从行人再识别技术说起
如何实现计算机视觉技术的原创突破

袁余锋/算法研究员

云从科技 CLOUDWALK

目录

- 云从介绍
- 行人再识别(ReID)的发展历史与困难
- 云从对ReID的探索与成果
- ReID应用场景与技术展望

云从科技

人工智能行业领导者

云从科技发展历程—— “十年磨一剑”

2006前

计算机视觉之父

Thomas S. Huang教授，

四院院士，20多年技术积累

2006-2011

周曦博士带领UIUC团队六次斩获世界模式识别、智能识别大赛冠军。

2011-2014

中科院人脸识别团队建立并发展，

专注科技技术产品化，

多项人脸识别应用级产品在此诞生

2015后

2015年4月云从科技成立，快速发展，

半年内推出多行业应用级产品及解决方案，

打造自主品牌

About us .

产业化国家队

——中科院人脸识别技术唯一代表

中科院旗下人工智能企业

- 中科院战略先导重大专项人脸识别团队唯一代表
- 白春礼院长称科大讯飞和**云从科技**是中国人工智能行业的标杆性企业

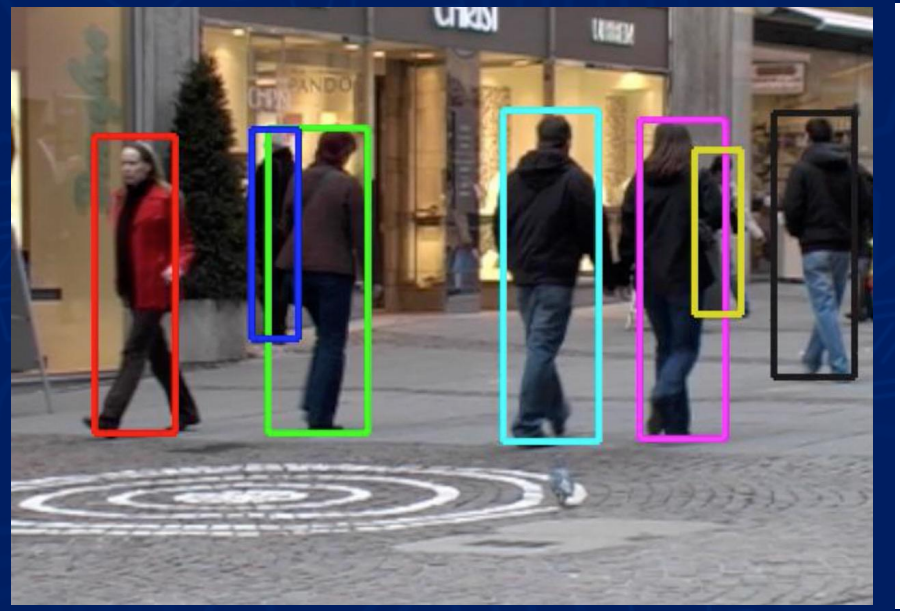
人脸识别标准制定者

唯一 一家同时受邀制定人脸识别**国家标准、公安部行业标准**的企业

About us

ReID的定义

行人智能认知 --人脸识别之后重要的研究方向

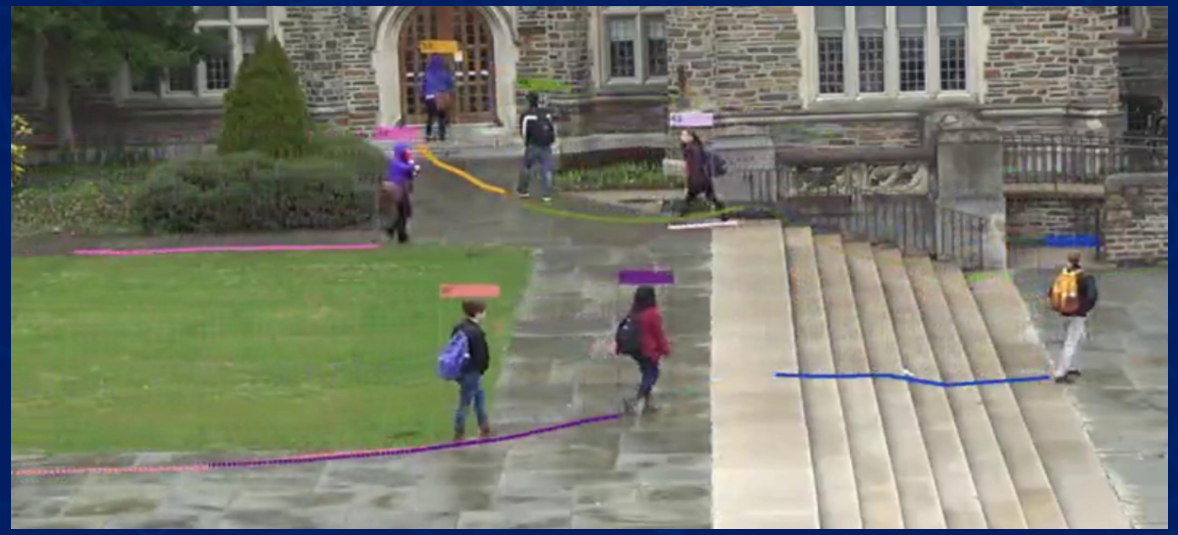


行人检测

行人分割&背景替换

骨架关键点检测
&姿态识别

行人智能认知 --人脸识别之后重要的研究方向



行人跟踪/MOT

动作识别

例如：闯红灯、突发事件、偷窃识别等

行人智能认知 --人脸识别之后重要的研究方向

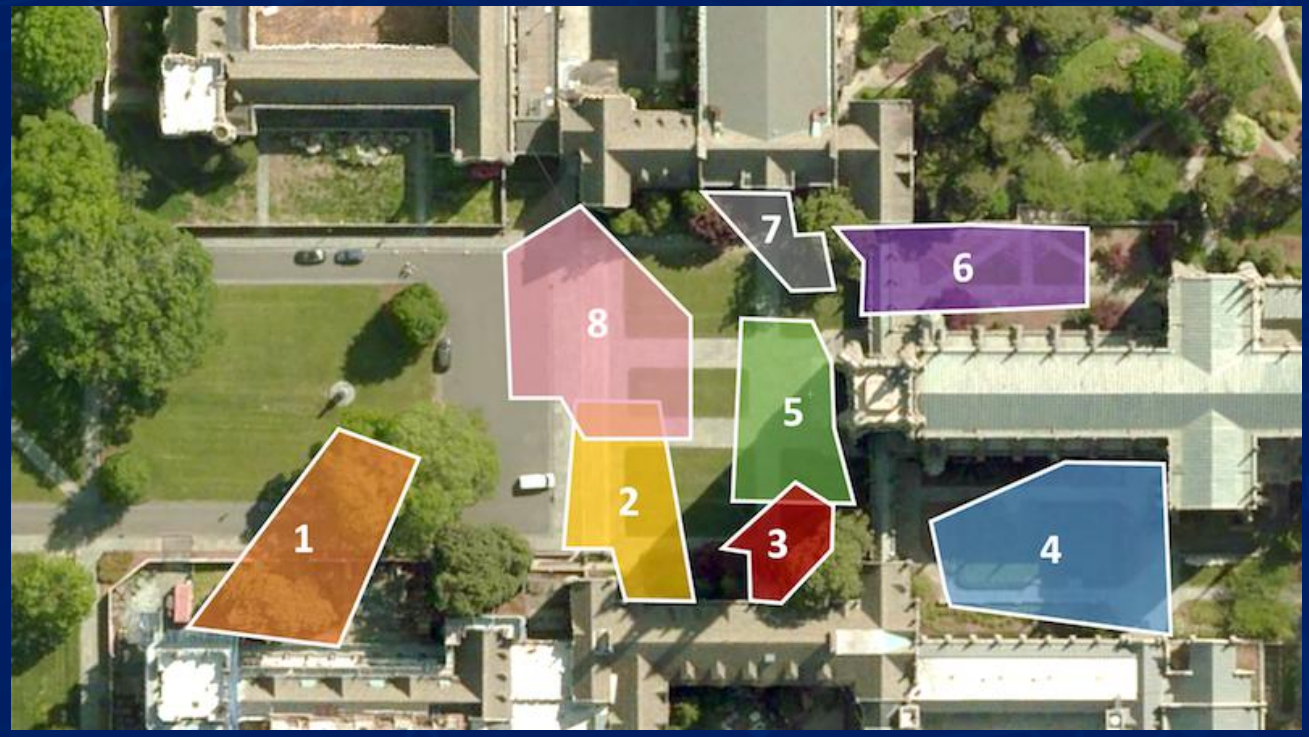
Attribute	Label
gender	2
top	1
backpack	2
bag	1
handbag	1
boots	2
shoes	1
handbag	1
upblack	2
downblue	2

行人属性结构化

跨镜追踪&行人再识别(ReID)

ReID定义

跨镜追踪（Person Re-Identification，简称 ReID）技术是现在计算机视觉研究的热门方向，主要解决跨摄像头跨场景下行人的识别与检索。该技术可以作为人脸识别技术的重要补充，可以对无法获取清晰拍摄人脸的行人进行跨摄像头连续跟踪，增强数据的时空连续性。



ReID的发展历史与困难

常用数据集

深度学习很简单！

只要有大数据就可以解决问题？

ReID迎刃而解？

常用数据集

在ReID中，也行！但仅仅是理论上的

实际操作上不行！！

WHY?

ReID技术难点

ReID实际应用场景下的数据非常复杂，会受到各种因素的影响，这些因素中是客观存在，ReID要尝试去解决的

无正脸照

姿态

配饰

遮挡

更多因素

- 相机拍摄角度
- 图片模糊不清楚
- 室内室外环境变化
- 行人更换服装饰配
- 冬季夏季风格差异
- 白天晚上光线差异

常用数据集

学术界研究中用于ReID算法训练以及评估的三个主流公开数据集

	Market1501	DukeMTMC-reID	CUHK03
拍摄地点	清华大学	Duke大学	香港中文大学
图片数量	32217	36441	13164
行人数量	1501	1812	1467
摄像头	6	8	10

ReID数据集

- 图片数量在几万张
- ID数量2千
- 摄像头10个以下
- 大多数是学生

VS

人脸数据集

- 上千万张照片
- 上百万ID
- 身份多样

常用数据集

Market1501

DukeMTMC-reID

CUHK03

常用数据集

ReID数据采集特点

- 必须跨摄像头数据
- 数据集规模小
- 影响因素复杂多样
- 连续视频截图
- 同一个人多张全身照片
- 互联网无法提供有效数据
- 大规模搜集涉及到隐私问题

数据获取难度大 → 算法挑战大

评价指标

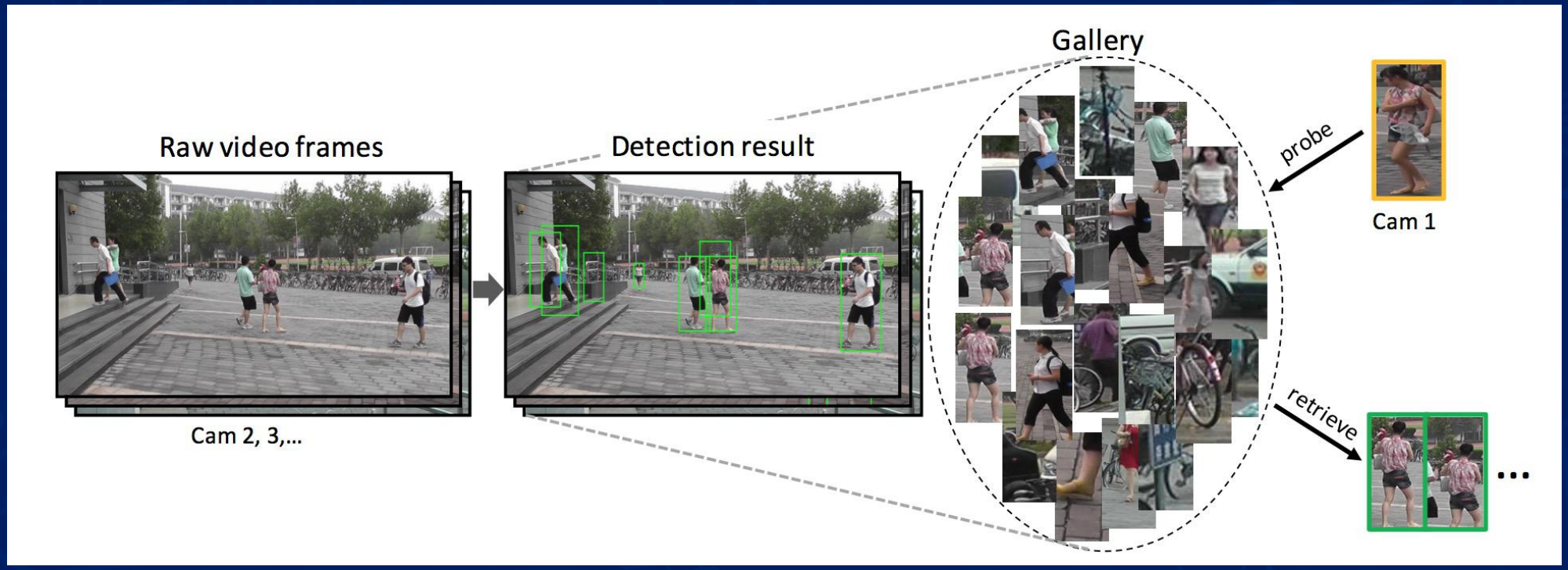
Rank1：首位命中率（Rank-1 Accuracy）

mAP：平均精度均值（Mean Average Precision，mAP）

ReID实现思路

ReID实现思路

基本流程



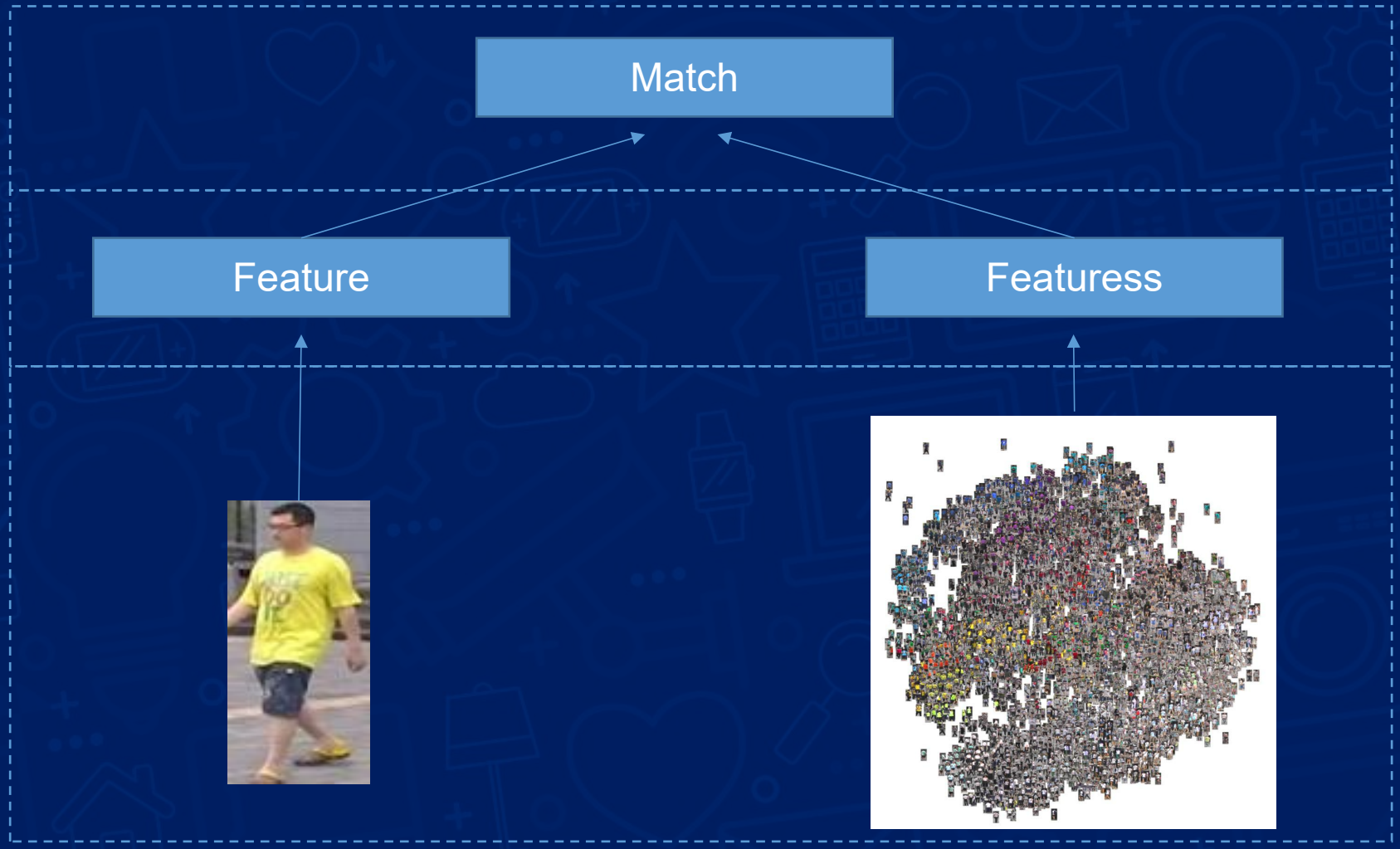
RawFrame → Detect → ReID

常用数据集

ReID实现思路

先抽取特征再进行比对

1. 检索图经过网络抽取图片特征（Feature）
2. 底库里的所有图片全部抽取图片特征（Feature）
3. 将检索图与地库图的特征计算距离（例如欧式距离）
4. 根据计算距离进行排序，排序越靠前表示是相似率越高



ReID实现思路--表征学习

基于SOFTMAXLOSS(分类损失)与 CONTRASTIVELOSS(对比损失)

- RANK1: 79.51%
- MAP: 59.87%

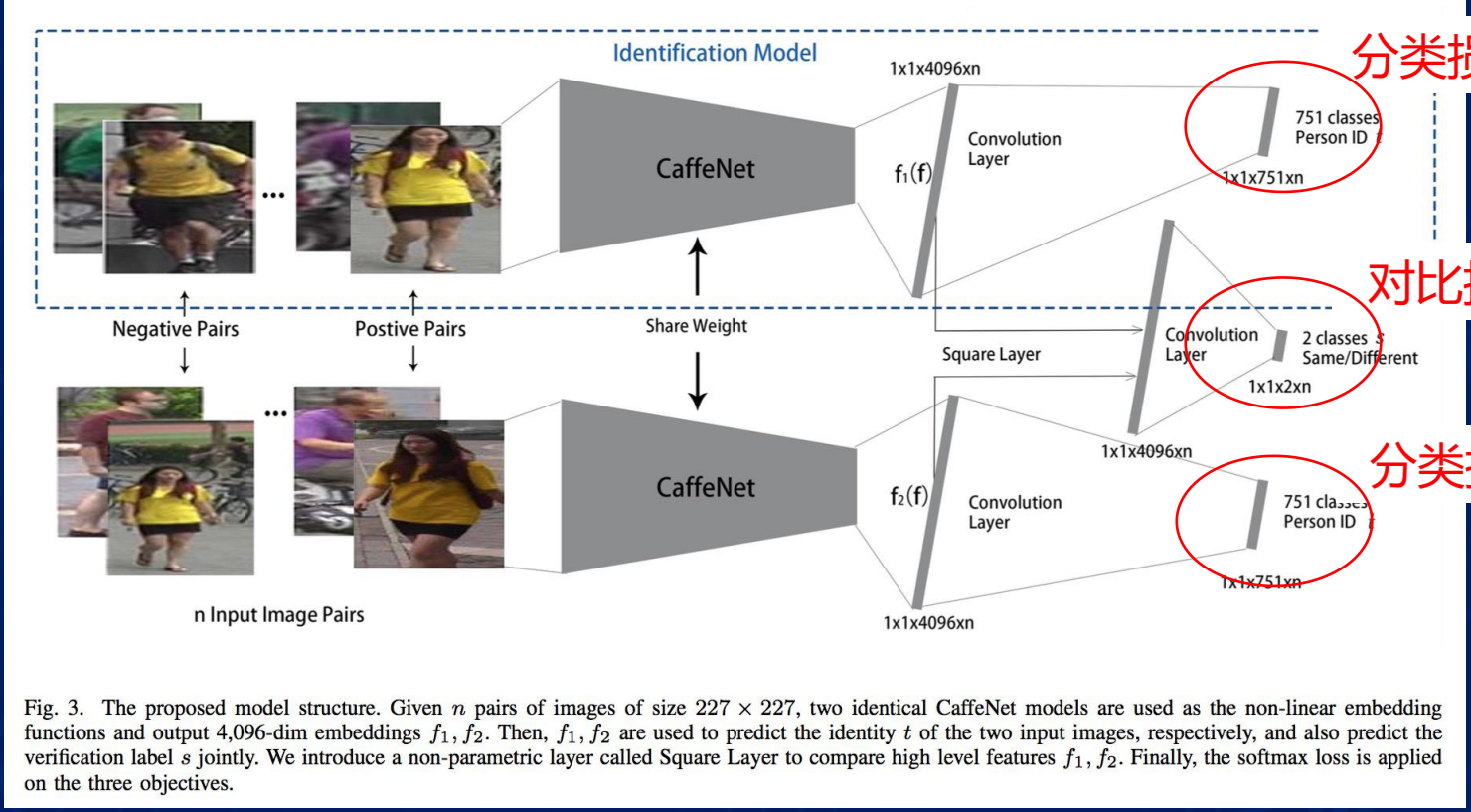


Fig. 3. The proposed model structure. Given n pairs of images of size 227×227, two identical CaffeNet models are used as the non-linear embedding functions and output 4,096-dim embeddings f_1, f_2. Then, f_1, f_2 are used to predict the identity t of the two input images, respectively, and also predict the verification label s jointly. We introduce a non-parametric layer called Square Layer to compare high level features f_1, f_2. Finally, the softmax loss is applied on the three objectives.

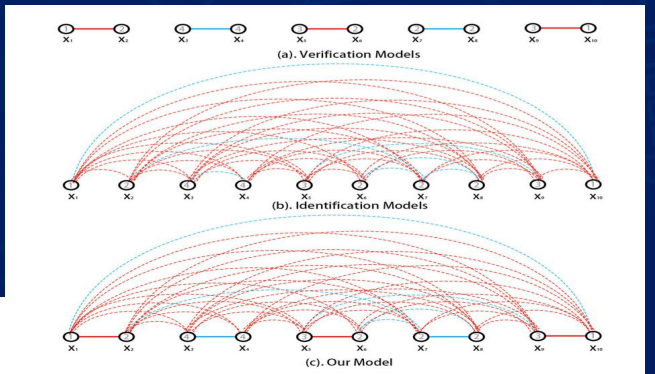


Fig. 2. Illustration for a training batch. The number in the circle is the identity label. Blue and red edges represent whether the image pair depicts the same identity or not. Dotted edges represent implicit relationships and solid edges represent explicit relationships. Our model combine the strengths of the two models.

CaffeNet-Basel. [26]	50.89	26.79
Ours(CaffeNet)	62.14	39.61
VGG16-Basel. [27]	65.02	38.27
Ours(VGG16)	70.16	47.45
ResNet-50-Basel. [28]	73.69	51.48
Ours(ResNet-50)	**79.51**	**59.87**

ReID实现思路—度量学习

基于TRIPLET LOSS（三元损失）的REID

- RANK1: 84.92%
- MAP: 69.14%

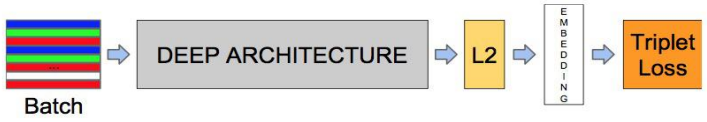


Figure 2. **Model structure.** Our network consists of a batch input layer and a deep CNN followed by L_2 normalization, which results in the face embedding. This is followed by the triplet loss during training.

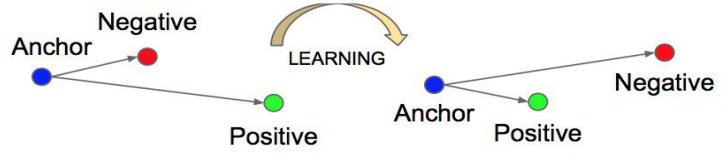


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

Triplet Loss拉近同类距离，拉开异类距离，直到满足margin

$$\mathcal{L}_{\text{tri}}(\theta) = \sum_{\substack{a,p,n \\ y_a=y_p\neq y_n}} [m + D_{a,p} - D_{a,n}]_+ .$$

$$\mathcal{L}_{\text{BH}}(\theta; X) = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \Big[m + \overbrace{\max_{p=1\ldots K} D\left(f_\theta(x_a^i), f_\theta(x_p^i)\right)}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(f_\theta(x_a^i), f_\theta(x_n^j)\right)}^{\text{hardest negative}} \Big]_+ , \quad (5)$$

Market-1501 SQ		
mAP	rank-1	rank-5
69.14	**84.92**	**94.21**
60.71	81.38	92.34
58.06	78.50	91.18

ReID中的Triplet选择Batchhard，随机抽取P个人，每个人K张图片形成一个Batch，每个人的K张图片之间形成K*(K-1)个anchor-positive(ap)对，再在剩下其他人里取一个与该ap距离最近的negtive，组成apn组将apn组按照上面式子中的公式取模型里进行训练，使得上面的式子值尽量小

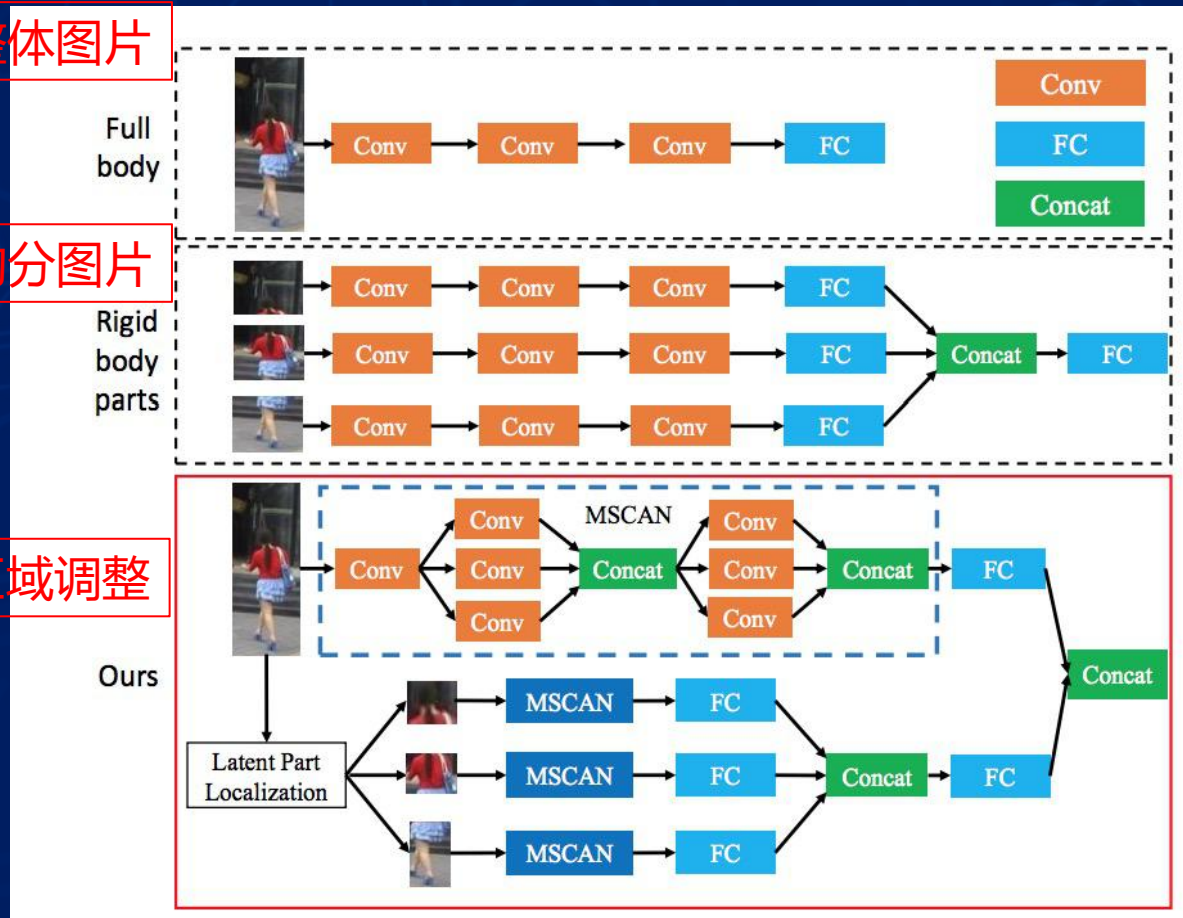
A. HERMANS, L. BEYER, AND B. LEIBE. IN DEFENSE OF THE TRIPLET LOSS FOR PERSON RE-IDENTIFICATION. *ARXIV PREPRINT* 3:07737, 2017.

ReID实现思路--局部特征学习

整体图片

均分图片

区域调整



基于局部区域调整的REID

- BANK1: 80.31%
- MAP: 57.53%

Figure 4. Six samples of original image, rigid predefined parts and learned latent pedestrian parts. Samples in each column are the same person with different backgrounds. Best viewed in color.

L. WANG, AND Y. ZHUANG. 2017. DEEPLY-LEARNED PART-ALIGNED REPRESENTATIONS FOR PERSON RE-

ReID实现思路--局部特征学习

基于姿态估计局部特征调整

- RANK1:84.14%
- MAP:63.41%

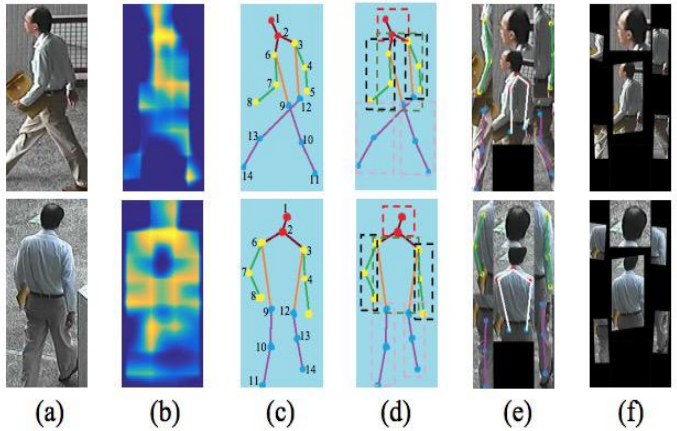


Figure 1. Illustration of part extraction and pose normalization in our Feature Embedding sub-Net (FEN). Response maps of 14 body joints (b) are first generated from the original image in (a). 14 body joints in (c) and 6 body parts in (d) can hence be inferred. The part regions are firstly rotated and resized in (e), then normalized by Pose Transform Network in (f).

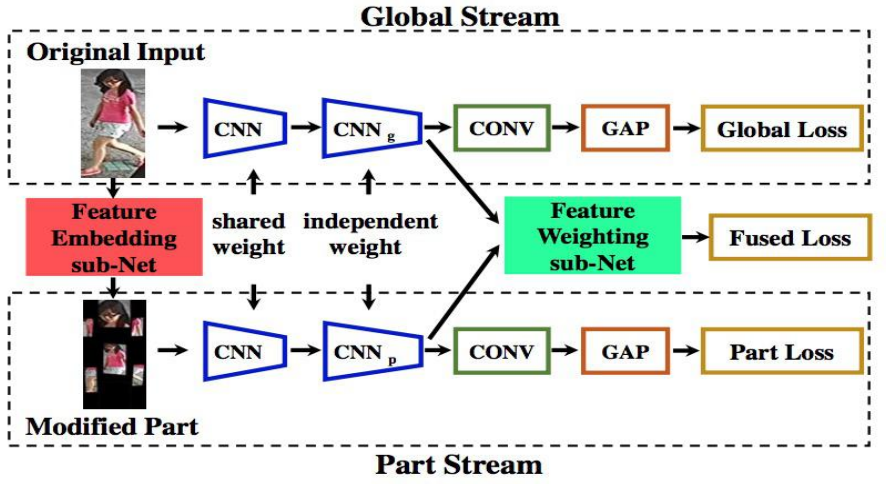


Table 2. The results on the *CUHK 03*, *Market 1501* and *VIPeR* datasets by five variants of our approach and the complete PDC.

dataset	CUHK03		Market1501		VIPeR
	labeled	detected			
method	rank1	rank1	mAP	rank1	rank1
Global Only	79.83	71.89	52.84	76.22	37.97
Part Only	53.73	47.29	31.74	55.67	22.78
Global+Part	85.07	76.33	62.20	81.74	48.42
Global+Part+FEN	87.15	77.57	62.58	83.05	50.32
Global+Part+FWN	86.41	77.62	62.58	82.69	50.00
PDC	**88.70**	**78.29**	**63.41**	**84.14**	**51.27**

HANG, J. XING, W. GAO, AND Q. TIAN. 2017. POSE-DRIVEN DEEP CONVOLUTIONAL MODEL FOR PERSON RE-IDENTICATION.

ReID实现思路--局部特征学习

简称PCB

- 2018年1月文章
- 现在较好的BASELIN
- RANK1:93.8%
- MAP:81.6%

- **设计思路**

设计模型关注局部的特征

- **设计方案**

将特征图谱在(b,24,8,2048)位置按照纵向分割为6部分，每部分单独进行Softmax分类

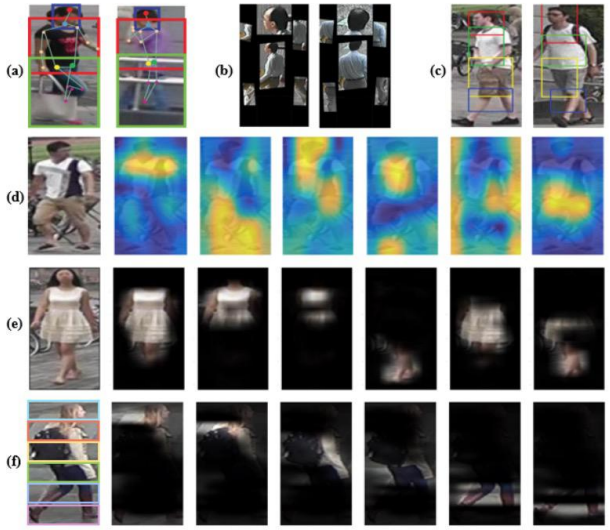


Figure 1. Partition strategies of several deep part models in person retrieval. (a) to (e): Partitioned parts by GLAD [31], PDC [27], DPL [35], Hydra-plus [22] and PAR [37], respectively, which are cropped from the corresponding papers. (f): Our method employs a uniform partition and then refines each stripe. Both PAR [37] and our method conduct "soft" partition, but our method differs significantly from [37], as detailed in Section 2.

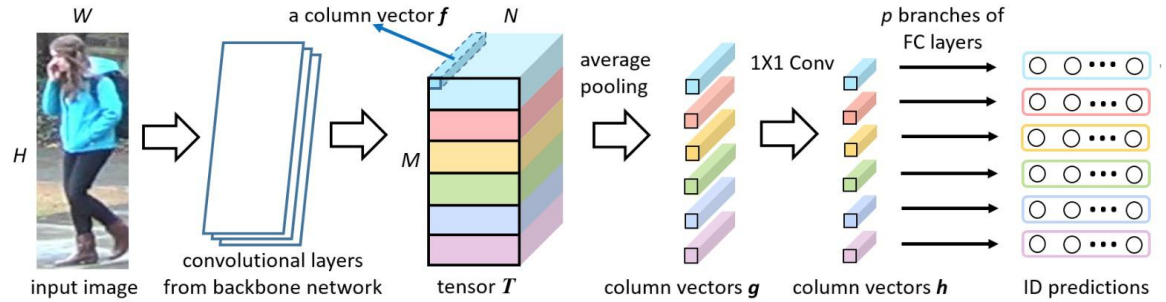


Figure 2. Structure of PCB. The input image goes forward through the stacked convolutional layers from the backbone network to form a 3D tensor $\boldsymbol{T}$. PCB replaces the original global pooling layer with a conventional pooling layer, to spatially down-sample $\boldsymbol{T}$ into p pieces of column vectors $\boldsymbol{g}$. A following 1×1 kernel-sized convolutional layer reduces the dimension of $\boldsymbol{g}$. Finally, each dimension-reduced column vector $\boldsymbol{h}$ is input into a classifier, respectively. Each classifier is implemented with a fully-connected (FC) layer and a sequential Softmax layer. During training, each classifier predicts the identity of the input image and is supervised by Cross-Entropy loss. During testing, either p pieces of $\boldsymbol{g}$ or $\boldsymbol{h}$ are concatenated to form the final descriptor of the input image.

Models	Feature	dim	Market-1501				DukeMTMC-reID				CUHK03			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
IDE	pool5	2048	85.3	94.0	96.3	68.5	73.2	84.0	87.6	52.8	43.8	62.7	71.2	38.9
IDE	FC	256	83.8	93.1	95.8	67.7	72.4	83.0	87.1	51.6	43.3	62.5	71.0	38.3
Variant 1	$\mathcal{G}$	12288	86.7	95.2	96.5	69.4	73.9	84.6	88.1	53.2	43.6	62.9	71.3	38.8
Variant 1	$\mathcal{H}$	1536	85.6	94.3	96.3	68.3	72.8	83.3	87.2	52.5	44.1	63.0	71.5	39.1
Variant 2	$\mathcal{G}$	12288	91.2	96.6	97.7	75.0	80.2	88.8	91.3	62.8	52.6	72.4	80.9	45.8
Variant 2	$\mathcal{H}$	1536	91.0	96.6	97.6	75.3	80.0	88.1	90.4	62.6	54.0	73.7	81.4	47.2
PCB	$\mathcal{G}$	12288	92.3	97.2	98.2	77.4	81.7	89.7	91.9	66.1	59.7	77.7	85.2	53.2
PCB	$\mathcal{H}$	1536	92.4	97.0	97.9	77.3	81.9	89.4	91.6	65.3	61.3	78.6	85.6	54.2
PCB+RPP	$\mathcal{G}$	12288	**93.8**	**97.5**	**98.5**	**81.6**	**83.3**	**90.5**	**92.5**	**69.2**	62.8	79.8	86.8	56.7
PCB+RPP	$\mathcal{H}$	1536	93.1	97.4	98.3	81.0	82.9	90.1	92.3	68.5	**63.7**	**80.6**	**86.9**	**57.5**

Table 1. Comparison of the proposed method with IDE and 2 variants. Both variants are described in Section 3.3. pool5: output of Pool5 layer in ResNet50. FC: output of the appended FC layer for dimension reduction. $\mathcal{G}$ ($\mathcal{H}$): feature representation assembled with column vectors $\boldsymbol{g}$ ($\boldsymbol{h}$). Both $\boldsymbol{g}$ and $\boldsymbol{h}$ are illustrated in Fig. 2.

C. Y. YANG, Q. TIAN, AND S. WANG. BEYOND PART MODELS: PERSON RETRIEVAL WITH REFINED PART POOLING. *ARXIV PREPRINT*

云从对ReID的探索与成果

云从科技：多粒度网络（MGN）

论文题目：
Learning Discriminative Features with Multiple Granularities For Person Re-Identifcation
学习多粒度显著特征用于跨镜追踪技术（行人再识别）

Learning Discriminative Features with Multiple Granularities for Person Re-Identification

Guanshuo Wang[1]*, Yufeng Yuan[2]*, Xiong Chen[2], Jiwei Li[2], Xi Zhou[1,2]
[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2] CloudWalk Technology Co., Ltd.
guanshuo.wang@sjtu.edu.cn
{yuanyufeng, chenxiong, lijiwei, zhouxi}@cloudwalk.cn

多粒度网络（MGN）

设计原理：兼顾全局与局部细节，端到端学习

三级粒度

三支路网络

整合所有特征

多粒度特征演示

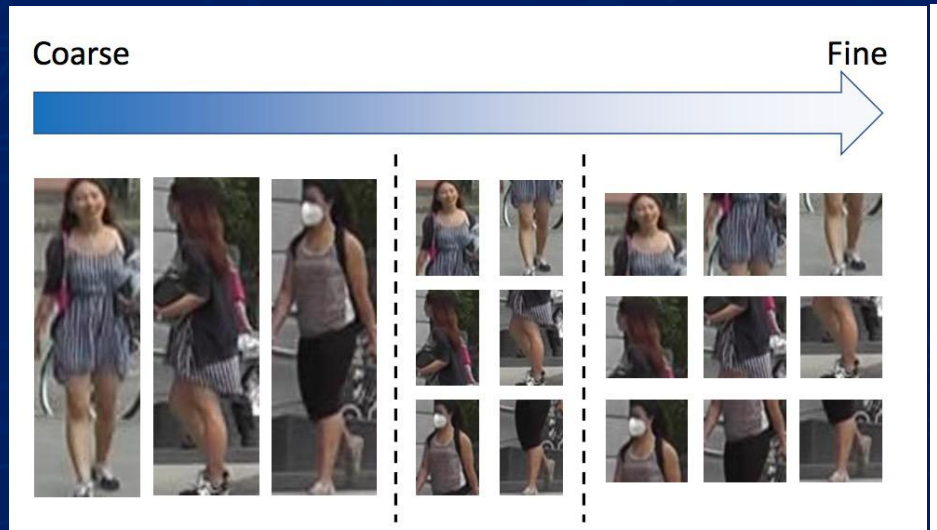


Figure 1. Body part partitions from coarse to fine granularities. We regard original pedestrian images with the whole body as the coarsest level of granularity in the left column. The middle and right column are respectively pedestrian partitions divided into 2 and 3 stripes from the original images. The more stripes images are divided into, the finer the granularity of partitions is.

图像纵向区分为不同粒度的区域

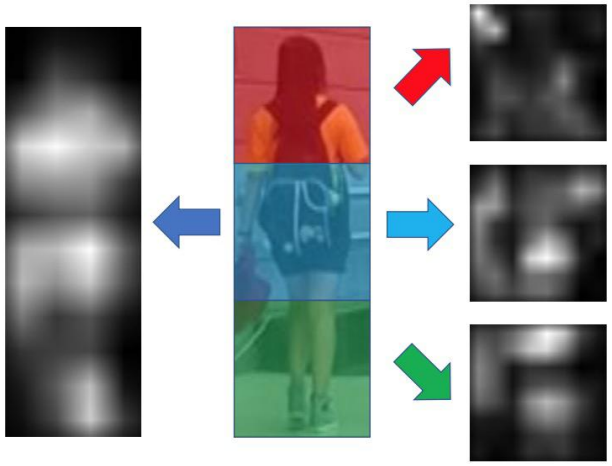
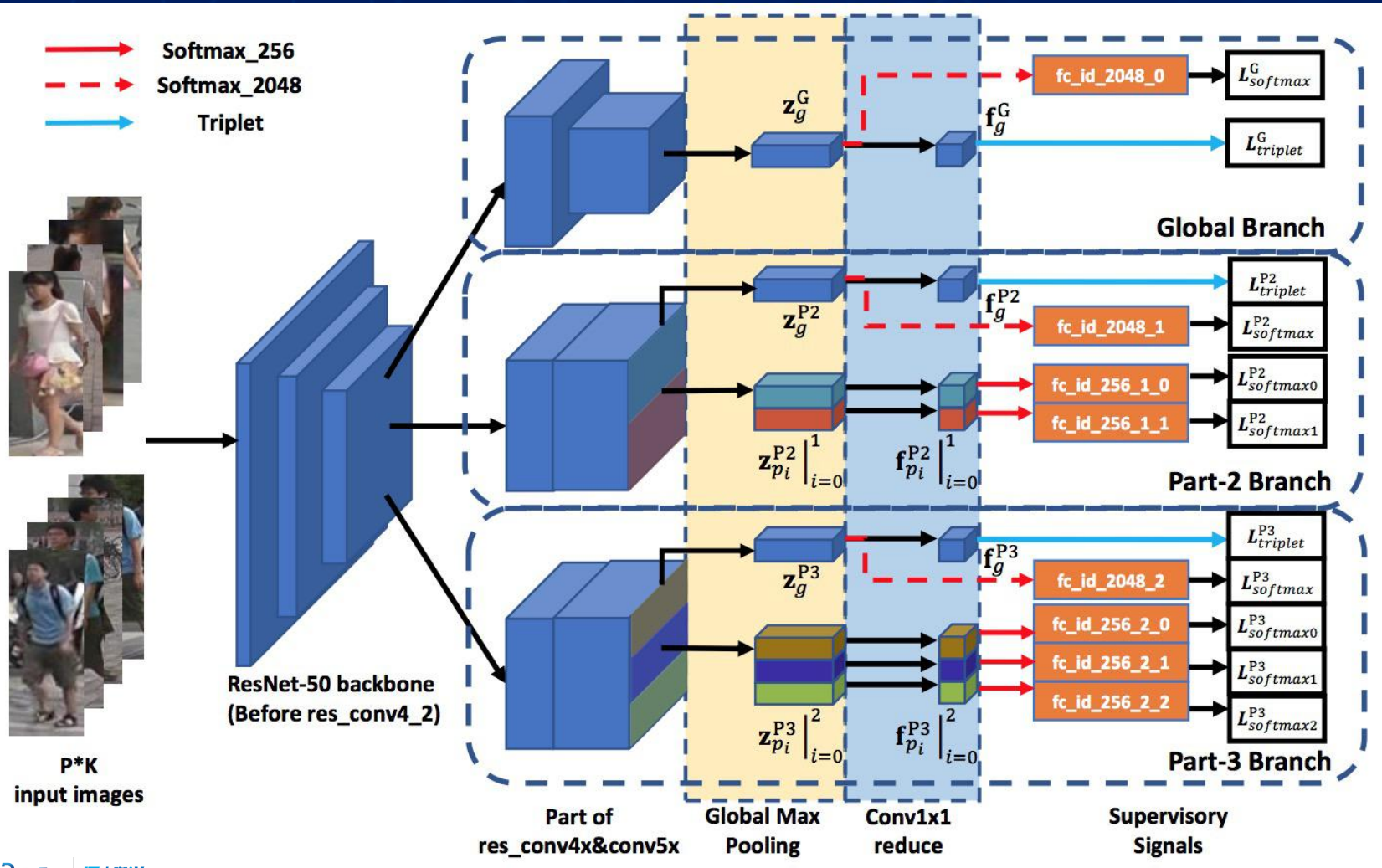


Figure 2. Attention maps in different granularities extracted from the last output of different models. Middle Column: a pedestrian image. Left Column: global attention map by IDE embedding. Right Column: three local attention maps corresponding to three split stripes of the origin image, extracted by part-based model. Best viewed in color.

不同粒度的网络关注点分布

多粒度网络（MGN）--网络结构

网络逻辑直观有效，易复现，易迁移！



STRUCTURE

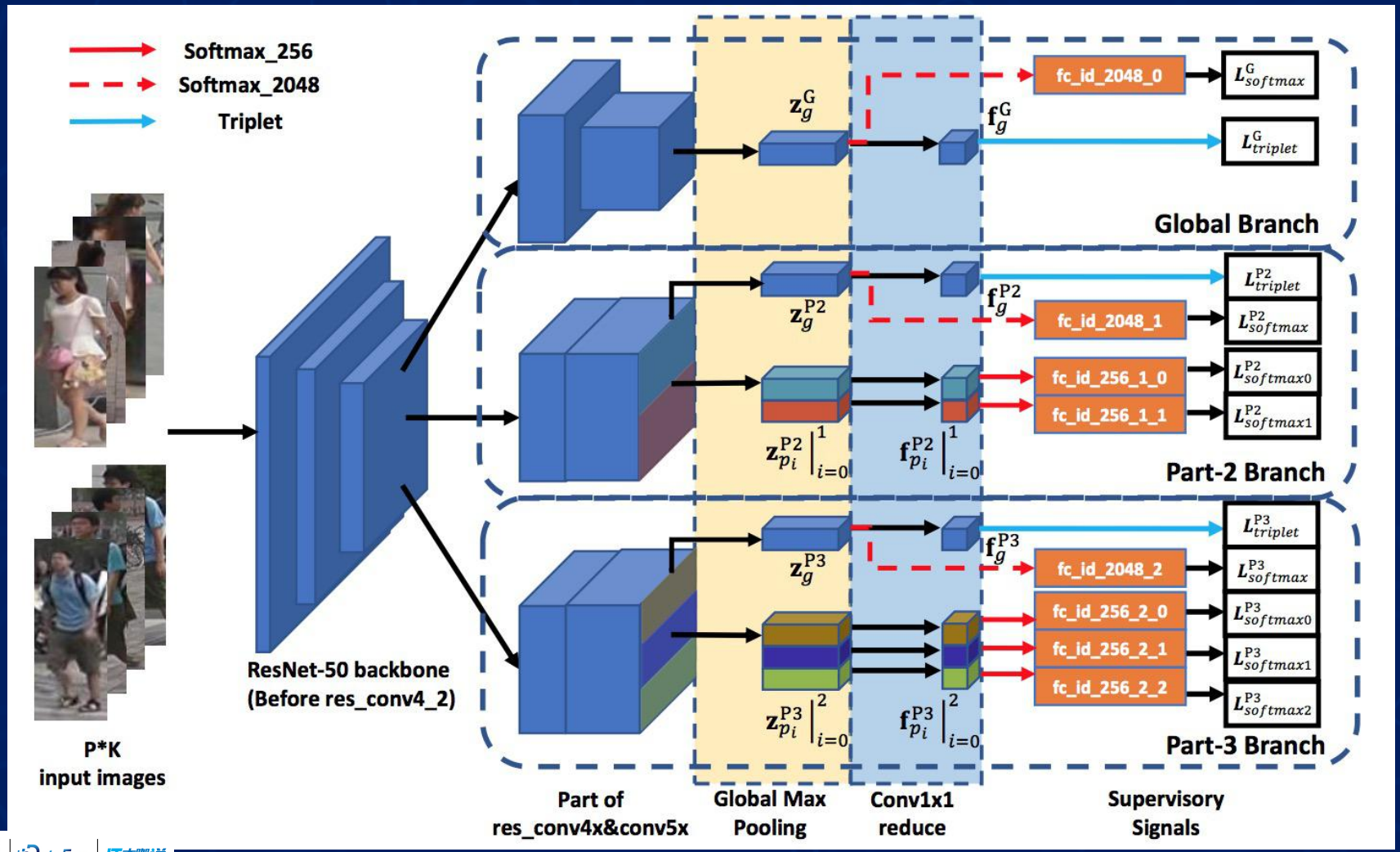
- INPUT: 384 * 128
- BASE: RESNET-50
- BRANCHES FROM RES_CONV4_1
- GLOBAL BRANCH & PART-N BRANCH
- STRIDE IN RES_CONV5_1 BLOCK
- FEATURE MAP SIZE (见下图)
- SPLIT FEATURES TO MULTI STRIPES
- MAX-POOLING

Branch	Part No.	Map Size	Dims
Global	1	12 × 4	256
Part-2	2	24 × 8	256*2+256
Part-3	3	24 × 8	256*3+256

- DIM: 8

多粒度网络（MGN）--Loss设计

网络逻辑直观有效，易复现，易迁移！



LOSS

- GLOABAL BRANCH

GLOBAL FEATURES: SOFTMAX & TRIPLET

- N-PART

GLOBAL FEATURES: SOFTMAX & TRIPLET

LOCAL FEATURES: SOFTMAX ONLY

8 Softmax + 3 Triplet

多粒度网络（MGN）--实验结果

STATE OF THE ART METHODS

Market1501

Rank1:95.7% +1.9%

mAP:86.9% +5.3%

RK:

Rank1:96.6%

mAP:94.2%

DukeMTMC-reID

Rank1:88.7% +3.5%

mAP:78.4% +5.6%

CUHK03

Rank1:68%

mAP:67.4%

Methods	Single Query		Multiple Query	
	Rank-1	mAP	Rank-1	mAP
MSCAN[18]	80.3	57.5	86.8	66.7
DLPA[40]	81.0	63.4	-	-
SVDNet[31]	82.3	62.1	-	-
PDC[30]	84.1	63.4	-	-
TriNet[14]	84.9	69.1	90.5	76.4
JLML[20]	85.1	65.5	89.7	74.5
DML[38]	87.7	68.8	91.7	77.1
DPFL[6]	88.6	72.6	92.2	80.4
HA-CNN[21]	91.2	75.7	93.8	82.8
GP-reid[2]	92.2	81.2	94.7	87.3
PCB [32]	92.3	77.4	-	-
Deep-Person[3]	92.3	79.6	94.5	85.1
Aligned-ReID[37]	92.6	82.3	-	-
PCB+RPP[32]	93.8	81.6	-	-
MGN(Ours)	**95.7**	**86.9**	**96.9**	**90.7**
TriNet(RK)[14]	86.7	81.1	91.8	87.2
GP-reid(RK)[2]	92.2	90.0	94.2	91.2
Aligned-ReID(RK)[37]	94.0	91.2	-	-
MGN(Ours, RK)	**96.6**	**94.2**	**97.1**	**95.9**

Table 2. Comparison of results on Market-1501 with Single Query setting (SQ) and Multiple Query setting (MQ). "RK" refers to implementing re-ranking operation.

Methods	Rank-1	mAP
PAN[43]	71.6	51.5
FMN[9]	74.5	56.9
SVDNet[31]	76.7	56.8
PSE[27]	79.8	62.0
HA-CNN[21]	80.5	63.8
Deep-Person[3]	80.9	64.8
PCB[32]	83.3	69.2
GP-reid[2]	85.2	72.8
MGN(Ours)	**88.7**	**78.4**

Table 3. Comparison of results on DukeMTMC-reID.

Methods	Labeled		Detected	
	Rank-1	mAP	Rank-1	mAP
BOW+XQDA[41]	7.9	7.3	6.4	6.4
LOMO+XQDA[22]	14.8	13.6	12.8	11.5
IDE[42]	22.2	21.0	21.3	19.7
PAN[43]	36.9	35.0	36.3	34.0
SVDNet[31]	40.9	37.8	41.5	37.3
HA-CNN[21]	44.4	41.0	41.7	38.6
MLFN[4]	54.7	49.2	52.8	47.8
PCB[32]	-	-	61.3	54.2
PCB+RPP[32]	-	-	63.7	57.5
MGN(Ours)	**68.0**	**67.4**	**66.8**	**66.0**

Table 4. Comparison of results on CUHK03 with evaluation protocols in [45].

在三个权威的Reid的数据集上，我们的结果都获得了现阶段最好的结果，无论是Rank1还是mAP都有明显的优势，特别是mAP有着非常巨大的gap，这个结果能够正面我们的设计方案在解决ReID这个课题时还是非常有效的

多粒度网络（MGN）

多粒度网络关注具有显著区分度的细节特征

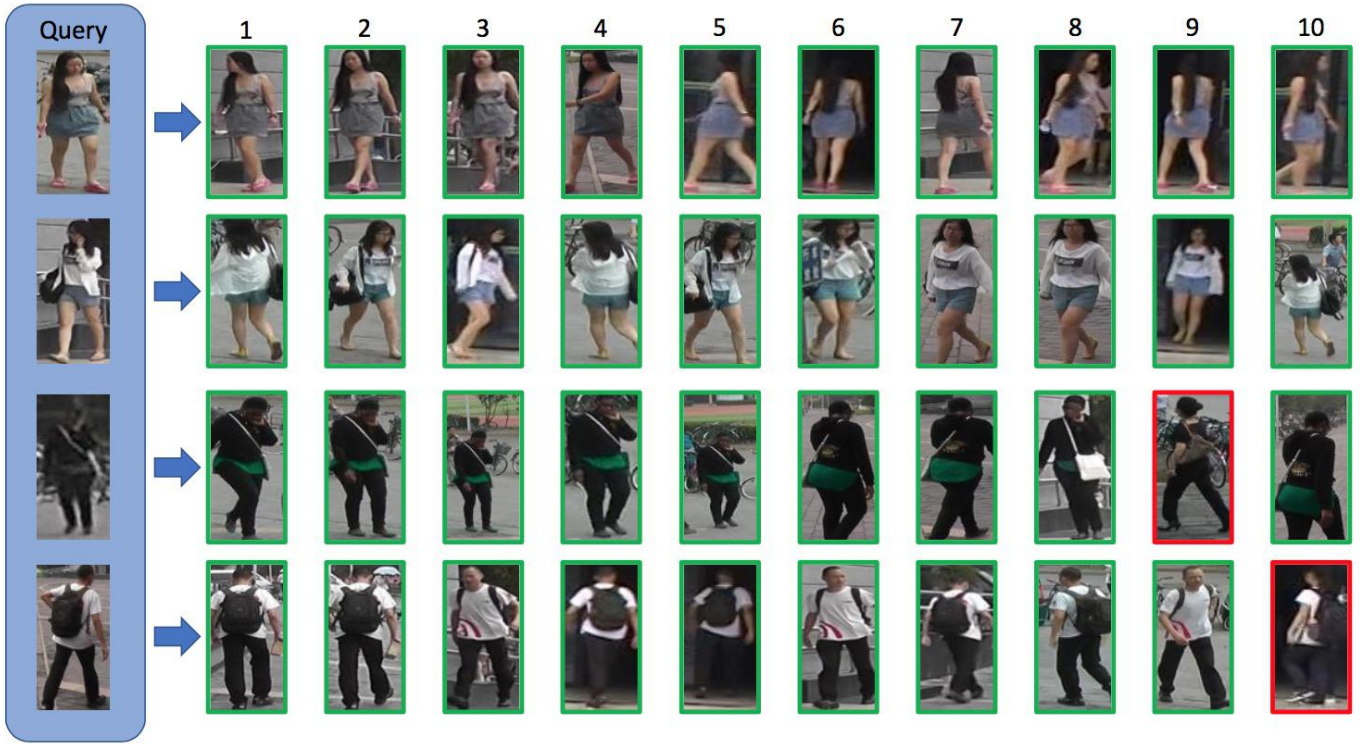


Figure 4. Top-10 ranking list for some query images on Market-1501 datasets by MGN. The retrieved images are all from in the gallery set, but not from the same camera shot. The images with green borders belong to the same identity as the given query, and that with red borders do not.

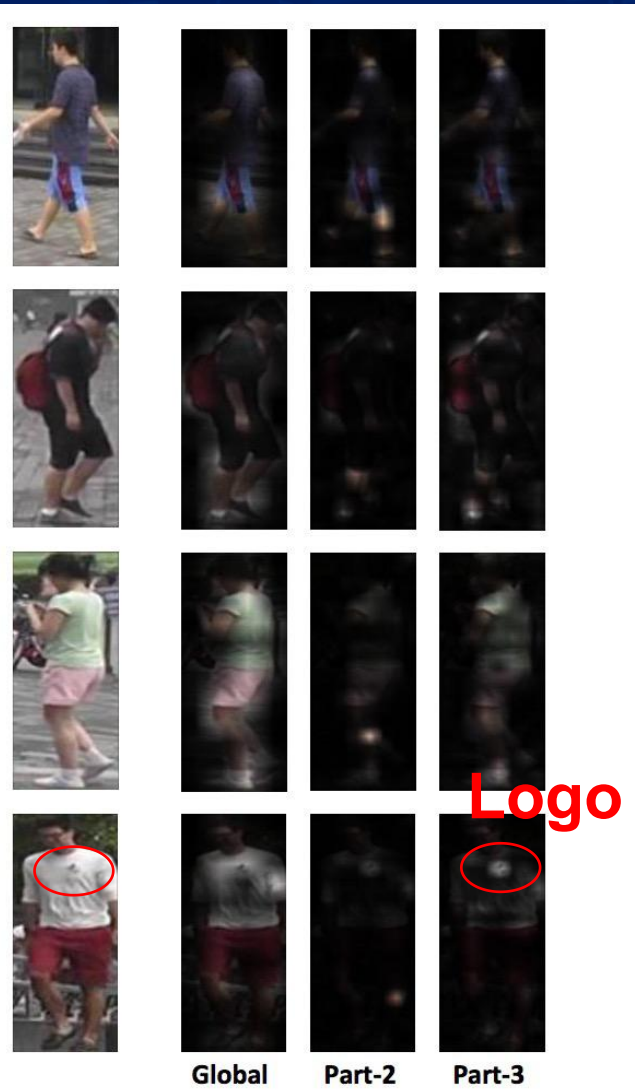


Figure 5. Attention maps extracted from output layers of every branches. First column: the original pedestrian images. Second column: attention maps from Global Branch. Third column: attention maps from Part-2 Branch. Fourth column: attention maps from Part-3 Branch. The brighter the area is, the more attention is paid.

应用场景与技术展望

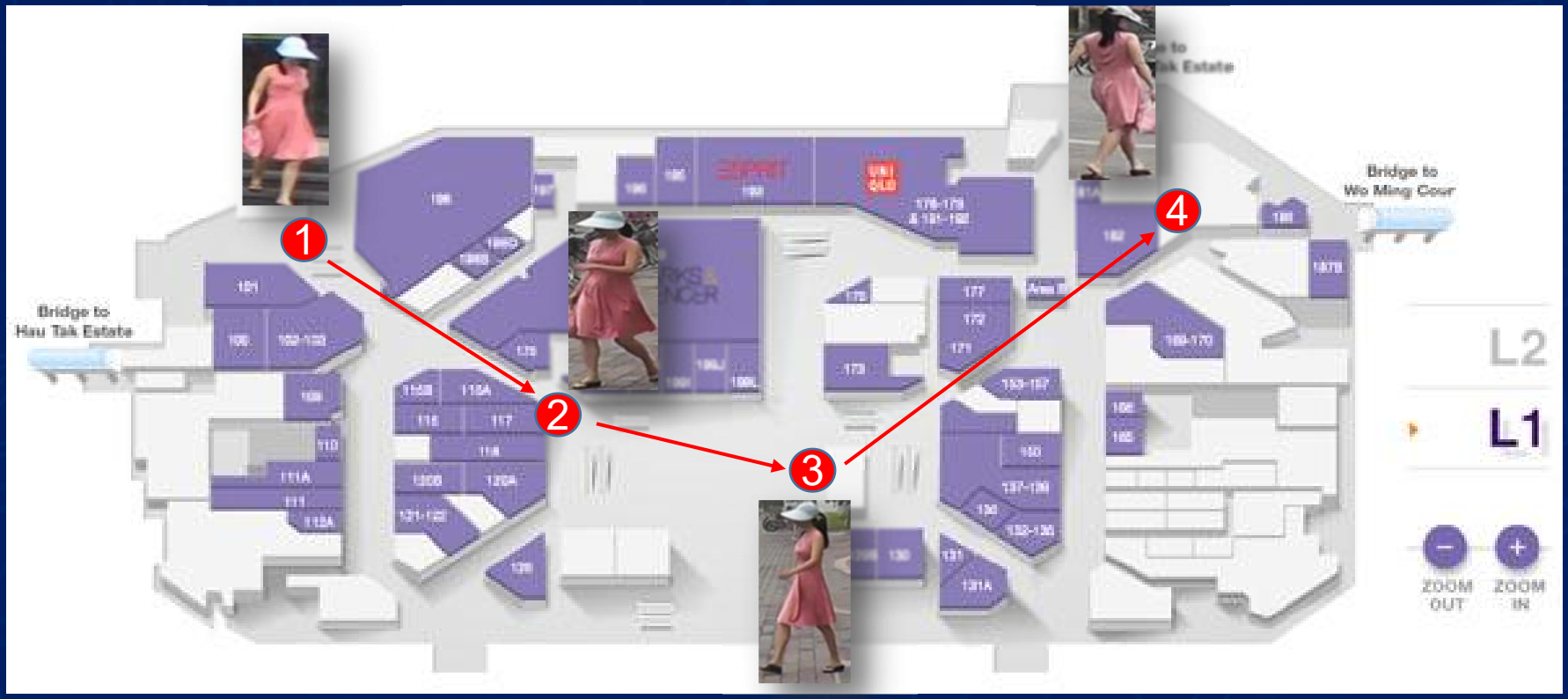
ReID的应用场景

智能安防

已知嫌疑人图片　　搜索历史监控视频获得更多嫌疑人图片

ReID的应用场景

智能商业—大型商场



技术点：用户行进与停留轨迹识别，智能理解用户需求（公共场合）

ReID的应用场景

智能商业—无人超市

AMAZON GO等

利用REID技术实现在超市内用户的跟踪与行为分析

ReID的应用场景

相册聚类

将相册里同一个人的照片归类到一起

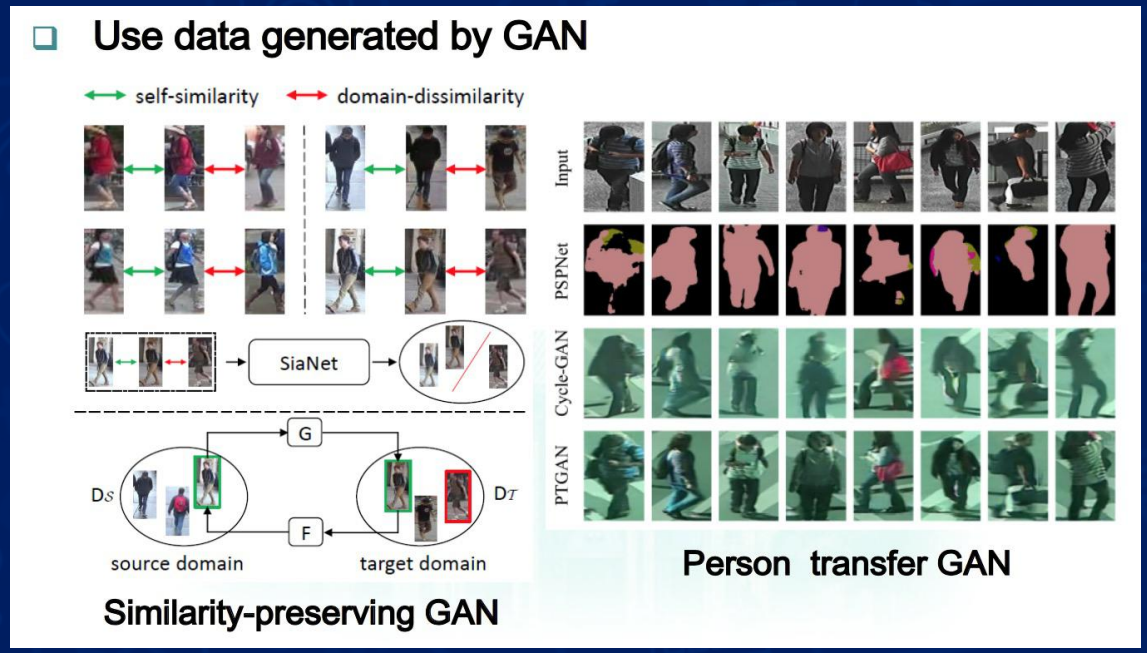
ReID的应用场景

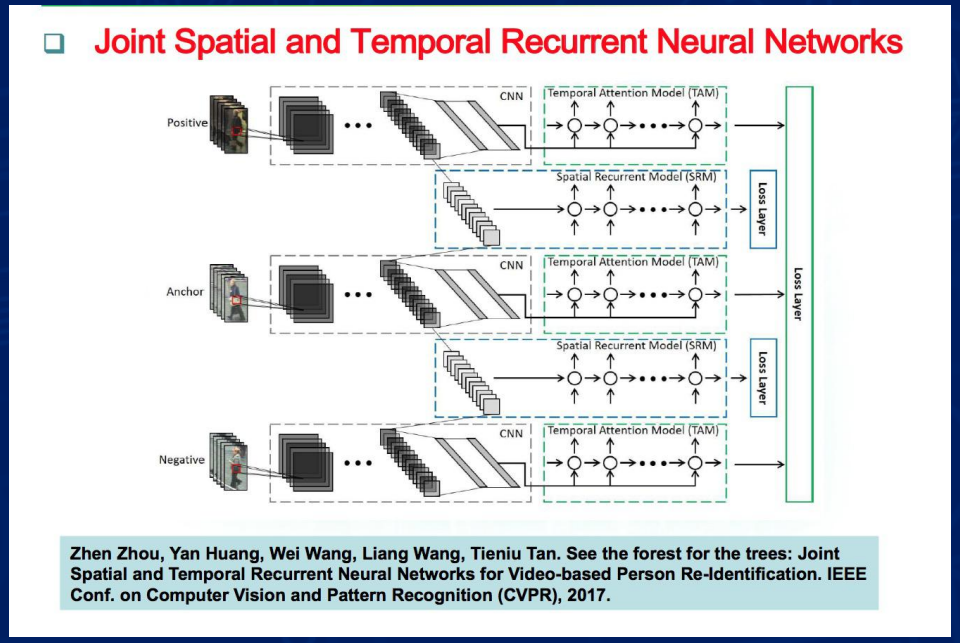
家庭机器人

通过衣着或者姿态认知主人，容易认知不容易跟丢，以便于做出互动跟随等动作

ReID的技术展望

- 应用无监督学习
- 基于视频的ReID



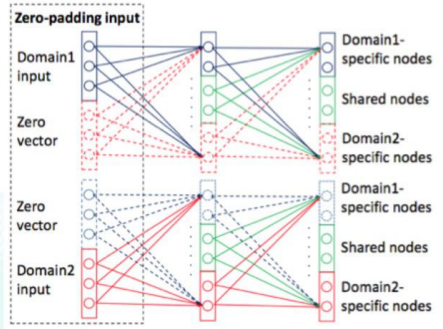


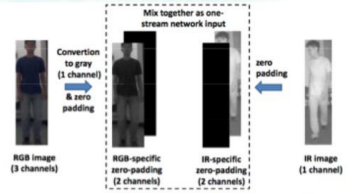
ReID的技术展望

- 跨模态的ReID
- 跨场景的迁移学习
- 应用系统设计

❑ RGB-IR Re-ID (ICCV2017)

○ Deep zero-padding

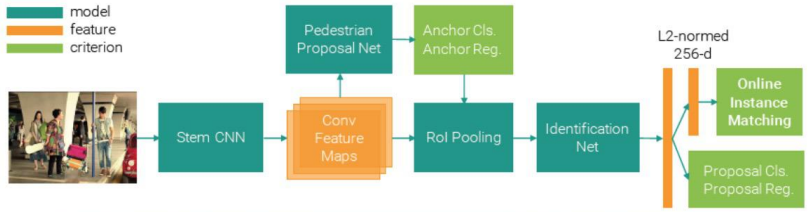




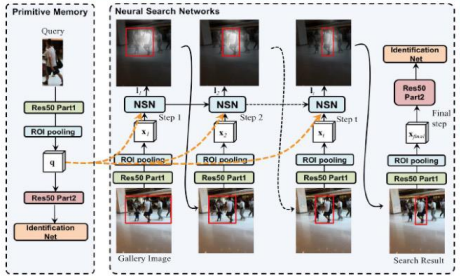
Ancong Wu, Wei-Shi Zheng*(PI), Hongxing Yu, Shaogang Gong, Jianhuang Lai. RGB-Infrared Cross-Modality Person Re-Identification. In IEEE Conf. on Computer Vision (ICCV), 2017.

Figure 6. Deep zero-padding for RGB and infrared (IR) images.

❑ Deep Learning for Person Search



Joint Detection and Identification Feature Learning (CVPR2017)



Neural Person Search Machines (ICCV2017)

Tong Xiao et al., "Joint Detection and Identification Feature Learning for Person Search", CVPR, 2017.
Hao Liu et al., "Neural Person Search Machines", ICCV, 2017.


云从科技
CLOUDWALK
THANK YOU
袁余锋
zhejiangyyf@163.com