962–971, October 2021a.

Jia, J., Huang, H., Yang, W., Chen, X., and Huang, K. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv*, 2020.

Jia, J., Huang, H., Chen, X., and Huang, K. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021b.

Jia, J., Gao, N., He, F., Chen, X., and Huang, K. Learning disentangled attribute representations for robust pedestrian attribute recognition. pp. 1069–1077. AAAI Press, 2022.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, D., Chen, X., and Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 111–115. IEEE, 2015.

Li, D., Zhang, Z., Chen, X., Ling, H., and Huang, K. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

Li, D., Chen, X., Zhang, Z., and Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. *IEEE*, 2017.

Li, S., Gong, K., Liu, C. H., Wang, Y., Qiao, F., and Cheng, X. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5212–5221, 2021.

Li, W., Cao, Z., Feng, J., Zhou, J., and Lu, J. Label2label: A language modeling framework for multi-attribute learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 562–579. Springer, 2022.

Liu, P., Liu, X., Yan, J., and Shao, J. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018.

Liu, W., Wen, Y., Yu, Z., and Yang, M. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., and Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pp. 350–359, 2017.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Lu, W.-Q., Hu, H.-M., Yu, J., Zhou, Y., Wang, H., and Li, B. Orientation-aware pedestrian attribute recognition based on graph convolution network. *IEEE Transactions on Multimedia*, 2023.

Nguyen, A. T., Lu, F., Munoz, G. L., Raff, E., Nicholas, C., and Holt, J. Out of distribution data detection using dropout bayesian neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7877–7885, 2022.

Specker, A., Cormier, M., and Beyerer, J. Upar: Unified pedestrian attribute recognition and person retrieval. *ArXiv*, abs/2209.02522, 2022.

Tan, Yang, Y., Wan, J., Guo, G., and Li, S. Z. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12055–12062, 2020a.

Tan, Z., Yang, Y., Wan, J., Guo, G., and Li, S. Z. Relation-aware pedestrian attribute recognition with graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12055–12062, 2020b.

Tang, Z. and Huang, J. Drformer: Learning dual relations using transformer for pedestrian attribute recognition. *Neurocomputing*, 497:159–169, 2022.

Wan, W., Zhong, Y., Li, T., and Chen, J. Rethinking feature distribution for loss functions in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9117–9126, 2018.

Wang, J., Zhu, X., and Gong, S. Discovering visual concept structure with sparse and incomplete tags. *Artificial Intelligence*, 250:16–36, 2017.

Wang, X., Zheng, S., Yang, R., Zheng, A., Chen, Z., Tang, J., and Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., and Wu, C. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Weng, D., Tan, Z., Fang, L., and Guo, G. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing*, 531:140–150, 2023.

Wu, J., Huang, Y., Gao, Z., Hong, Y., Zhao, J., and Du, X. Inter-attribute awareness for pedestrian attribute recognition. *Pattern Recognition*, 131:108865, 2022.

Xu, C., Zheng, Y., Zhang, Y., Sun, C., Li, G., and Zhu, Z. Adaptive class-balanced loss based on re-weighting. In *2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pp. 1–8. IEEE, 2022.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Zhang, D., Han, J., Cheng, G., and Yang, M.-H. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021b.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, S., Li, Z., Yan, S., He, X., and Sun, J. Distribution alignment: A unified framework for long-tail visual recognition (supplementary material). 2021c.

Zhang, Y., Wei, X.-S., Zhou, B., and Wu, J. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3447–3455, 2021d.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zheng, A., Wang, H., Wang, J., Huang, H., He, R., and Hussain, A. Diverse features discovery transformer for pedestrian attribute recognition. *Engineering Applications of Artificial Intelligence*, 119:105708, 2023.

Zheng, L., Shen, L., Tian, L., Wang, S., Bu, J., and Tian, Q. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015.

Zhou, Y. Rethinking reconstruction autoencoder-based out-of-distribution detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Zhou, Y., Hu, H.-M., Yu, J., Xu, Z., Lu, W., and Cao, Y. A solution to co-occurrence bias: Attributes disentanglement via mutual information minimization for pedestrian attribute recognition. *arXiv preprint arXiv:2307.15252*, 2023.

A. Overview of Appendix

The appendix is organized via the following contributions:

Appendix B (PAR Datasets) details the adopted datasets and explores our method on additional realistic PAR datasets.

- B.1 introduces important statistics about PETA, RAP, PA100k and **our results on an open-set PAR challenge**.
- B.2 discusses the data leakage of overlapped pedestrian identities between PETA training set and test set.
- B.3 **reports our performance on some realistic datasets**, where the data leakage is well-addressed.

Appendix C (Theoretical Analysis) elucidates the mathematical insights behind GOAT.

- C.1 **researches the regularization effect of high-density oriented feature translating regarding feature noise**.

Appendix D (Further Experiments) examines our method with further experimental results.

- D.1 explains our performance divergence between mA and F1.
- D.2 supplements the experimental settings that underlie our results in Figure 4(a) and Figure 5.
- D.3 **provides additional insight regarding feature augmentation from the Bayesian point of view**.

B. PAR Datasets

B.1. Basic of PETA, RAP and PA100k

PETA, RAP, and PA100k have emerged as three most pre-eminent datasets for PAR, and are widely adopted by leading methodologies in this domain (Jia et al., 2022; Tang & Huang, 2022; Wu et al., 2022; Weng et al., 2023; Zheng et al., 2023; Cao et al., 2023; Bao et al., 2023; Fan et al., 2023; Lu et al., 2023; Zhou et al., 2023). In Table 3, we present the statistics of these datasets.

	PETA	RAP	PA100k
# sample	19,000	41,585	100,000
# attribute	35 (60)	51 (72)	26
# scene	-	26 (indoor)	598 (outdoor)
# tracklet	-	-	18,206
data leakage	✓	✓	✘
resolution	17×39 to 169×365	36×92 to 344×554	50×100 to 758×454

Table 3. Details of the three adopted PAR datasets. The number of attributes denoted within the parentheses signifies the total number of attributes, whereas the numeral presented outside the parentheses represents the experimentally adopted attributes in popular benchmarks. Data leakage indicates whether there is an overlap of pedestrian identities between the training and test set.

PETA. PETA (PEdesTrian Attribute) is introduced by (Deng et al., 2014) as a comprehensive dataset encompassing 19,000 meticulously selected images. These images, sourced from ten publicly accessible small-scale datasets, are annotated with 61 binary attributes and four multi-class attributes. Due to the uneven distribution of certain attributes, only 35 of PETA attributes have been kept for evaluation purposes in popular benchmarks.

RAP. (Li et al., 2016) constructed RAP (Richly Annotated Pedestrian) dataset, specifically RAPv1, which comprises 41,585 pedestrian samples. These samples were captured from a real-world surveillance network consisting of 26 video cameras strategically positioned at a busy shopping mall. RAPv1 dataset features detailed annotations for 69 fine-grained attributes, along with annotations for three critical environmental factors: viewpoints, occlusion styles, and body parts. However, for evaluation purposes, only 51 attributes for RAP are chosen in popular benchmarks, based on their proportion of positive samples present.

PA100k. Further advancing the field, PA100k (Liu et al., 2017) is presented with staggering 100,000 images annotated with 26 attributes. PA100k is one of the most extensive pedestrian attribute dataset to date, making it an invaluable resource for a wide range of pedestrian analysis tasks.

UPAR. The UPAR (Cormier et al., 2023) dataset includes 40 crucial binary attributes spanning across 12 distinct attribute categories, and was integrated from four distinct datasets of PA100k, PETA, RAP and Market-1501 (Zheng et al., 2015). UPAR establishes an open-set benchmark for PAR, by training the models on a restricted set of data from specific datasets and subsequently evaluating their performance using data of previously unseen dataset, respecting the realistic deployment environment of PAR models. As the UPAR test set is not released, we use its training set to re-configure a new dataset **UPAR***: the UPAR re-labeled PA100k, Market-1501 and PETA are employed as training set, while the re-labeled RAP dataset is leaved out as test set. We report our method on UPAR* in Table 4.

Table 4. mA comparison of our method vs. some notable works of VAC (Guo et al., 2020), JLAC (Tan et al., 2020b), L2L (Li et al., 2022), OAGCN (Lu et al., 2023) and PARFormer-B (Fan et al., 2023) on three realistic datasets of UPAR*, PETAzs and RAPzs. We refer the scores of prior arts on PETAzs and RAPzs from (Jia et al., 2021b) and (Zhou et al., 2023). The results of PARFormer-B are produced by the public code from its original literature. We also denote as the subscript of our score the relative improvement over the highest existing method.

Dataset	VAC	JLAC	L2L	OAGCN	PARFormer	**Ours**
UPAR*	69.85	-	70.44	-	72.58	**79.40**$_{+6.82}$
PETA	84.58	86.88	87.07	88.21	**88.65**	88.59$_{-0.06}$
PETAzs	71.91	73.60	72.13	75.44	76.16	**79.10**$_{+2.94}$
RAP	80.27	81.51	81.93	86.02	83.84	**87.72**$_{+1.18}$
RAPzs	73.70	76.38	73.84	76.20	77.24	**83.69**$_{+6.45}$

B.2. Data Leakage in PETA

A notable constraint in PETA dataset pertains to the partitioning of training and test sets (Jia et al., 2021b). Specifically, PETA images are randomly allocated to either set without any regard for pedestrian identity. Consequently, this haphazard approach in both image acquisition and dataset division results in a great overlap of highly similar images between the training and test sets, with only minor variations in background and pose. This phenomenon, commonly known as "data leakage," poses a challenge in accurately assessing model generalization capabilities, and renders evaluated methods significantly over-estimated.

Similar issue also exists in RAP, however, with a relatively less proportion. While for PA100k dataset, it addressed this issue by assigning all images of a single pedestrian exclusively to either the training or test set.

B.3. Realistic Results on PETAzs and RAPzs

For reliable performance evaluations of PAR models on PETA and RAP dataset, (Jia et al., 2021b) undertook reorganization of them and created zero-shot variants dubbed PETAzs and RAPzs. PETAzs and RAPzs adhere strictly to the zero-shot paradigm for pedestrian identities, ensur-

ing no overlap between training and test sets in terms of identities. Subsequently, we have documented our methods mA on these re-configured realistic datasets in Table 4. Our method demonstrates significant superiority over previous works when the issues of data leakage in RAP and PETA are addressed, exhibiting notable margins of improvement.

C. Theoretical Analysis

C.1. GOAT Regularizes the Feature Noise of FRDL

For brevity, we confine following discussion to the recognition of a single attribute, and the conclusion drawn herein can be readily extrapolated to more conventional instances pertaining to multi-hot labeling. Formally, the *Stage#2* of FRDL re-samples features on a dataset $\{(\boldsymbol{f}_i, z_i)\}_{i=1}^N$, where $\boldsymbol{f}_i$ represents a cached feature to be sampled for the classifier fine-tuning, and z_i is its ground true binary label of a certain attribute. Practically, as z_i is unknown, we use the label y_i that corresponds to its input image $\boldsymbol{x}_i$ as a proxy, and apply a surrogate dataset $\{(\boldsymbol{f}_i, y_i)\}_{i=1}^N$ for the implemental feature re-sampling in FRDL. However, since features are extracted with a failure rate σ, there could be a faulty label assignment of $y_i \neq z_i$ to $\boldsymbol{f}_i$. Finally, the classifier is re-trained on a polluted dataset, and tends to exhibit poor generalization owing to being misled by the spurious samples that diverge from the true joint distribution of $\{(\boldsymbol{f}_i, z_i)\}_{i=1}^N$. Assuming that the feature noise ratio σ is only label-dependent, above process indicates a conditional probability distribution $P(Y|Z)$ with $P(Y = z|Z = z) = 1 - \sigma$ and $P(Y \neq z|Z = z) = \sigma$. Thus, if we minimize the BCE loss of PAR to re-train a classifier denoted by W, we have

$$
\begin{aligned}
&-\frac{1}{N}\sum_{i=1}^{N}\log P(Y=y_i|\boldsymbol{f}_i;W)\\
=&-\frac{1}{N}\sum_{i=1}^{N}\log\sum_{z_i\in\{y_i,\neq y_i\}}P(Y=y_i,Z=z_i|\boldsymbol{f}_i;W)\\
=&-\frac{1}{N}\sum_{i=1}^{N}\log\sum_{z_i\in\{y_i,\neq y_i\}}P(Y=y_i|Z=z_i)P(Z=z_i|\boldsymbol{f}_i;W)\\
=&-\frac{1}{N}\sum_{i=1}^{N}\log((1-\sigma)P(Z=y_i|\boldsymbol{f}_i;W)+\sigma P(Z\neq y_i|\boldsymbol{f}_i;W))\\
=&-\frac{1}{N}\sum_{i=1}^{N}\log((1-\sigma)P(Z=y_i|\boldsymbol{f}_i;W)\\
&\qquad\qquad+\sigma(1-P(Z=y_i|\boldsymbol{f}_i;W))).
\end{aligned}
\tag{6}
$$

By taking the derivatives of Eq.6, the BCE loss is minimized when $P(Z = y_i|\boldsymbol{f}_i; W) = 1 - \sigma$. It implies that, to relieve the classifier from further overfitting on the noisy features, the following training objective should be minimized w.r.t. W to encourage $P(Z = y_i|\boldsymbol{f}_i; W)$ to take $1 - \sigma$,

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}|-\log P(Z=y_i|\boldsymbol{f}_i;W)-(-\log(1-\sigma))|\\
=&\frac{1}{N}\sum_{i=1}^{N}|-\log P(Z=y_i|\boldsymbol{f}_i;W)\\
&\qquad\qquad-(-\log\mathbb{E}_{\boldsymbol{f}}[P(Z=y|\boldsymbol{f};W^*)])|\\
\geq&\frac{1}{N}\sum_{i=1}^{N}|-\log P(Z=y_i|\boldsymbol{f}_i;W)\\
&\qquad\qquad-\mathbb{E}_{\boldsymbol{f}}[-\log P(Z=y|\boldsymbol{f};W^*)]|.
\end{aligned}
\tag{7}
$$

As $1 - \sigma$ denotes the success rate of a feature extractor, it could be estimated as the highest $\mathbb{E}_{\boldsymbol{f}}[P(Z = y|\boldsymbol{f})]$ achievable by a classifier (denoted by W^*) on the corresponding features. Considering that neural models lean towards prioritizing fitting clean data before noisy ones (Zhang et al., 2021a), we simply apply the optimized classifier trained with early stopping in the *Stage#1* of FRDL as an approximate of W^*. After applying Jensen inequality, the optimum feature de-noising objective is exactly an upper bound of our high-density-translating loss in GOAT, regardless that the latter is represented for multi-hot labeling in Eq.4.

D. Further Experiments

D.1. Inconsistency between mA and F1

Table 1 highlights that our method does not produce F1 scores on par with those of mA. This discrepancy (or inconsistency) between the two evaluation metrics is not unique to our approach. As can be seen in Table 1, many methodologies that excel in mA also tend to exhibit lower F1 scores. A similar trend has also been summarized in the UPAR challenge (Cormier et al., 2023), where PAR methods prevailing in terms of mA often falter when assessed using F1. The underlying rationale behind this phenomenon is that mA assigns equal importance to both positive and negative samples when evaluating an attribute, whereas F1 primarily emphasizes the recognition precision of positive labels since it is an instance-based metric:

$$
\begin{aligned}
Prec &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i},\\
Recall &= \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i},\\
F1 &= \frac{2\cdot Prec\cdot Recall}{Prec+Recall},\\
mA &= \frac{1}{C}\sum_{j=1}^{C}\frac{1}{2}(\frac{TP^j}{TP^j+FN^j}+\frac{TN^j}{TN^j+FP^j}),
\end{aligned}
$$

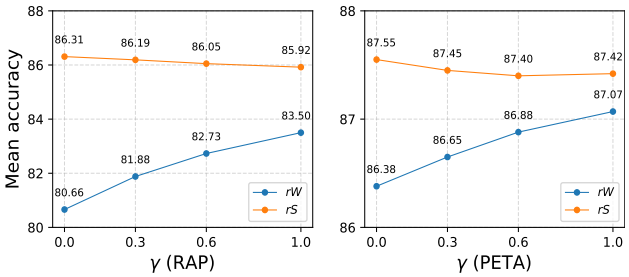


Figure 6. Representation quality of feature extractor as a function of the label-balancing ratio γ for RAP and PETA, respectively. Implementation details are identical to those for Figure 4(a).

where TP_i, FP_i, FN_i are the total number of true positive, false positive and false negative attributes of i-th sample, and TP^j, TN^j, FP^j, FN^j are the number of true positive, true negative, false positive and false negative samples of j-th attribute. Since every attribute should be regarded as equally important (so as its positive and negative label), mA is typically deemed to have greater practical implication.

D.2. Experimental Details in Ablations

In Figure 4(a), we gradually modified the instance-balanced feature extractor towards that trained with label-balancing technique, and by a comprehensive analysis of the variation of feature quality during this transition, we have concluded that the FRDL method serves as a superior alternative to the LIR approach. Specifically, we applied following weighted BCE loss in the training of all backbones:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{C} w_j^i(\boldsymbol{y}_i^j \log \boldsymbol{p}_i^j + (1-\boldsymbol{y}_i^j)\log(1-\boldsymbol{p}_i^j)),$$

$$s.t.\ \ w_j^i = \begin{cases} e^{1-(\gamma(r_j-0.5)+0.5)}, & \boldsymbol{y}_i^j = 1 \\ e^{\gamma(r_j-0.5)+0.5}, & \boldsymbol{y}_i^j = 0 \end{cases}$$

where $\boldsymbol{p}_i$ is the estimated attributes posterior of $\boldsymbol{x}_i$, r_j the label mean of attribute j, and γ the label balancing ratio that transitions from 0 to 1 to study the impact of label-balancing having on feature extractors. When γ is 0, it respects the instance-balanced learning as no label balancing would be exerted. In Figure 6, we present further experimental results on PETA and RAP datasets, and they exhibit similar trend of variation as that of PA100k in Figure 4(a).

For the features PDF in Figure 5, we followed (Zhou, 2022) to train a feature reconstructor on all activation vectors extracted from PA100k training set. Next, we fit two Weibull distributions on the tail of training-set features reconstruction residual and sigmoid confidence scores, respectively. Finally, we use the Weibulls product as the final feature normality measure. Details are identical to (Zhou, 2022; Chen et al., 2022).

D.3. Bayesian Inference as Feature Augmentation

Table 5. We compare GOAT with popular data augmentation approaches of AutoAug (Cubuk et al., 2019) (ImageNet policies), Mixup (Zhang et al., 2017) and Cutout (DeVries & Taylor, 2017b), to research its break-down effect as an independent data augmentation method for PAR. GOAT$^+$ represents GOAT enhanced by the dropout variational inference.

Dataset	Baseline	AutoAug	Mixup	Cutout	GOAT	GOAT$^+$
PA100k	82.45	82.55	82.07	81.09	84.16	**84.50**
PETAzs	75.01	75.22	74.50	73.83	76.41	**76.66**
RAPzs	76.14	76.43	75.79	73.56	77.98	**78.41**

Bayesian approaches quantify uncertainty by assigning a probability distribution to model parameters, and are prevalent in generative modeling frameworks that rely on variational inference (Kingma & Welling, 2013), or out-of-distribution detection and uncertainty estimation (Nguyen et al., 2022). Nevertheless, to our knowledge, this study represents a pioneering effort in utilizing the probabilistic characteristic of Bayesian inference to offer additional in-distribution variation for feature augmentation. In Table 5, we juxtapose GOAT against several prevalent data augmentation strategies. The findings underscore that image augmentation techniques, such as Mixup and Cutout, can potentially corrupt the fine-grained signatures of attributes in the pixel domain, thereby leading to decreased performance in PAR. Conversely, GOAT prevails by manipulating data within the latent space, affirming the indispensability of Bayesian feature augmentation in PAR.

Equipped with Bayesian perspective, the GOAT framework can be further extended by incorporating other Bayesian methodologies. In Table 5, we demonstrate the application of dropout variational inference (Gal & Ghahramani, 2016; Gal et al., 2017), which utilizes a spike and slab variational distribution to interpret dropout during the testing phase as an approximation of Bayesian inference, to offer further randomness over the process of feature augmentation.