局部图像匹配的深度学习方法

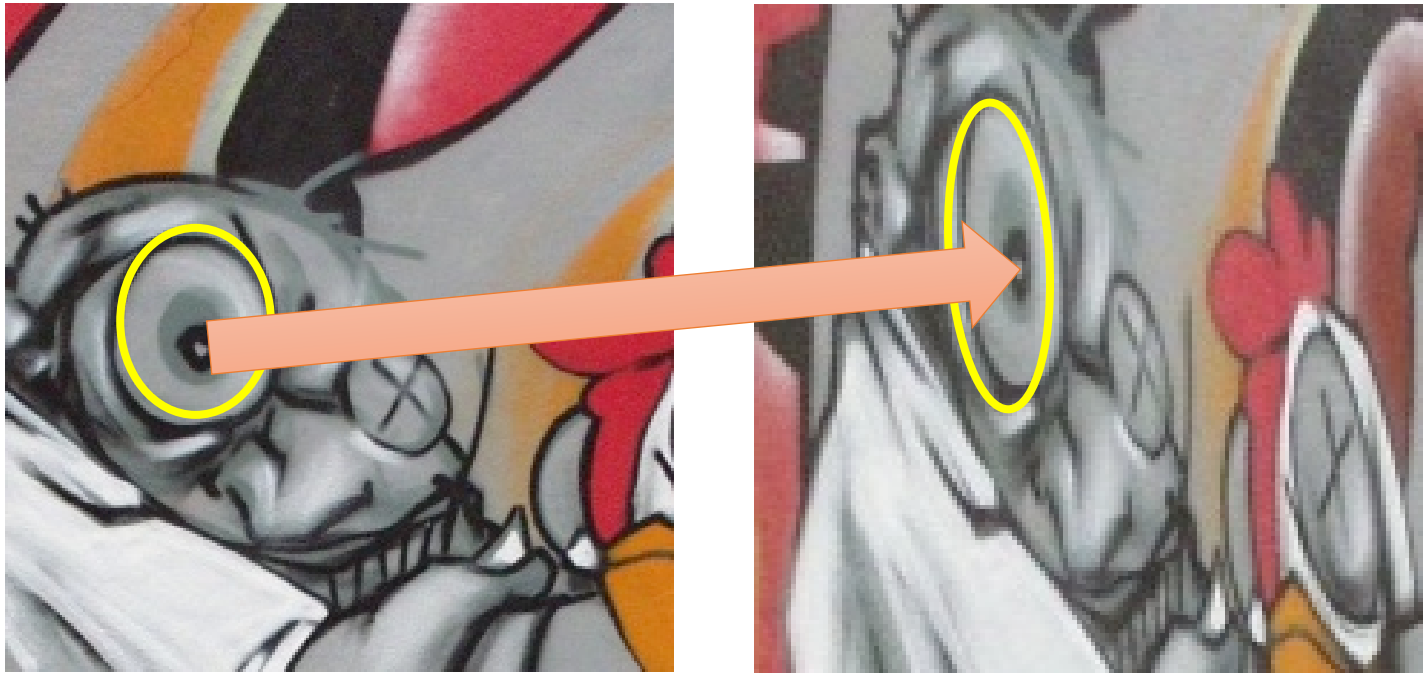
樊彬

模式识别国家重点实验室

中国科学院自动化研究所

2017. 6. 17

局部图像匹配：是什么？



解决从哪到哪的问题！

局部图像匹配：能干什么？

场景深度计算

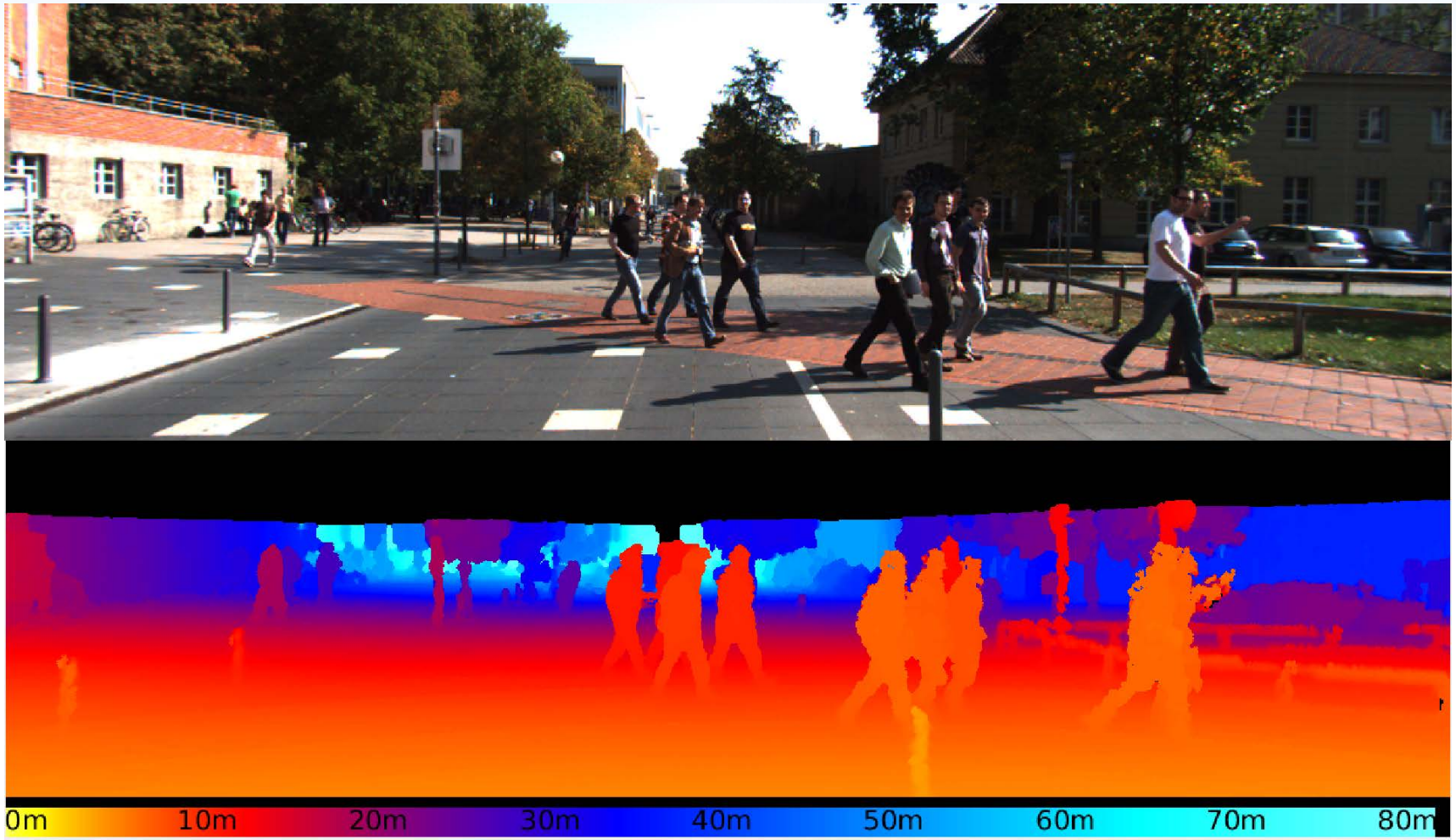
物体识别

图像拼接

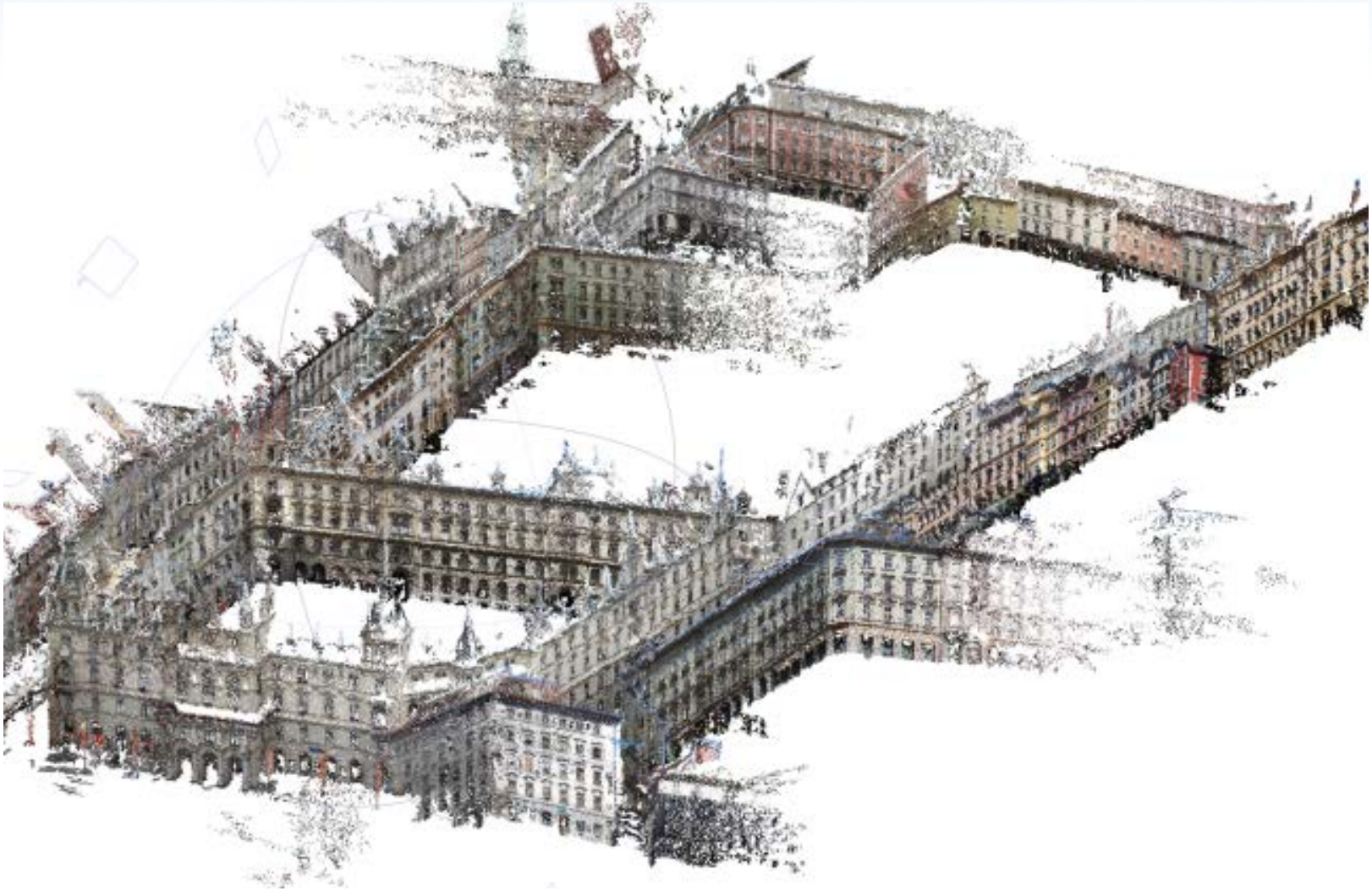
三维重建

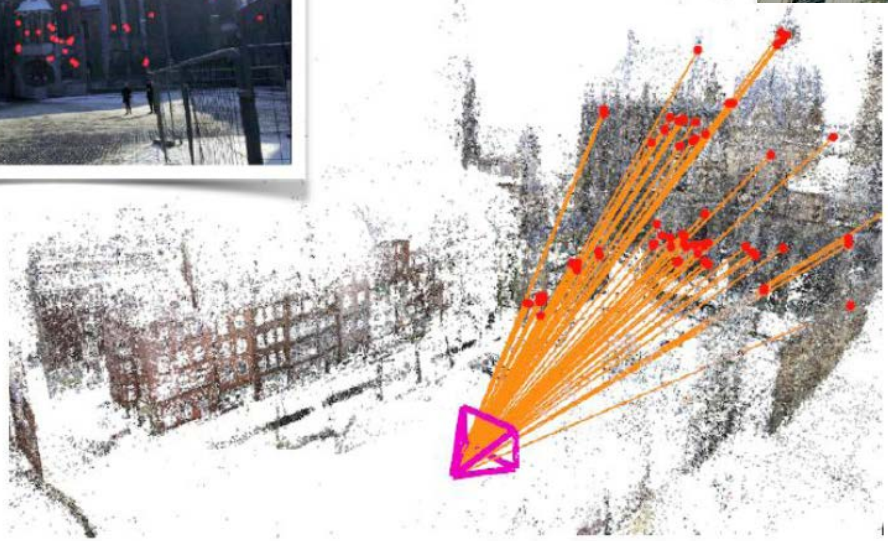
视觉定位

……


0m
10m
20m
30m
40m
50m
60m
70m
80m

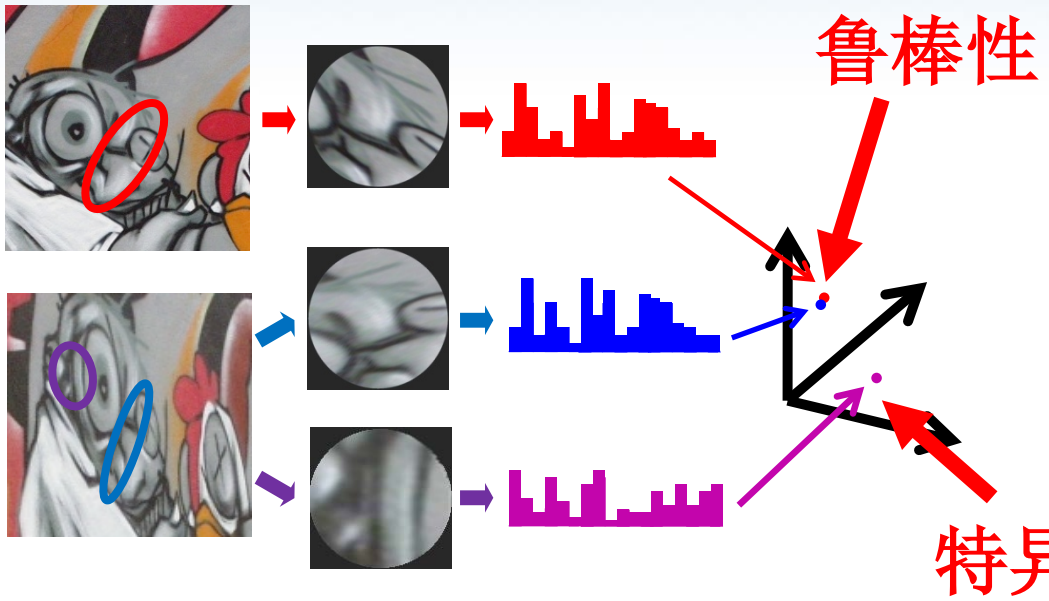
Database






模式识别国家重点实验室
National Laboratory of Pattern Recognition
NLPR
中国科学院自动化研究所
复杂系统管理与控制国家重点实验室

局部图像匹配：怎么做？



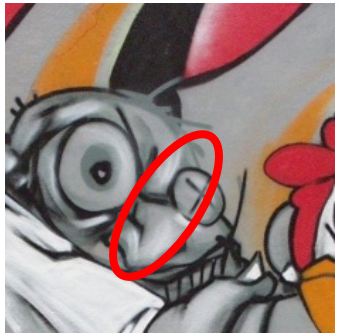
知识驱动的方法：2012年之前的主流方法，较为统一的框架。

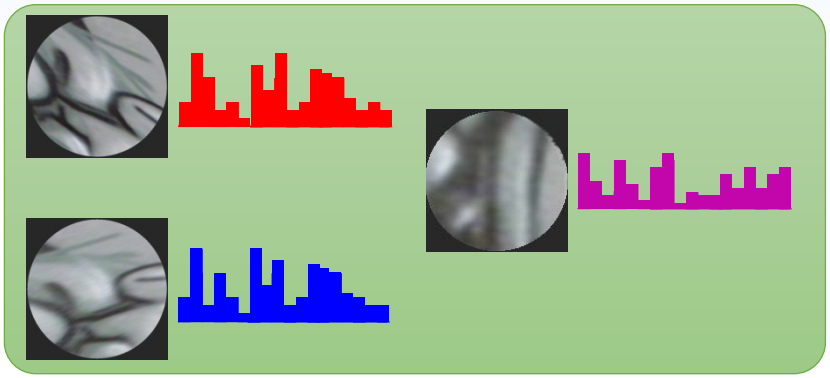
底层特征 ＋ 特征汇聚 ＋ 归一化 ＝ 描述子

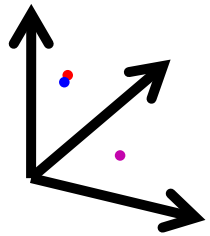
优点：通用性强，可解释性强

缺点：依赖设计者的知识，难度大，设计周期长

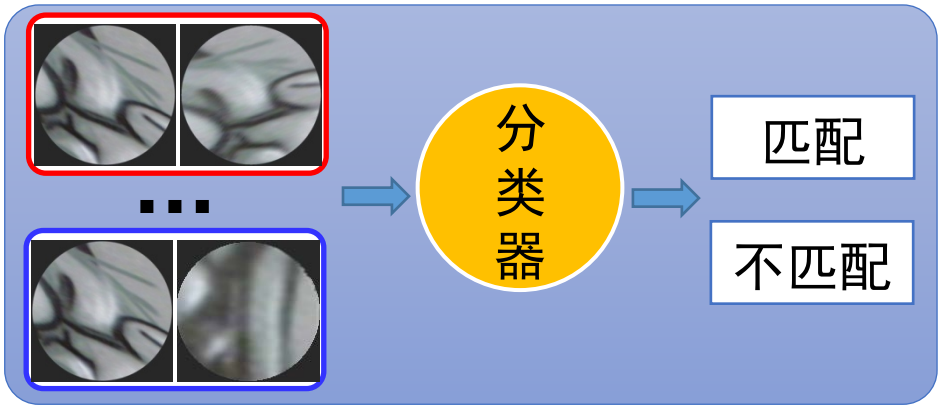
局部图像匹配：怎么做？







特征描述子学习



分类器学习

数据驱动的方法：近几年的主流方法，利用CNN网络直接匹配或者学习特征描述子。

局部图像匹配：方法

- 基于CNN的局部图像匹配

自带度量网络，直接输出是否匹配，但是特征描述与度量耦合在一起，使用不方便、计算量大，应用比较有限

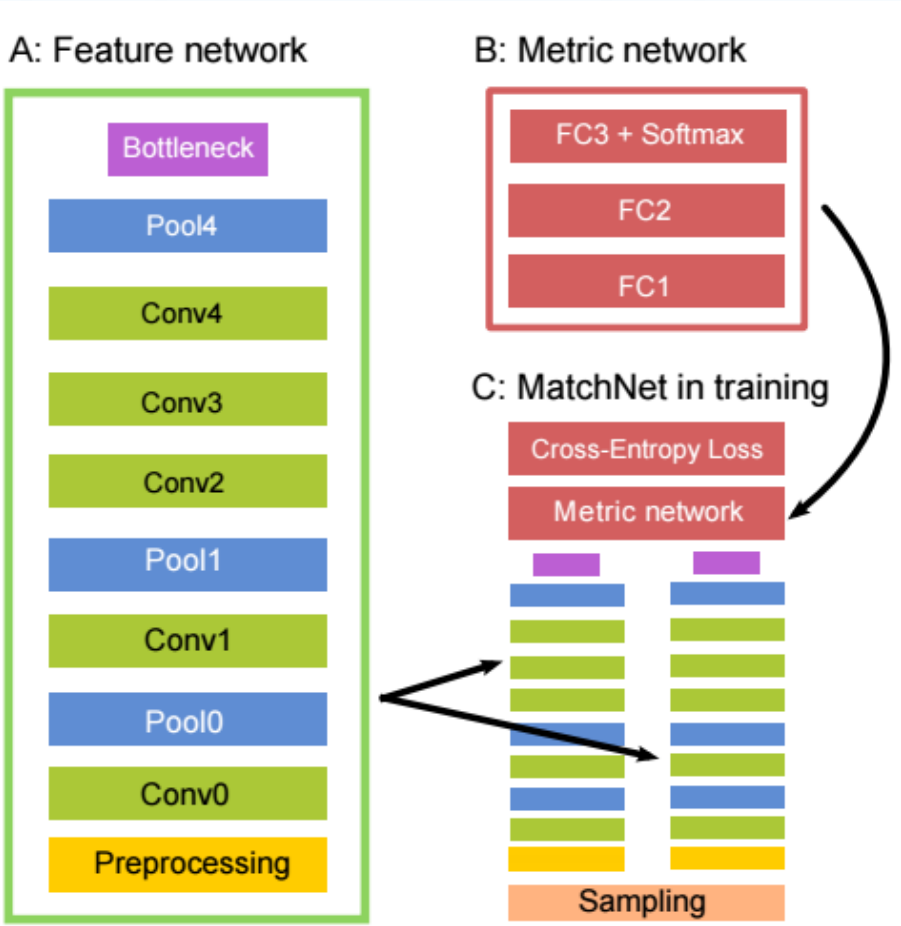
MatchNet[CVPR 2015]，DeepCompare[CVPR 2015]，GLoss Net[CVPR 2016]

- 基于CNN的局部图像特征描述子

网络输出即特征描述子，直接替代传统方法使用，应用广

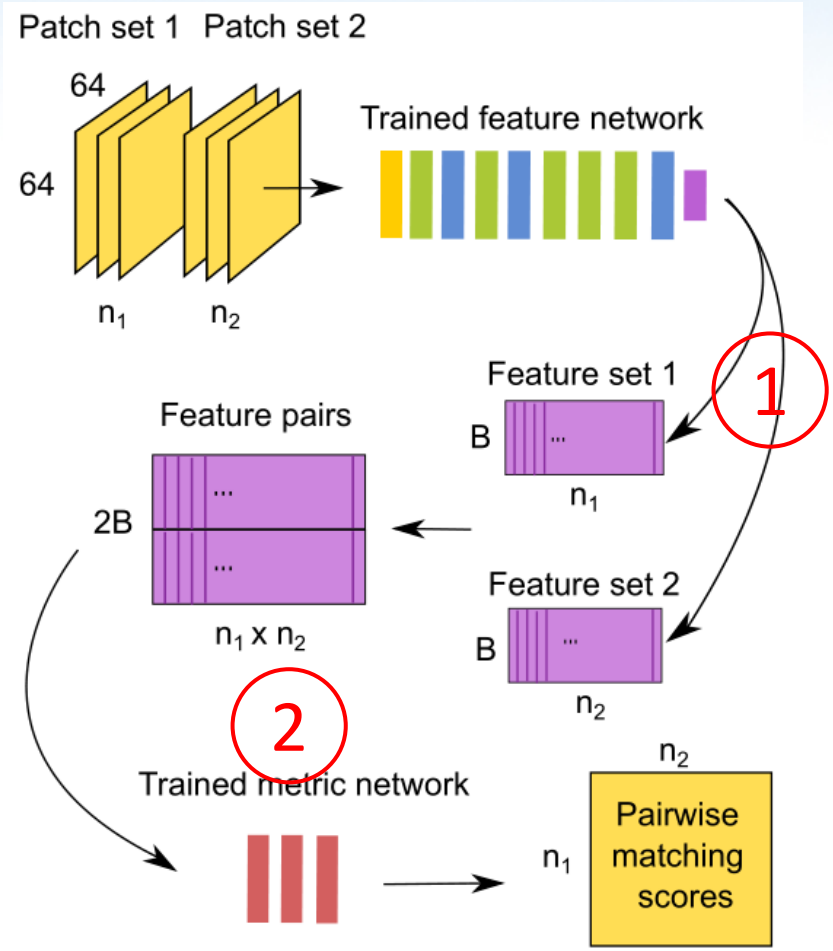
DeepDesc[ICCV 2015]，TFeat[BMVC 2016]，L2-Net[CVPR 2017]

MatchNet[CVPR 2015]



- 学习特征描述子的同时也学习了对应的度量
- 4个卷积层+1个全连接层 = 特征描述子网络
- 3个全连接层 + softmax = 距离度量网络
- 匹配问题->分类问题，交叉熵损失函数
- Bottleneck层的输出可作为特征描述子

X. Han, T. Leung, Y. Jia, R. Sukthankar, and A.C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. CVPR, 2015.



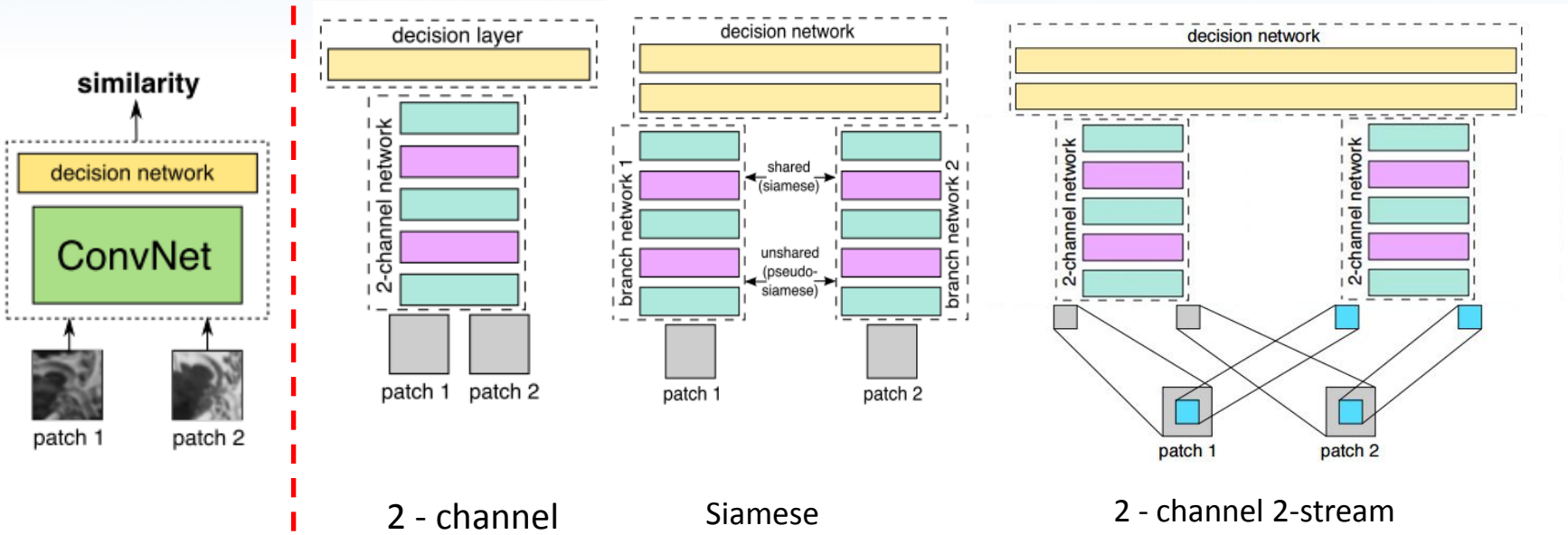
网络使用时（2步操作）

1. 先使用特征描述子网络对所有的patch提取描述子；

2. 成对联和描述子，并输入度量网络，计算匹配分数

相对于直接输入成对的patch计算匹配分数，可减少描述子网络的计算量。

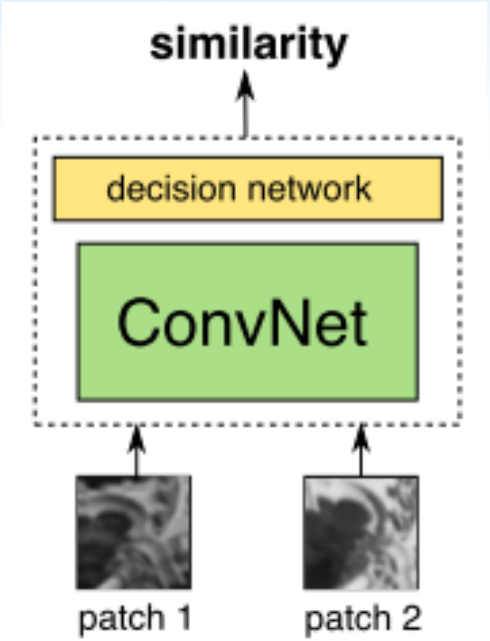
DeepCompare[CVPR 2015]

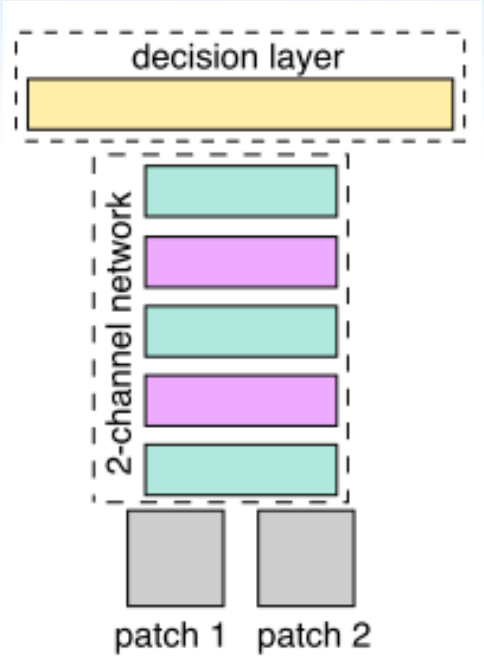


2 - channel　　Siamese　　2 - channel 2-stream

- 将局部图像匹配问题建模为图像对相似函数的学习问题
- 利用CNN网络与全连接度量网络相结合的结构来拟合相似函数
- 设计了多种CNN网络结构接收一对局部图像作为输入
- 在标准的Patch匹配数据集上，对这些网络的性能进行了实验分析

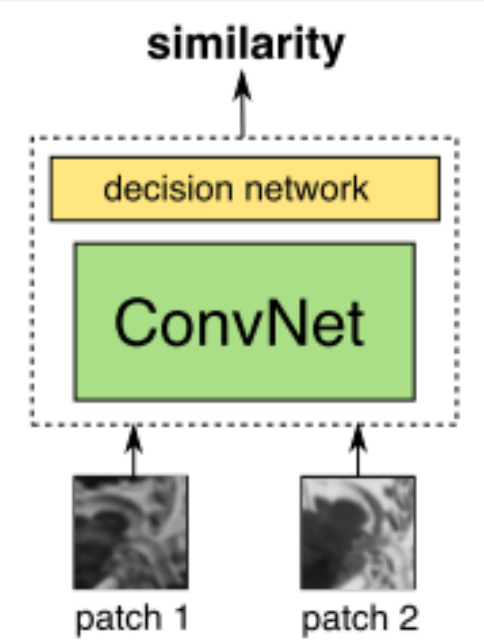
S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. CVPR, 2015.

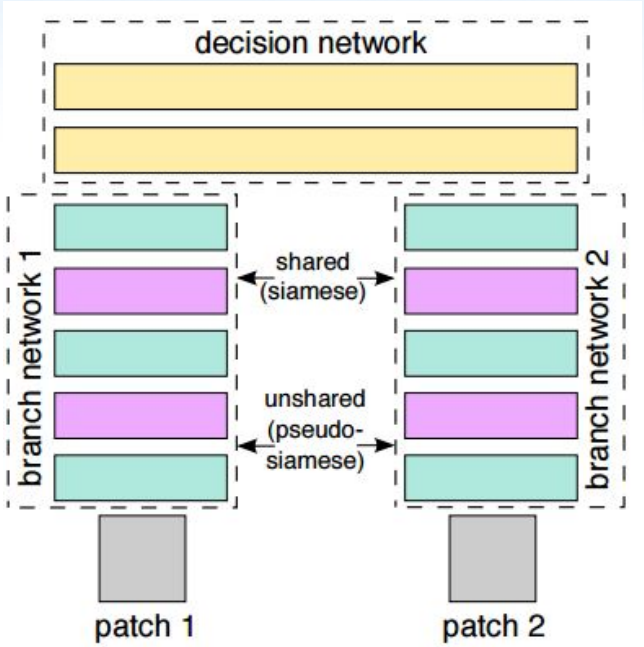




2 channel 结构（2ch）

- 输入的一对待匹配patch，直接叠加，作为CNN网络输入图像的2个channel
- CNN网络之后接全连接层作为度量网络，最终输出是否匹配
- 测试非常费时，需要穷举所有可能的组合！

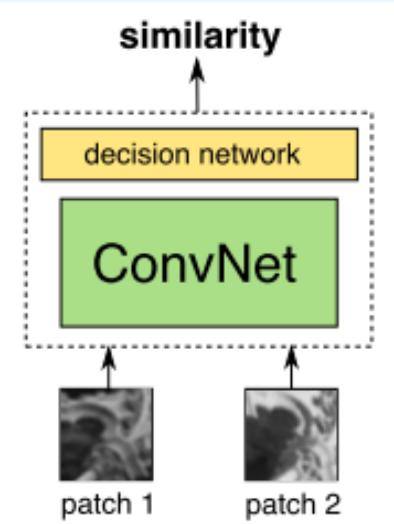


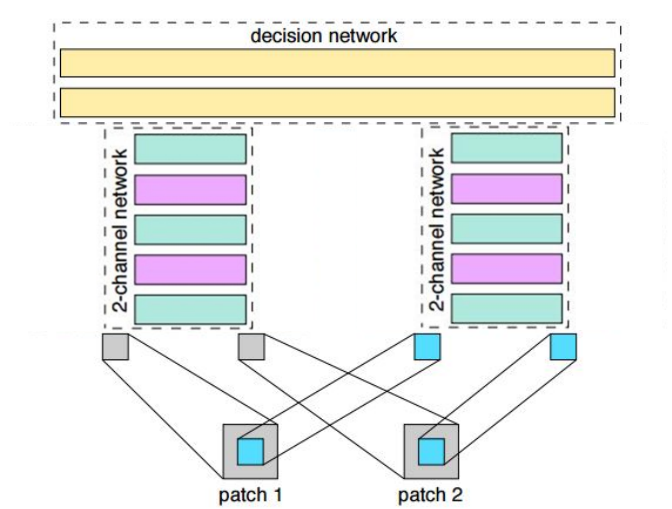


Siamese 结构

MatchNet

- 两个相同网络接受输入的patch对
- 两个CNN网络输出（提取的特征描述子）串连在一起，后接全连接层作为度量网络，最终输出是否匹配
- 和MatchNet一样，测试时可以采用2步策略来加快匹配速度





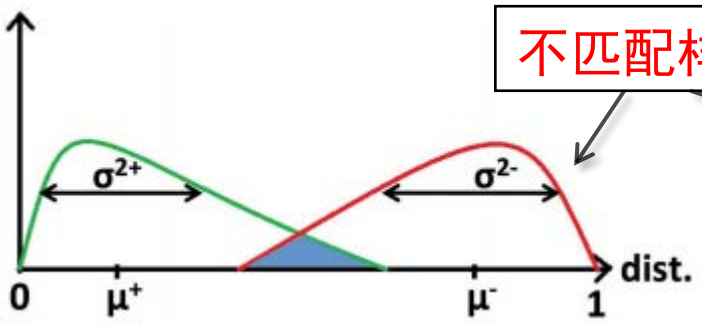
双通道、双数据流结构（2ch-2stream）

- 将一对patch分成两组数据流，一组为patch的中心部分，一组为所有的patch，这个结构也称为Center-Surround（CS结构）
- 每组数据流经过一个CNN网络结构（如2ch结构）
- 两组数据流的CNN输出串在一起，经过度量网络，输出是否匹配

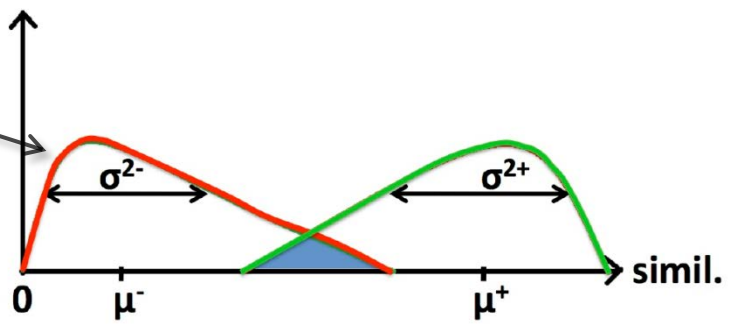
DeepCompare：总结

- 2 channel的最后匹配性能要优于Siamese网络结构
- Siamese网络结构需要和度量网络一起使用，否则性能下降
- CS结构能够有效提升网络性能
- 2 channel的结构不能输出每个patch相应的描述子
- 2 channel的结构进行匹配时计算复杂度高

GLoss-Net[CVPR 2016]



特征描述子的距离分布



特征描述子的相似性分布

核心思想：减少全局误识率（图中蓝色部分）

■ 学习目标（全局损失，Global Loss）

1. 最小化正负样本分布的方差
2. 优化正负样本均值之间的margin

V. Kumar B G, G. Carneiro, and I. Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. CVPR, 2016.

Triplet样本输入：$(x_i, x_i^+, x_i^-), i = 1, \cdots, N$，$(x_i, x_i^+)$匹配样本，$(x_i, x_i^-)$不匹配样本

①GLoss优化目标（针对特征描述子，L2距离）：

$$J_1 = \delta_+^2 + \delta_-^2 + \lambda \max(0, \mu_+ - \mu_- + t)$$

$$d_i^+ = \left\| f(x_i) - f(x_i^+) \right\|, d_i^- = \left\| f(x_i) - f(x_i^-) \right\|$$

δ_+^2, μ_+为d_i^+的方差和均值，δ_-^2, μ_-为d_i^-的方差和均值

②Triplet优化目标：

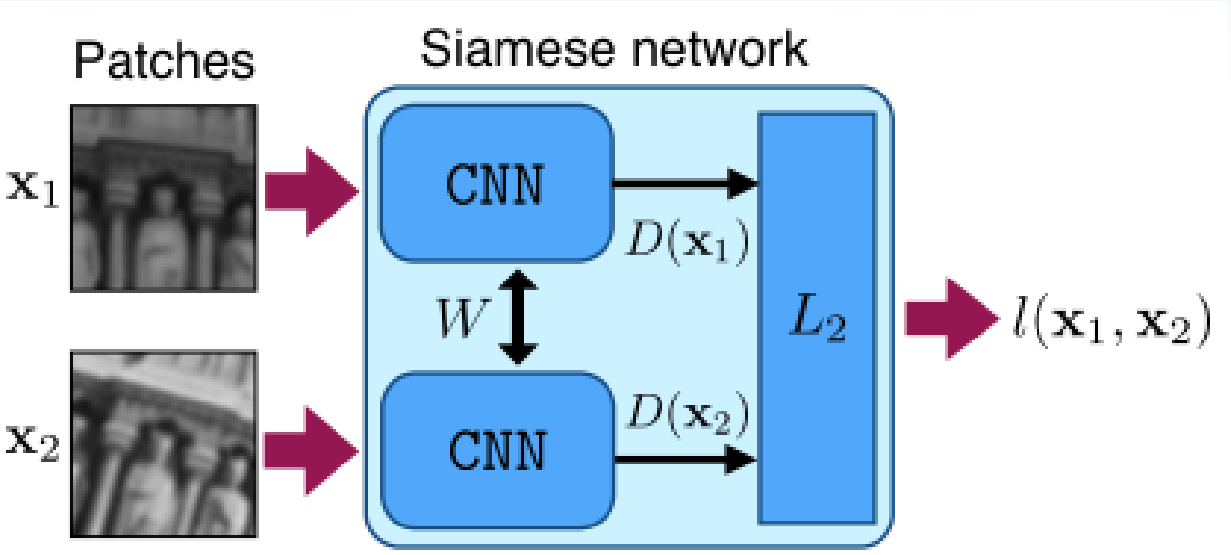
$$J_2 = \max(0, 1 - \frac{d_i^-}{d_i^+ + m})$$

③GLoss优化目标（基于网络的相似性输出）：

$$J_3 = \delta_+^2 + \delta_-^2 + \lambda \max(0, \mu_- - \mu_+ + m), d_i^+ = g(x_i, x_i^+), d_i^- = g(x_i, x_i^-)$$

- 有度量网络：SNet-GLoss（③）,CS SNet-Gloss（③, 2ch-2stream）
- 无度量网络 ：TNet-TGLoss（② + ①）,TNet-Tloss（②）

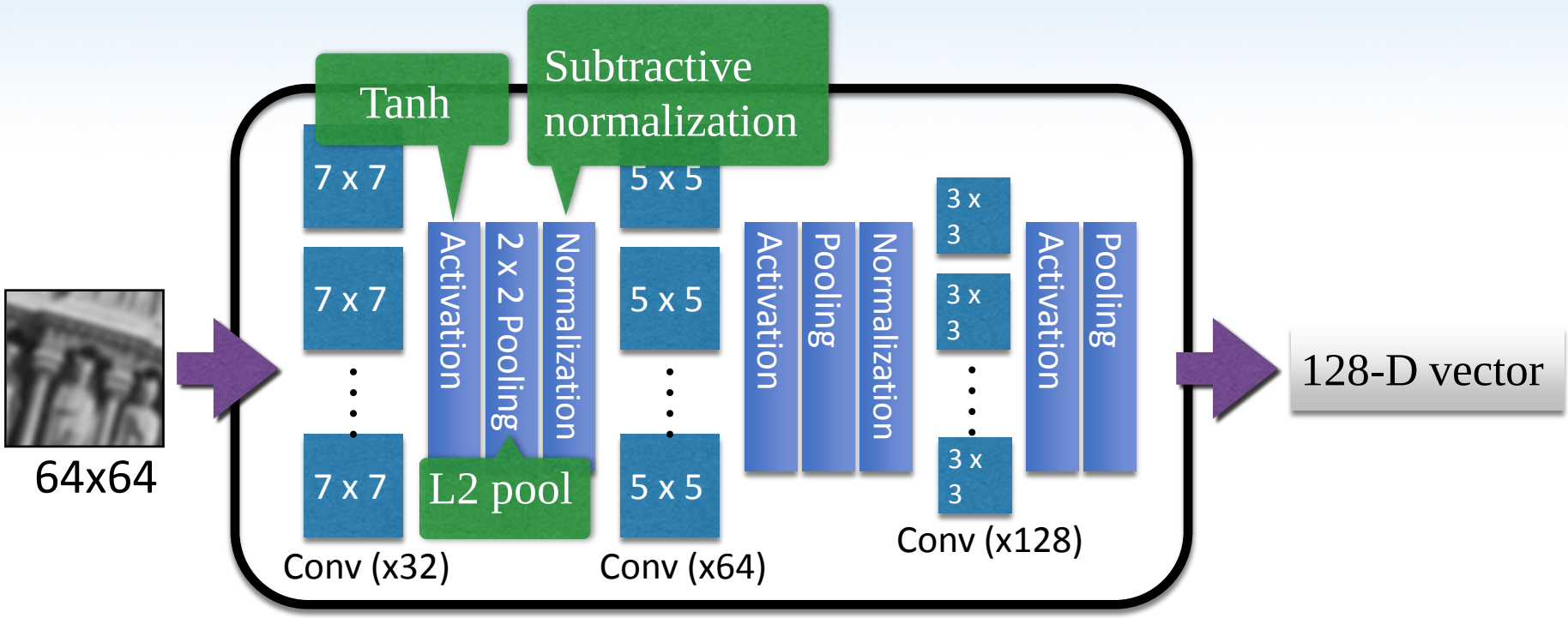
DeepDesc[ICCV 2015]



- 无度量网络，使用欧式距离，可以直接代替SIFT使用
- 网络损失函数：最小化pairwise的hinge loss

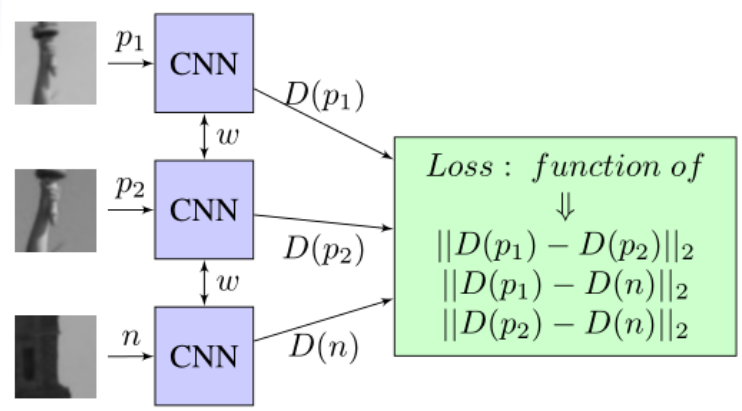
$$l(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \|D(\mathbf{x}_1) - D(\mathbf{x}_2)\|_2, & p_1 = p_2 \\ \max(0, C - \|D(\mathbf{x}_1) - D(\mathbf{x}_2)\|_2), & p_1 \neq p_2 \end{cases}$$

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. ICCV, 2015.



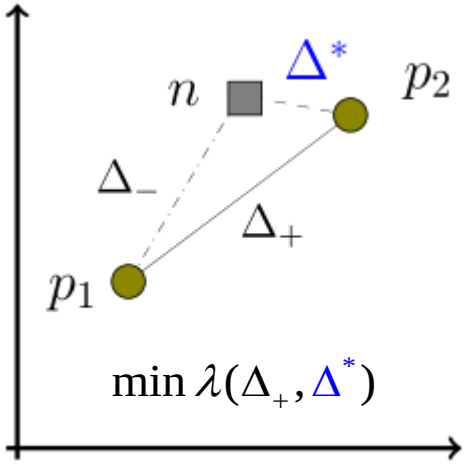
- 只有3个卷积层，没有全连接层
- 训练过程中，极度的样本不平衡，使用hard sample mining技术（核心贡献，获得好结果的关键！）
- Hard sample mining：训练网络时，Kf个样本，取其中Kb个训练loss最大的样本用于误差反传：Kf/Kb = 8

TFeat[BMVC 2016]



CNN Structure

Layer #	Description
1	Spatial Convolution(7,7) $\rightarrow$ 32
2	Tanh
2	MaxPooling(2,2)
3	Spatial Convolution(6,6) $\rightarrow$ 64
4	Tanh
5	Linear $\rightarrow \{128, 256\}$
6	Tanh



核心思想：三元组内最小的负样本距离应该大于正样本距离

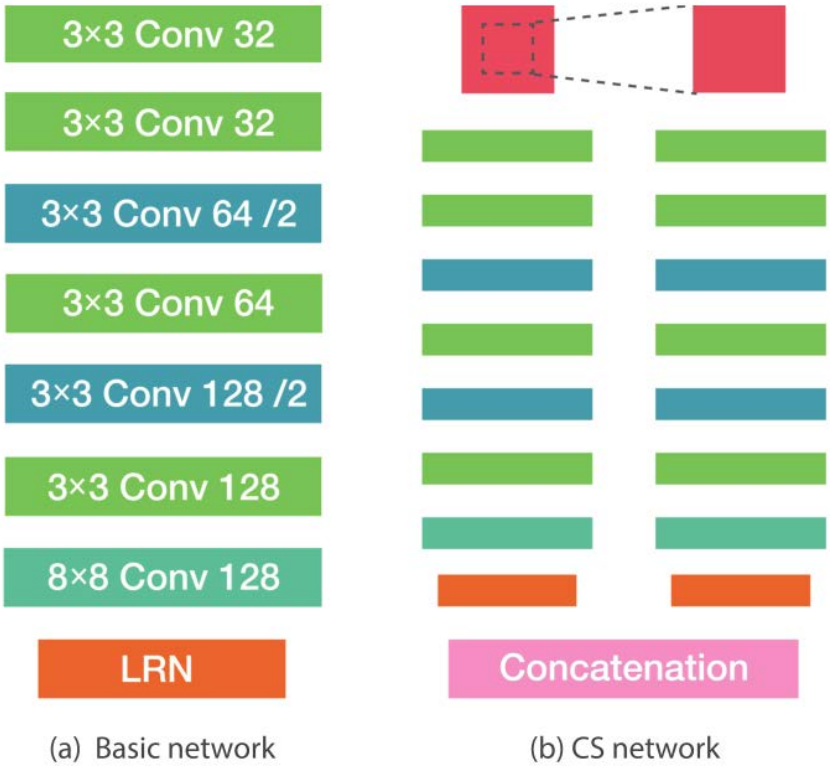
Ranking-based: $\lambda(\Delta_+, \Delta^*) = \max(0, \mu + \Delta_+ - \Delta^*)$

Ratio-based: $\lambda(\Delta_+, \Delta^*) = \left(\frac{\Delta_+}{\Delta_+ + \Delta^*}\right)^2$

V. Balntas, E. Riba, D. Ponsa and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. BMVC, 2016.

L2-Net[CVPR 2017]

网络（L2-Net）结构



- 每个卷积层之后接一个批归一化层（Batch Normalization）
- 输出之前连接Local Response Normalization层进行归一化
- 输入32x32大小patch，输出128维向量
- CS L2-Net：两个网络分别处理整个patch和中心部分，网络输出串联

Yurun Tian, Bin Fan, and Fuchao Wu. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. CVPR, 2017.

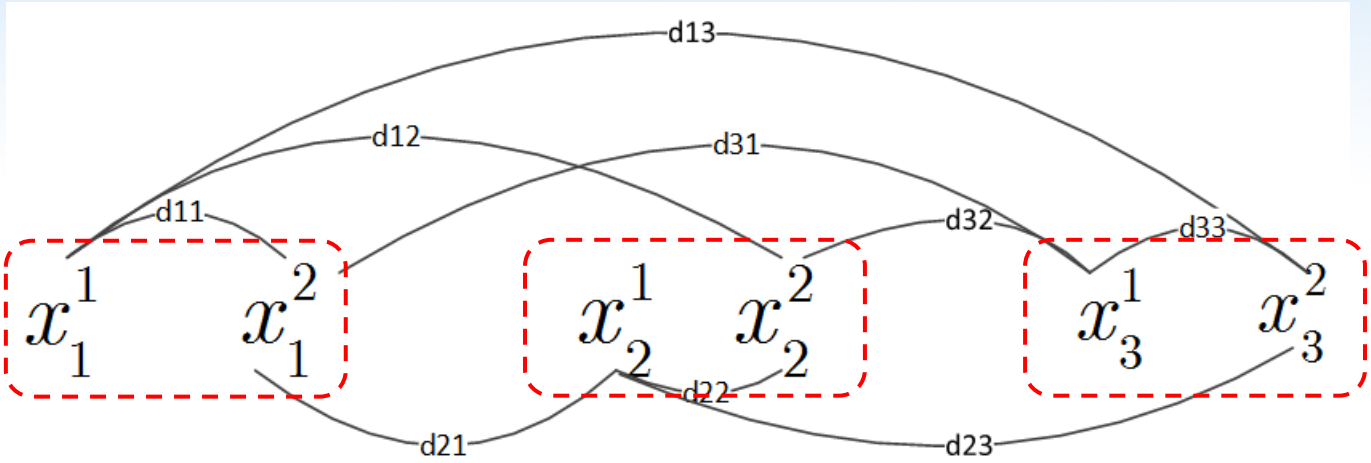
核心思想：

1. 渐进式采样策略：使得网络训练时在一个batch里可以访问大量的负样本、同时访问少量正样本，符合匹配的应用特性。

2. 将相对距离最小引入网络目标函数，符合匹配特性。

3. 将网络中间层的监督信息引入网络目标函数，提高泛化能力。

4. 在目标函数中对生成描述子的冗余性进行约束，减少过拟合。



$$D=\begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix}$$

渐进采样：在每个训练batch中，从每对匹配样本中随机抽取一个组成若干对不匹配样本，增加负样本数量。

优势：相比之前的pairwise，triplet的样本形式，可以利用更多的负样本信息。

L2-Net：学习目标

1. 欧式空间匹配样本距离相对最小
2. 输出特征冗余度低
3. 匹配样本的中间层特征相似性相对最大

学习目标一：匹配样本距离相对最小

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1i} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2i} & \cdots & d_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{j1} & d_{j2} & \cdots & d_{ji} & \cdots & d_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{p1} & d_{p2} & \cdots & d_{pi} & \cdots & d_{pp} \end{bmatrix}$$

$$\mathbf{Y_s} = \left[\mathbf{y}_1^s, \cdots, \mathbf{y}_i^s, \cdots, \mathbf{y}_p^s\right]_{q\times p}, (s = 1, 2)$$

$$\mathbf{D} = \sqrt{\mathbf{2}\left(\mathbf{1} - \mathbf{Y}_\mathbf{1}^\mathbf{T}\mathbf{Y_2}\right)}$$

描述子的欧式距离

理想情况下，d_{ii}应该是第i行中最小的，同时也是第i列中最小的。

$$s_i^c = \exp(2 - d_{ii}) / \sum_m \exp(2 - d_{mi})$$

$$s_i^r = \exp(2 - d_{ii}) / \sum_m \exp(2 - d_{im})$$

$$E_1 = -\frac{1}{2}\left(\sum_i \log s_i^r + \sum_i \log s_i^c\right)$$

学习目标二：描述子不同维度之间相关度低

$$\mathbf{Y_s} = \left[\mathbf{y}_1^s, \cdots, \mathbf{y}_i^s, \cdots, \mathbf{y}_p^s\right]_{q\times p}, (s = 1, 2)$$

$$\mathrm{Y}_s^T = [b_1^s, \cdots, b_i^s, \cdots, b_q^s]$$

输出向量第i维和第j维的相关性：

$$r_{ij}^s = \frac{(\mathbf{b}_i^s - \bar{b}_i^s)^T(\mathbf{b}_j^s - \bar{b}_j^s)}{\sqrt{(\mathbf{b}_i^s - \bar{b}_i^s)^T(\mathbf{b}_i^s - \bar{b}_i^s)}\sqrt{(\mathbf{b}_j^s - \bar{b}_j^s)^T(\mathbf{b}_j - \bar{b}_j^s)}}$$

最小化描述子不同维度之间的相关性：

$$E_2 = \frac{1}{2}\left(\sum_{i\neq j}(r_{ij}^1)^2 + \sum_{i\neq j}(r_{ij}^2)^2\right)$$

学习目标三：匹配样本的中间层特征相似性相对最大

对于中间层的feature map：

$$\mathrm{F}_s = [f_1^s, f_2^s, \cdots, f_p^s], s = 1, 2$$

Feature map之间的相似性（内积）：

$$\mathbf{G} = (\mathbf{F}_1)^{\mathbf{T}} \mathbf{F}_2$$

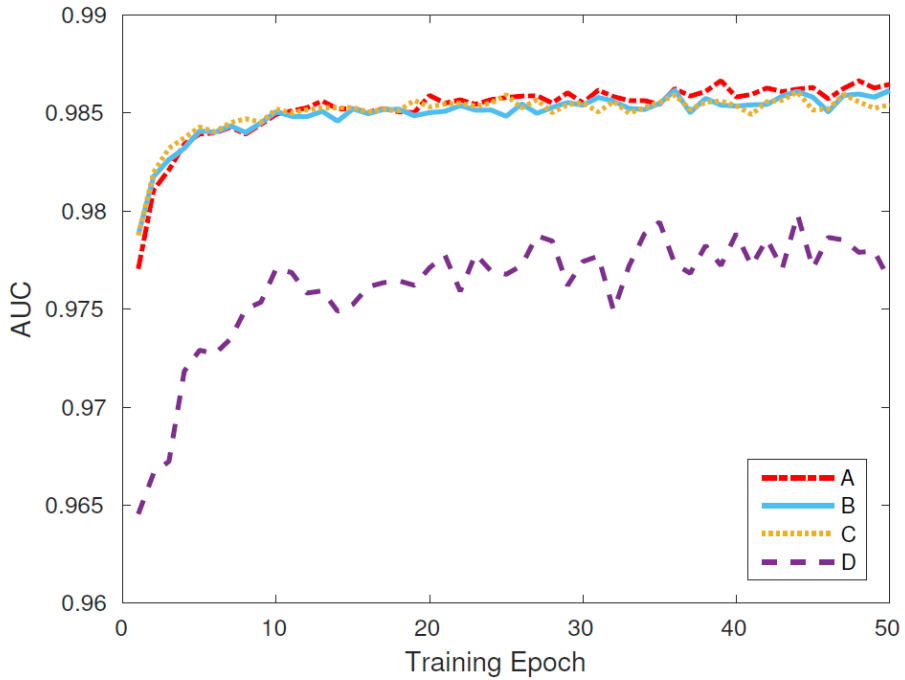
理想情况下，匹配样本的相似性应该最大，即：g_{ii}应该是第i行中最大的，同时也是第i列中最大的。

$$v_i^r = \exp(g_{ii}) / \sum_m \exp(g_{im})$$

$$v_i^c = \exp(g_{ii}) / \sum_m \exp(g_{mi})$$

$$E_3 = -\frac{1}{2}(\sum_i \log v_i^r + \sum_i \log v_i^c)$$

学习目标三：匹配样本的中间层特征相似性相对最大



A: E1+E2+E3（第一和最后一个BN层之后）

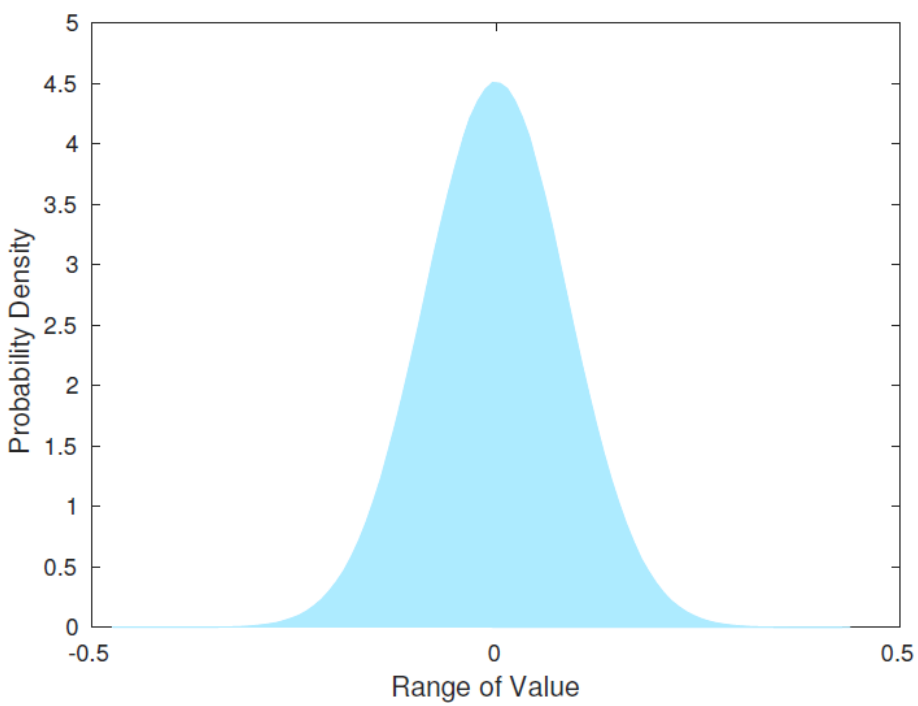
D: E1+E2+E3（所有BN层之后）

B: E1+E2

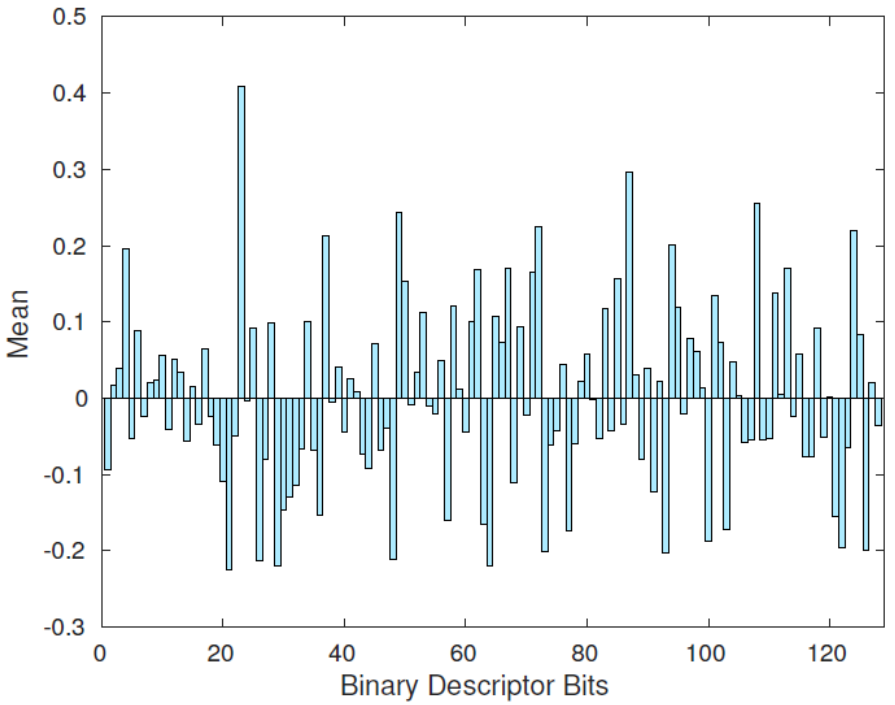
C: 对BN参数学习

注意：中间层监督信息不能滥用，仅在第一层和最后一层使用该监督信息。

Binary L2-Net



L2-Net的输出值分布



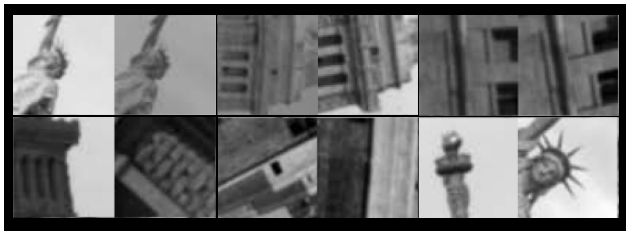
Binary L2-Net每一位的均值

L2-Net的输出值近似服从0均值的Gaussian分布，非常符合二进制描述子的要求。我们直接对L2-Net进行符号化（sign操作）得到二进制的L2-Net。

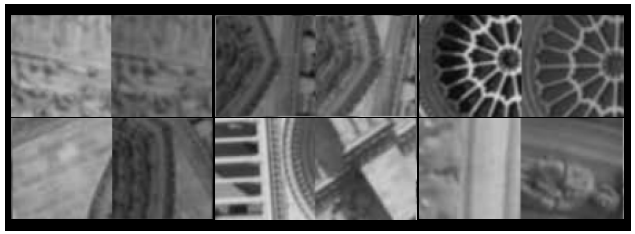
性能比较

Brown Dataset/Patch Dataset

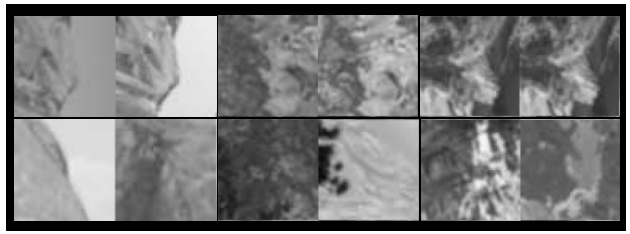
Liberty



Notre Dame



Yosemite



http://www.cs.ubc.ca/~mbrown/patchdata/patchdata.html

- 包含三个子集，每个对应一个场景
- 每个子集包含~150K个点，每个点包含2-10个局部图像
- 每个子集提供了~500K局部图像对，一半匹配的，一半不匹配的
- 广泛用于评测数据驱动的局部图像匹配方法

浮点型描述子性能比较

Training / Test	Feature Dim	Notredame Yosemite / Liberty		Liberty Yosemite / Notredame		Liberty Notredame / Yosemite		Mean
Metric Learning								
SIFT	128	29.84		22.53		27.29		26.55
MatchNet	4096	6.9	10.77	3.87	5.67	10.88	8.39	7.74
DeepCompare 2ch-2stream +	256	4.85	7.20	1.90	2.11	5.00	4.10	4.19
DeepCompare 2ch-deep +	256	4.55	7.40	2.01	2.52	4.75	4.38	4.26
SNet-GLoss +	256	6.39	8.43	1.84	2.83	6.61	5.57	5.27
CS SNet-GLoss +	384	3.69	4.91	0.77	1.14	3.09	2.67	2.71
Float Descriptors								
TNet-TGLoss +	256	9.91	13.45	3.91	5.43	10.65	9.47	8.8
TNet-TLoss +	256	10.77	13.90	4.47	5.58	11.82	10.96	9.58
TFeat	256	8.13	9.65	3.71	4.23	8.99	7.21	6.98
DeepDesc	128	10.9		4.40		5.69		6.99
L2-Net	128	3.64	5.29	1.15	1.62	4.43	3.30	3.23
L2-Net +	128	2.36	4.7	0.72	1.29	2.57	1.71	2.22
CS L2-Net	256	2.55	4.24	0.87	1.39	3.81	2.84	2.61
CS L2-Net +	256	**1.71**	**3.87**	**0.56**	**1.09**	**2.07**	**1.3**	**1.76**

“+”表示训练网络时对数据做了增强（旋转、对称翻转）

浮点型描述子性能比较

Training / Test	Feature Dim	Notredame Yosemite / Liberty		Liberty Yosemite / Notredame		Liberty Notredame / Yosemite		Mean
Metric Learning								
SIFT	128	29.84		22.53		27.29		26.55
MatchNet	4096	6.9	10.77	3.87	5.67	10.88	8.39	7.74
DeepCompare 2ch-2stream +	256	4.85	7.20	1.90	2.11	5.00	4.10	4.19
DeepCompare 2ch-deep +	256	4.55	7.40	2.01	2.52	4.75	4.38	4.26
SNet-GLoss +	256	6.39	8.43	1.84	2.83	6.61	5.57	5.27
CS SNet-GLoss +	384	3.69	4.91	0.77	1.14	3.09	2.67	2.71
Float Descriptors								
TNet-TGLoss +	256	9.91	13.45	3.91	5.43	10.65	9.47	8.8
TNet-TLoss +	256	10.77	13.90	4.47	5.58	11.82	10.96	9.58
TFeat	256	8.13	9.65	3.71	4.23	8.99	7.21	6.98
DeepDesc	128	10.9		4.40		5.69		6.99
L2-Net	128	3.64	5.29	1.15	1.62	4.43	3.30	3.23
L2-Net +	128	2.36	4.7	0.72	1.29	2.57	1.71	2.22
CS L2-Net	256	2.55	4.24	0.87	1.39	3.81	2.84	2.61
CS L2-Net +	256	**1.71**	**3.87**	**0.56**	**1.09**	**2.07**	**1.3**	**1.76**

“+”表示训练网络时对数据做了增强（旋转、对称翻转）

浮点型描述子性能比较

Training / Test	Feature Dim	Notredame Yosemite / Liberty		Liberty Yosemite / Notredame		Liberty Notredame / Yosemite		Mean
Metric Learning								
SIFT	128	29.84		22.53		27.29		26.55
MatchNet	4096	6.9	10.77	3.87	5.67	10.88	8.39	7.74
DeepCompare 2ch-2stream +	256	4.85	7.20	1.90	2.11	5.00	4.10	4.19
DeepCompare 2ch-deep +	256	4.55	7.40	2.01	2.52	4.75	4.38	4.26
SNet-GLoss +	256	6.39	8.43	1.84	2.83	6.61	5.57	5.27
CS SNet-GLoss +	384	3.69	4.91	0.77	1.14	3.09	2.67	2.71
Float Descriptors								
TNet-TGLoss +	256	9.91	13.45	3.91	5.43	10.65	9.47	8.8
TNet-TLoss +	256	10.77	13.90	4.47	5.58	11.82	10.96	9.58
TFeat	256	8.13	9.65	3.71	4.23	8.99	7.21	6.98
DeepDesc	128	10.9		4.40		5.69		6.99
L2-Net	128	3.64	5.29	1.15	1.62	4.43	3.30	3.23
L2-Net +	128	2.36	4.7	0.72	1.29	2.57	1.71	2.22
CS L2-Net	256	2.55	4.24	0.87	1.39	3.81	2.84	2.61
CS L2-Net +	256	**1.71**	**3.87**	**0.56**	**1.09**	**2.07**	**1.3**	**1.76**

“+”表示训练网络时对数据做了增强（旋转、对称翻转）

浮点型描述子性能比较

Training / Test	Feature Dim	Notredame Yosemite / Liberty		Liberty Yosemite / Notredame		Liberty Notredame / Yosemite		Mean
Metric Learning								
SIFT	128	29.84		22.53		27.29		26.55
MatchNet	4096	6.9	10.77	3.87	5.67	10.88	8.39	7.74
DeepCompare 2ch-2stream +	256	4.85	7.20	1.90	2.11	5.00	4.10	4.19
DeepCompare 2ch-deep +	256	4.55	7.40	2.01	2.52	4.75	4.38	4.26
SNet-GLoss +	256	6.39	8.43	1.84	2.83	6.61	5.57	5.27
CS SNet-GLoss +	384	3.69	4.91	0.77	1.14	3.09	2.67	2.71
Float Descriptors								
TNet-TGLoss +	256	9.91	13.45	3.91	5.43	10.65	9.47	8.8
TNet-TLoss +	256	10.77	13.90	4.47	5.58	11.82	10.96	9.58
TFeat	256	8.13	9.65	3.71	4.23	8.99	7.21	6.98
DeepDesc	128	10.9		4.40		5.69		6.99
L2-Net	128	3.64	5.29	1.15	1.62	4.43	3.30	3.23
L2-Net +	128	2.36	4.7	0.72	1.29	2.57	1.71	2.22
CS L2-Net	256	2.55	4.24	0.87	1.39	3.81	2.84	2.61
CS L2-Net +	256	**1.71**	**3.87**	**0.56**	**1.09**	**2.07**	**1.3**	**1.76**

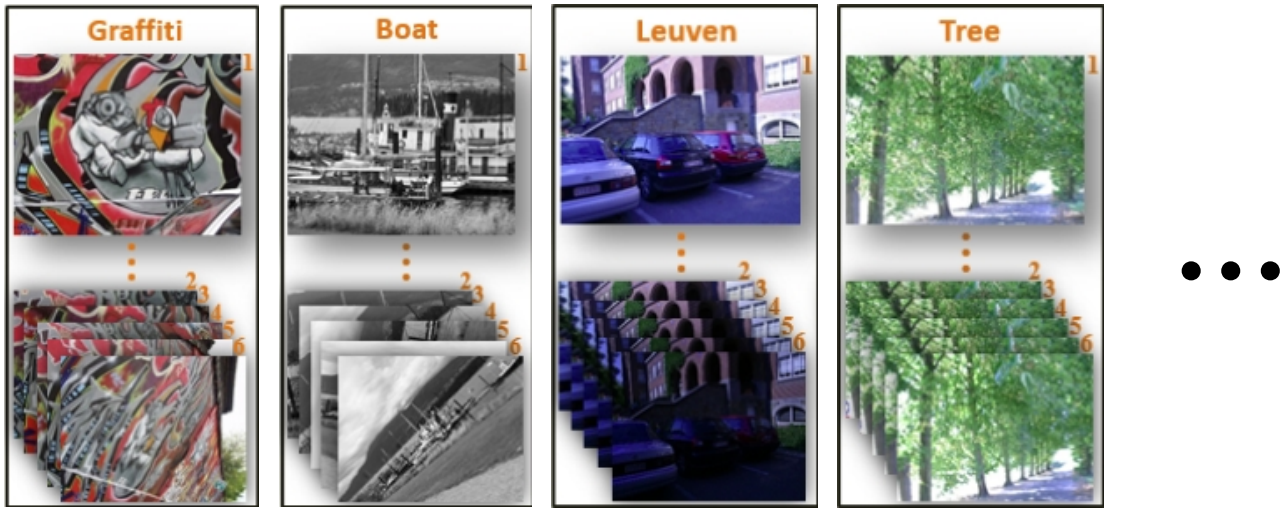
“+”表示训练网络时对数据做了增强（旋转、对称翻转）

二进制描述子性能比较

Training / Test	Feature Dim	Notredame Yosemite / Liberty		Liberty Yosemite / Notredame		Liberty Notredame / Yosemite		Mean
RFD_R	293-598	19.35	19.40	13.23	11.68	16.99	14.50	15.85
RFD_G	406-563	17.77	19.03	12.49	11.37	17.62	14.14	15.4
BinBoost	64	20.49	21.67	16.90	14.54	22.88	18.97	19.24
RMGD	1376-1600	15.09	17.42	10.15	10.86	14.46	13.82	13.63
Boixet al	1360	15.6	15.52	-	8.52	-	8.87	12.12
Binary L2-Net	128	10.3	11.71	6.37	6.76	13.5	11.57	10.03
Binary L2-Net +	128	7.44	10.29	3.81	4.31	8.81	7.45	7.01
Binary CS L2-Net	256	5.25	7.83	3.07	3.52	8.49	6.92	5.84
Binary CS L2-Net +	256	**4.01**	**6.65**	**1.9**	**2.51**	**5.61**	**4.04**	**4.12**

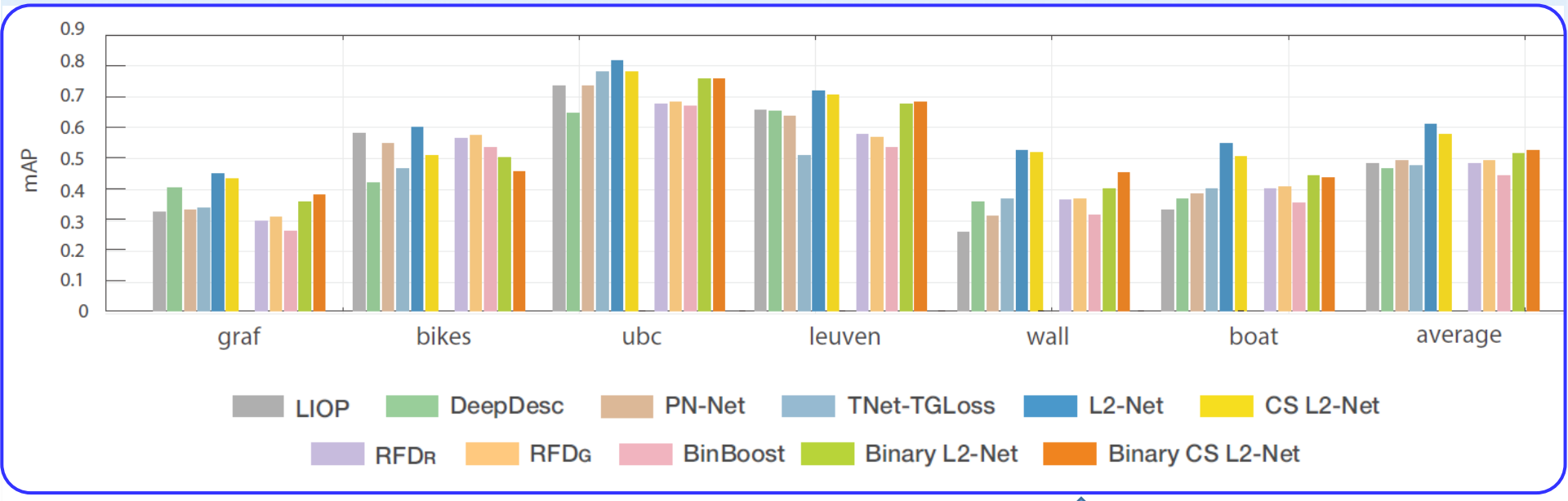
性能比较

Oxford Dataset/VGG Dataset

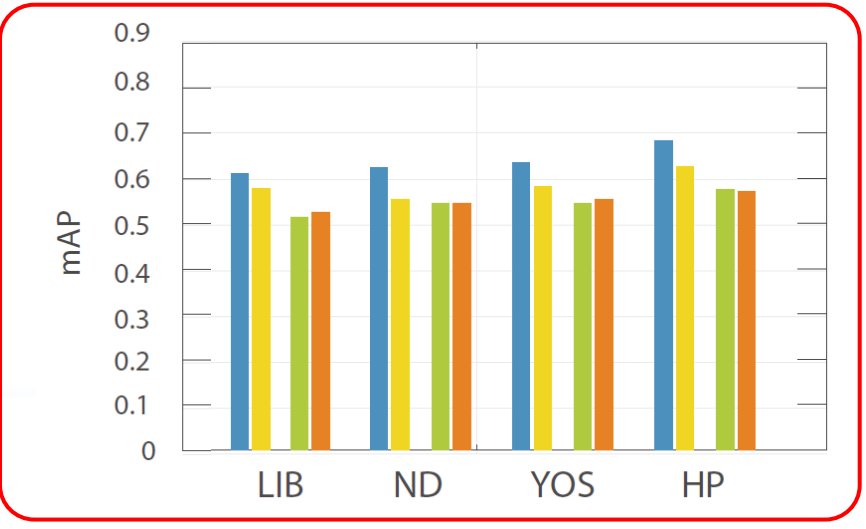


http://www.robots.ox.ac.uk/~vgg/research/affine/index.html

- 2005年提出，面向图像匹配，广泛使用
- 8个图像序列，每个包含6张图像
- 包含尺度、旋转、视角、图像模糊、光照、JPEG压缩等多种图像变换
- 给定了真实的变换对应关系



浮点特征描述子中，L2-Net要明显好于其他方法；二进制特征描述子中，L2-Net的二值版本也要优于现有的方法。



不同训练数据上得到的模型，在测试数据上性能类似，表明了L2-Net优秀的泛化能力。

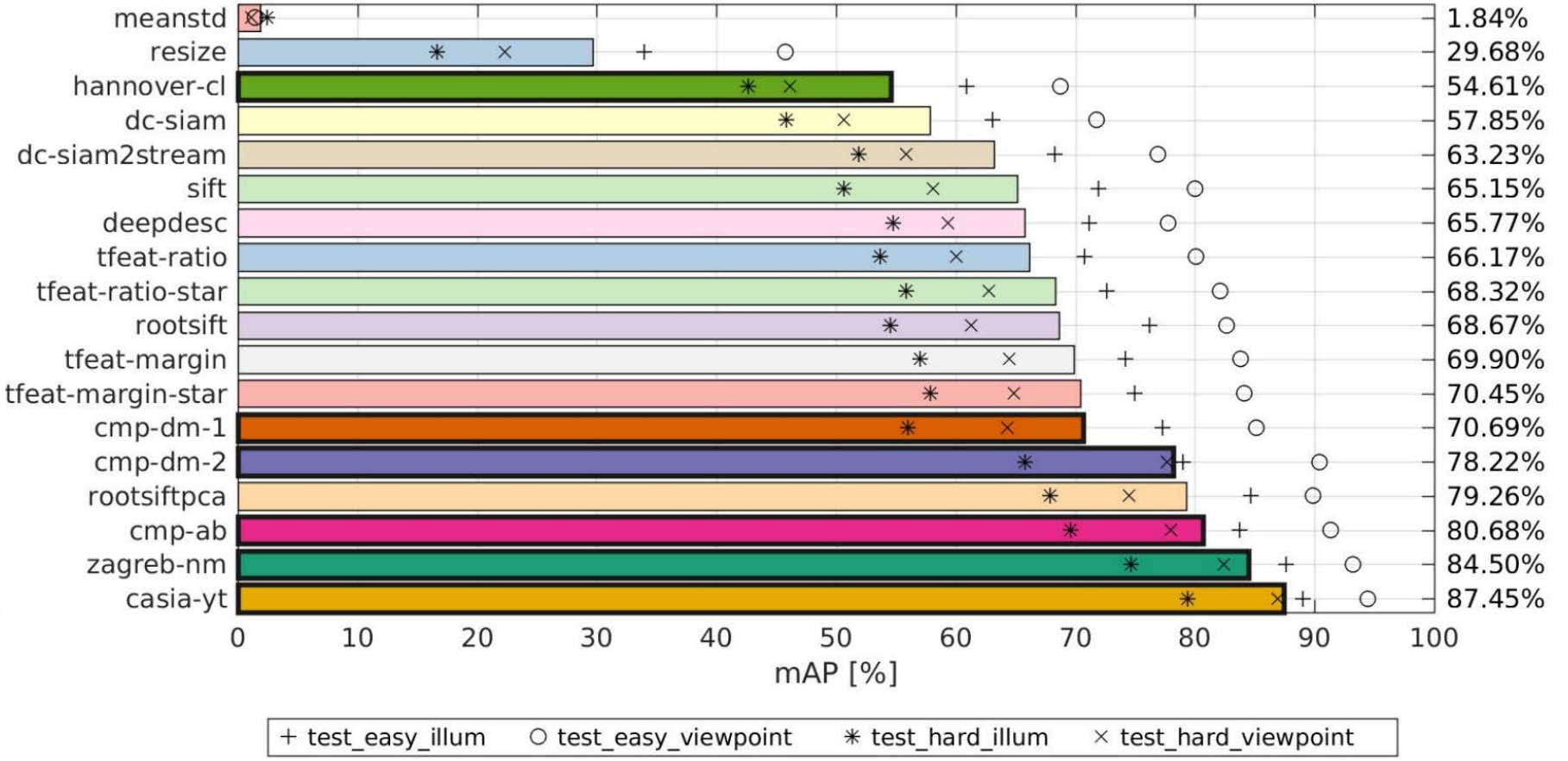
性能比较

HPacthes

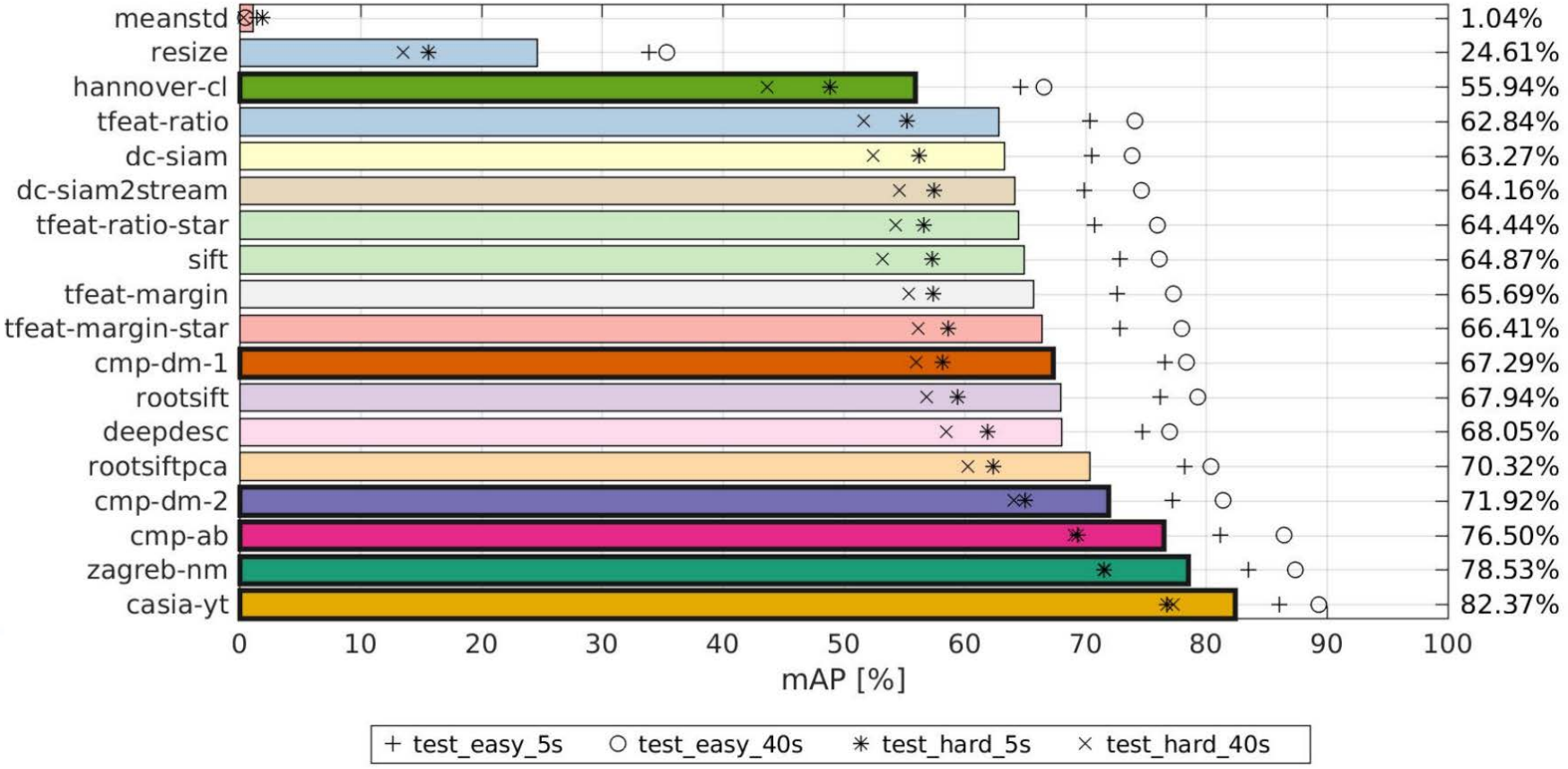
- 来自于116个图像序列的局部图像（Patch）级数据库。
- 每个序列包含同一场景的6张图像。
- 局部图像提取于多种不同的detector（Hessian，Harris，DoG）

	Train	Test
Num. of sequences	76	40
Num. of patches	581 706	353 256
AVG Patches / Image	1200	1400

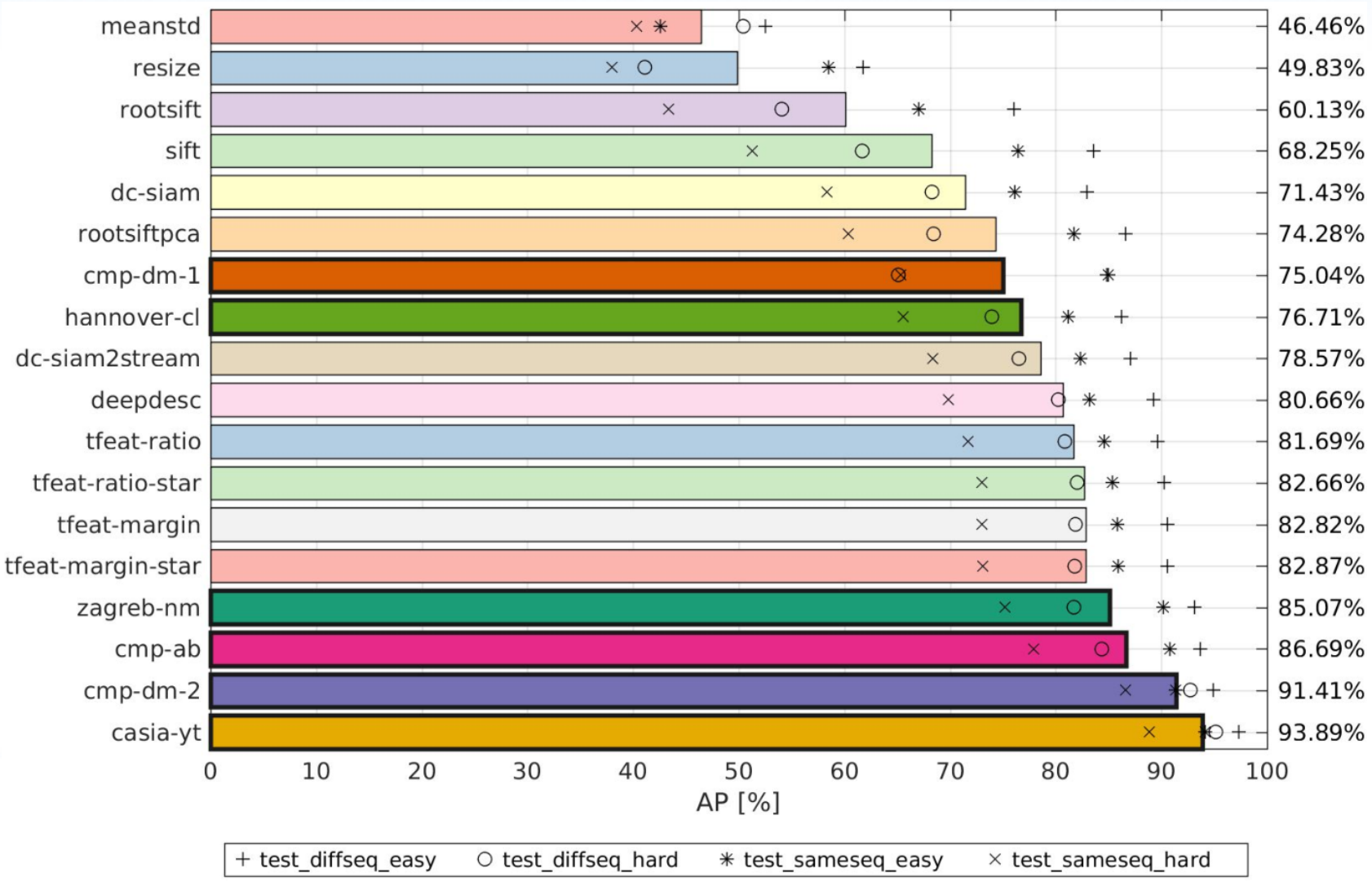
图像匹配任务：给定两幅图像中的局部图像集合，对参考图像中的每个局部图像，在目标图像中找到对应的局部图像。



检索任务：给定一个查询的局部图像，从一个非常大集合中返回匹配度最高的检索结果



分类任务：给定一对局部图像，判断它们是否匹配。



总结

	样本组织形式			应用形式		计算复杂度	
	Pairwise	Triplet	Global	Metric	L2	提取时间	训练时间
MatchNet	√			√		●●●	1-7天
Deep Compare	√			√		●●	2天
GLoss Net		√	√	√	√	●●	2天
DeepDesc	√				√	●●●	—
TFeat		√			√	●	10小时
L2-Net			√		√	●●	2-4小时

谢 谢！