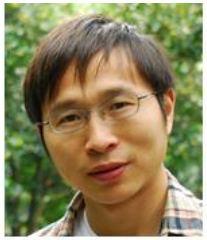
The 13th IEEE Conference on Automatic Face and Gesture Recognition

2018 FG Xi'an, China | May 15-19, 2018

Person Re-Id: Recent Advances and Challenges

Session 2:
Attribute Learning and Scalability

Shiliang Zhang
Peking University
Beijing, China



Jingdong Wang
Microsoft Research
Beijing, China

Qi Tian
University of Texas at
San Antonio, USA

Wen Gao
Peking University
Beijing, China

Longhui Wei
Peking University
Beijing, China

Institute of Digitial Media

Peking University

2018.5.15

北京大学

Outline

- Attribute Learning
  - Attribute embedding in multi-task learning
  - Attribute learning with deep models
- Scalable Person Re-ID
  - Deep feature extraction and compression
  - Off-line index optimization

Attribute

- Semantic description to an image
- a stable and robust mid-level description for pedestrian
  - “wear shorts”, “male”, “long hair”, “wear grey T-shirt”
- a compact binary feature
  - ‘1’ means presence, ‘0’ means absence


cam1


cam2


cam3


cam4

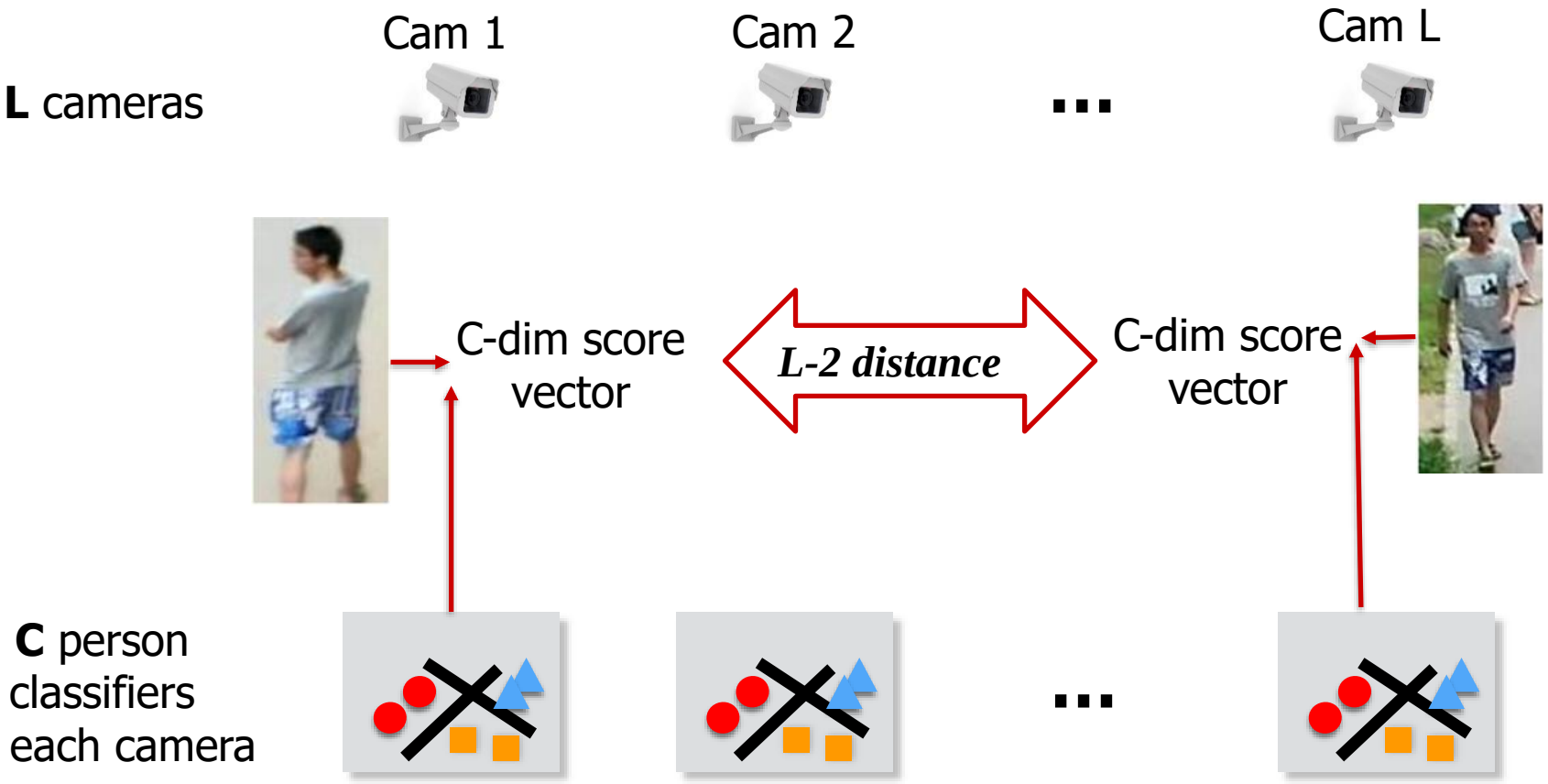

cam5

Issues in attribute learning

- Training data is hard to annotate
  - Expensive and time consuming for human annotation
- There are errors in attribute prediction
  - E.g., both ‘male’ and ‘female’ are ‘1’
- Strong correlations among attributes
  - E.g., ‘short’ and ‘bare leg’
  - ‘female’ and ‘long hair’

Attribute embedding in multi-task learning

- Basic idea



How to train *C* person classifiers

First issue:

- the classifier feature should be robust
- consider attributes and attribute correlations as features

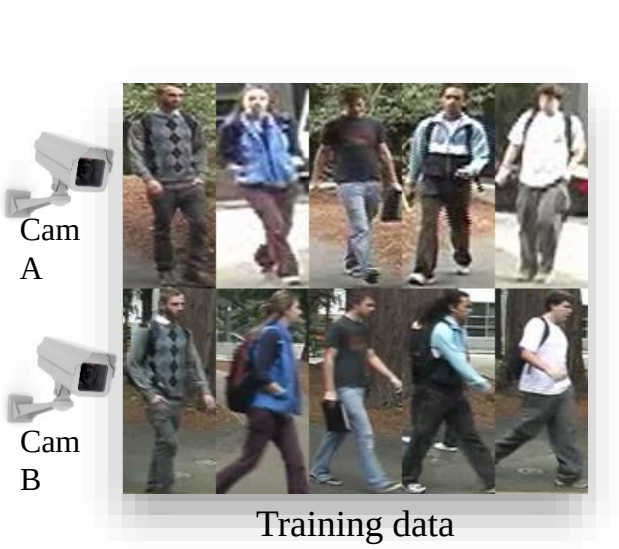
Second issue:

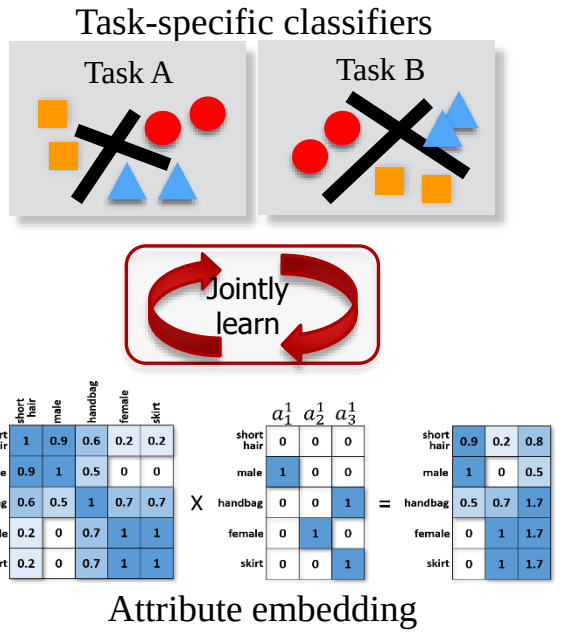
- different cameras corresponds to different classification models, due to different illumination, viewpoints, etc.
- the amount of training data on each camera is limited
- we treat classifier learning under different cameras as related tasks
  - Different cameras record the same set of persons

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Overview of our approach

- Multiple cameras are treated as related tasks to exploit their relationships
- Attributes correlations are embedded to improve the accuracy of predicted attributes and reduce noise
- Classifiers of multiple tasks and attribute embedding are jointly learned by alternating optimization

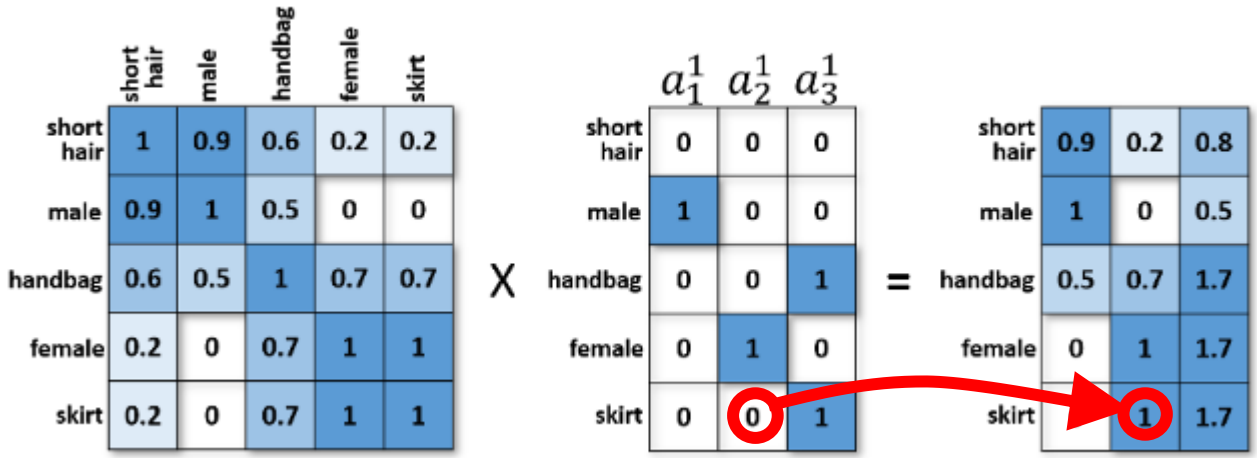




C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

attribute feature embedding

- Predict initial attributes with SVMs
- Learn attribute correlations to improve the accuracy



Correlations Matrix: $\mathbf{Z}$ Initial attributes: a_i^l Updated attributes:

$$\phi_{\mathbf{Z}}(\mathbf{a}_i^l) = \mathbf{Z}^\top \mathbf{a}_i^l$$

$$\text{s.t.} \quad \text{rank}(\mathbf{Z}) \leq r,$$

- Final feature:

$$\widetilde{\mathbf{x}}_i^l = [\mathbf{x}_i^l; \phi_{\mathbf{Z}}(\mathbf{a}_i^l)] \in \mathbb{R}^{d+k}$$

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Formulation of classifier learning

- Loss function for person classification under camera *l*:

$$\ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) = \frac{1}{2}(||y_i^l - \mathbf{w}^{l\top}\widetilde{\mathbf{x}}_i^l||^2 + \gamma||\mathbf{a}_i^l - \mathbf{Z}^\top \mathbf{a}_i^l||^2)$$

Person classification error

$$\widetilde{\mathbf{x}}_i^l = [\mathbf{x}_i^l; \phi_{\mathbf{Z}}(\mathbf{a}_i^l)] \in \mathbb{R}^{d+k}$$

:concatenation of low level features and the updated attributes

attribute embedding regularization

- Goal: learn classifier ***w*** and embedding matrix ***Z***

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Formulation of multi-task learning

- Low-rank embedding
  - the classifiers to be learned on *L* cameras:

$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^L] \in \mathbb{R}^{(d+k)\times L}$$

  - We define: **W** = **R** + **S**
  - **R**: a low-rank matrix shared by all tasks
  - **S**: task-specific sparse component

$$\begin{aligned} \min_{\mathbf{R},\mathbf{S},\mathbf{Z}} \quad & \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \tilde{\mathbf{x}}_i^l, \mathbf{Z}) + \lambda||\mathbf{S}||_0 \\ \text{s.t.} \quad & \mathbf{W} = \mathbf{R} + \mathbf{S}, \text{rank}(\mathbf{R}) \leq r_1, \text{rank}(\mathbf{Z}) \leq r_2 \end{aligned}$$

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Formulation of multi-task learning

- Relax the previous formulation into a smooth one

$$\min_{\mathbf{R},\mathbf{S},\mathbf{Z}} \quad \sum_{l=1}^{L}\sum_{i=1}^{n_l} \ell(y_i^l, \mathbf{a}_i^l, \widetilde{\mathbf{x}}_i^l, \mathbf{Z}) + \lambda||\mathbf{S}||_1$$

$$\text{s.t.} \quad \mathbf{W} = \mathbf{R} + \mathbf{S}, ||\mathbf{R}||_* \leq r_1, ||\mathbf{Z}||_* \leq r_2,$$

- This formulation can be optimized by an alternating optimization
  - Fix **Z** to update the classifier parameter **w** by **MixedNorm**
  - Fix **w** update **Z** by **proximal gradient method**

For details:

Multi-Task Learning with Low Rank Attribute Embedding for Person Re-identification, *ICCV*, 2015

Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

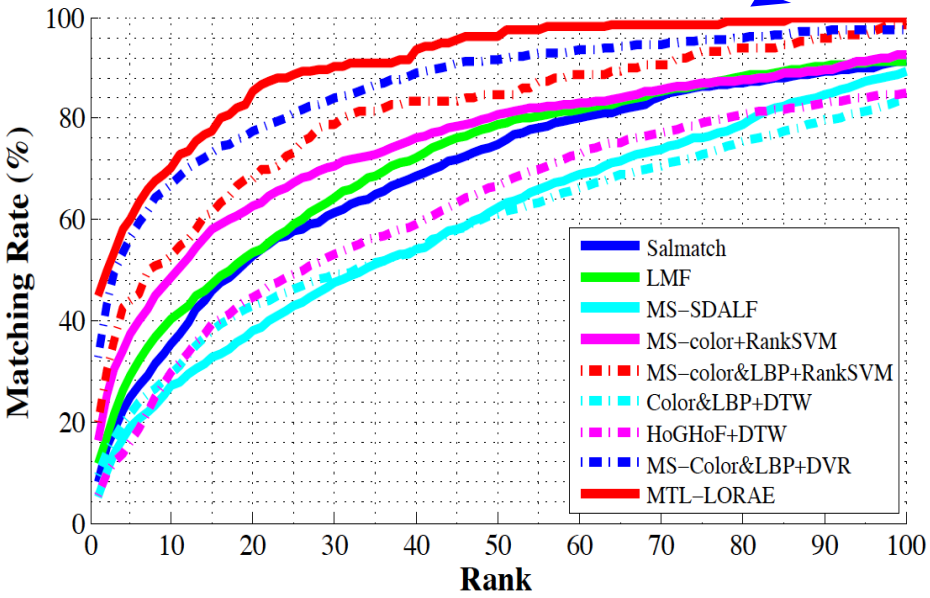
Datasets

- iLIDS-VID
  - two camera, multi-shot
  - 600 image sets of 300 persons
- PRID
  - two camera, multi-shot
  - 385 and 749 persons in cameras A and B, respectively. 200 persons appear in both cameras
- VIPeR
  - two camera, single-shot
  - 632 persons in two cameras
- Use a 2784-dimensional color and texture descriptor as low level feature
- learn binary SVMs to predict 20-bit attributes for PRID, 90-bit attributes for VIPeR. Learn 32-bit attributes for iLIDS-VID.

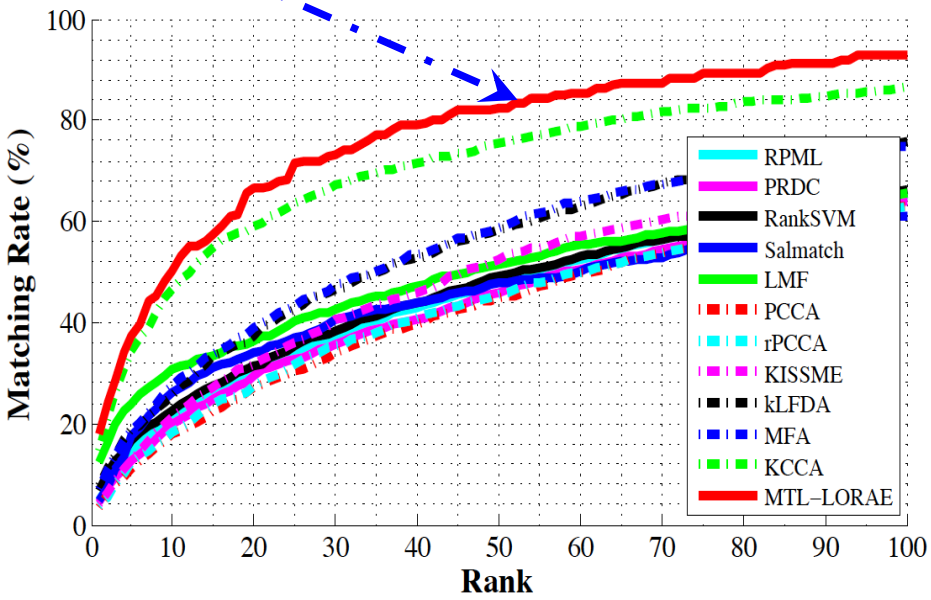
C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Performance of two camera multi-shot Re-ID task

Our approach gets very competitive performance



iLiDS-VID



PRID

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Performance of two camera single-shot Re-ID task

Viper dataset

TABLE 4
CMC Scores of Ranks from 1 to 20 on the VIPeR Dataset

Rank	1	5	10	20
KISSME [11]	19.6	47.5	62.2	77.0
kLFDA [12]	32.2	65.8	79.7	90.9
KCCA [77]	37.3	71.4	**84.6**	92.3
LOMO + XQDA [78]	40.0	68.9	81.5	91.1
TSR [79]	31.6	68.6	82.8	**94.6**
EPKFM [80]	36.8	70.4	83.7	91.7
MLAPG [81]	40.7	69.9	82.3	92.4
MTL-LORAE	**42.3**	**72.2**	81.6	89.6

Our approach gets best performance for Rank -1 and 5

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Performance of combining deep feature

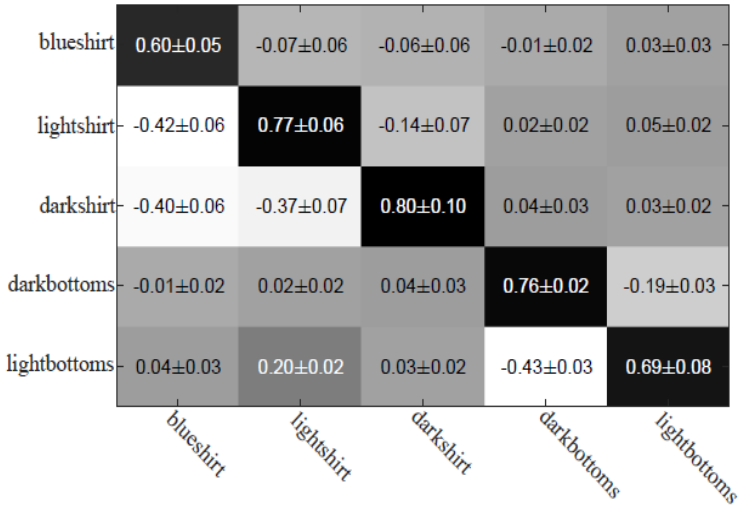
- VGG-FC7: use FC7 layer output of VGG-16
- Pre-trained on ImageNet, fine-tuned on cameras #1, #2, #4, #6 and #7 on SAIVT-SoftBio

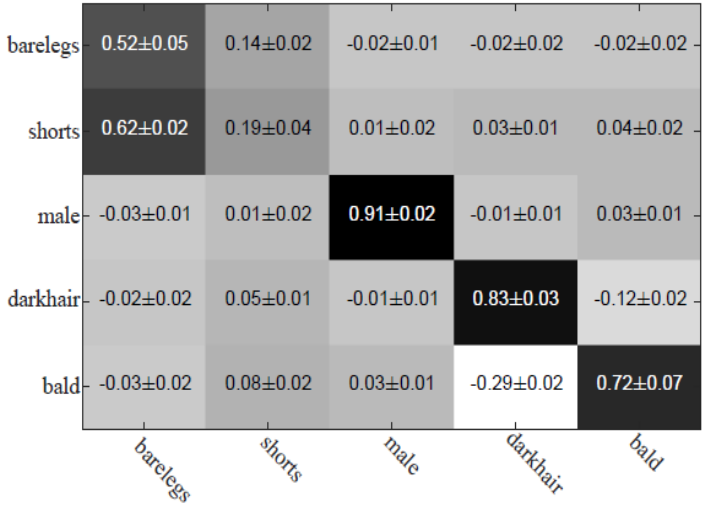
Methods		Rank 1	Rank 5	Rank 10	Rank 20
iLIDS-VID	VGG-FC7	24.1	43.6	52.8	65.6
	MTL-LOREA	43.0	60.0	70.2	85.3
	MTL-LOREA-FC7	56.4	69.0	78.4	87.4
PRID	VGG-FC7	19.8	28.5	42.4	53.9
	MTL-LOREA	18.0	37.4	50.1	66.6
	MTL-LOREA-FC7	21.0	44.0	55.9	68.7
VIPeR	VGG-FC7	25.1	39.8	48.5	60.6
	MTL-LOREA	42.3	72.2	81.6	89.6
	MTL-LOREA-FC7	45.4	76.6	85.3	91.7

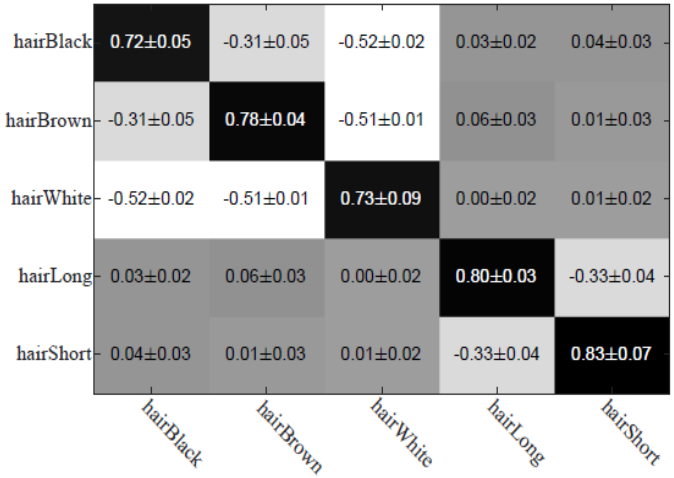
- Our approach outperforms deep features
- The performance further boosted by combining deep features

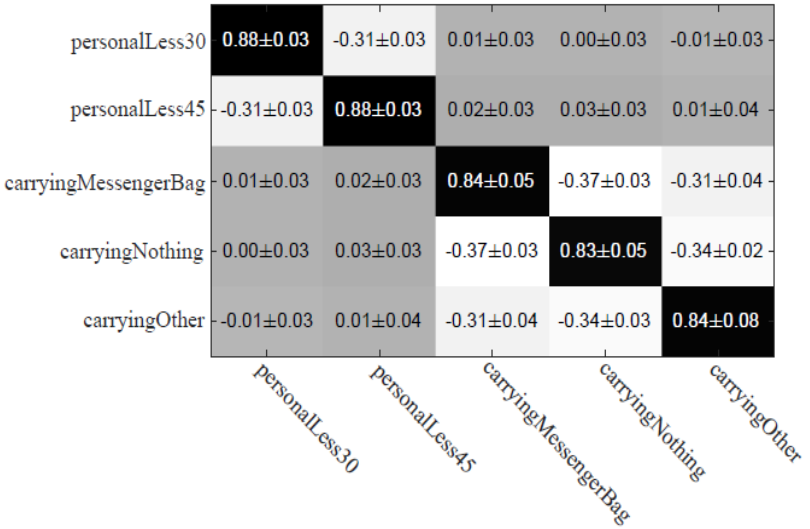
C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Examples of attribute correlations









Validity of MTL and attribute embedding

- STL: single task learning
- MTL-Att: without attribute embedding

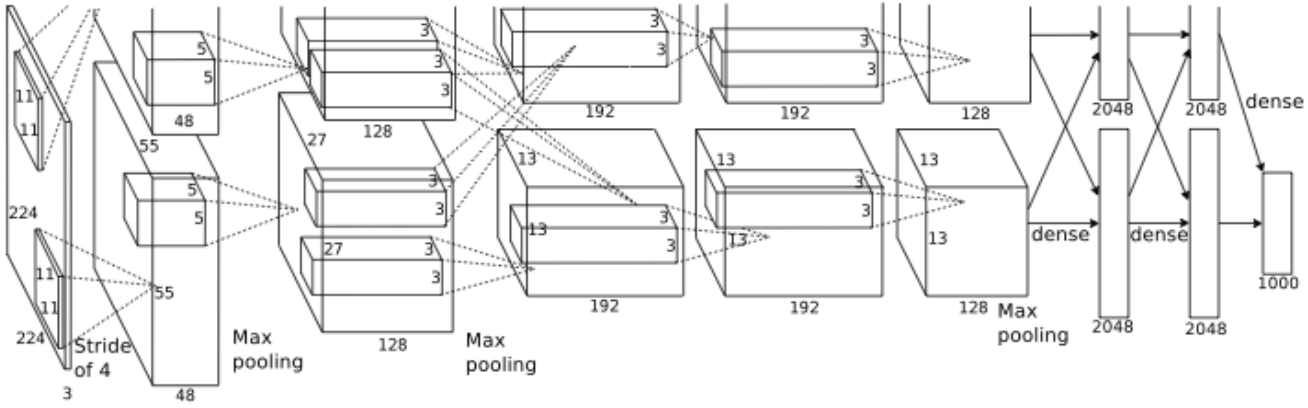
	iLIDS-VID					
Rank	1	5	10	20	30	50
STL	14.7	42.7	41.8	58.5	83.5	91.7
MTL-FR	37.7	54.0	47.4	64.9	85.3	92.5
MTL-Att	40.5	54.9	47.5	64.2	84.2	91.2
MTL-LOREA	**43.0**	**60.0**	**70.2**	**85.3**	**90.2**	**96.3**
	PRID					
Rank	1	5	10	20	30	50
STL	11.3	27.9	41.8	53.0	68.5	74.6
MTL-FR	11.3	34.1	47.4	61.1	69.8	79.0
MTL-Att	12.2	34.7	47.5	61.7	70.9	79.8
MTL-LOREA	**18.0**	**37.4**	**50.1**	**66.6**	**73.1**	**82.3**
	VIPeR					
Rank	1	5	10	20	30	50
STL	13.3	27.4	32.8	42.7	56.2	68.3
MTL-FR	35.3	63.3	75.6	83.8	89.9	94.4
MTL-Att	37.2	64.2	76.3	84.9	91.4	95.3
MTL-LOREA	**42.3**	**72.2**	**81.6**	**89.6**	**93.1**	**97.4**

- Multi-task learning boosts the performance
- Attribute embedding boost the performance

C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *T-PAMI*, 2018

Attribute learning with deep models

- Deep learning (deep Convolutional Neural Network, dCNN):
  - Powerful learning ability and good performance in pattern recognition tasks
  - Good generalization ability



The structure of AlexNet (NIPS'12)

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Issues & Solution

- Issues
  - dCNN needs a large amount of training data
  - The available labeled attribute data is limited
- Solution: Semi-supervised Deep Attribute Learning (SSDAL)
  - Train dCNN with a partially-labeled dataset
    - Attribute annotation: PETA
    - Another dataset with person ID: MOT Challenge
  - Ensure its discriminative power and generalization ability in the person ReID tasks

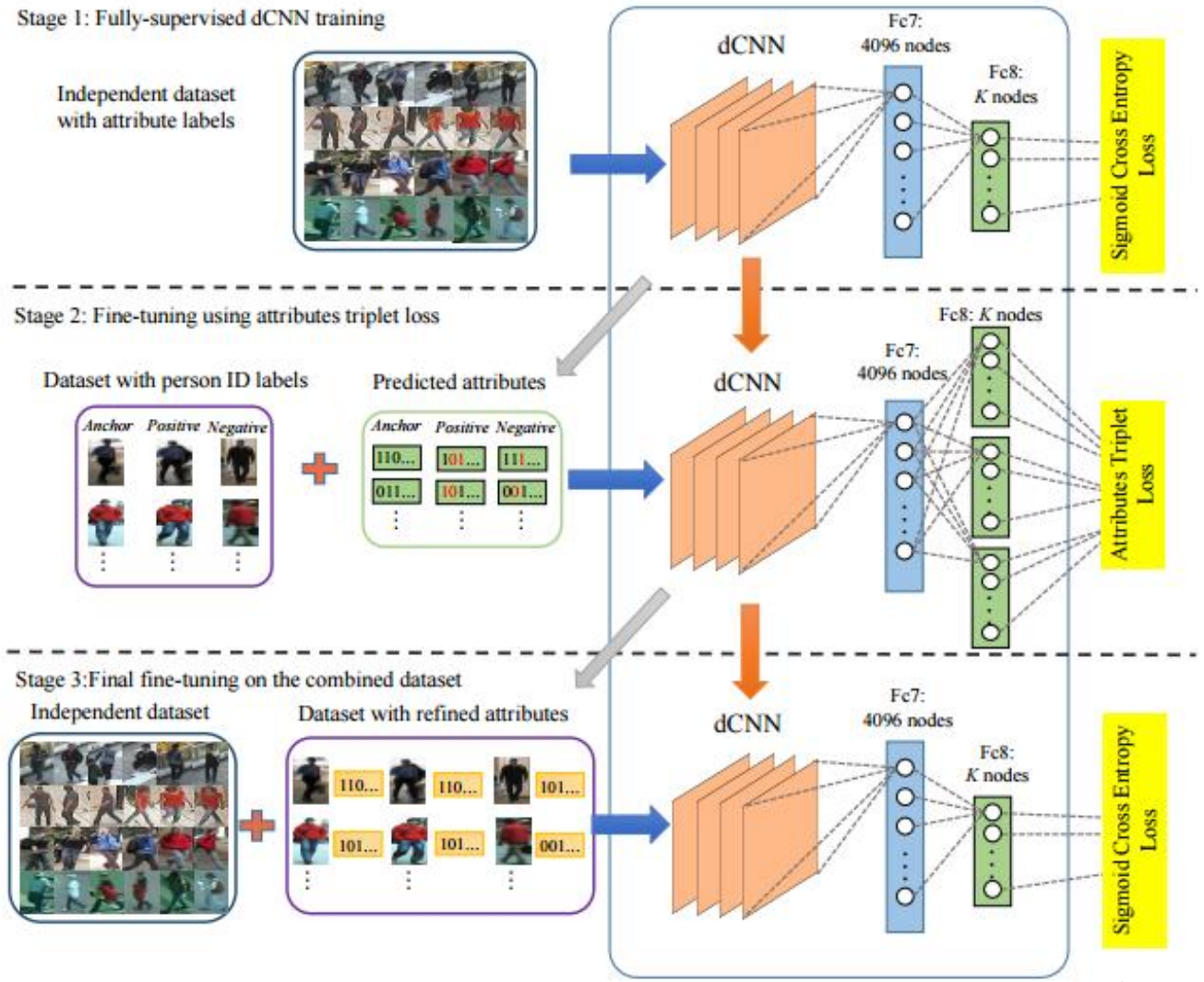
C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Framework

Stage-1

Stage-2

Stage-3



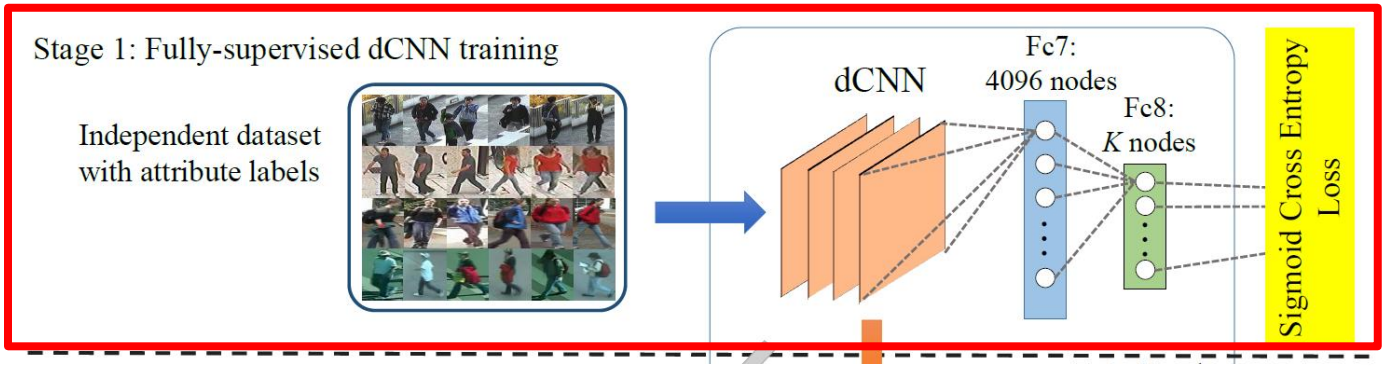
C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Our Approach

- Solution:
  - Three Stage:
    - Stage 1: a fully supervised training on an independent dataset labeled with human attributes
    - Stage 2: fine-tune on another target dataset labeled with person IDs using *Attributes Triplet Loss*
    - Stage 3: The updated dCNN predicts attribute labels for the target dataset, which is finally combined with the independent dataset for the second round of fine-tuning.

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Stage-1

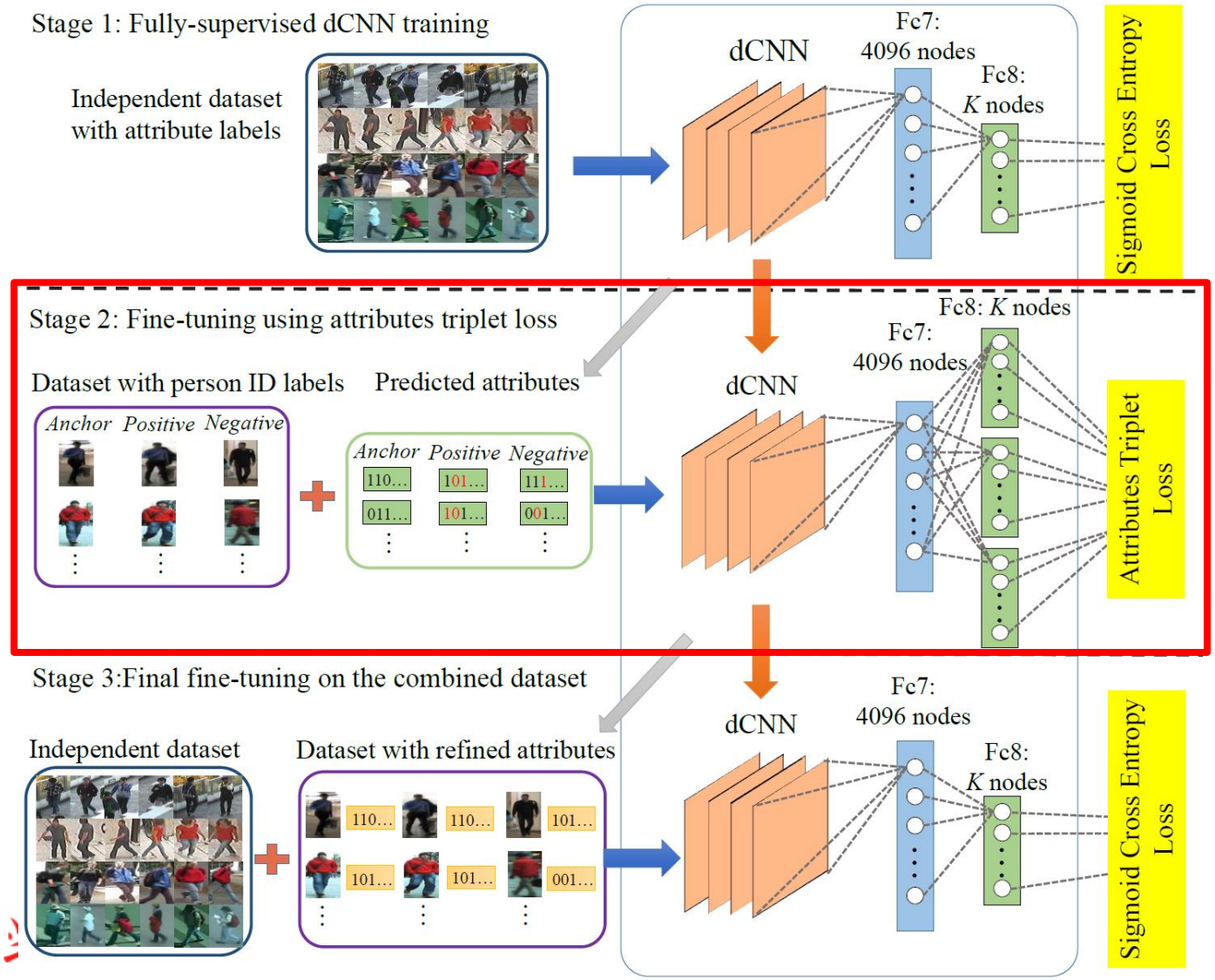


Fully-Supervised dCNN Training

- Train with a small independent dataset labeled with: image + attributes
- Use a sigmoid cross-entropy loss layer for multi-label prediction

Stage-2

Fine-Tune with Attributes Triplet Loss



Fine-Tune with Attributes Triplet Loss

- Goal: refine the predicted attributes to make same person share more similar attributes
- Network predicts the attributes $[A_{(a)}, A_{(p)}, A_{(n)}]$ of three images labeled with person ID
- Network loss:

distance between the same person

$$\mathcal{L} = \sum_{e}^{E} \left\{ \max\left(0, \mathbf{D}\left(A_{(a)}^{(e)}, A_{(p)}^{(e)}\right) + \theta - \mathbf{D}\left(A_{(a)}^{(e)}, A_{(n)}^{(e)}\right)\right) + \gamma \times \mathcal{E} \right\}$$

distance between different person

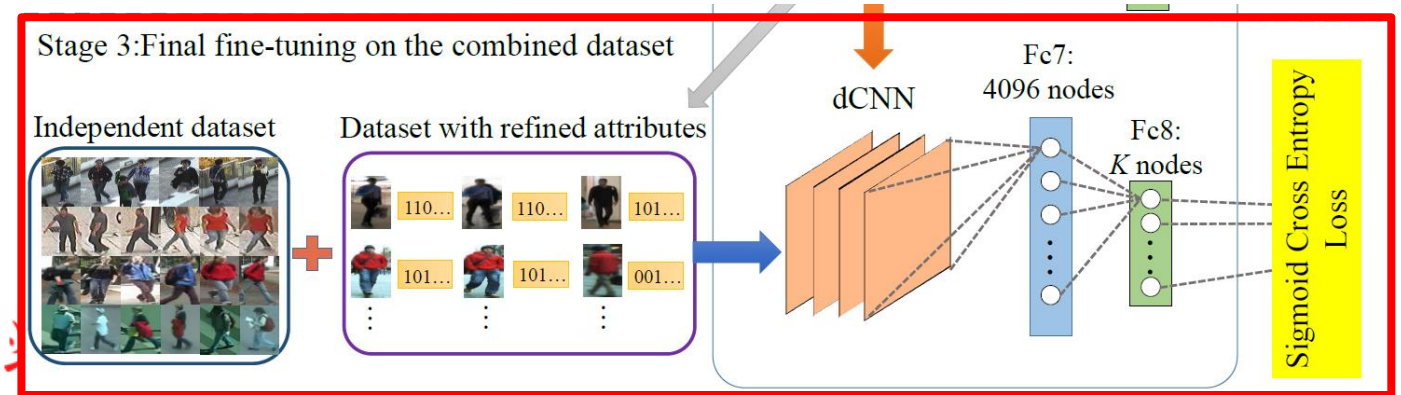
$$\mathcal{E} = \mathbf{D}(A_{(a)}^{(e)}, \tilde{A}_{(a)}^{(e)}) + \mathbf{D}(A_{(p)}^{(e)}, \tilde{A}_{(p)}^{(e)}) + \mathbf{D}(A_{(n)}^{(e)}, \tilde{A}_{(n)}^{(e)})$$

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Stage-3

Fine-Tunning on the Combined Dataset

- The labeled dataset from Stage-2 + the original independent dataset
- The attributes predicted by the final dCNN model are named as *deep attributes*
- *deep attributes + Cosine Distance* for Person Re-ID



Examples of Predicted Attributes

lowerBodyBlack
lowerBodyTrousers
hairShort
personalFemale
carryingMessengerBag

accessoryNothing
hairBlack

lowerBodyTrous
ers hairShort
hairGrey
upperBodyWhite

personalMale
carryingNothing
footwearLeatherShoes

upperBodyGrey
accessoryHat
lowerBodyTrousers
hairShort

upperBodyOther

upperBodyRed
lowerBodyJeans
personalLess30
lowerBodyBlue
accessoryNothing

footwearSneakers
hairShort

upperBodyWhite
hairShort
accessoryNothing
lowerBodyWhite
carryingNothing
lowerBodyGrey

lowerBodyWhite

lowerBodyTrousers
upperBodyBlue
lowerBodyGrey
carryingMessengerBag

accessoryNothing
hairBlack
personalMale

accessoryNothing
hairLong
carryingMessenger
Bag footwearBlack
lowerBodyTrousers

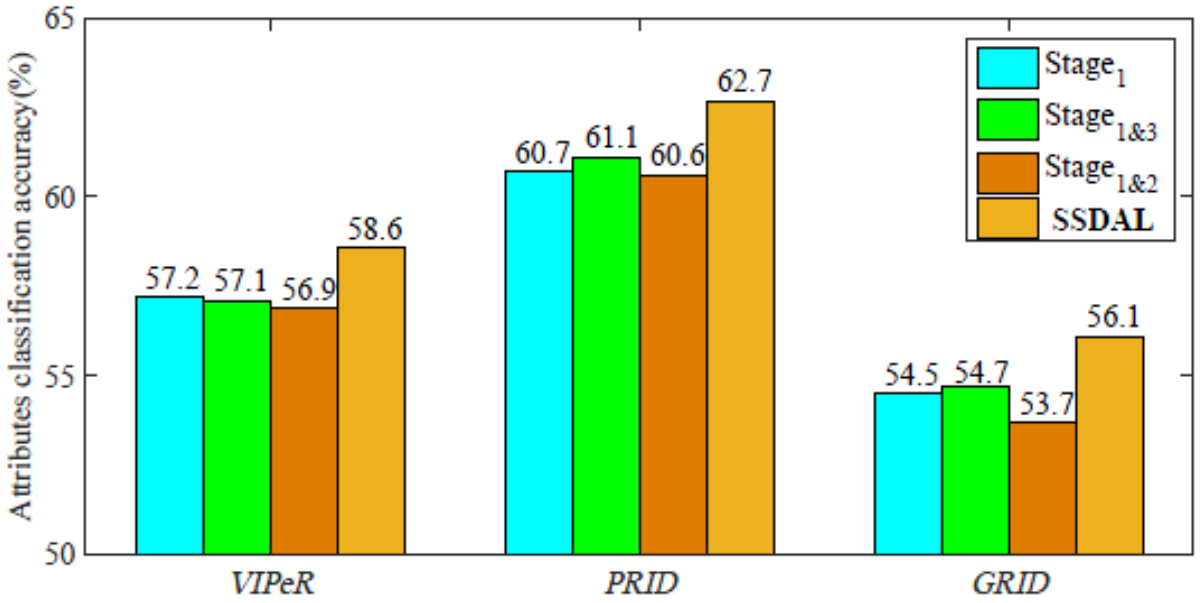
upperBodyWhite

lowerBodyShorts
upperBodyTshirt
hairShort
personalLess30
accessoryNothing
footwearWhite

upperBodyShortSleeve
footwearSandals

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Why using three stages



The three stage training gets the best performance

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Datasets for experiments

- Training dataset:
  - PETA: annotated with 105 attributes
  - MOT challenge: annotated with person ID
- Independent testing dataset:
  - VIPeR: 632 persons, each has two 48*128 images
  - PRID: 385 and 749 persons captured by camera A and B
  - GRID: 1025 persons taken by 8 non-adjacent cameras
  - Market: 25,000 images of 1501 persons taken by 6 cameras

The training data and testing data are totally independent

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Performance-1

- Two cams datasets:
  - VIPeR:

Methods		Rank 1	Rank 5	Rank 10	Rank 20
Metric Learning based ReID	RPML [10]	27.0	57.0	69.0	83.0
	Salmatch [48]	30.2	52.4	65.5	79.1
	LMF [49]	29.1	52.3	65.9	80.0
	KISSME [13]	19.6	47.5	62.2	77.0
	KCCA [50]	37.3	71.4	**84.6**	92.3
	kLFDA [14]	32.2	65.8	79.7	90.9
	LOMO + XQDA [20]	40.0	68.9	81.5	91.1
	CSL [22]	34.8	68.7	82.3	91.8
	MLAPG [23]	40.7	69.9	82.3	92.4
	TSR [51]	31.6	68.6	82.8	**94.6**
	EPKFM [19]	36.8	70.4	83.7	91.7
Traditional Attributes Learning based ReID	AIR [29]	18.0	38.8	51.1	71.2
	OAR [31]	21.4	41.5	55.2	71.5
	LORAE [34]	42.3	**72.2**	81.6	89.6
Deep Learning based ReID	IDLA [42]	34.8	54.3	76.5	87.6
	DML [41]	28.2	59.3	73.5	86.4
	Deep-RDC [24]	40.5	60.8	70.4	84.4
Proposed	$Stage_1$	34.5	63.9	73.1	87.0
	SSDAL	37.9	65.5	75.6	88.4
	SSDAL + XQDA	**43.5**	71.8	81.5	89.0

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Performance-2

- Two cams datasets:
  - PRID:

Methods	Rank 1	Rank 5	Rank 10	Rank 20
RPML [10]	4.8	14.3	21.6	30.2
PRDC [17]	4.5	12.6	19.7	29.5
RSVM [52]	6.8	16.5	22.7	31.5
Salmatch [48]	4.9	17.5	26.1	33.9
LMF [49]	12.5	23.9	30.7	36.5
PCCA [9]	3.5	10.9	17.9	27.1
KISSME [13]	4.1	12.8	21.1	31.8
kLFDA [14]	7.6	18.9	25.6	37.4
KCCA [50]	14.5	34.3	46.7	59.1
LOREA [34]	18.0	37.4	50.1	66.6
LOMO + XQDA [20]	15.3	35.7	41.2	53.8
MLAPG [23]	16.6	33.1	41.4	52.5
$Stage_1$	18.7	46.9	55.0	65.8
SSDAL	20.1	47.4	55.7	68.6
SSDAL + XQDA	**22.6**	**48.7**	**57.8**	**69.2**

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Performance-3

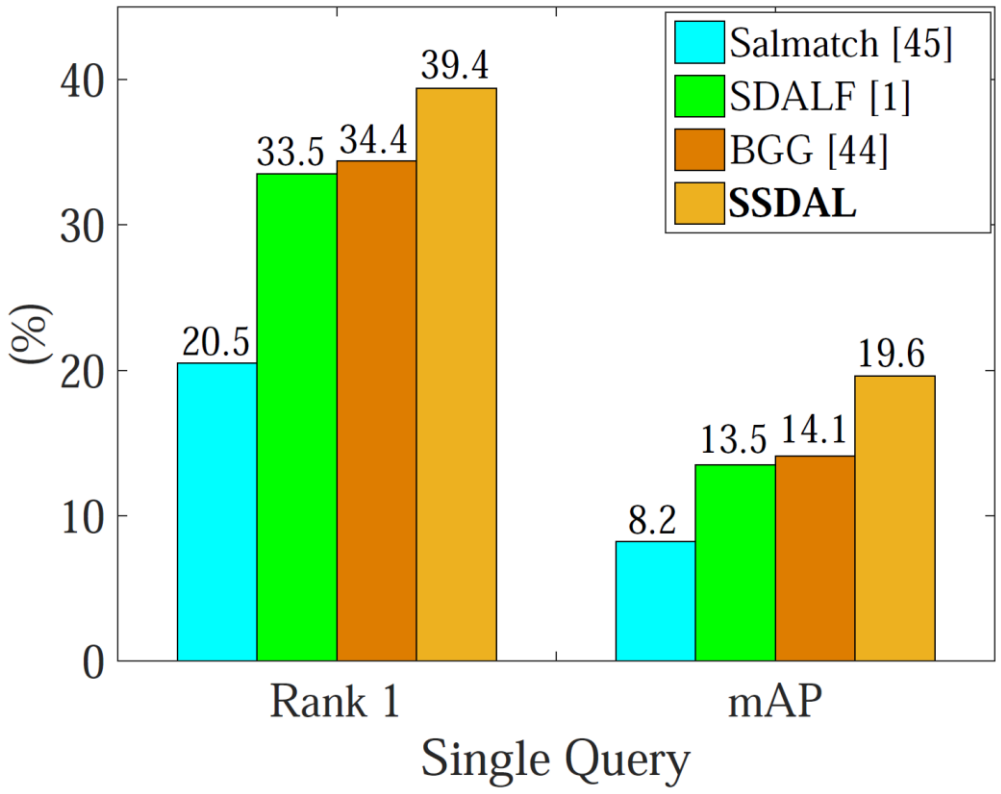
- Eight cams datasets:
  - GRID:

Methods	Rank 1	Rank 5	Rank 10	Rank 20
PRDC [17]	9.7	22.0	33.0	44.3
RSVM [52]	10.2	24.6	33.3	43.7
MRank-PRDC [28]	11.1	26.1	35.8	46.6
MRank-RSVM [28]	12.2	27.8	36.3	49.3
RQDA [53]	15.2	30.1	39.2	49.3
EPKFM [19]	16.3	35.8	46.0	57.6
LOMO + XQDA [20]	16.6	35.4	41.8	52.4
$Stage_1$	16.9	30.1	40.7	50.2
SSDAL	19.1	35.6	45.8	58.1
SSDAL + XQDA	**22.4**	**39.2**	**48.0**	**58.4**

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Performance-4

- Large-scale multi-cam datasets:
  - Market：



C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Issues of SSDAL

- Shallow network， 5 conv layers
- Has thresholds to select positive attributes
- Does not consider the correlations between attributes

C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *ECCV*, 2016.

Solution and Extension

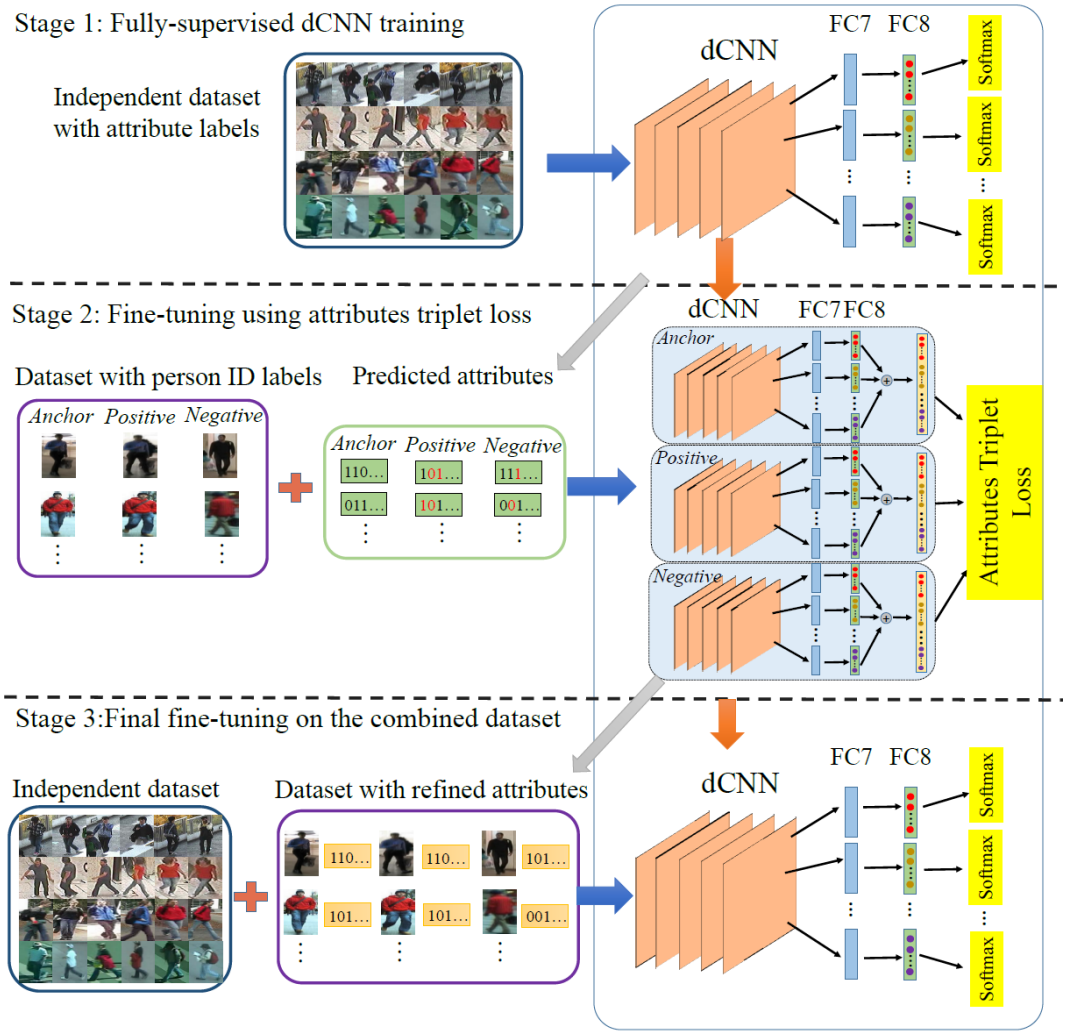
- Propose a Semi-supervised Multi-Type Attribute Learning (SMTAL)
- Uses a deeper network：VGG-16，13 conv layers
- Takes the correlations between attributes into consideration
  - Divide the attributes into sub-groups
  - in each sub-group only one attribute could be positive, e.g., a group contains: male and female

Attribute groups

Type	Attribute Label
Age	personalLess15, personalLess30, personalLess45, personalLess60, personalLarger60
Gender	personalFemale, personalMale
CarryObject	carryingBabyBuggy, carryingBackpack, carryingOther, carryingShoppingTro, carryingUmbrella, carryingFolder, carryingLuggageCase, carryingMessengerBag, carryingNothing, carryingPlasticBags, carryingSuitcase
AccessoryObject	accessoryHeadphone, accessoryHairBand, accessoryHat, accessoryKerchief, accessoryMuffler, accessoryNothing, accessorySunglasses
SleeveStyle	upperBodyNoSleeve, upperBodyShortSleeve, upperBodyLongSleeve
UpperStyle	upperBodyCasual, upperBodyFormal
UpperType	upperBodyJacket, upperBodyLogo, upperBodyPlaid, upperBodyThinStripes, upperBodySuit, upperBodySweater, upperBodyThickStripes, upperBodyTshirt, upperBodyOther, upperBodyVNeck
LowerStyle	lowerBodyCasual, lowerBodyFormal
LowerType	lowerBodyCapri, lowerBodyHotPants, lowerBodyJeans, lowerBodyLongSkirt, lowerBodyPlaid, lowerBodyThinStripes, lowerBodyShorts, lowerBodyShortSkirt, lowerBodySuits, lowerBodyTrousers
HairStyle	hairBald, hairShort, hairLong
FootStyle	footwearBoots, footwearLeatherShoes, footwearSandals, footwearShoes, footwearSneakers, footwearStocking
UpperColor	upperBodyBlack, upperBodyBlue, upperBodyBrown, upperBodyGreen, upperBodyGrey, upperBodyOrange, upperBodyPink, upperBodyPurple, upperBodyRed, upperBodyWhite, upperBodyYellow
LowerColor	lowerBodyBlack, lowerBodyBlue, lowerBodyBrown, lowerBodyGreen, lowerBodyGrey, lowerBodyOrange, lowerBodyPink, lowerBodyPurple, lowerBodyRed, lowerBodyWhite, lowerBodyYellow
HairColor	hairBlack, hairBlue, hairBrown, hairGreen, hairGrey, hairOrange, hairPink, hairPurple, hairRed, hairWhite, hairYellow
FootColor	footwearBlack, footwearBlue, footwearBrown, footwearGreen, footwearGrey, footwearOrange, footwearPink, footwearPurple, footwearRed, footwearWhite, footwearYellow

C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

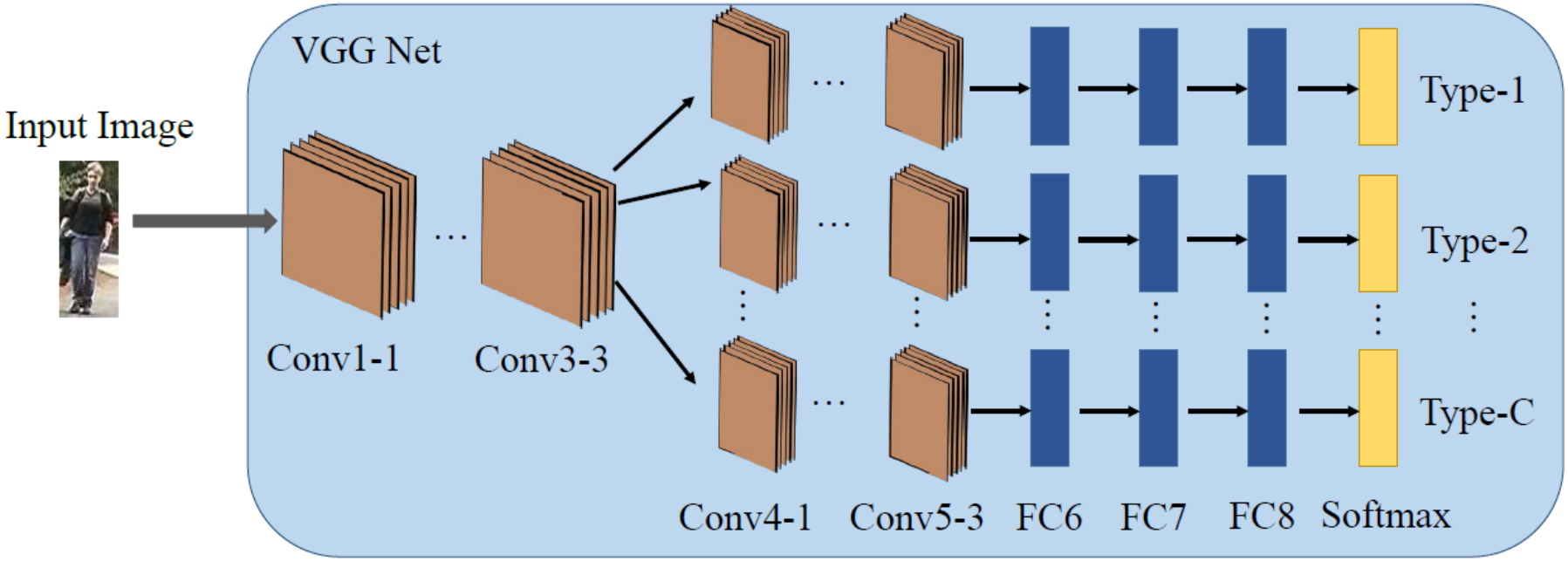
framework



C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Network structure

- Each group shares several conv layers, has independent classification layer



C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Attribute prediction accuracy

Types	Number		Validation(%)		VIPeR(%)		PRID(%)		GRID(%)	
			Top-1	Top-2	Top-1	Top-2	Top-1	Top-2	Top-1	Top-2
Age	5	$Stage_1$	86.41	97.03	44.34	80.42	48.13	77.35	42.56	78.35
		$Stage_{1\&2}$	84.56	95.18	41.09	77.85	46.03	75.45	39.87	76.42
		$Stage_{1\&3}$	**86.47**	**97.51**	45.02	80.33	48.23	78.24	42.45	77.99
		SMTAL	86.02	96.87	**46.99**	**81.84**	**50.43**	**79.13**	**45.13**	**81.84**
Sex	2	$Stage_1$	**96.53**	**100.00**	70.05	**100.00**	67.50	**100.00**	67.82	**100.00**
		$Stage_{1\&2}$	95.47	**100.00**	69.12	**100.00**	66.43	**100.00**	66.34	**100.00**
		$Stage_{1\&3}$	96.49	**100.00**	70.13	**100.00**	67.66	**100.00**	68.03	**100.00**
		SMTAL	95.83	**100.00**	**71.20**	**100.00**	**69.64**	**100.00**	**69.73**	**100.00**
CarryObject	11	$Stage_1$	**86.35**	**96.31**	26.46	46.82	48.13	77.35	26.96	41.28
		$Stage_{1\&2}$	84.75	95.92	25.12	47.03	46.03	75.45	25.47	41.96
		$Stage_{1\&3}$	86.32	95.69	27.19	46.71	48.23	78.24	26.88	41.63
		SMTAL	85.08	94.24	**29.27**	**49.01**	**31.71**	**54.82**	**28.74**	**44.53**
AccessoryObject	7	$Stage_1$	**91.86**	**97.54**	44.32	66.33	57.66	75.58	52.45	83.77
		$Stage_{1\&2}$	90.25	96.44	43.21	65.74	56.41	74.93	51.68	83.40
		$Stage_{1\&3}$	91.55	97.19	44.83	66.85	57.38	75.69	51.85	83.79
		SMTAL	90.62	97.40	**46.17**	**69.10**	**60.14**	**78.85**	**54.19**	**86.06**
SleeveStyle	3	$Stage_1$	99.14	**100.00**	80.17	93.73	85.37	95.62	42.71	77.96
		$Stage_{1\&2}$	98.75	**100.00**	79.64	93.46	84.33	94.23	42.22	77.68
		$Stage_{1\&3}$	**99.93**	**100.00**	79.38	93.63	84.96	95.73	43.09	76.93
		SMTAL	98.96	**100.00**	**85.76**	**98.73**	**87.09**	**97.88**	**45.13**	**81.84**
UpperStyle	2	$Stage_1$	**99.01**	**100.00**	90.17	**100.00**	82.64	**100.00**	80.53	**100.00**
		$Stage_{1\&2}$	97.85	**100.00**	89.25	**100.00**	82.01	**100.00**	80.11	**100.00**
		$Stage_{1\&3}$	98.69	**100.00**	90.33	**100.00**	82.49	**100.00**	81.36	**100.00**
		SMTAL	98.96	**100.00**	**92.96**	**100.00**	**84.10**	**100.00**	**83.63**	**100.00**
UpperType	10	$Stage_1$	85.31	99.36	62.57	83.71	57.48	79.13	55.97	76.53
		$Stage_{1\&2}$	84.27	99.17	61.48	83.42	57.39	79.11	55.48	75.68
		$Stage_{1\&3}$	**86.33**	**100.00**	61.34	83.66	57.24	78.54	54.76	76.42
		SMTAL	84.90	98.44	**64.34**	**85.93**	**60.81**	**81.56**	**58.99**	**79.83**

Types	Number		Validation(%)		VIPeR(%)		PRID(%)		GRID(%)	
			Top-1	Top-2	Top-1	Top-2	Top-1	Top-2	Top-1	Top-2
LowerStyle	2	$Stage_1$	96.27	**100.00**	90.23	**100.00**	81.97	**100.00**	79.19	**100.00**
		$Stage_{1\&2}$	95.69	**100.00**	89.96	**100.00**	81.05	**100.00**	78.56	**100.00**
		$Stage_{1\&3}$	**96.42**	**100.00**	90.15	**100.00**	82.01	**100.00**	79.08	**100.00**
		SMTAL	96.33	**100.00**	**93.24**	**100.00**	**84.44**	**100.00**	**81.70**	**100.00**
LowerType	10	$Stage_1$	**93.55**	**100.00**	70.39	90.14	64.32	89.44	41.75	64.35
		$Stage_{1\&2}$	92.29	99.42	70.11	90.27	63.57	89.28	40.19	63.70
		$Stage_{1\&3}$	93.27	99.33	69.74	89.36	64.68	88.93	41.38	63.77
		SMTAL	91.67	98.96	**74.92**	**95.94**	**66.20**	**91.66**	**44.04**	**67.03**
HairStyle	3	$Stage_1$	96.79	**100.00**	68.59	94.19	70.13	94.53	66.23	95.43
		$Stage_{1\&2}$	96.23	**100.00**	67.13	94.00	68.47	94.53	66.12	94.13
		$Stage_{1\&3}$	**96.93**	**100.00**	68.48	94.32	69.88	94.89	67.05	95.00
		SMTAL	96.88	**100.00**	**71.42**	**95.91**	**71.51**	**96.81**	**68.82**	**96.03**
FootStyle	6	$Stage_1$	71.53	95.06	35.38	64.63	45.34	72.53	30.15	55.91
		$Stage_{1\&2}$	71.59	94.69	33.62	63.92	44.43	72.04	29.96	56.82
		$Stage_{1\&3}$	**71.66**	**95.88**	34.96	64.55	46.02	73.10	30.05	56.39
		SMTAL	70.83	94.79	**38.36**	**66.25**	**47.75**	**74.14**	**33.80**	**60.13**
UpperColor	11	$Stage_1$	79.35	**93.23**	44.33	67.11	28.21	51.03	30.25	52.47
		$Stage_{1\&2}$	78.02	91.68	42.96	66.49	27.63	50.41	30.23	51.39
		$Stage_{1\&3}$	**79.44**	91.99	45.36	66.52	28.04	50.88	31.18	53.00
		SMTAL	78.12	91.15	**47.92**	**69.13**	**28.97**	**52.31**	**32.90**	**55.04**
LowerColor	11	$Stage_1$	**95.62**	98.96	47.76	72.36	38.06	70.53	50.91	76.35
		$Stage_{1\&2}$	94.18	98.78	46.33	71.45	37.82	69.65	50.13	76.42
		$Stage_{1\&3}$	94.37	**100.00**	48.25	71.88	38.27	70.48	49.97	75.99
		SMTAL	94.23	98.44	**51.75**	**75.69**	**40.99**	**72.82**	**52.18**	**77.89**
HairColor	11	$Stage_1$	96.39	**99.13**	55.37	76.14	48.11	80.37	44.35	64.50
		$Stage_{1\&2}$	95.44	98.27	54.23	75.38	46.27	81.04	43.19	65.39
		$Stage_{1\&3}$	**96.50**	98.69	56.30	76.23	48.09	82.10	43.96	66.42
		SMTAL	95.84	97.92	**57.74**	**79.30**	**50.80**	**84.49**	**47.09**	**70.71**
FootColor	11	$Stage_1$	**93.68**	**98.66**	44.28	68.13	62.45	83.69	30.30	60.13
		$Stage_{1\&2}$	92.71	96.75	43.19	67.48	60.97	82.47	28.56	59.48
		$Stage_{1\&3}$	93.07	97.58	45.07	68.25	62.37	83.02	30.66	59.88
		SMTAL	92.19	96.88	**46.03**	**70.84**	**67.14**	**84.81**	**34.48**	**63.80**

Examples of predicted attributes

personalLess45
personalMale
carryingMessengerBag
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyPlaid
lowerBodyCasual
lowerBodyTrousers
hairShort
footwearLeatherShoes
upperBodyGreen
lowerBodyBlack
hairYellow
footwearBrown

personalLess30
personalMale
carryingBackpack
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyThickStripes
lowerBodyCasual
lowerBodyTrousers
hairShort
footwearSneakers
upperBodyRed
lowerBodyGrey
hairGrey
footwearBrown

personalLess30
personalMale
carryingBackpack
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyPlaid
lowerBodyCasual
lowerBodyTrousers
hairShort
footwearSneakers
upperBodyBrown
lowerBodyGrey
hairGrey
footwearBrown

personalLess30
personalFemale
carryingNothing
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyOther
lowerBodyCasual
lowerBodyJeans
hairLong
footwearShoes
upperBodyRed
lowerBodyGrey
hairBlack
footwearBlack

personalLess30
personalMale
carryingNothing
accessoryHat
upperBodyNoSleeve
upperBodyCasual
upperBodySweater
lowerBodyCasual
lowerBodyTrousers
hairLong
footwearSneakers
upperBodyWhite
lowerBodyWhite
hairBlack
footwearWhite

personalLarger60
personalFemale
carryingPlasticBags
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyJacket
lowerBodyCasual
lowerBodyTrousers
hairShort
footwearLeatherShoes
upperBodyGrey
lowerBodyBrown
hairWhite
footwearBlack

personalLess30
personalFemale
carryingLuggageCase
accessoryNothing
upperBodyLongSleeve
upperBodyCasual
upperBodyThickStripes
lowerBodyCasual
lowerBodyTrousers
hairLong
footwearSneakers
upperBodyGrey
lowerBodyBrown
hairYellow
footwearBlack

personalLess30
personalMale
carryingBackpack
accessorySunglasses
upperBodyShortSleeve
upperBodyCasual
upperBodyTshirt
lowerBodyCasual
lowerBodyShorts
hairShort
footwearSneakers
upperBodyGreen
lowerBodyWhite
hairBlack
footwearWhite

C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Comparsions-1

- VIPeR：

Methods		Rank 1	Rank 5	Rank 10	Rank 20
Metric Learning based ReID	RPML [24]	27.0	57.0	69.0	83.0
	Salmatch [63]	30.2	52.4	65.5	79.1
	LMF [46]	29.1	52.3	65.9	80.0
	KISSME [27]	19.6	47.5	62.2	77.0
	KCCA [79]	37.3	71.4	**84.6**	92.3
	kLFDA [28]	32.2	65.8	79.7	90.9
	LOMO + XQDA [34]	40.0	68.9	81.5	91.1
	CSL [36]	34.8	68.7	82.3	91.8
	MLAPG [37]	40.7	69.9	82.3	92.4
	TSR [80]	31.6	68.6	82.8	**94.6**
	EPKFM [33]	36.8	70.4	83.7	91.7
Traditional Attributes based ReID	AIR [39]	18.0	38.8	51.1	71.2
	OAR [40]	21.4	41.5	55.2	71.5
	LORAE [87]	42.3	72.2	81.6	89.6
Deep Learning based ReID	IDLA [52]	34.8	54.3	76.5	87.6
	DML [51]	28.2	59.3	73.5	86.4
	Deep-RDC [53]	40.5	60.8	70.4	84.4
	Deep-TCP [54]	**47.8**	**74.7**	**84.8**	**91.1**
	Stage1	36.2	63.9	73.5	84.7
Proposed	**SSDAL**	37.9	65.5	75.6	88.4
	SSDAL + XQDA	43.5	71.8	81.5	89.0
	SMTAL	39.7	66.9	76.5	86.6
	SMTAL + XQDA	47.1	71.5	80.3	88.2

Comparsions-2

- PRID:

Methods	Rank 1	Rank 5	Rank 10	Rank 20
RPML [24]	4.8	14.3	21.6	30.2
PRDC [31]	4.5	12.6	19.7	29.5
RSVM [21]	6.8	16.5	22.7	31.5
Salmatch [63]	4.9	17.5	26.1	33.9
LMF [46]	12.5	23.9	30.7	36.5
PCCA [23]	3.5	10.9	17.9	27.1
KISSME [27]	4.1	12.8	21.1	31.8
kLFDA [28]	7.6	18.9	25.6	37.4
KCCA [79]	14.5	34.3	46.7	59.1
LOREA [87]	18.0	37.4	50.1	66.6
LOMO + XQDA [34]	15.3	35.7	41.2	53.8
MLAPG [37]	16.6	33.1	41.4	52.5
Deep-TCP [54]	22.0	-	47.0	57.0
$Stage_1$	19.6	46.7	55.1	66.4
SSDAL	20.1	47.4	55.7	68.6
SSDAL + XQDA	22.6	48.7	57.8	69.2
SMTAL	22.4	47.8	56.8	67.6
SMTAL + XQDA	**24.4**	**52.3**	**62.5**	**74.2**

C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Comparsions-3

- GRID：

Methods	Rank 1	Rank 5	Rank 10	Rank 20
PRDC [31]	9.7	22.0	33.0	44.3
RSVM [21]	10.2	24.6	33.3	43.7
MRank-PRDC [67]	11.1	26.1	35.8	46.6
MRank-RSVM [67]	12.2	27.8	36.3	49.3
RQDA [92]	15.2	30.1	39.2	49.3
EPKFM [33]	16.3	35.8	46.0	57.6
LOMO + XQDA [34]	16.6	35.4	41.8	52.4
$Stage_1$	17.5	34.5	42.8	55.3
SSDAL	19.1	35.6	45.8	58.1
SSDAL + XQDA	22.4	39.2	48.0	58.4
SMTAL	19.2	38.1	47.8	58.7
SMTAL + XQDA	**23.4**	**39.6**	**49.8**	**60.3**

C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Comparsions-4

- Market：

Single Query	Rank 1	mAP
Salmatch [63]	20.5	8.2
SDALF [13]	33.5	13.5
BGG [19]	34.4	14.1
KISSME [27]	40.5	19.0
MFA [28]	45.7	18.2
kLFDA [28]	**51.3**	24.4
LOMO + XQDA [34]	43.8	22.2
SSDAL	39.4	19.6
SMTAL	49.5	**29.2**
Multiple Query	Rank 1	MAP
BGG+MultiQ_max [19]	42.1	18.5
kLFDA [28]	52.7	27.4
LOMO + XQDA [34]	54.1	28.4
SSDAL	49.0	25.8
SMTAL	**56.6**	**31.2**

C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Pattern Recognition*, 2018.

Summary

- Attribute is robust, effective, and efficient for Person Re-ID
- Attribute correlation is important
- Our deep model effectively learns attributes with a limited number of labeled data
- It works well even when the training and testing data are independent

Reference:

1. C. Su, **S. Zhang**, J. Xing, Q. Tian, and W. Gao. Multi-type attributes driven multi-camera person re-identification, *Elsevier Pattern Recognition* (*PR*), 75: 77-89, 2018.
2. C. Su, F. Yang, **S. Zhang**, Q. Tian, L. S. Davis, and W. Gao. Multi-Task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (*T-PAMI*), 2018
3. C. Su, **S. Zhang**, J. Xing, W. Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. *European Conference on Computer Vision* (*ECCV*), 2016.
4. C. Su, F. Yang, **S. Zhang**, Q. Tian, L. Davis, W. Gao. Multi-Task Learning with Low Rank Attribute Embedding for Person Re-identification, *IEEE International Conference on Computer Vision* (*ICCV*), 2015.

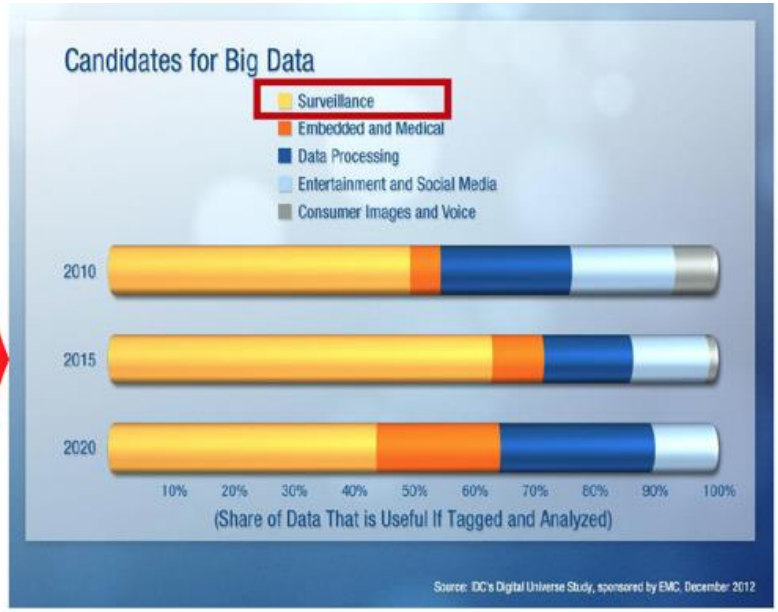
Outline

- Attribute Learning
  - Attribute embedding in multi-task learning
  - Attribute learning with deep models
- Scalable Person Re-ID
  - Deep feature extraction and compression
  - Off-line index optimization

Scalability is required for person re-id

- Camera is everywhere in smart city

Camera Networks in Smart City



More than half of all big data

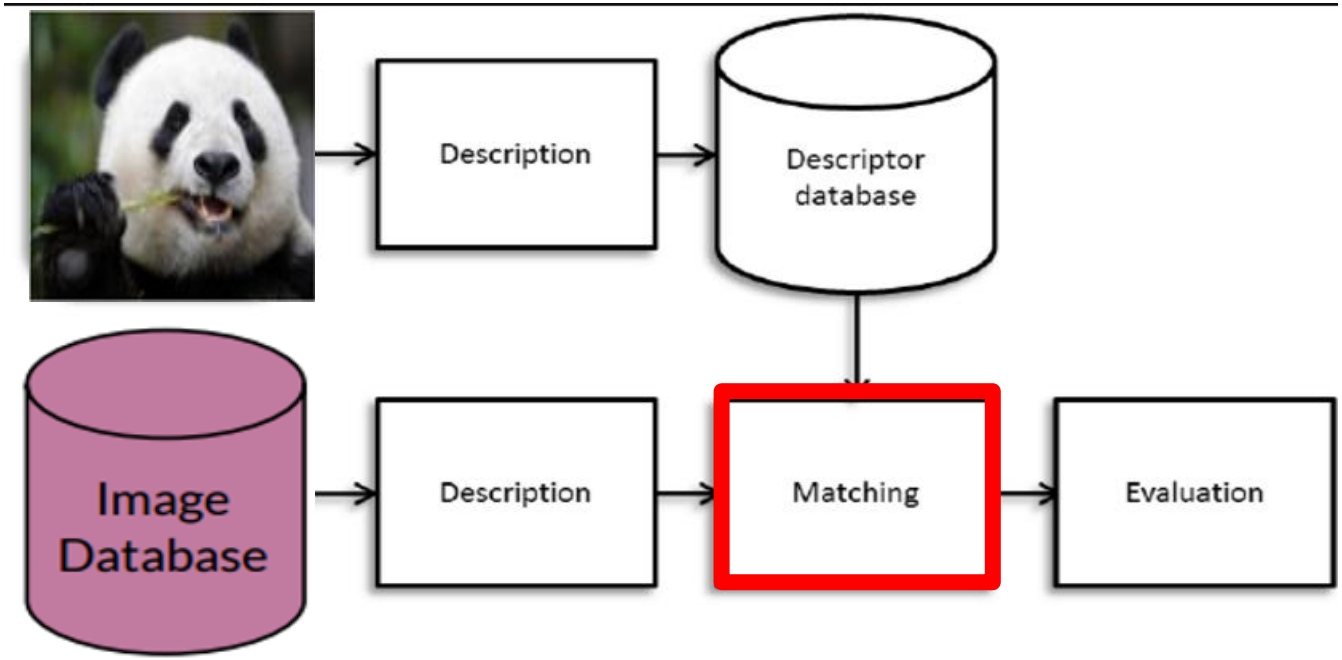
- Big data, fast growth
  - 30 millions cameras have been deployed in China
- Complicated content, rich Information
  - Cars, persons, others

Issues

- Most of existing algorithms are designed and evaluated on small-scale datasets,
  - CHUK03 consists of 14,096 images,1,467 identities,
  - Market1501 consists of 36,036 images, 1,501 identities
- Working on those datasets, existing methods focus on improving the accuracy, and pay less attention to the efficiency and generalization ability
- person Re-ID can be conducted by matching the query person image to database images and return only the images of the same person.
- This is similar to image retrieval

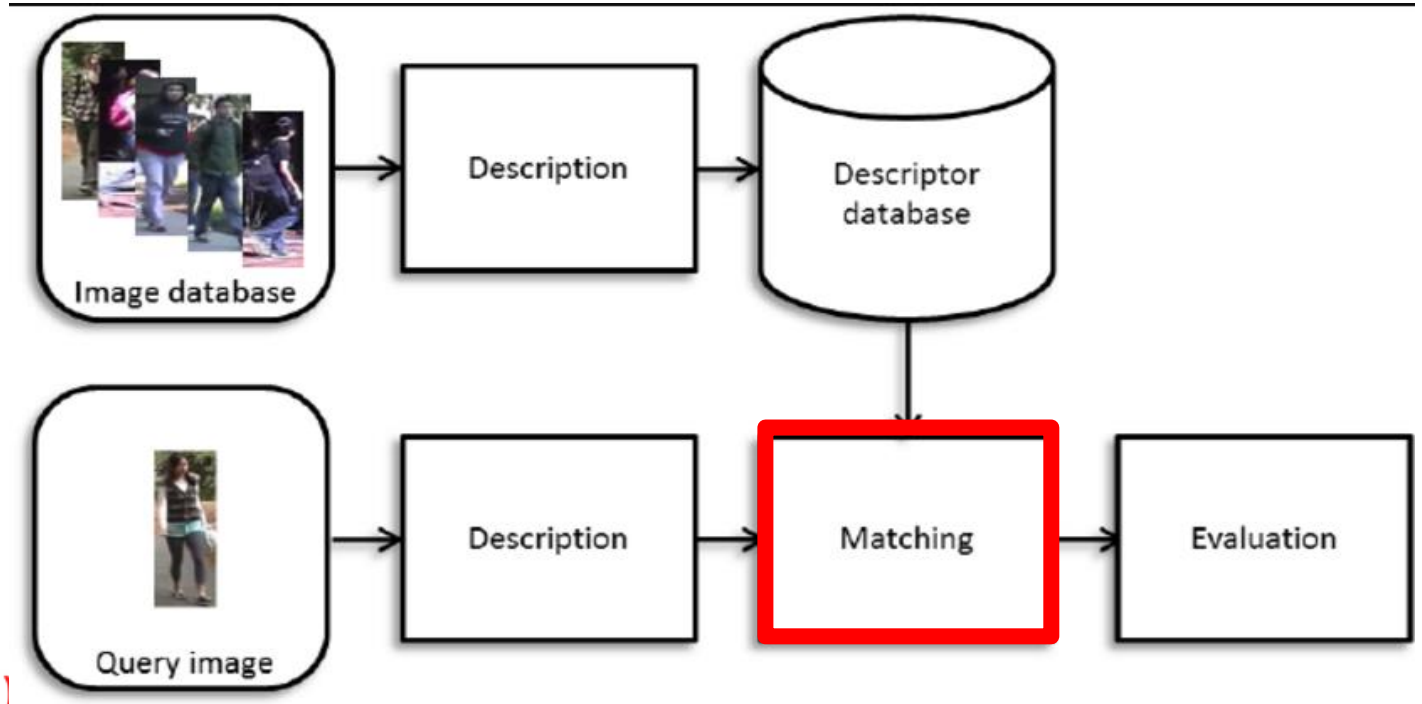
Large-Scale Person Re-ID as Retrieval

- Image Retrieval: Hash could be used for large-scale image retrieval to accelerate online matching.



Large-Scale Person Re-ID as Retrieval

- Person Re-ID is commonly conducted with linear search
- Time-consuming for large-scale dataset
- research efforts on image retrieval could make the large-scale person Re-ID matching more efficient



Challenges

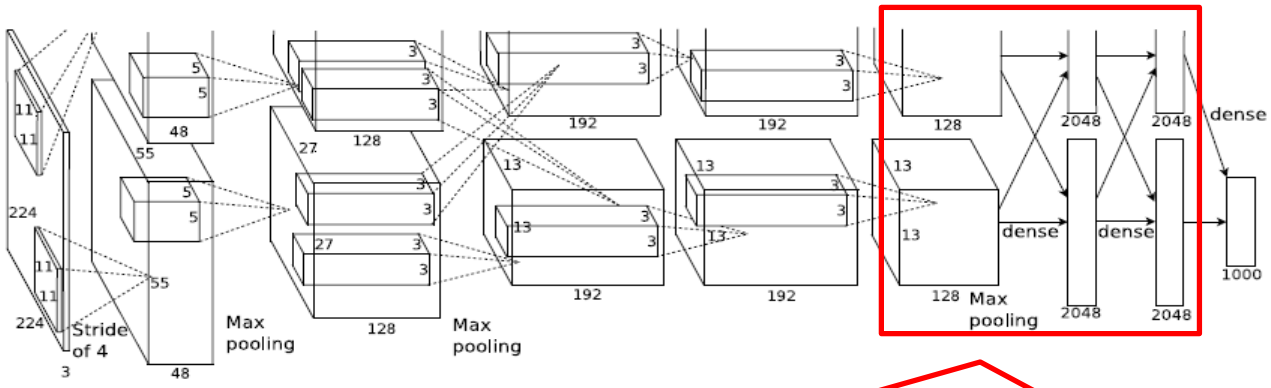
Fig. 1. Example images of four persons from Market1501 (first two) and CUHK03 (last two). The subtle differences among different persons and the large variance among images of the same person make person ReID challenging.

The key is image representation, which should be fast and discriminative for unseen persons

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Convolutional Neural Network

- AlexNet [Krizhevsky *etal*, NIPS,2012].
  - CNN =Feature Extraction + Classifier
  - Disadvantage of Fully-Connected Layer
    - Huge parameters make the network easy to overfitting.
    - Fully-connected operation makes the input for CNN is fixed size image.

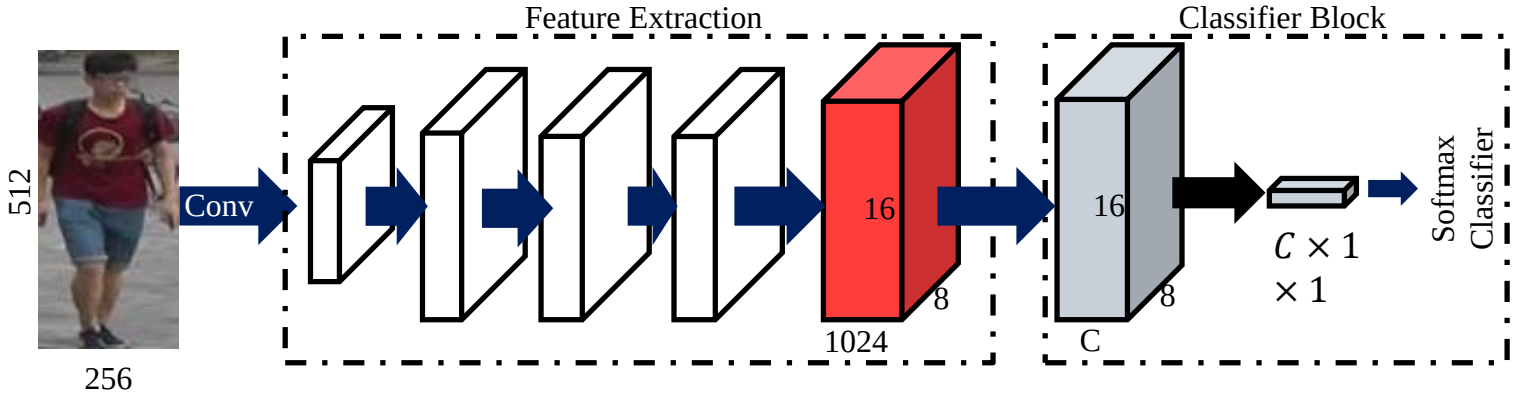


$$\text{Pool5->Fc6: } 9126 \times 4096 = 35.6M$$
$$\text{Fc6->Fc7:} 4096 \times 4096 = 16M$$

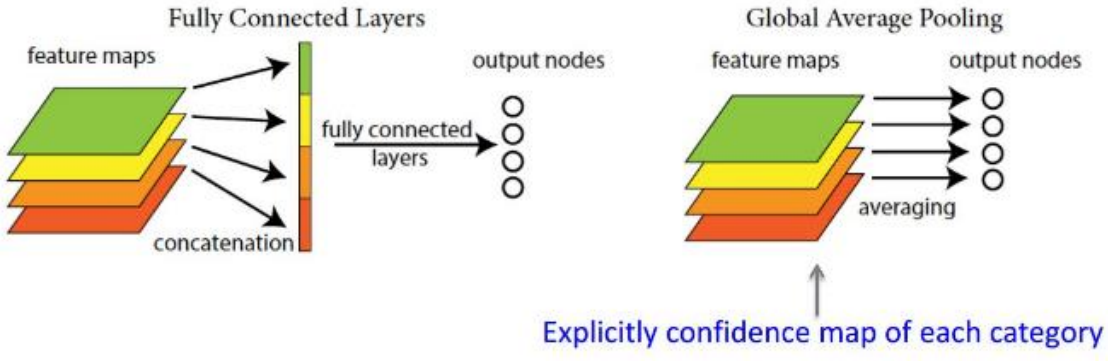
H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Conv-Net for feature learning

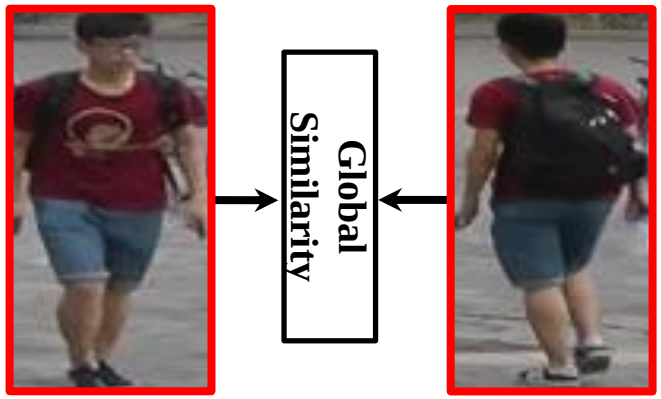
- Replace the Fully-Connected classifier with Convolutional Classifier Block.



- Convolutional Classifier Block= Convolutional Operation+ Global Average Pooling (GAP).



Global Feature Extraction



- Person Descriptor = Global Descriptor + Local Descriptor
  - Global descriptor is used to describe the global appearance of the person.
  - Global Descriptor： $\boldsymbol{f}^g$,

$$\boldsymbol{f}^g = [f_1, f_2, \ldots, f_K], f_k = \frac{1}{W \times H} \sum_{h=1}^{H} \sum_{w=1}^{W} \chi_{k,h,w}$$

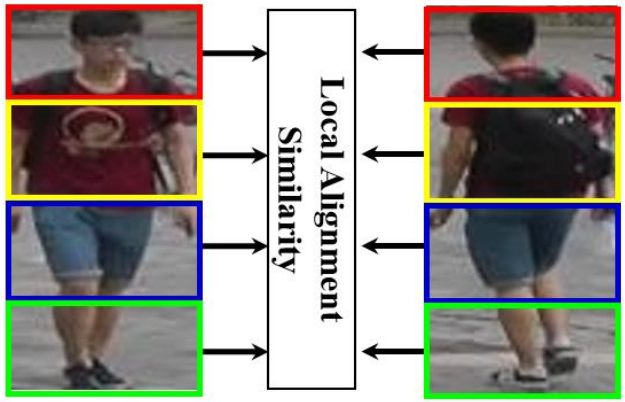
Where χ denotes the output of the last layer of feature extraction block.

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Local Feature Extraction

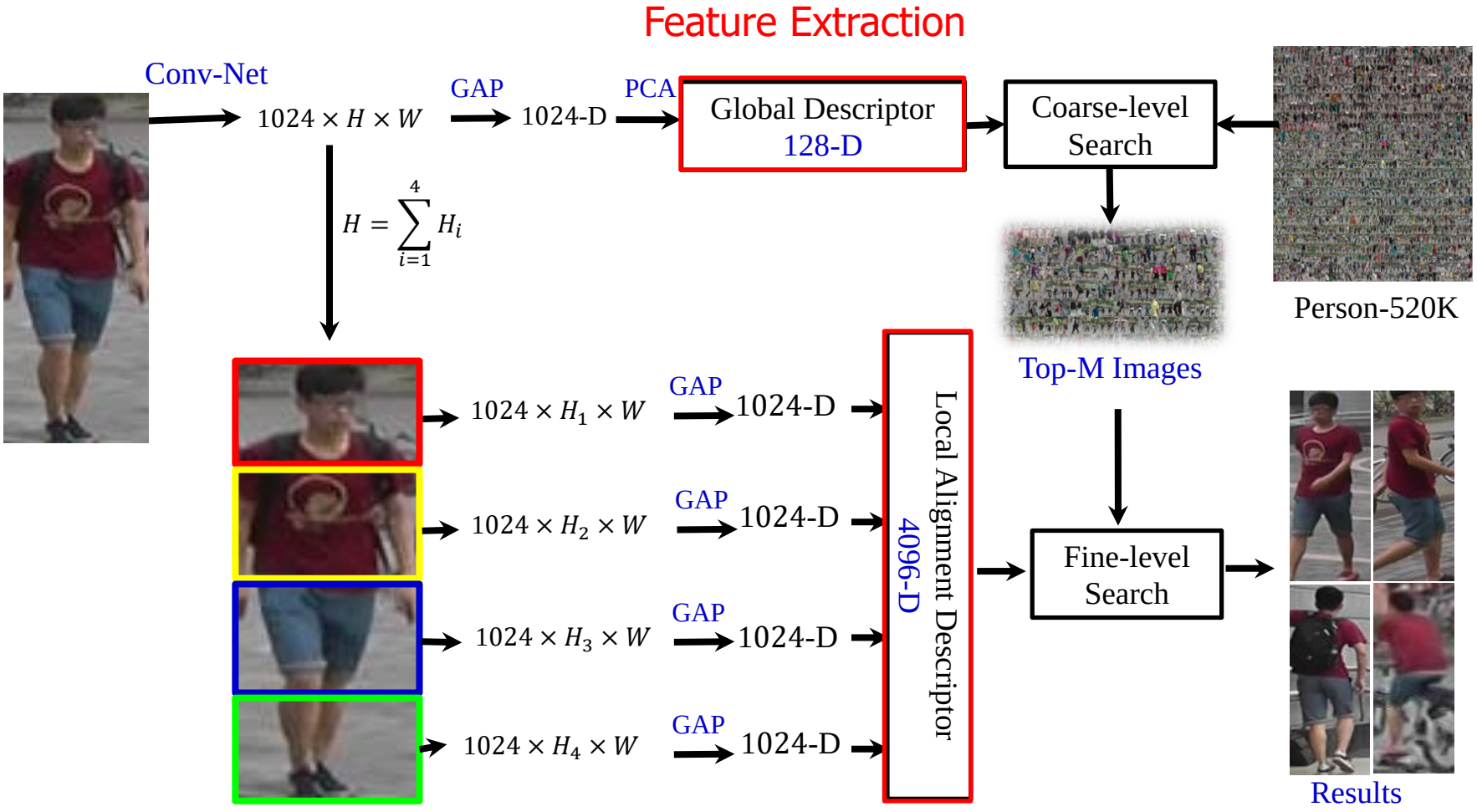
- Person Descriptor = Global Descriptor + Local Descriptor
  - Global descriptor lacks the ability to describe the local parts.
  - For person Re-ID images, most person images are weakly aligned.
  - Therefore, we divide the image into four stripes to generate local descriptors.

$$\boldsymbol{f}^{l} = \left[\boldsymbol{f}^{l_1}; \boldsymbol{f}^{l_2}; \boldsymbol{f}^{l_3}; \boldsymbol{f}^{l_4}\right],$$
$$\boldsymbol{f}^{l_i} = [f_1, f_2, \ldots, f_K],$$



$$f_k = \frac{1}{W \times \mathcal{H}} \sum_{h=(i-1)\times\mathcal{H}}^{i\times\mathcal{H}} \sum_{w=1}^{W} \chi_{k,h,w} \ , i \in [1,4] \ and \ \mathcal{H} = \frac{H}{4}$$

Proposed Approach



Online Retrieval

- Coarse-to-Fine Retrieval
  - Global Descriptor used for coarse retrieval.
    - Feature dim is reduced to 128-D
  - Local Descriptor used for fine re-ranking.
    - Use 4096-D feature for fine retrieval

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Dataset Construction

- A large-scale Person-520K dataset is constructed for evaluation.
- Combines CUHK03 and Market1501

Table 1. Details of Person-520K dataset.

	Training		Testing		Query	
	Identity	Image	Identity	Image	Identity	Image
CUHK03	736	7,029	731	5,606	731	1,461
Market1501	751	12,936	750	19,732	750	3,368
Distractors	-	-	-	500K	-	-
Total	**1,487**	**19,965**	**1,481**	**520K**	**1,481**	**4,829**

The code and Person-520K could be downloaded from:
http://www.yaohantao.com/

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Performance of Conv-Net

- Input size 224 * 224
- Convolutional Classifier block is better than fully-connected classifier.

Comparison on Market1501

Models	mAP(%)	Rank-1 (%)
AlexNet [13]	26.79	50.89
VGG16Net [14]	38.27	65.02
GoogleNet [16]	48.24	70.27
Res50Net [15]	51.48	73.69
Conv-Net	**54.86**	**76.36**

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Effect of network input size

Larger input size corresponds to better performance

Table 3. Effect of various input sizes on Market1501.

Scales	mAP(%)	Rank-1 (%)
224×224	54.86	76.36
128×64	44.57	68.27
256×128	53.50	75.89
384×192	56.04	77.64
512×256	61.9	81.5

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Performance on Person-520K

Table 6. The performance on Person-520K. "Time" denotes the retrieval time. Feature extraction takes about 50ms for Conv-Net on GPU.

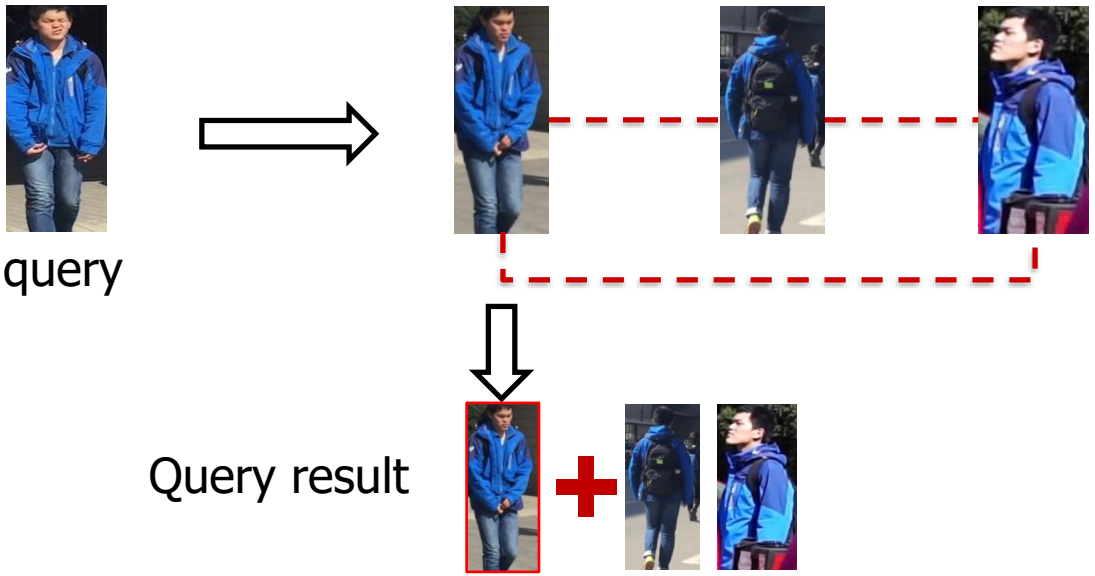
Methods	Dim	Time(ms)	mAP(%)	Rank-1(%)
AlexNet	4,096	3932	17.13	33.46
GoogleNet	1,024	960	36.38	56.05
Conv-Net_f^g	1,024	961	43.42	60.84
Conv-Net_f^g	512	500	41.06	61.79
Conv-Net_f^g	256	262	40.84	61.21
Conv-Net_f^g	128	148	39.99	59.3
Conv-Net_f	5,120	4746	46.95	64.60
coarse-level	128	150	39.99	59.3
C2F	-	180	46.74	64.58

- Reduce dimensionality with PCA drops the accuracy
- C2F achieves a good balance between accuracy and efficiency

H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. *ICME*, 2017.

Introduce indexing to ReID

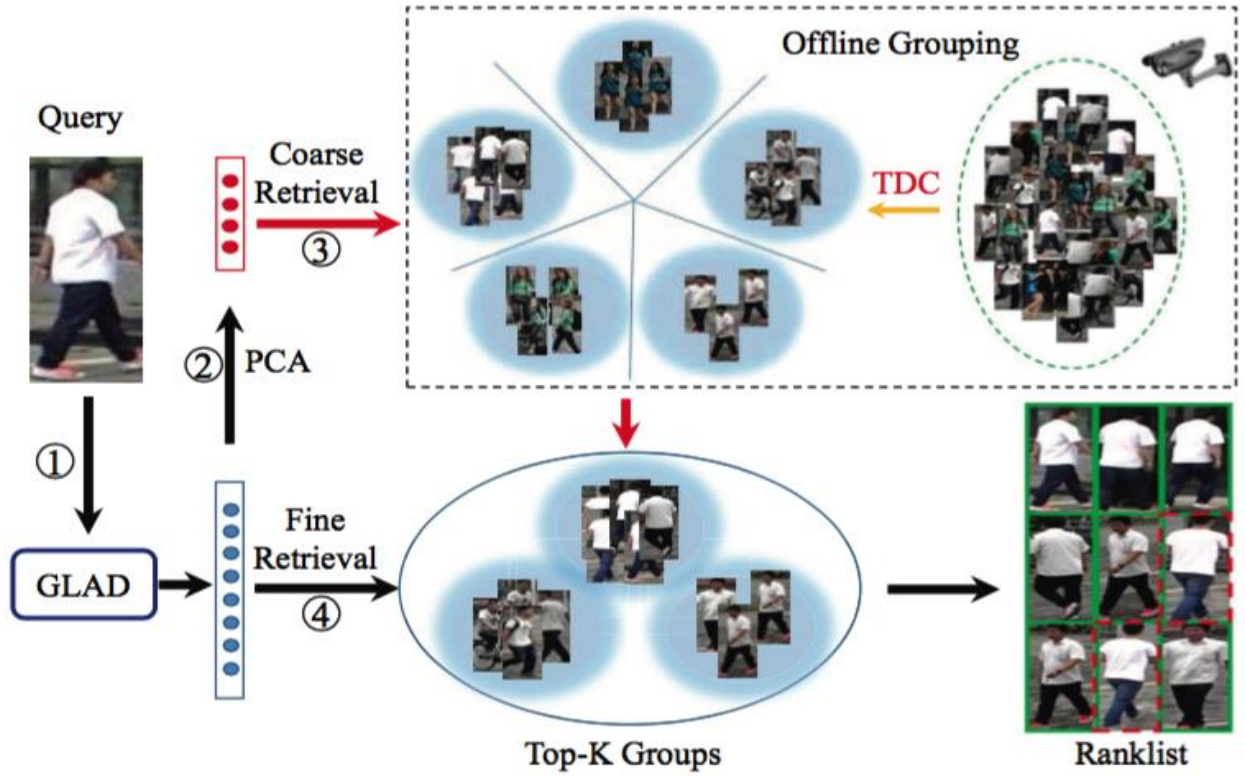
- Basic Idea
  - each person has multiple samples in the gallery set
  - Can we link those images during offline indexing?
  - The re-id accuracy can be improved, if the linkage is accurate and smart



L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. *ACM MM*, 2017.

Indexing & Retrieval Framework

- Generate image groups for indexing,
- Two retrieval stages, to retrieve groups and images, respectively.
  - Offline group generation can use more expensive features
  - Group number smaller than image number



L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. *ACM MM*, 2017.

Group Generation

- Compute the image distance with Jaccard distance

$$dis(I_i^a, I_j^b) = 1 - \frac{|\overline{KNN}(I_i^a, K) \cap \overline{KNN}(I_j^b, K)|}{|\overline{KNN}(I_i^a, K) \cup \overline{KNN}(I_j^b, K)|},$$

→ K-reciprocal neighbor set

- Propose a Two-fold Divisive Clustering (TDC)
  - Hierarchically divides image sets into two sets and ensure the dissimilarity within each group smaller than a threshold
  - Only need one parameter Θ, need not to specify the group number
- Use average pooling for group feature generation:

$$\mathbf{f}^{\mathcal{G}}(i) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{f}_j^{GLAD}(i),$$

L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. *ACM MM*, 2017.

Efficiency and Accuracy Boost

TABLE IX

RE-ID PERFORMANCE OF OUR RETRIEVAL FRAMEWORK ON MARKET1501 WITH DIFFERENT θ AND K. THE SECOND ROW DENOTES THE PERFORMANCE OF LINEAR SEARCH WITH THE ORIGINAL GLAD.

θ	K	Group Number	Dim	mAP	Rank-1	Times(ms)
-	-	19732	4096	73.9	89.9	352.0
0.2	10	16692	128	73.8	89.7	9.7
0.4	10	12785	128	74.8	[illegible]	7.7
0.6	10	8302	128	76.0	89.7	5.3
0.8	10	4911	128	76.4	89.8	4.5
0.8	20	3726	128	75.9	89.7	5.2
0.8	30	3532	128	75.4	[illegible]	6.2
0.8	40	3493	128	75.0	89.6	6.8
0.8	50	3433	128	74.7	89.6	6.5

100X speed up

Rank-1 Acc. improves 2.5%

L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. *ACM MM*, 2017.

Performance on Person520K

- Improved accuracy
- Improved efficiency

TABLE XI
RE-ID PERFORMANCE OF OUR RETRIEVAL FRAMEWORK ON PERSON-520K.

Method	Group Number	Dim	mAP	Rank-1	Speedup Ratio
[56]	$525K$	128	46.7	64.6	26X
GLAD+Linear	$525K$	4096	56.5	72.7	1X
GLAD+C2F	**199K**	128	**58.2**	**72.9**	**61X**

L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. *ACM MM*, 2017.

Conclusions

- Person ReID and image retrieval are essentially the same
- A coarse-to-fine framework is proposed for large-scale person ReID.
- Off-line indexing boosts both efficiency and accuracy

Reference:

1. H. Yao, **S. Zhang**, D. Zhang, Y. Zhang, J. Li, Y. Wang, Q. Tian. Large-Scale Person Re-Identification as Retrieval. ICME, 2017.
2. L. Wei, **S. Zhang**, H. Yao, W. Gao, and Q. Tian. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. ACM MM, 2017.

Thanks!

Homepage

www.pkuvmc.com

北京大学