Building a More Discriminative Deep Feature Space for Person Re-Identification

(submitted to IEEE TIP)

Alessandro Borgia – HWU & UoE

Prof Neil Robertson – QUB

Highlights

- Person Re-ID: context and motivation
- State-of-the-art results in the field
- Our proposed approach: revisiting metric learning technique
- Performance
- Advantages and limitations

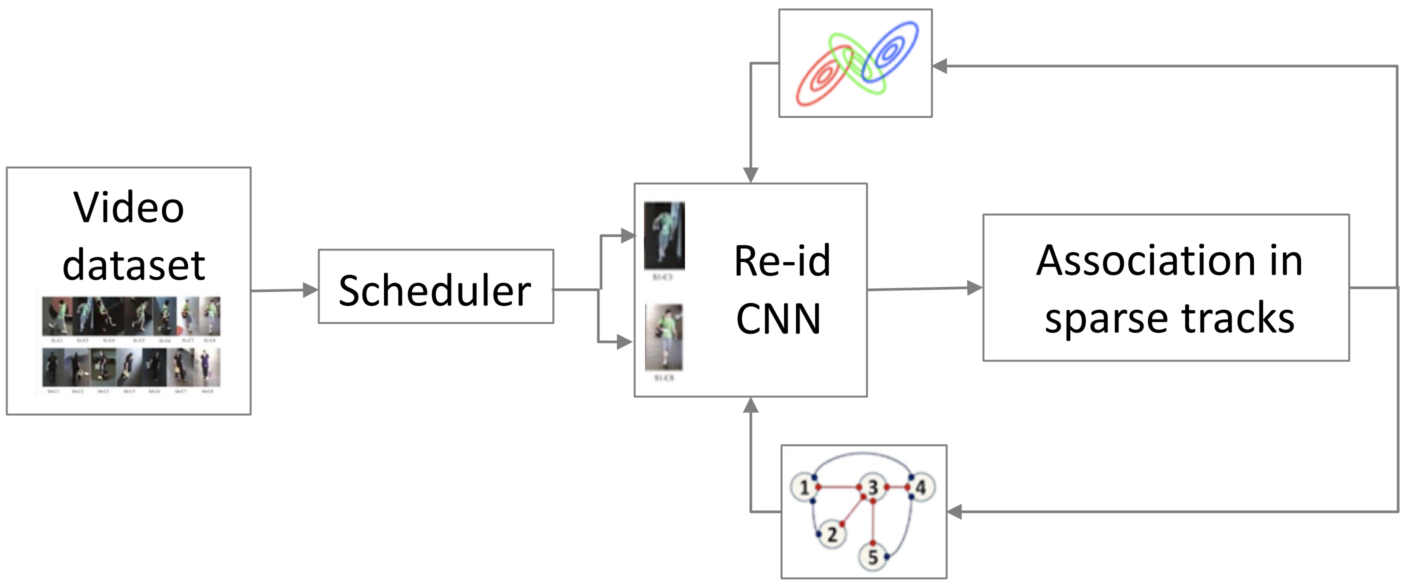
Motivation
and investigated context

Motivation

- **245 million** surveillance cameras active and operational globally (HIS, 2014)
- CCTV cameras on Britain's roads capture **26 million images every day** (The Guardian, Jan 23, 2014)
- London's subway attacks on July 7, 2005: It took investigators **thousands hours** to parse the city's CCTV footage (CNN, April 27, 2013)

Motivation

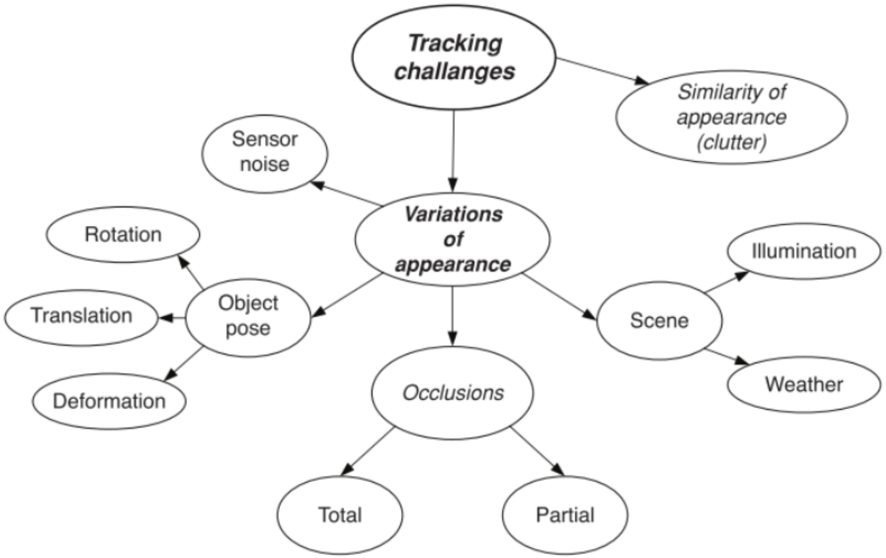
- Re-id capability critical when tracking across cameras
- Changing viewpoint: severe problem for re-id in multi-camera networks
- Deep learning pradigm
- Re-id evaluation following a ranking approach



Investigated context

- Investigated context
  - outdoor wide area surveillance network
  - non-calibrated, non overlapping CCTV cameras
  - unknown, unconstrained topology

- Factors affecting re-identification
  - lightings
  - **viewpoints**
  - poses
  - misalignments due to imperfect detections
  - long occlusions



Viewpoint problem

Viewpoint problem

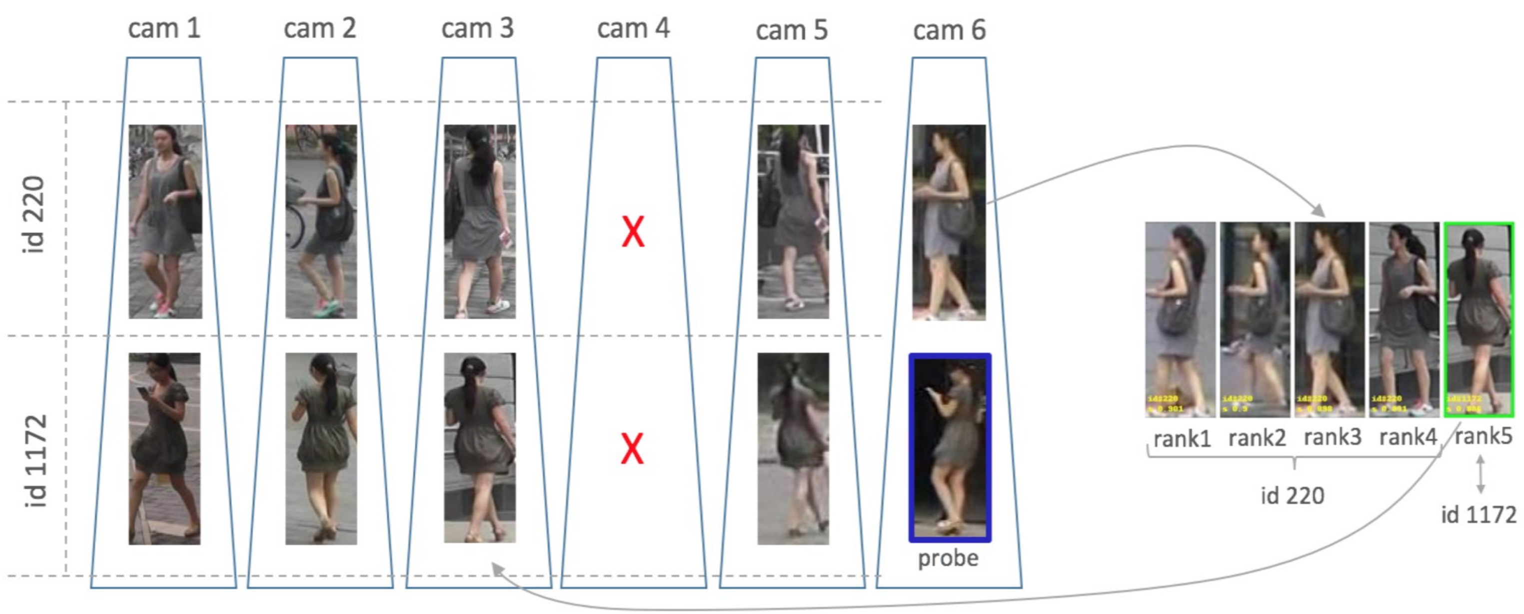
- Viewpoint variability effect:

$id_1\{I_{1A}, I_{1B},\ldots\} \rightarrow net(I_{1A}) = F_{1A}, \quad net(I_{1B}) = F_{1B}$

$id_2\{I_{2A},\ldots\} \rightarrow net(I_C) = F_{2A}$

$dist(F_{1A}, F_{2A}) < dist(F_{1A}, F_{1B}) \rightarrow$ wrong ranking event

- Quite recurrent when only softmax supervision is used



More examples of the viewpoint problem

What to do?

Options from the literature:

1. Feature design (net structure)
2. Side information (target allignment, pose estimation,…)
3. Metric learning (over the learned space)

- Advantages: - flexible applicability, network structure independent
  - better exploitation of available row data
- Limitation: operates on networks with fixed weights

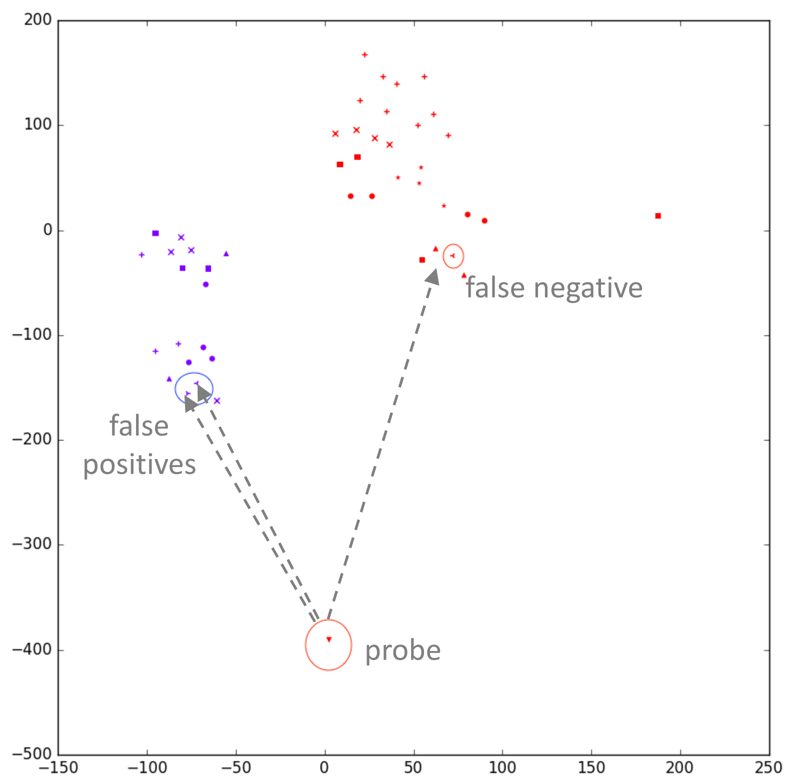
Our approach:

- Extending ML to the CNN training stage making it contextual with feature learning
- Influencing the construction itself of the featues space according to a metric, instead of just learning the metric afterwards disjointly, aiming to get:
  - increased inter-class separability
  - more discriminative features → intra-class compactness

<u>Improving the training objective</u>

Negative Euclidean distance match

- T-sne visualization tool [6]
- Red points → identity #1 Blue points → identity #2
  (different markers correspond to different cameras)
- The probe and the false positives share the same camera view



Datasets and CNN structure

Datasets

CUHK03 [7]

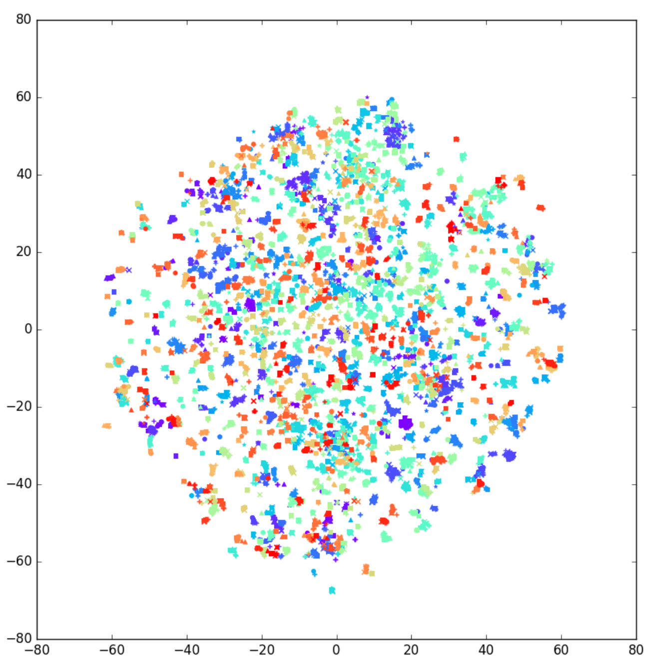
- 1360 identities (1160 for training, 100 for validation)
- Up to 10 imgs/id
- Each id seen under 1 pair of cameras (max 5 shots/cam)
- 3 camera pairs overall
- Reproduced setting/results in [7]: 20 test-sets, 100 imgs/testset

MARKET-1501 [5]

- Reproduced setting/results in [5]
- 1501 identities (751 for training)
- Up to 70 imgs/id
- Each id seen under up to 6 views
- Training set: 12936 imgs
- Test set: 13115 (including 2798 distractors)
- Query set: 3365 imgs belonging to 750 test ids (1shot/cam)

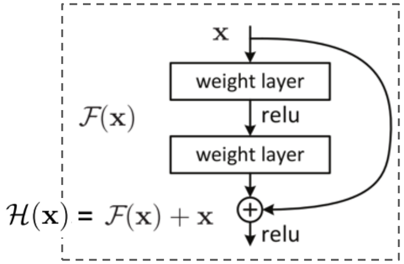
How it looks like

- Training set of Market-1501 dataset: 751 ids
- T-sne: visualization tool technique for the visualization of similarity data
    - Retains local structure of data
    - Reveals some important global structure (clusters at multiple scales)

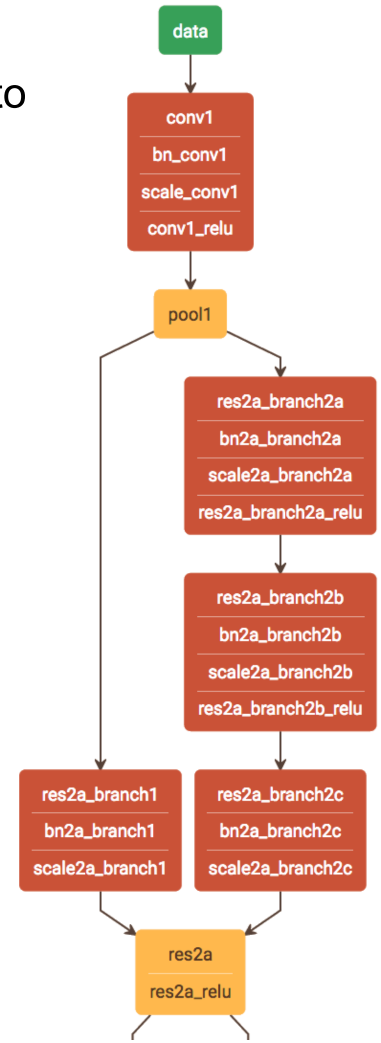


CNN structure

- ResNet50
- Addresses the performance degradation problem due to CNN depth
- Forces layers to fit a residual mapping $\mathcal{H}(\mathbf{x})$
- Dim. softmax output: CUHK03 → 1160
  Market-1501 →751
- Features size = (1, 2048, 1, 1)



Res net 50



One solution from face verification: center loss

Center loss

- In face verification [1]…

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

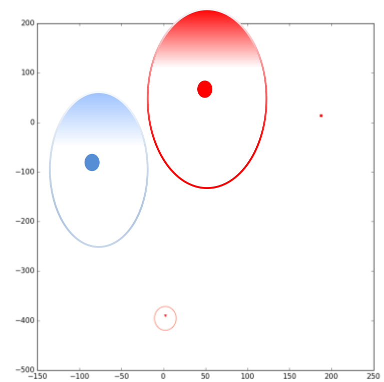
(center loss)

$$\frac{\partial \mathcal{L}_C}{\partial \boldsymbol{x}_i} = \boldsymbol{x}_i - \boldsymbol{c}_{y_i}$$

$$\Delta \boldsymbol{c}_j = \frac{\sum_{i=1}^{m} \delta(y_i = j) \cdot (\boldsymbol{c}_j - \boldsymbol{x}_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)}$$

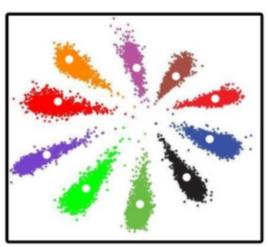
$$\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}}$$

(softmax)

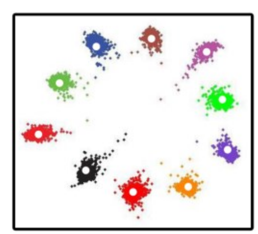


- Increased features intra-class compactness under the joint supervision

Softmax supervision



Softmax + center loss



State-of-the-art

- Our baseline:
  - net → ResNet50
  - training supervision → softmax loss
  - re-id features → from pooling layer 5 output

Method	Market-1501 rank1	Market-1501 mAP
PersonNet [44]	37.21	18.57
DADM [51]	39.40	19.60
Multiregion CNN [43]	45.58	26.11
Bow + HS [23]	47.25	21.88
Fisher Network [24]	48.15	29.94
SL [40]	51.90	26.35
DNS [46]	61.02	35.68
LSTM Siam. [8]	61.6	35.3
Gated S-CNN [10]	65.88	39.55
P2S [36]	70.72	44.27
Basel.(R, Pool5) [14]	73.02	47.62
CADL [45]	73.84	47.11
PIE(R, Pool5) [14]	78.65	53.87

Method	CUHK03 rank1
CDM [16]	40.91
Basel.(R, pool5) [14]	51.60
SI-CI [13]	52.17
DCNN [25]	54.74
DARI [38]	55.4
LSTM Siam. [8]	57.3
PIE(A, FC8) [14]	62.4
DeepDiff [52]	62.43
DNS [46]	62.55
Fisher Network [24]	63.23
Multiregion CNN [43]	63.87
PersonNet [44]	64.80
Gated S-CNN [10]	68.10

Our approach

Enabling considerations

- Our starting point:

1. Person re-id is affected by more severe viewpoint variability than face recognition
    - Center loss does not exploit camera information at all
2. Learning inter-camera relationships critical for enabling the viewpoint invariance.
    - Center loss only addresses intra-class compactness → no inter-camera relationships learned → no ML addressed
3. Non-DL ML techniques perform feature-metric learning sequentially → sub-optimal solution
4. There are some DL ML techniques performing feature-metric learning jointly: Siamese networks but… several drawbacks

Our approach vs traditional ML

- Non DL ML (disjoint): data → CNN learning by traditional losses → feature space → ML → sub-optimal discriminative space

- DL ML (Siamese): data →
  - build training samples
  - balance neg & pos samples
  - data mining (hard neg, moderate pos)

  (> training complexity)

  → CNN learning by contrastive loss (relies on weak re-id labels) ↓↑ ML → discriminative space

- Our approach: data → CNN learning by **our loss** ↓↑ ML → **more** discriminative feature space

Our discriminative model

Our new loss:

1. Additive with regards to the softmax loss
2. Trainable by gradient descent
3. Keeps the training complexity low (1 training sample → 1 input image): suitable to be easily integrated in a simple one branch shaped CNN
4. Scales well to large datasets → Suitable for fast search requirements
5. Produces embeddings discriminative enough that simple metrics (normalized Euclidean distance) can be applied for features points comparison

$$\boxed{L = L_{softmax} + \lambda_{SMC} \cdot L_{SMC} + \lambda_{ECD} \cdot L_{ECD}}$$

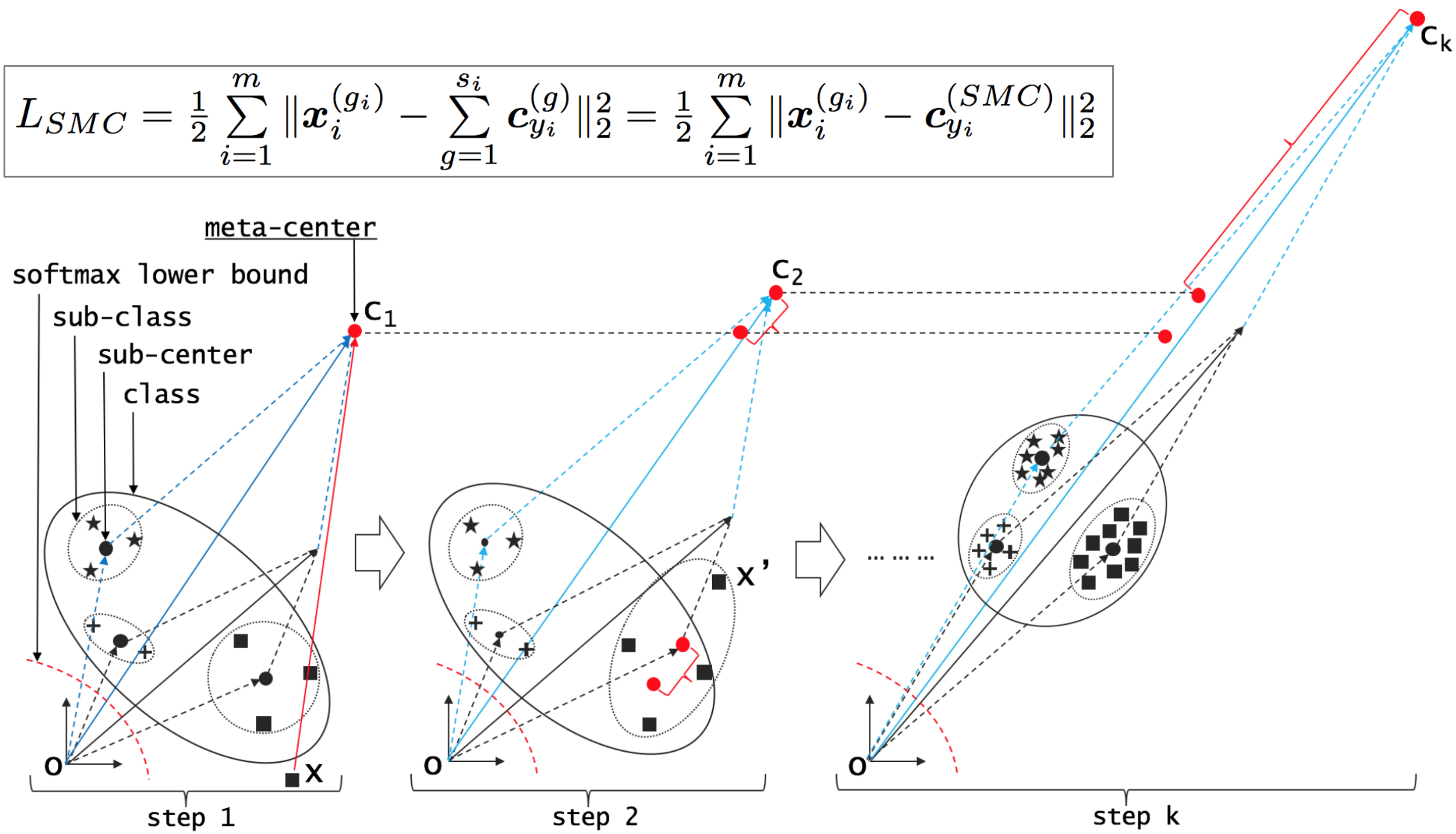
Steering Meta-Center loss term

Enhancing Certers Dispersion loss term

Steering Meta-Center (SMC) loss

- Addresses intra-class compactness AND inter-class dispersion
- Maps all the sub-centers of an identity to a unique *"meta-center"*
- Exploits the camera information

$$L_{SMC} = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i^{(g_i)} - \sum_{g=1}^{s_i} \boldsymbol{c}_{y_i}^{(g)}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(SMC)}\|_2^2$$



SMC loss vs center loss: geometrical meaning

- Meta-Center: scaled version of the unweighted mean of the sub-centers
- Unweighted mean of sub-centers accounts equally all sub-classes → viewpoint invriance property

$$\boldsymbol{c_c} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i = \frac{1}{N}\sum_{g=1}^{s}\sum_{i=1}^{N_g}\boldsymbol{x}_i^{(g)} = \frac{1}{N}\sum_{g=1}^{s}N_g\boldsymbol{c}_g$$

+ N_1=20 → # images camera view 1

▲ N_2=4 → # images camera view 2

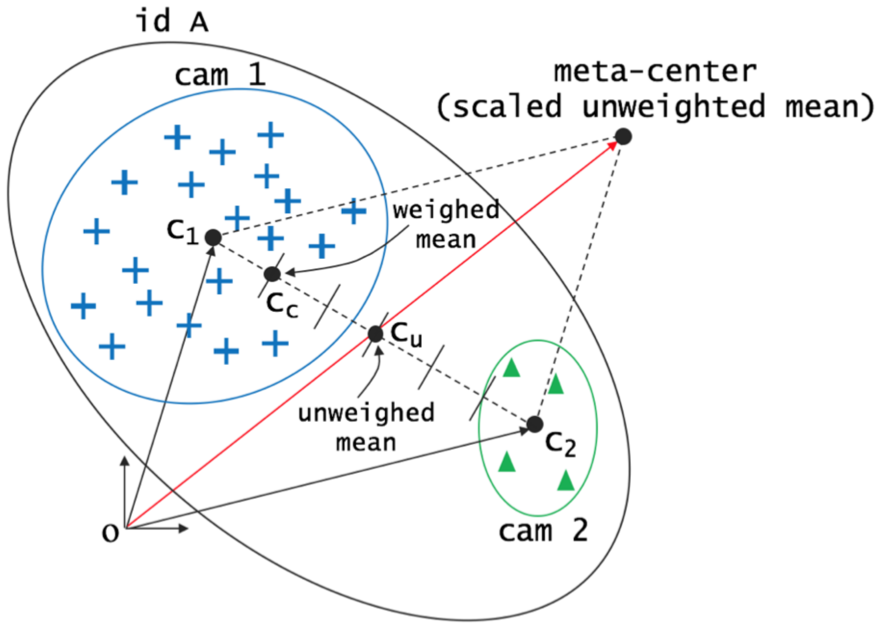
$N=N_1+N_2$ → # images of id A

c_1: center sub-class 1 (cam 1)

c_2: center sub-class 2 (cam 2)

c_c: center defined by center loss

$\mathrm{dist}(c_c,c_2) = (N2/N1)*\mathrm{dist}(c_1,c_c)$

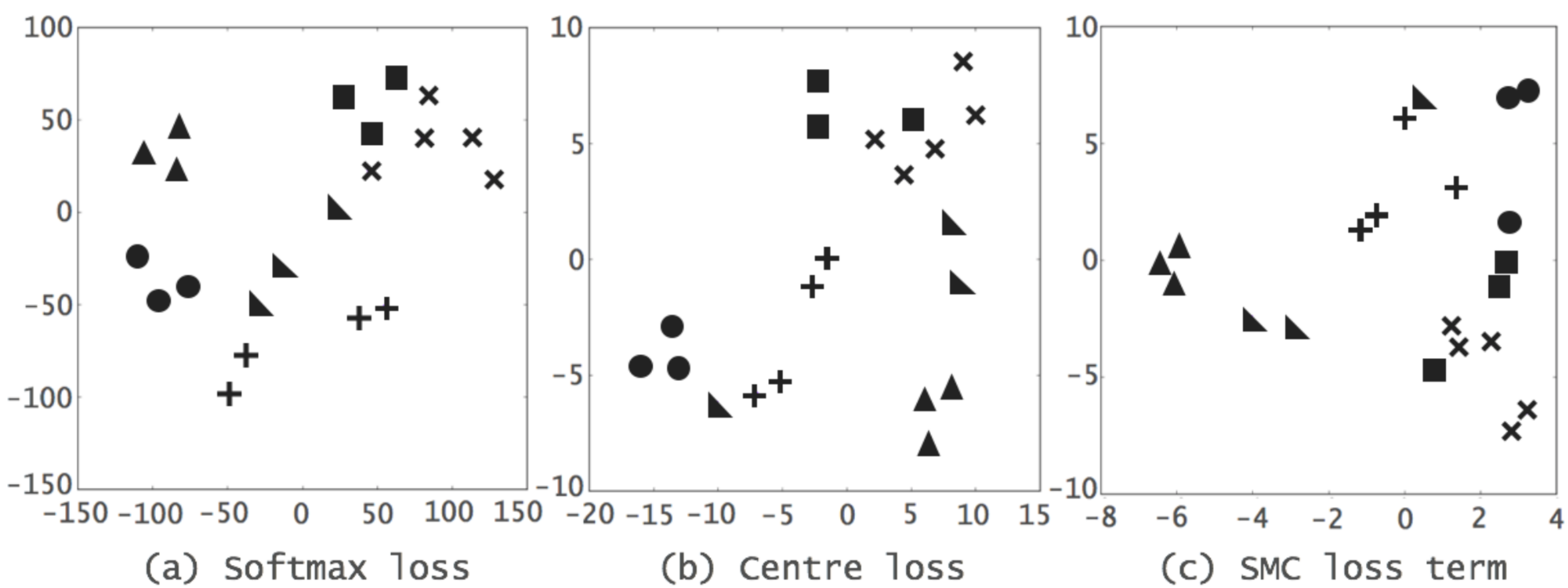


SMC loss effect

- 2D visualization of id 1322 with T-sne
- Enhanced compactness (~10 times):

  softmax → (range X, range Y) ~ (300, 250)

  softmax + SMC → (range X, range Y) ~ (35, 20)
- Less sub-clustered structure in (c) than in (a) → better invariance to viewpoint



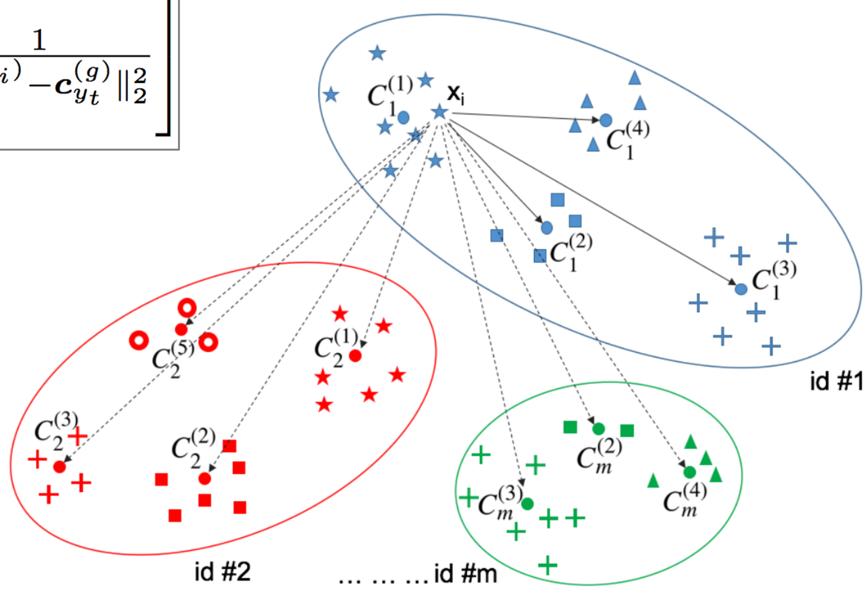
Enhancing Centers Dispersion (ECD) loss

- Relative constraint between intra-class and inter-class scope distances
- Penalizes the distances of x_i from each single sub-centre of the sub-classes belonging to the current training minibatch.
- The larger the number of sub-classes, the stronger the effect of this constraint

$$L_{ECD} = \frac{1}{2}\sum_{i=1}^{m}\left[\sum_{g=1}^{s_i}\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_i}^{(g)}\|_2^2 \cdot \sum_{\substack{t=1\\t\neq i}}^{m}\sum_{g=1}^{s_i}\frac{1}{\|\boldsymbol{x}_i^{(g_i)} - \boldsymbol{c}_{y_t}^{(g)}\|_2^2}\right]$$

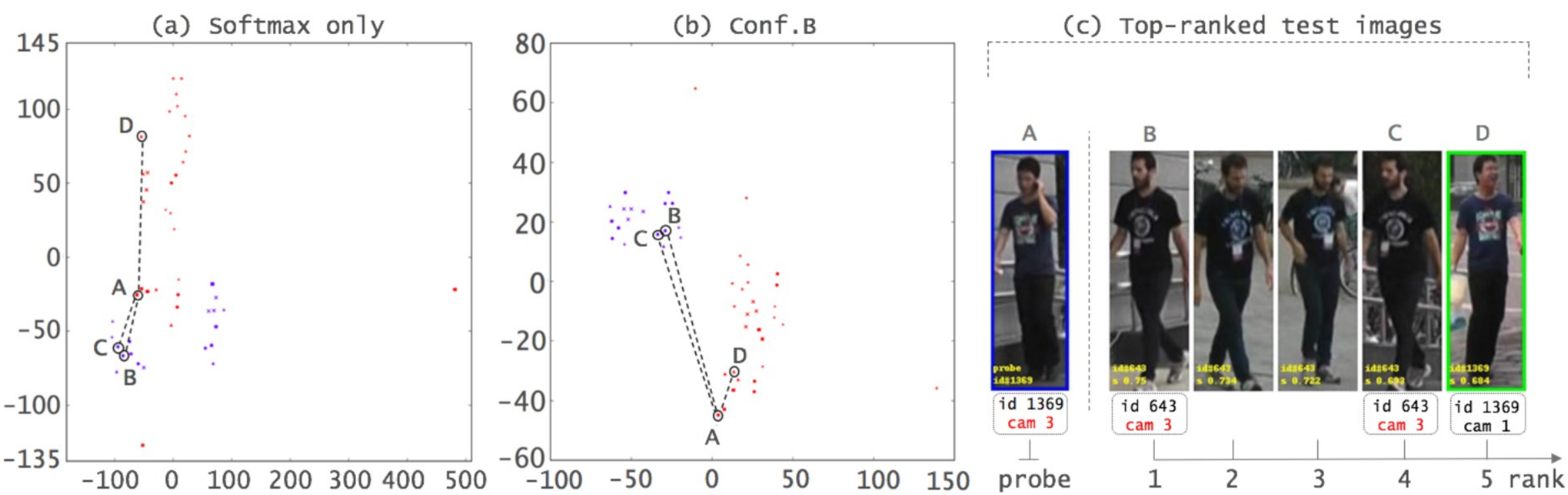
For each sub-center of the training minibatch:

$$ECD = \frac{\sum(solid\ line\ centers\ distances)}{d(C_1^{(1)}, C_k^{(cam_k)})}$$



ECS loss effect

- Learns a similarity/distance relation between inter-class pairs
- Reproduces at training time what non-DL ML methods do on top of a CNN already learned
- Under the softmax loss supervision (a) bboxes B and C represent occurrences of the viewpoint problem
- Under SMC+ECD loss supervision (b) the true positive bbox D is ranked 1st



Performance

Performance

	Market-1501	
Method	rank1	mAP
PersonNet [44]	37.21	18.57
DADM [51]	39.40	19.60
Multiregion CNN [43]	45.58	26.11
Bow + HS [23]	47.25	21.88
Fisher Network [24]	48.15	29.94
SL [40]	51.90	26.35
DNS [46]	61.02	35.68
LSTM Siam. [8]	61.6	35.3
Gated S-CNN [10]	65.88	39.55
P2S [36]	70.72	44.27
Basel.(R, Pool5) [14]	73.02	47.62
CADL [45]	73.84	47.11
PIE(R, Pool5) [14]	78.65	53.87
ours (single query)	**80.31**	59.68
(multiple query)	(85.63)	(67.28)

	CUHK03
Method	rank1
CDM [16]	40.91
Basel.(R, pool5) [14]	51.60
SI-CI [13]	52.17
DCNN [25]	54.74
DARI [38]	55.4
LSTM Siam. [8]	57.3
PIE(A, FC8) [14]	62.4
DeepDiff [52]	62.43
DNS [46]	62.55
Fisher Network [24]	63.23
Multiregion CNN [43]	63.87
PersonNet [44]	64.80
Gated S-CNN [10]	68.10
ours	**69.55**

SMC+ECD on Market-1501:

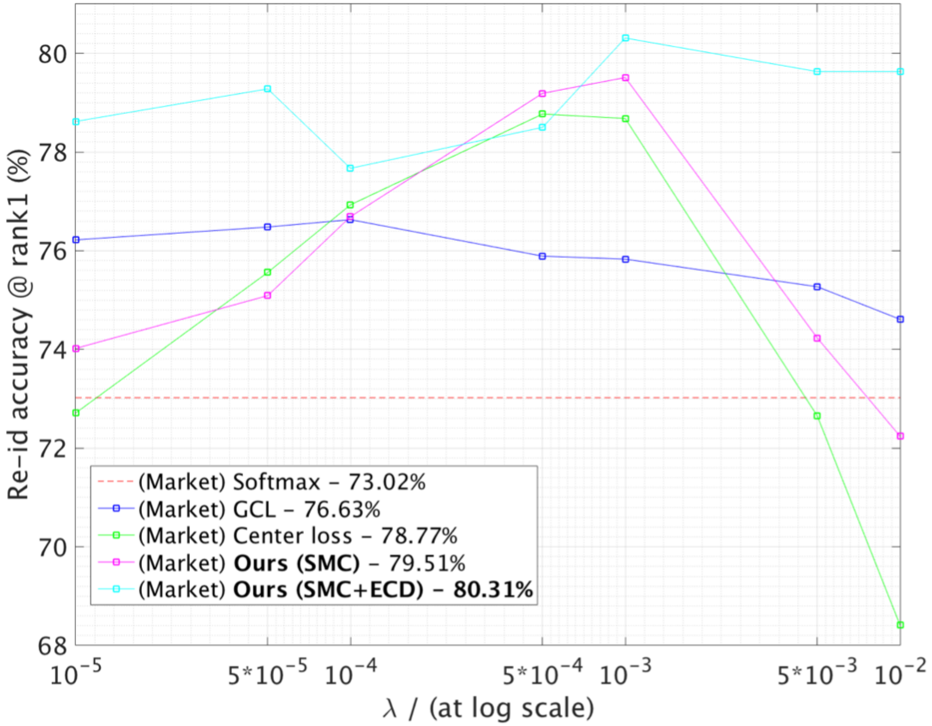
- Rank1 **+9.9%** baseline
- mAP **+25.3%** baseline

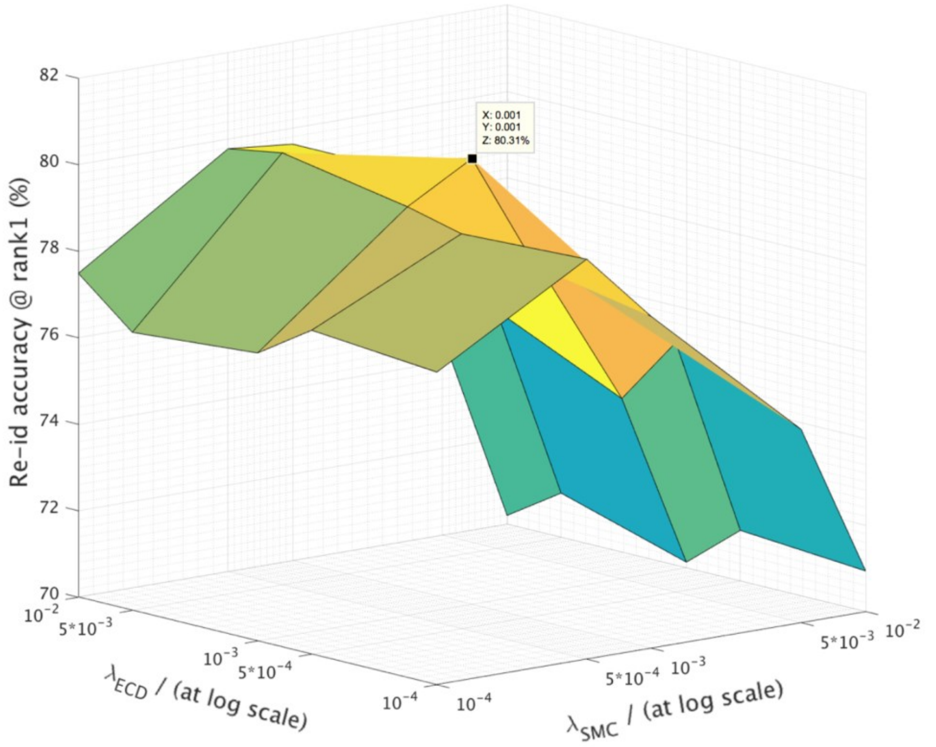
SMC+ECD on CUHK03:

- Rank1 **+34.8%** baseline

Parametric performance

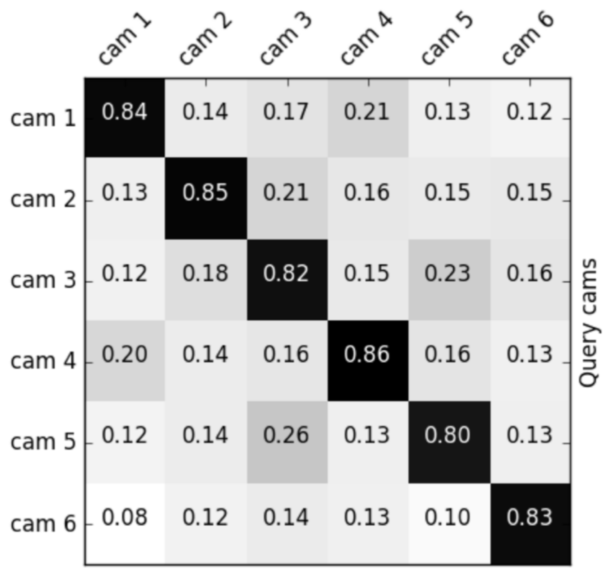
	Market-1501 [23]					CUHK03 [7]			
	mAP	rank 1	5	10	20	rank 1	5	10	20
Softmax	47.62	73.02	85.84	90.35	93.32	51.60	79.60	87.70	95.00
GCL	54.25	76.63	88.78	92.25	95.19	63.66	88.58	94.20	98.03
Center	57.76	78.77	90.14	93.62	95.72	66.19	90.65	96.06	98.73
SMC	58.28	79.51	90.59	93.74	95.90	**69.59**	92.62	96.86	98.91
SMC+ECD	**59.68**	**80.31**	91.27	94.09	96.02	69.55	90.96	95.07	97.54



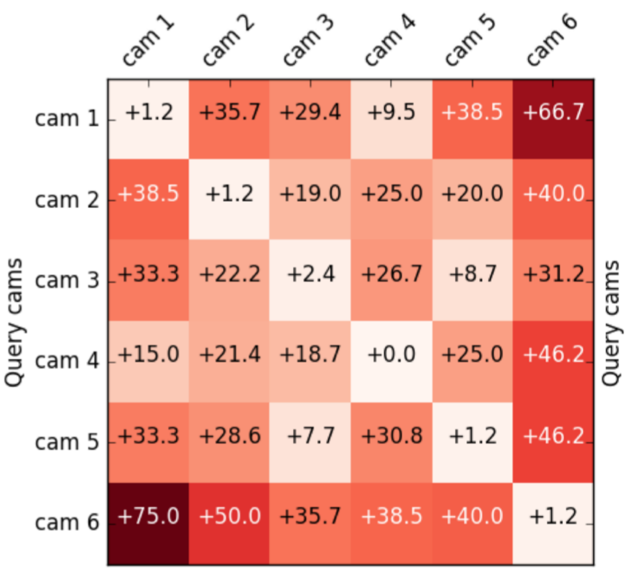


Ablation study

- Re-id performance between camera pairs: mAP confusion matrix



(a) Softmax



(b) SMC+ECD

- Fraction of the performance improvement which translates in tmprovement of the viewpoint problem, determined by negatives analysis: “Figure of merit”

	GCL	Center	SMC	SMC+ECD
F_{rank1}	15.5	23.4	24.3	**26.3**
F_{mAP}	33.4	35.7	47.6	**50.7**

Our approach vs Joint Bayesian

- perf(our approach) > perf(baseline + Joint-Bayesian)
- perf(our approach + Joint-Bayesian) > perf(our approach)

		Softmax	SMC	SMC+ECD
Market-1501	**rank 1**	77.06 (+5.5)	79.93 (+0.5)	80.38 (+0.1)
	mAP	53.76 (+12.9)	58.40 (+0.2)	59.73 (+0.1)
CUHK03	**rank 1**	65.03 (+26.0)	72.04 (+3.5)	71.76 (+3.2)

Final remarks

Advantages

1. More effective in learning in mitigating the changing viewpoint problem
2. Replicating the capability of Siamese networks to carry out a joint features-metric learning process
3. Not increasing training complexity (1 input image → 1 training sample)
4. Not employing extra training data or side information.
5. More effective than both DL ML techniques and non-DL ML techniques
6. Flexibility: our loss can be easily integrated in any architecture

Disadvantage

- Increased training time (for Market: baseline time +1h)

Novelty

- We re-interpret in person re-id the center loss introduced in face verification.
- We adapt the ML approach to the CNN training stage avoiding traditional ML drawbacks but retaining their capability to learn an inter-class similarity function.


[00148]
S0001
A0001
S0002
[00148]
103

Thank you!

Questions?