A Coarse-to-fine Pyramidal Model for Person Re-identification via Multi-Loss Dynamic Training

Feng Zheng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang

YouTu Lab, Tencent

Shanghai, China

f.zheng2016@gmail.com, {winfredsun, sevjiang, scorpioguo, quentinyu, garyhuang}@tencent.com

Abstract

Most existing Re-IDentification (Re-ID) methods are highly dependent on precise bounding boxes that enable images to be aligned with each other. However, due to the inevitable challenging scenarios, current detection models often output inaccurate bounding boxes yet, which inevitably worsen the performance of these Re-ID algorithms. In this paper, to relax the requirement, we propose a novel coarse-to-fine pyramid model that not only incorporates local and global information, but also integrates the gradual cues between them. The pyramid model is able to match the cues at different scales and then search for the correct image of the same identity even when the image pair are not aligned. In addition, in order to learn discriminative identity representation, we explore a dynamic training scheme to seamlessly unify two losses and extract appropriate shared information between them. Experimental results clearly demonstrate that the proposed method achieves the state-of-the-art results on three datasets and it is worth noting that our approach exceeds the current best method by 9.5% on the most challenging dataset CUHK03.

1. Introduction

Person Re-IDentification (Re-ID) aims to associate the images of the same person captured at different physical sites, facilitating cross-camera tracking techniques used in vision-based smart retail and security surveillance. In general, person Re-ID is considered to be the next high-level task after a pedestrian detection system, so the basic assumption of Re-ID is that the detection model can provide precise and highly-aligned bounding boxes. Despite the recent great progress, there are limited room for the performance improvement of existing methods due to the potential problems with part-based models and the difficulties in training.

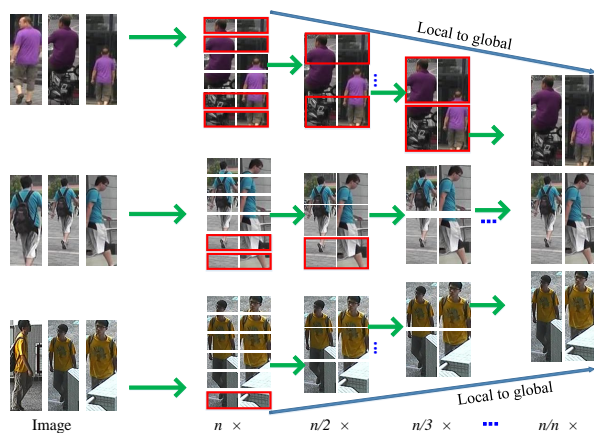**Drawbacks of part-based models:** As it is well-known,

Figure 1. The examples of part-based matching at different scales, when bounding boxes are not aligned or parts of human body have been occluded. The red bounding box indicates most cues in the two parts are varied. We can see that, in a finely partitioned way, a handful of horizontal stripes (left) cannot be well-matched due to different cues, while those stripes (right) in a more global view have more similar cues.

part-based models can generally achieve promising performance in many computer vision tasks, because these models are potentially robust to some unavoidable challenges such as occlusion and partial variations. Actually, the performance of person Re-ID in the real-world applications is severely affected by these challenges. Thus, the recent proposed Part-based Convolutional Baseline (PCB) can achieve the state-of-the-art results. PCB is simple but very effective and even can outperform other learned part models. Nevertheless, in PCB, directly partitioning the feature map of backbone networks into a fixed number of parts strictly limits the capacities of further improving the performance. It has at least two major drawbacks, but not limited to: 1) The overall performance seriously depends on that a powerful and robust pedestrian detection model outputs a precise bounding box otherwise the parts cannot be

well-aligned. However, in most cases of challenging scenes, current detection models are insufficient to do that. 2) The global information, which is also a very significant cue for recognition and identification, is completely ignored in this model whilst global features are normally robust to the subtle view changes and internal variations. Several examples are illustrated in Fg. 1 to show that the parts of diverse scales are equivalently important for matching.

Difficulties of multi-loss training: Recent studies demonstrate that multi-task learning has the capabilities to achieve advanced performance by extracting appropriate shared information between tasks. Without loss of generality, the terms "loss" and "task" will be used alternatively. In fact, many existing Re-ID methods also benefit from the multi-loss scheme to improve the performance. Generally, most multi-task methods choose to weight the losses using balancing parameters which are fixed during entire training process. 1) The performance highly relies on an appropriate parameter but choosing an appropriate parameter is undoubtedly a labor-intensive and tricky work. 2) The difficulty of different tasks actually change when the models are updated gradually, resulting in really varied appropriate parameters for different iterations. 3) More importantly, sampling strategies for different losses are generally diverse due to the specific considerations. For example, hard sample sampling for triplet loss would suppress the role of another task of identification loss.

To address above problems, in this paper, we specifically propose a novel coarse-to-fine pyramidal model based on the feature map extracted by a backbone network for person re-identification. First, the pyramid is actually a set of 3-dimensional sub-maps with a specific coarse-to-fine architecture, in which each member captures the discriminative information of different spatial scales. Then, a convolutional layer is used to reduce the dimension of features for each separated branch in the pyramid. Third, for each branch, the identification loss of a softmax function is independently applied on a fully connected layer which considers the features as the input. Furthermore, the features of all branches will be concatenated to form the identity representation, for which a triplet loss is defined to learn more discriminative features. To smoothly integrate the two losses, a dynamic training scheme with two sampling strategies is explored to optimize the parameters of deep neural networks. Finally, the learned identity representation will be used for person image matching, retrieval and re-identification.

In summary, the contribution of this paper is three-fold: 1) To relax the assumption of requiring a strong detection model, we propose a novel coarse-to-fine pyramid model that not only incorporates local and global information, but also integrates the gradual cues between them. 2) To maximally take advantage of different losses, we explore a dynamic training scheme to seamlessly unify two losses and

extract appropriate shared information between them for learning discriminative identity representation. 3) The proposed method achieves the state-of-the-art results on the three datasets and most significantly, our approach exceeds the current best method by 9.5% on dataset CUHK03.

2. Related Work

Most existing Re-ID methods ether particularly consider the local parts of person images or mainly explore the global information. Some methods [13, 26] aware of that integrating the local and global features can improve the performance but the information between them is also ignored. We observe that those cues in the transition process are significant as well.

Part-based algorithms: By performing bilinear pooling in a more local way, an embedding can be learned, in which each pooling is confined to a predefined region [25]. Inspired by attention models, in [16, 14, 21], the attention-based deep neural networks are proposed to capture multiple attentions and select multi-scale attentive features. Similarly, Zhao et al. [31] explore a deep neural network method to jointly model body part extraction and representation computation, and learn model parameters. Based on a L2 distance, [13] formulates a method for joint learning of local and global feature selection losses particularly designed person Re-ID. In [22], a pose-driven deep convolutional model, which leverages the human part cues to alleviate the pose variations, is designed to learn feature extraction and matching models. Furthermore, both the fine and coarse pose information of the person [19] are incorporated to learn a discriminative embedding. A part loss is proposed in [29], which automatically detects human body parts and computes the person classification loss on each part separately. Chen et al. [4] develop a CNN-based appearance model to jointly learn scale-specific features and maximize multiscale feature fusion selections. Several part regions are first detected and then deep neural networks are designed for representation learning on both the local and global regions [27]. Part-based Convolutional Baseline (PCB) [24] outputs a convolutional descriptor consisting of several part-level features and then a refined part pooling method is used to re-assign outliers in the parts. Based on PCB, Multiple Granularity Network (MGN) [26] explores a branch for global features and two branches for local representations for person re-identificaiton.

Non part-based methods: Recently, a completely synthetic dataset [1] and some adversarially occluded samples [8] are constructed to train the re-identification model. In [23], singular vector decomposition is used to iteratively integrate the orthogonality constraint in CNN training for image retrieval. A pedestrian alignment network [37] is built to learn discriminative embedding and pedestrian alignment without extra annotations. Geng et al. [6] propose a number

of deep transfer learning models to address the data sparsity problem and transfer knowledge from auxiliary datasets. [7] shows that a plain CNN with a triplet loss can outperform most recent published methods by a large margin. Learning binary representation for fast matching [33, 32] is also a promising direction of object re-identification. In [20], A group-shuffling random walk network is proposed to refine the probe-to-gallery affinities based on gallery-to-gallery affinities. The "local similarity" metrics for image pairs are learned with considering dependencies from all the images in a group, forming "group similarities" in [3].

Multi-task learning: Self-paced learning [10] and focal loss [15] both train models by diversely weighting the samples in different learning stages. Inspired by this, in [11], a task-oriented regularizer is designed to jointly prioritize both tasks and instances. The multiple loss functions [9] are weighted by considering the uncertainty of tasks in both classification and regression settings. Moreover, a routing network consisting of two components is introduced to dynamically select different functions in response to the input [18]. While, Chen et al. [5] propose a gradient normalization algorithm that automatically balances training in deep multitask models by dynamically tuning gradient magnitudes. We learn the embedding for person Re-ID by simultaneously minimizing a list-wise metric loss and a classification loss with two types of sampling strategies.

3. The Proposed Method

3.1. Coarse-to-fine Pyramidal Model

In this section, we propose a novel coarse-to-fine pyramidal model which can moderately relax the requirement of detection model and smoothly incorporate the global information, simultaneously. It is worth noting that, in our approach, not only local and global information is integrated but the gradual transition process between them is also incorporated.

3.1.1 Pyramidal Branches

Given a set of images $\mathbf{X} = \{I_1, \cdots, I_N\}$ containing persons captured by cameras in surveillance systems where N is the number of images, the task of person ReIDentification (Re-ID) is to associate the images of the same person at different times and locations. Our model is built on a feature map M extracted by a backbone network $\mathcal{BN}$. Thus, we have a 3-dimensional tensor $M = \mathcal{BN}(I)$ of the size $C \times H \times W$, where C is the number of encoded channels and W and H are the spatial width and height of the tensor, respectively.

First, we divide the feature map into n number of parts according to the spatial height axis and thus each basic part has the size of $C \times (H/n) \times W$. Suppose that H can be divisible by n. Thus, our pyramidal model is constructed according to the following rules: 1) In the bottom level ($l = 1$) of the pyramid, there are n number of branches in which one corresponds to a basic part. 2) The branches in higher level has one more adjacent basic part than that of previous lower level. 3) The sliding step for all levels is set to one. It means the number of branches in the current level is just one less than that of previous level. 4) In the top level ($l = n$) of the pyramid, there is only one branch which is just the original feature map M. Therefore, we assume that $\mathbf{P}\{l, k\}$ is the kth sub-map in the lth level of the pyramid model $\mathbf{P}$ defined as:

$$
\begin{aligned}
\mathbf{P} = \quad &\{M(1:C, st:ed, 1:W): \qquad (1)\\
&st = (k-1) * H/n + 1,\\
&ed = (k-1) * H/n + l * H/n,\\
&l = 1, \cdots, n, k = 1, \cdots, n-l+1\},
\end{aligned}
$$

where $1:C$ means that all elements from index 1 to index C are selected.

Obviously, $\mathbf{P}$ is a set of 3-dimensional sub-maps with a specific coarse-to-fine architecture, in which each member captures the discriminative information of different spatial scales. Moreover, the pyramidal model contains both the global feature map $M \in \mathbf{P}$ and the part-based model: PCB. While, it is easy to know that there are totally $\sum_{l=1}^{n} l$ number of components in $\mathbf{P}$ and the lth level has $n-l+1$ number of components, where the level index is in a fine-to-coarse fashion. The details of the proposed architecture are shown in Fig. 2.

3.1.2 Basic Operations

For each branch $\mathbf{P}\{l, k\}$ in pyramid $\mathbf{P}$, first, a global maximum pooling (GMP) and a global average pooling (GAP) is separately executed to capture the statistical properties of different channels in the sub-maps. Then, the two statistical variables are added to form a vector with the same size of encoded channels. Third, a convolutional layer followed by a batch normalization and a ReLU activation is explored to reduce the dimension and produce a feature vector $x(l, k) \in R^D$ for the re-identification task. Simply, we denote $x(l, k) = \mathcal{BO}(\mathbf{P}\{k, l\})$. Fourth, to make the feature vector capable of sufficient discriminativity, a softmax based identification loss will be used for a fully connected layer which considers the feature map as the input. At the same time, a triplet loss will be imposed on a vector $x = (x(1, 1)^T, \cdots, x(n, 1)^T)^T$ which concatenates all the feature vectors of different branches in the pyramid $\mathbf{P}$. The basic operations for each branch will be executed independently with respect to different components in the pyramid. Finally, all the parameters will be learned by simultaneously minimizing the two losses in an alternate way.
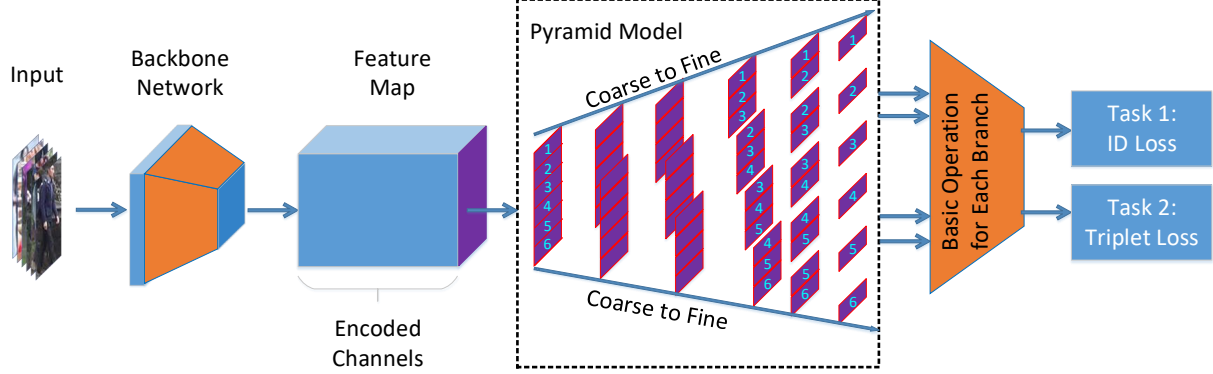
Figure 2. The architecture of our proposed pyramidal model for person re-identification. For better layout, only the spatial profile of the member branch in the pyramid, which is originally 3-dimensional tensors, is shown. We assume that the original feature map is divided into 6 basic sub-maps while other number of sub-maps can be used as well. The branch always consists of several consecutive basic sub-maps and the basic operation for each branch will be given in Fig. 3.
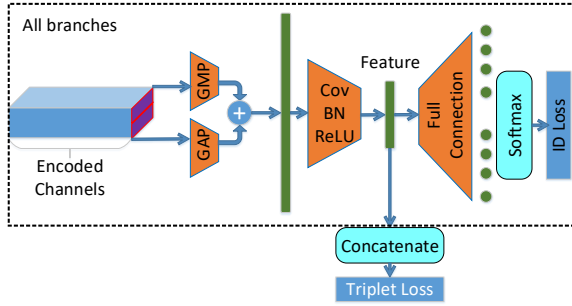

Figure 3. The illustration of basic operations for a branch consisting of two consecutive basic sub-maps, which include a global maximum pooling, a global average pooling, a convolutional filter, a batch normalization, a ReLu activation and a linear fully connected layer. These operations for different branches will be executed independently and the features of all branches will be finally concatenated for the triplet loss.

An example of a branch consisting of two consecutive basic parts is illustrated in Fig. 3.

If assume that $f(\cdot)$ refers to all the operations in the embedding including $\mathcal{BN}$, $\mathbf{P}$ and $\mathcal{BO}$s, thus we can denote the feature vector as $x = f(I)$ simply. In the inference stage, the re-identification task will be achieved by ranking the distances $\{d(x, x_i) : x = f(I), I_i \in \mathbf{G}\}$ between a query I and a gallery $\mathbf{G}$.

3.2. Multi-Loss Dynamic Training

Recent studies demonstrate that multi-task learning has the capabilities to achieve advanced performance by extracting appropriate shared information between tasks. The potential reason is that multiple tasks can benefit from each other by exploring the relatedness, leading to boosted generalization performance.

3.2.1 Two Tasks

To learn the discriminative features, we adopt two related tasks but emphasizing different aspects to learn the parameters of the embedding f, including an identification loss and a triplet loss. The first one is point-wise classification loss while the second one is the list-wise metric learning.

Identification loss: Generally, the identification loss is the same as the classification loss defined as:

$$
\begin{aligned}
L^{id} &= \frac{1}{N_{id}} \sum_i \sum_{k,l} \mathcal{S}((W_c^{kl})^T x_i(l, k)) \quad (2) \\
&= \frac{1}{N_{id}} \sum_i \sum_{k,l} -\log \frac{(W_c^{kl})^T x_i(l, k)}{\sum_j (W_j^{kl})^T x_i(l, k)},
\end{aligned}
$$

where N_{id} is the number of used images, c denotes the corresponding identity of the input image I_i, $\mathcal{S}$ is the softmax function and W_c^{kl} is the weight matrix of the fully connected layer for cth identity in the (l, k) branch.

Triplet loss: Given a triplet of samples (I, I_p, I_n) where I and I_p are of the same identity whilst I and I_n are the images for different identities, the aim of embedding is to learn a new feature space in which the distance between the sample pair I and I_p will be smaller than that between the pair I and I_p. Intuitively, a triplet loss can be defined as:

$$
L^{tp} = \frac{1}{N_{tp}} \sum_{I, I_p, I_n} [d(f(I), f(I_p)) \quad (3)
$$
$$
-d(f(I), f(I_n)) + \delta]_+, \quad (4)
$$

where δ is a margin hyper-parameter to control the distance differences, N_{tp} is the number of available triplets and

$[\cdot]_+ = \max(\cdot, 0)$ is the hinge loss.

3.2.2 Dynamic Training

The above two tasks are not novel and popularly used in various applications. While, how to integrate them is still an open problem.

Actually, from the general perspective, most multi-task methods normally weight the tasks using balancing parameters and treat some tasks as the regularization items. In the learning stage, the balancing parameters are fixed during entire training process. 1) The performance strictly lies in an appropriate parameter but choosing an appropriate parameter is undoubtedly a labor-intensive and tricky work. 2) The difficulty of different tasks actually changes when the models are updated gradually, resulting in really varied appropriate parameters for different iterations.

Furthermore, from the view of the re-identification task, to some extent, the two tasks are also conflicting when they are directly combined. On the one hand, effective triplets are rare, if the general random mini-batch sampling is used, making that the triplet loss contributes little in the learning procedure. This is because the number of identities is large but the number of images for each identity is small. On the other hand, to avoid the problem, we propose an ID-balanced sampling strategy to make sure triplets do exist in the mini-batches. However, this strategy suppresses the identification loss since fewer identities can be used in each mini-batch. Due to the sampling bias, it is possible that some images cannot be used all the time. Therefore, directly arithmetic weighting the losses would be very simple but obviously result in many difficulties in optimization.

Sampling: To solve the problem, alternatively, we choose to alternately minimize the two losses incorporating two sampling methods accordingly: random sampling and ID-balanced hard triplet sampling. Random sampling is easy to be implemented while ID-balanced hard triplet sampling is implemented according to the following steps. To build the effective triplets, we randomly select 8 number of identities for each mini-batch, in which 8 images of each identity are randomly chosen. Hence, this strategy definitely enables to use the hard positive/negative mining based on the largest intra-class (identity) distance and the smallest inter-class distance. However, the samples for different identities are unbalanced and those whose number of images is less than 8 will never be used.

Dynamic weighting: For each loss, we define a performance measure to estimate the likelihood of a loss reduction. Suppose L_τ^t be the average loss in the current training iteration τ for the task $t \in \{id, tp\}$. Thus, we can calculate k_τ^t to be an exponential moving average according to:

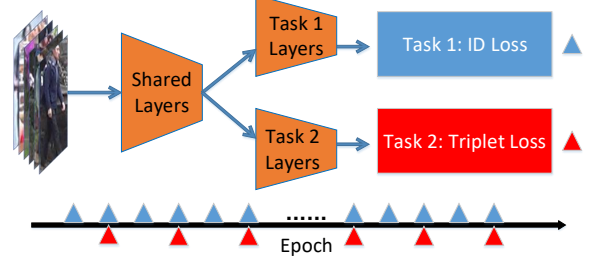$$k_\tau^t = \alpha L_\tau^t + (1-\alpha)k_{\tau-1}^t, \tag{5}$$

Figure 4. Dynamic training for two related tasks with two types of sampling strategy.

where $\alpha \in [0, 1]$ is a discount factor and $k_{-1}^t = L_0^t$. Based on the quantity k_τ^t, we defined a probability to describe the likelihood of a loss reduction as:

$$p_\tau^t = \frac{\min\{k_\tau^t, k_{\tau-1}^t\}}{k_{\tau-1}^t}. \tag{6}$$

In case of loss increasing occasionally, the function $\min$ is used to normalize p_τ^t to be 1. Obviously, $p_\tau^t = 1$ means the current optimization step didn't reduce the loss yet. The larger the value, the greater the probability that the optimization of the task t steps into a local minimum. Similar to the Focal Loss which down-weights easier samples and concentrates on hard samples, we define a measure $(FL(\cdot))$ to weight the losses:

$$FL(p_\tau^t, \gamma) = -(1 - p_\tau^t)^\gamma \log(p_\tau^t) \tag{7}$$

where γ is used to control the focusing intensity. $FL(p_\tau^t, \gamma)$ is designed to weight the tasks and choose the desire loss to be optimized. Actually, the overall objective function can be rewritten as:

$$L = \sum_{t \in \{id, tp\}} FL(p_\tau^t, \gamma)L_\tau^t. \tag{8}$$

Due to the different sampling strategies, we optimize the ID loss in Eq. 2 with randomly selected mini-batches, when $FL(p_\tau^{id}, \gamma)$ dominates the two tasks $(FL(p_\tau^{tp}, \gamma)/FL(p_\tau^{id}, \gamma) < \delta)$. Thus, we start our dynamic optimization system from simply minimizing the ID loss. Actually, $FL(p_\tau^{id}, \gamma)$ always dominates in the early optimization since each step can greatly reduce the ID loss. Moreover, because the model is currently in an immature status, all samples are equally difficult so that hard sampling-based triplet loss cannot play essential role for our optimization. This is similar to the scheme of self-paced (curriculum learning) leaning [10] in which easier samples are first trained and hard samples are considered later while here dynamically optimizing the two tasks plays the same role. In this case, the both losses $L_\tau^t : t \in \{id, tp\}$ in the objective Eq. 8 will be calculated.

When $FL(p_\tau^{tp}, \gamma)$ dominates in the optimization, the overall objective 8 considering both Eq. 2 and 3 will be directly optimized because ID-balanced hard triplet sampling will not influence the use of ID loss. This optimization successfully avoids the tortuous balancing-parameter tuning and seamlessly incorporates the ideas of both ID-balanced hard triplet sampling and curriculum learning to further improve the performance. The flowchart of alternate training is illustrated in Fg. 4 and details of training are given in Algorithm 1.

Algorithm 1   Multi-Loss Dynamic Training

Input: Dataset X, pretrained backbone network $\mathcal{BN}$ and hyper-parameters $(n, \alpha, \delta, \gamma, D)$.
Output: The embedding function $x = f(I)$.
Initialization:
Initiate the network parameters except for the backbone.
Set $p_\tau^{id} = 0$ and $p_\tau^{tp} = 1$.
for $\tau = 1, \cdots, N_\tau$
 Calculate $FL(p_\tau^{tp}, \gamma)$ and $FL(p_\tau^{id}, \gamma)$.
 if $FL(p_\tau^{tp}, \gamma)/FL(p_\tau^{id}, \gamma) < \delta$
  Perform random mini-batch sampling.
  Forward though $\mathcal{BN}$ and construct the pyramid $\mathbf{P}$.
  Apply batch-normalization and calculate $x = f_\tau(I)$.
  Optimize the objective in Eq. 2.
 else
  Perform ID-balanced hard triplet sampling.
  Forward though $\mathcal{BN}$ and construct the pyramid $\mathbf{P}$.
  Apply batch-normalization and calculate $x = f_\tau(I)$.
  Optimize the objective in Eq. 8.
 end if
 Backpropagate the gradients and update the parameters.
end for

---

4. Experiment

To validate the performance of the proposed method, we test it on three popular used person re-identification datasets: Market-1501 [34], DukeMTMC-reID [17] and CUHK03 [12].

4.1. Experimental Setting

Implementation details: All images are resized into a resolution of 384×128 which is the same as that of PCB. The ResNet model with the pretrained parameters on ImageNet is considered as the backbone network in our system. For the feature map, the number of encoded channels is 2048 while the feature will be reduced to a 128-dimensional vector using a convolutoinal layer. We set the number of basic parts to 6 so there are 21 branches in the pyramid according to construction rules. The margin in the triplet loss is 1.4 in all our experiments. We select a mini-batch of 64 images for each iteration. Stochastic gradient descent (SGD) with two sampling strategies is used in our optimization, where the momentum and the weight decay factor is set to 0.9 and 0.0005, respectively. Totally, the proposed model will be

trained 120 epochs. As for the learning rate strategy, the initial learning rate is set to 0.01 while the learning rate will be dropped by half every 10 epochs from epoch 60 to epoch 90. While, for the dynamic training, we set the parameters $\delta = 0.16$, $\alpha = 0.25$ and $\gamma = 2$, according to the suggestions in [15]. All the experiments in this paper will follow the same setting.

Evaluation metrics: To compare the re-identification performance of the proposed method with the existing advanced methods, we adopt the Cumulative Matching Characteristics (CMC) at rank-1, rank-5 and rank-10, and mean Average Precision (*mAP*) on all the datasets. It is worth noting that all our results are obtained in a single-query setting and, more importantly, re-ranking algorithm is not used to improve the *mAP* in all the experiments for our proposed method.

4.2. Datasets

Market-1501: In this dataset, $32,668$ images of $1,501$ identities with annotated bounding boxes detected using the pedestrian detector of Deformable Part Model (DPM) are collected. View overlapping exists among different cameras, including 5 high-resolution cameras, and a low-resolution camera. Following the setting of PCB, we divide the dataset into a training set with $12,936$ images of 751 persons and a testing set of 750 persons containing $3,368$ query images and $19,732$ gallery images.

DukeMTMC-reID: Following the protocol [36] of the Market-1501 dataset, this dataset is a subset of the DukeMTMC dataset specifically collected for person re-identification. In this dataset, $1,404$ identities appears in more than two cameras while 408 (distractor) identities appears in only one camera. We divide the dataset into a training set of $16,522$ images with 702 identities and a testing set which consists of $2,228$ query images of the other 702 identities and $17,661$ gallery images of 702 identities plus 408 distractor identities.

CUHK03: We follow the new protocol [38] similar to that of Market-1501, which splits the CUHK03 dataset into training set of 767 identities and testing set of 700 identities. From each camera, one image is selected as the query for each identity and the rest of images are used to construct the gallery set. This dataset has two ways of annotating bounding box including labelled by human or detected by a detector. The labelled dataset includes $7,368$ training, $1,400$ query and $5,328$ gallery images while detected dataset consists of $7,365$ training, $1,400$ query and $5,332$ gallery images.

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

In this section, we compare the proposed method called "Pyramid-ours" with 26 state-of-the-art methods, most of which is proposed in the last year, on the three datasets

Method	mAP	rank 1	rank 5	rank 10
Pyramid-ours	**88.2**	**95.7**	**98.4**	**99.0**
MGN [26]	86.9	95.7	-	-
PCB+RPP [24]	81.6	93.8	97.5	98.5
PCB [24]	77.4	92.3	97.2	98.2
GLAD* [27]	73.9	89.9	-	-
MultiScale [4]	73.1	88.9	-	-
PartLoss [29]	69.3	88.2	-	-
PDC* [22]	63.4	84.4	92.7	94.9
MultiLoss [13]	64.4	83.9	-	-
PAR [31]	63.4	81.0	92.0	94.7
HydraPlus [16]	-	76.9	91.3	94.5
MultiRegion [25]	41.2	66.4	85.0	90.2
DML [30]	68.8	87.7	-	-
Triplet Loss [7]	69.1	84.9	94.2	-
Transfer [6]	65.5	83.7	-	-
PAN [37]	63.4	82.8	-	-
SVDNet [23]	62.1	82.3	92.3	95.2
SOMAnet [1]	47.9	73.9	-	-

Table 1. Comparison results (%) on Market-1501 dataset at 4 evaluation metrics: *mAP*, *rank 1*, *rank 5* and *rank 10* where the bold font denotes the best method. "*" denotes that the method needs auxiliary part labels. We divide other compared methods into two groups: the methods exploring part-based features and the methods extracting global features. Our proposed pyramid model achieves the best results on all the 4 evaluation metrics.

Method	mAP	rank 1
Pyramid-ours	**79.0**	**89.0**
MGN [26]	78.4	88.7
SVDNet [23]	56.8	76.7
AOS [8]	62.1	79.2
HA-CNN [14]	63.8	80.5
GSRW [20]	66.4	80.7
DuATM [21]	64.6	81.8
PCB+RPP [24]	69.2	83.3
PSE+ECN [19]	75.7	84.5
DNN-CRF [3]	69.5	84.9
GP-reid [28]	72.8	85.2

Table 2. Comparison results (%) on DukeMTMC-reID dataset. The results of our proposed pyramid model at *rank 5* and *rank 10* are 94.7% and 96.3%, respectively. Our proposed pyramid model also achieves the best results.

including Market-1501, DukeMTMC-reID and CUHK03. For the comparison of each dataset, we detail the following.

Market-1501: For this dataset, we divide the compared methods into two groups: including part-based and global methods, and the comparisons are given in Table 1. The results clearly show that local-based methods generally get better evaluation scores than that of these methods extracting global features only. The PCB is a convolutional base-

Method	Labelled		Detected	
	mAP	rank 1	mAP	rank 1
Pyramid-ours	**76.9**	**78.9**	**74.8**	**78.9**
MGN [26]	67.4	68.0	66.0	68.0
PCB+RPP [24]	-	-	57.5	63.7
MLFN [2]	49.2	54.7	47.8	52.8
HA-CNN [14]	41.0	44.4	38.6	41.7
SVDNet [23]	37.8	40.9	37.3	41.5
PAN [37]	35.0	36.9	34	36.3
IDE [35]	21.0	22.2	19.7	21.3

Table 3. Comparison results (%) on CUHK03 dataset using the new protocol [38]. For the labelled set, the results of our model at *rank 5* and *rank 10* are 91.0% and 94.4%, respectively, while they are 90.7% and 94.5% for the detected dataset. This data set will be the most difficult set by comparing the average performance of all methods. The proposed pyramid model outperforms all other state-of-the-art methods with great advantages.

line that motivates our approach, but we have improved performance by 10.8% and 3.4% on metrics *mAP* and *rank 1*, respectively. MGN is a method considering multiple branches as well but it ignores the gradual cues between global and local information. Our method achieves the same result with MGN on metric *rank 1* but exceeds it 1.3% on metric *mAP*. In comparison, the performances of other algorithms are similar with each other on metric *rank 10* but all of them are much worse on metric *mAP* and *rank 1* than that of ours.

DukeMTMC-reID: From Table 2, we can see that our method also achieves the best results on this dataset at both metrics *mAP* and *rank 1*. Among the compared methods, MGN is the closest method to our method score, but still below 0.6% *mAP* score. PSE+ECN which is a method using a pose-sensitive embedding and the re-ranking procedure also performs worse than ours (75.7% vs. 79.0%). Similar to the comparison on Market-1501 dataset, our pyramid model exceeds PCB+RPP 9.8% and 5.7% at metrics *mAP* and *rank 1*, respectively. We provide the achievements of our method at metrics *rank 5* (94 : 7%) and *rank 10* (96 : 3%) for comparison in the future.

CUHK03: This dataset is the most challenging dataset under the new protocol and the bounding boxes are annotated using two ways. While, from Table 3, we can see that our proposed approach has achieved the most outstanding results for these two annotation ways. On this datasst, the pyramid model outperforms all other methods at least 8.8% and 10.9%, respectively.

Furthermore, on Market-1501 dataset, we compare the our model with PCB using the same sampling strategy and some retrieved examples are shown in Fg. 5. We can see that the PCB cannot respond well to the challenge of inaccurate bounding boxes. Taking the first query as an example,

Query   Top 11 retrieved images

Figure 5. Examples of retrieved images by two methods: Pyramid-ours and PCB, in case of the imprecise detection. For each query, the first row images are returned by our proposed method while the second row images are searched by PCB. The green/red rectangles indicate that images have the same/different identities as the query and the blue eclipse denotes the similar contexts in the images.

our model is able to find three images of the same identity in the top 11 results whilst PCB could not search anyone. While, from the second query, we can see that the lower-body parts (blue eclipse) of retrieved images match to the upper-body part of query, due to the imprecise detection.

In summary, our proposed pyramid model using the novel multi-loss dynamic training can always be superior to all other existing advanced methods, no matter which evaluation metric is used. Through the comparative experiment on the three datasets, it is easy to know that CUHK03 dataset with the new protocol would be the most challenging one because all methods make worse results on it. However, our method can consistently outperform all other algorithms by a large margin. Therefore, we can conclude that our method particularly specializes in challenging problems.

4.4. Component Analysis

To further investigate the contribution of every component in the pyramid model, we conduct comprehensive ablation studies on the performance of different sub-models. The comparison results at the metrics: *mAP*, *rank 1*, *rank 5* and *rank 10* are shown in Table 4 and each result is obtained with only one setting changed and the rest being the same as the default value.

First, we merely use the part of branches in the pyramid to test the function. In the term "Pyramid-000001", the left

Model	mAP	Rank 1	Rank 5	Rank 10
Pyramid-000001	82.1	92.8	97.3	98.2
Pyramid-100000	84.9	93.9	97.6	98.5
Pyramid-001111	86.7	94.8	98.4	98.8
Pyramid-110011	87.2	95.0	98.1	98.8
Pyramid-111100	87.5	94.8	98.3	98.9
Feature-64	86.9	94.5	97.8	98.6
Feature-256	87.8	95.3	98.2	98.9
No triplet loss	86.5	93.8	97.5	98.4
Pyramid-ours	**88.2**	**95.7**	**98.4**	**99.0**
PCB+RPP [24]	81.6	93.8	97.5	98.5

Table 4. Results (%) of sub-models on Market-1501 dataset. In the term "Pyramid-000001", '0' means the corresponding level of pyramid is not used while '1' means that is used. "Feature-64" denotes the dimension of features for each branch is set 64. "No triplet loss" refers to that only identification loss is optimized.

number denotes whether the branches in low level are used or not while the rightest number is for the global branch. For example, "Pyramid-000001" means only the global branch in the highest level of the pyramid is used. From this table, we can obtain: 1) The local branches in the lower levels play more important roles than that of the global branches. (84.9% vs. 82.1%) 2) The more branches we use, the better the performance. 3) The only global branch plus the proposed dynamic training strategy can achieve better results than that of PCA+RPP. It clearly shows the dynamic training strategy is able to potentially improve the capacities of the model.

Second, the features of different dimensions are also analyzed. Compared to the default dimension 128, the features of dimension 64 and 256 both achieve worse results. It shows that the redundant information plays negative influence to the performance while too short feature cannot provide sufficient discriminative cues. In summary, the performance is relatively ($\leq 1.3\%$) stable with the change of the feature dimension. While, in the case of resource-limited application, 64-dimensional feature is a more acceptable choice.

Finally, we fix the dynamic balance parameter to 0 and alternately execute the two sampling strategies to train the identification loss. It means that the triplet loss will never be used in this experiment. In one step, mini-batch is selected using random sampling while ID-balance hard sampling is adopted in the next step. We can see that the overall performance is a little bit lower than that of the default setting of our proposed model but still much higher than that of PCB-RPP. It demonstrates the new pyramid model and the dynamic sampling strategy contribute most for the performance improvement.

5. Conclusion

In this paper, we construct a coarse-to-fine pyramid model for person re-identification via a novel dynamic training scheme. Our model relaxes the requirement of detection models and thus achieves advanced results on the three datasets. Specially, our model outperforms the existing best method by a large margin on dataset CUHK03 which is the most challenging datasets under the new protocol. It is worth noting that all our results are achieved in a single-query setting and re-ranking algorithm is not used. In the future, it will be interesting to jointly learn the detection and re-identification models in an integrated training framework. The two tasks are highly related and Re-ID models can be improved by means of attention maps in detection models. Moreover, the features of middle layers in the backbone network can be incorporated into the proposed pyramid model as well to further improve the Re-ID performance.

References

[1] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *arXiv:1701.03153v1*, 2017. 2, 7

[2] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 7

[3] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, 2018. 3, 7

[4] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV Workshop*, 2017. 2, 7

[5] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 3

[6] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv:1611.05244v2*, 2016. 2, 7

[7] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737v4*, 2017. 3, 7

[8] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *CVPR*, 2018. 2, 7

[9] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3

[10] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 3, 5

[11] C. Li, J. Yan, F. Wei, W. Dong, Q. Liu, and H. Zha. Self-paced multi-task learning. In *AAAI*, 2017. 3

[12] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 6

[13] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017. 2, 7

[14] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2, 7

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 3, 6

[16] X. Liu, H. Zhao, M. Tian, L. S. J. Shao, S. Y. J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 2, 7

[17] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop*, 2016. 6

[18] C. Rosenbaum, T. Klinger, and M. Riemer. Routing networks adaptive selection of non-linear functions for multi-task learning. In *ICLR*, 2018. 3

[19] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 2, 7

[20] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, 2018. 3, 7

[21] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 2, 7

[22] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 2, 7

[23] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 2, 7

[24] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018. 2, 7, 8

[25] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv:1512.05300v5*, 2015. 2, 7

[26] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. *arXiv:1804.01438v3*, 2018. 2, 7

[27] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017. 2, 7

[28] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou. Towards good practices on building effective cnn baseline model for person re-identification. *arXiv:1807.11042v1*, 2018. 7

[29] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv:1707.00798v2*, 2017. 2, 7

[30] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In *CVPR*, 2018. 7

[31] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 2, 7

[32] F. Zheng, X. Miao, and H. Huang. Fast vehicle identification via ranked semantic sampling based embedding. In *IJCAI*, 2018. 3

[33] F. Zheng and L. Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, 2016. 3

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 6

[35] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Pastand present and future. *arXiv:1610.02984v1*, 2016. 7

[36] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 6

[37] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE TCSVT*, 2018. 2, 7

[38] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6, 7