**Deep Association Learning for Unsupervised Video Person Re-identification**

Yanbei Chen\(^1\)  Xiutian Zhu\(^2\)  Shaogang Gong\(^1\)

yanbei.chen@qmul.ac.uk  eddy@visionsemantics.com  s.gong@qmul.ac.uk

\(^1\)Queen Mary University of London  \(^2\)Vision Semantics Ltd., London, UK

---

**Introduction**

**Video Person Re-identification (ReID)**

A task to match person identities (ID) in the video tracklets (sequences) captured from disjoint surveillance camera views.

**Unsupervised Video Person ReID**

- **Problem**
  - How to formulate supervision signals without utilizing any pairwise ID matching information to guide model learning?

- **Main Contributions**
  - End-to-end deep unsupervised learning framework for video ReID: none manual labelled supervision is given for model training.
  - State-of-the-art results on various video ReID benchmark datasets.

---

**Methodology Overview**

**Deep Association Learning**

- **Overall idea**
  - Learning by (1) intra-camera image-to-tracklet association and (2) cross-camera tracklet-to-tracklet association

**Two types of association learning**

1) **Intra-Camera Association Learning**
   - image-to-tracklet association under the same camera view
   - exploit the inherent label information for supervision

2) **Cross-Camera Association Learning**
   - tracklet-to-tracklet association across disjoint camera views
   - mitigate the cross-camera domain gaps

---

**Intra-Camera Association Learning**

**Key Idea:**

intra-camera image-to-tracklet association

**Local Space-Time Consistency**

**Batch-wise three iterative steps**

1) **Learning Intra-Camera Anchors**
   - Represent each tracklet as a learnable anchor
   
   \[ x_{k,i}^{t+1} = x_{k,i}^t - \eta \cdot (f_2(x_{k,i}^t) - f_1(x_{k,i}^t)) \]
   
   if \( i = p \)

2) **Tracklet Association Ranking**
   - Rank all anchors in each camera view
   
   \[ D_{p,i} = \min_{i \in [1,N_i]} \| f_2(x_{k,i}^t) - f_1(x_{k,i}^t) \|_2 \]

3) **Intra-Camera Association Loss**
   - Associate each frame with its own source tracklet
   
   \[ L_i = \begin{cases} 
   [D_{p,i} - D_{p,j} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\
   [D_{p,i} - D_{j,i} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} 
   \end{cases} \]

---

**Cross-Camera Association Learning**

**Key Idea:**

cross-camera tracklet-to-tracklet association

**Batch-wise three iterative steps**

1) **Cyclic Ranking**
   - Discover highly associated tracklets across disjoint camera views
   
   \[ D_{c,i,j} = \min_{i \in [1,N_i]} \| f_2(x_{k,i}^t) - f_1(x_{k,j}^t) \|_2 \]

2) **Learning Cross-Camera Anchors**
   - Merge two highly associated tracklets as a cross-camera anchor
   
   \[ x_{k,i,j}^{t+1} = \frac{1}{2} \cdot \begin{cases} 
   f_2(x_{k,i}^t) + f_1(x_{k,j}^t) & \text{if } j = i \text{ (Cyclic ranking consistent)} \\
   x_{k,i}^t, & \text{others} 
   \end{cases} \]

3) **Cross-Camera Association Loss**
   - Associate each frame with the best-matched cross-camera anchor
   
   \[ L_c = \begin{cases} 
   [D_{p,i} - D_{p,j} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\
   [D_{p,p} - D_{j,i} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} 
   \end{cases} \]

---

**Experiments & Ablation Studies**

**Comparison to state-of-the-art unsupervised video ReID methods**

**Evolution of cross-camera tracklet association**

(a) Evolution on association rate.  
(b) Evolution on true-match rate.