

A Free Lunch to Person Re-identification: Learning from Automatically Generated Noisy Tracklets

Hehan Teng¹, Tao He¹, Yuchen Guo¹, Zhenhua Guo², Guiguang Ding¹

¹ School of Software, Tsinghua University, Beijing, China

² Alibaba Group

thss15.tenghh@163.com, {kevin.92.he, yuchen.w.guo, cszguo}@gmail.com, dinggg@tsinghua.edu.cn

Abstract

A series of unsupervised video-based re-identification (re-ID) methods have been proposed to solve the problem of high labor cost required to annotate re-ID datasets. But their performance is still far lower than the supervised counterparts. In the mean time, clean datasets without noise are used in these methods, which is not realistic. In this paper, we propose to tackle this problem by learning re-ID models from automatically generated person tracklets by multiple objects tracking (MOT) algorithm. To this end, we design a tracklet-based multi-level clustering (TMC) framework to effectively learn the re-ID model from the noisy person tracklets. First, intra-tracklet isolation to reduce ID switch noise within tracklets; second, alternates between using inter-tracklet association to eliminate ID fragmentation noise and network training using the pseudo label. Extensive experiments on MARS with various manually generated noises show the effectiveness of the proposed framework. Specifically, the proposed framework achieved mAP 53.4% and rank-1 63.7% on the simulated tracklets with strongest noise, even outperforming the best existing method on clean tracklets. Based on the results, we believe that building re-ID models from automatically generated noisy tracklets is a reasonable approach and will also be an important way to make re-ID models feasible in real-world applications.

1. Introduction

Person re-identification (re-ID) is to match persons across non-overlapping cameras. It is one of the core techniques in intelligent surveillance analysis. Due to the urgent demand for public safety, it has been an active research field over the years. Video-based person re-ID is the problem where subjects to be retrieved are presented as video sequences. Person re-ID has shown promising results in a fully supervised setting. This learning paradigm assumes

that there is a large number of labeled high-quality cross-camera training data. But it is of the high cost to collect such a large-scale dataset, due to the exponential labeling cost. Besides, a well-trained re-ID model has been proved to perform much worse in a new domain.

To overcome the drawbacks of supervised methods, in the last two years, several works have turned to study unsupervised or weakly supervised person re-ID. Specifically, we focus on unsupervised video-based person re-ID, where training data can be obtained without human labor by multiple object tracking [2] (MOT) algorithms, as shown in figure 1. Most of the existing unsupervised video re-ID methods still yield unsatisfactory results. Moreover, these methods operate on video re-ID datasets, such as MARS [28], iLIDS-VID [17] and PRID 2011 [7]. It should be noted that, as shown in figure 1, although these datasets are used in an unsupervised manner, i.e. video sequences without intra- and inter-camera ID association, the sequences themselves are clean and without noise, and the production of such clean sequences requires substantial human effort as well. The gap between such datasets and MOT-generated tracklets are the noise introduced by MOT algorithms, which is mainly ID fragmentation noise and ID switch noise, alongside with detection [13] noise. Some methods have been proposed to exploit tracklets to build a re-ID model, but only ID fragmentation noise is considered, while ID switch and detection noise are ignored, resulting in a wider gap from being applicable to real-life scenarios.

We propose a new tracklet-based clustering and fine-tuning framework to account for both ID fragmentation noise and ID switch noise in MOT-generated tracklets. By analyzing characteristics of aforementioned noise, a multi-stage clustering-based method is proposed to reduce noise in the tracklets before feeding them into the unsupervised training pipeline, resulting in significant performance boost. Since raw video of video re-ID datasets is generally unavailable, a novel algorithm is proposed to generate simulated tracklets from video re-ID datasets, to assist evaluating our method under various strength of noise.

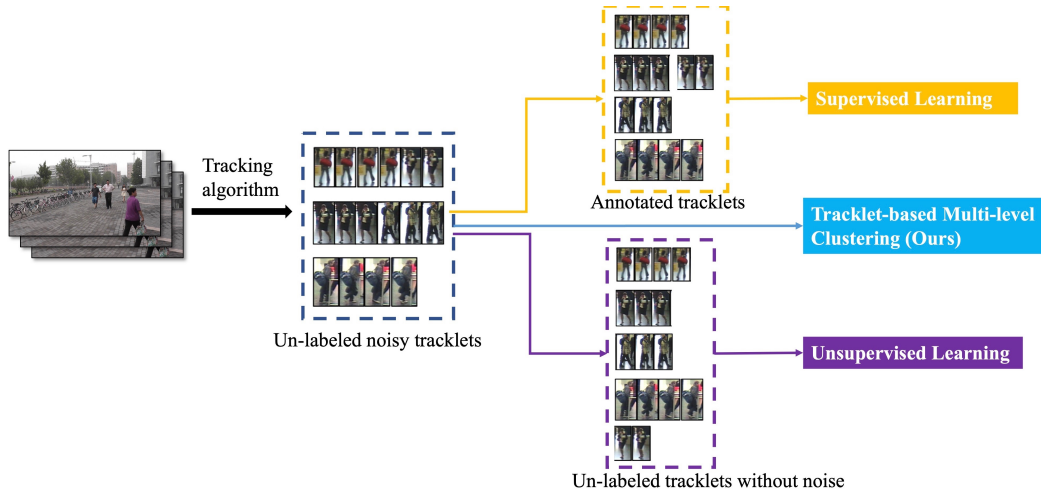


Figure 1: Different categories of person re-ID methods. Supervised methods require full annotation of video sequences. Datasets used by unsupervised video re-ID methods require human labor to eliminate noise within tracklets generated by MOT algorithms. Our method uses MOT tracklets directly.

We summarize our contribution as three-fold.

1. Firstly, we propose taking raw tracklets generated by MOT algorithms as input of our method, removing the requirement of human effort completely, resulting in nearly zero-cost training data preparation, moving a step closer into solving realistic problems.
2. Secondly, we analyze dominant noise categories in tracklets generated by MOT algorithms, i.e. ID fragmentation and ID switch noise, revealing characteristics which is exploited in our novel noise reduction processing. We combine the noise reduction techniques with a self-training mechanism in our method, named Tracklet-based Multi-level Clustering (TMC).
3. Thirdly, experiments show that our method achieves remarkable performance given that realistic tracklets with noise are used (mAP 55.3% rank-1 68.2% on simulated tracklets), and, if clean video re-ID datasets are used instead, outperforms existing unsupervised video re-ID methods.

2. Related Work

2.1. Unsupervised video-based re-ID

Although unsupervised video-based person re-ID is a relatively unexplored area compared to the supervised problem, it is gaining attention over the past few years. In this setting, unlabeled video sequences (tracklets) without noise are provided, and each person may have more than one tracklets under a certain camera. From MOT point of view, within each camera, these tracklets have ID fragmentation

noise, but not ID switch noise. Dynamic Graph Matching (DGM) [24] leverages the graph matching technique by constructing a graph for samples in each camera for label estimation, and iteratively update the graph to produce estimated label. To further improve the performance, global camera network constraints [23] are exploited for consistent matching. Riachy *et al.* [14] formulates the re-ID task as a set-based matching problem that they tackle using the NBNN classifier. DAL [1] proposed by Chenet *al.* is an end-to-end deep learning method which jointly optimize two association losses. EUG [21] is a step-wise one-shot learning method, gradually selecting a few candidates from unlabeled tracklets to enrich the labeled tracklet set. RACE [22] is robust anchor embedding method iteratively assigns labels to the unlabelled tracklets to enlarge the anchor video sequences set. Note that the EUG and RACE, requires additional information to initialize their learning process, which usually involves extra human labor.

2.2. Tracklet-based re-ID

Methods based on tracklets are developed to be independent of source data or model and also save human efforts. UTAL [10] is the first work to build a re-ID model from tracklets, which are automatically generated by multiple object tracking (MOT) algorithms. But only fragmentation noise is considered in this work, and the identity switch and detection noises, which commonly occur in the generated tracklets, are ignored in this work. Furthermore, the experiments on two commonly used datasets, in fact, are particularly designed to only contain the inter-camera fragmentation noise, without intra-camera noise. Their performance is also much lower than the domain adaptation or supervised counterparts. TSSL [19] is also built on tracklets but

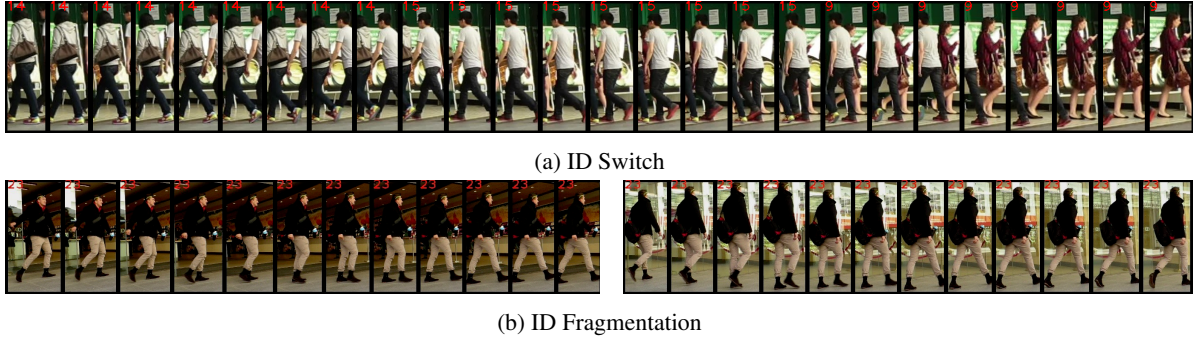


Figure 2: Examples of noise types in tracklets. (a) Example of ID switch. The tracklet each contains more than one (ground-truth) person. (b) Example of ID fragmentation. The occurrence of a (ground-truth) person was split up in more than one tracklet. Red numbers in the top-left corner of each frame denotes the ground-truth ID of the person tracked in that frame.

without camera view information. The experimental results on the Market are similarly low as UTAL. There are other works that require human efforts to label part of the dataset. MTML [29] assumes that the intra-camera instances are perfectly labeled, and it performs multi-task learning on multiple camera views. The progressive framework by Wu *et al.* [20] also requires the label of one example for each pedestrian. Although involved human efforts, the performance of these methods is still far behind the supervised ones, which hinders the application in the real world.

3. Method

3.1. Tracklet Noise Analysis

We present the definition of noise types in tracklets before analyzing their characteristics. In the evaluation metrics used in MOT benchmarks, such as MOT20 [3], the determination of ID fragmentation and ID switch requires a global tracklet-to-ground-truth assignment produced by optimal matching using Hungarian algorithm [9]. In this work we use a simplified and more intuitive definition as follows. ID fragmentation is said to occur if the occurrence of a (ground-truth) person was split up in more than one tracklet, and ID switch is said to occur if a tracklet contains more than one (ground-truth) person. Figure 2 shows examples for both types of noise.

MOT-generated tracklets cannot be used to train re-ID model directly since both types of noise have a negative impact on the learning process. ID fragmentation causes loss of information since during the learning process, multiple tracklets of a person would have been treated as different IDs. A supervised method will try to guide the model to differentiate these IDs while they are actually the same person, effectively misleading the model. On the other hand, an unsupervised method would have to discover and associate multiple tracklets of a person. Tracklets with ID switch have internal inconsistency, which affect both supervised and un-

supervised method in the same way. These tracklets guide the system into treating relevant people as same, and will harm the ability the model to distinguish different people.

3.2. Tracklet-based Multi-level Clustering

A multi-level clustering method is proposed to reduce both ID fragmentation and ID switch noise in the tracklets. A low-level clustering, named “intra-tracklet isolation” is first done within each tracklet, forming groups of images, to reduce ID switch noise. A high-level clustering, named “inter-tracklet association” is then conducted to mine association between tracklets to reduce ID fragmentation noise. **Intra-tracklet isolation** We propose an intra-tracklet isolation step to reduce ID switch noise in MOT-generated tracklets. Features of images in the tracklets are extracted by ImageNet-pretrained [4] ResNet50 and visualized with t-SNE [16] in figures 3 (a)(c). It is shown that features of different person in a tracklet are roughly distinguishable in the feature space, indicating that occurrences of ID switch do not result from similarities of people appearance, instead, are more of a result of indistinguishable movement trajectories, which is consistent with what is observed in real-life MOT algorithm results. Figures 3 (b)(d) show cosine similarity of features between adjacent frames in the tracklet. It can be shown in figure 3 (b) that similarity is significantly low when ID switch occurs (red points). This indicator, however, is unreliable as illustrated in figure 3 (d).

Therefore, intra-tracklet isolation was used to process tracklets into smaller but cleaner tracklets (figure 5). Specifically, DBSCAN [5] was used to cluster the features. Images from each resulting cluster will be composed into a smaller tracklet, while preserving the chronological order. Figure 4 shows the result of intra-tracklet isolation.

Inter-tracklet association While intra-tracklet isolation reduces ID switch noise, it introduces extra ID fragmentation noise, in addition to inherent ID fragmentation noise in tracking results. We propose using inter-tracklet association

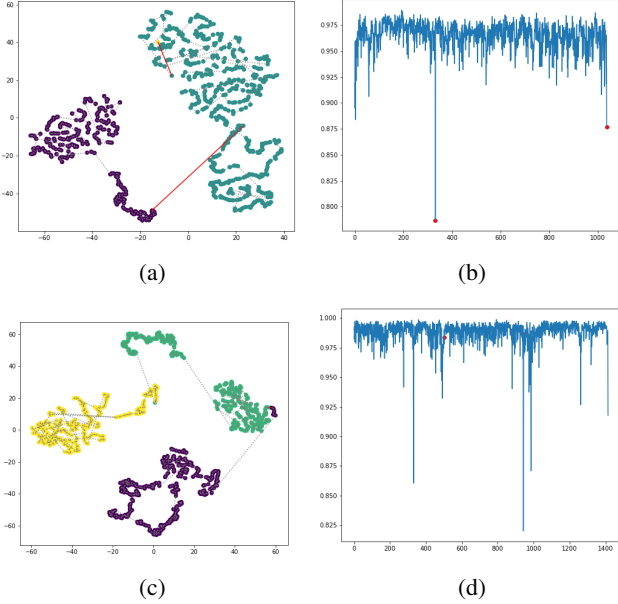


Figure 3: Illustration of tracklets with ID switch noise. Figures (a)(c) show the feature distribution of images in the tracklets, points colored according to ground-truth person ID. Figures (b)(d) show the cosine similarity of adjacent frames in the tracklet. Figures (a)(b), (c)(d) are from the same tracklet, respectively.

step to mine association of tracklets from the same person. Consecutive images of fixed length were sampled for each tracklet. Features of images in a tracklet are extracted using the updated model, followed by an average-pooling step to produce the feature of the tracklet. Features of all tracklets are then clustered with DBSCAN to produce hard pseudo labels.

3.3. Training pipeline

A simplification of MMT [6] [27] which keeps only one network (“student”) and its past temporal average model (“teacher”) was adopted as our overall training pipeline. The former is denoted as “Net” and latter as “Mean Net” [15] in figure 6.

Framework The tracklets with noise are first processed with intra-tracklet isolation to produce a new collection of tracklets with less noise. At the start of each training epoch, hard pseudo label are generated using inter-tracklet association as described above. In each iteration of an epoch, the same input images as Net were fed into Mean Net to produce soft pseudo labels. The Net is trained by utilizing both hard pseudo labels and soft pseudo labels. Updated Net is then used to update Mean Net in EMA (Exponential Moving Average) mode.

$$E^{(T)}[\theta] = \alpha E^{(T-1)}[\theta] + (1 - \alpha)\theta \quad (1)$$

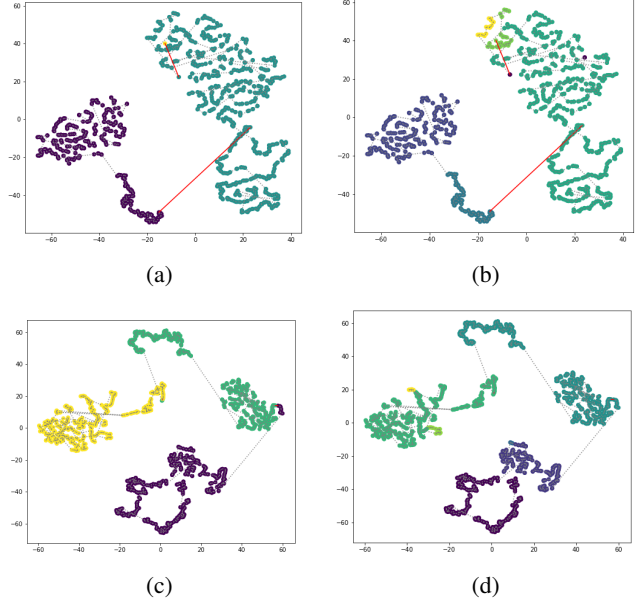


Figure 4: Results of intra-tracklet clustering. Figures show the feature distribution of images in the tracklets with ID switch. Points in figures (a)(c) colored according to ground-truth person ID. Points in figures (b)(d) colored according to clustering results.

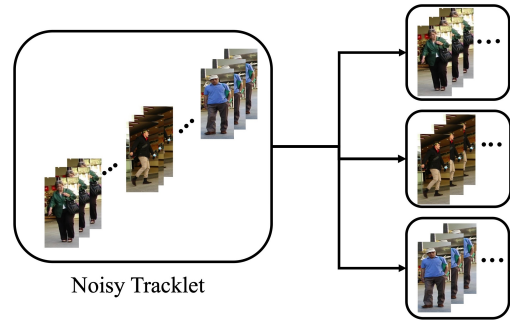


Figure 5: Intra-tracklet isolation. Tracklets are clustered into smaller but cleaner tracklets. Images from each resulting cluster will be composed into a smaller tracklet, while preserving the chronological order.

where $E^{(T-1)}[\theta]$ indicate the temporal average parameters of Net, i.e. Mean Net in the previous iteration, following the notation in [6]. The initial Mean Net parameters $E^{(0)}[\theta] = \theta$, and α is the ensembling momentum to be within the range $[0, 1)$.

Loss function In each iteration, the Net is trained by jointly optimizing the following loss functions: hard identity classification loss \mathcal{L}_{id} , hard triplet loss \mathcal{L}_{tri} , soft classification loss \mathcal{L}_{sid} , soft triplet loss \mathcal{L}_{stri} . The two hard loss functions are the same as typical loss in re-ID.

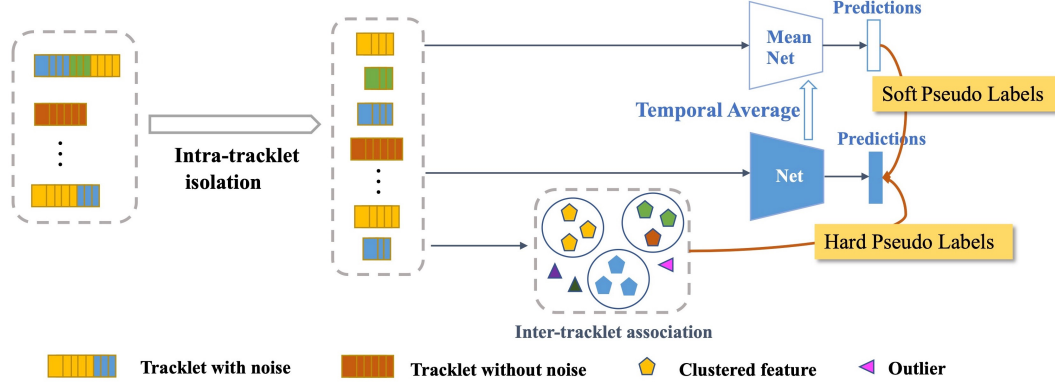


Figure 6: The overall pipeline. The tracklets with noise are first processed with intra-tracklet isolation to produce a new collection of tracklets with less noise. At the start of each training epoch, hard pseudo label are generated using inter-tracklet association. In each iteration of a epoch, the same input images as Net were feed into Mean Net to produce soft pseudo labels. The Net is trained by utilizing both hard pseudo labels and soft pseudo labels. Updated Net is then used to update Mean Net in EMA (Exponential Moving Average) [8] mode.

$$\mathcal{L}_{id}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{ce}(C(F(x_i|\theta)), \tilde{y}_i) \quad (2)$$

$$\mathcal{L}_{tri}(\theta) = \frac{1}{N} \sum_{i=1}^N \max(0, \|F(x_i|\theta) - F(x_{i,p}|\theta)\| + m - \|F(x_i|\theta) - F(x_{i,n}|\theta)\|) \quad (3)$$

where $\|\cdot\|$ denotes the L^2 -norm distance, x_i denotes the sampled consecutive images sampled from tracklet i at the current iteration, $F(x_i|\theta)$ is the feature of tracklet i output by current Net, $C(\cdot)$ is classifier, subscripts i,p and i,n indicate the hardest positive and hardest negative feature index in the batch, and $m = 0.5$ denotes the triplet distance margin.

The two soft loss functions use the soft pseudo labels generated by the Mean Net $E^{(T)}[\theta]$.

$$\mathcal{L}_{sid}(\theta) = -\frac{1}{N} \sum_{i=1}^N (C(F(x_i|E^{(T)}[\theta])) \cdot \log C(F(x_i|\theta))) \quad (4)$$

$$\mathcal{L}_{stri}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{bce}(\mathcal{T}_i(\theta), \mathcal{T}_i(E^{(T)}[\theta])) \quad (5)$$

where \mathcal{T}_i is the soft triplet labels generated by the Mean Net. We refer the readers to [6] for the detail definition of soft triplet loss.

The overall loss function $\mathcal{L}(\theta)$ combines four loss functions and is formulated as,

$$\mathcal{L}(\theta) = (1 - \lambda_{id}) \mathcal{L}_{id}(\theta) + \lambda_{id} \mathcal{L}_{sid}(\theta) + (1 - \lambda_{tri}) \mathcal{L}_{tri}(\theta) + \lambda_{tri} \mathcal{L}_{stri}(\theta) \quad (6)$$

where λ_{id} , λ_{tri} are the weighting parameters.

The detailed optimization process is summarized in Algorithm 1.

```

Require: Ensembling momentum  $\alpha$  for equation 1,
           weighting factors  $\lambda_{id}$ ,  $\lambda_{tri}$  for equation 6
Initialize  $\theta$  with ImageNet pre-trained ResNet-50;
Intra-tracklet isolation on raw tracklets;
for  $n$  in  $[1, num\_epochs]$  do
  Generate hard pseudo labels from inter-tracklet
  association;
  for each mini-batch, iteration  $T$  do
    Generate soft pseudo labels from the Mean
    Net;
    Update parameters  $\theta$  by the gradient descent
    of loss function equation 6;
    Update Mean Net weights following
    equation 1;
  end
end

```

Algorithm 1: Tracklet-based Multi-level Clustering (TMC) Training Pipeline

4. Experiments

4.1. Datasets

Video-based re-ID datasets Two publicly available video re-ID datasets are used in our experiments: PRID 2011 [7]

and MARS [28] dataset. The MARS dataset is the largest video dataset for the person re-ID task. The dataset contains 17,503 tracklets for 1,261 identities and 3,248 distractor tracklets, which are recorded by six cameras. This dataset is split into 625 identities for training and 636 identities for testing. The PRID 2011 dataset is collected from two cameras with significant color inconsistency. It contains 385 person tracklets in camera A and 749 person tracklets in camera B. Among all persons, 200 persons are captured in both camera views.

Simulation The need for simulated tracklets stems from two aspects. Firstly, video sequences in existing video-based re-ID dataset are clean and lack of noise, which makes them unrealistic. Using MOT algorithm to generate tracklets need original video as input, however the corresponding video for video re-ID dataset is generally unavailable. Secondly, by simulating tracklets we have more control over strength of different types of noise, giving us more insight on how different noise affect re-ID model training process.

Therefore, we propose a novel algorithm to simulate tracklets from video re-ID dataset. We first formulate noise strength as “rate of fragmentation” (r_{FM}) and “rate of switch” (r_{SW}). Given a collection of MOT-generated tracklets $\{S_1, S_2, \dots, S_N\}$ and the corresponding tracking ground-truth, one can use IOU matching to find P_{ij} , the ground-truth person ID for the j -th frame in tracklet S_i . Let L_i be the length of tracklet S_i , $Q_i = \{P_{ij} \mid j = 1, 2, \dots, L_i\}$ be the set of ground-truth person IDs in the tracklet. Assume the total number of ground-truth person IDs is M .

r_{FM} is defined to be the average number of tracklets each ground-truth person ID is in.

$$r_{FM} = \frac{1}{M} \sum_{k=1}^M \sum_{i=1}^N \mathbb{1}_{k \in Q_i} \quad (7)$$

r_{SW} is defined to be the average number of ground-truth person IDs each tracklet contains.

$$r_{SW} = \frac{1}{N} \sum_{i=1}^N |Q_i| \quad (8)$$

Given r_{FM} , r_{SW} , and the number of ground-truth person IDs, the number of tracklets with noise can be calculated. The exact number of person IDs contained by each noisy tracklet are determined by observed distribution. Person IDs are randomly assigned to noisy tracklets, after which the assignment of the pure tracklets and the “slots” in the noisy tracklets are randomly generated. Finally, the comprising pure tracklets within each simulated tracklet are shuffled.

Our generation algorithm has the ability to generate tracklets with various noise types (e.g., tracklets with ID

| No. | Name | r_{FM} | r_{SW} | #Tracklets |
|-----|--------------|----------|----------|------------|
| 1 | MARS_1.7_1.2 | 1.7 | 1.2 | 2835 |
| 2 | MARS_2.5_1.2 | 2.5 | 1.2 | 4023 |
| 3 | MARS_2.5_1.5 | 2.5 | 1.5 | 3272 |
| 4 | MARS | - | - | 8298 |

Table 1: Summary of generated simulation datasets

switch and fragmentation at the same time, tracklets with multiple ID switches). Please refer to the supplementary material for further details of the generation algorithm.

Videos with static camera from the training set of MOT17 [12] were used for estimating realistic values of r_{FM} and r_{SW} . Two MOT algorithms, DeepSORT [18] and FairMOT [26] were picked as representatives for baseline method and recent SotA method, respectively. For realistic upper bounds of r_{FM} and r_{SW} , the most difficult video (per person density provided by MOT17) and the less performant algorithm, DeepSORT, was used, yielding an r_{FM} of value 1.7 and r_{SW} of 1.2. For lower bounds, the easiest video (same above) and the better algorithm, FairMOT, was used, giving $r_{FM} = 2.5$, $r_{SW} = 1.2$. For the sake of comparison, we added another set of parameters $r_{FM} = 2.5$, $r_{SW} = 1.5$. Three MARS-derived datasets was generated based on these three set of (r_{FM}, r_{SW}) values, summarized in Table 1. For ease of reference, we name these datasets by their generating parameters (r_{FM}, r_{SW}), e.g. MARS_1.7_1.2.

4.2. Evaluation Protocol

For PRID 2011, we use the Cumulative Matching Characteristic (CMC) curve to evaluate the performance of each method. Both the averaged CMC and the mean Average Precision (mAP) are used to measure re-ID performance on MARS. Note, our method does not utilize any ID labels for model initialisation or training.

4.3. Implementation Details

Intra-tracklet clustering In this step, we utilize DBSCAN clustering algorithm to cluster the features extracted by ResNet-50 model initialized on ImageNet. The hyper-parameters eps is set to 0.6. Since ImageNet pre-trained model can already obtain good clustering results, in order to improve training efficiency, intra-tracklet isolation is only executed once before the end-to-end training process.

Inter-tracklet clustering For each tracklet generated in the first step, 32 consecutive images were randomly sampled, followed by feature extraction using up-to-date model. Then we utilize DBSCAN clustering algorithm to cluster the features to generate hard pseudo labels. The hyper-parameter eps is set to data-dependent. Inter-tracklet as-

| eps_{inter} | MARS | |
|---------------|-------------|-------------|
| | mAP | rank-1 |
| 0.600 | 27.4 | 33.6 |
| 0.650 | 26.2 | 32.9 |
| 0.670 | 57.9 | 70.4 |
| 0.676* | 61.8 | 72.4 |
| 0.680 | 57.1 | 69.2 |
| 0.700 | 56.4 | 68.3 |

Table 2: Ablation studies on the value of eps_{inter} in inter-tracklet clustering. * denotes eps_{inter} value calculated by data-dependent policy.

sociation is executed before each epoch to update the hard pseudo labels.

End-to-end training ResNet-50 backbone was used, initialized by weights pre-trained on ImageNet. The network is updated by optimizing loss with the loss weight $\lambda_{id} = 0.5$, $\lambda_{tri} = 0.8$. Adam optimizer is adopted with a weight decay of 0.0005. The temporal ensemble momentum α is set to 0.999. The learning rate is set to 0.00035 for 40 epochs.

4.4. Ablation Studies

Necessity of noise reduction methods We use intra-tracklet isolation and inter-tracklet association to reduce the ID switch and fragmentation. The necessity of inter-tracklet association has been proved in [6]. To investigate the contribution of intra-tracklet isolation, comparative experiments are conducted on MARS and the three simulated datasets, each with and without intra-tracklet isolation. As illustrated in Table 3, the experiments without intra-tracklet isolation result in much lower performances on the noisy datasets. Under the interference of ID switch, each tracklet could contain more than one person, so the hard pseudo label generated from part of the tracklet can not represent the whole tracklet. For dataset MARS without noise, the two results almost equal. In addition, the best results of the three noisy datasets are similar, proving that our method can handle varying degrees of noise.

Effectiveness of the clustering eps The two eps values used in intra- and inter-tracklet clustering are closely related. We denote the values eps_{intra} and eps_{inter} , respectively. The images eps_{intra} operates on are consecutive frames of a tracklet, therefore the images are much more similar and from a small number of distinct ground-truth IDs. eps_{inter} , however, operates on all tracklets, which are less similar and from a large number of IDs.

As illustrated in Table 2, the performance is sensitive to the eps_{inter} value chosen. The best performance is achieved when eps_{inter} is calculated from data distribution.

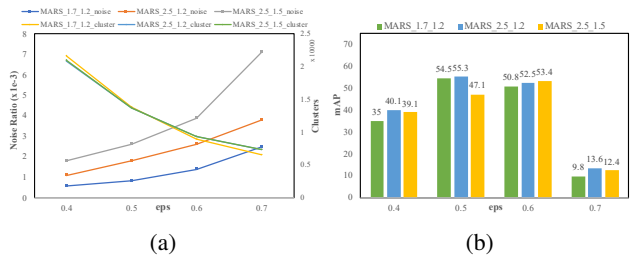


Figure 7: Experiments on eps value for the intra-tracklet clustering. (a) Noise ratio is larger and the number of total clusters generated is smaller with larger eps value. (b) An eps of value too small or too large results in less performant model learned.

For eps_{intra} of intra-tracklet isolation, we conducted experiments on the three noisy datasets using four different eps_{intra} values. In order to ensure that different tracklets use the same clustering standard, we set the eps_{intra} to a fixed value instead of data-dependent. As illustrated in Table 4, The most appropriate eps_{intra} for MARS_1.7_1.2 and MARS_2.5_1.2 is 0.5. the highest mAP value could reach 54.5% and 55.3%. 0.6 is the most appropriate eps_{intra} value for MARS_2.5_1.5, the mAP of which is 53.4%. **1.** As shown in figure 7 (a), for a larger eps_{intra} value, the density requirements for a cluster to be formed are weaker. Therefore, the total number of clusters generated is smaller, and the noise ratio of clusters generated is higher. **2.** For the three simulated noisy datasets, as shown in figure 7 (b), if eps_{intra} is too small, the total number of clusters would be too large, presenting difficulties during the inter-tracklet association step. On the other hand, an eps_{intra} value that is too large would weaken the effect of noise reduction in intra-tracklet isolation, hindering the performance of trained model. An appropriate eps_{intra} value gives the right balance between the two effects, yielding relatively higher mAP and rank-1. **3.** eps_{inter} operates on all tracklets, which are less similar and from a large number of IDs. Moreover, it is generally preferable to have ID fragmentation noise rather than ID switch in intra-tracklet isolation result, since the former can be corrected in the subsequent inter-tracklet association step, but the latter cannot. This intuition coincides with experiment outcomes in figure 7 (b), where an eps_{intra} smaller than optimal value causes smaller drop in performance than an eps_{intra} larger than optimal value. We conclude that eps_{intra} should be smaller than eps_{inter} .

4.5. Comparisons with State-of-the-Art

We compare our proposed method with six state-of-the-art methods on two video re-ID datasets, PRID 2011 and large-scale MARS dataset. The results are shown in Table 5. As for the two clean datasets without noise, our method significantly outperforms all existing video unsupervised and

| Intra-tracklet Isolation | Inter-tracklet Isolation | MARS_1.7_1.2 | | MARS_2.5_1.2 | | MARS_2.5_1.5 | | MARS | |
|--------------------------|--------------------------|--------------|--------|--------------|--------|--------------|--------|------|--------|
| | | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 |
| N | Y | 8.0 | 17.3 | 14.4 | 22.0 | 11.8 | 22.7 | 61.8 | 72.4 |
| Y | Y | 54.5 | 67.3 | 55.3 | 68.2 | 53.4 | 63.7 | 60.7 | 72.0 |

Table 3: Ablation studies on intra-tracklet isolation. The experiments without intra-tracklet isolation result in much lower performances on the noisy datasets. For dataset MARS without noise, the two results almost equal.

| eps_{intra} | MARS_1.7_1.2 | | | | MARS_2.5_1.2 | | | | MARS_2.5_1.5 | | | |
|---------------|--------------|-----------|-------------|-------------|--------------|-----------|-------------|-------------|--------------|-----------|-------------|-------------|
| | noise | #clusters | mAP | rank-1 | noise | #clusters | mAP | rank-1 | noise | #clusters | mAP | rank-1 |
| 0.4 | 0.57 | 21720 | 35.0 | 48.8 | 1.1 | 20919 | 40.1 | 53.8 | 1.8 | 20958 | 39.1 | 52.5 |
| 0.5 | 0.84 | 13847 | 54.5 | 66.2 | 1.8 | 13673 | 55.3 | 68.2 | 2.6 | 13689 | 47.1 | 57.7 |
| 0.6 | 1.4 | 8963 | 50.8 | 62.0 | 2.6 | 9383 | 52.5 | 66.5 | 3.9 | 9401 | 53.4 | 63.7 |
| 0.7 | 2.5 | 6541 | 9.8 | 15.7 | 3.8 | 7352 | 15.6 | 20.1 | 7.1 | 7306 | 12.4 | 19.9 |

Table 4: Ablation studies on the value of eps_{intra} . Experiments were conducted on noisy datasets. eps_{intra} is calculated by data-dependent policy for all experiments. “Noise” is the average noise ratio of the tracklets produced by intra-tracklet isolation. #Clusters is the total number of tracklets produced.

| Dataset | PRID 2011 | | | MARS | | | mAP |
|------------------------|-------------|------|------|-------------|------|------|-------------|
| | 1 | 5 | 10 | 1 | 5 | 10 | |
| DAL [1] | 85.3 | 97.0 | 98.8 | 46.8 | 63.9 | 71.6 | 21.4 |
| RACE [22] | 50.6 | 79.4 | 88.6 | 41.0 | 55.6 | 62.2 | 22.3 |
| DGM [24] | 61.6 | 89.0 | 94.8 | 43.8 | 60.1 | 67.4 | 24.0 |
| DGM+ [23] | 62.7 | 90.8 | 96.0 | 48.1 | 64.7 | 71.1 | 29.2 |
| UTAL [10] | 54.7 | 83.1 | 96.2 | 49.9 | 66.4 | 77.8 | 35.2 |
| EUG [21] | - | - | - | 62.7 | 74.9 | 82.6 | 42.5 |
| TMC- | 67.4 | 82.0 | 93.3 | 72.4 | 85.9 | 89.3 | 61.8 |
| TMC | 68.0 | 82.0 | 93.5 | 72.0 | 84.2 | 88.2 | 60.7 |
| STAN [†] [11] | 90.3 | 98.2 | - | 82.3 | - | - | 65.8 |
| SDM [†] [25] | 85.2 | 97.1 | - | 71.2 | 85.7 | - | - |

[†] Supervised method.

Table 5: Comparisons with state-of-the-art. We compare our proposed method with six unsupervised state-of-the-art methods and two supervised methods on two video re-ID datasets. “TMC-” is our method with intra-tracklet isolation step omitted.

tracklet-based approaches on MARS, which is the dataset closest to the realistic problem. Our method even achieved competitive results against some supervised methods. Our method ranked second on PRID 2011, which we presume to be caused by the size of PRID 2011. The dataset contains relatively small number of IDs and only two tracklets are provided for each ID. This creates gap between the dataset and realistic scenarios, limiting the performance of our method.

On MARS, the intra-tracklet isolation step in TMC might split the already-clean tracklets, having negative impact on inter-tracklet association. Therefore, results with

full TMC is slightly worse than TMC with intra-tracklet isolation removed.

In addition, the method also shorten the gap in performance between supervised and unsupervised methods. Apart from this, as shown in Table 4, the proposed method could achieve 55.3% mAP, 68.2% rank-1 on realistic tracklets with noise. Even on the tracklets with strongest noise, our method could achieve mAP 53.4% and rank-1 63.7%, outperforming existing unsupervised video re-ID methods even with clean tracklets. The results show that our framework outperforms existing unsupervised and weakly-supervised video re-ID methods in real-life scenarios.

5. Conclusion

In this work, we propose a clustering and fine-tuning framework to learn a person re-ID model from tracklets automatically generated from MOT algorithms. This is a setting that was less studied but closer to realistic problems. Our framework requires no human effort for labeling and also independent from data or model from other domains. Characteristics of dominant noise types of MOT tracklets, i.e. ID fragmentation and ID switch, are analyzed and utilized in the noise reduction process. Extensive experiment results on MARS and simulated tracklets of various noise show that our framework outperforms all existing unsupervised and weakly-supervised video person re-ID methods.

6. Broader Impact

Over the past few years, the technology of person re-ID has been deployed to benefit the society in areas such as intelligent security, smart city, and smart retail. It should

be notes that, when applied maliciously, person re-ID may cause the infringement of privacy or even crime. Governments and facilities are responsible to control the usage of this technology by legislation.

References

- [1] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. *arXiv preprint arXiv:1808.07301*, 2018.
- [2] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.
- [3] Patrick Dendorfer, Hamid Rezaatofghi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes, 2020.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2019.
- [7] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.
- [8] Frank Klinker. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58(1):97–107, 2011.
- [9] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [10] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [11] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.
- [12] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
- [13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [14] Chirine Riachy, Fouad Khelifi, and Ahmed Bouridane. Video-based person re-identification using unsupervised tracklet matching. *IEEE Access*, 7:20596–20606, 2019.
- [15] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.
- [16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [17] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer, 2014.
- [18] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. pages 3645–3649, 2017.
- [19] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI*, pages 12362–12369, 2020.
- [20] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wei Bian, and Yi Yang. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing*, 28(6):2872–2881, 2019.
- [21] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.
- [22] Mang Ye, Xiangyuan Lan, and Pong C Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–186, 2018.
- [23] Mang Ye, Jiawei Li, Andy J Ma, Liang Zheng, and Pong C Yuen. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Transactions on Image Processing*, 28(6):2976–2990, 2019.
- [24] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5142–5150, 2017.
- [25] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6781–6789, 2018.
- [26] Yifu Zhang, Chunyu Wang, Xinggong Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking, 2020.
- [27] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.
- [28] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.
- [29] Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification: A new benchmark. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.