深度人脸识别-综述

一、人脸识别是什么？

人脸识别和一般的目标分类任务有较大的区别，主要有下面两个方面的原因：

- 类间差别不大：因为每个人的脸都长得差不多
- 类内差别很大：同一个人在不同的姿态、光照、表情、年龄和遮挡下的人脸图片有着很大的变化

这些挑战促进了很多新结构和损失函数的产生，从而进一步提升深度模型的判别能力和泛化能力。一般来说，人脸识别内容可以分为下面两类：

1.1 Face Verification(人脸验证)

人脸验证需要解决的问题是1:1的比对，即判断两张图片里的人是否为同一个人。比如现在手机上常见的人脸解锁便是其应用场景之一，设备（手机）将摄像头采集的图片与用户事先注册的照片进行比对，从而判断是否是同一个人，从而完成人脸验证。

1.2 Face Recognition(人脸识别)

人脸识别需要解决的问题是1：N的比对，即判断系统当前输入的人脸图片，是不是事先存储在数据库中的N个人中的某一个。常见的应用场景有公司门禁、会场签到等。

两者在早期（2012~2015年）是通过不同的算法框架实现的，需要单独为它们训练一个神经网络。2015年Google的FaceNet论文的发表将两者统一到一个框架里，其基本思想是将人脸图片映射到128维的特征空间中，同一个人的图片在特征空间中的距离比与其他人的距离要近。对于人脸验证只要比对事先存储的用户照片和当前采集到的图片在特征空间中的距离，如果低于某个阈值，就判断为同一个人。对于人脸识别，需要将当前采集到的图片与数据库中所有人的图片在特征空间中一一比对（后面也会提到用SVM分类器来解决）。

1.3 当前系统的局限

在光照较差、遮挡、形变和侧脸等诸多条件下，神经网络很难提取出与“标准脸”相似的特征，从而导致验证和识别失败。目前研究中的一个方向就是如何提高CNN对这些场景的识别能力。为了保证系统的正常运行，通常有三种应对措施：

- 工程角度：研发质量模型，对检测到的人脸进行评价，质量较差则不验证/识别
- 应用角度：施加场景限制，要求用户在良好的光照环境中正对摄像头，以避免采集到质量很差的图片
- 算法角度：提升人脸识别模型性能，在训练数据中添加更多复杂场景和质量的图片，以增强模型的抗干扰能力

二、人脸识别的两种处理思路

首先给出一张图概括深度人脸识别的发展历程：

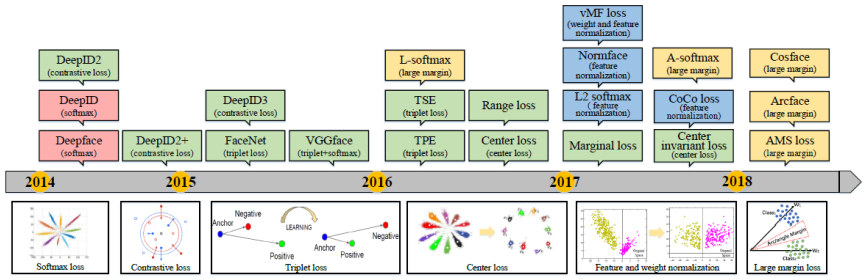


Fig. 5. The development of loss functions. It marks the beginning of deep FR that Deepface [195] and DeepID [191] were introduced in 2014. After that, Euclidean-distance-based loss always played the important role in loss function, such as contractive loss, triplet loss and center loss. In 2016 and 2017, L-softmax [126] and A-softmax [125] further promoted the development of the large-margin feature learning. In 2017, feature and weight normalization also begun to show excellent performance, which leads to the study on variations of softmax. Red, green, blue and yellow rectangles represent deep methods with softmax, Euclidean-distance-based loss, angular/cosine-margin-based loss and variations of softmax, respectively.

下表给出了上图中各种方案的表现：

TABLE IV
THE ACCURACY OF DIFFERENT VERIFICATION METHODS ON THE LFW DATASET.

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [195]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [187]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [188]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [176]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [124]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [149]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [225]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [218]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [126]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [261]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [157]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [206]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [130]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [75]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [43]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [125]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [155]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [205]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [207]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [42]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [272]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

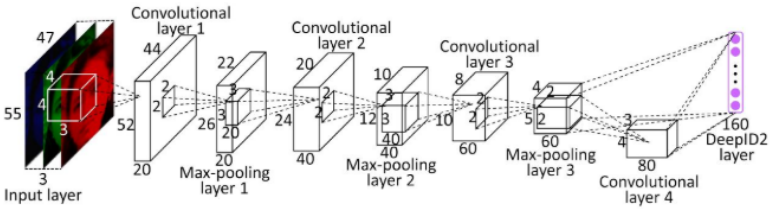
这些方案大致可以分为下面两种思路(下面的讨论只挑选了几种有代表性的方案，更多方案的细节需要读者去阅读原论文)：

2.1 思路一：Metric Learning

> *这种思路的缺点如下：*
>
> 1. *模型需要较长时间才能拟合（训练的样本数是$O(N^2)$或$O(N^3)$）*
> 2. *模型的好坏很依赖训练数据的选择（训练集很大时，遍历训练所有可能的样本计算成本太高）*

- **2.1.1 Contrastive Loss**

  代表论文DeepID2（2014），在同一个网络同时训练Verification和Classification，其中Verification Loss就是在特征层引入了Contrastive Loss。其基本思想是从数据集中选取若干对人脸二元组，属于同一个人则label为1，如果是不同人则label为-1

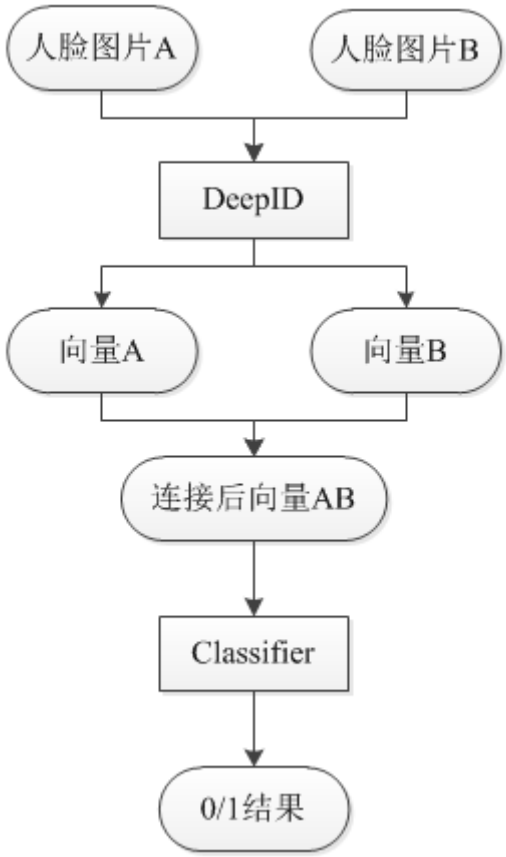


DeepID2网络模型如上图所示，其160维的DeepID2 layer的输入包括pool3和Conv4，其损失函数包括两部分，Softmax loss和L2/L1 loss，即前面所提到的识别损失和分类损失。

$$\text{Ident}(f, t, \theta_{id}) = -\sum_{i=1}^{n} p_i \log \hat{p}_i = -\log \hat{p}_t$$

$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

第一个公式是Classification loss，其主要目的是增加类间差距，从而区分不同人脸的图像；第二个公式是Verification loss，其主要目的是增加类间差距，降低类内差距。Y_{ij} =1时表示i和j属于同一类，这时f_i和f_j 越接近（类内差距越小）则loss越低；Y_{ij} =-1时表示i和j属于不同类，这时如果f_i 和f_j的距离大于某一个阈值时则loss为0，因此保证了当不同类别距离越大时loss越小。Verification loss可以从L1/L2/Cos loss中任选一种形式，同时两种损失的权重也可用超参数来控制。



上图是DeepID系列算法的整体流程，在上面的流程中，DeepID可以换为Hog、LBP等传统特征提取算法，Classifier可以是SVM、LR、NN等任意的分类算法。

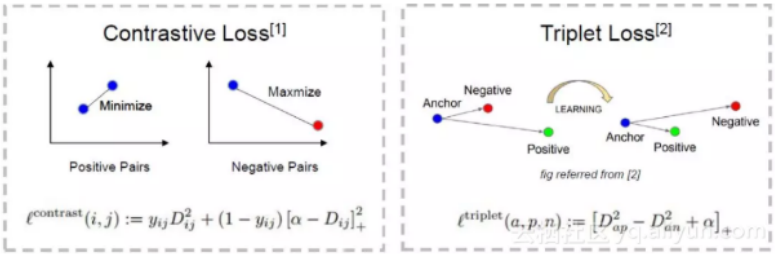
- **2.1.2 Triplet Loss**

代表论文FaceNet（2015），这篇论文提出了一个绝大部分人脸问题的统一解决框架，即识别、验证、搜索等问题都可以放到特征空间里去做，需要专注去解决的仅仅是如何将人脸更好的映射到特征空间。

Google在DeepID2的基础上，抛弃了分类层，将Contrastive Loss改进为Triplet LossT，从而让网络学到更好的feature。FaceNet的输入是三张图片（Triplet），分别为Anchor Face，Negative Face和Positive Face，则Triplet loss可以表示为：

$$\|x_i^a - x_i^p\|_2^2 + \alpha < \|x_i^a - x_i^n\|_2^2$$

直观的理解就是：在特征空间里Anchor与Positive的距离要小于Anchor与Negative的距离，同时为了保证网络不将所有的权重都学习为0，用α来限制。下图是与Contrastive Loss的对比：



- **2.1.3 一些改进**

a. Deep Face Recognition（2015）

这篇文章先用传统的Softmax训练人脸Classification模型，之后移除顶层的Classification layer，用Triplet Loss对模型进行特征层finetune，这篇文章还发布了人脸数据集VGG-Face。

b. In Defense of the Triplet Loss for Person Re-Identification

提出Squareed Euclidean Distance（L2 Distance）表现不如开方后的真实欧氏距离（L1 Distance）；

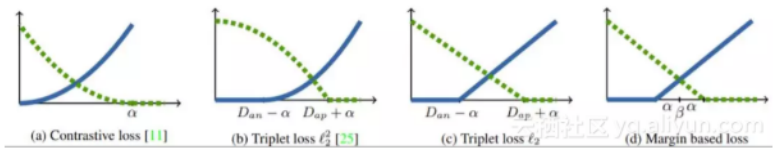
提出Soft-Margin损失公式替代原始的Triplet Loss表达式；

引进了Batch Hard Sampling。

c. Sampling Matters in Deep Embedding Learning

从导函数的角度解释了为什么b中提出的L1距离要比L2距离好，并在此基础上提出了Margin Based Loss（本质还是Triplet loss的变体）；

提出了Distance Weighted Sampling，解释说FaceNet中的Semi-hard Sampling，Deep Face Recognition中的Random Hard和b中的Batch Hard都不能轻易取到会产生大梯度（大loss，即对模型训练帮助更大的Triplets），所以从统计学的视角使用了Distance Weighted Sampling Method。



(a) Contrastive loss [11] (b) Triplet loss ℓ_2^2 [25] (c) Triplet loss ℓ_2 (d) Margin based loss

2.2 思路二：Margin Based Classification

> *Metric Learning是在feature层施加直观的强限制来计算损失，而这种思路是依然把人脸识别当成Classification任务来训练，通过对Softmax公式的改造，间接实现了对feature层施加margin的限制，从而使网络最终得到更有表现力的feature。*

- **2.2.1 Sphereface (A-Softmax Loss)**

先跟随作者的思路（下图截自原文）：

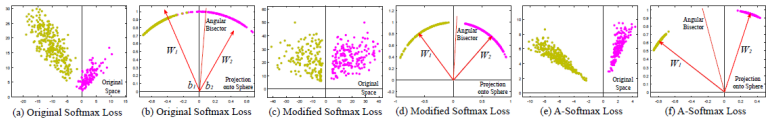


Figure 2: Comparison among softmax loss, modified softmax loss and A-Softmax loss. In this toy experiment, we construct a CNN to learn 2-D features on a subset of the CASIA face dataset. In specific, we set the output dimension of FC1 layer as 2 and visualize the learned features. Yellow dots represent the first class face features, while purple dots represent the second class face features. One can see that features learned by the original softmax loss can not be classified simply via angles, while modified softmax loss can. Our A-Softmax loss can further increase the angular margin of learned features.

原始的Softmax损失函数如下所示：

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

其中L_i 可以重新表示为角度的形式：

$$\begin{aligned}L_i&=-\log\left(\frac{e^{\boldsymbol{W}_{y_i}^T\boldsymbol{x}_i+b_{y_i}}}{\sum_j e^{\boldsymbol{W}_j^T\boldsymbol{x}_i+b_j}}\right)\\&=-\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\cos(\theta_{y_i,i})+b_{y_i}}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_{j,i})+b_j}}\right)\end{aligned}$$

图(a)是用原始Softmax损失函数训练出来的特征，图(b)是归一化的特征，不难发现原始损失函数不能通过角度来分类，归一化的损失函数则可以（把分类层的权重归一化，偏置置为0）得到了改进后的损失函数。

$$L_{\text{modified}}=\frac{1}{N}\sum_i -\log\left(\frac{e^{\|\boldsymbol{x}_i\|\cos(\theta_{y_i,i})}}{\sum_j e^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right)$$

不难看出对于特征x_i，该损失函数的优化方向是使得其向类别y_i 中心靠近，并且远离其它的类别中心，这个目标跟人脸识别目标是一致的，即最小化类内距离并且最大化类间距离。

为了保证人脸识别的正确性，还要保证最大类内距离小于最小类间距离，上面的损失函数并不能保证这一点，所以作者引入了margin的思想，这和Triplet Loss中引入α的思想是一致的。

上面损失函数中$\cos\left(\theta_{y_i,i}\right)$是样本特征与类中心的余弦值，我们的目标是缩小样本特征与类中心的角度，即增大这个值。换句话说，如果这个值越小，损失函数值越大，即我们对偏离优化目标的惩罚越大；这样就能进一步缩小类内距离和增大类间距离，最终的损失函数如下：

$$L_{\text{ang}}=\frac{1}{N}\sum_i -\log\left(\frac{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}+\sum_{j\neq y_i}e^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right)$$

原始的$cos(\theta)$被换成了$\psi(\theta)$,$\psi(\theta)$的最简单形式就是$cos(m\theta)$，文章中变得比较复杂是为了将定义域扩展到[0,2π]上，并保证在定义域内单调递减。

m便是增加的margin系数，当m=1时，$\psi(\theta)$等于$cos(\theta)$；m越大，惩罚力度越大（模型会表现的更好）。作者从数学上证明了m>3就能保证最大类内距离小于最小类间距离。

最后简单从两个方面总结一下：

a.为什么A-softmax有用？

angular marin可以直接与manifold(流形)联系起来，而人脸位于一个流形中，这更符合直觉；另一点是Softmax loss的分类结果中存在隐含的角度的某种分布，因此将angular和Softmax loss结合起来是更合理的。

b.与其它损失函数的对比

二元组(Contrastive loss)、三元组损失(Triplet loss)需要很好的策略从训练集选择合适的二元组和三元组，因此可以直观的理解它们是在这些二/三元组上找到规律，而A-softmax loss不需要数据选择，因此是在整个mini-batch上寻找规律，结果自然更具判别力和泛化能力。

- **2.2.2 NormFace**

本文主要的思想是在训练过程中将权重和特征都归一化。

Sphereface的效果很好，但是它并不优美。在测试阶段，Sphereface通过特征间的余弦值来衡量相似性，即以角度为相似性的度量。但在训练阶段，Sphereface的损失函数并不是在直接优化特征与类中心的角度，而是优化特征与类中心的角度再乘上一个特征的长度，这导致优化的方向有一部分是去增大特征的长度去了（简而言之就是Sphereface在训练时只归一化了权重）。

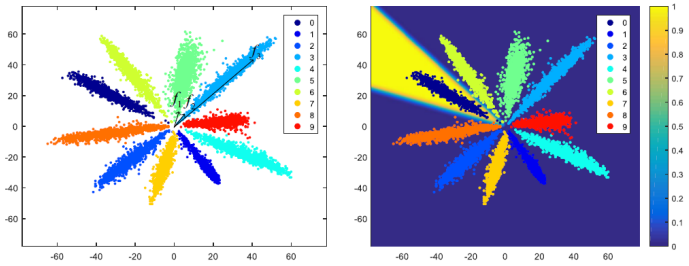


Figure 2: *Left:* The optimized 2-dimensional feature distribution using softmax loss on MNIST[14] dataset. Note that the Euclidean distance between f_1 and f_2 is much smaller than the distance between f_2 and f_3, even though f_2 and f_3 are from the same class. *Right:* The softmax probability for class 0 on the 2-dimension plane. Best viewed in color.

如上图所示，f1是一个类的特征，f2和f3是另一类的两个特征，但是f1和f2的距离明显小于f1和f3的距离，因此如果不进行特征归一化的话，很容易将f1和f2判定为同一类。

文中提到一个细节是Softmax分类不仅与W有关，也与b有关，而归一化W之后，b的影响不好补偿，因此去掉了偏置项。

另一个问题是如果直接通过归一化Softmax层的所有输入特征和权重构造损失函数来训练，网络会不收敛。文章为了解决这个问题，提出在归一化之后再乘以一个缩放参数L（假设每个类别的样本数量一致）。

最后文章给出了对Contrastive loss和Triplet loss的两个改进版公式，解决了mining pairs/triplets的问题（需要技巧且耗时），改进后的公式为：

$$\mathcal{L}_C, = \begin{cases} \left\|\tilde{\mathbf{f}}_i - \tilde{W}_j\right\|_2^2, & c_i = j \\ \max\left(0, m - \left\|\tilde{\mathbf{f}}_i - \tilde{W}_j\right\|_2^2\right), & c_i \neq j \end{cases}$$

$$\mathcal{L}_T, = \max\left(0, m + \left\|\tilde{\mathbf{f}}_i - \tilde{W}_j\right\|_2^2 - \left\|\tilde{\mathbf{f}}_i - \tilde{W}_k\right\|_2^2\right), \quad c_i = j, c_i \neq k$$

其中W_j是对应类别的权重，借鉴了Softmax是以W与X相乘表示余弦相似度的思路。

此外文中提到一个验证过程中的细节，将输入图片和它水平翻转后的图片的输出特征向量对应位置相加，再使用PCA降维，之后计算余弦相似度。

- **2.2.3 AM-softmax(CosFace-腾讯)**

这两篇文章其实是同样的内容。Normface用特征归一化解决了Sphereface训练和测试不一致的问题，但是却没有了margin的约束。AM-softmax在Normface的基础上引入了margin。损失函数如下：

$$\begin{aligned}\mathcal{L}_{AMS} &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(\cos\theta_{y_i}-m)}}{e^{s\cdot(\cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s\cdot\cos\theta_j}}\\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot\left(W_{y_i}^{T}f_i-m\right)}}{e^{s\cdot\left(W_{y_i}^{T}\boldsymbol{f}_i-m\right)}+\sum_{j=1,j\neq y_i}^{c}e^{sW_j^{T}\boldsymbol{f}_i}}\end{aligned}$$

CosFace的损失函数如下：

$$L_{lmc} = \frac{1}{N}\sum_{i} -\log\frac{e^{s\left(\cos\left(\theta_{y_i},i\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i},i\right)-m\right)}+\sum_{j\neq y_i}e^{s\,\cos(\theta_{j,i})}}$$

两篇论文思想和损失函数是一致的，相对于SphereFace来说，训练更加简单，且性能得到了提升（SphereFace为了加速训练和防止不收敛，也使用了Softmax loss帮助训练，用一个超参数来控制）。

- **2.2.4 ArcFace**

与AM-softmax/CosFace的区别在于，Arcface引入margin的方式不同。损失函数如下：

$$L_7 = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i}^{n}e^{s\,\cos\theta_j}}$$

注意m在余弦里面。文章指出尽管CosFace中的cosine margin可以一对一的映射到angular space，但本文提出的angular margin具有更清晰的几何解释（angular margin对应hypersphere manifold上的弧距）。

- **2.2.5 其它**

a. A discriminative feature learning approach for deep face recognition

提出了Center loss，加权整合进原始的softmax loss。通过维护一个欧式空间类中心，缩小类内距离，增强特征的表现力。

b. Large-margin softmax loss for convolutional neural networks

Sphereface作者的前一篇文章，未归一化权重，在softmax loss中引入了margin。这篇文章主要将L-Softmax loss用于人脸验证任务中去，后面的A-Softmax那篇论文是对权重做了归一化，并且应用到了人脸识别任务中去。

三、现状及展望

> *一个标准的人脸识别系统一般流程是：人脸检测(Face detection)及特征点检测(Landmarks detection)---->人脸对齐(Alignment)---->人脸识别(Classify)*

目前最流行的人脸及特征点检测方法是MTCNN。这篇文章提出了一个级联的多任务学习全卷积神经网络(FCN)，将人脸检测和对齐任务结合起来完成，两种任务都获得了更好的结果。整体网络由P-Net、R-Net和O-Net三级组成，每个网络都同时训练完成Face classification，Bounding box regression和Facial landmark localization三个任务，具体训练细节论文中有提到。论文的另一个亮点是提出了一种Online hard sample mining strategy，进一步提升了模型性能。

人脸对齐做的是将检测到的人脸和特征点通过几何变换，将五官变换到图中相对固定的位置，从而为后续的识别提供较为一致的输入。广泛使用的对齐方法为Similarrity Transformation，此外还有全仿射变换和带约束的仿射变换等。参考：https://zhuanlan.zhihu.com/p/29515986

值得讨论的一个问题是：人脸检测和对齐真的有必要吗？现实应用中常出现特征点无法检测的情况，没有特征点就不能对齐。相关的研究有Spatial Transform Network，文章提出让网络自己学习alignment。这方面的研究还不是很充分，实际应用中多数情况下还是使用了detection---->alignment这套流程。

目前主流的网络结构主要有DeepFace、DeepID系列、VGGFace系列、FaceNet和其它专门为FR设计的结构。其演变如下图所示：

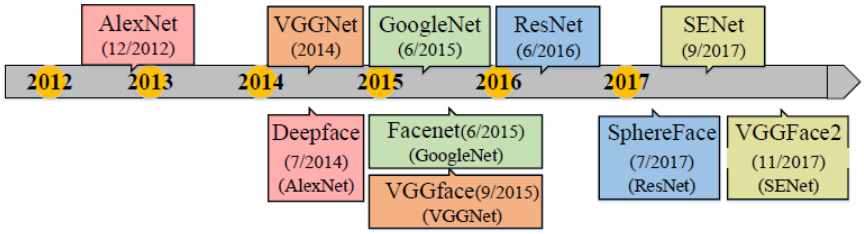


Fig. 8. The top row presents the typical network architectures in object classification, and the bottom row describes the well-known algorithms of deep FR that use the typical architectures and achieve good performance. The same color rectangles mean the same architecture. It is easy to find that the architectures of deep FR have always followed those of deep object classification and evolved from AlexNet to SENet rapidly.

四、常用的数据集

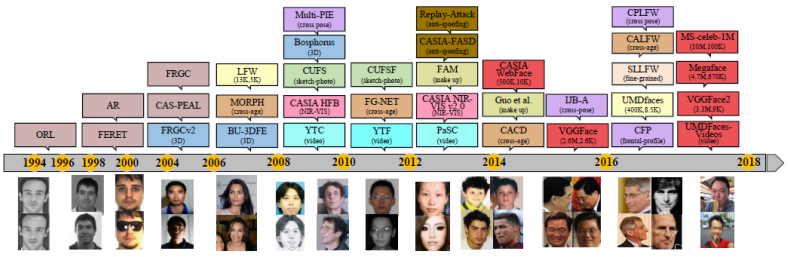


Fig. 15. The evolution of FR datasets. Before 2007, early work in FR focused on controlled and small-scale datasets. In 2007, LFW [90] dataset was introduced which marks the beginning of FR under unconstrained conditions. Since then, more testing databases with different tasks and scenes are designed. And in 2014, CASIA-Webface [243] provided the first widely-used public training dataset, large-scale training datasets begun to be hot topic. Red rectangles represent training datasets, and other color rectangles represent testing datasets with different task and scenes.

部分数据集介绍见下表：

TABLE VI
THE COMMONLY USED FR DATASETS FOR TRAINING

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Key Features
MS-Celeb-1M (Challenge 1)[69]	2016	10M 3.8M(clean)	100,000 85K(clean)	100	breadth; central part of long tail; celebrity; knowledge base
MS-Celeb-1M (Challenge 2)[69]	2016	1.5M(base set) 1K(novel set)	20K(base set) 1K(novel set)	1/-/100	low-shot learning; tailed data; celebrity
MS-Celeb-1M (Challenge 3) [2]	2018	4M(MSv1c) 2.8M(Asian-Celeb)	80K(MSv1c) 100K(Asian-Celeb)	-	breadth;central part of long tail; celebrity
MegaFace [105], [145]	2016	4.7M	672,057	3/7/2469	breadth; the whole long tail;commonalty
VGGFace2 [22]	2017	3.31M	9,131	87/362.6/843	depth; head part of long tail; cross pose, age and ethnicity; celebrity
CASIA WebFace [243]	2014	494,414	10,575	2/46.8/804	celebrity
UMDFaces-Videos [10]	2017	22,075	3,107	–	video
VGGFace [149]	2015	2.6M	2,622	1,000	depth; celebrity; annotation with bounding boxes and coarse pose
CelebFaces+ [187]	2014	202,599	10,177	19.9	private
Google [176]	2015	>500M	>10M	50	private
Facebook [195]	2014	4.4M	4K	800/1100/1200	private

[1] The min/average/max numbers of photos or frames per subject