This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

LTReID: Factorizable Feature Generation with Independent Components for Long-Tailed Person Re-Identification

Pingyu Wang, Zhicheng Zhao, Fei Su, Hongying Meng

***Abstract*—With the rapid increase of large-scale and real-world person datasets, it is crucial to address the problem of long-tailed data distributions, *i.e.*, head classes have large number of images while tail classes occupy extremely few samples. We observe that the imbalanced data distribution is likely to distort the overall feature space and impair the generalization capability of trained models. Nevertheless, this long-tailed problem has been rarely investigated in previous person Re-Identification (ReID) works. In this paper, we propose a novel *Long-Tailed Re-Identification* (LTReID) framework to simultaneously alleviate class-imbalance and hard-imbalance problems. Specifically, each real feature is decomposed into multiple independent components with two decorrelation losses. Then these components are randomly aggregated to generate more fake features for tail classes than head ones, resulting in the class-balance between head and tail classes. For the hard-balance between easy and hard samples, we utilize adversarial learning to generate more hard features than easy ones. The proposed framework can be trained in an end-to-end manner and avoids increasing the space and time complexity of inference models. Moreover, comprehensive experiments are conducted on the four ReID datasets so as to validate the effectiveness of the overall framework and the advantage of each module. Our results show that when trained with either balanced or imbalanced datasets, the LTReID achieves superior performance over the state-of-the-art methods.**

***Index Terms*—Person Re-Identification, Long-Tailed Distribution, Feature Factorization, Feature Generation**

I. Introduction

PERSON Re-Identification (ReID) aims to match images of the same person across non-overlapping cameras. It plays an important role in various video surveillance applications such as suspect tracking and missing elderly or children retrieval. With the advent of *Convolutional Neural Networks* (CNNs), the current deep feature learning based methods [1–17] have significantly outperformed a variety of traditional feature learning based approaches [18–24]. As deep learning is an essentially data-driven algorithm, the success of deep ReID models is undoubtedly inseparable to large-scale person ReID datasets, *e.g.*, Market1501 [25], CUHK03 [26], DukeMTMC [27] and MSMT17 [28].

Pingyu Wang, Zhicheng Zhao and Fei Su are with Beijing Key Laboratory of Network System and Network Culture, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. (e-mail: applewangpingyu@bupt.edu.cn; zhaozc@bupt.edu.cn; sufei@bupt.edu.cn)

Hongying Meng is with the College of Engineering, Design, and Physical Sciences, Brunel University London, Uxbridge, United Kingdom. (e-mail: hongying.meng@brunel.ac.uk)

This work is supported by Chinese National Natural Science Foundation (62076033, U1931202).

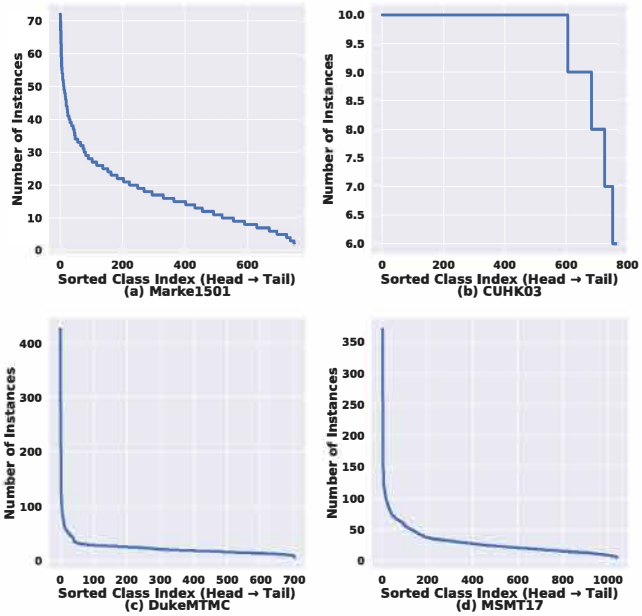


Fig. 1: The instance distribution of training data on Market1501, CUHK03, DukeMTMC and MSMT17 datasets. All classes are sorted in descending order by the number of instances. The instances on Market1501, DukeMTMC and MSMT17 datasets obey a long-tailed distribution, while the instances on CUHK03 dataset are almost uniformly distributed.

However, in contrast with commonly used visual recognition datasets (*e.g.*, CIFAR [29, 30] and ImageNet [31]) that exhibit roughly uniform distributions of class labels, most person ReID datasets always possess long-tailed distributions. As illustrated in Fig. 1, head classes claim most of samples, while tail classes are represented by relatively few examples. Since head classes contain more intra-class variations than tail classes, head classes prefer occupying a much larger spatial span than tail classes in the feature space [32]. This imbalanced feature distribution twists the overall feature space, which undermines the generalization capability of ReID models. Regrettably, this long-tailed problem has been infrequently investigated in previous person ReID works. Except for the ReID task, several studies [33–38] have proposed various imbalance learning algorithms for other specific tasks. These methods often improve the performance reasonably yet the improvement deteriorates when tail classes are severely under-represented. Specifically, they are often designed to adjust the decision boundary to reduce the bias introduced by imbalanced classes. Nonetheless, without enriching the diversities of tail classes, it becomes challenging to find the right direction to refine the decision boundary. Therefore, we focus on distilling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

variation-related knowledges to augment the examples of tail classes for long-tailed datasets.

In this work, we propose a novel *Long-Tailed ReID* (LTReID) framework to solve the long-tailed ReID problem. In particular, we construct an *Independent Component Factorization* (ICF) method to decompose each person feature into an identity-related component and K variation-related components. The identity-related component purely encodes person identities, while the K variation-related components exclusively represent different image variations (*e.g.*, body poses, part occlusions and background clutters). To preserve the component independence, the two decorrelation losses, *i.e.*, *Identity Variation Decorrelation* (IVD) loss and *Multiple Variation Decorrelation* (MVD) loss, are proposed to eliminate the correlations among the $K+1$ components. For augmenting training data, an effective *Factorizable Feature Generation* (FFG) method is proposed to synthesize fake features by randomly aggregating the $K+1$ components. More importantly, we generate more fake features for tail classes than head classes and then both real and fake features are merged to train ReID models, resulting in the class-balance between head and tail classes.

Apart from the class-imbalance problem, we also take into account the hard-imbalance between easy and hard samples. Specifically, hard samples usually account for the tiny minority of training data, while the vast majority belong to easy samples. During the training phase, hard samples can produce gradients with large magnitudes, while the gradients from easy samples are close to zero. Accordingly, the hard-imbalance may engender early training saturation, which prematurely stops contributing significant gradients to back-propagation. In order to relieve this training saturation, we put forward an *Adversarial Feature Generation* (AFG) method to encourage hard feature generation with adversarial learning, which is conducive to retain the hard-balance and output significant gradients.

To sum up, this paper makes the following contributions:

- We propose the ICF method to decompose each person feature into an identity-related component and K variation-related components. The IVD and MVD losses are introduced to reduce component correlations.
- We put forward the FFG method to generate diverse fake features to relieve the class-imbalance between head and tail classes.
- We construct the AFG method to synthesize hard fake features to mitigate the hard-imbalance between easy and hard samples.
- The proposed end-to-end LTReID framework achieves state-of-the-art performance on the four person ReID datasets without increasing the time and space complexity of inference models.

II. Related Works

A. Person Re-Identification

Person ReID frameworks roughly consist of two major components, *i.e.*, representation learning and metric learning. Some conventional ReID methods [18, 24, 39, 40] primarily employ handcrafted features such as color and texture histograms. For instance, Yang *et al.* [39] put forward salient color names based on color descriptor to form a feature representation for person matching. Recently, deep learning based person ReID methods [1–17, 41] have achieved great success through simultaneously learning person features and similarities within one framework. For example, based on PCB [2, 7], some following works, *i.e.*, MGN [9], PyramidNet [10] and HPM [8], extract both global and local person representations by dividing convolutional feature maps horizontally into multi-grained patches. Additionally, some works adopt deep metric learning methods with softmax loss [42–44], triplet loss [45–47] or quadruplet loss [48] for discriminative feature learning.

Nevertheless, previous ReID works have seldomly provided flexible and effective solutions to the long-tailed data distribution. Although some ReID systems [49–52] use a *Generative Adversarial Network* (GAN) [53] to synthesize unseen person images, these methods are not explicitly concerned with the long-tailed distribution problem. With regard to imbalanced person ReID, Liu *et al.* [32] transfer the angular margin of head classes to tail classes for margin-based softmax losses [54, 55]. Besides, Liu *et al.* [56] introduce a memory-based jitter to augment tail classes with higher diversities. Different from these studies, the proposed LTReID decomposes each person feature into an identity-related component and K variation-related components. These $K+1$ components are randomly aggregated to generate more fake features for tail classes than head ones. To the best of our knowledge, this is the first attempt to manipulate multiple component factorization for fake feature generation in the long-tailed ReID task.

B. Imbalance Learning

Imbalance learning has received increasing attention due to its wide applications for many real-world problems. Current works leverage data re-sampling, cost-sensitive learning, margin learning and data generation to cope with imbalanced datasets. Firstly, for data re-sampling methods, training samples are either over-sampled (increasing sampling frequencies for tail classes) [33, 34] or under-sampled (decreasing sampling frequencies for head classes) [35, 36] to achieve a more balanced data distribution. Whereas, over-sampling duplicated samples might lead to over-fitting upon tail classes, while under-sampling informative samples will certainly lose important information of head classes. Secondly, for cost-sensitive learning methods, training loss functions are weighted at class level by multiplying different weights on different classes to promote the influence of tail classes [57, 58] or at instance level by multiplying different weights on different training samples for more fine-grained control [37, 59–62]. However, cost-sensitive learning is not capable of handling the large-scale and real-world scenarios of long-tailed data and tend to cause optimization difficulty [63]. Thirdly, for margin learning methods [64–67], a larger inter-class margin is assigned to tail classes than head classes in margin-based loss functions [54, 55, 68, 69]. Unfortunately, it is non-trivial to select appropriate and universal margins for different long-tailed data. Finally, for data generation methods [70–73], additional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

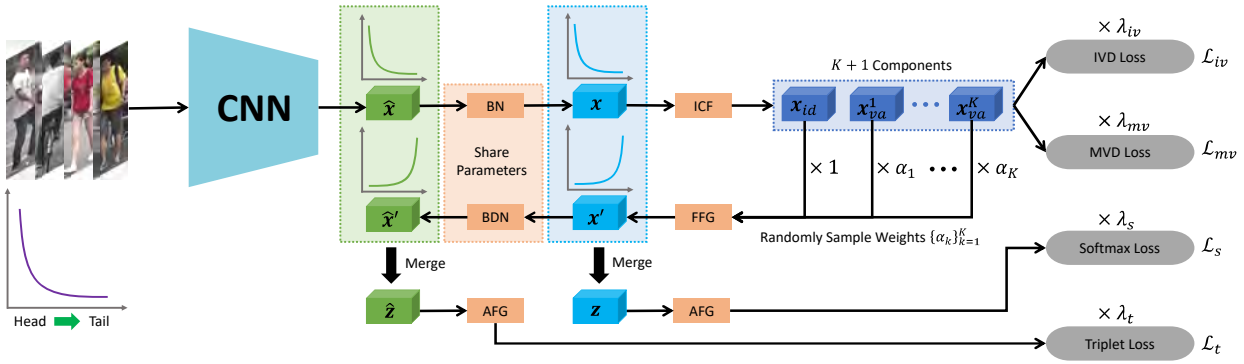


Fig. 2: Overview of the proposed LTReID framework. It consists of six key modules: (1) The backbone network aims at extracting deep person features. (2) The BN layer transforms unnormalized features into normalized ones. (3) The ICF decomposes each real feature into an identity-related component and K variation-related components. (4) The FFG generates fake features by randomly aggregating the $K+1$ components. (5) The BDN layer transforms normalized features into unnormalized ones. (6) The AFG encourages hard feature generation via an adversarial learning strategy. Note that, we merge both real and fake features to compute the softmax and triplet losses for training, while we employ the real features from the BN layer to compute cosine similarities for testing. Since the ICF, FFG, BDN and AFG modules are discarded during evaluation, the LTReID does not introduce additional time and space costs of inference models.

fake samples or features are synthesized by generative models to equilibrate long-tailed distributions. Notwithstanding, these data generation methods only consider the class-imbalance between head and tail classes, but ignore the hard-imbalance between easy and hard samples. Unlike the former studies, our method intends to generate hard fake features to overcome both class-imbalance and hard-imbalance problems in the long-tailed ReID task.

III. Proposed Method

A. Overview of LTReID Framework

Network Structure: As shown in Fig. 2, the proposed LTReID framework consists of six key modules. The first module is a backbone network $\mathcal{X}(\cdot;\boldsymbol{\theta}_{\mathcal{X}})$, where $\boldsymbol{\theta}_{\mathcal{X}}$ denotes the backbone parameter. Formally, given an input image $\boldsymbol{I}_x$, this module outputs an one-dimensional feature vector,

$$\widehat{\boldsymbol{x}} = \mathcal{X}(\boldsymbol{I}_x;\boldsymbol{\theta}_{\mathcal{X}}), \tag{1}$$

where $\widehat{\boldsymbol{x}} \in \mathbb{R}^D$ and D is the feature length. The second module uses a *Batch Normalization* (BN) layer [74] to ensure the stable convergence [75] of softmax loss and triplet loss by transforming the unnormalized feature $\widehat{\boldsymbol{x}}$ into the normalized feature $\boldsymbol{x}$ as follows,

$$\boldsymbol{x} = \gamma\frac{\widehat{\boldsymbol{x}}-\mu}{\sigma} + \beta, \tag{2}$$

where μ and σ denote the mean and standard deviation of input features, while γ and β are trainable scale and shift parameters. For the third module, the proposed *Independent Component Factorization* (ICF) decomposes the real feature $\boldsymbol{x}$ into an identity-related component $\boldsymbol{x}_{id}$ and K variation-related components $\{\boldsymbol{x}_{va}^k\}_{k=1}^K$. The fourth module named *Factorizable Feature Generation* (FFG) synthesizes diverse fake features $\boldsymbol{x}'$ by randomly aggregating the $K+1$ independent components. Contrary to the BN layer, the fifth module employs a *Batch DeNormalization* (BDN) layer to transform the normalized fake feature $\boldsymbol{x}'$ to the unnormalized fake feature $\widehat{\boldsymbol{x}}'$. The last module named *Adversarial Feature Generation* (AFG) promotes hard feature generation via an adversarial training method.

Merged Feature: Given a mini-batch of unnormalized real features $\{\widehat{\boldsymbol{x}}_i\}_{i=1}^N$ and normalized real features $\{\boldsymbol{x}_i\}_{i=1}^N$, we generate their corresponding fake features $\{\widehat{\boldsymbol{x}}_i'\}_{i=1}^G$ and $\{\boldsymbol{x}_i'\}_{i=1}^G$, where N and $G=\rho N$ represent the number of real and fake features within a mini-batch. The generation ratio ρ controls the quantity of generated fake features. Subsequently, both real and fake features are merged into a new mini-batch to counteract the long-tailed distribution,

$$\begin{aligned}\{\widehat{\boldsymbol{z}}_i\}_{i=1}^Q &= \{\widehat{\boldsymbol{x}}_i\}_{i=1}^N \bigcup \{\widehat{\boldsymbol{x}}_i'\}_{i=1}^G,\\ \{\boldsymbol{z}_i\}_{i=1}^Q &= \{\boldsymbol{x}_i\}_{i=1}^N \bigcup \{\boldsymbol{x}_i'\}_{i=1}^G,\end{aligned} \tag{3}$$

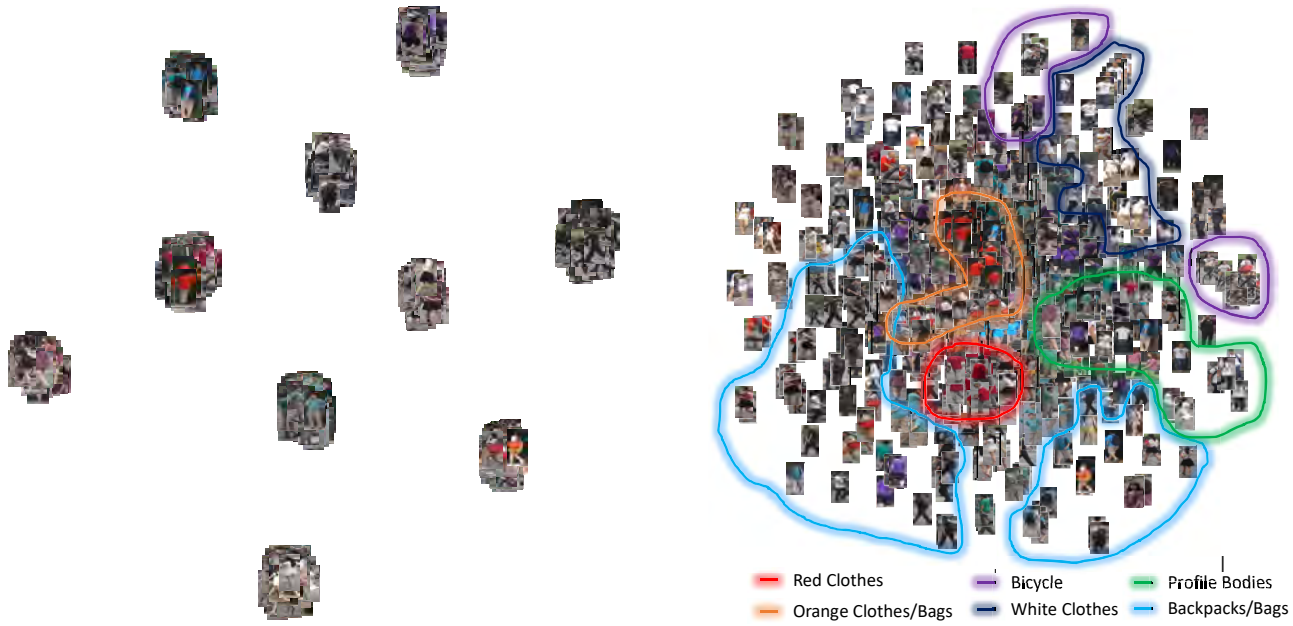
where $\widehat{\boldsymbol{z}}_i$ and $\boldsymbol{z}_i$ denote the i-th unnormalized and normalized merged feature, respectively. $Q=N+G$ denotes the number of merged features within a mini-batch. For the purpose of mitigating the inconsistence between the learning goals of softmax loss and triplet loss [75], unnormalized and normalized features are individually input into the triplet loss and softmax loss. During the testing phase, we use the backbone network and BN layer to extract real features $\boldsymbol{x}$ of query and gallery images for computing cosine similarities, while the other modules are removed to make the inference model more computationally efficient. Therefore, the LTReID framework does not introduce additional computing time and space costs for applications.

Loss Function: For the softmax loss, we build an identity classifier $\mathcal{C}$ with a learnable matrix $\boldsymbol{W} \in \mathbb{R}^{D\times C}$, where C is the identity number. This identity classifier uses a softmax function to map the normalized feature $\boldsymbol{z}$ into an identity distribution $\mathcal{C}(\boldsymbol{z};\boldsymbol{W}) = \texttt{softmax}(\boldsymbol{W}^T\boldsymbol{z})$. The softmax loss adopts cross entropy to enforce the predicted distribution to approximate the ground truth distribution by,

$$\mathcal{L}_s = -\frac{1}{Q}\sum_{i=1}^{Q}\boldsymbol{y}_i^T\log\left(\mathcal{C}(\boldsymbol{z}_i;\boldsymbol{W})\right), \tag{4}$$

where $\boldsymbol{y}_i \in \mathbb{R}^C$ denote the one-hot identity target of $\boldsymbol{z}_i$. The triplet loss aims at preserving the rank relationship among a triplet of samples with a large margin, which increases the inter-class distance and reduces the intra-class one. Formally,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,



(a) Identity Distribution **(b)** Variation Distribution

Fig. 3: The image visualization of identity and variation distributions. We use t-SNE [76] to transform $\boldsymbol{x}$ and $\boldsymbol{x}_{va}$ into two-dimensional vectors. Then, the two-dimensional vectors are viewed as spatial coordinates and we plot the input images at the coordinates of their features.

the unnormalized features are used to compute the Euclidean distance for the batch hard triplet loss [77],

$$\mathcal{L}_t = \frac{1}{Q}\sum_{i=1}^{Q}\left[\mathcal{D}\left(\widehat{\boldsymbol{z}}_i, \widehat{\boldsymbol{z}}_i^{+}\right) - \mathcal{D}\left(\widehat{\boldsymbol{z}}_i, \widehat{\boldsymbol{z}}_i^{-}\right) + m\right]_{+}, \tag{5}$$

where $[\cdot]_+$ indicates the hinge loss, $\mathcal{D}(\cdot)$ computes the Euclidean distance and m is a distance margin. $\widehat{\boldsymbol{z}}_i$, $\widehat{\boldsymbol{z}}_i^{+}$ and $\widehat{\boldsymbol{z}}_i^{-}$ denote the anchor, hard-positive and hard-negative features.

B. Independent Component Factorization

Image Visualization: For a given input image $\boldsymbol{I}_x$, we extract the corresponding person feature $\boldsymbol{x}$ and the variation feature $\boldsymbol{x}_{va}$. As shown in Fig. 3a, the images of the same person obey a Gaussian distribution. Since the Gaussian distribution is determined by mean and variance values, the person feature can be approximately decomposed into an identity-related component and a variation-related component. Interestingly, Fig. 3b shows that person variations are related to many aspects, *e.g.*, the color of clothes/bags, background objects and body poses. Besides, the samples with similar variations are easily grouped together in the feature space. As a result, the variation component can be further decomposed into multiple variation-related subcomponents.

Component Factorization: According to the previous observation that identity and variation knowledges are mixed together in the feature space, we put forward an *Independent Component Factorization* (ICF) method to decompose each person feature into an identity-related component and K variation-related components. Specifically, the identity-related component purely encodes person identity information, while the K variation-related components exclusively capture different image variations (*e.g.*, body poses, part occlusions and background clutters). With a proper combination, these $K+1$ components are able to reconstruct the original person feature. Formally, given a feature $\boldsymbol{x}$, we have

$$\boldsymbol{x} = \boldsymbol{x}_{id} + \sum_{k=1}^{K}\alpha_k \boldsymbol{x}_{va}^{k}, \ \text{s.t.} \sum_{k=1}^{K}\alpha_k = 1, \tag{6}$$

where $\boldsymbol{x}_{id} \in \mathbb{R}^D$ denotes the identity-related component, while $\boldsymbol{x}_{va}^{k} \in \mathbb{R}^D$ refers to the variation-related component of the k-th image variation. In addition, $\alpha_k \geq 0$ is a non-negative weight parameter controlling the ingredient proportion of the k-th image variation. Note that we set the summation of weights as 1 in order to balance the contributions of identity-related and variation-related components.

Component Distillation: Current studies have argued that the center feature of each class encodes identity information and abates variation information [78]. Therefore, we employ the center feature $\boldsymbol{x}_{id}^{b}$ of each person within a mini-batch to represent the identity-related component,

$$\boldsymbol{x}_{id}^{b} = \frac{1}{\left|\mathcal{S}\left(\boldsymbol{x}_{id}^{b}\right)\right|}\sum_{i\in\mathcal{S}\left(\boldsymbol{x}_{id}^{b}\right)}\boldsymbol{x}_i, \tag{7}$$

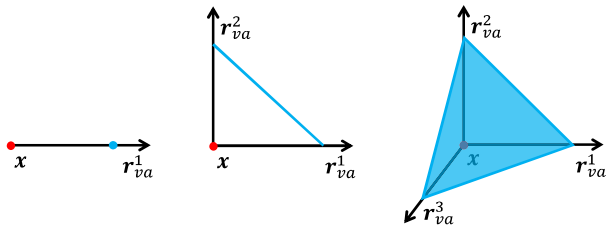
where $\mathcal{S}\left(\boldsymbol{x}_{id}^{b}\right)$ denotes the index set of person features with the same identity as $\boldsymbol{x}_{id}^{b}$ and $|\cdot|$ returns the size of a set. Following BN [74], we adopt an exponential moving average strategy to estimate the identity-related component $\boldsymbol{x}_{id}$ of all training samples belonging to the same person by,

$$\boldsymbol{x}_{id} = (1-\eta)\boldsymbol{x}_{id} + \eta\boldsymbol{x}_{id}^{b}, \tag{8}$$

where we set the momentum parameter $\eta = 0.005$ as default. After that, we acquire the overall variation-related component $\boldsymbol{x}_{va} \in \mathbb{R}^d$ by subtracting $\boldsymbol{x}_{id}$ from $\boldsymbol{x}$ as follows,

$$\boldsymbol{x}_{va} = \boldsymbol{x} - \boldsymbol{x}_{id}. \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,



(a) $K = 1$ (b) $K = 2$ (c) $K = 3$

Fig. 4: The spatial distribution of generated features with different K. (a) The feature distribution is a point (0-simplex) when $K = 1$. (b) The feature distribution is a line segment (1-simplex) when $K = 2$. (c) The feature distribution is a triangle (2-simplex) when $K = 3$. It is worth noting that the norm and direction of $\boldsymbol{r}_{va}^k$ ($k \in \{1, 2, 3\}$) are dynamically changed during training, so the synthetic features are able to cover the whole feature space.

Since discrepant image variations are likely to distribute in different feature space, multiple variation-related modules $\{\phi_k\}_{k=1}^K$ are constructed to further disintegrate $\boldsymbol{x}_{va}$ into K sub-components $\{\boldsymbol{x}_{va}^k\}_{k=1}^K$. Inspired by the recent works on residual learning [79, 80], we have noticed that learning residual variations $\boldsymbol{r}_{va}^k = \phi_k(\boldsymbol{x}_{va}; \boldsymbol{\theta}_{\phi_k})$ is beneficial to excavate potential variation knowledges,

$$\boldsymbol{x}_{va}^k = \boldsymbol{x}_{va} + \boldsymbol{r}_{va}^k = \boldsymbol{x}_{va} + \phi_k(\boldsymbol{x}_{va}; \boldsymbol{\theta}_{\phi_k}), \tag{10}$$

where $\boldsymbol{\theta}_{\phi_k}$ denotes the parameter of the k-th variation-related module. In practice, we adopt K different *Fully-Connected* (FC) layers to model these modules $\{\phi_k\}_{k=1}^K$.

Component Decorrelation: Through component purification, it is crucial for feature decomposition that $\boldsymbol{x}_{id}$ should be identity preserving and necessarily unrelated to image variations. We design an *Identity Variation Decorrelation* (IVD) loss to maintain the independence between identity-related and variation-related components. If $\boldsymbol{x}_{id}$ and $\boldsymbol{x}_{va}$ are mutually uncorrelated, the predicted identity distribution $\mathcal{C}(\boldsymbol{x}_{va}; \boldsymbol{W}) = \texttt{softmax}(\boldsymbol{W}^T \boldsymbol{x}_{va})$ should be a uniform distribution. Based on the maximum entropy principle, it is equivalent to minimizing the negative entropy of the predicted identity distribution as follows,

$$\mathcal{L}_{iv} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}(\boldsymbol{x}_{va,i}; \boldsymbol{W})^T \log\left(\mathcal{C}(\boldsymbol{x}_{va,i}; \boldsymbol{W})\right), \tag{11}$$

where the classifier weight $\boldsymbol{W}$ is shared with Eq. 4. Furthermore, since different variation-related components are probably associated with each other, we construct a *Multiple Variation Decorrelation* (MVD) loss to eliminate the correlation among the K variation-related components. Specifically, the MVD loss reduces the cosine similarity between pairs of residual variation-related components,

$$\mathcal{L}_{mv} = \frac{1}{2NK(K-1)} \sum_{i=1}^{N} \sum_{k_1 \neq k_2}^{K} \left| \left\langle \boldsymbol{r}_{va,i}^{k_1}, \boldsymbol{r}_{va,i}^{k_2} \right\rangle \right|, \tag{12}$$

where $\boldsymbol{r}_{va,i}^k$ denotes the k-th residual variation-related components of the i-th person features. $\langle \cdot \rangle$ computes the cosine similarity and $|\cdot|$ returns the absolute value.

C. Factorizable Feature Generation

Feature Generation: In order to enhance the varieties of training data, we propose a new *Factorizable Feature Generation* (FFG) method to synthesize fake features by randomly

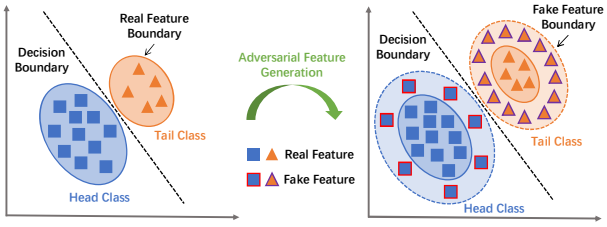


Fig. 5: A toy example of the *Adversarial Feature Generation* (AFG) method. The number of fake features in tail classes is more than head classes. The AFG generates hard fake features by pushing fake features close to the decision boundary and therefore real features would be moved away from the decision boundary. In other words, the hard fake features contribute to maximizing intra-class similarities and inter-class differences of real features.

aggregating the $K+1$ independent components. When Eq. 10 is substituted into Eq. 6, the generated fake normalized feature $\boldsymbol{x}' \in \mathbb{R}^D$ is formulated as follows,

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{x}_{id} + \sum_{k=1}^{K} \alpha_k \left(\boldsymbol{x}_{va} + \boldsymbol{r}_{va}^k\right) \\ &= \boldsymbol{x} + \sum_{k=1}^{K} \alpha_k \boldsymbol{r}_{va}^k, \ \text{s.t.} \sum_{k=1}^{K} \alpha_k = 1, \end{aligned} \tag{13}$$

where the weights $\{\alpha_k\}_{k=1}^K$ are randomly sampled so as to generate dissimilar features. As reflected in Fig. 4, the fake normalized feature can also be viewed as the combination of the original feature $\boldsymbol{x}$ and K residual variation-related components $\{\boldsymbol{r}_{va}^k\}_{k=1}^K$. More interestingly, it seems that it's a trend to generate richer features along with the increase of K. For unnormalized feature generation, we use a *Batch De-Normalization* (BDN) layer to transform normalized features $\boldsymbol{x}'$ to unnormalized features $\widehat{\boldsymbol{x}}'$ as follows,

$$\widehat{\boldsymbol{x}}' = \sigma \frac{\boldsymbol{x}' - \beta}{\gamma} + \mu, \tag{14}$$

where μ, σ, γ and β are shared with the BN layer in Eq. 2. With the BDN layer, we do not need to employ the ICF to decompose the real unnormalized features into $K+1$ components for generating fake unnormalized features.

Reversed Generation: However, we can not guarantee that learned person features are robust to class imbalance if we only uniformly generate fake features for each class without considering the long-tailed distribution of training data. Our solution is to design a reversed feature generation strategy to synthesize more fake features for tail classes and fewer fake features for head classes. Motivated by [56, 63], the inverse class frequency is adopted to formulate the proportional distribution of generated fake features,

$$\pi_c = \frac{N_c^{-\tau}}{\sum_{l \in \mathcal{S}_C} N_l^{-\tau}}, \ \forall c \in \mathcal{S}_C, \tag{15}$$

where $\mathcal{S}_C$ denotes the class index set within a mini-batch and N_c refers to the c-th class frequency in entire training data. Besides, the generation exponent $\tau \geq 0$ controls the quantity distribution of generated features. Specifically, the higher τ leads to lower entropy, which makes the FFG concentrate on generating more features for tail classes. The number of the fake features in c-th class is $G_c = \lceil \pi_c G \rceil$, where $\lceil \cdot \rceil$ is a ceiling operator. As seen in Fig. 2, since the distributions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

TABLE I: Comparisons with state-of-the-art methods on Market1501, CUHK03, DukeMTMC and MSMT17 datasets. CUHK03-L and CUHK03-D use labeled and detected bounding boxes to crop person images on CUHK03, respectively. Note that we adopt the strong baseline model in [75], where the normalized and unnormalized features are trained by softmax loss and triplet loss, respectively.

Method	Market1501		CUHK03-L		CUHK03-D		DukeMTMC		MSMT17	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
PDC [81]	84.14	63.41	-	-	-	-	-	-	58.00	29.70
GLAD [82]	89.90	73.90	-	-	-	-	-	-	61.40	34.00
PAB [83]	91.70	79.60	-	-	-	-	84.40	69.30	-	-
Mancs [84]	93.10	82.30	69.00	63.90	65.50	60.50	84.90	71.80	-	-
PCB + RPP [2]	93.80	81.60	63.70	57.50	-	-	83.30	69.20	68.20	40.40
MGN [9]	95.70	86.90	68.00	67.40	66.80	66.00	88.70	78.40	-	-
HPM [8]	94.20	82.70	-	-	63.90	57.50	86.60	74.30	-	-
DGNet [51]	94.80	86.00	-	-	-	-	86.60	74.80	77.20	52.30
CASN [43]	94.40	82.80	73.70	68.00	71.50	64.40	87.70	73.70	-	-
IANet [85]	94.40	83.10	-	-	-	-	83.10	73.40	75.50	46.80
DSAReID [86]	95.70	87.60	78.90	75.20	78.20	73.10	86.20	74.30	-	-
PyramidNet [10]	95.70	88.20	78.90	76.90	78.90	74.80	89.00	79.00	-	-
MHN [87]	95.10	85.00	77.20	72.40	71.70	65.40	89.10	77.20	-	-
OSNet [88]	94.80	84.90	-	-	72.30	67.80	88.60	73.50	78.70	52.90
BFE [89]	95.30	86.20	79.40	76.70	76.40	73.50	88.90	75.90	78.80	51.50
SAN [90]	96.10	88.00	80.10	76.40	79.40	74.60	87.90	75.50	79.20	55.70
RGA-SC [13]	96.10	88.40	81.10	77.40	79.60	74.50	-	-	80.30	57.50
LEAP-CF [32]	93.50	84.20	-	-	-	-	87.80	74.20	76.70	50.80
GASM [15]	95.30	84.70	-	-	-	-	88.30	74.40	79.50	52.50
PISNet [91]	95.60	87.10	-	-	-	-	88.80	78.70	-	-
ISP [14]	95.30	88.60	76.50	74.10	75.20	71.40	89.60	80.00	-	-
CSPR [12]	94.20	84.80	64.70	62.80	-	-	83.50	71.90	75.30	50.80
ReID-NAS [92]	95.10	85.70	-	-	-	-	88.10	74.60	79.50	53.30
Occluded-ReID [93]	92.70	81.30	-	-	-	-	86.20	72.60		
RFC [94]	95.20	89.20	-	-	81.10	78.00	90.70	80.70	82.00	60.20
APNet [95]	**96.20**	90.50	**87.40**	**85.30**	83.00	81.50	90.40	81.50	**83.70**	**63.50**
ResNet50 + Baseline	94.60	86.52	77.07	74.57	73.43	70.60	88.06	76.59	78.10	53.91
ResNet50 + LTReID	95.89	89.02	82.14	80.33	81.81	79.07	90.48	80.35	81.02	58.64
ResNet101 + Baseline	94.63	87.55	79.50	77.56	79.21	75.44	89.95	78.59	80.71	57.34
ResNet101 + LTReID	96.12	**90.65**	87.14	85.24	**85.57**	**83.63**	**91.20**	**82.09**	83.33	62.71

of real and fake features are reversed, the joint of these two features is able to equilibrate the sample distribution, which contributes to learning balanced features.

D. Adversarial Feature Generation

The FFG method only considers the class-imbalance of class frequencies but ignores the hard-imbalance between easy and hard samples. For the sake of solving this hard-imbalance problem, an *Adversarial Feature Generation* (AFG) method is proposed to generate enough hard fake features. Our goal is to learn discriminative features and generate hard features simultaneously. Mathematically, the training process is formulated as a minimax game, leading to an adversarial learning problem [53] as,

$$\min_{\boldsymbol{\theta}_D} \max_{\boldsymbol{\theta}_H} \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t, \tag{16}$$

where λ_s and λ_t are the loss weights of $\mathcal{L}_s$ and $\mathcal{L}_t$. In addition, the optimization of $\boldsymbol{\theta}_D = \{\boldsymbol{\theta}_{\mathcal{X}}, \boldsymbol{W}, \gamma, \beta\}$ aims at learning discriminative features, which are beneficial to acquire reliable variation-related components. Contrastingly, the optimization of $\boldsymbol{\theta}_H = \{\boldsymbol{\theta}_{\phi_k}\}_{k=1}^{K}$ focuses on generating hard features, which would enlarge loss values and produce gradients with large magnitudes to back-propagation.

As depicted in Fig. 5, the impact of the AFG method is attributed to three aspects. (1) The AFG helps to augment enough hard features and therefore balances the quantity distribution of easy and hard samples. (2) The AFG prefers pushing hard features as close as possible to decision boundaries and makes feature learning concentrate on hard samples which have large overlaps with neighboring classes. (3) Since the LTReID tries to classify hard features, the AFG contributes to reducing intra-class distances and enlarging inter-class distances. Therefore, the AFG not only equilibrates the hard-imbalance between easy and hard samples, but also facilitates discriminative feature learning.

E. Overall Loss Function

We optimize the LTReID framework in an end-to-end manner by cascading $\mathcal{L}_s$, $\mathcal{L}_t$, $\mathcal{L}_{iv}$ and $\mathcal{L}_{mv}$ as follows,

$$\min_{\boldsymbol{\theta}_D} \max_{\boldsymbol{\theta}_H} (\lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t) + \min_{\boldsymbol{\theta}_D, \boldsymbol{\theta}_H} (\lambda_{iv} \mathcal{L}_{iv} + \lambda_{mv} \mathcal{L}_{mv}), \tag{17}$$

where λ_{iv} and λ_{mv} are loss weights of $\mathcal{L}_{iv}$ and $\mathcal{L}_{mv}$, respectively. The training process includes two sub-process: (1) fix $\boldsymbol{\theta}_H$ and update $\boldsymbol{\theta}_D$; (2) fix $\boldsymbol{\theta}_D$ and update $\boldsymbol{\theta}_H$. The alternative learning process is formulated as follows,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_D &= \operatorname*{argmin}_{\boldsymbol{\theta}_D} \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t + \lambda_{iv} \mathcal{L}_{iv} + \lambda_{mv} \mathcal{L}_{mv}, \\ \widehat{\boldsymbol{\theta}}_H &= \operatorname*{argmax}_{\boldsymbol{\theta}_H} \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t - \lambda_{iv} \mathcal{L}_{iv} - \lambda_{mv} \mathcal{L}_{mv}, \end{aligned} \tag{18}$$

where $\widehat{\boldsymbol{\theta}}_D$ and $\widehat{\boldsymbol{\theta}}_H$ denote the optimal solutions of $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_H$, respectively. Note that we set $\lambda_s = 1.0$, $\lambda_t = 1.0$, $\lambda_{iv} = 0.1$ and $\lambda_{mv} = 10.0$ as default in our experiments.

IV. Experiments

A. Dataset

For the purpose of justifying the effectiveness of our method, we evaluate the proposed LTReID framework on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

TABLE II: Comparisons with state-of-the-art methods on long-tailed Market1501 and DukeMTMC. Different long-tailed datasets are constructed by varying the number of head class. H is the number of head classes and S denotes the instance number per tail class. Note that all methods use ResNet50 as the backbone.

Long-Tailed	Method	Market1501		DukeMTMC	
		Rank1	mAP	Rank1	mAP
$\langle H100, S5\rangle$	LEAP-CV [32]	86.50	68.70	74.80	55.60
	LEAP-AV [32]	87.30	69.80	76.50	57.90
	MBJ [56]	88.40	72.60	78.60	60.80
	Baseline	90.97	77.66	82.32	67.04
	LTReID	**92.84**	**82.60**	**84.78**	**70.96**
$\langle H50, S5\rangle$	LEAP-CV [32]	84.90	67.30	73.00	53.10
	LEAP-AV [32]	84.60	67.10	73.50	54.40
	MBJ [56]	86.20	68.80	74.40	56.70
	Baseline	89.79	75.26	80.92	65.41
	LTReID	**91.54**	**80.28**	**82.50**	**68.07**
$\langle H20, S5\rangle$	LEAP-CV [32]	83.20	64.10	72.70	52.40
	LEAP-AV [32]	82.20	64.30	73.70	54.20
	MBJ [56]	84.80	66.70	75.50	57.90
	Baseline	87.32	70.72	79.37	64.70
	LTReID	**90.56**	**77.95**	**82.67**	**66.76**

four holistic datasets, *i.e.*, Market1501 [25], CUHK03 [26], DukeMTMC [27] and MSMT17 [28].

Market1501 [25]: It contains 32,668 images of 1,501 persons captured by six camera views. The whole dataset is divided into a training set containing 12,936 images of 751 persons and a testing set containing 19,732 images of 750 persons. For each person in testing set, we select one image from each camera as a query image, forming 3,368 queries following the standard setting in [25].

CUHK03 [26]: It contains 14,097 images of 1,467 persons, captured by six camera views. Two types of person images are provided: manually labeled person bounding boxes (Labeled) and automatically detected bounding boxes (Detected). We use the settings of both labeled and detected person images on the splits in [96], where 767 and 700 persons are used for training and testing, respectively.

DukeMTMC [27]: It contains 36,411 images of 1,812 persons captured by 8 cameras, where only 1,404 persons appeared in more than 2 cameras. The other 408 persons are regarded as distractors. The training set contains 16,522 images of 702 persons while the testing set contains 2,228 query images of 702 persons and 17,661 gallery images.

MSMT17 [28]: It contains manually annotated 126,441 bounding boxes of 4,101 persons, which is currently the largest person ReID dataset. All images are captured by the 15-camera network deployed in campus, which contains 12 outdoor cameras and 3 indoor cameras. The training set contains 32,621 bounding boxes of 1,041 persons, and the testing set contains 93,820 bounding boxes of 3,060 persons. From the testing set, 11,659 bounding boxes are randomly selected as query images and the other 82,161 bounding boxes are used as gallery images.

B. Implementation Details

Network Architecture: We take the ResNet-50/101 [79] initialized with the parameters pretrained on ImageNet [97] as the backbone network. Following the work [8], the last fully-connected layer and global average pooling layer are removed and the stride of the last residual block *Conv4_1* is set from 2 to 1 for increasing the feature map size.

TABLE III: Comparisons with state-of-the-art methods on long-tailed Market1501 and DukeMTMC. Different long-tailed datasets are constructed by varying the instance number per tail class. Note that all methods use ResNet50 as the backbone.

Long-Tailed	Method	Market1501		DukeMTMC	
		Rank1	mAP	Rank1	mAP
$\langle H20, S5\rangle$	LEAP-CF [32]	83.40	65.20	72.80	52.70
	LEAP-AF [32]	83.20	63.90	73.60	54.20
	Baseline	87.32	70.72	79.37	64.70
	LTReID	**90.56**	**77.95**	**82.67**	**66.76**
$\langle H20, S4\rangle$	LEAP-CF [32]	76.80	54.70	63.00	42.60
	LEAP-AF [32]	77.90	56.50	64.40	44.20
	Baseline	84.83	67.11	76.75	59.53
	LTReID	**88.63**	**74.33**	**79.22**	**63.25**
$\langle H20, S3\rangle$	LEAP-CF [32]	67.20	43.50	51.10	33.20
	LEAP-AF [32]	66.10	44.10	53.30	34.30
	Baseline	80.76	60.36	70.51	52.59
	LTReID	**86.49**	**69.95**	**74.96**	**57.66**

Data Processing: In order to obtain enough context information from person images and a proper size of feature map for the proposed LTReID framework, we first resize training images to 384×128. Then we randomly crop each training image with scale in the interval $[0.64, 1.0]$ and aspect ratio $[2, 3]$. Third, we resize these cropped images back to 384×128. Following the work [84], the training images are augmented with horizontal flipping and random erasing [98]. Before it is sent to the network, each image is subtracted from the mean values $[0.485, 0.456, 0.406]$ and divided by the standard deviations $[0.229, 0.224, 0.225]$ according to normalization procedure when using the pretrained model on ImageNet.

Training/Testing Configurations: Since triplet loss is used to learn person features, we need to adopt an appropriate triplet sampling strategy. To simplify this procedure, triplets are generated using the $\mathcal{PK}$ sampling method [77], which randomly samples $\mathcal{P}$ classes and then randomly selects $\mathcal{K}$ images for each person to form a mini-batch with the size $\mathcal{P} \times \mathcal{K}$. In a mini-batch, we use all possible $\mathcal{PK}\,(\mathcal{PK}-\mathcal{K})\,(\mathcal{K}-1)$ combinations of triplets for triplet loss. For all datasets, $\mathcal{P}$ and $\mathcal{K}$ are set to 16 and 4, respectively. Following the work [75], we warm up the model for 10 epochs with a linearly growing learning rate from 3.5×10^{-5} to 3.5×10^{-4}. Then, the learning rate is deceased by a factor 0.1 at 40th and 70th epoch. We observe that 120 epochs are enough for model converging. The batch size is set to 64 and Adam method is adopted to optimize the model. All our methods are implemented on PyTorch [99]. All experiments run on a server with 2 Intel(R) Xeon(R) E5-2620 v4@2.10GHz CPUs, 4 GeForce GTX 1080 Ti GPU and 128G RAM.

C. Comparison with State-of-the-Art Methods

Original Dataset: We compare the LTReID framework with other state-of-the-art methods on Market1501, CUHK03, DukeMTMC and MSMT17 datasets. The comparisons are reported in Table I. Experimental results show that our baseline has surpassed many advanced methods and the LTReID further improves performances compared with the baseline [75]. In addition, compared with other methods, the results show that the proposed LTReID achieves superior or competitive retrieval accuracies. Compared with other datasets, the MSMT17

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

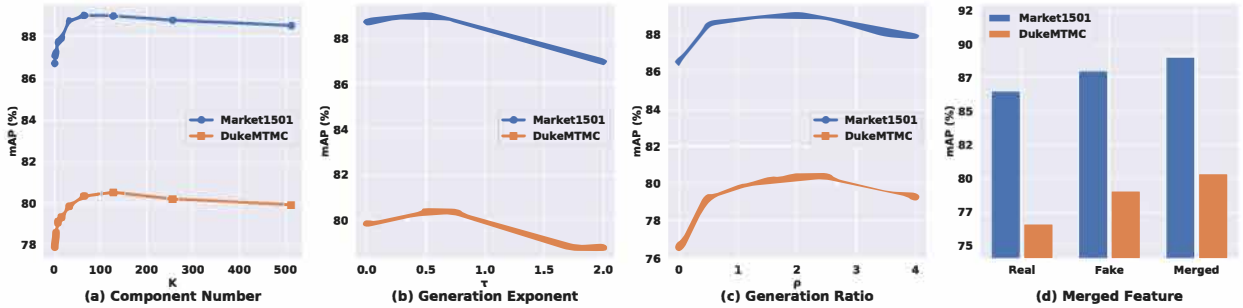


Fig. 6: Ablation studies on Market1501 and DukeMTMC datasets. (a) Analyzing the impact of the variation-related component number K of Eq. 13. (b) Analyzing the sensitivity of the generation exponent τ of Eq. 15. (c) Analyzing the sensitivity of the generation ratio ρ. (d) Analyzing the effect of the merged features. Note that all ReID models adopt ResNet50 as the backbone network.

TABLE IV: Ablation studies on Market1501, DukeMTMC and MSMT17 datasets. "Res" denotes the residual variation is used for the ICF. "UFG" denotes the unnormalized feature generation while "NFG" denotes the normalized feature generation. All methods use ResNet50 as the backbone.

Method			mAP		
Res	UFG	NFG	Market1501	DukeMTMC	MSMT17
✘	✘	✘	86.52	76.59	53.91
✘	✔	✔	88.20	78.41	60.27
✔	✘	✔	88.00	79.51	62.53
✔	✔	✘	87.30	77.63	59.90
✔	✔	✔	**89.02**	**80.35**	**63.50**

TABLE V: Ablation studies of the all framework on Market1501 and DukeMTMC. Note that all methods use ResNet50 as the backbone.

Method				mAP	
FFG	AFG	IVD	MVD	Market1501	DukeMTMC
✘	✘	✘	✘	86.52	76.59
✔	✘	✘	✘	86.97	77.49
✔	✔	✘	✘	87.88	78.80
✔	✔	✔	✘	88.23	79.45
✔	✔	✘	✔	88.81	79.94
✔	✔	✔	✔	**89.02**	**80.35**

dataset presents the following challenges: (1) large number of person identities, bounding boxes and cameras; (2) complex scenes and backgrounds; (3) multiple time slots with severe lighting changes. Apart from APNet [95], the LTReID still achieves higher retrieval accuracies than the other methods. This clearly demonstrates that the LTReID is able to achieve a satisfactory generalization on the large-scale dataset.

Long-Tailed Dataset: To interpret the impact of long-tailed distributions on training ReID models, we construct different long-tailed datasets based on the original datasets. Following the work [32], the classes are ranked by their number of samples and then the top classes $H \in \{100, 50, 20\}$ are marked as the head classes, respectively. The remaining classes are treated as the tail classes and the number of samples is reduced to $S \in \{5, 4, 3\}$ for each tail class. In order to study the impact of head classes, we fix $S = 5$ and vary $H \in \{100, 50, 20\}$ in Table II. The results show that the baseline model [75] has higher Rank1 and mAP accuracies than LEAP [32] and MBJ [56] on the same long-tailed setting. This indicates that the baseline model has a stronger robustness for the long-tailed problem. Moreover, the proposed LTReID achieves vastly superior results over the strong baseline. This indicates that the LTReID framework is effective to relieve imbalanced distributions. In order to explore the impact of tail classes, we fix $H = 20$ and vary $S \in \{5, 4, 3\}$ in Table III. The results show that the performance of all ReID models drops dramatically when the samples of tail classes are gradually reduced. In addition, the proposed LTReID still improves ReID performance over the baseline with significant margins. By analyzing the results in Table II and III, we observe an interesting phenomenon. As the long-tailed distribution is more serious, the performance improvement of our method becomes even more significant.

D. Ablation Study

Component Number K: We investigate the sensitivity of the variation-related component number K of Eq. 13. When $K = 0$, the resulting system is equivalent to the baseline model without using feature generation. As reported in Fig. 6(a), the LTReID reaches the best performance when $K = 64$. With the increasing of the number K, the mAP scores are significantly improved by 2.30% on Market1501 and 2.46% on DukeMTMC from $K = 1$ to $K = 64$. This indicates that the increasing of K contributes to enriching the diversity of generated fake features. We also try the larger K settings, *i.e.*, $K > 64$. However, the too large K would incur additional computational costs without leading to any observable performance improvements. Therefore, we recommend to set $K = 64$ as default in this work.

Generation Exponent τ: We study the sensitivity of the generation exponent τ of Eq. 15, which is associated with the distribution of fake features. Specifically, the LTReID uniformly generates fake features when $\tau = 0$, while the LTReID generates more features for tail classes and fewer features for head classes when $\tau > 0$. Two interesting phenomena are observed in Fig. 6(b). First, a larger generation exponent τ benefits person ReID performance and the LTReID reaches the best performance when $\tau = 0.5$. Second, increasing the generation exponent ($\tau > 0.5$) significantly degrades ReID accuracies. To some extent, this is because the large generation exponent dramatically decreases the number of generated features for head classes. Therefore, the distribution of merged features may not be well balanced between head and tail classes. In this work, we recommend $\tau = 0.5$ as it strikes a satisfactory balance between the data equilibrium and retrieval performance.

Generation Ratio ρ: We explore the sensitivity of the generation ratio ρ, which is associated with the quantity of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

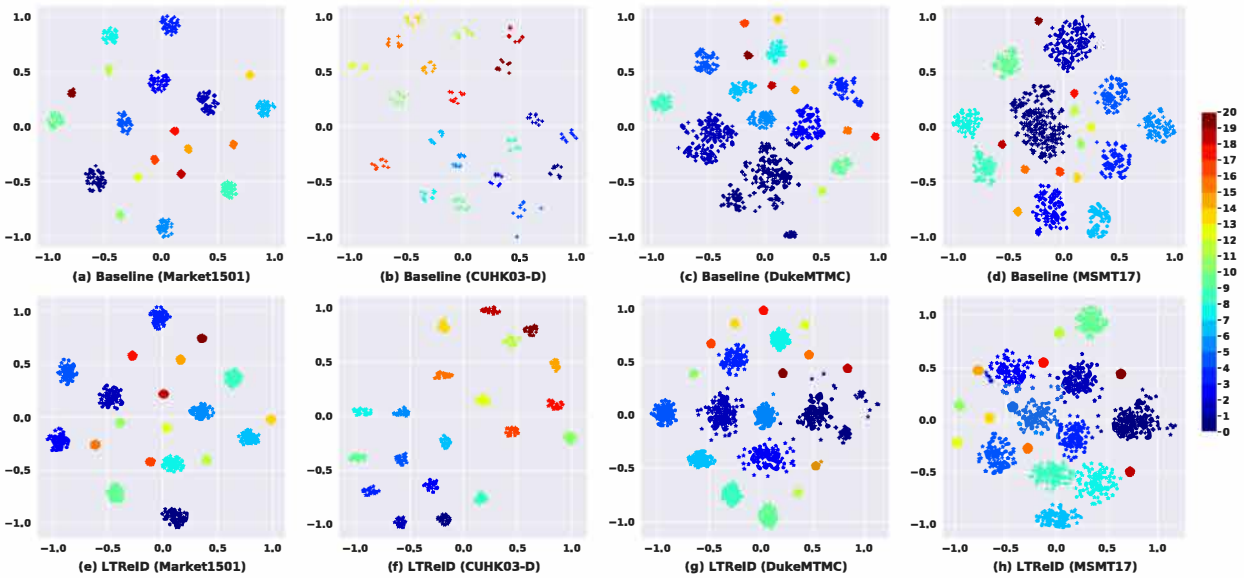


Fig. 7: The t-SNE [76] visualization of 10 head classes (from 0 to 9) and 10 tail classes (from 10 to 19) on on Market1501 [25], CUHK03 [26], DukeMTMC [27] and MSMT17 [28] datasets. The different classes are distinguished by different colors. The real and fake features are marked as "+" and "★", respectively.

generated fake features. When $\rho = 0$, the resulting system is equivalent to the baseline model without using generated features. As shown in Fig. 6(c), the LTReID reaches the best performance when $\rho = 2.0$. Furthermore, we also observe that simply using too small or large ρ is not an ideal parameter setting, resulting in poor ReID performance. The main reason is that the appropriate number of generated features may be helpful to relieve the long-tailed distributions. Unfortunately, too many generated features are likely to bring unpleasant noise and hamper training convergence, which reduces the generalization capability of ReID models. Given the above, we set $\rho = 2.0$ in this work.

Merged Feature: We analyze the contributions of the real, fake and merged features on network optimization. In Fig. 6(d), the results show that training models with the fake features consistently achieves superior mAP scores than the real features. This phenomenon indicates that the fake features are more suitable than the real features in the long-tailed ReID task. In order to show the effectiveness of the joint learning of the two features, we merge them into a new mini-batch to obtain complete merged features for training balanced ReID models. Interestingly, it is observed that the merged features significantly outperform either of these two features. Accordingly, we would recommend to use merged features for the LTReID framework.

Residual Variation: In this part, we study the effectiveness of residual variations in Eq. **10**. From the results in Table IV, decomposing residual variations achieves superior ReID performance over decomposing original variations on the three datasets. This suggests that it is easier to decompose residual variations than to decompose the original variations. From the perspective of feature generation, the original FFG randomly aggregates identity-related components and K variation-related components, while the residual FFG randomly aggregates original features and K residual variation-related components. Therefore, compared to the original FFG, the residual FFG is able to maintain the identity-related knowledges, even when some identity-related information is improperly assigned to the variation-related components. Accordingly, the residual variation decomposition enhances the predictive and stable behavior of the training process, which is more suitable for training data with large variations.

Feature Generation: As fake feature generation is key to balancing feature distributions, it is worth exploring the impact of unnormalized and normalized fake feature generation. As shown in Table IV, the normalized fake features are observed to perform significantly better than the unnormalized ones. This observation suggests that the normalized feature generation plays a more critical role in equilibrating the long-tailed distributions than the unnormalized feature generation. More interestingly, it is found that leveraging both unnormalized and normalized feature generation significantly outperforms either of them on the three datasets. Although the normalized feature generation has some advantages over the unnormalized feature generation, they are complementary to each other. Hence, we would recommend to use both unnormalized and normalized feature generation for the LTReID framework.

Overall Framework: Finally, we examine key design choices in the proposed LTReID framework. The person ReID performance on different datasets is reported in Table V. The results show that the LTReID model with the FFG outperforms the baseline model on each dataset. This is because fake features generated by the FFG are able to relieve the class-imbalance of training data. Moreover, the joint of the FFG and AFG further improves the ReID performance with significant margins, which indicates that hard feature generation is conducive to mitigate the hard-imbalance between easy and hard samples. Interestingly, the IVD and MVD loss functions consistently achieve a significant performance improvement for the LTReID model. This improvement demonstrates that reducing the correlations between different components is beneficial to boost the generalization capability of the LTReID framework.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TMM.2022.3179902,

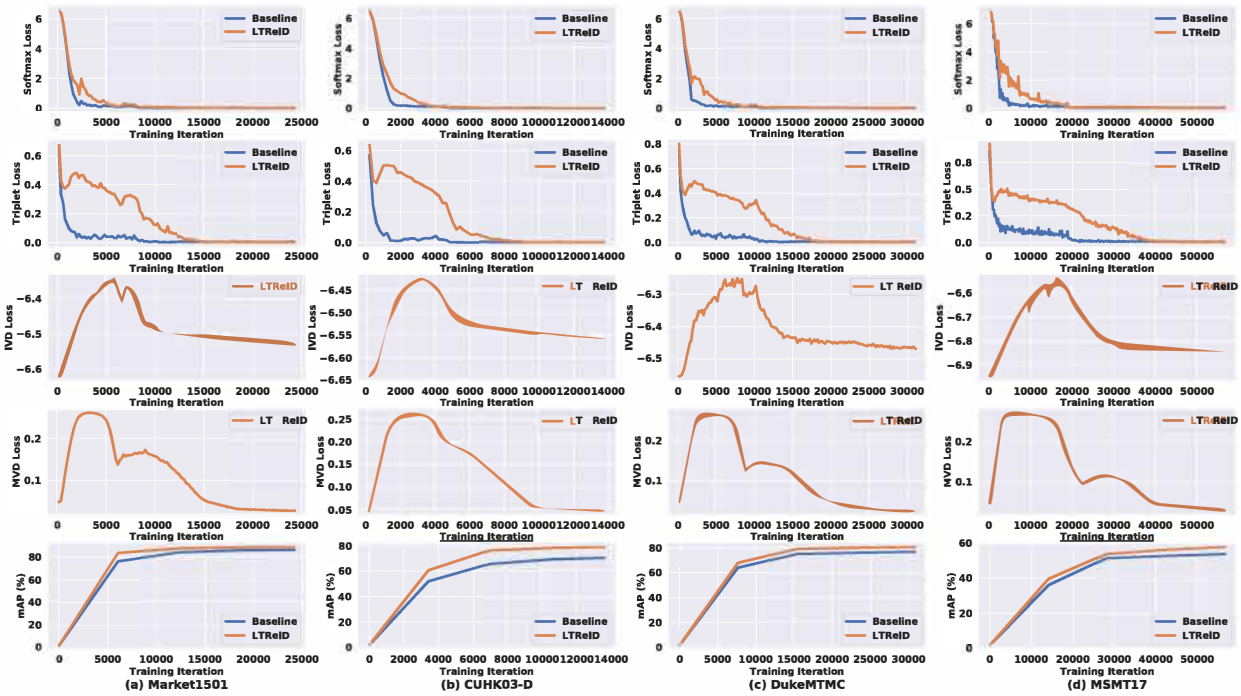


Fig. 8: The curves of softmax loss, triplet loss, IVD loss, MVD loss and mAP on Market1501, CUHK03, DukeMTMC and MSMT17 datasets.

E. Visualization Analysis

Feature Visualization: For demonstrating the effectiveness of the proposed method, we visualize the person features extracted from the baseline and LTReID models, respectively. As shown in Fig. 7, the features of different classes on CUHK03 dataset exhibit more uniform distributions than other datasets. Moreover, the features of head classes have a much larger spatial span than tail classes on MSMT17 dataset. This is because MSMT17 dataset has a more imbalanced data distribution than other datasets. Worse still, the long-tailed distribution seriously distorts the overall feature space and it is hard for tail classes to be separated from other classes. Consequently, the long-tailed distribution compromises the discriminative ability of the learned features. Moreover, since fake features expand the space of tail classes, it is more effective to push tail classes away from other classes. That is to say, the LTReID is able to reduce intra-class distances and enlarge inter-class differences, which is beneficial to relieve the long-tailed ReID problem.

Training Saturation: With regard to the influence of the LTReID on the training procedure, we show the curves of softmax loss, triplet loss, IVD loss, MVD loss and mAP in Fig. 8. From the results, it can be observed that the softmax and triplet losses of the baseline model decline quickly to a relatively low level (almost zero), implying that the early training saturation is serious. This is because the number of easy samples is much larger than hard samples. Since the baseline model is optimized with gradient-based methods such as SGD and Adam, the training saturation may prematurely stops contributing gradients to back-propagation due to negligible gradients. As the number of easy samples increases and the number of hard samples decreases, the baseline model has few chances to move around and is more likely to converge at a local minima. Therefore, the baseline model easily suffers from over-fitting and requires extra hard samples to recover. In short, the hard-imbalance between easy and hard samples causes the early training saturation and introduces short-lived gradient propagation which is not enough to help models converge at a global minima or a better local minima. In contrast, the softmax and triplet losses of the LTReID model decline slowly and are much larger than the baseline model, verifying that the early saturation behavior is significantly avoided. From the results in Fig. 8, it can be observed that the LTReID outperforms the baseline with significant margins. Note that, after 2,500 iterations, the LTReID model achieves better mAP scores on Market1501 but higher loss values. This demonstrates that feature generation gives the LTReID chances to traverse more portions of parameter space for optimal solution. In the later period of training, the LTReID will spend more efforts to explore this region and converge at a better local minima with a finite number of steps.

V. Conclusion

In this work, we propose a new LTReID framework to simultaneously solve the class-imbalance and hard-imbalance problems for the long-tailed ReID task. Specifically, the LTReID generates fake features with multiple independent components to relieve the class-imbalance between head and tail classes. Moreover, the LTReID encourages hard feature generation with adversarial learning to mitigate the hard-imbalance between easy and hard samples. Extensive empirical analysis demonstrates that the proposed LTReID contributes to learning both discriminative and balanced features on long-tailed training data. For the future work, we will extend this work to the fields of attribute recognition, face recognition and vehicle re-identification, where the long-tailed problem is prevalent.

REFERENCES

[1] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1412–1424, 2018.

[2] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[3] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[4] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.

[5] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2891–2902, 2019.

[6] F. Yang, Z. Zhong, Z. Luo, S. Lian, and S. Li, "Leveraging virtual and real person for unsupervised person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2444–2453, 2019.

[7] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3. IEEE, 2021, pp. 902–917.

[8] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8295–8302.

[9] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 274–282.

[10] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8514–8522.

[11] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," in *Pattern Recognition*, vol. 94. Elsevier, 2019, pp. 53–61.

[12] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua, "Concentrated local part discovery with fine-grained part representation for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1605–1618, 2019.

[13] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.

[14] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *European Conference on Computer Vision*, 2020, pp. 346–363.

[15] L. He and W. Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *European Conference on Computer Vision*, 2020, pp. 357–373.

[16] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3064–3074, 2020.

[17] H. Luo, W. Jiang, X. Fan, and C. Zhang, "Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2905–2913, 2020.

[18] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2015.

[19] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*. Springer, 2008, pp. 262–275.

[20] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2360–2367.

[21] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7. IEEE, 2012, pp. 1622–1634.

[22] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[23] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5. IEEE, 2014, pp. 1095–1108.

[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptors with application to person re-identification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9. IEEE, 2020, pp. 2179–2194.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[27] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[28] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[29] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images." Citeseer, 2009.

[30] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11. IEEE, 2008, pp. 1958–1970.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," in *International Journal of Computer Vision*, vol. 115, no. 3. Springer, 2015, pp. 211–252.

[32] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2970–2979.

[33] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.

[34] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, 2020, pp. 9719–9728.

[35] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, vol. 11. Citeseer, 2003, pp. 1–8.

[36] M. Peng, Q. Zhang, X. Xing, T. Gui, X. Huang, Y.-G. Jiang, K. Ding, and Z. Chen, "Trainable undersampling for class-imbalance learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4707–4714.

[37] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5409–5418.

[38] Y. Duan, J. Lu, and J. Zhou, "Uniformface: Learning deep equidistributed representation for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3415–3424.

[39] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision*. Springer, 2014, pp. 536–551.

[40] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[41] P. Wang, Z. Zhao, F. Su, Y. Zhao, H. Wang, L. Yang, and Y. Li, "Deep multi-patch matching network for visible thermal person re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1474–1488, 2021.

[42] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1249–1258.

[43] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5735–5744.

[44] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1389–1398.

[45] Y. Zhang, Q. Zhong, L. Ma, D. Xie, and S. Pu, "Learning incremental triplet margin for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 01, Jul. 2019, pp. 9243–9250.

[46] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 188–204.

[47] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 657–13 665.

[48] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 403–412.

[49] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 650–667.

[50] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems*, 2018, pp. 1230–1241.

[51] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.

[52] C. Eom and B. Ham, "Learning disentangled representation for robust person re-identification," in *Advances in Neural Information Processing Systems*, 2019, pp. 5297–5308.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[54] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[55] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[56] J. Liu, J. Zhang, W. Li, C. Zhang, and Y. Sun, "Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory," in *arXiv preprint arXiv:2008.09809*, 2020.

[57] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

[58] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[59] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[60] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8577–8584.

[61] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *Advances in Neural Information Processing Systems*, 2019, pp. 1919–1930.

[62] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *European Conference on Computer Vision*. Springer, 2020, pp. 162–178.

[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[64] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, 2019, pp. 1567–1578.

[65] D. Cao, X. Zhu, X. Huang, J. Guo, and Z. Lei, "Domain balancing: Face recognition on long-tailed domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5671–5679.

[66] T. Dutta, A. Singh, and S. Biswas, "Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir," in *European Conference on Computer Vision*. Springer, 2020, pp. 349–364.

[67] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li, "Balanced meta-softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 4175–4186.

[68] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *International Conference on Machine Learning*, 2016, pp. 507–516.

[69] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 212–220.

[70] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*. Springer, 2016, pp. 579–596.

[71] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5704–5713.

[72] Q. Yu and W. Lam, "Data augmentation based on adversarial autoencoder handling imbalance for learning to rank," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 411–418.

[73] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 694–710.

[74] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[75] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[76] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," in *Journal of Machine Learning Research*, vol. 9, no. Nov, 2008, pp. 2579–2605.

[77] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," in *arXiv preprint arXiv:1703.07737*, 2017.

[78] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[80] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[81] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.

[82] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global–local-alignment descriptor for scalable person re-identification," in *IEEE Transactions on Multimedia*, vol. 21, no. 4. IEEE, 2018, pp. 986–999.

[83] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[84] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 365–381.

[85] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.

[86] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 667–676.

[87] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 371–381.

[88] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3702–3712.

[89] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3691–3701.

[90] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 173–11 180.

[91] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, X. Guo, W.-S. Zheng, N. Sang, and X. Sun, "Do not disturb me: Person re-identification under the interference of other pedestrians," in *European Conference on Computer Vision*, 2020, pp. 647–663.

[92] Q. Zhou, B. Zhong, X. Liu, and R. Ji, "Attention-based neural architecture search for person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[93] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[94] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[95] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Transactions on Image Processing*, vol. 30, pp. 7663–7676, 2021.

[96] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

[97] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[98] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[99] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.