计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：一种端到端的人脸对齐方法
作者：康智慧，王全玉，王战军
DOI：10.19678/j.issn.1000-3428.0059225
网络首发日期：2020-10-15
引用格式：康智慧，王全玉，王战军．一种端到端的人脸对齐方法．计算机工程.
https://doi.org/10.19678/j.issn.1000-3428.0059225

网络首发时间：2020-10-15 10:26:23
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.tp.20201014.1821.006.html

计 算 机 工 程
Computer Engineering

一种端到端的人脸对齐方法

康智慧[1]，王全玉[1]，王战军[2]

(1.北京理工大学，计算机科学与技术学院，北京 100081；2.北京理工大学，人文与社会科学学院，北京 100081)

摘　要：人脸对齐在人脸研究中扮演着重要的角色。现存的人脸对齐方法大多是非端到端的，中间过程需要大量的人工干预，导致方法具有不稳定性。提出了一种端到端的基于深度学习的人脸对齐方法，该方法的网络基于 MobileNet 系列网络的子模块，采用类 VGG 结构的方式进行搭建。该方法将整张图片作为输入，采用基于深度可分离卷积模块进行特征提取，采用改进的倒残差结构避免网络训练过程的梯度消失的同时减少了特征的损失。采用眼间距离作为正规化方法，设计的网络在 300W 人脸数据集上进行测试，并与先进的方法进行比较，实验结果证明了算法在保证良好的精度的同时也具有良好的实时性。

关键词：人脸对齐；人脸特征点；特征提取；深度可分离卷积

开放科学（资源服务）标志码（OSID）：

An end-to-end face alignment method

Kang Zhihui[1], Wang Quanyu[1], Wang Zhanjun[2]

(1. School of Computer Science and Technology, Beijing Institute of Technology University, Beijing 100081, China

2. School of Humanities and Social Sciences, Beijing Institute of Technology University, Beijing 100081, China)

【Abstract】Face alignment plays an important role in face research. Most of the existing face alignment methods are not end-to-end, and the intermediate process requires a lot of manual intervention, which leads to instability of the method. An end-to-end face alignment method based on deep learning is proposed. The network of this method is based on the sub-modules of the MobileNet series network and is built in a similar VGG structure. This method takes the entire picture as input, uses a depth-based separable convolution module for feature extraction, and uses an improved inverted residual structure to avoid the disappearance of gradients in the network training process and reduce the loss of features. Using the distance between eyes as a normalization method, the designed network is tested on a 300W face dataset and compared with advanced methods. The experimental results prove that the algorithm has good real-time performance while ensuring good accuracy.

【Key words】face alignment; facial landmark; feature extraction; depth separable convolution

DOI:10.19678/j.issn.1000-3428.0059225

0　概述

人脸对齐是在给定的图像中确定人脸主要器官（如：眼睛、鼻子、嘴巴等）的轮廓点位置。这些轮廓点在人脸研究中被称为人脸特征点或人脸关键点。它对人脸研究具有重要意义，在人脸验证、人脸表情识别、人机交互以及人脸动画技术方面起着不可代替的作用。

人脸关键点检测研究由来已久，许多优秀的方法不断被提出。其中根据是否有参数分为参数化方法和非参数化的方法。参数化方法中具有代表性的人脸关键点检测方法有：基于主动形状模型（Active Shape Model, ASM）[1]的方法，该方法是基于局部特征的，独立考虑每个关键点周围的变化，通过训练学习到的全局形状来检测人脸关键点；基于主动外观模型（Active Appearance Model, AAM）[2]的方法，该方法是对 ASM 的一种改进和优化，

基金项目：国家自然科学基金(71834001)

作者简介：康智慧，1993，女，硕士研究生，主研方向为人机交互、深度学习；王全玉，副教授、博士；王战军，教授、博士。**E-mail：18811708090@163.com**

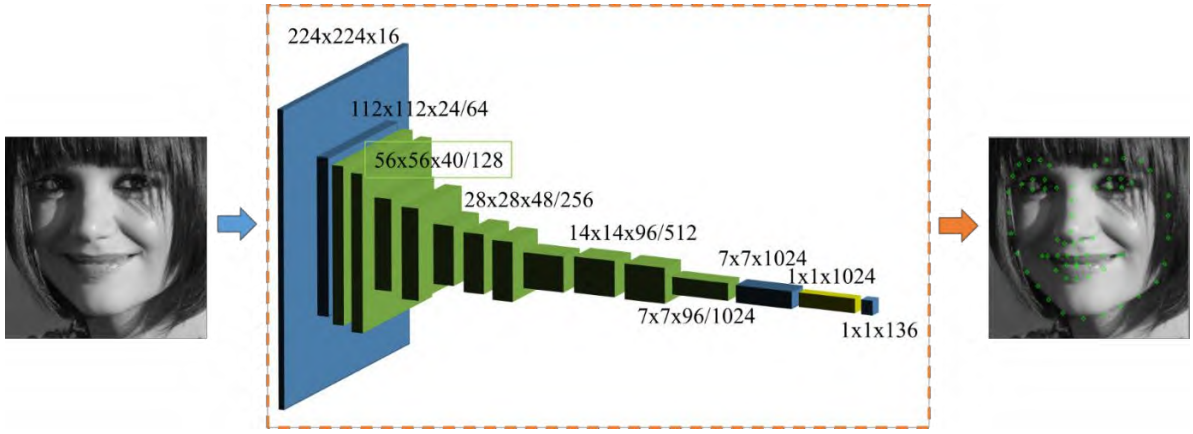


图 1 人脸对齐网络结构图

Fig.1 The structure of face alignment network

同时考虑面部形状和纹理，以便更精确地检测人脸关键点。基于非参数化的方法有：基于图模型的马尔可夫随机场的建模；基于级联回归的方法，该方法采用从粗略估计到精细估计的方式对人脸关键点进行直接估计，并不需要对任何模型进行学习和建模；基于深度学习的方法，随着深度学习研究的不断深入，其在人脸关键点检测方面的应用也随之增加。这种方法完全通过对训练数据的学习自动地生成人脸关键点检测模型，无需人工干预。这种超强的学习能力，使其成为近年来广泛使用的方法。然而，即使已经存在大量先进的人脸关键点检测算法，人脸关键点检测任务仍然面临很多挑战。首先，不同的人脸表情、不同的头部姿势以及遮挡、光线等外在条件都会影响人脸关键点的位置和外观特征，从而影响人脸关键点检测的准确性和可靠性。其次，现存的人脸关键点检测方法大多不是端到端的检测，中间过程需要大量的人工干预，使得模型不具有良好的稳定性。再次，现存的大多数方法输入的是人脸的局部特征，关键点定位不具有整体的稳定性。很多方法虽然具有良好的定位精度，但是其模型过大，在实时性方面还有待提高。

为克服以上提到的困难，本文的主要工作如下：

1.我们采用了整张图像作为网络的输入，这样的设计可以保证人脸对齐具有全局性。

2.设计一种端到端的网络结构，以减少中间过程的人工干预带来的不确定性。

3.为了使人脸对齐结果具有较好的准确性和实时性，采用基于深度可分离卷积[3]模块构建一个类VGG[4]结构的网络进行人脸特征提取与关键点定位。

1 相关工作

在计算机视觉领域，人脸对齐有很长的历史。在研究初期，人脸关键点检测大都基于传统机器学习，其中的经典方法是基于 AAM[2]的算法，其采用的是人脸形状和外观两种特征进行人脸关键点检测，随后相关文献[5-6]在 AAM 基础上进行优化。其中主要有两个优化方向：一方面是对关键点准确率进行提升，另一种是对拟合的速度进行提升。

随着深度学习的普及以及计算机性能的提升，人们开始采用深度学习的方法对人脸关键点进行检测，2013 年 Sun 等人[7]首次提出采用深度学习方法对人脸关键点进行检测和跟踪，该算法采用了三层级联卷积神经网络（Convolutional Neural Network, CNN）的结构对人脸的 5 个关键点进行检测，在当时取得了比较先进的检测结果。

文献[8]采用由粗到精的深度学习的方法对人脸的 68 个关键点进行检测，该方法的贡献在于检测的关键点更多，而且降低了传统卷积网络的网络复杂性和减轻训练模型的负担。文献[9]提出，人脸关键点检测不是一个独立的问题，对人脸关键点位置的估计会受到许多因素的影响，因此提出了一种基于多任务学习的人脸关键点检测算法（Multi-task convolutional neural networks, MTCNN）。当人脸有遮挡或者人脸姿势变化较大时，该算法确实获得了较高的准确率。为了能够更好地克服头部姿势运动带来的困扰，2017 年 Kowalski 等人[10]提出了（Deep Alignment Network, DAN）人脸关键点检测算法，该方法在人脸关键点检测的整个过程中都采用整张

脸作为输入，这使得它对于头部运动较大变化时关键点的检测都很稳定，这也是本文的创新动机来源之一。除了卷积神经网络，递归神经网络（Recurrent Neural Network, RNN）也被用于人脸关键点的检测与跟踪。

2 用于人脸对齐的端到端网络

在这部分，我们首先对提出的用于人脸对齐的端到端的网络模型进行总体的概述，然后分别对每个子模块进行详细介绍。子模块主要包含深度可分离卷积模块、改进的倒残差结构和 Squeeze-and-Excitation 结构[11]。最后，介绍本文设计的网络结构的具体实现细节。

2.1 方法概述

在本文中，设计了一种端到端的网络模型对图像中的 N 个人脸关键点进行定位。图 1 展示了该方法的整体结构。

本文采用了基于深度可分离卷积的方法对图像中的人脸关键点进行定位。采用该网络结构的主要原因是：深度可分离卷积可以采用不同尺寸的视野域，不同的视野域可以提取出不同的图像特征。其次，在计算量一定的情况下，与传统的全卷积网络相比，深度可分离卷积可以被设计为更深层次的网络，因此其采集到的图像特征会更加丰富。采用深度可分离卷积神经网络可以减少网络模型中的参数个数，缩短计算时间，从而提升效率。由于 VGG 结构在目标检测中具有良好的表现，因此采用类 VGG 的结构进行网络构建，来提高人脸对齐的精度。

2.2 深度可分离卷积结构

在特征提取网络中，主要是从图像的像素信息中提取与关键点定位相关的特征，本文采用基于深度可分离卷积的网络结构对图像信息进行提取。深度可分离网络是 Howard 等人[3]在 2017 年提出的。视野域在深度卷积操作中对应的就是卷积核，选择不同尺寸的卷积核进行操作，就意味着考虑的图像周围的环境不同，因此提取到的特征就会不同。一个标准的卷积既可以卷积又可以将输入合并为一组新的输出，而深度可分离卷积包含两部分：一个专门用于卷积的层称为深度卷积层，一个专门用于特征生成的称为点式卷积层。深度卷积层将卷积按照图像通道数均匀分解，点式卷积层采用 1*1 的卷积实现。深度可分离卷积结构图如图 2 所示。

假设输入特征图为 $D_1\times D_1\times M$，输出的特征图为 $D_2\times D_2\times N$，卷积核的大小为 K，若采用普通的卷积操作，则计算成本为 $K\times K\times M\times N\times D_1\times D_1$。若采用深度可分离卷积操作，深度卷积的计算成本为 $K\times K\times M\times D_1\times D_1$，1*1 卷积操作的计算成本为 $M\times N\times D_1\times D_1$，因此深度可分离卷积的总的计算成本为 $K\times K\times M\times D_1\times D_1+M\times N\times D_1\times D_1$。仅一次卷积操作，在计算成本上，采用深度可分离卷积仅为普通卷的 $1/N+1/(K^2)$，由此可见采用深度可分离结构比普通的卷积网络的计算成本低，因此在计算量一定的情况下，深度可分离卷积能够提取到更深层次的图像特征。因此，本文设计的网络结构在设备的计算能力有限或者对实时性要求较高的场景下具有一定的优势。

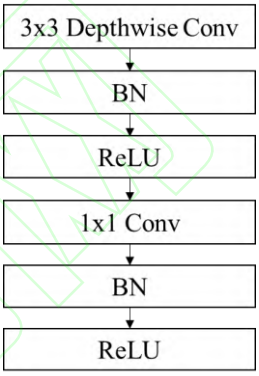


图 2 深度可分离卷积

Fig.2 Depth separable convolution

为了加速模型收敛和防止过拟合，在每个卷积分支的卷积后都会进行一次批量正规化（batch normalization），采用的激活函数是 ReLU6，ReLU6 的计算方式如式（1）：

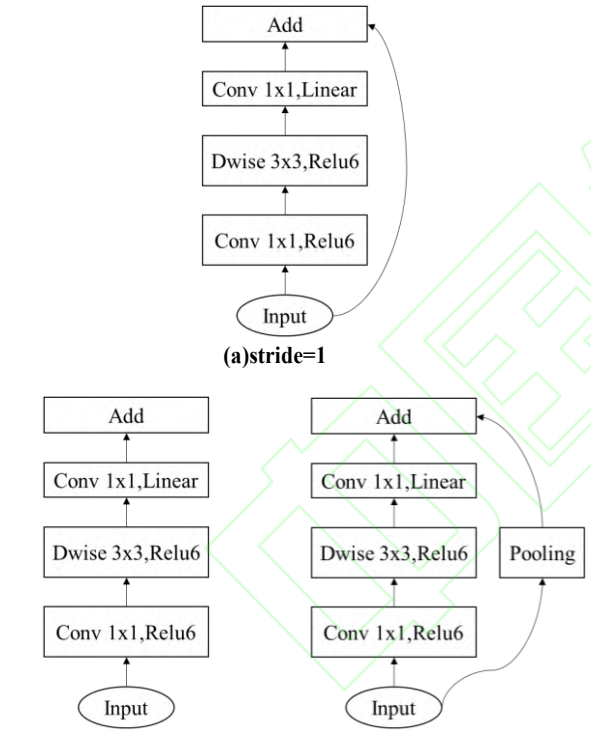
$$\mathrm{Re}LU6[x]=\begin{cases}0 & (x<0)\\ x & (0\le x<6)\\ 6 & (x\ge 6)\end{cases} \tag{1}$$

2.3 改进的倒残差结构

为了避免训练阶段出现梯度消失等情况，采用 MobileNet 系列中的一种称为“倒残差”[11]模块，即在每次深度卷积之后再与此次深度卷积之前的图像特征做合并，作为下一次深度卷积的输入。但是这一“倒残差”的变换与传统的残差网络的变换过程有所不同，由于深度可分离卷积不能改变通道数，通道数量越多采集到的特征也就越多，因此为了提取到更多的特征，在进行深度卷积之前需要先增加通道数，因此“倒残差”结构的变换过程在通道数量上的变化恰好与传统的残差结构的变换过程相反，即倒残差的变换过程是“扩展-深度分离

卷积-压缩”。

原始的倒残差结构在输入尺寸与深度卷积后的尺寸相同的情况下直接合并通道，若两者尺寸不同则直接采用卷积后的特征作为下一模块的输入，这在一定程度上损失了图像特征。为了最大限度地避免图像特征的丢失，本文对输入尺寸与深度可分离卷积后的尺寸不同的情况做了改进，即将输入的尺寸经过池化变换后生成与深度可分离卷积输出尺寸相同的特征图，然后将两者合并，作为下一次卷积的输入。改进前后的倒残差结构在两种情况下的结构如图 3 所示，图 3(a)表示卷积的步长 stride=1 时的情况，即直接将输入与卷积之后的输出合并，图 3(b)是原始倒残差结构卷积步长为 stride=2 时的情况，即直接将卷积后的输出作为下一卷积的输入，图 3(c)是卷积步长 stride=2 时改进后的结构，将原始的输入进行池化操作后与卷积后的输出进行合并。



(a)stride=1

(b)stride=2 原始结构　　(c)stride=2 改进后的结构

(b)stride=2 Origin structure　　(c)stride=2 Improved structure

图 3　改进前后的倒残差结构

Fig.3 Improved inverse residual

很明显，当 stride=2 时，在改进后的结构中，不仅包含了原始结构的卷积操作的输出特征，另外增加了对输入进行池化后的特征，池化后的特征在一定程度上保留着原始输入的特征，与原始模型相比较，用于下一次卷积的特征内容更加丰富。

2.4 Squeeze-and-Excitation 结构

Squeeze-and-Excitation 结构由 Hu 等人[12]提出，该结构能够学习图像通道之间的关系。图 4 展示了 Squeeze-and-Excitation 模块详细结构，$X \in R^{H' \times W' \times C'}$ 为输入，F_{tr} 为普通的卷积操作，$U \in R^{H \times W \times C}$ 为 X 经过 F_{tr} 卷积后的输出，$F_{sq}(\cdot)$ 为全局的平均池化操作，该操作是 Squeeze 过程，$F_{ex}(\cdot,W)$ 为两个连续的全连接操作，全连接的输出维度为 1×1×C，该过程称为 Excitation，$F_{scale}(\cdot,\cdot)$ 为采用 hard_sigmoid 激活函数的激活层，目的是将最后的输出值限定在[0,1]之间，并将该值作为每个通道的系数乘以特征 U，使得到的特征中重要的特征增强，不重要的特征减弱，最终提取到的特征指向性更强。

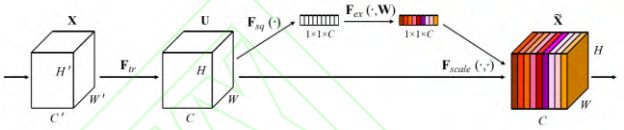


图 4 Squeeze-and-Excitation 模块结构[11]

Fig.4 The structure of Squeeze-and-Excitation

在卷积操作 F_{tr} 中，输入为 $X \in R^{H' \times W' \times C'}$，卷积核表示为 $V=[v_1, v_2, \ldots, v_c]$，卷积操作的输出表示为 $U=[u_1, u_2, \ldots, u_c]$。其中 v_c 是第 c 个卷积核参数，对应的输出 u_c 可以表示为下式（2）：

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \qquad (2)$$

在上式中 * 表示卷积操作，$v_c = \left\lfloor v_c^1, v_c^1, \cdots, v_c^{C'} \right\rfloor$，$X = \left\lfloor x^1, x^2, \cdots, x^{C'} \right\rfloor$，$u_c \in R^{H \times W}$。$v_c^s$ 是二维空间卷积核，其代表着 v_c 的一个通道，对应于 X 的单个通道。根据式中的表达可知输出是由所有通道之和产生的，通道之间的依赖关系隐藏在 v_c 中。

每个卷积核都只能对局部区域进行操作，因此输出的 u_c 都无法利用该区域以外的上下文信息。为了克服这一问题，采用全局的平均池化的方法将全局信息压缩到一个通道中，生成通道的统计信息。统计信息 $z \in R^C$ 是通过将 U 的空间维度减小成 $H \times W$ 实现的。因此 z 的第 c 个统计信息如式（3）所示。

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \qquad (3)$$

在获得统计信息后，为了捕捉通道之间的依赖关系，采用下式（4）进行全连接操作。这一过程也就是 Excitation 操作。

$$s=F_{ex}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1z)) \quad (4)$$

在上式中 δ 表示 ReLU 函数，$W_1\in R^{\frac{C}{r}\times C}$，$W_2\in R^{\frac{C}{r}\times C}$，$\sigma$ 表示 sigmoid 激活函数。运算的流程图如图 5 所示。

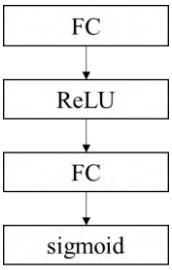


图 5 Excitation 结构

Fig.5 Excitation structure

2.5 人脸对齐网络

根据前面的分析可以得出：深度可分离卷积结构与传统的卷积操作相比具有计算成本低的特点，因此当计算成本一定的情况下，采用深度可分离卷积可以提取到更深层次的网络结构。图像通道数越多，提取到的图像特征会越多，但是深度可分离卷积又不能改变图像的通道数，因此采用改进的倒残差结构对图像的通道数进行增加，同时在原始深度卷积的特征的基础上增加了对输入的池化输出部分特征，使得用于下一次卷积的输入特征增加。采用 Squeeze-and-Excitation 可以学习到不同通道之间的关系，更加有利于最终人脸关键点的定位。搭建类 VGG 结构是由于 VGG 结构在目标检测中表现良好，说明这样的结构是利于特征提取的。

因此本文基于深度可分离卷积结构、改进的倒残差机构和 Squeeze-and-Excitation 结构构建了一个类 VGG 结构的人脸对齐网络。

在人脸特征提取网络中，输入是人脸图像 $X\in R^{W'\times H'\times C'}$，W'为图像的宽度，H'为图像的高度，C'为图像的通道(RGB)。在本文中，我们使用的是 224x224x1 的二维图像，经过多次的深度可分离卷积后提取出丰富的人脸特征，用于最终的人脸关键点定位。

本文设计的网络输出为对人脸的 N 个关键点进行定位，采用(x,y)表示人脸关键点坐标位置，最终输出的结果为（N，2）结构。在本文中，我们对人脸的 68 个关键点进行测试时 N 为 68，当仅对人脸内部关键器官眼睛、鼻子、嘴巴进行测试时，N 为各个器官的轮廓关键点数目。

3 相关实验

3.1 数据集

文献[13-17]介绍了用于人脸关键点定位的各种数据集，300W[18]数据集是来自文献[13-17]中介绍的 LFPW、HELEN、AFW、IBUG 和 300W 私有测试集这五个数据集的集合。300W 数据集具有图像内容广泛、数据量大等优点，数据集对图像中的人脸标记了 68 个关键点的坐标，根据文献[10]的划分方法将数据集划分为训练集和测试集。

训练集部分包括 AFW 数据集以及 LFPW 和 HELEN 的训练子集，共计 3148 张图像。测试数据由其余数据集组成：IBUG，300W 专用测试集，LFPW 和 HELEN 测试子集。为便于与现存的方法进行比较，我们将该测试数据分为四个子集：

(1) 普通数据集，包括 LFPW 和 HELEN 测试子集，共计 554 张图像，该测试集的特点是图像均为正面人脸，可以很容易地定位到人脸关键点位置。

(2) 具有挑战性数据集，包括 IBUG 数据集，共有 135 张图像，该测试集的特点是这类图像或者为侧面人脸，或者是光线不佳时的人脸，该数据集中的人脸关键点不易被定位。

(3) 由 (1)、(2) 共同构成的 300W 公共测试集，共计 689 张图像。

(4) 300W 专用测试集，共 600 张图像。

其中普通数据集的样例如图 6(a)所示，具有挑战性的数据集样例如图 6(b)所示。

(a) 普通数据集样例　　(b) 挑战性数据集样例

图 6 300W 测试集样例

Fig.6 Samples of 300W test dataset

3.2 评估方法

对于人脸关键点的检测，近来的有关研究中，针对单个面部图像的几种面部特征点检测误差的度量有如下几种方法：

1）预测关键点和真实关键点之间的平均距离除以眼间距离（外眼角之间的距离），如图 7 所示。

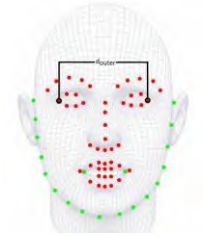


图 7 两眼间外侧距离

Fig.7 Outer distance between eyes

2）预测关键点和真实关键点之间的平均距离除以瞳孔间距离（眼中心之间的距离）；

3）预测关键点和真实关键点之间的平均距离除以边界框的对角线。

本文采用第（1）种归一化方法，以便与先进的算法进行比较。均方误差的计算方法如式（5）所示。

$$RMSE=\frac{\sum_{i=1}^{N}\sqrt{\left(x_i^f-x_i^g\right)^2+\left(y_i^f-y_i^g\right)^2}}{d*N} \tag{5}$$

在上式中，$\left(x_i^f,y_i^f\right)$表示第 i 个关键点的预测坐标，$\left(x_i^g,y_i^g\right)$表示第 i 个关键点的实际坐标值，N 表示预测的关键点总数，d 为两眼外眼角之间的欧氏距离。在本文的研究中，当对整体人脸关键点进行评估时，N 为 68，当要对面部的每个器官分别进行评估时，N 取相应的值，对单个眼睛轮廓的关键点进行评估时，N 为 8，对鼻子轮廓关键点进行评估时，N 为 9，对嘴巴轮廓关键点进行评估时，N 为 18。

另外，本文还采用累积误差分布（CED）曲线下的面积（$AUC_{0.08}$）和失败率进行结果评估。

3.3 实验安排及数据集处理

本文主要采用平均误差、失败率等对实验结果进行评估。首先，对本文设计的网络结构进行 68 个人脸关键点的定位评估，并与先进的人脸对齐方法进行比较。

其次，分别对人脸的眼睛、鼻子和嘴巴这三个主要器官的轮廓进行评估，并与现存的眼睛、鼻子、嘴巴的定位方法进行比较。

为了提高训练模型的性能，本文进行了数据增强，即对原始数据进行平移，放大，缩小和旋转等操作，最终将每个原始图像扩充为 10 张，这样获得的训练图像共计 31480 张，测试图像是原来的 10 倍。

3.4 实验结果

采用 300W 训练集进行模型的训练，并分别在 300W 的 4 个子测试集上进行测试，在以下的实验结果中，平均误差和失败率。

（1）68 个关键点定位结果

首先在 300W 的公共测试集的普通数据集和具有挑战性数据集上对设计的网络结构分别进行了测试。表 1 记录了先进的人脸对齐方法和本文设计的人脸对齐方法的平均误差的测试结果。

表 1 人脸对齐方法在 300W 公共测试集上的平均误差（%）

Table 1 The average error of the face alignment method on the 300W public test set（%）

方法	普通测试集	挑战性测试集	公共测试集
CDM [19]	10.10	19.34	11.94
DRMF [20]	6.65	19.79	9.22
CFAN [21]	5.50	16.78	7.58
ESR [22]	5.28	17.00	7.58
SDM [23]	5.57	15.40	7.50
CFSS [24]	4.73	9.98	5.76
TCDCN [25]	4.80	8.60	5.54
DAN [10]	3.19	5.24	3.59
本文方法	3.97	7.24	4.61

由表 1 中的数据可知，我们提出的方法在所有列出的关键点定位方法中仅次于 DAN 算法的结果，但是我们提出方法的模型简单，参数数量相较于 DAN 较少。

表 2 人脸对齐方法在 300W 公共测试集上的 AUC 和失败率（%）

Table 2 AUC and failure rate of face alignment method on 300W public test set（%）

方法	$AUC_{0.08}$	失败率
ESR [22]	43.12	10.45
SDM [23]	42.94	10.89
CFSS [24]	49.87	5.08
MDM [26]	52.12	4.21
DAN [10]	55.33	1.16
本文方法	44.28	6.38

然后在 300W 公共测试集上采用 AUC 和错误率对设计的网络模型进行评估，其中将可接受的误差设置为 0.08，实验结果如表 2 所示。

通过表 2 中的数据可以得知，我们提出的算法对 68 个关键点的 $AUC_{0.08}$ 仅比 ESR 和 SDM 稍高一些，但是失败率却低于这两种算法，在此种评估方法中我们设计的模型处于居中偏上的水平。

（2）人脸内部关键器官轮廓点的定位结果

同时，我们分别对眼睛、鼻子、嘴巴这三个主要的

面部器官的轮廓进行了测试，并与现有的面部器官的定位结果进行比较，结果记录在表 3 中。

表 3.本文算法在 300W 公共测试集上测试的平均误差（%）

Table 3 The average error of the algorithm tested on the 300W public test set（%）

人脸器官	普通数据集	挑战性数据集	整体数据集
眼睛（双眼）	2.74	5.55	3.29
鼻子	2.60	5.65	3.19
嘴巴	3.03	5.59	3.53
内部关键点	2.79	5.60	3.34

从对人脸的关键器官的关键点定位来看，在普通数据集上这三个器官的定位给误差会比最先进的算法误差还要小，在挑战性数据集上，只比 DAN 的结果稍差一点，在整体数据集上的测试结果都优于其余方法。由此推断出，本文所提出算法的较大误差时存在于人脸外轮廓的定位上的，因此本文所提出的方法适用于对眼睛鼻子嘴巴定位精度较高且对人脸外轮廓定位精度相对不高的人脸任务中。

为了评估提出算法的稳定性，在 300W 的专用的私有测试集上对人脸内部关键点的平均误差、$AUC_{0.08}$ 和失败率进行评估，与先进方法的比较如下表 4 所示。

表 4. 300W 私有测试集上的平均误差 AUC 和失败率（%）

Table 4 Average error AUC and failure rate on 300W private test set（%）

方法	平均误差	$AUC_{0.08}$	失败率
MDM[26]	5.05	45.32	6.80
DAN [10]	4.30	47.00	2.67
内部点	3.99	50.76	2.00

由表中数据可以看出，仅对内部 51 个关键点进行评估时，$AUC_{0.08}$ 的值要远高于最好的算法 DAN 的 $AUC_{0.08}$ 值，失败率比 DAN 算法降低了 0.67，比 MDM 方法降低了 4.8，说明本文设计的方法有良好的关键点定位效果。

表 4 中的结果与在 300W 公共数据集上的得到的结论一致，充分说明了本文设计算法的有效性和稳定性。

（3）人脸对齐性能对比

本文采用 python 语言实现的算法在 NVIDIA GeForce RTX 2060 GPU 笔记本电脑上的人脸对齐速度为 65fps，为了证明本文提出算法在性能方面的优势，在同样的硬件条件下对 python 实现的 DAN 算法进行了性能评测，其人脸对齐速度为 50fps，这一结果充分说明了本文提出方法在性能上优于 DAN 算法。

4 结束语

本文基于 MobileNet 系列的子模块设计了一种端到端的用于人脸对齐的网络，该网络是基于深度可分离卷积构建的，对倒残差模块进行了改进，尽量减少特征的损失。实验表明，本文设计的算法对人脸的 68 个关键点的定位，在定位精度上优于大部分先进算法；对面部主要器官的 51 个轮廓关键点的定位误差明显小于最先进的算法的定位误差，在性能方面具有良好的实时性。因此算法更适用于对眼睛、鼻子、嘴巴定位精度较高且对人脸外轮廓定位精度相对不那么高的人脸任务中，如卡通人脸动画。因此，后续的优化内容是提高本算法对人脸外部轮廓关键点的定位精度，使算法适用于更广的人脸研究相关领域。

参考文献

[1] Cootes T F, Taylor C J, Cooper D H, et al. Active shape models—their training and application[J]. Computer Vision and Image Understanding, 1995, 61(1): 38-59.

[2] Cootes T F, Edwards G J, Taylor C J, et al. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 681-685.

[3] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[5] Saragih J, Goecke R. A nonlinear discriminative approach to AAM fitting[C]//2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-8.

[6] Tzimiropoulos G, Pantic M. Optimization Problems for Fast AAM Fitting in-the-Wild[C]. international conference on computer vision, 2013: 593-600.

[7] Sun Y, Wang X, Tang X, et al. Deep Convolutional Network Cascade for Facial Point Detection[C]. computer vision and pattern recognition, 2013: 3476-3483.

[8] Zhou E, Fan H, Cao Z, et al. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade[C]. international conference on computer vision, 2013: 386-391.

[9] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional

Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[10] Kowalski M, Naruniec J, Trzcinski T P, et al. Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment[C]. computer vision and pattern recognition, 2017: 2034-2043.

[11] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]. computer vision and pattern recognition, 2018: 4510-4520.

[12] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 1-1.

[13] Jesorsky O, Kirchberg K J, Frischholz R, et al. Robust Face Detection Using the Hausdorff Distance[J]. Lecture Notes in Computer Science, 2001: 90-95.

[14] 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.

[15] Kostinger M, Wohlhart P, Roth P M, et al. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization[C]. international conference on computer vision, 2011: 2144-2151.

[16] Belhumeur P N, Jacobs D W, Kriegman D J, et al. Localizing parts of faces using a consensus of exemplars[C]. computer vision and pattern recognition, 2011: 545-552.

[17] Zhu X, Ramanan D. Face detection, pose estimation, and landmark localization in the wild[C]. computer vision and pattern recognition, 2012: 2879-2886.

[18] Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. A Semi-automatic Methodology for Facial Landmark Annotation[C]. computer vision and pattern recognition, 2013: 896-903.

[19] Xiong X, La Torre F D. Supervised Descent Method and Its Applications to Face Alignment[C]. computer vision and pattern recognition, 2013: 532-539.

[20] Asthana A, Zafeiriou S, Cheng S, et al. Robust Discriminative Response Map Fitting with Constrained Local Models[C]. computer vision and pattern recognition, 2013: 3444-3451.

[21] Zhang J, Shan S, Kan M, et al. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment[C]. european conference on computer vision, 2014: 1-16.

[22] Cao X, Wei Y, Wen F, et al. Face alignment by Explicit Shape Regression[C]. computer vision and pattern recognition, 2012: 2887-2894.

[23] Xiong X, La Torre F D. Supervised Descent Method and Its Applications to Face Alignment[C]. computer vision and pattern recognition, 2013: 532-539.

[24] Zhu S, Li C, Loy C C, et al. Face alignment by coarse-to-fine shape searching[C]. computer vision and pattern recognition, 2015: 4998-5006.

[25] Zhang Z, Luo P, Loy C C, et al. Facial Landmark Detection by Deep Multi-task Learning[C]. european conference on computer vision, 2014: 94-108.

[26] Trigeorgis G, Snape P, Nicolaou M A, et al. Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment[C]. computer vision and pattern recognition, 2016: 4177-4187.