Cross domain Residual Transfer Learning for Person Re-identification

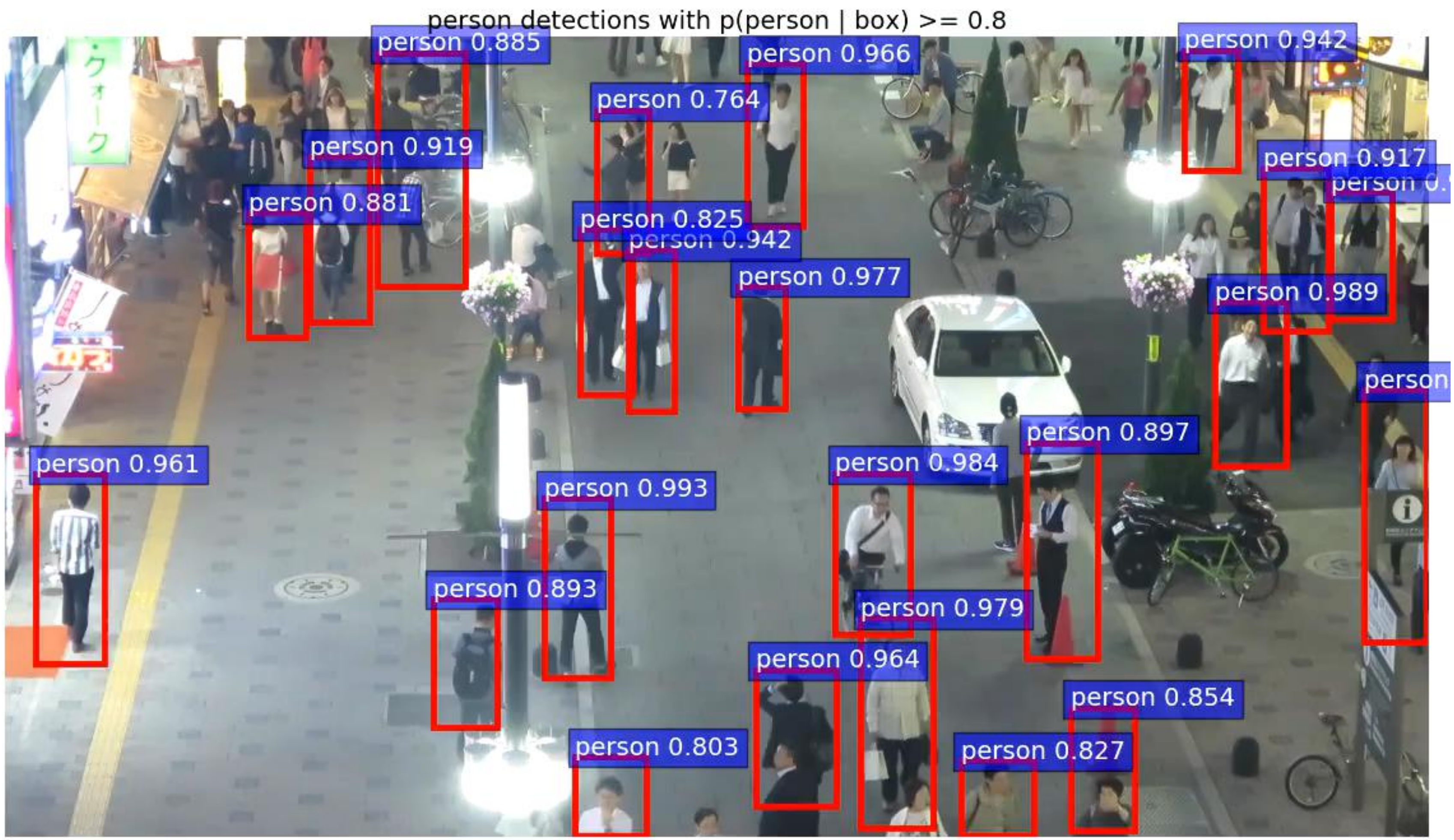
INRIA Sophia Antipolis – **STARS team**

Institut National Recherche Informatique et Automatisme

Furqan Khan and Francois.Bremond@inria.fr

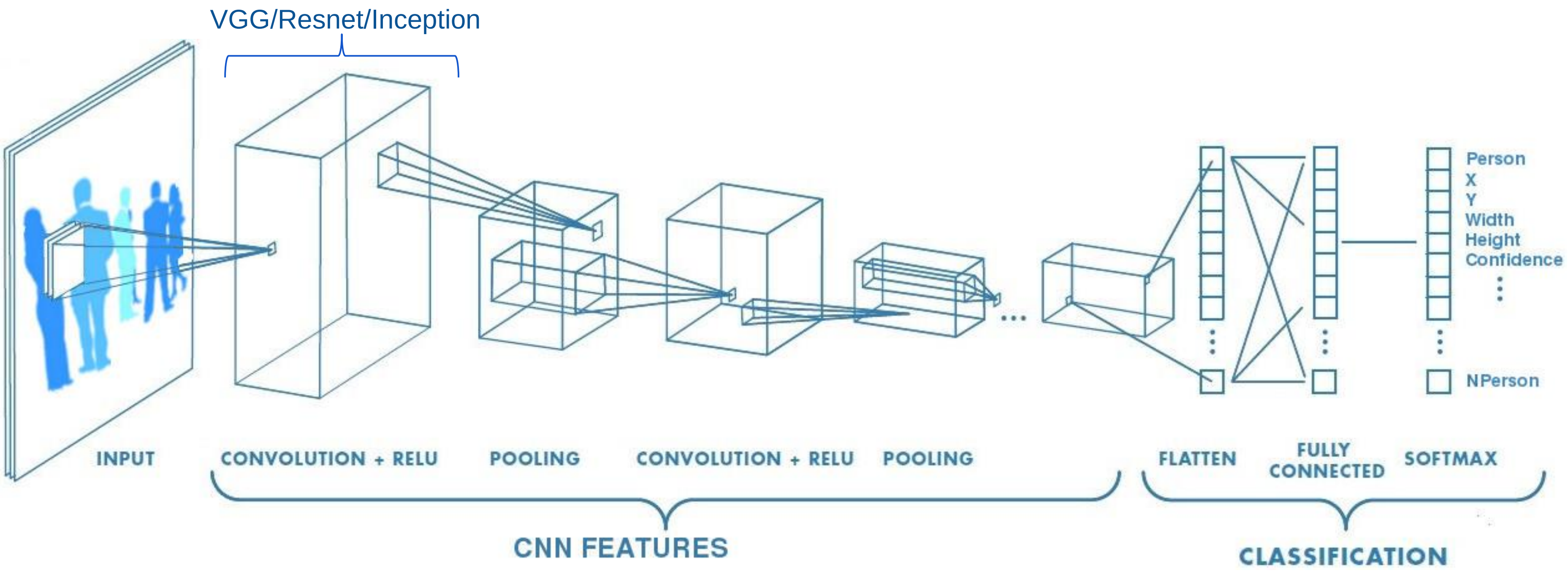
http://www-sop.inria.fr/members/Francois.Bremond/

People detection : Faster-RCNN on MOT Video Protection



CNN Architecture: RPN - RCNN - SSD

Define the deep learning people detection architecture



People Tracking (MOT)

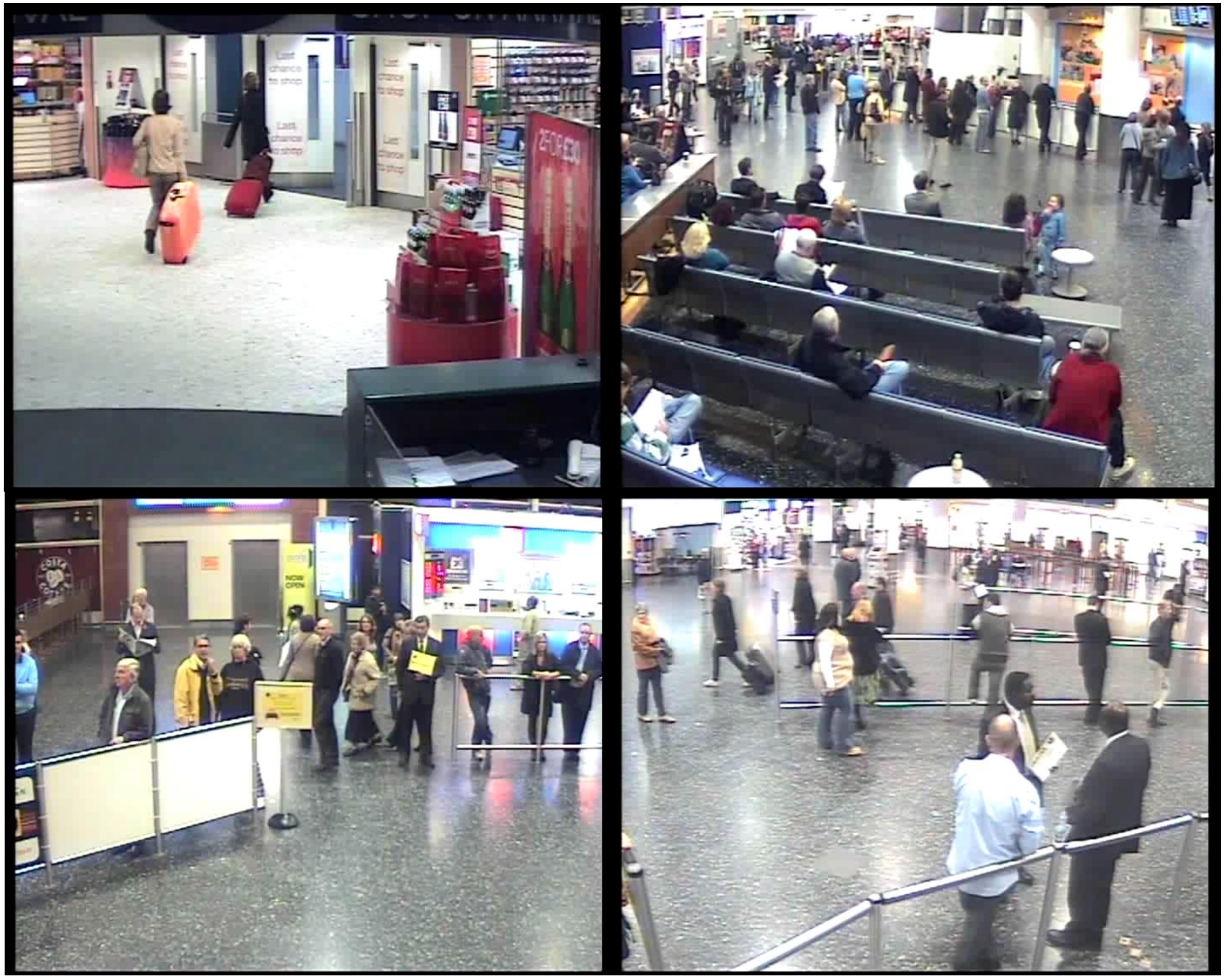
Multiple Object Tracking (MOT17) challenge:

- Our online tracker based on Residual Learning Transfer has the best performance for online trackers [AVSS18] for Mostly Tracked (MT) metric
- Results in progress, but still challenging (e.g. Objects are too small to track)

MOT17-07-SDP

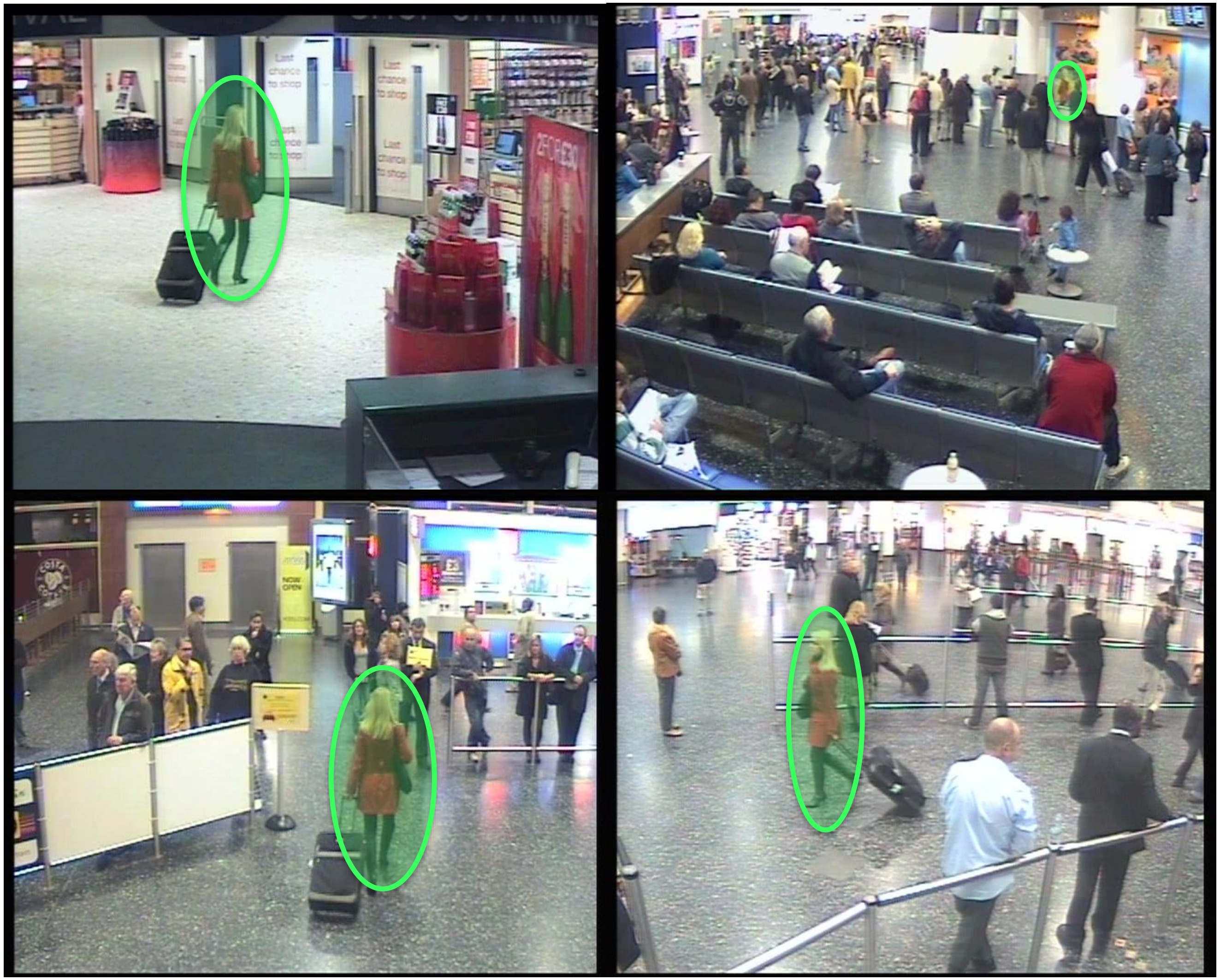
MT ↑	ML ↓	MOTA ↑	MOTP ↑	FP ↓	FN ↓	IDSw ↓	Frag ↓
18%	37%	**46%**	76%	#	#	#	#

Person Re-identification (Slawomir BAK)

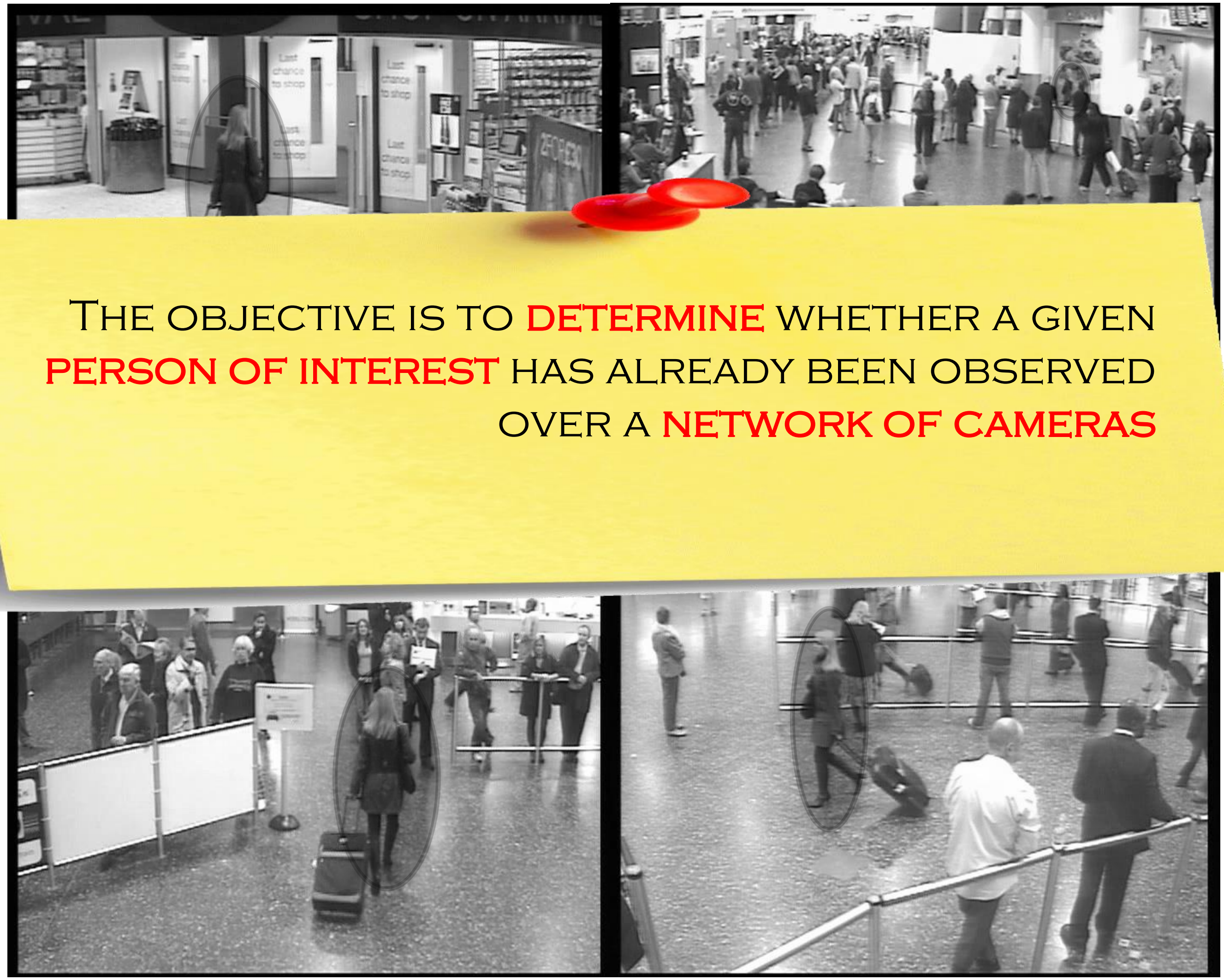


CCTV cameras, UK: 4M, London: 1.8M
Human can not perform efficient surveillance after 12minutes

Person Re-identification



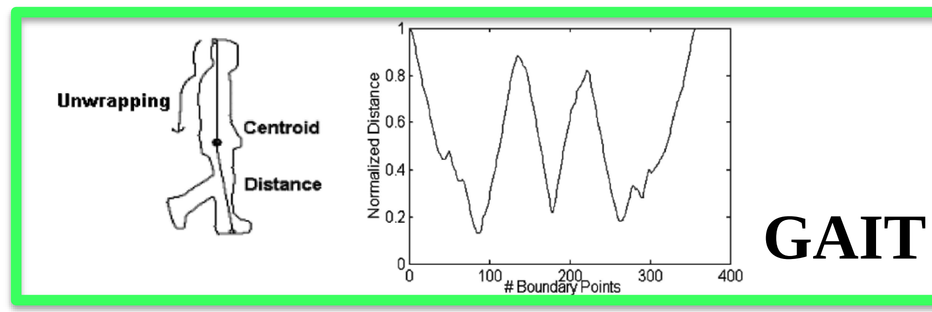
Person Re-identification Problem



The objective is to **determine** whether a given **person of interest** has already been observed over a **network of cameras**

Person Re-identification

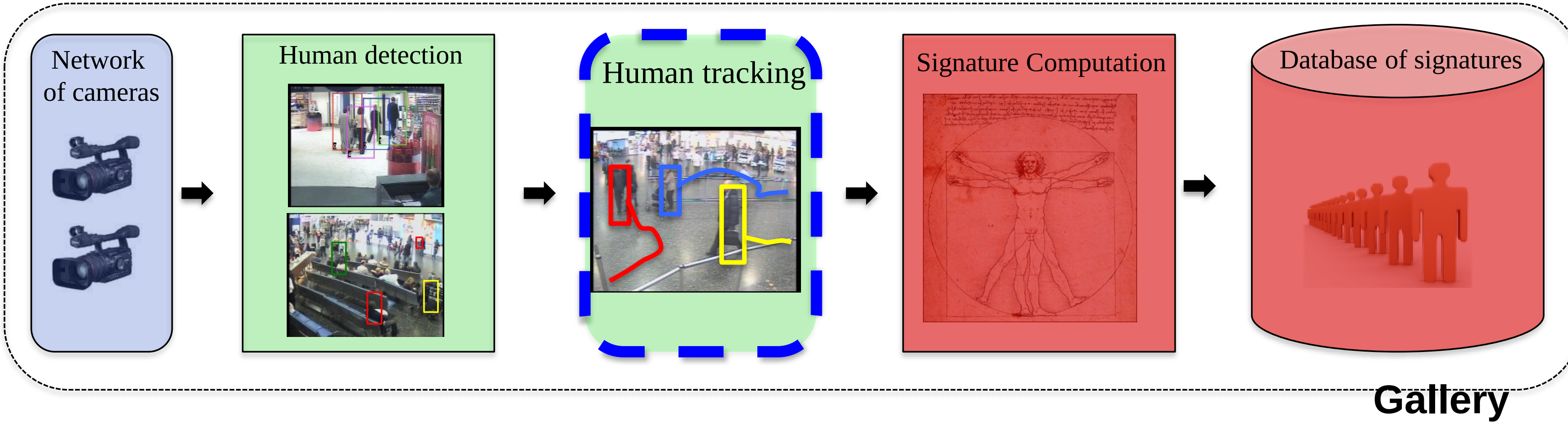

Levels



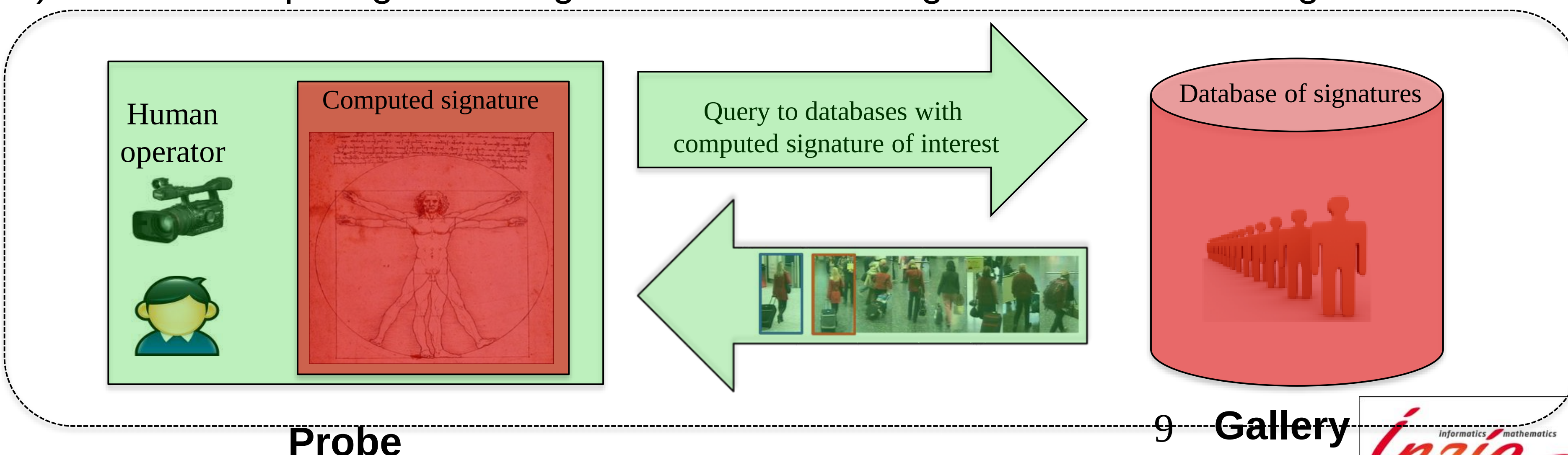
Global Appearance

How? In two steps: 1) Gallery and 2) Probe

1) **Gallery** : computing the visual signatures in the database



2) **Probe** : computing a new signature and retrieving it from the stored signatures



Main Challenges

Color

Viewpoint

Occlusion

Detection

Discriminative features

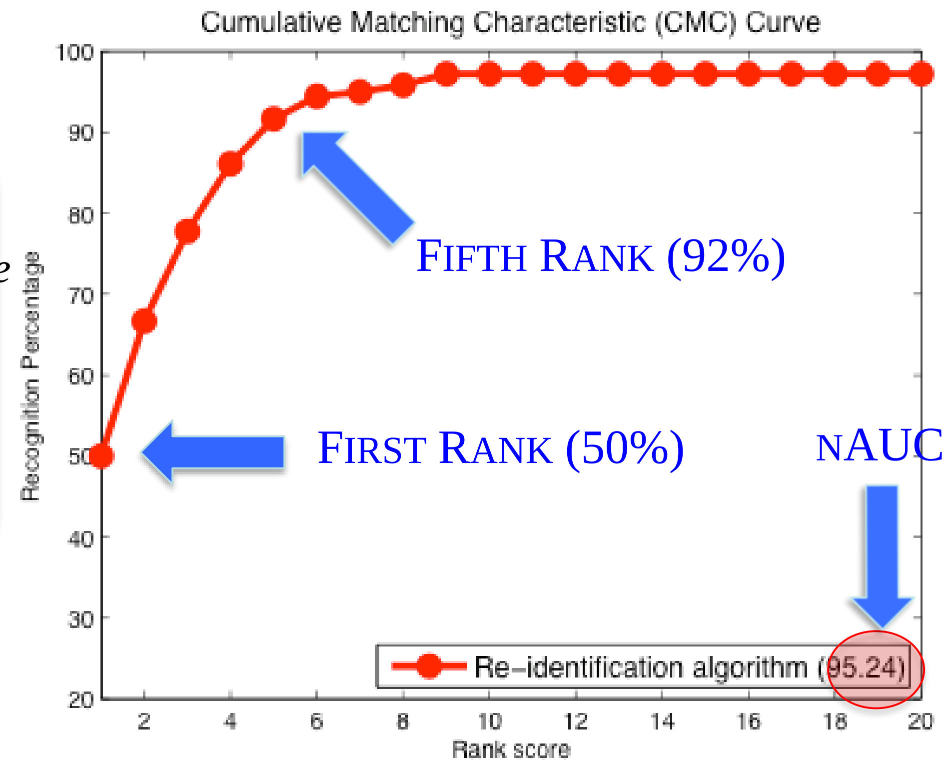
Inspired by **human memory** and in particular – **recognition memory**

People Re-identification (ReID): Performance Evaluation

Evaluation metrics

(1) ***C****umulative* ***M****atching* ***C****haracteristic curve*

(2) ***n****ormalized* ***A****rea* ***U****nder* ***C****urve* (**nAUC**) – a quantitative scalar appraisal of CMC



Evaluation scheme based on **queries**

Probe / Gallery

Comparison with state-of-the-art ReID

Bak et al, *"Boosted human re-identification using Riemannian manifolds"*, Image and Vision Computing 2011

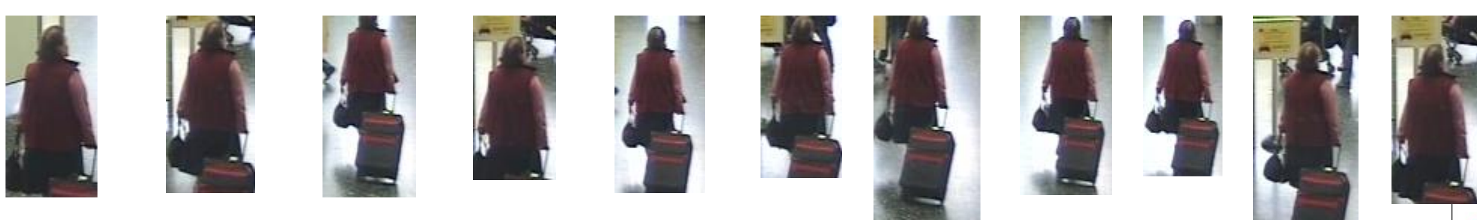
i-LIDS-MA
-40 individuals
- in average 46 images per camera
-**manually** detected

i-LIDS-AA
-100 individuals
-in average 50 images per camera
-**automatically** detected

People Re-identification (ReID) – F. Khan

Practical issues – towards real-world

- Imperfection of automated detection and tracking systems
  - Misalignment
  - Partial visibility
  - ID switches – when one track has images from two people at different time intervals (corrupted tracklets)
- Many Metrics for Multi-shot – set based metrics
  - Minimum/Average Pointwise Distance, Local metric fields, collaborative coding
- Metric (Supervised) learning improves performance BUT requires data annotation – limits scalability



People Re-identification (ReID) [Khan AVSS16]

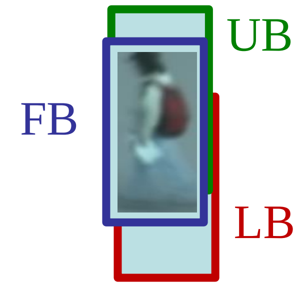
Signature representation :

- Signature = **Part Appearance Mixture** (PAM) or **Multi Channel Means** (MCM)
  - Each GMM mixture represents distribution of several feature descriptors in the image cells
- **Feature descriptors:**
  - shape - Histogram of Gradients (HoG) [Dalal05];
  - color - Color Spatio-histogram (CSH) [Zeng CVPR-W15];
  - texture – Brownian Covariance (BCov) [Bak ICIP12].
    - For re-scaled and histogram equalized images of 64 x 192 pixels
    - Separately over 3 x 11 overlapping rectangular grid
  - Local Maximal Occurrence (LOMO) [Liao CVPR15], HSCD [Zeng CVPRW15]
  - Deep Features from Conv4 or Conv5 of VGG16 Fine-Tuned
- **Metrics**
  - Minimum Pointwise Distance (MPD) => M = I, Euclidian dist. (no training)
  - KISSME => M trained using manually annotated data (supervised training)
  - UnKISSME => M trained using automated data (unsupervised training)

M: Mahalanobis Distance for KissMe

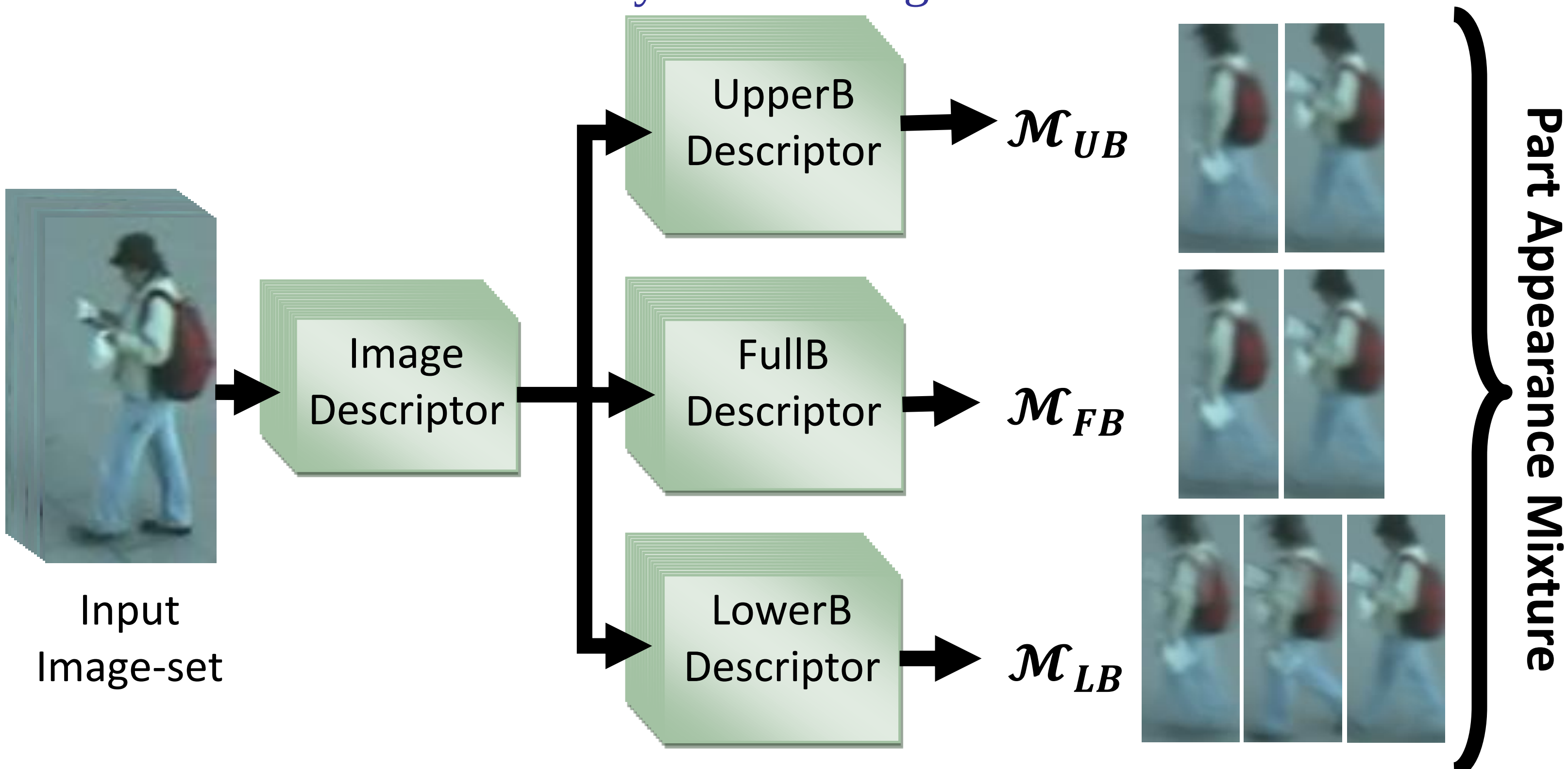
Part Appearance Mixture (PAM)

Three Bounding Boxes:



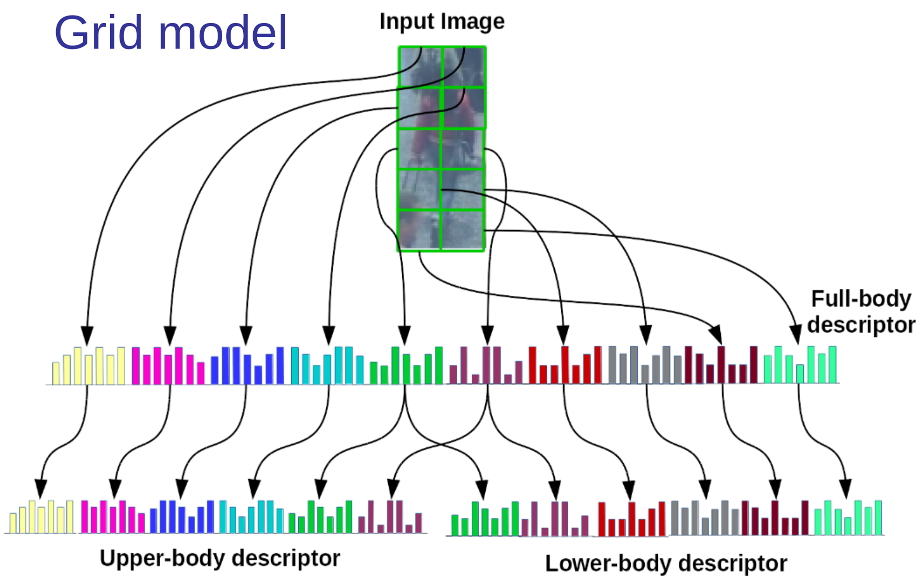
PAM model for appearance

- Three part models
- Each part a multi-modal parametric distribution
- Simultaneous mode-discovery and learning

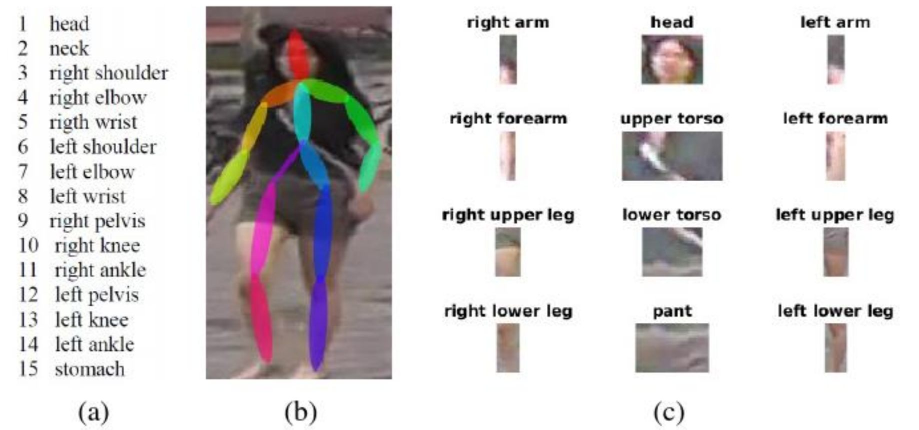


Person Re-identification
Visual Signature

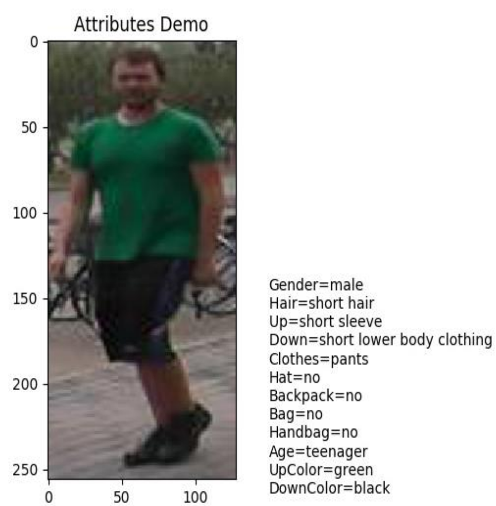
Grid model

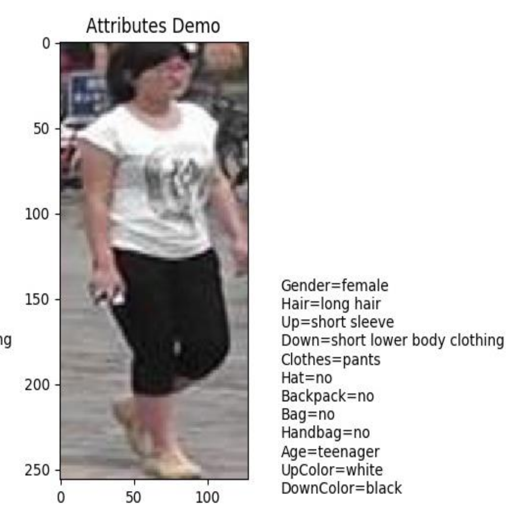


Body Part model



Semantic Attribute model





Parts share computation

Features:

- HOG – 8 bins unsigned, 11x3 grid over 192x64 image, RGB channels
- LOMO – 3 scales, HSV + SILTP histograms, max-pool horizontally

Comparison with state-of-the-art ReID

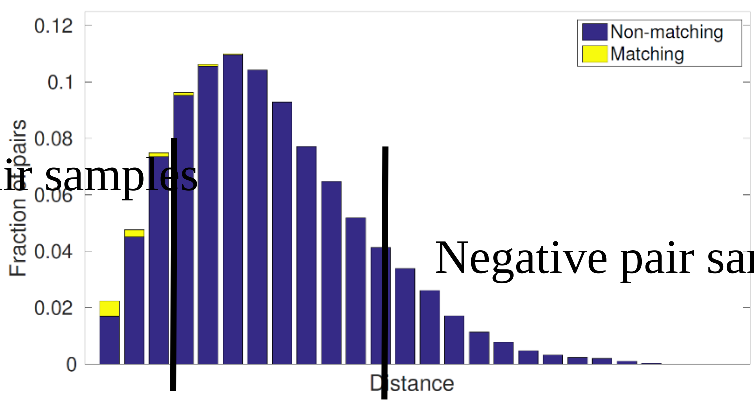
- Comparison with **supervised** methods using recognition rate at rank r in % : PRID 2011

Method	r=1	r=5	r=10	r=20
Color+DVR [Wang14]	41.8	63.8	76.7	88.3
ColorLBP+DVR [Wang14]	37.6	63.9	75.3	89.4
ColorLBP+RSVM [Wang14]	34.3	56.0	65.5	77.3
DVR [Wang14]	28.9	55.3	65.5	82.8
DSVR [Wang16]	40.0	71.7	84.5	92.2
Salience+DVR [Wang14]	41.7	64.5	77.5	88.8
SDALF+DVR [Wang14]	31.6	58.0	70.3	85.3
STFV3D+KISSME [Liu15]	64.1	87.3	89.9	92.0
MCM+KISSME[AVSS16]	**[64.3]**	86.1	**[94.5]**	**[98.0]**
PAM+LOMO+KISSME [WACV17]	**92.5**	**99.3**	**100**	**100**
Conv4	96%			

Performance of unsupervised learning

Recognition rate at different ranks in % when using **MCM (PAM)** representation under different modes of learning

Method	PRID 2011				iLIDS-VID				iLIDS-AA			
	r=1	*r=5*	*r=10*	*r=20*	*r=1*	*r=5*	*r=10*	*r=20*	*r=1*	*r=5*	*r=10*	*r=20*
MCM + MPD	53.6	83.1	91.0	96.9	34.3	61.5	74.4	83.3	56.5	79.7	90.9	95.2
MCM + KISSME	64.3	86.1	94.5	98.0	40.3	69.9	79.0	87.5	62.9	84.7	93.4	97.0
MCM + UnKISSME	59.2	81.7	90.6	96.1	38.2	65.7	75.9	84.1	61.2	85.1	92.8	96.0



MPD distribution of MCM on PRID 2011

Qualitative Results on PRID2011

MCM-MPD

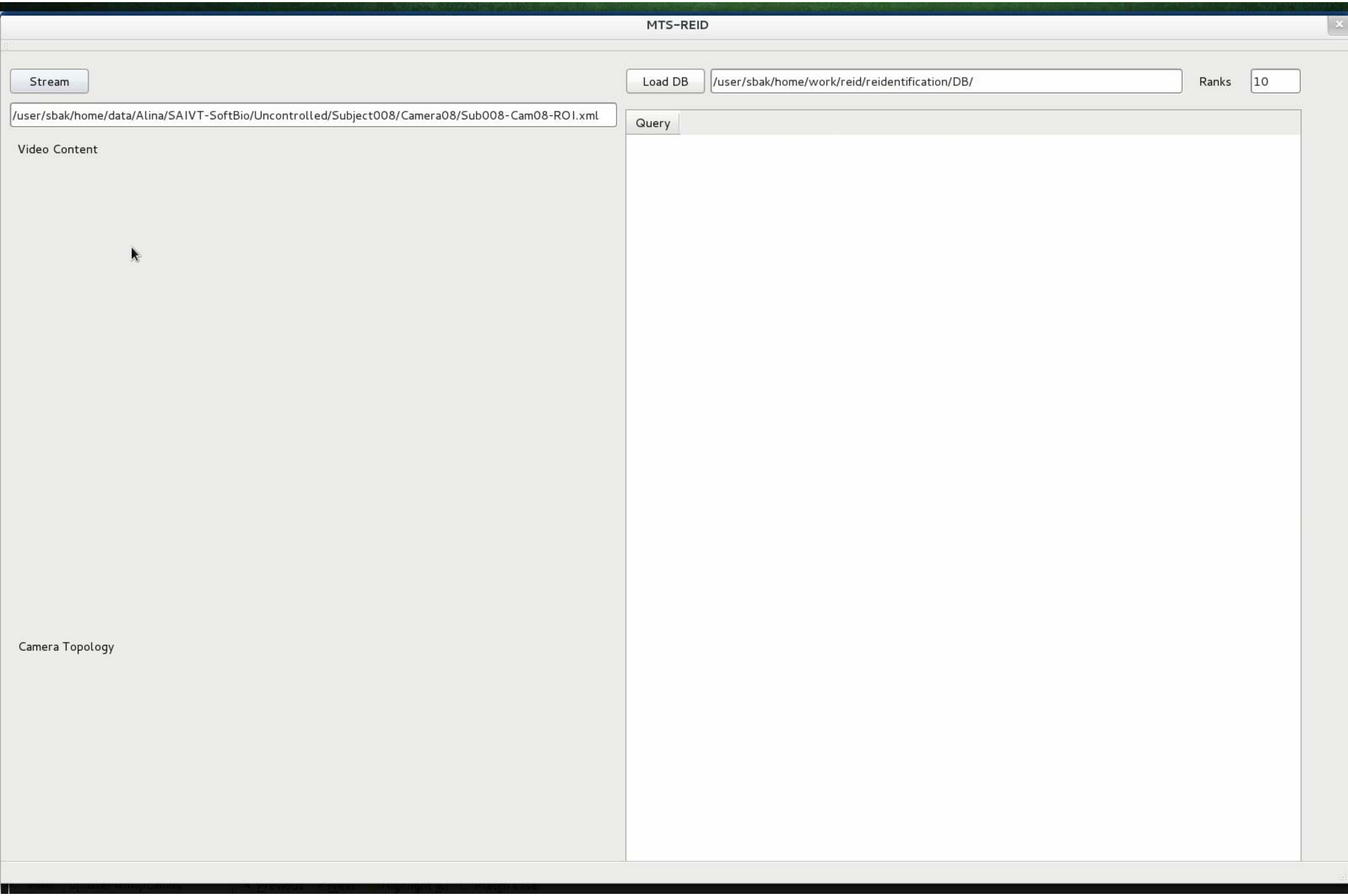
MCM-UnKISSME

Comparison with state-of-the-art ReID

- Comparison with **unsupervised** methods using recognition rate at rank r in % : PRID 2011

Method	r=1	r=5	r=10	r=20
Color+LFDA [Pedagadi13]	43.0	73.1	82.9	90.3
SDALF [Farenzena10]	5.2	20.7	32.0	47.9
Salience [Zhao13]	25.8	43.6	52.6	62.0
FV2D [Ma12]	33.6	64.0	76.3	86.0
FV3D [Liu15]	38.7	71.0	80.6	90.3
DVDL [Karanam15]	40.6	69.7	77.8	85.6
STFV3D [Liu15]	42.1	71.9	84.4	91.6
MCM+UnKISSME[AVSS16]	**[59.2]**	**[81.7]**	**[90.6]**	**[96.1]**

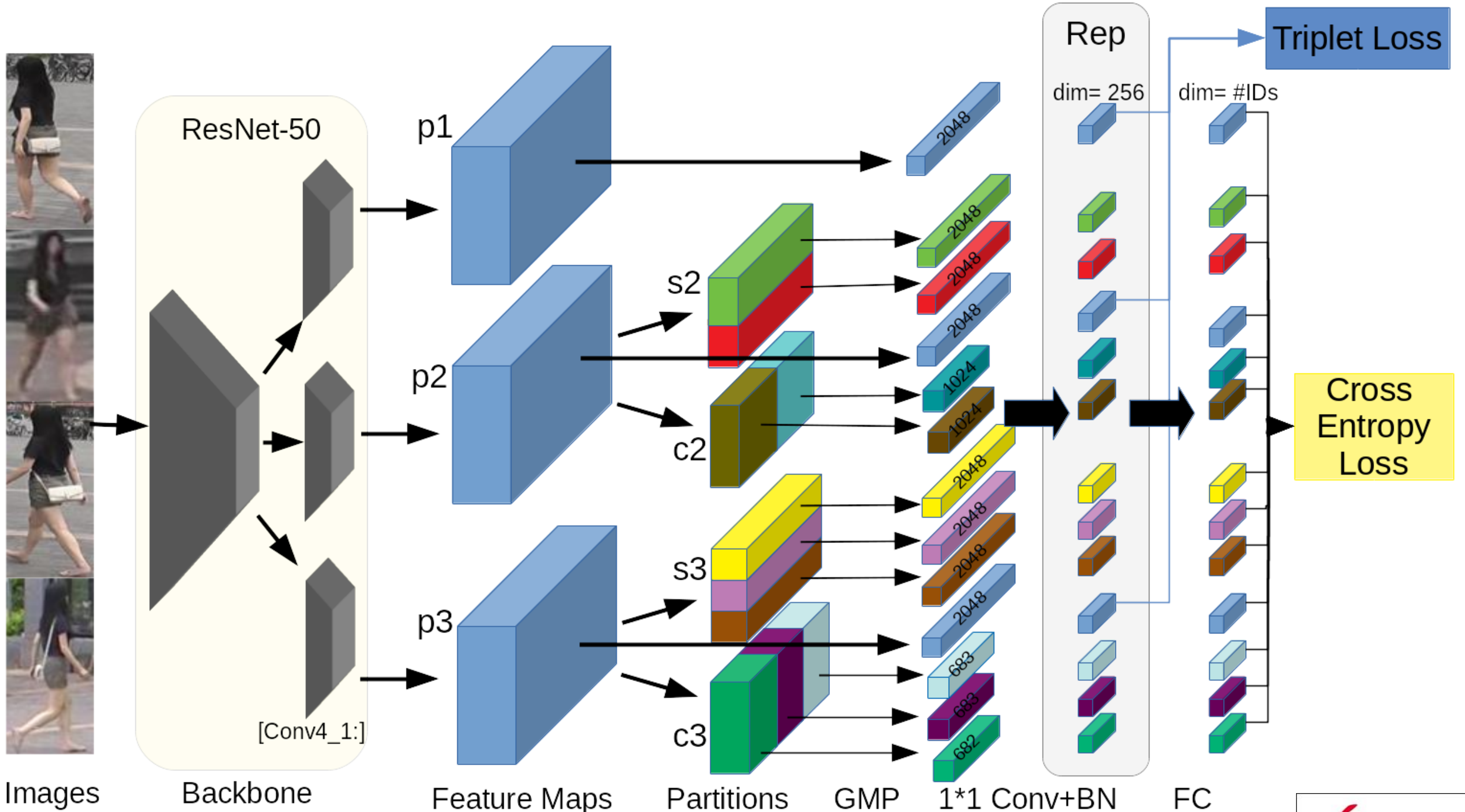
We lose only 5.1%



Queensland University of Technology, Brisbane, Australia
150 humans (400 frames) through up to eight camera views

Spatial and Channel partition CNN Representations (SCR) for Person Re-Identification & Large Dataset (Hao)

General Architecture of SCR:



Spatial and Channel partition Representations (SCR) for Person Re-Identification

Results

Method	Market-1501			
	Single Query		Multiple Query	
	Rank1	mAP	Rank1	mAP
TriNet [10]	84.9	69.1	90.5	76.4
HA-CNN [18]	91.2	75.7	93.8	82.8
GSRW [24]	92.7	82.5	-	-
DNN_CRF [2]	93.5	81.6	-	-
Mancs [27]	93.1	82.3	95.4	87.5
PCB+RPP [26]	93.8	81.6	-	-
SCPNet-a [4]	94.1	81.8	-	-
HPM [7]	94.2	82.7	-	-
MGN [29]	**95.7**	86.9	**96.9**	90.7
CPM [35]	**95.7**	88.2	-	-
SCR(ours)	**95.7**	**89.0**	96.7	**92.2**
SCR(ours)+RR	**96.4**	**94.7**	**97.0**	**96.0**

Table 6. Comparison of results (%) on Market-1501 dataset under Single Query and Multiple Query setting where the bold font denotes the best method. RR stands for Re-Ranking [40].

Method	DukeMTMC-reID	
	Rank1	mAP
HA-CNN [18]	80.5	63.8
GSRW [24]	80.7	66.4
DNN_CRF [2]	84.9	69.5
Mancs [27]	84.9	71.8
PCB+RPP [26]	83.3	69.2
SCPNet-a [4]	84.4	68.5
HPM [7]	86.6	74.3
MGN [29]	88.7	78.4
CPM [35]	89.0	79.0
SCR(ours)	**91.1**	**81.4**
SCR(ours)+RR	**92.9**	**91.1**

Table 7. Comparison of results (%) on DukeMTMC-reID dataset where the bold font denotes the best method. RR stands for Re-Ranking [40].

Method	CUHK03			
	Labelled		Detected	
	Rank1	mAP	Rank1	mAP
HA-CNN [18]	44.4	41.0	41.7	38.6
PCB+RPP [26]	-	-	63.7	57.5
HPM [7]	-	-	63.9	57.5
MGN [29]	68.0	67.4	68.0	66.0
DaRe(R) [30]+RR	72.9	73.7	69.8	71.2
CPM [35]	78.9	76.9	78.9	74.8
SCR(ours)	**83.8**	**80.4**	**82.2**	**77.6**
SCR(ours)+RR	**88.6**	**89.4**	**88.3**	**88.5**

Table 8. Comparison of results (%) on CUHK03 dataset using the new protocol [40] where the bold font denotes the best method. RR stands for Re-Ranking [40].

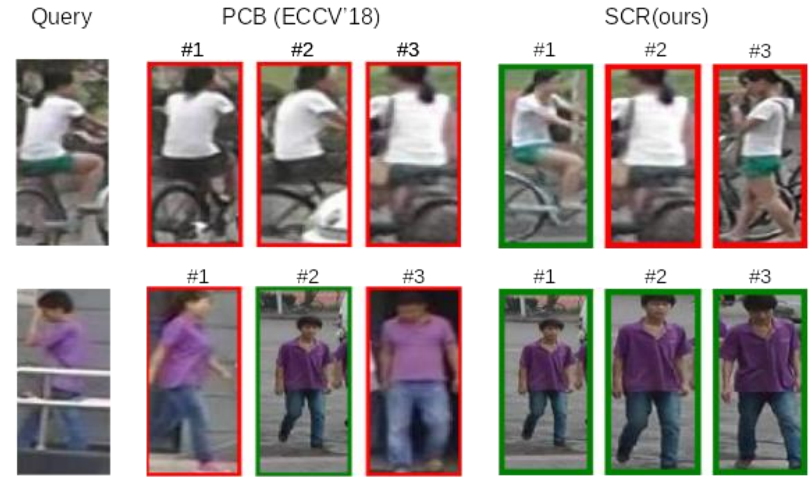
Method	MARS	
	Rank1	mAP
IDE+Kissme [36]	68.3	49.3
TriNet [10]	79.8	67.7
DRSTA [16]	82.3	65.8
M3D [15]	84.4	74.0
SCR(ours)	**87.3**	**81.3**
SCR(ours)+RR	**88.1**	**87.4**

Table 9. Comparison of results (%) on MARS dataset. RR stands for Re-Ranking [40].

Spatial and Channel partition Representations (SCR) for Person Re-Identification

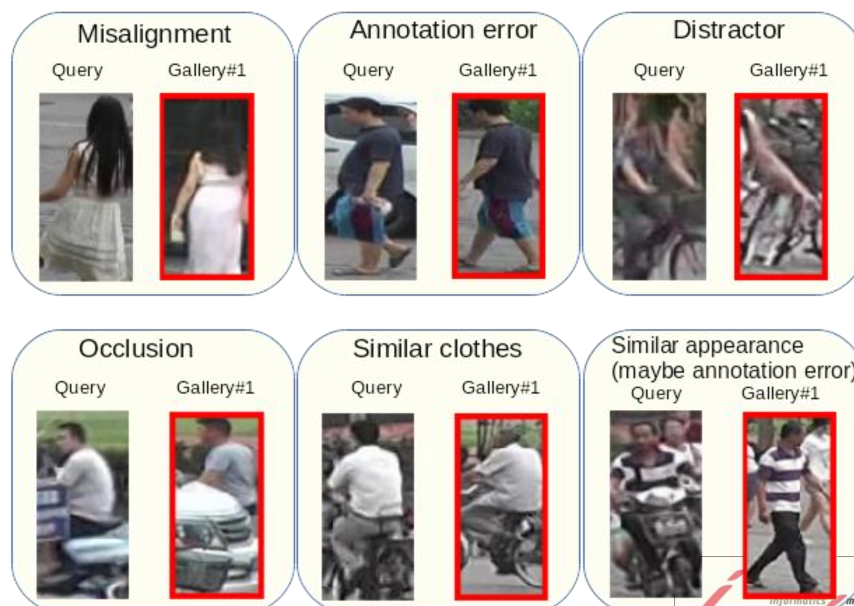
Examples on Market-1501

- Success cases



High accuracy, but SCR requires **large** amount of labeled training data

- Failure cases



Cross domain Residual Transfer Learning for Person Re-identification

Objective

Build Person Re-Identification (Re-ID) models using CNN with small amount of labeled training data

Cross domain Residual Transfer Learning for Person Re-identification

Motivation

- Labeling data for person Re-ID is a recurring and onerous process
- Deep learning struggles with low-amount of training data
- Models do not generalize well across Re-ID datasets
- Hand crafted features perform better on smaller datasets

Cross domain Residual Transfer Learning for Person Re-identification

Residual Learning framework to transfer knowledge from one domain to another

Objective: Minimize residue in network's optimal and current performance

Fine-tuning - Learned parameters are modified

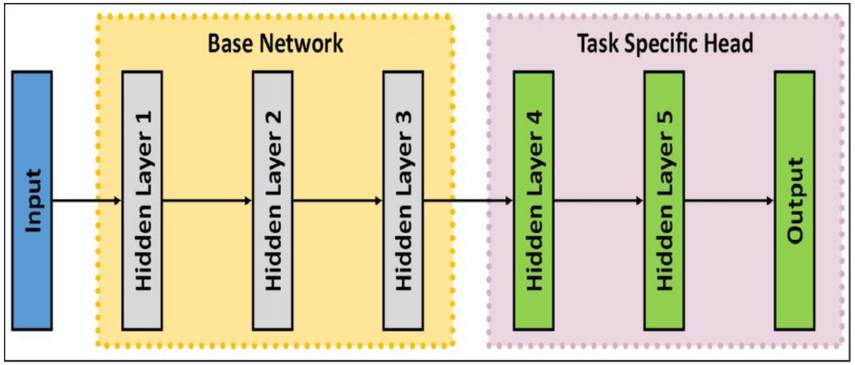
Network, except "head" is fixed

RTL - Add bottleneck layers and modify new parameters

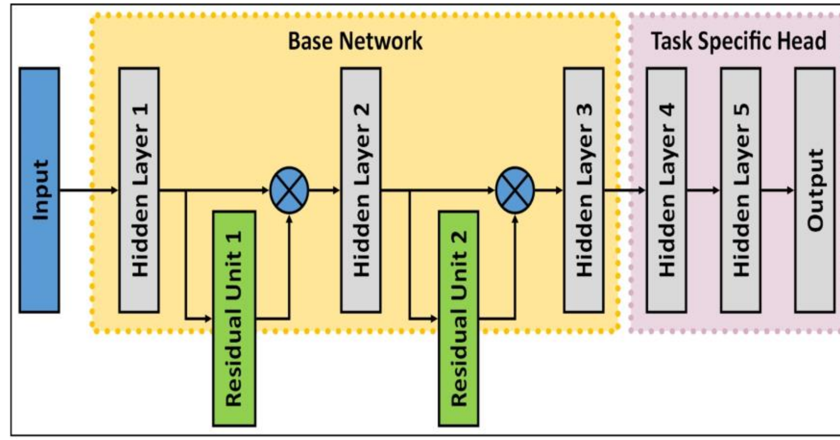
More flexible: bottlenecks may have different quantity and architecture from input layers

Residual Transfer Learning (RTL)

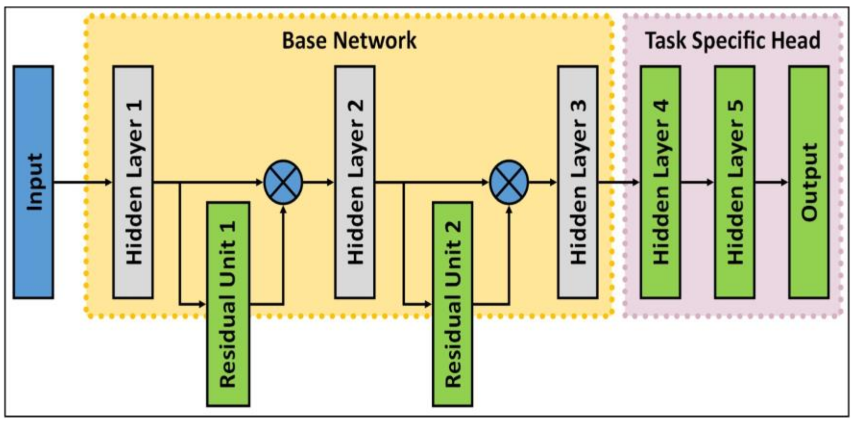
Stage 1



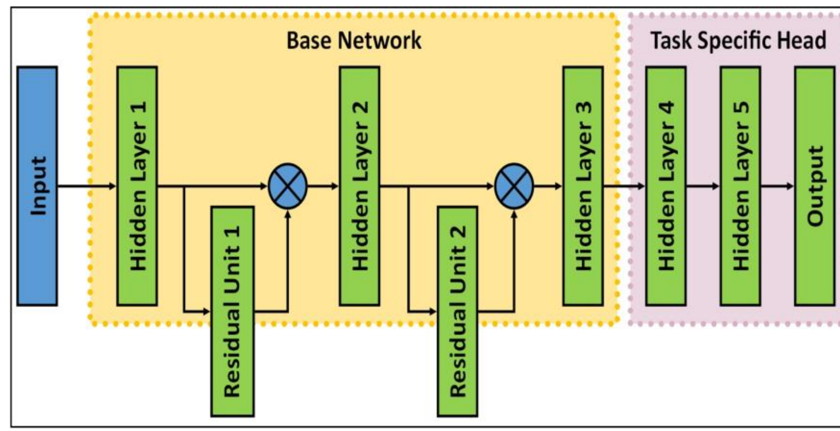
Stage 2



Stage 3



Stage 4 (Optional)



4 Stage Learning Process

Cross domain Residual Transfer Learning for Person Re-identification

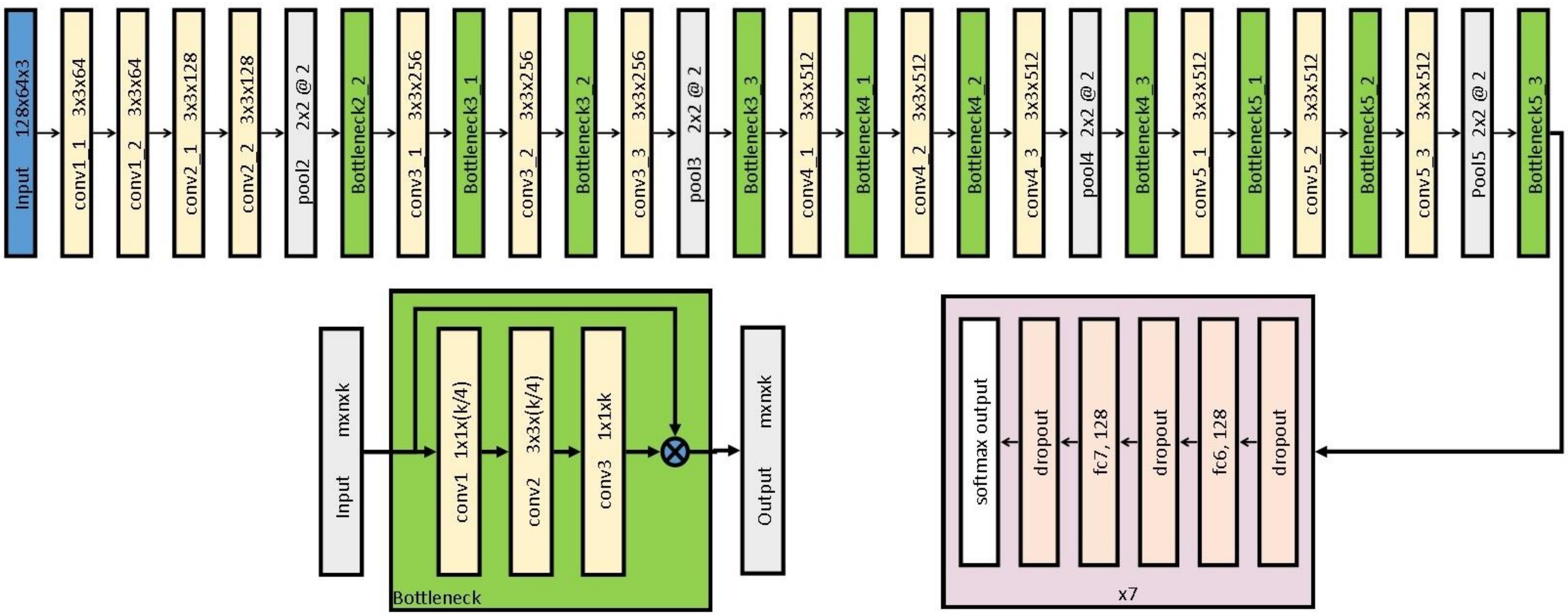
Experimentation : RTL for Person Re-ID

Base Network: **VGG16**

7x Task specific heads, 1 for each local region

Train the network for Identity Discriminative Embedding (IDE)

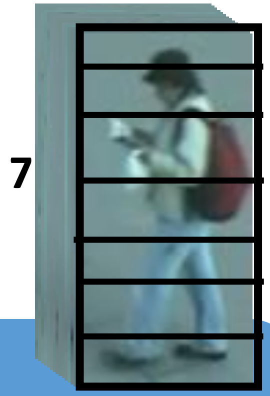
RTL for Person Re-Identification



RTL for Person Re-Identification

Hybrid Modeling for Person Re-ID		
	Metric Learning	Learn embedding space that increases intra-class similarity and reduces inter-class similarity for input features – KISS, XQDA, etc
	Deep Learning	Task Specific Head imitates learned metric, i.e., for IDE, it embeds input features into a class discriminative space
	Hybrid	Train for IDE, then discard Task Specific Head and use XQDA

RTL for Person Re-Identification



Appearance description

8 descriptors per image

7 regions + 1 whole

For multi-shot Re-ID, aggregate descriptors by mean/max pooling

RTL for Person Re-Identification

Experiments

Datasets:

iLIDS-VID, PRID and MARS

Models

B7 ≡ RTL-fc7 + Eucl

B5 ≡ RTL-pool5 + Eucl

B4 ≡ RTL-pool4 + Eucl

H7 ≡ RTL-fc7 + XQDA

H5 ≡ RTL-pool5 + XQDA

H4 ≡ RTL-pool4 + XQDA

RTL for Person Re-Identification

Results

iLIDS-VID						
Stage	B7	B5	B4	H7	H5	H4
1	34	18	32	46	42	69
2	48	41	42	55	56	76
3	53	50	45	57	58	77
4	54	51	46	56	58	79

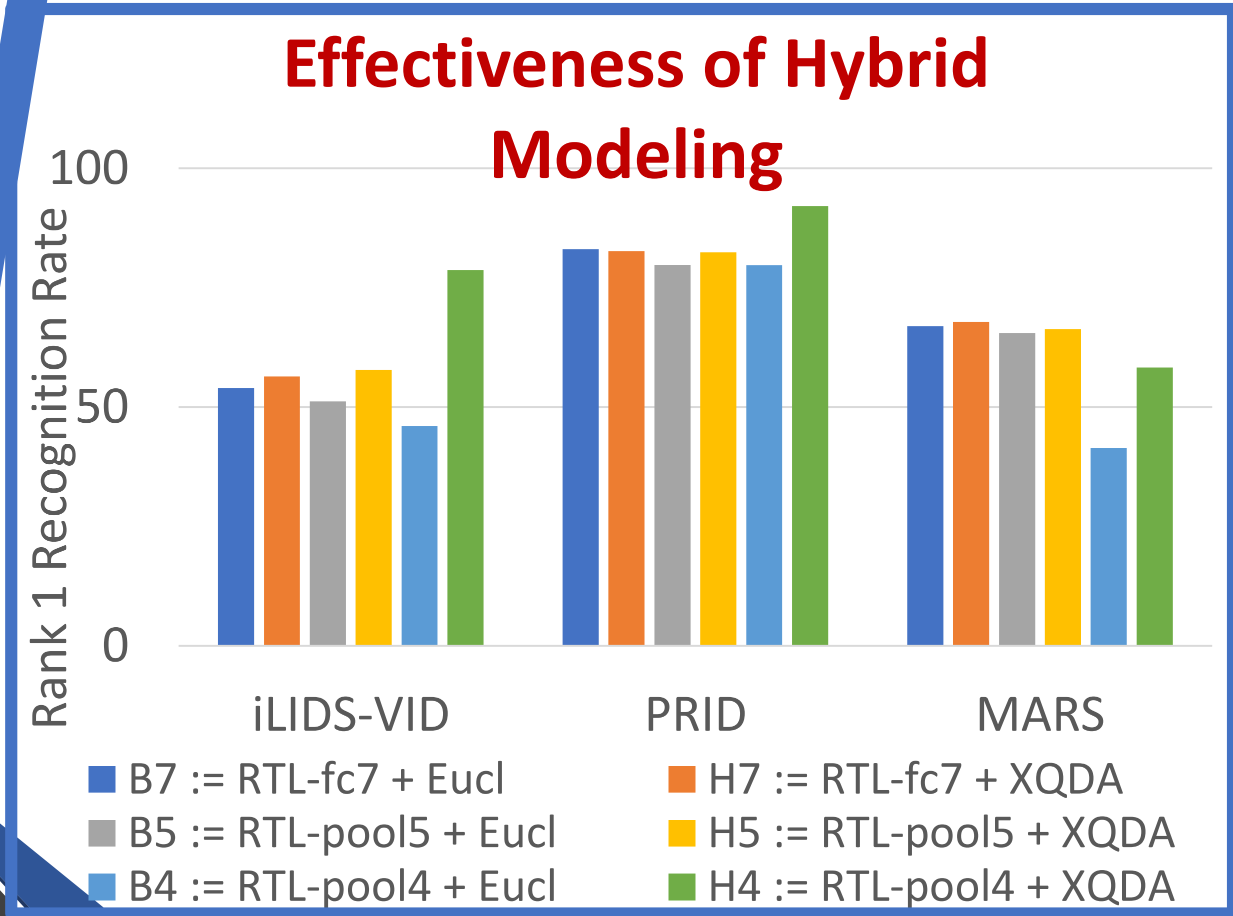
PRID						
Stage	B7	B5	B4	H7	H5	H4
1	75	55	63	74	62	82
2	83	73	75	83	79	91
3	85	77	77	83	82	92
4	83	80	79	83	82	92

MARS						
Stage	B7	B5	B4	H7	H5	H4
1	35	21	19	42	27	30
2	54	49	32	57	56	49
3	66	64	39	66	65	55
4	67	65	41	68	66	58

Rank 1 Recognition Rate after each stage of RTL

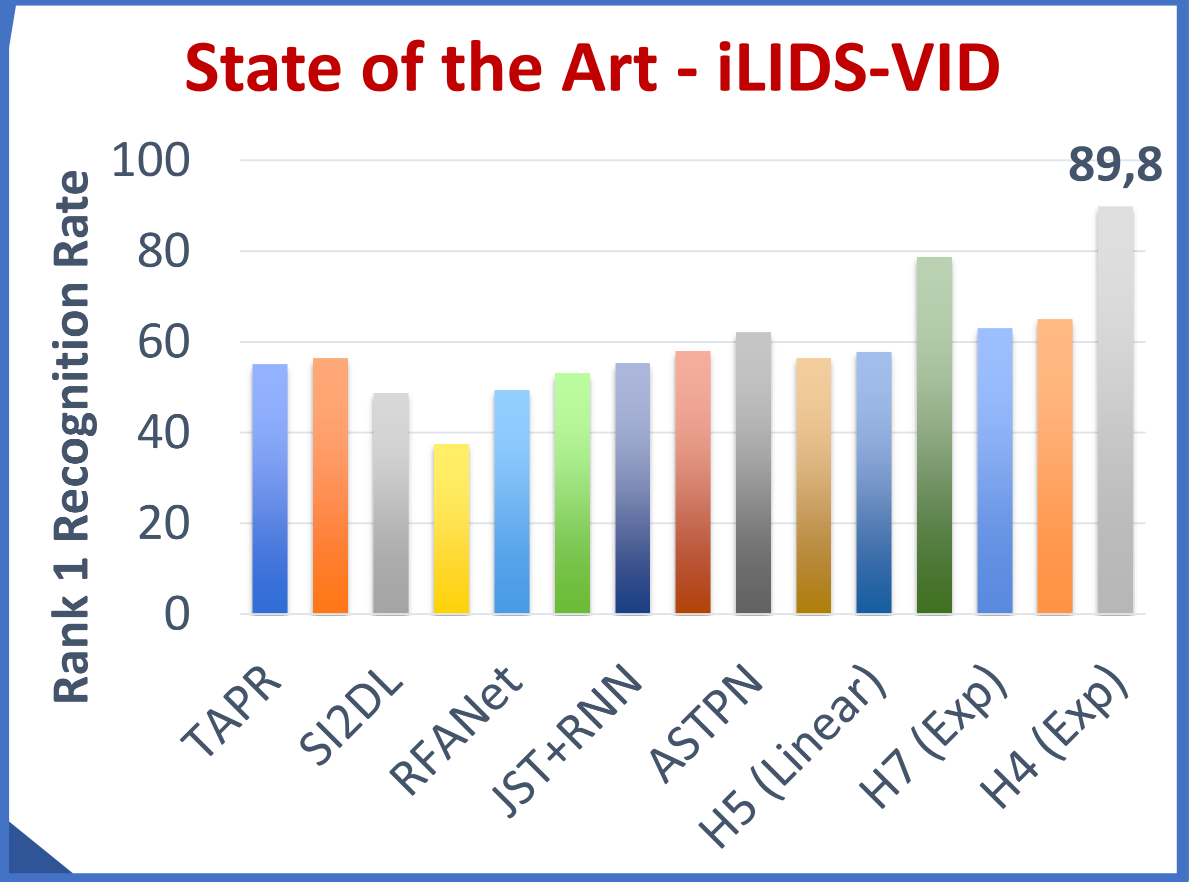
RTL for Person Re-Identification

Results



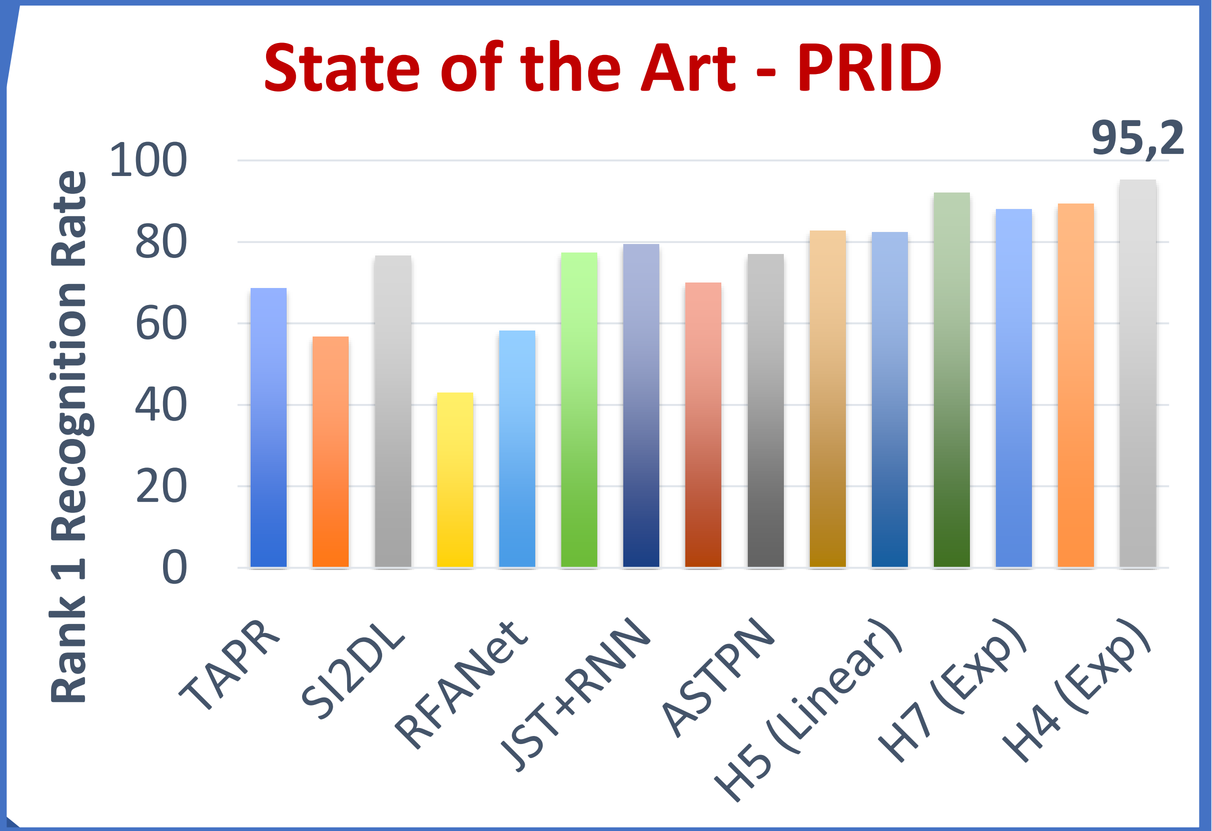
RTL for Person Re-Identification

Results



RTL for Person Re-Identification

Results



Conclusion – Person Re-Identification

Small Gallery dataset : 100 – 200 ID

- **No annotation** : using handcrafted/CNN features + (un)supervised learning
- **Very few annotation (20% annotation)** : using handcrafted/CNN features + Metric Learning
- **Few annotation (50% annotation)**: CNN features using RTL Learning + Metric Learning

Big Gallery dataset (with annotation) : 1000 – 5000 ID

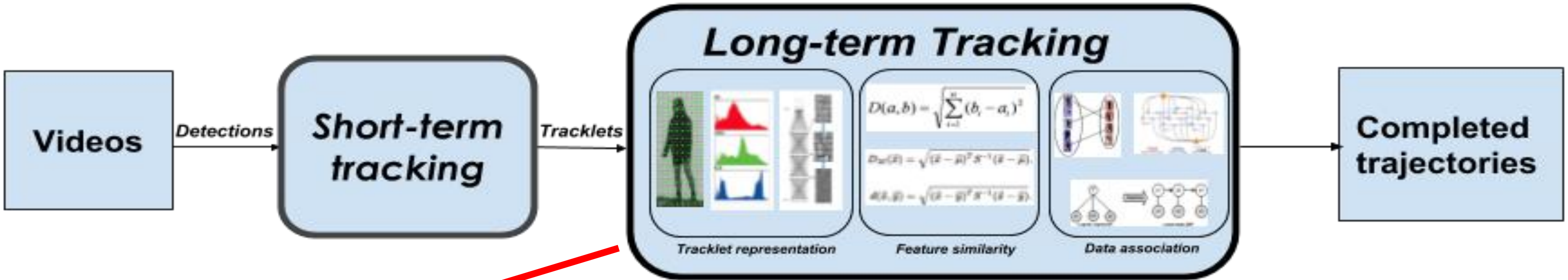
- CNN features trained
  - On Triplet – loss and ID – loss
  - With Partitions on width/length (spatial) + channel + temporal

Perspectives for cross-dataset ReID:

- Disentangling pose from appearance using GAN
- Signature based on semantic attributes

Thank you!

People Tracking: Long Term Tracking



1/ Feature extraction
- Handcrafted features
- Deep features

2/ Feature similarity computation
- Euclidean distance
- Bhattacharyya distance
- Kullback-leibler distance
- Online Re-id metric learning (KISSME)

3/ Data association
- Local data association
- Global data association
- Hybrid data association

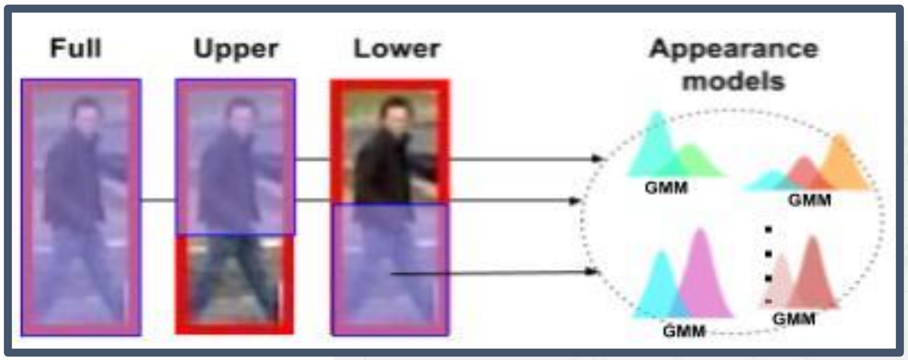
Feature extraction and selection

- Online tuning the feature weights to increase discriminative power (no training) [AVSS2016]
- Online retrieving optimal tracking parameters [AVSS2017]
- Extending powerful features for Re-Id to MOT (Handcrafted and CNN features) [AVSS2017-18]
  1. RBT(HF) = Re-id Based Tracker (Handcrafted Features)
  2. RBT(RTL) = Re-id Based Tracker (Residual Learning Transfer)

People Tracking: RBT
Re-id Based Tracker [AVSS 2017]

Objectives:

- Show that features (handcrafted and learned features) which are powerful in Re-ID domain are effective in MOT domain
- Extend the metric learning proposed for offline Re-ID to online MOT



KISSME (*Keep It Simple and straightforward Metric*)

- Simplicity
- Low computation cost
- effective under challenging conditions
- do not need a large number of training data (only two hundreds of pairs)

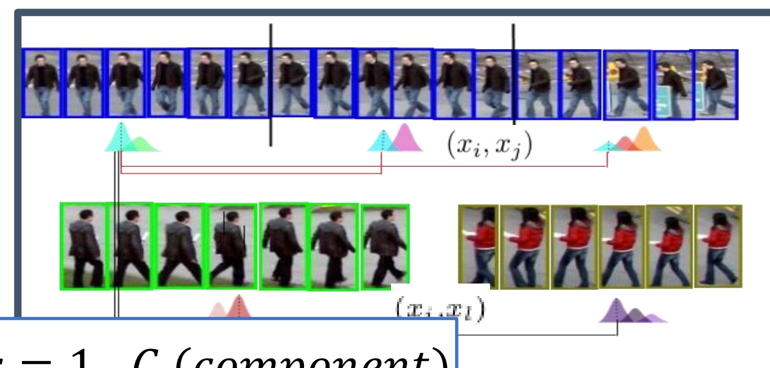
- HOG
- Color Histogram
- LOMO
- MCSH
- CNN

Tracklet representation

$$\boxed{\nabla_{Tr_i} = \{M_i^{p,f} \,|\, p \in \mathbb{P}, f \in \mathbb{F}\}}$$

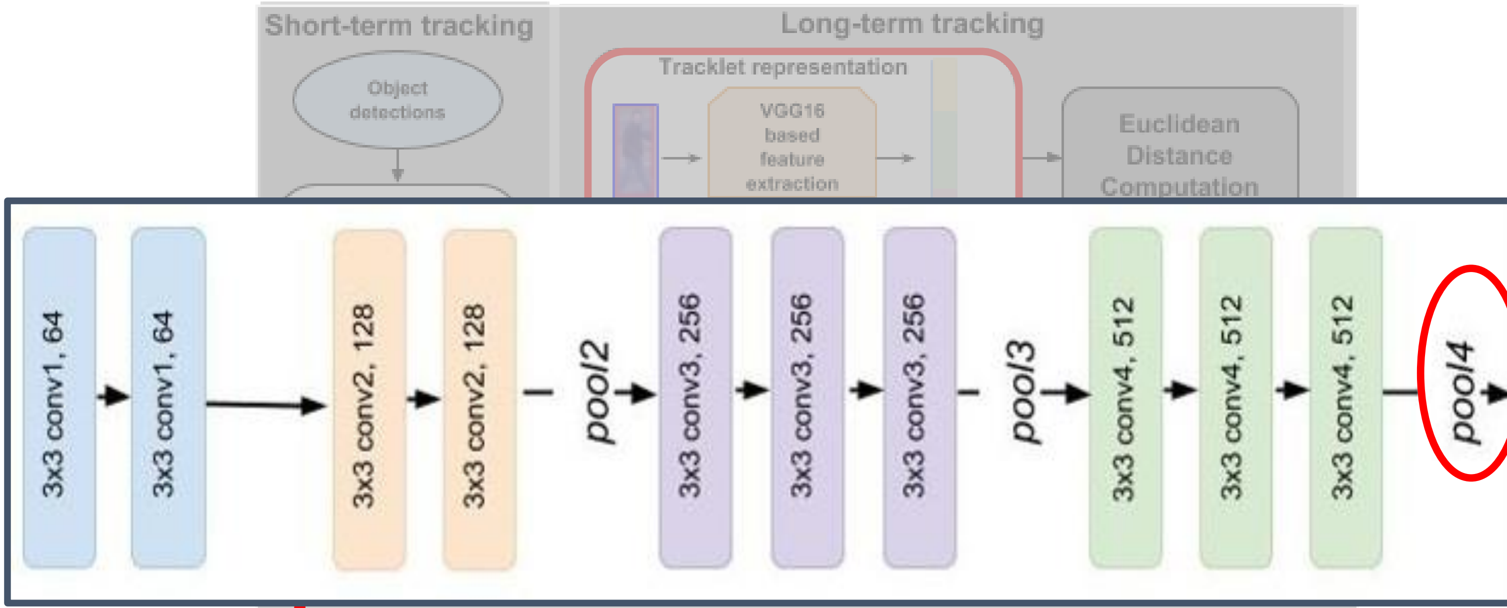
Appearance model

$$M_i^{p,f}(GMM) = \left\{\left(\mu_{i,c}^{p,f}, \sigma_{i,c}^{p,f}\right)^c\right\} \; c = 1..C \; (component)$$



Metric learning sampling

People Tracking: RBT
Learned features (VGG16)



Feature vectors

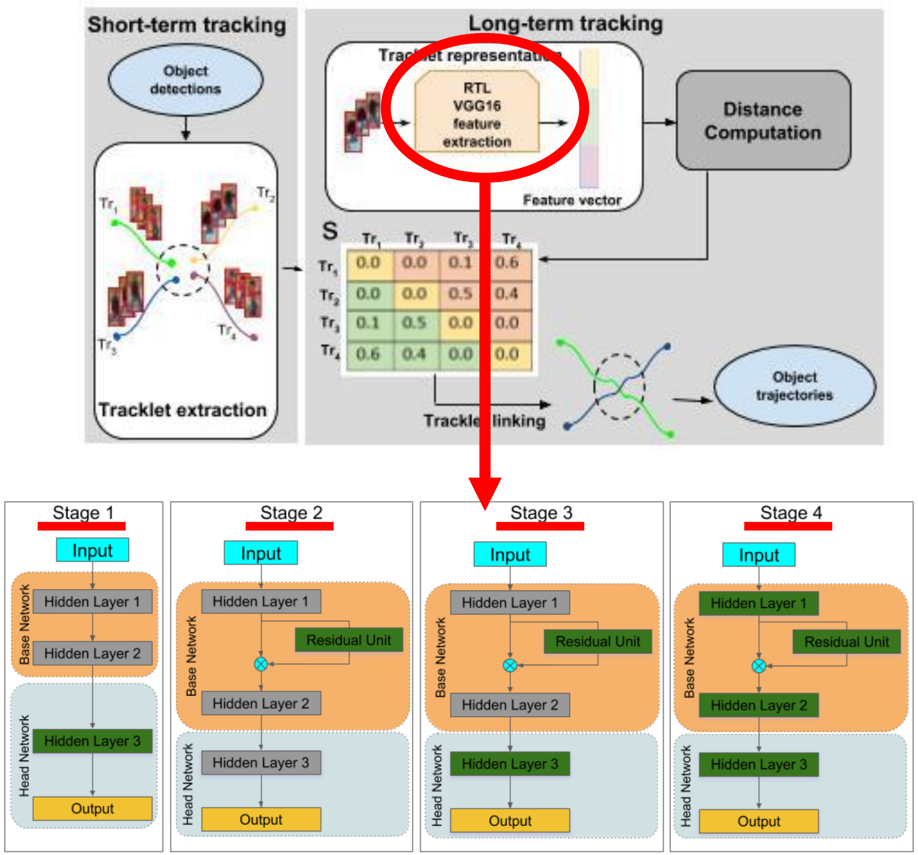
The pretrained-VGG16 feature extractor

CBT(CNN) - Framework

Classification

Tracking

People Tracking: RBT
Residual Transfer Learning [AVSS 2018]



Residual Transfer Learning (4 step training)

*Learning part is marked by **green***

Stage 1:

- learns the high level representations
- trains only the new head (initialize it randomly) and keeps the network's base unchanged.

Stage 2:

- Learns low level representations
- adds the residual units between the convolutional layers and initialize them randomly.
- fixes the base and head of the network and train only the residual units.

Stage 3:

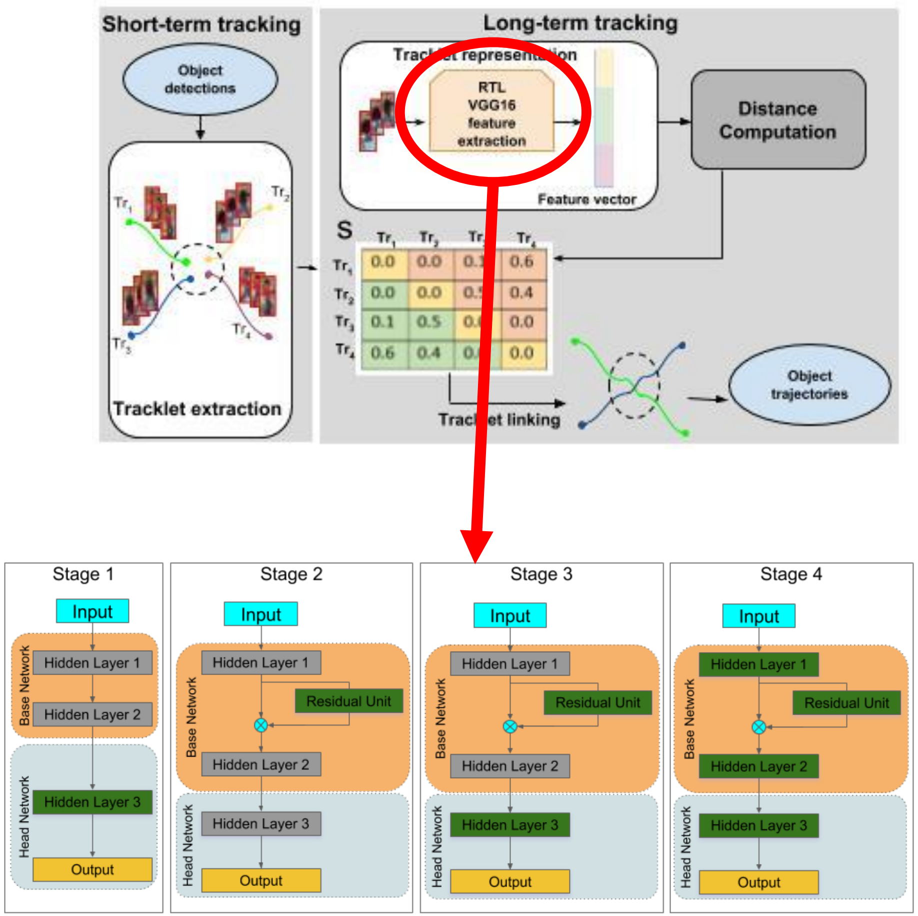
- trains the head and the residual units conjointly.
- The value of loss function is low enough

Stage 4 (optional):

- Further improvement performance can be achieved by training the whole network.

People Tracking: RBT

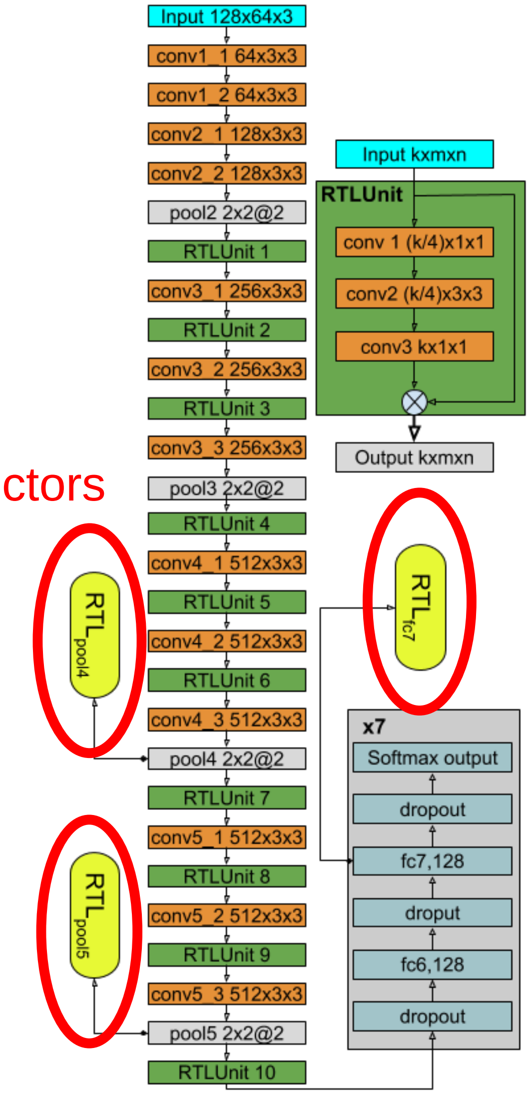
Residual Transfer Learning



Residual Transfer Learning (4 step training)

Learning part is marked by ***green***

Feature vectors



Network architecture

People Tracking: Experiments – MOT Metrics

Metric	Description	Note
MT (%)	Mostly tracked (> 80% of GT trajectory is tracked)	↑
ML (%)	Mostly lost (< 20% of GT trajectory is tracked)	↓
MOTA (%)	Multiple Object Tracking Accuracy	↑
MOTP (%)	Multiple Object Tracking Precision	↑
FP (#)	The total number of false positives	↓
FN (#)	The total number of false negatives	↓
IDSw (#)	The total number of identify switches	↓
Frag (#)	The total number of times a trajectory is fragmented	↓

$$MOTA = 1 - \frac{\sum_t (fn_t + fp_t + IDSw_t)}{\sum_t gt}$$

MT: evaluates in term of object trajectory
MOTA: punishes more on detection error

$\boldsymbol{fn_t}$: false negatives, $\boldsymbol{fp_t}$: false positives, $\boldsymbol{IDSw_t}$: ID Switches

People Tracking
Experiments: State-of-the-art Comparison

MOT15

- ✓ 22 challenging video sequences with only one provided detection
- ✓ 11 training and 11 testing sequences
- ✓ A diversity of outdoor scenarios:
  - strong and frequent person-person occlusions
  - crowded environment
  - captured by fixed or moving camera
  - low illumination

Training sequences

Testing sequences

SoA Tracking performances on MOT 2015

https://motchallenge.net

Trackers	Methods	MT(%)	ML(%)	MOTA(%)	MOTP (%)	FP (#)	FN (#)	IDSw (#)	Frag (#)
CNNTCM (CVPR-2016)	Offline	**11.2±13.0**	44.0	**29.6±13.9**	**71.8**	7,786	34,733	**712**	**943**
CEM (TPAMI-2014)		8.5±8.08	46.5	19.3±17.5	70.7	14,180	**34,591**	813	1,023
SiameseCNN (CVPR-2016)		8.5±20.3	48.4	29.0±15.1	71.2	**5,160**	37,798	639	1,316
ELP (WACV-2015)		7.5±6.3	**43.8**	25.0±10.8	71.2	7,345	37,344	1,369	1,804
TBD (PAMI-2014)		6.4±13.4	47.9	15.9±17.6	70.9	14,943	34,777	1,939	1,963
Moticon(CVPR-2014)		4.7±8.6	52.0	23.1±16.4	70.9	10,404	35,844	1,018	1,061
RBT(HC) Ours	**Online**	**9.0±17.4**	**36.9**	20.6±18.7	70.3	15.161	**32,212**	1,387	2,375
SCEA (CVPR-2016)		8.9±6.6	47.3	**29.1±12.2**	71.1	**6,060**	36,912	604	1,182
OMT_DFH (OSA journal-2017)		7.1±11.3	46.5	21.2±17.2	69.9	13,218	34,657	563	1,255
RNN_LSTM (AAAI-2017)		5.5±9.9	45.6	19.0±15.2	71.0	11,578	36,706	1,490	2,081
EAMTTpub (ECCV-2016)		5.4±7.5	52.7	22.3±14.2	**70.8**	7,924	38,982	833	1,485
RMOT (WACV-2015)		5.3±9.8	53.3	18.6±17.5	69.6	12,473	36,835	684	1,282
TC_ODAL (CVPR-2014)		3.2±7.9	55.8	15.1±15.0	70.5	12,790	38,538	637	1,716
GSCR (ICIP-2015)		1.8±2.14	61.0	15.8±10.5	69.4	7,597	43,633	**514**	**1,010**

The best performances are marked in **bold**

People Tracking : Long TermTracking

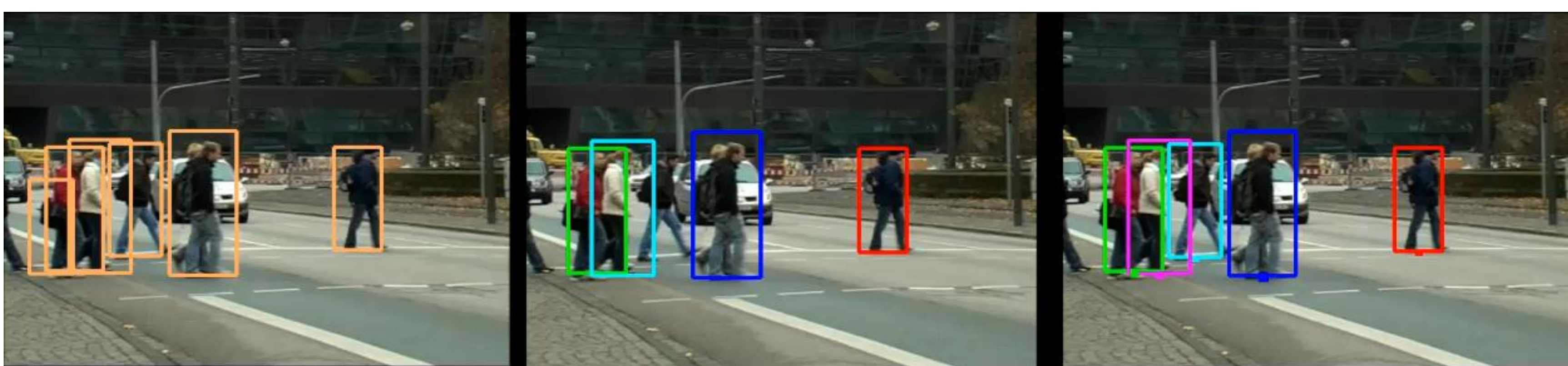
Multiple Object Tracking (MOT15) challenge:

Our online tracker : RBT(HF) = Re-id Based Tracker (Handcrafted Features) has the best performance [AVSS17] for Mostly Tracked (MT) metric

Sequences	Trackers	Methods	MT ↑	ML ↓	MOTA ↑	MOTP ↓	FP ↓	FN ↓	IDSw ↓	Frag ↓
TUD Crossing	CNNTCM	Offline	46.2	23.1	60.5	73.7	66	352	17	14
	RBT(HF)-Ours	Online	61.5	7.7	72.1	73.0	55	230	22	43
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TUD-Crossing

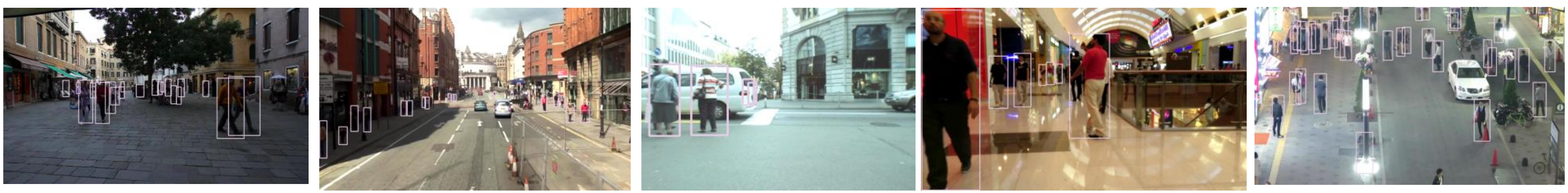
Detection **CNNTCM** **RBT(HC)-Ours**



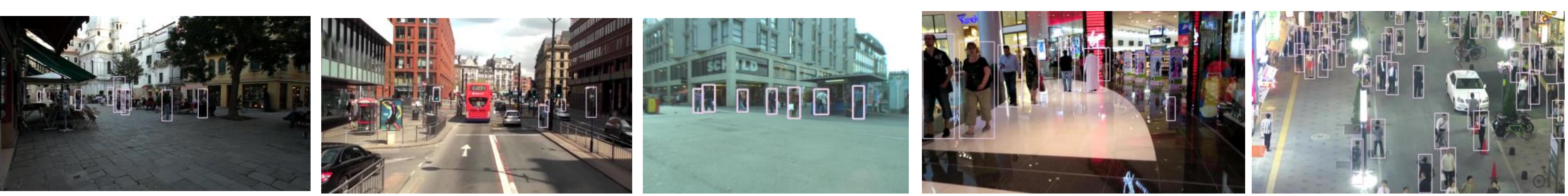
People Tracking
Experiments: State-of-the-art Comparison

MOT17

- ✓ 14 challenging video sequences with 3 detections are provided: DPM, SDP, FRCNN
- ✓ 21 sequences for training and 21 sequences for testing
- ✓ A diversity of outdoor scenarios:
  - Strong and frequent occlusions
  - Low illumination
  - Fixed and moving camera
  - High object density



Training sequences



Testing sequences