Robust Anchor Embedding for Unsupervised Video Person Re-Identification in the Wild

Mang Ye, Xiangyuan Lan, and Pong C. Yuen*

Department of Computer Science, Hong Kong Baptist University
{mangye, pcyuen}@comp.hkbu.edu.hk, xiangyuanlan@life.hkbu.edu.hk

Abstract. This paper addresses the scalability and robustness issues of estimating labels from imbalanced unlabeled data for unsupervised video-based person re-identification (re-ID). To achieve it, we propose a novel Robust AnChor Embedding (RACE) framework via deep feature representation learning for large-scale unsupervised video re-ID. Within this framework, anchor sequences representing different persons are firstly selected to formulate an anchor graph which also initializes the CNN model to get discriminative feature representations for later label estimation. To accurately estimate labels from unlabeled sequences with noisy frames, robust anchor embedding is introduced based on the regularized affine hull. Efficiency is ensured with kNN anchors embedding instead of the whole anchor set under manifold assumptions. After that, a robust and efficient top-k counts label prediction strategy is proposed to predict the labels of unlabeled image sequences. With the newly estimated labeled sequences, the unified anchor embedding framework enables the feature learning process to be further facilitated. Extensive experimental results on the large-scale dataset show that the proposed method outperforms existing unsupervised video re-ID methods.

Keywords: unsupervised person re-id · robust anchor embedding

1 Introduction

Person re-identification (re-ID) addresses the problem of searching specific persons across disjoint camera views [54,55]. Video-based re-ID has gained increasing attention in recent years due to its practicality [53], where the video sequences can be trivially obtained by effective pedestrian detection and tracking algorithms in practical applications [17,18]. Impressive progress has been reported with advanced deep learning methods [22, 34, 51]. However, the annotation difficulty limits the applicability of supervised methods in large-scale camera network, which motivates us to investigate an unsupervised solution with deep neural networks for video re-ID.

We follow the cross-camera label estimation approach to mine labels from the unlabeled image sequences [27, 29, 47], where existing supervised methods can be subsequently used to learn discriminative re-ID models. Thus this approach owns good flexibility and applicability [47]. However, most previous unsupervised

Fig. 1: Practical imbalanced unlabeled data in re-ID task for unsupervised training.

learning methods adopt the same training set as in supervised methods, where all persons appear in both cameras [16,27,29]. In practical unsupervised settings, only a small portion of persons appear in both cameras due to the imbalanced unlabeled data, *i.e.*, most persons only appear in one camera as illustrated in Fig. 1. As a result, large amount of false positives would be introduced and significant performance drop is inevitable. Therefore, their performances are somewhat over-estimated for practical wild settings. Moreover, most of these methods suffer from the scalability issues, thus cannot be applied to large-scale applications [9, 43, 53]. In this paper, we propose a scalable solution with deep neural networks for unsupervised video re-ID under wild settings.

The proposed method is designed on top of the *application-specific characteristics* existing in video re-ID. Specifically, we assume that several training video sequences representing different persons could be collected as *anchor sequences* for initialization. It's reasonable since persons appear in different non-overlapping cameras at the same time interval could be treated as different persons [27, 28]. Thus the anchor sequences could be easily collected without manually label efforts in practical applications. In addition, the image frames within each video sequence could be roughly assumed to represent the same person identity, which provides abundant weakly labeled images by treating each sequence as a class (person identity). Therefore, the easily collected anchor sequences provide abundant training samples to initialize the CNN model to obtain discriminative feature representations, which ensures the later label estimation performance from unlabeled sequences. With the learnt feature representations, we propose a novel Robust AnChor Embedding (RACE) framework to estimate labels from unlabelled sequences for large-scale unsupervised video re-ID.

RACE measures the underlying relationship between unlabelled sequences and anchor sequences with embedding process. To address the scalability and efficiency problem, we propose to perform anchors embedding with k-nearest neighbor instead of the whole anchor set under manifold assumptions. To handle the noisy frames within sequences and achieve a more robust label estimation under unsupervised settings, a novel constraint based on regularized affine full is incorporated to suppress the negative effects of noisy frames. With the learnt embedding weights, a robust and efficient top-k counts label prediction strategy is subsequently proposed to predict labels of the unlabeled image sequences. It does not require compulsory label assignment and reduces the false positives,

guaranteeing the robustness under wild settings. The main idea is that if two video sequences share the same label, they should be very similar under different measure dimensions. With the newly estimated labeled sequences, the feature learning process is further facilitated. Compared to existing unsupervised re-ID methods, the proposed method is robust and efficient for large-scale video re-ID in the wild. The main contributions are summarized as follows:

- We propose an unsupervised deep feature representation learning framework for large-scale video re-ID under wild settings. It is built on the *application-specific characteristics* existing in video re-ID task.
- We present a novel robust anchor embedding method to measure the underlying similarity relationship between the unlabelled sequences and anchors for better label estimation. The outlier-insensitive affine hull regularization is integrated to handle noisy frames in sequences to enhance the robustness.
- We introduce a robust and efficient top-k counts label prediction strategy to reduce false positives. It considers both visual and intrinsic similarity, achieving higher label estimation accuracy and slightly better efficiency.

2 Related Work

Unsupervised Re-ID. Several unsupervised re-ID methods have been developed in recent years. Unsupervised transfer learning approach learns re-ID models on the unlabelled target dataset with labelled source dataset [28,32]. Salience learning has also been investigated in early years [36,52]. Besides that, other attempts adopted dictionary learning with graph regularization constraints to learn shared feature representations [16,32]. Additionally, Yu *et al.* [49] introduced a cross-view asymmetric metric learning method to learn the distance measures. Meanwhile, Ye *et al.* [47] and Liu *et al.* [27] solved unsupervised video re-ID problem by estimating labels with hand-craft feature representations, and then adopted existing supervised learning methods to learn the re-ID models. Most of previous methods suffer from scalability issues, and it is hard for them to be applied to the large-scale applications [13, 15, 21].

Unsupervised Deep Learning. Unsupervised deep learning has been widely investigated in general image recognition tasks [3, 20, 50]. Some approaches attempt to design a self-supervision signal [50], but they do not explicitly aim to learn discriminative features, which are unsuitable for re-ID task due to large intra-class variations. Some other methods adopt ranking [19] or retrieval [4] based label assignment strategies, but they are easy to suffer from the collapsing problem that most unlabeled samples might be assigned to the same class [3]. Additionally, several clustering based unsupervised deep learning methods are introduced for re-ID [9]. However, they are hard to be applied on large-scale person re-ID applications due to time-consuming clustering procedure. Other approaches utilize the graph theory to exploit the relationship among different samples [3, 20]. However, large cross-camera variations in person re-ID may introduce lots of false positives which depresses the effectiveness of these methods.

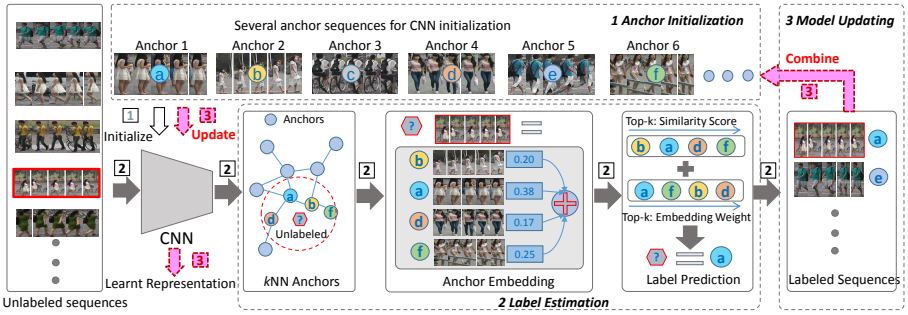


Fig. 2: The proposed RACE framework. It contains three main steps: (1) *Anchor Initialization*, several anchor sequences representing different persons are selected for CNN initialization; (2) *Label Estimation*, label estimation of unlabeled sequences via robust anchor embedding and top-k counts label prediction; (3) *Model Updating*, the deep feature representation is updated with newly labeled sequences and anchor sequences.

Deep Learning for Re-ID. Existing deep learning re-ID methods can be roughly categorized into three categories according to the learning objectives: triplet loss [5, 11, 48], contrastive/verification loss [6, 35, 45] and classification/identity loss [33, 56]. Moreover, some works combine them together to improve the performance [42, 48]. In addition, some CNN-RNN related network structures are also designed for video re-ID task [26, 30, 44]. All these methods can be configured in our framework to learn discriminative re-ID models.

Semi-supervised Learning. The proposed method is also related to the anchor graph based semi-supervised learning approaches [25,37,38] since we randomly select the sequences for anchor initialization. Similarly, they also contain the anchor embedding process to measure the relationship between the anchors and unlabelled samples. Different from previous methods, they utilize the graph regularization to estimated labels while we introduce a novel top-k counts strategy to estimate labels, which is more robust and efficient. In addition, we modify the anchor embedding procedure by considering the characteristics of video re-ID tasks under practical wild scenarios.

3 Proposed Method

3.1 Overview

Our goal is to accurately estimate labels with large amount of unlabeled tracking sequences collected from different cameras, where discriminative re-ID models can be subsequently learnt. The proposed framework contains three main steps as shown in Fig. 2: (1) *Anchor Initialization*, several anchor sequences are randomly selected for CNN initialization to get discriminative feature representations for better label estimation (Section 3.2). Meanwhile, the selected anchor sequences are consequently used for later label estimation of unlabelled image

sequences. (2) *Label Estimation*, with the learnt representation, label estimation of unlabeled sequences via robust anchor embedding and top-k counts label prediction is introduced (Section 3.3). Specifically, a robust anchor embedding is introduced to reconstruct any unlabelled sequences with their nearest anchor sequences to ensure efficiency. Meanwhile, each image sequence is represented by its regularized affine hull to reduce the impact of outlier frames. After that, the top-k counts label prediction strategy with the learnt embedding weights is conducted to predict the labels of unlabelled sequences. (3) *Model Updating*, with the newly estimated sequences and anchor sequences, we update the deep feature representation learning with more training samples (Section 3.4).

3.2 Anchor Initialization

It is well recognized that good model initialization is essential for deep feature representation learning systems. In this paper, we design an effective model initialization strategy according to the characteristics of video re-ID task. We firstly randomly select m anchor sequences ($\mathcal{A}$) to fine-tune the pre-trained ImageNet model [10], where the m anchor sequences are assumed to represent different persons[1]. The assumption is reasonable since the same person cannot be presented at the same instant under different non-overlapping cameras [27,28]. Under this assumption, the anchor sequences can be trivially obtained without human label annotation efforts in real applications. Accordingly, image frames within each sequence are assumed to belong to the same person identity, which could be ensured with effective tracking algorithms [7]. In this manner, the video sequence provides abundant training samples for each person by treating different person identities as different classes. Therefore, these selected anchor sequences could be adopted to initialize the deep neural network to learn discriminative feature representations for label estimation. In this paper, we adopt classification loss (IDE [54]) as the baseline structures, since it is effective for training and has shown good convergency [53]. Correspondingly, an anchor set $\mathcal{A}$ is constructed for these initialized anchor sequences. Denoted by

$$\mathcal{A} = \{A_l \mid l = 1, 2, \cdots, m\} \tag{1}$$

Each node A_l represents a set of frame-level feature vectors from the lth anchor sequence. l denotes the corresponding *initialized pseudo-label* assigned to anchor A_l. Then these anchors are utilized for label estimation from unlabeled sequences.

3.3 Label Estimation

Robust Anchor Embedding. With the initialized CNN representation, we could extract the feature representations of the unlabelled sequences $\mathcal{X} = \{X_i \mid i = 1, 2, \cdots, n\}$ and anchor sequences $\mathcal{A}$ for label estimation. In video re-ID,

[1] Anchors are assumed to represent different identities, but somehow it's unrestricted to this assumption (two anchors may belong to the same person) as shown in Fig. 6.

each sequence contains several different frame-level feature vectors, typical way to represent the sequence is adopting mean-pooling or max-pooling to transform multiple frame vectors into single feature vector [30, 53]. However, it may deteriorate the label estimation performance by introducing the noisy frames within sequences, which are usually caused by tracking or detection errors. Indeed, there are some methods trying to learn a better video sequence representation [6,44], but they do not explicitly consider the outlier frames existing within sequences or the efficiency issues. Thus, we adopt the efficient regularized affine hull (RAH) [58] to reduce the impact of outlier frames when measure the sequence to sequence similarity. It can handle arbitrary sequence length thus owns good flexibility. For sequence X_i, its RAH is denoted by

$$\mathbf{x}_i^{\mathcal{H}} = \{\sum \alpha_j \mathbf{x}_{i,j} \Big| \sum \alpha_j = 1, \|\alpha\|_{l_p} \leq \delta\} \tag{2}$$

where $\|\cdot\|_{l_p}$ (e.g, l_2 norm) could make the representation robust to outlier frames by suppressing unnecessary components for the final video sequence representation. The RAH transforms the original set of frame-level feature vectors of each sequence to a single feature vector with the learnt coefficients [58]. For simplification, the RAH of an image sequence i is represented by a d-dimensional feature vector hereinafter, termed as $\mathbf{x}_i^{\mathcal{H}}$ with a superscript $^{\mathcal{H}}$.

For the unlabeled sequences label prediction, we firstly aim at learning an embedding vector $\mathbf{w}_i$ that measure the underlying relationship between the unlabelled sequence $\mathbf{x}_i^{\mathcal{H}}$ and anchor set $\mathcal{A}^{\mathcal{H}}$ represented with RAHs. To ensure the efficiency, we learn the embedding weights of the unlabelled sequence i with its nearest (k) anchors instead of all anchors. It is reasonable that distant sequences are very likely to have different labels and contiguous sequences may have similar labels under the manifold assumptions [25, 40, 41]. This strategy greatly reduces the unnecessary computational cost since $k \ll m$. Therefore, an unlabeled sequence $\mathbf{x}_i^{\mathcal{H}} \in \mathbb{R}^{d\times 1}$ is formulated as a convex combination of its k nearest anchors ($\mathcal{A}_{\langle i\rangle}^{\mathcal{H}} \in \mathbb{R}^{d\times k}$). We formulate the coefficient learning problem as Robust AnChor Embedding (RACE) represented by:

$$\begin{aligned} &\min_{\mathbf{w}_i\in\mathbb{R}^k} f(\mathbf{w}_i) = \left\|\mathbf{x}_i^{\mathcal{H}} - \mathcal{A}_{\langle i\rangle}^{\mathcal{H}}\mathbf{w}_i\right\|^2 + \lambda \left\|d_{\langle i\rangle} \odot \mathbf{w}_i\right\|^2 \\ &s.t. \quad \mathbf{1}^T\mathbf{w}_i = 1, \mathbf{w}_i \geq 0 \end{aligned} \tag{3}$$

where the k entries of the vector $\mathbf{w}_i$ represent the corresponding embedding weights of unlabeled sequence $\mathbf{x}_i^{\mathcal{H}}$ to its k closest anchors $\mathcal{A}_{\langle i\rangle}^{\mathcal{H}}$. $d_{\langle i\rangle}$ is a vector that represents the visual similarity between the unlabeled sequence $\mathbf{x}_i^{\mathcal{H}}$ and the anchors $\mathcal{A}_{\langle i\rangle}^{\mathcal{H}}$. $\odot$ denotes the element-wise multiplication. λ is a trade-off factor to balance two terms. RACE contains two separate terms, the first *embedding term* aims at reconstructing the unlabelled sequence with its nearest neighbor anchors. The second *smoothing term* constrains the learnt coefficients such that larger weights should be assigned to the anchors with smaller distance. RACE transforms the high-dimensional CNN representation into low-dimensional embedding weight vector to reduce the computational cost.

Fig. 3: Noisy frames within the image sequence on MARS dataset.

Smoothing term. Since the original LAE in [25] does not have any constraint between the embedding weights and sequence to anchor distance. From the manifold assumption perspective, it is reasonable that nearby sequences tend to have similar labels. That is, nearby anchors should have larger reconstruction weights while the distant anchors should be assigned with smaller weights. Correspondingly, we define $d_{\langle i \rangle}$ by

$$d_{\langle i \rangle}(k) = exp(-\frac{\|\mathbf{x}_i^{\mathcal{H}} - \mathcal{A}_{\langle i \rangle}^{\mathcal{H}}(k)\|^2}{\sigma}) \tag{4}$$

where σ is a balancing parameter, and is usually defined by the average distance of $\mathbf{x}_i^{\mathcal{H}}$ to its nearest anchors $\mathcal{A}_{\langle i \rangle}^{\mathcal{H}}$.

Optimization. After transforming the multiple frame-level feature vectors in each sequence to RAH with an approximate solution in [58], the optimization problem in Eq. 3 becomes the standard quadratic programming problem. To accelerate the optimization and ensure the sparsity of the learnt weights, we adopt the projected gradient method [8, 25] to optimize Eq. 3. The updating rule is expressed by

$$\begin{aligned} \mathbf{w}_i^{(t+1)} &= \mathcal{P}_{\mathbb{S}}(\mathbf{w}_i^{(t)} - \eta_t \nabla f(\mathbf{w}_i^{(t)})) \\ \mathcal{P}_{\mathbb{S}}(\mathbf{w}) &= arg \min_{\mathbf{w}_i \in \mathbb{S}} \|\mathbf{w}' - \mathbf{w}\| \end{aligned} \tag{5}$$

where t denotes the iteration step, $\mathcal{P}_{\mathbb{S}}$ is a simplex projection to ensure the nonnegative normalization constraints in Eq. 3. η_t is a positive step size, $\nabla f(\mathbf{w})$ denotes the gradient of f at $\mathbf{w}$. Details can be found in [8, 58]. The embedding weights measure the intrinsic similarity between the unlabeled sequences and anchors, which are subsequently used for label estimation.

Label Prediction with Top-k Counts. A straightforward solution for label estimation is to design the graph Laplacian and conduct graph regularization as done in many anchor-graph based semi-supervised learning methods [25, 31, 37]. However, it is unsuitable for our scenario due to the following reasons:

– Under semi-supervised settings, they usually assume that every unlabeled sample must be assigned with a label according the anchor labels. However, for video re-ID, the identities of the anchor set are usually only a subset of all possible identities. Compulsory label assignment may produce large amount of false positives especially for wild settings, which would deteriorate the later feature representation learning.

– To the best of our knowledge, most graph based learning methods suffer from high computational complexities. Specifically, the graph Laplacian step is $O(m^2n)$, and the graph regularization process is $O(m^2n+n^3)$. In large-scale camera network applications, both m and n might be extremely large, which makes these methods incapable.

To address above robustness and efficiency issues, we design a simple but effective top-k counts strategy for the label prediction. The main idea is that if two image sequences belong to the same person identity, they should be very close to each other under different measure dimensions [14]. Specifically, if unlabeled sequence $\mathbf{x}_i$ is assigned with label l of $\mathcal{A}_l$, it should satisfy two principles: (1) $\mathcal{A}_l$ should be within the nearest ($k' \leq k$) anchors of $\mathbf{x}_i$, denoted by $\mathcal{N}_{\langle i,k'\rangle}$. It means that the sequence $\mathbf{x}_i$ should be extremely visual similar to anchor $\mathcal{A}_l$. This principle guarantees that only visual similar samples could share the same label, it measures the visual similarity. (2) The embedding weight of $\mathbf{w}_{i,l}$ should be large enough since embedding process measures the intrinsic underlying relationship between the unlabelled sequences and anchors, it acts as the intrinsic similarity. Mathematically, we formulate the label prediction as

$$\hat{y}_i = \begin{cases} 0 & , if \mathcal{A}^{\mathcal{H}}_{\langle i,k'\rangle} \cap \mathcal{N}_{\langle i,k'\rangle} = \emptyset \\ \arg\max\limits_{l \in \mathcal{A}^{\mathcal{H}}_{\langle i,k'\rangle}} \frac{\mathbf{w}_{i,l}}{\mathcal{R}(A_l)} & , others \end{cases} \tag{6}$$

where $\mathcal{R}(A_l)$ represents the ranking order of A_l in $\mathcal{N}_{\langle i,k'\rangle}$ according to the visual similarity, which jointly considers the embedding weights and visual similarity scores. Our label prediction strategy has two main advantages: (1) we could avoid the compulsory label assignment of uncertain sequences, thus could reduce large amount of false positives under wild settings. Smaller k' means stricter constraints. (2) it is quite efficient. The first criteria could be efficiently done with and-or operation, and the second label prediction step only needs to compute less than k' ($k' \leq k \leq m$) times for each unlabelled sequence. The computational complexity of the label prediction stage is $O(kn + k' \lg k' n)$ (intersection operation + ranking operation), which is much lower than the graph models [25, 37] with $O(m^2n+n^3)$. Experiments show that the proposed method produces higher label estimation performance with slightly better efficiency for video re-ID.

3.4 Model Updating

With the newly estimated sequences together with the anchor sequences, we could adopt existing supervised methods (e.g. IDE [53], QAN [26], ASTPN [44]) to update the deep feature representation learning. The learnt feature representation is improved with more training samples. Additionally, self-training strategy could also be adopted to refine the label estimation process and feature representation learning. Moreover, with the newly estimated labels by RACE together with anchor set, we could learn an improved similarity measurement. Therefore, the anchor embedding could be updated to get more accurate label prediction results and training samples. With iterative updating, better label estimation performance and feature representation would be achieved.

4 Experimental Results

4.1 Experimental Settings

Datasets. Three publicly available video re-ID datasets are selected for evaluation: two small-scale datasets, PRID-2011 dataset [12], iLIDS-VID dataset [39] and one large-scale MARS dataset [53]. The PRID-2011 dataset is collected from two disjoint surveillance cameras with significant color inconsistency. It contains 385 person video sequences in camera view A and 749 person sequences in camera view B. Among all persons, 200 persons are recorded in both camera views. The iLIDS-VID dataset is captured by two non-overlapping cameras located in an airport arrival hall, 300 person video sequences are sampled in each camera. The MARS dataset is a large-scale dataset, it contains 1,261 different persons whom are captured by at least 2 cameras, totally 20,715 image sequences which are automatically achieved by DPM detector and GMCCP tracker.

Evaluation Protocol. Different from previous unsupervised settings on PRID-2011 and iLIDS-VID datasets [27, 47], they adopt the same training set as in supervised methods, which is impractical for real applications. We modify the training settings for *wild evaluation*. For the PRID-2011 dataset, there are totally 600 person sequences from two cameras (300 sequences in each camera) for training, only 100 persons appear in both cameras. For anchor initialization, 300 image sequences representing different persons are randomly selected from two cameras. For the iLIDS-VID dataset, there are totally 300 person sequences from two cameras (100 sequences in each camera) for training, only 50 persons appear in both cameras. For anchor initialization, 100 image sequences representing different persons are randomly selected from two cameras. For the MARS dataset, 625 sequences from 625 persons are randomly selected as anchors for initialization. The anchors are assumed to represent different persons, but somehow it's unrestricted to this assumption (*two anchors may belong to the same person*) as illustrated in Fig. 6. In testing procedure, the Euclidean distance of two sequences is adopted. Rank-k matching rates and mAP values are both reported in the testing phase.

Implementation details. We use ResNet-50 [10] pre-trained on ImageNet as our basic CNN model. Specifically, we insert a fully connected layer with 512 units after the pooling-5 layer, followed by batch normalization, ReLU and Dropout [33]. The dropout probability is set to 0.5 for all datasets. All images are resized to 128×256. Standard data augmentation methods are adopted. Batch size is set to 256 for MARS dataset, and 64 for both PRID-2011 and iLIDS-VID datasets. We use stochastic gradient descent to optimize the neural networks. We adopt the default Normal function of MxNet for the variables initialization. The initial learning rate is set to 0.003 for MARS dataset and 0.01 for both PRID-2011 and iLIDS-VID datasets, it is decreased by 0.1 after 20 epochs. The total training epochs are set to 30 for all datasets unless otherwise specified. The kNN graph construction k in Eq. 3 is set to 15, and the label prediction k' is set to 1. The smoothing parameter λ is set to 0.1. RAH is optimized with [58]. The default experimental results are with 1-round label estimation.

Methods	Recall	Precision	F-Score
1NN	41.76	41.76	41.76
AGR [25]	43.30	43.30	43.30
DGM [47]	42.40	59.64	49.57
RACE	40.87	66.22	**50.54**

Table 1: Evaluation of label estimation performance (%) on MARS dataset. The label estimation time for DGM [47] is about 2 hours, and RACE is only 183s.

Settings	Recall	Precision	F-Score
w/o Top-k	47.84	47.84	47.84
w/o RAH	37.20	68.18	48.14
w/o Smooth	42.75	59.22	49.65
RACE	40.87	66.22	**50.54**

Table 2: Evaluation of different components in the proposed RACE. Label estimation performance (%) on MARS dataset.

4.2 Detailed Analysis

Evaluation of label estimation. We adopt the general precision, recall and F-score as the evaluation criteria of label estimation performance. The results on the MARS dataset are shown in Table 1. 1) **Effectiveness**. Results illustrate that the proposed method can improve the precision and F-score by a large margin compared with 1NN (nearest neighbor) and AGR [25] baselines. Specifically, we could achieve 66.22% label estimation accuracy on the large-scale MARS dataset, and the F-score is about 50.54%. 2) **Efficiency**. Compared to the state-of-the-art DGM method in ICCV17 on the large-scale MARS dataset, the proposed method is much more efficient than DGM in terms of the label estimation time (Ours: $\sim$183s, DGM [47]: $\sim$2 hours). Meanwhile, better label estimation performance is also achieved. Compared to AGR [25] with 185s, the proposed RACE is also more efficient in terms of the label estimation process. Furthermore, considering that both methods contain the embedding process with about 157s, the advantage of our top-k counts label prediction is more obvious.

Evaluation of each component. We evaluate each component of the proposed method by removing the corresponding component. The experimental results shown in Table 2 could verify the effectiveness of each component. "w/o Top-k" means that we directly conduct the label estimation according to the maximum embedding weights without top-k counts label prediction. It shows that the top-k counts label prediction improves the label estimation precision from 48% to 66%. Additionally, RAH mainly benefits the recall criterion, since RAH is more robust to outlier frames within sequences than the simply pooling method. Moreover, smoothing term also improves the label estimation performance further with the smoothing similarity constraint in the embedding process. Overall, the F-score is increased by integrating three main components.

Parameters analysis. Three important parameters: (1) k, the number of nearest anchors selected for RACE, (2) k', the parameter of top-k counts label prediction in Eq. 6, (3) λ, the trade-off parameter balances the embedding and smoothing term in Eq. 3, are evaluated in Fig. 4. (1) For the kNN anchor graph construction, larger k usually could bring better performance as shown in Fig. 4(a). However, it also increases the computational time in later steps. Moreover, we could see that the performance becomes stable when k is up to 15. Therefore,

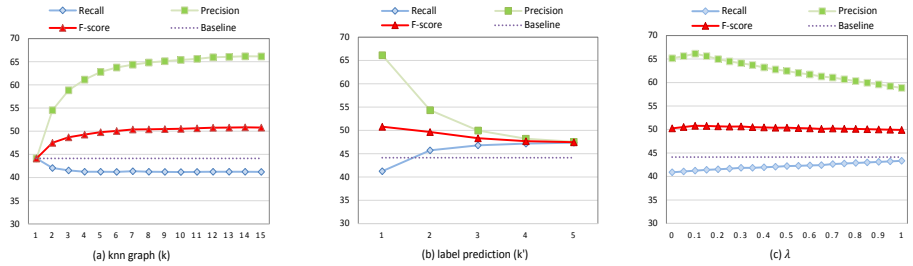


Fig. 4: Parameters Analysis on MARS dataset. (a) The number of nearest anchors (k) selected for RACE; (b) The parameter k' of label prediction in Eq. 6; (c) λ, the trade-off parameter in Eq. 3. '*Baseline*' provide a lower bound of 1NN with RAH representation.

Datasets	PRID-2011				iLIDS-VID				MARS				
Rank at r	1	5	10	20	1	5	10	20	1	5	10	20	mAP
Supervised	64.7	91.2	94.6	98.7	45.6	69.9	78.4	87.5	69.3	85.8	89.4	92.8	49.8
Baseline	45.3	72.5	86.6	90.4	13.6	33.7	44.3	58.1	33.2	47.7	54.7	62.0	15.5
RACE	50.6	79.4	88.6	91.8	19.3	39.3	53.3	68.7	41.0	55.6	62.2	67.2	22.3

Table 3: Comparison to baseline systems (IDE [53] + Resnet50 [10]). Person re-identification performance (%) at rank-1, 5, 10, 20 and mAP on three datasets."Baseline" means the performance of initialized feature representation.

we choose $k = 15$ in our experiments. (2) In terms of k', smaller k' means stricter constraints between the visual similarity and the intrinsic similarity, it also may result in smaller recall values. Since larger recall values means more noisy labels would be encountered for the feature representation learning procedure, we prefer a better label precision performance, so k' is set to 1 in our experiments. (3) Sensitivity to λ, it could also illustrate the improvement of the smoothing term. Besides, larger λ means stricter constraints for the similarity scores between two embedded anchors. Obviously, if λ is large enough, the proposed RACE would be degenerated to nearest neighbor method. Overall, a proper choice of λ would improve the overall performance as illustrated in Fig. 4(c).

Evaluation of re-identification. We evaluate the re-ID performance with the estimated labels on three datasets as shown in Table 3. Note that our evaluation protocols simulate the wild settings, which are slightly different from the standard supervised settings on PRID-2011 and iLIDS-VID datasets. Table 3 illustrates that the proposed RACE improves the baseline feature representation learning method consistently on all three datasets with one-round label estimation and feature learning. Specifically, we improve the rank-1 matching rates from 45.32% to 50.64% on the PRID-2011 dataset, 13.6% to 19.33% on the iLDIS-VID dataset and 33.2% to 41.0% on the MARS dataset. We suppose that the performance would be further boosted with iterative updating. Note that the performance of feature learning process might be improved with other advanced deep learning [26, 44] or re-ranking methods [1, 2, 46, 57].

Datasets	PRID-2011				iLIDS-VID				Ref.
Rank at r	1	5	10	20	1	5	10	20	
Salience [52]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9	CVPR13
LOMO [23]	40.6	66.7	79.4	92.3	9.2	20.0	27.9	46.9	CVPR15
STFV3D [24]	42.1	71.9	84.4	91.6	**37.0**	**64.3**	**77.0**	**86.9**	ICCV15
GRDL [16]	41.6	76.4	84.6	89.9	21.7	42.9	56.2	71.6	ECCV16
SMP [27]	38.7	68.1	79.6	90.0	16.0	31.8	43.8	56.8	ICCV17
DGM [47]	48.2	78.3	83.9	**92.4**	23.1	46.7	58.3	71.2	ICCV17
RACE (Round1)	**50.6**	**79.4**	**84.8**	91.8	19.3	39.3	53.3	68.7	-

Table 4: Comparison with state-of-the-art unsupervised methods on *small-scale* PRID-2011 and iLIDS-VID datasets *under wild settings*. Rank-k matching rates (%).

Rank at r	1	5	10	20	mAP	Ref.
LOMO [23]	14.9	27.4	33.7	40.8	5.5	ICCV15
GRDL [16]	19.3	33.2	41.6	46.5	9.6	ECCV16
SMP [27]	41.2	55.6	-	66.8	19.7	ICCV17
DGM [47]	36.8	54.0	61.6	**68.5**	21.3	ICCV17
RACE (Round1)	41.0	55.6	61.9	67.2	22.3	-
RACE (Round2)	**43.2**	**57.1**	**62.1**	67.6	**24.5**	-

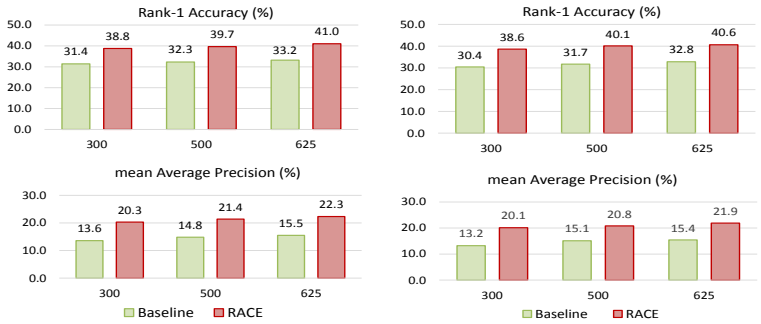
Table 5: Comparison with state-of-the-art unsupervised methods on the *large-scale* MARS datset. Rank-k matching rates (%) and mAP (%).

4.3 Comparison with State-of-the-arts

This subsection demonstrates the comparison with other state-of-the-art unsupervised re-ID methods, including Salience [52], LOMO [23], STFV3D [24], GRDL [16], DGM [47] and SMP[2] [27]. Note that our evaluation settings on PRID-2011 and iLIDS-VID datatsets are different from the original DGM [47] and SMP [27], where they assume that all persons appear in both cameras. The comparisons on three datasets are shown in Table 4 and 5.

Results illustrate that we could achieve the best performance under wild settings on PRID-2011 dataset and the large-scale MARS dataset. Specifically, the rank-1 accuracy is about 50.6% on the PRID-2011 dataset as shown in Table 4 under wild settings. For the large-scale MARS dataset, 625 persons randomly appear in 6 cameras, thus it is more related to the practical multi-camera networks. Correspondingly, we could achieve the state-of-the-art performance, the rank-1 matching rate is 43.2% and mAP is 24.5% with 2-round training as shown in Table 5. However, Table 4 shows that our results on iLIDS-VID dataset are lower than the state-of-the-art unsupervised methods, it can be attributed to the limited training data for deep feature representation learning. We suppose that the proposed method can be applied to practical applications where a large amount of unlabeled tracking sequences could be collected for unsupervised deep feature representation initialization and learning.

[2] GRDL, DGM and SMP are implemented with the released code.



(a) Number of anchor persons (b) Number of different anchors

Fig. 5: Sensitivity to anchor selection on the MARS dataset. (a) Number of anchor persons. All these anchors represent different persons. (b) Number of different anchors. Anchors are randomly selected, two anchors may belong to the same person.

We also observe that the performance can be further improved with further label estimation/refinement as shown in Table 5. Specifically, it is around 2% improvement for both rank-1 accuracy (41.0% to 43.2%) and mAP values (22.3% to 24.5%) on the large-scale MARS dataset in round 2. With iterative updating scheme, the performance can be further improved by scarifying the efficiency.

4.4 Robustness in the Wild

In this section, we evaluate RACE under more challenging settings, which are 1) *Sensitivity to anchor selection*, different anchor initialization strategies. 2) *Sensitivity to imbalance ratios*, different imbalance ratios of the training set.

Sensitivity to anchor selection. The anchor initialization is very important in our proposed method, especially for the number of selected anchors. Two sets of different anchor selection experiments are evaluated as shown in Fig. 5. (1) *it's hard to know the specific number of person identities in an open environment.* Therefore, we randomly select different number of initialized anchor sequences to test the performance variations on the large-scale MARS dataset as shown in Fig. 5(a). The results illustrate that the proposed method can improve the baseline feature representations consistently with different number of initialized anchors. Specifically, the overall performances are slightly decreased compared to 625 sequences initialization, but they are still competitive to the current state-of-the-art unsupervised methods. 2) *it's hard to ensure the selected anchors truly represent different persons.* Therefore, we relax the assumption, where anchors are randomly selected, thus two anchors may belong to the same person. Fig. 5(b) shows that the proposed method still achieves satisfactory performance with slight decrease even without the assumption.

Sensitivity to imbalance ratios. We adopt the additional 734 person sequences together with 200 person training sequence pairs in PRID-2011 dataset

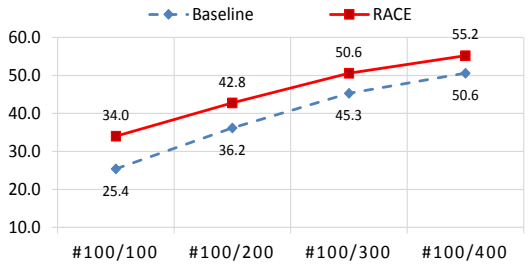


Fig. 6: Rank-1 matching rates (%) of different imbalance ratios on the PRID-2011 dataset. "#m/n" means that m persons out of n different person sequences appear in both cameras.

to simulate different imbalance ratios as shown in Fig. 6. "#100/400" means that only 100 persons appear in both cameras while each camera contains 400 different person sequences. Specifically, the additional person sequences are simply treated as different persons to initialize the deep feature representation learning. Fig. 6 demonstrates that RACE improves the deep feature representation learning performance consistently under different wild settings. Moreover, since deep feature representation learning could benefit from more training data, RACE achieves even better performance with more anchor sequences on PRID-2011 dataset even with lower positive ratio. Compared with DGM [47], RACE is more robust to lower positive ratios, while DGM drops quickly with low positive ratios. Overall, RACE is superior in the following aspects: 1) it is scalable for large-scale scenarios, which learns discriminative deep feature representations without any manually labeled information. 2) it is efficient in terms of the label estimation procedure. 3) it is robust under wild settings, thus could be applied to real applications with highly imbalanced unlabeled training data.

5 Conclusion

This paper proposes an efficient and scalable unsupervised deep feature representation learning framework for video re-ID under wild settings. To accurately estimate labels from unlabelled sequences, a robust anchor embedding method is designed for this task, regularized affine hull together with a manifold smoothing term is integrated into the embedding process. A novel top-k counts label prediction strategy is then introduced to reduce false positives. Deep feature representation learning could be updated with newly estimated unlabeled sequences. Experimental results on large-scale datasets under wild settings demonstrate the superiority of the proposed method.

Acknowledgments

This work is partially supported by Hong Kong RGC General Research Fund HKBU (12254316), and National Natural Science Foundation of China (61562048).

References

1. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2530–2539 (2017)
2. Bai, S., Sun, S., Bai, X., Zhang, Z., Tian, Q.: Smooth neighborhood structure mining on multiple affinity graphs with applications to context-sensitive similarity. In: European Conference on Computer Vision (ECCV). pp. 592–608 (2016)
3. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. ICML (2017)
4. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: ECCV (2014)
5. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR (2016)
6. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: ICCV (2017)
7. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: efficient convolution operators for tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21–26 (2017)
8. Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T.: Efficient projections onto the l 1-ball for learning in high dimensions. In: ICML (2008)
9. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. arXiv preprint arXiv:1705.10444 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. In: ICCV (2017)
12. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis. pp. 91–102 (2011)
13. Jianming, L., Weihang, C., Qing, L., Can, Y.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Jin, S., Su, H., Stauffer, C., Learned-Miller, E.: End-to-end face detection and cast grouping in movies using erdos-rényi clustering. In: International Conference on Computer Vision (ICCV). vol. 2, p. 8 (2017)
15. Jingya, W., Xiatian, Z., Shaogang, G., Wei, L.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Person re-identification by unsupervised l1 graph learning. In: European Conference on Computer Vision (ECCV). pp. 178–195 (2016)
17. Lan, X., Ma, A.J., Yuen, P.C., Chellappa, R.: Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Transactions on Image Processing (TIP) **24**(12), 5826–5841 (2015)
18. Lan, X., Zhang, S., Yuen, P.C., Chellappa, R.: Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. IEEE Transactions on Image Processing **27**(4), 2022–2037 (2018)

19. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: IEEE International Conference on Computer Vision (ICCV). pp. 667–676 (2017)
20. Li, D., Hung, W.C., Huang, J.B., Wang, S., Ahuja, N., Yang, M.H.: Unsupervised visual representation learning by graph-based consistent constraints. In: ECCV (2016)
21. Li, J., Ma, A.J., Yuen, P.C.: Semi-supervised region metric learning for person re-identification. International Journal of Computer Vision pp. 1–20 (2018)
22. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
23. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2197–2206 (2015)
24. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: IEEE International Conference on Computer Vision (ICCV). pp. 3810–3818 (2015)
25. Liu, W., He, J., Chang, S.F.: Large graph construction for scalable semi-supervised learning. In: ICML (2010)
26. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR (2017)
27. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: IEEE International Conference on Computer Vision (ICCV). pp. 2429–2438 (2017)
28. Ma, A.J., Li, J., Yuen, P.C., Li, P.: Cross-domain person reidentification using domain adaptation ranking svms. IEEE Trans. Image Processing (TIP) **24**(5), 1599–1613 (2015)
29. Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K.M., Zhong, Y.: Person re-identification by unsupervised video matching. Pattern Recognition (PR) **65**, 197–210 (2017)
30. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1325–1334 (2016)
31. Nie, F., Zhu, W., Li, X.: Unsupervised large graph embedding. In: AAAI (2017)
32. Peng, P., Xiang, T., Wang, Y., et, a.: Unsupervised cross-dataset transfer learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1306–1315 (2016)
33. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV (2017)
34. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling. arXiv preprint arXiv:1711.09349 (2017)
35. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV (2016)
36. Wang, H., Gong, S., Xiang, T.: Unsupervised learning of generative topic saliency for person re-identification. In: BMVC (2014)
37. Wang, M., Fu, W., Hao, S., Tao, D., Wu, X.: Scalable semi-supervised learning by efficient anchor graph regularization. IEEE TKDE (2016)
38. Wang, Q., Yuen, P.C., Feng, G.: Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions. Pattern Recognition **46**(9), 2576–2587 (2013)

39. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European Conference on Computer Vision (ECCV). pp. 688–703 (2014)
40. Wang, Z., Hu, R., Chen, C., Yu, Y., Jiang, J., Liang, C., Satoh, S.: Person reidentification via discrepancy matrix and matrix metric. IEEE transactions on cybernetics (2017)
41. Wang, Z., Hu, R., Liang, C., et al.: Zero-shot person re-identification via cross-view consistency. IEEE Transactions on Multimedia (TMM) **18**(12), 2553–2566 (2016)
42. Wang, Z., Ye, M., Yang, F., Bai, X., Satoh, S.: Cascaded sr-gan for scale-adaptive low resolution person re-identification. In: IJCAI. pp. 3891–3897 (2018)
43. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
44. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: ICCV (2017)
45. Ye, M., Lan, X., Li, J., Yuen, P.C.: Hierarchical discriminative learning for visible thermal person re-identification. In: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI) (2018)
46. Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., Hu, R.: Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Transactions on Multimedia **18**(12), 2553–2566 (2016)
47. Ye, M., Ma, A.J., Zheng, L., Li, J., Yuen, P.C.: Dynamic label graph matching for unsupervised video re-identification. In: IEEE International Conference on Computer Vision (ICCV). pp. 5142–5150 (2017)
48. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI. pp. 1092–1099 (2018)
49. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
50. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: CVPR (2017)
51. Zhao, J., Xiong, L., Cheng, Y., Cheng, Y., et al.: 3d-aided deep pose-invariant face recognition. In: IJCAI (2018)
52. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3586–3593 (2013)
53. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: European Conference on Computer Vision (ECCV). pp. 868–884 (2016)
54. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv (2016)
55. Zheng, L., Yang, Y., Tian, Q.: Sift meets cnn: A decade survey of instance retrieval. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **40**(5), 1224–1244 (2018)
56. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)
57. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3652–3661 (2017)
58. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: Extend the learning of distance metrics. In: ICCV (2013)