Paper Reading

2018-11-24

谢乔康

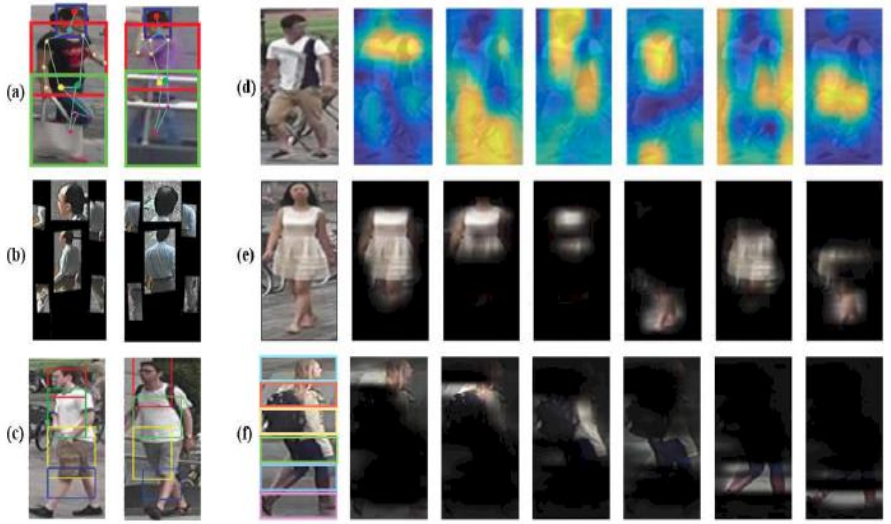
Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

Yifan Sun[1], Liang Zheng[2], Yi Yang[3], Qi Tian[4], and Shengjin Wang[1⋆]

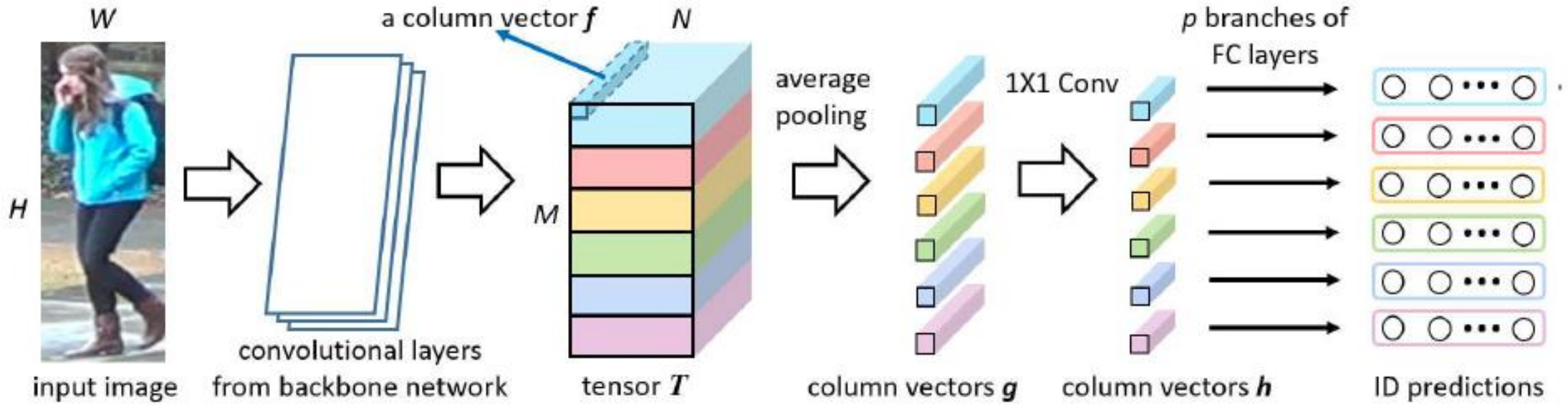
[1] Department of Electronic Engineering, Tsinghua University, China
[2] Research School of Computer Science, Australian National University, Australia
[3] Centre for Artificial Intelligence, University of Technology Sydney, Australia
[4] (1) Huawei Noah's Ark Lab (2) University of Texas at San Antonio
sunyf15@mails.tsinghua.edu.cn, wgsgj@tsinghua.edu.cn

- Motivation
  - A prerequisite of learning discriminative part features is that parts should be precisely located. Various strategies have been employed for accurate part discovery.
  - Rethink the problem of what makes well-aligned parts
    - Partitions based on pose estimation or human parsing may offer stable cues to good alignment but are prone to noisy pose detections.
    - This paper speculate that the consistency of the context within each part is vital to precise partition.
    - So given coarsely partitioned parts, e.g., the uniform stripes, they aim to refine them by reinforcing within-part consistency.



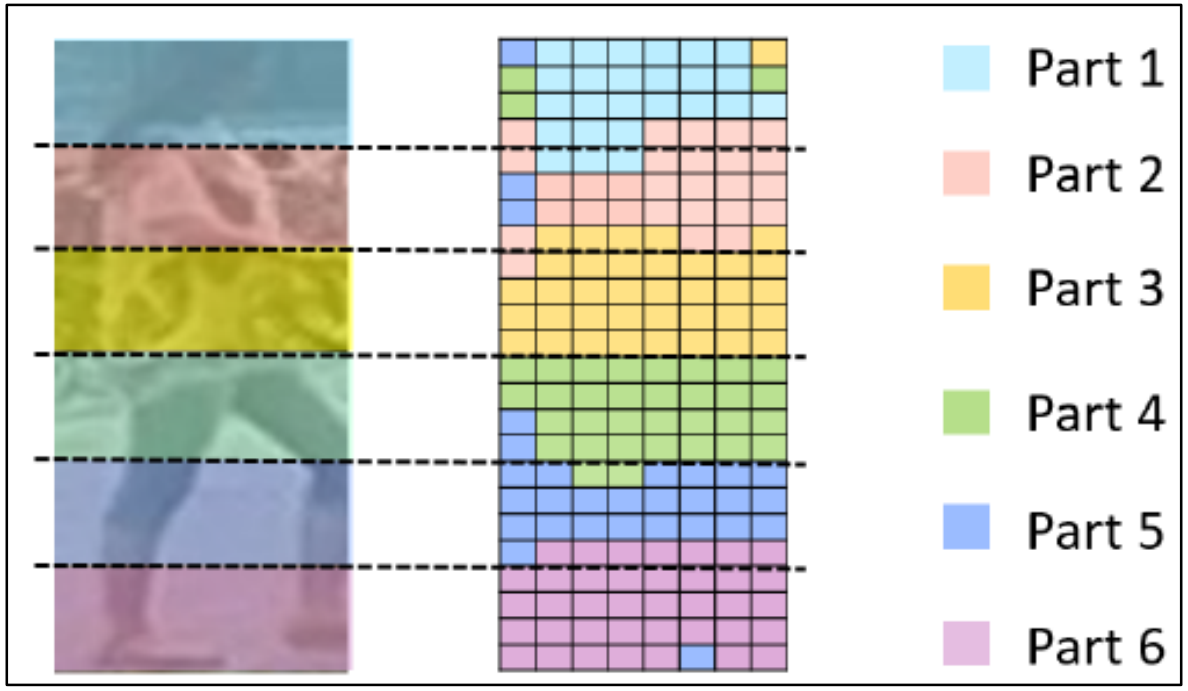
Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

- Part-based Convolutional Baseline (PCB)
  - PCB employs uniform partition on the feature maps
  - **Training:** each branch of the part features is supervised by the ID labels, respectively.
  - **Testing:** all the part features are concatenated to form the learned descriptor.
  - PCB already achieves state of the art on several re-ID benchmarks.



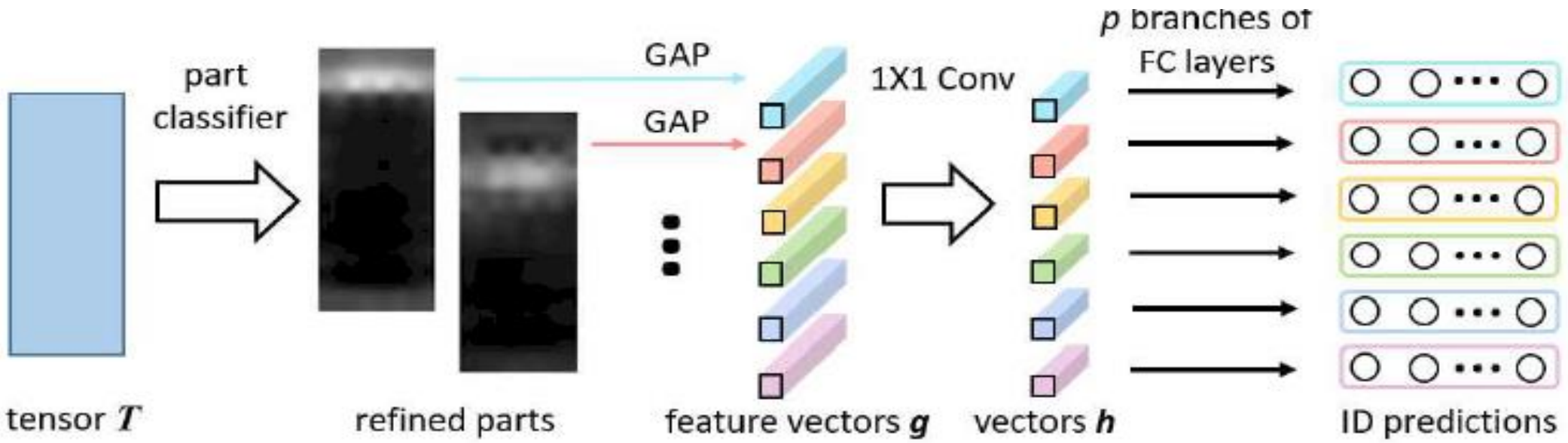
Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

- Partition errors
  - Some column vectors, while designated to a specified part during training, are more similar to another part after the model converges. The existence of these outliers indicates inappropriate partion.



Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

- Refined Part Pooling (RPP)
  - First predicts the similarities between a column vector and all the parts.
  - Then assigns the column vector to each part with corresponding similarity value as the weights.
  - The key point of RPP is to train a part classifier which predicts the similarity between column vectors and all the parts. The training requires no part labels and is induced by the knowledge learned from uniformly partitioned parts.



Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)

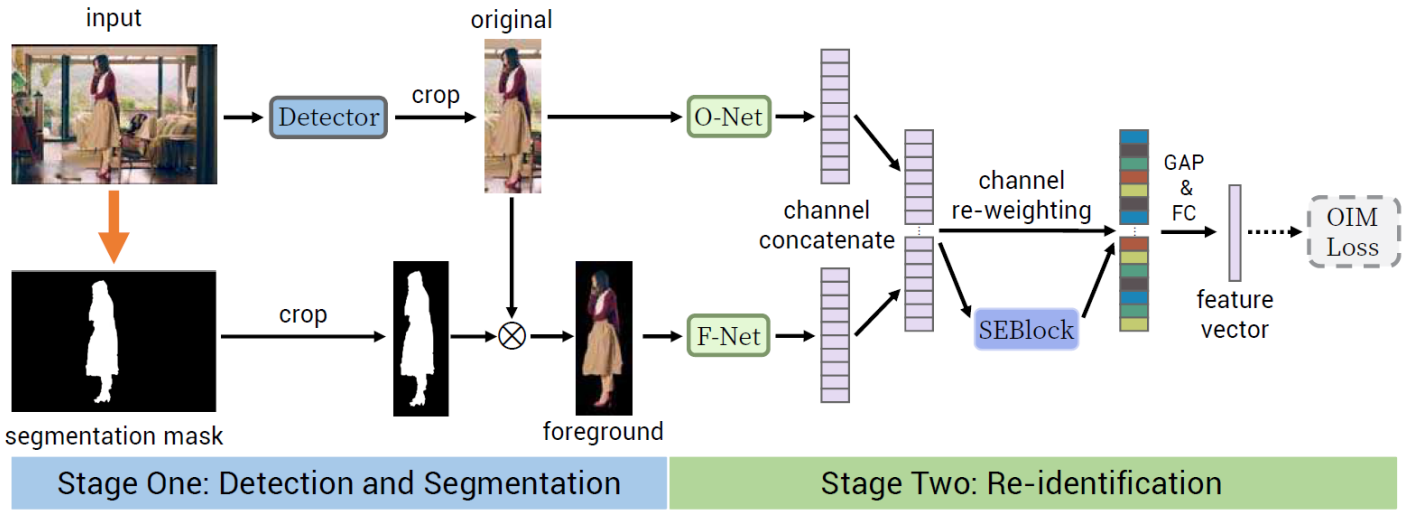
Algorithm 1: Induced training for part classifier

Step 1. A standard PCB is trained to convergence with uniform partition.
Step 2. A p-category part classifier is appended on the tensor $\boldsymbol{T}$.
Step 3. All the pre-trained layers of PCB are fixed. Only the part classifier is trainable. The model is trained until convergence again.
Step 4. The whole net is fine-tuned to convergence for overall optimization.

- Contributions
  - A very concise Part-based Convolutional Baseline (PCB) which achieves state of the art on re-ID simply employing uniform partition on feature maps.
  - Refined Part Pooling (RPP) to reduce partition errors, which requires no auxiliary part labels and allows PCB to gain another round of performance boost.

Person Search via A Mask-guided Two-stream CNN Model

- Motivation
  - It is not appropriate to share representations between the detection and re-ID tasks, as their goals contradict with each other.
  - It is more suitable to consider a compromised strategy of paying extra attention on the foreground person while also using the background as a complementary cue.
- Mask-guided Two-Stream CNN Model



- Two stages are trained **separately**
- RoI expansion by a ratio γ is conducted while cropping proposals
- **Detector:** Faster R-CNN based on VGG16
- **Segmentation Mask:** FCIS pre-trained on COCO
- **O-Net and F-Net:** ResNet-50

Person Search via A Mask-guided Two-stream CNN Model

- Separation > Integration

Method	Joint	AP(%)	Recall(%)
OIM-ours	✓	69.5	75.6
CNN [3]	✗	78.0	75.7

Method	Joint	mAP(%)	top-1(%)
GT + OIM [3]	✓	77.9	80.5
GT + IDNetOIM	✗	78.5	81.7

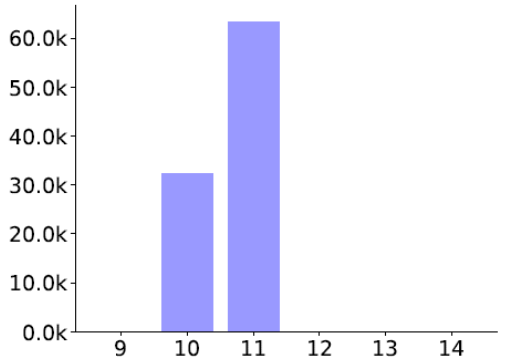
- Visual Component Study

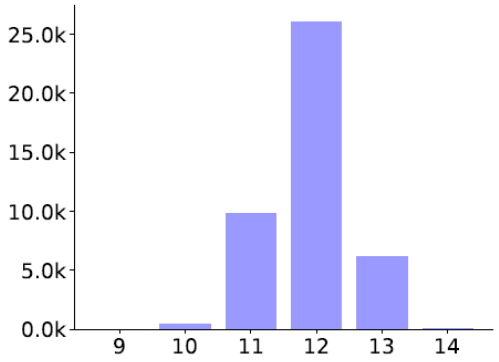
O	F	B	E	mAP(%)	top-1(%)
✓				78.5	81.7
	✓			75.3	78.7
		✓		34.2	35.9
✓			✓	77.7	81.1
		✓	✓	38.7	40.0
✓	✓		✓	89.1	90.0

- O: Original image
- F: Forground person only
- B: Background only
- E:Expand RoI by a ratio of γ
  - Hard discarding BG hurts
  - Hard expansion on BG hurts
  - Two-stream modeling boosts a lot

SEBlock Weights Inspection

- Average weights for sample i:
  - $Avg_i(F) > Avg_i(O)$
- Number of F stream weights among the top 20: $N^{20}(F)$
  - Most information cues are from the foreground patch
  - Context information contained in the original image patch is helpful



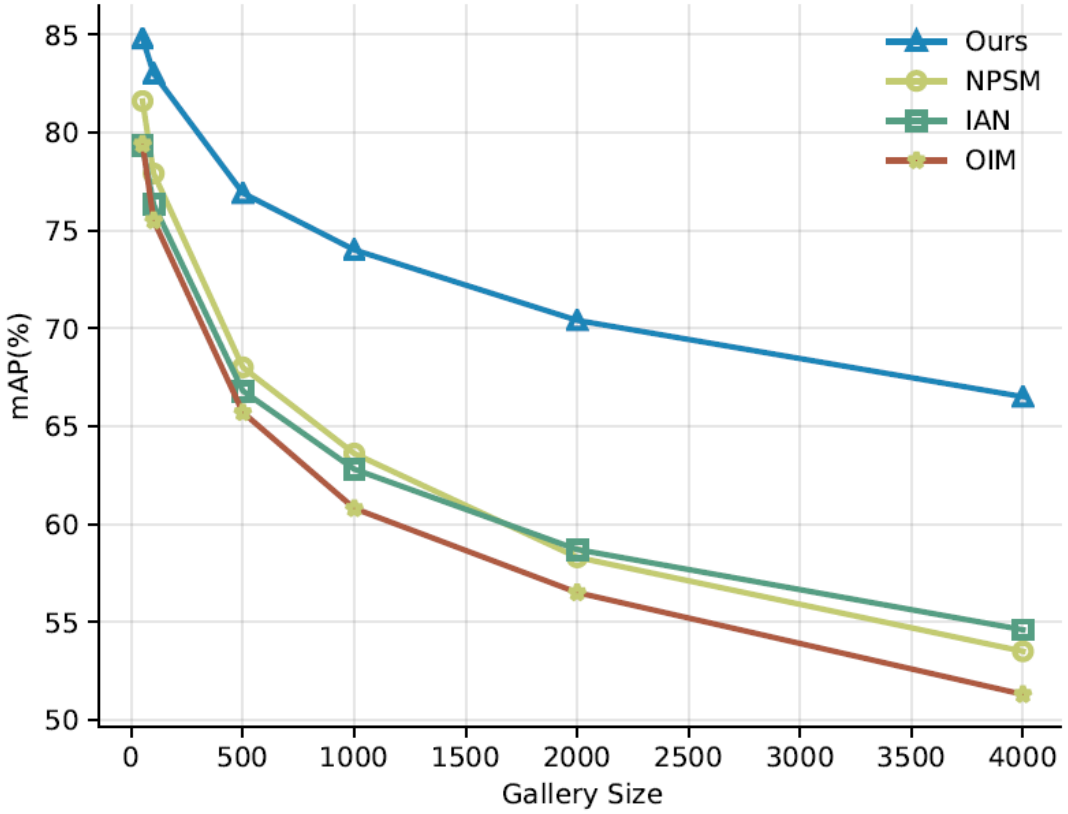


Person Search via A Mask-guided Two-stream CNN Model

- Comparison with State-of-the-Art Methods
  - Comparison of results on CUHK-SYSU with gallery size of 100

Method	mAP(%)	top-1(%)
CNN + DSIFT + Euclidean	34.5	39.4
CNN + DSIFT + KISSME	47.8	53.6
CNN + BoW + Cosine	56.9	62.3
CNN + LOMO + XQDA	68.9	74.1
CNN + IDNet	68.6	74.8
OIM [3]	75.5	78.7
IAN [4]	76.3	80.1
NPSM [5]	77.9	81.2
Ours(CNN_v + IDNetOIM)	75.8	79.5
Ours(CNN_v + MGTS)	**83.0**	**83.7**

  - Performance comparison on CUHK-SYSU with varying gallery sizes

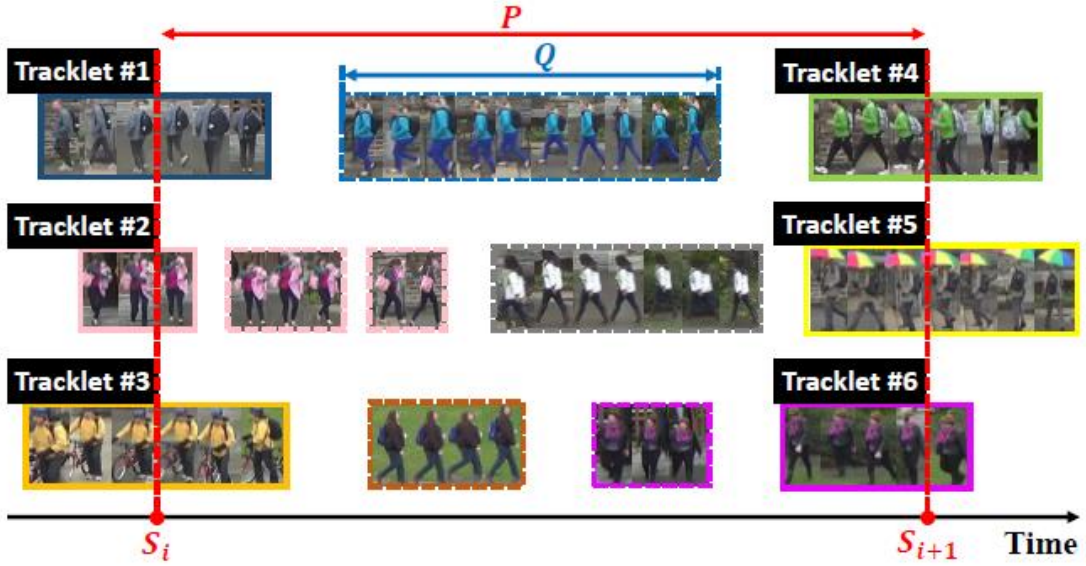


Unsupervised Person Re-identification by Deep Learning Tracklet Association

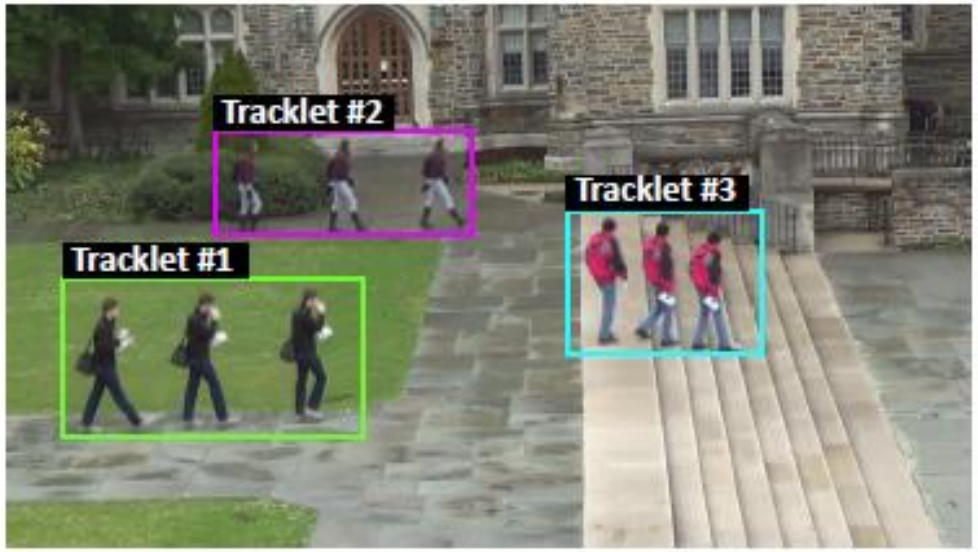
- Limitation of existing methods
  - Supervised learning, unscalable due to the need for exhaustive manually labelled ID matching pairs for every camera pair of every target camera network
- Key Idea
  - Unsupervised deep learning of auto-extracted person tracklet data
  - Self-discover person re-id knowledge in tracklets across cameras
- Contributions
  - **Tracklet Association Unsupervised Deep Learning (TAUDL)**
    - Per-Camera Tracklet Discrimination learning (PCTD)
    - Cross-Camera Tracklet Association learning (CCTA)
  - **Sparse Space-Time Tracklet** sampling
    - Minimise per-camera tracklet ID duplication to support TAUDL

Unsupervised Person Re-identification by Deep Learning Tracklet Association

- Sparse Space-Time Tracklet Sampling (SSTT)
  - **Temporal** sampling gap P > the view transit time Q
  - Tracklets **spatially** far away to each other



(a) Temporal sampling



(b) Spatial sampling

Unsupervised Person Re-identification by Deep Learning Tracklet Association

- Approach Overview
  - Multi-camera multi-task deep learning of tracklet labels

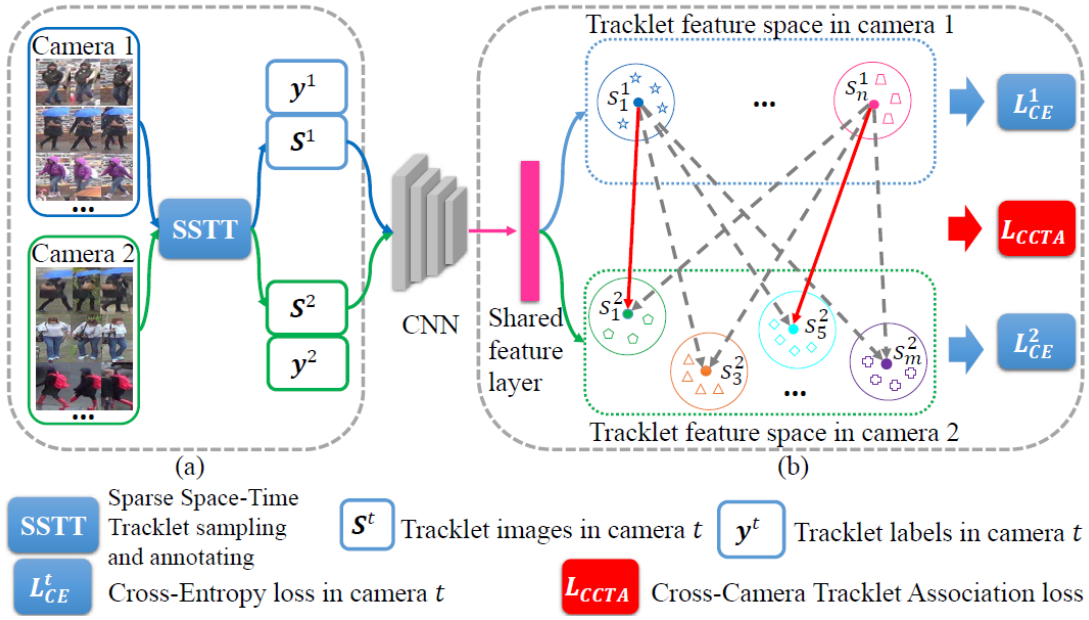


Fig. 1. An overview of Tracklet Association Unsupervised Deep Learning (TAUDL) re-id model: (a) Per-camera unsupervised tracklet sampling and label assignment; (b) Joint learning of both within-camera tracklet discrimination and cross-camera tracklet association in an end-to-end global deep learning on tracklets from all the cameras.

Loss Functions

- Per-Camera Tracklet Discrimination (PCTD)

$$\mathcal{L}_{\text{ce}} = -\log\Big(\frac{\exp(\boldsymbol{W}_y^\top \boldsymbol{x})}{\sum_{k=1}^{M_t}\exp(\boldsymbol{W}_k^\top \boldsymbol{x})}\Big),$$

$$\mathcal{L}_{\text{pctd}} = \frac{1}{N_{\text{bs}}}\sum_{t=1}^{T}\mathcal{L}_{\text{ce}}^t,$$

- Cross-Camera Tracklet Association (CCTA)

$$\mathcal{L}_{\text{ccta}} = -\log\frac{\sum_{\boldsymbol{z}_k\in\mathcal{N}_i^t}\exp(-\frac{1}{2\sigma^2}\parallel \boldsymbol{s}_i^t - \boldsymbol{z}_k \parallel_2)}{\sum_{t'=1}^{T}\sum_{j=1}^{n_j}\exp(-\frac{1}{2\sigma^2}\parallel \boldsymbol{s}_i^t - \boldsymbol{s}_j^{t'} \parallel_2)},$$

- Joint Loss Function

$$\mathcal{L}_{\text{taudl}} = (1-\lambda)\mathcal{L}_{\text{pctd}} + \lambda\mathcal{L}_{\text{ccta}},$$