Pyramid Loss for Person Re-identification

Yuanyuan Wang, Zhijian Wang and Mingxin Jiang

***Abstract*—**Person re-identification (ReID) is an important task in computer vision, meanwhile attracted the attention of industry. Person ReID focuses on identifying person among multiple different cameras. A key under-addressed problem is to learn a good metric for measuring the similarity among images. Recently, deep learning networks with metric learning loss has become a common framework for person ReID, such as triplet loss and its variants. However, the previous method mainly uses the distance to measure the similarity, and the distance measure is more sensitive when the scale changes. In this paper, we propose pyramid loss to learn better similarity metric for the person ReID. Our approach uses the angular relationship in triangles as a measure of similarity, minimizing the angle at the negative point of the triangle. Pyramid loss can learn better similarity metric and can achieve a higher performance on the person ReID benchmark datasets. The experimental results show that, our method yields competitive accuracy with the state-of-the-art methods.

***Keywords*—**person re-identification, metric learning, pyramid loss.

I. INTRODUCTION

IN recent years, person re-identification (ReID) has attracted significant attention due to its wide applications in video surveillance, such as surveillance security and retrieval of suspects. Person ReID aims at matching persons observed in non-overlapping camera views with visual features and finding a person-of-interest (query) among a gallery of person image dataset [1]. Owing to the various difficulties including changed lighting, large variations of body pose, background environments and view angles, occlusions and low-resolution images, the similarity among different persons increases the difficulty. Person ReID is still a challenging problem [2].

In the person ReID task, deep learning has attained better results than the traditional approach recently [3], [4], [5], [6]. The existing methods mainly consists of two stages, extracting discriminative features from person images and computing the distance of samples by feature comparison. The convolutional neural network (CNN) is frequently used for feature representation, extracts discriminative features from the query and the gallery images [7], [8], [9]. The first stage mainly considers extracting more robust features. The second stage involves the feature learning.

In supervised learning, representation learning [4] and metric learning [10], [11] are two types of methods in terms of the target loss. For the representation learning, person ReID is considered as a verification or identification task. Recently, several person ReID approaches show that identification loss combined with verification loss can learn more discriminative person embedding [4], [7], [12], [13], [14]. However, the parameters of identification loss grow when the number of identities increases, many parameters are discarded after retraining. The representation learning method focuses on the dissimilarity between different classes and ignores the similarities between pairs of same persons. In addition, the use of verification loss to determine the similarity of two persons images, the efficiency is relatively low, especially when the number of pedestrian categories is large or uncertain. Therefore, different metric learning methods, such as triplet loss [15], improved triplet loss [10], quadruplet loss [16], margin sample mining loss [2], lifted structured loss [8] and multi-class N-pair loss [17] are proposed. These methods get better performance than representation learning methods. The metric learning problem finds a mapping function, minimizes the same person distance and maximizes the different persons distance. This work focuses on the distance metric learning in person ReID.

Nevertheless, the previous methods mainly consider the optimization of similarity (*e.g.* contrastive loss [4]) or the relative similarity (*e.g.* triplet loss [15]). They mainly use the distance to measure the similarity, and the distance measure is more sensitive when the scale changes. The difficulty of training partly comes from the limitation by defining the purpose samples only in distance [18]. Moreover, the selection of margin threshold is obviously not suitable for different intra-class [5]. All above referred losses are defined in term of distances of sample points, and only a few has considered other probably forms of loss (*e.g.* angular loss [18], clustering loss [11] and smart mining loss [19]).

Person ReID is mostly viewed as an image retrieval problem and the angular loss [18] is proposed to solve the image search task. The experimental results show that, the angular loss method [18] achieves competitive results in birds dataset and cars dataset. Commonly used triplet loss presents problem in solving unbalanced intra-class change. The latest deep metric learning method multi-class N-pair loss [17] improves upon the triplet loss on the task of image retrieval. However, the N-pair loss compares multiple negative samples, the dimension will be

This work was supported in part by the Science and Technology Projects of Huaian(HAG201602), Six talent peaks project in Jiangsu Province under Grant 2016XYDXXJS-012, the Natural Science Foundation of Jiangsu Province under Grant BK20171267, 533 talents engineering project in Huaian under Grant HAA201738.

Yuanyuan Wang is with the College of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, Jiangsu, China. She is a Ph.D student from the College of Computer and Information, Hohai University, China. (corresponding author; e-mail: zhfwyyzjx@gmail.com).

Zhijian Wang is with the College of Computer and Information, Hohai University, Nanjing 211100, Jiangsu, China.

Mingxin Jiang is with the College of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian 223003, Jiangsu, China.

high and the computation is huge if multiple negative samples are introduced into angular loss. Angular loss [18] uses the relationship of angles rather than distances as a measure of similarity. Angular in the triangle has rotation invariance and scale invariance. Meanwhile, third-order potentials can be similarity-invariant by comparing angles of triangles [20]. Therefore, we propose a more stable pyramid structure for person ReID.

In this paper, we propose a novel pyramid loss which introduces the idea of quadruplet network [16] to angular loss geometry structure [18]. Our method constructs a triangular pyramid structure using angular relations as the measurement. The pyramid loss is expanded from the angular loss which introduces another negative person sample in structure. The performance of the quadruplet loss on the testing set can be improved by further reducing the intra-class variations and enlarging the inter-class variations [16]. We use angular relationships in a triangular pyramid structure to achieve this requirement. The structure of pyramid includes four samples, an anchor image, a positive image and two different negative images. Our method limits the angle $\angle n$ between the negative point of the triplet triangles, while constraining another angle $\angle \beta$ between the sides of the negative sample and its adjacent plane in Fig. 1. Similar to [16], the constraining $\angle \beta$ is helpful to improve the generalization ability of the trained models on testing dataset and get better performance on testing dataset in our experiment. Hyperparameters *n* and β are easier to choose than the distance-based triplet loss with unbalanced intra-class change margin. Then it produces a stronger push between positive and negative pairs. Our method improves the robustness of the target to the feature differences, captures the additional local structure of triplet triangles. Additionally, our method describes its local structure more accurately than triplet loss based on distance. The main advantage of our idea is the introduction of a triangular pyramid structure that can be further resistant to scale changes than the angular loss [18]. It is noted that hard samples are feedback to train the model, inappropriate training samples will produce gradients that are close to zero, cause a slow convergence and achieve a suboptimal solution [16], [19]. The methods of distance-based deep metric learning explore variety of hard negative / positive mining methods [2], [16], [19]. In this work, we do not need to select margin threshold which has a larger range of values the same as triplet loss. We only need to determine the angle relationship between the samples in a pyramid structure. By the nature of the triangle and the relationship between the person samples, it is clear that the values of the hyperparameters $\angle n$ and $\angle \beta$ must be less than 60°. We determine the value range of the hyperparameters through the experimental statistical data, and finally optimize the hyperparameters by the hyperparameter optimizer. Our method illustrates performance on common person ReID datasets *i.e.*, Market1501 [21], CUHK03 [1] and MARS [3]. The performance is comparable to the state-of-the-art methods on the person ReID problem. To the best of our knowledge, this is the first work to explore pyramid loss for person ReID.

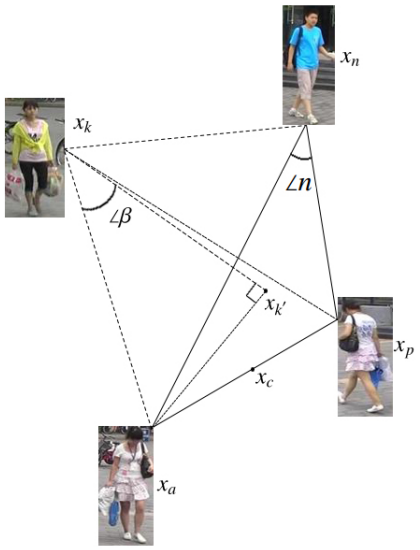

Fig. 1. Illustration of the pyramid loss. The triple loss $\{x_a, x_p, x_n\}$ constitutes plane $\mathcal{P}$, x_k is the vertex of the pyramid structure, $x_{k'}$ is the projection of x_k on plane $\mathcal{P}$. x_a and x_p are the same person, x_n is a different negative sample, x_k is another different negative sample, x_n and x_k are not the same person. The angles of $\angle n$ and $\angle \beta$ are expected to be less than pre-defined hyperparameter θ and δ respectively.

II. Relate Work

In the past, traditional metric learning methods learn a Mahalanobis distance in Euclidean space, such as Keep It Simple and Straightforward Metric Learning (KISSME) [22] applied in person ReID. Recently, deep metric learning methods usually extract features from CNN or other model, and then compute the feature distances in Euclidean space. Triplet loss function [5], [17], [19], [23] is used to investigate the relative similarity of different pairs of person images. It has been widely used in person ReID retrieval [4] and face recognition [15]. In deep metric learning, a positive pair are two images of the same person whereas a negative pair are two of different persons. A triplet is made up of three persons samples, which comprise a positive pair and a negative pair. The positive pair distance is enforced to be smaller than the negative pair, and the triplet loss is motivated by the threshold between positive and negative pairs. However, the triplet loss needs mining hard samples for efficient mining of similar features, otherwise training process will stagnate, training unstable and time-consuming [10], [19]. It will lead to training process shocks, unable to converge if the sample is too difficult.

To address above problems, some variants of the triplet loss and hard negative/positive mining methods are proposed [2], [5], [10], [24], [25], [26]. Wu *et al*. [26] propose DeepLDA method which bring in LDA objective function using fisher vectors. However, it seems more difficult to train. Cheng *et al*. [5] propose ImpTrpLoss to reduce the same class variations in person ReID. They further restrict the distance between pairs belonging to the same type based on triplet loss to be less than a

preset value. Unfortunately, the method partly neglects the relative relationships between pairs. Ding *et al*. [24] propose batch all triplet loss which count all possible when calculating loss. Song *et al*. [8] propose the method which fill the batch with triplets. It considers all except the anchor-positive pair as negatives, meanwhile optimize the smooth boundary of loss. Hermans *et al*. [10] propose a generalization of the lifted embedding loss which considers all anchor-positive pairs based on [8] and [24]. Chen *et al*. [16] propose the quadruplet loss for person ReID which introduces a new constraint. It extends the triplet loss by adding a different negative pair. Sohn [17] proposes multi-class N-pair loss which further extends the triplet loss by pushing away multi negative samples. Xiao *et al*. [2] propose margin sample mining loss (MSML). MSML introduces the idea of hard sample mining which only picks out the hardest positive sample pair and the hardest negative sample pair to calculate the loss. Ben Harwood *et al*. [19] propose a method which automatically mining hyperparameters and accelerate the convergence. Wang *et al*. [18] use triangular geometry to capture the local structure of triplet loss.

Then our work introduces structure of quadruplet into a pyramid structure. The proposed method limits the relationship between samples through the angular relationship in the pyramid structure. The optimized hyperparameters make the proposed model easier to achieve convergence.

III. Methodology

In this section, we propose pyramid loss applied in person ReID. Our method is designed based on the angular loss [18] and quadruplet network [16]. We present the pyramid loss by rebuilding a triangle pyramid structure. Then we discuss the optimization of the loss function in a batch.

A. The Triplet Loss

Traditional triplet loss, the triplet loss variants or contrastive loss are based on distance measurement, which cannot solve the problem of scale change. It is difficult to select an appropriate global distance margin α or β in Eq. 2 due to the distance of intro-class vary sharply. Wang *et al*. [18] propose the angular loss, constraining the angle of the triplet's negative sample point. The angular loss has scale invariance and improves the robustness of the objective function to counter the feature change. The angular loss essentially adds third-order geometric constraints, which can capture additional local structures in comparison with triplet loss or contrastive loss.

The goal of metric embedding learning is to learn a function $f(x):\mathbb{R}^F \to \mathbb{R}^D$ which maps semantically similar instances from the data manifold in $\mathbb{R}^F$ onto metrically close points in $\mathbb{R}^D$ [10]. The triplet loss has been proved effective in learning discriminative image features compared to softmax loss for classification, which widely used in person ReID [2] and face recognition [15]. There are many literatures that have proposed many methods to improve triplet loss for higher performance on the testing set. Each of triplet loss $\{x_a, x_p, x_n\}$ contains an anchor x_a, a positive x_p and a negative x_n in an iteration of the batch. x_a and x_p are images from the same person, and x_n is from another different person. The philosophy of triplet loss function is try to minimize the distance between an anchor and a positive person sample meanwhile maximize the distance between the anchor and the negative pairs. The triplet of ℓ_2-normalized features $\{f_a, f_p, f_n\}$ is used to calculate distances and the commonly used triplet loss can be formulated as following:

$$L_{triplet} = \sum_{a,p,n}^{m} \left[\overbrace{\left\| f(x_a) - f(x_p) \right\|_2}^{\text{minimize}} - \overbrace{\left\| f(x_a) - f(x_n) \right\|_2}^{\text{maxmize}} + \alpha_{triplet} \right]_+ \quad (1)$$

where threshold $\alpha_{triplet}$ is a distance margin distinguish the positive pairs with the negative. $f(x_a)$, $f(x_p)$, $f(x_n)$ represents normalized highly-embed features and $[x]_+ = \max(x, 0)$.

B. The quadruplet loss

The triplet loss [15] works well on the training set. However, the ability to generalize from training set to testing set is weaker with poor performance. The quadruplet loss [16] improves the triplet loss by adding a different negative pair. A quadruplet loss function involves four different images $\{x_a, x_p, x_s, x_t\}$, where x_a and x_p are images of the same person while x_s and x_t are images of another two persons separately. The quadruplet loss is formulated as following:

$$\begin{aligned} L_{\text{quad}} &= \sum_{a,p,s}^{N} \left[\overbrace{\left\| f(x_a) - f(x_p) \right\|_2 - \left\| f(x_a) - f(x_s) \right\|_2 + \alpha}^{\text{relative distance}} \right]_+ + \\ &\quad \sum_{a,p,s,t}^{N} \left[\overbrace{\left\| f(x_a) - f(x_p) \right\|_2 - \left\| f(x_s) - f(x_t) \right\|_2 + \beta}^{\text{absolute distance}} \right]_+ \\ &= \left[d_{a,p} - d_{a,s} + \alpha \right]_+ + \left[d_{a,p} - d_{s,t} + \beta \right]_+ \end{aligned} \quad (2)$$

where α and β are the values of the margins, β is set to be less than the marginal α to achieve the relative perimeter constraint, so the first term plays a major role. In Eq.2, the first term is the same as Eq.1 which produces a strong push between positive and negative samples for the same probe image. The second term provides a relatively weaker push to reduce the inter-class variations for different probe images.

Furthermore, if we ignore the effects of the parameters α and β, we can represent the quadruplet loss in a more general form as following:

$$\begin{aligned} L_{\text{quad}} &= \sum_{a,p,s,t}^{N} \left[\left\| f(x_a) - f(x_p) \right\|_2 - \left\| f(x_s) - f(x_t) \right\|_2 + \alpha \right]_+ \\ &= \left[d_{a,p} - d_{s,t} + \alpha \right]_+ \end{aligned} \quad (3)$$

where *s* and *t* are a pair of negative samples, *s* and *a* may be either a pair of positive samples or a pair of negative samples.

C. The Pyramid Loss

The quadruplet loss further enlarges inter-class variations and reduces intra-class variations via introducing another

negative samples. It provides an auxiliary relatively weaker constraint from the perspective of different probe images. The pyramid loss introduces quadruplet network to angular loss for further constraints which pushes away negative pairs from positive pairs.

In a triangle, we construct the anchor sample, positive sample and negative sample as angular losses which form the triangle Δ_{apn}. The edges are denoted as e_{na}, e_{np} and e_{ap}, shown in Fig. 1. It is obviously that the purpose is to make the distance of anchor and negative as large as possible, the distance of anchor and positive as small as possible. Intuitively, angular loss [18] specifies that the angle of the negative point $\angle n$ in triplet loss is less than a certain value $\angle\theta$, shown in Fig. 1. x_a and x_p are the same person, x_n is a different negative sample, x_k is another different negative sample, yet x_n and x_k are not the same person. Depending on the nature of the triangle, we want to keep negative samples away from anchor and positive samples, then we make $\angle n$ smaller. Using the tangent theorem and the defined angle *n*, the angular loss [18] consists of minimizing the following hinge loss $L_{Angular}$, where circumcircle passing through x_a and x_p, centered at the middle $x_c = (x_a + x_p)/2$. In Eq. 4, it constrains the angle $\angle n$ closed by the edge of e_{na} and e_{np} to be less than a pre-define upper bound θ. The hyper parameter θ works well which is set between 25º - 35º in the experiment.

$$L_{Angular} = \left[\left\|x_a - x_p\right\|_2^2 - 4\tan^2\theta\left\|x_n - x_c\right\|_2^2\right]_+ = \left[d_{a,p}^2 - 4\tan^2\theta \cdot d_{n,c}^2\right]_+ \quad (4)$$

It is noted that further reducing intra-class variations and enlarging inter-class variations can decrease the generalization error of trained models and improve the performance of the triplet loss on the testing set [16]. Then we introduce a plane $\mathcal{P}$, x_a, x_p and x_n are in this plane. The equation of plane $\mathcal{P}$ is $Ax_a + Bx_p + Cx_n + D = 0$. Our pyramid loss extends the triplet loss by adding a different negative pair. The structure of pyramid contains four different person images $\{x_a, x_p, x_n, x_k\}$, where x_a and x_p are images of the same persons while x_n and x_k are images of another two persons respectively ($x_k \neq x_a$, $x_k \neq x_n$). We construct a triangular pyramid □*kapn* using x_k, x_a, x_p and x_n. $x_{k'}$ is the projection of x_k on plane $\mathcal{P}$, $\angle\beta$ is the angle between the side edge e_{ka} and $e_{kk'}$. In the triangular pyramid □*kapn*, we push away x_k from the plane $\mathcal{P}$, then $\angle\beta$ would be smaller. In experiment, we set $\angle\beta < \angle\delta$, δ is pre-define hyper parameter. We formulate Eq. 5 to constrain the angle β to be less than δ. It works well which is set 15º - 25º in the experiment. x_k is pushed away from the plane $\mathcal{P}$, both $\angle\beta$ and $\angle n$ are getting smaller and smaller. Then we argue that the length between $e_{ak'}$ and $e_{an}/2$ is similar, $x_{k'} \approx (x_a + x_n)/2$. The distance from x_k to plane $\mathcal{P}$ is formulated by Eq. 6, where the coordinates of point $x_{k'}$ are (x_{a0}, x_{p0}, x_{n0}).

$$\tan\beta = \frac{\left\|x_a - x_{k'}\right\|_2}{\left\|x_k - x_{k'}\right\|_2} \approx \frac{\left\|x_a - x_n\right\|_2}{2\left\|x_k - x_{k'}\right\|_2} \leq \tan\delta \quad (5)$$

$$\left\|x_k - x_{k'}\right\|_2 = \frac{\left|Ax_{a0} + Bx_{p0} + Cx_{n0} + D\right|}{\sqrt{A^2 + B^2 + C^2}} \quad (6)$$

Inspired by the angular loss [18] and quadruplet loss [16], we seek for the optimum embedding such that the samples of different classes can be separated well. Our triangular pyramid loss consists of minimizing the following hinge loss.

$$L_{Pyramid} = \left[\left\|x_a - x_p\right\|_2^2 - 4\tan^2\theta\left\|x_n - x_c\right\|_2^2\right]_+ + \left[\left\|x_a - x_n\right\|_2^2 - 4\tan^2\delta\left\|x_k - x_{k'}\right\|_2^2\right]_+ \quad (7)$$

In order to improve the overall of the performance, we combine pyramid loss with the traditional distance metric loss, one of the latest work MSML [2]. In Eq.1, the triplet loss uses the Euclidean distance to measure the similarity of extracted features from two input images. Based on the method proposed by Wang [27], we use the learning metric $g(x_a, x_p)$ instead of the Euclidean distance to improve the robustness. Wang [27] *et al*. use a fully connected layer with a one-dimensional output to learn the value $g(x_a, x_p)$. Regardless of the threshold $\alpha_{triplet}$, the model can multiply $g(x_a, x_p)$ and $g(x_a, x_n)$ by appropriate values to meet the boundary threshold requirement. At the same time, a softmax constraint is added to obtain the similarity of [0,1]. The optimized MSML is shown in Eq. 9.

$$L_{MSML'} = \sum_{a,p,s,t}^{N}\left[\max_{a,p}\left\|g(x_a) - g(x_p)\right\|_2 - \min_{s,t}\left\|g(x_s) - g(x_t)\right\|_2 + \alpha\right]_+ \quad (8)$$

We simultaneously consider the relationship of distance and angle among samples in Eq. 10.

$$L_{Pyramid'} = \mu L_{Pyramid} + L_{MSML'} \quad (9)$$

where μ is a trade-off weight between the pyramid loss and MSML. In the experiment, we always set $\mu = 2$.

Inspired by [8], [17], [18], we adjust the upper bound of the smoothness in Eq. 8 in our experiment. It is assumed that the feature has a unit length in Eq. 8, We use Eq. 11 to represent the pyramid loss of a batch Ψ.

$$L_{\text{Pyramid}}(\Psi) = \frac{1}{N}\sum_{x\in\Psi}\left\{\log\left[1 + \sum\exp(g_{a,p,n,k})\right]\right\} \quad (10)$$

$g_{a,p,n,k}$ is shown as Eq. 11.

$$\begin{aligned} g_{a,p,n,k} &= \left\|x_a - x_p\right\|^2 - 4\tan^2\theta\left\|x_n - x_c\right\|^2 + \left(\left\|x_a - x_n\right\|^2 - 4\tan^2\delta\left\|x_k - x_{k'}\right\|^2\right) \\ &= -2x_a^T x_p - 4\tan^2\theta(-2x_n^T x_c + x_c^T x_c) + (-2x_a^T x_n - 4\tan^2\delta(-2x_k^T x_{k'})) \\ &= 4\tan^2\theta(x_a + x_p)^T x_n - 2(1 + \tan^2\theta)x_a^T x_p + 4\tan^2\delta \cdot x_k^T(x_a + x_n) - 2x_a^T x_n \end{aligned} \quad (11)$$

where $x_c = (x_a + x_p)/2$, $x_{k'} \approx (x_a + x_n)/2$.

In summary, compared with other metric learning losses, the proposed method has advantages as following. Our method applies the pyramid loss to the person ReID, constraining the

angle of the negative point of the triplet. Our pyramid loss introduces another negative sample in angular loss producing a stronger push between positive and negative pairs. The method can be used to counter the scale and feature variation. Taking into account the relative distance and absolute distance, it combines the idea of hard sample pairs mining with the pyramid loss.

IV. Experiments

In this section, we mainly evaluate the proposed method using three common benchmark datasets of person ReID, *i.e.*, Market1501 [21], CUHK03 [1] and MARS [3]. We report the results trained by two network structures, GoogLeNet and ResNet-50. Then we compare the proposed approach with state-of-the-art methods.

A. Datasets and Implementation

The Chainer package is used throughout the experiment. Chainer [28] is an open-source deep learning framework featuring the define-by-run approach. Each image is normalized to 256 × 256 pixels before processing with data enhancement such as random horizontal flip, random crop and zoom. The final feature dimensions of GoogLeNet and ResNet-50 are transformed to 1024 through a fully-connected layer. Adam optimizer is used and the initial learning rate is set to 0.0001. Hyper parameters optimizer sklearn is used to search hyperparameter space to find the most reasonable hyperparameter θ and δ for learner model. Parameter Sampler is chosen for model_selection. We use SGD with 20k training iterations and 60 mini-batch sizes consuming 6GB of memory. The HDF5 format is used to read and write data files, data of different types can be embedded in a HDF5 file.

The experiment is conducted on three datasets including Market1501 [21], CUHK03 [1] and MARS [3]. Market1501 [21] is one of the most widely used datasets in the person ReID field. It contains 32,668 annotated bounding boxes of 1,501 identities collected from six cameras. It contains 19,732 images for testing and 12,936 images for training. There are 17.2 images per identity in the training set. The images are automatically detected by the deformable part model (DPM) instead of using hand-drawn boxes which is closer to the realistic setting.

CUHK03 [1] contains 14,097 images of 1,467 identities collected in the CUHK campus. Each identity is captured by two cameras and has 4.8 images in average for each view. The CUHK03 dataset contains two kinds of bounding boxes. We evaluate our model on the bounding boxes detected by DPM, which is closer to the realistic setting. There are 9.6 images per identity in the training set. We report the averaged single-shot results after training/testing 15 times on the datasets.

MARS [3] (Motion Analysis and Recognition Set) is an expanded version of the Market1501 dataset. This is a large video-based person ReID dataset. Since all bounding boxes and trajectories are automatically generated. MARS has a total of 20,478 tracklets, including 1,261 identities for 6 camera views.

B. Experiment Results

We evaluate our method with rank-1, rank-5, rank-10 accuracy and mean average precision (mAP). We compared our method with the representative person ReID methods on several benchmark datasets. The results are shown in Table 1, Table 2 and Table 3 separately.

As shown in Table 1 and Table 2, BoW + KISSME [21] is the baseline experiment. The method of pyramid loss obtains 86.26% and 85.95% rank-1 accuracy by ResNet-50, respectively on Market1501 and CUHK03. The pyramid which employs hard sample pairs mining strategies further improves rank-1 accuracy on Market1501 and CUHK03. The improvement can be observed on ResNet-50 network architecture. Similarly, we observe 86.32% and 86.46% baseline rank-1 accuracy on Market1501 and CUHK03 in single-shot setting. The Fig. 2 shows the comparison of different values on θ for angular loss for Market1501 dataset (r=1). As shown in Table 3, MARS [3] is the baseline experiment. The proposed pyramid loss with hard sample pairs mining consistently achieves better accuracy on most experimental datasets. The Fig. 3 shows the comparison of queries and top-5 retrievals among Triplet Loss (TL), Angular Loss (AL) and Pyramid Loss (PL). We show two examples for the Market1501 dataset. The retrieved person images adding red border are the ones that belong to the same class as the query. Table 4 shows the effect of the setting of two hyperparameters θ and δ on accuracy results. θ and δ are set by hyper parameters optimizer sklearn. These results show that our method can handle the network and improve their results.

TABLE 1 Comparison on Market1501 with single query.

Method	mAP	r=1	r=5	r=10
BoW + KISSME [21]	20.76	44.42	63.90	72.18
Multiregion CNN [29]	41.17	66.36	85.01	90.17
Past(ResNet-50) [7]	47.80	73.90	87.68	91.54
Spindle Net [30]	-	76.90	91.50	94.60
Triplet(ResNet-50) [15]	54.80	75.90	89.60	-
Unlabeled [31]	56.23	78.06	-	-
TOMM(ResNet-50) [4]	59.87	79.51	90.91	94.09
Quad(ResNet-50) [16]	61.10	80.00	91.80	-
Pose-driven [32]	63.41	84.14	92.73	94.92
Context-aware [13]	57.53	80.31	-	-
Deep Joint [14]	65.50	85.10	-	-
Defense(ResNet-50) [10]	69.14	84.92	94.21	-
MSML(ResNet-50) [2]	69.60	85.20	93.70	-
Angular(GoogLeNet)	69.73	85.53	93.09	97.29
Pyramid(GoogLeNet)	**69.91**	**85.60**	**93.23**	**97.67**
Pyramid′ (GoogLeNet)	**70.36**	**85.73**	**93.39**	**97.99**
Pyramid(ResNet-50)	**70.83**	**86.26**	**93.85**	**98.25**
Pyramid′ (ResNet-50)	**71.01**	**86.32**	**93.87**	**98.77**

TABLE 2 Comparison on CUHK03 with single query.

Method	mAP	r=1	r=5	r=10
KISSME [1]	-	19.90	49.30	64.70
BoW + KISSME [21]	-	24.30	-	-
SI-CI [27]	-	52.20	84.30	94.80
Ensembles [25]	-	62.10	89.10	94.30
DeepLDA [26]	-	63.23	89.95	92.73
Triplet(ResNet-50) [15]	-	73.00	92.00	96.00
Joint [33]	-	77.50	-	-
Quad(ResNet-50) [16]	-	79.10	95.30	97.90
TOMM(ResNet-50) [4]	86.40	83.40	97.10	98.70
Context-aware [13]	-	74.21	94.33	97.54
Defense(ResNet-50) [10]	-	79.50	95.00	98.00

MSML(ResNet-50) [2]	-	84.00	96.70	98.20
Unlabeled [31]	87.40	84.60	97.60	98.90
Angular(GoogLeNet)	87.59	85.01	97.97	98.87
Pyramid(GoogLeNet)	**87.67**	**85.83**	**98.76**	**99.01**
Pyramid′(GoogLeNet)	**87.71**	**86.32**	**98.99**	**99.12**
Pyramid(ResNet-50)	**87.85**	**85.95**	**98.39**	**99.13**
Pyramid′(ResNet-50)	**87.95**	**86.46**	**99.04**	**99.21**

TABLE 3 Comparison on MARS with single query.

Method	mAP	r=1	r=5	r=10
MARS [3]	42.40	60.00	77.90	87.90
Context-aware [13]	56.05	71.77	86.57	-
Triplet(ResNet-50) [15]	62.10	76.10	89.60	-
Quad(ResNet-50) [16]	62.10	74.90	88.90	-
Defense(ResNet-50) [10]	71.30	82.50	92.10	-
MSML(ResNet-50) [2]	72.00	83.00	92.60	-
Angular(GoogLeNet)	72.60	83.95	92.64	93.66
Pyramid(GoogLeNet)	**72.67**	**85.53**	**92.93**	**94.29**
Pyramid′(GoogLeNet)	**72.72**	**85.92**	**91.42**	**98.01**
Pyramid(ResNet-50)	**72.93**	**85.90**	**91.56**	**97.85**
Pyramid′(ResNet-50)	**73.01**	**86.13**	**92.03**	**98.71**

TABLE 4 Comparison of different values on θ and δ for pyramid (GoogLeNet) on Market1501 dataset.

pyramid-θ	pyramid-δ	mAP	r=1	r=5	r=10
θ = 44.52°	δ = 26.35°	68.67	83.04	89.44	95.49
θ = 37.67°	δ = 23.49°	69.80	84.96	90.16	96.88
θ = 28.54°	**δ = 20.27°**	**69.91**	**85.60**	**91.23**	**97.67**
θ = 21.36°	δ = 17.69°	69.92	84.35	90.91	96.57
θ = 17.82°	δ = 16.05°	69.84	84.14	89.76	96.21
θ = 39.16°	δ = 25.38°	68.02	82.86	88.46	95.10
θ = 33.45°	δ = 22.27°	69.35	84.13	89.73	96.04
θ = 29.39°	δ = 19.62°	69.89	85.01	90.92	97.24
θ = 24.32°	δ = 17.19°	69.11	84.13	89.89	96.04
θ = 21.31°	δ = 15.67°	68.98	83.94	89.15	95.95

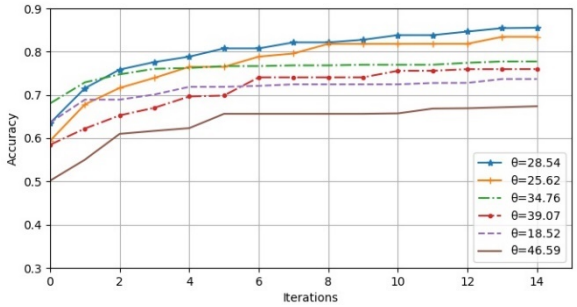

Fig. 2. Comparison of different values on θ for angular loss for Market1501 dataset (r=1).

Fig. 3. Comparison of queries and top-5 retrievals among Triplet Loss (TL), Angular Loss (AL) and Pyramid Loss (PL). We show two examples for the Market1501 dataset. The retrieved person images adding red border are the ones that belong to the same class as the query.

V. Conclusion

In this work, we combine pyramid loss with hard sample pairs mining applied for person ReID. Depending on the nature of the triangle, we use angular loss to constrain the angle of the negative point of triplet, which can be used to combat scale and feature changes. The pyramid loss introduces another negative sample in angular loss producing a stronger push between positive and negative pairs, so as to improve the robustness of the model. Our method takes both distance and the angular relation of samples into consideration. We use GoogLeNet and ResNet-50 as base model to do some contrast experiments with different metric learning losses. On several benchmark datasets, including Market1501 [21], CUHK03 [1] and MARS [3], the results show that our approach improve the performance of person ReID on testing datasets.

REFERENCES

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

[2] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," *arXiv preprint arXiv:1710.00478*, 2017.

[3] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person reidentification," in *European Conference on Computer Vision*, 2016, pp. 868–884.

[4] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2017.

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. 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*, 2016, pp. 1335–1344.

[6] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person reidentification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.

[7] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person reidentification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[8] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[9] Y. Yang, L. Wen, S. Lyu, and S. Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification." in *Association for the Advance of Artificial Intelligence*, vol. 1, 2017, p. 2.

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

[11] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, "Deep metric learning via facility location," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5382–5390.

[12] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification." in *Association for the Advance of Artificial Intelligence*, vol. 1, no. 2, 2017, p. 3.

[13] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[14] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2017.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

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

[17] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.

[18] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *IEEE International Conference on Computer Vision*, 2017, pp. 2593–2601.

[19] B. Harwood, V. K. BG, G. Carneiro, I. Reid, and T. Drummond, "Smart mining for deep metric learning," in *IEEE International Conference on Computer Vision*, 2017, pp. 2821– 2829.

[20] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce, "A tensor-based algorithm for high-order graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2383–2395, 2011.

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

[22] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2288–2295.

[23] 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.

[24] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[25] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[26] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person reidentification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[27] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.

[28] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a nextgeneration open source framework for deep learning," in *Advances in Neural Information Processing Systems*, vol. 5, 2015.

[29] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Advanced Video and Signal Based Surveillance*. IEEE, 2017, pp. 1–6.

[30] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[31] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *IEEE International Conference on Computer Vision*, 2017.

[32] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *IEEE International Conference on Computer Vision*, 2017, pp. 3980–3989.

[33] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3376–3385.

Yuanyuan Wang was born on Oct. 25, 1981**.** She is currently pursuing the Ph.D degree in Computer Science and Technology from Hohai University of China. She received the M.S. degree in Computer Technology from Nanjing University of Science and Technology, in 2010. She is currently a lecturer with the College of Computer and Software Engineering of Huaiyin Institute of Technology. Her current research focuses on computer vision.

Zhijian Wang was born on Jul. 1, 1958**.** He received the M.S. and Ph.D. degree in Computer Science from Nanjing University, China. He is currently a Professor in the College of Computer and Information, Hohai University, China. His research interests include machine learning and computer application.

Mingxin Jiang was born on May. 10, 1979**.** She received a Ph.D. degree in Signal and Information Processing, Dalian University of Technology, China, in 2013. She was a post-doctoral researcher with the Department of Electrical Engineering in Dalian University of Technology from 2013 to 2015. She is currently an associate professor in College of Electronic Information Engineering at Huaiyin Institute of Technology. Her research interests include multi-object tracking, video content analysis and vision sensors for robotics.