Fair Training with Zero Inputs

Wenjie Pan, Jianqing Zhu*, Huanqiang Zeng

College of Engineering, Huaqiao University, Quanzhou 362021, China
panwj@stu.hqu.edu.cn, {jqzhu, zeng0043}@hqu.edu.cn

Abstract

There are two manifestations of classification fairness. One is the preference for head classes with more instances due to the long-tail (LT) distribution of training data. The other is the clever Hans (CH) effect, where non-discriminative features are mistakenly used for classification. In this paper, we find that using category-agnostic zero-valued data can simultaneously reveal both types of unfairness. Based on this, we propose a zero uniformity training (ZUT) framework to optimize classification fairness. The ZUT framework inputs category-agnostic zero-valued data into the model in parallel and uses zero uniformity loss (ZUL) to optimize classification fairness. The ZUL loss mitigates bias towards specific classes by unifying the classification features corresponding to zero-valued data. The ZUT framework is compatible with various classification-based tasks. Experiments show that the ZUT framework can improve the performance of multiple state-of-the-art methods in image classification, person re-identification, and semantic segmentation.

Code — https://github.com/asd123pwj/ZUT

Introduction

The fairness of models is crucial for their effective application in the real world. Explicit unfairness usually manifests as the long-tailed (LT) distribution of training data (Cao et al. 2019), where some classes have significantly fewer samples than others. Implicit unfairness, on the other hand, is caused by model biases towards non-discriminative features (Anders et al. 2022), such as when a model trained with black dogs and orange cats tends to use color as a criterion.

The unfairness of LT training data causes models to favor head data (categories with more samples) while ignoring tail data (categories with fewer samples) because training tend to minimize empirical risk (Vapnik 1991). To address this unfairness, some methods study fairness in data distribution. One strategy (Shi et al. 2023) is to improve data sampling methods so that the number of samples in different categories is similar during training. Another approach is to use data augmentation (Wang et al. 2024) to enrich tail data. Some methods explore fairness in training intensity, such as designing loss functions (Lin et al. 2017) to weight the impact of different samples.

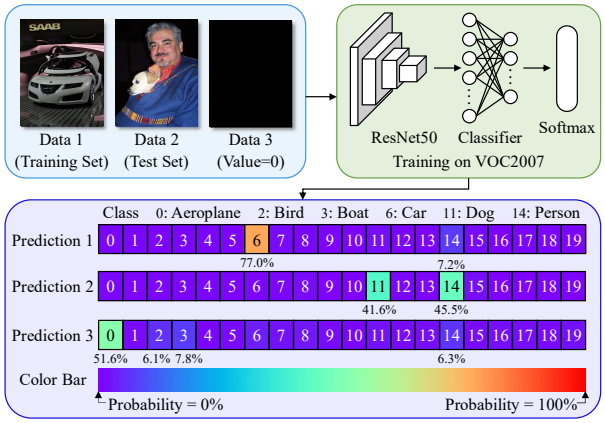

Figure 1: ResNet-50 (He et al. 2016) trained on VOC2007 (Everingham et al. 2010) is used to predict three data: Data 1 from the training set, Data 2 from the test set, and Data 3 consisting entirely of zero values.

The bias of models toward non-discriminative features, known as the clever Hans (CH) effect (Anders et al. 2022), poses challenges for explainable AI. Hans was a horse believed to perform arithmetic but was merely reacting to subtle body language cues. To identify discriminative features, class activation mapping (CAM) (Zhou et al. 2016) is used to visualize classifier activations on input images. Studies (Zeiler and Fergus 2014) investigate the effect of occluding different parts of objects on classification accuracy. To counteract the CH effect, human explanations can be integrated into the model (Rieger et al. 2020), such as reducing color influence in the loss function for black dogs and orange cats. Another approach (Anders et al. 2022) is to suppress activations related to spurious features.

We discovered a method to simultaneously display two types of unfairness by using category-agnostic zero-valued data for classification. As shown in Figure 1, Data 1 and Data 2 are from the VOC2007 (Everingham et al. 2010) training set and test set, respectively, and Data 3 is a category-agnostic data with zero-valued values. Here, Data 3 is termed category-agnostic due to its lack of texture varia-

*Corresponding author.

tions found in real-world objects. When these three data are classified with a model trained on VOC2007, two unfairness can be observed. First, the model bias caused by the LT distribution of data, where Data 1 and Data 3 are more likely to be misclassified as 'person' compared to other negative categories. This may be due to nearly half of the data containing the 'person' category om VOC2007, leading to a model bias towards the 'person' category. Second, the CH model's bias towards non-discriminative features, where Data 3 is more likely to be misclassified as 'airplane,' 'ship,' or 'bird.' This may be because these three categories have less background texture (i.e., sky and sea), causing the model to erroneously associate background texture with target categories. Data 3, as a category-agnostic zero-valued data, would not be classified by humans into any category, and the first layer weights of the model would not activate on it. This suggests that the two types of unfairness displayed by Data 3 are closely related to the model's inherent biases. And data 3 can be utilized to enhance model fairness.

In this paper, we introduce a zero uniformity training (ZUT) framework to enhance model classification fairness. The ZUT framework operates parallel to existing training frameworks, where it inputs a category-agnostic zero-valued data into the model and uses the classifier's output features for zero uniformity loss (ZUL) computation. The ZUL loss mitigates model's inherent biases by unifying classification features, addressing unfairness from LT distributions and the CH effect. The ZUT framework is applicable to most classification-based training frameworks, including image classification, person re-identification, and semantic segmentation. Experiments demonstrate that the ZUT framework improves the performance of various SOTA methods.

The contributions of the paper are summarized as follows.

- We propose using category-agnostic zero-valued data to make classification features reflect unfairness caused by long-tailed distributions and the clever Hans effect.
- We introduce the zero uniformity training (ZUT) framework, which involves training with zero-valued data and using zero uniformity loss (ZUL) to suppress model inherent biases, thereby optimizing model fairness.
- The ZUT framework improves the performance of SOTA models in image classification, person re-identification, and semantic segmentation.

Related Work

Long-Tailed Distribution

Long-tailed (LT) distribution (Cao et al. 2019; Kang et al. 2020) refers to a dataset where the number of instances for some classes is much smaller than for others. Models trained on LT data tend to exhibit bias in classification, performing better on head data (classes with more instances) and worse on tail data (classes with fewer instances). This unfairness is inevitable in training processes that optimize for empirical risk minimization (Vapnik 1991).

To mitigate the unfairness caused by LT data, data resampling (Chawla et al. 2002) is a classic strategy. Data resampling changes the distribution of training data, giving tail data and head data as consistent training resources as possible. This means that effectively utilizing the discriminative information in tail data is very important. Therefore, GradCAM (Selvaraju et al. 2017) is used to extract background masks, allowing some more discriminative images to be used for training (Shi et al. 2023). Since data augmentation can change the form of data, it can also enhance the generalization ability to tail data. However, changing the form of data may damage some discriminative features of tail data, so there are methods (Wang et al. 2024) to dynamically apply data augmentation by class. Considering that the loss function aims to optimize LT data towards empirical risk minimization (Vapnik 1991), setting different weights for the losses of different classes based on the number of instances is very effective (Lin et al. 2017).

Clever Hans Effect

The clever hans (CH) effect (Anders et al. 2022) refers to the model incorrectly associating non-discriminative features with classes. This non-discrimination usually arises from erroneous connections made with limited data. In traditional person re-identification (Xie et al. 2024, 2023), since each person always wears the same clothing, the model shows high activation intensity in the clothing area. This presents a significant issue as the model retrieves clothing rather than the person. Additionally, the CH effect is a key issue in achieving explainable AI (Weber et al. 2023).

Class activation mapping (Zhou et al. 2016; Selvaraju et al. 2017) is an excellent visualization method that maps the activated data corresponding to the class, allowing analysis of whether discriminative features are learned. Furthermore, some studies (Chen and Sun 2023) have extended this to non-discriminative features of the target, enabling a classification method to exhibit segmentation-like results. Redundant data can trigger the CH effect, so some methods (Zhong et al. 2020; Zeiler and Fergus 2014) also investigate the impact of data erasure on classification results. Some data that trigger the CH effect can be perceived by humans, so introducing human explanations for classification (Rieger et al. 2020) can effectively suppress the model bias caused by these non-discriminative data. There are also studies (Anders et al. 2022) that identify these non-discriminative features within the model, avoiding manual annotation cost.

Fairness of Classification

The classifier is crucial in linking data with labels and is a key layer for ensuring fairness. It exists in various tasks, fulfilling different roles. In image classification, it includes multi-class tasks (Krizhevsky and Hinton 2009) that assign a single image to one category and multi-label tasks (Everingham et al. 2010) that classify a single image into multiple categories. Due to the varying instance numbers for different classes, classes with more instances achieve better learning outcomes (unfairness in LT data). Additionally, due to the different appearances of data in various classes, the model might use some non-discriminative features unrelated to the class as a basis for classification (unfairness from the CH effect). Clothing-changing person re-identification (CC-ReID) (Yang, Wu, and Zheng 2019; Huang et al. 2019; Wan et al.

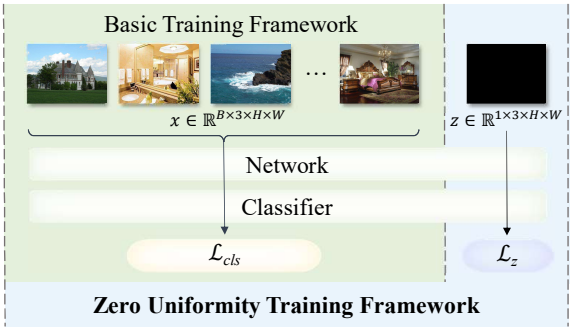

Figure 2: The structure of zero uniformity training (ZUT) framework. Here, B denotes batch size; H and W denote the height and the width of data, respectively; $\mathcal{L}_{cls}$ denotes classification loss; $\mathcal{L}_z$ denotes zero uniformity loss.

2020; Gu et al. 2022) is a task that largely avoids these two types of unfairness. When collecting datasets, a similar number of photos are usually taken for each person, and the diversity in clothes reduces the impact of clothes on classification. Moreover, classification of semantic segmentation (Zhou et al. 2017) includes positional information, meaning spatial distribution fairness must be considered.

The category-irrleated zero-valued data shown in Figure 1 makes the model reflect two types of unfairness. Therefore, we choose to optimize classification fairness through additional zero-valued data. It is well compatible with various classification forms, including image classification, person re-identification, and semantic segmentation.

Approach

Zero Uniformity Training Framework

The zero uniformity training (ZUT) framework is designed to incorporate category-agnostic zero-valued data into training to enhance the model's classification fairness. Figure 2 shows the structure of the ZUT framework. The ZUT framework performs parallel computation on zero-valued data z in the classification part of the existing training framework. The left side of Figure 2 shows the basic training framework. Tensor-formatted image data x passes through the network and classifier sequentially, and the classification loss $\mathcal{L}_{cls}$ is calculated. On the right side of Figure 2 is a parallel process, where an zero-valued data z is fed into the network, and the zero uniformity loss (ZUL) $\mathcal{L}_z$ is computed.

The ZUT framework is used only during training, with a single loss weight α as a hyperparameter, and can be easily integrated into existing networks. Thus, we applied the ZUT framework in three tasks that use classifiers: image classification, person re-identification (ReID), and semantic segmentation. The algorithm for applying the ZUT framework in image classification is shown in Algorithm 1. Since the training framework for ReID is consistent with image classification, the ZUT framework is applied in the same manner. For semantic segmentation, the classifier categorizes pixels instead of images, so the ZUT framework must consider the spatial position's impact on ZUL loss $\mathcal{L}_z$. Further analysis of this application will be provided in the next section.

Algorithm 1: ZUT Framework in Classification

Input: Image tensors $x \in \mathbb{R}^{B\times3\times H\times W}$, Label GT
Parameter: Loss weight α

$z \leftarrow$ Tensor.zeros(1, 3, H, W)
$y \leftarrow$ Concat(z, x)
$f \leftarrow$ Classifier(Network(y))
$v, u \leftarrow f[0], f[1:]$
$\mathcal{L} \leftarrow \alpha \cdot \mathcal{L}_z(v) + \mathcal{L}_{cls}(u, GT)$

Zero Uniformity Loss

The zero uniformity loss (ZUL) aims at using classification features corresponding to zero-valued data to improve the model's classification fairness. As described in Figure 1, classification features corresponding to zero-valued data should not show any bias toward a specific class. However, two types of unfairness are observed: unfairness caused by long-tailed (LT) data and the model's clever hans (CH) effect. Therefore, the optimization direction of the ZUL loss is to ensure that the classification features corresponding to zero-valued data have a uniform probability distribution.

Three types of ZUL losses are designed. First, the classification features corresponding to zero-valued data are driven towards the mean by minimizing standard deviation, as shown in Eq. (1). Second, the classification features corresponding to zero-valued data are driven towards zero by minimizing the mean, as shown in Eq. (2). Third, classification probabilities of zero-valued data are driven towards uniformity by maximizing entropy, as shown in Eq. (3).

$$\mathcal{L}_z^{std} = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(v_i - \bar{v})^2} \tag{1}$$

$$\mathcal{L}_z^{mean} = \frac{1}{C}\sum_{i=1}^{C}|v_i| \tag{2}$$

where $v \in \mathbb{R}^C$ denotes the classification features of zero-valued data; C denotes the number of classes.

$$\begin{aligned} p_i &= \frac{e^{v_i}}{\sum_{j=1}^{C} e^{v_j}}, \\ \mathcal{L}_z^e &= \sum_{i=1}^{C} p_i \log(p_i) - \sum_{i=1}^{C} \frac{1}{C}\log(\frac{1}{C}) \end{aligned} \tag{3}$$

For image classification and ReID, the classification features are one-dimensional ($v \in \mathbb{R}^C$), allowing direct application of ZUL loss. However, for semantic segmentation, the classification features are three-dimensional ($v \in \mathbb{R}^{C\times H\times W}$) and cannot be directly used with ZUL loss. To handle spatial features, we initially considered optimizing spatial uniformity rather than class uniformity, but this could disrupt object spatial distributions. We visualized the ground truth distribution in the most commonly used semantic segmentation dataset, ADE20K (Zhou et al. 2017), as shown in Figure 3. It can be seen that different objects have obvious spatial distributions. Additionally, a common application scenario for semantic segmentation is autonomous driving, where the camera perspective is fixed. This means that

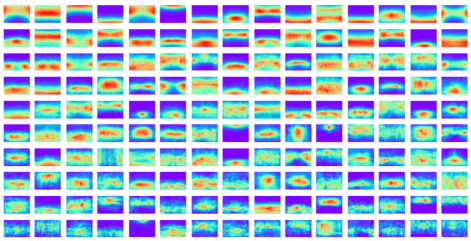

Figure 3: Ground truth distribution map of ADE20K.

vehicles and pedestrians in autonomous driving typically appear in fixed regions. Thus, spatial uniformity is not feasible. Instead, we use global average pooling (GAP) for each class, as shown in Eq. (4), ensuring class-wise uniform while allowing spatial biases within classes. The classification features processed by GAP are fed into the ZUL loss.

$$\mathrm{GAP}(v) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} v_{c,i,j} \tag{4}$$

Discussion: Input Type. The ZUT framework uses category-agnostic zero-valued data as input to unify the classifier's output. The purpose of zero-data is to train the model to recognize category-agnostic data as understood by human cognition. This means that any category-agnostic data can be used in the ZUT framework. For example, some variants used in random erasing (Zhong et al. 2020), one-valued data, mean value during normalization, and random values.

Efficiency Since the ZUT framework introduces only a new loss and participates in training using parallel computation, its performance overhead is approximately $\frac{1}{B}$ of the training performance overhead. Additionally, since ZUT does not add new layers and does not participate in the testing process, it does not affect inference performance (e.g., model size and inference speed).

Experiments

All experiments were conducted on RTX 3090 GPU. Pretrained on ImageNet (Russakovsky et al. 2015). Each experiment was repeated three times, and the mean and sample standard deviation were reported. Depending on the dataset, the evaluation metrics included mean average precision (mAP), accuracy (Acc), rank-1 (R1) of the cumulative match characteristic (CMC), mean accuracy (mAcc), and mean intersection over union (mIoU). Higher values indicate better performance. The zero uniformity training (ZUT) framework is applied using the zero uniformity loss (ZUL) $\mathcal{L}_z^{std}$ that leverages standard deviation.

Image Classification

Datasets VOC2007 (Everingham et al. 2010) is a multi-label classification dataset that includes 9,963 images across 20 categories, such as people, animals, vehicles, and furniture. Nearly half of the images contain pedestrians. CIFAR100-LT (Cao et al. 2019) is a long-tailed distribution dataset created by performing long-tailed sampling on CIFAR100 (Krizhevsky and Hinton 2009). Following (Cao et al. 2019), we used an imbalance ratio of 100, meaning the most frequent category has 100 times more samples than the least frequent category.

Dataset	Methods	M. (%)
VOC2007	RNN-frequency (Lyu et al. 2019)	85.6
	Eva02 (Fang et al. 2023)	84.6
	Riformer (Wang et al. 2023)	86.0
	ZUT-ResNet (Ours)	92.0
CIFAR100-LT	CR-CE (Ma et al. 2023)	40.5
	CSA (Shi et al. 2023)	46.6
	Riformer (Wang et al. 2023)	47.4
	ZUT-Riformer (Ours)	48.9

Table 1: Comparisons with SOTA methods on multi-label classification dataset VOC2007 and long-tail distribution classification dataset CIFAR100-LT. Here, M. denotes the metric, for VOC2007, it is mAP, and for CIFAR100-LT, it is accuracy.

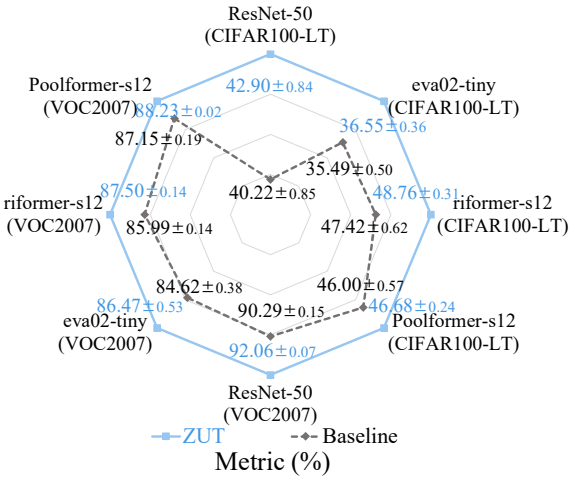

Figure 4: Ablations on two classification datasets.

Setups All experiments are performed using MMPretrain (Contributors 2023). Each experiment runs for 100 epochs with a batch size of 64. The AdamW (Loshchilov and Hutter 2017) optimizer is employed, with a learning rate set at 1e-4 and a weight decay of 0.3. The learning rate decays by cosine annealing (Loshchilov and Hutter 2016) with a minimum learning rate of 1e-5. The classification loss incorporates label smoothing (Szegedy et al. 2016) at 0.1. For the VOC2007 dataset, images are randomly resized and cropped to 256×256 during training, with a 50% probability of random horizontal flipping. During testing, images are resized to ensure the shortest side is 256 pixels, followed by a center crop of 256×256. For the CIFAR100-LT dataset, training involves random cropping to 32×32 with a 4-pixel padding and a 50% probability of random horizontal flipping.

Comparisons with State-of-the-Art Methods Table 1 shows the comparisons between ZUT and state-of-the-art (SOTA) methods for classification tasks. We evaluated our method on VOC2007 using a ResNet-50 (He et al. 2016).

Datasets	Methods	R1 (%)
PRCC	CVSL (Nguyen et al. 2024)	57.5
	CLIP3DReID (Liu et al. 2024)	60.6
	ZUT-ResNet (Ours)	62.2
VCClothes	3DSL (Chen et al. 2021)	79.9
	GI-ReID (Jin et al. 2022)	64.5
	ZUT-ResNet (Ours)	80.9
Celeb-ReID	AFDNet (Xu et al. 2021)	52.1
	ACID (Yang et al. 2023)	52.5
	ZUT-ResNet (Ours)	53.3
CCVID	3DInvarReID (Liu et al. 2023)	84.3
	CLIP3DReID (Liu et al. 2024)	82.4
	ZUT-ResNet (Ours)	85.4

Table 2: Comparisons with SOTA on CC-ReID datasets.

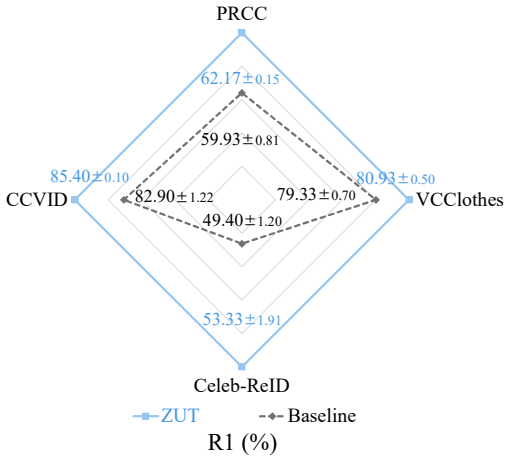

Figure 5: Ablations on four CC-ReID datasets.

On CIFAR100-LT, we used a Riformer-s12 (Wang et al. 2023). Our method demonstrated comparable performance with SOTA methods, surpassing DAN by 0.9% mAP on VOC2007 and CSA by 2.3% Acc on CIFAR100-LT.

Ablation We applied the ZUT framework to various models, including ResNet-50 (He et al. 2016), eva02-tiny (Fang et al. 2023), riformer-s12 (Wang et al. 2023), and poolformer-s12 (Yu et al. 2022). The ablation experiments on VOC2007 and CIFAR100-LT are shown in Figure 4. These data indicate that the ZUT framework can be applied to different models, including convolutional neural networks and transformer-based networks. For example, the ZUT framework improved the mAP of ResNet-50 by 1.77% on VOC2007 and the accuracy of riformer-s12 by 1.34% on CIFAR100-LT. This suggests that the ZUT framework improves the model's fairness, thereby enhancing its understanding of discriminative features.

Clothes Changing Person Re-Identification

Datasets PRCC (Yang, Wu, and Zheng 2019) is an image-based clothes changing person re-identification (CC-ReID) dataset, comprising 33,698 images of 221 individuals. Celeb-ReID (Huang et al. 2019) is another image-based CC-ReID dataset, consisting of 34,185 street photos of 1,052 celebrities. VCClothes (Wan et al. 2020) is a synthetic image-based CC-ReID dataset, including 19,060 images of 512 individuals. CCVID (Gu et al. 2022) is a video-based CC-ReID dataset, consisting of 2,856 tracklets of 226 individuals, with a total of 347,833 images. The common challenge across these four datasets is that re-identification can mostly only be performed using facial features, body shape, and posture. This implies that the datasets themselves have minimal inherent bias.

Setups All experiments are implemented according to CAL (Gu et al. 2022), with training and testing strategies consistent with CAL, except that we do not use random cropping as a data augmentation. Simply put, the ZUT framework is applied to ResNet-50. Images are resized to 384×192 for image-based datasets (i.e., PRCC, Celeb-ReID, and VCClothes) and 256×128 for video-based datasets (i.e., CCVID).

Comparisons with State-of-the-Art Methods Comparisons on PRCC, VCClothes, Celeb-ReID, and CCVID are shown in Table 2. Our method demonstrates comparable performance with SOTA methods. For example, on image-based dataset PRCC, ZUT-ResNet improves R1 by 4.7% compared to CVSL (Nguyen et al. 2024). On video-based dataset CCVID, ZUT-ResNet improves R1 by 3.0% compared to CLIP3DReID (Liu et al. 2024). This demonstrates that our method achieves SOTA performance.

Ablations Ablation studies of the ZUT framework on ResNet-50 were conducted on four CC-ReID datasets, as shown in Figure 5. It can be seen that the ZUT framework performs well on all datasets. For example, on the image-based dataset PRCC, the use of the ZUT framework resulted in a 2.24% improvement in R1, and on the video-based dataset CCVID, it resulted in a 2.50% improvement in R1. This demonstrates the effectiveness of the ZUT framework.

Semantic Segmentation

Dataset ADE20K (Zhou et al. 2017) is a large-scale semantic segmentation dataset for scene understanding, comprising 25,210 images across 150 categories. This dataset exhibits a long-tail distribution, where the most numerous category is approximately 700 times more prevalent than the least numerous category.

Setups All experiments are implemented using MMSegmentation (Contributors 2020). Training involved 80,000 iterations with a batch size of 16. AdamW (Loshchilov and Hutter 2017) is utilized as the optimizer with a learning rate of 1e-4 and weight decay of 1e-4. The PolyLR scheduler with $\gamma = 0.9$ is employed. Cross-entropy is used as the classification loss function. Semantic FPN (Kirillov et al. 2019) served as the decoder. During training, images are randomly scaled and cropped to 512×512, with a 50% probability of random flipping. For testing, images are scaled to ensure the short side was 512 pixels.

Comparisons with State-of-the-Art Methods Table 3 presents the comparison of the ZUT framework with SOTA methods on ADE20K. The ZUT framework was applied to

Methods	mAcc (%)	mIoU (%)
ResNet-50 (He et al. 2016)	46.38	35.65
Segformer-b0 (Xie et al. 2021)	48.44	37.41
VAN-b0 (Guo et al. 2023)	49.41	37.46
ZUT-VAN-b0	52.56	37.87

Table 3: Comparisons with SOTA methods on semantic segmentation dataset ADE20K.

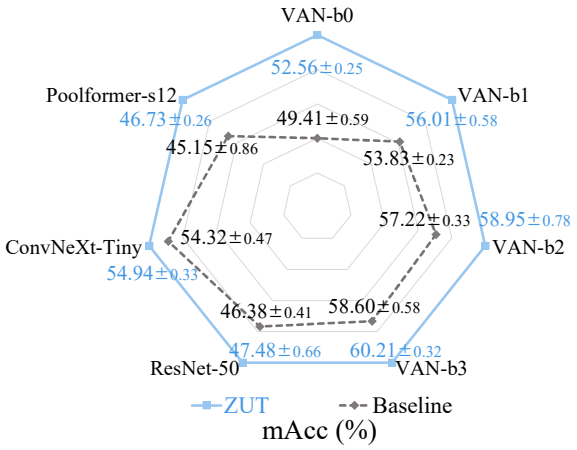

Figure 6: Ablations on semantic segmentation dataset.

VAN-b0. The results show that ZUT-VAN-b0 achieved performance comparable to SOTA methods. For example, ZUT-VAN-b0 exceeded Poolformer-s12 by 7.41% in mACC and 2.33% in mIoU.

Ablations Ablation experiments were conducted on ADE20K using the ZUT framework. The backbone networks used were ResNet-50 (He et al. 2016), ConvNeXt-tiny (Liu et al. 2022), Poolformer-s12 (Yu et al. 2022), and VAN-{b0, b1, b2, b3} (Guo et al. 2023). The results are shown in Figure 6. The ZUT framework significantly improved the performance of these SOTA methods. For example, the ZUT framework increased ResNet-50's mAcc by 1.10% and VAN-b0's mAcc by 3.15%. Additionally, the ZUT framework proved effective across different model sizes. It enhanced the performance of all four VAN models. These findings demonstrate that the ZUT framework is suitable for various models and different model sizes.

Visualization Classification of semantic segmentation contains spatial distribution information. Figure 7 visualizes classification results on VAN-b0. Figure 7 (a) and Figure 7 (b) display the probability distribution of classification features for zero-valued data on baseline and the ZUT framwork, respectively. Baseline shows high activation in the first row's third category, while the ZUT framwork corrects this unfairness, indicating improved fairness for category-agnostic data. Figure 7 (c) presents the feature value histogram, showing baseline's discrete distribution with a nearly 60-unit range between highest and lowest values. The ZUT framework mitigates this dispersion effectively. These results indicate that the ZUT framework suppresses the model's bias towards category-agnostic data.

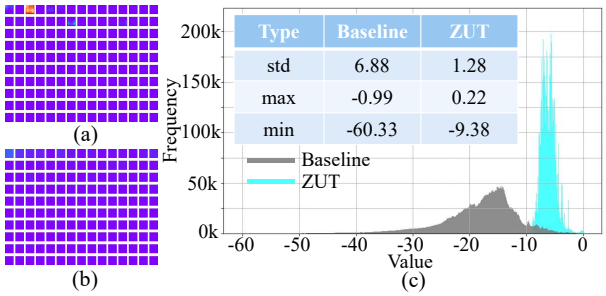

Figure 7: Visualizations of the probability distribution of classification features for zero-valued data on (a) Baseline and (b) ZUT. (c) Histogram of feature values. The baseline is VAN-B0 trained on ADE20K.

Analysis

Influence of Loss Weight We evaluated the impact of loss weights α across three tasks, as shown in Figure 8. It can be observed that the optimal performance varies with different loss weights for each task. For instance, in multi-label classification, $\alpha = 500$ achieves optimal performance, whereas for CC-ReID, the optimal weight is $\alpha = 0.01$. This variation may stem from inherent dataset biases, as CC-ReID discriminative information includes only facial, pose, and body shape, while multi-label classification involves more complex discriminative features. Fortunately, regardless of the task, the range of effective loss weights is broad. For example, α of 10 to 2000 are effective for multi-target classification, 0.005 to 0.5 for CC-ReID, and 0.001 to 100 for semantic segmentation. This flexible range of α underscores the superiority of the ZUT framework.

Influence of Inputs We evaluated the influence of different inputs on three tasks, as shown in Table 4. It can be observed that across all three tasks, almost all types of category-agnostic data can improve model performance. This suggests that category-agnostic data can enhance fairness by mitigating the model's bias towards non-discriminative features. However, there was an exception observed in CC-ReID, where using mean value of normalization as input resulted in decreased performance. This could be attributed to the use of random erasing data augmentation (Zhong et al. 2020) (which utilizes mean value of normalization) in this task, causing confusion between category-agnostic data and pedestrian data, thereby inhibiting the model's learning capability.

Evaluation of Loss Type We evaluated the impact of three types of losses on CC-ReID, as shown in Table 5. It can be seen that all losses can improve model performance under specific weight settings. For example, $\mathcal{L}_z^{std}$ ($\alpha = 0.01$) improves 2.24% of R1, $\mathcal{L}_z^{mean}$ ($\alpha = 0.1$) improves 1.80% of R1, and $\mathcal{L}_z^{e}$ ($\alpha = 0.05$) improves 1.37% of R1. This suggests that unifying the classification features of zero-valued data is a broadly effective approach. Furthermore, Figure 9 shows the variation curves of the three ZUL losses at their optimal

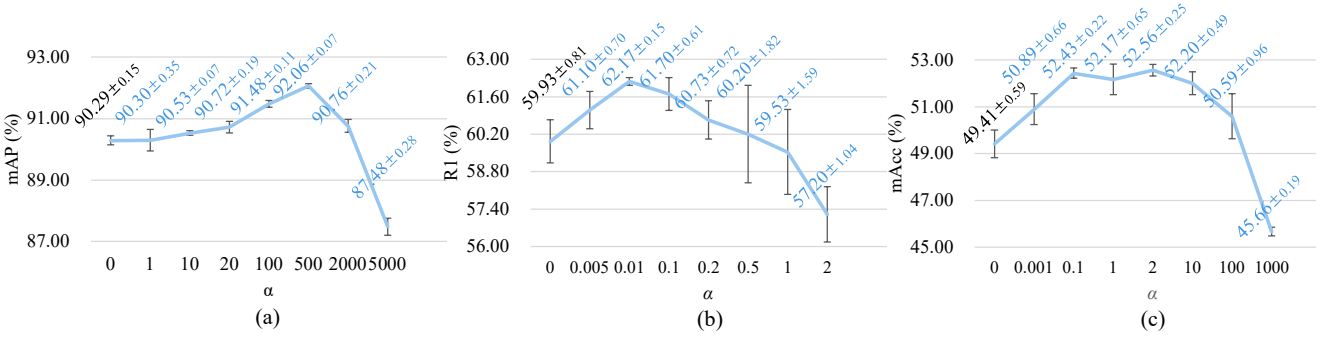

Figure 8: The effect of loss weights α on (a) VOC2007, (b) PRCC, and (c) ADE20K.

Datasets	Input Type	Metric (%)
VOC2007	N/A	90.29±0.15
	Zero	92.06±0.07
	One	91.91±0.05
	Mean	92.01±0.13
	Random	91.97±0.16
PRCC	N/A	59.93±0.81
	Zero	62.17±0.15
	One	60.77±0.51
	Mean	59.00±1.08
	Random	60.53±0.99
ADE20K	N/A	49.41±0.59
	Zero	52.17±0.65
	One	51.81±0.13
	Mean	52.03±0.79
	Random	51.15±1.17

Table 4: The impact of input types on the multi-label classification dataset VOC2007, the CC-ReID dataset PRCC, and the semantic segmentation dataset ADE20K. Here, "N/A" denotes baseline. For VOC2007, the metric is mAP. For PRCC, the metric is R1. For ADE20K, the metric is mAcc.

Loss Type	Loss Weight	R1 (%)
$\mathcal{L}_z^{std}$	0	59.93±0.81
	0.005	61.10±0.70
	0.01	62.17±0.15
	0.1	61.79±0.61
	1	59.53±1.59
$\mathcal{L}_z^{mean}$	0	59.93±0.81
	0.05	59.83±1.42
	0.1	61.73±1.15
	1	60.37±0.85
	2	58.17±0.81
$\mathcal{L}_z^{e}$	0	59.93±0.81
	0.01	60.60±1.54
	0.02	61.07±1.70
	0.05	61.30±1.73
	0.1	60.03±1.31

Table 5: Evaluation on CC-ReID dataset PRCC of standard deviation-based ZUL loss $\mathcal{L}_z^{std}$, mean-based ZUL loss $\mathcal{L}_z^{mean}$, and entropy-based ZUL loss $\mathcal{L}_z^{e}$.

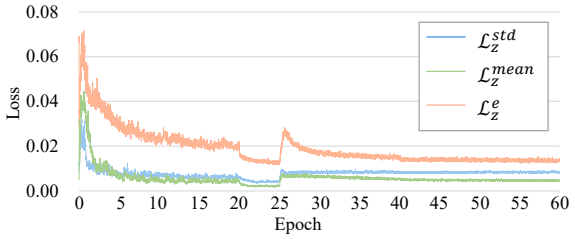

Figure 9: Variation curves of three ZUL losses.

weights. Larger loss values indicate a higher degree of classification bias on zero data across different categories, which signifies increased unfairness. All three losses exhibit a similar trend: an initial increase followed by a decrease. The initial rise in loss values suggests an increase in unfairness of the model, while the subsequent decline implies that the ZUL loss mitigates this unfairness. These findings demonstrate that the ZUL loss can suppress the model's unconscious unfairness during training. Furthermore, fluctuations in Figure 9 are observed at epochs 20, 25, and 40. Epochs 20 and 40 correspond to learning rate reductions, and epoch 25 marks the introduction of a new loss function by baseline. These fluctuations are not attributable to the ZUL loss.

Conclusion

Classification tasks face two types of unfairness from long-tail distribution and the clever Hans effect. In this paper, we use category-agnostic zero-valued data as input to make the model exhibit both types of unfairness. Based on this, we propose a zero uniformity training (ZUT) framework to optimize model fairness. The ZUT framework processes category-agnostic zero-valued data in parallel and sends its corresponding classification features to the zero uniformity loss (ZUL). The ZUL loss aims to unify the features of all categories, thereby suppressing internal biases in the model. The ZUT framework can effectively integrate with multiple classification-based tasks. Experiments show that ZUT improves the performance of state-of-the-art models across various tasks. In the future, we will explore a strategy for automatic learning of loss weights to select the optimal training strategy for different tasks within the ZUT framework.

Acknowledgments

This work was supported in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under Grant 2022J06023, Fujian Province Science and Technology Empowering Police Research Initiative Under Grant 2024Y0064, and in part by the High-level Talent Innovation and Entrepreneurship Project of Quanzhou City under Grant 2023C013R.

References

Anders, C. J.; Weber, L.; Neumann, D.; Samek, W.; Müller, K. r.; and Lapuschkin, S. 2022. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77: 261–295.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32.

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

Chen, J.; Jiang, X.; Wang, F.; Zhang, J.; Zheng, F.; Sun, X.; and Zheng, W. S. 2021. Learning 3D shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8146–8155. Tennessee, USA.

Chen, Z.; and Sun, Q. 2023. Extracting class activation maps from non-discriminative features as well. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3135–3144. Vancouver, Canada.

Contributors, M. 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark.

Contributors, M. 2023. OpenMMLab's Pre-training Toolbox and Benchmark.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Gu, X.; Chang, H.; Ma, B.; Bai, S.; Shan, S.; and Chen, X. 2022. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1060–1069. Louisiana, USA.

Guo, M.; Lu, C.; Liu, Z.; Cheng, M.; and Hu, S. 2023. Visual attention network. *Computational Visual Media*, 9(4): 733–752.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. Nevada, USA.

Huang, Y.; Xu, J.; Wu, Q.; Zhong, Y.; Zhang, P.; and Zhang, Z. 2019. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10): 3459–3471.

Jin, X.; He, T.; Zheng, K.; Yin, Z.; Shen, X.; Huang, Z.; Feng, R.; Huang, J.; Chen, Z.; and Hua, X. S. 2022. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14278–14287. Louisiana, USA.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*. Addis Ababa, Ethiopia.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6399–6408. California, USA.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Report, University of Toronto.

Lin, T.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988. Venice, Italy.

Liu, F.; Kim, M.; Gu, Z.; Jain, A.; and Liu, X. 2023. Learning clothing and pose invariant 3D shape representation for long-term person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 19617–19626. Paris, France.

Liu, F.; Kim, M.; Ren, Z.; and Liu, X. 2024. Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 256–266. Washington, USA.

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

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; and Hutter, F. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Lyu, F.; Wu, Q.; Hu, F.; Wu, Q.; and Tan, M. 2019. Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks. *IEEE Transactions on Multimedia*, 21(8): 1971–1981.

Ma, Y.; Jiao, L.; Liu, F.; Yang, S.; Liu, X.; and Li, L. 2023. Curvature-balanced feature manifold learning for long-tailed classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15824–15835. Vancouver, Canada.

Nguyen, V. D.; Khaldi, K.; Nguyen, D.; Mantini, P.; and Shah, S. 2024. Contrastive viewpoint-aware shape learning for long-term person re-identification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 1041–1049. Hawaii, USA.

Rieger, L.; Singh, C.; Murdoch, W.; and Yu, B. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, 8116–8126. Vienna, Austria: PMLR. ISBN 2640-3498.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. Venice, Italy.

Shi, J.; Wei, T.; Xiang, Y.; and Li, Y. 2023. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826. Nevada, USA.

Vapnik, V. 1991. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4.

Wan, F.; Wu, Y.; Qian, X.; Chen, Y.; and Fu, Y. 2020. When person re-identification meets changing clothes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 830–831. Washington, USA.

Wang, B.; Wang, P.; Xu, W.; Wang, X.; Zhang, Y.; Wang, K.; and Wang, Y. 2024. Kill Two Birds with One Stone: Rethinking Data Augmentation for Deep Long-tailed Learning. In *International Conference on Learning Representations*. Vienna, Austria.

Wang, J.; Zhang, S.; Liu, Y.; Wu, T.; Yang, Y.; Liu, X.; Chen, K.; Luo, P.; and Lin, D. 2023. Riformer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14443–14452. Washington, USA.

Weber, L.; Lapuschkin, S.; Binder, A.; and Samek, W. 2023. Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, 92: 154–176.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, Y.; Lin, Y.; Cai, W.; Xu, X.; Zhang, H.; Du, Y.; and He, S. 2024. D3still: Decoupled differential distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17181–17190. Washington, USA.

Xie, Y.; Zhang, H.; Xu, X.; Zhu, J.; and He, S. 2023. Towards a smaller student: capacity dynamic distillation for efficient image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16006–16015. Vancouver, Canada.

Xu, W.; Liu, H.; Shi, W.; Miao, Z.; Lu, Z.; and Chen, F. 2021. Adversarial feature disentanglement for long-term person re-identification. In *International Joint Conference on Artificial Intelligence*, 1201–1207. Montreal, Canada.

Yang, Q.; Wu, A.; and Zheng, W. S. 2019. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 2029–2046.

Yang, Z.; Zhong, X.; Zhong, Z.; Liu, H.; Wang, Z.; and Satoh, S. 2023. Win-win by competition: Auxiliary-free cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 32: 2985–2999.

Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. Metaformer is actually what you need for vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 10819–10829. Louisiana, USA.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 818–833. Zurich, Switzerland: Springer. ISBN 3319105892.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008. New York, USA.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929. Nevada, USA.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 633–641. Hawaii, USA.