In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation

Ye Yuan[1], Wuyang Chen[1], Yang Yang[2] and Zhangyang Wang[1]
{*ye.yuan, wuyang.chen, atlaswang*}*@tamu.edu*, *yang.yang2@walmart.com*
[1]Texas A&M University, [2]Walmart Technology
https://github.com/TAMU-VITA/FAT

Abstract

*The comparative losses (typically, triplet loss) are appealing choices for learning person re-identification (ReID) features. However, the triplet loss is computationally much more expensive than the (practically more popular) classification loss, limiting their wider usage in massive datasets. Moreover, the abundance of label noise and outliers in ReID datasets may also put the margin-based loss in jeopardy. This work addresses the above two shortcomings of triplet loss, extending its effectiveness to large-scale ReID datasets with potentially noisy labels. We propose a fast-approximated triplet (**FAT**) loss, which provably converts the point-wise triplet loss into its upper bound form, consisting of a point-to-set loss term plus cluster compactness regularization. It preserves the effectiveness of triplet loss, while leading to linear complexity to the training set size. A label distillation strategy is further designed to learn refined soft-labels in place of the potentially noisy labels, from only an identified subset of confident examples, through teacher-student networks. We conduct extensive experiments on three most popular ReID benchmarks (Market-1501, DukeMTMC-reID, and MSMT17), and demonstrate that FAT loss with distilled labels lead to ReID features with remarkable accuracy, efficiency, robustness, and direct transferability to unseen datasets.*

1. Introduction

Person re-identification (ReID) has attracted tremendous attention owing to its vast applications in video surveillance, public safety, and so on. Given a person image spotted by one camera, ReID aims to accurately match that probe image against a large amount of gallery images, taken by other cameras and timestamps. The dramatic visual appearance variations of the same person, as caused by different poses, view angles, illuminations, and backgrounds, constitute serious challenges for learning robust identity representations.

Most existing ReID algorithms use a classification loss to train their feature learning backbones [47, 40, 41, 18, 3, 44]. However, ReID is essentially an "open-ended" retrieval problem rather than closed-set classification, *e.g.*, the training and testing sets usually have no overlapped identity classes. The learned feature extractor should be able to generalize to matching unseen identities. The testing performance is evaluated by the precision and recall of the matching instances, rather than classification accuracy. Therefore, the classification-driven learning could be misaligned with the end goal. Instead, the comparative losses [30, 7, 24, 48], which compares the distances between two sample pairs, are naturally better choices, as empirically validated by a handful of works [20, 18, 4, 39, 5]. Among many, the triplet loss [12], which maximizes the margin between the intra-class distance and the inter-class distance, has been mostly used in ReID, in order to explicitly embed the relative orders

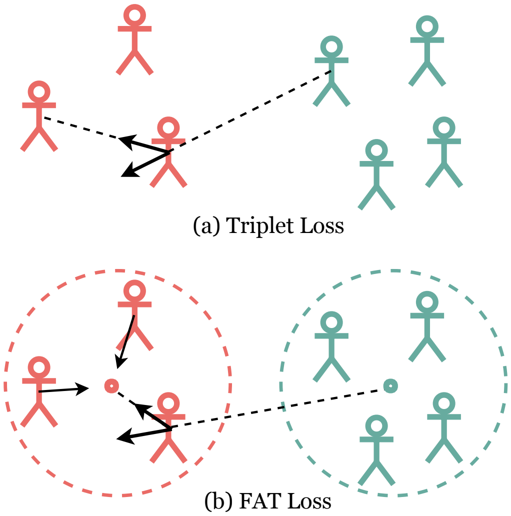

Figure 1: Illustrative comparison of standard triplet loss and FAT loss. The former compares point-to-point distances, while the latter compares point-to-set distances while regularizing all cluster sets to be compact. The solid arrows depict the "push and pull" effect of triplet loss and the point-to-set term of FAT loss. The dash arrows represents the compactness regularization of FAT loss. See details in Section 3.

between right and wrong matches (*i.e*., the correct matches should always be closer to the query than the wrong ones).

However, an important downside of triplet loss lies in its **computational expensiveness**, which prohibits its wide usage in the large-scale ReID applications. A naive triplet loss that compares every possible pair of training samples will incur cubic complexity w.r.t. the training set size [12]. Also, triplet loss relatively quickly learns to correctly map most trivial triplets, rendering a large fraction of all triplets uninformative. Applying triplet loss with randomly selected triplets can accelerate training but quickly stagnates, or becomes difficult to converge. Hard sample mining [42, 45] has recently become the standard practice in using triplet loss, to select only "informative" (a.k.a. hard) pairs rather than all pairs to enforce the loss. However, it runs the risk of causing sample bias [42], and often appears fragile to outliers. The vanilla triplet loss needs to calculate over all $PK(K-1)(PK-K)$ possible triplets, where K denotes average number of images per identity and P identities in total [12]. The time complexity can be reduced to $PK(PK-1)+PK$ when hard sample mining is used.

In this paper, we will propose a new fast-approximated triplet (**FAT**) loss to trim down the computational cost of triplet loss without hampering its effectiveness. Viewing all images belonging to the same identity class as a cluster, the proposed FAT loss re-defines a triplet to include an anchor, its corresponding cluster centroid, and the centroid of another cluster. The main idea of FAT loss is to replace point-to-point distances with point-to-cluster distances, through an upper bound relaxation of the triplet form. Such a relaxation simultaneously requires the query to be closest to its ground-truth-cluster centroid, and enforces each cluster to have a compact radius. The FAT loss thus has a linear complexity w.r.t. the training set size.

Another downside of triplet loss, as well as many other margin-based losses, lies in their **fragility to label noise**. Unfortunately, ReID datasets are notorious to have many noisy labels and outliers, such as label flipping, mislabel, and multi-person coexistence, due to the tedious manual annotation process. The proposed FAT loss can alleviate the label noise to some extent, by averaging all samples within the same cluster. To provide further improved robustness, we consider a distillation network to first generate soft pseudo labels for each sample, associated with its confidence. Then we use those soft labels in place of the original labels to feed into the FAT loss, where each individual sampleâs contribution to the model update will be re-weighted by their label confidence.

In sum, we strive to make triplet loss a more effective, efficient, and robust choice for ReID, via multi-fold efforts:

- We propose a fast-approximated triplet (FAT) loss to remarkably improve the efficiency over the standard triplet loss, with linear complexity to the training set size. It is derived by relaxing triplet loss to its upper bound form, and operates without hard sample mining.
- We are the first to demonstrate that explicitly considering and handling label noise can further boost ReID performance. A distillation network is presented to assign soft labels for samples in place of the original (potentially noisy) hard labels. Combined with FAT loss, a more robust re-ID feature can be learned.
- We conduct extensive experiments on three most popular ReID benchmarks, and demonstrate that FAT loss with learned soft labels lead to comparable or superior ReID performance than using triplet loss and other state-of-the-art baselines, with remarkably higher efficiency than triplet loss. We also observe improved robustness and direct transferability to unseen data.

2. Related Work

Triplet Loss and Hard Sample Mining The triplet loss was first introduced in FaceNet [30] by Google to train face embeddings for the recognition task, where softmax cross entropy loss failed to handle a variable number of classes. The goal of triplet loss is to maximize the inter-class variation while minimizing the intra-class variation. Triple loss is formulated as (1) below, where the triplet is defined as an anchor sample a, a positive sample p from the same class and a negative sample n from a different class (y_a, y_p, y_n denote class labels for a, p, n, respectively):

$$L_{\text{tri}} = \sum_{\substack{a,p,n \\ y_p = y_a \\ y_n \neq y_a}} \max\{d(a,p) + m - d(a,n), 0\} \tag{1}$$

FaceNet picked a random negative for every pair of anchor and positive, which was very time-consuming. Later on, [12] improved the efficiency of triplet loss for the ReID task, by proposing two triplet selection strategies: batch all and batch hard. The batch all strategy selects all valid triplets and averaged the loss. The batch hard strategy selects the hardest positive and negative samples within the batch when forming the triplets. The author suggested that batch hard strategy with soft margin to yield better performance. [42] found that selecting the hardest triplets often led to bad local minima. They argued that the bias in the triplet selection degraded the performance of learning with triplet loss, and proposed a new variant of triplet loss that adaptively corrects the distribution shift on the selected triplets.

Besides, there are many other successful practices in applying triplet loss to ReID task. [6] proposed a multi-channel convolutional neural network to learn global-local parts features and improved the triplet loss requiring the intra-class feature distances to be less than a predefined threshold. [4] extended the triplet loss to a quadruplet form and required the intra-class variations to be smaller than any inter-class variations. [43] generalized the point-to-point (P2P) triplet loss to the point-to-set (P2S) form by assuming a positive set (to which the anchor belongs) and a negative set (including all other clusters) for each anchor. It then penalizes the difference between the distance from the anchor to the positive set centroid and the anchor-to-negative-centroid distance. The model was also trained in a soft hard-mining scheme with greater weights to harder samples.

Being related to previous works [12, 43], FAT loss differs substantially in the following ways:

- FAT loss has linear time complexity w.r.t training dataset size: $\mathcal{O}(PK)$ or $\mathcal{O}(PK^2)$ (depending on the choice of negative set), where K denotes the average image number per identity and P the number of identities. Previous triplet losses have either cubic (vanilla) and quadratic (with hard sample mining) time complexity w.r.t training dataset size.
- FAT loss is analytically derived from the upper bound of standard triplet loss. It consists of a P2S loss term and intra-class compactness regularization. Up to our best knowledge, all previous approximations or accelerations for triplet loss, e.g., [6, 43], are only empirical.
- We studied different choices of the negative cluster/centroid, and compared their impacts. Note that FAT loss chooses the negative on "cluster" level, and does not refer to any individual sample mining.

Learning from Noisy Labels The growing scale of training datasets embrace the potential of a more powerful model, but introduces sample outliers and label noise during data collection and annotation. [35] observed that a face recognition model trained with only a subset 30% manually cleaned-label samples can achieve comparable performance with models trained on the full dataset. To overcome the negative effect of noisy labels, [28] proposed a bootstrap technique to modify the labels on-the-fly by augmenting the prediction objective with a notion of consistency. [21] extended [25] and proposed a re-weighting method that can be combined with any surrogate loss function for classification, to handle class-conditional random label flipping. [32] introduced an extra noise layer to absorb the label noise by adapting the network outputs to the noisy label distribution. [10] further augmented the correction architecture by adding a softmax layer on top to explicitly connect the correct labels to noisy ones. [26] provided a forward-and-backward loss correction method given a class-condition label flipping probability. [34] proposed a generic conditional random field (CRF) model as a robust loss to be plugged into any existing network for label space smoothness and therefore noise resistance. [36] designed a Siamese network to distinguish clean labels from noisy labels and to simultaneously give clean labels more emphasis.

Interestingly, various label noise, such as class-conditional or sample-conditional label flipping, mislabeling, and multi-person co-existence, are extensively found in ReID dataset. Yet to our best knowledge, few previous works have formally studied how to handle them, and how that may improve ReID performance.

Network Distillation Network distillation was first developed in [13] to transfer the knowledge in an ensemble of models to a single model, using a soft target distribution produced by the former models. [2] used distillation to train a more efficient and accurate predictor. [22] unified distillation and privileged information into one generalized distillation framework to learn better representations. [27] further extended data distillation to omni-supervised learning by ensemble of predictions from multiple transformations of unlabeled data to generate new training annotations using a single network. [23, 9] applied data distillation to multi-modal training, while the testing sets might have noisy or missing modalities. As a relevant work, [19] argued that noisy labels contains useful "side information" and shall not be discarded. The authors proposed a distillation approach to learn from noisy data guided by a knowledge graph.

Our proposed distillation algorithm to learn from noisy labels differs from previous ones in the following respects:

- We are free from the assumption of the existence of a manually-cleaned set. Instead, we train the teacher network with the entire noisy dataset, but only use most confident samples within a batch to update the parameters. We observed that the model updated based on a subset of confident samples can achieve similar or better performance, compared to the model trained with all noisy-labeled samples.
- We investigate different loss functions for distillation; the teacher network is trained with cross entropy loss with the purpose of providing pseudo soft label associated with a confidence; the student network is trained with FAT loss using the soft pseudo labels generated by the teacher network. Hence instead of mimicking a similar softmax classifier as the teacher network, the student network has the capability to "innovate" on a different task with the help of FAT loss, and eventually outperforms the teacher network.

3. Method

3.1. Fast Approximated Triplet (FAT) Loss

Given an anchor image a with the identity label y_a, the triplet loss attempts to find a positive sample p with the same identity label $y_p = y_a$ and a negative sample n with a different label $y_n \neq y_a$, and then maximizes the difference of distances between the positive pair $d(a, p)$ and the negative pair $d(a, n)$ by a margin m. We typically use the euclidean distance (or cosine similarity) between learned ReID features $f_E(a), f_E(p), f_E(n)$ as distance metrics.

However, computing triplet loss exhaustively over all

Algorithm 1 Derivation of FAT loss as an upper bound for triplet loss (1).

```
$L_{\text{tri}} = \max\{0, d(a,p) + m - d(a,n)\}$
    $\leq \max\{0, d(a,c_a) + d(c_a,p) + m - \max\{0, d(a,c_n) - d(c_n,n)\}\}$ ▷ refer to both inequalities in (2)
    $= \max\{0, d(a,c_a) + d(c_a,p) + m - d(a,c_n) + \min\{d(c_n,a), d(c_n,n)\}\}$ ▷ move d(a, $c_n$) out of inner $\max$ then reverse sign
    $= \max\{0, d(a,c_a) + m - d(a,c_n) + d(c_a,p) + \min\{d(c_n,a), d(c_n,n)\}\}$
    $= \max\{0, d(a,c_a) + m - d(a,c_n)\} + d(c_a,p) + \min\{d(c_n,a), d(c_n,n)\}$ ▷ move non-negative sums out of $\max$
    $\leq \max\{0, d(a,c_a) + m - d(a,c_n)\} + d(c_p,p) + d(c_n,n)$ ▷ $c_a$ = $c_p$; min{d($c_n$, a), d($c_n$, n)} $\leq$ d($c_n$, n)
    $\leq \underbrace{\max\{0, d(a,c_a) + m - d(a,c_n)\}}_{\textbf{anchor-dependent point-to-set loss}} + \underbrace{R(a) + R(n)}_{\textbf{cluster compactness}}$ ▷R() defines the radius of the cluster adnd can be pre-computed
```

possible pairs is too expensive to be practical. We propose a relaxation of the triplet loss 1 into its upper bound form. We first have the following two triangle inequalities:

$$\begin{aligned}\max\{0, d(a,c_a) - d(c_a,p)\} \leq d(a,p) \leq d(a,c_a) + d(c_a,p)\\ \max\{0, d(a,c_n) - d(c_n,n)\} \leq d(a,n) \leq d(a,c_n) + d(c_n,n)\end{aligned} \tag{2}$$

where c_a, c_n are defined as the centroids (average) of the clusters that a, n belong to, respectively. Their proofs are self-evident, given that $d()$ is a well-defined distance function in some metric space. Notice that although we use Euclidean distance for $d()$ by default, our derivations are applicable to other distances too.

We next expand our derivation as in the outline (1). Interestingly, the upper bound consists of two terms: a point-to-set (P2S) term which depends on the anchor point; plus a penalty term on the cluster compactness, defined as the largest cluster "radius" among all clusters, whose value is decided by the entire dataset and is agnostic to the anchor. We minimize this upper bound instead, and name it as the *fast approximated triplet* (**FAT**) loss:

$$L_{\text{FAT}} = \sum_{\substack{a,n\\ n\neq y_a}} \max\{0, d(a,c_a) + m - d(a,c_n)\} + R(a) + R(n). \tag{3}$$

As the name suggests, the new loss will give rise to similarly competitive ReID performance compared to the full triplet loss, but with tremendously better efficiency. We now analyze FAT loss w.r.t. the triplet loss from two aspects.

As can be obviously seen from its form, FAT loss greatly accelerates the cubic/quadratic time complexity of computing triplet loss, to linear complexity, w.r.t. the training set size. We also examine how tight it approximates the original tripelet loss. Observing (1), three relaxations take place in the second, sixth and seven lines. For the first one, the equality in (2) could be met when: a, c_a, p are co-linear with a, p on the same side of c_a; while a, c_n, n are also co-linear with a, n on different sides of c_n. The second relaxation becomes tight if and only if $d(a, c_n) \geq d(n, c_n)$, which implies that a is sufficiently far away from the cluster of c_n. For the last one, the exact equality can only be taken in a very special case, when every cluster has the same radius and every sample in a cluster distributes on a circle. In sum, when clusters are well-separated and balanced in size, FAT loss can provide a relatively tighter approximation for triplet loss. However, it is always reasonable to expect that minimizing this upper bound would lead to suppressing the original triplet loss value too.

Normalized FAT Loss As a margin loss, FAT loss, as well as triplet loss, is sensitive to input scales. Given the fact that ReID features are also scale-sensitive: neighboring features in the normalized space can be far away from each other in the original feature space, the learned feature are often normalized before feeding into the evaluation metrics. That could be reflected in a normalized FAT loss:

$$\begin{aligned}L_{\text{FATnorm}} = &\max\{0, d(\frac{a}{||a||}, c'_a) + m - d(\frac{a}{||a||}, c'_n)\}\\ &+ R'(a) + R'(n),\end{aligned} \tag{4}$$

where R' is similarly defined as the radius of the normalized sample set. In practice, we empirically find that adding a cross entropy (CE) loss L_{CE} term will help stabilize training with FAT or Normalized FAT loss notably. That leads to minimizing a hybrid loss ($L_{\text{CE-FAT}}$ can be replaced to $L_{\text{FAT-N}}$; λ is a scalar):

$$L_{\text{CE-FAT}} = L_{\text{FAT}} + \lambda * L_{\text{CE}} \tag{5}$$

Choices of Centroids The choice of cluster centroids is also found to be critical to the effectiveness of FAT loss. Four options of cluster centroids are available: i) mean of cluster features; ii) mean of normalized cluster features; iii) normalized mean of cluster features; and iv) normalized mean of normalized cluster features. Mathematically:

$$\begin{aligned}C_{i1} &= \frac{1}{N_i}\sum_{y_k=i} f_E(X_k), \qquad C_{i2} = \frac{1}{N_i}\sum_{y_k=i}\frac{f_E(X_k)}{\|f_E(X_k)\|}\\ C_{i3} &= \frac{\sum_{y_k=i} f_E(X_k)}{\|\sum_{y_k=i} f_E(X_k)\|}, \quad C_{i4} = \frac{\sum_{y_k=i}\frac{f_E(X_k)}{\|f_E(X_k)\|}}{\|\sum_{y_k=i}\frac{f_E(X_k)}{\|f_E(X_k)\|}\|}\end{aligned} \tag{6}$$

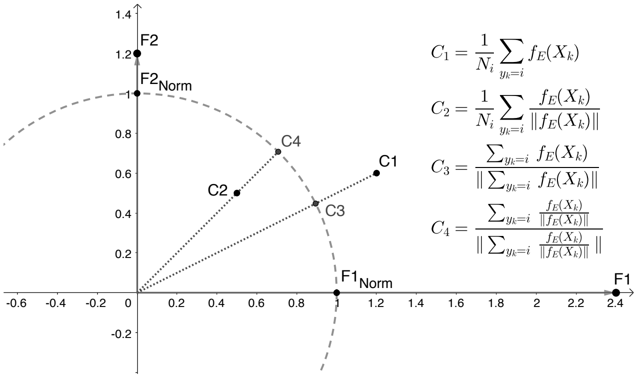
F2
F2Norm
C4
C1
C2
C3
F1Norm
F1
$C_1 = \frac{1}{N_i} \sum_{y_k=i} f_E(X_k)$
$C_2 = \frac{1}{N_i} \sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|}$
$C_3 = \frac{\sum_{y_k=i} f_E(X_k)}{\| \sum_{y_k=i} f_E(X_k)\|}$
$C_4 = \frac{\sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|}}{\| \sum_{y_k=i} \frac{f_E(X_k)}{\|f_E(X_k)\|} \|}$
1.4
1.2
1
0.8
0.6
0.4
0.2
-0.2
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4

Figure 2: Example of four different centroid options.

A visual comparison of the four options are in Figure 2.

Since the original FAT loss (3) is calculated based on un-normalized features, only the first centroid option C_{i1} makes sense for it. The remaining three options can all be utilized for the normalized FAT loss (4). Our experiments indicate that the normalized mean of normalized cluster features C_{i4} works best with the normalized FAT loss.

3.2. Distillation for Noisy Label Robustness

Typically, there are three common label noises in ReID datasets: i) label flip, i.e., an image is assigned to a wrong identity class; ii) mislabeling, i.e., an image does not belong to any known identity class; iii) multiple identities co-exist in one image. Similar to other margin-based losses, triplet loss is highly sensitive to label noise. Since the proposed FAT loss has a P2S term where all samples within the same cluster are averaged, hence alleviating noisy labels to some extents. We hereby propose a label distillation approach based on a teacher-student model, to improve FAT loss robustness to label noise further, using "soft labels" predicted from another teacher model, trained with a loss that is less sensitive to label noise, *e.g.*, cross entropy. The pipeline is plotted in the supplementary, with details explained below.

We first use a self-bootstrapping approach to learn the teacher model robustly. The teacher net is first trained with cross entropy loss on classifying all samples (including noisy labels) for 5 epochs. It was previously observed that the network would be more inclined to learning with high confidence for "easy samples", within the early stage of training [16, 15]. Those confident, easy samples are hypothesized to have labels that are semantically consistent and correct, less confusing and ambiguous, and therefore more reliable. We identify those most confidently predicted samples based on the entropy of their currently predicted softmax vectors. We then resume training for another 5 epochs; but now in each epoch, we will keep using those identified confident samples, while not using or only partially using the others that are more likely to contain label noise or outliers. We periodically repeat the above process, and each time we may gradually enlarge the pool of confident examples as the training continues. More details will be presented in Section 4.1.

After the teacher model is trained, its predictions are treated as soft labels to replace the original labels, for training the student model with FAT loss. Only the "confident" labels eventually selected by the teacher net will participate in averaging to estimate the cluster centroids. If we use the hybrid FAT loss (5), then soft labels are the prediction targets for the cross entropy (softmax) loss too.

4. Experiment

4.1. Datasets and Implementation

We evaluate the proposed method on three most popular large-scale benchmarks: Market-1501 [46], DukeMTMC-reID [29, 49], and MSMT17 [37].

Market-1501 comprises 32,668 labeled images of 1,501 identities captured by six cameras. Following [46], 12,936 images of 751 identities are used for training, while the rest are used for testing. Among the testing data, the test probe set is fiCEd with 3,368 images of 750 identities. The test gallery set also includes 2,793 additional distractors.

DukeMTMC-reID is a subset of the DukeMTMC dataset [29] for person ReID. This dataset contains 36,411 images of 1,812 identities, cropped from the videos every 120 frames. These images are captured by eight cameras, among which, 1,404 identities appear in more than two cameras and 408 identities (distractors) who appear in only one camera. The 1,404 identities are randomly divided, with 702 identities for training and the others for testing. In the testing set, one query image for each ID in each camera is chosen for the probe set, while all remaining images including distractors are in the gallery.

MSMT17 is the current largest publicly-available ReID dataset. It has 126,441 images of 4,101 identities captured by 15-camera network (12 outdoor, 3 indoor). We follow the training-testing split of [37].The video is collected with different weather conditions at three time slots (morning, noon, afternoon). All annotations, including camera IDs, weathers and time slots, are available. MSMT17 is **significantly more challenging** than the other two, due to its massive scale, more complex and dynamic scenes, and severe **label noise** (see examples in the supplementary).

Implementation of FAT Loss We implement our FAT loss in PyTorch deep learning framework. In the training phase, all images are resized to 144×432 and then randomly cropped into 128×384 sub-images. Standard horizontal flipping is adopted for data augmentation. In the test phase, all images are re-sized to 128×384 and no data augmentations are applied. All images have the training set

Table 1: Evaluation results on Market1501 and transfer results from Market1501 to DukeMTMC-reId. We use Resnet50 as our default backbone and trained on Market1501, with only one exception indicated by * using DenseNet161 backbone.

Settings			Test on Market1501				Transfer to DukeMTMC-reID			
loss	negative	margin	top1	top5	top10	mAP	top1	top5	top10	mAP
Histogram Loss [33]	NA	NA	59.5	80.7	86.9	-	-	-	-	
Multi-loss class [18]	NA	NA	83.9	-	-	64.4	-	-	-	-
Point to Set Similarity [51]	NA	NA	70.7	-	-	44.3	-	-	-	-
Triplet loss [12]	NA	1	84.9	94.2	-	69.1	-	-	-	-
Support Neighbor Loss [17]	NA	NA	88.3	-	-	73.4	-	-	-	-
CycleGAN [8]	NA	NA	-	-	-	-	38.5	54.6	60.8	19.9
CE-FAT	ctrdAll	1	89.1	95.0	96.7	71.6	34.4	51.5	57.6	18.9
CE-FAT	ctrdAvg	1	89.2	95.3	97.0	72.4	35.1	51.2	57.6	19.2
CE-FAT	ctrdHM	1	87.1	94.7	96.3	69.9	34.3	50.8	56.9	18.0
CE-FAT	batchNeg	1	89.4	95.6	97.1	73.1	37.3	52.3	58.4	20.3
CE-P2S	ctrdAll	1	87.4	95.0	96.7	68.9	27.6	42.9	50.0	14.1
CE-P2S	batchNeg	1	87.2	94.6	96.7	67.0	28.1	42.6	49.2	14.3
CE-P2Snorm	batchNeg	0.1	87.5	95.3	96.8	68.1	27.8	41.7	48.7	13.6
CE-FATnorm	batchNeg	0.1	88.6	95.1	96.7	69.7	35.0	50.6	57.4	18.9
CE-FAT* (DenseNet161)	batchNeg	1	**91.4**	**96.6**	**97.7**	**76.4**	**40.8**	**57.1**	**63.2**	**23.4**

Table 2: Evaluation results on DukeMTMC-reID and transfer results from DukeMTMC-reID to Market1501. We use Resnet50 as our backbone, and trained on DukeMTMC-reID, with only one exception indicated by * using DenseNet161 backbone.

Settings			Test on DukeMTMC-reID				Transfer to Market1501			
loss	negative	margin	top1	top5	top10	mAP	top1	top5	top10	mAP
Deep-Person [1]	NA	NA	**80.9**	-	-	**64.8**	-	-	-	-
CycleGAN [8]	NA	NA	-	-	-	-	48.1	66.2	72.7	20.7
CE-P2Snorm	batchNeg	0.1	76.5	87.3	90.6	57.3	46.5	63.9	71.0	19.9
CE-FATnorm	batchNeg	0.1	77.9	87.8	91.4	58.3	49.8	65.8	73.2	21.2
CE-P2S	batchNeg	1	78.2	88.5	91.8	59.5	47.0	64.6	71.4	19.7
CE-FAT	batchNeg	1	78.8	88.7	91.5	60.8	49.1	67.1	73.9	21.8
CE-FAT* (DenseNet161)	batchNeg	1	80.8	**89.5**	**92.0**	63.1	**54.7**	**70.8**	**77.4**	**25.2**

Table 3: Evaluation results on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17 with different negative sets.

loss	negative set	Test on MSMT17				Transfer to DukeMTMC-reID				Transfer to Market1501			
PDC [31]	NA	58.0	73.6	79.4	29.7	-	-	-	-	-	-	-	-
GLAD [38]	NA	61.4	76.8	81.6	34.0	-	-	-	-	-	-	-	-
HHL [50]	NA	-	-	-	-	45.0	59.4	64.4	23.0	**56.0**	**75.8**	**81.2**	26.7
CE-P2Snorm	batchNeg	64.8	78.3	83.0	33.8	49.1	64.9	70.6	29.2	51.6	68.9	75.5	23.9
CE-FATnorm	batchNeg	66.2	79.4	83.7	33.1	**51.2**	**66.1**	71.1	29.5	54.8	70.9	76.5	25.1
CE-P2S	batchNeg	65.2	78.5	82.9	33.7	49.9	67.6	**74.5**	22.9	48.7	63.5	69.3	**28.5**
CE-FAT	ctrdAll	68.8	81.4	85.4	39.1	50.9	65.0	70.2	**30.7**	51.5	69.4	75.9	24.4
CE-FAT	ctrdAvg	67.0	80.2	84.6	37.4	45.0	61.7	67.0	25.4	48.3	65.6	73.0	21.5
CE-FAT	ctrdHM	67.7	80.2	84.5	36.2	50.1	64.4	70.2	28.4	48.4	66.0	72.5	21.5
CE-FAT	batchNeg	**69.4**	**81.5**	**85.6**	**39.2**	49.2	64.8	69.6	28.7	50.6	68.0	74.9	23.6

mean subtracted and then normalized by the training set standard deviation, before feeding into the network.

Following a standard ReID protocol, we use ResNet [11] or Densenet [14] backbone as the feature extractor f_E towards learning a pedestrian representation directly supervised by FAT loss L_{fat}. The cluster centroids are computed at the beginning of each epoch, using C_{i1} for FAT loss and C_{i4} for normalized FAT loss in Equation 6. Besides, we also compare four different options of choosing the negative cluster c_n for computing FAT loss each time: i) *ctrdAll:* identity classes that are different from the one a belong to; ii) *ctrdAvg:* consider all other classes, except the one that a

Table 4: Evaluation results of the Teacher Net on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17.

Method	Test on MSMT17				Tranfer to DukeMTMC-reID				Tranfer to Market1501			
	top1	top5	top10	mAP	top1	top5	top10	mAP	top1	top5	top10	mAP
whole set	**65.1**	78.2	82.8	**34.5**	48.2	63.8	69.9	29.0	51.1	68.3	74.2	23.5
hard threshold	64.5	77.8	82.2	33.7	46.5	62.8	69.0	27.4	49.9	66.2	73.3	23.0
soft threshold	64.8	**78.3**	**83.0**	34.2	48.2	63.5	69.0	28.9	49.6	67.3	74.1	23.1
hard percentage	64.2	77.5	82.1	34.2	49.3	64.4	69.8	29.8	52.0	69.2	**76.5**	**24.8**
soft percentage	62.9	76.1	80.9	32.6	**50.5**	**66.0**	**71.0**	**30.3**	**52.4**	**69.6**	76.0	24.6

Table 5: Evaluation results of the Student Net on MSMT17, DukeMTMC-reID, and Market1501. We use ResNet50 as our backbone and trained on MSMT17.

loss	negative set	Test on MSMT17				Transfer to DukeMTMC-reID				Transfer to Market1501			
HHL [50]	NA	-	-	-	-	45.0	59.4	64.4	23.0	**56.0**	**75.8**	**81.2**	**26.7**
CE-FAT	batchNeg	**69.4**	**81.5**	**85.6**	**39.2**	49.2	64.8	69.6	28.7	50.6	68.0	74.9	23.6
CE-FAT-distillation	batchNeg	66.2	79.2	83.6	36.5	**50.9**	**66.6**	**72.2**	**31.3**	52.8	69.2	75.9	25.4

belongs to, as one cluster and obtain one negative centroid by computing the average of all negative centroids, which is similar to [43] but differs in the way of calculating all negative samples' mean; iii) *ctrdHM:* find a hard negative cluster (in terms of closest centroid to the one that a belongs to), from all classes of the whole dataset; iv) *batchHM:* find a hard negative sample on "batch level", *e.g.*, from all classes that are sampled by the current batch.

Implementation of Label Distillation The heavy label noise on MSMT17 further motivates us to conduct label distillation experiments on it. Following the basic routine described in Section 3.2, we further study four different modes of identifying confident samples: i) *hard threshold*: select all samples whose softmax entropy values are below a preset threshold t as the trusted training subset, and discard all un-selected samples; ii) *soft threshold*: select all samples whose softmax entropy values are below a pre-set threshold $t/2$, and then randomly select 50% of the remaining (unselected) samples to add into the trusted training subset; iii) *hard percentage*: always select 50% samples with lowest softmax entropy values, as the trusted training subset; iv) *hard percentage*: always select 25% samples with lowest softmax entropy values first, and then randomly select another $1/3$ from the remaining 75% (unselected) samples to add into the trusted training subset.

The important difference between "threshold" and "percentage" methods lies in whether we keep a constant or dynamic size of the trusted training subset for the teacher model. For the first two threshold-based methods, even sticking to the same t throughout one training, the portion of samples selected into the trusted set will be dynamic, as more samples might become better confident as training continues. Figure 3 visualizes this trend: given $t \leq 0.1$, the final training stage will always have considered all training samples as trusted; while a larger t may lead to more "conservative" selection. We choose $t = 0.1$ as the empirical default value found in experiments for i) and ii). Also, for the two "soft" strategies ii) and iv), our hope is to utilize a larger set of samples while letting the stochastic selection "smooth out" the impacts from noisy labels.

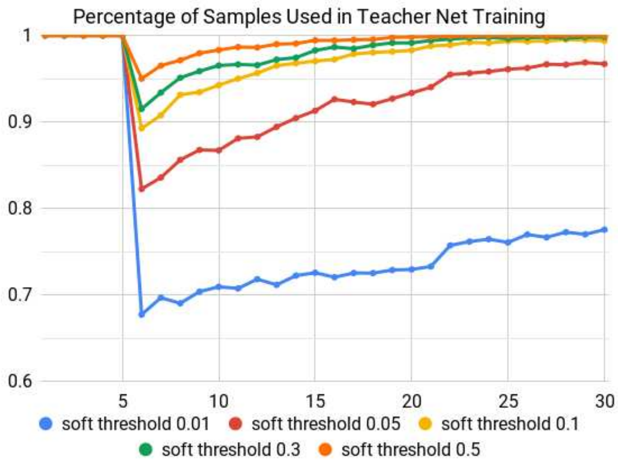

Figure 3: The number of samples actually used as the trusted training subset, when training the ResNet50 teacher model with different soft threshold t values, on the Market1501 dataset.

4.2. Comparison Analysis on FAT loss

We first present a comprehensive ablation study on the effectiveness of FAT loss in Table 1, using the Market1501 dataset. By default, we use the CE-FAT loss defined in (5), with $\lambda = 1$, as it consistently improves over either FAT or CE loss alone. The margin m is chosen as 1 for FAT loss and 0.1 for normalized FAT loss, as validated to be effective in experiments. We study on the four choices of the negative cluster (only ctrdAvg was previously explored in a similar form [43]), as well as the FAT loss hyperparameter (margin m). We also compare CE-FAT with CE-P2S, the latter defined by removing the cluster compactness term in FAT loss; as well as the normalized versions for both, denoted as CE-FATnorm and CE-P2S norm, respectively.

We evaluate different methods in terms of their top-1/top-5/top-10 accuracy and mean average precision (mAP) values obtained on the Market1501 testing set. Moreover, we use the **direct transfer** performance of the Market1501-trained feature extraction to the DukeMTMC-reID dataset, as an additional performance criterion, to avoid overfitting small ReID datasets. A few popular ReID loss options proposed in previous works [33, 18, 51, 12] are also included into comparison, so is a CycleGAN [8] baseline for transfer evaluation. Note that CycleGAN is a domain adaption method that demands re-training on the target domain, while the direct transfer needs no extra re-training.

First, comparing CE-FAT with ctrdAll, ctrdAvg, ctrdHM, and batchNeg, it is clear that batchNeg outperforms the other three. Second, comparing CE-P2S with CE-FAT in fair settings, we show the necessity of cluster compactness regularization in addition to the P2S loss; for example, without the compactness term, we will see 1.8% (ctrdAll) and 2.2% (batchNeg) top-1 accuracy drops on the Market1501 test case, and 7.5% (ctrdAll) and 9.2% d(batchNeg) top-1 accuracy drops on the transfer case to DukeMTMC-reID. The performance gaps clearly differentiate FAT loss from previous empirical P2S losses, thanks to our more rigorous upper-bound derivation. Third, no performance gain has been observed on Market1501, when using normalized features for FAT/P2S. Finally, CE-FAT outperforms all state-of-the-art losses trained with the same ResNet50, on the Market1501 testing set. Furthermore, after we replace the backbone into DenseNet161, CE-FAT achieves not only further boosted Market1501 testing results, but also impressive direct transfer performance to DukeMTMC-reID, even surpassing Cycle-GAN domain adaption [8] that is re-trained with the target domain data.

Tables 2 and 3 report similar experiments using DukeMTMC-reID ad MSMT17 datasets, respectively. With most observations aligned with the Market1501 cases, we find the training behavior on MSMT17 to slightly differ from the other two (much) smaller datasets. In particular, while batchNeg remains effective for its own testing set, ctrdAll becomes the best option when it comes to the feature transferability evaluation. That might be attributed to the heavier label noise on MSMT17, that likely benefits from averaging the triplet effects between with current one and all other clusters. Also, we observe CE-FATnorm to outperform CE-FAT, when transferring from MSMT17 to the other two datasets. That implies that normalization may become essential to overcome feature scale variances on large datasets. Finally, training ResNet50 with CE-FAT loss and batchNeg has surpassed the state-of-the-art performance [37] ever reported on MSMT17.

4.3. Effect of Label Distillation

To overcome the noisy label issue on MSMT17, we next investigate label distillation to further unleash the power of FAT loss. Both teacher and student nets adopt the same ResNet50 backbone for simplicity.

As shown in Table 4, for the training of the teacher net, the soft threshold/percentage methods appear to outperform their hard counterparts, as they can learn with a wider variety of samples (while hard methods may tend to select too many similar easy samples), meanwhile smoothing out the negative impacts of potential noisy samples due to stochastic sampling/averaging effects. In comparison, soft threshold seems to produce superior results on the same MSMT17 testing set, whereas soft percentage leads to better feature transferability. It implies that soft percentage suffers from less overfitting, because of its curriculum-style learning (as Figure 3 shows) that progressively takes into account the entire dataset information. To our surprise, our teacher net trained with only the trusted subsets by soft threshold/percentage yield competitive or even superior performance than the one trained with the whole dataset, in particular on transfer cases. That proves that the teacher net learns effectively and without being misled by noisy labels.

We then pick the teacher net trained with soft percentage, due to its best transfer performance, to provide soft pseudo labels for training the student net. The training of student net is supervised by the CE-FAT loss with the batchNeg strategy, using the soft pseudo labels in place of original one-hot labels for both CE and FAT terms. The new model in Table 5, dubbed CE-FAT-distillation, does not lead to better test results on MSMT17 than our best result (CE-FAT with batchNeg) in Section 4.2. However, it produces state-of-the-art **direct transfer** performance from MSMT17 to DukeMTMC-reID. Its transfer performance to Market1501 largely surpasses that of CE-FAT without distillation, and shows competitiveness to state-of-the-art HHL domain adaption [50]. To re-iterate, direct transfer does not re-train on target domain data as domain adaption has to.

5. Conclusion

This work proposes the fast-approximated triplet (**FAT**) loss, which remarkably improve the efficiency over the standard triplet loss in ReID models. Instead of using point-to-point distances, the FAT loss uses a point-to-set distances with cluster compactness regularization, which is derived rigorously as an upper bound of standard triplet loss, with linear complexity to the training set size. A distillation network is also designed to assign soft labels for samples in place of potentially noisy hard labels. Extensive experiments demonstrate the high effectiveness and promise of the proposed FAT loss along with label distillation.

References

[1] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. Deep-person: Learning discriminative deep features for person re-identification. *arXiv preprint arXiv:1711.10658*, 2017.

[2] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. Dropout distillation. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 99–107, 2016.

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abdnet: Attentive but diverse person re-identification. *ICCV*, 2019.

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPRW), Faces "in-the-wild" Workshop/Challenge*, volume 4, 2017.

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[10] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[15] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018.

[16] Hao Li and Maoguo Gong. Self-paced convolutional neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, pages 2110–2116, 2017.

[17] Kai Li, Zhengming Ding, Kunpeng Li, Yulun Zhang, and Yun Fu. Support neighbor loss for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1492–1500. ACM, 2018.

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[19] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017.

[21] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016.

[22] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[23] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[24] Zuheng Ming, Joseph Chazalon, Muhammad Muzzamil Luqman, Muriel Visani, and Jean-Christophe Burie. Simple triplet loss based on intra/inter-class metric learning for face verification. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 1656–1664. IEEE, 2017.

[25] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[26] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[27] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *The European Conference on Computer Vision (ECCV)*, September 2016.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[31] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

[33] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pages 4170–4178, 2016.

[34] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5596–5605. Curran Associates, Inc., 2017.

[35] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[36] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 420–428, New York, NY, USA, 2017. ACM.

[39] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017.

[40] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang 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*, pages 1249–1258, 2016.

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2, 2016.

[42] Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[43] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[44] Ye Yuan, Wuyang Chen, Tianlong Chen, Yang Yang, Zhou Ren, Zhangyang Wang, and Gang Hua. Calibrated domain-invariant learning for highly generalizable large scale re-identification. In *WACV*, 2019.

[45] Yiru Zhao, Zhongming Jin, Guo-jun Qi, Hongtao Lu, and Xian-sheng Hua. An adversarial approach to hard triplet generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–517, 2018.

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[47] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[48] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Feiyue Huang, and Yanhua Yang. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In *IJCAI*, pages 1226–1233, 2018.

[49] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[50] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[51] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person reidentification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, 2017.