Pedestrian Re-identification

Previous

- ECCV2016
  - Benchmark
    1. MARS: A Video Benchmark for Large-Scale Person Re-identification (video)
  - Non-typical approach
    1. Human-In-The-Loop Person Re-Identification
    2. Person Re-identification by Unsupervised L1 Graph Learning
    3. Human Re-identification in Crowd Videos using Personal, Social and Environmental Constraints (video)
  - Typical approach
    1. Embedding Deep Metric for Person Re-identification A Study Against Large Variations
    2. .Person Re-Identification via Recurrent Feature Aggregation (video)
    3. Temporal Model Adaptation for Person Re-Identification
    4. Deep Attributes Driven Person Re-identification
  - Siamese Network
    1. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification
    2. A Siamese Long Short-Term Memory Architecture for Human Re-Identification

Supplement

- Benchmark
    1. PATE：Pedestrian Attribute Recognition At Far Distance

       (http://mmlab.ie.cuhk.edu.hk/projects/PETA.html)
- Non-typical approach
    1. Hierarchical Gaussian Descriptor for Person Re-Identification (not CNN) (CVPR2016)
    2. Video-Based Pedestrian Re-Identification by Adaptive Spatio-Temporal Appearance Model. (IEEE Transactions on Image Processing, 2017.)
    3. Enhancing Person Re-identification in a Self-trained Subspace (ACM TOMM2017)
    4. Relevance Subject Machine: A Novel Person Re-identification Framework (IEEE.PRML)
- Typical approach（Presentation, 2017/4/28）
    1. Person Re-Identification Using CNN Features Learned from Combination of Attributes (ICLR2016)
    2. Improving Person Re-identification by Attribute and Identity Learning (L Zheng)
    3. Person Re-Identification by Multi-Channel Parts-Based CNN with ImprovedTriplet Loss Function (CVPR2016)
    4. Beyond triplet loss: a deep quadruplet network for person re-identification (CVPR2017)
    5. Person Search with Natural Language Description (Sensetime)
    6. Video-based Person Re-identification with Accumulative Motion Context (颜水成)
    7. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro
    8. Pose Invariant Embedding for Deep Person Re-identification
    9. SVDNet for Pedestrian Retrieval
    10. Re-ranking Person Re-identification with k-reciprocal Encoding (CVPR2017)

PETA 数据集：http://mmlab.ie.cuhk.edu.hk/projects/PETA.html

Composition of PEdesTrian Attribute (PETA) dataset

The PETA dataset consists of 19000 images, with resolution ranging from 17-by-39 to 169-by-365 pixels. Those 19000 images include 8705 persons, each annotated with 61 binary and 4 multi-class attributes. The detail composition can be seen from the table below.

Datasets	#Images	Camera angle	View point	Illumination	Resolution	Scene
3DPeS	1012	high	varying	varying	from 31x100 to 236 x 178	outdoor
CAVIAR4REID	1220	ground	varying	low	from 17x39 to 72x141	outdoor
CUHK	4563	high	varying	varying	80x160	indoor
GRID	1275	varying	frontal & back	low	from 29x87 to 169x365	indoor
i-LIDS	477	medium	back	high	from 32x76 to 115x294	outdoor
MIT	888	ground	back	high	64x128	outdoor
PRID	1134	high	profile	low	64x128	outdoor
SARC3D	200	medium	varying	varying	from 54x187 to 150x307	outdoor
TownCentre	6967	medium	varying	medium	from 44x109 to 148x332	outdoor
VIPeR	1264	ground	varying	varying	48x128	outdoor
Total = PETA	**19000**	**varting**	**varying**	**varying**	**varying**	**varying**

Sample Images in PETA Dataset

- Person Re-Identification Using CNN Features Learned from Combination of Attributes(ICLR2016)

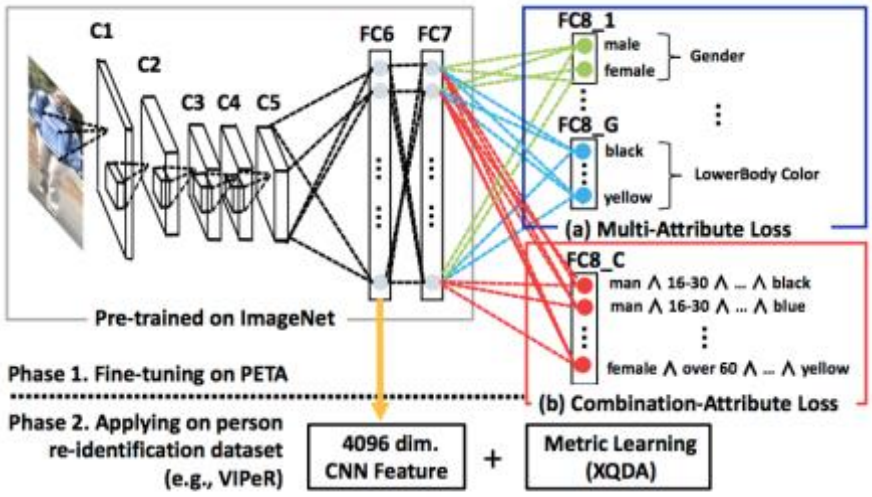


Fig. 2. The proposed CNN features. For a CNN fine-tuning, new fully connected layers and softmax loss layers for classifying multi-attributes and combination-attributes are attached to the FC7 layers of AlexNet.

论文核心思想是对 Alexnet 进行 fine-tuning，Alexnet 模型是在 ImageNet 数据集上预训练的。Fine-tuning 分为两步，第一步是在相关数据集上对整个网络进行训练，所有的 attributes 都会计算 loss 用于训练所有层，第二步是把训练的网络的 FC6 层作为输出，得到的特征向量再次进行一次 metric learning 来实现 Re-id 的任务。

- 标签损失
  - 全标签 L^C，集成所有属性：属性 1 类别×属性 2 类别数×…×属性 N 类别数
  - 第 G 个属性的标签损失 L^G，
- 损失函数
  - 全标签的交叉熵损失和每个属性的交叉熵损失平均的加权

$$L = \alpha L^C + (1-\alpha)\frac{1}{G}\sum_{g=1}^{G} L^g,$$

- Label

TABLE I
GROUP OF MUTUALLY EXCLUSIVE ATTRIBUTES.

Group (g)	Attributes	$K^{(g)}$
Gender	male, female	2
Age	less 15, 15-30, 31-45, 46-60, over 60	5
Luggage	backpack, other, folder, luggage case nothing, plastic bags, suitcase	7
UpperBody Clothing	sweter, tshit, suit, jacket, no sleeve, other	6
UpperBody Color	black, blue, brown, green, grey orange, pink, purple, red, white, yellow	11
LowerBody Clothing	suit, shorts, shirt skirt, long skirt trousers, hot pants, jeans, capri	8
LowerBody Color	black, blue, brown, grey, pink, red, white, yellow	8

- Improving Person Re-identification by Attribute and Identity Learning (L Zheng)

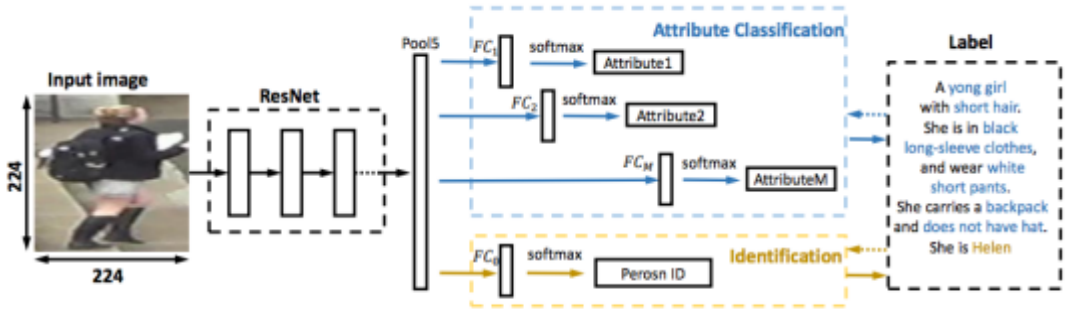


Figure 2. An overview of the APR network. During training, it predicts M attribute labels and an ID label. The weighted sum of the individual losses is back propagated. During testing, we extract the Pool5 (ResNet-50) or FC7 (CaffeNet) descriptors for retrieval.

这篇论文和上一篇论文几乎完全一样，最主要的不同点在于计算损失的时候 Loss 增加了 ID 的分类 loss，另外一个小的不同点是 CNN 使用的网络也不一样。

- ID loss

$$L_{\mathrm{ID}}(f,d) = -\sum_{k=1}^{K} log(p(k))q(k).$$

- Attributes loss

$$L_{\mathrm{att}}(f,l) = -\sum_{j=1}^{m} log(p(j))q(j),$$

- Final loss

$$L = \lambda L_{\mathrm{ID}} + \frac{1}{M}\sum_{i=1}^{M} L_{\mathrm{att}},$$

- Result

Methods	rank-1	rank-5	rank-10	rank-20	mAP
DADM[34]	39.4	-	-	-	19.6
MBC[37]	45.56	67	76	82	26.11
SML[15]	45.16	68.12	76	84	-
DLDA[40]	48.15	-	-	-	29.94
SL[4]	51.9	-	-	-	26.35
DNS[45]	55.43	-	-	-	29.87
LSTM[39]	61.6	-	-	-	35.3
S-CNN[38]	65.88	-	-	-	39.55
2Stream[53]*	79.51	90.91	94.09	96.23	59.87
GAN[54]*	79.33	-	-	-	55.95
Pose[50]*	78.06	90.76	94.41	96.52	56.23
Deep[9]*	83.7	-	-	-	**65.5**
B1 (C, 651)	52.13	73.33	80.84	86.90	27.29
B1 (R, 651)	70.51	86.40	90.82	93.91	48.19
B1 (R, 751)	73.69	88.15	91.80	94.83	51.48
B2 (R, 651)	49.76	70.07	77.76	83.87	23.95
APR (C, 651)	57.54	78.26	85.03	90.38	32.85
APR (R, 651)	82.98	92.81	95.30	96.94	61.98
APR (R, 751)	**84.29**	**93.20**	**95.19**	**97.00**	64.67

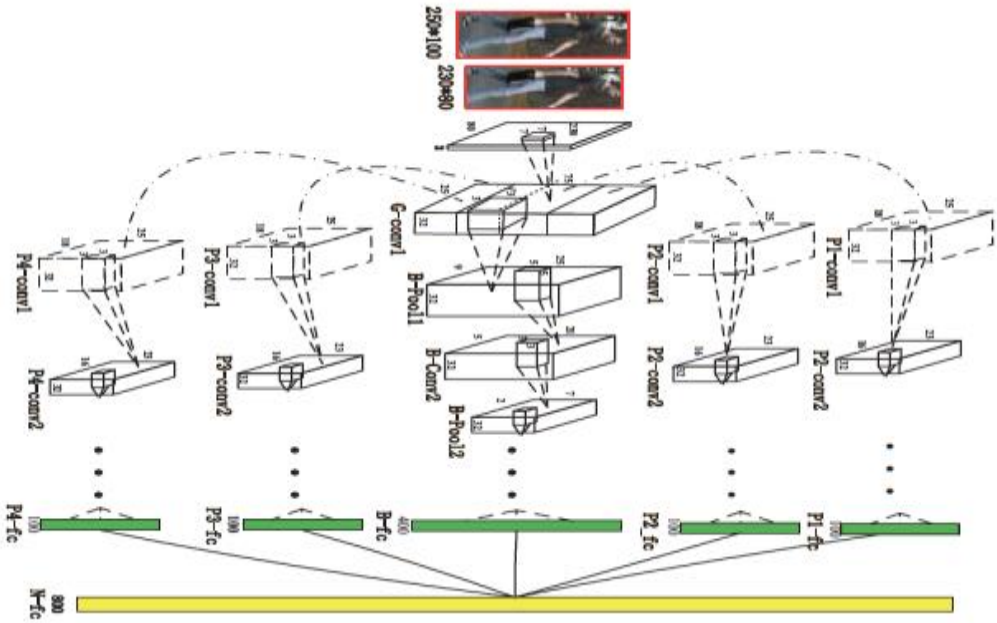
Table 1. Comparison with state of the art on Market-1501. "B1"

Methods	rank-1	mAP
BoW+kissme [51]	25.13	12.17
LOMO+XQDA [24]	30.75	17.04
GAN (R, 702) [54]	67.68	47.13
B1 (R, 702)	64.22	43.50
B2 (R, 702)	52.91	31.23
APR (R, 702)	**70.69**	**51.88**

Table 2. Comparison with the state of the art on DukeMTMC-reID.

- Person Re-Identification by Multi-Channel Parts-Based CNN with ImprovedTriplet Loss Function (CVPR2016)

思路使用典型的 Triplet loss，一个原始样本，一个正样本和一个负样本，共享 CNN 网络层。改进点在于 CNN 层融合了多个 CNN 网络层进行 Ensemble learning。另外通过 crop 的方式进行了 data argumentation。最大改进点在于设计了 loss function，同时考虑了类内损失和类间损失。结果显示基本都提高了 5~10%



- Loss function
  - Inter-class-constraint

$$d^n(I_i^o, I_i^+, I_i^-, \mathbf{w}) = \\ d(\phi_{\mathbf{w}}(I_i^o), \phi_{\mathbf{w}}(I_i^+)) - d(\phi_{\mathbf{w}}(I_i^o), \phi_{\mathbf{w}}(I_i^-)) \leq \tau_1. \quad (1)$$

  - Intra-class-constraint

$$d^p(I_i^o, I_i^+, \mathbf{w}) = d(\phi_{\mathbf{w}}(I_i^o), \phi_{\mathbf{w}}(I_i^+)) \leq \tau_2. \quad (2)$$

  - L2-norm distance

$$d(\phi_{\mathbf{w}}(I_i^o), \phi_{\mathbf{w}}(I_i^+)) = ||\phi_{\mathbf{w}}(I_i^o) - \phi_{\mathbf{w}}(I_i^+)||^2. \quad (4)$$

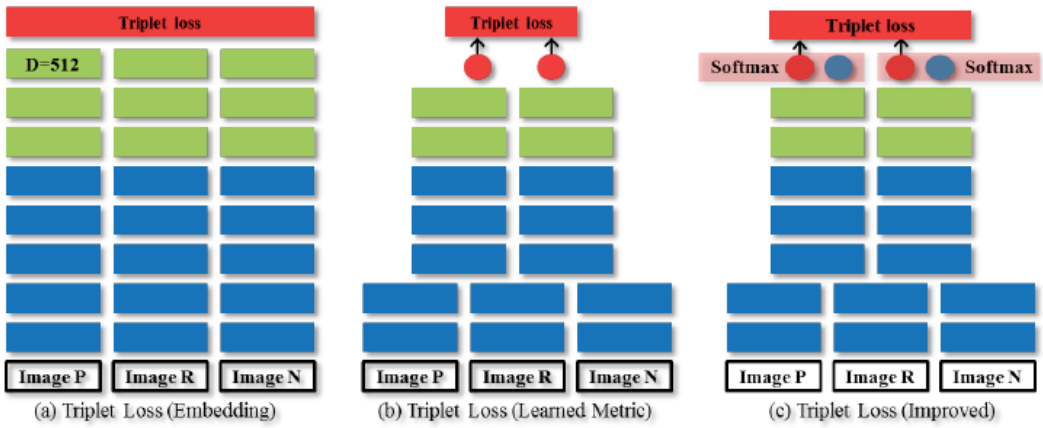
  - Improved triplet loss function

$$L(I, \mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\max\{d^n(I_i^o, I_i^+, I_i^-, \mathbf{w}), \tau_1\}}_{inter-class-constraint} \\ + \underbrace{\beta \max\{d^p(I_i^o, I_i^+, \mathbf{w}), \tau_2\}}_{intra-class-constraint}), \quad (3)$$

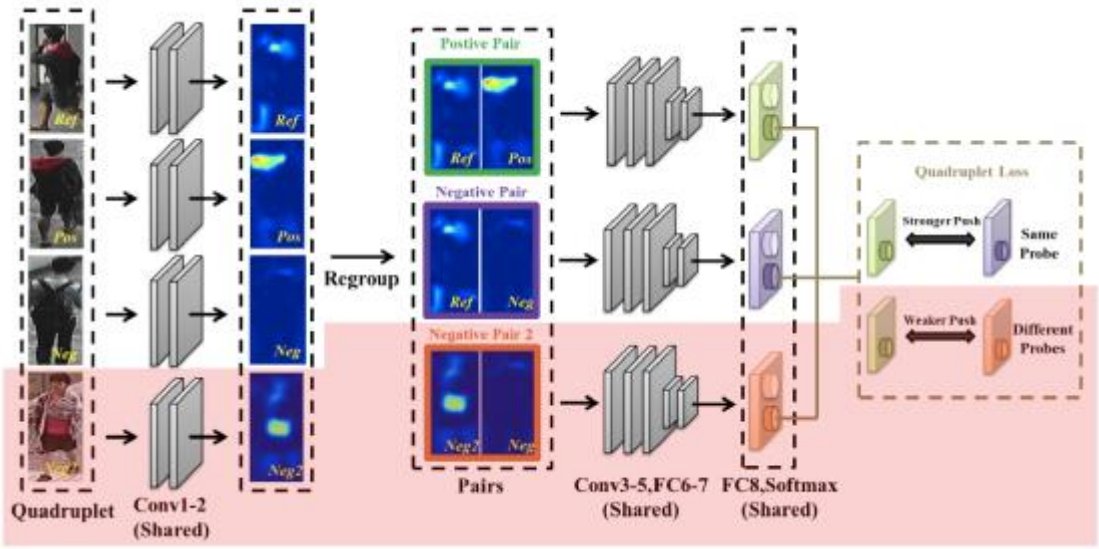
- Beyond triplet loss: a deep quadruplet network for person re-identification (CVPR2017)

本文主要贡献在于设计了一个新的 loss function，改变了传统的 triplet loss，设计了一个新的 quadruple loss，并且声称击败了绝大部分的 state-of-the-art。

- Loss 流程，Triplet Loss (Improved)为论文提出



- Network



- Quadruplet loss

$$L_{quad} = \sum_{i,j,k}^{N} [g(x_i, x_j)^2 - g(x_i, x_k)^2 + \alpha_1]_+ + \sum_{i,j,k,l}^{N} [g(x_i, x_j)^2 - g(x_l, x_k)^2 + \alpha_2]_+$$

$$s_i = s_j, s_l \neq s_k, s_i \neq s_l, s_i \neq s_k$$

其中 i,j 是正样本，k,l 分别是两个不一样的负样本

- Distance visualization

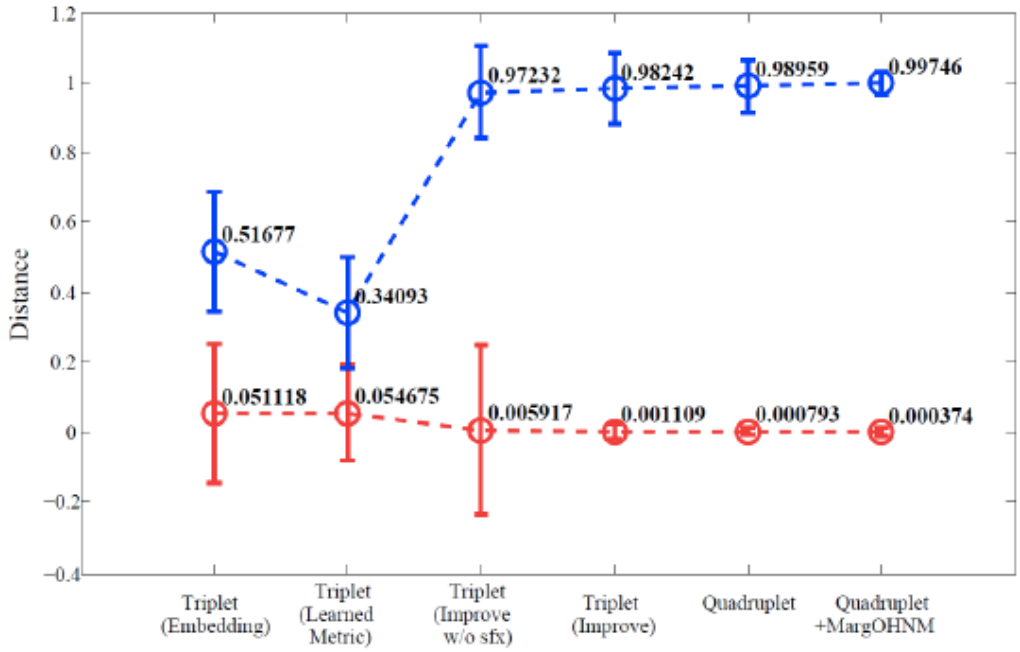


Figure 5. The distributions of intra- and inter- class distances from different models on CUHK03 training set. The red and blue lines indicate intra- and inter- distance respectively.

- Result

Table 1. The CMC performance of the state-of-the-art methods and different architectures in our method on three representative datasets.

Method	CUHK03			CUHK01(p=486)			CUHK01(p=100)			VIPeR		
	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$
ITML [6]	5.53	18.89	29.96	15.98	35.22	45.60	17.10	42.31	55.07	-	-	-
eSDC [43]	8.76	24.07	38.28	19.76	32.72	40.29	22.84	43.89	57.67	26.31	46.61	58.86
KISSME [14]	14.17	48.54	52.57	-	-	-	29.40	57.67	62.43	19.60	48.00	62.20
FPNN [19]	20.65	51.00	67.00	-	-	-	27.87	64.00	77.00	-	-	-
mFilter [44]	-	-	-	34.30	55.00	65.30	-	-	-	29.11	52.34	65.95
kLFDA [38]	48.20	59.34	66.38	32.76	59.01	69.63	42.76	69.01	79.63	32.33	65.78	79.72
DML [41]	-	-	-	-	-	-	-	-	-	34.40	62.15	75.89
IDLA [1]	54.74	86.50	94.00	47.53	71.50	80.00	65.00	89.50	93.00	34.81	63.32	74.79
SIRCIR [32]	52.17	85.00	92.00	-	-	-	72.50	91.00	95.50	35.76	67.00	82.50
DeepRanking [2]	-	-	-	50.41	75.93	84.07	70.94	92.30	96.90	38.37	69.22	81.33
DeepRDC [7]	-	-	-	-	-	-	-	-	-	40.50	60.80	70.40
NullReid [42]	58.90	85.60	92.45	64.98	84.96	89.92	-	-	-	42.28	71.46	82.94
Ensembles [24]	62.10	89.10	94.30	53.40	76.30	84.40	-	-	-	45.90	77.50	**88.90**
DeepLDA [36]	63.23	89.95	92.73	-	-	-	67.12	89.45	91.68	44.11	72.59	81.66
GOG [23]	67.30	91.00	96.00	57.80	79.10	86.20	-	-	-	**49.70**	**79.70**	88.70
GatedSiamese [30]	68.10	88.10	94.60	-	-	-	-	-	-	37.80	66.90	77.40
ImpTrpLoss [4]	-	-	-	53.70	84.30	91.00	-	-	-	47.80	74.70	84.80
DGD [37]	**80.50**	94.90	97.10	**71.70**	**88.60**	**92.60**	-	-	-	35.40	62.30	69.30
BL1: Triplet(Embedding)	60.13	90.51	95.15	44.24	67.08	77.57	63.50	80.00	89.50	28.16	52.22	65.19
BL2: Triplet(Learned Metric)	61.60	92.41	97.47	58.74	80.35	88.07	77.00	94.00	97.50	40.19	70.25	82.91
Triplet(Improved w/o sfx)	70.25	95.97	98.10	58.85	82.61	88.37	77.50	95.00	96.50	44.30	72.47	80.06
Triplet(Improved)	72.78	95.97	97.68	59.26	82.41	88.27	78.00	95.50	98.00	44.30	71.84	81.96
Quadruplet	74.47	**96.62**	98.95	62.55	83.02	88.79	79.00	96.00	97.00	48.42	74.05	84.49
BL3: Classification	68.35	93.46	97.47	58.74	79.01	87.14	76.50	94.00	97.00	44.30	69.94	81.96
Quadruplet + MargOHNM	75.53	95.15	**99.16**	62.55	83.44	89.71	**81.00**	**96.50**	**98.00**	49.05	73.10	81.96

- Person Search with Natural Language Description (Sensetime)

这篇文章和别的文章最重要的不同在于不再使用离散的属性标签，而是使用信息量更加丰富同时也更加杂乱的 NLP 标签，这样的标签更加符合人类的描述，但是也同时处理难度更大，更加适合用于互联网搜索、智能客服机器人等场景。本文给出了商汤自己标注的语义 Person Re-ID 数据集 CHUK-PEDES，总共有 80000+语句，40000+图片，13000+人。本文的算法使用的是带 Attention model 的 RNN 网络 GNA-RNN。论文里提高还特意屏蔽了语句中的某些单词，来观测哪些单词的重要度更高，从而达到提升标准性能的目的。

- 总体性能

  Attributes: Top1:33.3%; Top5:74.7%; time:81.84s

  Description: Top1:58.7%; Top5:92.0%; time:62.18s

- 数据集示例

The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing yellow sneakers, white socks with blue stripes on the top of them, black athletic shorts and a yellow with blue t-shirt. He has short black hair.

The man has dark hair and is wearing glasses. He has on a pink shirt, blue shorts, and white tennis shoes. He has on a blue backpack and is carrying a re-useable tote.

The girl is wearing a pink shirt with white shorts, she is wearing black converse, with her hair in a pony tail.

The woman has long light brown hair, is wearing a black business suit with white low-cut blouse with large, white cuffs, a gold ring, and is talking on a cellphone.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

- GNA-RNN model

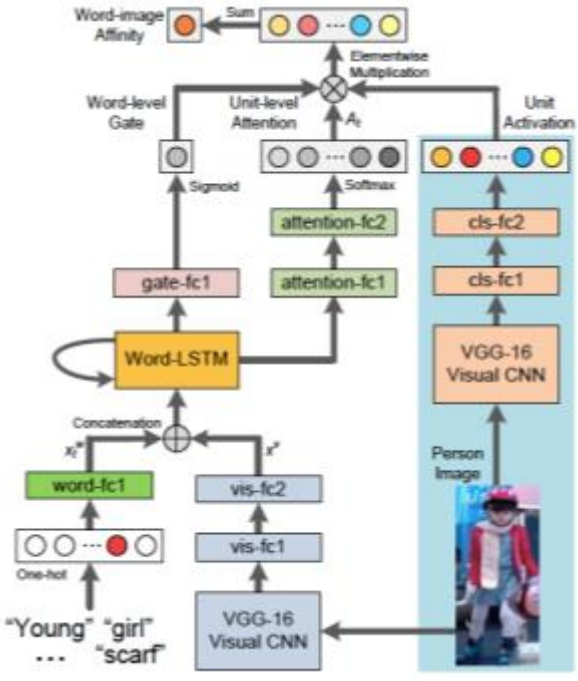


Figure 5. The network structure of the proposed GNA-RNN. It consists of a visual sub-network (right blue branch) and a language sub-network (left branch). The visual sub-network generates a series of visual units, each of which encodes if certain appearance patterns exist in the person image. Given each input word, The language sub-network outputs world-level gates and unit-level attentions for weighting visual units.

- Visual sub-network

  视觉网络是上图中蓝色的部分，用的是 VGG-16 和 512 维输出的 FC7，之后添加了两个 FC1,FC2 进行 fine-tuning，输出一个 visual feature。

- Language sub-network

  NLP 网络的输入时 word vector 和 image feature，网络是 LSTM 结合 attention unit，每次输入一个单词都会在 image 上更新 attention，网络输出为两个通道，一个为 attention vector，和视觉网络输出的 visual feature 的维度一样，另一个输出是 word gate level，表示当前单词和图像的联系度，比如 a,the 这类单词可能值就很小。

- 融合

  最后把 word gate level 和 attention vector 和 visual feature 相乘，最后得到的 feature 放到最后相加然后进行 end to end 训练，最后就可以得到每个单词的 affinity value，然后把所有单词的 affinity value 加起来就可以得到最后 sentence 的 affinity value，这个值在 0~1 之间，最后和 label（0 表示不匹配，1 表示匹配）计算交叉熵损失，训练网络。

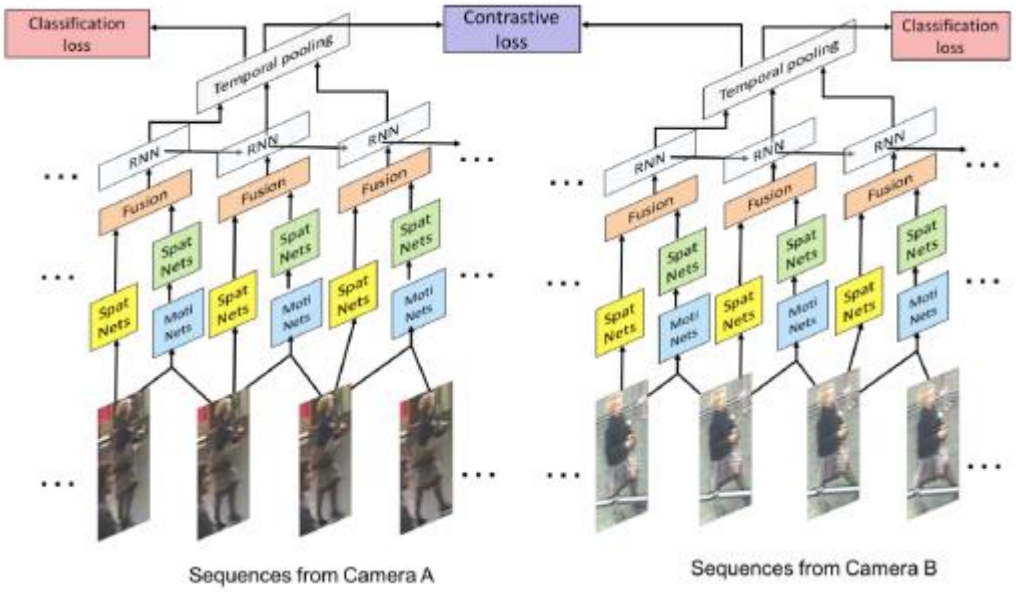
$$\hat{a}_t = g_t \sum_{n=1}^{512} A_t(n) v_n, \qquad \hat{a} = \sum_{t=1}^{T} \hat{a}_t.$$

- GNA-RNN model

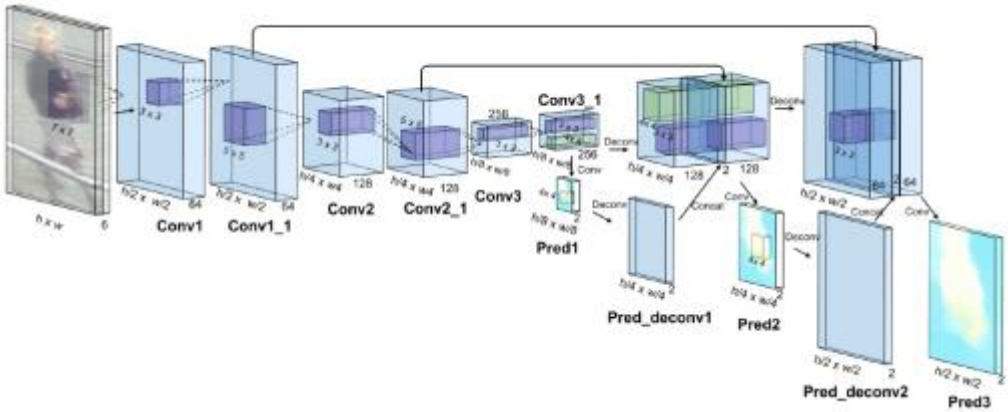
Figure 7. Images with the highest activations on 4 different visual units. The 4 units are identified as the one with the maximum average attention values in our GNA-RNN with the same word ("backpack", "sleeveless", "pink", "yellow") and a large number of images. Each unit determines the existence of some common visual patterns.

- Video-based Person Re-identification with Accumulative Motion Context (颜水成)

这篇文章最主要的思想是融合的视频的上下文信息，用了一个空间网络提取单帧图像特征，用了一个运动网络提取相邻帧运动信息，运动网络用的是 FLownet (当然运动网络也可以用光流图+CNN)，运动网络的监督信号来自于传统光流法，最后把融合这些特征放到 LSTM 上去实现上下文特征提取（水滴科技一直致力于步态辅助 Re-Id 的研究)，当然增加了 LSTM 层之后会带来计算时间的增加。

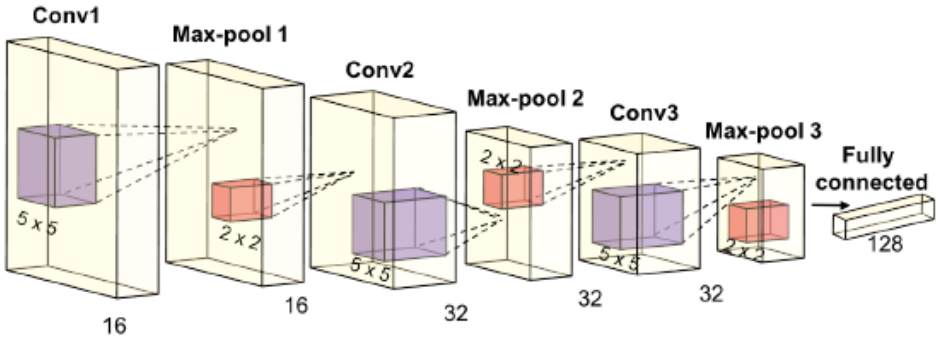


- Moti Nets (Flownet)



FLownet 有两种模式，一种是将图像叠层输入到网络，另外一种是作为两幅图像输入到双通道的网络。本文使用的是第一种模式，并且用了三个尺度的光流图做最终的训练。

- Spat Nets：一个简单的三层 CNN 网络



- Fusion：三种 fusion 模式，实验显示 concatenation fusion 在 max-pool2 层的准确度最高

- **Concatenation fusion** This fusion operation stacks the two feature maps at the same spatial locations i, j across the feature channels d:

$$y^{\text{cat}}_{i,j,2d} = x^{\text{A}}_{i,j,d}, y^{\text{cat}}_{i,j,2d-1} = x^{\text{B}}_{i,j,d}, \quad (1)$$

where $\mathbf{x}^A, \mathbf{x}^B \in \mathbb{R}^{H\times W\times D}, \mathbf{y}^{\text{cat}} \in \mathbb{R}^{H\times W\times 2D}$ and $1 \le i \le H, 1 \le j \le W$.

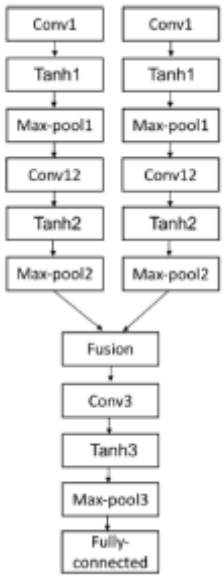
- **Sum fusion** The sum fusion computes the sum of the two feature maps at the same spatial locations i, j and channels d:

$$y^{\text{sum}}_{i,j,d} = x^{\text{A}}_{i,j,d} + x^{\text{B}}_{i,j,d}, \quad (2)$$

where $\mathbf{y}^{\text{cat}} \in \mathbb{R}^{H\times W\times 2D}$.

- **Max fusion** Similarly, max fusion takes the maximum of the two feature maps:

$$y^{\text{max}}_{i,j,d} = \max\{x^{\text{A}}_{i,j,d}, x^{\text{B}}_{i,j,d}\}. \quad (3)$$



- Loss：融合了 contrastive loss 和 classification loss

$$L_{multi}(\mathbf{u}_{(a)}, \mathbf{u}_{(b)}) = L_{con}(\mathbf{u}_{(a)}, \mathbf{u}_{(b)}) + L_{class}(\mathbf{u}_{(a)}) + L_{class}(\mathbf{u}_{(b)}).$$

- Result

Dataset	iLIDS-VID				PRID-2011			
Methods	Rank1	Rank5	Rank10	Rank20	Rank1	Rank5	Rank10	Rank20
Baseline + LK-Flow [10]	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
Baseline + EpicFlow [36]	59.3	87.2	92.7	98.2	76.2	97.5	98.2	99.0
AMOC + LK-Flow	63.3	85.3	95.1	96.4	76.0	96.5	97.4	99.6
AMOC + EpicFlow	65.5	93.1	97.2	98.7	82.0	97.3	99.3	99.4
end-to-end AMOC + LK-Flow	65.3	87.3	96.1	98.4	78.0	97.2	99.1	99.7
end-to-end AMOC + EpicFlow	**68.7**	**94.3**	**98.3**	**99.3**	**83.7**	**98.3**	**99.4**	**100**

- 训练细节

用了 crop 和 mirror 等操作进行数据增广，网络初始化细节可见原论文。

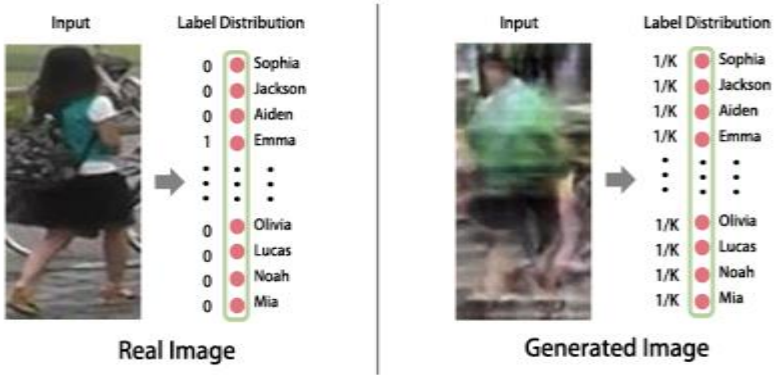
- Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro

这篇文章主要思路是用 GAN 网络进行数据生成，另外因为用的是无条件的随机 GAN 生成，所以论文提出给生成的数据用一个标签平滑正则化的技术（LSRO），所谓 LSRO 技术根据公式推导，最后的结果就是每个标签最后的值是 1/K。CNN 网络用的在 ImageNet 上预训练的 ResNet50，然后进行 fine-tuing。

- LSRO (Label & Loss)

$$q_{LSR}(k) = \begin{cases} \frac{\varepsilon}{K} & k \neq y \\ 1 - \varepsilon + \frac{\varepsilon}{K} & k = y \end{cases},$$

$$l_{LSR} = -(1-\varepsilon)\log(p(y)) - \frac{\varepsilon}{K}\sum_{k=1}^{K}\log(p(k)).$$

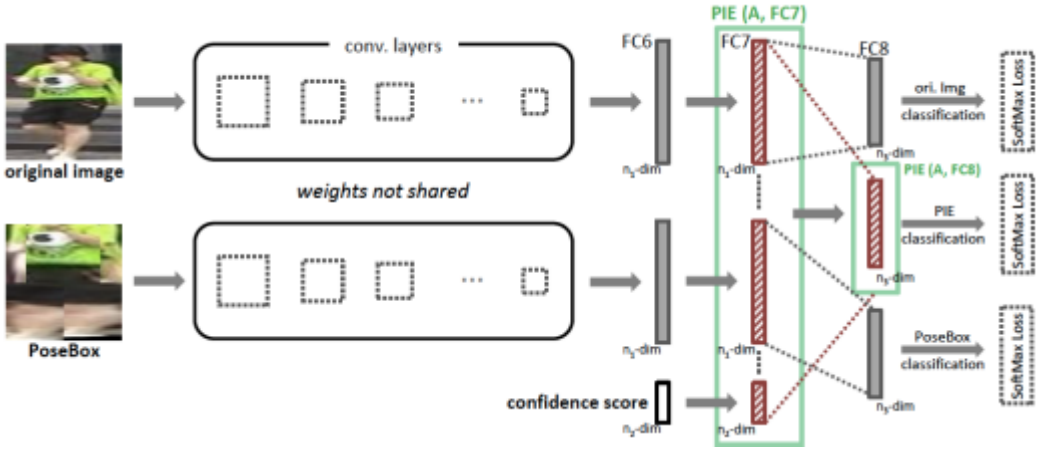


ε等于 0 的话，那么就是真实图像的 one-hot 标签，ε等于 1 的话就是随机的平滑标签，即 1/K。

- Result

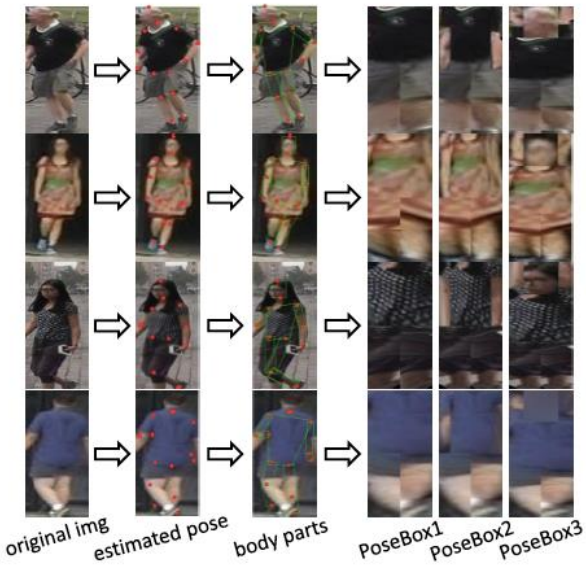
- Pose Invariant Embedding for Deep Person Re-identification

这篇文章主要思想是作者认为影响 re-ID 很重要的一个因素是行人的姿态不对齐，如果能够对姿态进行一个标准的对齐，就能够提高 re-ID 的准确度。姿态估计采用 convolutional pose machines (CPM)，然后对图片进行了仿射变换，CNN 网络是用 ResNet-50 和 Alexnet 的 fine-tuning，原始图片是 Alexnet，pose 图片是 ResNet-50，另外 pose 置信度得分的 14 维向量和 FC7 的输出向量合并。



训练网络时候用了三个 loss，原始图片 CNN 的 loss，pose 图片 CNN 的 loss，融合三个输入的 FC8 层 loss，测试的时候只用了 PIE 分类时的 feature，用了欧拉距离计算相似度。

- Result



PoseBox1: torso + legs;

PoseBox2:PoseBox1 + arms;

PoseBox3: PoseBox2 + head

14 点 pose，pose 图片进行了分块的放射变换

实验显示通常 PoseBox2 的准确度最高

- 结果显示基本在各大数据集都得到了最高准确度，验证了此方法可以一定程度上解决因相机变化而带来的姿态变化的问题，但是该方法错误多半出现在人体侧面无法定位到较好 pose 的情况下。

- SVDNet for Pedestrian Retrieval

这篇文章最主要的就是对网络进行了一个 SVD 分解，压缩了网络，并且有一点主成分分析的感觉，既压缩了模型提高了速度，也同时提高了检测准确度。简而言之，就是对网络进行了一个 PCA。

- 公式：对权重 W 矩阵进行 SVD 分解

$$
\begin{aligned}
D_{ij} &= \|\vec{f_i} - \vec{f_j}\|_2 = \sqrt{(\vec{f_i} - \vec{f_j})^{\mathrm{T}}(\vec{f_i} - \vec{f_j})} \\
&= \sqrt{(\vec{h_i} - \vec{h_j})^{\mathrm{T}} W W^{\mathrm{T}} (\vec{h_i} - \vec{h_j})} \\
&= \sqrt{(\vec{h_i} - \vec{h_j})^{\mathrm{T}} U S V^{\mathrm{T}} V S^{\mathrm{T}} U^{\mathrm{T}} (\vec{h_i} - \vec{h_j})}, \quad (2)
\end{aligned}
$$

where U, S and V are defined in Eq. 1. Since V is a unit orthogonal matrix, Eq. 2 is equal to:

$$
D_{ij} = \sqrt{(\vec{h_i} - \vec{h_j})^{\mathrm{T}} U S S^{\mathrm{T}} U^{\mathrm{T}} (\vec{h_i} - \vec{h_j})} \quad (3)
$$

- 初始实验：对 W 矩阵进行了五次替换（证明直接替换会丢失准确度，于是要进行下一步“主成分提取”）

Methods	$Orig$	US	U	UV^{T}	QD
rank-1	63.6	63.6	61.7	61.7	61.6
mAP	39.0	39.0	37.1	37.1	37.3

- 成分贡献度：论文最后保留贡献度大于 0.0072 的部分

$$
G = W^{\mathrm{T}} W = \begin{bmatrix} \vec{w_1}^{\mathrm{T}}\vec{w_1} & \vec{w_1}^{\mathrm{T}}\vec{w_2} & \cdots & \vec{w_1}^{\mathrm{T}}\vec{w_k} \\ \vec{w_2}^{\mathrm{T}}\vec{w_1} & \vec{w_2}^{\mathrm{T}}\vec{w_2} & \cdots & \vec{w_2}^{\mathrm{T}}\vec{w_k} \\ \vec{w_k}^{\mathrm{T}}\vec{w_1} & \vec{w_k}^{\mathrm{T}}\vec{w_2} & \cdots & \vec{w_k}^{\mathrm{T}}\vec{w_k} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1k} \\ g_{21} & g_{22} & \cdots & g_{2k} \\ g_{k1} & g_{k2} & \cdots & g_{kk} \end{bmatrix}
$$

$$
S(W) = \frac{\sum_{i=1}^{k} g_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} |g_{ij}|}.
$$

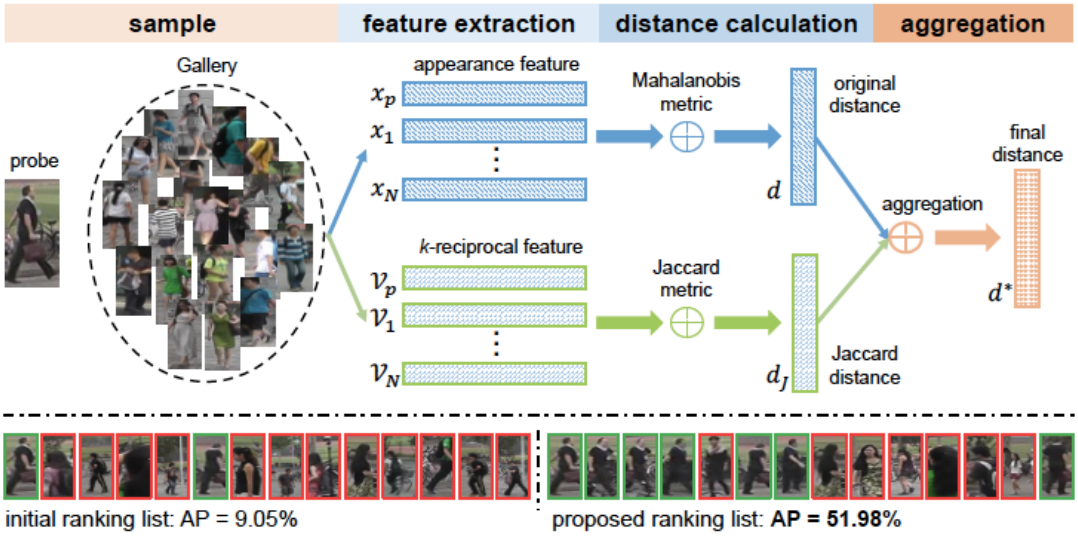
- Result：以 CaffeNet 和 ResNet-50 为基准

Models & Features	dim	Market-1501				CUHK03				DukeMTMC-reID			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Baseline(C) FC6	4096	55.3	75.8	81.9	30.4	38.6	66.4	76.8	45.0	46.9	63.2	69.2	28.3
Baseline(C) FC7	4096	54.6	75.5	81.3	30.3	42.2	70.2	80.4	48.6	45.9	62.0	69.7	27.1
SVDNet(C) FC6	4096	**80.5**	**91.7**	**94.7**	**55.9**	**68.5**	**90.2**	**95.0**	**73.3**	**67.6**	**80.5**	**85.7**	**45.8**
SVDNet(C) FC7	1024	79.0	91.3	94.2	54.6	66.0	89.4	93.8	71.1	66.7	80.5	85.1	44.4
Baseline(R) Pool5	2048	73.8	87.6	91.3	47.9	66.2	87.2	93.2	71.1	65.5	78.5	82.5	44.1
Baseline(R) FC	N	71.1	85.0	90.0	46.0	64.6	89.4	95.0	70.0	60.6	76.0	80.9	40.4
SVDNet(R) Pool5	2048	**82.3**	**92.3**	**95.2**	**62.1**	**81.8**	**95.2**	97.2	**84.8**	**76.7**	**86.4**	**89.9**	**56.8**
SVDNet(R) FC	1024	81.4	91.9	94.5	61.2	81.2	95.2	**98.2**	84.5	75.9	86.4	89.5	56.3

Table 2: Comparison of the proposed method with baselines. C: CaffeNet. R: ResNet-50. In ResNet Baseline, "FC" denotes the last FC layer, and its output dimension N changes with the number of training identities, *i.e.*, 751 on Market-1501, 1,160 on CUHK03 and 762 on DukeMTMC-reID. For SVDNet based on ResNet, the Eigenlayer is denoted by "FC", and its output dimension is set to 1,024.

- Re-ranking Person Re-identification with k-reciprocal Encoding (CVPR2017)

这篇文章最主要的工作就是在计算相似度的距离里面引入了 jaccard distance，另外一个工作是在搜索环节引入了 K 近邻算法，降低了搜索的计算复杂度。



- Origin distance (Mahalanobis distance): M 是个半正定矩阵，通常就是单位阵

$$d(p, g_i) = (x_p - x_{g_i})^\top \mathbf{M}(x_p - x_{g_i})$$

- K-reciprocal Nearest Neighbors

  p 最近的 K 个样本：

$$N(p,k) = \{g_1^0, g_2^0, ..., g_k^0\}, |N(p,k)| = k$$

  定义关系集：

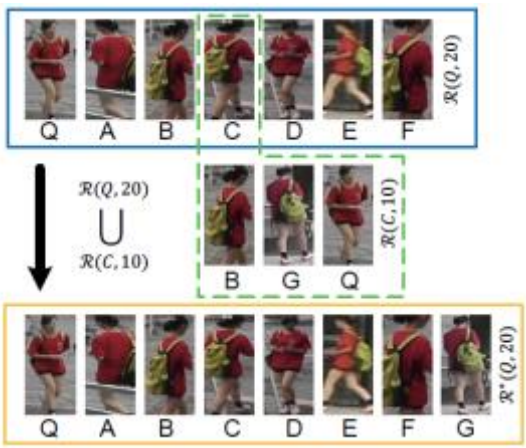
$$\mathcal{R}(p,k) = \{(g_i \in N(p,k)) \cap (p \in N(g_i,k))\}$$

  定义更加鲁棒的关系集，因为背景、角度、姿态等的变化：

$$\mathcal{R}^*(p,k) \leftarrow \mathcal{R}(p,k) \cup \mathcal{R}(q, \tfrac{1}{2}k)$$

$$s.t. \left|\mathcal{R}(p,k) \cap \mathcal{R}(q, \tfrac{1}{2}k)\right| \geqslant \tfrac{2}{3}\left|\mathcal{R}(q, \tfrac{1}{2}k)\right|,$$

$$\forall q \in \mathcal{R}(p,k)$$



- Jaccard distance

  原始的公式

$$d_J(p,g_i) = 1 - \frac{|\mathcal{R}^*(p,k) \cap \mathcal{R}^*(g_i,k)|}{|\mathcal{R}^*(p,k) \cup \mathcal{R}^*(g_i,k)|}$$

  但是这个公式是很耗时并且重复计算的，所以要改进，先定义一个向量，有 hard 模式和 soft 模式，hard 模式不考虑内容相似度，soft 模式考虑内容相似度，所以论文最终使用 soft 模式

  k-reciprocal feature: $\mathcal{V}_p = [\mathcal{V}_{p,g_1}, \mathcal{V}_{p,g_2}, ..., \mathcal{V}_{p,g_N}]$

$$\mathcal{V}_{p,g_i} = \begin{cases} 1 & \text{if } g_i \in \mathcal{R}^*(p,k) \\ 0 & \text{otherwise.} \end{cases} \qquad \mathcal{V}_{p,g_i} = \begin{cases} e^{-d(p,g_i)} & \text{if } g_i \in \mathcal{R}^*(p,k) \\ 0 & \text{otherwise.} \end{cases}$$

  之后 Jaccard distance 近似为：

$$|\mathcal{R}^*(p,k) \cap \mathcal{R}^*(g_i,k)| = \|\min(\mathcal{V}_p, \mathcal{V}_{g_i})\|_1$$
$$|\mathcal{R}^*(p,k) \cup \mathcal{R}^*(g_i,k)| = \|\max(\mathcal{V}_p, \mathcal{V}_{g_i})\|_1$$

$$d_J(p,g_i) = 1 - \frac{\sum_{j=1}^{N} \min(\mathcal{V}_{p,g_j}, \mathcal{V}_{g_i,g_j})}{\sum_{j=1}^{N} \max(\mathcal{V}_{p,g_j}, \mathcal{V}_{g_i,g_j})}$$

  进一步化简，可以用子集中的 feature 平均来近似作为 probe p 的 feature

$$\mathcal{V}_p = \frac{1}{|N(p,k)|} \sum_{g_i \in N(p,k)} \mathcal{V}_{g_i}$$

- Final distance

$$d^*(p,g_i) = (1-\lambda) d_J(p,g_i) + \lambda d(p,g_i)$$

- 在 KISSME 和 XQDA 度量方法，在 LOMO 和 IDE 特征在 CaffeNet 和 ResNet-50 上都做了对比实验，在 Market-1501，CUHK03，Mars，PRW 数据集上准确度均有提高

Method	Rank 1	mAP
LOMO + KISSME	30.86	15.36
LOMO + KISSME + Ours	31.31	22.38
LOMO + XQDA [23]	31.82	17.00
LOMO + XQDA + Ours	33.99	23.20
IDE (C) [53]	61.72	41.17
IDE (C) + AQE [5]	61.83	47.02
IDE (C) + CDM [14]	62.05	44.23
IDE (C) + Ours	62.78	51.47
IDE (C) + KISSME	65.25	44.83
IDE (C) + KISSME + Ours	66.87	56.18
IDE (C) + XQDA	65.05	46.87
IDE (C) + XQDA + Ours	67.78	57.98
IDE (R) [53]	62.73	44.07
IDE (R) + AQE [5]	63.74	49.14
IDE (R) + CDM [14]	64.11	47.68
IDE (R) + Ours	65.61	57.94
IDE (R) + KISSME	70.35	53.27
IDE (R) + KISSME + Ours	72.32	67.29
IDE (R) + XQDA	70.51	55.12
IDE (R) + XQDA + Ours	**73.94**	**68.45**