深度人脸识别

一、 数据

1) 多对一

i. 正面化（编解码器重构）

1. Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations

主要思想：

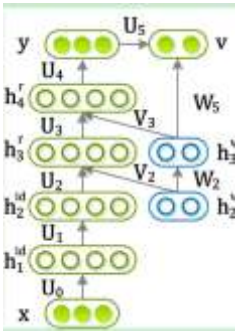
- MVP 可以从输入图像中分离身份和视图表示，也可以生成输入图像的全光谱视图。
- 与目前最先进的方法相比，MVP 的身份特征在人脸识别方面具有更好的性能。
- 将视图因子建模为连续变量，使得 MVP 能够在视点下插值出在训练数据中没有出现过的预测图像。

主要方法：

训练数据是一组图像对，$\mathcal{I}=\{x_{ij},(y_{ik},v_{ik})\}_{i=1,j=1,k=1}^{N,M,M}$，其中$x_{ij}$是第 i 个身份在 j 个视图下的输入图像。$y_{ik}$表示第 k 视图中相同身份的输出图像，而$v_{ik}$是输出的视图标签。$v_{ik}$是一个 m 维二元向量，k-元素为 1，其余为零。MVP 是从训练数据中学习得到的，给定一个输入 x，它可以输出不同视图中相同标识的图像 y 和它们的视图标签 V。

$$\mathbf{v}=F(\mathbf{y},\mathbf{h}^v;\Theta),\ \mathbf{y}=F(\mathbf{x},\mathbf{h}^{id},\mathbf{h}^v,\mathbf{h}^r;\Theta)+\epsilon, \quad (1)$$

其中 f 是一个非线性函数，Θ是一组需要学习的权重和偏差。隐神经元有h^{id}、h^v和h^r三种类型，分别提取特征特征、视图特征和特征来重建输出人脸图像。ϵ表示噪声变量。



MVP 的网络结构，它有六层，其中三层只有确定性的神经元（权重参数为 U0、U1、U4），和三层具有确定性和随机神经元（即权重 U2、V2、W2、U3、V3、U5、W5）。绿色和蓝色的节点分别表示确定性和随机神经元。y 和 v 的生成过程从 x 开始，经过h^{id}提取特征，再与隐藏视图表示h^v相结合，生成特征h^r进行人脸恢复。然后，h^r生成 y，同时，h^v和 y 联合生成 v，h^{id}和h^r是确定的二元隐神经元，而h^v是从$q(h^v)$采样的随机二进制隐式神经元。不同的采样产生不同的 y，使得多视图的感知成为可能。h^v通常具有低维数，大约为十。

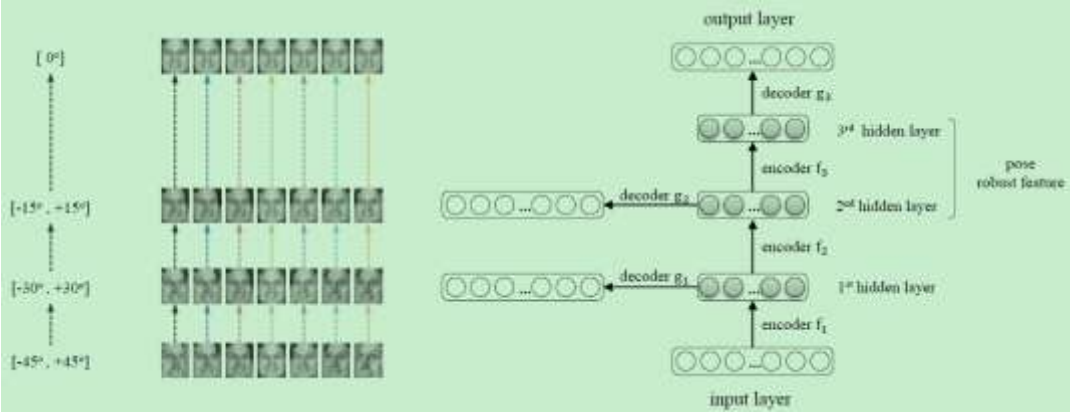
2. Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses

主要思想：

- 提出通过以渐进方式从非正面图像到正面图像的复杂非线性变换来学习姿势稳健特征，所述渐进方式被称为 stacked progressiveauto-encoders (SPAE)。
- 层叠网络的每一个浅层渐进式自动编码器设计成将大姿态的人脸图像映射到较小尺寸的虚拟视图，同时保持这些图像在较小的姿态不变。然后，叠加多个浅部自动编码器，可以将非正面人脸图像逐步转换为正面图像，这意味着将姿态变化逐步缩小到零。
- 测试图像不需要姿态估计，这与以往的许多测试方法相比，具有很大的优势。

主要步骤：

如图对于图中显示的第一个渐进式 AE，大于 30 的被转换为 30，而其他小于 30 的人脸图像被映射到自己。在测试阶段，给定图像，将其输入 SPAE 网络，并输出姿态变化很小的最顶层隐藏层被用作人脸识别的姿态鲁棒特征。



假设面部姿势被分为 2L+1 个姿态[-V,V]，-V 和 V 是左右两边最大姿态角。0 表示正面姿势。例如，在 V=45，L=3 的情况下，V 则是{-45；-30；-15；0；15；30；45}。P 表示每个 progressive AE 的目标姿态。

第 k 个渐进自动编码器尝试将角度超过 P(k)的图像转换为 P(k)：

$$\left[\mathbf{W}_k^*, \mathbf{b}_k^*, \hat{\mathbf{W}}_k^*, \hat{\mathbf{b}}_k^*\right] = \arg\min_{\mathbf{W}_k, \mathbf{b}_k, \hat{\mathbf{W}}_k, \hat{\mathbf{b}}_k} \sum_{i=1}^{N} \sum_{j \in \mathbf{V}} \|\mathbf{x}_{il} - \mathbf{g}_k(\mathbf{f}_k(\mathbf{z}_{ij}^{k-1}))\|_2^2, \quad (4)$$

$$l_{ij}^k = \begin{cases} -\mathbb{P}(k) & \text{if } l_{ij}^{k-1} < -\mathbb{P}(k) \\ +\mathbb{P}(k) & \text{if } l_{ij}^{k-1} > \mathbb{P}(k) \\ l_{ij}^{k-1} & \text{if } |l_{ij}^{k-1}| \le \mathbb{P}(k) \end{cases}, \quad (5)$$

在实现了每个浅层渐进式自动编码器之后，整个网络通过优化所有层的方式进行了精细的调整：

$$\left[\mathbf{W}_k^*|_{k=1}^L, \mathbf{b}_k^*|_{k=1}^L, \hat{\mathbf{W}}_L^*, \hat{\mathbf{b}}_L^*\right] = \arg\min_{\mathbf{W}_k|_{k=1}^L, \mathbf{b}_k|_{k=1}^L, \mathbf{W}_L, \hat{\mathbf{b}}_L} \sum_{i=1}^{N} \sum_{j \in \mathbf{V}} \|\mathbf{x}_{i,0^\circ} - \mathbf{g}_L(\mathbf{f}_L(\mathbf{f}_{L-1}(\cdots \mathbf{f}_1(\mathbf{x}_{ij}))))\|_2^2. \quad (7)$$

当姿态变化逐层减少时，最顶层 fl 的表示几乎应该有没有姿态变化。在较低层，例如 fl-1；fl-2，不是姿态不变，而是只包含非常小的姿态变化。大多数识别方法对小姿态都具有鲁棒性。使用最上面隐藏的几个层都是姿势健壮的特征，表示为 F(X)，计算为：

$$\mathbf{F}(\mathbf{x}) = [\mathbf{f}_L(\mathbf{z}^{L-1}); \mathbf{f}_{L-1}(\mathbf{z}^{L-2}); \cdots; \mathbf{f}_{L-k}(\mathbf{z}^{L-k-1})], \quad (8)$$

这些小姿态变化比只有纯粹的姿势不变特征具有更强的鲁棒性，结合 Fisher 线性判别分析（FLD）进行人脸识别。

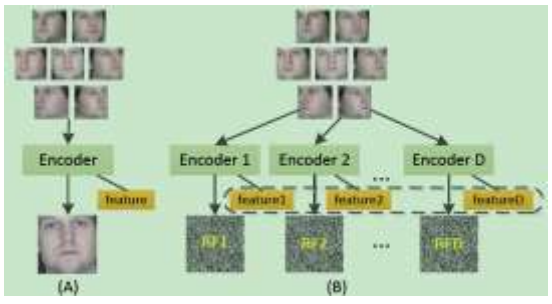
3. Random Faces Guided Sparse Many-to-One Encoderfor Pose-Invariant Face Recognition

主要思想：

- 提出了一种高阶特征学习方案，用于提取人脸识别的姿态不变特征。
- 首先，我们建立了一个单隐层神经网络稀疏约束，利用监督的方式提取 pose-invariant 特征。其次，利用多个随机人脸作为多编码器的目标值，进一步增强了该特征的判别能力。

主要步骤：

包括两个组成部分，即“稀疏多对一编码器”（SME）和“随机面孔”（RF）。稀疏多对一编码器映射不同的姿态图像到正面图像，因此在 SNN 网络的隐含层产生一个 pose-free 的特征表示。另一方面，随机面孔为 s-nn 输出许多选项，人为地在两个身份之间产生许多随机的共享结构。



- Sparse Many-to-One Encoder

为了克服过拟合问题，常在权重 w 上使用正则化项。在 W1 和 W2 权值上引入 L1 范数，使它们稀疏化。稀疏的原因有两个。首先，并不是所有的特征都同样重要，特别是对于人脸具有明显的结构的情况，稀疏可以选择最关键的特性。第二，避免过度拟合。

$$\min_{W_1,W_2,b_1,b_2} \frac{1}{2N}\sum_{i,j}\|x_i^1 - h(x_i^j)\|_2^2 + \lambda_1\|W_1\|_1 + \lambda_2\|W_2\|_1,$$

- Random Faces

对于每个主题 i，我们生成了 D 个随机人脸$y_i^d \in \mathbb{R}^n, 1 \le d \le D$，其中每个像素为都为独立同分布，且服从 0，1 均匀分布。显然，这些“随机面孔”在外观上甚至不是脸，但在训练编码器方面，它们扮演着与代表相同的正面脸的角色。对于每个输入 xi（为了简单起见，我们省略了姿态索引），我们训练了 d 个不同的编码器，因此，隐藏层的输出是 d 个，$[\tau_i^1, \tau_i^2, \ldots, \tau_i^D]$，我们将所有这些向量垂直放置，从而得到最终的姿态不变性。

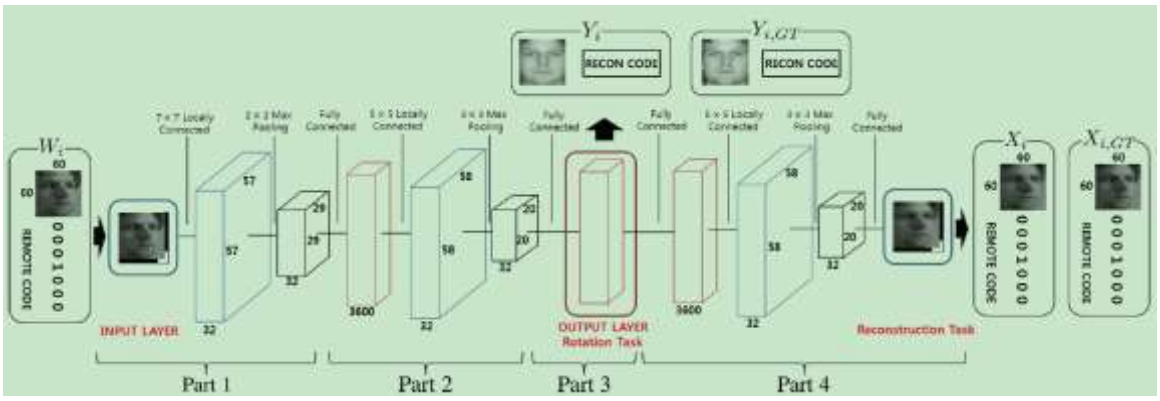
4. Rotating Your Face Using Multi-task Deep Neural Network

主要思想：

- 提出了一种新的基于多任务学习的深度结构，它可以从任意姿态和光照图像转换到指定的姿态，同时保持目标的身份。
- 训练一个深度神经网络(DNN)，接受人脸图像和编码目标姿态的二进制代码，我们称之为远程代码，并按照远程代码指示生成一个目标姿态且具有相同身份的人脸图像。
- 引入辅助 DNN 和辅助任务，该辅助任务要求主 DNN 的串联至主 DNN（产生期望的位姿图像），辅助 DNN 重建原始输入图像。

主要步骤：

该模型包括四个主要部分：特征提取部分、特征旋转部分、成像部分和辅助任务重构部分。



模型使用图像$M \in R^{N\times N}$和远程代码$C \in \{0, 1\}^{2N+1}$进行拼接（C 作为最后一行和右边一行依附在图像外延），作为输入图像$M \in R^{(N+1)\times(N+1)}$，定义为：

$$W_{(x,y)} = \begin{cases} M_{(x,y)} & \text{if } 1 \le x, y \le N \\ C^{N+1-x+y} & \text{otherwise} \end{cases}, \quad (1)$$

- Remote Code

远程代码，$C_i, i = 1, \ldots, n,$指示输入图像以相同的身份转换为 n 个姿势中的第 i 个姿态。远程代码是一种简单的重复代码，$C_i \in \{0, 1\}^l$，其总长度为 l，定义为：

$$C_i^j = \begin{cases} 1 & \text{if } (i-1)\times k < j \le i \times k \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

其中C_i^j是码C_i的第 j 位，$k = \lfloor l/n \rfloor$。(l，n)设为(121，7)和(65，9)。

从主输入的输出层开始的辅助系统不仅需要输入图像的姿态信息，而且需要输入图像的光照信息来重建输入图像。我们设置输出层代码，称为 recon 代码，$\{Q_i, S_t\}, i = 1, \ldots, n, t = 1, \ldots, m,$

类似于远程代码，我们设置了姿势代码$Q_i \in \{0, 1\}^l$

$$Q_i^j = \begin{cases} 1 & \text{if } (i-1)\times k < j \le i \times k \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

其中Q_i^j是码Q_i的第 j 位，$k = \lfloor l/n \rfloor$。(l，n)设为(49，7)和(72，9)。

此外，光照代码，$S_t \in \{0, 1\}^l$

$$S_t^j = \begin{cases} 1 & \text{if } (t-1) \times k < j \leq t \times k \\ 0 & \text{otherwise} \end{cases}, \qquad (4)$$

其中S_i^j是码S_i的第 j 位，$k = \lfloor l/n \rfloor$。(l，n)设为(80，20)和(60，20)。

- Multitask Learning

我们以平方 L2 范数作为这两个任务的代价函数。

① 对于第一项任务，对于输出层构造新的姿态图像和 recon 码构造 L2 成本函数，定义为：

$$E_c = \sum_{i=1}^{N} \|Y_{i,GT} - Y_i\|_2^2,$$

其中，$Y_{i,GT}$和Y_i分别是真实图像和生成的图像，其中包含变化的姿态图像，以及输入图像的姿态和光照信息。

② 第二个任务的成本函数，即重建输入图像和远程代码，定义为：

$$E_r = \sum_{i=1}^{N} \|X_{i,GT} - X_i\|_2^2,$$

$X_{i,GT}$和X_i是真实图像和重构图像包含输入图像和远程代码。

③ 总目标函数

$$E = \lambda_c E_c + \lambda_r E_r,$$

5. Weakly-supervised Disentangling withRecurrent Transformations for 3D View Synthesis

主要思想：

- 提出了一种新的递归卷积编解码网络，该网络利用端到端的训练由单个图像开始进行对象旋转。
- 提出的 curriculum training 方法是通过逐步增加训练序列的轨迹长度，为姿态不变识别提供更好的图像外观和更多的鉴别特征。

主要步骤：

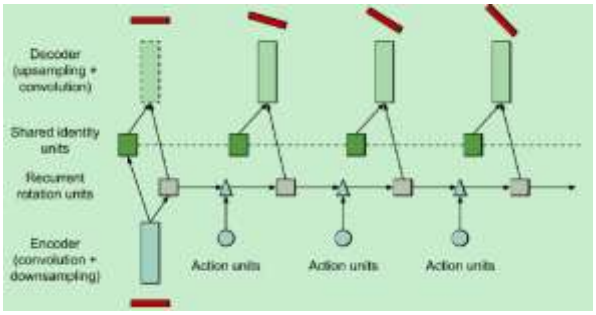
编码器网络采用 5*5 个卷积-relu 层，具有 stride 2 和 2-pixel padding，然后是两个完全连接的层。在瓶颈层中，我们定义了一组单元来表示所需转换的姿态(姿态单元)。另一组单元表示在转换过程中不会更改的内容，称为身份单位。解码器和编码器的结构是对称的。

所需的转换由动作单元反映。我们用一个 1-of-3 编码，其中[100]编码顺时针旋转，[010]编码一个空操作，[001]编码的逆时针旋转。三角形表示一个张量积作为姿态单元和动作单元的输入，并生成变换的姿态单元。等效地，动作单元选择将输入姿态单元转换成输出姿态单元的矩阵。

我们使用递归的姿态单元来建立一步一步的姿态流形变化模型。由于我们假设所有训练序列只保留身份，而只改变姿势，所以所有的身份单元都是在所有时间步骤中共享的。

$$\mathcal{L}_{rnn} = \sum_{i=1}^{N} \sum_{t=1}^{T} \|y^{(i,t)} - g(f_{pose}(x^{(i)}, a^{(i)}, t), f_{id}(x^{(i)}))\|_2^2 \qquad (1)$$

其中$a^{(i)}$是 T 动作的序列，$f_{id}(x^{(i)})$产生对所有时间步长不变的恒等身份特征，$f_{pose}(x^{(i)}, a^{(i)}, t)$在时间步长 T 内产生的转换的姿态特征，g(.;.)是给定$f_{id}(.)$和$f_{pose}(.;.;.)$时产生的图像输出，x(i)是第 i 个图像，$y^{(i,t)}$是步骤 t 处的第 i 训练图像目标。



ii. 正面化（CNN）

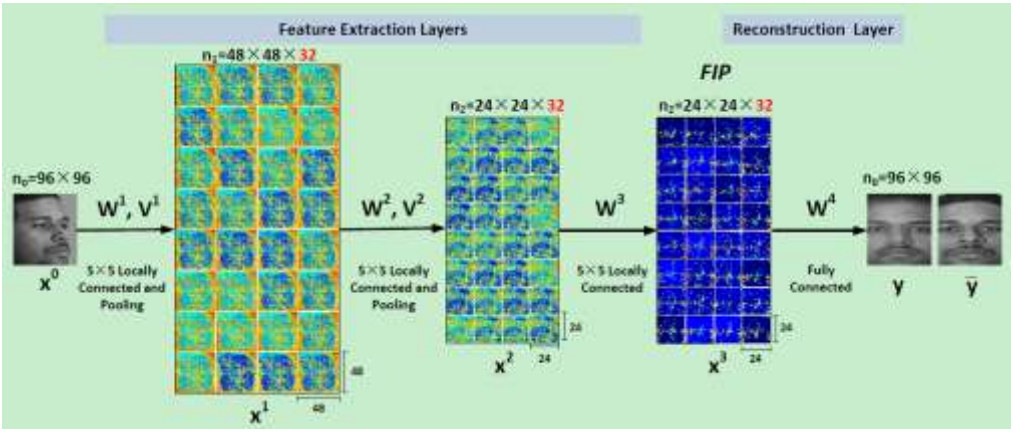
6. Deep Learning Identity-Preserving Face Space

主要思想：

- 提出了一种新的学习的人脸表示：face 身份保持（FIP）特征。与传统的人脸描述，FIP 特征可以显著降低内部身份差异，同时保持 discriminativeness。此外，从图像中提取任何姿态和光照下的 FIP 特征可以在规范的角度重构人脸图像。
- 设计了一个结合特征提取层和重建层的深层网络。前者编码的人脸图像为 FIP 特征，后者转换他们在标准视图图像。

主要步骤：

它结合了特征提取层和重建层。特征提取层包括三个局部连接层和两个汇聚层。它们编码的输入面 x0 为 FIP 特征 X3。X1，x2 是第一和第二局部连接层的输出特征映射。FIP 特征可以用来恢复人脸图像 in the canonical view。y 是目标图片。



① 首先，通过特征提取层对输入图像进行编码，特征提取层包括三个局部连接层和两个交替叠加的池层。每一层捕捉不同尺度的人脸特征。

第一局部连接层输出 32 特征图，主要捕获姿势信息和一些面部区域内高响应，捕捉面部结构（红色）表示大响应，蓝色表示无响应）。在第二局部连接层的特征图中，人脸区域外的高响应显著降低，表明在保留人脸结构的同时丢弃了大部分的姿态变化。第三层局部连接层输出的 FIP 特征，是稀疏的且专注于身份信息。

② 其次，FIP 特征使用全连接的重建层在 canonical view 中恢复人脸图像。

首先，基于 least squaredictionary learning 初始化参数，然后通过反向传播，利用重构图像与真实图片之间的求和平方重建误差来更新所有参数。

$X^3 = \{x_i^3\}_{i=1}^m$是 FIP 特征，$\bar{Y}$是目标图片。

$$\arg\min_{W^1} \| \bar{Y} - OW^1X^0 \|_F^2, \quad (7)$$

$$\arg\min_{W^2} \| \bar{Y} - PW^2X^1 \|_F^2, \quad (8)$$

$$\arg\min_{W^3} \| \bar{Y} - QW^3X^2 \|_F^2, \quad (9)$$

$$\arg\min_{W^4} \| \bar{Y} - W^4X^3 \|_F^2, \quad (10)$$

$X^0 = \{x_i^0\}_{i=1}^m$是输入图像，所以，W^1X^0在输出 32 个特征图。O 是一个固定的二进制矩阵，将 32 个特征图相同位置的像素相加。特征映射，这使得OW^1X^0与$\bar{Y}$尺寸相同。$X^1 = \{x_i^1\}_{i=1}^m$是第一个局部连接层的输出。P 也是一个固定的二进制矩阵，它将相应的像素加和并将结果重新变为与$\bar{Y}$相同的大小。先优化 W1，在优化 W2，重复这个过程，直到所有矩阵都被初始化为止。

通过最小化重构误差的损失函数来更新初始化之后的所有权重矩阵：

$$E(X^0; \mathbf{W}) = \| \bar{Y} - Y \|_F^2, \quad (11)$$

其中 Y 是重构图片，$\bar{Y}$是目标图片。

7. Recover Canonical-View Faces in the Wild with Deep Neural Networks

主要思想：

- 开发了一种从个体照片中自动选择/合成 canonical-view（标准正脸图片）的方法。
- 在应用方面，该人脸恢复方法已经应用于人脸验证。
- 该文章的亮点在于：一，新的检测/选择 canonical-view 的方法；二，训练深度神经网络来重建人脸正面标准图片（canonical-view）。

主要步骤：

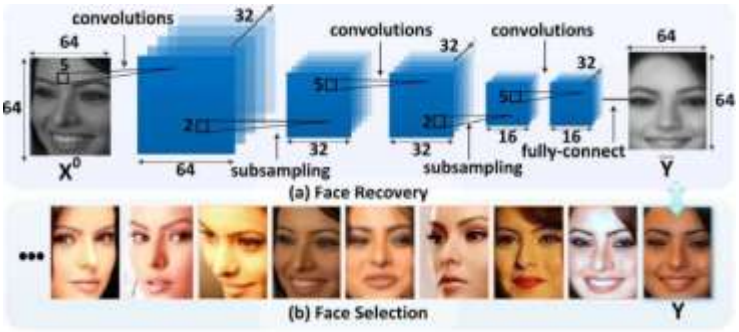
- canonical view 选择方法

设计了基于矩阵排序和对称性的人脸正面图像检测方法。按照以下三个标准来采集个体人脸图片：一，人脸对称性（左右脸的差异）进行升序排列；二，图像锐度进行降序排列；三，一和二的组合。

$$M(\mathbf{Y}_i) = \| \mathbf{Y}_i\mathbf{P} - \mathbf{Y}_i\mathbf{Q} \|_F^2 - \lambda \| \mathbf{Y}_i \|_*, \quad (1)$$

P、Q 是两个常数矩阵，$\mathbf{P} = diag([\mathbf{1}_{32}, \mathbf{0}_{32}])$ and $\mathbf{Q} = diag([\mathbf{0}_{32}, \mathbf{1}_{32}])$的第一项测量脸部的对称性，即左半边和右半边的区别，第二项测量面部的秩。值越小表明，更可能是正面视图。

- 人脸重建



我们通过训练深度神经网络来进行人脸重建。loss 函数为：

$$E(\{\mathbf{X}_{ik}^0\}; \mathbf{W}) = \sum_i \sum_k \| \mathbf{Y}_i - f(\mathbf{X}_{ik}^0; \mathbf{W}) \|_F^2, \quad (2)$$

i 为第 i 个个体，k 为第 i 个个体的第 k 张样本。X0 和 Y 为训练图像和目标图像。

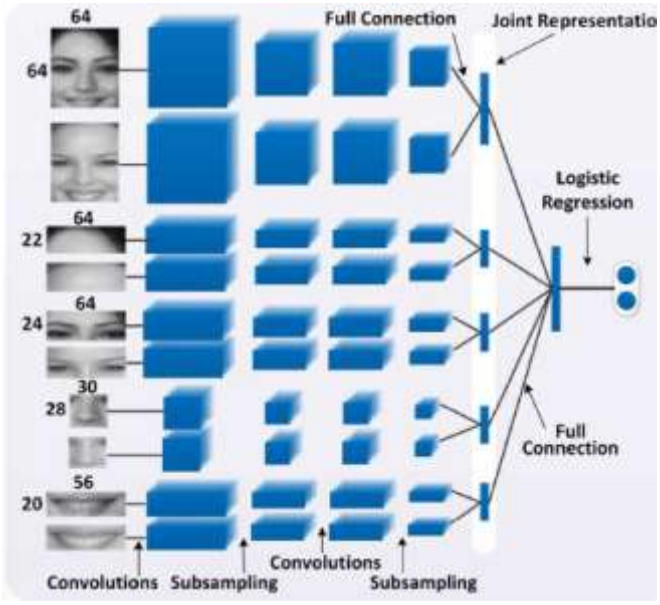
深层网络包含三个卷积层。前两个之后是最大的池层，最后一个是完全连接层。不同于传统 CNN 的过滤器分享权重，我们的过滤器是局部和不共享的权重，因为假设不同脸区域应该采用不同的特性。

- Face Verification

网络包含五个 cnn，每一个都需要一双全脸或面部组件作为输入。整张脸的大小，额头，眼睛，鼻子和嘴巴都是 64 x 64，22 x 64，24 x 64，28 x 30 和 20 x 56。首先，CNN 的学会联合表示。逻辑回归层然后连接所有联合表征特性预测两个脸图像是否属于同一身份。

$$Err = \mathbf{y} \log \bar{\mathbf{y}} + (1 - \mathbf{y}) \log(1 - \bar{\mathbf{y}})$$

测试时，采用 PCA 降维，SVM 进行人脸验证。



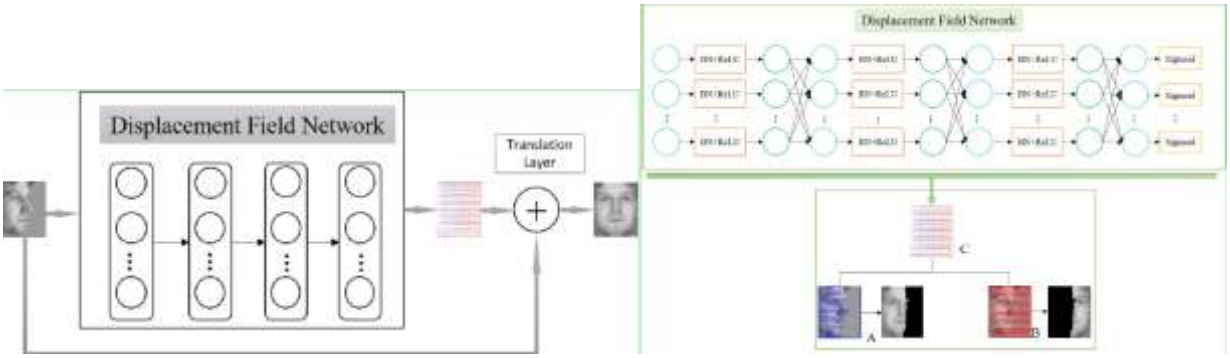
8. LDF-Net: Learning a Displacement Field Network for Face Recognition

主要思想：

- 该方法通过学习位移场网络(Ldfnet)，将非正面人脸图像直接转化为正面人脸图像，然后利用变换后的图像进行识别。
- 三维方法可能会导致一些像素在变换后的正面图像中的不可见性，而 2d 方法则可能导致变换后的正面图像中的像素与原始的非正面图像之间的差异。我们提出的 LDF-Net 方法可以通过学习变换后的正面图像中每个像素的可移动场来处理这两个问题。
- LDF-net 学习位移场，它反映了非正面人脸图像和变换后的正面人脸图像像素之间的移动关系。LDF-net 可以实现正面图像，尽可能地保留原始图像中的信息信息，并且没有不可见的像素。

主要步骤：

LDF 网主要分为两个部分。主体部分，位移场网络 F，学习目标正面图像与输入非正面图像之间的像素的位移场。另一部分，平移层 T，由 F 出的位移场,入的非正面人脸图像转换成正面图像。



- Displacement Field Network

位移场模拟了两个像素之间的移动关系，即转换后的正面图像中的一个像素和原始非正面图像中的相应像素之间的移动关系。

$$D_k = F_W(I_k)$$

位移场 $D_k \in \mathbb{R}^{h\times w\times 2}$ 由变换后的图像中每个像素的二维位移组成。设 $(\Delta^h_{kij}, \Delta^w_{kij}) \triangleq (D_k(i,j,1), D_k(i,j,2))$ 表示位于 h 和 w 轴上(i，j)像素的平移距离。位移场网络可以是任何一种深层次的网络结构，如 cnn，也可以是完全连接的网络。

- Translation Layer

利用位移场 dk，平移层通过移动输入图像中的像素，将输入图像 ik 转换为正面图像 I_k^{est}，

$$I_k^{est} = T(I_k, D_k) = T\left(I_k, F_W(I_k)\right).$$

若 Δ^h_{kij} and Δ^w_{kij} 为整数，则 $I_k^{est}(i,j) = I_k(\tilde{i},\tilde{j})$，其中 $\tilde{i} \triangleq i + \Delta^h_{kij}, \tilde{j} \triangleq j + \Delta^w_{kij}$。

若不为整数，可以采用四个相邻像素的加权和，而不是四舍五入为整数：

$$I_k^{est}(i,j) = \sum_{m=\lfloor\tilde{i}\rfloor}^{\lceil\tilde{i}\rceil}\sum_{n=\lfloor\tilde{j}\rfloor}^{\lceil\tilde{j}\rceil} I_k(m,n)(1-|\tilde{i}-m|)(1-|\tilde{j}-n|).$$

- Overall Objective

$$\begin{aligned} W &= \arg\min_W \sum_{k=1}^{n} ||I_k^{gt} - I_k^{est}||_2^2 \\ &= \arg\min_W \sum_{k=1}^{n} ||I_k^{gt} - T(I_k, D_k)||_2^2 \\ &= \arg\min_W \sum_{k=1}^{n} \left|\left|I_k^{gt} - T\left(I_k, F_W(I_k)\right)\right|\right|_2^2. \end{aligned} \tag{8}$$

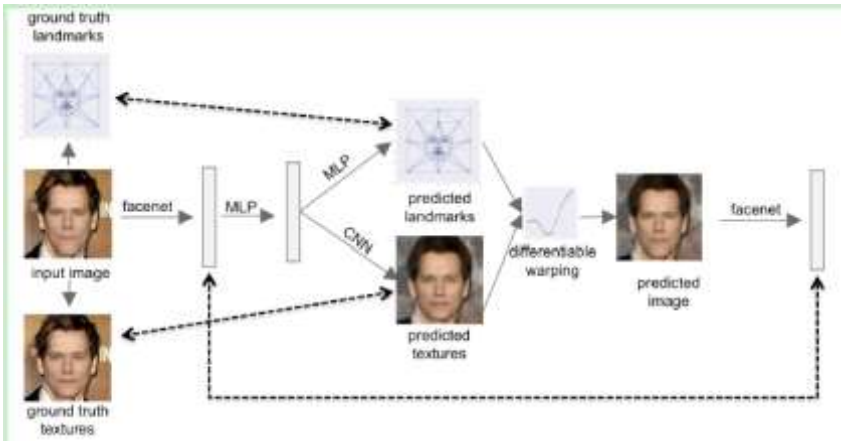
9. Synthesizing Normalized Faces from Facial Identity Features

主要思想：

- 提供一种方法，用于合成给定输入面部照片的人面部正面、中立表情图像。这是通过学习从面部识别网络中提取的特征生成面部地标和纹理来实现的。
- 编码器提取特征向量在很大程度上不受光照、姿态和面部表情的影响。然后，解码器独立地预测 landmarks 和纹理，并使用一种可微图像扭曲操作来组合它们。

主要步骤：

利用面部识别特征的不变性来进行姿势、照明的表达，然后从特征向量映射到均匀分布的、正面的、中和表情的面部。



首先使用 FaceNet 加上多层感知器(MLP)层（即完全连接和 relu 层）将图像编码为一个小的特征向量。然后，利用深度卷积网络(CNN)生成 landmarks 向量，利用 MLP 分别生成纹理映射。这些组合使用可微性扭曲来产生最终渲染的图像。每条虚线连接在损失函数中比较的两个项。纹理使用平均绝对误差，地标用均方误差，FaceNet 嵌入使用负余弦相似性。

- Differentiable Image Warping

图像扭曲：左：开始地标位置，中左：理想的最终位置，包括零位移边界条件，中右：通过样条插值获得的密集流场，右：流在图像中的应用。

- Data Augmentation using Random Morphs

给定一个种子面 A，我们首先选择一个随机的 k＝200 个最近邻中的一个来选择目标面。我们测量 a 和 b 之间的距离：

$$d(A,B) = \lambda\|L_A - L_B\| + \|T_A - T_B\|, \qquad (2)$$

给定 a 和随机邻域 b，我们独立地对它们的地标和纹理进行线性插值，其中插值权重是从[0; 1]均匀分布。

变形倾向于保存面部的细节，在那里的地标是准确的，但不能捕捉头发和背景细节。为了使增强后的图像更加逼真，我们使用梯度域编辑技术将变形的脸粘贴到原始背景上。

考虑到变形人脸图像 TF 和目标背景图像 TB 的纹理，我们构造了对输出纹理 TO 的梯度和颜色的约束，以达到如下目的：

$$\begin{aligned}
\frac{\partial}{\partial x}T_o &= \frac{\partial}{\partial x}T_f \circ M + \frac{\partial}{\partial x}T_b \circ (1-M) \\
\frac{\partial}{\partial y}T_o &= \frac{\partial}{\partial y}T_f \circ M + \frac{\partial}{\partial y}T_b \circ (1-M) \\
T_o \circ M &= T_f \circ M,
\end{aligned} \qquad (3)$$

iii. 正面化（GAN）

10. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis

主要思想：

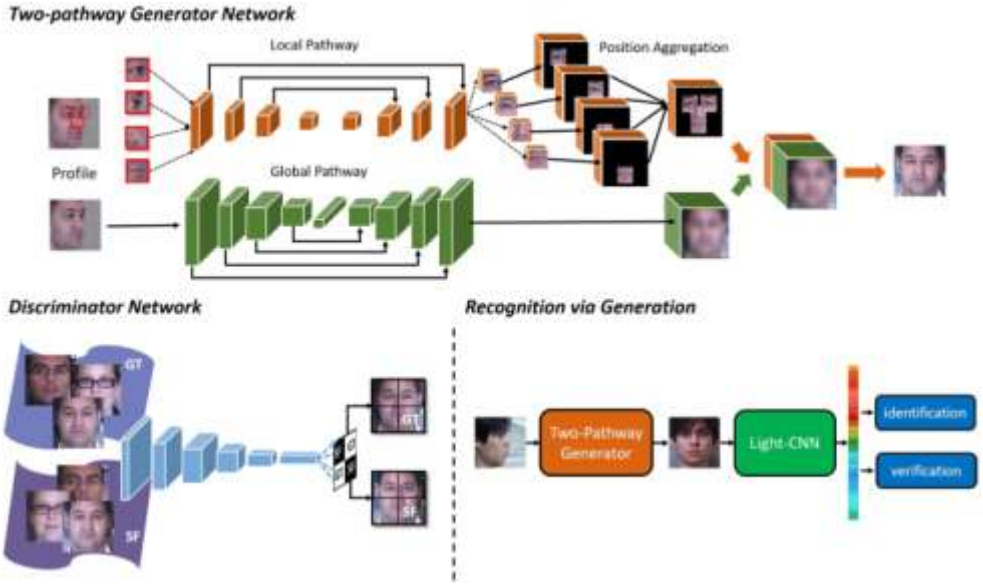
- 介绍了一种新提出了一种 a Two-Pathway Generative Adversarial Network (TP-GAN)，生成逼真的正面视图合成的同时保持整体结构和局部细节。
- 除了全局的编码器解码网络外，还提出了四个 Four landmark located patch networks 来关注局部纹理。
- 通过引入 adversarial loss, symmetry loss 和 identity preserving loss，使这种大姿态问题得到了很好的约束。
- 对抗性损失可以忠实地发现和引导合成驻留在正面人脸的数据分布中。对称损失可以显式利用对称性，以减轻大姿态情况下的自遮挡效应。此外，身份保留损失使生成的合成结果不仅是视觉上的吸引力，但也容易适用于准确的人脸识别。

主要步骤：

这里采用双通道网络，Two Pathway Generator，一个是 local pathway,另一个是 global pathway.：

local pathway：用于解决人脸的细节问题，输入侧脸的四个特征图像块：分别是两个眼睛、鼻子、嘴巴。输出正脸的对应四个图像块。 基于编码器解码器结构，但它没有完全连接的瓶颈层。四个局部路径的特征张量（多特征映射）最后映射到一个单一特征张量（然后引入最大融合策略来减少重叠区域上的拼接伪影），它与全局特征张量具有相同的空间分辨率。

global pathway：用于生产人脸大的结构，缺少细节，输入完整的侧脸图像输出完整的模糊的正脸图像。global networkG_{θ^g}由下采样编码器$G_{E_\theta^g}$和上采样 θ 解码器$G_{D_\theta^g}$，额外的 skip 层引入多尺度特征融合。在中间的瓶颈层输出的 256 维特征向量 V，用于身份分类。



- Adversarial Networks

$$\min_{\theta_G}\max_{\theta_D}\mathbb{E}_{I^F\sim P(I^F)}\log D_{\theta_D}(I^F)+ \mathbb{E}_{I^P\sim P(I^P)}\log(1-D_{\theta_D}(G_{\theta_G}(I^P))) \quad (2)$$

输出的 2×2 概率图，而不是一个标量值。每个概率值现在对应于某一区域而不是整个脸，

- Pixel-wise Loss

$$L_{pixel}=\frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|I_{x,y}^{pred}-I_{x,y}^{gt}| \quad (3)$$

pixel wise loss 作用在全局和 landmark located patch 网络的输出以及他们的融合结果。

- Symmetry Loss

$$L_{sym}=\frac{1}{W/2\times H}\sum_{x=1}^{W/2}\sum_{y=1}^{H}|I_{x,y}^{pred}-I_{W-(x-1),y}^{pred}| \quad (4)$$

- Adversarial Loss

$$L_{adv}=\frac{1}{N}\sum_{n=1}^{N}-\log D_{\theta_D}(G_{\theta_G}(I_n^P)) \quad (5)$$

- Identity Preserving Loss

$$L_{ip}=\sum_{i=1}^{2}\frac{1}{W_i\times H_i}\sum_{x=1}^{W_i}\sum_{y=1}^{H_i}|F(I^P)_{x,y}^i-F(G(I^{pred}))_{x,y}^i|$$

在紧致深特征空间中，身份保持损失使得预测与地面真实距离很小。

- 总表达式

$$\hat{\theta}_G = \frac{1}{N}\underset{\theta_G}{\operatorname{argmin}} \sum_{n=1}^{N} \{L_{syn}(G_{\theta_G}(I_n^P), I_n^F) + \alpha L_{cross\_entropy}(G_{\theta_g^y}(I_n^P), y_n)\}$$

$$L_{syn} = L_{pixel} + \lambda_1 L_{sym} + \lambda_2 L_{adv} + \lambda_3 L_{ip} + \lambda_4 L_{tv}$$

11. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition

主要思想：

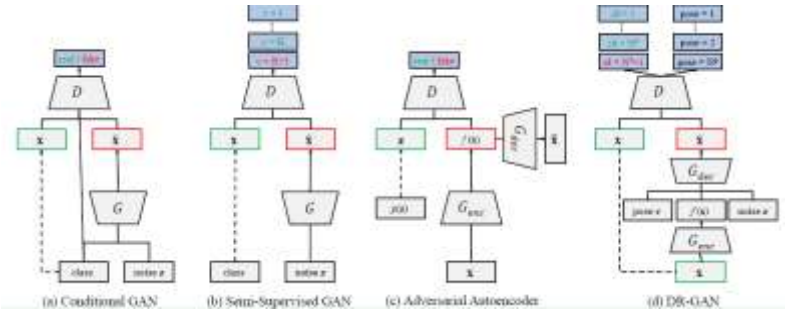
- 首先，发生器的编解码结构允许 DR-GAN 学习一种 generative 和 discriminative 的表示，同时完成图像合成。
- 第二，通过提供给解码器的姿态码和鉴别器中的姿态估计，显式地将这种表示与其他人脸变化(如姿态)分离。
- 第三，DR-GAN 可以以一幅或多幅图像作为输入，并与任意数量的合成图像一起生成一个统一的表示。

主要步骤：

该网络主要实现两个功能 1）学习的姿态不变的身份表示，2）合成的人脸图像$\hat{x}$具有相同身份的y^d但根据姿势代码 C 指定一个不同的姿势

- Single-Image DR-GAN

生成对抗网络由 G 和 D 生成样本，GAN 的主要目标是图像合成，建立了一个带有编码-解码结构的 G，编码器的输入是任意姿态的人脸，将人脸表示成一个与姿态无关的特征表示，解码器输入是特征表示，姿态编码 c 及一个随机噪声 z（串联），输出是人造的指定姿态的人脸。D 不仅用来分辨真实和人造图像，还用来预测 ID 及姿态。



D 是一个多任务的 CNN，包含两个部分：$D = [D^d, D^p]$。D^d用来识别身份，D^p用来识别姿态。

$$\max_D V_D(D,G) = E_{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y})}[\log D^d_{y^d}(\mathbf{x}) + \log D^p_{y^p}(\mathbf{x})] + E_{\substack{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{N^d+1}(G(\mathbf{x},\mathbf{c},\mathbf{z})))], \quad (4)$$

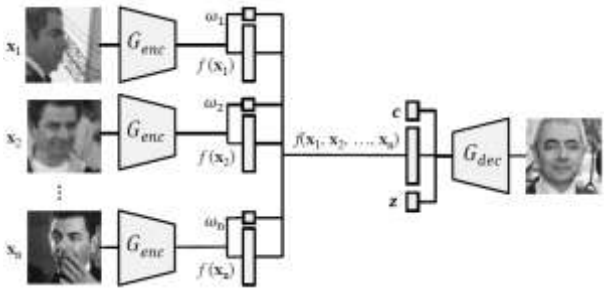
第一项是最大化 x 被归类为真实身份和姿势的概率。第二项是最大化$\hat{x}$被归类为伪类的概率。

G 由编码器G_{enc}和解码器G_{dec}构成。G_{enc}的目的是从人脸图像 x 中学习身份表示：f(x)=G_{enc}(x)。G_{dec}旨在合成人脸图像$\hat{x}$= G_{dec}(f(x)、C、Z）在保持身份y^d和目标姿态 C 的情况下，其中 Z 是除了身份或带来其他方差建模的噪声。

$$\max_G V_G(D,G) = E_{\substack{\mathbf{x},\mathbf{y}\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x},\mathbf{c},\mathbf{z})))+ \log(D^p_{y^t}(G(\mathbf{x},\mathbf{c},\mathbf{z})))]. \quad (5)$$

- Multi-Image DR-GAN

Genc 输入多幅图像，生成每个图像的特征表示及系数，所有表示综合成一个表示，Gdec 使用这个表示合成人脸。Multi-Image GAN 的生成器如下图所示：



对于不用样本有不同的 Genc，但是只有一个 Gdec 和 D。此外，Genc 生成特征表示的同时，还为每个图像产生一个置信度，融合表示是所有表示的加权平均值。

$$f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) = \frac{\sum_{i=1}^{n} \omega_i f(\mathbf{x}_i)}{\sum_{i=1}^{n} \omega_i}. \quad (6)$$

具体来说，在 Genc 的末尾，我们在 avgpool 之前的层中再增加一个卷积通道，来估计系数。使用 sigmoid 激活函数来约束ϖ在[0, 1]范围内。

此时，G 的优化函数为：

$$\begin{aligned}\max_G V_G(D,G) = & \sum_{i=1}^{n}[E_{\substack{\mathbf{x}_i,\mathbf{y}_i\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x}_i,\mathbf{c},\mathbf{z}))) + \\ & \log(D^p_{y^p}(G(\mathbf{x}_i,\mathbf{c},\mathbf{z})))]] + \\ & E_{\substack{\mathbf{x}_i,\mathbf{y}_i\sim p_d(\mathbf{x},\mathbf{y}),\\ \mathbf{z}\sim p_z(\mathbf{z}),\mathbf{c}\sim p_c(\mathbf{c})}}[\log(D^d_{y^d}(G(\mathbf{x}_1,...,\mathbf{x}_n,\mathbf{c},\mathbf{z}))) + \\ & \log(D^p_{y^p}(G(\mathbf{x}_1,...,\mathbf{x}_n,\mathbf{c},\mathbf{z})))]. \quad (7)\end{aligned}$$

12. UV-GAN: Adversarial Facial UV Map Completion for Pose-invariant Face Recognition

主要思想：

- 提出了一种训练深层卷积神经网络(DCNN)的框架，以完成从wild图像中提取的面部UV图，解决从一幅图像中生成面部UV图的不完整问题（由于自遮挡）。
- 从3 DMM到2D图像，生成不完全的面部UV图。为了完成uv图，将本地和全局对抗性网络结合起来学习保持身份的完全uv纹理。
- 在人脸识别训练中，可以丰富训练数据的姿态变化，而不需要手工标记所有姿态的大数据集。对于识别试验，我们的方法可以缩小验证图相对的姿态差异，从而获得更好的性能。

主要方法：

- 3D Morphable Model Fitting

三个参数模型是需要解决的：形状模型、纹理模型和摄像机模型：

$$\begin{aligned}\mathcal{S}(\mathbf{p}) &= \bar{\mathbf{s}} + \mathbf{U}_s\mathbf{p}, \\ \mathcal{T}(\boldsymbol{\lambda}) &= \bar{\mathbf{t}} + \mathbf{U}_t\boldsymbol{\lambda}, \\ \mathcal{W}(\mathbf{p},\mathbf{c}) &= \mathcal{P}(\mathcal{S}(\mathbf{p}),\mathbf{c}),\end{aligned}$$

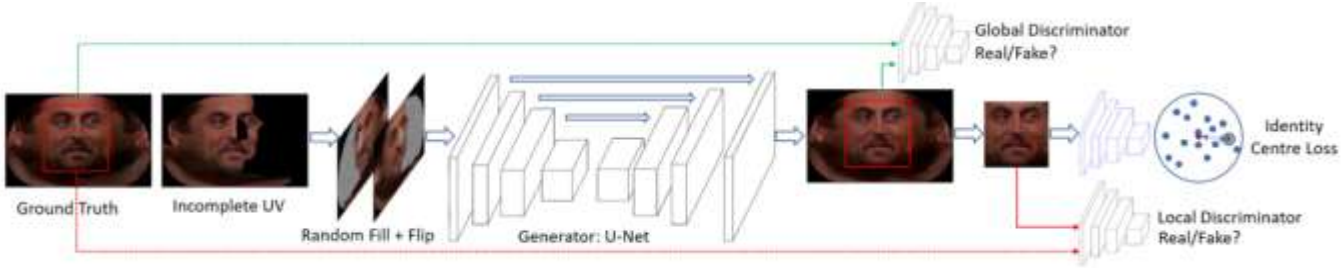
3 DMM fitting的总损失函数如下：

$$\begin{aligned}\arg\min_{\mathbf{p},\boldsymbol{\lambda},\mathbf{c}} & \|\mathbf{F}(\mathcal{W}(\mathbf{p},\mathbf{c})) - \mathcal{T}(\boldsymbol{\lambda})\|^2 + \alpha_l\|\mathcal{W}_l(\mathbf{p},\mathbf{c})) - \mathbf{s}_l\|^2 \\ & + \alpha_s\|\mathbf{p}\|^2_{\boldsymbol{\Sigma}_s^{-1}} + \alpha_t\|\boldsymbol{\lambda}\|^2_{\boldsymbol{\Sigma}_t^{-1}}, \quad (5)\end{aligned}$$

基于fitting结果，我们从Casia数据集中对人脸图像进行分类，分成13个姿态组。

- UV Texture Completion

提出一种用于 UV 完成的生成对抗网络(我们称为 UV-GaN)，它包括一个 UV 生成模块、两个区分器和一个附加模块以保持脸部特征。



生成器采用编解码器架构，其中，在编码器和解码器堆栈中的镜像层之间建立 kip-connection。我们将不完全的 UV 纹理与随机噪声相结合，并以其镜像作为发生器的输入。

$$L_{gen} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}|I_{i,j} - I^*_{i,j}|, \quad (6)$$

d 由两个判别器组成：全局判别器和局部鉴别器。全局鉴别器来确定整个 UV 图的真实性。此外，还设计了一种局部鉴别器，它专注于人脸中心。引入局部判别器有两个原因：(1)对于野外的 UV 人脸，外部人脸(如耳朵、额头)通常比较嘈杂，不可靠；(2)内部人脸作为身份识别的信息要丰富得多。与全局鉴别器相比，局部模块(图 5(D))以更小的噪声和更锐化的边界增强了中心人脸区域。全局鉴别器与局部鉴别器相结合的优点是：全局鉴别器保持人脸图像的上下文，而局部鉴别器则强制生成的纹理在中心人脸区域内具有更多的信息量。

$$L_{adv} = \mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x}),\mathbf{y}\sim p_d(\mathbf{y})}[\log D(\mathbf{x},\mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z}),\mathbf{y}\sim p_d(\mathbf{y})}[1-\log D(G(\mathbf{z},\mathbf{y}),\mathbf{y})], \quad (7)$$

利用 center loss 提高 UV-GaN 的身份保护能力。具体地，我们根据 ResNet-27 平均池层后的激活来计算(resnet-27 是预先训练，在 CASIA 数据集使用 Softmax 损失分类 10k 身份)：

$$L_{id} = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2,$$

总损失函数：

$$L = L_{gen} + \lambda_1 L_{adv\_g} + \lambda_2 L_{adv\_l} + \lambda_3 L_{id}.$$

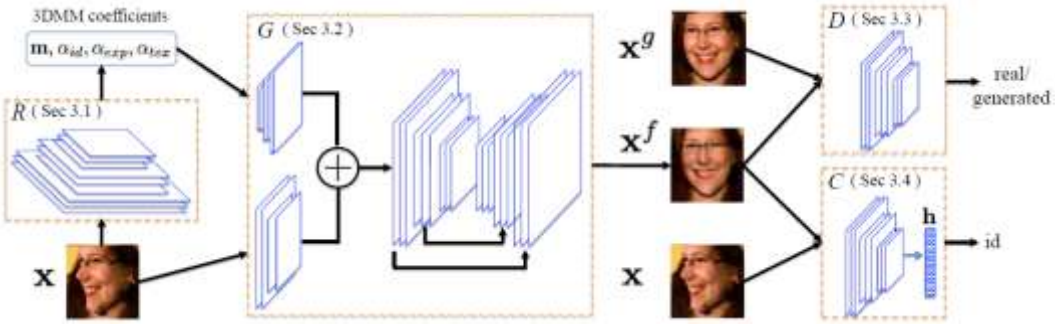
13. Towards Large-Pose Face Frontalization in the Wild

主要思想：

- 提出了一种新的深度三维人脸模型(3DMM)---条件性人脸前端生成辅助网络(GAN)，称为 FF-GAN，在所有姿态范围内，包括极端轮廓视图下，用于人脸的正面化。
- 在 GaN 结构中加入 3 DMM 为快速收敛提供了形状和外观的先导，训练数据少，同时也支持端到端的训练。
- 识别引擎调整所生成的图像以保持身份。
- 不仅采用了判别器和生成器 loss，而且还采用了一种新的掩蔽对称损耗来保持遮挡下的视觉质量，同时还采用了身份损失来恢复高频信息。

主要步骤：

R 是 3 dmm 系数估计的重构模块。G 是合成正面脸的生成模块。D 是做出真实决策或生成决策的判别模块。C 是身份分类的识别模块。



- Reconstruction Module

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp},$$
$$\mathbf{T} = \bar{\mathbf{T}} + \mathbf{A}_{tex}\alpha_{tex}, \quad (1)$$

其中 S 是以平均形状$\bar{S}$、形状偏移量A_{id}和表情偏移量A_{exp}的线性组合计算的三维形状坐标，而 T 是纹理，平均纹理$\bar{T}$和纹理偏移量A_{tex}的线性组合。应用 3dmm 进行人脸对齐，其中使用弱透视投影模型将三维形状投影到 2d 空间。因此基于俯仰、偏航、滚转、比例尺和二维平移，优化了投影矩阵 m∈R2×4，以表示输入人脸图像的姿态。R 则是产生$\mathbf{p} = \{\mathbf{m}, \alpha_{id}, \alpha_{exp}, \alpha_{tex}\}$是 3DMM 的系数。

我们使用基于 CASIA-NET 训练的 CNN 模型进行该回归任务。

$$\min_{\mathbf{p}} L_R = (\mathbf{p}-\mathbf{p}^g)^\top \mathbf{W}(\mathbf{p}-\mathbf{p}^g),$$

- Generation Module

通过编解码网络融合两个输入到生成器 g 的特征，以合成正面 xf=g(x，p)。

$$L_{G_{rec}} = \|G(\mathbf{x},\mathbf{p}) - \mathbf{x}^g\|_1.$$

为了减少块效应，我们使用空间总变异损耗来鼓励发电机输出的平滑性：

$$L_{G_{tv}} = \frac{1}{|\Omega|}\int_{\Omega} |\nabla G(\mathbf{x},\mathbf{p})| dw, \tag{4}$$

基于观察人的面孔在左右两半有自相似性，我们明确地对一个对称损失，我们要求对原始输入图像和翻转版本生成的正面图像应该在各自的掩码中相似：

$$\begin{aligned} L_{G_{sym}} &= \|\mathcal{M}\odot G(\mathbf{x},\mathbf{p}) - \mathcal{M}\odot G(\mathbf{x}_{flip},\mathbf{p}_{flip})\|_2 \\ &+ \|\mathcal{M}_{flip}\odot G(\mathbf{x},\mathbf{p}) - \mathcal{M}_{flip}\odot G(\mathbf{x}_{flip},\mathbf{p}_{flip})\|_2. \end{aligned} \tag{5}$$

其中 xflip 是水平翻转图像输入 x，pflip 是 xflip 的 3DMM 系数，⊙表示对应元素的乘法。

- Discrimination Module

$$\min_D L_D = -\mathbb{E}_{\mathbf{x}^g\in\mathcal{R}}\log(D(\mathbf{x}^g)) - \mathbb{E}_{\mathbf{x}\in\mathcal{K}}\log(1-D(G(\mathbf{x},\mathbf{p}))).$$

$$L_{G_{gan}} = -\mathbb{E}_{\mathbf{x}\in\mathcal{K}}\log(D(G(\mathbf{x},\mathbf{p}))).$$

- Recognition Module

我们使用 CASIA-NET 结构用于识别引擎 C，具有用于训练 C 的交叉熵损失，以将图像 X 与真实身份 Y 进行分类：

$$\min_C L_C = \sum_j -y_j\log(C_j(\mathbf{x})), \tag{8}$$

iv. template 聚合

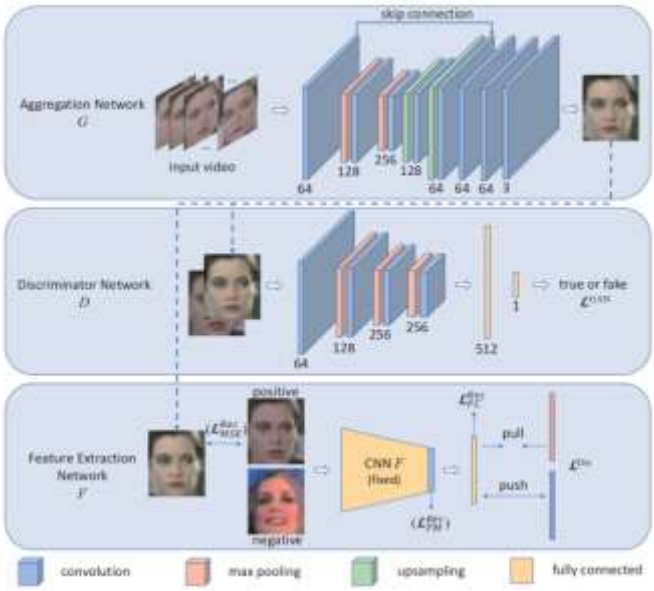
14. Learning Discriminative Aggregation Network for Video-based Face Recognition

主要思想：

- 提出了一种用于视频人脸识别的识别聚合网络 we propose a discriminative aggregation network (DAN)方法，该方法旨在有效地、有效地集成视频帧的信息。
- 与现有的聚合方法不同，我们的方法直接对原始视频帧进行聚合，而不是处理获得的特征。
- 通过结合度量学习和对抗学习的思想，我们学习了一个聚合网络，与原始输入帧相比，产生更多的判别合成图像。
- 减少了要处理的帧的数量，并且大大加快了识别过程。此外，劣质的帧包含误导信息过滤和去噪中聚集的过程，使系统具有更强的鲁棒性和区分性

主要步骤：

DAN 由 3 个子网络组成。将它们定义为聚合（生成器）网络 g、鉴别器网络 d 和特征提取网络 F。我们将整个视频表示为 V。实施的难易程度，在每一次我们将 V 的一个子集 S 聚合到一个单一的图像，所以 G 的输入是一个子集 S，输出是一个单一的判别图像 X。鉴频器 D 试图判断图像是由 G 或选择产生的原始视频，形成对抗学习 G.特征发生器网络提取特征集合图像，并试图使特征在特征空间的判别。



- Discriminative Loss

使用（x，p）的正对和（x，n）为负对，其中 x 是聚集图像，p 和 n 分别是从其他视频剪辑中随机选择的正和负样本。

$$\mathcal{L}^{Dis} = \begin{cases} (||F(X) - F(P)||^2 - \alpha)_+ & y = 1 \\ (\beta - ||F(X) - F(N)||^2)_+ & y = 0 \end{cases} \quad (3)$$

and

$$\alpha = \min_{A \in S} ||F(A) - F(P)||^2 \quad (4)$$

其中 y 是表示正或负对的 1 或 0，f 是特征提取网络。s 是聚合的子集剪辑，a 是其中的一个帧。我们使用欧氏距离度量两个特征表示之间的距离。α 是 s 和 p 中所有帧之间最小的距离。β 是一种手动设置。在 f 的特征空间中，聚集图像 x 比来自原始视频子集 S 的任何其他帧更接近于 p。反之，如果考虑负样本，我们希望生成的 x 和 n 之间的距离大于一定的余量。

- Reconstruction Loss

① Pixel-wise MSE loss

$$\mathcal{L}_{MSE}^{Rec} = \frac{1}{N_I} ||I - X||_{\mathcal{F}}^2 \quad (5)$$

I 是原始图像，x 是重建图像。NI 是图像中总像素的个数。

② 特征图谱的差异

$$\mathcal{L}_{FM}^{Rec} = \frac{1}{N} \sum_{i=1}^{n} ||\phi_i(I) - \phi_i(X)||_{\mathcal{F}}^2 \quad (6)$$

我们不能直接地定义上述两种形式的损失，因为在输入端有多个图像用于执行，我们根据以下规则选择 i：

选择正样本中距离最近的，负样本中距离最远的

$$I = \begin{cases} \text{argmin}_{A \in S} ||F(A) - F(P)||^2 & y = 1 \\ \text{argmax}_{A \in S} ||F(A) - F(N)||^2 & y = 0 \end{cases} \quad (7)$$

以上损失保证视觉特征而不是语义信息或辨别力：

$$\mathcal{L}_{FC}^{Rec} = ||F(X) - \text{mean}(F(V^m))||^2 \quad (8)$$

f 是如上所述的特征提取网络，Vm 是由 m 帧组成的原始视频。我们希望重建图像的特征接近于每帧 V 提取的特征均值，以减少类内距离。

- Adversarial Loss

$$\begin{aligned} \mathcal{L}^{GAN} &= \mathbb{E}_{A \sim p_{\text{train}}(A)}[\log D(A)] \\ &+ \mathbb{E}_{V^m \sim p_{\text{train}}(V^m)}[\log(1 - D(G(V^m))] \end{aligned} \quad (9)$$

- 总损失

$$\mathcal{L} = \lambda \mathcal{L}^{Dis} + \eta \mathcal{L}^{Rec} + 0.01 \mathcal{L}^{GAN} \quad (2)$$

15. Pooling Faces: Template based Face Recognition with Pooled Face Images

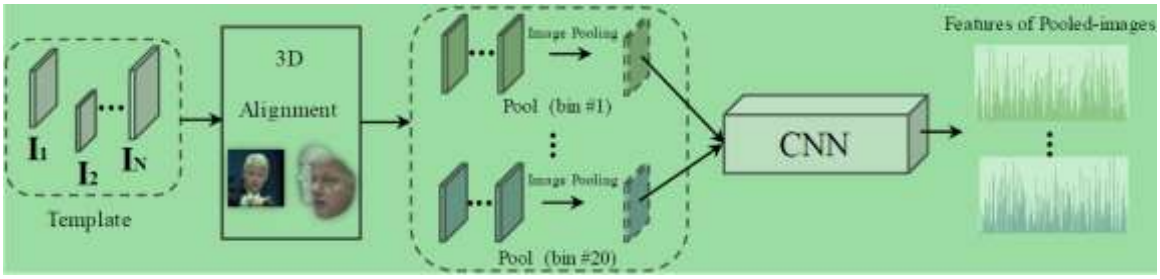
主要思想：

- 本文将介绍一些最完善的图像处理和计算机视觉原理——减少存储和计算的图像平均，以及提高图像质量——寻求一种更简单的方法来表示人脸图像集。
- 通过在 3D 中对齐人脸并根据头部姿态和图像对它们进行分区，pooling 提供了一种令人惊讶的有效且计算效率高的表示和匹配人脸 template 的方法。

主要步骤：

给出了人脸模板 f，我们对其图像进行三维对齐，然后根据图像的姿态和质量对对齐图像进行装箱。掉进同一个 bin 的图像采用下式进行，合并的图像采用卷积神经网络(CNN)进行编码。最后，我们使用 cnn 的这些特性来匹配模板。

$$\mathbf{F} \doteq avg(\mathcal{F}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}_i$$



- Binning by head pose

给出输入图像，检测 68 个 landmarks。同时检测通用 3d 模型的渲染图像中的 landmarks。获得通用模型与呈现视图之间的三维坐标之间的对应关系。因此，给定所检测到输入图像和渲染图像的 68 个点，可以得到通用模型上的 68 个对应的点。

计算 a camera matrix M，得到头部的偏角、俯仰角和滚角。这三个角度分别用于滚动补偿、头部姿态量化和姿态消除。滚动补偿只是指面内对齐，使两眼之间的线是水平的。

姿态量化则把姿态划分为$\{(0^\circ \le |\theta| < 20^\circ), (20^\circ \le |\theta| < 40^\circ), (40^\circ \le |\theta| < 60^\circ), (60^\circ \le |\theta|\}$.

头部姿态消除将所有 bin 中的图像都在 3d 内对齐，以消除任何剩余的姿态变化。

- Binning by image quality

标准化的 SSEQ 评分将图像划分为五个图像质量 bin:

$$Q(\mathbf{I}) = \begin{cases} 0, & \text{if } -\infty < SSEQ(\mathbf{I}) < 0.45 \\ 1, & \text{if } 0.45 \le SSEQ(\mathbf{I}) < 0.55 \\ 2, & \text{if } 0.55 \le SSEQ(\mathbf{I}) < 0.65 \\ 3, & \text{if } 0.65 \le SSEQ(\mathbf{I}) < 0.75 \\ 4, & \text{if } 0.75 \le SSEQ(\mathbf{I}) < \infty, \end{cases} \quad (2)$$

- Representing and comparing templates

使用 vgg-19 编码人脸图像。这 19 层网络在 ILSVRC 初始化。然后在分两次微调这个网络的权重：首先在 CASIAWebFace images 上进行微调，随着识别 10，575 身份标签。第二次微调使用 CASIA 图像再次执行。这一次，使用 pooling 图像进行训练。CASIA 没有模板定义，所以使用身份标签：从同一主题获取随机子集的图像，然后将每个子集视为模板。

给出 probe 模板 p 中的图像 IP 和 gallery 模板 g 中的 Ig，通过它们特征向量的 normalized cross correlation (NCC)，计算它们的相似性$s(\mathbf{x}_p^{fc7}, \mathbf{x}_g^{fc7})$。

在计算了所有成对的pooling图像的相似性分数之后，这些值被使用softmax$s_\beta(\mathcal{P},\mathcal{G})$进行融合：

$$s_\beta(\mathcal{P},\mathcal{G}) = \frac{\sum_{p\in\mathcal{P},g\in\mathcal{G}} w_{pg}\, s(\mathbf{x}_p,\mathbf{x}_g)}{\sum_{p\in\mathcal{P},g\in\mathcal{G}} w_{pg}}, \quad w_{pg} \doteq e^{\beta\, s(\mathbf{x}_p,\mathbf{x}_g)} \quad (3)$$

16. Neural Aggregation Network for Video Face Recognition

主要思想：

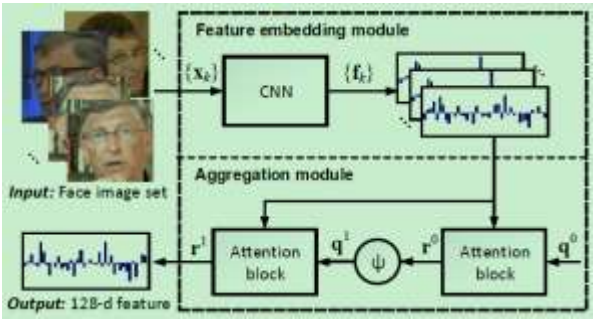
- 提出了一种用于视频人脸识别的神经聚合网络(NAN)。该网络以一个人的面部视频或人脸图像集作为输入，并生成一个紧凑的固

定维特征表示。

- 整个网络由两个模块组成。特征嵌入模块是一种深度卷积神经网络(CNN)，它将人脸图像映射成一个特征向量。聚合模块由两个由内存驱动的注意块组成，存储所有提取的特征。

主要步骤：

所有输入的人脸图像$\{x_k\}$由一个带有 CNN 的特征嵌入模块处理，产生一组特征表示$\{f_k\}$。这些特征被传递到聚合模块，为输入视频脸生成 128 维向量表示 r1。这种紧凑的表示法用于识别。



- Feature embedding module

采用 GoogLeNet，利用 Batch Normalization (BN)。Googlet 产生 128 维的图像特征，首先将其归一化为单位矢量，然后将其馈送到聚合模块中。

- Aggregation module

我们的目标是利用视频中的所有特征向量生成一组线性权重$\{a_k\}_{k=1}^{t}$，从而使聚合的特征表示成为：

$$\mathbf{r} = \sum_k a_k \mathbf{f}_k.$$

首先，模块应该能够处理不同数量的图像。其次，聚合应该与图像顺序无关。第三，在标准的人脸识别训练任务中，模块要适应输入人脸，并具有可通过监督学习进行参数训练的功能。

- Attention blocks

让$\{f_k\}$作为人脸特征向量，然后注意块通过点积用核 q 对它们进行滤波，得到一组相应的意义$\{e_k\}$。然后，将它们传递给 Softmax 运算符以生成权重$\{a_k\}$。

$$e_k = \mathbf{q}^T \mathbf{f}_k$$
$$a_k = \frac{\exp(e_k)}{\sum_j \exp(e_j)}$$

单注意块-通用人脸特征质量测量。我们首先尝试使用一个注意块进行聚合。在这种情况下，向量 q 是要学习的参数。它与单一特征 f 具有相同的大小，是衡量面部特征质量的通用先验工具。

级联两个注意块-内容感知聚合。设q^0为第一注意块的核，r^0为q^0的聚合特征。我们自适应地计算q^1：

$$\mathbf{q}^1 = \tanh(\mathbf{W}\mathbf{r}^0 + \mathbf{b})$$

由q^1生成的特征向量r^1将是最终的聚合结果。因此，（q^0；W；b）是聚集模块的训练参数。

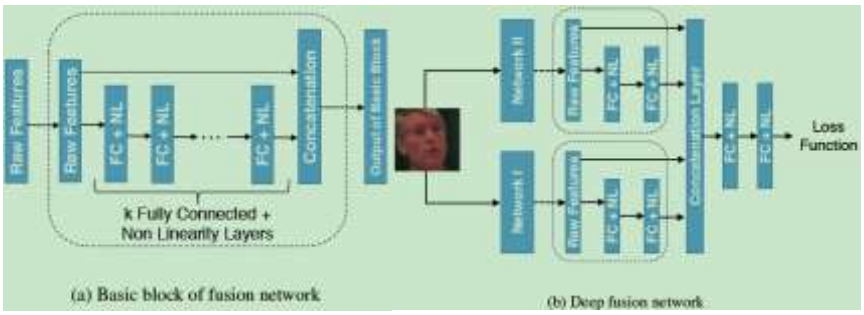
17. Deep Heterogeneous Feature Fusion for Template-Based Face Recognition

主要思想：

- 提出了一种深度异构特征融合网络，利用不同深度卷积神经网络(Dcnns)生成的特征中的互补信息进行基于模板的人脸识别。
- 该方法有效地融合了不同深度特征的判别信息通过，1）联合学习深层特征的非线性高维投影,；2)生成了一种更具有鉴别性的模板表示，保留了特征空间中深层特征的固有几何特征。

主要步骤：

网络一和网络二是两个预先训练的 dcnn，产生两个特征称为原始特征($\mathbf{x} \in \mathbb{R}^{d_1 \times 1}$和$\mathbf{y} \in \mathbb{R}^{d_2 \times 1}$，其中 d1 和 d2 是由网络一和二产生的特征的尺寸)。网络一作为利用紧密的边框作为输入，而网络二则以松的边界框中的面孔作为输入，从而包含更多的背景上下文信息。利用深度神经网络将这两个互补的原始特征融合到非线性高维空间中。深度融合网络产生了融合特征向量$\mathbf{z} \in \mathbb{R}^{d \times 1}$，其中 d 是融合特征向量的维数。



- Basic Building Block of Fusion Network

该基本融合块以原始 dcnn 特征作为输入，由全连接层和非线性激活函数执行一系列非线性高维投影后，生成级联特征。对于这种网络，输入是从预先训练的 dcnn 中提取的原始特征 x，非线性激活函数后隐藏层的输出是输入 x 的非线性高维投影ϕ(x)。如果 x 和(X)是连在一起的，则级联特征向量 z(其中$\mathbf{z}^\intercal = [\mathbf{x}_i^\intercal \quad \phi(\mathbf{x}_i^\intercal)]$)的对偶形式对应于核的组合：

$$\begin{aligned} k(\mathbf{z}_i, \mathbf{z}_j) &= \mathbf{z}_i^\intercal \mathbf{z}_j \\ &= [\mathbf{x}_i^\intercal \quad \phi(\mathbf{x}_i^\intercal)] \begin{bmatrix} \mathbf{x}_j \\ \phi(\mathbf{x}_j) \end{bmatrix} \\ &= k_1(\mathbf{x}_i, \mathbf{x}_j) + k_2(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) \end{aligned}$$

其中$k_1(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\intercal \mathbf{x}_j$是对应于原始特征的核，$k_2(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) = \phi(\mathbf{x}_i)^\intercal \phi(\mathbf{x}_j)$是对应于原始特征的非线性高维投影的核，是这两个核的组合。然后利用 hinge loss 来训练这三层网络，以及从级联中提取的结果特征可以用来人脸识别。

- Heterogeneous Fusion of Deep Features

每个原始 dcnn 特征首先通过一个基本块来生成一个级联的特征向量。基本块的输出被进一步连接，以生成一个高维特征向量。如果每个基本块都有 k 个完全连通的层，那么所有的 k 输出都可以通过将它们作为高维投影来连接。这里是一个超参数，在我们的例子中，我们把它设为 1，也就是说，我们只连接每个基本块的最后一个完全连通层的输出。

我们将两个网络的特征进行融合，并将相应的四个特征串联成高维表示。

$$\mathbf{z}^\intercal = [\mathbf{x}^\intercal \quad \mathbf{y}^\intercal \quad \phi_1(\mathbf{x}^\intercal) \quad \phi_2(\mathbf{y}^\intercal)]$$

考虑两个版本的融合网络：1)直接在级联层之上训练 hinge loss(HL)的网络；2)在级联层顶部增加两个完全连接的层和使用 Softmax loss(SI)训练的网络。

- Template Representative Fused Feature

由于最终融合向量是通过深度神经网络学习的，因此它可以用输入特征 x 和 y 的函数表示为 z=f(x，y)，其中 f 是非线性函数(即本文中的融合网络)。考虑一个有 n 个人脸(图像加帧)的模板。$\mathbf{x}_1, \ldots, \mathbf{x}_n$是网络一的原始特征，$\mathbf{y}_1, \ldots, \mathbf{y}_n$是网络二的原始特征，首先对原始特征进行 average pooling，然后对平均原始特征进行融合，然后利用平均集合特征 x 和 y 作为融合网络的输入，得到具有模板代表性的融合特征向量 z。

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i, \bar{\mathbf{y}} = \sum_{i=1}^{n} \lambda_i \mathbf{y}_i, \bar{\mathbf{z}} = f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \qquad (4)$$

为了评估两个模板 A 和 B 之间的相似性，我们计算了各自的模板表示融合特征$\bar{z}_a$和$\bar{z}_b$之间的余弦距离。

2) 一对多

i. 利用 CNN 合成 3D 模型（或传统 3D 合成再用 CNN 识别）

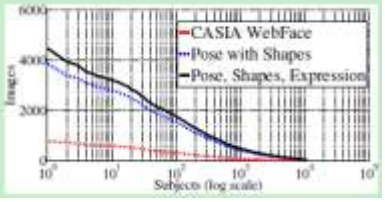
18. Do We Really Need to Collect Millions of Facesfor Effective Face Recognition?

主要思想：

- 采用了新的人脸数据扩增方法。对现有公共数据库人脸图像，从 pose，shape 和 expression 三个方面合成新的人脸图像，极大的扩增数据量。
- 在 LFW 和 IJB-A 数据集上取得了和百万级人脸数据训练一样好的结果。

主要步骤：

Dataset	#ID	#Img	#Img/#ID
Google [20]	8M	200M	25
Facebook [19]	4,030	4.4M	1K
VGG Face [24]	2,622	2.6M	1K
MegaFace [12]	690,572	1.02M	1.5
CASIA [34]	10,575	494,414	46
Aug. pose+shape	10,575	1,977,656	187
Aug. pose+shape+expr	10,575	2,472,070	234

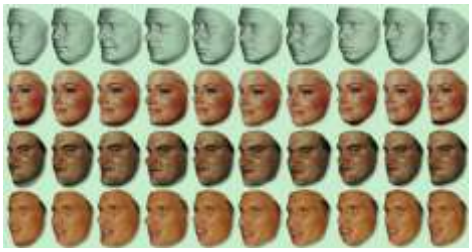


- Pose（姿态，文章中为人脸角度，即通过 3d 人脸模型数据库合成图像看不见的角度，生成新的角度的人脸）。

首先，通过人脸特征点检测，获取人脸特征点。根据人脸特征点和开放的 Basel 3D face set 数据库的人脸模板合成 3d 人脸。

- shape（脸型）

首先，通过 Basel 3D face 获取 10 种高质量 3d 面部扫描数据。再将图像数据与不同 3d 脸型数据结合，生成同一个人不同脸型的图像。



- expression（表情，本文中，将图像的张嘴表情替换为闭口表情）

采用中性嘴型将图像中的开口表情换位闭口表情。根据三维人脸模型和二维人脸图像在姿态和表情上的对齐，基于图像的纹理映射，将人脸纹理映射到模型上。

- Augmented training data
  1. 原始 CASIA 图像：大致正面图（30 度到负 30 度），使用理想的正面模板上的 9 个特征点对齐；而侧脸图像图像(所有其他偏角)则使用可见的眼睛和鼻尖对齐。
  2. CASIA 中的每一幅图像生成不同的视角。
  3. 随机选择的 3D 通用人脸模型的人脸形状产生的基础，从而增加形状变化。
  4. 表情变化也加入到训练。
- CNN fine-tuning

  使用 19 层的 VGGNET，采用 softmax loss，在学习权重上使用标准 L2 范数的随机梯度下降(SGC)进行优化。
- General matching process

  给定两个输入图像 IP 和 IQ，它们的相似性 $s(\mathbf{x}_p, \mathbf{x}_q)$ 就是它们的特征向量的归一化互相关(NCC)。当有多幅图像计算相似性的时候，采用 SoftMax operator，相比于取最大值、最小值和平均值来说，效果较好。

$$s_\beta(\cdot,\cdot) = \begin{cases} \max(\cdot) & \text{if } \beta \to \infty \\ \text{avg}(\cdot) & \text{if } \beta = 0 \\ \min(\cdot) & \text{if } \beta \to -\infty \end{cases} \quad \text{and } s_\beta(\mathcal{P},\mathcal{Q}) = \frac{\sum_{p\in\mathcal{P},q\in\mathcal{Q}} s(\mathbf{x}_p,\mathbf{x}_q) e^{\beta\, s(\mathbf{x}_p,\mathbf{x}_q)}}{\sum_{p\in\mathcal{P},q\in\mathcal{Q}} e^{\beta\, s(\mathbf{x}_p,\mathbf{x}_q)}}.$$

$$s(\mathcal{P},\mathcal{Q}) = \frac{1}{21}\sum_{\beta=0}^{20} s_\beta(\mathcal{P},\mathcal{Q}).$$

19. Rapid Synthesis of Massive Face Sets for Improved Face Recognition

主要思想：

- 生成人脸的新视图在无约束视图中。
- 将使用通用的三维人脸并对固定视图进行渲染的过程所需的大部分计算工作作为预处理，可以在预处理时执行渲染人脸。允许

快速生成巨大的人脸集，为 CNN 提供外观变化。

主要步骤：

- Precomputing output projections

最耗时的步骤之一是 ray casting：计算穿过每个输出像素的射线与脸部表面之间的交点位置。当使用固定的通用人脸形状和输出视图时，这些步骤只需要在预处理时执行一次。后续的人脸合成使用相同的形状和姿态，可以跳过这一步，并且需要与标准图像在 2d 内进行纹理映射一样的计算量。

使用标准渲染引擎在期望的输出视图 j 上执行一般人脸形状 f 的光线投射。为每个输出的 2d 像素位置 pi∈j 存储投影到该像素上的 3D 坐标 Pi∈f。此信息存储在查找表 u 中，只需将其定义为：

$$\mathbf{U}(\mathbf{p}_i) = \mathbf{P}_i$$

- Rendering with precomputed projections

给定一个输入图像 i，使用 u 将其呈现到所需的新视图。首先估计面部的三维姿态，提供了一个摄像机矩阵 Mi，将 f 表面的 3D 点与 i 中的像素相关联。

$$\bar{\mathbf{q}} = \mathbf{M}^I \mathbf{U}_i$$

- Preparing generic 3D heads and backgrounds

通过将 10 个三维模型拼接到一个包含头部、耳朵和颈部的通用三维结构中，并添加一个平面表示平面背景来实现的。

- Training with rendered face images

使用 vggnet 在大规模图像识别基准(ILSVRC)上进行预训练。使用 CASIA 对 cnn 进行了微调，通过使用相似变换(即平面内对齐)对齐人脸，并合成三种视图。采用 softmax 进行优化。

- Pooling across Synthesized Views

pooling 从图像 i 及其呈现的视图中获得的特性。

20. 3D Face Reconstruction by Learningfrom Synthetic Data（没太明白）

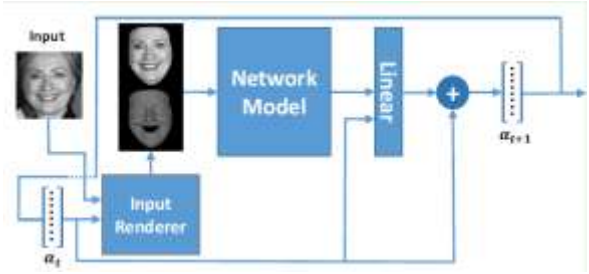
它们的网络只能产生粗几何，必须给出一个对齐的模板模型作为初始化。这些限制迫使它们的解决方案依赖于外部算法的姿态对齐和细节细化。（见 Learning Detailed Face Reconstruction from a Single Image）

主要思想：

- 提出一种基于学习的单一图像三维人脸重建方法
- 采用了一个用合成数据训练的迭代卷积神经网络。作为一个可选的细节细化步骤，应用了一种形状自着色算法。
- 先利用传统方法将 2D 生成 3D 图像，作为监督信号，然后将生成的图像和 t 时刻的 shading 图像联合输入 CNN，通过与监督 3DMM 信号 MSE 损失，进行梯度下降优化。将输出的 3DMM 参数与 t 的参数对比，生成 t+1 时刻的 3DMM 参数，然后将新的参数生成 t+1 时刻的 shading 图像，在进行梯度下降优化，过程重复三次。等于分三阶段学习，逐步优化。

主要步骤：

迭代误差反馈(IEF)的核心思想是使用辅助输入信道将前一个网络的输出表示为图像。然后，可以根据原始输入和辅助信道对网络进行训练以纠正先前的预测。



在计算过程中，我们使用平均形状初始化几何参数，设置 α0=0，然后利用该几何参数创建阴影图像，并从背景中屏蔽输入图像。每次迭代时，都会预测出一个新的几何向量 αt，用于更新阴影图像和隐藏输入图像。该程序重复 3 次，迭代地改进了掩蔽和重建。在输入图像中掩蔽人脸的目的是简化数据生成过程。这样我们只需要精确合成脸部本身。

- Training Criterion

采用 MSE 准则：

$$L(x,y) = \left\| [A_{id}|A_{exp}]\, x - [A_{id}|A_{exp}]\, y \right\|_2^2, \quad (3)$$

- Data Generation

给定纹理的三维几何，可以渲染合成图像。首先，随机阴影需要绘制，以创造照片真实感的结果.。我们使用 Phong 反射率用于模拟阴影。在我们的配方中，光泽常数固定在 10，而周围的、漫射的和镜面的常数通常是在平均值[0.5，0.7，0.05]。光的方向是从均匀分布中画出的，覆盖了所有可能的前角。

然后将阴影几何投影到图像平面上，并进行平行弱透视投影。

$$\begin{bmatrix} p_x \\ p_y \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \end{bmatrix} [R|t] \begin{bmatrix} P_x \\ P_y \\ P_z \\ 1 \end{bmatrix}, \quad (4)$$

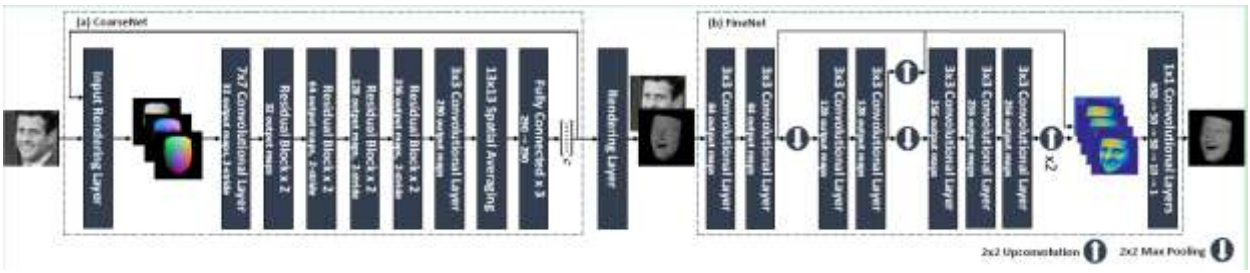
21. Learning Detailed Face Reconstruction from a Single Image

主要思想：

- 提出了一种基于单幅图像的细节人脸重建的端到端方法。
- 该方法由两个主要模块组成，一个用于恢复对人脸几何形状的粗略估计的网络，然后是一个精细细节重构网络。前者采用合成图像训练，后者采用端到端无监督训练方案中的真实人脸图像训练。
- 为了将这两个网络连接起来，引入了一个 differentiable rendering 层。

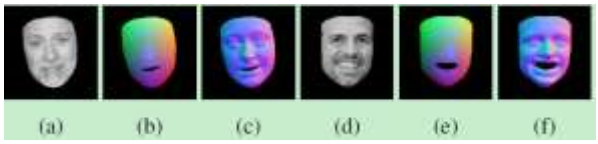
主要步骤：

首先，每一幅通过 CoarseNet 四次迭代，创建粗糙的几何表示。然后，渲染层将 3dmm 表示转换为深度图。最后，将深度图与原始输入图像一起通过 Finenet 进行传播，得到密集更新的深度图。



- The CoarseNet Training Framework

采用迭代网络，将网络的反馈通道添加到表示一个特征映射集的网络以前的输出，然后训练网络基于原始输入和反馈通道对其预测进行改进。特别地,希望反馈通道能够显示地表示粗糙面部几何的当前状态,采取的反馈的特征图包括两个:the corresponding PNCCs of the network’ soutput and normal maps。



输入层包括反馈通道，并且使用了灰度图像。应用几何均方误差(GMSE)

$$L(\hat{\alpha},\alpha) = \left\| [A_{id}|A_{exp}]\, \hat{\alpha} - [A_{id}|A_{exp}]\, \alpha \right\|_2^2, \quad (4)$$

- Acquiring the Data

通过绘制几何和姿态的随机表示来创建一个合成数据集，然后使用随机纹理、光照和反射率来呈现r_{gt}。这个过程提供了一个二维图像数据集，其中的姿态和相应的几何形状是已知的。然后，通过绘制另一组参数r_t，在r_{gt}和随机参数集r_{nd}之间采样，模拟迭代精化过程。

$$r_t = \beta \cdot r_{gt} + (1-\beta) \cdot r_{rnd}, \qquad 0 \le \beta \le 1, \quad (3)$$

网络进行训练预测真实标签，r_{gt}，从目前的一个表示特征，r_t。

- The Rendering Layer

Rendering Layer 输入为几何和姿态的特征表达，输出为 200*200 的深度图像。

- FineNet Framework

复杂的面部特征如皱纹和凹坑难以由 3DMM 的低维空间来表示，主要是由于它们的高维。FineNet Framework 的主要思想是生成一个每个像素的特征映射，它既包含了结构数据，也包含了语义数据。这是通过连接网络路径上来自几个卷积层的输出响应来实现的。

由于池层，内部层的输出映射大小不匹配输入图像的大小，因此，它们被内插回原来的大小，以创建一个密集的每像素体积的特征。然后，由几个 1*1 的卷积层处理这个卷，以创建最终的预测。

提出了一个无监督学习过程，将输出深度图与 2d 图像相关联的测量方法。为此目的，我们采取从阴影形状(SFS)。

22. End-to-end 3D face reconstruction with deep neural networks

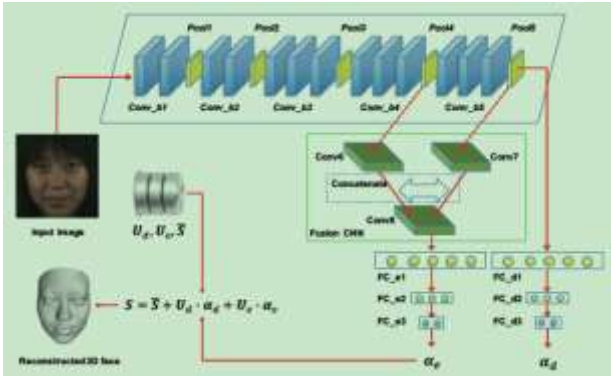
（写了 3DMM 的局限性）

主要思想：

- 该方法以端到端推理方案代替迭代模型参数更新方案，消除了对三维形状绘制或初始模型参数的依赖，增加了模型的输入，使框架得到了极大的简化。
- 还将两个关键组件引入到我们的框架中,即融合-CNN 和多任务学习丢失。两个组件,我们将 3D 人脸重建分为两个子任务,即中性 3D 面部形状重建和表情 3D 面部形状重建,并在针对这两个特定任务的单个 DNN 模型中训练不同类型的神经层。

主要步骤：

它是基于 VGG 人脸模型，由 13 个卷积层和 5 个 pooling 层。增加了两个关键部件：fusion-CNN，融合 VGG 脸中间层特征；一个多任务学习损失函数，用于身份参数预测和表达式参数预测。因此，可以在一个单一的 DNN 架构训练三种类型的神经层：第一类包括那些低于第四 pooling 层（pool4），其中学习通用特征对应的低层次的面部结构，如棱角。这些层由两个任务共享。第二类神经层包括 fusion-CNN 和之后的完全连接层。这些层学习特定的表情特征。第三类包括第四 pooling 层（pool4）以上学习身份特征。



- 多任务学习损失函数

$$E_e = \|U_e \cdot \hat{\alpha}_e - U_e \cdot \alpha_e\|_2^2 , \quad (2)$$

$$E = \lambda_d E_d + \lambda_e E_e , \quad (3)$$

- 合成数据

提出了用真实的二维图像和合成的二维图像来训练深层神经网络。用真实的二维图像初始化深层神经网络，用合成的二维图像进行微调。

使用两种三维面部形状模型，即 BFM 模型和 AFM 模型。使用随机参数创。为 BFM 和 AFM 三维人脸模型创建 10000 个中性三维人脸及其相应的面部纹理，对应于 10000 个身份。对于每个三维人脸，我们合成了 25 幅不同的面部姿态、光照和表情的图像。在多个 2d 人脸数据库上收集了一组非常大的表达式参数并随机抽样，通过改变表情参数生成各种面部表情。

在 3D 渲染过程中，适当地控制相机参数和照明。我们使用透视相机模型并将相机视场随机设置为在$[15°, 35°]$;的范围内。因此，相机与对象之间的距离设置为 1,900mm 和 500mm。

使用 Phong 反射模型合成光照。从二维人脸数据库收集了一大套光泽参数并随机样本。

对于环境参数、漫射参数和镜面参数，我们在[0.2，0.4]，[0.6，0.8]，[0.1，0.2]范围内使用随机值。合成图像的面部姿势是随机生成的。偏航、俯仰和滚转在[-90，90]，[-30，30)，和[-30%，30%]范围内均匀分布。合成图像的背景也是随机生成的。

23. 3DFaceNet: Real-time Dense Face Reconstructionvia Synthesizing Photo-realistic Face Images

解释了为什么不直接使用合成 3D 图片，而使用 CNN 进行建模

主要思想：

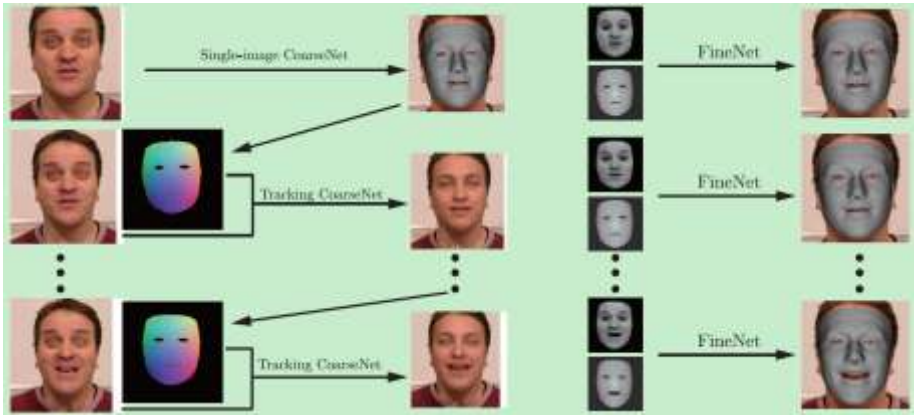
- 合成视频三维重建
- 最先进的技术是使用粗糙的面部模型来合成这些数据，然而，这种模型很难生成面部(有皱纹)的详细照片写实图像。本文提出

了一种新的人脸数据生成方法。具体来说，我们在逆向渲染的基础上渲染了大量具有不同属性的照片真实感人脸图像.。此外，通过将不同尺度的细节从一幅图像转移到另一幅图像，构造了一个精细的人脸图像数据集。通过模拟真实视频数据的分布，构造了大量的视频类型相邻帧对。

- 利用这些构造良好的数据集，我们提出了一个由三个卷积网络组成的粗到精细的学习框架。该网络被训练用于从单目视频以及从单个图像中实时精细的重建三维面部。

主要步骤：

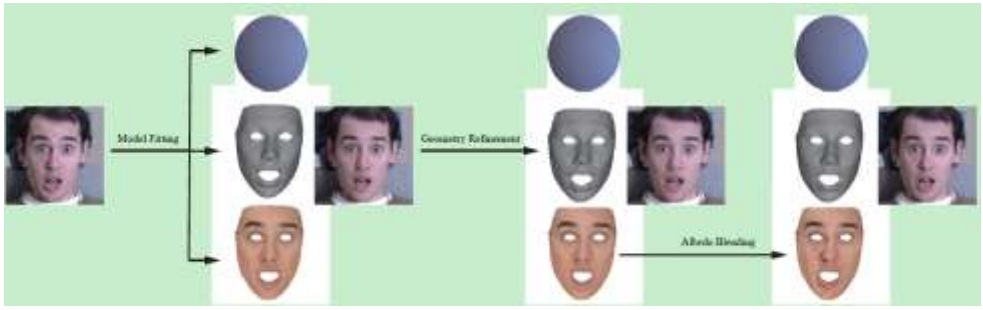
每帧使用两个 CNN，即 CoarseNet 和 FineNet。第一个是估计粗尺度几何、反照率、照明和姿态参数，第二个是重建像素级编码的精细几何图形。有两种 CoarseNet：Single-image CoarseNet 和 TrackingCoarseNet。TrackingCoarseNet 利用前一个帧的预测参数，而 Single-image CoarseNe 用于第一个帧情况，其中没有前一个帧可用。请注意，整个框架可以很容易地退化为从单一的图像重建三维人脸，通过组合 Single-image CoarseNet 和 Finenet。



- Inverse rendering for Single-image CoarseNet

逆绘制由参数化人脸模型拟合、几何细化和反照率混合三个阶段组成。第一阶段是基于参数面模型恢复光照、粗糙几何和反照率。第二阶段是进一步恢复几何细节。第三阶段是混合反照率，使渲染的图像更接近输入图像。

从数据集 300W 选择不遮挡人脸的 4000 张人脸图像。对于 4000 幅图像，我们采用基于优化的逆绘制方法获得参数集，通过随机改变姿态参数 p 和表达式参数 exp 来呈现新的人脸图像。



- Single-image CoarseNet

对我们的单图像粗图像的输入是面部图像,并且输出是与 3D 面部和投影的形状相关的参数,即$T = \{\alpha_{\mathrm{id}}, \alpha_{\mathrm{exp}}, s, pitch, yaw, roll, t_x, t_y\}$。该网络基于 RESNET-18，将完全连接层的输出数修改为 185(100 表示同一性，79 表示表示，3 表示旋转，2 表示平移，1 表示缩放)。

我们使用一个损失函数，在每个像素级别计算真实参数 TG 和网络输出参数 TN 之间的距离：

$$Proj(\mathcal{T}) = \Pi R(\bar{\mathbf{p}}_q + \mathbf{A}_{q,\mathrm{id}}\boldsymbol{\alpha}_{\mathrm{id}} + \mathbf{A}_{q,\mathrm{exp}}\boldsymbol{\alpha}_{\mathrm{exp}}) + \mathbf{t}. \quad (11)$$

$$\mathcal{D}(\mathcal{T}_g, \mathcal{T}_n) = \|Proj(\mathcal{T}_g) - Proj(\mathcal{T}_n)\|_2^2$$

为了更好地收敛，我们进一步分离了方程中的损失：

$$\mathcal{L}_{\mathrm{pose}} = \|Proj(\mathcal{T}_g) - Proj(\mathcal{T}_{n,\mathrm{pose}}, \mathcal{T}_{g,\mathrm{geo}})\|_2^2,$$

$$\mathcal{L}_{\mathrm{geo}} = \|Proj(\mathcal{T}_g) - Proj(\mathcal{T}_{n,\mathrm{geo}}, \mathcal{T}_{g,\mathrm{pose}})\|_2^2,$$

- Tracking CoarseNet

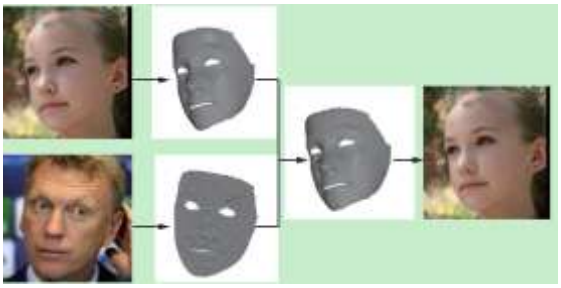
输入到跟踪网络的是 k 帧人脸，k-1 帧地标和投影归一化坐标码(Pncc)（从 k-1 帧中姿态参数生成）和平均脸参数 p。跟踪网络的输出是参数$\{\alpha_{\mathrm{id}}^k, \alpha_{\mathrm{exp}}^k, \alpha_{\mathrm{alb}}^k, \delta^k(s), \delta^k(pitch), \delta^k(yaw), \delta^k(roll), \delta^k(\mathbf{t}), \mathbf{t}^k\}$，其中δ()表示当前帧与前一个帧之间的差异。

- Training data generation for Tracking CoarseNet

模拟相邻的视频帧，即为单图像粗化训练中使用的 80000 幅合成图像中的每一帧生成前帧。假设参数$\delta^k(\chi_p) = \chi_p^{k-1} - \chi_p^k$和$\delta^k(\chi_t) = \chi_t^{k-1} - \chi_t^k$符合正态分布。我们从 300-vw 视频数据集中提取了大约 160000 个相邻帧对。并利用我们的单图像粗化得到拟合正态分布的参数。最后，根据得到的正态分布，对于 80000 幅合成的图像，我们可以通过生成$\tilde{\chi}_p^{k-1}$ and $\tilde{\chi}_t^{k-1}$来模拟其上一帧。

- Constructing FineNet Training Data

利用一个没有很多几何细节的目标人脸图像(左上角)和一个布满皱纹的源人脸图像(左下角)。我们首先将我们开发的反绘制应用于两幅图像上，以获得目标面的投影几何图形(中上角)和源面的位移图(中下角)。然后将源面的位移图转换为目标面的几何形状。最后，我们渲染更新的几何形状，以获得一个新的人脸图像(右上)，其中包含与源脸相同类型的皱纹。



24. Regressing Robust and Discriminative 3D Morphable Models with a very Deep Neural Network

主要思想：

- 当在真实场景中应用 3d 模拟来增加人脸识别精度，存在两类问题：要么 3d 模拟不稳定，导致同一个个体的 3d 模拟差异较大；要么过于泛化，导致大部分合成的图片都类似。
- 采用了卷积神经网络（CNN）来根据输入照片来调节三维人脸模型的脸型和纹理参数。
- 文章的关键点有两个：一，3D 重建模型训练数据获取；二，3D 重建模型训练 。

主要步骤：

- Single image 3DMM fitting

对于图像 I，我们估计α^*和β^*来表示与输入图像 I 类似的图像。采用了目前最好的人脸特征点检测器（CLNF）来检测 K=68 个人脸特征点和置信值ω。其中，脸部特征点用于在 3DMM 坐标系中初始化输入人脸的角度。角度表达为 6 个自由度：角度$r = [r_\alpha, r_{\beta\alpha}, r_{\gamma\alpha}]$和平移$t = [t_X, t_Y, t_Z]$然后再对脸型，纹理，角度，光照和色彩进行处理。

$$\mathbf{S}' = \widehat{\mathbf{s}} + \mathbf{W}_S\,\boldsymbol{\alpha} \quad , \quad \mathbf{T}' = \widehat{\mathbf{t}} + \mathbf{W}_T\,\boldsymbol{\beta}. \qquad (1)$$

- Multi image 3DMM fitting

多图像 3DMM 生成通过 pool 单个个体不同图片生成的 3DMM 的脸型和纹理参数来实现。$\gamma_i = [\alpha_i, \beta_i]$，置信值ω

$$\widehat{\gamma} = \sum_{i=1}^{N} w_i \cdot \gamma_i \quad \text{and} \quad \sum_{i=1}^{N} w_i = 1, \qquad (2)$$

- 3D 重建模型训练

对于数据集中每一个个体，有多张图片以及单个 pool 的 3DMM。我们将该数据用于训练模型，使模型可以根据同一个体不同的图片来生成类似的 3DMM 特征向量。

采用了 101 层的 deep ResNet 网络来进行人脸识别。神经网络的输出层为 198 维度的 3DMM 特征向量。然后，使用 CASIA 图像生成的 pooled 3DMM 作为目标值对神经网络进行 fine-tuned。

- The asymmetric Euclidean loss

使用 Euclidean loss 会导致输出 3d 人脸缺少细节，引入了 asymmetric Euclidean loss。

$$\mathcal{L}(\gamma_p, \gamma) = \lambda_1 \cdot \underbrace{\|\gamma^+ - \gamma_{\max}\|_2^2}_{\text{over-estimate}} + \lambda_2 \cdot \underbrace{\|\gamma_p^+ - \gamma_{\max}\|_2^2}_{\text{under-estimate}}, \quad (3)$$

using the element-wise operators:

$$\gamma^+ \doteq \text{abs}(\gamma) \doteq \text{sign}(\gamma) \cdot \gamma; \quad \gamma_p^+ \doteq \text{sign}(\gamma) \cdot \gamma_p, \quad (4)$$

$$\gamma_{\max} \doteq \max(\gamma^+, \gamma_p^+). \quad (5)$$

其中，γ为目标 pooled 3DMM 值，γ_p为输入，$\gamma_1\gamma_2$为平衡 over 和 under estimation errors 的值。

- 3D-3D recognition

使用 3DMM 参数γ_p作为描述符。最后，计算两个人脸的相似度（γP1，γP2），计算其余弦得分：

$$s(\gamma_1,\gamma_2)=\frac{\gamma_{p1}\cdot\gamma_{p2}^T}{\|\gamma_{p1}\|\cdot\|\gamma_{p2}\|}. \qquad (6)$$

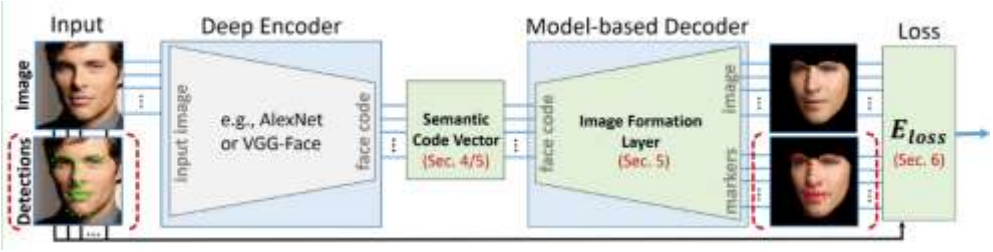
25. MoFA: Model-based Deep Convolutional Face Autoencoderfor Unsupervised Monocular Reconstruction

主要思想：

- 解决了从单一的 wild 图像重建三维人脸的高度挑战性的问题。
- 基于 cnn 和基于模型的人脸重建相结合的新方法，将卷积编码器网络与专家设计的生成模型相结合（解码器）。
- 译码器以一个定义了精确语义的代码向量作为输入，该矢量编码详细的人脸姿态、形状、表情、皮肤反射率和场景光照。
- 首次，cnn 编码器和专家设计的生成模型可以以无监督的方式进行端到端的训练

主要步骤：

基于深度模型的人脸自动编码器实现了对诸如姿态、形状、表情、皮肤反射率和光照等语义参数的无监督端到端学习。一种可选的基于地标的 surrogate loss 可以使更快的收敛和改进的重建结果。这两种情景都不需要在训练期间对语义参数进行监督。



- Semantic Code Vector

语义代码向量$\mathbf{x}\in\mathbb{R}^{257}$统一地将面部表情$\boldsymbol{\delta}\in\mathbb{R}^{64}$、形状$\boldsymbol{\alpha}\in\mathbb{R}^{80}$、皮肤反射率$\boldsymbol{\beta}\in\mathbb{R}^{80}$、摄像机旋转$\mathbf{T}\in SO(3)$和平移 t∈$\mathbb{R}^3$以及场景照明$\boldsymbol{\gamma}\in\mathbb{R}^{27}$统一地参数化。

$$\mathbf{x}=(\underbrace{\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\beta}}_{\text{face}},\underbrace{\mathbf{T},\mathbf{t},\boldsymbol{\gamma}}_{\text{scene}}),$$

$$\mathbf{V}=\hat{\mathbf{V}}(\boldsymbol{\alpha},\boldsymbol{\delta})=\mathbf{A}_s+\mathbf{E}_s\boldsymbol{\alpha}+\mathbf{E}_e\boldsymbol{\delta}$$

$$\mathbf{R}=\hat{\mathbf{R}}(\boldsymbol{\beta})=\mathbf{A}_r+\mathbf{E}_r\boldsymbol{\beta}$$

- Parametric Model-based Decoder

① Perspective Camera

通过刚性变换给出了相机在世界空间中的位置和方位，并基于旋转$\mathbf{T}\in SO(3)$和全局平移 t∈$\mathbb{R}^3$对其进行了参数化。

② Illumination Model

我们用 Spherical Harmonics (SH)表示场景照明：$C(\mathbf{r}_i,\mathbf{n}_i,\boldsymbol{\gamma})=\mathbf{r}_i\cdot\sum_{b=1}^{B^2}\gamma_b\mathbf{H}_b(\mathbf{n}_i)$

前向传播：使用所提出的摄像机和光照模型来绘制真实感场景的图像。为此，在前向通道 f 中，我们计算屏幕空间位置 ui(X)和相关联的像素颜色 ci(X)：

$$\begin{aligned}\mathcal{F}_i(\mathbf{x}) &= [\mathbf{u}_i(\mathbf{x}),\mathbf{c}_i(\mathbf{x})]^T\in\mathbb{R}^5,\\ \mathbf{u}_i(\mathbf{x}) &= \Pi\circ\Phi_{\mathbf{T},\mathbf{t}}(\mathbf{V}_i(\boldsymbol{\alpha},\boldsymbol{\delta})),\\ \mathbf{c}_i(\mathbf{x}) &= C(\mathbf{R}_i(\boldsymbol{\beta}),\mathbf{Tn}_i(\boldsymbol{\alpha},\boldsymbol{\delta}),\boldsymbol{\gamma})\end{aligned}$$

反向传播：以支持训练，我们实现了反向传递图像形成：

$$\mathcal{B}_i(\mathbf{x})=\frac{\mathrm{d}\mathcal{F}_i(\mathbf{x})}{\mathrm{d}(\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\beta},\mathbf{T},\mathbf{t},\boldsymbol{\gamma})}\in\mathbb{R}^{5\times 257},$$

- Loss Layer

$$E_{\text{loss}}(\mathbf{x})=\underbrace{w_{\text{land}}E_{\text{land}}(\mathbf{x})+w_{\text{photo}}E_{\text{photo}}(\mathbf{x})}_{\text{data term}}+\underbrace{w_{\text{reg}}E_{\text{reg}}(\mathbf{x})}_{\text{regularizer}}$$

我们实施 Sparse Landmark Alignment，Dense Photometric Alignment 和 Statistical Regularization。

Dense Photometric Alignment：编码器的目标是预测模型参数，使合成的人脸图像与所提供的输入图像相匹配

$$E_{\text{photo}}(\mathbf{x})=\frac{1}{N}\sum_{i\in\mathcal{V}}\left\|\mathcal{I}(\mathbf{u}_i(\mathbf{x}))-\mathbf{c}_i(\mathbf{x})\right\|_2$$

Sparse Landmark Alignment：我们使用了 46 个地标的子集(在 66 个中)，见图 1。给定子集$\mathcal{L}=\{(\mathbf{s}_j,c_j,k_j)\}_{j=1}^{46}$检测到的地标$\mathbf{s}_j\in\mathbb{R}^2$且具有置信度$c_j\in[0,1]$和相应的模型顶点$k_j\in\{1,\ldots,N\}$，我们强制执行投影的三维顶点接近 2d 检测

$$E_{\text{land}}(\mathbf{x})=\sum_{j=1}^{46}c_j\cdot\left\|\mathbf{u}_{k_j}(\mathbf{x})-\mathbf{s}_j\right\|_2^2$$

Statistical Regularization：这种约束通过选择接近平均值的值(线性人脸模型的基础已经用标准差来衡量)来加强合理的面部形状、表情和皮肤反射率。

$$E_{\text{reg}}(\mathbf{x})=\sum_{k=1}^{80}\alpha_k^2+w_\beta\sum_{k=1}^{80}\beta_k^2+w_\delta\sum_{k=1}^{64}\delta_k^2$$

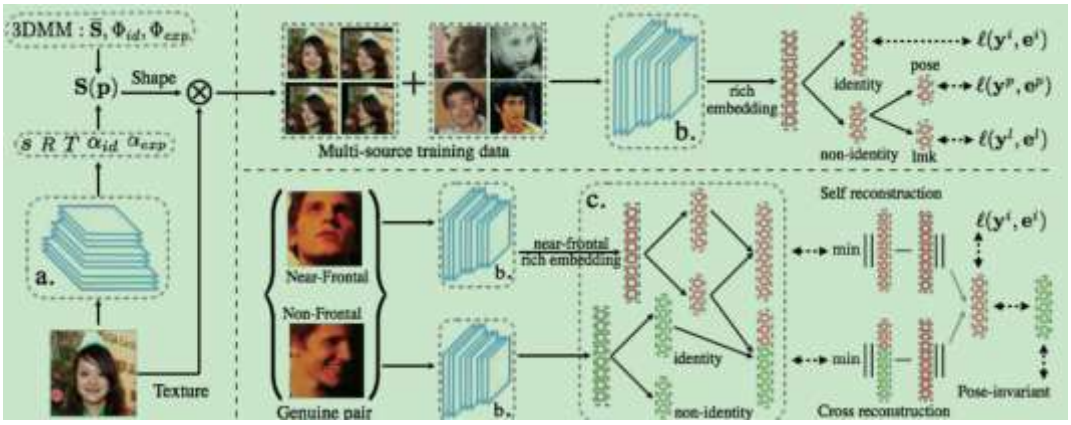
26. Reconstruction-Based Disentanglement for Pose-invariant Face Recognition

主要思想：

- 从一个正面人脸生成非正面视图，以增加训练数据的多样性，同时保留对身份识别至关重要的准确的面部细节。
- 寻找一个丰富的嵌入编码身份特征，包括非身份特征，如姿态和 landmark locations。
- 提出了一种新的特征重建度量学习方法，通过不同的身份特征和姿态特征的组合，要求特征重构之间的对齐性，从而显式地描述身份和姿态。

主要步骤：

（a）从一个近正面的面部重建三维形状，以生成新的人脸图像。（b）丰富的功能嵌入是使用多源监督特征共同学习的身份和非身份。（c）最后，通过重构从非身份信息中解耦身份特征，实现鲁棒姿态不变的表示



- PosevariantFace Generation

从一个近正面的面部重建三维形状，以生成新的人脸图像。

$$\mathbf{S}(\mathbf{p})=sR(\bar{\mathbf{S}}+\mathbf{\Phi_{id}}\alpha_{id}+\mathbf{\Phi_{exp}}\alpha_{exp})+T,\qquad(1)$$

我们采用三维形态模型(3dmm)。[2]学习非线性映射 f(·s)：→将 x 嵌入低维参数空间的 R 235。3dmm 参数 p 控制刚性仿射变换和非刚性变形，从三维平均形状 s 到实例形状 s。$\mathbf{p}=\{s,R,T,\alpha_{id},\alpha_{exp}\}$包括缩放 S、旋转 R、平移 T、身份系数$\alpha_{id}$和表达式系数$\alpha_{exp}$。

- Rich Feature Embedding

仅使用标识监督可能不足以实现不变表示。在此基础上，我们提出利用多源监督来学习一个丰富的特征表示$\mathbf{e}^r$，该特征可以“显式”地划分为一个身份特征$\mathbf{e}^i$和一个非身份特征$\mathbf{e}^n$。$\mathbf{e}^n$可以进一步分支为$\mathbf{e}^p$和$\mathbf{e}^l$，以表示姿态和 landmark 的提示。

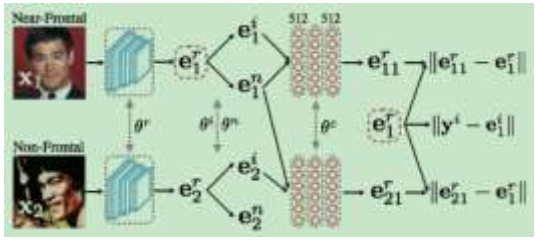
$$\begin{aligned}&\mathbf{e}^i=f(\mathbf{x};\theta^r,\theta^i),\ \mathbf{e}^n=f(\mathbf{x};\theta^r,\theta^n),\\&\mathbf{e}^p=h(\mathbf{e}^n;w^p)=f(\mathbf{x};\theta^r,\theta^n,w^p),\\&\mathbf{e}^l=h(\mathbf{e}^n;w^l)=f(\mathbf{x};\theta^r,\theta^n,w^l),\end{aligned}$$

$$\begin{aligned}\underset{\theta^r,\theta^i,\theta^n,w^l,w^p,w^l}{\operatorname{argmin}}\sum_{image}&-\lambda^i[\mathbf{y}^i\log softmax(w^{iT}\mathbf{e}^i))]\\&+\lambda^p\|\mathbf{y}^p-\mathbf{e}^p\|_2^2+\lambda^l\|\mathbf{y}^l-\mathbf{e}^l\|_2^2,\quad(2)\end{aligned}$$

采用交叉熵和 l2 损失来监督训练。

- Disentanglement by Feature Reconstruction

以上的身份特征和非身份特征是在不同的监督下共同学习的。然而，由于不存在对解耦过程的监督，无法保证身份特征与非身份性特征完全分离。



① 根据其绝对角度将图像分为两组：近正面(≤5)和非正面(>5)。

② 将一对图像{x_k：k=1，2}输入网络，以输出相应的身份和非身份特征。

$$\mathbf{e}_k^i = f(\mathbf{e}_k^r;\theta^i) = f(\mathbf{x}_k;\theta^r,\theta^i),$$
$$\mathbf{e}_k^n = f(\mathbf{e}_k^r;\theta^n) = f(\mathbf{x}_k;\theta^r,\theta^n).$$

③ 在交叉熵约束下，通过最小化自重构损失和交叉重构损失，可以从丰富特征嵌入中重新平衡身份特征和非恒等特征。

$$\mathbf{e}_{11}^r = g(\mathbf{e}_1^i,\mathbf{e}_1^n;\theta^c),\ \mathbf{e}_{21}^r = g(\mathbf{e}_2^i,\mathbf{e}_1^n;\theta^c)$$

其中 er 11 表示近正脸特征的自重构，er 21 表示非正脸特征的交叉重构。

$$\underset{\theta^i,\theta^n,\theta^c}{\operatorname{argmin}} \sum_{pair} -\gamma^i[\mathbf{y}_1^i \log softmax(w^{iT}\mathbf{e}_1^i)] + \gamma^s\|\mathbf{e}_{11}^r - \mathbf{e}_1^r\|_2^2 + \gamma^c\|\mathbf{e}_{21}^r - \mathbf{e}_1^r\|_2^2, \quad (3)$$

只需要微调{θ^i,θ^n}(以及θ^c)来重新平衡身份和非身份特征，同时保持θ^r不变。

ii. 姿态（GAN）

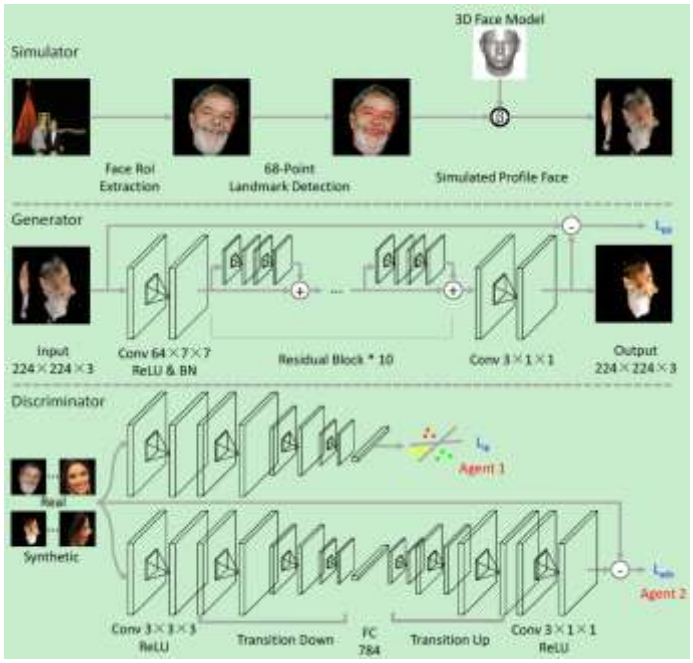
27. Dual-Agent GANs for Photorealistic and IdentityPreserving Profile Face Synthesis

主要思想：

- 提出了一种 Dual-Agent 生成的对抗网络(DA-GAN)模型，该模型可以提高真实人脸仿真器输出的真实感，同时保持真实感细化过程中的身份信息。
- DA-GAN 用全卷积网络作为生成器产生高分辨率图像，利用自动编码器作为鉴别器。
- 除了新的结构外，我们还对标准 GaN 进行了几个关键的修改，以保持姿态和纹理，保持身份和稳定训练过程：(1)姿态感知损失；(2)身份感知损失；(3)具有边界平衡正则项的对抗性损失。

主要步骤：

模拟器提取人脸感兴趣区域（ROI），定位 68 个地标点，然后利用最小二乘拟合估计了三维模型(3D Mm)中检测到的 2d 地标与相应的地标之间的变换矩阵。最后，对不同姿态下的轮廓人脸图像进行了模拟，产生任意姿态的合成人脸，输入到 GAN 进行真实感细化。GAN 使用一个带有 skip-net 的全卷积网络作为生成器和一个自动编码器作为鉴别器。鉴别器既注重识别真假(最大限度地减少L_{adv})，又保护身份信息(最小化L_{ip})。



生成器 $\tilde{x} := G_\theta(x).$

L_{pp} 是一种像素级别的 L1 loss，它是为了增强合成轮廓人脸图像在经过 da-gan 细化前后的姿态(即偏航角)的一致性而引入的：

$$\mathcal{L}_{pp} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} |x_{i,j} - \tilde{x}_{i,j}|.$$

用 L_{adv} 训练，使带边界平衡正则化项的 Wasserstein 距离最小化，以保持生成器和鉴别器损失之间的平衡。

$$\mathcal{L}_{adv} = \sum_{j} |y_j - D_\phi(y_j)| - k_t \sum_{i} |\tilde{x}_i - D_\phi(\tilde{x}_i)|,$$

$$k_{t+1} = k_t + \alpha(\gamma \sum_{j} |y_j - D_\phi(y_j)| - \sum_{i} |\tilde{x}_i - D_\phi(\tilde{x}_i)|),$$

用 L_{ip} 训练，以保持改进后的人脸图像的身份鉴别能力。基于鉴别器的瓶颈层输出定义了具有多类交叉熵损失的 L_{ip}

$$\begin{aligned}\mathcal{L}_{ip} = &\frac{1}{N} \sum_{j} -(Y_j log(D_\phi(y_j)) + (1 - Y_j) log(1 - D_\phi(y_j))) \\ &+ \frac{1}{N} \sum_{i} -(Y_i log(D_\phi(\tilde{x}_i)) + (1 - Y_i) log(1 - D_\phi(\tilde{x}_i))),\end{aligned}$$

交替更新生成器和鉴别器：

$$\begin{cases} \mathcal{L}_{D_\phi} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{ip}, \\ \mathcal{L}_{G_\theta} = (-\mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{ip}) + \lambda_2 \mathcal{L}_{pp}. \end{cases}$$

28. SSPP-DAN: Deep Domain Adaptation Network for Face Recognition with Single Sample Per Person

主要思想：

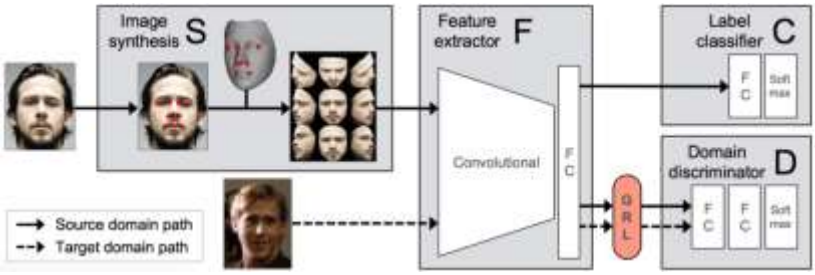
- 每个类一个训练样本的SSPP特征不足以训练深度网络。为了克服这种不足，我们使用3D面部模型来生成各种各样的合成图像来扩充源域图像
- 引入了一个SSPP域适应网络(SSPP-DAN)。利用domain-adversarial训练结合了域适应、特征抽取和分类深度网络

主要方法：

图像合成用于增加源域中的样本数量。通过与梯度反转层(GRL)的对抗训练，利用该特性提取器和两个分类器用于弥补源域之间的差异(例如：稳定的图像)和目标域(例如不稳定图像)。

$$\begin{aligned} L &= \sum_{i \in S} L_C^i + \sum_{i \in S \cup T} L_D^i \qquad \text{when update } \theta_D \\ L &= \sum_{i \in S} L_C^i - \lambda \sum_{i \in S \cup T} L_D^i \qquad \text{when update } \theta_F, \theta_C \end{aligned} \tag{1}$$

L_C^i 和 L_D^i 代表第 i 个样本的标签预测损失和领域预测损失



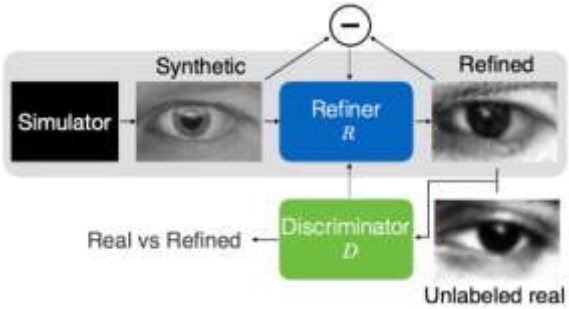
29. Learning from Simulated and Unsupervised Images through AdversarialTraining

主要思想：

- 提出 Simulated+Unsupervised (S+U) learning，使用对抗网络类似于生成对抗网络法（GANs），但使用合成图像而不是随机矢量作为输入
- 对标准的 GAN 算法进行几个关键的修改，以保存标记，避免伪影，并稳定训练:(i)自正则化，(2)局部的对抗性损失，(3)用历史的改善的图像更新识别器

主要方法：

用一个改善器的神经网络—R 来优化模拟器的输出，优化函数是最小化局部对抗损失和自正则化项的组合。对抗性的损失企图混淆一个判别器网络-D，判别图像为真实或改善的。自正则化可以最小化合成图像和改善图像之间的图像差异。改善器网络和鉴别器网络是交替更新的。



- **S+U Learning with SimGAN**

使用一组无标记的真实图像 y 来学习一个提炼合成图像 x 的提炼者$R_\theta(x)$，其中函数参数为θ。

在保留来自模拟器的注释信息的同时，改善的图像$\tilde{x}$应该看起来像一个真实的图像。

$$\mathcal{L}_R(\theta) = \sum_i \ell_{\text{real}}(\theta; \mathbf{x}_i, \mathcal{Y}) + \lambda \ell_{\text{reg}}(\theta; \mathbf{x}_i), \quad (1)$$

第一部分是真实的，l_{real} 增加了人工合成的现实性，而第二部分，l_{reg}，保留了注释信息。

- **Adversarial Loss with Self-Regularization**

使用对抗的鉴别器网络D_ϕ，训练分类真实的和改善,其中ϕ是判别器的参数

$$\mathcal{L}_D(\phi) = -\sum_i \log(D_\phi(\tilde{\mathbf{x}}_i)) - \sum_j \log(1 - D_\phi(\mathbf{y}_j)).$$

$$\ell_{\text{real}}(\theta; \mathbf{x}_i, \mathcal{Y}) = -\log(1 - D_\phi(R_\theta(\mathbf{x}_i))). \quad (3)$$

- **self-regularization loss**

使用自正则化损失，将合成和精炼图像的特征转换的差异最小化。$\ell_{\text{reg}} = \|\psi(\tilde{\mathbf{x}}) - \mathbf{x}\|_1$

$$\mathcal{L}_R(\theta) = -\sum_i \log(1 - D_\phi(R_\theta(\mathbf{x}_i))) + \lambda \|\psi(R_\theta(\mathbf{x}_i)) - \psi(\mathbf{x}_i)\|_1. \quad (4)$$

- **Updating Discriminator using a History of Refined Images**

B 是缓冲区的大小，b 是使用的 mini-batch 的大小。在鉴别器训练的每一次迭代中，我们通过从当前的改善器网络中采样 b/2 图像，并从缓冲区中取样额外的 b/2 图像，以更新鉴别其参数。我们保持缓冲区的大小 B 固定的，在每次训练迭代之后，我们随机地用新生成的改善图像来替换缓冲中的 b/2 样本。

二、 算法

1) Loss 函数

30. DeepFace: Closing the Gap to Human-Level Performance in Face Verification

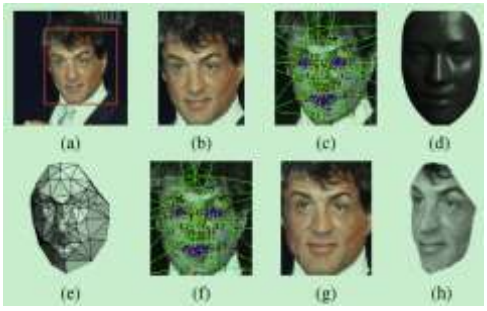
主要思想：

- 依旧是沿着“检测-对齐-人脸表示-分类”这一人脸识别技术路线来的，其贡献在于对人脸对齐和人脸表示环节的改进。
- 在人脸对齐环节，引入了 3D 人脸模型对有姿态的人脸就行分片的仿射对齐。
- 在人脸表示环节，利用一个 9 层的深度卷积在包含 4000 人、400 万张人脸的数据集上学习人脸表示，这个 9 层的 DCNN 网络有超过 1.2 亿个参数。

主要步骤：

- Face Alignment

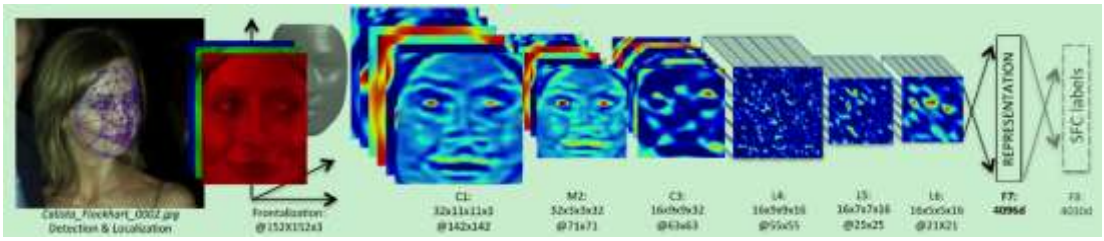
人脸特征点对齐的方法大致可分为：(1)采用脸部的分析 3D 模型;(2)从外部数据集搜索类似的基准点配置以推断；(3)为像素找到相似变换的无监督方法



(a)检测到的面，具有 6 个初始基准点。(b)诱导的 2D 对准作物。(c)在具有相应 Delaunay 三角剖分的 2D 对准作物上的 67 个基准点，我们在轮廓上添加三角形以避免不连续。(d)将参考 3D 形状变换到 2D 对准的作物图像平面。(e)三角可见性 w.r.t.到拟合的 3d-2d 相机；较暗的三角形不那么可见。(f)由 3D 模型诱导的 67 个基准点，该 3D 模型用于指导逐块仿射变换。(g)最终的前端作物。(h)由 3D 模型(未在本文中使用)生成的新视图。

其基本步骤是：1. 检测六个面部关键点 2. 基于六个关键点进行人脸全局仿射变换。3. 检测 67 个面部关键点，并对人脸进行三角剖分。4. 将 3D 人脸转到当前对齐人脸同一视角并获得三角块的可见性。5. 利用 3D 模型产生新的 67 个关键点位置及其三角剖分 6. 分片仿射变换得到的正面人脸。

- Representation



利用了深度卷积神经网络，技术细节需要注意的有两点：1. 激励函数使用的是 ReLu ，优化目标是 cross-entroy Loss，SGD 算法优化。2.为了提高对光照的鲁棒性，最终的特征还进行了二范数归一。

- Verification Metric

最终每张人脸用 F7 层 4096 维的特征向量来表示，采用了两种方法：

① 加权卡方距离，用线性 SVM 学习权重

$$\chi^2(f_1, f_2) = \sum_i w_i(f_1[i] - f_2[i])^2/(f_1[i] + f_2[i])$$

② Siame 网络

一旦进行了学习，一个人脸识别网络（除去最顶层）会被复制两次（一个输入图像使用一个网络）并且特征会被使用于直接预测两张输入图像是否属于同一个人。上述过程通过（a）特征之间采用绝对差异，（b）再将最高的全连接层映射到一个单独的逻辑单元（相同或不同）上来实现。Siamese 网络中的参数被误差的标准交叉熵与反向传播方法所训练。

$$d(f_1, f_2) = \sum_i \alpha_i |f_1[i] - f_2[i]|$$

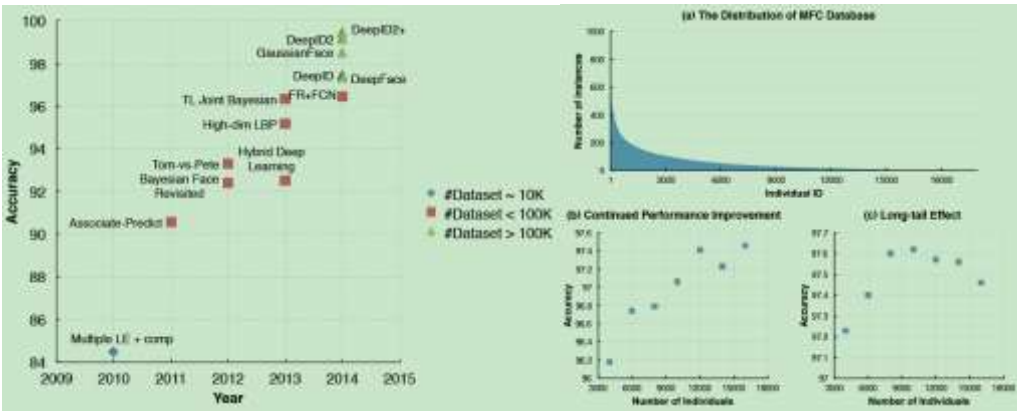
31. Naive-Deep face Recognition: Touching the Limit of LFW Benchmark or Not?

主要思想：

- 分析了大数据如何影响识别性能的观察结果。
- 建立了 Megvii 人脸识别系统，它在 lfw 基准上达到 99.50%的精度。
- 分析在真实的安全认证场景中的性能。机器与人类识别之间仍然存在着明显的差距。

主要步骤：

从 2010 年到 2014 年，数据量扩大了 100 倍，从而获得了巨大的性能提升。



- 长尾效应

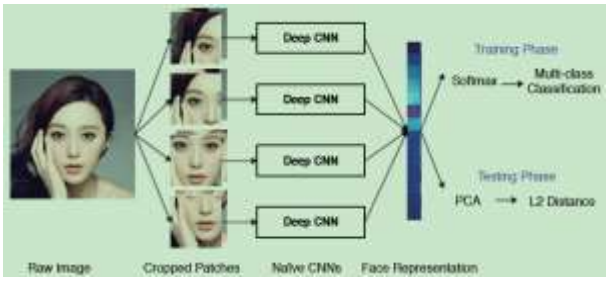
MFC 数据库的分布情况。所有个人都按实例数量排序。(B)在不同数量的训练数据下的性能。随着数据大小的增加，LFW 的精度呈线性上升趋势。每个子训练集从 MFC 数据库中随机选择个体。(C)在不同数量的训练数据下的表现，同时每个子数据库都选择当前数量最多的个体。当个体数大于 10000 时，长尾效应就会出现：保持增加个体数量，每个人只增加几个实例，无助于提高绩效。

- Traditional tricks fade as data increasing

Joint Bayesian, Multi-stage features（deepID2+），Clustering, Joint identification and verification 所有这些复杂的方法都会给系统带来额外的超参数，这使得训练变得更加困难。但是，将这些方法应用到 MFC 数据库中时，根据实验，获得的增益很小。

- 人脸识别

我们设计了一个简单的 10 层深卷积神经网络用于识别。四个人脸区域被裁剪出来表示抽取。我们在传统的多类分类框架下对 MFC 数据库进行网络训练。在测试阶段，采用 PCA 模型进行特征降维，并用简单的 L2 范数来测量测试面对。



- 问题：

在真实场景测试中（Chinese ID （CHID）），该系统的假阳性率（FP=10^{-5} ）非常低。但是，真阳性率仅为 0.66，没有达到真实场景应用要求。而在该测试标准(CHID)下，人类表现的准确率大于 0.90.

因此，该文章提出未来进一步研究的方向。方向一：从视频中提取训练数据。视频中人脸画面接近于现实应用场景（变化的角度，光照，表情等）；方向二：通过人脸合成方法增加训练数据。因为单个个体不同的照片很困难（比如，难以搜集大量的单个个体不同年龄段的照片，可以采用人脸合成的方法（比如 3D 人脸重建）生成单个个体不同年龄段的照片）。

32. Bayesian Face Revisited: A Joint Formulation

（1）隐变量。一个人脸由两部分组成：**μ**用来区分不同人，**ε**是同一个人在不同姿态下的差异。

$$x = \mu + \varepsilon, \qquad (2)$$

这两个潜在变量μ和ε分布服从两个高斯分布：N(0, Sμ)和N(0, Sε)。

（2）用 *x1*，*x2* 分别表示两张图片，**Hi** 表示这两张图片为 intra-personal（同一个人），用 **He** 表示 extra-personal（不同人）。

P(x1, x2|Hi) 和 P(x1, x2|He)分布的协方差矩阵为：

$$\Sigma_I = \begin{bmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{bmatrix} \Sigma_E = \begin{bmatrix} S_\mu + S_\varepsilon & 0 \\ 0 & S_\mu + S_\varepsilon \end{bmatrix}.$$

（3）计算对数似然比 **r**(x1,x2)。，ratio 值很大，则判断为同一个人； ratio 值很小，则判断为不同人。

$$r(x_1, x_2) = \log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)} = x_1^T A x_1 + x_2^T A x_2 - 2x_1^T G x_2, \qquad (4)$$

where

$$A = (S_\mu + S_\varepsilon)^{-1} - (F + G), \qquad (5)$$

$$\begin{pmatrix} F+G & G \\ G & F+G \end{pmatrix} = \begin{pmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{pmatrix}^{-1}. \qquad (6)$$

（4）构建一个类似 EM 算法估计两个协方差矩阵 Sμ和 Sε 。

S-step:

h = [μ；ε1; ……; εm] and x = [x1; ……; xm]之间的关系为：

$$\mathbf{x} = \mathbf{P}\mathbf{h}, \quad \text{where } \mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}. \qquad (7)$$

给定 x，则隐变量的期望值为：

$$E(\mathbf{h}|\mathbf{x}) = \Sigma_h \mathbf{P}^T \Sigma_x^{-1} \mathbf{x}.$$

M-step:

更新参数值**Θ**={Sμ，Sε}:

$$S_\mu = \mathbf{cov}(\mu)$$
$$S_\varepsilon = \mathbf{cov}(\varepsilon)$$

这里的**μ**和**ε**是 **E-step** 阶段的结果。不断重复 E-step 跟 M-step 过程，直到 Sμ，Sε 收敛

33. Hybrid Deep Learning for Face Verification

主要思想：

- 在我们的模型中的深层 ConvNets 模拟初级视觉皮层的联合提取局部关系的视觉特征，从两人脸图像学习的滤波器对相比。这些关系特征通过多个层进行进一步处理，以提取高级和全局特征。
- ConvNets 的多组构造为了实现鲁棒性和相似性，从不同的角度刻画的脸。
- 顶层 RBM 从不同的 ConvNet 组提取的推理，进行最后的预测。

主要方法：

- 我们的混合模型的下部有 12 组,每组五 ConvNets。在不同的组 ConvNets 输入区域对区域范围和不同颜色通道使他们的预测补充。同时每个输入区域对通过交换两个区域并水平翻转每个区域产生八种模式。
- 每一个 ConvNet 以一对对准人脸区域作为输入。它的四个卷积层（接着是 max 池）分层地提取关系特性。最后，提取的特征通过一个全连接层和完全连接到一个单一的神经元层 L0，这是否表明两地区属于同一个人。层 L 0 包含所有 5*12 个 ConvNets 的输出，因此有 8*5*12 个神经元。

$$y_j^r = \max\left(0, b_j^r + \sum_i k_{ij}^r * x_i^r\right), \qquad (1)$$

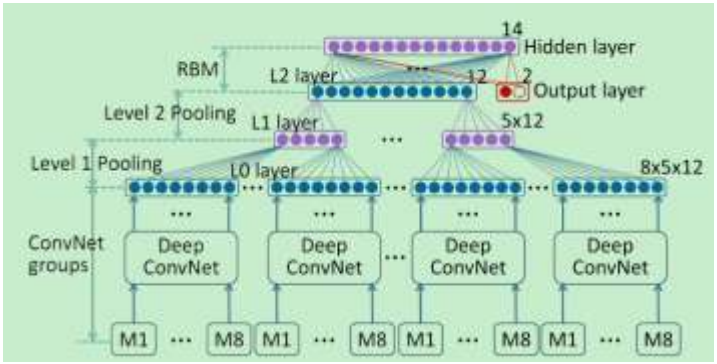
- L1 层(含 5 12 个神经元)是通过 8 种不同输入模式对同一 ConvNet 的 8 种预测进行平均而形成的。L2 层(12 个神经元)是通过将 L1 中与同一组相关的 5 个神经元平均而形成的。

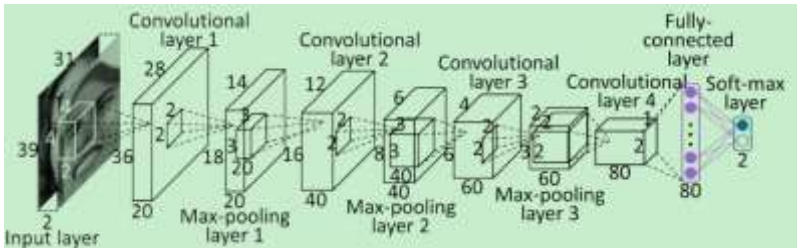
$$x_n = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K} C_m^n(I_k^n),$$

其中设 N 和 M 分别为群数和每组 ConvNet 的个数，$C_m^n(.)$()是 n 个组中 m 个 ConvNet 的输入-输出映射。设$\{I_k^n\}_{k=1}^K$是 n 组的 k 个可能的输入模式。

- 图 2 中我们模型的顶层是一个分类 RBM。它将 L2 中的 12 组输出进行合并，从而给出最终的预测结果。

$$p(y_c \mid x) = \frac{e^{d_c}\prod_j\left(1+e^{c_j+U_{jc}+\sum_k W_{jk}x_k}\right)}{\sum_i e^{d_i}\prod_j\left(1+e^{c_j+U_{ji}+\sum_k W_{jk}x_k}\right)}, \qquad (2)$$





因此，我们首先分别训练每个 ConvNet。然后，通过固定所有的 ConvNet，对 RBM 进行训练。最后，整个网络由 backpropagating 从顶层 RBM 到所有 ConvNets 微调。

34. Deep Learning Face Representation from Predicting 10,000 Classes（Deep ID1）

主要思想：

（1）Deep ID1 通过深度学习学习到一组高级特征表示集合用于人脸验证（对比两个人脸确认是否是同一个人）。

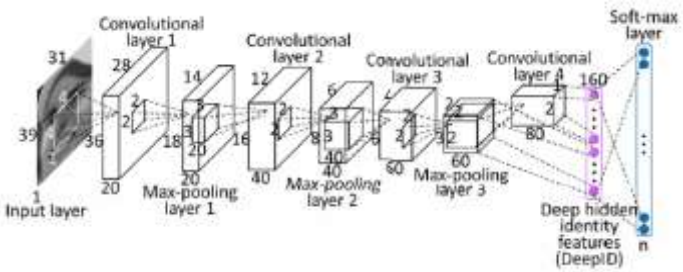
（2）用多分类代替二分类。

（3）关于网络。沿着特征抽取的层次，逐渐在高层形成紧致的和只有少量隐藏神经元的特征。从多个人脸区域中抽取特征，形成相互补充和超完备的表示。

（4）DeepID 和其他的人脸分类器（如 Joint Bayesian)相整合。

主要步骤：

（1）卷积网络架构



一共 4 个卷积层，其中前三个卷积层后都有一个 max-pooling layer。DeepID 层（160 维）同时全连接到最后一个卷积层和第三个卷积层（池化后），这样就可以同时学到高层特征和中层特征。

第三层卷积的神经单元的参数在 2*2 的局部区域内共享；第四层卷积则是全连接，参数在神经单元之间不共享。

最后的 soft-max 层（全连接）是学习类别分布的，若类别数是 10000，则该层维度是 10000。

其中，卷积公式为（采用 ReLU 激活器）：

$$y^{j(r)} = \max\left(0,\ b^{j(r)} + \sum_i k^{ij(r)} * x^{i(r)}\right), \quad (1)$$

池化公式为：

$$y^i_{j,k} = \max_{0 \le m,n < s} \left\{ x^i_{j \cdot s+m,\, k \cdot s+n} \right\}. \quad (2)$$

Deep ID 层公式为：

$$y_j = \max\left(0,\ \sum_i x^1_i \cdot w^1_{i,j} + \sum_i x^2_i \cdot w^2_{i,j} + b_j\right), \quad (3)$$

Softmax 输出：

$$y_i = \frac{\exp(y'_i)}{\sum_{j=1}^{n} \exp(y'_j)}, \quad (4)$$

（2）特征提取

通过 5 个 landmarks 对齐人脸，从人脸图像中选取 10 个区域，3 个尺度，RGB 和 gray 图像块共 60 个块，每个块及其水平翻转抽取 160 维的特征。这样整个 DeepID 的长度是 160*60*2。

（3）人脸验证

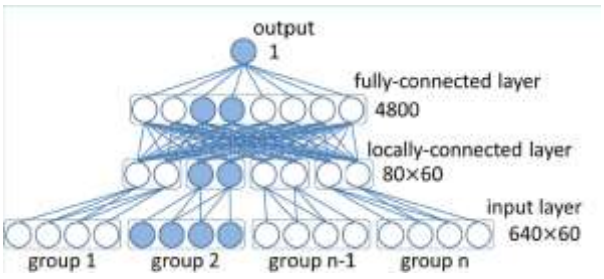
分别采用 Joint Bayesian 和构建深度网络分别进行。实验表明 Joint Bayesian 较好。

- Joint Bayesian

通过计算对数似然比 ratio 判断是否为同一人脸（具体可见论文 Bayesian Face Revisited A Joint Formulation）

$$r(x_1, x_2) = \log \frac{P(x_1, x_2 \mid H_I)}{P(x_1, x_2 \mid H_E)}, \quad (8)$$

- 深度网络



其中隐藏层采用 ReLU 函数，而输出层采用 sigmoid 函数。

35. Deeplearningfacerepresentationbyjointidentificationverification（Deep ID2）

主要思想：

同时使用 face identification 和 verification 信号进行监督学习。face identification 增大类间的变化，face verification 减少类内变化。在 LFW 数据库上得到了 99.15%人脸确认率。

原本的卷积神经网络最后一层 softmax 使用的是 Logistic Regression 作为最终的目标函数，也就是识别信号；但在 DeepID2 中，目标函数上添加了验证信号，两个信号使用加权的方式进行了组合。

主要步骤：

（1）特征提取。使用 SDM 算法抽取人脸上的 21 个标记对齐人脸。通过变化位置、尺度、颜色通道，得到 200 个 face patch，对每个 face patch，使用该 patch 及其水平反转的图像进行（400）特征学习。所以，一共需要 200 个深度卷积神经网络。

（2）学习特征。构建 200 个 DeepID2 来学习上一步得到的 patch。每个 DeepID2 都将输入图像表示成一个 160 维的向量。

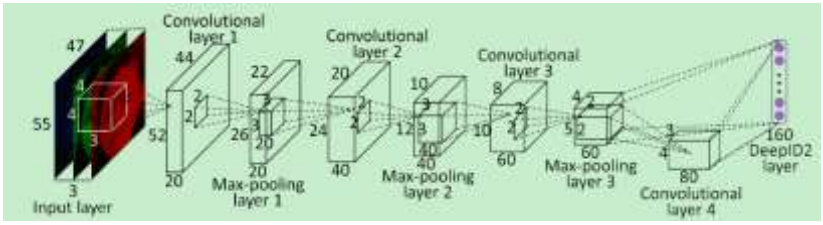
（3）使用前向后向贪心算法来选取一些有效且互补的 DeepID2 向量。

（4）选中 25 个向量后，每张图像的维度是 25*160=4000 维。使用 PCA 进行降维，降维后大约有 180 维。

（5）对于输出后的向量，联合贝叶斯模型来进行分类。

其中：

（1）卷积网络架构



因为添加了类间差距，所以最终层不能再成为是 softmax 层。

（2）两种监督信号

识别信号公式为：

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

验证信号公式为：

- L2 范数

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

当两个样本相同时，则需要最小化它们之间的距离；当两个样本不同时，则需要最小化 m 与它们的距离值之差，m 是一个需要手动调整的参数。

- 余弦值

$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \frac{1}{2}\left(y_{ij} - \sigma(wd + b)\right)^2 , \tag{3}$$

在最终组合目标函数时，将 Ident 与 Verif 加权。

（3）训练过程

input: training set $\chi = \{(x_i, l_i)\}$, initialized parameters θ_c, θ_{id}, and θ_{ve}, hyperparameter λ, learning rate $\eta(t)$, $t \leftarrow 0$

while not converge **do**

$t \leftarrow t+1$ sample two training samples (x_i, l_i) and (x_j, l_j) from χ

$f_i = \text{Conv}(x_i, \theta_c)$ and $f_j = \text{Conv}(x_j, \theta_c)$

$\nabla\theta_{id} = \frac{\partial \text{Ident}(f_i, l_i, \theta_{id})}{\partial \theta_{id}} + \frac{\partial \text{Ident}(f_j, l_j, \theta_{id})}{\partial \theta_{id}}$

$\nabla\theta_{ve} = \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial \theta_{ve}}$, where $y_{ij} = 1$ if $l_i = l_j$, and $y_{ij} = -1$ otherwise.

$\nabla f_i = \frac{\partial \text{Ident}(f_i, l_i, \theta_{id})}{\partial f_i} + \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial f_i}$

$\nabla f_j = \frac{\partial \text{Ident}(f_j, l_j, \theta_{id})}{\partial f_j} + \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial f_j}$

$\nabla\theta_c = \nabla f_i \cdot \frac{\partial \text{Conv}(x_i, \theta_c)}{\partial \theta_c} + \nabla f_j \cdot \frac{\partial \text{Conv}(x_j, \theta_c)}{\partial \theta_c}$

update $\theta_{id} = \theta_{id} - \eta(t) \cdot \nabla\theta_{id}$, $\theta_{ve} = \theta_{ve} - \eta(t) \cdot \nabla\theta_{ve}$, and $\theta_c = \theta_c - \eta(t) \cdot \nabla\theta_c$

end while

output θ_c

36. Deeply learned face representations are sparse, selective, and robust（deepID2+）

主要思想：

- 通过增加隐藏层的维度并增加对早期卷积层的监督，对 deepID2 进行改进。
- 发现其深层的神经激活的高性能关键的三个性质：稀疏性、选择性和鲁棒性。

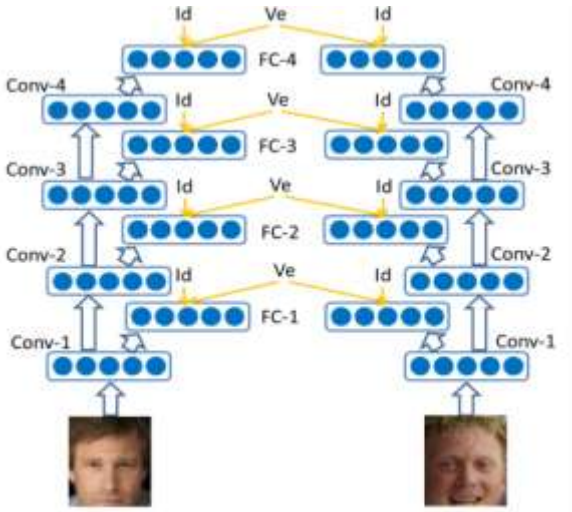
主要步骤：

- **deepID2+**

deepID2+网在 deepID2 上做了 3 个改进：

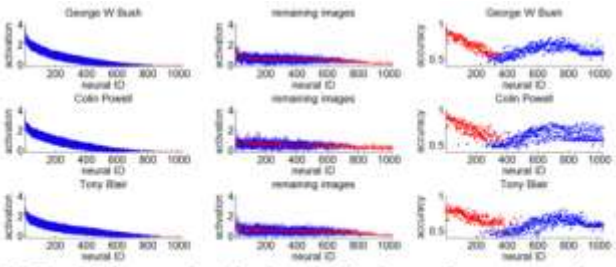
1) 首先，深度 id2+网在四个卷积层的每一层都有 128 个特征图。最终的特性表示也会增加到 512 维。
2) 其次，我们的训练数据是通过合并 CelebFaces+数据集、WDRef 数据和一些从 LFW 中获得的新收集的身份集进行扩展。更大的 DeepID2+网接受了来自 12 000 个身份的大约 29 万张面孔图像，相比之下，从 8000 个用于训练 deepID2 网络的身份信息中，有 16 万张图片。
3) 第三，我们通过将一个 512 维的全连通层连接到四个卷积层的每一个层(在最初的三层卷积层之后，将其连接起来，从而加强了对其的监督。

cov-n 表示第 n 层卷积层(with max-汇聚)。fc-n 表示第 n 个完全连接层。Id 和 Ve 表示识别和验证监视信号。蓝色箭头表示前方传播。黄色箭头表示监视信号。左边和右边的网是相同的深度 id2+网，有不同的输入。

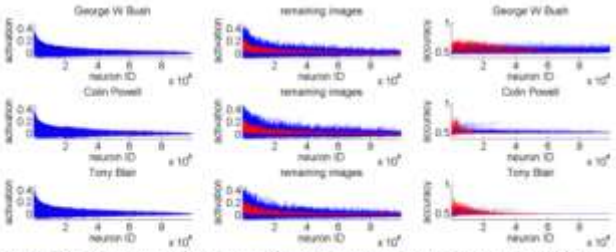


- **Selectiveness on identities and attributes**

对于每个给定的身份，神经元都有强烈的兴奋(例如，神经 ID 小于 200 的神经元)或抑制(例如，神经 ID 大于 600 的神经元)。对于兴奋的神经元，它们的激活分布在更高的值上，而其他图像在这些神经元上的平均值则要低得多。因此，兴奋性神经细胞可以很容易地分辨出一个人的身份，通过在右边的图中所示的红点上显示的红点的高分类精度来验证。



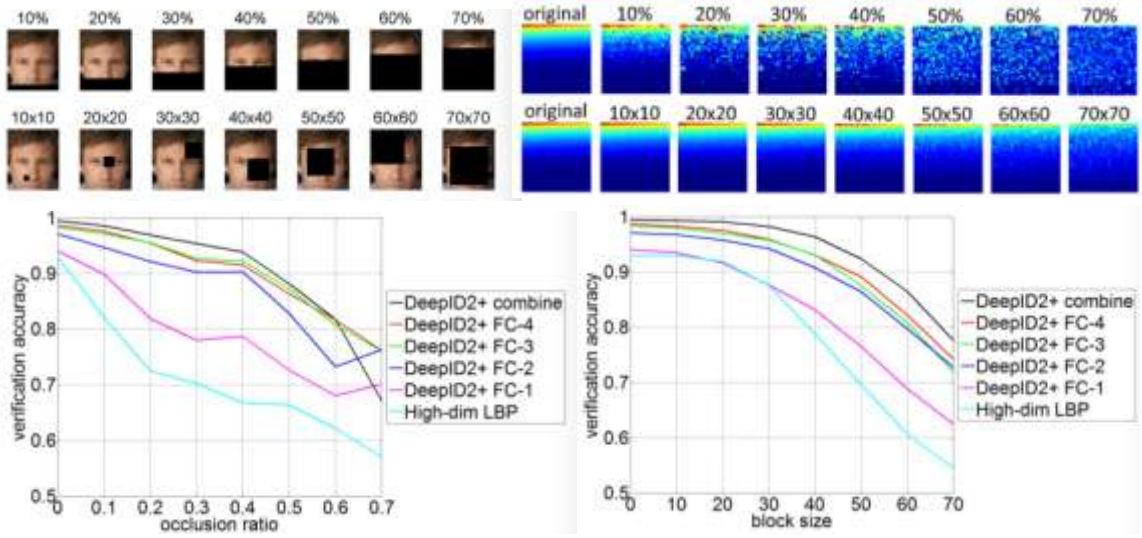
(a) DeepID2+ neural activation distributions and per-neuron classification accuracies.



(b) LBP feature activation distributions and per-feature classification accuracies.

- **Robustness of DeepID2+ features**

在第一个设置中，脸的 10%到 70%被部分遮挡；在第二个设置中，面部被随机的 10× 10 到 70×70 像素块遮挡。



37. DeepID3: Face Recognition with Very Deep Neural Networks

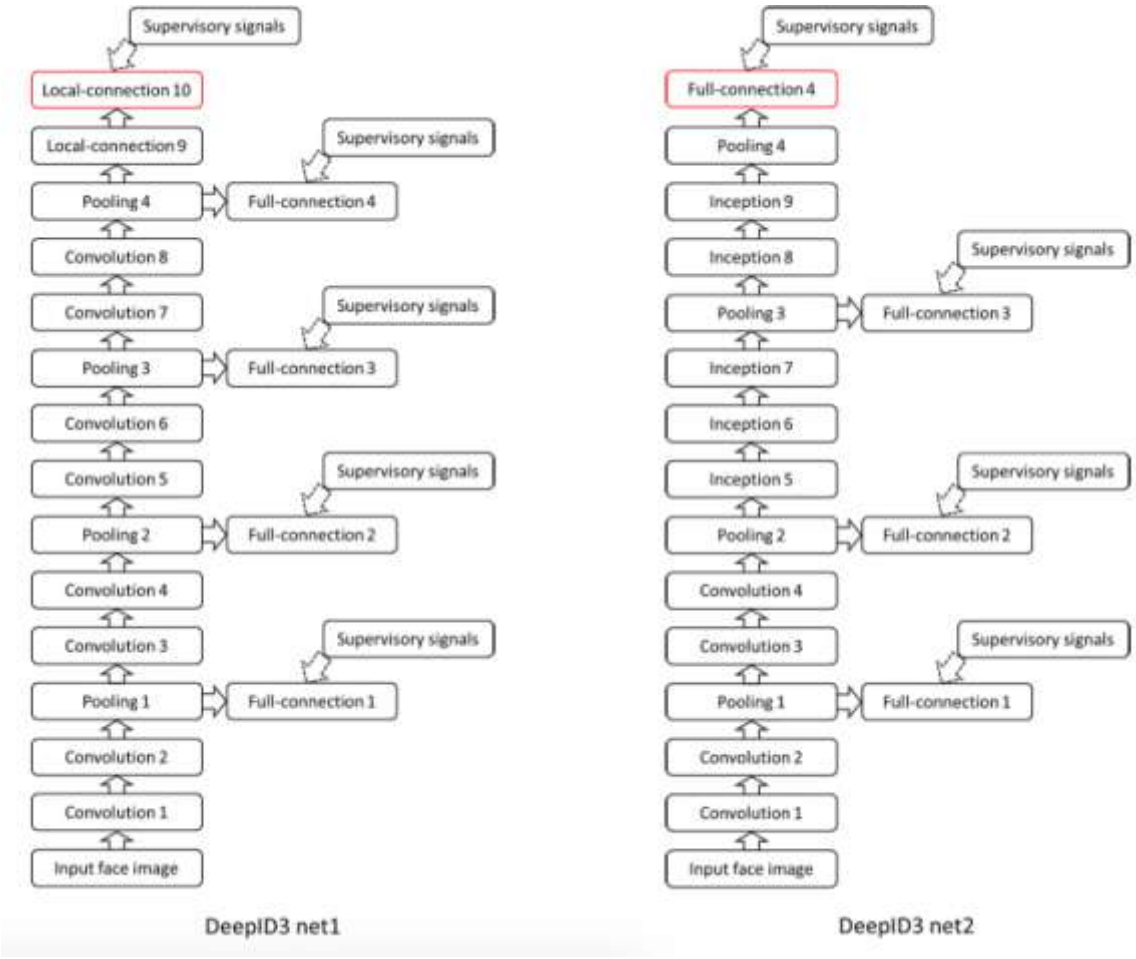
主要思想：

- 提出了两种非常深的神经网络结构，称为 DeepID3，用于人脸识别。这两种架构是 VGG 和 google net，以使它们适合于人脸识别。
- 在训练过程中，在中间和最后的特征提取层中加入了联合面孔识别的监控信号。

主要步骤：

DeepID3 net1 在每个池化层之前都有两个连续的卷积层。与传统 VGG 网络相比，我们在许多从中间层扩展的全连接层中添加了额外的监控信号。顶部的两个卷积层被 local 连接的层所取代。最后一个 local 连接层被用来提取最终的特性，而不需要额外的全连接层。

DeepID3 net2 从每两个连续的卷积层开始，紧接着有一个池化层，与 deepID3 net1 相同。同时在后期的特征提取阶段开始使用 inception 层:在第三个池层之前，有三个连续的 inception 层，在第四个池层之前，有两个连续的 inception 层。



38. A Discriminative Feature Learning Approachfor Deep Face Recognition

主要思想：

- 提出了一种新的用于人脸识别任务的 center loss 信号。center loss 同时学到的每一个类的功能中心，惩罚与类中心距离较远的特征。
- 随着 Softmax 损失和中心损失的共同监督，我们可以训练一个强大的 CNN 获得的两个主要学习目标的深层特征，类间分散和类内紧凑尽可能。
- softmax 损失迫使不同类别的深层特征分开。Center loss 明显地将同一类别的深度特征吸引到他们的中心。

主要方法：

- center loss

  $c_{y_i} \in \mathbb{R}^d$ 表示深部特征的y_i类中心：

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2 \tag{2}$$

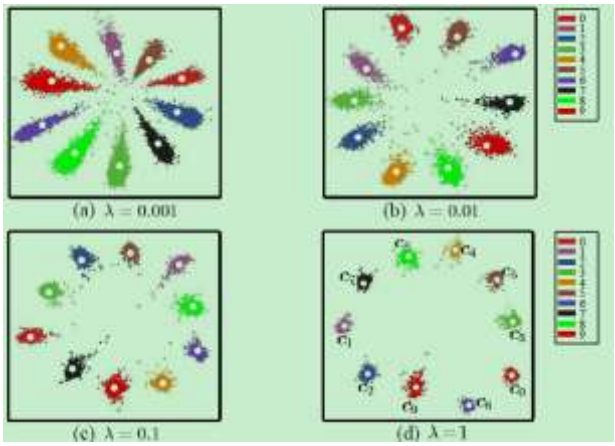
首先，我们没有更新整个训练集的中心，而是基于小批处理执行更新。在每次迭代中，中心都是通过平均相应类的特征来计算的(在这种情况下，某些中心可能不会更新)。第二，为了避免由少量错误标记样本引起的大扰动，我们使用标量来控制中心的学习速度。

$$\frac{\partial \mathcal{L}_C}{\partial x_i} = x_i - c_{y_i} \tag{3}$$

$$\Delta c_j = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(c_j - x_i)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} \tag{4}$$

梯度$\mathcal{L}_c$是基于 xi 的，而c_{y_i}计算如下：

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_S + \lambda\mathcal{L}_C \\ &= -\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2\end{aligned} \tag{5}$$



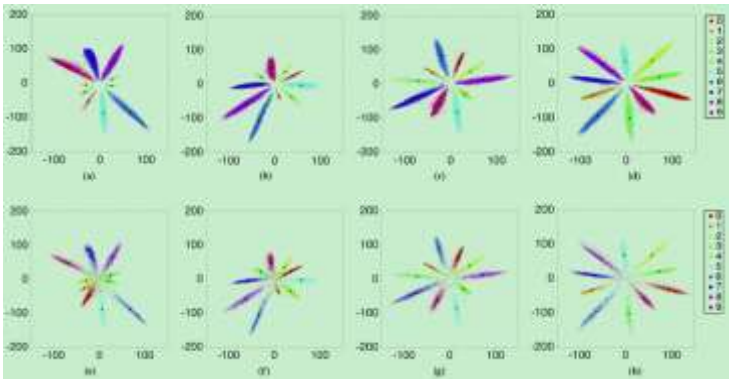
39. Deep Face Recognition with Center Invariant Loss

主要思想：

- 提出了一种新的损失函数，即中心不变损失，通过惩罚大类和小类之间的差异，提高深入学习特征的泛化能力。
- 验证了中心不变损失可以帮助深入学习特征，对训练数据极不平衡的所有类都能平等地分离特征空间，并提高分类性能。
- 通过对 Softmax 损失、中心不变损失和中心损失的联合监督，我们可以训练出一个鲁棒 CNN，它可以平等地对待每个类，而不考虑类样本的数量。

主要步骤：

下图 dh 中，十类样本数量均匀分布（都是 10000），abcefg 中有五类样本数量较少（200,500,1000），在图 2a 中，虽然这些特性看起来是可分离的，但实际上并不是所有类都很好地占用了特征空间。与其他类相比，样本数量少的类占用的面积小。此外，我们可以观察到，五类小样本的中心比其他五个类别更接近特征空间的起源。特征空间分割不好，导致深入学习的特性泛化能力差，如图 2e、2f 和 2g 所示。

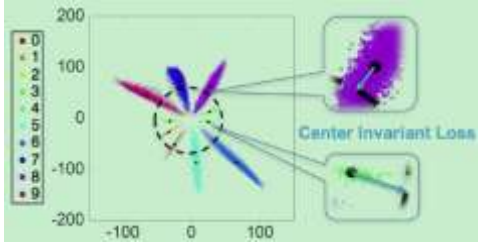


带有标签 y 的 i 样本的中心不变损失可以形式化为：

$$L_I = \frac{1}{4}(||c_y||_2^2 - \frac{1}{m}\sum_{k=1}^{m}||c_k||_2^2)^2.$$

$$c_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i.$$

该公式对每个类的中心间的差异进行了惩罚，中心在特征空间中分布在不同的位置，中心不变损失的目的是使这些中心具有相同的欧几里德范数。圆的半径是所有中心欧氏范数的平均值。我们可以看到，中心不变损失将中心在圆内的点拉出，并将中心在园外的点推入。



目标函数由三个部分组成：Softmax 损失、中心不变损耗和中心损耗。Softmax 损失的目标是最大限度地利用类间的变化.。中心不变损耗对给定不平衡数据的每一类进行正则化处理。这两种损失都处理了类间的分散。同时，中心损失试图将类内变化最小化，这对于学习鉴别特征也是必不可少的。

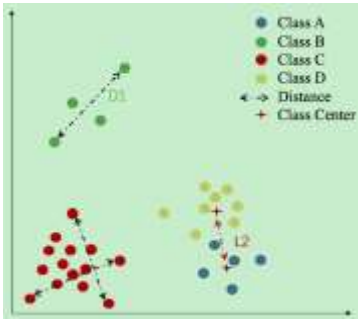
40. Range Loss for Deep Face Recognition with Long-tail

主要思想：

- 长尾数据对人脸识别造成了一定的影响，拥有更多样本的类将对特征学习过程产生更大的影响。并相反地削弱了整个模型的特征提取能力的尾部部分数据。
- 提出一种新的损失函数，称为 range loss，有效地利用训练过程中的长尾数据。更具体地说，面对极不平衡数据，range loss 的目的是在一个 mini batch 内减少整体内在变化同时扩大类间差异。
- 距离损失的最优目标是一类中 k 个最大距离的 harmonic mean values 和最短的类间距离。

主要步骤：

这个小批中有 4 个类，B 类代表一个典型的差类。D1 表示 B 类的最大类内距离。D 类和 A 类之间的 L2 表示这两个类的中心距离。Range loss 的目标可以看作是最短中心距离(这 4 类中的 L2)和每个类中 k 类最大距离的 harmonic mean value（B 类的 D1)。



Range loss 可以表述为：

$$\mathcal{L}_R = \alpha \mathcal{L}_{R_{intra}} + \beta \mathcal{L}_{R_{inter}}$$

$\mathcal{L}_{R_{intra}}$ 指的是类内损失，对每个类的最大 harmonic range 进行惩罚：

$$\mathcal{L}_{R_{intra}} = \sum_{i \subseteq I} \mathcal{L}^i_{R_{intra}} = \sum_{i \subseteq I} \frac{k}{\sum_{j=1}^{k} \frac{1}{\mathcal{D}_j}} \qquad (2)$$

I 表示这个 mini batch 中完整的类/标识集。dj 是第 j 个的最大距离。例如 $\mathcal{D}_1 = \|x_1 - x_2\|_2^2$ 且 $\mathcal{D}_2 = \|x_3 - x_4\|_2^2$，输入 x1 和 x2 表示两个距离最长的人脸样本,同样,输入 x3 和 x4 是第二长距离的样本。等效地,总成本是每个类别中前 k 个最大范围的 harmonic mean value。经验表明，k＝2 带来了良好的性能。

$$\begin{aligned} \mathcal{L}_{R_{inter}} &= \max(m - \mathcal{D}_{Center}, 0) \\ &= \max(m - \|\bar{x}_Q - \bar{x}_R\|_2^2, 0) \end{aligned} \qquad (3)$$

其中，D_{center} 是类中心之间最短的距离，定义为该类中所有输出特性的算术平均值。在一个小批处理中，Q 类中心到 R 类中心之间的距离是所有类中心的最短距离。M 表示超参数作为最大优化间隔，在计算损失时排除大于此间隔的中心。

41. Marginal Loss for Deep Face Recognition

主要思想：

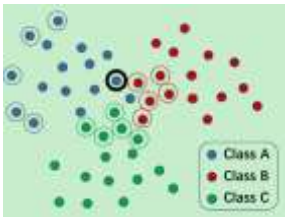
- 通过聚焦于边缘样本，margin loss 同时最小化了类内方差，并使类间距离最大化。
- 将 softmax loss 和 margin loss 联合使用，得到了较好的结果。
- 该方法在跨年龄识别时表现较好，但由于训练数据中年龄差异的限制，当存在较大的年间隔时（大于三十），该方法的性能下降。

主要步骤：

当 xi 和 xj 来自同一类时，它们的距离 $\|x_i - x_j\|_2^2$ 应该接近阈值θ，而当 xi 和 xj 来自不同的类时，它们的距离 $\|x_i - x_j\|_2^2$ 应该比阈值更远。

$$L_m = \frac{1}{m^2 - m} \sum_{i,j,i \neq j}^{m} \left(\xi - y_{ij} \left(\theta - \left\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \right\|_2^2 \right) \right)_+$$

选择最远的类内样本和最近的类间样本来计算损失，既能减小类内方差，又能保持类间距离。



类间损失并不像类内损失那么明显，边际损失退化为中心损失仅仅控制类内方差。由于相似的人很难随机聚集在一起，我们以离线的方式计算每个身份的特征中心，并对训练数据的读取序列进行重新排序。对于每一步，我们根据现成的特征中心随机选择一个标识及其 15 个最近邻标识，从而增加有效的类间边际损失的概率。

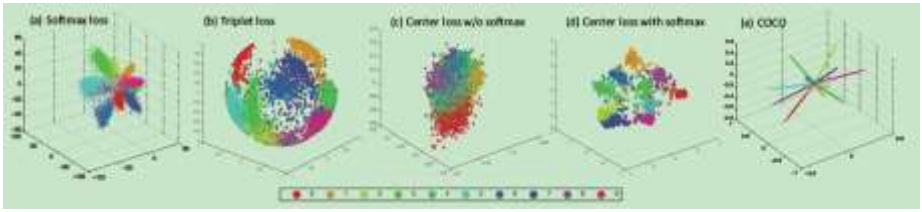
42. Rethinking Feature Discrimination andPolymerization for Large-scale Recognition

总结了 large margin 和度量学习的优缺点。同时证明了归一化特征之后，scale 的下界。

主要思想：

- 提出了同时优化数据间余弦相似度的 congenerous cosine (COCO)算法。它继承了 Softmax 属性，使类间特征具有区分性，并在度量学习中共享类质心的思想。

主要步骤：



聚合内部类特征距离的一个有效解决方案是在网络中提供一系列截断：充当每个类的质心，从而加强了围绕这些中心学习的功能。。为此，我们将 k 类的质心定义为一个 mini-batch 的特征平均值。$c_k = \frac{1}{N_k} \sum_{i \in \mathcal{B}} \delta(l_i, k) f^{(i)} \in \mathbb{R}^D$，其中 Nk 是批次中属于 k 类的样本数。

$$\mathcal{L}^{revise} = \sum_{i \in \mathcal{B}} \frac{\exp \mathcal{C}(f^{(i)}, c_{l_i})}{\sum_{m \neq l_i} \exp \mathcal{C}(f^{(i)}, c_m)},$$

$$\mathcal{C}(f^{(i)}, f^{(j)}) = f^{(i)T} \cdot f^{(j)} / \|f^{(i)}\| \|f^{(j)}\|.$$

分子确保示例 i 离其类中心l_i足够近，分母对其他类中的样本强制足够远。指数算子 exp 是将余弦相似性转移到归一化的概率输出。为了在实践中优化上述损失，我们首先用 l2 范数对特征和质心进行规范化，然后再将特征进行缩放，再将其输入损失层。

$$\hat{c}_k = \frac{c_k}{\|c_k\|}, \quad \hat{f}^{(i)} = \frac{\alpha f^{(i)}}{\|f^{(i)}\|}, \quad p_k^{(i)} = \frac{\exp(\hat{c}_k^T \cdot \hat{f}^{(i)})}{\sum_m \exp(\hat{c}_m^T \cdot \hat{f}^{(i)})},$$

最后优化函数为：

$$\mathcal{L}^{COCO}(f^{(i)}, c_k) = -\sum_{i \in \mathcal{B}, k} t_k^{(i)} \log p_k^{(i)} = -\sum_{i \in \mathcal{B}} \log p_{l_i}^{(i)},$$

- 特征缩放因子的下界

  给定优化损失$\mathcal{L}$有一个上界，即$\mathcal{L} < \epsilon$l；且神经网络有一个 k 的类数，则输入特征上的缩放因子有一个下边界：

$$\alpha > \frac{1}{2} \log \frac{K-1}{\exp \epsilon - 1},$$

43. FaceNet: A Unified Embedding for Face Recognition and Clustering

主要思想：

- 通过 CNN 将人脸映射到欧式空间的特征向量上，计算不同图片人脸特征的距离，通过相同个体的人脸的距离，总是小于不同个体的人脸这一先验知识训练网络。

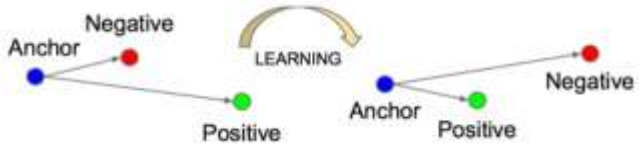
- FaceNet 通过基于 triplet 的损失函数，直接将其输出训练为一个紧凑的 128-d 嵌入。
- 我们提出了一种新颖的在线负样本挖掘策略，它能确保不断增加 triplet loss 的难度。

主要步骤：

Facenet 的结构如下图：

前面就是一个传统的卷积神经网络，然后在求 L2 范数之前进行归一化，就建立了这个嵌入空间，最后的损失函数，就是本文的最大亮点。

Triplet loss 相同身份之间的特征距离要尽可能的小，而不同身份之间的特征距离要尽可能的大（LDA 思想）。



用公式来表示就是：左边类内的距离（加上边际）要小于右边类间的距离，这个约束需要在所有的 Triplet 图像对上都成立：

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 , \quad (1)$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \mathcal{T} . \quad (2)$$

转换一下，它的损失函数就变为上式所示：即 最小化 （类内距离-类间距离+边际）。

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ .$$

一个算法的优劣，还要通过他的时间复杂度来判断，这里一定要确保他的收敛速度。

- Triplet Selection

选择最难区分的图像对：

给定一张人脸图片，我们要挑选其中的一张 hard positive：即另外同身份图像中，跟它最不相似的图片。同时选择一张 hard negative：即在不同身份图像中，跟它最为相似的图片。挑选 hard positive 和 hard negative 有两种方法，offline 和 online 方法，具体的差别只是在训练上。

实际采用:采用在线的方式,在 mini-batch 中挑选所有的 anchor-positive 图像对,同时,依然选择最为困难的 anchor-negative 图像对。

但在实践中，选择最困难的消极因素在训练中会导致糟糕的局部极小值，特别是它会导致崩溃的模型。为了避免这个问题，在选择 negative 的时候，使其满足下式：

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2 . \quad (4)$$

44. Learning Face Representation from Scratch

主要思想：

- 提出了一种从 Internet 收集人脸图像的半自动方法，并建立了一个包含约 10,000 个身份和 500,000 个图像的大型数据集，称为 CASIA-WebFace。
- 使用 17 层 cnn 来学习区分表示，网络集成了最流行的组件，如 ReLU, dropout, 低维表示，identification+ verification cost，小的滤波器和很深的架构，并在 LFW 和 YTF 上获得较好结果。

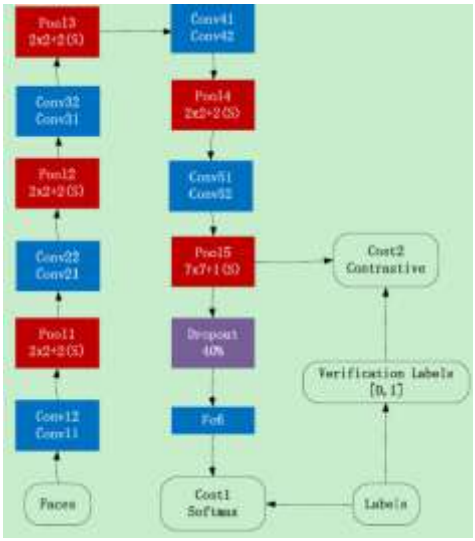
主要步骤：

- 构建数据库

首先，一些名人的名字从网站上被爬取，然后下载他们的网页中的照片。提出了一个简单和快速聚类的方法在照片中标注人脸的身份（通过预训练提取每张人脸的特征模板；使用每个名人的“主照片”作为其种子；使用包含 1 个面的图像来增强每个名人的种子

图像；对于“照片库”中的剩余图像，通过相似性和名称标记查找对应人脸与名字之间的约束）。确保在数据集对象不重叠 LFW，我们使用编辑距离来检查重复的名字。最后，我们扫描整个数据集进行手动和纠正错误的注释。

Dataset	#Subjects	#Images	Availability
LFW [7]	5,749	13,233	Public
WDRef [4]	2,995	99,773	Public (feature only)
CelebFaces [24]	10,177	202,599	Private
SFC [26]	4030	4,400,000	Private
CACD [3]	2,000	163,446	Public (partial annotated)
CASIA-WebFace	10,575	494,414	Public



- 网络架构

① 输入层的尺寸为 100*100*1，灰度图像。输入图像通过两个特征点进行定位，归一化后两个点之间的距离为 25 像素。由于人脸具有近似对称结构，利用镜像操作将训练集增加了一倍。

② 10 个卷积层、5 个池层和 1 个完全连接层，所有滤波器尺寸为 3*3，前四个池层使用 max 运算符，最后一个池层是平均值。池 5 层用作人脸表示。

③ 将 Softmax (identification) 和 Contrastive(verification) cost 相结合，构造目标函数。pool5 作为 Contrastive 的成本函数的输入。FC6 作为 softmax 成本函数的输入。

④ 除了卷积层 52 外，在所有卷积层之后都使用 relu 神经元。由于卷积层 52 是平均组合生成的低维人脸表示，因此它们应该是密集的、紧凑的。REU 容易产生稀疏向量，因此将其应用于人脸表示会降低性能。

45. Deep Face Recognition

主要思想：

- 提供了一种能用有限的人力资源获取规模较大的人脸数据的方法（大概 2.6M 张图像，有超过 2.6k 个人）。

主要步骤：

- 数据收集：

Stage1. 获取数据集的名单：从 IMDB 名人名单里选取 2500 个男性和 2500 个女性，去掉正脸太少的名人，去跟标准基准数据集重合的数据，人工筛选同名的或者图像不足的，再删掉出现在 LFW 和 YTF 中的。这样我们得到 2622 人的名单。

Stage2. 获取图片：在 Google 或者 Bing 搜人名，每个搜索引擎每人各下载 500 张图，再搜人名+演员，每搜索引擎每人各下载 500 张图，得到每人 2000 张图。

Stage3. 用自动过滤器提升纯度：用前 50 张 Google 搜索得到的图片作为正训练样本，以线性 SVM 和 Fisher 向量描述器评估其他人脸与这 50 张人脸的相似度，每个人留下 1000 张图。

Stage4. 移除复制的图片：计算每张图的 VLAD 描述，用非常苛刻的阈值对 1000 张图片聚类，能聚到一类说明可能是同一张图或者是同一张图做了一点颜色调整变换的，聚类结果中每一类只取一张图。

Stage5. 人工过滤：用训练过的 AlexNet 中的 softmax 层判断一组人脸中是否有错误的人脸，每组 200 张图送入网络，如果纯度达到 95%以上则认为这 200 张图片是比较纯净的。最后留下 982803 张图，其中有 95%的正脸和 5%的侧脸。

Stage	Aim	Type	# of persons	# of images per person	total # images	Annotation effort	100%-EER
1	Candidate list generation	A	5,000	200	1,000,000	–	–
2	Image set expansion	M	2,622	2,000	5,244,000	4 days	–
3	Rank image sets	A	2,622	1,000	2,622,000	–	96.90
4	Near dup. removal	A	2,622	623	1,635,159	–	–
5	Final manual filtering	M	2,622	375	982,803	10 days	92.83

- 网络结构和训练

本文用了 A，B，D 三种网络结构，A 的网络结构如上图，B 的网络结构比 A 多 2 个卷积层，D 的网络结构比 A 多 5 个卷积层。网络输入 224*224 的图片，A 是从头开始训练的，B 和 D 是在 A 的基础上 finetune 的。

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmax
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

① Softmax 对数损失：

将每张图片送入网络提取特征，再用输出为 N（N=2622）的全连接层（在 A 中就是 fc8）把每个特征对应的分数算出来组成向量，和各个类的独热向量比较，计算 softmax 对数损失。然后提取特征，通过欧氏距离进行对比相似度。

② triplet 损失：

提完特征后，将它 l2 标准化并投影到维数更低的空间$\mathbf{x}_i = W'\phi(\ell_i)/\|\phi(\ell_i)\|_2, W' \in \mathbb{R}^{L\times D}$，再计算 triplet 损失。其中投影向量$W'$使 triplet 损失最小化：

$$E(W') = \sum_{(a,p,n)\in T} \max\{0, \alpha - \|\mathbf{x}_a - \mathbf{x}_n\|_2^2 + \|\mathbf{x}_a - \mathbf{x}_p\|_2^2\}, \quad \mathbf{x}_i = W'\frac{\phi(\ell_i)}{\|\phi(\ell_i)\|_2}. \qquad (1)$$

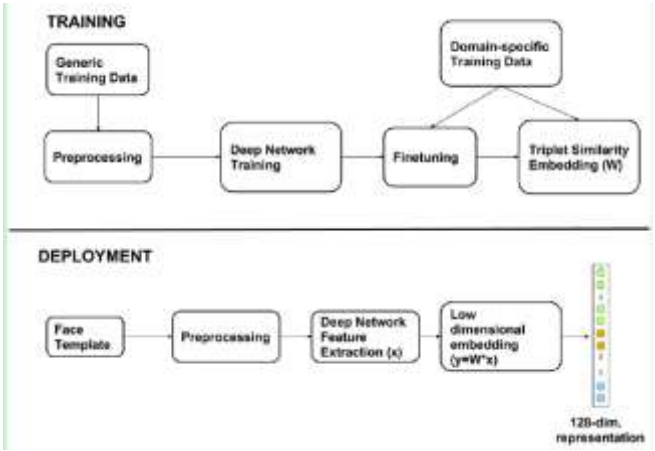
测试时，用是否大于阈值来判断是不是同一个人，这个阈值是用与训练集不同的验证集计算出来的。

46. Triplet Similarity Embedding for Face Verification

主要思想：

- 所提出的算法结合了基于 CNN 的深层方法，并利用三元组相似性约束以大幅度的方式学习低维区分嵌入。
- 除了提高性能，这种嵌入在内存和后处理操作方面提供了显著优势。

主要步骤：



在训练过程中，我们使用通用数据集（独立于 IJB-A 数据集）来训练我们的深层架构。为了进一步提高性能，我们使用与 IJB-A 数据一起提供的训练数据来微调 DCNN 模型。从 FinetunedDCNN 模型中提取的特征，使用相同的训练集来学习所提出的三元组相似性嵌入。在测试阶段，给出了人脸模板，经过预处理后，利用细化模型提取了人脸的深层特征。利用训练过程中学习到的嵌入矩阵将深层特征投影到一个低维空间中。

该工作主要由两个部分组成：深度网络作为特征提取器和学习过程，将输入特征投影到一个有区别的低维空间。

- NETWORK ARCHITECTURE

体系结构与 alexnet 的体系结构密切相关。具有以下差异：f/c 层参数较少，使参数数减少了一半以上。使用 Parametric Rectifier Linear units (p-relu)代替 relu，因为它们允许基于学习阈值的输出值为负值，并且已经证明可以提高收敛速度。

使用在 ImageNet 挑战数据集上训练的 alexnet 模型的权重初始化卷积层权重。为了学习更多特定领域的信息，我们添加了一个额外的卷积层(卷积 6)，并从头开始初始化完全连接的层 fc6-fc8。由于网络被用作特征提取器，所以在部署过程中删除了最后一层 fc8，从而将参数数减少到 15M。特征是从 fc 7 层中提取出来的，从而产生了 512 的维数。

Deep Network Architecture		
Layer	Kernel Size/Stride	#params
conv1	11x11/4	35K
conv2	5x5/2	614K
conv3	3x3/2	885K
conv4	3x3/2	1.3M
conv5	3x3/1	885K
conv6	3x3/1	590K
fc6	1024	9.4M
fc7	512	524K
fc8	10575	10.8M
Softmax Loss		Total: 25M

- LEARNING A DISCRIMINATIVE EMBEDDING

考虑一个三元组{a，p，n}，其中一个 a 和 p（正）来自同一个类，但是 n（负）属于另一个类。我们的目标是从数据中学习线性映射 w，下面的约束：

$$(\mathbf{W}a)^T \cdot (\mathbf{W}p) > (\mathbf{W}a)^T \cdot (\mathbf{W}n)$$

通过选择 W 的维数为 D×512，d<512，我们实现了除了性能更好的降维。对于我们的工作，我们基于交叉验证确定 D=128。给定一组标记数据点，我们解决以下优化问题：

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{a,p,n \in \mathbb{T}} max(0, \alpha + a^T \mathbf{W}^T \mathbf{W} n - a^T \mathbf{W}^T \mathbf{W} p)$$

W 更新如下：

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta * \mathbf{W}_t * a * (n-p)^T \qquad (3)$$

47. Triplet Probabilistic Embedding for Face Verification and Clustering（这篇相比于上篇发表在前）

主要思想：

- 与上一篇基本思想相同，都是结合了基于 CNN 的深层方法，并利用三元组相似性约束以大幅度的方式学习低维区分嵌入。但是架构稍有不同，triplet embedding 的优化函数也略有不同

主要步骤：

- Network Architecture

由 7 个不同内核大小的卷积层组成。初始层具有较大的尺寸，快速地对图像进行采样和参数的减小，而后一层则由小的滤波器尺寸组成。使用 Parametric Rectifier Linear units（PReLUs）。前三层的卷积层（conv1-conv3）直接采用在 ImageNet 数据集上训练的 alexnet 模型初始化权重。因此，为了了解更多的领域特定信息，我们添加了 4 个卷积层，每个层由 512 个 3*3 核组成。Conv4-conv7 层卷积不对输入进行下采样，从而学习更复杂的高维表示。用随机高斯分布从头开始初始化卷积 4-卷积 7 和完全连通层 fc6-fc8。网络使用 Softmax 损失函数进行训练。

Layer	Kernel Size/Stride	#params
conv1	11x11/4	35K
pool1	3x3/2	
conv2	5x5/2	614K
pool2	3x3/2	
conv3	3x3/2	885K
conv4	3x3/2	1.3M
conv5	3x3/1	2.3M
conv6	3x3/1	2.3M
conv7	3x3/1	2.3M
pool7	6x6/2	
fc6	1024	18.8M
fc7	512	524K
fc8	10548	10.8M
Softmax Loss		Total: 39.8M

- Learning a Discriminative Embedding

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{(v_i, v_j, v_k) \in \mathbb{T}} \max\{0, \alpha + (v_i - v_j)^T \mathbf{W}^T \mathbf{W}(v_i - v_j) - (v_i - v_k)^T \mathbf{W}^T \mathbf{W}(v_i - v_k)\} \qquad (5)$$

W 更新如下：

$$\mathbf{W}_{r+1} = \mathbf{W}_r - \eta * \mathbf{W}_r * ((v_i - v_j)(v_i - v_j)^T - (v_i - v_k)(v_i - v_k)^T) \qquad (6)$$

48. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding

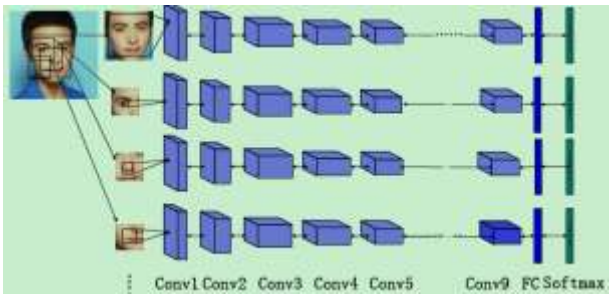
主要思想：

- 提出了一种两步学习方法，结合 mutil-patch deep CNN 和 deep metric learning，实现脸部特征提取和识别。
- 通过 1.2million（18000 个个体）的训练集训练，该方法在 LFW 数据集上取得了 0.9977 的成绩。

主要步骤：

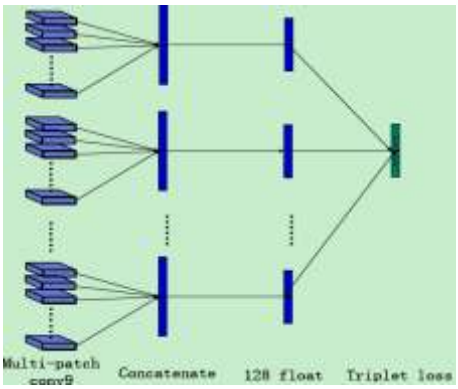
- multi-patch deep CNN

人脸不同区域通过深度卷积神经网络分别进行特征提取。简单地使用具有 9 个卷积层的网络结构和在端部用于受监督的多类学习的 softmax 层。网络的输入是 2D 对准的 RGB 面部图像。pooling 和归一化层位于一些卷积层之间，相同的结构用于提取以不同面部部区域为中心的重叠图像面片。对每个网络上的卷积层的输出作为人脸表征和连接在一起形成一个高维的特征。



- deep metric learning

深度卷积神经网络提取的特征再经过 metric learning 将维度降低到 128 维度。Triplet loss 旨在缩短相同身份的样本的 L2 距离,并在来自不同的样本之间扩大其之间的距离。

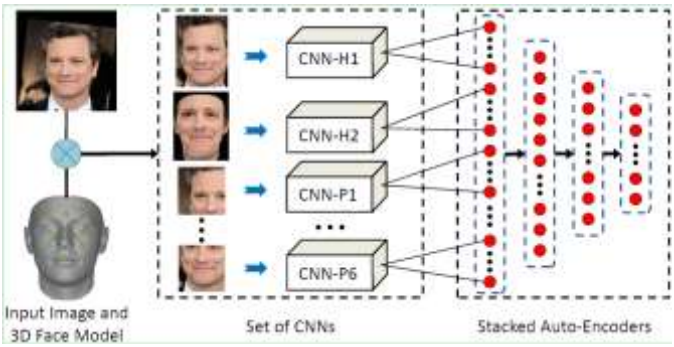


49. Robust Face Recognition via Multimodal DeepFace Representation

主要思想：

- 采用深互补信息的多模态图像数据提升人脸识别性能。
- 第一，通过仔细集成一些已发布的或自己开发的技巧，如深层结构、小过滤器、谨慎使用 REU 非线性、积极的数据增强、droupout 和使用不同损失函数进行多阶段训练，L2 标准化，来提高每一个 CNN 的识别能力。
- 其次，采用不同 CNN 网络从原始的整体人脸图像中、通过三维模型对呈现的正面姿态图像和均匀采样的图像块提取多模态信息。
- 第三，采用堆叠式自动编码器进一步提取融合的 CNN 网络特征，有利于学习的非线性降维。

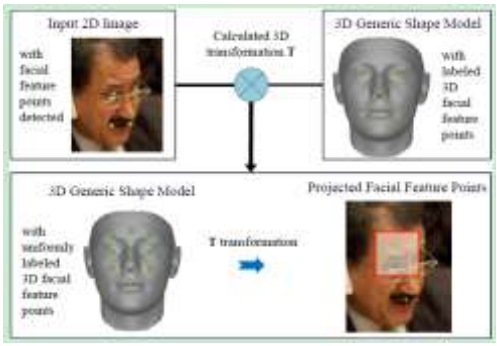
主要步骤：



- Single CNN Architecture

① 采用仿射变换，根据五个稀疏人脸特征点的坐标将所有人脸图像首先归一化为 230*230 像素。一张 165*120 像素的整体人脸图像，六张 100*100 像素的 patch 被采样，另一张整体的人脸图像由三维姿态归一化获得。

注：相比于随机采样 6 个 patch，给出一幅二维人脸图像，首先利用正交投影对通用的三维人脸模型，借助五个人脸特征点对齐。然后，在通用的 3D 面部模型上手动标记 9 个 3D 标记，将预先标记的三维标记投影到 2d 图像中。最后，一个大小为 100*100 像素的贴片被裁剪在每一个投影的 2d 标记周围。



② CNN 从整体人脸图像中提取特征为 CNN-H1，从 3D 图像中提取特征为 CNN-H2，从六个 patch 中提取特征为 CNN-P1，……，P6。除了最后的卷积层外，所有层都利用 ReLU 激活函数，还消除了 fc6 之后的 ReLU。

③ 首先，训练 CNN 为多类分类问题，即使用 Softmax 损失。第二，我们采用最近提出的 triplet loss 进行微调。然后一个一个地训练三个自动编码器。在测试阶段，我们从原始图像和水平翻转图像中提取深度特征。

- Combination of CNNs using Stacked Auto-Encoder

与传统的降维方法(如 PCA)相比，SAE 在学习非线性特征变换方面具有优势.。在本文中，我们采用了三层 SAE。三种自动编码器的神经元数分别为 2048、1024 和 512.。最后一个编码器的输出作为人脸图像的紧凑表示。使用了两种激活函数，即 ReLU 和 tanh 函数，且 tanh 函数的效果更好。

- 人脸验证

在无监督模式下，余弦距离用来测量 Y1 和 Y2 之间的相似度 S。

$$s(y_1, y_2) = \frac{y_1^T y_2}{\|y_1\| \|y_2\|}. \quad (5)$$

对于监督模式，采用联合贝叶斯（JB）模型

$$r(x_1, x_2) = log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)}, \quad (7)$$

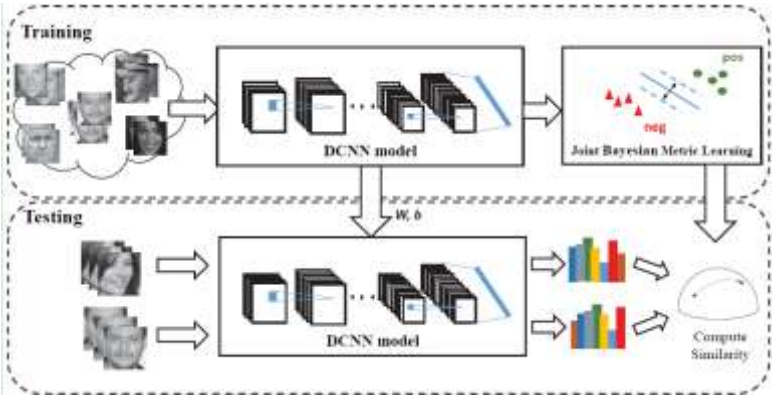
50. Unconstrained Face Verification using Deep CNN Features

主要思想：

- 基于深度卷积特征无约束人脸验证算法和评估它在新发布的 IARPA Janus 基准数据集（IJB-A）以及传统的标记在 LFW 数据集。

主要步骤：

我们的方法包括训练阶段和测试阶段。为了训练，我们首先在 CASIA-WebFace 数据集上执行面部和标记检测，在 IJB-A 上对每个面部进行局部化和对齐。接下来，我们在 CASIA-WebFace 上训练我们的 DCNN，并用 JIB-A 数据集和 DCNN 特征导出联合贝叶斯度量。然后，在给定一对测试图像集的情况下，基于它们的 DCNN 特征和学习的度量来计算相似度分数。



- Preprocessing

每个脸部对齐到标准坐标，通过使用 7 个标志点(即两个左眼角、两个右眼角、鼻尖和两个口角)的相似性变换，在对准之后，面部图像分辨率为 100×100 像素，并且两个眼睛的中心之间的距离为大约 36 像素。

- Deep Face Feature Representation

① 输入图像是 100×100×1 的灰度图像。

② 网络包括 10 个卷积层、5 个 pooling 层和 1 个完全连接的层。

③ 每个卷积层后面跟着一个校正的线性单元(Relu)，除了最后一个卷积层。我们使用参数 relu(Prelu)代替使用 relu 对零的所有负响应进行压缩，允许负面响应，从而提高网络性能。

④ 在 CONV12 和 CONV22 之后添加两个局部归一化层，以减轻照明变化的影响。

⑤ 前四个池层使用 max pooling，最后一层使用 average pooling。

⑥ 在度量学习阶段之前，提取的特征进一步被 L2 归一化为单位长度。

- Joint Bayesian Metric Learning

$$r(\mathbf{x}_i,\mathbf{x}_j)=\log\frac{P(\mathbf{x}_i,\mathbf{x}_j|H_I)}{P(\mathbf{x}_i,\mathbf{x}_j|H_E)}=\mathbf{x}_i^T\mathbf{M}\mathbf{x}_i+\mathbf{x}_j^T\mathbf{M}\mathbf{x}_j-2\mathbf{x}_i^T\mathbf{R}\mathbf{x}_j,$$

我们不使用 EM 算法来估计 Sμ 和 Sθ，而是如下所述优化大边缘框架中的距离：

$$\underset{\mathbf{M},\mathbf{B},b}{\operatorname{argmin}}\sum_{i,j}max[1-y_{ij}(b-(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)+2\mathbf{x}_i^T\mathbf{B}\mathbf{x}_j),0],$$

其中，B=R-M。其中 b 是阈值。如果 Person i 和 j 是相同的，yij=1，否则，yij=-1。

$$\mathbf{M}_{t+1}=\begin{cases}\mathbf{M}_t, & \text{if } y_{ij}(b_t-d_{\mathbf{M},\mathbf{B}}(\mathbf{x}_i,\mathbf{x}_j))>1\\ \mathbf{M}_t-\gamma y_{ij}\boldsymbol{\Gamma}_{ij}, & \text{otherwise,}\end{cases}$$

$$\mathbf{B}_{t+1}=\begin{cases}\mathbf{B}_t, & \text{if } y_{ij}(b_t-d_{\mathbf{M},\mathbf{B}}(\mathbf{x}_i,\mathbf{x}_j))>1\\ \mathbf{B}_t+2\gamma y_{ij}\mathbf{x}_i\mathbf{x}_j^T, & \text{otherwise,}\end{cases}$$

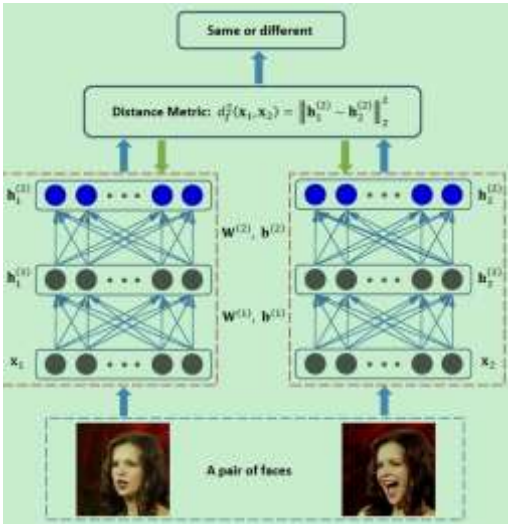
$$b_{t+1}=\begin{cases}b_t, & \text{if } y_{ij}(b_t-d_{\mathbf{M},\mathbf{B}}(\mathbf{x}_i,\mathbf{x}_j))>1\\ b_t+\gamma_b y_{ij}, & \text{otherwise,}\end{cases}$$

51. Discriminative Deep Metric Learning for Face Verification in the Wild

主要思想：

- 提出一种 discriminative deep metriclearning (DDML)方法,用于在野外进行人脸验证,不同于现有的基于度量学习的人脸验证方法,其目的在于学习马氏距离度量,以使类间变化最大化,同时最小化类内变化
- 文章用了两个全连接层进行 DDML，输入的是传统特征，设计一个损失函数。

主要步骤：



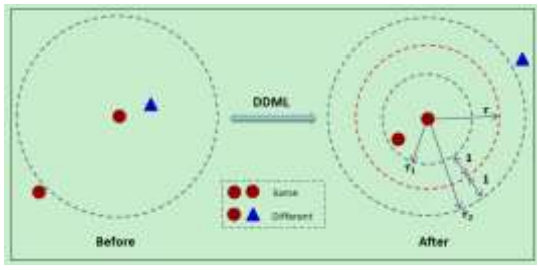
- Mahalanobis Distance Metric Learning

$$\begin{aligned} d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) &= \sqrt{(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)} \\ &= \sqrt{(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{W}^T\mathbf{W}(\mathbf{x}_i-\mathbf{x}_j)} \\ &= \|\mathbf{W}\mathbf{x}_i-\mathbf{W}\mathbf{x}_j\|_2 \end{aligned} \quad (3)$$

学习一个马氏距离度量是相当于寻找一个线性变换 W，将个样本映射到低维子空间，并在变换空间中两样本欧氏距离等于在原始空间马氏距离度量。

- DDML

在原始特征空间中有三个人脸样本，用于生成两对人脸图像，其中两对构成正对，两个分别形成负对。在原始人脸特征空间中，正对之间的距离大于负对之间的距离。当应用 DDML 方法时，正对的距离小于一个较小的阈值τ_1，负对的距离大于一个较大的阈值τ_2。



使用一个阈值$\tau\ (\tau > 1)$连接τ_1和τ_2，并强制$d_f^2(\mathbf{x}_i,\mathbf{x}_j)$和$\tau$之间的间隔大于 1：

$$\ell_{ij}\left(\tau - d_f^2(\mathbf{x}_i,\mathbf{x}_j)\right) > 1.$$

where $\tau_1 = \tau - 1$ and $\tau_2 = \tau + 1$，$l_{ij} = 1$表示正对，$l_{ij} = -1$表示负对。

优化函数为：

$$\begin{aligned} \arg\min_f J &= J_1 + J_2 \\ &= \frac{1}{2}\sum_{i,j} g\left(1 - \ell_{ij}\left(\tau - d_f^2(\mathbf{x}_i,\mathbf{x}_j)\right)\right) \\ &+ \frac{\lambda}{2}\sum_{m=1}^{M}\left(\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2\right) \end{aligned} \quad (7)$$

52. Face Search at Scale: 80 Million Gallery

主要思想：

- 除了提高性能，这种嵌入在内存和后处理操作方面提供了显著优势。在级联的框架，联合 COTS 匹配器，提出了一种结合快速人脸搜索系统。
- 给定一个 probe 的脸，我们的第一个 filter 使用从卷积神经网络生成特征从 gallery 里找到 top-k 最相似的面孔。K 个候选照片再结合深度特征相似性和 COTS 匹配重新排名。

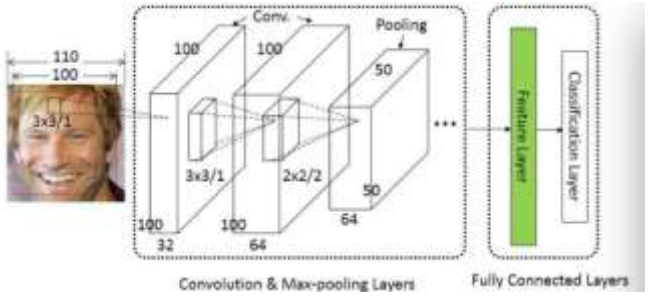
主要步骤：

人脸搜索架构包括三个主要步骤：i）从 N 个 gallery 图像和 probe 图像中提取特征；II）对比 probe 和 gallery 图像的特征，利

用 product quantization 选出 k 个最像的候选图像（iii）利用深度网络和 COTS face matcher 的融合分数对 k 个候选图像重新排序。

- Template Generation

输入到网络是彩色图像而不是灰度图像；ii）鲁棒的人脸对齐程序；iii）增加数据增强步骤，从 110×110 彩色输入图像随机 crop100×100 区域；iv）删除 contrastive 层。



人脸对齐步骤：1）探测 68 个特征点 2）旋转人脸在图像平面上使它基于眼睛的位置直立；3）找到脸部左右的一个中心点，以及眼睛的中心点和嘴；4）将左右中心点置于图像 x 轴中间；5）固定 X 轴位置，沿 y 轴放置眼睛和嘴中心点，使得从图像的顶部到眼睛中心点占 45%和嘴到底部站图像 25%；6）调整图像的分辨率为 110×110。

利用几个图像变换操作来扩充训练集:通过随机地对 110×110 对齐人脸图像进行水平重选和随机裁剪,得到输入图像的转换版本。

在输入层之后，有 10 个卷积层，4 个最大池层和 1 个平均池层。第四组卷积层依次是窗口大小为 2×2 和步长为 2 的最大汇聚层，而最后一组卷积层则是窗口大小为 7×7 的平均池层。

- Fusion Method

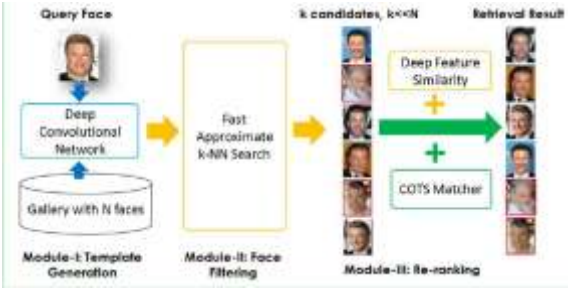
所提出的级联搜索系统的另一个重要问题是来自深度特征(DF)和 COTS 的相似性得分的融合。我们根据经验对以下策略进行了评估：

① DFCOTS：基于深度特征和 COTS 匹配器的相似性分数级融合，无需任何筛选。

② DF→COTS：使用深度特征对图库进行过滤，然后基于深度特征和 COTS 评分之间的分数级别融合对候选列表进行重新排序。

③ DF→COTS_{only}：只使用 COTS 匹配器的相似分数来对 k 个候选面孔进行排序。

④ DF →COTS 排名：将所有的 k 个候选对象分别用 COTS 和深层特征评分进行排序，然后使将这两个排名列表组合起来。

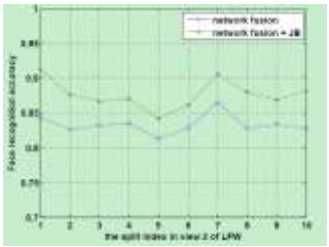


53. When Face Recognition Meets with Deep Learning: an Evaluation ofConvolutional Neural Networks for Face Recognition

- 网络融合可以显著提高人脸识别性能，因为不同的网络捕获来自不同区域和尺度的信息，形成强大的人脸表示

	Accuracy
single network	0.7882 ± 0.0037
network fusion	0.8333 ± 0.0042

- 采用联合贝叶斯方法的度量学习可以极大的提高人脸识别率。



- 倒装、镜像图像水平产生两个样本，是一种常用的人脸识别数据增强技术。一对测试图像可以产生 2 个新镜像的图像。这 4 张图像可以生成 4 对，而不是一对原件。为了结合这 4 幅图像/对，在这项工作中实现了两种融合策略（特征和分数融合）。对于

特征融合，测试图像和镜像图像的学习特征被连接到一个特征，然后将其用于分数计算。对于分数融合，从 4 对中产生的 4 个分数平均为一个分数。镜像图像提高人脸识别的性能。此外，特征融合的作品略优于评分融合，然而，改善不显著。

	Accuracy
no flip on test set	0.7882 ± 0.0037
feature fusion	0.7895 ± 0.0036
score fusion	0.7893 ± 0.0035

- 余弦和相关的识别率最好，但余弦的标准差小于相关的标准偏差。因此，余弦距离是这些距离中最好的。

Distance	Accuracy
euclidean	0.6898±0.0092
city block	0.6892±0.0088
chebychev	0.6692±0.0088
cosine	0.7882±0.0037
correlation	0.7882±0.0040
spearman	0.7878±0.0031

- 基于灰度和彩色图像的人脸识别精度分别为 0.7830±0.0077 和 0.7882±0.0118。使用灰色和彩色图像的性能非常接近。虽然彩色图像包含更多的信息，它们并没有带来重大的改善。

54. A Comprehensive Analysis of Deep Learning Based Representation for Face Recognition

主要思想：

- 提出了一种基于深度学习的人脸识别表示法（vgg-faces 和 light CNN）在姿态、光照、遮挡和错位等条件下的综合评价。
- 这些实验使用了五个著名的人脸数据集，即 AR 脸数据库分析遮挡效应，CMU PIE 和 Extended Yale dataset B 用于分析光照变化，Color FERET 数据库评估姿态变化的影响，以及 FRGC 数据库评估对齐失败的影响。

主要步骤：

① 尽管深度学习为人脸识别提供了一种强有力的表示，但它无法实现针对姿态、光照和隐蔽性的最先进的结果。为了使深度学习模型能取得更好的效果，在训练过程中要考虑到这些变化，对姿态光照等进行归一化处理。

② 基于深度学习的人脸表示法对不对齐具有较强的鲁棒性，能够容忍高达 10%的人脸特征定位误差。

③ VGG-face model 与 light cnn 模型相比，它具有更好的可移植性。

- The AR Face Database – Face Occlusion

Testing Set	VGG-Face		Lightened CNN
	FC6	FC7	
Sunglasses Session 1	33.64	**35.45**	5.45 (A)
Scarf Session 1	86.36	**89.09**	12.73 (A)
Sunglasses Session 2	**29.09**	28.18	7.27 (B)
Scarf Session 2	**85.45**	83.64	10.00 (A)

深度人脸表示法与最先进的遮挡鲁棒人脸识别算法相比，在戴太阳镜而造成的上脸遮挡时得到的结果较低。这些结果表明，除非对大量有遮挡的数据进行特殊训练，否则，当面部遮挡存在时，基于 cnn 的深度表示可能无法正常工作。在相同的实验中，VGG-人脸模型也被发现比简化的 CNN 模型对面部遮挡更有鲁棒性。

- CMU PIE Database – Illumination Variations

	VGG-Face		Lightened CNN
	FC6	FC7	
Accuracy	**93.16**	92.87	20.51 (A)

利用 VGG-面得到的深度表示对光照条件具有较强的鲁棒性.。然而，与现有的光照鲁棒人脸识别方法相比，所获得的识别精度略低。且 fc6 获得的结果比 fc7 好。

- Extended Yale Dataset – Illumination Changes

Testing Set	VGG-Face		Lightened CNN
	FC6	FC7	
Subset 2	**100**	**100**	82.43 (A)
Subset 3	88.38	**92.32**	18.42 (B)
Subset 4	46.62	**52.44**	8.46 (B)
Subset 5	13.85	**18.28**	4.29 (B)
Preprocessed Subset 4	71.80	**75.56**	26.32 (A)
Preprocessed Subset 5	73.82	**76.32**	24.93 (A)

深层人脸表示对子集 2 和 3 中存在的小光照变化具有较强的鲁棒性。但是，随着光照强度的增加，系统的性能会有明显的下降。解决这一问题的一个方法是在应用深度 CNN 模型进行特征提取之前，使用预处理。

- Color FERET Database - Pose Variations

Testing Set	VGG-Face		Lightened CNN
	FC6	FC7	
Quarter Left	**97.63**	96.71	25.76 (A)
Quarter Right	**98.42**	98.16	26.02 (A)
Half Left	**88.32**	87.85	6.08 (B)
Half Right	**91.74**	87.85	5.98 (A)
Profile Left	40.63	**43.60**	0.76 (B)
Profile Right	43.95	**44.53**	1.10 (B)

VGG-脸模型能够处理高达 67.5 度的姿态变化.。然而，采用姿态归一化方法可以进一步提高识别结果。当用 PRO LE 图像对系统进行测试时，系统的性能明显下降。

- The FRGC Database - Misalignment

基于深度 cnn 的人脸表示对错误对准具有较强的鲁棒性，即可以容忍人脸特征定位系统中多达 10%的眼间距离误差。

- Facial Bounding Box Extension

数据集的每一张图像都对齐并裁剪成一个扩展的方形，包括头部的所有部分，即耳朵、头发等。发现使用整个头部走识别的性能比仅用脸部好。

55. Large-Margin Softmax Loss for Convolutional Neural Networks

主要思想：

- 提出了卷积神经网络的 large margin softmax，large margin 定义了一个具有可调节裕度的灵活学习任务。
- 可以设置参数 M 来控制裕度，随着 M 的增大，类间的决策裕度也变大。
- large margin softmax 具有非常明显的直觉和几何解释。

主要步骤：

初始的 softmax 的目的是使得 WT1x>WT2x，即||W1||||x||cos(θ1)>||W2||||x||cos(θ2)，从而得到 x（来自类别 1）正确的分类结果。作者提出 large-magrin softmax loss 的动机是希望通过增加一个正整数变量 m，从而产生一个决策余量，能够更加严格地约束上述不等式，即：

$$\begin{aligned}\|\boldsymbol{W}_1\|\|\boldsymbol{x}\|\cos(\theta_1) &\geq \|\boldsymbol{W}_1\|\|\boldsymbol{x}\|\cos(m\theta_1)\\ &> \|\boldsymbol{W}_2\|\|\boldsymbol{x}\|\cos(\theta_2),\end{aligned} \tag{3}$$

其中 0≤θ1<π/m。如果 W1 和 W2 能够满足//W1//·//x//cos(mθ1)>//W2//·//x//cos(θ2)，那么就必然满足//W1//·//x//cos(θ1)>//W2//·//x//cos(θ2)。这样的约束对学习 W1 和 W2 的过程提出了更高的要求，从而使得 1 类和 2 类有了更宽的分类决策边界。

按照上节的思路，L-Softmax loss 可写为：

$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}+\sum_{j\neq y_i} e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right) \tag{4}$$

在这里，ψ(θ) 可以表示为：

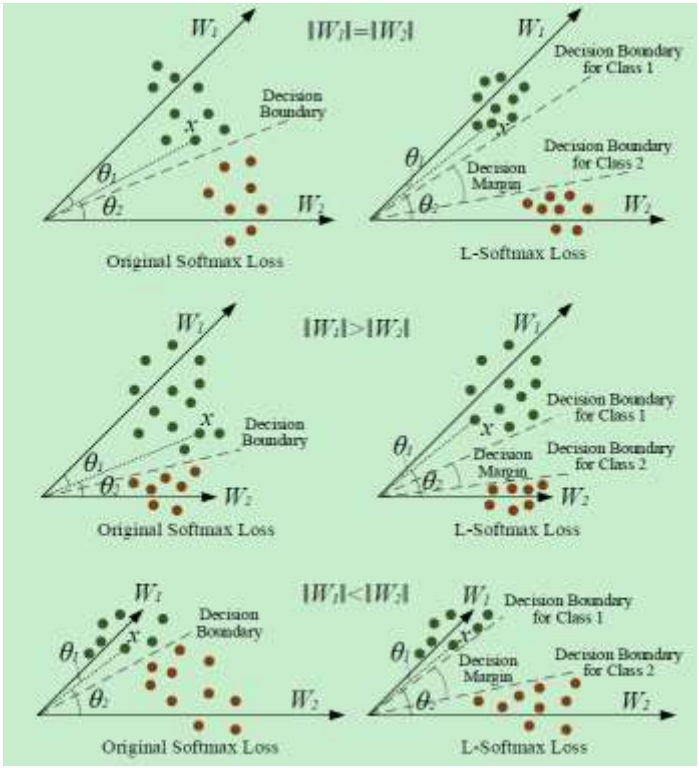
$$\psi(\theta) = \begin{cases}\cos(m\theta), & 0 \leq \theta \leq \frac{\pi}{m}\\ \mathcal{D}(\theta), & \frac{\pi}{m} < \theta \leq \pi\end{cases} \tag{5}$$

当 m 越大时，分类的边界越大，学习难度当然就越高。同时，公式中的 D(θ)必须是一个单调减函数且 D(π/m)=cos(π/m)， 以保证 ψ(θ) 是一个连续函数。

$$\psi(\theta) = (-1)^k \cos(m\theta) - 2k, \quad \theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}] \qquad (6)$$

其中 k 是一个整数且 k∈[0,m－1]。

在训练过程中，当 W1=W2 时，softmax loss 要求 θ1<θ2，而 L-Softmax 则要求 mθ1<θ2,我们从图中可以看到 L-Softmax 得到了一个更严格的分类标准。当 W1>W2 和 W1<W2 时，虽然情况会复杂些，但是同样可以看到 L-Softmax 会产生一个较大的决策余量。



56. SphereFace: Deep Hypersphere Embedding for Face Recognition

主要思想：

- 我们提出了角的 softmax（a-softmax）损失，使卷积神经网络（CNN）能够学习具有角度的鉴别特征。
- 在几何上，a-softmax 损失可以被看作是在一个超球面流形上设定区分性的限制，这本质上与人脸也位于一个流形的先验知识相匹配。
- 此外，可以通过参数 m 对角度隔离区大小进行定量调整，并进一步推导出近似理想特征准则的特定 m。。

主要步骤：

Loss Function	Decision Boundary
Softmax Loss	$(\boldsymbol{W}_1-\boldsymbol{W}_2)\boldsymbol{x}+b_1-b_2=0$
Modified Softmax Loss	$\|\boldsymbol{x}\|(\cos\theta_1-\cos\theta_2)=0$
A-Softmax Loss	$\|\boldsymbol{x}\|(\cos m\theta_1-\cos\theta_2)=0$ for class 1 $\|\boldsymbol{x}\|(\cos\theta_1-\cos m\theta_2)=0$ for class 2

- softmax 损失的问题

① softmax 损失仅学习分辨性不够强的特征,一些方法结合 softmax loss 和 contrastive loss,center loss。FaceNet 使用了 triplet losss。center loss 仅能使得类内紧凑。

② contrastive loss 和 triplet loss 还需要 pair/triplet 挖掘过程，耗时。

③ 所有的这些方法都使用欧式距离

- 改进的 softmax 损失

softmax 损失学习到的特征呈角度分布，说明欧式距离损失和 softmax 损失相容度不好。论文提出了角度距离。以二分类问题为例，softmax 损失决策边界为：

(W1-W2)x+b1-b2=0

论文约束||W1||=||W2||=1,b1=b2=0，决策边界成为：

$$||x||(\cos(\theta 1)-\cos(\theta 2))=0,$$

其中 θi 是 Wi 和 x 之间的 夹角，新的决策边界只依赖 θ1 和 θ2，改进后的 softmax 损失可以直接优化角度，CNN 可以学到呈角度分布的特征。这样第 i 类的特征相比其他类具有较小的 θi。

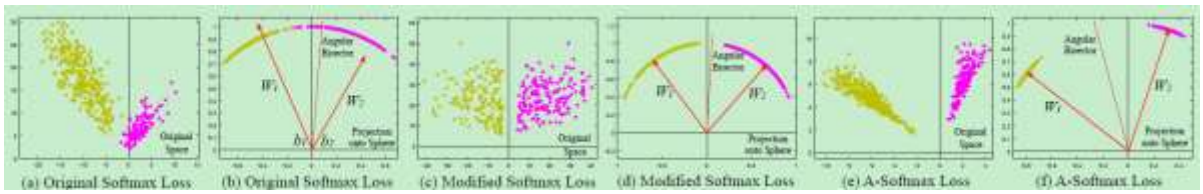
- A-softmax

接下来将损失改到 angular softmax，引入整数 m，量化决策边界。对二分类问题，类别 1 和类别 2 的决策边界不一样，分别为：

$$||x||(\cos(m\theta 1)-\cos(\theta 2))=0,\ 类\ 1:\theta_1<\frac{\theta_2}{m}$$

$$||x||(\cos(\theta 1)-\cos(m\theta 2))=0,\ 类\ 2:\theta_2<\frac{\theta_1}{m}$$

m 控制角度距离的尺寸，二分类问题可扩展到多分类问题，通过优化 A-Softmax，决策区域分的更开，拉大了类间距离，压缩了类内距离。几种损失函数学到的特征分布如下图所示，



定义 1：m_{min} 被定义为当 m>m_{min} 时有类内间的最大角度特征距离小于类间的最小角度特征距离。

性质 2：在二分类问题中：$m_{min}>2+\sqrt{3}$

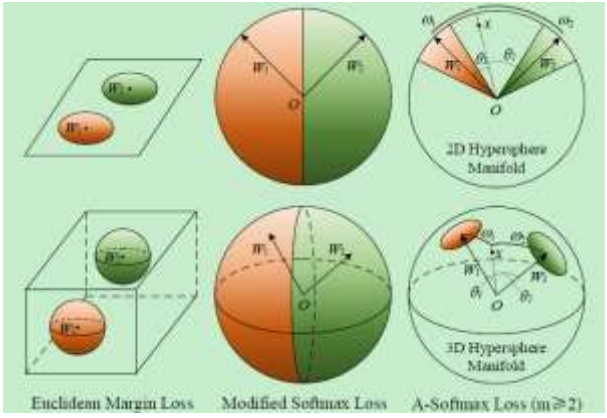
$$\underbrace{\frac{\theta_{12}}{m-1}}_{\text{max intra-class angle}}+\frac{\theta_{12}}{m+1}\leq\underbrace{\frac{(m-1)\theta_{12}}{m+1}}_{\text{min inter-class angle}},\ \theta_{12}\leq\frac{m-1}{m}\pi \quad (8)$$

$$\underbrace{\frac{2\pi-\theta_{12}}{m+1}}_{\text{max intra-class angle}}+\frac{\theta_{12}}{m+1}\leq\underbrace{\frac{(m-1)\theta_{12}}{m+1}}_{\text{min inter-class angle}},\ \theta_{12}>\frac{m-1}{m}\pi \quad (9)$$

有多分类问题中：$m_{min}\geq 3$。

$$\underbrace{\frac{\theta_i^{i+1}}{m+1}+\frac{\theta_{i-1}^{i}}{m+1}}_{\text{max intra-class angle}}\leq\underbrace{\min\left\{\frac{(m-1)\theta_i^{i+1}}{m+1},\frac{(m-1)\theta_{i-1}^{i}}{m+1}\right\}}_{\text{min inter-class angle}} \quad (10)$$

https://www.cnblogs.com/heguanyou/p/7503025.html



57. Additive Margin Softmax for Face Verification

主要思想：

- 提出了一种在特征和权值归一化的情况下，对Softmax损失引入加性margin策略。

主要步骤：

- A-softmax loss

$$\mathcal{L}_{AS}=-\frac{1}{n}\sum_{i=1}^{n}log\frac{e^{\|\boldsymbol{f}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{f}_i\|\psi(\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{c}e^{\|\boldsymbol{f}_i\|\cos(\theta_j)}},$$

$$\psi(\theta)=\frac{(-1)^k\cos(m\theta)-2k+\lambda\cos(\theta)}{1+\lambda},$$
$$\theta\in[\frac{k\pi}{m},\frac{(k+1)\pi}{m}].$$

- Additive Margin Softmax

我们进一步提出了一个具体的$\varphi(\theta_{y_i})$，它引入了一个附加的 margin 到 Softmax 损失函数。这个公式是：

$$\psi(\theta)=cos\theta-m.$$

与 L-Softmax 和 A-Softmax 中定义的$\varphi(\theta_{y_i})$进行比较，我们的定义更简单，可以在不调整过多的超参数的情况下，将较大的裕度属性引入到功能中。

由于我们使用余弦作为相似度来比较两种人脸特征，将特征归一化和权重归一化应用于内积层，以建立余弦层。然后，我们使用超参数 s 缩放余弦值。最后，损失函数变成：

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

显然，我们改进的 Softmax 损失函数是对余弦相似度进行优化，而不是对角度进行优化。如果我们使用传统的 Softmax 损失，这可能不是问题，因为这两种形式的决策边界是相同的$(\cos\theta_1=\cos\theta_2\Rightarrow\theta_1=\theta_2)$。然而，当我们试图推展边界时，我们将面临这样一个问题：这两个相似点(距离)具有不同的密度。余弦值在夹角接近于 0 或π时更密集，如果要优化角度，在得到内积W^Tf值后，可能需要 arccos 操作，在计算上它可能会更昂贵。

58. ArcFace: Additive Angular Margin Loss for Deep Face Recognition

Related work 里面有不同架构有哪些文章，以及对于 loss 函数的分类，对 Feature Normalisation 的文章进行了综述

主要思想：

- 提出了一种新的监控信号—additiveangular margin (ArcFace)，它比目前提出的监督信号具有更好的几何解释能力。具体而言，本文提出的 ArcFace$\cos(\theta+m)$基于 L2 归一化权值和特征，直接使角(弧)空间中的判决边界最大化。与乘法角余量$\cos(m\theta)$ cos(M)和加性余弦$\cos\theta-m$比较，弧面可以获得更具判别性的深度特征。
- 探索不同的网络设置，并分析精度和速度之间的权衡（explore MobileNet, Inception-Resnet-V2, Densely connected convolutional networks(DenseNet), Squeeze and excitation networks(SE) and Dual path Network (DPN)）

主要步骤：

- SphereFace

$$L_3=-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\|x_i\|\cos(m\theta_{y_i})}}{e^{\|x_i\|\cos(m\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{n}e^{\|x_i\|\cos\theta_j}},$$

其中$\theta_{y_i}\in[0,\pi/m]$，为了消除这个限制，构造了另一个函数：

$$L_4=-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{\|x_i\|\psi(\theta_{y_i})}}{e^{\|x_i\|\psi(\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{n}e^{\|x_i\|\cos\theta_j}},$$

其中$\psi(\theta_{y_i})=(-1)^k\cos(m\theta_{y_i})-2k,\theta_{y_i}\in\left[\frac{k\pi}{m},\frac{(k+1)\pi}{m}\right],k\in[0,m-1],\ m\geqslant1$，然而，在 SphereFaces 的实施过程中，采用了 Softmax 监督来保证训练的收敛性，并通过一个动态的超参数λ来控制训练的权重。这个附加的动态超参数使得训练比较棘手。

- Additive Cosine Margin

特征和重量归一化可以消除径向变化，并推动每个特征分布在超球面流形上。我们把$\|x_i\|$修正为 L2 规范，再把$\|x_i\|$重新设为 s。

$$L_5=-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}. \tag{8}$$

(1) 易于实现，不需要复杂的超参数；(2)在没有 Softmax 监督的情况下，更清晰、更能收敛；(3)性能有明显的改善。

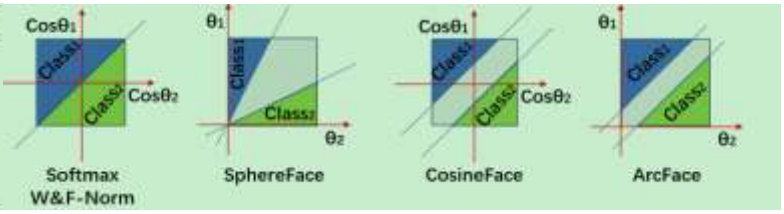
- Additive Angular Margin

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

$$W_j = \frac{W_j}{\|W_j\|}, x_i = \frac{x_i}{\|x_i\|}, \cos\theta_j = W_j^T x_i.$$

如果我们扩展建议的加性角余量$\cos(\theta+m)$，则得到$\cos(\theta+m)=\cos\theta\cos m-\sin\theta\sin m$。与相加余弦余弦余量$\cos(\theta)-m$相比，ArcFace相似，但$\sin\theta$导致边缘是动态的。

Loss Functions	Decision Boundaries
Softmax	$(W_1-W_2)x+b_1-b_2=0$
W-Norm Softmax	$\|x\|(\cos\theta_1-\cos\theta_2)=0$
SphereFace [23]	$\|x\|(\cos m\theta_1-\cos\theta_2)=0$
F-Norm SphereFace	$s(\cos m\theta_1-\cos\theta_2)=0$
CosineFace [44, 43]	$s(\cos\theta_1-m-\cos\theta_2)=0$
ArcFace	$s(\cos(\theta_1+m)-\cos\theta_2)=0$



59. L2-constrained Softmax Loss for Discriminative Face Verification

主要思想：

- 对 softmax loss 进行 L2-限制实现人脸验证系统。该约束强制特征位于以参数α为固定半径的超球面上。我们
- 还提供了实现一致性能的值的界限

主要步骤：

softmax loss 优点：可以很容易地使用公开可用的工具箱中内置的函数来实现。对输入 batch 大小没有任何限制，并快速收敛。

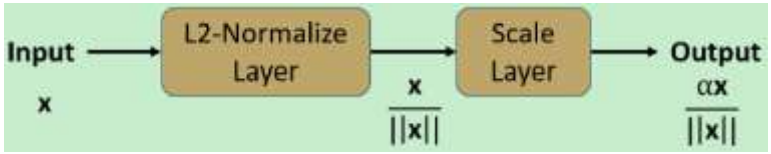
softmax loss 缺点：对样本分布有偏向，对高品质的面孔适合，忽略了训练中少有的困难面孔。高质量正面脸的特征具有较高的 L2-范数，而姿态极端的模糊脸具有较低的 L2-范数。没有优化保持正对和负对之间距离较远的验证要求。

提出一种 L2-Softmax 损失，它在训练过程中增加了对特征的约束，使它们的 L2-范数保持不变。换句话说，把特征限制在一定半径的超球面上。首先，它提供了同样的关注好的和坏的质量的面孔，因为所有的特征都有相同的 L2-规范。其次，它通过使同一主题特征在归一化空间中更加接近，不同主题特征之间的距离较远来增强验证信号。

$$\begin{aligned}&\text{minimize} \quad -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C}e^{W_j^T f(\mathbf{x}_i)+b_j}} \qquad (3)\\ &\text{subject to} \quad \|f(\mathbf{x}_i)\|_2=\alpha,\ \forall i=1,2,...M,\end{aligned}$$

该模块加在 CNN 的倒数第二层，作为特征描述符。L2-normalize layer 标准化输入特征 x 到一个单位向量。scale layer 将输入单元向量缩放到由参数给定的固定半径

$$\mathbf{y}=\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \quad \mathbf{z}=\alpha\cdot\mathbf{y}$$



- Bounds on Parameterα

执行 L2 约束的方法有两种：1)在整个训练过程中保持α固定，2)让网络学习参数α。但网络学习的参数很高，这就导致了 L2-约束的放松。而α价值很低，训练不能收敛。给定数据集的类数 C，我们可以得到下界α，从而得到概率分数 p。

$$\alpha_{low}=\log\frac{p(C-2)}{1-p}$$

60. DeepVisage: Making face recognition simple yet with powerful generalizationskills

对于架构，损失函数等的分析特别好

主要思想：

- CNN 模型利用了剩余学习框架。此外，它还使用标准化的特征来计算损失
- 给定一个 probe 的脸，我们的第一个 filter 使用从卷积神经网络生成特征从 gallery 里找到 top-k 最相似的面孔。K 个候选照

片再结合深度特征相似性和 COTS 匹配重新排名。

主要步骤：

- Convolutional networks

CNN 模型由 27 个卷积(COV)、4 个池(池)和 1 个完全连接(FC)层组成。每个卷积使用一个 3*3 核，然后是一个 PReLU 激活函数。CNN 利用 2*2 最大池层降低空间分辨率，同时逐渐增加特征图的数量，从 32 幅增加到 512 幅。在最后一层之后，我们使用了一个由 512 个神经元组成的 FC 层。我们将这个 fc 层的输出规范化并将其视为输入图像的所需特征表示。最后，我们使用 Softmax 层来计算和优化训练过程中的损失。

- Residual learning framework

- Loss function

$$\mathcal{L}_{Softmax} = -\sum_{i=1}^{N} log \frac{e^{W_{y_i}^T f_i + b_{y_i}}}{\sum_{j=1}^{K} e^{W_j^T f_i + b_j}}$$

- Feature normalization (FN)

将归一化特征f^{Nr}提供给 Softmax 损失如下：$f^{Nr} = \frac{f^{Pr} - \mu}{\sqrt{\sigma^2}}$，μ和σ是均值和方差。

在训练过程中，我们通过计算每个 batch 中样本的μ和σ进行规范化.。此外，我们保持μ和σ，并使用它们对测试样本进行了规范化。

- Pre-processing

(A) 使用检测器 mtcnn 探测人脸和 landmarks；(B)应用二维相似变换对人脸图像进行归一化。转换参数由图像上检测到的地标位置和 11296 图像帧中的预置坐标计算；(C)转换为灰度。

61. NormFace: L 2 Hypersphere Embedding for Face Verification

主要思想：

- 在对分类模型进行训练时，我们建议对最后一个内积层的特征和权重进行 L2 归一化运算。我们从解析和几何两个角度解释了规范化运算的必要性。
- 提出了两种损失函数来训练归一化特征。一种是在余弦分数和损失之间插入一个带有 scale 的重构的 Softmax 损失。另一种是受度量学习启发而设计的。为了避免选择硬样本挖掘的需要，我们提出了一种 Agent 策略。

主要步骤：

大部分 works 中，在训练过程中，没有采用特征提取后的归一化。但是在测试阶段，所有的方法都使用归一化相似度，例如余弦，来比较两个特征。在测试过程中，特性规范化似乎是获得良好性能的关键一步。

Similarity	Before Normalization	After Normalization
Inner-Product	98.27%	98.98%
Euclidean	98.35%	98.95%

the Cosine Loss[17], vMFMM[21]和我们建议的损失函数同时规范化了特征和权重，而 L2-softmax[24]只对特征规范化和 SphereFace 只对权重进行规范化。

- Reformulating Softmax Loss

如果直接将特征和权重值归一化为 1,则 softmax loss 不会收敛。通过将特征和权重列标准化为较大的值 s 而不是 1,softmax loss 可以继续减少。在实践中，我们可以通过在余弦层之后直接添加一个缩放层来实现这一点。(后来实验指出，缩放层主要是针对特征，权重为 1 不受影响)

$$\mathcal{L}_S = -\frac{1}{m}\sum_{i=1}^{m} log \frac{e^{s\hat{W}_{y_i}^T \hat{f}_i}}{\sum_{j=1}^{n} e^{s\hat{W}_j^T \hat{f}_i}},$$

- Reformulating METRIC LEARNING

然而，由于度量学习模型中可能存在的输入对或三重态，即 O(N2)组合用于配对，O(N3)组合用于三胞胎，其中 N 是训练样本的数

量，因此，度量学习比分类更有价值。在训练过程中几乎不可能处理所有可能的组合，因此通常需要采样和硬挖掘算法。相比之下，在分类任务中，我们通常将数据迭代地输入到模型中，即输入数据的顺序是O(N)。在本节中，我们尝试重新构造一些度量学习损失函数来完成分类任务，同时保持它们与规范化特征的兼容性。

Softmax 可以写成，因为 $\|\tilde{\mathbf{x}}-\tilde{\mathbf{y}}\|_2^2 = 2-2\tilde{\mathbf{x}}^T\tilde{\mathbf{y}}$.

$$\mathcal{L}_S = -\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{s\tilde{W}_{y_i}^T\tilde{f}_i}}{\sum_{j=1}^{n} e^{s\tilde{W}_j^T\tilde{f}_i}} = -\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{-\frac{s}{2}\|\tilde{f}_i-\tilde{W}_{y_i}\|_2^2}}{\sum_{j=1}^{n} e^{-\frac{s}{2}\|\tilde{f}_i-\tilde{W}_j\|_2^2}},$$

我们将其中一个特征修改为加权矩阵 W∈Rd×n 的一列，其中 d 是特征的维数，n 是类数。我们称列 Wi 为第 i 类的“代理”。权重矩阵 W 可以像内积层一样通过反向传播来学习。

$$\mathcal{L}_C = \begin{cases} \|\tilde{f}_i-\tilde{W}_j\|_2^2, & c_i=j \\ \max(0, m-\|\tilde{f}_i-\tilde{W}_j\|_2^2), & c_i\neq j \end{cases}, \quad (10)$$

and the triplet loss,

$$\mathcal{L}_T = max(0, m+\|\tilde{f}_i-\tilde{W}_j\|_2^2-\|\tilde{f}_i-\tilde{W}_k\|_2^2), \quad c_i=j, c_i\neq k. \quad (11)$$

实验结果：

Table 2: Results on LFW 6,000 pairs using Wen's model[36]

loss function	Normalization	Accuracy
softmax	No	98.28%
softmax + dropout	No	98.35%
softmax + center[36]	No	99.03%
softmax	feature only	98.72%
softmax	weight only	98.95%
softmax	Yes	99.16% ± 0.025%
softmax + center	Yes	99.17% ± 0.017%
C-contrasitve	Yes	99.15% ± 0.017%
C-triplet	Yes	99.11% ± 0.008%
C-triplet + center	Yes	99.13% ± 0.017%
softmax + C-contrastive	Yes	99.19% ± 0.008%

Table 5: Results on YTF with Wen's model[36]

loss function	Normalization	Accuracy
softmax + center[36]	No	93.74%
softmax	Yes	94.24%
softmax + HIK-SVM	Yes	94.56%
C-triplet + center	Yes	94.3%
C-triplet + center + HIK-SVM	Yes	94.58%
softmax + C-contrastive	Yes	94.34%
softmax + C-contrastive + HIK-SVM	Yes	94.72%

62. von Mises-Fisher Mixture Model-based Deeplearning: Application to Face Verification

列举了一些文章，使用不同的网络架构

主要思想：

- 以 vmf 混合模型为理论基础，提出了一种统计特征表示(SFR)模型。接着，我们
- 提出了一种有效的方向特征学习方法，称为 vmf-fl，它构造了一个新的损失函数 vmfml。
- VMFML 损失和 softmax 的关系：(A)vmfml 使用单位归一化特征：$\mathbf{x}=\frac{\mathbf{r}}{\|\mathbf{r}\|}$；(B)平均参数μ与 Softmax 权重的关系为：$\mu=\frac{\mathbf{w}}{\|\mathbf{w}\|}$；(C)它没有偏差，(D)它有一个附加参数 k。可以说，vmfml 通过对权值和特征向量进行归一化，并将加性偏差项替换为乘性标量项，从而提供了 Softmax 损失的另一种形式。

主要步骤：

- Statistical Features Representation (SFR) Model

vmf 分布是用平均方向(以实线表示)和浓度(表示特征点从固体线上的扩展)来参数化的。对于 ith 图像特征 xi，让我们称之为 vmf 特征，我们定义了带有 m 类的 sfr 模型为：

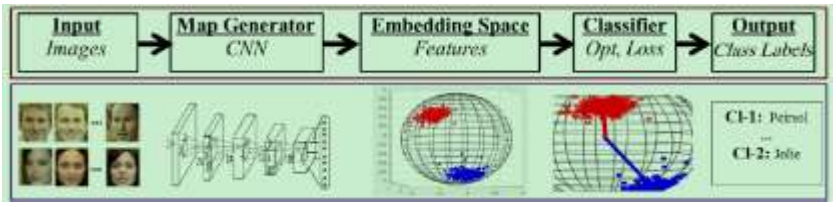
$$SFR(\mathbf{x}_i|\Theta_M)=\sum_{j=1}^{M}\pi_j V_d(\mathbf{x}_i|\mu_j,\kappa_j)$$

其中π_j,μ_j和k_j分别表示 j 类的混合比例、平均方向和浓度值。θ_M是模型参数的集合，vd(.)是 vmf 分布的密度函数。

SFR 模型对类进行了相同的特权假设，即每一个 j 类都有相同的出现概率，并且具有相同的浓度值。这个假设对于鉴别学习很重要，以确保监督分类器不偏倚于任何特定的类，而不管样本数和变化量在。另一方面，j 在保持其各自空间中的每个标识方面起着重要作用。

- vMF Features Learning (vMF-FL) Method

vmf-fl 方法由两个子任务组成：(1)利用 CNN 模型将输入的 2d 对象图像映射到 vmf 特征；(2)基于 SFR 模型的判别视图，将特征分类到相应的类。它通过将 SFR 模型和 CNN 模型相结合来求解一个优化问题，并通过最小化分类损失来学习参数。CNN 模型一般采用 Softmax 函数，而该方法利用 von Mises-Fisher Mixture Loss (vMFML)进行优化。



- SFR model and von Mises-Fisher Mixture Loss (vMFML)

  对于 d 维随机单位向量 $\mathbf{x} = [x_1, \ldots, x_d]^T \in S^{d-1} \subset \mathbb{R}^d \, (i.e., \|\mathbf{x}\|_2 = 1)$，vmf 分布的密度函数定义为

$$V_d(\mathbf{x}|\mu, \kappa) = C_d(\kappa) \exp(\kappa \mu^T \mathbf{x})$$

  其中，$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ 是归一化常数，其中，$I_\rho(.)$ 是第一类修正的贝塞尔函数。

  因此，我们在 SFR 模型的等特权假设基础上重写了后验概率：

$$p_{ij} = \frac{\exp(\kappa \mu_j^T \mathbf{x}_i)}{\sum_{l=1}^{M} \exp(\kappa \mu_l^T \mathbf{x}_i)}$$

  可以利用后验/条件概率最小化交叉熵并定义损失函数，称为 vmfml：

$$\mathcal{L}_{vMFML} = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log(p_{ij}) = -\sum_{i=1}^{N} \log \frac{\exp(\kappa \mu_j^T \mathbf{x}_i)}{\sum_{l=1}^{M} \exp(\kappa \mu_l^T \mathbf{x}_i)} = -\sum_{i=1}^{N} \log \frac{e^{z_{ij}}}{\sum_{l=1}^{M} e^{z_{il}}} \quad \left[z_j = \kappa \mu_j^T \mathbf{x}_i\right]$$

63. Face Recognition via Centralized CoordinateLearning

主要思想：

- 同时对 w 和 x 进行归一化
- 利用分类向量的 L2 范数对分类向量进行归一化处理，并将人脸特征的各个维数集中到零均值，具有单位方差。此外，还定义了一个自适应角 margin，以进一步提高相邻类的可分性。

主要步骤：

- Centralized Feature Learning

  对于权重 w 进行归一化

$$\mathcal{L}_{sf} = \sum_{i}^{N} -\log\left(\frac{\exp\left(\frac{\mathbf{w}_{y_i}^T \Phi(\mathbf{x}_i)}{\|\mathbf{w}_{y_i}\|}\right)}{\sum_{k=1}^{K} \exp\left(\frac{\mathbf{w}_k^T \Phi(\mathbf{x}_i)}{\|\mathbf{w}_k\|}\right)}\right).$$

$$\mathcal{L}_{sf} = \sum_{i}^{N} -\log\left(\frac{\exp(\|\Phi(\mathbf{x}_i)\| \cos(\theta_{y_i,i}))}{\sum_{k=1}^{K} \exp(\|\Phi(\mathbf{x}_i)\| \cos(\theta_{k,i}))}\right),$$

  如果 $\|\Phi(\mathbf{x}_i)\|$ 小，SOFTMAX 预计所有样品的概率将相近，这样的损失将不具有区分性。如果 $\|\Phi(\mathbf{x}_i)\|$ 大，概率可能很大，使 DNN 不稳定的学习。另一方面，在测试阶段，计算出两个人脸特征向量的余弦相似度，用于人脸识别。理想情况下，ϕ（x）分布在整个坐标空间中，使得两个不同主题的人脸特征向量在大角度下更容易分离。

  在学习过程中，我们建议将人脸特征集中到空间的原点。具体来说，对于特征向量 x 的每个维 j，我们将ϕ(x(j))定义为：

$$\Phi(\mathbf{x}(j)) = \frac{\mathbf{x}(j) - \mathbf{o}(j)}{\boldsymbol{\sigma}(j)},$$

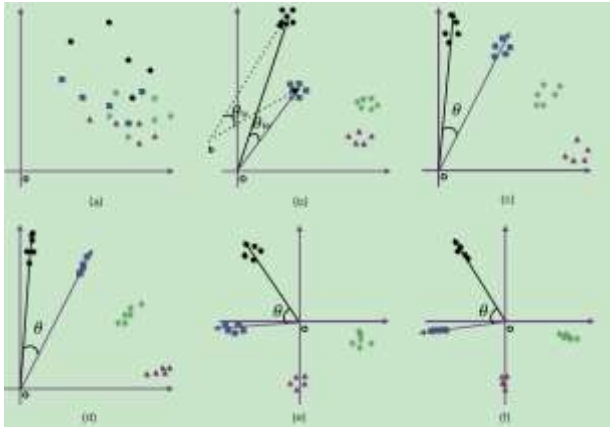
  O 是均值，σ是方差。将 x 的每个维集中到原点，使特征ϕ（x）跨越坐标空间的所有象限。同时，ϕ（x）的每个维数都有相同的单位方差，因此每个维都将同样地有助于人脸的识别，而不是仅仅使用几个强维来进行人脸识别。

- Adaptive Angular Margin

$$p_{y_i}^{AAM} = \frac{\exp(\|\Phi(\mathbf{x}_i)\| \cos(\eta\theta_{y_i,i}))}{\exp(\|\Phi(\mathbf{x}_i)\| \cos(\eta\theta_{y_i,i}) + \sum_{k \neq y_i} \exp(\|\Phi(\mathbf{x}_i)\| \cos(\theta_{k,i}))}, \quad (20)$$

where η is an adaptive parameter and it is set based on the value of $\theta_{y_i,i}$:

$$\eta = \begin{cases} 1, & \pi/3 < \theta_{y_i,i} <= \pi; \\ \frac{\pi/3}{\theta_{y_i,i}}, & \pi/30 < \theta_{y_i,i} <= \pi/3; \\ 10, & \theta_{y_i,i} <= \pi/30. \end{cases} \quad (21)$$



(A)原始数据分配。(B)使用原始的 Softmax 损失来融合人脸特征。(C).使用 a-Softmax 损失来聚合人脸特征。(D)通过使用 SphereFace loss 来融合面部特征。(E)通过使用 CCL 损失来融合面部特征（四个象限）。(F)利用 CCL 损失与 AAM 融合面部特征。

64. Noisy Softmax: Improving the Generalization Ability of DCNN via Postponingthe Early Softmax Saturation

主要思想：

- 首先强调了 Softmax 的早期饱和行为会阻碍 SGD 的探索，这有时是模型收敛于差的局部极小值的原因之一。
- 提出在每次迭代过程中，通过在 Softmax 中注入退火噪声来缓解这一早期个体饱和问题。这种基于噪声注入的方法旨在推迟早期饱和，并进一步引入连续梯度传播，从而显着地鼓励 sgd 解算器进行更多的探索，帮助寻找更好的局部极小值。

主要步骤：

- Early Individual Saturation

  传统的 softmax 得到交叉熵损失和偏导数如下：

$$L = -\frac{1}{N}\sum_i \log P(y_i|x_i) = -\frac{1}{N}\sum_i \log \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \quad (1)$$

$$\frac{\partial L}{\partial f_j} = P(y_i = j|x_i) - 1\{y_i = j\} = \frac{e^{f_j}}{\sum_k e^{f_k}} - 1\{y_i = j\}$$

当使用基于梯度的方法(如 sgd)优化 cnn 时，过早饱和的个体由于可忽略的梯度而停止了对反向传播的贡献，$P(y_i = 1|x_i) \approx 1, \frac{\partial L}{\partial f_{y_i}} \approx 0$，随着饱和个体数的增加，贡献数据的数量减少，SGD 几乎没有机会移动，并且更有可能在局部极小值处收敛，因此，很容易过度拟合，需要额外的数据才能恢复。

- Noisy Softmax

  因此，我们减缓早期饱和的技术是在 Softmax 输入 fyi 中注入适当的噪声，由此产生的噪声相关噪声如下：

$$f_{y_i}^{noise} = f_{y_i} - n, \quad n = \mu + \sigma\xi, \xi \sim N(0,1)$$

使$f_{y_i}^{noise}$比f_{y_i}小，因为如果$f_{y_i}^{noise} > f_{y_i}$，会加速饱和。所以需要噪声 n 始终是正的：

$$f_{y_i}^{noise} = f_{y_i} - \sigma|\xi|$$

而加噪声的目的是打算推迟 xi 的早期饱和，而不是不允许它饱和，这意味着最初需要较大的噪声来提高勘探能力，而随后则需要相对较小的噪声才能使模型收敛。因此，

$$f_{y_i}^{noise} = f_{y_i} - \alpha\|W_{y_i}\|\|X_i\|(1 - \cos\theta_{y_i})|\xi| \quad (5)$$

最后，noisy softmax 损失定义为：

$$L = -\frac{1}{N}\sum_i \log \frac{e^{f_{y_i} - \alpha\|W_{y_i}\|\|X_i\|(1-\cos\theta_{y_i})|\xi|}}{\sum_{j\neq y_i} e^{f_j} + e^{f_{y_i} - \alpha\|W_{y_i}\|\|X_i\|(1-\cos\theta_{y_i})|\xi|}}$$

65. Deep Hyperspherical Learning

主要思想：

- 提出了一种新的学习框架-超球面卷积(SphereConv)，它给出了超球面上的角表示。
- 引入了 SphereNet，deep hyperspherical convolution networks 不同于传统的基于内积的卷积网络。尤其是，SphereNet 采用 SphereConv 作为其基本卷积算子，并采用 angular softmax loss 作为损失函数。

- Sphere 卷积有一个好处是能够保证输出分布一致(自归一化能力)，所以可以去掉 Batch Norm 层。从实验来看似乎比传统卷积收敛快且更稳定。

主要步骤：

- 网络架构

传统的 Conv 是内积的形式，$\mathcal{F}(w,x)=w^{\top}x+b_{\mathcal{F}}$，其中 w 是卷积滤波器，x 表示底部特征映射的局部块，bf 表示偏差。这里的矩阵乘法实质上是计算局部块和滤波器之间的相似性。因此，标准的卷积层可以看作是逐片矩阵乘法。与标准卷积算子不同，超球面卷积算子计算超球面上的相似性：

$$\mathcal{F}_s(w,x)=g(\theta_{(w,x)})+b_{\mathcal{F}_s},$$

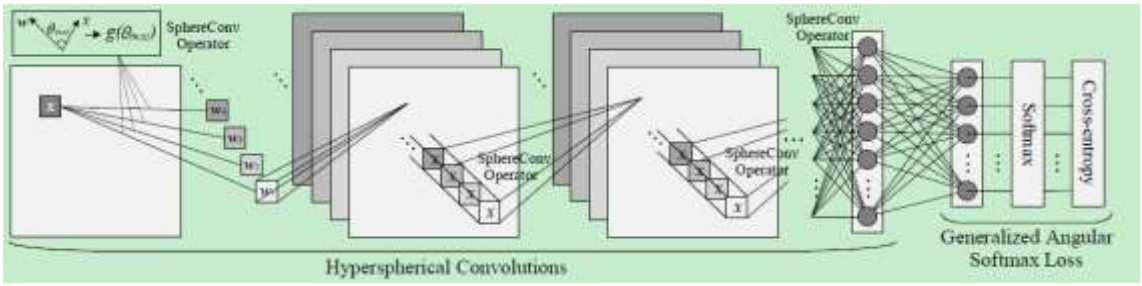
$\theta_{(w,x)}$ 为卷积核与输入patch x 的角度，g($\theta_{(w,x)}$)是$\theta_{(w,x)}$的函数 ，文中主要提出了三种 Sphere Conv 形式:Linear Cosine Sigmoid，其实都是上面的广义形式的具体化，其中所有 Sphere Conv 输出为[0,1]。

Linear SphereConv: $g(\theta_{(w,x)})=a\theta_{(w,x)}+b,$

Cosine SphereConv: $g(\theta_{(w,x)})=\cos(\theta_{(w,x)}),$

Sigmoid SphereConv：当 k 接近 0 时，g（$\theta_{(w,x)}$）将近似于阶跃函数。当 k 变大时，g（$\theta_{(w,x)}$）更像线性函数，即线性球函数。SigmoidSpheconv 是参数 Spheconv 族的一个实例。随着参数的增加，参数球具有更丰富的表示能力。

$$g(\theta_{(w,x)})=\frac{1+\exp(-\frac{\pi}{2k})}{1-\exp(-\frac{\pi}{2k})}\cdot\frac{1-\exp\left(\frac{\theta_{(w,x)}}{k}-\frac{\pi}{2k}\right)}{1+\exp\left(\frac{\theta_{(w,x)}}{k}-\frac{\pi}{2k}\right)},$$



- Loss 函数

Weight-normalized Softmax Loss：

$$L_i=-\log\left(\frac{e^{\|x_i\|g(\theta_{y_i,i})}}{\sum_j e^{\|x_i\|g(\theta_{j,i})}}\right),$$

Generalized Angular Softmax Loss：

$$L_i=-\log\left(\frac{e^{\|x_i\|g(m\theta_{y_i,i})}}{e^{\|x_i\|g(m\theta_{y_i,i})}+\sum_{j\neq y_i}e^{\|x_i\|g(\theta_{j,i})}}\right),$$

66. Deep Convolutional Neural Network Features and the Original Image

主要思想：

- 首先，dcnntop-level features 保留了大量关于原始输入图像的信息。Yaw，pitch 和媒体类型在顶级 dcnn 代码中很容易获得，并且可以以高精度进行分类。
- 在 dcnn 空间中找到了一个的图像质量指标：与空间远点的距离。低质量的图像在原点周围聚集，随着距离的增加图像质量随之增加。

主要步骤：

- Predicting Yaw, Pitch, Media Type

Network	Yaw	Pitch
A	+/-8.06 degs. (*sd.* 0.078)	77.0% correct
B	+/-8.59 degs. (*sd.* 0.071)	71.5% correct

Network	Media Type
A	87.1% (sd. 0.004)
B	93.3 % (sd. 0.002)

- poor “quality”

2) 迁移学习

67. Face Recognition Using Deep Multi-Pose Representations

主要思想：

- 提出采用 multiple pose-aware 深度学习模型进行人脸识别的方法和系统。
- 人脸图像由几个特定位置的深卷积神经网络(Cnn)模型处理，以产生多个特定于姿态的特征。
- 不同的 CNN 网络由 AlexNet 和 VGG 通过迁移学习得到。
- 三维渲染是用来从输入图像生成多个人脸姿态的。

主要步骤：

- Facial Landmark Detection and Face Alignment

面部 landmarks 检测算法 lmd，获取人脸图像 x 并估计 n 个预定义的关键点，如眼角、嘴角、鼻尖等。且根据定义的面部关键点的数量，可以粗略地将 landmarks 分为两类：(1)稀疏 landmarks，如 5 点。[23]和(2)密集 landmarks，例如在 68 点。

$$\mathrm{lmd}(X) = \begin{bmatrix} P_x^1 & P_y^1 \\ P_x^2 & P_y^2 \\ \vdots & \vdots \\ P_x^n & P_y^n \end{bmatrix}$$

文中使用非反射相似变换来进行平面内 2D 对准，而对于面外 3D 对齐使用透视变换。

2D：$[\mathrm{lmd}(X)\,|\,1]$是 LMD(X)加入一个全 1 向量，T 是一个齐次矩阵由旋转角度θ、缩放因子 s 和平移矢量[TX；TY]组成

$$T^* = \underset{T}{\operatorname{argmin}} \|T\,[\mathrm{lmd}(X)\,|\,\mathbf{1}]^t - [\mathrm{lmd}(R)\,|\,\mathbf{1}]\|_2^2 \quad (2)$$

$$T = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

3D：与 2d 比对不同，我们的 3D 对齐依赖于三维通用人脸形状模型，一旦成功地将通用的三维人脸形状模型与给定的人脸图像相匹配，我们就可以用任意的偏航俯仰滚动参数来绘制人脸图像。

- Face Representation

Pip. Acronym	Alignment	Ref. Model	Feature	Rep. Dim.
HLBP	in-plane	avg-face	HDLBP	100,000
ALEX-AF	in-plane	avg-all-face-lmd	AlexNet	4,000
ALEX-FF	in-plane	avg-frontal-face-lmd	AlexNet	4,000
ALEX-PF	in-plane	avg-profile-face-lmd	AlexNet	4,000
ALEX-FY0	out-of-plane	gene-face-yaw@0	AlexNet	4,000
ALEX-FY45	out-of-plane	gene-face-yaw@45	AlexNet	4,000
VGG16-AF	in-plane	avg-face-lmd	VGG16	4,000
VGG19-AF	in-plane	avg-all-face-lmd	VGG19	4,000
VGG19-FF	in-plane	avg-frontal-face-lmd	VGG19	4,000
VGG19-PF	in-plane	avg-profile-face-lmd	VGG19	4,000
VGG19-FY0	out-of-plane	gene-face-yaw@0	VGG19	4,000
VGG19-FY45	out-of-plane	gene-face-yaw@45	VGG19	4,000
VGG19-FY75	out-of-plane	gene-face-yaw@75	VGG19	4,000

HDLBP 代表高维局部二进制模式。alexnet,vgg16 和 vgg19 都是指 cnn。Ref. Model 表示用于人脸对齐的参考模型,"avg-all-fac-lmd "表示使用所有训练数据的平均地标向量，而" gene-fac-yaw@45 "则表示使用通用的三维人脸模型在 45 度 yaw and 0 度 pitch。

- Transfer Learning

使用 ilsvrc 2014 图像分类任务中公开可用的模型初始化我们的 cnn。保留了所有 CNN 层的所有权重，除了最后一层，因为最后一层的输出节点数量必须对应于 WebFaces 中的身份书，并使用随机权重重新初始化该层。

然后我们应用微调来学习特定的姿态特征。CNN 微调的实质是总是向前迈进一步。比如使用 alex-AF 模型作为我们的基本模型，并进一步训练 alex-FF 和 alex-PF，它们分别关注于近额面和近轮廓面。基于 alex-FF cnn，可以进一步用生成的 0yaw 的人脸微调它，获得 ALEX-FY0 CNN。

Pip. Acronym	Learning Type	Base CNN Model	Training Partition
ALEX-AF	transfer	AlexNet	all real
ALEX-FF	finetune	ALEX-AF	real frontal
ALEX-PF	finetune	ALEX-AF	real profile
ALEX-FY0	finetune	ALEX-FF	rendered yaw0
ALEX-FY45	finetune	ALEX-PF	rendered yaw45
VGG16-AF	transfer	VGG16	all real
VGG19-AF	transfer	VGG19	all real
VGG19-FF	finetune	VGG19-AF	real frontal
VGG19-PF	finetune	VGG19-AF	real profile
VGG19-FY0	finetune	VGG19-FF	rendered yaw0
VGG19-FY45	finetune	VGG19-PF	rendered yaw45
VGG19-FY75	finetune	VGG19-FY45	rendered yaw75

- MultiModalRepresentation for Recognition

(1) 比较来自同一表示管道的特征之间的相似性评分，(2)融合不同表示管道之间的相似性分数：

$$\mathrm{sim}(X,Y)=\mathrm{fuse}\left(\{\mathrm{rsim}(\mathrm{rep}_j(X),\mathrm{rep}_j(Y)\}|_{j=1}^{k}\right) \quad (5)$$

$$\mathrm{rsim}(\mathrm{rep}(X),\mathrm{rep}(Y))=\frac{<\mathrm{rep}(X),\mathrm{rep}(Y)>}{\|\mathrm{rep}(X)\|\cdot\|\mathrm{rep}(Y)\|}$$

$$\mathrm{fuse}(\{s_1,s_2,\cdots,s_k\})=\frac{\sum_{i=1}^{k}s_i\cdot\exp(\beta\cdot s_k)}{\sum_{i=1}^{k}\exp(\beta\cdot s_k)}$$

使用余弦相似度度量来量化两个特征之间的相似性，并使用 Softmax 权重来融合不同的分数。因为 IJB-A 数据集需要比较两个模板 X 和 Y，而不是两个图像。因此，再使用一步 Softmax 融合两个模板中图像的得分。

$$\mathrm{tsim}(\mathbf{X},\mathbf{Y})=\mathrm{fuse}\left(\{\mathrm{sim}(X,Y)|X\in\mathbf{X},Y\in\mathbf{Y}\}\right)$$

68. Template Adaptation for Face Verification and Identification

主要思想：

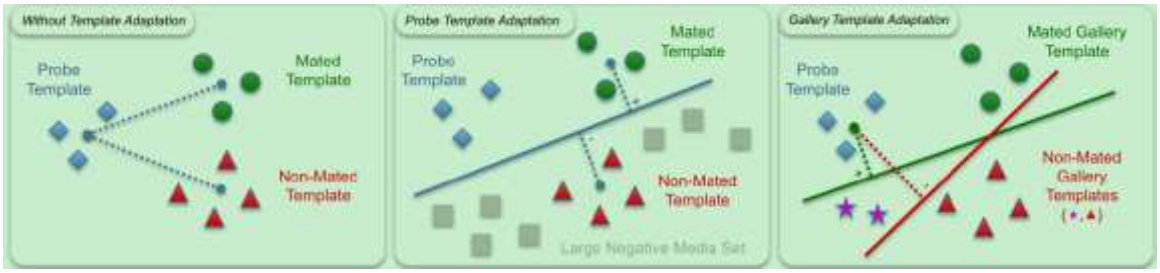
- 研究模板适配问题，结合了深度卷积网络特性和模板特定的线性 svms，这是一种将学习转移到模板中的媒体集合的形式。该策略可应用于现有网络，以提高性能。
- 性能在很大程度上取决于模板中可用媒体的数量。模板包含单个媒体时，其性能会降低 19%。
- 当 probe 或 gallery 模板丰富且至少有一个模板包含大于三个媒体时，性能很快就会饱和，并主宰着最先进的技术。
-

主要步骤：

模板自适应是一种转移学习的形式，它结合了在多个标记人脸的源域上训练的深卷积网络特征，以及利用模板中的媒体在目标域上训练的模板特定线性 svm。模板自适应可以进一步分解为用于人脸验证的 probe 自适应和用于人脸识别的 gallery 自适应。

probe 自适应:从模板到大的负特征集对正特征进行最大边缘分类的问题。蓝色 probe 模板与匹配绿色模板之间的相似性是绿色特征编码到决策面的边缘(虚线)。观察到这个边距是正的，而红色分类器的边距是负的，因此蓝/绿的相似性比所需的蓝/红大得多。

Gallery 自适应：最大边缘分类的问题，其中画廊模板的负面特征集是由其他 gallery 模板定义的。观察到，添加红色特征编码会导致红绿分类器的决策面发生移位，从而提高探针的边缘得分。



图像编码 z=f(X)是从图像 x 到具有维数 d 的编码 z 的映射。平均编码$\bar{z}=1/m\sum_x f(x)$是媒体中图像/帧编码的平均值，例如视频

中所有帧的编码。模板 X 是一组编码的媒体$X = \{f(x_1), f(x_2), \ldots, f(x_k)\}$。Gallery 是一组媒体$G = \{(X_1, y_1), (X_2, y_2), \ldots, (X_m, y_m)\}$，其中 y 是身份标签。

probe 自适应是对 probe 模板 P 和参考模板 Q 的相似函数 s(P,Q)的训练，使用 P 中媒体的单位归一化平均编码作为正特征，一个大的特征集作为负特征（大的负值集包含一个特征编码，用于许多主题标识，因此这个集合很可能与 probe 模板不相关联），训练 P 的线性 svm。同样，训练一个 Q 的线性的 svm，使用 Q 的单位归一化平均编码的媒体在作为正特征和一个大的特征集作为负特征。然后用 P 的 SVM 评估 q 得到 P(q)，用 Q 的 SVM 评估 p 得到 Q（p)。最终相似性评分是使用线性组合$s(P,Q) = \frac{1}{2}P(q) + \frac{1}{2}Q(p)$融合两个分类器边缘。

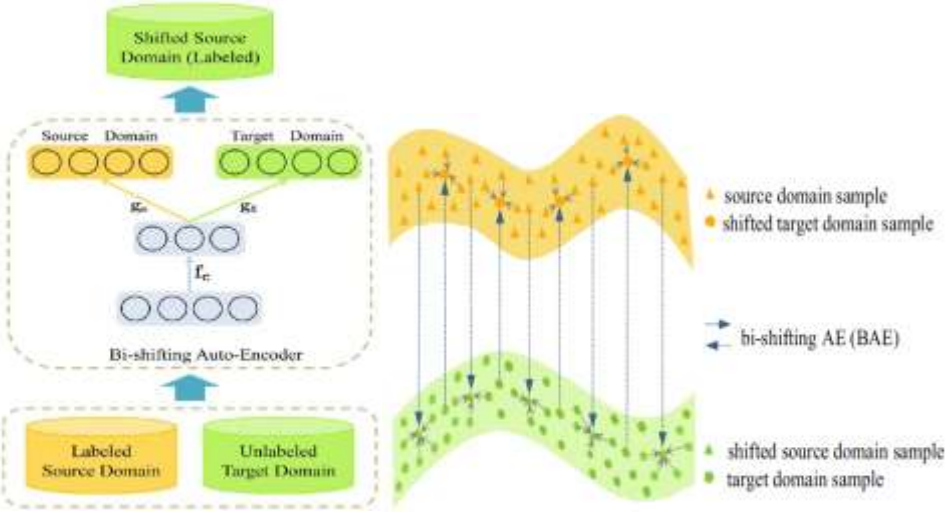
Gallery 自适应是对从 probe 模板 P 到图库 Q 的相似函数 s(P,Q)的训练。图库包含模板$G = \{X_1, X_2, \ldots, X_m\}$，而 Gallery 自适应训练所有对 s(p; xi)的线性 svm，并遵循 probe 自适应的方法。Gallery 自适应不同于 probe 自适应，因为模板 xi 的负值集都是所有 G 中其他模板的单元规范化媒体编码，但不包括 xi。

69. Bi-shifting Auto-Encoder for Unsupervised Domain Adaptation

主要思想：

- 编码器和解码器的非线性映射函数保证了将样本从一个域转移到另一个域的可行性
- 稀疏重构约束则保证移位域和理想域遵循相似分布

主要方法：



- Auto-Encoder

通过 BAE，源域样本可以转换为目标域样本，并通过稀疏和线性重建的目标域样本表示；同样的，目标域样本可以转化为源域，也可以由源域样本进行稀疏线性重建。

双移位自动编码器网络由一个编码器f_c和两个解码器，即g_s和g_t组成，它们可以分别将输入样本转换为源域和目标域。

$$\mathbf{z} \triangleq \mathbf{f}_c(\mathbf{x}) = \sigma(\mathbf{W}_c\mathbf{x} + \mathbf{b}_c) \tag{4}$$

$$\begin{aligned} \mathbf{g}_s(\mathbf{z}) &= \sigma(\mathbf{W}_s\mathbf{z} + \mathbf{b}_s), \\ \mathbf{g}_t(\mathbf{z}) &= \sigma(\mathbf{W}_t\mathbf{z} + \mathbf{b}_t), \end{aligned} \tag{5}$$

- sparely reconstructed

如果每个移动源域样本（绿色三角）可以由目标域（绿色圈）的几个局部邻居稀疏重构，它们往往遵循相似的局部结构，这意味着整个相似分布。同样，每个移动目标域样本（黄色圆圈）被限制为稀疏重构的源域邻居（黄色三角形），强制他们遵循类似的分布。

$$\mathbf{g}_t(\mathbf{f}_c(\mathbf{x}_i^s)) = \mathbf{X}_t\boldsymbol{\beta}_i^t, \quad s.t., |\boldsymbol{\beta}_i^t|_0 < \tau, \tag{6}$$

$g_t(f_c(x_i^s))$表示源域通过编码器f_c和解码器g_t重构的目标域样本，β_i^t是一个与邻域相关的非零值的稀疏向量。

则最终优化目标为：

$$\begin{aligned} \min_{\mathbf{f}_c,\mathbf{g}_s,\mathbf{g}_t,\mathbf{B}_s,\mathbf{B}_t} & ||\mathbf{X}_s - \mathbf{g}_s(\mathbf{f}_c(\mathbf{X}_s))||_2^2 + ||\mathbf{X}_t\mathbf{B}_t - \mathbf{g}_t(\mathbf{f}_c(\mathbf{X}_s))||_2^2 \\ & + ||\mathbf{X}_s\mathbf{B}_s - \mathbf{g}_s(\mathbf{f}_c(\mathbf{X}_t))||_2^2 + ||\mathbf{X}_t - \mathbf{g}_t(\mathbf{f}_c(\mathbf{X}_t))||_2^2 \\ & + \gamma\left(\sum_{i=1}^{n_s}|\boldsymbol{\beta}_i^s|_1 + \sum_{i=1}^{n_t}|\boldsymbol{\beta}_i^t|_1\right). \end{aligned} \tag{9}$$

其中，γ 是控制稀疏性的一个参数，即较大的 γ 导致稀疏重建所需的样本较少，而较小的 γ 导致更多的样本被选择用于稀疏重建。

因此，标记的源域样本（X_s, y_s）可以转移到目标域$(\mathbf{G}_t, \mathbf{y}_s), \mathbf{G}_t \triangleq \mathbf{g}_t(\mathbf{f}_c(\mathbf{X}_s))$。映射的源域样本$G_t$与目标域有相似的分布，所以任何监督的方法可以应用到学习在目标域分类的分类器。

70. Unsupervised Domain Adaptation for Face Recognition in Unlabeled Videos

主要思想：

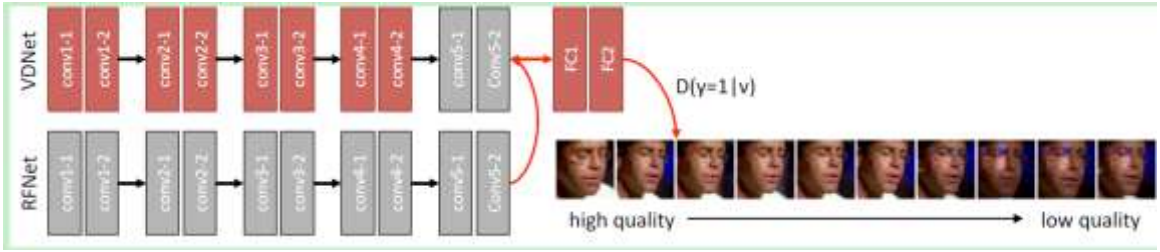
- 利用大规模无标记的视频数据来缩小不同领域之间的差距，将知识从大规模的标签上转移
- 自适应实现：（i）通过特征匹配从网络提取知识到视频自适应网络；（ii）通过合成数据增强进行特征恢复；（iii）通过域对抗鉴别器学习域不变特征。
- 提出了一种具有识别功能的特征融合方法，可以有效地将多个帧的特征集合起来，并根据它们对人脸识别的适应性进行有效的排序，剔除不想似的帧。

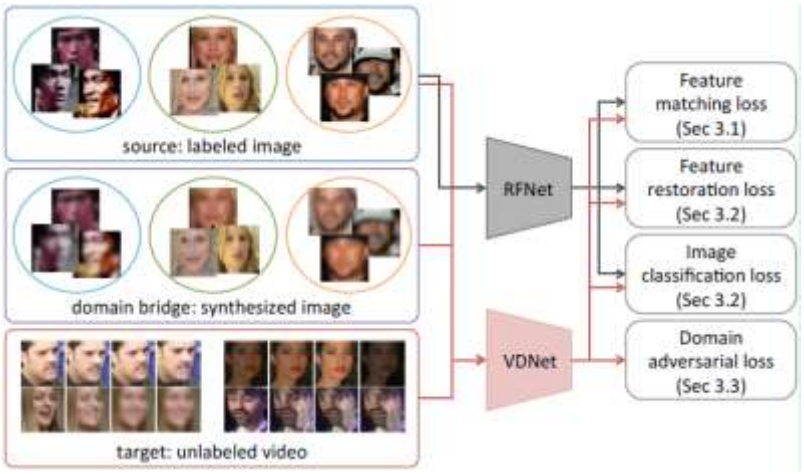
主要方法：

由于视频中的模糊、压缩、运动和其他伪影，静止图像和视频之间还是具有一定的差距。

为了利用有标签的网络人脸图像，我们通过一个人脸识别引擎来训练 VDNet，该引擎预先利用 web 人脸数据集预训练，，我们称它为参考网络(RFNet)。

红色和灰色的块分别表示可训练和固定的模块。VDNet 不仅与 RFNet 共享网络体系结构，而且使用相同的网络参数进行初始化。一旦经过训练，鉴别器 D 就可以在视频序列中对帧进行排序，通过指示帧是否与人脸识别引擎相匹配，并拒绝那些极其不适合人脸识别的帧。





- feature matching (FM) loss

VDNet 的特征提取函数为 $\phi(\cdot)$：$R^D \to R^K$，RFNet 的特征提取器为 $\varphi(\cdot)$：$R^D \to R^K$，静止图像数据集为 $\mathcal{L}$，视频数据集为 $\mathcal{V}$。

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{|\mathcal{I}|}\sum_{x\in\mathcal{I}} \|\phi(x) - \psi(x)\|_2^2 \qquad (1)$$

- feature restoration (FR) loss

训练 VDNet“还原”源域图像的特征表示（没有数据增强）

$$\mathcal{L}_{\mathrm{FR}} = \frac{1}{|\mathcal{I}|}\sum_{x\in\mathcal{I}} \mathbb{E}_{B(\cdot)}\left[\|\phi(B(x)) - \psi(x)\|_2^2\right] \qquad (2)$$

其中 B 是转换，在这项工作中，我们考虑以下三种类型的图像转换：

① 线性运动模糊：内核长度随机选择（5，15）和内核角选择在（10，30）

② 尺度变化：我们重新调整图像 6 1 的原始图像尺寸小

③ JPEG 压缩：质量参数是随机设置的（30，75）

- Image classification loss（N-pair loss）

$$\mathcal{L}_{\mathrm{IC}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\phi(B_i(x_i^+))^\top\psi(x_i))}{\sum_{n=1}^{N}\exp(\phi(B_i(x_i^+))^\top\psi(x_n))} \quad (3)$$

- Domain Adversarial Loss

① Two-way D

使用 two-way softmax 分类器 D 区分源域（y=1）和目标域的合成图像和视频（y=2）。原始图像来自图像域，但图像的合成图像和随机视频帧都被训练成属于同一域的

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = &-\mathbb{E}_{x\in\mathcal{I}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \\ &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=2|\phi(x))\big] \quad (6)\\ \mathcal{L}_{\mathrm{Adv}} = &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \quad (7)\end{aligned}$$

② Three-way D

使用 three-way softmax 分类器 D 区分源域（y=1）和合成图像（y=2）和视频（y=3）

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = &-\mathbb{E}_{x\in\mathcal{I}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \\ &-\mathbb{E}_{x\in B(\mathcal{I})}\big[\log\mathcal{D}(y=2|\phi(x))\big] \quad (8)\\ &-\mathbb{E}_{x\in\mathcal{V}}\big[\log\mathcal{D}(y=3|\phi(x))\big] \\ \mathcal{L}_{\mathrm{Adv}} = &-\mathbb{E}_{x\in B(\mathcal{I})\cup\mathcal{V}}\big[\log\mathcal{D}(y=1|\phi(x))\big] \quad (9)\end{aligned}$$

- 总体目标

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \alpha\mathcal{L}_{\mathrm{FR}} + \beta\mathcal{L}_{\mathrm{IC}} + \gamma\mathcal{L}_{\mathrm{Adv}} \quad (10)$$

- Discriminator-Guided Feature Fusion

鉴别器已经被训练来区分静止的图像和模糊的图像或视频帧，因此它的输出可能已经被用来作为一个高质量图像的分数。

有了鉴别器的得分，视频 V 的聚合特征向量被表示为特征向量的加权平均值，如下所列

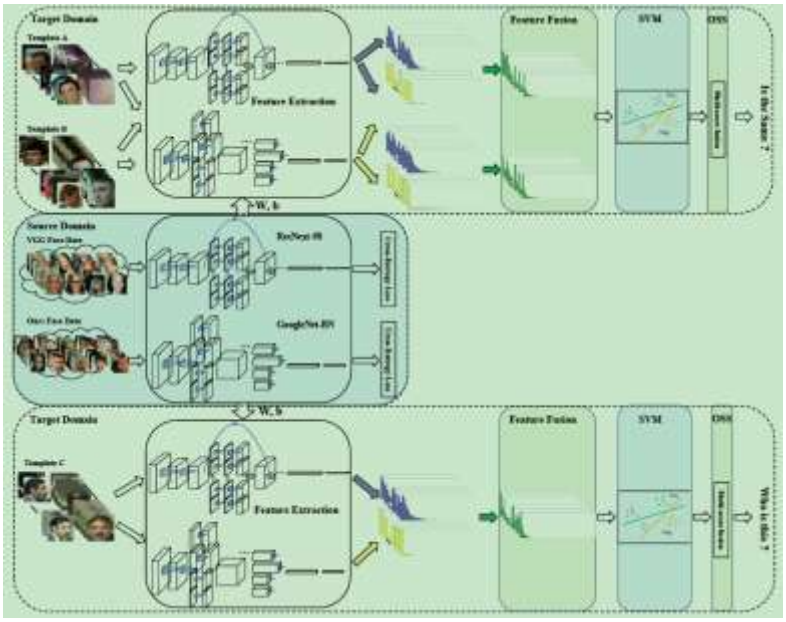
$$\phi_V = \frac{\sum_{v\in V}\mathcal{D}(y=1|\phi(v))\cdot\phi(v)}{\sum_{v\in V}\mathcal{D}(y=1|\phi(v))}. \quad (11)$$

71. A Good Practice Towards Top Performance of FaceRecognition: Transferred Deep Feature Fusion

主要思想：

- 在迁移学习的启发下，我们在源域用两个不同的大数据集训练了两个深卷积神经网络(Dcnn)。
- 通过研究两种不同特征点之间的互补性，在目标域进行特征提取。
- 设计了两阶段融合，一个用于特征，另一个用于相似度。后，利用深度特征融合。然后，
- 在目标域中训练 One-vs-rest template specific linear SVMs，将多个匹配分数对应的不同模板合并为最终结果

主要步骤：



- Deep feature learning in source domain

用两个不同的大人脸数据集对两个不同的 Dcnn 模型进行训练，在源域用中间分量表示。使用 VGG-Face dataset，利用残差网络作为第一个网络，googlenet 作为第二个网络。

- Template-based unconstrained face recognition

对于残差网络，倒数第二个全局平均池层作为特征提取层。它有 2048 输出大小。因此，特征维数为 2048。对于 GoogleNet，7*7 平均池层被视为特征提取层，特征维数为 1024。在第一阶段融合中，fr(Xi)和 fg(Xi)连接成 $f_F(\mathbf{x}_i) \in \mathbb{R}^{d_1}$，维数是 3072。

在特征融合之后，为了在目标域中训练出更多的判别模型，One-vs-rest template specific linear SVMs:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda_+ \sum_{i=1}^{N_+} \max\left[0, 1 - y_i \mathbf{w}^T f_F(\mathbf{x}_i)\right]^2 + \lambda_- \sum_{i=1}^{N_-} \max\left[0, 1 - y_j \mathbf{w}^T f_F(\mathbf{x}_j)\right]^2$$

针对视频帧的特点，我们计算了平均媒体编码量，将视频的每个帧聚合：$t_j^V = \frac{1}{N_j^V}\sum_{i=1}^{N_j^V} f_F(\mathbf{x}_i)$

因此，对于第 a 个模板，人脸深表示可以表示为：$T_a = \{t_1^I, ..., t_{N_a}^V\}$。

对于相似性计算，参考《Template Adaptation for Face Verification and Identification》。所有媒体编码都需要执行单元规范化。为了 vertification，阳性样本是 probe 模板，大规模的阴性样本包括整个训练集。对于 recognition，probe template specificSVMs 采用整个训练集作为大规模的负样本；而对于 gallery template specific SVMs，我们采用其他库模板和整个训练集作为大规模的负样本。$s(p,q) = \frac{1}{2}\mathcal{P}(q) + \frac{1}{2}\mathcal{Q}(p)$计算相似性。

在第二阶段融合中，所得到的多个匹配分数应该被组合成一个对模板对的单个匹配分数。

$$s(T_a, T_b) = \frac{\sum_{t_i \in T_a, t_j \in T_b} s(t_i, t_j) e^{\beta s(t_i, t_j)}}{\sum_{t_i \in T_a, t_j \in T_b} e^{\beta s(t_i, t_j)}}$$

3) 结构

72. Sparsifying Neural Network Connections for Face Recognition

主要思想：

- 一种压缩网络的方法，压缩的模型是 DeepID2+，压缩的核心：剪枝+再训练。
- 稀疏的 ConvNets 以迭代的方式学习，每次一个附加层稀疏化，整个模型用之前的迭代学习的初始权值再训练。
- 直接从零开始训练稀疏 ConvNet 无法找到很好的人脸识别解决方案，而使用先前学习到的密度更高的模型来正确初始化稀疏化模型对于继续学习有效的人脸识别特征至关重要。
- 提出了一种新的基于神经相关的权重选择准则，并在经验上验证了它在从每个迭代中从先前学习的模型中选择信息性连接方面

的有效性。

主要步骤：

整个算法的流程很简单：从网络的最后一层开始，根据一定规则对该层进行剪枝，然后 retrain 网络，循环上述过程。

剪枝的实现方法，就是为权重施加一个相同大小的 Mask，Mask 中只有激活的地方才是 1，其余全 0.

- 剪枝准则

剪枝的目标就是只保留重要的权重。

① 全连接层剪枝

首先，对于如全连接和局部连接这些没有权值共享的层，我们可以很简单的计算神经元之间的相关性：

假设 ai 是当前层的一个神经元，上一层有 K 个神经元,则此时 ai 与上一层之间应该有 K 个连接,即 K 个权重参数:bi1,bi2…biK。于是我们可以用下式计算 ai 与每一个 bik 的相关系数：

$$r_{ik} = \frac{E[a_i - \mu_{a_i}][b_{ik} - \mu_{b_{ik}}]}{\sigma_{a_i}\sigma_{b_{ik}}}, \quad (1)$$

其中 μ 和 σ 分别是在验证集上计算得到的均值与方差。

正相关和负相关同样重要，而且实验发现保留一些相关性较小的权重也会提高实验效果。

于是，作者首先将所有 K+个正相关的 rik 降序排列，然后均分为两部分，在前一部分随机采样 λSK+个，在后面一部分随机采样(1-λ)SK+个，其中 S 为事先确定的稀疏度，λ 文中设定为 0.75 。对负相关采取同样操作。

② 卷积层剪枝

卷积层剪枝稍微复杂一点，因为存在权值共享。

设 aim 是当前层第 i 个 feature map 中的第 m 神经元，该 feature map 中的共有 M 个神经元。显然，根据卷积规则，这 M 个神经元都只与一个卷积核有关，即 K 个权值有关 (K 为 filter size)。

最后，相关系数通过平均的方式计算：

$$r_{ik} \triangleq \sum_{m=1}^{M} \left| \frac{E[a_{im} - \mu_{a_{im}}][b_{mk} - \mu_{b_{mk}}]}{\sigma_{a_{im}}\sigma_{b_{mk}}} \right|, \quad (2)$$

input: network structure T; layers to be sparsified $L_1, L_2, \ldots, L_M$; degrees of sparsity $S_{L_1}, S_{L_2}, \ldots, S_{L_M}$

train baseline network N_0 with structure T
for m from 1 to M **do**
  calculate dropping matrix D_{L_m} of layer L_m according to the neural correlations in network N_{m-1} and the sparsity degree S_{L_m}
  initialize network N_m with structure T and weights of network N_{m-1}
  while not converge **do**
    update weights in layers $L_1, L_2, \ldots, L_m$ by dot-multiplying them with dropping matrices $D_{L_1}, D_{L_2}, \ldots, D_{L_m}$, respectively
    forward- and back-propagation one mini-batch of training samples in network N_m and update weights in network N_m
  end while
end for
output network N_M with sparsified connections specified by dropping matrices $D_{L_1}, D_{L_2}, \ldots, D_{L_M}$

73. A Lightened CNN for Deep Face Representation

主要思想：

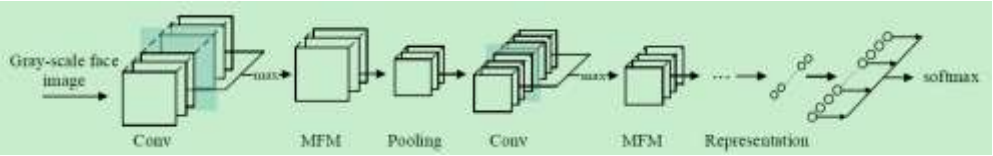
Related work 写的很好

- 将全连接层中的 MAXOUT 引入卷积层，形成了一个新的激活函数，称为最大特征映射(MFM)。与广泛使用的 ReLU 相比，MFM 可以同时捕获紧凑的表示和丰富信息。
- 两个模型：一个浅层的 CNN 模型由 4 个卷积层构成，完全包含 4M 个参数。另一个是通过减少卷积层的内核大小和在卷积层之间基于前一层的网络层构造 Network IN Network（NIN）网络。

主要步骤：

用 CNN 进行人脸验证分为三种。一种是使用人脸分类的任务训练 CNN 提取特征（softmax），然后用分类器判断是不是同一个人。第二种是直接优化验证损失（pairwise/triplet loss）。第三种是将人脸识别和验证任务同时进行。本文框架是属于第一种。

- 网络架构



本文网络结构如上图所示，和 DeepID 一样，在训练时使用人脸分类的任务进行训练，最后得到 256 维的人脸特征。网络最后一层是 Sofmax 层，实现分类的目的，fc1 的结果就是人脸的特征。

A 模型是由 4 层卷积层，MFM 激活函数，4 层最大池化层和 2 层全连接崩构成，灵感来自 alexnet。B 模型包含 5 个卷积层，4 个 NIN 层，MFM 激活函数，4 个最大池层和两个全连接层构成。

A				B			
Name	Filter Size /Stride	Output Size	#params	Name	Filter Size /Stride, Pad	Output Size	#params
input	-	144 × 144 × 1	-	input	-	144 × 144 × 1	-
crop	-	128 × 128 × 1	-	crop	-	128 × 128 × 1	-
conv1_1	9 × 9/1	120 × 120 × 48	3.8K	conv1_1	5 × 5/1, 2	128 × 128 × 48	1.2K
conv1_2	9 × 9/1	120 × 120 × 48	3.8K	conv1_2	5 × 5/1, 2	128 × 128 × 48	1.2K
mfm1	-	120 × 120 × 48	-	mfm1	-	128 × 128 × 48	-
pool1	2 × 2/2	60 × 60 × 48	-	pool1	2 × 2/2	64 × 64 × 48	-
				conv2_a	1 × 1/1	64 × 64 × 48	0.01K
conv2_1	5 × 5/1	56 × 56 × 96	2.4K	conv2_1	3 × 3/1, 1	64 × 64 × 96	0.8K
conv2_2	5 × 5/1	56 × 56 × 96	2.4K	conv2_2	3 × 3/1, 1	64 × 64 × 96	0.8K
mfm2	-	56 × 56 × 96	-	mfm2	-	64 × 64 × 96	-
pool2	2 × 2/2	28 × 28 × 96	-	pool2	2 × 2/2	32 × 32 × 96	-
				conv3_a	1 × 1/1	32 × 32 × 96	0.09K
conv3_1	5 × 5/1	24 × 24 × 128	3.2K	conv3_1	3 × 3/1, 1	32 × 32 × 192	1.7K
conv3_2	5 × 5/1	24 × 24 × 128	3.2K	conv3_2	3 × 3/1, 1	32 × 32 × 192	1.7K
mfm3	-	24 × 24 × 128	-	mfm3	-	32 × 32 × 192	-
pool3	2 × 2/2	12 × 12 × 128	-	pool3	2 × 2/2	16 × 16 × 192	-
				conv4_a	1 × 1/1	16 × 16 × 192	0.19K
conv4_1	4 × 4/1	9 × 9 × 192	3K	conv4_1	3 × 3/1, 1	16 × 16 × 128	1.1K
conv4_2	4 × 4/1	9 × 9 × 192	3K	conv4_2	3 × 3/1, 1	16 × 16 × 128	1.1K
mfm4	-	9 × 9 × 192	-	mfm4	-	16 × 16 × 128	-
pool4	2 × 2/2	5 × 5 × 192	-	pool4	2 × 2/2	8 × 8 × 128	-
				conv5_a	1 × 1/1	8 × 8 × 128	0.12K
				conv5_1	3 × 3/1, 1	8 × 8 × 128	1.1K
				conv5_2	3 × 3/1, 1	8 × 8 × 128	1.1K
				mfm5	-	8 × 8 × 128	-
				pool5	2 × 2/2	4 × 4 × 128	-
fc1	-	256	1.179K	fc1	-	256	524K
fc2	-	10,575	2.707K	fc2	-	10,575	2.707K
loss	-	10,575	-	loss	-	10,575	-
total			13.61K				5.244K

- MFM 激活函数

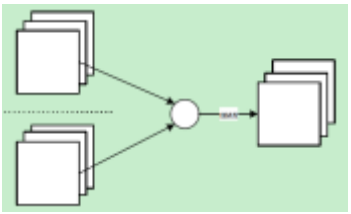
在输入的卷积层中，选择两层，取相同位置较大的值。

$$f_{ij}^{k} = \max_{1 \le k \le n} (C_{ij}^{k}, C_{ij}^{k+n}) \quad (1)$$

输入的卷积层为 2n 层，取第 k 层和第 k+n 层中较大的值作为输出，MFM 输出就变成了 n 层。激活函数的梯度为

$$\frac{\partial f}{\partial C^{k'}} = \begin{cases} 1, \text{ if } C_{ij}^{k} \ge C_{ij}^{k+n} \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

这样激活层有一半的梯度为 0，MFM 可以得到稀疏的梯度。MFM 激活函数相比于 ReLU 函数，ReLU 函数得到的特征是稀疏高维的，MFM 可以得到紧实（compact）的特征，还能实现特征选择和降维的效果。



74. A Light CNN for Deep Face Representation with Noisy Labels

主要思想：

- 提出了一种新的激活函数 Max-Feature-Map（MFM 不仅能区分开噪声数据和信息数据，而且在特征选择方面起着重要的作用）
- 提出了通过减少卷积层的内核大小和在卷积层之间基于前一层的网络层构造 Network IN Network（NIN）网络。
- 通过有引导的预训练可以处理更大规模有噪声的数据

主要步骤：

- MFM 2-1 激活函数

在输入的卷积层中，选择两层，取相同位置较大的值。

$$f_{ij}^{k}=\max_{1\le k\le n}(C_{ij}^{k},C_{ij}^{k+n}) \quad (1)$$

输入的卷积层为 2n 层，取第 k 层和第 k+n 层中较大的值作为输出，MFM 输出就变成了 n 层。激活函数的梯度为

$$\frac{\partial f}{\partial C^{k'}}=\begin{cases}1, \text{ if } C_{ij}^{k}\ge C_{ij}^{k+n}\\ 0, \text{ otherwise}\end{cases} \quad (2)$$

这样激活层有一半的梯度为 0，MFM 可以得到稀疏的梯度。MFM 激活函数相比于 ReLU 函数，ReLU 函数得到的特征是稀疏高维的，MFM 可以得到紧实（compact）的特征，还能实现特征选择和降维的效果。

- MFM 3-2 激活函数

输入三个特征图，删除最小的一个：

$$\begin{cases}\hat{x}_{ij}^{k_1}=\max(x_{ij}^{k},x_{ij}^{k+N},x_{ij}^{k+2N})\\ \hat{x}_{ij}^{k_2}=\text{median}(x_{ij}^{k},x_{ij}^{k+N},x_{ij}^{k+2N})\end{cases} \quad (4)$$

- Semantic Bootstrapping for Noisy Label

首先，我们在原始噪声标记的数据集上训练轻 CNN 模型。

其次，利用训练后的模型对含噪声训练样本的标签进行预测。然后我们设定一个阈值来决定是否接受或拒绝预测根据条件概率 $p(t_i|f(x))$。

最后，我们对 CNN 模式的重新标记的训练数据集。

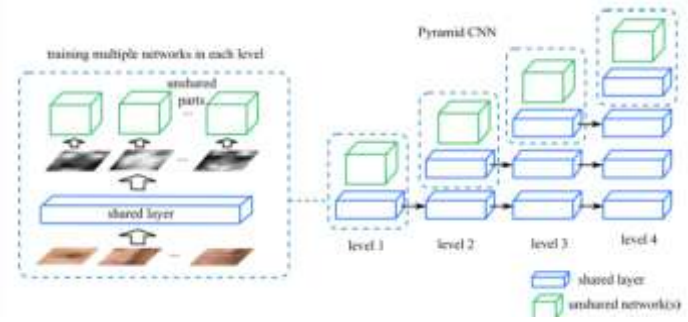
75. Learning Deep Face Representation

主要思想：

- 基于一种新的深层网络结构（称为 Pyramid CNN）。Pyramid CNN 采用贪婪滤波和降采样操作，使训练过程是非常快速和高效的计算效率。
- Pyramid CNN 的结构可以自然地将多尺度人脸表示中的特征共享结合起来，提高了结果表示的判别能力。

主要步骤：

基本的网络结构不是传统上的 CNN 结构，而是“Siamese”网络，它的特点是它接收两个图片作为输入，而不是一张图片作为输入。



网络的主要特点:它们是由多个金字塔组成，分为不同个 level 级别的特征，每一个 level 的网络由两部分组成，一部分是共享的层，它由它的前一个 level 的网络层组成；另一部分是一个非共享层，在每一层中训练只需要训练非共享层就可以，而共享层是由前一个 level 的的网络层共享过来，每一个神经网络中，非共享层用来进行数据的预处理，比如卷积下采样等。（由于网络的共享层，使得每一个 level 的训练速度不会随着网络层数的增多而急剧提高的训练时间，说白了也就是说每次只需要训练网络的最后一层就可以了，前面的层可以保持固定。）

采用金字塔的原因在于：1.加快网络的训练速度；2.可以提取多尺度人脸结构特征；

注意每一个 level 有多个神经网络，分别对应于每个输入图像的 patch。每个网络的训练目标函数都是下面的损失函数；

$$L=\sum_{I_1,I_2}\log(1+\exp(\delta(I_1,I_2)D(I_1,I_2))) \quad (1)$$

$$D(I_1,I_2)=\alpha\cdot d(f_\theta(I_1),f_\theta(I_2))-\beta \quad (2)$$

在$\theta(I_1,I_2)$指示两图像 I 1 和 I 2 属于同一个人。f_θ代表由神经网络进行计算，和 D 是一个函数度量两向量之间的距离。θ 代表网络中的权重。

76. One-to-many face recognition with bilinear CNNs

主要思想：

- 将 bilinear CNN (B-CNN)应用于具有挑战性的新人脸识别基准，IARPA Janus Benchmark A(IJB-A)。
- 从 alexnet 网络开始，在 ImageNet 上进行预训练，展示 B-CNN 模型的性能。然后，我们展示了使用中等大小的公共外部数据库 facescrb 进行微调的结果。进一步利用 IJB-A 训练集在进行微调。微调的 B-CNN 模型显示出比标准 CNN 有很大的改进。
- 一个标准的 cnn（VGG）可以转换为 b-cnn，而不需要任何额外的功能培训。

主要步骤：

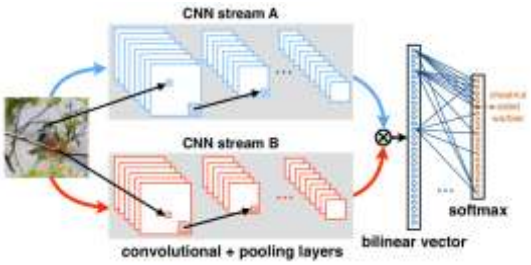
该结构由两个特征提取器产生，两个输出是图像每一个位置的外积(outer product)，然后进行 pool，得到最终的图像描述算子。一个 Bilinear 模型 B 由一个四元组组成：B=(fA,fB,P,C)。其中，fA，fB 代表特征提取函数，即图中的网络 A、B，P 是一个池化函数（Pooling Function）， C 则是分类函数。

特征提取函数 f（.）的作用可以看作一个函数映射，$\mathcal{L} \times \mathcal{I} \rightarrow \mathbb{R}^{K \times D}$将输入图像 I 与位置区域 L 映射为一个 K*D 维的特征。而两个特征提取函数的输出，可以通过一个双线性操作进行汇聚，得到最终的 Bilinear 特征：

$$\text{bilinear}(l, I, f_A, f_B) = f_A(l, I)^T f_B(l, I).$$

其中池化函数的作用是将所有位置的 Bilinear 特征汇聚成一个特征。Bilinear CNN 中所采用的池化函数是将所有位置的 Bilinear 特征累加起来：

$$\Phi(I) = \sum_{l \in \mathcal{L}} \text{bilinear}(l, I, f_A, f_B) = \sum_{l \in \mathcal{L}} f_A(l, I)^T f_B(l, I).$$



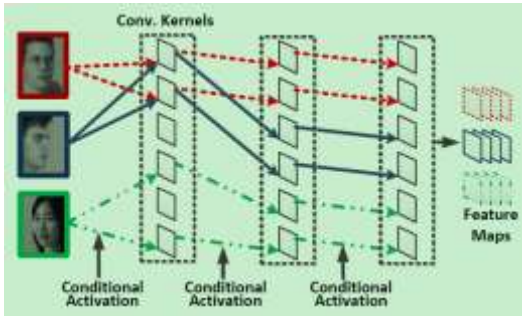
77. Conditional Convolutional Neural Network for Modality-aware Face Recognition

主要思想：

- 提出了一种条件卷积神经网络(C-CNN)来处理多模态人脸识别问题。
- 与传统 CNN 采用固定卷积核的方法不同，c-CNN 中的样本采用动态激活的核集进行处理。特别是，每一层中的卷积核只有在样本通过网络时才会被稀疏激活。对于给定的样本，卷积核在某一层中的激活取决于其当前的中间表示和下层的激活状态。
- 与大多数现有方法相比，拟议框架不依赖于事先对模式的任何了解。
- 为了充实通用框架，我们引入了一个特殊的 c-cnn 案例，结合决策树的条件路由，通过多模态多视点人脸识别和遮挡人脸验证两个问题对其进行了评价。

主要步骤：

每一张图像都会与特定于模式的路径一起传递，该路径由相应的彩色箭头指示。只有沿途的内核被激活并被用来提取特征。通过路径以粗到细的方式定义了分裂，例如红色虚线和蓝色实线的模式，可能在初始层共享某些内核。



- Conditional Convolutional Neural Network

对于 c-cnn，第 i 层内核的激活是由当前第 i 层输入表示$X_n^{(i)}$和低层通过路径$\{\theta_n^{(j)}, j=0,\ldots,i-1\}$共同决定的。

以 n 作为输入样本的指标，并建立了相应的前向函数：

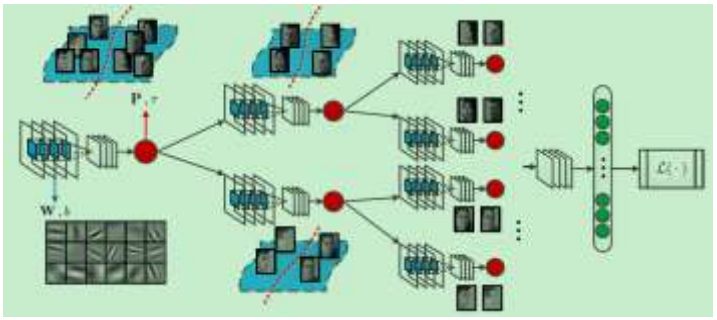
$$X_{n,k}^{(i+1)} = g_{n,k}^{(i)} \cdot \sigma(\widetilde{W}_k^{(i)} * X_n^{(i)} + b^{(i)}),$$

其中$X_{n,k}^{(i+1)}$是第 i+1 层 n 个样本的第 k 核映射，$g_{n,k}^{(i)}$表示 k-卷积核$\widetilde{W}_k^{(i)}$的激活指示。$g_{n,k}^{(i)}$服从 Bernoulli 分布，$g_{n,k}^{(i)} \sim B(1, p_{n,k}^{(i)})$

$$p_{n,k}^{(i)} = Pr(\theta_{n,k}^{(i)} | X_n^{(i)}, \theta_n^{(i-1)}, \ldots, \theta_n^{(0)}),$$

基于决策树的方法可以看作是一个简化的案例。该网络包括两个部分：模态感知投影树和卷积神经网络分支。

- ModalityawareProjection Tree



模态感知投影树(MPT)的目的是在样本空间中定义一个硬划分，使同一模式的样本落入同一叶节点。通过学习树的每个节点的分裂函数来探索模式。树的节点表示为 v(i，j)，其中 i 是树中层的索引，j 是第一层中叶节点的索引。在节点 v(i，j)中，样本的传递路径由拆分函数决定$\varphi : \mathcal{S} \to \{\mathcal{S}^{\mathcal{L}}, \mathcal{S}^{\mathcal{R}}\}$，如果我们表示该节点的整个输入集为 s，则两个子节点的子集分别表示为$S^{\mathcal{L}}$和$S^{\mathcal{R}}$。

$$x = \begin{cases} \mathcal{S}^{\mathcal{L}}, & \varphi(x) \geq 0 \\ \mathcal{S}^{\mathcal{R}}, & \varphi(x) < 0 \end{cases}$$

分裂函数由投影向量 p(i，j)和偏置 τ(i，j)定义，如下所示

$$\varphi(x) = x^T \cdot P^{(i,j)} + \tau^{(i,j)},$$

对每个节点施加无监督约束，使两个子簇的质心距离最大化。相应的节点级损失表示为：

$$\mathcal{L} = \frac{\frac{1}{N}\sum_{x \in \mathcal{S}} \varphi(x)^2}{(\frac{1}{N_{\mathcal{L}}}\sum_{x \in \mathcal{S}^{\mathcal{L}}} \varphi(x) - \frac{1}{N_{\mathcal{R}}}\sum_{x \in \mathcal{S}^{\mathcal{R}}} \varphi(x))^2},$$

- Convolutional Neural Branch

类似模式的样本应该比远距离模式的样本处理得更相似。

$$\begin{cases} X_n^{(i+1,2j)} & = \mathbb{1}(\varphi(\widetilde{X}_n^{(i,j)}) \geq 0) \cdot \widetilde{X}_n^{(i,j)} \\ X_n^{(i+1,2j+1)} & = \mathbb{1}(\varphi(\widetilde{X}_n^{(i,j)}) < 0) \cdot \widetilde{X}_n^{(i,j)} \end{cases}, \quad (7)$$

$$\widetilde{X_n^{(i,j)}} = \sigma(W^{(i,j)} * X_n^{(i,j)} + b^{(i,j)}),$$

- Joint Learning of MPT and CNN Branch

$$\mathcal{L} = \sum_n \mathcal{J}(x_n, y_n) + \beta \sum_i \sum_j \mathcal{L}^{(i,j)},$$

其中，第一项表示 n-class 分类问题的 Softmax 损失，第二项表示节点级损失，β 是一个标度因子。

78. MobileNets: Efficient Convolutional Neural Networks for Mobile VisionApplications

主要思想：

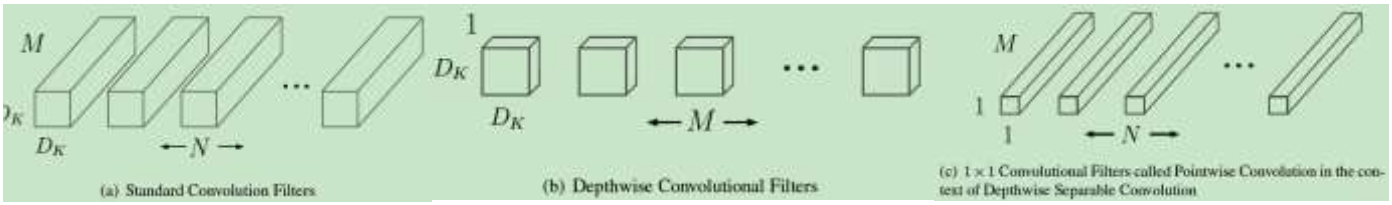
- MobileNets 基于精简架构，使用 depthwiseseparable convolutions 神经网络建立轻量级网络架构。我们
- 引入了两个简单的全局参数，有效地在延迟和精度之间权衡。这些超参数允许模型生成器根据问题的限制选择合适大小的应用程序模型。

主要步骤：

- Depthwise Separable Convolution

MobileNet 基于 depthwise separableconvolutions，它是分解卷积的一种形式，它将标准卷积分解为 depthwiseconvolution 深度卷积和 1*1 卷积称为点态卷积。对于 MobileNet，深度卷积将单个滤波器应用于每个输入信道。然后，点态卷积应用 1*1 卷积来组

合输出深度卷积。一种标准的卷积，将滤波和输入合并成一组新的输出这两个步骤在一步内完成。depthwise separableconvolutions 将其分为两层，一层用于滤波，另一层用于组合。

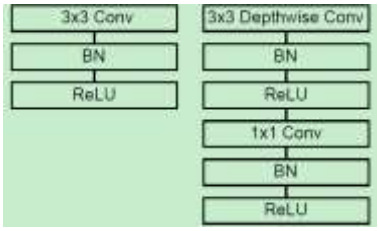


标准卷积层以$D_F \times D_F \times M$特征映射 F 作为输入，并生成$D_G \times D_G \times N$特征图 G，D_F是平方输入特征映射的空间宽度和高度，M 是输入通道数(输入深度)，D_G是平方输出特征映射的空间宽度和高度，N 是输出通道数(输出深度)。

标准卷积层用卷积核表示，$D_K \times D_K \times M \times N$其中 k 是核的空间维数，m 是输入通道数，n 是前面定义的输出通道数。

标准卷积的计算成本是$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$。

使用 depthwise separable convolutions，在每个输入通道(输入深度)应用一个单一滤波器。点态卷积，一个简单的 1*1 卷积，然后被用来创建深度层输出的线性组合。移动网对两个层都使用 BN 和 RELU。Depthwise separable convolutions 的计算成本是$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$。



- Width Multiplier: Thinner Models

  输入信道数 M 变为αM，输出信道数 N 变为αN。α＜1，典型取值为 1，0.75，0.5，0.23，计算成本变为：$D_K \cdot D_K \cdot \alpha M \cdot D_F \cdot D_F + \alpha M \cdot \alpha N \cdot D_F \cdot D_F$ (6)。
- Resolution Multiplier: Reduced Representation

  将其应用到输入图像中，每一层的内部表示随后都会减少。ρ＜1，通过取值将分辨率降为 224, 192, 160 or 128。计算成本变为：$D_K \cdot D_K \cdot \alpha M \cdot \rho D_F \cdot \rho D_F + \alpha M \cdot \alpha N \cdot \rho D_F \cdot \rho D_F$ (7)

4) 多网络结构（多 patch，多 pose，多任务，视频）

79. Multi-view Deep Network for Cross-view Classification

主要思想：

- 提出了一种 multi-view deep network (Mvdn)，它寻求多个视图之间共享的非线性的鉴别性和姿态不变表示。
- 由两个子网络组成：view-specificsub-network 一个是试图移除特定视图的变化，紧随的 common sub-network 试图实现所有视图共享的公共表示。
- 作为 MvDN 网络的目标函数，从所有视图的样本中计算 Fisher 损失以指导整个网络的学习。

主要步骤：

对于来自i^{th}视图的任何样本$\boldsymbol{x}_j^i$，其 MvDN $\mathbf{y}_j^i$的表示都是通过经过i^{th}视图特定的子网络并通过公共子网络生成的，其表示形式如下：

$$\mathbf{y}_j^i = \mathbf{g}_c\left(\mathbf{f}_i\left(\mathbf{x}_j^i\right)\right).$$

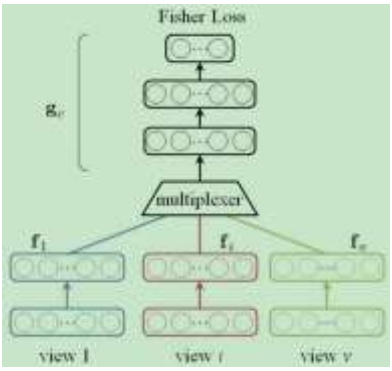
为了确保来自 MVDN 的表示$\mathbf{y}_j^i$具有判别性和对不同视角的鲁棒性，将来自所有视图的样本的 Rayleigh 商用作如下的目标函数：

$$[\mathbf{g}_c^*, \mathbf{f}_1^*, \cdots, \mathbf{f}_v^*] = \arg\min_{\mathbf{g}_c, \mathbf{f}_1, \cdots, \mathbf{f}_v} Tr\left(\frac{\mathbf{S}_W^y}{\mathbf{S}_B^y}\right)$$

其中 tr(•)表示矩阵的迹，S_W^y表示来自所有 v 视图的样本的类内散射，S_B^y表示来自所有 v 视图的样本的类间散射。

$$\mathbf{S}_W^y = \sum_{k=1}^{c}\sum_{i=1}^{v}\sum_{j=1}^{n_{ki}} \left(\mathbf{y}_{jk}^i - \boldsymbol{\mu}_k\right)\left(\mathbf{y}_{jk}^i - \boldsymbol{\mu}_k\right)^T,$$

$$\mathbf{S}_B^y = \sum_{k=1}^{c} n_k \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)^T,$$



80. Pose-Aware Face Recognition in the Wild（没细看）

主要思想：

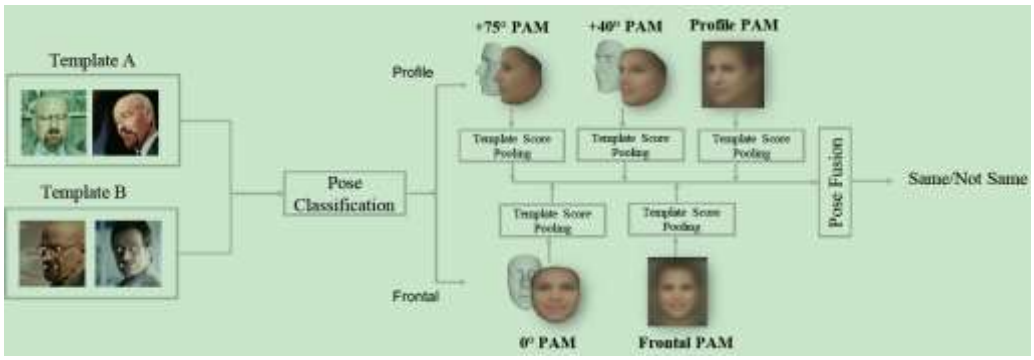
- 展示了我们如何不仅依靠单一的正面模型，而且依靠半轮廓和全轮廓模型在 wild 环境下进行人脸识别。

主要步骤：

针对人脸，我们采用了 multi-alignment 策略：

① 2D in-plane alignment：平面内主要解决尺度、平移、旋转问题。这里我们使用了 两个 CNN 模型，一个对应正脸，一个对应侧脸。

② 3D out-of-plane alignment:平面外旋转是通过在特定偏角下绘制图像以调整姿态和消除姿态变化。这里使用了三个 CNN 模型：+75°、+40°、 0° 。

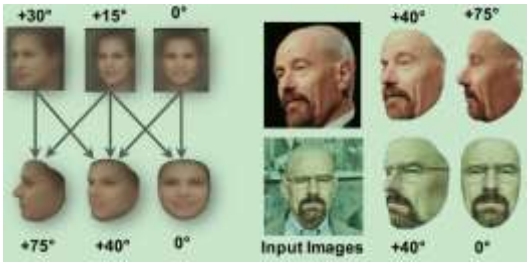


- 首先计算出训练数据的姿态分布，可以发现数据库中的人脸偏向于正脸
- PAMs for in-plane alignment

首先检测 landmarks 并使用$\mu_{profile}$ 对姿态进行分类。如果$|\psi| \leq \mu_{profile}$则为正脸，如果大于则为侧脸。分别训练两个 CNN 模型， PAM in-f and PAM in-p。

- PAMs for out-of-plane alignment

这里我们通过 rendering 将每个人脸映射到对应的姿态。一共分为三类。+75° PAM；+40° PAM；0° PAM 。这三类数据可以训练三个 CNN 模型。如果脸是正面的，我们可以渲染到正面和半轮廓的视图。如果图像远不是正面的，我们就避免正面化



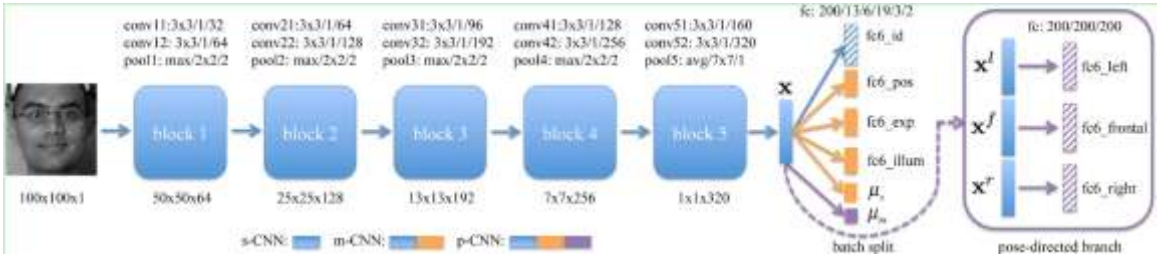
81. Multi-Task Convolutional Neural Network forPose-Invariant Face Recognition

主要思想：

- 首先，我们提出了一种多任务卷积神经网络(CNN)用于人脸识别，其中身份分类是主要任务，姿态、光照和表情估计是次要任务。
- 其次，我们提出了一种动态加权方案，将损失权重自动分配给次要任务。
- 第三，我们提出了一种面向姿态的多任务 cnn，它通过分组不同的姿势来学习特定姿势的身份特征，同时跨越所有的姿势。在测试阶段，我们提出了一种随机路由方案来融合一般的身份特征和特定于人脸的特征，这对估计误差更为鲁棒。
- 最后，我们提出了一种基于能量的权重分析方法，以探讨基于 cnn 的 MTL 是如何工作的。

主要步骤：

首先，我们提出了一种动态权重的多任务 CNN（m-cnn）用于人脸识别的（主要任务）和姿态不变学习（侧任务）。其次，我们提出了一种姿态定向多任务 CNN（p-cnn），将姿态分离成不同的组，每个组共同学习特定姿态下的身份特征。



- Multi-Task CNN

对 CASIA-NET 进行三个修改。首先，批归一化（BN）被应用于加速训练过程。第二，不用 contrastive loss。第三，根据不同的任务，改变全连接层的尺寸

给定一个训练集 D 包括 N 个图像和对应标签：$\mathbf{D} = \{\mathbf{I}_i, \mathbf{y}_i\}_{i=1}^N$，其中$I_i$是图像和$y_i$是一个向量组成的标签，$y_i^d$身份特征标签（主要任务）和三个辅助任务标签，包括姿势（y_i^p）、光照（y_i^l）、表情（y_i^e）。

$$softmax(\mathbf{y}^d)_n = p(\hat{y}^d = n|\mathbf{x}) = \frac{\exp(\mathbf{y}_n^d)}{\sum_j \exp(\mathbf{y}_j^d)}, \quad (3)$$

$$L(\mathbf{I}, y^d) = -\log(p(\hat{y}^d = y^d|\mathbf{I}, \boldsymbol{\Theta}, \mathbf{W}^d, \mathbf{b}^d)). \quad (5)$$

其他的任务的损失函数与主任务类似，最后总损失函数为：

$$\begin{aligned}\underset{\boldsymbol{\Theta},\mathbf{W}}{\operatorname{argmin}} \quad & \alpha_d \sum_{i=1}^N L(\mathbf{I}_i, y_i^d) + \alpha_p \sum_{i=1}^N L(\mathbf{I}_i, y_i^p) + \\ & \alpha_l \sum_{i=1}^N L(\mathbf{I}_i, y_i^l) + \alpha_e \sum_{i=1}^N L(\mathbf{I}_i, y_i^e),\end{aligned} \quad (6)$$

- Dynamic-Weighting Scheme

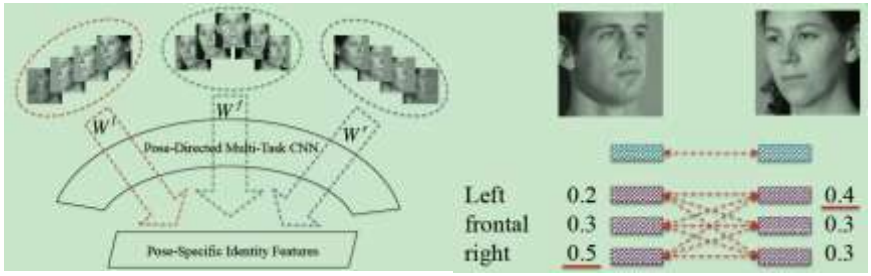
首先，我们将主任务的权重设为 1，即α_d=1。其次，通过验证集中的蛮力搜索，我们找出了所有辅助任务的总损失权重，即α_ss=$\alpha_p + \alpha_l + \alpha_e$，而不是为每个任务找到损失权重。然后我们的 m-cnn 学会分配三项任务的权重。

$$\boldsymbol{\mu}_s = softmax(\boldsymbol{\omega}_s^{\mathsf{T}}\mathbf{x} + \boldsymbol{\epsilon}_s),$$

其中，$\boldsymbol{\mu}_s = [\mu_p, \mu_l, \mu_e]^{\mathsf{T}}$是辅助任务的动态权重百分比，$\alpha_p + \alpha_l + \alpha_e$=1。

$$\begin{aligned}\underset{\boldsymbol{\Theta},\mathbf{W},\boldsymbol{\omega}_s}{\operatorname{argmin}} \quad & \sum_{i=1}^N L(\mathbf{I}_i, y_i^d) + \varphi_s\Big[\mu_p \sum_{i=1}^N L(\mathbf{I}_i, y_i^p) + \\ & \mu_l \sum_{i=1}^N L(\mathbf{I}_i, y_i^l) + \mu_e \sum_{i=1}^N L(\mathbf{I}_i, y_i^e)\Big] \\ s.t. \quad & \mu_p + \mu_l + \mu_e = 1,\end{aligned} \quad (8)$$

- Pose-Directed Multi-Task CNN



P-CNN 是建立在 m-CNN 之上，增加了姿态导向分支(PDB)。PDB 组通过批量分割操作，面对具有类似姿势的图像来学习特定姿势的身份特征。我们根据姿势标签将训练分为三组：左侧轮廓(G^l)、正面(G^f)和右侧轮廓(G^r)。p-cnn 旨在学习两种身份特点：W^d是提取

通用的身份特征；$W^l W^f W^r$的权重矩阵提取姿态特定身份特征，在小姿态范围内具有鲁棒性。这两项任务都被认为是我们的主要任务。与 m-cnn 中的动态加权方案类似，我们也使用动态权重来组合我们的主要任务。

$$\boldsymbol{\mu}_m = softmax(\boldsymbol{\omega}_m{}^{\intercal}\mathbf{x} + \boldsymbol{\epsilon}_m). \tag{10}$$

$$\begin{aligned}\underset{\boldsymbol{\Theta},\mathbf{W},\boldsymbol{\omega}}{\operatorname{argmin}} \quad & \varphi_m\Big[\mu_d \sum_{i=1}^{N} L(\mathbf{I}_i, y_i^d) + \mu_g \sum_{g=1}^{G}\sum_{i=1}^{N_g} L(\mathbf{I}_i, y_i^d)\Big] + \\ & \varphi_s\Big[\mu_p \sum_{i=1}^{N} L(\mathbf{I}_i, y_i^p) + \mu_l \sum_{i=1}^{N} L(\mathbf{I}_i, y_i^l) + \mu_e \sum_{i=1}^{N} L(\mathbf{I}_i, y_i^e)\Big] \\ & s.t. \quad \mu_d + \mu_g = 1, \quad \mu_p + \mu_l + \mu_e = 1,\end{aligned} \tag{11}$$

- Stochastic Routing

  一副人脸图像之间(I1 和 I2)的距离 C 通过计算通用身份特征之间的距离$(\mathbf{y}_1^d, \mathbf{y}_2^d)$和姿态特定身份特征的加权距离($\{\mathbf{y}_1^g\}, \{\mathbf{y}_2^g\}$)：

$$c = \frac{1}{2}h(\mathbf{y}_1^d, \mathbf{y}_2^d) + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} h(\mathbf{y}_1^i, \mathbf{y}_2^j) \cdot p_1^i \cdot p_2^j, \tag{12}$$

  其中 h()是度量两个特征向量之间距离的余弦距离度量。

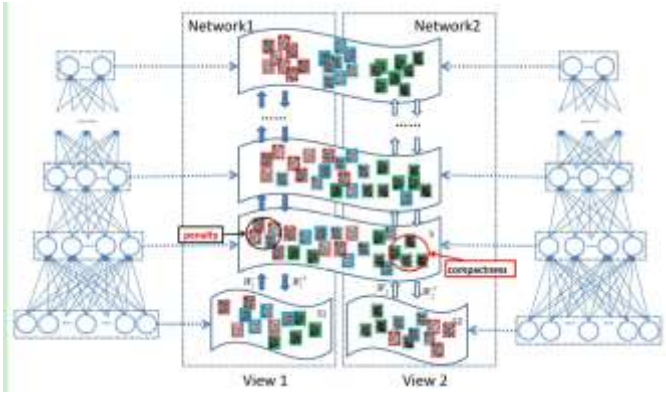
82. Deeply Coupled Auto-encoder Networks forCross-view Classification

主要思想：

- 提出了一种简单而有效的耦合神经网络，称为深度耦合自编码网络(DCAN)，该网络旨在建立两个相互对应的深层神经网络。
- 在 dcan 中，每一个深层结构都是通过叠加多分辨耦合自动编码器来实现的，这是一个由类内紧凑性和类间惩罚组成的最大 margin 准则训练的去噪自动编码器。

主要步骤：

dcan 试图学习两个非线性变换$f_x : X \longrightarrow H_x$ and $f_y : Y \longrightarrow H_y$，该方法可以将两个视图的样本分别投影到一个判别的公共空间中，其中每个视图的局部邻域关系和类分离性都得到很好的保持。这种类似于自动编码器的结构在保持局部一致性方面表现突出，而去噪形式增强了学习表示的鲁棒性。然而，可区分性并没有被考虑在内。因此，我们对去噪的自动编码器进行了改进，增加了一个由类内紧致性和类间惩罚组成的最大 margin 准则。



编解码器重构损失如下：

$$L(X,\Theta) = \sum_{x \in X^o} \mathbb{E}_{\hat{x} \sim P(\hat{x}|x)} \|\hat{x} - x\| \tag{3}$$

$$L(Y,\Theta) = \sum_{y \in Y^o} \mathbb{E}_{\hat{y} \sim P(\hat{y}|y)} \|\hat{y} - y\| \tag{4}$$

Large margin:

约束项$G_1(\cdot) - G_2(\cdot)$用于实现耦合，因为无论来自哪个视图，相同类的样本都被类似地对待。假设 s 是来自同一类的样本对的集合，d 是来自不同类的样本对的集合。请注意，来自两个视图的对应物自然地被添加到 s；d 中，因为被认为是类而不是视图。

$$G_1(H) = \frac{1}{2N_1}\sum_{I_i, I_j \in S} \|h_i - h_j\|^2,$$

$$G_2(H) = \frac{1}{2N_2}\sum_{\substack{I_i, I_j \in D \\ I_j \in KNN(I_i)}} \|h_i - h_j\|^2, \tag{9}$$

在 G2 中，Ij 属于具有不同类别标签的 Ii 的 k 个最近邻区。

三、 数据库

83. Labeled Faces in the Wild: Updates and New Reporting Procedures

LFW 数据集（Labeled Faces in the Wild）是目前用得最多的人脸图像数据库。该数据库共 13233 幅图像，其中 5749 个人，其中 1680 人有两幅及以上的图像，4069 人只有一幅图像。图像为 250*250 大小的 JPEG 格式。绝大多数为彩色图，少数为灰度图。该数据库采集的是自然条件下人脸图片，目的是提高自然条件下人脸识别的精度。该数据集有 6 中评价标准：

1) Unsupervised.
2) Image-restricted with no outside data.
3) Unrestricted with no outside data.
4) Image-restricted with label-free outside data.
5) Unrestricted with label-free outside data.
6) Unrestricted with labeled outside data.

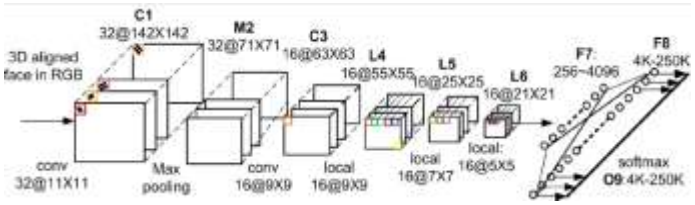
Protocol	Same/Different Labels for LFW training pairs allowed?	Identity info for LFW training images allowed?	Annotations for LFW training data allowed?	Non-LFW images allowed?	Non-LFW annotations allowed?	Same/Different labels for non-LFW pairs allowed?	Identity info for non-LFW images allowed?
Unsupervised	**no**	**no**	yes	yes	yes	**no**	**no**
Image-Restricted, No Outside Data	yes	**no**	**no**	**no**	**no**	**no**	**no**
Unrestricted, No Outside Data	yes	yes	**no**	**no**	**no**	**no**	**no**
Image-Restricted, Label-Free Outside Data	yes	**no**	yes	yes	yes	**no**	**no**
Unrestricted, Label-Free Outside Data	yes	yes	yes	yes	yes	**no**	**no**
Unrestricted With Labeled Outside Data	yes	yes	yes	yes	yes	yes	yes

84. Web-Scale Training for Face Identification

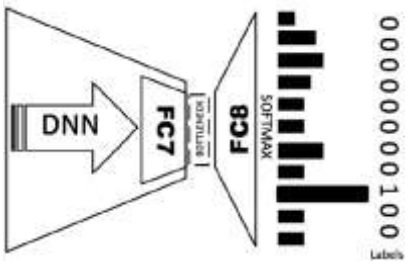
主要思想：

- 首先，我们发现并解释了网络瓶颈作为训练集特异性与通用性之间的一个重要规则的作用。网络的泛化能力可以通过控制的全连接层的维数来提高。
- 第二，我们已经确定了性能的饱和点，因为训练样本的数量超出了过去所探索的范围，并且提供了一种有效的方法，通过修改随机子采样训练集的共同实践来减轻这一点。

主要步骤：

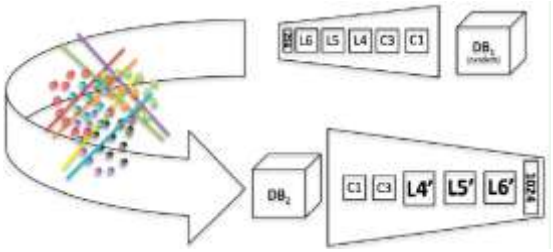


一个狭窄的瓶颈增加了从零开始训练时优化网络的难度。我们能够有效地训练出低到 1024，但不能更低。对于较小的维度，训练误差在早期停止下降。然而，我们注意到一个有用的特性：通过预先加载初始网络的权重，除了最后一层之外，我们能够有效地学习更小的嵌入。值得注意的是，在所有配置中，仅 256 个维度的瓶颈就使目标域的精度有了令人信服的提高。



- Semantic Bootstrapping

我们用一个分类器来表示每个类，也就是说，对于每个标识，我们学习了一个在二元分类设置中训练的超平面，其中正实例(表示)是相同的，而负数是其他恒等式的随机子集。我们随机选择 100 个身份，作为种子，在 1000 万个模型中。每一个种子，我们寻找最近的 1000 个模型，其中任何两个模型之间的相似性 H1；H2 被定义为相关的超平面之间的角度的余弦 $S(h_1,h_2) =< h_1,h_2 > /(\|h_1\|\|h_2\|)$ 所有检索到的图片身份结合构成新的数据集 DB2，总共包含 55000 个标识。注意， DB2 由简单和硬组成。种子之间的分离和以前一样容易，但是在每一个种子的附近，通过构造更加坚硬。



85. The MegaFace Benchmark--1 Million Faces for Recognition at Scale

motivation :

提出了 Megaface 数据集，包括了一百万张图片，其中包括了 690000 个人。然后应用了现有的人脸检测模型，查看现有模型在这个数据集上的结果。其中，Megaface 作为 gallery ,FaceScrub 和 FG-NET 作为 probe set.

LFW 包含十个，而 Megaface 包含一百万个 distrator(即出现在 gallery 上)。对比这些算法在这两个数据集上的表现，作者发现了以下几点：

1. 在 LFW（10 个 distrator）上识别率为 95%以上的算法，只达到了 35~75%的识别率（一百万个 distrator）
2. 训练集尺寸大效果大多较好
3. 年龄差距越大，测试效果越差
4. 姿态变化越大，测试效果越差

Megaface 数据集里面的图片还包括以下标签：①代表性的图片和方框 ②flicker 里的标签 ③GPS 地址 ④摄像机类型 ⑤3D 位置 ⑥每张图里脸的数量 ⑦脸部的解析度

Megadace challenge 包括以下两个挑战：①identification: 给定一个测试图，以及一个 gallery，里面包含相同的人的至少一张图，算法按照相似度排列 gallery 中与 probe 图相似的图片。②verification:判别给定的图片对是否是同一个人，图片对来自 probe dataset 和 Megaface distractor .

Dataset / Stats	MegaFace (this paper)	CASIA-WebFace	LFW	PIPA	FaceScrub	YouTube Faces	Parkhi et al.	CelebFaces	DeepFace (Facebook)	NTechLab	FaceNet (Google)	WebFaces Wang et al.	IJB-A IAPRA
#photos	1,027,060	494,414	13K	60K	100K	3425 videos	2.6M	202K	4.4M	18.4M	>500M	80M	25,813
#subjects	**690,572**	10,575	5K	2K	500	1595	2.6K	10K	4K	200K	>10M	N/A	500
Source of photos	**Flickr**	Celebrity search	Yahoo News	**Flickr**	Celebrity search	Celebrities on YouTube	Celebrity search	Celebrity search	Internal	Internal	Internal	Web crawling	Internal
Public/private dataset	**Public**	**Public**	**Public**	**Public**	**Public**	**Public**	Private	Private	Private	Private	Private	Private	**Public**

Group/algorithm / Training data used	LBP	JointBayes	3DiVi	BareBonesFR (UMD)	FaceAll Beijing	Ntech Lab small model	Ntech Lab large model	FaceNet (Google)
#photos	0	494,414	240,000	365,495	838,776	494,414	18,435,445	>500M
#unique people	0	10,575	5,000	5,772	17,452	10,575	200K	>10M
Public/private dataset	N/A	Public (CASIA)	Private	Private	Private	Private	Private	Private

Algorithm / Probe set	Google FaceNet v8	NTechLAB FaceNLarge	Faceall Beijing 1600Norm	Faceall Beijing 1600	NTechLAB FaceNSmall	Barebones cnn	3DiVi tdvm6
FaceScrub	70.49	**73.30**	64.80	63.97	58.2	59.36	33.70
FGNET	**74.59**	52.71	25.02	26.15	29.16	12.31	15.77

86. Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPAJanus Benchmark A

IJB-一个数据集的主要特征是：

(1) 全姿态变化，

	(a)	(b)	(c)
Types of Images:	Frontal, Cooperative subject, Controlled environment	Near frontal, uncooperative, minimal environment variations (e.g., LFW)	Full variation in pose, illumination, environment
Automated detection ability:	Human performance	Near human performance	Cannot detect consistently
Automated recognition ability:	Human performance	Near human performance	Cannot recognize

(2)图像和视频的混合，

(3)身份类别更广泛的地理变化，

Continent	# of subjects
Asia	89
Oceania	7
North America	135
South America	50

Continent	# of subjects
Europe	149
Middle East	29
Africa	41

(4)支持开放集识别(1：n搜索)和验证(1：1比较)的协议。

Protocol	Applications	Accuracy Metrics
Compare	1:1 match; Access control; Re-identification	TAR @ FAR of 0.1, 0.01, and 0.001; ROC plot (TAR vs. FAR)
Search	De-duplication; Watch list; Forensic	FNIR @ FPIR of 0.1 and 0.01; Rank 1 and 5 accuracy; CMC plot; DET plot (FNIR vs. FPIR)

(5)允许对 gallery 建模

(6)地面真实眼睛和鼻子位置

(7)为每个图像实例提供主题性别和肤色、遮挡(眼睛、嘴/鼻和额头)、面部听觉和粗糙的姿态信息。

87. MS-Celeb-1M: A Dataset and Benchmark forLarge-Scale Face Recognition

Task :识别 1M 个明星 from their face images。提出要建立**知识库**。

因为，首先，知识库上的每个人实体是 unique，并且清楚地定义，而没有歧异，使得可以定义这样的大规模面部识别任务。第二，每个实体自然具有多个属性（例如性别，出生日期，职业），为数据收集，清洗和多任务学习提供丰富且有价值的信息。

- Training dataset

从 1M 个名人中,根据他们的受欢迎程度,选择 100K 个。然后,利用搜索引擎,给 100K 个人,每人搜大概 100 张图片。共 100K*100=10M 个图片。

- 测量集

建立了一个测量集，其中包括一组经过仔细标记的图像，并将另一组随机选取的人脸图像作为干扰物。在我们的测量集中，大约有 25%的名人来自我们的百万名人榜的最后 90%位。当建立测量集之后，为每个名人提供两张图片。

随机集：该子集中的图像是随机从标记图像中选择的。每位名人一张照片。这一套揭示了有多少名人真正被测试的模型所覆盖。

困难集：该子集中的图像(来自标记图像)是训练数据集中任何图像中最不同的图像。每位名人一张照片。该集合是评价模型的泛化能力。

- Evaluation Protocol

对于 i 图像，让 g(xi)表示真值标号。对于任何要被测试的模型，我们假设模型输出$\{\hat{g}(x_i), c(x_i)\}$图像的预测实体键，并给出相应的预测可信度。我们允许模型执行拒绝。也就是说，如果 c(xi)<t,当 t 是一个预设的阈值时，图像 xi 的识别结果将被忽略

Precision：

$$P(t)=\frac{|\{x_i|\hat{g}(x_i)=g(x_i)\wedge c(x_i)\geq t, i=1,2,...,m\}|}{|\{x_i|c(x_i)\geq t, i=1,2,...,m\}|}$$

Coverage：

$$C(t)=\frac{|\{x_i|c(x_i)\geq t, i=1,2,...,m\}|}{m}$$

88. A Cross Benchmark Assessment of A Deep Convolutional Neural Network for Face Recognition

报告了视觉几何组(VGG)-人脸算法的性能在八项 NIST benchmark。VGG-人脸算法是一种在 LFW 基准上的高性能算法.。这八个 benchmark 涵盖了一系列场景，从在演播室环境中获取的面部图像到用数码点和摄影相机拍摄的面部图像。我们对第一时间的分析将在一系列条件下评估基于 dcn 的算法的精度，将结果与每个基准的既定精度进行比较，并将 DCNN 算法的精度与人类的精度进行比较。在分析的基础上，我们建议在评估人脸识别算法的准确性方面进行改进。

对于其中的七个 benchmark，这些图像是用数字单镜头相机获取的。这些照片是在工作室的照明和走廊和室外的环境照明中拍摄的。人类会认为图像质量很高。一个基准中的图像是用数码点和摄影相机获取的。8 项 benchmark 的总体人口组成为 59%名男性和 41%名女性；71%名白种人和 10%名东亚人；92%名 18 至 29 岁。

Good, Bad, and Ugly (GBU)：都是在 ambient lighting conditions 条件下获得的，包括室外或室内中庭和走廊。

EFCT：为了测量专业的面部检查人员的感知能力，建立了 EFCT。EFCT 由 GBU 的 bad 和 ugly 的图像组成。

FRGC 和 FRVT2006：一个图像在工作室环境下，一个图像是 ambient lighting conditions

PaSC：包含静态图像和视频。这些图像和视频是用数码点和摄影相机拍摄的，特别是手机里的手持相机

extremely-difficult：选择 50 个相同的身份对和 50 种不同的身份对，所有相同的身份对相似性得分低于所有不同的身份对。较高的相似性分数意味着更大的可能性。因此，FRVT 2006 融合算法的性能是 100%不正确的，而这些被称为 extremely-difficult 人脸对

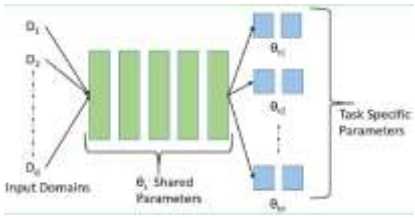
四、 多任务或全流程

89. An All-In-One Convolutional Neural Network for Face Analysis

主要思想：

- 提出了一种基于单一深卷积神经网络(CNN)的人脸检测、人脸对准、姿态估计、性别识别、微笑检测、年龄估计和人脸识别的多用途算法。

主要步骤：



将具有共享参数 s 和对于给定的任务 ti，其参数包括共享参数θ_s和任务特定参数θ_{t_i}的代价函数设为$J_i(\theta_s,\theta_{t_i};D)$，其中 d 是输入数据。对于各任务独立学习：

$$(\theta_s^*,\theta_{t_i}^*)=\arg\min_{(\theta_s,\theta_{t_i})} J_i(\theta_s,\theta_{t_i};D)$$

对于 MTL，可以通过使每个任务的损失函数的加权和最小化来获得任务 TI 的最优参数：

$$(\theta_s^*,\theta_{t_i}^*)=\arg\min_{(\theta_s,\theta_{t_i})} \alpha_i J_i(\theta_s,\theta_{t_i};D)+\sum_{j\neq i}^{n}\alpha_j J_j(\theta_s,\theta_{t_j};D)$$

由于其他任务只有助于学习共享参数，因此可以将它们解释为与给定任务 ti 有关的正则化 Ri。

$$(\theta_s^*,\theta_{t_i}^*)=\arg\min_{(\theta_s,\theta_{t_i})} J_i(\theta_s,\theta_{t_i};D)+\lambda R_i(\theta_s;D) \quad (3)$$

- Network Architecture

网络由七个卷积层和三个完全连接的层组成。我们利用它作为训练人脸识别任务的骨干网络，并将其七个卷积层的参数与其他与人脸相关的任务共享。采用参数整流器线性单元(PReLU)作为激活函数。

将这些任务分为两大类：1)独立于身份的任务，包括人脸检测、关键点定位和可见度、姿态估计和微笑预测；2)身份相关的任务，包括年龄估计、性别预测和人脸识别。

我们融合了第一三五层来训练独立于身份的任务，因为它们更多地依赖于来自网络底层的局部信息。我们分别在这些层上附加了两个卷积层和一个池层，以获得一个一致的特征映射大小为 6*6。增加了一个降维层，将特征映射的数目减少到 256。它后面是一个完全连接的维度 2048 层，它构成了独立于身份的任务的通用表示形式。此时，特定的任务被划分为每个维度 512 的完全连接层，然后分别是输出层。

年龄估计和性别分类的身份相关任务在执行最大池操作后从主干网络的第六层卷积层中分离出来，得到的全局特征输入到三层全连接网络。我们把第七的卷积层共享适应具体的人脸识别任务。

- Face Detection, Key-points Localization and Pose Estimation

Face Detection 使用 Softmax 损失函数。Key-points Localization and Pose Estimation 作为回归任务用欧式距离。

- Gender Recognition

$$L_G = -(1-g)\cdot\log(1-p_g) - g\cdot\log(p_g),$$

- Smile Detection

$$L_S = -(1-s)\cdot\log(1-p_s) - s\cdot\log(p_s),$$

- Age Estimation:

当年龄的标准偏差给出时，高斯损失比表观年龄估计要好得多。然而，当预测年龄远离真实年龄时，高斯损失的梯度接近零，从而减慢了训练过程。因此，我们使用这两个损失函数的线性组合。

$$L_A = (1-\lambda)\frac{1}{2}(y-a)^2 + \lambda\left(1 - exp(-\frac{(y-a)^2}{2\sigma^2})\right)$$

- Face Recognition

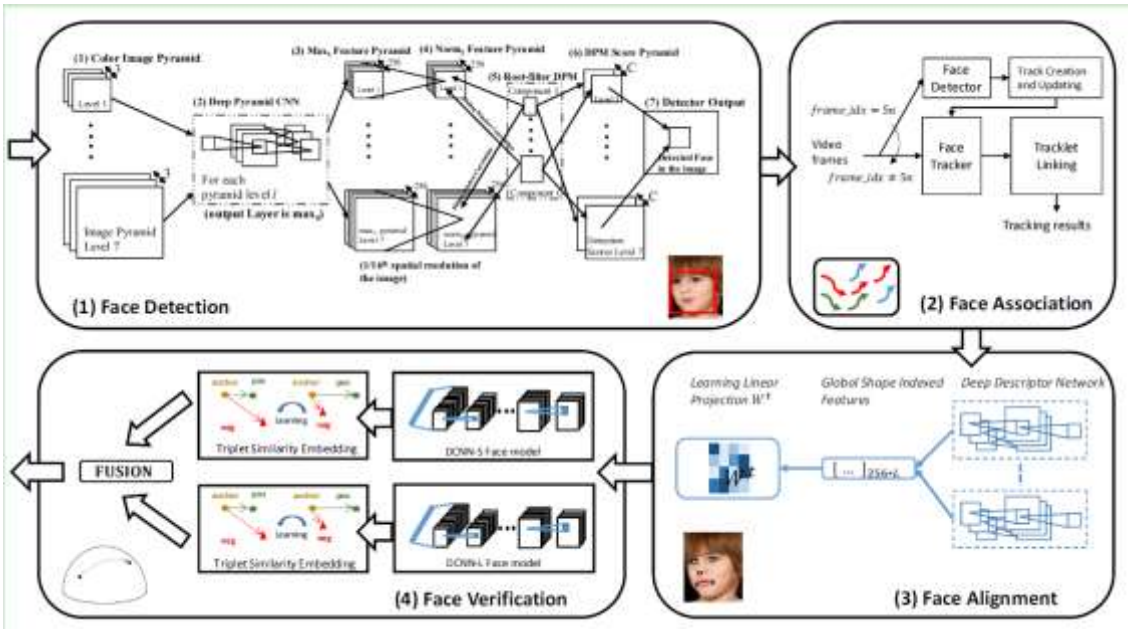
$$L_R = \sum_{c=0}^{10547} -y_c\cdot\log(p_c),$$

90. An End-to-End System for Unconstrained Face Verificationwith Deep Convolutional Neural Networks

主要思想：

- 提出了一个端到端的自动化人脸验证系统。基于离散的 dcnn 模型构建了系统的各个组成部分。
- 一个典型的端到端人脸识别系统由以下几个部分组成：(1)人脸检测，(2)人脸关联定位，(3)人脸标记检测，(4)验证被试身份的人脸验证。

主要步骤：



- Face Detection

图像/视频帧中的所有人脸都是使用一种基于 dcnn 的人脸检测器来检测的，该模型称为 DeepPyramid Deformable Parts Model for Face Detection(DP2MFD)。

- Face Association

使用 Kanade-Lucas-Tomasi (KLT) feature tracker 进行人脸的跟踪，从而将视频中同一个人脸进行关联。

- Facial Landmark Detection

将特征点检测看做一个回归问题。使用 CNN 特征

- Face Representation

我们训练了两个深卷积网络。一种是使用紧脸部包围框(DCNN$_S$)，另一种是使用包含更多上下文信息的大边界框(DCNN$_L$)。

Name	Type	Filter Size/Stride	#Params
conv11	convolution	3×3 / 1	0.84K
conv12	convolution	3×3 / 1	18K
pool1	max pooling	2×2 / 2	
conv21	convolution	3×3 / 1	36K
conv22	convolution	3×3 / 1	72K
pool2	max pooling	2×2 / 2	
conv31	convolution	3×3 / 1	108K
conv32	convolution	3×3 / 1	162K
pool3	max pooling	2×2 / 2	
conv41	convolution	3×3 / 1	216K
conv42	convolution	3×3 / 1	288K
pool4	max pooling	2×2 / 2	
conv51	convolution	3×3 / 1	360K
conv52	convolution	3×3 / 1	450K
pool5	avg pooling	7×7 / 1	
dropout	dropout (40%)		
fc6	fully connected	10548	3296K
loss	softmax	10548	
total			5M

Table 1 The architectures of DCNN$_S$.

Name	Type	Filter Size/Stride	#Params
conv1	convolution	11×11 / 4	35K
pool1	max pooling	3×3 / 2	
conv2	convolution	5×5 / 2	614K
pool2	max pooling	3×3 / 2	
conv3	convolution	3×3 / 2	885K
conv4	convolution	3×3 / 2	1.3M
conv5	convolution	3×3 / 1	885K
conv6	convolution	3×3 / 1	590K
pool6	max pooling	3×3 / 2	
fc6	fully connected	1024	9.4M
dropout	dropout (50%)		
fc7	fully connected	512	524K
dropout	dropout (50%)		
fc8	fully connected	10548	5.5M
loss	softmax	10548	
total			19.8M

Table 2 The architecture of DCNN$_L$.

- Triplet Similarity Embedding

采用 Triplet Similarity Embedding (TSE)或 Triplet Distance Embedding (TDE)

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{a,p,n\in\mathbb{T}} max(0, \alpha + a^T\mathbf{W}^T\mathbf{W}n - a^T\mathbf{W}^T\mathbf{W}p)$$

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{a,p,n\in\mathbb{T}} max\{0, \alpha + (a-p)^T\mathbf{W}^T\mathbf{W}(a-p) - (a-n)^T\mathbf{W}^T\mathbf{W}(a-n)\}$$

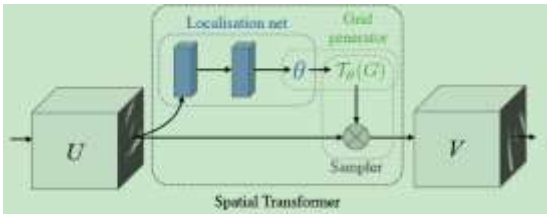
91. Spatial Transformer Networks

主要思想：

- 介绍了一种新的可学习模块——Spatial Transformer，它明确地允许在网络内对数据进行空间操纵
- 为执行输入特征地图的扭曲，参数化采样网格的每个输出像素是通过采样内核在一个特定的位置输入特征图中心计算的。

主要步骤：

Spatial Transformer 分为三个部分。localisation network 获取输入特征图，通过一些隐藏层输出应该应用到特征映射的空间转换的参数，这就给出了对输入的转换条件。然后，预测的变换参数被用来创建采样网格，这是一组输入采样映射，以产生转换输出。这是由网格生成器完成的。最后,特征图和采样网格作为采样器的输入，产生输出图。



- Localisation Network

localisation network 采用输入特征图$U \in \mathbb{R}^{H \times W \times C}$，宽度为 w、高度 h 和 c 通道并输出θ，θ是$\mathcal{T}_\theta$变换参数作用在特征图上的。$\theta = f_{\text{loc}}(U)$。θ的尺寸取决于参数化的转换类型。localisation network 函数 Floc()可以采取任何形式，例如完全连接的网络或卷积网络，但是应该包括一个最终的回归层来生成转换参数。

- Parameterised Sampling Grid

为了执行输入特征映射的扭曲，每个输出像素通过输入特征映射中的特定位置为中心的采样内核来计算。

a 2D affine transformation:（定义的转换允许对输入特征映射应用裁剪、平移、旋转、缩放和倾斜）

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

(x_i^t, y_i^t)是输出特征映射中规则网格的目标坐标。(x_i^s, y_i^s)是定义示例点的输入图中的源坐标。$A_\theta$affinetransformation matrix。

- Differentiable Image Sampling

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \quad \forall i \in [1 \dots H'W'] \quad \forall c \in [1 \dots C] \qquad (3)$$

$\Phi_x\Phi_y$是泛型取样内核k()的参数，它定义了图像插值(例如双线性)。U_{mn}^c是c通道（m，n）点位置的输入。V_i^c是c通道（x_i^t, y_i^t）点的输出。注意，采样对于输入的每个通道都是相同的，因此每个通道都以相同的方式进行转换（这保持了通道之间的空间一致性）。

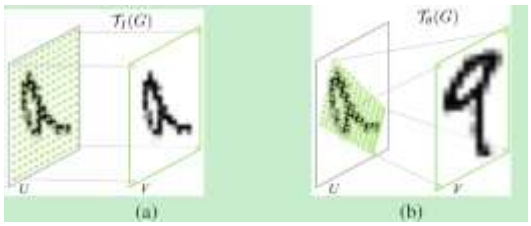
采用integer sampling kernel:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \delta(\lfloor x_i^s + 0.5 \rfloor - m) \delta(\lfloor y_i^s + 0.5 \rfloor - n)$$

这个采样内核等同于将（x_i^s, y_i^s）最近像素处的值复制到到输出位置（x_i^t, y_i^t）。

采用bilinear sampling kernel:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$



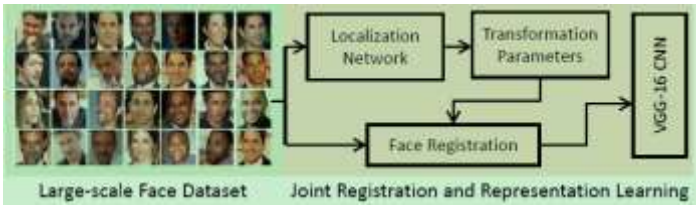
92. Joint Registration and Representation Learning for Unconstrained FaceIdentification

主要思想：

- 针对极端姿态的挑战，本文提出了一种基于 CNN 的数据驱动方法，该方法学习同时注册和表示人脸
- 注册模块通常包括用最少的背景裁剪最相关的面部区域，并在裁剪区域应用变形操作，将其转换为规范的正面视图。
- 学习表示模块以获得面部特征编码。
- 与现有的方法将所有模板媒体信息在特征层进行合成不同，本文建议保持模板介质完整。相反，我们用经过训练的一对其他的 SVM 模型来表示 gallery 模板，然后在决策层使用 Bayesian Classifier Combination (BCC) model，该策略能最佳地融合模板中所有媒体的决策。

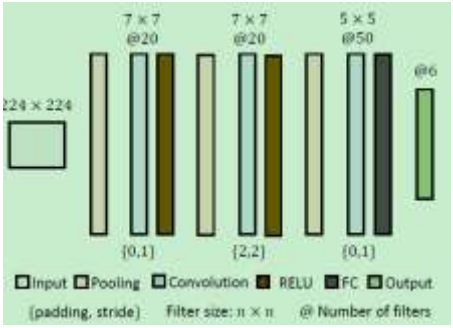
主要步骤：

首先，注册模块学习一组变换参数以最佳地注册脸部图像。第二，表示模块学习注册脸部图像的独特特征编码。



- Registration Module

本文是基于 CNN 并部署 Spatial Transformer Network，有三个部分：一个 localization network 来回归一组注册参数。然后由网格生成器使用这些参数，该网格生成器输出一个采样网格。最后，一个采样器将输入图像映射到生成的网格上。下图是 localization network 的网络结构。注意，第一个 pooling 层实现了平均池，而其余的层执行 max 操作。



对于给定的输入图像，localization network 输出一组仿射变换的六个参数(剪切、平移、旋转、缩放和倾斜)，用于生成采样网格。

- Representation Module

为了学习面部特征编码，我们使用 vgg-16，它由 8 个卷积层和 3 个完全连接的层组成，每个层后面是一个或多个非线性 relu 层然后，利用 Parkhi 等人公开的脸部数据集对整个网络进行训练。

- PersonSpecificDiscriminative Models

Gallery 包含 N 个 templates {T1, T1, T1, . . . TN} 对应 N 个注册人身份。每个 Ti = {x1, x2, . . . xM} 具有 M 个媒体（一个媒体是一个图像或者一个视频帧）。

训练一个 one-vs-rest 二进制 SVM 分类器。具体来说，学习一个人的模型参数，将该人的所有模板介质视为正类，而其余受试者的编码被认为是负类。

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_t \left(\max\left(0, 1 - \ell_t \mathbf{w}^T \mathbf{x}_t\right)\right)^2, \quad (1)$$

- Query Template Classification

利用参数 w，可以计算出 M 模板媒体属于与受试者 i 的 decision valued_i^m。

$$d_i^m = \frac{1/\left(1+\exp^{-\mathbf{w}_i^T\mathbf{x}_m}\right)}{\sum_{i=1}^{N} 1/\left(1+\exp^{-\mathbf{w}_i^T\mathbf{x}_m}\right)} \quad (2)$$

有两种方法综合 m 个模板的 decision value：

① $$y_q = \arg\max_i \sum_m d_i^m. \quad (3)$$

② Bayesian Classifier Combination (BCC) model

假设真实标签y_i是由参数为 p 的多项式分布生成的，$p(y_i = j|\mathbf{p}) = p_j$，$p_j$表示类概率。

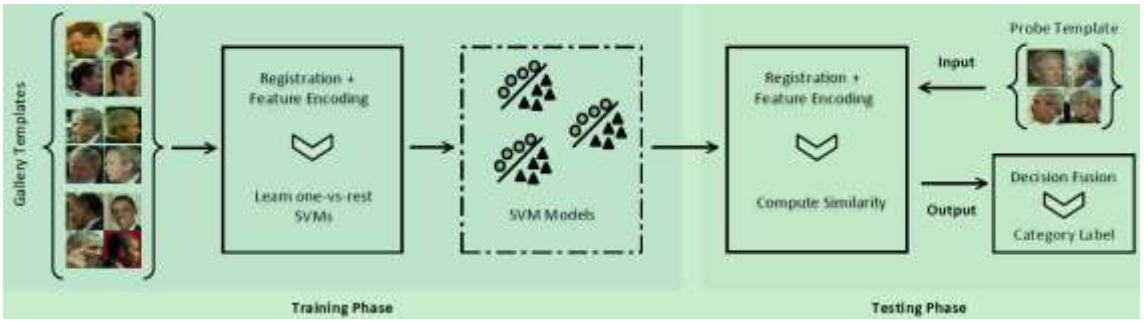
假设每个媒体的决策也都是由参数为π_j^m的多项式分布产生的，$p(d_i^m = k|y_i = j) = \pi_{j,k}^m$。$\pi_j^m$表示与每个媒体表示对应的混淆矩阵$\pi^m$的行。

参数 p 和π_j^m的先验分布是具有超参数αβ的狄利克雷分布模型：

$$p(\boldsymbol{\pi}_j^m|\boldsymbol{\alpha}_j^m) = \mathrm{Dir}(\boldsymbol{\pi}_j^m; \boldsymbol{\alpha}_j^m) \quad (4)$$
$$p(\mathbf{p}|\boldsymbol{\beta}) = \mathrm{Dir}(\boldsymbol{p}; \boldsymbol{\beta}) \quad (5)$$

则最后综合 m 个模板的 decision value 为：

$$p(\boldsymbol{y}, \mathbf{p}, \boldsymbol{\pi}|\boldsymbol{d}) \propto \prod_{i=1}^{N}\left\{p_{y_i}\prod_{m=1}^{M}\pi_{y_i,d_i^m}^m\right\} p(\mathbf{p}|\boldsymbol{\beta})p(\boldsymbol{\pi}|\boldsymbol{\alpha})$$



93. Recursive Spatial Transformer (ReST) for Alignment-Free Face Recognition

主要思想：

- 在cnn中引入了递归Spatial Transformer (ReST) (REST)模块，允许以端到端的方式与人脸识别共同学习人脸对齐。
- REST的递归结构使检测到的人脸被逐步对齐，这意味着在每个递归中的任务更容易，即使是那些变化很大的面。
- 层次结构的REST将非rigid转换分散为多个rigid转换，从而实现更精确的对齐。

主要步骤：

包括两个部分,递归空间转换(REST)和 dcnn 分类。在端到端方案中,将 REST 和 dcnn 与一个分类目标(如 Softmax 损失)一起优化。

- DCNN with ReST

REST 采用递归结构，包括卷积层 C、局部化网络 F 和空间转换层 T。卷积特征映射作为 C(X)通过卷积层实现，定位层以特征映射 C(X)为输入，预测空间变换参数$\boldsymbol{\theta} \in \mathbb{R}^{2\times 3}$：

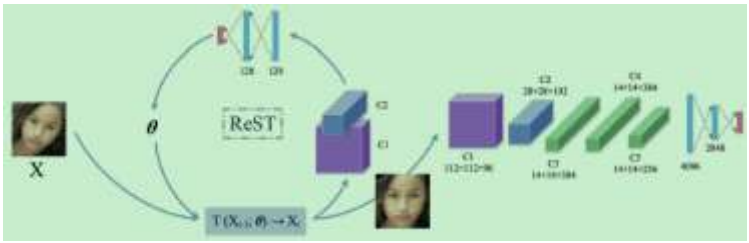
$$\boldsymbol{\theta} = \mathbf{F}(\mathbf{C}(\mathbf{X})).$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}$$

最后，空间变换层根据空间变换参数从输入 x 中采样产生变换后的特征映射。通过将新的人脸图像再一次输入到这条管道中，形成递归结构 REST，可以进一步转换新的人脸图像。

$$\mathbf{X}_i = \mathbf{T}(\mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}) \quad \boldsymbol{\theta}_i = \mathbf{F}(\mathbf{C}(\mathbf{X}_i))$$

- DCNN with Hierarchical ReST



将整个人脸划分为几个层次区域，每个区域都配备一个 REST，用于确定 rigid 区域及其相应的空间变换参数。hirest-9 是典型的层次结构，每层三区，共两层。第二层可以直接从第一层作为输入，以对齐的图像，但一般的二层也可以对齐的图像卷积特征图作为输入。hirest-3 是降级版只有一层，有三个区域。

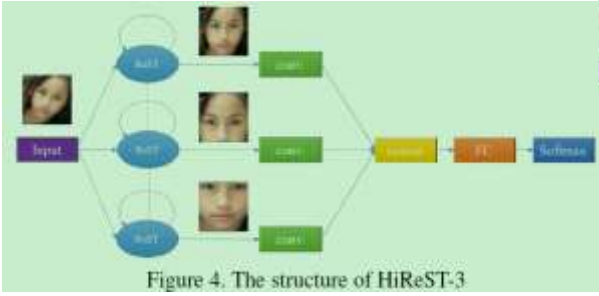


Figure 4. The structure of HiReST-3

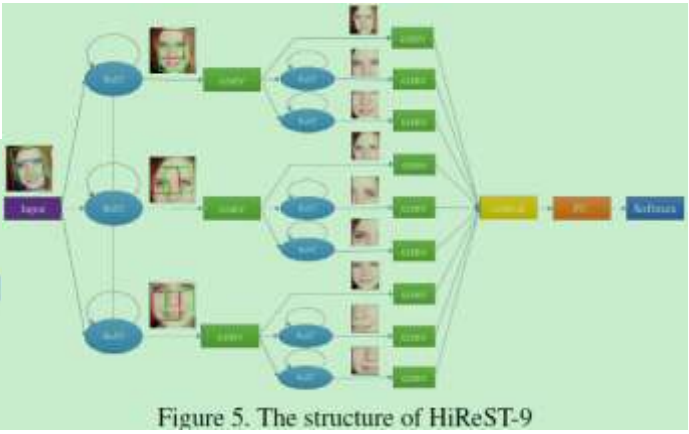


Figure 5. The structure of HiReST-9

- Fast ReST

① 首先，共享REST和后续的dcnn的前两层卷积层。例如，在REST中，dcnn的前两个卷积层与其余层共享。

② 其次，在前几次递归中，例如k(1<k<K)，空间变换直接应用于卷积特征映射，而不是人脸图像。在第一次k次递归中，卷积只计算一次，即c(X)，空间变换直接应用于卷积特征映射c(X)上，空间变换参数在k次递归后累加到输入图像x上。在后续的(>k)递归中，每个递归实现的空间转换直接应用于先前递归的转换图像，而不积累，因为后期递归中的空间转换通常较小。

$$\mathbf{X}_i = \begin{cases} \mathbf{T}(\mathbf{X}, \boldsymbol{\theta}_i), & i = k+1 \\ \mathbf{T}(\mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}), & i > k+1 \end{cases}$$

$$\boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}_{i-1} * \begin{bmatrix} \mathbf{F}(\mathbf{T}(\mathbf{C}(\mathbf{X}), \boldsymbol{\theta}_{i-1})) \\ 0 \quad 0 \quad 1 \end{bmatrix}, & 0 < i \le k+1 \\ \mathbf{F}(\mathbf{C}(\mathbf{X}_i)), & K > i > k+1 \end{cases}$$

当递归的深度为 0 时，即没有递归，建议的 REST 方法退化为典型的 cnn。当递归深度为 1 时，Spatial Transformer 可视为 REST 的特例。

五、 应用

1) 视频

94. Trunk-Branch Ensemble Convolutional NeuralNetworks for Video-based Face Recognition

主要思想：

- 首先，为了处理图像模糊，我们利用人工模糊技术对静止图像进行处理，丰富了 CNN 训练数据，弥补了现实视频训练数据的不足。由静止图像及其模糊版本组成的训练集鼓励 cnn 学习模糊健壮的表示。
- 其次，为了有效地提取姿态和遮挡鲁棒表示，我们提出了一种新的 CNN 体系结构 TBE-CNN（主干网络和多个分支网络）。TBE-CNN 通过共享不同 CNN 的低层和中间层，有效地提取了整体人脸图像和部分脸部的表示。
- 最后，为了进一步提高 TBE-CNN 的鉴别能力，我们提出了一种新的深度度量学习方法 MDR-TL，它的性能远远超过了广泛采用的 triplet loss。

主要步骤：

- Artificially Simulated Video Data

  模拟监视或移动相机成像过程中的两个挑战：运动模糊和焦距模糊。

  运动模糊：

$$k_m(i, j; L, \theta) = \begin{cases} \frac{1}{L}, & \text{if } \sqrt{i^2 + j^2} \le \frac{L}{2} \text{ and } \frac{i}{j} = -\tan\theta, \\ 0, & \text{otherwise.} \end{cases}$$

  焦距模糊：

$$k_o(i, j) = \begin{cases} C \cdot \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right), & \text{if } i \le \frac{R}{2} \text{ and } j \le \frac{R}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

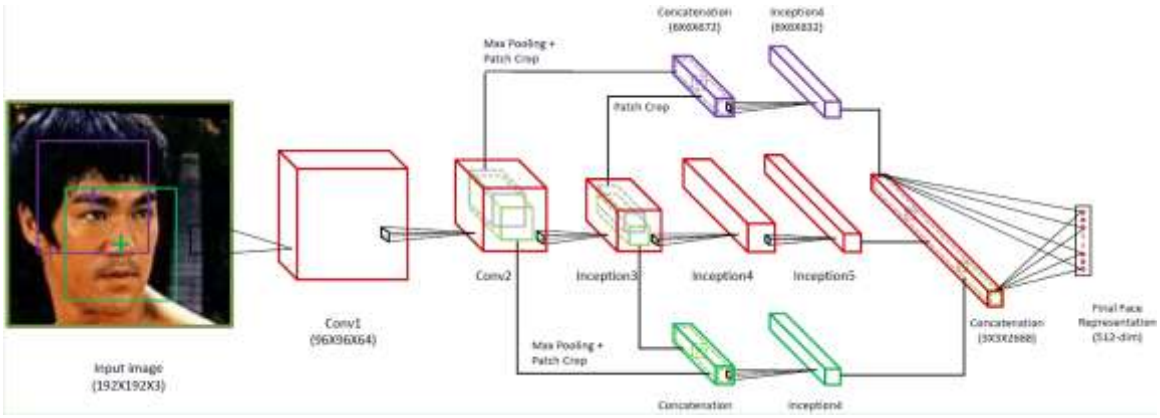
  给定一个静止的人脸图像I_s和一个模糊的核 k，通过卷积可以得到模拟的视频帧I_v: $I_v = I_s * k$.。

从每幅静止图像中得到一幅模糊图像，所以我们得到了两条大小相等的训练数据流，即一条由原始静止图像组成的流，另一条由相同数目的模糊图像组成的流。对于 CNN 培训，我们同时向 CNN 提供两种训练数据流。由于我们鼓励每个静止图像及其模糊版本被分类为同一类，CNN 会自动学习模糊的人脸表示。

- Trunk-Branch Ensemble CNN

对主干网络进行训练以学习全局人脸图像的人脸表示，并且每个分支网络被训练以学习从一个面部组件中裁剪的图像块的人脸表示。主干网络的实现是基于 GoogleNet。将 Google 网层划分为三个层次：低层、中层和高层。由于中低端特征代表局部信息，干线网络和分支网络可以共享低层和中层。高级特性代表抽象的和全局的信息；因此，不同的模型应该有单独的高层次。

从 Conv 2 输出中裁剪的 feature map 的大小被 max pooling 减少了一半，然后与从 Inception 3 输出中裁剪的 feature ma 连接起来。连接的特征映射构成分支网络的输入。为了提高效率，每个分支网络只包含一个 Inception 4 模块。



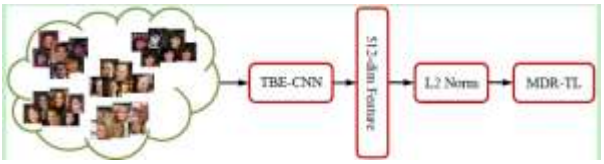
- TBE-CNN TRAINING

① 在第一阶段，主干网是使用 Softmax 损失作为惩罚单独训练的。

② 在第二阶段，主干网参数是固定的，每个分支网络都使用 Softmax 损耗进行训练。

③ 在对主干网和所有支路网络进行预训练后，对整个模型进行微调，并采用 Softmax loss 对主干和分支网络进行融合。

④ 最后，为了提高学习后的人脸表示的分辨能力，我们提出了一种新的深度度量学习方法--MDR-TL，用于对整个网络进行精细调整。



MDR-TL 中包含两个约束。一是，训练 batch 中应满足 triplet 约束：

$$\|f(x^a) - f(x^p)\|_2^2 + \beta < \|f(x^a) - f(x^n)\|_2^2,$$

第二，要将不同身份的平均表征分离开来，以确保不同身份的样本均匀分布。我们通过强制一个主题的平均表示$\overline{u_c}$与其最近的平均表示$\overline{u_c^n}$之间的边界α来实现这一约束：

$$\|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2 > \alpha, \hat{\mu}_c = \frac{\mu_c}{\|\mu_c\|}, \ \mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} f(x_{ci}),$$

因此，总损失函数为：

$$\min_f L(f) = L_{triplet}(f) + L_{mean}(f), \qquad (8)$$

$$L_{triplet}(f) = \frac{1}{2N}\sum_{i=1}^{N}[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \beta]_+,$$

$$L_{mean}(f) = \frac{1}{2P}\sum_{c=1}^{C} max(0, \alpha - \|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2),$$

95. Using Deep Autoencoders to Learn Robust Domain-Invariant Representationsfor Still-to-Video Face Recognition

主要思想：

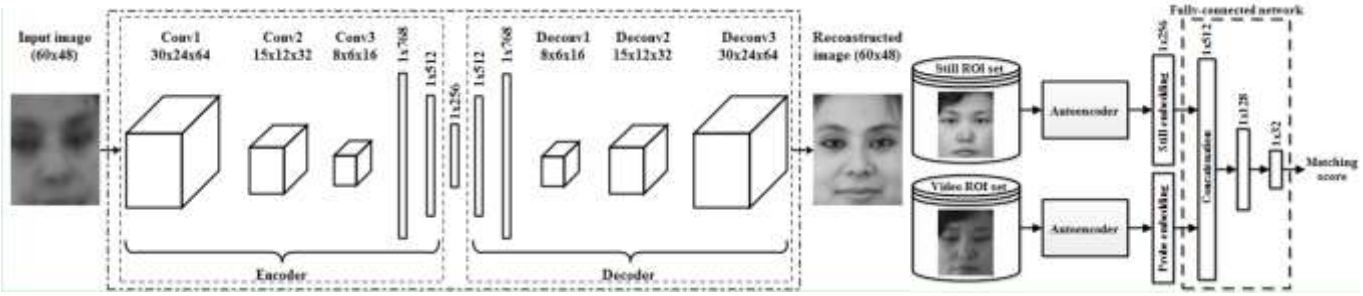
- 提出了一种高效的 CanonicalFace Representation CNN (CFR-CNN)，用于每个人的单个样本中进行静止图像-视频人脸识别，其中静态和视频 ROIs 是在不同的条件下捕获的。
- 给定在无约束视频条件下捕捉到的面部 ROI，CRF-CNN 将其重构为一个高质量的标准 ROI，用于匹配静止图像的 ROI 的条件
- 深自编码网络是使用一种新的加权损失函数，能够有效地为相同的身份类别产生相似的人脸特征。然后，使用全连接的分类网络精确地匹配静止和视频图像对。
- 该方法参数较少，高效实时

主要步骤：

① 该自动编码器网络使用一种新的加权像素级损失函数进行训练，该函数专门针对 sspp 问题，并允许重建高质量的标准人脸图片(正面、光照好、模糊程度较小、表情中立)，以匹配与参考条件相对应的仍为 ROI 的人脸。

② 该自动编码器的中间层被设计为生成对相同个体相似的人脸特征表示，并且对无约束的真实世界视频场景中通常观察到的变化具有很强的鲁棒性。

③ 训练一个完全连接的分类网络，使用从深层自动编码器中提取的人脸特征表示来执行人脸匹配，并准确地确定静止和视频 ROI 对是否对应于同一个个体。



- Autoencoder Network

输入图像是使用监视摄像机捕获的 probe 视频 ROI，而输出是重建图像。该网络由(1)三个卷积层组成，每个层接着一个最大池层来提取鲁棒卷积映射，(2)一个生成 256 维面嵌入的两层全连接网络。解码器逆转这些操作。然后利用中间层的输出作为人脸表示，该表示不受无约束监视环境中常见的不同干扰因素的影响。最后，利用完全连通的分类网络进行人脸匹配。

该自动编码器网络的参数采用了一种新的加权均方误差(MSE)准则进行优化，其中有一个 T 形区域被认为对眼睛、鼻子和嘴巴等有区别的面部成分给予了更高的重视。

$$L_{CFR-CNN} = \sum_{i \in rows} \sum_{j \in cols} \tau_{i,j} \left\| X^2 - \hat{X}^2 \right\| \quad (1)$$

$$\tau_{i,j} = \begin{cases} \alpha & \text{if (i,j) belongs to T} \\ \beta & \text{if (i,j) otherwise} \end{cases}$$

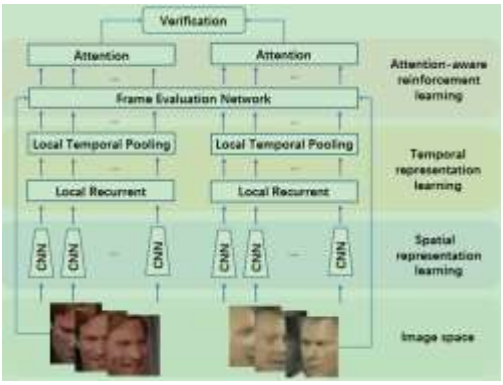
96. Attention-aware Deep Reinforcement Learning for Video Face Recognition

主要思想：

- 提出了一种基于注意感知的深度强化学习(ADRL)方法，该方法旨在消除人脸识别中存在的误导性和混淆性帧，寻找人脸识别中注意的焦点。
- 将视频注意力的发现过程描述为马尔可夫决策过程，并通过深度强化学习框架对注意模型进行训练，而不使用额外的标签。与已有的注意模型不同，该方法以图像空间和特征空间的信息为输入，更好地利用了特征学习过程中丢弃的人脸信息。

主要步骤：

我们的框架由两部分组成：特征学习和注意力学习。特征学习部分是一个以一对人脸视频为输入的网络。该网络用一个卷积神经网络(CNN)、一个递归层和一个时间池层来处理整个视频，分别产生视频中每个帧的时间表示。注意部分是一个帧评价网络，它是为了产生帧的值而设计的。这些值被用来寻找最有代表性的帧，这是视频关注的焦点。将过程描述为马尔可夫决策过程(MDP)，并引入了一种增强学习方法来训练评价网络。帧评价网络的输入信息既来自图像空间，也来自特征空间。



- Temporal Representation Learning

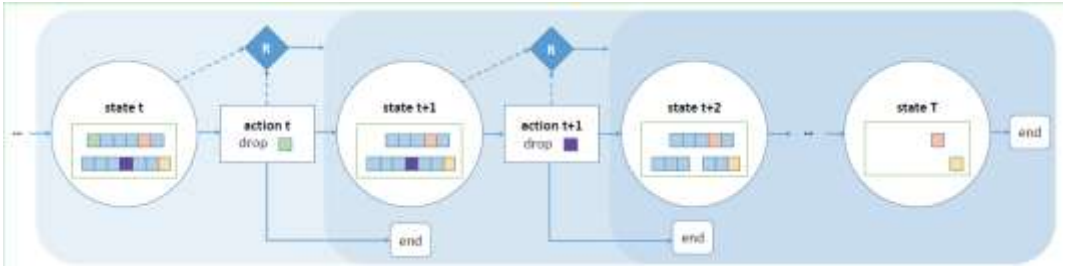
引入了一个更灵活的局部双向递归层和局部时态池层，它们将几个相邻帧组合成一个时态表示，并将其他帧视为无关帧。

C1(X)是 cnn 特征表示，每帧x_i^A都有相应的卷积特征表示$f_i^A = C_1(x_i^A)$。采用了广泛使用的长时记忆(Lstm)作为递归层和均值池策略来组合特征，从而使帧x_i^A的时间表示成为：

$$h_i^A = \frac{1}{1+2r}\sum_{k=i-r}^{i+r} m_k^A, \quad [m_{i-r}^A, ..., m_i^A, ..., m_{i+r}^A] = \mathcal{R}([f_{i-r}^A, ..., f_i^A, ..., f_{i+r}^A]),$$

其中，R 是 LSTM，r 是相邻帧的范围。CNN 模型和递归层分别进行训练。使用为静止图片训练 CNN 模型。

- Attention-aware Deep Reinforcement Learning



提出了一种深度模型$Q_i = C_2(I_i, M_i)$，以图像空间的信息 Ii 和特征空间的 Mi 作为输入，生成值 Qi，从而评价了 Xi 的帧情，确定了注意的权重。并设计了一种用 C1 来教 c2 的算法。

我们将 t 次下降后的剩余帧表示为 st 状态，即丢弃帧的作用。删除帧可能导致两种状态：St+1 和终止。对(st，at)的的评价反馈 ri 由专家 C1 决定。在实践中，我们使用均值池特征计算的余弦相似性作为两个不同视频的度量。

$$S(X^A, X^B|s_t) = \cos(p^A|s_t, p^B|s_t),$$

将 ri 定义为在(st，at)时，对于评价精度的提升：

$r_t = l_{AB}(S(X^A, X^B|s_{t+1}) - S(X^A, X^B|s_t))$.且当 $r(s_t, a_t) < 0, \forall a_t$.则终止。

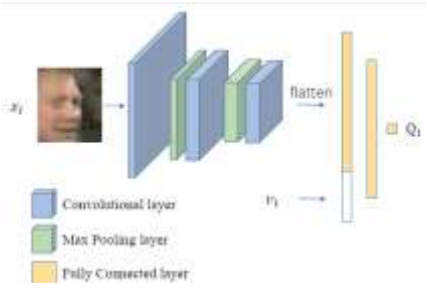
Qi 作为在 St 做出 at 的行动的期望：

$$Q_i = Q(s_t, a_t) = \max_{\pi} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|\pi]. \quad (7)$$

使用深度网络 C2 来得到 $Q^*(s_t, a_t) = C_2(s_t, a_t)$.

帧评价网络采用了丢弃帧 xi 和向量 vi，描述了特征空间中丢帧和两个视频之间的几何关系。帧评价网络首先用卷积网络将 xi 表示为深特征，然后将特征与向量 vi 连接，将级联特征转化为完全连通的网络，生成 Q(st，at)。

$$\min_{\theta} \mathcal{H} = \mathbb{E}[r(s_i, a_i) + \gamma \max_{a_{i+1}} Q(s_{i+1}, a_{i+1}) - Q(s_i, a_i)]^2,$$



2) 人脸攻击

97. Adversarial Generative Nets: Neural NetworkAttacks on State-of-the-Art Face Recognition

主要思想：

- 展示了如何创建眼镜，当戴上眼镜时，可以成功地进行目标攻击(模拟)或非目标攻击(躲避)
- (1)不明显，这是我们通过用户研究来测试的；(2)攻击对拟议的防御措施的稳健性；(3)将眼镜与将佩戴眼镜的人分离开来的可伸缩性，即创造一套“通用”眼镜，以便于分类错误。

主要步骤：

在模拟(或目标)攻击中，对手试图被错误地归类为特定的其他人。在躲避(或非目标)攻击时，对手试图被错误地归类为任意其他人员。

- Generative Adversarial Networks

$$Loss_G(Z, D) = \sum_{z \in Z} \lg\left(1 - D(G(z))\right)$$

$$Gain_D(G,Z,data) = \sum_{x \in data} \lg\left(D(x)\right) + \sum_{z \in Z} \lg\left(1 - D(G(z))\right)$$

- Attack Framework

训练神经网络来生成眼镜的图像，当戴上眼镜时，会引起闪避或冒充。为了达到不引人注目的目的，我们要求这些神经网络生成的眼镜类似于真实的眼镜设计。

AGN 训练涉及三个神经网络：一个发生器 G；一个判别器 D；和一个预先训练的系统，其分类函数用 F()表示。

给定 DNN 的输入 x，G 被训练以产生欺骗 F()并且不显眼的输出。

$$Loss_G(Z,D) - \kappa \cdot \sum_{z \in Z} Loss_F(x + G(z))$$

$Loss_G$以与上面的方式进行训练——最小化它的目的是产生误导 D 的真实的(即不明显的)输出。$Loss_F$是定义在 DNN 的分类上的损失函数，当训练 G 时它被最大化。$Loss_F$的定义取决于攻击者是否旨在实现躲避或模拟。为了躲避，我们使用：

$$Loss_F(x + G(z)) = \sum_{i \neq x} F_{c_i}(x + G(z)) - F_{c_x}(x + G(z))$$

对于模拟，我们使用：

$$Loss_F(x + G(z)) = F_{c_t}(x + G(z)) - \sum_{i \neq t} F_{c_i}(x + G(z))$$

为了躲避，正确类 CX 的概率降低；而对于模拟，目标类 ct 的概率增加。

3) 近红外

98. Heterogeneous Face Recognition with CNNs

主要思想：

- 利用 CNN 现在可见光图片上进行预训练，然后使用 CASIA NIR-VIS dataset 进行微调，用来进行异构的人脸识别（可见光图像和红外图像）
- 探索不同的度量学习策略，以减少不同模式之间的差异。

主要方法：

- Learning a deep CNN model

使用预训练的 CNN 从 Pool3 到 softmax 层的各个层提取图像特征。实验证明在这两个领域中使用 P4 特征得到最好的结果。

	S	P5	C52	C51	P4	C42	C41	P3
Raw features	63.1	62.7	63.8	51.0	29.4	26.8	18.8	14.8
Domain adapt. [3]	63.1	62.7	64.2	51.8	31.8	28.6	19.1	13.7
Our approach	72.6	75.3	80.6	82.9	**85.9**	84.8	83.5	79.5

- Metric learning to align modalities

采用马氏距离

- Shared vs. separate projection matrices

为每个域学习一个单独的投影矩阵，将其映射到公共子空间，在每个领域中提取不同维度的领域特定特征。

或者使用相同的投影矩阵

- Inter-domain and Intra-domain pairs

度量学习中的另一种设计选择是用于训练的对。同时对域内和域间图像对进行区分。我们的目标是将 gallery 集中一模态的图像和 probe 集的另一个模态的图像相匹配，而跨域对直接反映这一点。域内对与我们任务的多模态特性无关，但正如我们在实验中所展示的，它们提供了一种正规化的形式。

		S	P5	C52	C51	P4	C42	C41	P3
Inter+Intra	Shared	72.6	75.3	**80.6**	**82.9**	**85.9**	**84.8**	**83.5**	**79.5**
	Separate	66.6	70.4	78.6	80.0	82.4	80.7	76.6	69.2
Inter	Shared	70.0	74.3	79.8	81.7	83.6	82.0	78.6	72.3
	Separate	**73.0**	**75.7**	77.9	76.8	76.91	74.7	63.1	52.9

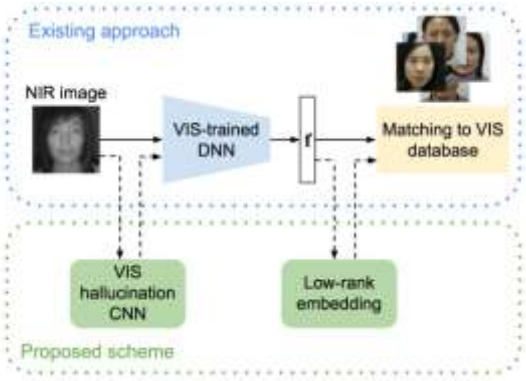
99. Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-spectral Hallucination and Low-rank Embedding

主要思想：

- 调整一个经过预先训练的最先进的 dnn（只为 VIS 人脸图像设计的），为 VIS 和 NIR 人脸图像生成鉴别特征，而无需再培训 dnn。
- 该方法由两部分组成， cross-spectral hallucination 和低秩嵌入，分别适应 DNN 输入和输出的交叉谱识别。
- cross-spectral hallucination 使用 cnn 预处理近红外图像，执行交叉光谱转换近红外图像为 VIS 光谱。
- 低秩嵌入为同一类别的交叉光谱特征恢复低秩结构，同时增加不同的类别的距离。

主要方法：

一个简单的 NIR-VIS 人脸识别系统是利用仅对 VIS 图像进行训练的深度神经网络(DNN)从 NIR 图像中提取特征向量 f，并将其用于 VIS 数据库的匹配。首先，我们通过从 NIR 样本中产生 VIS 图像来修改输入。其次，在输出端对 DNN 特征进行低秩嵌入.。每一种方法在多模态识别性能上产生重要的改善，如果一起应用，则会产生更大的改进。



- Cross-spectral Hallucination

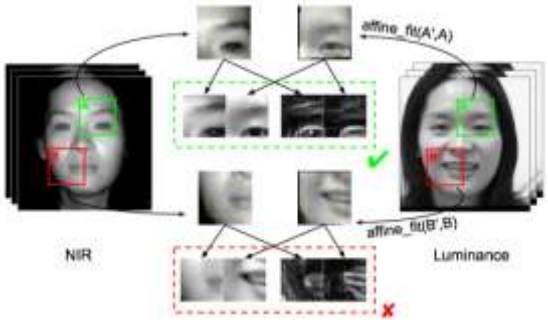
cross-spectral hallucination CNN 在 pairs of corresponding NIR-VIS patches 上训练，这些 patches 是从公开数据库中挖掘出来的。

在 VIS 领域，我们在亮度-色度空间中工作，因为它将重要的图像细节集中在亮度通道中，并且最小化了通道之间的相关性，从而使学习更加有效。我们用 YCbCr 空间进行训练，取得了最佳效果，我们训练了三个不同的网络。由于亮度通道 Y 包含了主体的大部分信息，因此我们利用一个较大的网络来处理这个通道，而对于两个颜色则使用一个较小的网络。此外，由于蓝色成分在表面上变化很小，一个更小的网络就足以实现蓝差色度 CB。

Ch.	layers	first and last	intermediate	skip-connections
Y	11	148x11x11 st. 1, pad 5 PReLU	36x11x11 st. 1, pad 5 PReLU	input to last layer
Cb	7	66x3x3 st. 1, pad 1 PReLU	32x3x3 st. 1, pad 1 PReLU	none
Cr	8	148x5x5 st. 1, pad 2 PReLU	48x5x5 st. 1, pad 2 PReLU	none

- Mining for NIR-VIS patches

对同一类别比较每一个近红外图像与每一个 VIS 图像的亮度通道（NIR 和 VIS 图像是在不同的姿势和面部表情下拍摄的）。使用面部标志对齐提取 224x224crop 进行初步对齐。从两幅图像中在同一位置滑动的 60x60crop 进行提取。VIScrop 利用相似变化对齐至 NIRcrop。然后，我们 patch 裁剪一个 40x40 区域。如果这两个 patch 及其梯度的相关性（利用 cnn 计算）超过一个阈值，我们就保留这对。在这个例子中，补丁 A 和 A‘形成一个合格对，而 B 和 B’被丢弃。



- Post-processing

  利用原始的 nir 图像和 cnn 的输出链接起来，从而弥补 cnn 输出中的 small artifacts。

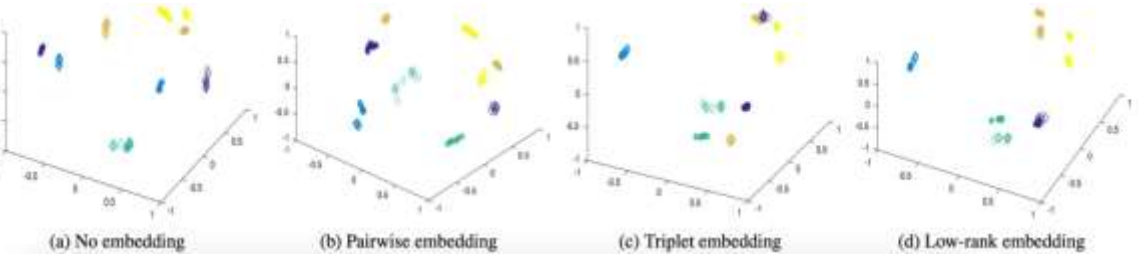
$$Y = \hat{Y} - \alpha \cdot G_\sigma^2 * (N_{ir} - \hat{Y}),$$

$\hat{Y}$是交叉谱 cnn 的输出，Nir 是近红外图像，Gσ 是 σ=1 的高斯近似，代表卷积。参数 α 平衡了从 NIR 图像中保留的信息量和 cnn 获得的信息，并允许移除 cnn 引入的一些工件(在我们的实验中 α=0.6)。

- Low-rank Embedding

设 Yc 表示由 y 的位于 c-类中的列构成的子矩阵。一个 d×d 低秩变换 T 被学习到最小化。

$$\sum_{c=1}^{C} ||\mathbf{TY}_c||_* - ||\mathbf{TY}||_*,$$



同一颜色空心和实心的点被聚合到一起，因此低秩变换 T 可以有效地恢复对同一类别的低等级结构，即使 YC 含有混合的近红外和可见光的训练数据。

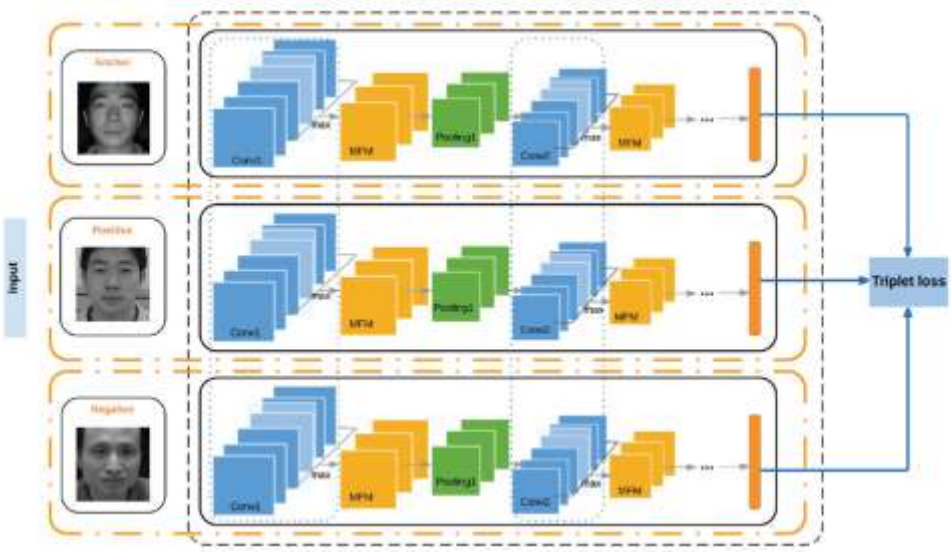
100. Transferring Deep Representation for NIR-VIS Heterogeneous Face Recognition

主要思想：

- 提出了一种用于 Nir-Vis 人脸识别的深度传递 Nir-Vis 人脸识别网络。
- 首先，为了利用大量未配对的人脸图像，利用激活函数(max-Feature-map)选择判别特征，使模型具有鲁棒性和轻量化。
- 第二，通过两种 triplet loss 的微调将这些模型转移到 Nir-Vis 域。triplet loss 不仅减少了类内 Nir-Vis 的变化，而且增加了正训练样本对的数目。它使得在一个小数据集上微调深层模型成为可能。

主要方法：

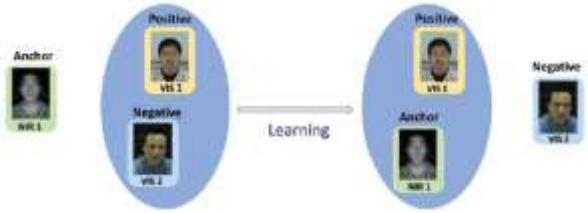
CNN 的输入是 triplets，三个通道具有相同的参数。在最终的完全连通层中提取特征后，将三层的高层次特征输入到三重态损耗层，从而连接 NIR 和 Vis 域之间的距离。



- Triplets Formation

$$loss = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

首先，我们设置了一个近红外图像作为 anchor，一个相同 ID 的可见图像作为一个正例，另一个不同 ID 的可见图像作为一个负例子。然后，利用 VIS 作为 anchor，相同 ID 的近红外图像作为正例，不同 ID 的近红外图像作为反例，构建另一个 triplets。无论人脸图像属于哪一种形态，如果它们来自同一个身份，那么人脸特征就会变得更加接近，如果它们属于不同的身份，则会彼此远离。



- Hard NIR-VIS Triplets Selection

将所有训练图像(NIR 和 VIS 图像)输入测试网络，计算相似度矩阵。以 NIRanchor 为例(锚也可以是 VIS 图像)，计算其与所有 VIS 人脸图像的相似度。分数较低但与 anchor 相同的图像被视为 hard positive 样本,而分数较高但 ID 不同的图像被标记为 hard negative 样本。困难的 triplets 由 probe NIR 照片、hard positiveVIS 图像和 hard negativeVIS 图像组成。

- CNN with Ordinal Measures

$$f_{ij}^k = \max(C_{ij}^k, C_{ij}^{k+n}), k \in \{1, \ldots, n\}$$

- Deep Transfer Learning

利用 CASIA WebFace Dataset 预训练网络。

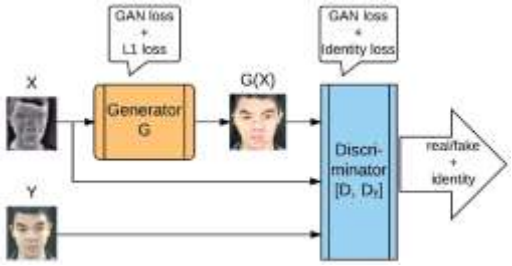
101. TV-GAN: Generative Adversarial Network Based Thermal to Visible FaceRecognition

主要思想：

- 借鉴 Pix2Pix 和 DR-GAN 的思想
- 提出一种 Thermal-to-Visible Generative Adversarial Network (tv-gan)，它能够将热人脸图像转换成相应的 vld 图像，同时保持身份信息，这足以使现有的 vld 人脸识别模型能够进行识别。
- 通过在鉴别器中插入一个封闭集人脸识别任务损失，使得生成器能够学习到保留了足够身份信息的转换函数。

主要方法：

我们的鉴别器不仅提供假和真实的识别，而且执行一个封闭集的人脸识别任务。生成器网络 g 的目标是生成能够“欺骗”鉴别器的图像。多任务鉴别器是由两个具有共享权重的网络组成的，[D, D_y]，其中 D 判别 y 是否为真，$\mathrm{D}_y : \boldsymbol{X} \times \boldsymbol{Y} \mapsto \{0,1\}^{N+1}$执行封闭集人脸识别任务。



$$\mathcal{L}_{TV-GAN}(\mathrm{G},\mathrm{D},\mathrm{D}_{\boldsymbol{y}}) = \mathcal{L}_{cGAN}(\mathrm{G},\mathrm{D}) + \lambda_1\mathcal{L}_{\ell_1}(\mathrm{G}) + \lambda_2\mathcal{L}_{id}(\mathrm{G},\mathrm{D}_{\boldsymbol{y}}),$$

其中$\mathcal{L}_{\text{cGAN}}$是传统 GAN 的优化函数：

$$\mathcal{L}_{cGAN}(\mathrm{G},\mathrm{D}) = \mathbb{E}_{\boldsymbol{X}\sim P_{data}(\boldsymbol{X})}[\log \mathrm{D}(\boldsymbol{X},\boldsymbol{Y})] + \mathbb{E}_{\boldsymbol{X}\sim P_{data}}[\log 1-\mathrm{D}(\boldsymbol{X},\mathrm{G}(\boldsymbol{X}))]$$

$\mathcal{L}_{l1}$借鉴 Pix2Pix 的思想：

$$\mathcal{L}_{\ell_1}(\mathrm{G}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}\sim P_{data}}[\|\boldsymbol{Y}-\mathrm{D}(\boldsymbol{X})\|_1].$$

$\mathcal{L}_{\text{id}}$是借鉴 DR-GAN 的思想：

$$\mathcal{L}_{id}(\mathrm{G},\mathrm{D}_{\boldsymbol{y}}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}\sim P_{data}}[\log(\mathrm{D}_{\boldsymbol{y}}(\boldsymbol{X},\boldsymbol{Y})]$$

102. Learning Invariant Deep Representation for NIR-VIS Face Recognition

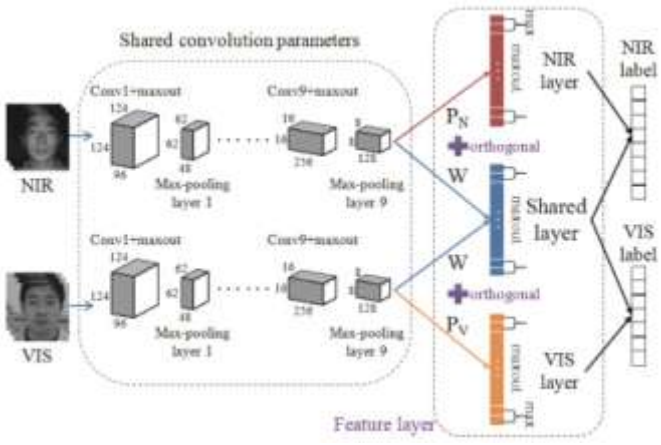
主要思想：

- 本文提出了一种深卷积网络方法，该方法只使用一个网络将近红外和可见光图像映射到紧致欧氏空间。
- 这个网络的低层只对大规模的 VIS 数据进行培训。每个卷积层都是用最简单的 Maxout 算子来实现的。
- 将高阶层划分为两个正交子空间，分别包含模态不变的身份信息和模态变异的谱信息。

主要方法：

在基本 VIS 网络的基础上,提出了一种用于 NIR-Vis 人脸识别的模态不变卷积神经网络。低层卷积层由预训练的基本网络初始化.。我们实现了两个共享参数的 CNN 信道，分别输入 NIR 和 VIS 图像。然后，我们设计了一个特征层，其目的是将低层特征投影到两个正交特征子空间中。这样，我们就可以利用 NIR 和 VIS 共享的身份信息，并加强这两种模式的领域特性。最后，softmax 函数分别用于近红外和可见光作为监控信号的表示。

使用 light cnnb 作为基础网络。该网络包括 9 个卷积层，4 个最大池层，然后是全连接层。以 softmax 作为损失函数。对人脸图像进行归一化训练，根据人脸点将训练值压缩到 144×144。为了丰富输入数据，我们随机地将输入图像裁剪成 128×128。MS-Celeb-1M 数据集作为训练集。



CNN 特征提取过程表示为 $X_i = Conv(I_i,\Theta)$.（i∈{N, V}），其中 Conv()是 ConvNet 的特征提取函数。

然后引入三个映射矩阵(i.e., $W, P_i \in \mathrm{R}^{d\times m}$)来建模身份不变信息和变谱信息。因此，特征表示可以按照

$$F_i = \begin{bmatrix} F_{\text{shared}} \\ F_{\text{unique}} \end{bmatrix} = \begin{bmatrix} WX_i \\ P_iX_i \end{bmatrix} \ (i\in\{N,V\})$$

进一步对它们施加正交约束以使它们彼此不相关：$P_i^TW = 0\ (i\in\{N,V\})$

最后优化函数：

$$\mathcal{L}(F,c,\Theta,W,P)=\sum_{i\in\{N,V\}}\text{softmax}\{F_i,c,\Theta,W,P_i\}+\sum_{i\in\{N,V\}}\lambda_i\|P_i^TW\|_F^2 \quad (4)$$

如果采用梯度下降法来最小化(4)，则需要更新参数θ，W、PI，对于 CNN 参数θ，则采用传统的反向传播方法进行更新。W 和 PI 的梯度包含两个可以表示为

$$\frac{\partial\mathcal{L}}{\partial W}=\sum_{i\in\{N,V\}}\frac{\partial\text{softmax}(F_i,c,\Theta_i,W,P_i)}{\partial W}+\sum_{i\in\{N,V\}}\lambda_iP_iP_i^TW \quad (5)$$

$$\frac{\partial\mathcal{L}}{\partial P_i}=\frac{\partial\text{softmax}(F_i,c,\Theta_i,W,P_i)}{\partial P_i}+\lambda_iWW^TP_i \quad (6)$$

103. Wasserstein CNN: Learning Invariant Featuresfor NIR-VIS Face Recognition

描述了 NIR-VIS 的难点

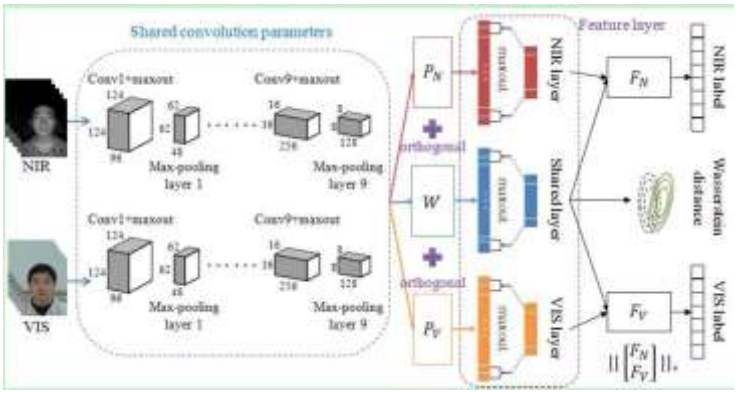
主要思想：

- 本文提出了一种新的方法，即 WassersteinCNN(简称 Wcnn)，用于学习近红外和视觉人脸图像之间的不变特征(即 Nir-Vis 人脸识别)。
- Wcnn 的低层层是利用视觉光谱中广泛可用的人脸图像进行训练的。高层层分为三部分，即 NIR 层、Vis 层和 Nir-Vis 共享层。前两层的目标是学习特定模态的特征，而 Nir-vis 共享层的目的是学习模态不变的特征子空间。
- 在 Nir-Vis 共享层中引入 Wasserstein 距离，以度量异构特征分布之间的差异。因此，w-CNN 学习的目的是将 Nir 分布与 Vis 分布之间的 Wserstein 距离最小化，以实现异构人脸图像的不变深度特征表示。
- 为了避免小尺度非均匀人脸数据的过拟合问题，在 Wcnn 网络的全连通层上引入了相关先验，以减少参数空间。该先验是通过端到端网络中的 low-rank constraint 来实现的。

主要方法：

- Network Structure

使用 light CNN 作为基础网络，低层卷积层由预训练的基本网络初始化。我们实现了两个具有共享参数的 cnn 信道，分别输入 Nir 和 Vis 图像。



- Modality Invariant Subspace

将低层特征投影到两个正交特征子空间的特征层：

$$f_i=\begin{bmatrix}f_{\text{shared}}\\f_{\text{unique}}\end{bmatrix}=\begin{bmatrix}WX_i\\P_i^TX_i\end{bmatrix}\ (i\in\{N,V\}), \quad (1)$$

$$P_i^TW=0\ (i\in\{N,V\}).$$

- The Wasserstein Distance

$$W_2(X,Y)^2=\frac{1}{2}\left[\|m_N-m_V\|_2^2+(c_N+c_V-2\sqrt{c_Nc_V})\right]=\frac{1}{2}\left[\|m_N-m_V\|_2^2+\|\sigma_N-\sigma_V\|_2^2\right],$$

- low-rank constraint

Wcnn 的全连通层由两个矩阵组成：FN 和 FV 分别对应于 NIR 和 VIST。设$\mathrm{M}=\begin{bmatrix}F_N\\F_V\end{bmatrix}$是高度相关的，因此$M^T$M是块对角矩阵。相关的 M 会减少估计参数空间，自然地缓解过拟合问题。我们利用 m 上的矩阵核范数

$$\|\mathrm{M}\|_*=\mathrm{tr}(\sqrt{\mathrm{M}^T\mathrm{M}}).$$

- Optimization Method

首先，我们通过传统的反向传播更新参数来优化 cnn。然后，我们固定 CNN 参数，并通过它们自己的梯度更新矩阵 W、Pi、Fi。

$$\mathcal{L}=\beta_1\mathcal{L}_{ida}+\beta_2\mathcal{L}_{dist}+\beta_3 R+\sum_{i\in\{N,V\}}\lambda_i\|P_i^TW\|_F^2,\quad(16)$$

$$\mathcal{L}_{ida}=\sum_{i\in\{N,V\}}\mathrm{softmax}(F_i,c,\Theta,W,P_i)+\sum_{i\in\{N,V\}}\lambda_i\|P_i^TW\|_F^2,$$

$$\mathcal{L}_{dist}=\frac{1}{2}\left[\|m_N-m_V\|_2^2+\|\sigma_N-\sigma_V\|_2^2\right].$$

Algorithm 1: Training the Wasserstein CNN network.

Require: Training set X_i, learning rate γ and lagrange multipliers λ_i.
Ensure: The CNN parameters Θ and the mapping matrix W.

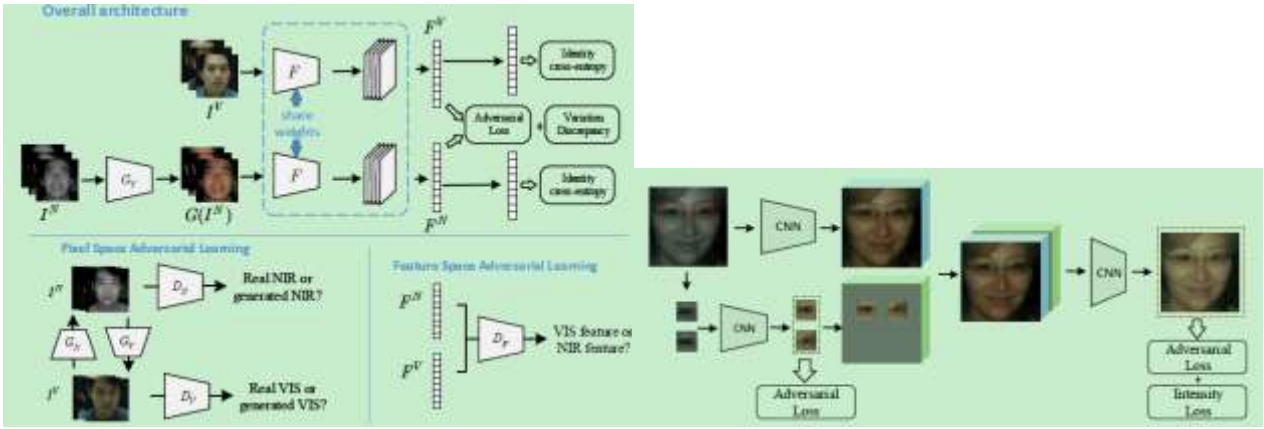
1: Initialize parameters Θ by pre-trained model and the mapping matrices W, P_i, F_i by Eq.(26);
2: **for** $t=1,\ldots,T$ **do**
3: CNN optimization:
4: Update Θ, W, P_i, F_i via back-propagation method;
5: Fix Θ:
6: Update W according to Eq.(20);
7: Update P_i according to Eq.(21);
8: Update F_i according to Eq.(12);
9: **end for**;
10: **Return** Θ and W;

104. Adversarial Discriminative Heterogeneous Face Recognition

主要思想：

- 提出了一种基于原始像素空间和紧凑特征空间的对抗性特征学习框架，通过对抗性学习来缩小 VIS 和 NIR 的差距。该框架将交叉光谱人脸幻觉和鉴别特征学习集成到端到端对抗性网络中。
- 在像素空间中，我们利用生成的对抗性网络来实现交叉光谱的人脸幻觉。提出了一种双路径模型，该模型既考虑了全局结构，又考虑了局部纹理。
- 在特征空间中，分别用一个对抗性损失和一个高阶方差差异损失来度量两个异质分布之间的全局和局部差异。

主要方法：



- Cross-spectral Face Hallucination

建立基于 cycle GAN 框架实现交叉光谱人脸幻觉模型。一对生成器$G_V: I^N \to I^V$ and $G_N: I^V \to I^N$来实现反向变换，从而构造 VIS 和 NIR 域之间的映射循环。与这两种产生器相关联的D_V and D_N旨在区分真实图像 i 和生成的图像 g(I)。

$$L_{G-adv} = -\mathbb{E}_{I\sim P(I)} \log D(G(I)), \quad (1)$$

$$L_{D-adv} = \mathbb{E}_{I'\sim P(I')} \log D(1-I') + \mathbb{E}_{I\sim P(I)} \log D(G(I)), \quad (2)$$

cycle consistencyloss 来保证输入图像和重构图像之间的一致性：

$$L_{cyc} = E_{I\sim P(I)} \| I - F(G(I)) \|_1,$$

由于眼周区域显示了Nir图像与与其他面部区域不同的Vis图像之间的特殊对应关系,所以我们在眼睛周围添加了一条局部路径，以便精确地恢复眼睛周围区域的细节。我们选择在 ycbcr 空间中表示输入和输出图像，其中亮度分量 y 编码大部分的结构信息以及身份信息。在全局路径中采用亮度保持项来保证结构的一致性：

$$L_{intensity} = \mathbb{E}_{I\sim P(I)} \| Y(I) - Y(G(I)) \|_1$$

因此生成器总体损失函数为：

$$L_G = L_{G-adv} + \alpha_1 L_{cyc} + \alpha_2 L_{intensity}$$

- Adversarial Discriminative Feature Learning

① Adversarial Loss：额外的鉴别器 DF 被用来对抗特征提取器：$L_{F-adv} = -\mathbb{E}_{I^N\sim P(I^N)} \log D_f(F(G_V(I^N)))$

② Variance Discrepancy：考虑到同一主题的特征分布应尽可能接近

$$\sigma(F) = \mathbb{E}\left((F - \mathbb{E}(F))^2\right), \quad (7)$$

$$L_{CVD} = \sum_{c=1}^{C} \mathbb{E}\left(\| \sigma(F_c^V) - \sigma(F_c^N) \|_2\right) \quad (8)$$

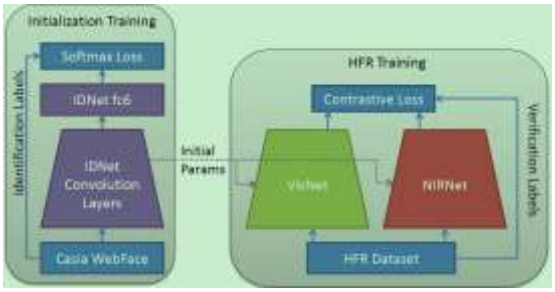
③ Cross-Entropy Loss：保持身份的可区分性 $L_{ds} = \frac{1}{|N|+|V|} \sum_{i\in(N,V)} \mathcal{L}(WF_i, y_i) \quad (9)$

105. Seeing the Forest from the Trees: A Holistic Approach to Near-infraredHeterogeneous Face Recognition

主要思想：

- 使用小卷积滤波器训练两个网络(名为 VisNet 和 Nirnet)，分别用于从可见和近红外图像中提取特征。
- 并通过创建一个具有 contrastive loss 的 Siamese network 来耦合这两个网络的输出特性。

主要方法：



首先在一个大的可见人脸数据集上训练一个深卷积神经网络来进行识别。然后，我们使用经过训练的网络初始化两个特定模态的两个网络，用于从可见和近红外图像中提取特征。最后，我们进一步优化了 HFR 训练数据上的 HFR 网络，使其能够匹配输出特征，适合于跨模态人脸识别。

- IDNet

采用 GoogleNet，在 CASIAWebFace 数据库上预训练。结构如下：

	Name	Type	Filter Size	Stride	Output Size	Params
Section 1	conv11	Convolution	$3 \times 3 \times 32$	1	$100 \times 100 \times 32$	288
	relu11	ReLU			$100 \times 100 \times 32$	0
	conv12	Convolution	$3 \times 3 \times 64$	1	$100 \times 100 \times 64$	18.4K
	relu12	ReLU			$100 \times 100 \times 64$	0
	pool1	Max Pooling	2×2	2	$50 \times 50 \times 64$	0
Section 2	conv21	Convolution	$3 \times 3 \times 64$	1	$50 \times 50 \times 64$	36.7K
	relu21	ReLU			$50 \times 50 \times 64$	0
	conv22	Convolution	$3 \times 3 \times 128$	1	$50 \times 50 \times 128$	73.7K
	relu22	ReLU			$50 \times 50 \times 128$	0
	pool2	Max Pooling	2×2	2	$25 \times 25 \times 128$	0
Section 3	conv31	Convolution	$3 \times 3 \times 96$	1	$25 \times 25 \times 128$	111K
	relu31	ReLU			$25 \times 25 \times 128$	0
	conv32	Convolution	$3 \times 3 \times 192$	1	$25 \times 25 \times 192$	166K
	relu32	ReLU			$25 \times 25 \times 192$	0
	pool3	Max Pooling	2×2	2	$13 \times 13 \times 192$	0
Section 4	conv41	Convolution	$3 \times 3 \times 128$	1	$13 \times 13 \times 128$	221K
	relu41	ReLU			$13 \times 13 \times 128$	0
	conv42	Convolution	$3 \times 3 \times 256$	1	$13 \times 13 \times 256$	295K
	relu42	ReLU			$13 \times 13 \times 256$	0
	pool4	Max Pooling	2×2	2	$7 \times 7 \times 256$	0
Section 5	conv51	Convolution	$3 \times 3 \times 160$	1	$7 \times 7 \times 160$	369K
	relu51	ReLU			$7 \times 7 \times 160$	0
	conv52	Convolution	$3 \times 3 \times 320$	1	$7 \times 7 \times 320$	461K
	relu52	ReLU			$7 \times 7 \times 320$	0
	pool5	Avg Pooling	7×7	1	$1 \times 1 \times 320$	0
	fc6	Fully Connected			10575	3.38M
	cost	Softmax			10575	0

- HFR Networks

培训两个跨模态网络：VisNet(用于可见图像)和 Nirnet(用于近红外图像)。我们将这些网络用 IDnet 初始化，不包括完全连接的 Softmax 分类器。

VisNet 和 Nirnet 通过创建一个 Siamese network 来耦合它们的输出特性。虽然用相同的值初始化，但并不强迫网络的两个部分在训练期间共享权重。用对比损失作为 VIS-net 和 NERNET 的网络损失函数。

$$L(\mathbf{x},\mathbf{y}) = \begin{cases} \|\mathbf{x}-\mathbf{y}\|_2^2 & \text{if } l_x = l_y \\ \max(0,(p-\|\mathbf{x}-\mathbf{y}\|_2))^2 & \text{otherwise} \end{cases} \quad (1)$$

106. Coupled Deep Learning for Heterogeneous Face Recognition

主要思想：

- 本文通过将 low-rank relevance constraint 和 cross modal ranking 引入到 CNN 中，提出了一种面向异构人脸识别的耦合深度学习(CDL)框架。
- low-rank relevance constraint 作为正则化器，在完全连通层上提出了一种新的迹范数，它不仅可以增强不同模式之间的相关性，而且还可以约束参数空间，特别是对于少量未配对的异源样本，可以减少过拟合。
- cross modal ranking 进一步提高了 CDL 的分辨能力，且有效地扩大训练数据，利用有限数量的异质样本之间的信息。

主要方法：

- Relevance Constraint

$$\mathcal{J}_{\text{relevance}}(X_i, W_i) = \sum_{i \in N,V} \text{softmax}(X_i, W_i) + \lambda \| [W_N \ W_V] \|_* \quad (2)$$

其中，$\|W\|_*$是迹范数。经过推导可得梯度公式为：

$$\frac{\partial \mathcal{J}_{\text{relevance}}}{\partial W_N} = \frac{\partial \text{softmax}(X_i, W_N)}{\partial W_N} + W_N(\Gamma^{-1} + (\Gamma^{-1})^T)$$
$$\frac{\partial \mathcal{J}_{\text{relevance}}}{\partial W_V} = \frac{\partial \text{softmax}(X_i, W_V)}{\partial W_V} + W_V(\Gamma^{-1} + (\Gamma^{-1})^T) \quad (8)$$

$$\Gamma = (W_N W_N^T + W_V W_V^T + \mu \mathrm{I})^{\frac{1}{2}} \quad (7)$$

- Cross Modal Ranking

$$\mathcal{J}_{\text{ranking}} = \sum_{i=0}^{N} \max(0, m + \|x_i^a - x_i^p\|^2 - \|x_i^a - x_i^n\|^2) \quad (14)$$

锚样本x_a^i和负样本x_n^i是从相同的模态中选择的，正样本x_p^i的模式不同于锚和负样本。

一个简单的三重态集很容易满足方程中的三重态约束，对训练贡献较少，使网络收敛缓慢。因此，采用半硬三重态选择方法进一步提高了交叉模态匹配性能。样本选择原则如下：

$$\begin{cases} \|x_i^a - x_i^p\|^2 + m > \|x_i^a - x_i^n\|^2 \\ \|x_i^a - x_i^p\|^2 < \|x_i^a - x_i^n\|^2 \\ x_i^a, x_i^n \in X_N(X_V),\ x_i^p \in X_V(X_N) \end{cases} \quad (15)$$

4) 3D 人脸

107. Deep 3D Face Identification

主要思想：

- 从接受过 2D 人脸图像的 CNN 转移学习可以通过用相对少量的 3D 面部扫描来完成 CNN 来有效地处理 3D 面部识别。
- 还提出了一种 3D 面部增强技术，其从单个 3D 面部扫描合成多个不同的面部表情。

主要方法：

① 应对有限数量的可用 3D 数据具有挑战性。利用现有的网络，训练为 2d 人脸识别，并利用少量的 3D 扫描微调他们，以执行三维到三维表面匹配。

② Expression Generation：对于人脸，表情的变化会影响三维结构，并会降低识别系统的性能。为了解决这一问题，我们提出用综合的三维人脸数据来扩充我们的三维人脸数据库，并考虑到面部表情。为了增加训练数据，我们使用了多线性三维形态模型，其中形状来自于 Basel 人脸模型，表情来自于 FaceWarehouse。

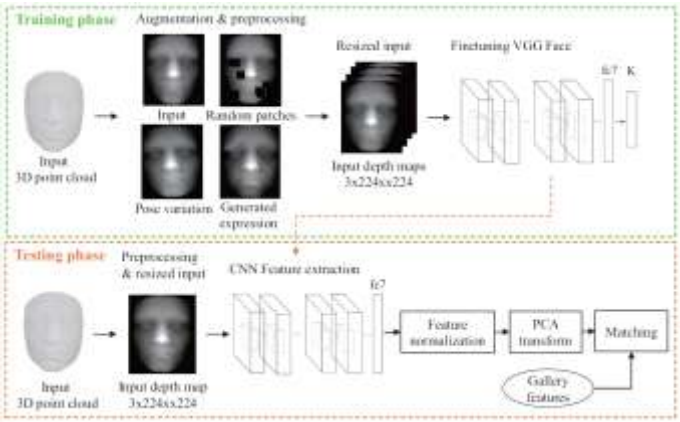
$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{P}_s\alpha + \mathbf{P}_e\beta,$$

③ 通过随机变化的参数值来表达的 3DMM 拟合 3dmm 添加随机表达式：$(-0.05 < \beta_i < 0.05)$。为了使合成的表情更加平滑，建议使用三维形态模型来计算原始点云与表达式增强点云之间的形变场。

④ Pose Variations：我们简单地对三维点云应用随机生成的刚性变换矩阵$(\mathbf{M} = [\mathbf{R}\ \mathbf{t})$。其中 R 是由不同过的 yaw, pitch and row rotations$(\mathbf{R} = \mathbf{R}_z(\theta_3)\mathbf{R}_y(\theta_2)\mathbf{R}_x(\theta_1))$生成的$(-10^\circ < \theta_1, \theta_2, \theta_3 < 10^\circ)$。$\mathbf{t} = [x, y, z]^T$，-10<x, y, z<10。

⑤ 为了将我们的 3D 数据传递给 2d 训练的 cnn，将三维扫描进行正面化，生成 2.5D 深度图，提取深度特征来表示三维表面，并匹配特征向量进行三维人脸识别。

⑥ 为了使我们的系统对小的对准误差有鲁棒性，每个三维形状都通过刚性变换来增强：3D 旋转和投影前的平移。还在三维数据中添加了一些随机补丁，以模拟随机遮挡。



108. Learning from Millions of 3D Scans for Large-scale 3D Face Recognition

主要思想：

- 训练数据：提出了一种生成大量标记三维人脸数据的方法，用于 CNN 的训练。我们的数据集包含 310 万个三维扫描，100 k 身份包含高度丰富的形状变化。我们的训练数据不包括公共数据集。
- 大规模测试数据：将现有最具挑战性的公共三维人脸数据集进行合并，提出了一种基于图库中单个特征样本的大规模人脸识别协议。测试数据包含 31860 个对 1853 个身份的三维扫描。据我们所知，这是最大的三维人脸库的大小，其中的人脸识别结果已经有报道。

- 深度三维人脸识别网络(Fr3dnet)：我们提出了第一个专门针对三维人脸识别的深度 CNN，训练对象为 310 万张 3D 人脸，然后在 1853 个 gallery 上进行微调。

Modality	Model \ Technique	Input Size	Training IDs	Training Scans	Testing IDs	Testing Scans	Testing Dataset	NW Param
2D	VGG-Face [45]	224 × 224	2.6K	2.6M	5K	13K	LFW	134M
	DeepFace [58]	152 × 152	4K	4.4M	5K	13K	LFW	120M
	FaceNet [53]	220 × 220	8M	200M	5K	13K	LFW	140M
	MF2 [42]	-	672K	4.7M	690K	1M	MegaFace	-
3D	MMH [35]	-	-	-	0.46K	4K	FRGCv2	-
	K3DM [34]	-	-	-	0.46K	4K	FRGCv2	-
	Kim *et al.* [29]	224 × 224	0.7K	123K	0.1K	4.6K	Bosphorus	140M
3D	**FR3DNet**	160 × 160	100K	3.1M	1.85K	31K	LS3DFace	**29M**

主要方法：

- Proposed Data Generation for Training

① Dense correspondence model

使用基于关键点的算法，在该数据集的人脸上建立 15k 个三维顶点上的密集对应关系。为了确保该对的身份尽可能“不同”，我们选择了最大非刚性形状差的人脸对。

$$\mathbf{D}(i,j)=\frac{\gamma_{ij}+\gamma_{ji}}{2}$$

其中$\gamma_{i,j}$是变形三维面从F_i到F_j所需的弯曲能量。$\gamma(i,j)=\mathbf{x}^T\mathcal{B}\mathbf{x}+\mathbf{y}^T\mathcal{B}\mathbf{y}+\mathbf{z}^T\mathcal{B}\mathbf{z}$

从可能的角度选择了 90100 对最大形状差 D(i,j)的三维面。由于每一对中的三维人脸彼此稠密对应，所以每一对（i，j）的线性空间生成一个新的面：$\bar{F}=\frac{[x_i^p,y_i^p,z_i^p]^T+[x_j^p,y_j^p,z_j^p]^T}{2}$。

② Synthetic

训练数据的第二个 3D 人脸来源是一个商业软件，它生成不同面部形状、种族和表情的密集对应的面孔。我们生成 300 个身份，每一个在四个不同的表达式中有三个强度级别，并遵循上面的协议，从 44850 对中创建 9950 个新的身份。然而，在这种情况下，我们选择了“相似”且人与人之间距离较小的对。

Table 2. Details of the dataset generated for training *FR3DNet*.

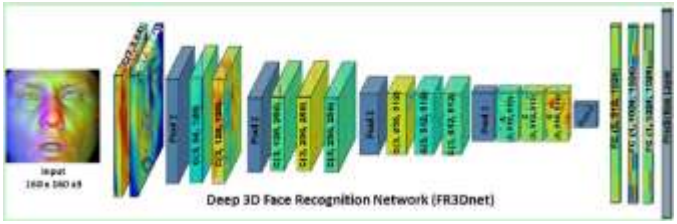
Type	IDs	Expressions	Poses	Total Scans
Dense Correspondence Model	90,100	2	15	*1,680,900
Real 3D Faces	1,785	1	15	26,775
Synthetic	8,120	12	15	1,461,600
Total	**100,005**	**12**	**15**	**3,169,275**

*Randomly selected from 2,703,000 scans.

最后，我们通过在三维人脸前面的一个半球上部署 15 个合成摄像机，模拟了每次 3D 扫描中的姿态变化和大遮挡。摄像机被部署在。经度[-90；90]关于和[-30,30]的纬度上；每隔 15 取一次。应用隐点去除算法，从摄像机视点中去除自遮挡的三维点

- FR3DNet

训练数据中每个扫描的三维点云被用来生成一个三通道图像。第一个通道是深度图像，它是通过使用网格匹配算法将 z(x；y)形式的曲面拟合到三维点云生成的。原始点云的表面法线是在球面坐标(θ；φ)中计算的，其中θ 和 φ是法向量的方位角和仰角。将θ(x；y)和φ(x；y)形状的表面拟合成方位角和仰角，使我们用来训练网络的三维图像表示的第二和第三通道。



我们的目标是通过使softmax层后的the average predictionlog-loss最小化，学习设计的网络参数，将N=100005进行分类。在网络训练

之后，我们删除掉softmax层。来自fc7的长度1024的嵌入特征向量可以通过最小化特征空间中probe与gallery之间的余弦距离来进行人脸识别。同时还在测试数据gallery的库上进行微调，并将其表示为FR3DNetft。

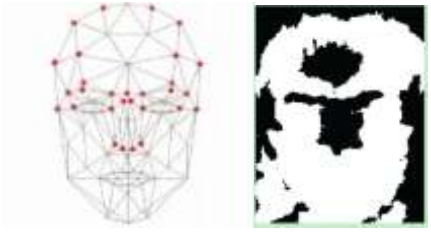
109. Research of 3D Face Recognition Algorithm Based on Deep Learning StackedDenoising Autoencoder Theory

主要思想：

- 首先，从 Candide-3 人脸模型中提取出 30 个特征点来表征人脸差异，提高训练效率。在对人脸三维数据进行初步分析后，利用 sdae 网络对人脸深度数据进行无监督的预训练，并进行有监督的微调。

主要方法：

脸由深方向的非平坦区和平坦区组成，非平坦区包括口、鼻、眼、眉毛、头盖骨等，扁平区包括额头、颧骨等。113 个点的 Candide-3 面模型基本覆盖平坦和非平坦区域。在人脸识别过程中起着关键作用的是非平坦区域。从 113 个点中选取 30 点，主要分布在非平坦区和颅骨边缘，如鼻尖端部、鼻梁、眼角、眉脊等。对于嘴唇上的特征点，由于每个人的面部深度和位置信息在转换表达或说话时变化很大，且变化不规则，我们将这些点视为噪声，以避免过度拟合。30 个特征点如下：



首先，通过空间旋转变换，将人脸姿态转换为 Xoy 平面的正方向。然后进行深度量化的过程。点与 xoy 平面之间的距离是深度值。采用深度图像，利用编解码器重构损失进行无监督预训练，然后利用数据标签进行微调。

$$E=\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|h_{w,b}(x_i(k))-x_i(k)\right\|^2\right) \quad (7)$$

5） 低分辨率

110. Low Resolution Face Recognition Using aTwo-Branch Deep Convolutional Neural NetworkArchitecture

主要思想：

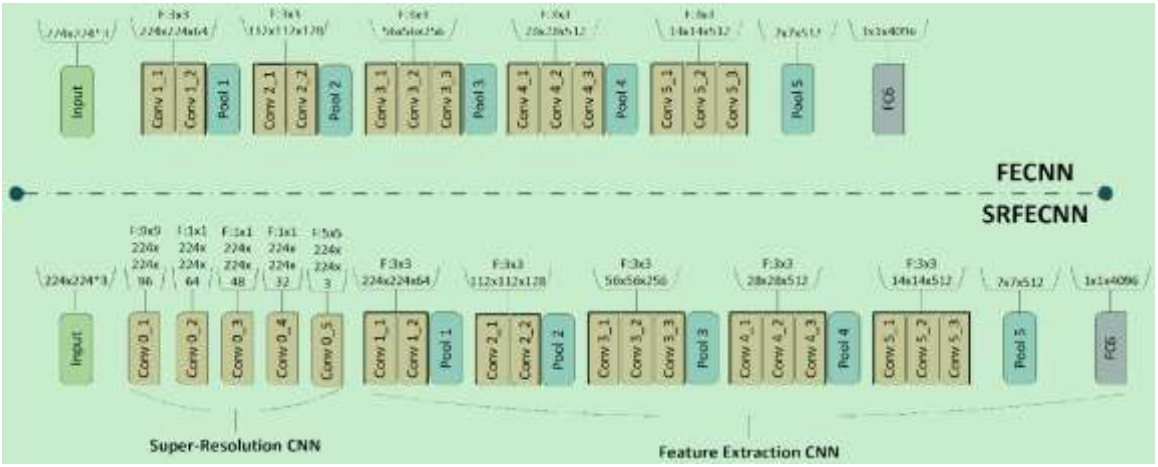
- 该体系结构由两个 Dcnns 分支组成，用于将高分辨率和低分辨率人脸图像映射到具有非线性变换的公共空间中。
- 与高分辨率图像转换对应的分支由 14 层组成，另一分支将低分辨率人脸图像映射到公共空间，包括连接到 14 层映射网络和 5 层超分辨率网络。

主要方法：

- Networks Architecture

有两个分支结构，一个是将高分辨率图像投影到公共空间，另一个是将低分辨率图像映射到这个公共空间。采用 vggnet 结构，有十六层，十三层卷积层和三层完全连接层。去掉了这个 vggnet 中最后两个完全连接的层，并称之为特征提取卷积神经网络(Fecnn)。输入图像为 224 维的高分辨率图像(当输入图像大小不同于 224 维时，我们使用传统的双三次插值方法获得所需的尺寸)。最后一层的输出是一个包含 4096 个元素的特征向量。

底部分支第一个子网是超分辨率网(SRNET)，用于生成高清图像。第一子网的输出被输入第二子网 FECNN，用于将图像映射到公共子空间，两个合成 SRFECNN。结构如下：

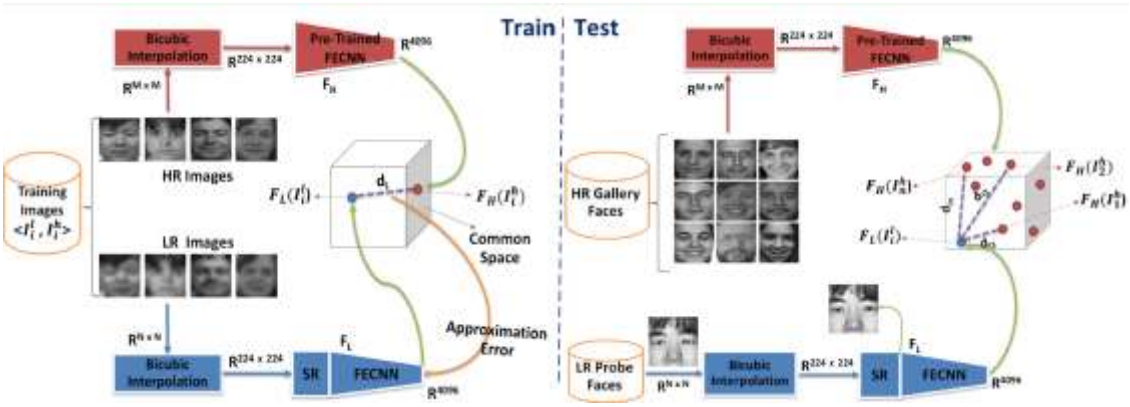


- Common Subspace Learning

① 在人脸数据集上使用经过训练的 vggnet，然后删除最后两个完全连接的层，并将其用于我们架构的顶部和底部分支。

② 在第二步中，我们使用高分辨率和低分辨率人脸图像对的数据库来训练底部分支的 SRnet。

③ 第三步是主要训练阶段。我们合并了两个子网，即 SRnet 和 FEcnn，并将一个包含相同人的低分辨率和高分辨率对的训练数据库分别输入到底层和顶部的分支中。



6) 移动端

111. Learning a Metric Embeddingfor Face Recognitionusing the Multibatch Method

主要思想：

- 减少训练时间
- 采用度量学习

主要步骤：

- Learn a Metric

整个网络学习一个从输入 x 到输出 fw(x)∈Rd 的一个映射，学习规则如下：

$$\begin{aligned} y = y' \quad &\Longrightarrow \quad \|f_w(x) - f_w(x')\|^2 < \theta - 1 \\ y \neq y' \quad &\Longrightarrow \quad \|f_w(x) - f_w(x')\|^2 > \theta + 1 \end{aligned}$$

现在，我们先据此定义 one pair 的 Loss，其中训练集定义为：

$$l(w, \theta;\, x_i, x_j, y_{ij}) = \left(1 - y_{ij}\left(\theta - \|f_w(x_i) - f_w(x_j)\|^2\right)\right)_+$$

其中，yij∈±1,+1 表示 xi 和 xj 属于同一身份，另外 $(u)_+ := \max(u, 0)$。

则整体的 Loss 为：

$$\mathrm{L}(w, \theta) = \frac{1}{m^2 - m} \sum_{i \neq j \in [m]} l(w, \theta;\, x_i, x_j, y_{ij}) .$$

另一方面，文章从理论和实验上证明了上面的 Loss 比 hinge-loss 或者 softmax-loss 等多分类 Loss 要更难收敛。

所以，Google 的 Facenet 中有很大一部分工作就是在如何选择和设计 Triplet 三元组，因此本文后面就参考该思想设计了新的训练方法。

- The Multi-Batch Estimator（没看明白）

（全文思想）The Multibatch method firstgenerates signatures for a mini-batch of k face images and then constructsan unbiased estimate of the full gradient by relying on all k2-k pairs fromthe mini-batch.

（网上微博解释，但觉得不太靠谱）这一部分在原文中占比挺多，可惜感觉完全在水，总之实现的方法就是：

假如 batch_size=K,那么两两配对的话总共会有 K×(K－1)种可能。(不过实际程序实现的时候,应该只有 K×(K－1)2,因为(xi,xj)或者(xj,xi)在 BP 时是一样的)。

于是，我们实际每一个 batch 中都进行类似遍历，最后将所有 pair 的 loss 加和即为一个 batch 的 Loss。

原文中作者的实验参数配置如下：batch_size=256,共 16 个人每人 16 张图片；训练图片 2.6M；模型大小为 1.3M；输入图像为 112x112 的 RGB 图像，编码长度 128Bit；学习率固定 0.01，最后一个 epoch 降为 0.001。

7) 跨年龄

112. Latent Factor Guided Convolutional Neural Networks for Age-Invariant FaceRecognition

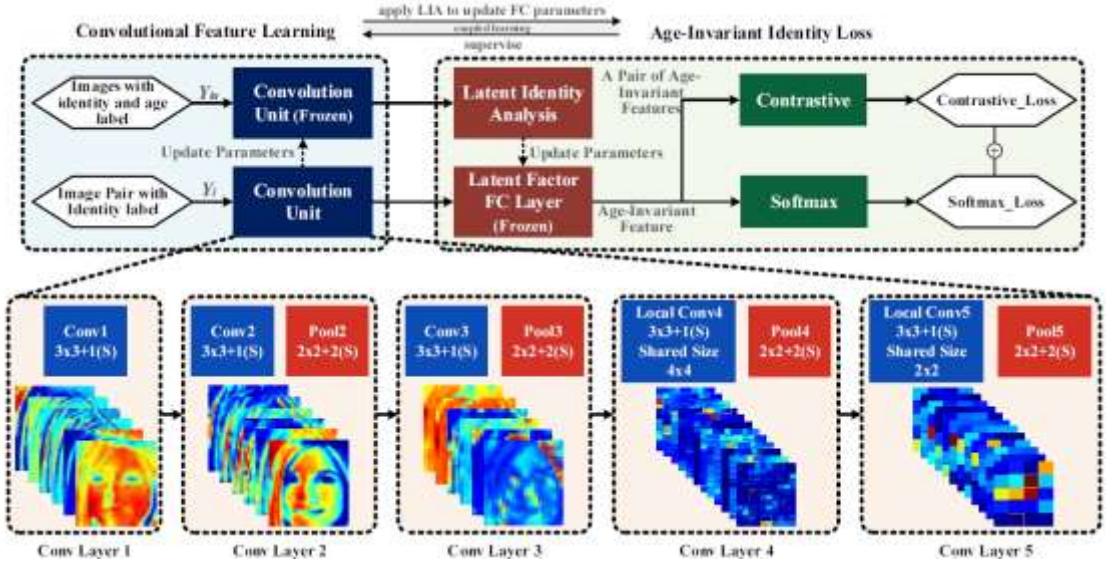
主要思想：

- 这是首次证明深度 CNNs 在跨年龄识别中的有效性，并取得了迄今为止最好的效果。
- 相比于深度学习模型直接应用于跨年龄识别，构建了一个潜在的身份分析模块，用于指导 CNN 参数的学习。该模型通过对 CNNS 参数和 RIA 参数的耦合学习，提取出适合于 AIR 任务的年龄不变的深层人脸特征。

主要步骤：

该架构称为 LF-CNN，通过精心设计的全连通层(LF-FC)从卷积特征中提取年龄不变的深层特征。为此，我们发展了一个潜在变量模型，称为 LIA，将年龄特征与卷积特征中的身份相关成分分离开来。利用 LIA 模型的参数更新 LF-FC 层的参数。此外，LIA 模型和 cnn 中的损失函数构成了年龄不变的身份损失，用于指导 lf-cnn 的学习。

卷积单元映射的原始输入图像 Fimg 到卷积特征 Fconv,Fconv=f(Fimg),然后如果 LF-FC 层计算年龄不变特征 Ffc,$F_{fc} = g(F_{conv})$,然后年龄不变特征用于识别。



- The LF-CNNsModel

在 lf-cnn 中，卷积单元的结构遵循典型的 cnn，交替叠加卷积层、非线性层和 pooling 层。

lf-FC 层的构造：FC 层等价于矩阵乘法：$F_{fc} = WF_{conv} + b$，其中F_{fc}是 FC 层的输出，Fconv 是卷积特征，w，b 是 FC 层的参数。我们利用这种等价性设计了一组 w，b，可以从 fconv 中提取年龄不变特征。与其用随机梯度下降(Sgd)迭代更新 lf-cnn 中的所有参数，我们还设计了一种潜在恒等式分析(LIA)方法来学习 lf-cnn 模型的 w 和 b。

- Latent Identity Analysis

根据不同的监督信号，将每个人脸特征看作是不同分量的组合。在跨年龄人脸识别中，我们通常将人脸特征分解为两个潜在的分量和一个噪声变量。$v = U_{id}x_{id} + U_{ag}x_{ag} + U_e x_e + \bar{v}$其中，$x_{id}$和$x_{ag}$满足标准高斯分布$N(0, I)$，$x_e$满足$N(0, \sigma^2 I)$。

通过 EM 算法学习参数$\theta=\{U_{id}, U_{ag}, \sigma^2, \bar{v}\}$，得到

$$
\begin{aligned}
U_{id} &= (C - DB^{-1}E)(A - FB^{-1}E)^{-1} \\
U_{ag} &= (D - CA^{-1}F)(B - EA^{-1}F)^{-1} \\
\sigma^2 &= \frac{1}{Nn}\sum_{i,j}\{(v_i^j - \bar{v} - U_{id}\mu_1(x_{id,i}) \\
&\quad - U_{ag}\mu_1(x_{ag,j}))^T(v_i^j - \bar{v})\}
\end{aligned} \tag{11}
$$

in which

$$
A=\sum_{ij}\mu_2(x_{id,i}, x_{id,i}),\ B=\sum_{i,j}\mu_2(x_{ag,j}, x_{ag,j}),
$$

$$
C=\sum_{ij}(v_i^j-\bar{v})(\mu_1(x_{id,i}))^T,\ D=\sum_{ij}(v_i^j-\bar{v})(\mu_1(x_{ag,j}))^T,
$$

$$
E=\sum_{ij}\mu_2(x_{ag,j}, x_{id,i}),\ F=\sum_{i,j}\mu_2(x_{id,i}, x_{ag,j}).
$$

通过参数θ计算可得 W，b，其中特征 v 来自卷积单元（相当于 Fconv）：

$$
W = U_{id}^T\Sigma^{-1},\ b = -U_{id}^T\Sigma^{-1}\bar{v} \tag{13}
$$

- Learning parameters for LF-FC layer

  我们使用具有年龄和身份标签的训练数据 Yia 来训练 LF-FC 层。具体来说，将卷积特征 Ffc 作为 LIA 中的观测特征 v。LIA 模型学习参数$\theta=\{U_{id}, U_{ag}, \sigma^2, \bar{v}\}$，然后计算 W 和 b。

- Learning parameters for convolution unit

  固定 lf-FC 层的参数 g。然后用只包含身份信息的数据对 lf-CNN 进行训练，采用 SGD。请注意，我们同时使用了 Softmax 损失和对比损失来加强对学习的监督。

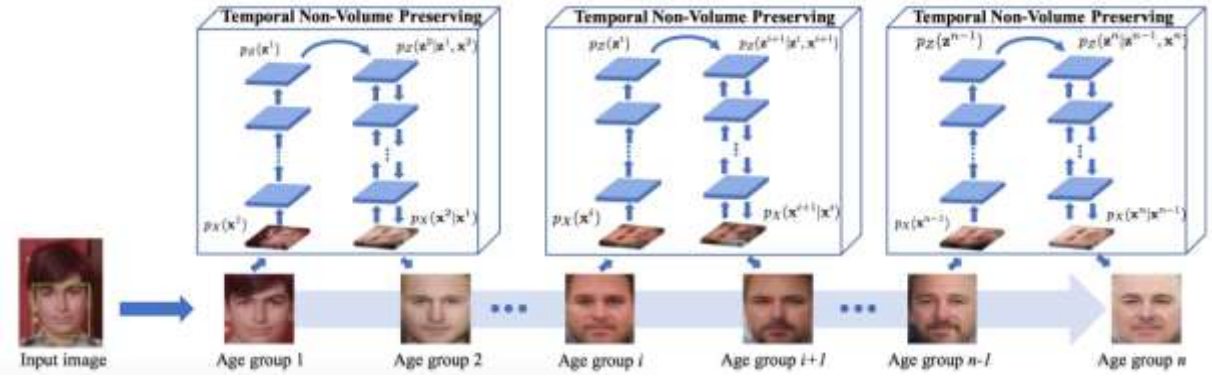
113. Temporal Non-Volume Preserving Approach to Facial Age-Progression and Age-Invariant Face Recognition

主要思想：

- 第一阶段将老化过程分解为多个短期阶段。然后，提出了一种新的生成概率模型--Temporal Non-Volume Preserving (TNVP) transformation，用于模拟面部各阶段的老化过程。
- 该模型不仅在捕捉各个阶段的非线性年龄相关方差方面具有优势，而且在人脸间的年龄推进过程中也产生了平滑的合成。
- 利用概率图形模型的优点，避免规则重构损失函数，从而产生更好的综合质量，以及深度残差网络(ResNet)改进了高度非线性特征的生成。

主要步骤：

体系结构包括三个主要步骤。(1) 预处理；(2) 通过映射函数进行人脸变异建模；(3) 老化变换嵌入。对映射函数的结构，我们的 tnvp 模型是高度非线性的。这是优化利用对数似然函数，产生更清晰的重建模型规范。



- Preprocessing

给定一幅图像后，根据四个地标点(即两眼和两个嘴角)的对应位置，简单地检测和排列面部区域。通过避免复杂的预处理步骤，我们提出的体系结构有两个优点。

- Face Aging Modeling

设$\mathcal{I}\subset\mathbb{R}^D$为图像域，$\{\mathbf{x}^t, \mathbf{x}^{t-1}\}\in\mathcal{I}$分别为编码 t 和 t-1 年龄组人脸图像纹理的变量。为了嵌入这些人脸之间的老化变换，我们建立了从图像空间 I 到隐空间 Z 的双射映射函数，并对这些潜在变量之间的关系进行了建模。形式上，让$F:\mathcal{I}\to\mathcal{Z}$从观测变量 x 到对应的潜变量 z 的双射。$G:\mathcal{Z}\to\mathcal{Z}$是一个老化变换函数，它模拟了潜空间中变量之间的关系。

$$\mathbf{z}^{t-1} = \mathcal{F}_1(\mathbf{x}^{t-1};\theta_1)$$
$$\mathbf{z}^{t} = \mathcal{H}(\mathbf{z}^{t-1},\mathbf{x}^{t};\theta_2,\theta_3)$$
$$= \mathcal{G}(\mathbf{z}^{t-1};\theta_3) + \mathcal{F}_2(\mathbf{x}^{t};\theta_2)$$

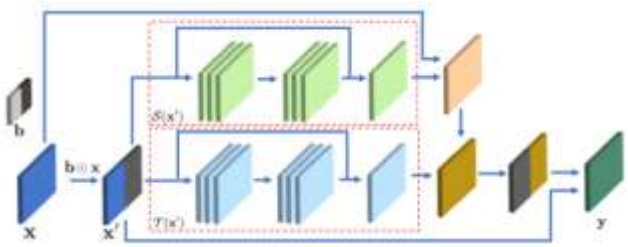
H 表示 G(zt-1; θ3)和 F2(xt; θ2)之和, θ={θ1, θ2, θ3}分别表示函数 F1、F2 和 G 的参数。

- Mapping function as CNN layers

给定一个输入 x，一个单位$f:\mathbf{x}\rightarrow\mathbf{y}$从 x 到中间潜伏状态 y 的映射

$$\mathbf{y} = \mathbf{x}' + (1-\mathbf{b})\odot[\mathbf{x}\odot\exp(\mathcal{S}(\mathbf{x}')) + \mathcal{T}(\mathbf{x}')]$$

其中，$\mathbf{x}' = \mathbf{b}\odot\mathbf{x}$; ⊙表示 Hadamard 乘积；$\mathbf{b} = [1,\cdots,1,0,\cdots,0]$是一个二进制掩码，b 的前 d 个元素设为 1，其余的为 0，S 和 T 分别表示 scale 函数和平移函数。其中 s 和 t 分别采用残差网络。



$$\left|\frac{\partial f}{\partial \mathbf{x}}\right| = \prod_j \exp(s_j) = \exp\left(\sum_j s_j\right)$$

反函数为：

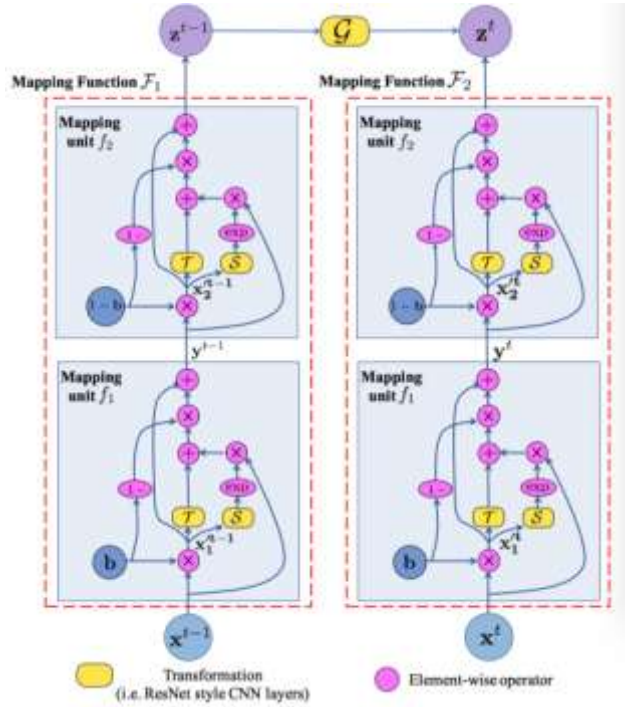
$$\mathbf{x} = \mathbf{y}' + (1-\mathbf{b})\odot[(\mathbf{y} - \mathcal{T}(\mathbf{y}'))\odot\exp(-\mathcal{S}(\mathbf{y}'))] \quad (6)$$

- Mapping function

双射映射函数 f 是制定通过组合序列$\mathcal{F} = f_1\circ f_2\circ\cdots\circ f_n$

$$\left|\frac{\partial \mathcal{F}}{\partial \mathbf{x}}\right| = \left|\frac{\partial f_1}{\partial \mathbf{x}}\right|\cdot\left|\frac{\partial f_2}{\partial f_1}\right|\cdots\left|\frac{\partial f_n}{\partial f_{n-1}}\right|$$
$$\mathcal{F}^{-1} = (f_1\circ f_2\circ\cdots\circ f_n)^{-1} = f_1^{-1}\circ f_2^{-1}\circ\cdots\circ f_n^{-1}$$

我们也可以在序列中将二进制掩码 b 更改为 1-b，这样 x 的每个组件都可以通过映射过程连接起来。



- The aging transform embedding

$$\mathcal{G}(\mathbf{z}^{t-1};\theta_3) = \mathbf{W}\mathbf{z}^{t-1} + \mathbf{b}_{\mathcal{G}}$$
$$\mathbf{z}^{t-1} \sim \mathcal{N}(0,\mathbb{I})$$
$$\mathcal{F}_2(\mathbf{x}^t,\theta_2) = \bar{\mathbf{z}}^t \sim \mathcal{N}(0,\mathbb{I})$$
$$p_{Z^t,Z^{t-1}}(\mathbf{z}^t,\mathbf{z}^{t-1};\theta) \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{b}_{\mathcal{G}}\\ 0\end{bmatrix},\begin{bmatrix}\mathbf{W}^T\mathbf{W}+\mathbb{I} & \mathbf{W}\\ \mathbf{W} & \mathbb{I}\end{bmatrix}\right)$$

则总优化函数为：

$$\theta_1^*, \theta_2^*, \theta_3^* = \arg\max_{\theta_1,\theta_2,\theta_3} \log p_{X^t}(\mathbf{x}^t|\mathbf{x}^{t-1};\theta_1,\theta_2,\theta_3) \quad (12)$$

From Eqn. (2), the log-likelihood can be computed as

$$\begin{aligned}\log p_{X^t}(\mathbf{x}^t|\mathbf{x}^{t-1};\theta) &= \log p_{Z^t}(\mathbf{z}^t|\mathbf{z}^{t-1},\theta) + \log\left|\frac{\partial \mathcal{F}_2(\mathbf{x}^t;\theta_2)}{\partial \mathbf{x}^t}\right| \\ &= \log p_{Z^t,Z^{t-1}}(\mathbf{z}^t,\mathbf{z}^{t-1};\theta) \\ &\quad - \log p_{Z^{t-1}}(\mathbf{z}^{t-1};\theta_1) + \log\left|\frac{\partial \mathcal{F}_2(\mathbf{x}^t;\theta_2)}{\partial \mathbf{x}^t}\right|\end{aligned}$$

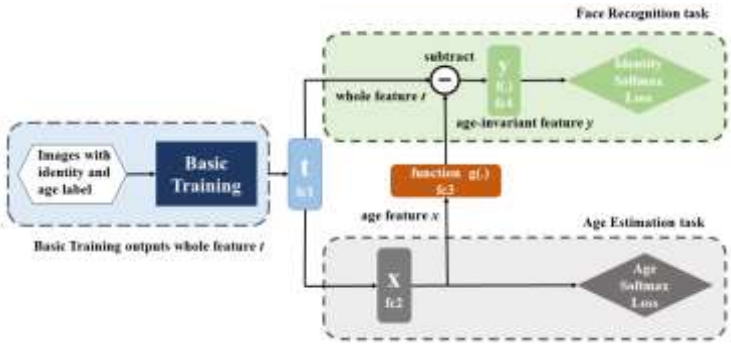
114. Age Estimation Guided Convolutional Neural Network for Age-Invariant FaceRecognition

主要思想：

- 提出了一种新的深度人脸识别网络- age estimation guided convolutional neural network (AE-CNN)，将老化引起的变化与人特有的稳定特征分离开来。
- 考虑到人脸识别任务获得的特征往往包含年龄相关因素，直接获取特定人特征是很困难的，因此我们增加了年龄估计任务来获得年龄特征，并从整个特征中减去年龄因子。

主要步骤：

我们使用的配方为 y=f(t-g(x))，t 为整个特征，其中 x 是在年龄估计任务中获得的年龄特征,y 是用于年龄不变的身份特定特征，面部识别，g(.)是获得年龄因素的函数，该年龄因素使面部识别的性能下降。F（）是功能可更好地处理整个功能、年龄特征和身份特定功能之间的关系。年龄估计任务和人脸识别任务同时更新网络中的参数。



- The AE=CNNFramework

该算法由卷积层和完全连通层组成，以获得年龄不变特征。以 4 层卷积网络的 light CNN 为基础网络，全连接层 fc1 输出脸部的全部特征，fc2 输出年龄特征。Fc3 起到特征函数 g（）的作用，并且从 fc1 的全部特征中减去年龄因素，最后 fc4 起到了 f（）函数的作用，输出身份特定的特征用于人脸识别

$$\begin{aligned} y &= f(t - g(x)), \\ f(x) &= W_1 x + b_1, \\ g(x) &= W_2 x + b_2, \end{aligned}$$

- Basic Training:

先用没有年龄估计得网络进行身份识别，得到 fc1

$$L_1 = -log(\frac{e^{t_i^y}}{\sum_{j=1}^{m} e^{t_i^j}}),$$

- Separation

然后将人脸识别和年龄估计任务一起训练，得到 fc4

$$L_2 = -log(\frac{e^{t_i^y}}{\sum_{j=1}^{m} e^{t_i^j}}) - \alpha(log(\frac{e^{a_i^x}}{\sum_{k=1}^{n} e^{a_i^k}})), \quad (5)$$

其中第一项是人脸识别任务，第二项是年龄估计任务，α是年龄估计任务的权重。

115. A Deep Joint Learning Approachfor Age Invariant Face Verification

主要思想：

- 训练一个深度卷积网络来同时学习特征、距离度量和阈值函数。
- 我们的目的不仅在于保持同一人在不同年龄间的相似性，同时区分不同的个体，而且同时学习隐式自适应阈值

主要步骤：

- Optimization Objective

  如果在匹配问题中使用度量学习，则需要一个阈值来决定 x 和 y 是否匹配。我们将其表述如下

$$(x-y)^t M(x-y) \le d, \quad M \succeq 0.$$

但是，对于固定的阈值 D 来说，这是不合适的，因为类内的距离可能比类间的距离要大建议自适应学习阈值，d 是与(x，y)相关的函数，而不是常数。

$$f(x,y) = d(x,y) - (x-y)^t M(x-y) \begin{cases} \ge 0 & \text{if } c(x) = c(y) \\ < 0 & \text{otherwise} \end{cases}. \qquad (2)$$

假设 d 为 $d(x,y) = \frac{1}{2}x^t \bar{A}x + \frac{1}{2}y^t \bar{A}y + x^t \bar{B}y + c^t(x+y) + b.$，则

$$\begin{aligned} f(x,y) &= \frac{1}{2}x^t L_A^T L_A x + \frac{1}{2}y^t L_A^T L_A y - x^t L_B^T L_B y + c^t(x+y) + b \\ &= \frac{1}{2}(L_A x)^t (L_A x) + \frac{1}{2}(L_A y)^t (L_A y) - (L_B x)^t (L_B y) + c^t x + c^t y + b \end{aligned} \qquad (5)$$

通过上述变换，将人脸识别转化为计算上述决策函数，对于一个人 p 的年龄实例 z，我们希望学习一个重新识别模型，能够成功地识别出同一人的另一个年龄实例 z’。这可以通过学习度量 la、lb 和向量 c 来实现，而 f(z，z)的值对于同一个人尽可能大，而对于不同的人则尽可能小。我们将$(L_A, L_B, c)^T$设为 W，$\Omega = \{\Omega_k = (z_i, z_j)\}$若两个为同一个人，则$l(\Omega_k) = -1$则 hingle loss 为

$$H(W) = \sum_{\Omega} \max\{0, l(\Omega_k) \times f(\Omega_k) + 1\} \quad k = 1, 2 \ldots N^2$$

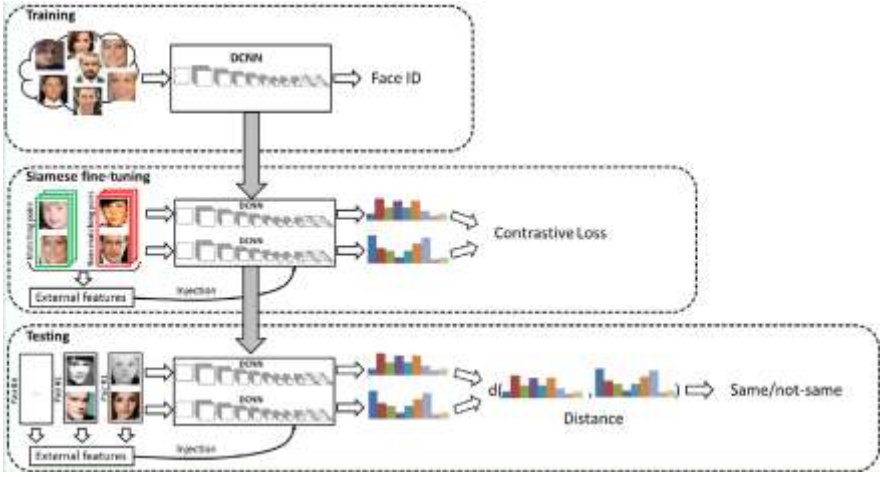
116. Large Age-Gap face verification by feature injection in deep networks

主要思想：

- 该方法利用了对大数据集上人脸识别任务进行预训练的深卷积神经网络(Dcnn)，并对大年龄间隔的人脸验证任务进行了微调。微调是在 Siamese architecture 中使用对比损失函数进行的。
- 进一步提高了网络的识别能力，包括特征注入层，该特征注入层将外部计算的特征注入到 dcnn 的最深层。

主要步骤：

第一，人脸和地标检测是在 CASIA-WebFace 和 Large Age-Gap (LAG)数据库上进行的。接下来，dcnn 在 CASIAWebFace 数据库上进行人脸识别预训练。然后，dcnn 学习的知识从人脸识别的源任务转移到大年龄间隔的人脸验证的目标任务，通过微调实现知识边缘转移，在一个 Siamese architecture 中使用对比损失，在完全连接的层中注入预计算的外部特征，从而实现 dcnn 的知识转移。

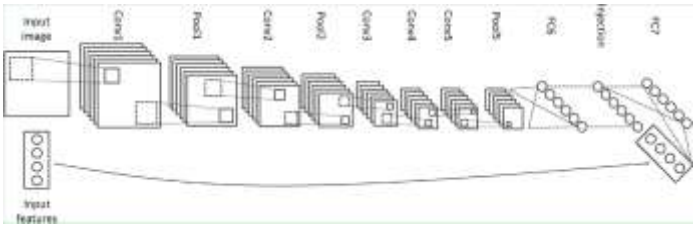


- Deep face feature representation

架构是 alexnet。dcnn 在 CASIAWebFace 数据库上进行人脸识别预训练。输入层的维数为 200×200×1 灰度图像.。网络包括 5 个卷积层、3 个池层和 3 个完全连通层.。每个卷积层后面跟着一个经校正的线性单元(Relu)。从第二层到最后一层全连通层(FC 7)提取的特征，经过 L2 归一化步骤后用于人脸表示。

- Feature injection

L2 归一化 FC 7 特征可以作为一系列人脸验证方法的输入。它们中的每一个作为输出，都提供了一对人脸图像同一性或不同一性的距离或信心分数，$d_i, i = 1, \ldots, n,$，将其堆叠成向量 d。除了 dcnn 特征外，还添加了一个特征注入层，以将外部计算的特征 d 与 dcnn 最深层的激活相融合。特征 d 被注入到第一个完全连接的层中。特征注入以外部特征 d 与 FC 6 激活的级联形式执行，



117. Face Aging With Conditional Generative Adversarial Networks

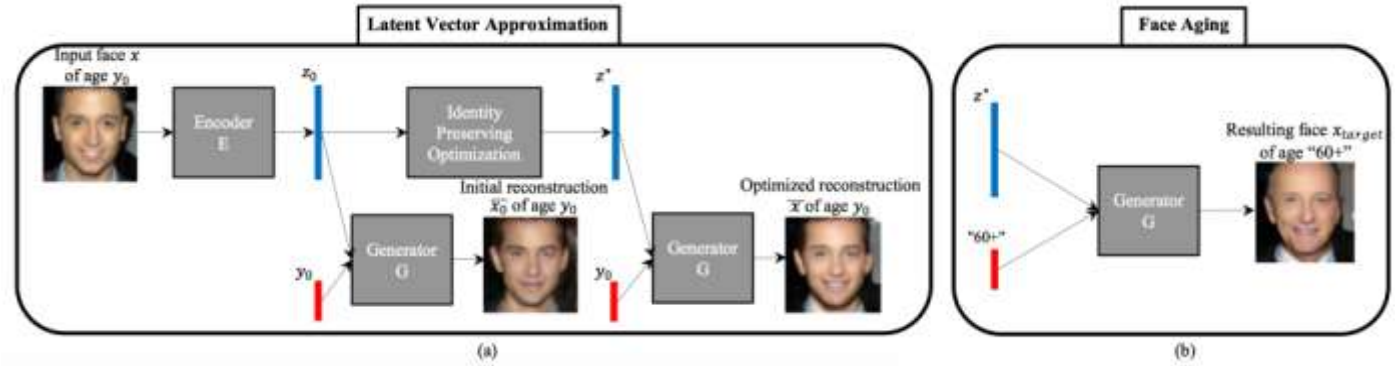
主要思想：

- 提出了一种 Age Conditional Generative Adversarial Network (acGAN)用于合成人脸衰老图片。
- 该方法由两个步骤组成：(1)输入面重构，并得到最优的潜在向量z^*；(2)在生成器输入时，通过条件 y 的简单改变来实现人脸老化。
- 提出一种新的“保持身份”的潜在向量优化方法，允许在重构过程中保留原人的身份。

主要步骤：

① 给出年龄为 y0 的输入人脸 x，寻找最优的潜向量z^*，使重构的人脸$\tilde{x} = G(z^*, y_0)$尽可能接近初始人脸

② 给定目标年龄 y 目标，通过简单切换生成器输入时的年龄，生成目标图像 x 目标$x_{target} = G(z^*, y_{target})$



- Age Conditional Generative Adversarial Network

条件 GaN(cGAN)扩展了 GaN 模型，允许生成具有某些属性的图像(“条件”)。在实际应用中，条件 y∈Ny 可以是与目标人脸图像相关的任何信息：光照水平、人脸姿态或面部属性。

$$\min_{\theta_G} \max_{\theta_G} v(\theta_G, \theta_D) = \mathbf{E}_{x,y\sim p_{data}}[\log D(x, y)] + \mathbf{E}_{z\sim p_z(z),\tilde{y}\sim p_y}[\log(1 - D(G(z, \tilde{y}), \tilde{y}))]$$

- Initial Latent Vector Approximation

与自动编码器相反，cGANs 没有明确的机制将属性为 y 的输入图像 x 映射到图像重建（X=G(z，y)）所必需的潜在向量 z:。通过训练编码器 E，一个近似逆映射的神经网络来解决这个问题。

为了训练 E，我们生成 100K 对的合成数据集(xi，g(zi，yi))，其中，zi~N(0，1)是随机的潜在载体，y~U 分布在 6 个年龄类别之间的随机年龄条件，g(z，y)是事先训练的 Acgan 的发生器，并且 xi=g(zi，yi)是合成脸部图像。训练 E 以使估计的潜向量 E(xi)与地面真实潜向量 zi 之间的欧几里得距离最小化。

E 产生初始潜在向量 Z0，但是却不足以保持身份信息，使得人脸识别精度仅达到 50%，但足以作为我们的优化算法的初始值。

- Latent Vector Optimization

给定一个能够识别输入人脸图像 x 中一个人身份的神经网络 FR，原始图像和重建图像 x 和 x 中的身份之间的差异可以表示为对应嵌入 FR(X)和 FR(X)之间的欧几里德距离。因此，最小化这种距离应该改善重建图像中的身份保持，从而优化 z。

$$z^*_{IP} = \underset{z}{\operatorname{argmin}} ||FR(x) - FR(\bar{x})||_{L_2}$$

118. Boosting Cross-Age Face Verification via Generative Age Normalization

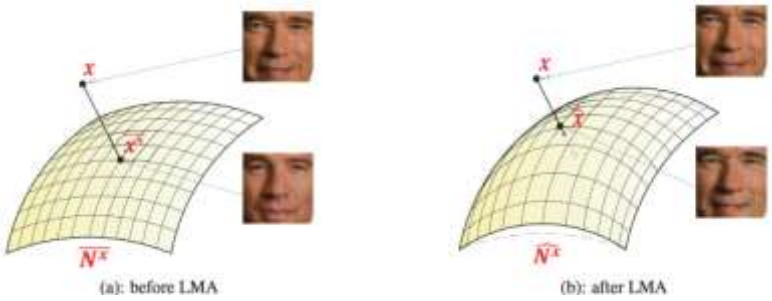
主要思想：

- 基于 age-cGAN，由于其不能直接用于人脸识别，因为它在老年/年轻化人脸中无法完全保留身份信息。因此，我们提出了局部流形适应(LMA)方法，解决了 age-cGAN 的问题，从而产生了新的年龄-cGAN 老化/再生方法。

主要步骤：

- Local Manifold Adaptation

局部流形自适应(LMA)方法改善重建人脸的身份保留。(a)通过将输入面 x 投影到合成的流形$\overline{N^x}$。(b)LMA 局部改善合成流形$\overline{N^x}$变换为新流形$\widehat{N^x}$的一个局部模函数。因此，新流形上的初始面 x 及其投影$\hat{x}$比 LMA 前更接近。



(a): before LMA　　(b): after LMA

由于是由发生器合成流形$\overline{N^x}$的，所以 LMA 是相对于输入图像 x 对由发生器 G 进行轻微的改动来执行的。LMA 的核心思想是：对通用生成器 G 进行了自定义，以使输入图像 x0 和(已知的)y 0 更好地获得一个新的生成器 Gx0。在 LMA 之后，Gx0 可以产生输入面 x0 的拟完全重构 x0=Gx0(z，y0)。我们的直觉表明，如果 LMA 重建 x 比 age-cGAN 重构 x 更接近 0，则老年/年轻化面部 x1=Gx0(z=z，Y1)也比通过通用生成器 G 获得的标识保留得更好。

采用方程$||FR(x) - FR(G(z^*, y))||_{L_2}$的优化目标，但在 LMA 情况下，固定 z 和优化 G。

8) 单样本

119. One-shot Face Recognition by Promoting Underrepresented Classes

主要思想：

- One-shot 分类泛化能力差的主要原因是数据不平衡问题，multi-nomial logistic regression 无法有效地解决。揭示了在多项式线性回归中，特征空间中的类划分体积与 one-shot 类权向量的范数之间存在着密切的联系，且 one-shot 中多项 Logistic 回归的不足与多项式 Logistic 回归中的权向量范数有关。
- 提出了一种新的监督信号，称为 underrepresented-classes promotion (UP) loss，它将 one-shot 类的权重向量的范数与正常类别对齐。以有效地解决 one-shot 中的数据不平衡问题。

主要步骤：

构建的数据库包含 21000 个人。分为两个 set，在 base set 中，有 20，000 人，每个人都有 50-100 张训练图像，5 张用于测试。在 novel set 中，有 1000 人，每个人有一个训练图像和 20 个用于测试的图像。主要关注 novel set 的分类性能，仅举一个例子来评价泛化能力。同时，还监视 base set 上的性能，以确保不通过牺牲 base set 上的性能来获得 novel set 上的性能增益。

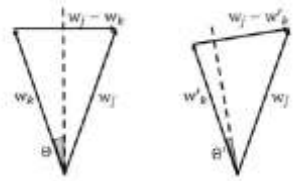
我们的方法包括以下两个阶段。第一阶段是表征学习。在这个阶段，我们建立了人脸表示模型使用 base set 的所有训练图像。第二阶段是 one-shot learning。在这一阶段，基于第一阶段学习的表示模型，训练了一个多分类器来识别 base set 和 novel set 中的

人。

- Representation learning

在 softmax 的监督下训练具有交叉熵损失的深度卷积神经网络(ConvNet)。采用了标准的 34 层残差网络。从最后一个池层提取特征作为人脸表示。

- OneshotLearning with UP



$$\frac{p_j(x)}{p_k(x)} = \frac{\exp(\mathbf{w}_j^T\phi(x))}{\exp(\mathbf{w}_k^T\phi(x))} = \exp[(\mathbf{w}_j - \mathbf{w}_k)^T\phi(x)]$$

上图表示w_k范数与 k 类分区体积大小的关系。破折线表示分隔相邻两个类的超平面(垂直于 wj-wk)。结果表明,当 wk 范数减小时,k 类在特征空间中具有较小的体积大小。

在损失函数中引入一个新的项，使得 novel set 中的人和 base set 中的人在特征空间中对应的分区平均体积大小相似：

$$\mathcal{L}_{up} = \sum_n -t_{k,n}\log p_k(x_n) + \frac{1}{|C_n|}\sum_{k\in C_n}\|\|\mathbf{w}_k\|_2^2 - \alpha\|_2^2,$$

其中 α 是 base set 的权向量平方范数的平均值$\alpha = \frac{1}{|C_b|}\sum_{k\in C_b}\|\mathbf{w}_k\|_2^2$。将 novel set 中权向量的平方范数的平均值推广到 base set 权向量的平方范数的平均值。

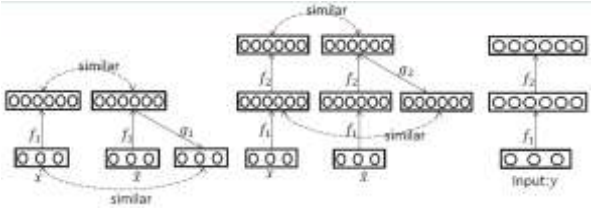
120. Single Sample Face Recognition via Learning DeepSupervised Auto-Encoders

主要思想：

- 本文针对人脸识别中单个训练样本的学习问题，提出了一种鲁棒的图像表示方法。
- 第一，我们强制使用带有 variants 的人脸映射到标准人脸，例如，具有中性表情和正常光照的正面人脸；第二，我们强制执行对应于同一人的特征是相似的。
- 通过与常用的基于稀疏表示的分类方法相结合，我们提出的基于叠加监督自动编码的人脸表示方法明显优于常用的基于单个样本的人脸识别图像表示。

主要步骤：

左图：基本的监督自动编码器，由干净/“污染”的人脸，特征(隐藏层)，以及使用“污染人脸”重建的干净人脸组成。中间图：利用先前隐藏层的输出作为输入，训练下一个有监督的自动编码器。我们多次重复这样的训练，直到达到所需的隐藏层数。本文只使用两个隐层。右图：一旦网络被训练，给定任何输入面，最后一个隐藏层的输出作为图像表示的特征。



类比于 *Denoising Auto-Encoder*，给出了一组包含 gallery 图像(干净数据)、probe 图像(损坏数据)及其标签的数据，利用它们训练出一种用于特征提取的深层神经网络。我们将该数据集中的每个 probe 图像表示为$\tilde{x_i}$，并将其对应的 gallery 图像表示为x_i。$\tilde{x_i}$和x_i的表示应该是相似的。

$$\min_{W,b_f,b_g} \frac{1}{N}\sum_i \left(\|x_i - g(f(x_i))\|_2^2 + \lambda\|f(x_i) - f(\hat{x}_i)\|_2^2\right) + \alpha\left(\mathrm{KL}(\rho_x\|\rho_0) + \mathrm{KL}(\rho_{\hat{x}}\|\rho_0)\right) \quad (2)$$

where

$$\rho_x = \frac{1}{N}\sum_i \frac{1}{2}(f(x_i)+1),$$

$$\rho_{\hat{x}} = \frac{1}{N}\sum_i \frac{1}{2}(f(\hat{x}_i)+1), \quad (3)$$

$$\mathrm{KL}(\rho\|\rho_0) = \sum_j \left(\rho_0 \log(\frac{\rho_0}{\rho_j}) + (1-\rho_0)\log(\frac{1-\rho_0}{1-\rho_j})\right).$$

第一项是重构损失，第二项身份特征相似性损失。由于隐层的输出作为特征，f(x_i)和 f($\hat{x}_i$)对应于同一人的特征。第三项和第四项的 Kullback-Leiber 散度(KL 散度)项在隐层中引入稀疏性。由于我们使用的激活函数是双曲正切，它的输出在-1 到 1 之间，其中-1 的值被认为是非激活的。因此，我们将编码器的输出映射到方程(3)中的范围(0，1)。通过选择一个小的ρ0，kl 散度正则化器强制只激活少数几个神经元。将ρ0 设为 0.05。

基于稀疏表示的分类方法用于人脸识别。在稀疏编码前，对特征进行归一化，使它们的 2 范数等于 1。

9) 化妆

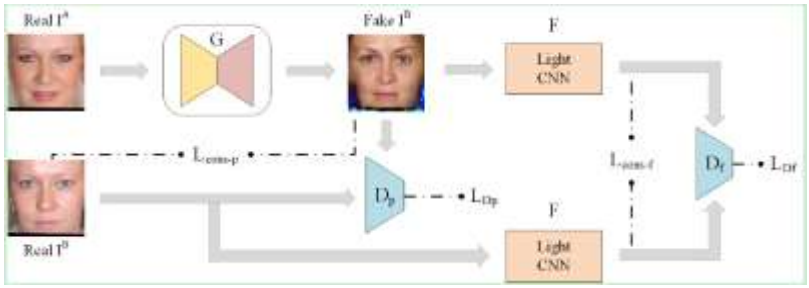
121. Anti-Makeup: Learning A Bi-Level Adversarial Network for Makeup-InvariantFace Verification

主要思想：

- 通过引入双层对抗网络(BLAN)，提出了一种基于生成的合成不变人脸验证方法。为了减轻化妆带来的负面影响，我们首先从化妆图像中生成非化妆图像，然后利用合成的非化妆图像进行进一步验证。
- BLAN 中的两个对抗性网络集成在一个端到端的深层网络中，一个在像素级重建吸引人的面部图像，另一个在特征层上保存身份信息。

主要步骤：

I^A是一个输入的化妆图像，而I^B代表相应的非化妆图像。生成器 G 学习欺骗两个鉴别器，其中 DP 在像素级，DF 在特征级别。



生成器 G 接收四种参数更新损失：两个重建损失L_{cons-p} and L_{cons-f}和两次对抗性损失L_{D_p} and L_{D_f}

$$G^* = \frac{1}{N}\arg\min_G \sum_{n=1}^{N} L_{cons-p} + \lambda_1 L_{D_p} + \lambda_2 L_{cons-f} + \lambda_3 L_{D_f}$$

- 图像级别的重构损失损失L_{cons-p}由像素级损失、对称损失和一阶损失组成：

① 像素级损失：$L_{pel} = \mathbb{E}_{(I^A,I^B)\sim p(I^A,I^B)}\|G(I^A) - I^B\|_1$ (4)

② 一阶损失：一阶损失也可称为边缘损失，因为它的目的是充分探索在$\boldsymbol{I^B}$中提供的梯度先验。

$$L_{edg} = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}\{ \|\,|G(I^A)_{i,j} - G(I^A)_{i,j+1}| - |I^B_{i,j} - I^B_{i,j+1}|\,\|_1 + \|\,|G(I^A)_{i,j} - G(I^A)_{i+1,j}| - |I^B_{i,j} - I^B_{i+1,j}|\,\|_1 \} \quad (5)$$

其中，$G(I^A)_{i,j}$表示合成图像$G(I^A)$的(i，j)像素。

③ 对称损失：$L_{sym} = \frac{1}{h\times w/2}\sum_{i=1}^{h}\sum_{j=1}^{w}\|G(I^A)_{i,j} - G(I^A)_{i,w-j+1}\|_1$

- 图像级别的对抗性损失：

$$L_{D_p} = \mathbb{E}_{\{I^A\}\sim p(I^A)}[-\log D_p(G(I^A))] \quad (7)$$

- 特征级别的对抗性损失，light CNN 作为特征提取器：

$$L_{D_f} = \mathbb{E}_{\{I^A\}\sim p(I^A)}[-\log D_f(F(G(I^A)))].$$

- 特征级别的重构损失

$$L_{cons-f} = \mathbb{E}_{\{I^A,I^B\}\sim p(I^A,I^B)}\|F(G(I^A)) - F(I^B)\|_1$$

判别器的优化函数如下：

$$D_p^* = \arg\max_D \mathbb{E}_{I^B\sim p(I^B)}\log D(I^B) + \mathbb{E}_{I^A\sim p(I^A)}\log(1 - D(G(I^A))) \quad (2)$$

$$D_f^* = \arg\max_D \mathbb{E}_{I^B\sim p(I^B)}\log D(F(I^B)) + \mathbb{E}_{I^A\sim p(I^A)}\log(1 - D(F(G(I^A)))) \quad (3)$$

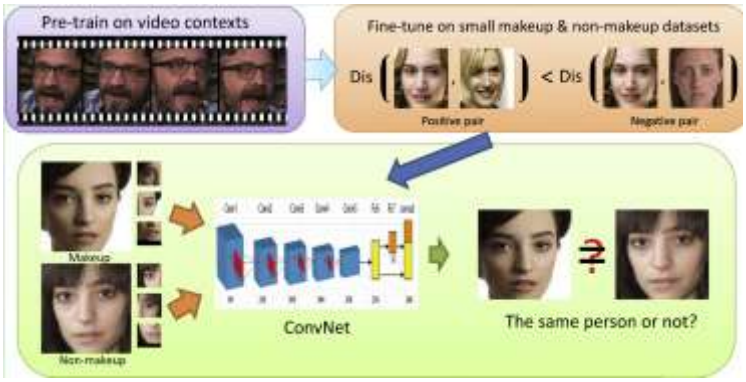
122. A weakly supervised method for makeup-invariant face verification

主要思想：

- 人脸验证模型是大规模视频数据中预先训练，并在小规模的化妆和非化妆数据集微调。
- 为了充分利用视频上下文和有限的化妆和非化妆数据集，许多技术被用来提高性能。定义了一种新的三重项和两对项的损失函数，并利用所提出的投票策略将多个人脸部件组合在一起，以获得更好的验证结果。

主要步骤：

首先，我们对网络进行了视频上下文的预训练，这些视频上下文很容易从公共视频中获取.。接下来，我们对化妆和非化妆图像进行微调，在第二阶段对这些图像的几个部分进行微调。在最后阶段，我们采用投票的方法，总结了从整个人脸图像和人脸各部分中得到的验证结果，并对输入的两幅图像是否属于同一身份做出了最终的判断。



- The triplet network

骨干网是基于 alexnet 框架。为了避免网络过度拟合，这两个完全连接的层(fc6 和 fc7)的神经元数目分别被截断为 256 和 32。将 fc6 和 fc7 层的特征串联在一起，并在最终的特征空间中得到一个 288 维特征向量。

在我们的网络中使用三种 loss 函数。一个是 triplet loss，另两个是 pairwise loss。所有这些损失函数都是基于标准的余弦距离。

$$l_{rank}(f(X), f(X^p), f(X^n)) = \max[0, d(f(X), f(X^p)) - d(f(X), f(X^n)) + \alpha].$$

三重态秩损失函数不限制 f(X)与 f(Xp)的相似程度以及 f(Xn)与 f(X)的区分程度。为了更严格地约束距离 d(f(X),f(Xp))和 d(f(X),f(Xn))，我们使用了两个对偶损失函数。

$$l_{pos}(f(X), f(X^p), f(X^n)) = \max[0, d(f(X), f(X^p)) - (\delta - \alpha/2)],$$

and

$$l_{neg}(f(X), f(X^p), f(X^n)) = \max[0, (\delta + \alpha/2) - d(f(X), f(X^n))],$$

最终优化函数：

$$l = l_{rank} + \lambda_{pos} \cdot l_{pos} + \lambda_{neg} \cdot l_{neg},$$

除了对整个人脸图像进行训练外，我们还在部分图像上训练。这三个部分分别包括左眼、右眼和嘴。为了集成对部件进行训练的的结果，我们使用了一种投票方法。

$$\sigma(X, Y, \gamma) = \begin{cases} 1, & \text{if } d(f(X), f(Y)) \le \gamma, \\ -1, & \text{if } d(f(X), f(Y)) > \gamma. \end{cases}$$

and calculate

$$v = \lambda\sigma(X, Y, \gamma) + \sum_i \sigma(X_i, Y_i, \gamma_i).$$

10） Set-based 人脸识别

123. Multi-Prototype Networks for Unconstrained Set-based FaceRecognition

介绍了 image-set 的分类，并详细介绍了几个评价指标

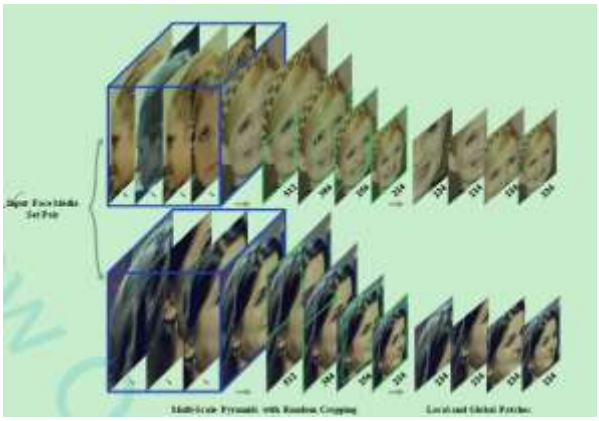
主要思想：

- 现有策略大致分为两种：一种，从集合中的每一个面部介质中学习一组图像级人脸表示，并使用所有信息进行后续人脸识别，计算量太大，具有冗余。另一种策略是通过平均或最大 pooling 聚合整个集合的面表示，并为每个集合生成单个表示，显然遭受了信息损失。
- 提出了一种新的多原型网络(MPNET)模型，该模型能自适应地从媒体集合中学习多个原型人脸表示。每个学习的原型在特定的姿态、光照和媒体形态条件下的，具有一致性。不同的原型进行聚合，这样既可以减少信息损失，也可以减少计算量。
- MPNet 引入了一种新的密集子图(DSG)学习子网进行原型的划分，它隐式地解开了不一致的媒体，并学习了许多基于无约束集的人脸识别的原型，而不是手工分区。
- 带有 DSG 子网的 MPNet 是端到端可训练的。

主要步骤：

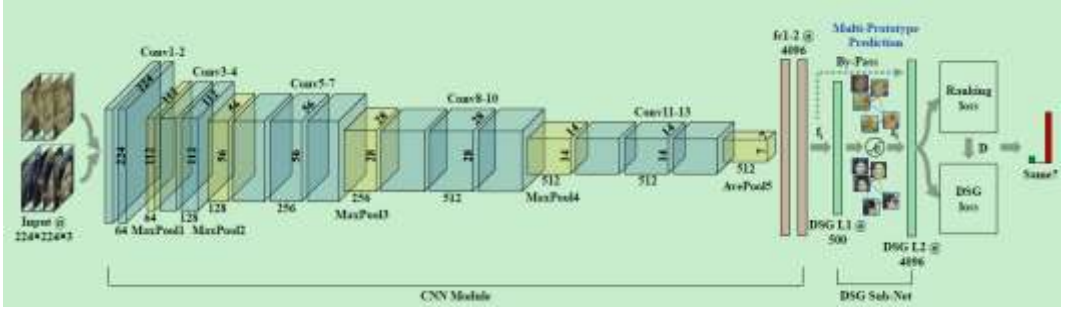
- Deep Set-based Facial Representation Learning

MPNET 在多尺度上学习人脸表示，以获得对真实人脸的尺度方差增强的鲁棒性。具体来说，对于面部媒体集中的每一种媒体，通过将图像或视频帧调整为四个不同的尺度，可以构造一个多尺度的金字塔。mpnet 执行随机裁剪，从固定大小的多尺度金字塔的每个尺度上收集局部和全局 patch。



为了处理真实脸数据的不平衡(例如，一些被试从有限的图像中加入稀缺媒体，而一些对象使用来自重复视频帧的冗余媒体)，通过重采样来调整每个集合内的数据分布。稀缺媒体的集合(即小于经验设置的预定义参数 r)通过复制和翻转图像来增强。具有冗余介质(即大于 r)的大集合被下采样为 r 的大小。

该网络采用 a Siamese CNN 架构，两个分支权重共享，每个分支包括 13 卷积层，5 层 pooling 层和 2 层全连接层。学习到的每个媒体集的深度面部表示表示为$\{f_1, f_2, \cdots, f_R\}$。



- Multi-Prototype Discriminative Learning

  原型被定义为在特定条件下代表一个主题脸的相似面部媒体的集合。每个面部媒体集都被隐式地分解成一定数量的原型。

  Dense SubGraph Learning:

  每个子图具有较高的内部相似度和与外界介质的不相似性。每个子图为输入主题面提供一个原型。然后，我们在原型层执行人脸识别，这是简洁的，也足够翔实。

  假设图 g 与 A 相关联，并且每个元素编码两个面媒体之间的相似性：$a_{ij} = \mathrm{aff}(f_i, f_j)$。设 k 是原型的数目(或等价的，稠密子图的数目)，$Z = [z_1, \ldots, z_K] \in \mathbb{R}^{n \times K}$，Zik=1 表示第一介质在第 k 原型中。

  DSG 的目的是通过优化 z 找到有代表性的子图：

$$\max_{Z} \mathrm{tr}(Z^\top AZ), \ \text{s.t.}, z_{ik} \in \{0,1\}, Z\mathbf{1} = \mathbf{1},$$

  Dense SubGraph Sub-Net：

  该子网由两层组成，它以基于深度集的人脸表示 FI 作为输入，输出重构的判别特征。

  第一层是原型预测层。给定输入深面表示 Fi，该层通过动态投影每个输入图到 K 原型，输出其指示符 $z_i \in \{0,1\}$，$z_i = \sigma(W^\top f_i)$。在给定预测的 Zi 和输入 Fi 的情况下，第二层计算以下 DSG 损失，以确保重建的表示 $\tilde{f}_i$ 能够形成合理的紧凑和鉴别的原型。$\tilde{f}_i$ 是通过将 DSG 子网的第二层的输出与从第一层得到的指示符 z 相乘得到的。

- 损失函数

  Ranking loss:

  通过欧式距离测量 $\tilde{f}_i$ 和 $\tilde{f}_j$ 之间相似性：$d_{ij} = \|\tilde{f}_i - \tilde{f}_j\|_2^2$。Ranking loss 如下：

$$\mathcal{L}_{Ranking}(\tilde{f}_i) \triangleq \min_{E} \{(1 - y^p)E + y^p \max(0, \tau - E)\}$$

$$E = \frac{\sum_{i,j}^{R} d_{ij} \times \exp(\beta d_{ij})}{\sum_{i,j}^{R} \exp(\beta d_{ij})},$$

  Dense SubGraph loss：

$\mathcal{L}_{\mathrm{DSG}}(\tilde{f}_i) \triangleq \left\{ \min_{Z} \mathrm{tr}(Z^\top DZ), \ \text{s.t.} \ z_{ij} \in \{0,1\}, Z\mathbf{1} = \mathbf{1} \right\}$，其中 D 为 $\tilde{f}_i$ 和 $\tilde{f}_j$ 的欧式距离

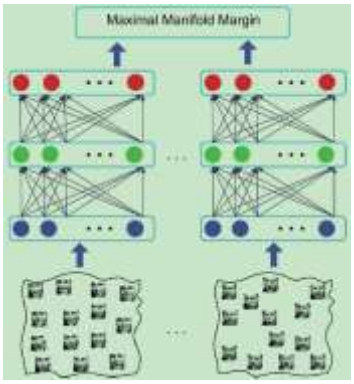
124. Multi-Manifold Deep Metric Learning for Image Set Classification

主要思想：

- 该方法通过学习多组非线性变换(每类一组)，将多组图像实例非线性映射到一个共享的特征子空间中，使不同类的流形边缘最大化，从而利用判别性、类特异性和非线性信息进行分类。

主要步骤：

我们提出的图像集分类方法的基本思想是，对每个图像集建模为一个流形，并将其传递到多个深层神经网络中，将每个流形非线性地映射到另一个特征空间。具体来说，深度网络是特定于类的，因此不同类别的网络中有不同的参数。在这些网络的顶层，使用最大流形 margin 准则。在测试阶段，我们使用这些特定于类的深度网络计算测试图像集与所有训练集之间的相似性，并使用最小距离进行分类。



深度网络得到样本的非线性映射：$h_{ci}^L = s(W_c^L h_{ci}^{L-1} + b_c^L)$

C 流形的每个样品h_{ci}^L，我们计算两个平方距离$D_1(h_{ci}^L)$和$D_2(h_{ci}^L)$，它测量了这个样本和类内和类间邻近样本的相似性

$$D_1(h_{ci}^L) = \frac{1}{K_1}\sum_{p=1}^{K_1}\|h_{ci}^L - h_{cip}^L\|_2^2$$
$$D_2(h_{ci}^L) = \frac{1}{K_2}\sum_{q=1}^{K_2}\|h_{ci}^L - h_{ciq}^L\|_2^2$$

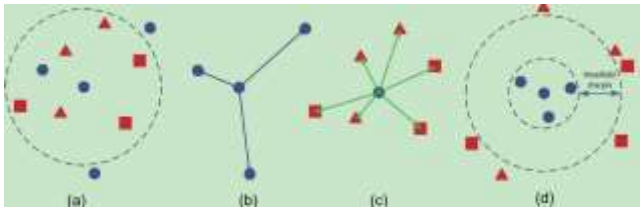
h_{cip}^L和h_{ciq}^L是第 p 个 intra-manifold 和第 q 个 inter-manifold 的多个邻居的顶层的特征表示形式。□1 和□2 是定义邻域大小的两个参数。提出以下优化问题，以最大限度地提高c流形和其他流形的 margin:

$$\min_{f_c}\sum_{i=1}^{N_c}(D_1(h_{ci}^L) - D_2(h_{ci}^L))$$

最终的优化函数为：

$$\min_{f_1,f_2,\cdots,f_C} H = H_1 + \frac{\lambda}{2}H_2$$
$$= \sum_{c=1}^{C}\sum_{i=1}^{N_c} g\left(D_1(h_{ci}^L) - D_2(h_{ci}^L)\right)$$
$$+ \frac{\lambda}{2}\sum_{c=1}^{C}\sum_{l=1}^{L}\left(\|W_c^l\|_F^2 + \|b_c^l\|_2^2\right) \quad (5)$$

其中 g（a）是一个广义逻辑损失函数，它平滑地逼近 hinge 损耗函数。$g(a) = \frac{1}{\rho}\log(1+\exp(\rho a))$。我们首先用适当的值初始化网络参数，计算类内和类间邻居，然后通过（5）更新这些参数直到收敛为止。



给定测试图像集。$X^q = [x_1^q, x_2^q, \cdots, x_{N_q}^q]$，其中$x_j^q$是第 j 个图像，$N_q$是这个集合中的图像数，我们计算测试集$X_q$和每个训练集$X_c$之间的距离。并指定一个标签$L_q$到测试集：

$$L_q = \arg\min_c d(X_q, X_c), \quad 1 \le c \le C.$$

计算距离 d（）：每个样本x_j^q，我们首先使用 c 类训练集的深度网络将其映射到特征空间$h_c(x_j^q)$。然后，我们计算出$h_c(x_j^q)$和每个训练样本h_{cj}的欧氏距离，最小的距离被选为h_c和 c 类流形的距离。最后，我们将所有这些点到流形的距离平均为流形X^q和X^c之间的距离。

125. Unconstrained Face Recognition Using A Set-to-SetDistance Measure

列举了 feature pooling 和 score pooling 的一些论文，并很好的解释了 feature pooling 和 score pooling 的原理

主要思想：

- S2S 距离采用 KNN- average pooling 对在两组中计算的所有媒体上的相似分数，使得识别比传统的特征平均池和分数平均池更不容易受到不良表示(异常值)的影响。
- 此外，我们还表明各种度量可以嵌入到我们的 S2S 距离框架中

主要步骤：

对于 probe 集中的每一个媒体项目$\tilde{z}_i^P$，我们首先计算了在 gallery 集中$\tilde{z}_i^P$与其 k 近邻$\tilde{z}_{nnk}^G$之间的相似性，并对它们进行了求和：

$$sim_{\tilde{z}_i^P} = \sum_{j=1}^{k} K(\tilde{z}_i^P, \tilde{z}_{NN_j}^G),$$

然后取所有分数的平均值作为从 probe 到 gallery 的相似性：

$$sim_{P-G_{NN}} = \frac{1}{\|\mathbf{P}\|}\sum_i sim_{\tilde{z}_i^P},$$

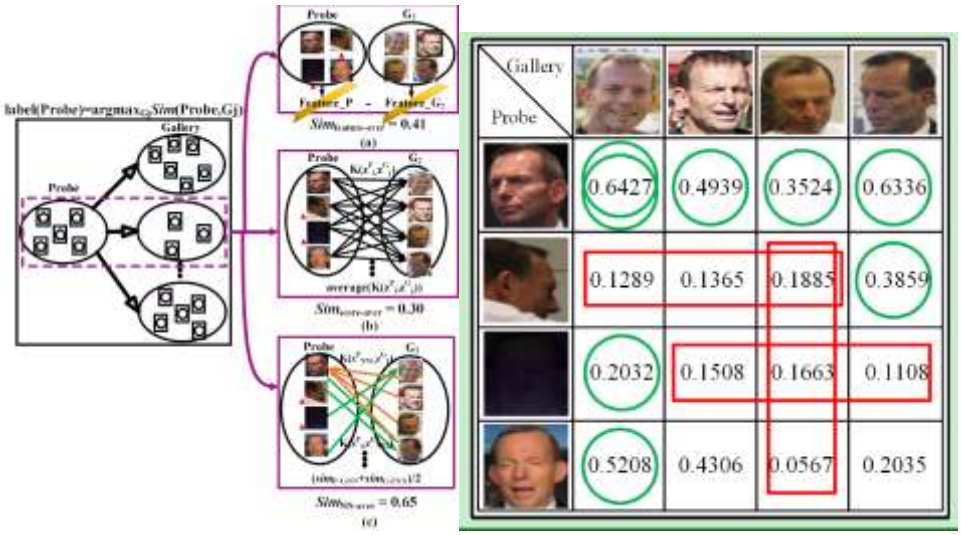
同样，我们对 gallery 集也进行了相同的操作，以获得从图库到探针的相似性：

$$sim_{s_j^G} = \sum_{i=1}^{k} K(z^P_{NN_i}, z^G_j),$$
$$sim_{G-P_{NN}} = \frac{1}{\|\mathbf{G}\|} \sum_j sim_{s_j^G}.$$

最后，这两个集合之间的相似性是

$$sim_{kNN\_aver} = (sim_{P-G_{NN}} + sim_{G-P_{NN}})/2. \quad (4)$$

一般情况下，特征平均池和分数平均池可以看作是比较两组的分布。但是，在比较的两组不包含大量媒体或它们包含严重噪声的情况下，可能会得到一个低的相似分数。

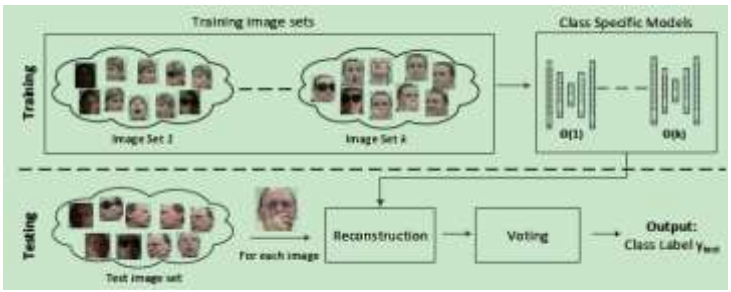


126. Learning Non-Linear Reconstruction Models for Image Set Classification

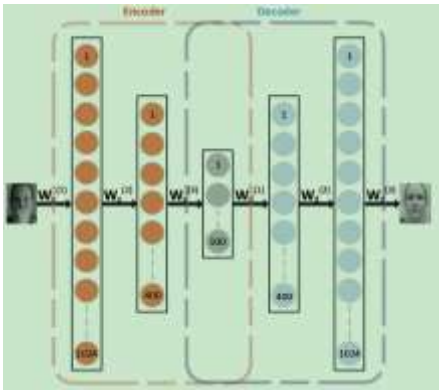
主要思想：

- 定义了一个自适应深网络模板(Adnt)，其参数通过使用高斯限制 Boltzmann 机器(Grbms)进行无监督的分层预训练来初始化，预初始化后的 adnt 被分别训练为每个类的图像，并学习特定于类的模型。基于学习到的类特定模型的最小重构误差，采用多数投票策略进行分类。

主要步骤：



为了使这样一个深层次的网络运行良好，需要适当初始化权重。我们使用高斯限制 Boltzmann 机器，以贪婪层明智的方式进行预训练，初始化 adnt 的权重。然后，对训练集的每一个 k 类分别对具有预初始化权值的 adnt 进行微调。微调深网络模型，每个模型对应于 k 类中的一个。然后将这些精细调整的模型用于图像集分类。



- The Adaptive Deep Network Template (ADNT)

adnt 是一个自动编码器(Ae)，由两个部分组成：编码器和解码器。编码器和解码器各有三个隐藏层，有一个共享的第三层(中央隐藏层)。

$$\begin{aligned}\mathbf{h} &= s(\mathbf{W}_e^{(3)}\mathbf{h}_2 + \mathbf{b}_e^{(3)})\\ \mathbf{h}_2 &= s(\mathbf{W}_e^{(2)}\mathbf{h}_1 + \mathbf{b}_e^{(2)})\\ \mathbf{h}_1 &= s(\mathbf{W}_e^{(1)}\mathbf{x} + \mathbf{b}_e^{(1)})\end{aligned}, \quad \begin{aligned}\tilde{\mathbf{x}} &= s(\mathbf{W}_d^{(3)}\mathbf{x}_2 + \mathbf{b}_d^{(3)})\\ \mathbf{x}_2 &= s(\mathbf{W}_d^{(2)}\mathbf{x}_1 + \mathbf{b}_d^{(2)})\\ \mathbf{x}_1 &= s(\mathbf{W}_d^{(1)}\mathbf{h} + \mathbf{b}_d^{(1)})\end{aligned}$$

- Image Set Classification Algorithm

给定 k 个训练图像集$\{X_c\}_{1\times k}$及其相应的类标签$y_c \in [1,2,\cdots k]$，其中属于 c 类的图像集$X_c = \{\mathbf{x}^{(t)}\}_{1\times N_c}$有 Nc 个图像$\mathbf{x}^{(t)} \in \mathbb{R}^{d_x \times d_y}$，则图像集分类问题可以表述如下：给定一个测试图像集$X_{test} = \{\mathbf{x}^{(t)}\}_{1\times N_{test}}$，找到属于哪个类的$y_{test}$？

类特定模型θ(c)的学习是通过反向传播进行随机梯度下降，以最小化重建误差，对训练图像集 Xc 的所有示例$x^{(t)}$：

$$J\left(\theta_{\text{ADNT}}; \mathbf{x}^{(t)} \in X_c\right) = \sum_{\mathbf{x}^{(t)}} \left\|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}^{(t)}\right\|^2$$

总优化函数为：

$$J_{\text{reg}}\left(\theta_{\text{ADNT}}; \mathbf{x}^{(t)} \in X_c\right) = \sum_{\mathbf{x}^{(t)}} \left\|\mathbf{x}^{(t)} - \hat{\mathbf{x}}^{(t)}\right\|^2 + \lambda_{\text{wd}} J_{\text{wd}} + \lambda_{\text{sp}} J_{\text{sp}}$$

$$J_{\text{wd}} = \sum_i^3 \left\|\mathbf{W}_e^{(i)}\right\|_F^2 + \sum_i^3 \left\|\mathbf{W}_d^{(i)}\right\|_F^2$$

$$\begin{aligned} J_{\text{sp}} &= \sum_i^5 \sum_j \text{KL}\left(\rho \| \bar{\rho}_j^{(i)}\right) \\ &= \sum_i^5 \sum_j \rho \log \frac{\rho}{\bar{\rho}_j^{(i)}} + (1-\rho)\log\frac{1-\rho}{1-\bar{\rho}_j^{(i)}} \end{aligned} \tag{11}$$

Classification：

我们从所有类别模型$\theta(c), c = 1\cdots k$中分别重建测试集每个图像$\mathbf{x}^{(t)} \in X_{test}$：

$$r^{(t)}(c) = \left\|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}^{(t)}(c)\right\|_2$$

在计算了所有 k 个模型的重建误差后，根据最小重建误差准则，对图像 x(T)的类进行了判定。

$$y^{(t)} = \arg\min_c r^{(t)}(c)$$

投票：计算测试集的所有N_{test}映像的类标签。然后将测试集X_{test}的标签y_{test}定义为X_{test}的所有图像中最重复出现的标签。

$$y_{test} = \arg\max_c \sum_t \delta_c(y^{(t)}) \text{ ,where}$$

$$\delta_c(y^{(t)}) = \begin{cases} 1, & y^{(t)} = c \\ 0, & \text{otherwise} \end{cases}$$