基于深度网络的面部关键点联合头部三维姿势预测算法

贺珂珂[1)] 薛向阳[1)]

[1)](复旦大学上海智能信息处理重点实验室， 上海 200433)

摘 要 在现有的人脸对齐流程中，当头部有较大非平面的旋转时，仅利用面部关键点的二维坐标信息，难以满足对齐的需求。为此，本文提出了一种新颖的基于深度学习的面部关键点定位联合头部三维姿势预测的算法。在深度学习的框架下，通过端到端的学习模式，在整个人脸区域上抽取层次性特征，直接预测出所有的关键点和三维姿势，在利用人脸的整体特征的同时，隐含关键点的几何约束。这种方法的优势在于：1. 相比传统的定位方法，能较好处理较大头部姿势下的定位问题。2. 能在几乎不增加计算代价的条件下，同时得到关键点和三维姿势。3. 模型轻量级，速度可以达到 55 fps。同时此算法在面部关键点的评测数据集上取得了领先的效果。

关键词 深度学习; 面部关键点定位; 头部姿势估计; 人脸对齐; 多任务训练

中图法分类号 **** DOI 号 *投稿时不提供 DOI 号* 分类号

A deep convolutional framework for joint facial landmark localization and head pose estimation

Keke He[1)] Xiangyang Xue[1)]

[1)](Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China)

Abstract

In current face alignment process, it is difficult to align the face just with 2D landmark coordinates when head has out-of-plane rotations. We propose a novel joint facial landmark localization and head pose estimation algorithm using deep neural network. Based on the end-to-end deep learning, we extract hierarchical facial features on the entire face, then predict all the landmarks and head pose at the same time. This method has several advantages: 1. It can handle faces with large head pose.2. It can obtain the extra head pose with little computation. 3. The model is lightweight, and can achieve 55 fps. Our method also achieves comparable result on 300-W benchmark.

Key words deep learning; facial landmark localization; head pose estimation; face alignment; multi-task learning

———————————————

收稿日期：年-月-日；最终修改稿收到日期：年-月-日 *投稿时不填写此项*. 本课题得到… …基金中文完整名称(No.项目号)、… …基金中文完整名称(No.项目号)、… … 基金中文完整名称(No.项目号)资助.作者名1(**通讯作者**)，性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********.作者名2（**通讯作者**），性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********. 作者名3（**通讯作者**），性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********.(**给出的电子邮件地址应不会因出国、毕业、更换工作单位等原因而变动。请给出所有作者的电子邮件**)

第1作者手机号码(投稿时必须提供，以便紧急联系，发表时会删除): … …, E-mail: … …*此部分6号宋体*

1 引言

随着生物认证技术的发展，人脸作为最为自然和普遍的身份特征，吸引了大量的研究。人脸识别由于具有操作便捷、可交互性强等独特的优势，易于为用户所接受，具有广泛的应用前景。面部关键点定位，是一个人脸研究领域中的关键问题，其目标是在人脸图像中定位眼、嘴、鼻、人脸外轮廓等位置。由于关键点定位与人脸识别，人脸追踪，人脸属性分析，3D 人脸建模的紧密联系，其在过去几十年引起了大量的关注。然而，由于头部的姿势，表情变化，遮挡，光照等因素，如图 1，极大增加了面部关键点定位的难度，面部关键点定位依旧是一个较大的挑战。

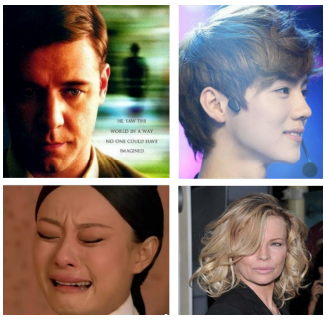

图 1 关键点定位的挑战（左上：光照，左下：表情，右上：姿势，右下：遮挡）

现有的关键点定位方法大致可以分为 3 类。第一类是基于 ASM 和 AAM 的对全局表观进行建模的方法。然而这些方法需要不断地迭代，计算代价较大，同时依赖一个好的关键点分布初值。这类方法通常把训练数据的平均关键点作为初始点，而这些初始点可能跟目标点有较大的差距，造成定点不精确。

第二类的方法是基于级联回归的方法。这类方法使用回归模型，直接学习从人脸表观到关键点坐标的映射函数。级联回归方法包含多个回归模型，先获得一个较粗的位置，然后利用后续的模型根据前面预测的位置，继续进行优化，以获得更精确的结果。根据模型所采用的特征的不同，又可以细分为非深度特征的级联回归方法和基于深度特征的级联回归方法。在非深度特征的级联回归方法中，如 Xiong[1]提出了 SDM 算法，使用 SIFT 特征，求解非线性最小二乘问题。Ren[2]利用随机森林，学习 LBF 特征，处理关键点定位问题。由于深度网络模型在计算机视觉的诸多问题，如物体分类，目标跟踪，图像分割等任务中获得了突破性的进展，显示了深度特征的鲁棒性。在面部关键点定位任务上，一系列基于深度特征的级联回归算法被提出来。Sun[3]提出了三阶段的级联深度网络从粗到细来定位关键点。Zhang[4]提出一种由粗到精的自编码器网络来描述从人脸表观到关键点位置的复杂非线性映射过程。然而这些方法需要为每个关键点集合训练一个模型，计算代价随着集合的数目和级联的层次递增。例如，Sun[3]提出的这个级联 CNN 方法需要训练 23 个独立的模型。

近来，基于多任务的关键点定位框架引起了关注。多任务的方法利用其它的人脸先验信息，例如微笑来辅助关键点定位。Zhang[5]提出了一种用微笑等相关信息辅助定位的方法。然而这类方法需要额外的标注信息。Yang[6]提出了一种多任务辅助的方法，先学习到头部姿势，然后利用头部姿势选择合适的面部关键点分布作为算法的初值。

考虑到现有的人脸对齐中，当人脸头部有非平面的旋转时，仅利用关键点的二维坐标信息，不能满足对齐的需求。如果能在获得关键点坐标的同时，得到头部的三维姿势(Pitch，Yaw 和 Row)，将更有助于人脸对齐工作。基于这个思考，本文提出一种新颖的基于深度学习的三维姿势联合面部关键点的多任务人脸定点方法。与现有的方法相比，本文的方法不要多个独立的模型，同时不需要额外的手工标注数据。更为重要的是，本文的方法能够在得到关键点的同时得到头部的三维姿势。

本文的算法在深度学习的框架下，同时将面部关键点和头部的三维姿势作为监督信息，通过端到端的联合学习模式，学习到一系列的非线性特征，利用了整体的人脸信息。在预测时，直接将原始的图片作为输入，经过已学习得到的非线性函数的映射，得到关键点和三维姿势。

在准确率上，本文的算法不仅能够有效地处理人脸头部姿势问题，同时不需要级联，较为轻量级，仅单模型就能在 300-W 评测数据集上达到领先水平。在速度上，可达到 55 fps。本文的代码可在 https://github.com/qiexing/face-landmark-localization 下载。

2 面部关键点联合头部姿势预测算法

通过对现有的部分算法的分析，发现当头部势偏转较大时，关键点的定位会有较大的误差。同

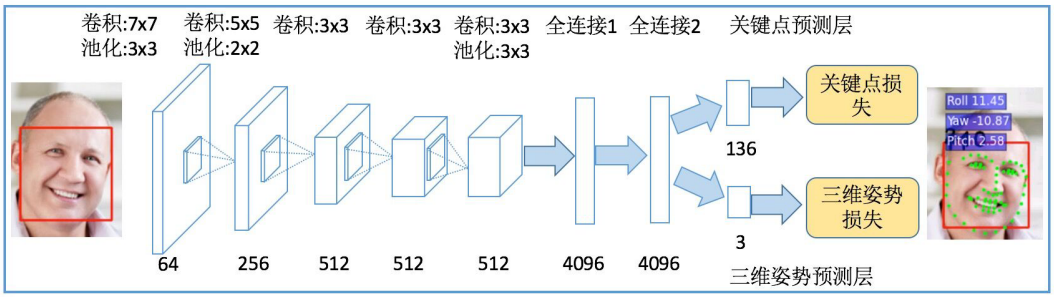

图 2 算法框架

时，头部三维姿势本身是一个重要的信息，表明了头部的朝向。考虑到头部姿势的重要性，本文将面部关键点和头部三维姿势的联合优化，能在几乎不增加现有的关键点定位算法复杂度的条件下，同时得到关键点和三维姿势。此外，由于三维姿势和关键点的信息联合，同时提高了关键点的精度。

2.1 三维姿势数据采集

现有的数据集中，标注的头部姿势往往是个大概的方向（如朝左，正面），并没有很精确的度数。同时标注面部关键点和头部三维姿势的数据更是几乎没有。所以如何获得三维姿势（Pitch，Yaw 和 Roll）的标注数据成了一个难题。本文生成标注的三维姿势的数据的方法来自[1]，从技术上来说，头部姿势是通过在给定 3 维空间到 2 维空间的映射情况下，根据人脸的模型，求解映射的过程。在这里，本文利用现有标注的面部关键点来计算三维姿势。

2.2 模型结构

图 2 为算法框架。网络的输入为根据人脸框截取的彩色图像，经过数个卷积层和池化层后，跟着 2 层全连接层，后续阶段的网络由 2 部分组成的。网络的第一部分为关键点定位，关键点定位是根据输入图像直接回归点的坐标。每个关键点的坐标由 2 个数字组成（表示平面内的 x，y 坐标），而这里需要预测 68 个关键点，因此设置关键点层的维度为 136。网络的第二部分为三维头部姿势的预测，本文将预测的头部三维姿势看成一个回归问题，头部姿势层的维度为 3，分为对应于 yaw，roll，pitch。关键点层的参数和头部姿势层的参数是各自独立的。

2.3 算法优化

网络的前向传播

网络共享层的前向计算过程跟普通的网络相同，逐层进行计算

$$x^0 \xrightarrow{h((w^0)^T x^0)} x^1 \xrightarrow{h((w^1)^T x^1)} x^2 \ldots \longrightarrow x^f$$

其中w^i表示每层的参数，h 称为激活函数，在这里使用$relu$函数。关键点子层的计算方式为：$f(x^f;w^l)=h((w^l)^T x^f)$，头部姿势子层的计算方式为：$f(x^f;w^p)=h((w^p)^T x^f)$。这 2 层的参数$w^l$, w^p 是独立的。

网络的反向求导

此深度网络的损失函数可以表示为关键点的损失和头部姿势的损失之和：

$$E=\underset{w_l,w_p}{\operatorname{argmin}}\sum_{i=1}^{N} l^l(y_i^l, f(x_i;w^l))+\lambda\sum_{i=1}^{N} l^p(y_i^p, f(x_i;w^p))$$

其中N表示样本的数量，$l^l(.)$表示关键点定位的损失函数，y_i^l表示标注的关键点值，$f(x_i;w^l)$表示网络预测的标注点的值；$l^p(.)$表示三维姿势预测的损失函数，y_i^p表示标注的三维姿势值，$f(x_i;w^p)$表示网络预测的三维姿势值，λ表示头部姿势任务的权重，当λ为 0 时，表示头部姿势任务不参与训练，当λ为 1 时，表示头部姿势任务共同参与训练。

关键点的坐标为连续值，损失函数选择用平方损失函数，头部三维姿势的角度在（-90°，90°）间，也是连续值，同样选择用平方损失函数。此深度网络的损失函数可以扩展表示为：

$$E=\underset{w_l,w_p}{\operatorname{argmin}}\frac{1}{2}\sum_{i=1}^{N}||y_i^l-f(x_i;w^l)||^2+\frac{1}{2}\lambda\sum_{i=1}^{N}||y_i^p-f(x_i;w^p)||^2$$

E 中的f函数可以详细表示为$f(x;w)=h((w)^T x)$网络的目标是根据训练样本，学习得到最优的参数w，使得损失最小。

根据经典的反向传导算法，权重的更新策略为：$\Delta w=-r\frac{\partial E}{\partial w}$，即求损失函数相对于权重的偏导，其中的$r$表示学习率，控制权重更新的幅度 。

关键点层的权重更新为$\Delta w^l=-r\frac{\partial E^l}{\partial w^l}$，$E^l$表示关键点的损失。

头部姿势层的权重更新为$\Delta w^p=-r\frac{\partial E^p}{\partial w^p}$，$E^p$

表示头部姿势的损失。

对于全连接层 2(简称为$fc2$)，作为关键点层和头部姿势层的共同输入，在反向传播时候，误差共同回传，联合调整$fc2$层的权重，所以，$fc2$层的权重的更新策略可以表示为：

$$\Delta w^{fc2} = -r\frac{\partial E}{\partial w^{fc2}}$$

在$fc2$层之前的每一层，跟一般网络的反向传导算法一致，逐层向前传播误差，逐层更新参数。

3 实验

本文选用 300-W[7]数据集进行实验。本文将在 3.1 详细介绍数据集，将在 3.2 说明网络的训练过程，将在 3.3，3.4，3.5 介绍实验的结果，包括不加头部姿势的训练和加头部姿势联合训练的差异，同时也会与其他公开的结果进行比较，最后在 3.6 节展示关键点定位的效果。

3.1 数据集

300-W (Faces in-the-Wild) [7] 是一个面部关键点定位问题的经典数据集。它包括了现有的 LFPW [8]，AFW [9]，Helen [10]和一个新推出的 IBUG 数据集。

跟其他论文的做法一致，本文将整个 AFW，LFPW 的训练集和 Helen 的训练集作为训练集，总共有 3148 张图片。选用 IBUG，LFPW 的测试集和 Helen 的测试集，作为测试集，总共有 689 张图片。将 IBUG 下的测试结果称为挑战集的结果，在 LFPW 测试集和 Helen 测试集下的结果称为普通集的结果，所有 689 张图片的测试结果称为所有集下的测试结果。

数据增强

对于深度卷积网络模型的训练来说，训练数据的质量和数量都是非常重要的，数据量过少，网络容易过拟合，在测试集上表现较差，所以本文采用了 2 种不同的策略来扩充数据集。第一种数据增强的策略是图像旋转。本文对图片进行随机的旋转，旋转的角度在(-15°，15°)间。第二种数据增强的策略是图像平移，本文对图片进行随机的上，下，左，右平移，平移的范围为人脸检测框的高或宽的-0.05 到 0.05 间。在图片进行变换的同时，将标注的坐标同样进行相应的变换。最终，通过数据增强策略，本文将原始的 3148 张图片扩充到了 39561 张图片。

评估策略

本文采用通用的平均误差来评估关键点定位的精确性。平均误差的衡量是通过计算预测的关键点坐标和标注的坐标的距离，然用内眼距离归一化得到的。具体采用的公式为：

$$err = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{M}\sum_{j=1}^{M}|p_{i,j}-g_{i,j}|_2}{|l_i - r_i|_2}$$

其中，N为测试样本数，M为点的个数，p为预测的点的坐标，g为标注的点的坐标，l和r为左眼和右眼的坐标。

3.2 网络的训练过程

本文的方法利用开源的深度学习框架 caffe[11]实现。首先是训练数据的准备，跟其它论文的做法一致，人脸框用标注的关键点计算得到。接着，根据得到的人脸框对图片进行上，下，左，右扩大，扩大的范围为人脸的长，宽的 0.05 倍，然后得到截取的人脸图像。之后，将图片缩放到 224 x 224 大小，将标注的关键点坐标和头部三维姿势归一化到(-1，1) 间。本文利用随机梯度算法进行训练。网络模型的参数用在 ImageNet 上预训练过的 VGG-S[12]模型作为初始化。本文将训练过程中的学习率设置为 0.0001，mini-batch 大小设置为 70，动量设置为 0.9，weight decay 设置为 0.0001。网络迭代 140 万轮后停止训练。

3.3 误差分析

首先进行头部三维姿势的评估。利用训练好的深度网络模型在 300-W 的 689 张测试集上进行测试。利用预测的三维姿势和标注的三维姿势计算平均的绝对误差，得到 pitch，yaw，roll 的误差分别为 3.7°， 2.3° 和 1.3°。图 3 为本文的头部姿势的误差与 Yang[6]的比较结果，本文的方法在 3 个维度上的误差都有降低。

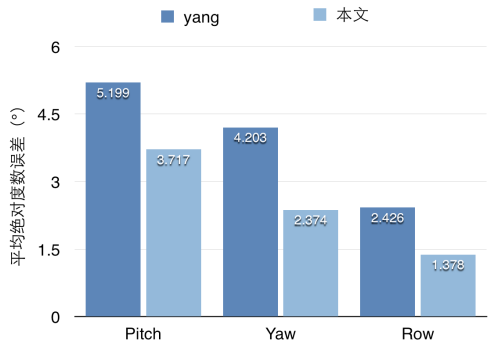

图 3 头部三维姿势误差对比

之后进行关键点定位的评估。在多任务的训练

中，由于不同任务关注点的不同，某一个任务的准确性很可能被另一任务所影响。因此，在进行三维关键点联合头部姿势的训练时候，我们比较加不加头部姿势，会不会影响关键点的定位准确性。实验的结果显示，关键点定位的准确性并没有因为头部姿势预测的模型而变低，反而，还有一定的促进作用。图 4 比较了有无头部姿势的网络在挑战集上的误差曲线。

同时，我们将得到的深度网络模型和非深度特征的模型作为比较。如图 5, 当有较大的头部姿势变化时，ERT 算法容易有较大的点的偏移，而本文的模型定位结果显得更加稳定。

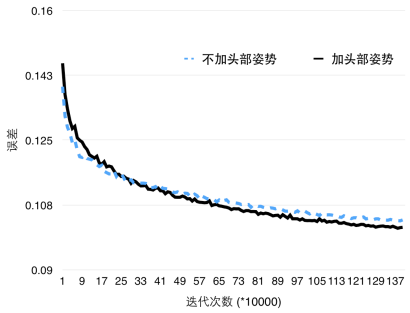

图 4 在测试集上，比较有无头部姿势的关键点定位误差曲线

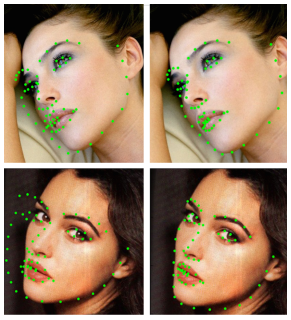

图 5 效果对比（第一列为 ERT[13]，第二列为本文效果）

3.4 效率分析

本文的实验在一台 cpu 为 Inter Xeon E5-2690 v2 3.00GHZ， gpu 为 Tesla K20 的机器上进行。每张图的网络前向计算平均为 18ms，在测试的时候，可以达到 55 fps。

3.5 与公开的算法比较

表 1 为本文与现有的公开的算法的比较结果，所比较的方法包括 PCRR[14]，GN-DPM[15]，CFAN[4]，ESR[16]，SDM[1]，ERT[13]，LBF[2]。可见本文的算法在 300-W 的 3 个评测子集上的误差较现有算法都有降低，特别是在挑战集上。这个结果更表明了本文基于深度网络的三维头部姿势联合关键点定位的算法的准确率和鲁棒性。

表 1 在 300-W 评测集上的结果

方法	普通集	挑战集	所有集
PCRR[14]	6.18	17.26	8.35
GN-DPM[15]	5.78	-	-
CFAN[4]	5.50	16.78	7.69
ESR[16]	5.28	17.00	7.58
SDM[1]	5.57	15.4	7.5
ERT[13]	-	-	6.4
LBF[2]	4.95	11.98	6.32
本文	**4.97**	**10.33**	**6.02**

3.6 算法结果可视化

下面是本文的算法在测试集上的结果。图 7 为在 IBUG，LFPW 测试集和 Helen 测试集上的结果。图 6 为网络上的随机检索到的图片的测试结果(这些图片利用 dlib[17]进行人脸检测）。

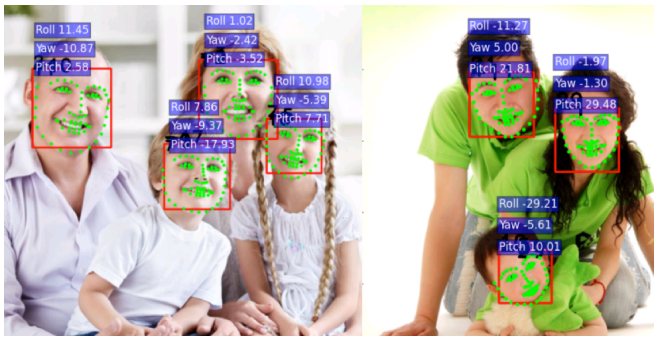

图 6 网络上的图片的定位结果

4 结论

本文提出一种新颖的基于深度学习的三维姿态联合面部关键点的多任务人脸定点方法。将原始图像的像素信息作为输入，将标注的面部关键点和头部三维姿势作为监督信息，通过端到端的联合模式，学习人脸图像到关键点坐标的一系列的非线性映射。这种方法利用整体的人脸上下文信息和关键点之间的内在约束，较为鲁棒，能够处理有较大头部姿势的关键点定位；同时轻量级，速度可达 55 fps；最为重要的是能够在得到关键点的位置同时，在几乎不增加计算代价的情况下获得头部的三维姿势。后期的研究工作可以是将本文的方法应用于三维人脸对齐中去。

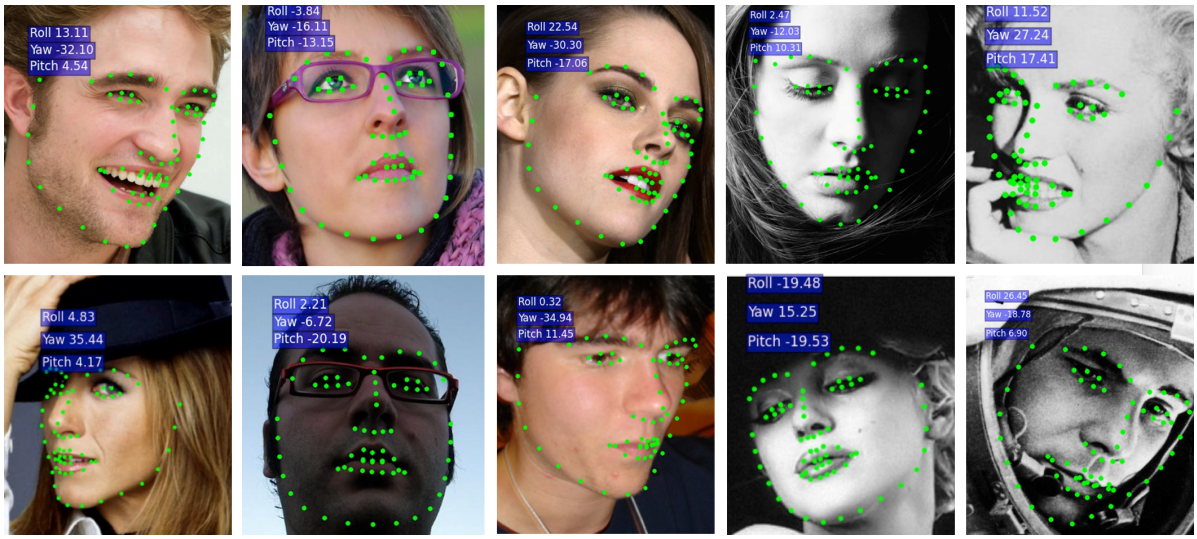

图 7 在 IBUG，LFPW 测试集和 Helen 测试集上的定位结果

参 考 文 献

[1] Xiong X, Torre F. Supervised descent method and its applications to face alignment[C]. Proceedings of the IEEE conference on computer visionand pattern recognition, 2013: 532-539.

[2] Ren S, Cao X, Wei Y, et al. Face alignment at 3000 fps via regressing local binary features[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1685-1692.

[3] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection[C]. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013: 3476-3483.

[4] Zhang J, Shan S, Kan M, et al.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, Computer Vision–ECCV 2014: Springer, 2014: 1-16.

[5] Zhang Z, Luo P, Loy C C, et al.: Facial landmark detection by deep multi-task learning, Computer Vision–ECCV 2014: Springer, 2014: 94-108.

[6] Yang H, Mou W, Zhang Y, et al. Face Alignment Assisted by Head Pose Estimation[J]. arXiv preprint arXiv:1507.03148, 2015.

[7] Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. 300 faces in-the-wild challenge: The first facial landmark localization challenge[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013: 397-403.

[8] Belhumeur P N, Jacobs D W, Kriegman D J, et al. Localizing parts of faces using a consensus of exemplars[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2930-2940

[9] Zhu X, Ramanan D. Face detection, pose estimation, and landmark localization in the wild[C]. Computer Vision and Pattern Recognition, 2012: 2879--2886.

[10] Le V, Brandt J, Lin Z, et al. Interactive facial feature localization[C]. European Conference on Computer Vision, 2012: 679-692

[11] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of the ACM International Conference on Multimedia, 2014: 675-678.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. ICLR, 2015.

[13] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1867-1874.

[14] Burgos-Artizzu X P, Perona P, Doll\'A R, Piotr. Robust face landmark estimation under occlusion[C]. International Comference on Computer Vision, 2013: 1513-1520.

[15] Tzimiropoulos G, Pantic M. Gauss-newton deformable part models for face alignment in-the-wild[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1851-1858.

[16] Cao X, Wei Y, Wen F, et al. Face alignment by explicit shape regression[J]. International Journal of Computer Vision, 2014, 107(2): 177-190.

[17] Dlib. http://dlib.net/.