Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss

Cheng YAN

Guansong PANG
Singapore Management University, gspang@smu.edu.sg

Xiao BAI

Changhong LIU

Xin NING

See next page for additional authors

Citation

YAN, Cheng; PANG, Guansong; BAI, Xiao; LIU, Changhong; NING, Xin; and ZHOU, Jun. Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss. (2022). *IEEE Transactions on Multimedia*. 24, 1665-1677.
Available at: https://ink.library.smu.edu.sg/sis_research/7023

Author

Cheng YAN, Guansong PANG, Xiao BAI, Changhong LIU, Xin NING, and Jun ZHOU

Beyond Triplet Loss: Person Re-identification with Fine-grained Difference-aware Pairwise Loss

Cheng Yan*, Guansong Pang*, Xiao Bai, Jun Zhou, Lin Gu

***Abstract*—Person Re-IDentification (ReID) aims at re-identifying persons from different viewpoints across multiple cameras. Capturing the fine-grained appearance differences is often the key to accurate person ReID, because many identities can be differentiated only when looking into these fine-grained differences. However, most state-of-the-art person ReID approaches, typically driven by a triplet loss, fail to effectively learn the fine-grained features as they are focused more on differentiating large appearance differences. To address this issue, we introduce a novel pairwise loss function that enables ReID models to learn the fine-grained features by adaptively enforcing an exponential penalization on the images of small differences and a bounded penalization on the images of large differences. The proposed loss is generic and can be used as a plugin to replace the triplet loss to significantly enhance different types of state-of-the-art approaches. Experimental results on four benchmark datasets show that the proposed loss substantially outperforms a number of popular loss functions by large margins; and it also enables significantly improved data efficiency.**

***Index Terms*—Person Re-Identification, Fine-grained Difference, Representation Learning, Triplet Loss, Pairwise Loss**

I. Introduction

Person re-identification (ReID), aiming at re-identifying people from viewpoints across multiple cameras, is a critical computer vision task due to its crucial applications in video surveillance, multi-camera tracking and forensic search. Although person ReID has attracted extensive research attentions in recent years, one largely unsolved challenge is how to effectively capture the fine-grained appearance differences of different persons. This problem is crucial to person ReID, because in real-world ReID applications images of different identities can often be differentiated only when looking into these fine-grained differences. This issue manifests itself in popular person ReID benchmarks such as CUHK03 [1], Market1501 [2] and DukeMTMC [3]. To provide a straightforward illustration, we explore and visualize the distribution of average pairwise distances on these datasets. The results are shown in Figure 1. It is clear that *inter-person distances*[1] (i.e., distance between an image pair of different persons) can be rather small due to fine-grained differences between these images, e.g., the demonstrated CUHK03 anchor image and the negative sample at the right bottom in the first row in Figure 1 have only small differences in the bags and glasses the two persons carry. On the other hand, *intra-person distances* (i.e., distance between an image pair of the same person) can be large due to the fine-grained differences, e.g., the background object in the positive sample at the left bottom in the first row of Figure 1. Consequently, the identified persons may contain a large number of false positive errors. Similar results can also be observed in the Market1501 [2] and DukeMTMCC [3]. Therefore, the ability to capture those fine-grained appearance differences is the key to accurate person ReID.

Inspired by the tremendous success of deep learning, many methods [4], [5], [6] have been introduced to learn deep expressive representations for person ReID and achieved state-of-the-art performance. Typically, most of these methods [7], [8], [4], [9], [5], [10], [11], [12], [6], [13], [14], [15], [16], [17], [18], [19], [20] employ a triplet loss [7], [5], [13] or its combination of a classification loss [10], [11], [12] as the driving force to extract relevant features. Under this generic framework, several approaches have been developed to learn semantically-rich and/or local features, such as the global feature-based approach [14], [15], data augmentation-based approach [6], [13] and striping approach [21], [10].

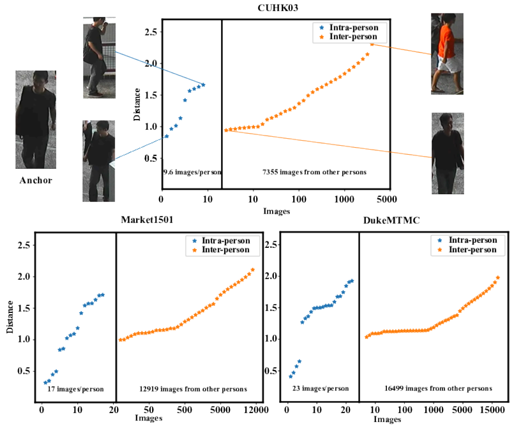


Fig. 1. Distribution of average distances between an anchor image and other images from the same person or different persons. Many image pairs have small inter-person distances in popular ReID benchmarks (see Table I in Section V for detailed statistics). The distances are calculated using features extracted from ResNet50.

However, the triplet loss, which enforces that inter-person distances are larger than intra-person distances by a predefined margin, is less effective in learning the fine-grained differences

*Cheng Yan and Guansong Pang contributed equally to this work. Cheng Yan's contribution was made when visiting the University of Adelaide

Xiao Bai is the corresponding author.

Cheng Yan and Xiao Bai are with School of Computer Science and Engineering, Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University

Guansong Pang is with School of Computer Science, University of Adelaide

Jun Zhou is with School of Information and Communication Technology, Griffith University

Lin Gu is with National Institute of Informatics, University of Tokyo.

[1] Images of each person are normally treated as samples from an individual class; so class and person/identity are used interchangeable in this study.

due to two main reasons: (i) as shown in Figure 2, the triplet loss function is dominated by infinitely increasing penalization on large differences between images of the same identity; (ii) it does not enforce sufficient penalization on the images of small differences. For example, triple loss enforces no penalization on the small intra-person differences and imposes a linearly increased penalization on the small inter-person differences. As a result, the triplet loss can only capture the high-level similarities and differences, and thus, it is ineffective in scenes where the fine-grained differences are the key to person ReID.

To address the aforementioned issues, we propose a novel fine-grained difference-aware (FIDI) loss for person ReID. This fine-grained difference-aware property refers to the capability of our loss in adaptively penalizing small inter-person or intra-person appearance differences. Particularly, the FIDI loss enforces an exponentially large penalization on images of those fine-grained differences while at the same time imposing a bounded penalization on their counterparts, i.e., images of large inter-person or intra-person differences. The exponential penalization drives the model to be sensitive to small differences, while the bounded penalization effectively reduces the bias towards large differences. The resulting models can balance expressive features learned from both large and small differences. Additionally, due to the fine-grained difference-aware property, our loss can also leverage the training data more efficiently than the triplet loss.

A number of studies [9], [22], [12] have dedicated to exploring loss functions other than the triplet loss function for more effective and/or efficient person ReID. Contrastive loss [23] is a well-known pairwise loss that learns features for face recognition or re-identification. However, it has similar weaknesses as the triplet loss. Additionally, the single predefined hard margin in these losses also makes it hard to adaptively penalize distance distributions within different person identities.Quadruplet loss [9] equips a quadruplet deep network with quadruplet inputs to replace the triplet loss. However, it is limited to specific network structures and is hard to be extended. Batch-hard triplet loss [22], [12] is another widely used person ReID loss that optimizes the margin between the most dissimilar intra-person distance and the most similar inter-person distance in each batch. The batch-hard operation is also explored to improve the contrastive loss for the ReID task [10]. The recently proposed circle Loss [24] combines the triplet loss with a softmax cross-entropy loss and re-weights each similarity to highlight the less-optimized similarity scores. However, although its batch-wise loss helps regularize the feature learning, it is built upon the triplet loss and thus exhibits similar behaviors in handling the fine-grained feature issues.

In summary, this paper makes the following four main contributions.

- We reveal that the widely-used triplet loss function, arguably currently the most popular ReID loss, has inherent difficulties in handling fine-grained appearance differences. This loss is ineffective in challenging ReID cases where different identities can be only distinguished by the fine-grained differences.
- We introduce a novel pairwise relationship-based loss function, termed fine-grained difference-aware (FIDI) loss. This FIDI loss enforces exponentially large penalization on small appearance differences while at the same time imposing bounded penalization on large differences. As a result, the FIDI-enabled models can effectively learn expressive features from both large and small appearance differences.
- The fine-grained difference-aware property also empowers the FIDI loss to harness the image samples more effectively and is thus substantially more data-efficient than the triplet loss.
- We demonstrate that the FIDI loss can be used as a plugin to replace the triplet loss and work effectively in different types of state-of-the-art approaches.

Experimental results on four benchmark datasets show that the FIDI loss substantially improves the triplet loss by a large margin, *e.g.*, typically 10%-20% improvement in effectiveness. We also show the FIDI loss based models can also largely outperform state-of-the-art vehicle ReID models.

The rest of our paper is organized as follows: In Section II, we review the related works for person ReID. Then we provide corresponding research background and discuss relevant loss functions for person ReID in Section III. Section IV introduces the proposed FIDI loss function. Experimental results, visualization and ablation studies are presented in Section V. Finally, the conclusions are given in Section VI.

II. Related Work

Many studies [8], [25] learn feature representations for person ReID by fine-tuning convolutional networks with a classification loss. Different approaches have been introduced to further improve the performance, including data augmentation, striping, and global feature approaches. In this section, we review three types of person ReID approaches.

A. Data Augmentation-based Approach

Data augmentation is an effective way to improve the feature learning capacity for CNN. There are generative adversarial networks (GANs) [6], [13], pose estimation [14], [15], random erasing [26] in this category.

GAN based approaches use GAN to generate more data for training. Mask or pose guided frameworks obtain the semantic information from pose estimation or segmentation models. These methods use other networks to generate image to increase the number of input images or improve the mask of input for augmentation. However, the benefit comes from the help of other networks with extra semantic information. By contrast, random erasing randomly selects a rectangle region and assigns random values either on image [26] or CNN feature maps [10]. Among these data augmentation-based methods, random erasing is arguably the simplest yet highly effective method without extra computation cost.

B. Striping-based Approach

Striping based methods aim at enforcing the learner to pay more attention to different parts of the identities by combining striping local features. Part based networks are widely adopted in these methods [25], [21], [10] to separate feature maps into several parts. They build multi-branch neural networks to learn local features in each of the predefined parts of the identities with one-branch network dedicated to one part, and then they concatenate these features to perform ReID during inference.

PCB [21] is the first part-based deep learning methods for person re-identification. It replaces the original global pooling layer with a spatial conventional pooling layer to separate the last convolution into several pieces of column vectors for independent pooling, in which each part refers to a body part of person. These feature then are concatenated for final feature learning. To further improve the accuracy, both global feature and part-based local feature are learned and used [27], [10]. The added features often result in accuracy improvement.

Though these striping methods are often the best performers on different benchmark datasets, they often involve more network parameters and expensive computation than single-branch network-based methods.

C. Global Feature-based Approach

Global feature-based approach focuses on a single network structure as the backbone to learn global identity features. These methods work on the sampling process [28], [29], loss design [22] or learning process [12]. Among them, the loss function is very crucial for feature learning and most relevant to our work.

The combination of classification loss and ranking loss such as triplet loss is one of the most widely-used loss functions for person ReID [7], [8], [4], [9], [5]. The triplet loss often works better than contrastive loss, since the triplet inputs provides a better guarantee of the distance margin than the pairwise contrastive loss. However, the triplet constraint is loose in the sense that it ignores the triplets when the predefined margin is met. The triplet loss is also cumbersome as the triplet sample space is often excessively large for large-scale data. A few studies attempt to address these issues for person ReID. One such example is a quadruplet loss with quadruplet network [9], but it is limited to specific network structures. Circle loss [24] is another closely related work that re-weights each similarity to highlight the less-optimized similarity scores, but the weighting factors are defined in a self-paced manner and need more calculation. Other methods [22], [12] avoid the explicit generation of hard triplet samples. Instead they work with batch-wise hard triplet loss, which optimizes the margin between the most dissimilar intra-person distance and the most similar inter-person distance in each batch. This enhanced triplet loss becomes more sensitive to small appearance differences than the basic triplet loss. However, its inherent penalization mechanism does not change.

III. Research Background

This section introduces person re-identification problem and a widely-used state-of-the-art frameworks to illustrate how our proposed loss could be plugged in.

A. Problem Formulation

In a person ReID system, let $\mathcal{X} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be a set of N training samples, where $\mathbf{x}_i$ is an image sample and y_i is its identity/class label. The person ReID algorithm learns a mapping function $\phi : \mathcal{X} \mapsto \mathcal{F}$ which projects the original data points $\mathcal{X}$ to a new feature space $\mathcal{F}$. This space $\mathcal{F}$ should shrink the intra-person distance while push the inter-person distance as large as possible. Given a query image $\mathbf{q}$ and ϕ, the ReID algorithm first computes this distance between $\phi(\mathbf{q})$ and every image $\phi(\mathbf{x})$ from a gallery image set G, and then returns the images that have the smallest distance. It should be noted that, for the sake of real-world applications, the gallery image set and the training image set have no overlapping, *i.e.*, the query person does not appear in the training set. Therefore, is is also regarded as a zero-shot problem. This largely distinguishes person ReID from general image retrieval tasks.

B. Triplet Loss-based Approach

The triplet loss is a widely-used loss function which takes a collection of triplet samples to learn feature representations space where the inter-class distances are greater than intra-class distances by at least a predefined margin m. A triplet is composed of three samples $\mathbf{x}_a$, $\mathbf{x}_p$ and $\mathbf{x}_n$, where $\mathbf{x}_a$ is an anchor sample. $\mathbf{x}_p$ is a positive sample that comes from the same person as $\mathbf{x}_a$, while $\mathbf{x}_n$ is a negative sample taken from an identity different from that of the anchor. The generic triplet loss (TL) is given as follows:

$$L_{tl} = [d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m]_+, \tag{1}$$

where $\mathbf{z} = \phi(\mathbf{x})$ denotes the learnt feature representation of $\mathbf{x}$. $d(\cdot, \cdot)$ is the distance of two samples. m is a predefined margin and $[\cdot]_+$ represents $\max(\cdot, 0)$. Contrastive loss can be regarded as a special case of triplet loss where $d(\mathbf{z}_a, \mathbf{z}_n) + m$ is 0 for similar pairs and $d(\mathbf{z}_a, \mathbf{z}_p)$ is 0 for dissimilar pairs. Convolutional networks are often employed to instantiate the ϕ function. The triplet loss is the key ingredient here, but Eqn.(1) requires the high-quality triplets as input. An advanced triplet loss, termed batch triplet loss (BTL) that is widely-used in person ReID, incorporates hard triplet mining into the loss calculation in each batch [7], [4], [9], [5]. BTL is defined as follows:

$$L_{btl} = [\max_{p=1\ldots B_p} d(\mathbf{z}_a, \mathbf{z}_p) - \min_{n=1\ldots B_n} d(\mathbf{z}_a, \mathbf{z}_n) + m]_+, \tag{2}$$

where $\max_{p=1\ldots B_p} d(\mathbf{z}_a, \mathbf{z}_p)$ represents the maximum distance between anchor and all B_p positive samples in a batch. $\min_{n=1\ldots B_n} d(\mathbf{z}_a, \mathbf{z}_n)$ represents the minimum distance between anchor and all B_p negative samples in the batch.

To complement the triplet-based local features, a classification loss is used in recent methods [27], [10], [12] to work together with the triplet loss for a global constraint in the optimization. This helps learn class-level global features effectively. The classification loss is defined as:

$$L_{cla} = \sum_{i=1}^{N} \mathbb{E}(\mathbf{z}_i^\intercal \mathbf{W}, \mathbf{y}_i), \tag{3}$$

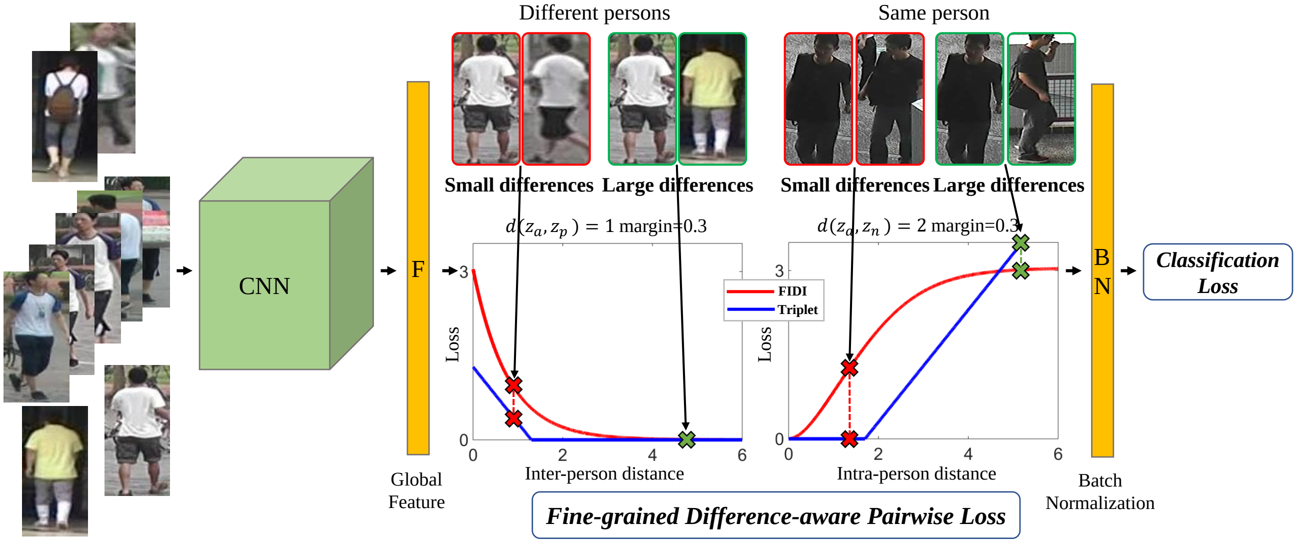


Fig. 2. An overview of the fine-grained difference-aware pairwise loss-based framework. It consists of a deep CNN-based network backbone, our proposed FIDI loss and a classification loss. This framework is exactly the same as the widely-used triplet loss-based framework except that the triplet loss is replaced with our FIDI loss. The network backbone can be various CNN architectures. Unlike the triplet loss that neglects small appearance differences due to the potential dominance of unbounded penalization on images of large intra-person differences, our FIDI loss can effectively capture the fine-grained intra-person/inter-person appearance differences, *e.g.*, image pairs having small appearance difference as in the red boxes above. We achieve this by enforcing exponentially large penalization on images of small differences and bounded penalization on images of large differences.

where $\mathbb{E}(\cdot)$ is the cross entropy loss. $\mathbf{W}$ is the weight matrix to map $\mathbf{z}_i$ to classification labels $\mathbf{x}_i$ encoded as one-hot vector $\mathbf{y}_i$ in the output layer. This classification loss is added in the output classification layer. A batch normalization layer is normally employed between the triplet loss-enabled feature layer and the output layer to speed up training and stabilize the performance.

This framework works effectively in different benchmark datasets. Recent advances incorporate data augmentation or striping strategies [27], [10] to achieve new state-of-the-art performance. However, the triplet loss, either L_{tl} or L_{btl}, fails to learn expressive features from fine-grained differences. This is because: (i) the triplet loss is not sensitive to small differences, i.e., it enforces no penalization on small intra-person differences or small penalization on small inter-person differences; (ii) the loss grows linear infinitely w.r.t. the increasing intra-person distances and has no upper bound. As a result, the optimization may be dominated by large intra-person differences.

IV. Fine-grained Difference-aware (FIDI) Loss

This section introduces our fine-grained difference-aware (FIDI) loss to address the bottleneck issue with the triplet loss.

A. The Proposed Framework

Our proposed framework aims to leverage the capability of the FIDI loss in capturing fine-grained differences to learn well discriminative and generalized features for the person ReID task. Specifically, as shown in Figure 2, our framework is composed of three modules: deep convolutional network-based feature mapping, the FIDI loss and the classification loss. We use exactly the same framework as the triplet loss approach except that the triplet loss is replaced with our FIDI loss. Note that we use this setting to facilitate a straightforward comparison with triplet loss-based approaches in our empirical studies. As discussed in Section V-D3, The FIDI loss could also improve other frameworks.

The procedure of our framework is as follows. It first uses a convolutional neural network to map image data into a low-dimensional space. Compared to quadruplet loss [9], here this backbone network is not limited to any specific deep convolutional network structures. Then the proposed FIDI loss enforces a pairwise constraint to the projected features by applying exponentially increasing penalization to small differences and bounded loss to large differences. This enables the learner to adaptively capture the fine-grained differences while enforce a desired margin between the feature representations of different identities. Finally, we use a batch normalization layer and a fully connected layer without bias as the classifier, which is optimized using the cross entropy loss in Eqn.(3).

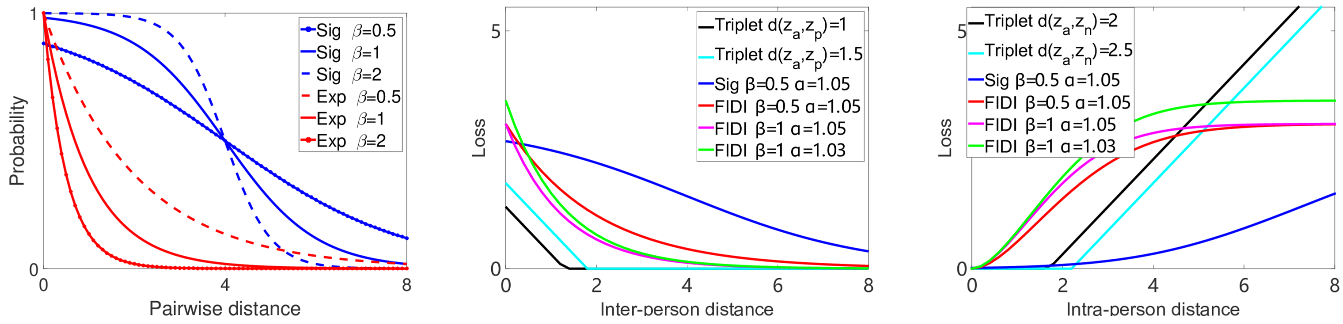
Particularly, the FIDI loss is built upon relative entropy [30], a measure of the distance between two distributions. Let $\mathcal{K}$ be a known distribution of training image pairs, *i.e.*, the ground truth identity labels, and $\mathcal{U}$ be an unknown distribution we aim to learn, then the FIDI loss is defined as follows:

$$L_{fidi} = D(\mathcal{U}||\mathcal{K}) + D(\mathcal{K}||\mathcal{U}), \tag{4}$$

where

$$D(\mathcal{U}||\mathcal{K}) = \sum_{p_{ij} \in \mathcal{P}} u_{p_{ij}} \log \frac{\alpha u_{p_{ij}}}{(\alpha - 1)u_{p_{ij}} + k_{p_{ij}}}, \tag{5}$$

where $p_{ij} = \{\mathbf{x}_i, \mathbf{x}_j\}$ is a pair of image samples and $\mathcal{P}$ is a collection of image pairs; $k_{p_{ij}} \in \mathcal{K}$ and $k_{p_{ij}} = 1$ if the image pair $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same identity, and $k_{p_{ij}} = 0$ otherwise; $u_{p_{ij}}$ is taken from an unknown distribution $\mathcal{U}$, which is the distribution of feature level relationship of image



(a) Probability (b) Loss $(k=0)$ (c) Loss $(k=1)$

Fig. 3. (a) Exponential vs. sigmoid distance-to-probability functions, (b) Loss w.r.t. inter-person distance and (c) Loss w.r.t. intra-person distance. One desired property of exponential distribution based distance-to-probability function is that its probability is exponentially sensitive to changes within a small distance. As shown in (b) and (c), our loss function exponentially punishes the similar/dissimilar pairs that have small distance whle imposing bounded loss to large distances.

pairs in $\mathcal{P}$; and $\alpha > 1$ is a parameter to control the scale of L_{fidi}. Since $\mathcal{K}$ is the supervised information and is known *a priori*, our target is to learn $u_{p_{ij}} \in \mathcal{U}$ such that the distribution $\mathcal{U}$ is close to $\mathcal{K}$ as much as possible.

The original relative entropy is one of the most popular and effective losses used in different learning tasks. However, it could not effectively reflect the true distance between distributions in our task due to the asymmetric and the lack of fine-grained difference-aware characteristic. Our L_{fidi} enhances it to achieve the following two main advantages:

- Our loss is a symmetric metric with a desired inter-class margin.
- Our loss enforces fine-grained difference-aware penalization on small differences and bounded loss on large differences

B. Exponential Loss on Images of Fine-grained Differences

One key ingredient in Eqn.(4) is the distance-to-probability function η that maps the distance in the representation space, $d(\mathbf{z}_i, \mathbf{z}_j)$, to the probability distribution $\mathcal{U}$, *i.e.*, $u_{p_{ij}} = \eta(d(\mathbf{z}_i, \mathbf{z}_j))$. In FIDI loss, we introduce an exponential distribution-based distance-to-probability function η to effectively penalize hard samples. Particularly, η is defined as follows:

$$u_{p_{ij}} = e^{-\beta d(\mathbf{z}_i, \mathbf{z}_j)}, \tag{6}$$

where β is a parameter to control the scale of the probability distribution. We have $u_{p_{ij}} \rightarrow 0$ with increasing pairwise distance, and $u_{p_{ij}} \rightarrow 1$ in the opposite.

We use the exponential distribution-based η because it is more sensitive and imposes more meaningful penalization on small differences compared to the commonly-used sigmoid function $\frac{1}{1+e^{-d}}$ or its advanced variant $\frac{1}{1+e^{-\beta d}}$ [4], [31], [29], where d denotes the pairwise distance and the parameter β controls the scale of the distribution shape.

Specifically, as shown in Figure 3(a), the exponential distribution shape is significantly more sensitive to the distance than the sigmoid distribution shape, especially when the pairwise distance is small. As a result, as shown in Figure 3(b-c), the exponential distribution based η results in exponentially varying relative entropy loss w.r.t. both the intra- and inter-person distances, whereas the sigmoid distribution-based loss applies rather conservative penalization in such cases.

One main benefit brought by the exponentially sensitive penalization is the capability in learning the fine-grained difference of the image pairs. Specifically, as shown in Figure 3(b), for image pairs that come from different persons but with small distances, the FIDI loss applies penalization inversely exponential to the distance and applies nearly zero loss to the pairs that have large inter-person distance; by contrast, the triplet loss may enforce no penalization on image pairs which have very small inter-person distance. In a similar sense, as shown in Figure 3(c), for image pairs that come from the same person, no penalization is enforced by the triplet loss on the image pairs that have small intra-person distance; by contrast, the FIDI loss also applies exponential penalization to such cases.

The resulting FIDI loss-based model effectively learns fine-grained feature representations that are significantly improved over the triplet loss. The fine-grained difference-aware ability also enables the FIDI loss-based model to leverage the labeled data substantially more efficiently than its counterpart, resulting in more data-efficient learning.

C. Bounded Loss on Images of Large Differences

Unlike triplet loss that has an infinitely linearly increasing penalization w.r.t. images of large appearance differences, the FIDI loss has a bounded loss on the large differences, which effectively prevents the dominance of images of large differences in the optimization. Specifically, the bounded loss of the FIDI loss can be provided as follows.

$$\begin{aligned} &\lim_{u \to 0} L_{fidi} = 0, \text{ when } k = 0; \\ &\lim_{u \to 0} L_{fidi} = \log \frac{\alpha}{(\alpha - 1)}, \text{ when } k = 1. \end{aligned} \tag{7}$$

This states that for image pairs from different identities, i.e., $k = 0$, we have a lower loss bound of zero with u approaching to zero. Recall that the pairwise distance increases as $u \rightarrow 0$. In other words, similar to the triplet loss, the FIDI loss does not penalize the image pairs if they are from different identities with a large distance in the new space. On the other hand, for image pairs from the same identity, while the triplet loss has an infinitely increasing loss, the FIDI loss imposes an upper loss bound of $\log \frac{\alpha}{(\alpha-1)}$ w.r.t. increasing intra-person distance. This upper bound is controlled by the hyperparameter α and can be easily tuned during training.

As shown in Figure 3(c), the punishment of triplet loss can be very large, given image pairs with very large intra-person distances. This hinders the triplet loss to learn the fine-grained differences from image pairs that have small intra-person distances. By contrast, the FIDI loss treats these samples equally and enforces a bounded penalization, preventing the domination of the large appearance differences over the counterpart small differences.

D. Symmetric Metric with a Desired Margin

Different from the original relative entropy that is asymmetric, our loss in Eqn.(4) is symmetric, as it is easy to see that we get the same results when switching $u_{p_{ij}}$ and $k_{p_{ij}}$. This characteristic eases the optimization of the feature learning and also helps learn more meaningful features.

Although the FIDI loss does not explicitly define a margin between intra- and inter-person image pairs as in the triplet loss, the FIDI loss can still achieve some implicit margins. This is because Eqn. (4) enforces the substantially small intra-person distances while at the same time encourages large inter-person distances, resulting in some implicit margins between intra- and inter-person image pairs. However, the margins are not directly predefined as in the triplet loss, but they are controlled by the parameter β in Eqn. (6).

V. Experiments

A. Datasets

We evaluate the performance on four widely used person ReID datasets, including Market1501 [2], DukeMTMC-ReID [3], CUHK03-D and CUHK03-L [1], and two vehicle datasets, VeRi-776 [32] and VehicleID [33].

Market1501 is a large person ReID dataset containing 12,936 images from 751 identities in the training data, and 3,368 query images and 19,732 gallery images from 750 identities in the testing data. These images were captured from 6 different camera viewpoints with manual bounding boxes. There are about 17 images for each identity.

DukeMTMC-ReID is a subset of DukeMTMC [34] for person ReID. The images are cropped by hand-drawn bounding boxes. The data was taken from 8 cameras of 1,404 identities with respective 16,522, 2,228 and 17,661 images in the training, query and gallery sets.

CUHK03-D and **CUHK03-L** contain the same image set with 14,096 images from 1,467 identities captured from two cameras in CUHK campus, but their identity-bounding box were created by different methods. CUHK03-D used pedestrian detectors to create the bounding boxes while that of CUHK03-L was manually labeled. The pedestrian detector-based method is more challenging than the manually labeled one since the former is less accurate.

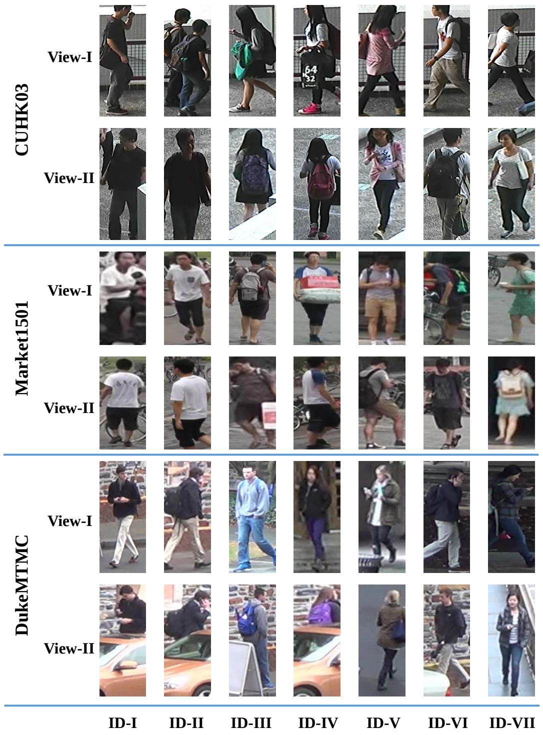


Fig. 4. Images from three person reid datasets. We give two images from a same person with different views. There are many hard/easy examples from different/same person in these datasets. For example, the images of ID-I and ID-II in View-II look very similar. However, the images of same person of ID-IV and ID-VII in different views look very different. These datasets also contain many images with occlusion.

VeRi-776 is a vehicle dataset in which all the images were captured in natural and unconstrained traffic environment. It contains about 50,000 images of 776 vehicles across 20 surveillance cameras with different orientations. This dataset is widely used in vehicles re-identification tasks because each image is captured from 2 to 18 viewpoints with different illuminations and resolutions. These images are also labeled with bounding boxes over the whole vehicle body.

VehicleID is a large-scale vehicle dataset that contains 221,763 images with 26,267 vehicles. All the images were captured from multiple surveillance cameras with no overlapping. There are three test subsets with different sizes and we use the largest test set which contains 20,038 images of 2,400 vehicles.

Note that the person/vehicle identities in the training and testing sets have no overlapping in all the used datasets. An image example is given in Figure. 4, in which we give two images from the same person with different views. There are many hard/easy examples from different/same person in these datasets. For example, the images of ID-I and ID-II in View-II look very similar and the images of same person of ID-IV and ID-VII in different views look very different. These datasets also contain many images with occlusion.

B. Evaluation Protocol

Following the standard protocol in [31], [29], [21], [35], [36], we use Cumulated Matching Characteristics (CMC) and mean average precision (mAP) to evaluate the performance on all datasets. We report the cumulated matching accuracy at rank 1 (R-1 for short) and the mAP value of the retrieval performance. Specifically, for all queries, we compute

$$R1 = \sum_{q=1}^{Q} r_1/Q, \tag{8}$$

where Q is the number of queries and r_1 is defined as

$$r1 = \begin{cases} 1, & \text{the first top-ranked sample is the query identity} \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

The mean average precision (mAP) is defined as

$$mAP = \sum_{q=1}^{Q} AveP(q)/Q, \tag{10}$$

where $AveP(q)$ is the average precision (AP) for a given query q.

Note that all the reported results here do not involve re-ranking which may be used as an extra step to further improve the accuracy.

C. Understanding the Resulting Feature Representations

We aim to understand the effectiveness of feature representations by looking into the fidelity and the saliency map of learned features.

1) Fidelity of Feature Representations: The feature representations fidelity measures how faithful the obtained feature represents the expectation, *i.e.* intra-person distances should be larger than inter-person distances. To efficiently evaluate this type of fidelity, we consider the number of erroneous cases where (i) anchor images have smaller inter-person distances than the their maximal intra-person distances, termed Error-I, or (ii) anchor images have larger intra-person distances than their minimal inter-person distances, termed Error-II. We count these two types of erroneous cases using feature representations of four different methods, including pre-trained features extracted from a pre-trained ResNet50[2] (PF) and features obtained by fine-tuning ResNet50 using respectively batch-hard constrastive loss (BCL), batch-hard triplet loss (BTL) and our proposed loss (FIDI). The statistics of erroneous cases on three person ReID benchmarks are reported in Table I.

It is clear from Table I that pre-trained features would result in a large number of erroneous cases, especially the Error-I cases. This indicates that most images of difference identities exhibit large similar appearance in both training and testing data, leading to small inter-person distances. The datasets also contain some Error-II cases that may be seen as outliers, because intra-person distances is rarely larger than minimal inter-person distances. To address these issues, models should be able to effectively learn the small appearance differences while prevent the impact of the outlying cases. After fine-tuning the models using either BCL, BTL or FIDI, the number of erroneous cases is significantly reduced in both training and testing data. In training data, BCL and BTL perform very well in enforcing intra-person distances to be smaller than inter-person distances, often achieving smaller error rates than FIDI. However, they perform significantly less effective than FIDI in the testing data, especially on the Error-I measure. This may indicate that both BCL and BTL overfit the training data rather than capturing the fine-grained appearance differences to distinguish the inter-person images. By contrast, with exponentially large penalization, FIDI enforces the models to learn any possible fine-grained appearance differences in the training data. Since the fine-grained differences are typically very difficult to learn, for some cases, even for humans, FIDI does not perform as well as BCL and BTL in the training data. However, its capability of discriminating the fine-grained differences pays off in the testing data.

2) Attention Maps: We further examine the resulting attention maps of our loss and the competing loss functions. We focus on comparing our FIDI loss to the BTL loss, because BTL is generally more effective and is much more widely-used than BCL in person ReID. Specifically, these two losses are plugged into one of the best ReID models, Baseline [12]. The attention maps are then obtained by applying the Grad-CAM visualization method [37] with Baseline to create pix-wise gradient visualizations. The attention maps on the last output feature maps are shown on Figure 5. The BTL-enabled Baseline highlights single discriminative parts only, which may correspond to the parts that have large appearance differences to other images. In contrast, our FIDI loss-enabled Baseline can effectively attend to diverse large and small discriminative parts in different cases, *e.g.*, shoes and heads in identity images taken different angles, different accessories and occluded identities. For example, in the 1st row in Figure 5, despite different angles and identities, our method can consistently pay attention to both small (shoes and heads) and large (the main body dress) discriminate parts, while the competing method focuses on a small discriminative region of the main body only. This demonstrates that the BTL loss-based models can be dominated and swayed by large appearance differences. Therefore, their attention is normally on single highly discriminative parts. In contrast, our loss can effectively drive the ReID models to pay attention on different body parts by enforcing the importance of distinguishing fine-grained differences.

3) Summary of Comparison: Overall, by enforcing exponentially large penalization on images of small appearance differences and bounded penalization on images of large differences, our FIDI pairwise loss brings in two major benefits compared to existing widely-used pairwise and triplet losses. First, the FIDI-enabled models can effectively capture fined-grained appearance differences, where the competing methods fail. This significantly improves the feature representations as demonstrated by significantly small errors in testing data in Table I. Second, as illustrated in Figure 5, our loss effectively pushes the ReID models to attend to diverse discriminative parts since fine-grained differences may appear in different body parts. This is important for distinguishing different

[2] https://github.com/kaiminghe/deep-residual-networks

TABLE I
Average erroneous distance cases over all images of each dataset in four feature spaces. Error-I refers to the average number of anchor images which have smaller inter-person distances than their maximal intra-person distances, while Error-II is the average number of anchor images which have larger intra-person distances than their minimal inter-person distances. PF refers to Pre-trained Features extracted from a pre-trained ResNet50. BCL, BTL and FIDI are feature spaces resulted by fine-tuning ResNet50 using respectively batch-hard constrastive loss, batch-hard triplet loss and our proposed loss. The average number of images per identity is 9.6 in CUHK03, 17 in Market1501 and 23 in DukeMTMC. The best results are boldfaced in each group.

Data	**Method**	**CUHK03**		**Market1501**		**DukeMTMC**	
		9.6 images/ID		17 images/ID		23 images/ID	
		Error-I	Error-II	Error-I	Error-II	Error-I	Error-II
Training Data	PF	4316	7	9596	15	13677	19
	BCL	0.008	0.010	1.977	0.397	6.360	4.420
	BTL	**0.005**	**0.008**	**1.905**	0.295	**3.310**	3.655
	FIDI	0.252	0.023	2.072	**0.261**	3.721	**1.807**
Testing Data	PF	2973	7.617	11687	506.7	13258	24.15
	BCL	85.06	6.600	276.3	19.18	925.7	20.89
	BTL	95.11	6.611	261.5	18.36	910.6	20.61
	FIDI	**45.02**	**6.310**	**229.1**	**16.50**	**819.3**	**19.54**

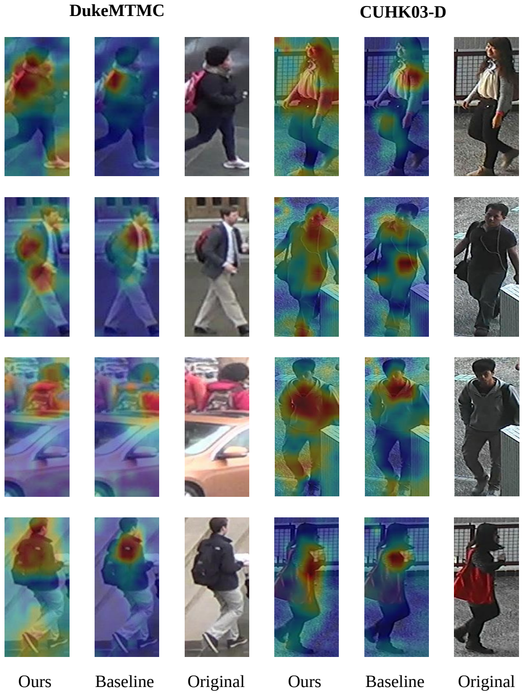


Fig. 5. Visualization of attention maps of our FIDI loss-enabled model (Ours) and the batch-hard triplet loss-enabled model (Baseline). Our method learns diverse important attention, but Baseline only focuses on small discriminative parts. The diverse attention maps from Ours span over the whole person rather than some local areas in Baseline.

identities with some similar appearances, *e.g.*, in dress, shoes and/or accessories. Models embodied with our loss would enjoy above two factors that are critical to accurate person ReID.

D. Enabling Different Type of Person ReID Models in Real-world Datasets

To have a comprehensive evaluation on real-world datasets, the FIDI loss is used to replace the batch-hard triplet loss in three types of recent state-of-the-art approaches, including data augmentation, global feature and striping approaches. Specifically, we choose the best performer(s) in each type of these approaches and them simply replace the triplet loss with our proposed FIDI loss, with all the other modules unchanged. The batch size and the number of identities in each batch are respectively set to 128 and 8 by default. The hyperparameters α and β in the FIDI loss are tuned via cross validation for each data set.

1) Enabling Data Augmentation Methods: This section compares our loss to several data augmentation-based methods, including GAN-based methods [38], [13] and segmentation-based masking methods [14], [39]. Note that, these methods employ other networks to generate images to obtain semantic information, which brings extra computational consumption. Baseline1 [12] without data augmentation, *i.e.*, random erasing, is the best performer. Therefore we plugged the FIDI loss into this method. Note that Baseline1 contains a center loss and we discard this loss in our Baseline1 model by replacing the triplet loss with our FIDI.

The comparison results are shown in the second row in Table II. Although Baseline1 has significantly outperformed all the other competing methods in this category, the FIDI loss-enabled Baseline1 can still consistently beat the original Baseline1 in both mAP and R-1 across all the four datasets. Particularly, the improvement is significantly larger on the challenging datasets than the relatively easy ones, *e.g.*, the improvement can be as large as 20.9%-21.3% in mAP and 22.1%-22.7% in R-1 on the two CUHK03 datasets whereas it is 2.7%-3.4% in mAP and 0.7%-1.3% in R-1 on Market1501/DukeMTMC. This demonstrates that the FIDI loss-enabled Baseline1 not only inherits the superior capability as in the original Baseline1 but also leverages the fine-grained difference-aware ability of the FIDI loss to learn extra discriminative information from the hard samples. This is especially true for the two challenging CUHK03 datasets in which we have much less images and the triplet loss-based models become overfitting (see Table I for detail).

2) Enabling Global Feature-based Methods: We then examine the plugging of the FIDI loss into the global feature-based methods. There are seven methods for comparison,

TABLE II
MAP AND R-1 OF DIFFERENT METHODS ON FOUR BENCHMARK DATASETS. BCL AND BTL RESPECTIVELY DENOTE BATCH CONTRASTIVE LOSS AND BATCH TRIPLET LOSS. THE BEST PERFORMANCE PER GROUP IS BOLDFACED.

Type	Methods	Market1501		DukeMTMC		CUHK03-D		CUHK03-L	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Data Augumentation	SPReID [14]	81.3	92.5	71.0	84.4	-	-	-	-
	Camstyle [13]	68.7	88.1	53.5	75.3	-	-	-	-
	PN-GAN [38]	72.6	89.4	53.2	73.6	-	-	-	-
	SVDNet [39]	62.1	82.3	56.8	76.7	37.3	41.5	37.8	40.9
	Baseline1 (BTL) [12]	82.3	93.5	71.0	84.9	52.5	54.2	55.3	56.3
	Baseline1 (FIDI)	**84.5**	**94.2**	**73.4**	**86.0**	**63.5**	**66.1**	**67.1**	**69.1**
Global Feature	TriNet [22]	69.1	84.9	-	-	-	-	-	-
	AWTL [40]	75.7	89.5	63.4	79.8	-	-	-	-
	AOS [28]	70.4	86.4	62.1	79.1	43.3	47.1	-	-
	GSRW [41]	82.5	92.7	66.4	80.7	-	-	-	-
	Mancs [29]	82.3	93.1	71.8	84.9	60.5	65.5	63.9	69.0
	BCL [42]	67.6	86.4	58.6	78.2	-	-	-	-
	CL [24]	84.9	94.2	-	-	-	-	-	-
	Baseline2 (BTL)[12]	85.9	**94.5**	76.4	86.4	58.2	60.5	60.2	62.1
	Baseline2 (BCL)	84.0	92.3	74.5	83.6	60.5	63.3	64.9	66.4
	Baseline2 (FIDI)	**86.8**	**94.5**	**77.5**	**88.1**	**69.1**	**72.1**	**73.2**	**75.0**
Striping	AlignedReID [43]	77.7	90.6	67.4	81.2	-	-	-	-
	MLFN [11]	70.4	86.4	62.1	79.1	47.8	52.8	49.2	54.7
	PCB [21]	77.4	92.3	65.3	81.9	53.2	59.7	-	-
	IANet [44]	83.1	94.4	73.4	87.1	-	-	-	-
	PL-Net [45]	69.3	88.2	-	-	-	-	-	-
	MCG [46]	78.3	92.6	69.4	84.7	-	-	55.3	61.7
	BDB [10]	84.3	94.2	72.1	86.8	69.3	72.8	71.7	73.6
	BDB (FIDI)	85.2	94.8	74.5	88.6	71.7	74.5	73.8	76.9
	MGN [27]	**86.9**	**95.7**	78.4	88.7	66.0	66.8	67.4	68.0
	MGN (FIDI)	**86.9**	95.4	**79.8**	**89.7**	**73.0**	**76.1**	**76.3**	**78.9**

including TriNet [22], AWTL [40], AOS [28], GSRW [41], Mancs [29], BCL [42] and Baseline2 [12]. These methods only employ simple single branch structure for training, which have less parameters to learn. All methods have only one pipeline with the basic ResNet50 as the backbone and use the feature representations obtained after global pooling. Note that, Baseline2 [12] is the Baseline1 with random erasing data augmentation. We also discard the centre loss and replace the triplet loss by our FIDI loss. We not only report the results of BCL [42] but also the results of Baseline2 (BCL), which is the Baseline2 [12] with BTL function being replaced by the BCL function from [42].

The results are given in the third row in Table II. It is clear that the FIDI loss-enabled Baseline2 consistently enhances the best performer in this group of methods, the original Baseline2, with significant improvement on the two CUHK03 datasets by 18.8%-20.3% in mAP and 19.2%-20.9% in R-1. This is because the FIDI loss-enabled Baseline2 can still gain the full benefits brought by a bag of different tricks used in Baseline2 while at the same time significantly improving Baseline2 when the datasets become more challenging.

3) Enabling Striping-based Methods: Lastly the FIDI loss is evaluated with the stripe-based methods, including some recent promising methods BDB [10] and MGN [27][3]. These methods are typically much more difficult to train and are computationally expensive than the other methods, because they involve multi-branch complex network structures. Since there is no consistent superiority of MGN and BDB over each other, we plug our FIDI loss into both methods.

The results are shown in the last row in Table II. We can see that the performance of both BDB and MGN is substantially improved in nearly all cases on the four datasets. Particularly, the FIDI loss consistently enhances BDB in both mAP and R-1 across all cases, especially lifting its mAP performance by 1.1% on Market1501, 4.2% on DukeMTM, 3.4% on CUHK03-D and 3.7% on CUHK03-L. The FIDI loss significantly improves MGN by 10.1%-13.2% in mAP and 13.9%-16.0% in R-1 on the two complex CUHK03 datasets. The FIDI loss-enabled MGN only works comparably well to, or less effectively than, the original MGN on Market1501. This may be due to that Market1501 is a simple and small dataset while MGN is a model with very complex architecture. Therefore, training MGN with our FIDI loss may lead to overfitting on this dataset.

4) Summary of Comparison: Overall, three main observations can be drawn across all the comparisons in Table II. First, our FIDI loss consistently and substantially improves all three types of recently proposed triplet loss-based state-of-the-art methods by a large margin on DukeMTMC and the two CUHK03 datasets. It is especially true on the complex CUHK03 datasets where the plugin of the FIDI loss typically results in 10%-20% improvement in both mAP and R-1, but the FIDI loss may not have clear advantages over the triplet loss on simple and/or small datasets such as Market1501. Second, by using the FIDI loss, simple models can perform

[3] DSA [47] also achieves state-of-the-art results on CUHK03-D and CUHK03-L datasets, but we cannot plug our loss into it since its source code is not available.

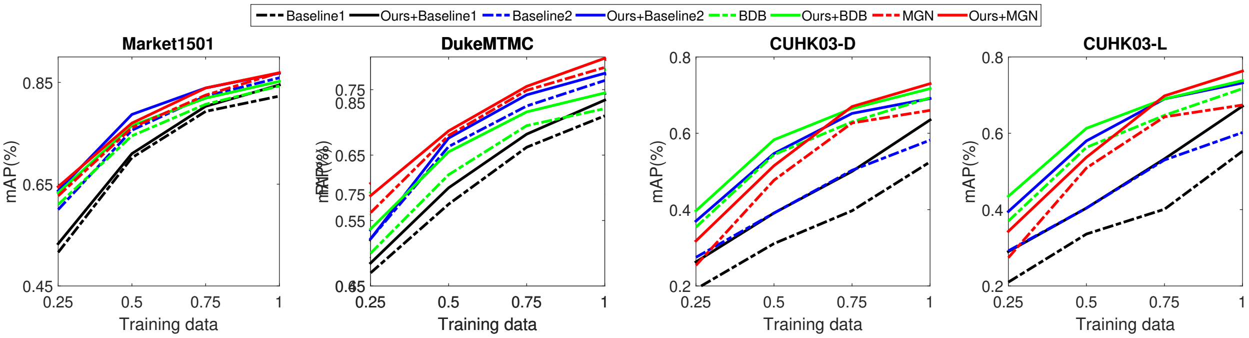


Fig. 6. mAP results on four datasets with varying percentage of training data.

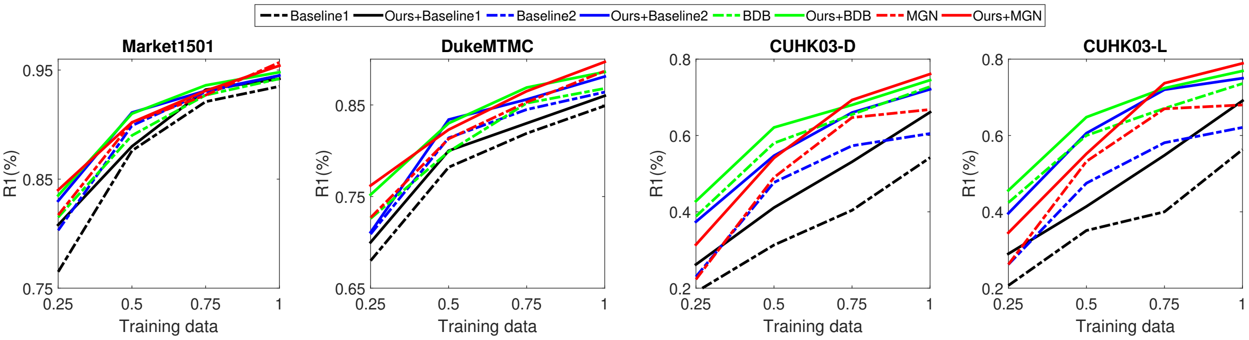


Fig. 7. R1 results on four datasets with varying percentage of training data.

substantially better than the complex models that use the triplet loss, e.g., 'Baseline2 (FIDI)' vs. BDB on Market1051, DukeMTMC and CUHK03-L. Third, the superiority of the FIDI loss sets new state-of-the-art results on DukeMTM, CUHK03-D and CUHK03-L, achieving 3.1%-7.2% improvement in mAP and 2.3%-4.6% improvement in R-1 over the prior best performance on the last two datasets.

E. Enhancing Data Efficiency

This section evaluates the data efficiency of the FIDI loss-enabled models. To do this, we reduce the training data by randomly removing 25% identities each step. The mAP results are given in Figure 6 and Figure 7.

It is clear that the FIDI loss-enabled models outperform their corresponding counterparts in all the training data settings across the four datasets, with substantial improvement in most cases. The performance of the proposed loss on the easy datasets Market1501 and DukeMTMC is mainly due to its shared key similar properties as the triplet loss, e.g., having an inter-class margin, while the superiority of our loss on the challenging datasets CUHK03-D and CUHK03-L is due to its fine-grained difference-aware capability and the bounded loss for easy samples. It is very impressive that even when three FIDI loss-enabled models use 25% less training data, they still can perform substantially better than the same models that use the triplet loss by a margin of at least 7.3% on CUHK03-D and CUHK03-L. This indicates that when handling challenging data, using a fine-grained difference-aware loss function is a much more cost-effective way than increasing the training data.

F. Beyond Person ReID: Enabling Vehicle ReID

To further evaluate the capability of our proposed loss, we evaluate the performance of the Baseline2 (FIDI) on two vehicle ReID datasets, VeRi-776 [32] and VehicleID [33].

1) Comparison with State-of-the-art Vehicle ReID Methods: We compare our method with 11 state-of-the-art vehicle ReID methods, including S-CNN [48], AAVER [49], VAMI [50], PROVID [18], MSVR [51], FDA-NET [52], OIFE [53], RAM [54], FACT [32], P-R [55] and Baseline2 [12]. The P-R and Baseline2 are more recent methods that have better performance than others.

The results are shown on Table III. We can see from the results that the Baseline2 (FIDI) outperforms most vehicle ReID methods by a large margin. Compared to Baseline2, our method achieves 1.3% - 2.6% improvement on mAP and and 0.6% - 0.7% improvement on R-1. This demonstrates that the proposed FIDI loss can effectively generalize from person ReID to vehicle ReID.

2) Visualization of Ranking Results: To provide a more straightforward illustration of the effectiveness, we present a set of visual image ranking results for vehicle ReID on VeRi-766 in Figure 8. We only show the results of Baseline2 (FIDI) and Baseline2 [12] because Baseline2 has better performance than all the other competing methods. As shown in Figure 8, with the increase of returned images, the accuracy of Baseline2 decreases. For example, the 15-th and 20-th returned images of

TABLE III
MAP AND R-1 PERFORMANCE ON VEHICLE REID DATASETS.

Methods	VeRi-776		VehicleID	
	mAP	R-1	R-1	R-5
S-CNN [48]	58.3	83.5	-	-
AAVER [49]	66.4	90.2	63.5	85.6
VAMI [50]	50.1	77.0	-	-
PROVID [18]	53.4	81.6	-	-
MSVR [51]	49.3	88.6	63.0	73.1
FDA-NET [52]	55.5	84.3	55.5	74.7
OIFE [53]	51.4	92.4	67.0	82.9
RAM [54]	61.5	88.6	67.7	84.5
FACT [32]	27.8	61.4	-	-
P-R [55]	74.3	94.3	74.2	86.4
Baseline2 [12]	75.7	95.2	77.5	91.0
Baseline2 (FIDI)	**77.6**	**95.7**	**78.5**	**91.9**

Baseline2 in the 2nd row are highly similar to the query image but they are actually different vehicles. This happens because Baseline2 fails to distinguish the fine grained appearance differences in the front of the vehicles. In contrast, Baseline2 (FIDI) can effectively capture the fined-grained differences and thus is able to return the images of the same vehicles taken from different viewpoints rather than the vehicles that are different from the query vehicle but share large similar appearance.

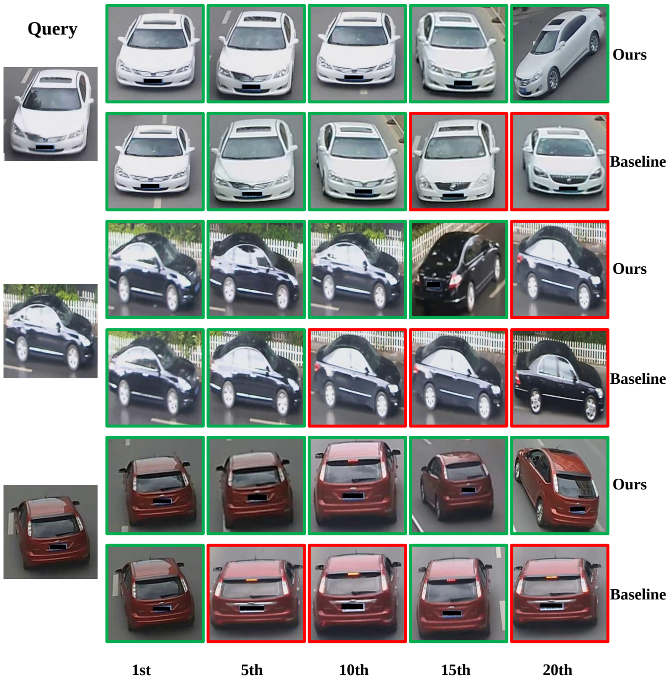


Fig. 8. Exemplar Ranking Results from Our Model, Baseline2 (FIDI), and the Original Baseline2.

G. Implication for Parameter Tuning

The two hyperparameters α and β, which respectively control the loss bound for easy samples and the sensitivity of the FIDI loss w.r.t. the pairwise distance, can be well tuned via cross validation. This section aims at providing some starting points of the parameter tuning based on our empirical results. Here $\alpha = 1.05$ and $\beta = 0.5$ are used by default and we vary one parameter with the other one fixed to examine its impact on the performance. Due to the page limit, we only present the mAP results of Baseline2 (FIDI) in Figure 9.

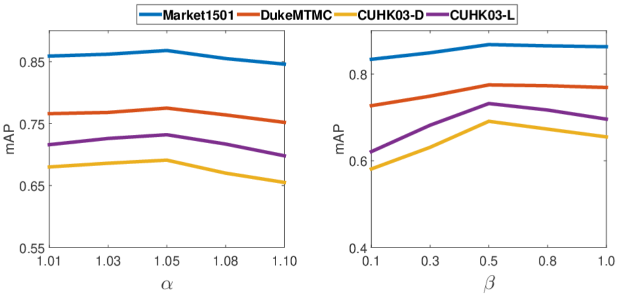


Fig. 9. mAP results w.r.t α and β in the FIDI loss

In general, FIDI is not sensitive to α unless it is too large. When the α is set to a small value, the punishment on images of small differences reduces, which decreases the final performance. On the right panel, we can see that it is beneficial to set a large β for challenging datasets CUHK03-D and CUHK03-L since in such cases our loss becomes more sensitive to the pairwise distance. Our loss imposes exponentially larger penalization on images of small differences that results in performance improvement. On the other hand, a relatively small β is more plausible for handling easier datasets like Market1501 and DukeMTM.

VI. Conclusions

This paper introduces a novel loss function called fine-grained difference-aware (FIDI) pairwise loss for the person ReID task. The FIDI loss not only ensures a similar inter-class margin as the triplet loss, but more importantly, also effectively penalizes images of both fine-grained and large appearance differences, especially on images of fine-grained differences. This delivers a significant improvement of three types of recent state-of-the-art ReID models in terms of both effectiveness and data efficiency. The improvement is particularly remarkable on complex datasets on which most current methods fail to work effectively. Also, our FIDI loss is simple and can replace the triplet loss as a plugin. All these characteristics make the FIDI loss a substantially more effective alternative to the widely-used triplet loss. We are performing large-scale studies to examine the applicability of replacing the triplet loss with the FIDI loss in other critical computer vision tasks.

VII. Acknowledgements

Cheng Yan and Xiao Bai were supported by the NSFC project no. 61772057 and BNSF project no. 4202039.

References

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159.

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1116–1124.

[3] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3754–3762.

[4] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 475–491.

[5] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[6] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 79–88.

[7] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person reid," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 3685–3693.

[8] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1249–1258.

[9] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 403–412.

[10] Z. Dai, M. Chen, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *International Conference on Computer Vision (ICCV)*, 2019.

[11] X. Chang, T. M. Hospedales, T. Xiang, and X. Chang, "Multi-level factorisation net for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2109–2118.

[12] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, pp. 1–10, 2020.

[13] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, 2018.

[14] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1062–1071.

[15] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 420–429.

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

[17] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, 2020.

[18] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2017.

[19] 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*, 2020.

[20] Q. Xie, W. Zhou, G.-J. Qi, Q. Tian, and H. Li, "Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization," *IEEE Transactions on Multimedia*, 2020.

[21] Y. Su, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

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

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1735–1742.

[24] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.

[25] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss," in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1335–1344.

[26] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

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

[28] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5098–5107.

[29] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person reid," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 365–381.

[30] T. M. Cover and J. A. Thomas, "Entropy, relative entropy and mutual information," *Elements of Information Theory*, vol. 2, pp. 1–55, 1991.

[31] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1179–1188.

[32] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 869–884.

[33] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2167–2175.

[34] 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 (ECCV)*, 2016, pp. 17–35.

[35] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[36] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, and S. Wang, "Perceive where to focus: Learning visibility-aware part-level features for partial person reid," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 393–402.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[38] 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 *European Conference on Computer Vision (ECCV)*, 2018, pp. 650–667.

[39] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3800–3808.

[40] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6036–6046.

[41] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2265–2274.

[42] G. Zhang and J. Xu, "Discriminative feature representation for person re-identification by batch-contrastive loss," in *Asian Conference on Machine Learning (ACML)*, 2018, pp. 208–219.

[43] X. Zhang, H. Luo, and X. Fan, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[44] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9317–9326.

[45] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[46] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In defense of the classification loss for person reid," in *Computer Vision and Pattern Recognition (CVPR)W*, 2019, pp. 50–58.

[47] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person reid," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 667–676.

[48] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 1900–1909.

[49] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual path modelwith adaptive attention for vehicle re-identification," *arXiv preprint arXiv:1905.03397*, 2019.

[50] Y. Zhou and L. Shao, "Aware attentive multi-view inference for vehicle re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6489–6498.

[51] A. Kanacı, X. Zhu, and S. Gong, “Vehicle re-identification in context,” in *Pattern Recognition*, 2018, pp. 377–390.

[52] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, “Veri-wild: A large dataset and a new method for vehicle re-identification in the wild,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3235–3243.

[53] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, “Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification,” in *International Conference on Computer Vision (ICCV)*, 2017, pp. 379–387.

[54] X. Liu, S. Zhang, Q. Huang, and W. Gao, “Ram: a region-aware deep model for vehicle re-identification,” in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[55] B. He, J. Li, Y. Zhao, and Y. Tian, “Part-regularized near-duplicate vehicle re-identification,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3997–4005.