A Solution to Co-occurence Bias: Attributes Disentanglement via Mutual Information Minimization for Pedestrian Attribute Recognition

Yibo Zhou[1] , **Hai-Miao Hu**[1,2] , **Jinzuo Yu**[1] , **Zhenbo Xu**[2] , **Weiqing Lu**[1] and **Yuran Cao**[1]

[1]State key laboratory of virtual reality technology and systems, Beihang University
[2]Hangzhou Innovation Institute, Beihang University
{ybzhou, hu, 17377133, 18241004}@buaa.edu.cn, xuzhenbo@mail.ustc.edu.cn, 574168985@qq.com

Abstract

Recent studies on pedestrian attribute recognition progress with either explicit or implicit modeling of the co-occurence among attributes. Considering that this known a prior is highly variable and unforeseeable regarding the specific scenarios, we show that current methods can actually suffer in generalizing such fitted attributes interdependencies onto scenes or identities off the dataset distribution, resulting in the underlined bias of attributes co-occurence. To render models robust in realistic scenes, we propose the attributes-disentangled feature learning to ensure the recognition of an attribute not inferring on the existence of others, and which is sequentially formulated as a problem of mutual information minimization. Rooting from it, practical strategies are devised to efficiently decouple attributes, which substantially improve the baseline and establish state-of-the-art performance on realistic datasets like PETAzs and RAPzs.

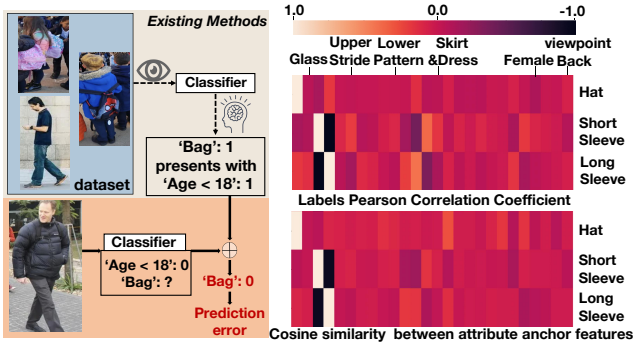

Figure 1: Left: Illustration of a prediction failure resulting from biased attributes interdependency. Right: The Pearson correlation coefficient (Upper) of training set labels, and the cosine similarity between anchor features (Lower), for hat and short/long sleeve *vs.* other annotated attributes in PA100k. It reveals that the network precisely captured the data selection bias that Glass & ShortSleeve, Viewpoint-Back & ShortSleeve and Hat & Female, etc. are not prone to present simultaneously, and uses such biased correlations for attributes inference. The weights of the last fully-connected (FC) layer in a Resnet-50 are applied as attribute anchor features.

1 Introduction

Pedestrian attribute recognition (PAR), as a key component of the pedestrian analysis stemming from development of the ubiquitous video surveillance, targets to determine the soft-biometrics of local or semantic attributes of a person given its captured main-body image. To date, researches investigate PAR basically along the analogous routine of discriminative deep multi-label classification. They can be basically abstracted as emphasizing on better attribute localization to mitigate the accuracy drop from inferring on irrelevant area [Liu *et al.*, 2018; Jia *et al.*, 2022], or involving additional supervision or information, like pose keypoints [Liu *et al.*, 2018] and pedestrian video clip [Chen *et al.*, 2019; Ji *et al.*, 2020], to guide PAR under explicit assumptions regarding body topological structure or temporal context, etc..

For work striving to enhance the attribute localization, [Fabbri *et al.*, 2017; Li *et al.*, 2017] split the body image vertically, by a manually defined fixed strategy, into three parts and feeds each part into an individual network for feature extracting. As a further step, [Liu *et al.*, 2017] proposed the HydraPlus-Net, learning to locate attributes of different scale with a multi-directional attention modules. Instead of learning to generate the attention map, [Liu *et al.*, 2018] produces attribute-specific features by means of the class activation map (CAM) [Zhou *et al.*, 2016], which is initially designed to build a generic localizable deep representation. However, same technique is applied in [Jia *et al.*, 2021b] to demonstrate that, even without explicitly modeling the attribute-specific area, network is still able to locate attributes precisely, implying that the fundamental task for PAR should be instead to explore better feature learning.

Seeking for better feature learning of information exchanges, a thread of literature [Wang *et al.*, 2016; Wang *et al.*, 2017; Zhu *et al.*, 2017; Fan *et al.*, 2020; Li *et al.*, 2022] aims to enhance the modeling of interdependencies among attributes, under the hypothesis that such correlations provide a contextual constraint complementary to visual attributes recognition. To this aim, Graph-based methods are often used to explicitly model the attributes co-occurrence and estimate the joint label probability. However, the strong variability and unpredictability exhibited by attributes co-occurrence actually cast doubts on the robustness of these methods for sce-

narios out of the datasets, which are typically in practical applications. Worse, given the observation in Figure.1 that, such unreliable interdependency can be memorized by network, in a form of bias, even without explicit modeling of it, current methods would suffer in generalizing well onto realistic PAR, as evidenced in [Jia *et al.*, 2021b].

Equipped with such perspective, this paper evolves in a novel spirit that we resort to infer attributes while entirely discarding their relations. Actually, it is a mechanism of PAR that accords with us human, say, we do not infer one's hat color by referring to even a single clue from its gender, instead, we look just at the hat for robustness, so should intelligent models. To embody this philosophy in stark contrast to the fragile mechanism of deep models for PAR, our attribute-disentangled learning for PAR is formalized as that the attribute-specific feature learned for predicting one attribute should not use information pertinent to other attributes, i.e., the mutual information between one attribute and other attributes' specific feature should be minimized.

To cope with the non-triviality of a direct mutual information minimization, an equivalent optimization-friendly training objective of it is deduced, which guarantees that the variation of other attributes' information results in no shift on the estimated posterior of a given attribute. Both mathematical insights and experimental evidence are provided, indicating that under this setting of the posterior-invariant learning, the proposed disentangled attribute learning can be attained, and thus tackle the issue of attribute co-occurrence bias.

Sequentially, we also propose an efficient training strategy for the posterior-invariant learning, which enables it implemented in a manner of fast convergence. Practically, unlike most work in this field building on large networks that deteriorates the applicability, our method's plug-and-play nature makes it cater to various models, with almost no extra computational burden thereon. Albeit lightweight, our method establishes the-state-of-art performance on various realistic datasets like the PETAzs and RAPzs [Jia *et al.*, 2021b], with considerable margins over previous approaches.

Our contribution is summarized as three folds:

- With solid experimental evidences, we establish a novel perspective for understanding the learned attributes interdependency bias as the current bottleneck of PAR from achieving robustness. It is not covered by existing work and can serve as a mindset of future researches for rethinking PAR.
- We propose one direction of improvement as to infer the attributes by disregarding their correlations. From it, a lightweight prescription of information-theoretic attributes-disentangled feature learning is developed.
- Along with ablation studies, we present analytical experiments on various realistic PAR datasets to demonstrate our generic proposal's efficacy and a spectrum of superiorities, validating that the proposed method might serve as a foothold of robust PAR.

2 On the Attributes Co-occurence Bias

Fundamentally, although there exists certain pattern of interdependencies among attributes, such a phenomenon es-

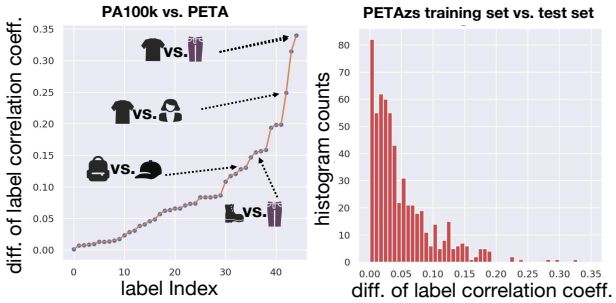

Figure 2: Left: For the 10 attributes common between PA100k and PETA, the absolute difference of labels Pearson correlation coefficients computed on these two datasets. For better viewing, all values are arranged in an increasing order. Right: On the PETAzs dataset of zero-shot setting (no overlapped identities of training set in test set), distribution of the difference between attributes Pearson correlation coefficients on training set and test set.

sentially roots from **conditioned statistical relevance** rather than **causality**, implying that it can vary from scene to scene, individual to individual drastically. Thus, such variability and unpredictability inherent to the attributes co-occurrence make it hardly a universal known a prior to safely rely on, and could be immensely biased regarding the statistics of limited scenarios, e.g., some datasets exclusively involve indoor or outdoor scenes and constant season or whether condition, race or culture, etc. [Li *et al.*, 2016; Liu *et al.*, 2017].

Such a claim is supported by the observations in Figure.2 that, the correlations among attributes are likely to fluctuate between different datasets like PA100k [Liu *et al.*, 2017] and PETA [Deng *et al.*, 2014], or even mutate on different groups of identities from the same scenes. For instances, ShortSleeve and Trouser lean to not co-occure in PA100k, as images in this dataset are mostly captured during summer time. While for PETA, ShortSleeve and Trouser appear simultaneously with a significantly larger ratio, resulting in the interdependency discrepancy of these two attributes. Also, even for two groups of pedestrians from a same dataset PETA, about 1/6 attributes correlations exhibit distinct characteristic (absolute difference between Pearson correlation coefficients ≥ 0.1).

As a result, empirical risk dominates in the way that, existing work would have difficulty in generalizing the explicitly or implicitly leveraged attributes co-occurrence to other circumstances with different pattern of attributes interdependencies. Since an ideal dataset collection of pedestrian images, which captures global facets of the non-static myriad population distribution of attributes co-occurence, to soften such underlined bias, can be intractable, to enhance a robust PAR, a disentangled and discriminative feature learning for each attribute can be indispensable and consequential.

3 Method

Attributes Disentanglement by Mutual Information Minimization. Here, we establish the theoretical framework of our methods. Formally, it is supposed that there is a distribution X characterized by all of the pedestrian images. Under certain conditions, some data points $\{\boldsymbol{x}_i\}_{i=1}^{N}$

are sampled from X, with their corresponding annotations $\{\boldsymbol{a}_i\}_{i=1}^N$ of some predefined attributes, to form the train set $D = \{\boldsymbol{x}_i, \boldsymbol{a}_i\}_{i=1}^N$, where $\boldsymbol{a}_i \in \{0,1\}^C$ and C is the total number of attributes. Specifically, if the latent embedding output from the feature extractor is denoted as $\boldsymbol{f} \in R^K$, we hope to decompose it as $\boldsymbol{f} = \boldsymbol{f}^1 + \boldsymbol{f}^2 + ... + \boldsymbol{f}^C$, where$\boldsymbol{f}^s \in R^K, s = 1, 2, ..., C$, is the attribute-specific feature learned for predicting the attribute indicating random variable y^s. From the information theory point of view, the mutual information $\mathcal{I}(y^k; \boldsymbol{f}^s)$ measures the knowledge that could be told from the random variable $\boldsymbol{f}^s$ about the random variable y^k. Thus, to ensure the prediction of each attribute independent to the existence of others, for every $\boldsymbol{f}^s$, it should be satisfied that the mutual information $\mathcal{I}(y^k; \boldsymbol{f}^s) = 0$, for any $k = 1, 2, ..., C, k \neq s$, which can be factorized as

$$\begin{aligned} &\mathcal{I}(y^k; \boldsymbol{f}^s) = \mathcal{H}(y^k) - \mathcal{H}(y^k|\boldsymbol{f}^s) \\ =&\mathbb{E}_{\boldsymbol{f}^s \sim \mathcal{F}^s}[\int P(y^k|\boldsymbol{f}^s) \log P(y^k|\boldsymbol{f}^s)\, dy^k] \\ &- \int P(y^k) \log P(y^k)\, dy^k = 0, \end{aligned} \tag{1}$$

where $\mathcal{F}^s$ is the marginal distribution of $\boldsymbol{f}^s$. Noticably, if $P(y^k|\boldsymbol{f}^s) = P(y^k)$, Eq.1 naturally holds since $\int P(y^k) \log P(y^k)\, dy^k$ is independent of $\boldsymbol{f}^s$. Obviously, for any $\boldsymbol{f}^s$ drawn form $\mathcal{F}^s$, if it satisfies that $P(y^k|\boldsymbol{f}^s) = \mathcal{Q}$ and $\mathcal{Q}$ is a constant, we have

$$\begin{aligned} P(y^k) &= \int P(\boldsymbol{f}^s) P(y^k|\boldsymbol{f}^s)\, d\boldsymbol{f}^s \\ &= \int P(\boldsymbol{f}^s) \mathcal{Q}\, d\boldsymbol{f}^s = \mathcal{Q}. \end{aligned} \tag{2}$$

Eq.2 reveals that, to satisfy $P(y^k) = P(y^k|\boldsymbol{f}^s)$ is equivalent to ensure that for any $\boldsymbol{f}_a^s$ and $\boldsymbol{f}_b^s$ sampled from $\mathcal{F}^s$, it holds that $P(y^k|\boldsymbol{f}^s = \boldsymbol{f}_a^s) = P(y^k|\boldsymbol{f}^s = \boldsymbol{f}_b^s)$. Note that

$$\begin{aligned} &P(y^k|\boldsymbol{f}^s = \boldsymbol{f}_a^s) - P(y^k|\boldsymbol{f}^s = \boldsymbol{f}_b^s) \\ =&\int (P(y^k|\boldsymbol{f}_a^s + \sum_{l=1, l\neq s}^{C} \boldsymbol{f}^l) - P(y^k|\boldsymbol{f}_b^s + \sum_{l=1, l\neq s}^{C} \boldsymbol{f}^l)) \\ &\cdot P(\boldsymbol{f}^1, \boldsymbol{f}^2, .., \boldsymbol{f}^{s-1}, \boldsymbol{f}^{s+1}, .., \boldsymbol{f}^C) \prod_{l=1, l\neq s}^{n} d\boldsymbol{f}^l. \end{aligned}$$

With this formulation, if the estimate of $P(y^k|\boldsymbol{f}) = P(y^k|\sum_{l=1}^C \boldsymbol{f}^l)$ is invariant of $\boldsymbol{f}$'s component $\boldsymbol{f}^s$, for any $s \neq k$, then we have $\mathcal{I}(y^k; \boldsymbol{f}^s) = 0$. Thus, a statement can be declared as follows

> **Goal**: attribute-disentangled learning can be enabled as long as the probability estimate of $P(y^k|\boldsymbol{f})$ varies only with the y^k's specific feature $\boldsymbol{f}^k$.

Approach of Posterior-invariant Leanring. In practice, for the feature of attributes hybrid, extracted from a given input $\boldsymbol{x}_i$, we translate it first to obtain each attribute-specific components $\boldsymbol{f}_i^s, s = 1, 2, ..., C$, by a FC layer followed with nonlinearities, and $\boldsymbol{f}_i$ is obtained by adding them up as $\boldsymbol{f}_i = \sum_{s=1}^C \boldsymbol{f}_i^s$. Sequentially, we randomly build C mappings $\mathcal{G}_i^s(\cdot) : \mathcal{F}^s \mapsto \mathcal{F}^s$ to transform each $\boldsymbol{f}_i^s$ into an arbitrary data point $\tilde{\boldsymbol{f}}_i^s$ within $\boldsymbol{f}^s$'s domain of distribution $\mathcal{F}^s$, and generate a new feature vector of $\tilde{\boldsymbol{f}}_i = \sum_{s=1}^C \tilde{\boldsymbol{f}}_i^s$, where $\tilde{\boldsymbol{f}}_i^s = \mathcal{G}_i^s(\boldsymbol{f}_i^s)$. We require $P(y^s = y_i^s|\tilde{\boldsymbol{f}}_i - \tilde{\boldsymbol{f}}_i^s + \boldsymbol{f}_i^s) = P(y^s = y_i^s|\boldsymbol{f}_i)$ to meet the goal of the proposed attribute-disentangled learning that $P(y^s|\boldsymbol{f})$ should be tolerant to the variation of other attributes' specific features, and therefore to ensure no factors for decreasing the uncertainty of y^s are conveyed by $\boldsymbol{f}^k, \forall k \neq s$. Under the context of supervised PAR, $P(y^s = y_i^s|\boldsymbol{f}_i)$ is simply the ground truth label a_i^s coupled with the input $\boldsymbol{x}_i$, and thereby the underlined attribute-disentangled learning can be enabled by leveraging the information-theoretic regularizer of

$$\begin{aligned} \min -\sum_{i=1}^{N}\sum_{s=1}^{C} a_i^s \log \widehat{P}(y^s = y_i^s|\tilde{\boldsymbol{f}}_i - \tilde{\boldsymbol{f}}_i^s + \boldsymbol{f}_i^s), \\ s.t.\ \tilde{\boldsymbol{f}}_i = \sum_{s=1}^{C} \tilde{\boldsymbol{f}}_i^s,\ and\ \tilde{\boldsymbol{f}}_i^s = \mathcal{G}_i^s(\boldsymbol{f}_i^s). \end{aligned} \tag{3}$$

$\widehat{P}$ is the estimate of $P(y^k|\boldsymbol{f})$ given by a discriminative classifier, which is a common technique for approximating a posterior, and plugged right onto $\boldsymbol{f}$ with MLP operations in it. This framework is graphed in Figure.3a.

To specify the form of $\mathcal{G}_i^s(\cdot)$, we opt for simplicity by adopting the convex linear combination between $\boldsymbol{f}_i^s$ and $\boldsymbol{f}_{r(i,s)}^s$, which is the extracted feature from a randomly picked training sample $\boldsymbol{x}_{r(i,s)}$, as a plausible form among possible variants that satisfy the closureness of a transform, i.e., the mapped feature still lies within the domain of $\mathcal{F}^s$. Here, $r(i, s)$ is simply a function that maps different $i\&s$ to a different sample index. It might be noticed that, with the convex combination of two attribute-specific features from training samples, points off of the training distribution can be generated. However, [Mikolov *et al.*, 2013] has shown that this linear interpolation between hidden states is an effective way of transiting between learned factors to produce new in-distribution semantics, making the combined features still informative of attributes thus reside within the $\mathcal{F}^s$. Specifically,

$$\mathcal{G}_i^s(\boldsymbol{f}_i^s) = \begin{cases} (\beta^s \|\boldsymbol{f}_i^s\| + (1-\beta^s)\|\boldsymbol{f}_{r(i,s)}^s\|) \\ \cdot(\alpha^s \frac{\boldsymbol{f}_i^s}{\|\boldsymbol{f}_i^s\|} + (1-\alpha^s)\frac{\boldsymbol{f}_{r(i,s)}^s}{\|\boldsymbol{f}_{r(i,s)}^s\|}) & if\ \ a_i^s = a_{r(i,s)}^s \\ \\ \alpha^s \boldsymbol{f}_i^s + (1-\alpha^s)\boldsymbol{f}_{r(i,s)}^s & otherwise. \end{cases}$$

α^s and β^s are independently drawn from the uniform distribution $\mathcal{U}(0, 1)$ for each attribute. To fully explore $\mathcal{F}_i$ and produce more of the points with semantics potentially to be presented in the capricious test environment, for attribute-specific features with identical label, we do the linear combinations to their norms and direction vectors, respectively. The rationale is that, since feature norm has been validated to be an effective measure of data uncertainty [Shalev *et al.*, 2018;

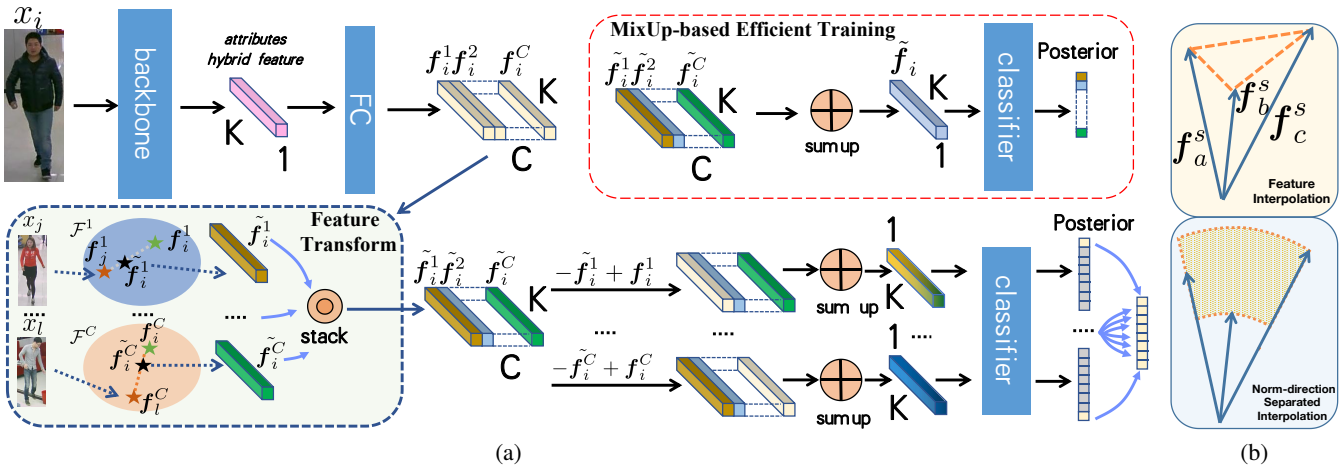

Figure 3: (a): Overall pipeline of the proposed method for attribute-disentangled feature learning. For Eq.4, the transformed attribute-specific features are fed into another branch within the red-dashed box. (b): Comparison of direct feature interpolation (Upper) and the proposed norm-direction separated interpolation (Lower), in term of the domain exploration of feature distribution. Given three attribute-specific features of same label, by direct feature interpolation only points on the orange line are regarded as possible variations of them. Whilst for norm-direction separated interpolation, in a reliable manner, the potential variations are expanded into the whole orange-shaded area.

Meng *et al.*, 2021; Zhou, 2022] that is prone to be fluctuated by bad image quality (often in PAR), the style semantics of attribute (color, size, shape, etc.) might be mostly encoded into feature's direction, i.e., it's normalized unit feature [Wang *et al.*, 2018]. Hence, such a norm-direction-separated linear combination can generate features from attributes of one style but shot under various uncertainties, exploiting more of the domain of $\mathcal{F}_s$, exemplified in Figure.3b.

Efficient MixUp-based Alternative. Notably, directly optimizing over Eq.3 necessitates C inferences though the classifier for each sample in every update, which can be computationally inefficient when the number of attributes grows larger. Regardless of the ease for implementing $\mathcal{G}(\cdot)$ as linear combination, such a design enjoys another merit as it actually supplies us with an approach to integrate the training objective of Eq.3 into a desirable single-inference-every-instance manner of

$$\begin{aligned} \min -\sum_{i=1}^{N}\sum_{s=1}^{C}(\alpha^s a_i^s + (1-\alpha^s)a_{r(i,s)}^s) \\ \cdot \log \widehat{P}(y^s = \alpha^s y_i^s + (1-\alpha^s)y_{r(i,s)}^s | \tilde{\boldsymbol{f}}_i), \qquad (4) \\ s.t.\ \ \tilde{\boldsymbol{f}}_i = \sum_{s=1}^{C} \tilde{\boldsymbol{f}}_i^s, \quad and \quad \tilde{\boldsymbol{f}}_i^s, \alpha^s, a_{r(i,s)}^s = \mathcal{G}_i^s(\boldsymbol{f}_i^s). \end{aligned}$$

Such a simplification from Eq.3 to Eq.4 is inspired by the well-known data augmentation method MixUp [Zhang *et al.*, 2017], which shows that the linear combination of images is actually aligned with exactly the same linear combination of their ground truth in the label space. Here, we adopt an analogous idea into Eq.4 to reduce the training cost.

Please note that Eq.4 differs from MixUp in two aspects. First, during every update, MixUp uses a single α and a single sample-to-interpolate for each input, purely for the sake of data augmentation. Whereas in Eq.4, for each attribute's specific feature, we use feature interpolation of different α^s, β^s and $\boldsymbol{f}_{r(i,s)}^s$ to guarantee that the transform imposed over it is totally independent thus differentiated from others. By requiring the variation of a certain attribute's posterior aligned exclusively to the variation of its specific feature, we enforce other attribute's specific feature acquires no factors informative to the given attribute's posterior, exactly the same way for realizing attribute-disentangled learning by Eq.3. Second, instead of pixel space, the linear combination is conducted in the space of decoupled attribute-specific features. One might argue that such a strategy can bring non-trivial boost in performance by somewhat working as a data augmentation of MixUp. We admit this concern. However, we will present in the experiment section that it is actually not the case here for achieving SOTA performance as MixUp can not foster any accuracy increase over the baseline. More importantly, as will be presented in the experiment section, the equivalent Eq.3 suffices to deliver similar performance w.r.t Eq.4, which does not optimize over the interpolated feature by its correspondingly interpolated label.

In a nutshell, attribute-disentangled learning is secured by our method from two sides. One, given the specific feature of all attributes, we render the posterior estimate of a certain attribute exclusively associated to the feature component of its own, tasking the rest specific features, on which other attributes are inferred, to preserve no clues informative about this given attribute. Second, serving as an additional virtue of Eq.4, the biased attribute interdependencies are strongly dismissed from classifier learning owing to the scheme that the ground truth employed in Eq.4 is randomly generated for each attribute, and therefore can be of any possible pattern in term of attributes co-occurence, making classifier no way to trace on the attribute correlations embedded in the limited dataset, and thus generalize better. Efficiently, both sides are

Method	Backbone	PA100k			RAP			RAPzs			PETA			PETAzs		
		mA	Recall	F1	mA	Recall	F1	mA	Recall	F1	mA	Recall	F1	mA	Recall	F1
Baseline ('21)	Resnet-50	80.38	87.01	87.05	80.32	79.89	79.46	72.32	76.62	76.75	84.42	85.08	85.97	71.62	70.33	71.68
MsVAA ('18)	Resnet-50	80.41	86.52	86.80	78.86	79.15	79.27	72.04	75.81	75.74	84.35	85.51	86.09	71.53	69.42	71.94
MTMS ('19)	Resnet-50	-	-	-	82.45	80.44	65.33	-	-	-	86.23	87.22	85.85	-	-	-
VAC ('20)	Resnet-50	79.16	86.26	87.59	80.27	79.77	78.36	73.70	76.97	76.12	83.63	85.45	86.23	71.91	70.64	70.90
*JLAC ('20)	Resnet-50	82.31	87.77	87.61	83.69	82.40	**80.82**	76.38	79.20	76.05	86.96	87.09	**87.45**	73.60	**72.41**	72.05
SSC ('21)	Resnet-50	81.87	89.10	86.87	82.77	**87.49**	80.43	-	-	-	86.52	87.12	86.99	-	-	-
DAFL ('22)	Resnet-50	83.54	89.19	**88.90**	83.72	83.39	80.29	-	-	-	87.07	87.03	86.40	-	-	-
Label2Label ('22)	Resnet-50	82.24	88.57	87.08	-	-	-	73.84	78.15	77.75	-	-	-	72.13	71.10	71.74
Ours	Resnet-50	84.53	89.13	87.01	83.26	83.31	79.88	76.71	80.18	77.93	86.41	86.80	87.03	74.35	72.09	72.39
***Ours**	Resnet-50	**85.12**	**89.40**	87.85	**83.88**	83.65	80.73	**77.49**	**80.42**	**78.30**	**87.18**	**87.29**	87.41	**75.13**	71.98	**72.49**
Baseline ('22)	ConvNeXt-base	82.2-	-	88.5-	-	-	-	76.54	81.47	80.25	86.1-	-	88.1-	75.21	74.43	75.62
ALM ('19)	BN-inception	80.68	88.84	86.46	81.87	86.48	80.16	74.28	80.73	76.65	84.24	85.60	85.41	73.01	73.69	71.53
MTA-Net ('20)	Resnet-152	-	-	-	77.62	78.44	79.07	-	-	-	84.62	86.42	86.04	-	-	-
AR-BiFPN ('20)	EfficientNet-B3	81.45	89.46	87.94	82.37	**87.23**	**82.33**	-	-	-	87.69	**89.20**	88.32	-	-	-
Ours	ConvNeXt-base	**88.11**	**91.51**	**89.13**	**85.05**	84.11	81.12	**80.18**	**83.51**	**80.36**	**88.12**	88.76	**88.54**	**78.54**	**76.34**	**75.91**

Table 1: Benchmark results in RAP, RAPzs, PETA, PETAzs and PA100k. Our method is compared with various notable SOTA methods. To make a fair comparison, for RAP, PETA and PA100k, we adpot the baseline results of Resnet-50 and ConvNeXt-base respectively from [Jia *et al.*, 2021b] and [Specker *et al.*, 2022], and the results reported in the original literature for each prior work. For RAPzs and PETAzs, we refer scores from the datasets work [Jia *et al.*, 2021b] with priority, next to use public code of the preivous work, if any, to reproduce the results. Since there is no baseline reported for ConvNeXt-base on RAPzs and PETAzs, corresponding results are based on our experiments. If no result is reported as a certain setting or the public code is not available for convincible testing on RAPzs and PETAzs, it is marked as $-$. *results are produced with additional data augmentations. All values are percentages and the highest scores are marked by bold fonts.

achieved in a holistic, end-to-end manner.

4 Experiments

Data and Evaluation Metric. For the benchmark datasets, PETA [Deng *et al.*, 2014], along with the two largest public pedestrian attribute datasets RAP [Li *et al.*, 2016] and PA100k [Liu *et al.*, 2017], are adopted for evaluation. Detailed dataset information and usage are consistent to those in [Jia *et al.*, 2021b]. We also test our methods on two realistic datasets of RAPzs and PETAzs stated and released in [Jia *et al.*, 2021b]. As for the evaluation metrics, the label-based metric mean Accuracy (mA), which takes an average over all attributes' classification accuracy on the positive and negative samples, and two instance-based metrics Recall and F1-score (F1) are considered. we do not present Precision since it can be basically inferred when Recall and F1 are told.

Network and Training Details. We adopt Resnet-50 [He *et al.*, 2016] and ConvNeXt-base [Liu *et al.*, 2022] as backbones to study the efficacy and compatibility of our method under feature extractors of both classical and up-to-date designs. For the training details, Adam solver is applied without Nesterov momentum. The learning rate starts at 1e-4 and decays by a factor of 10 in a manner of multistep. If not specifically stated, the results in this section are produced by Eq.4 with efficiency. We refer readers to the code of this work in the supplementary material for further details.

Benchmark Results. Following the benchmark protocol provided in [Jia *et al.*, 2021b], our method is compared with recent notable SOTA approaches MsVAA [Sarafianos and Kakadiaris, 2018], MTMS [Gao *et al.*, 2019], ALM [Tang *et al.*, 2019], MTA-Net [Ji *et al.*, 2020], AR-BiFPN [Tan *et al.*, 2020], VAC [Guo *et al.*, 2020], JLAC [Tan *et al.*, 2020], SSC [Jia *et al.*, 2021a], DAFL [Jia *et al.*, 2022] and Label2Label [Li *et al.*, 2022]. Since we start at a training setup that involves less data augmentation methods, the baseline performances on these datasets could be about 1% mA inferior to some of the compared methods. To mitigate this gap, we also present experimental results of applying the data augmentation settings akin to JLAC (additional random scaling, rotation, translation, cropping, erasing and adding random gaussian blurs). The overall results are reported in Table.1. It highlights that across all settings, our prescription achieves performance at least comparable to others. Please note that our method is highly efficient since for test samples, the extra operations over the baseline models, are only one single FC layer applied to produce attribute-specific features, in stark contrast to the prior arts paying a premium in term of computational cost. Therefore, one potential of it is that, one could just make full use of the model parameters being largely saved by our lightweight framework, to plug a better-but-wider backbone, like ConvNeXt for further significant improvements.

Basically, one might notice that our proposal works on PA100k better than RAP and PETA. It is not surprising, since, as pointed out by [Jia *et al.*, 2021b], about 31.5% and 57.7% of pedestrian identities in the test set of RAP and PETA are identical to those in their respective training set. Thus, for PETA and RAP, memorizing pattern of the biased attributes co-occurence in training set can be conducive for model's test-set performance, making existing approaches overestimated on them. For this, we also report in Table.1 the experimental results on the PETAzs and RAPzs, which are respectively formed from PETA and RAP to follow the zero-shot setting of pedestrian identities, i.e., no overlapped identities between their training and test sets. Therefore, the results on PETAzs and RAPzs, along with the results on PA100k, are much convincible [Jia *et al.*, 2021a], and thus should be attached of more practical implications, on which our method outperforms others with considerable margins. Also,

it should be noticed that the concepts of existing works only fuel trivial boost on these realistic datasets, implying that previous development of PAR on PETA and RAP may partly come from better modeling of the common bias in training and test sets. Overall, our method excels in both the recognition performance and practical applicability of realistic PAR.

Ablation Studies. In this sub-section, we investigate the effectiveness and characteristics of each technical contribution. To better discern between attributes classifiers, we use PA100k and ResNet-50 in the following ablation studies.

Posterior-invariant learning vs. MixUp-based efficient learning. In this paper, we present two pipelines for attributes-disentangled feature learning, the one is directly derived from the goal of posterior-invariant learning to minimize the mutual information between one certain attribute and its irrelevant features specific to other attributes, as described in Eq.3. While effective as shown in the first row of Table.2, it requires multiple inferences for a single sample, which hinders the training efficiency and increases the memory usage when the number of attributes is large. To address this issue, a second pipeline for efficient single-inference-every-instance is introduced in Eq.4. It extends the concept behind MixUp to significantly lower the training cost, and achieves performance similar to Eq.3, as can be seen in the third row of Table.2. During our multiple rounds of experiments, we find that Eq.4 can always yield slightly better mA, and we credit such improvement into that the biased attribute correlation within dataset is totally discarded during the training of classifier in Eq.4, since the labels used in it are attribute-wise randomly generated thus can form any possible pattern of attribute co-occurence, making classifier hardly to memorize the limited attribute co-occurence presenting in the dataset, and thereby generalize better, as shown in Figure.4.

Norm-direction separated interpolation vs. plain feature interpolation. We apply the proposed norm-direction separated interpolation in Eq.3 and Eq.4 to fully exploit the possible variations of attribute-specific features, as graphed in Figure.3b. The results in Table.2 clearly demonstrate that this trick, in tandem with both pipelines, outperforms the plain interpolation used in previous work and thus can be applied in other related fields of research like recognition with unstable image quality for enhanced robustness and performance.

Mixup vs. Proposed Methods. Eq.4 directly optimizes over the interpolated features with their correspondingly interpolated labels, in a similar way as that of Mixup. Therefore, it might be questioned that our method's superiority can fundamentally come from tailoring on MixUp. However, as shown in Table.2, sample interpolation is not the case here for bringing significant performance boost, since Mixup delivers even negative increase in mA. Moreover, our pipeline of Eq.3, which does not optimize over the augmented samples, can score comparatively with Eq.4. It implies that, what essentially works for Eq.4 is the strategy that we use feature interpolation to enable a random and differentiated transform over each attribute's specific feature, and empower model with the consequential mutual information minimization learning by Eq.3 or Eq.4. On a higher level, data augmentations are methods creating data points beyond the training set to reduce the various bias embedded in dataset. Under such perspective, it brings no bad to comprehend our method as a data augmentation trick that augments the pattern of attributes co-occurence, out of the less representative dataset.

Our model Pipeline	NDSI	mA	Recall	F1
Eq.3		83.20	89.27	86.96
Eq.3	✓	83.79	89.51	87.06
Eq.4		83.83	89.04	86.85
Eq.4	✓	84.53	89.13	87.01
Method		**mA**	**Recall**	**F1**
Baseline		80.38	87.01	87.05
MixUp		79.73	85.22	86.75

Table 2: The breakdown effect for each technical component of the introduced method, in which NDSI is short for the norm-direction separated interpolation depicted in Figure.3b. Also, the comparison of our method against MixUp.

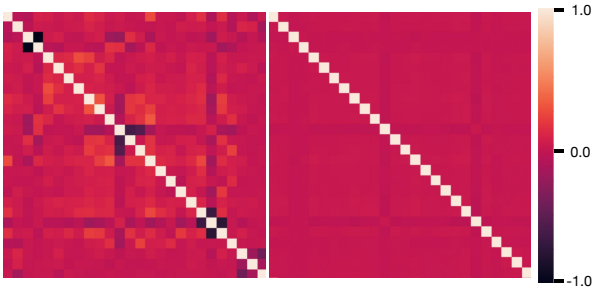

Figure 4: The matrix of cosine similarities between PA100k attribute anchor features of Left: baseline model and Right: our method. The results clearly support that our work suppresses the undesirable modeling of biased attributes co-occurence in classifier. Same to that in Figure.1, weights in the last FC layer of a trained Resnet-50 are employed as the attribute anchor features.

Analysis on Improvements. In Figure.5, we draw the per-attribute accuracy results on PA100k dataset of the proposal and the baseline work. We find that our work is capable of sizably raising scores on almost all attributes. It might serve as a relief to the likely concern that discarding totally the use of attributes correlation could be detrimental to the recognition of some certain type of attributes, for which a further study is presented in the next parts. Generally, it can be seen that improvements are considerable on attributes with positive ratio closer to 1 or 0. It is expectable as attributes with inadequate chances to appear or disappear are likely to be limited in the captured or non-captured scenes and identities in term of the attributes interdependency. By our framework, on one side the inference dependence of an imbalanced attribute to its frequently or hardly co-occurred attributes is greatly undermined, reducing the bias within feature learning. On another side, aforementioned for Figure.4, the memorization effect of attributes correlation in classifier is also repressed. Thanks to these two factors, the recognition of attributes with most potential to be correlated can be facilitated utterly, whilst for attributes appearing evenly, the improvements can be trivial.

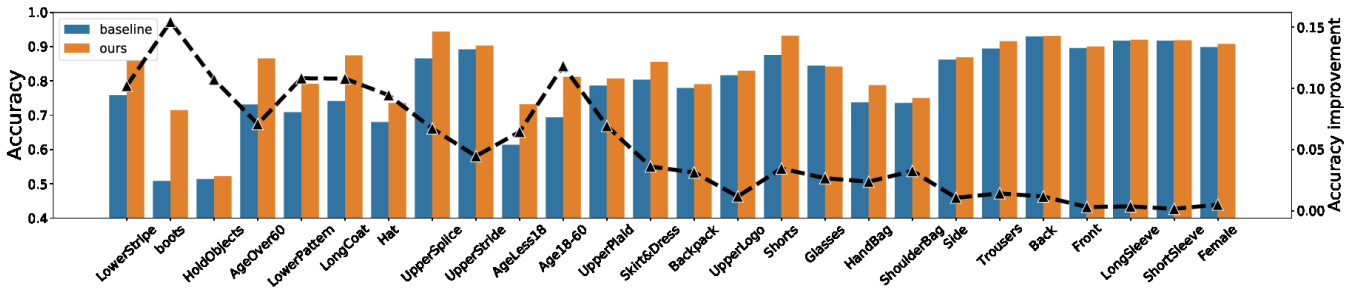

Figure 5: Accuracy results comparison of our method with the baseline work on all attributes of PA100k. From left to right, the attributes are arranged in a decreasing order of the distance between their respective positive ratio and 0.5. The black dashed line marks the variation of accuracy increase on each attribute, which is mean smoothed with the windows size of 2 for better visualization.

It also accounts for the reason that our method works better on label-based metric than instance-based metric, since the growth of mA is actually bottlenecked by the imbalanced attributes, while instance-based metrics are relatively not.

On the Attribute Interdependency Modeling. Intuitively, for attributes not causally independent to each other, discovering and utilizing the relations among them might be conducive for robust information exchanges and propagation in PAR. For this, turning back to seek help from previous work that precisely models the attributes interdependency seems suboptimal, since without elaborate expert knowledge use, there is no feasibilities of learning, for these methods to discern the welcomed attributes interdependency from the intricate attributes co-occurence bias, and exclusively dismiss use of the latter. Noting in Figure.5, the baseline model is error-prone not only for attributes appearing barely, but those occurring often like 'Age18-60' (positive ratio above 0.9) so, such an indiscriminative modeling of all attributes correlations is overall detrimental to most imbalanced attributes, making PAR strongly bottlenecked.

Moreover, there is another drawback of learning with attribute interdependencies - even the causally robust ones like the mutual exclusiveness among ages groups ('age $<$ 18', '18 $\leq$ age $\leq$ 60' and 'age $>$ 60' in PA100k are of a multi-class relation rather than that of multi-label), for which we call the infestation of attribute independency. As shown in Figure.4, two attributes tending to (not to) co-occur with each other would also prefer to (not to) co-occur with other attributes, representing as the dark or light stripes spanning across the anchor feature similarity matrix. Whereas these traversing lines do not show up in the Pearson correlation coefficient matrix of labels, it is actually by-produced during learning of the attributes correlations. This inclination of hallucinating new interdependency bias from the dataset attributes interdependency is smoothed out by ours, greatly yet not totally, as there are still lines can be eyed in the heatmap.

Towards Robust PAR. Ideally, a PAR of robustness should be capable of inferring on a-prior concrete attributes interdependencies while disregarding the others. Solutions of ease towards this aim can stand on the base of attributes-disentangled feature learning. Taking the age groups in PA100k as example, on a trained model of attributes-disentangled learning, one could train an additional classifier head of multi-class only for age prediction (namely, Robust-A), or conduct certain back-end processings onto the outputs to rectify or unify the age predictions (Robust-B). Here, we realize both in Table.3. For Robust-B, we output only the age attribute of highest confidence score. Compared to our method, these robust PARs yield negligible accuracy increase in test set. However, we still emphasize on the importance of such designs since they secure the reliability of PAR, and might make difference in realistic conditions.

Age attributes	Baseline	Ours	Robust-A	Robust-B
AgeLess18	61.36	75.67	75.72	75.67
Age18-60	69.36	82.21	81.95	82.22
AgeOver60	73.15	84.87	84.94	84.87

Table 3: On three causally related age attributes, we compare the accuracy results of baseline, our work and the described robust PARs. For Robust-A, we use weighted loss to soften the class imbalance.

5 Conclusion

We present a novel method of attribute-disentangled feature learning to enhance the robustness of PAR. Without any sophisticated or time-consuming frameworks, leveraging a simple information-theoretic regularizer, it is ensured that the pedestrian attributes are inferred without considering the biased attributes interdependency inherent to dataset, in order to enable an attribute recognition mechanism respecting us human. The comprehensive experiments demonstrate that our proposal reaches SOTA performance with appealing merits like better generalizability and applicability in realistic scenarios. Importantly, our theorized perspectives of attribute disentanglement learning differs from the paradigms of previous methods, even advancing in an opposite direction, heuristically exploring a promising avenue for future work.

Acknowledgments

This work was partially supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2023C01030), the National Natural Science Foundation of China (No.62122011, U21A20514), and the Fundamental Research Funds for the Central Universities. (Corresponding Author: Hai-Miao Hu)

References

[Chen *et al.*, 2019] Zhiyuan Chen, Annan Li, and Yunhong Wang. A temporal attentive approach for video-based pedestrian attribute recognition. In *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II 2*, pages 209–220. Springer, 2019.

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

[Fabbri *et al.*, 2017] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. Generative adversarial models for people attribute recognition in surveillance. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.

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

[Gao *et al.*, 2019] Lian Gao, Di Huang, Yuanfang Guo, and Yunhong Wang. Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1340–1348, New York, NY, USA, 2019. Association for Computing Machinery.

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

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE*, 2016.

[Ji *et al.*, 2020] Z. Ji, Z. Hu, E. He, J. Han, and Y. Pang. Pedestrian attribute recognition based on multiple time steps attention. *Pattern Recognition Letters*, 138(2), 2020.

[Jia *et al.*, 2021a] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 962–971, October 2021.

[Jia *et al.*, 2021b] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021.

[Jia *et al.*, 2022] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1069–1077. AAAI Press, 2022.

[Li *et al.*, 2016] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

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

[Li *et al.*, 2022] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework formulti-attribute learning. *Springer, Cham*, 2022.

[Liu *et al.*, 2017] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[Liu *et al.*, 2018] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[Meng *et al.*, 2021] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, June 2021.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[Sarafianos and Kakadiaris, 2018] Nikolaos Sarafianos and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV 2018*, 2018.

[Shalev *et al.*, 2018] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, pages 7375–7385, 2018.

[Specker *et al.*, 2022] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. *ArXiv*, abs/2209.02522, 2022.

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

[Tang *et al.*, 2019] C. Tang, L. Sheng, Z. X. Zhang, and X. Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[Wang *et al.*, 2016] J. Wang, X. Zhu, and S. Gong. Video semantic clustering with sparse and incomplete tags. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[Wang *et al.*, 2017] Jingya Wang, Xiatian Zhu, and Shaogang Gong. Discovering visual concept structure with sparse and incomplete tags. *Artificial Intelligence*, 250:16–36, 2017.

[Wang *et al.*, 2018] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, pages 926–930, 2018.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

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

[Zhu *et al.*, 2017] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image Vision Computing*, 58(FEB.):224–229, 2017.