第 11 卷 第 1 期
2022 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 11 No. 1
Jan. 2022

引文格式：
张昭辉, 张吉光, 徐士彪, 等. 基于特征混合聚类和关键点检测的智能人脸搜索 [J]. 集成技术, 2022, 11(1): 52-65.
Zhang ZH, Zhang JG, Xu SB, et al. Intelligent face search based on mixed feature clustering and keypoint detection [J]. Journal of Integration Technology, 2022, 11(1): 52-65.

基于特征混合聚类和关键点检测的智能人脸搜索

张昭辉[1,2] 张吉光[1] 徐士彪[3*] 孟维亮[1] 程章林[4] 张晓鹏[1]

[1]（中国科学院自动化研究所 北京 100190）
[2]（中国科学院大学 北京 100049）
[3]（北京邮电大学 北京 100876）
[4]（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 在信息产业急剧膨胀的时代背景下，主流数字媒体产生了由文字到图片再到视频的演化，如何快速有效地获取视频中人物的关键信息，成为各大互联网娱乐和大数据分析领域争相研究的话题。然而，现有的人物信息获取方法还有极大的局限性，无法在视频界面直接获取信息。为了解决这一问题，该文提出了一种新的“由粗到细”的基于特征混合聚类和关键点检测的智能人脸搜索框架，实现了对互联网视频数据的实时检测与高鲁棒的视频人脸数据智能搜索。该文将大数据下人脸数据实时搜索工作细分，首先，通过基于多尺度深度特征混合聚类的人脸检测算法，使用 *Softmax* 函数实现数据分类，并运用中心损失函数 *center loss* 形成聚类中心，随后通过对中心点的回归矫正，达成人脸的粗筛选；然后，通过基于脸部关键点检测算法，提取 68 个人脸关键特征点，生成易于计算处理的标准化特征码。此外，该文还构造了两个影视类人脸数据集，为后续相关互联网行业、娱乐多媒体提供大数据分析。基于该文章整体实验结果表明，在人脸快速检测方面，与现有的主流方法相比，该文方法在识别精度和效率上，都具有一定的提升，其中，基于多尺度深度特征混合聚类算法实验的识别效率提升 31.2%，假阳性样本辨别力提升 3 倍，整体运行效率达标，具有一定的实用价值。

关键词 深度学习；人脸检测；混合聚类；爬虫技术
中图分类号 TP 391.41 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20211001001

收稿日期：2021-10-01 **修回日期**：2021-11-25

作者简介：张昭辉，硕士研究生，研究方向为人工智能、计算机视觉；张吉光，助理研究员，研究方向为人工智能、计算机视觉；徐士彪(通讯作者)，教授，研究方向为人工智能、计算机视觉，E-mail：shibiaoxu@bupt.edu.cn；孟维亮，副研究员，研究方向为人工智能、计算机视觉；程章林，研究员，研究方向为人工智能、计算机视觉；张晓鹏，研究员，研究方向为人工智能、计算机视觉。

Intelligent Face Search Based on Mixed Feature Clustering and Keypoint Detection

ZHANG Zhaohui[1,2]　ZHANG Jiguang[1]　XU Shibiao[3*]　MENG Weiliang[1]
CHENG Zhanglin[4]　ZHANG Xiaopeng[1]

[1](*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*Beijing University of Post and Telecommunication*, *Beijing* 100876, *China*)
[4](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[*]*Corresponding Author*: *shibiaoxu@bupt.edu.cn*

Abstract　The information industry is developing rapidly and the mainstream digital media has produced an evolution from text to pictures to videos. How to quickly and effectively extract the key points of interest of characters in videos has become a hot topic in the field of Internet entertainment and big data analysis. However, existing methods for acquiring character information usually have significant limitations in obtaining information directly from the video interface. To address this problem, this paper proposes a novel "coarse to fine" intelligent face search framework based on feature hybrid clustering and key point detection. The real-time search of face data under big data is subdivided. First, the face detection algorithm based on multi-scale depth feature hybrid clustering uses the *Softmax* function to achieve data classification, and then uses the *central loss* function to form clustering centers that are modified by the regression of centroids to achieve coarse screening of faces. Then, based on the face key point detection algorithm, 68 individual face key feature points are extracted to generate standardized features that are easy to calculate and process to realize the fine search of faces under big data. This enables real-time and highly robust intelligent face search from Internet video data. Notably, this paper also constructs two film and television face datasets to provide big data analysis for subsequent related Internet industry and entertainment multimedia. System's overall experimental results prove that this paper has a certain improvement in recognition accuracy and efficiency compared with existing mainstream face detection methods, including a 31.2% improvement in recognition efficiency and 3 times improvement in discrimination of false-positive samples, and the overall operation efficiency meets the standard and has certain practical value.

Keywords　deep learning; face search; mixed clustering; crawler technology

1　引　言

在大数据环境下，信息技术得到了快速发展。目前，人们正处在一个信息过载的时代，如何在海量数据中获取所需内容，已成为迫在眉睫的问题。在此背景下，基于关键字检索技术的搜索引擎应运而生。随着网络技术的发展，网络传输带宽变得越来越大，可传输媒体种类随之增加。搜索引擎不仅可以搜索文本，还可以搜索图片等媒体，但由于部分图像比较难以使用自然语言去量化的描述[1]，基于内容的图片搜索技术(Content-Based Image Retrieval，CBIR)自此走

上了历史舞台。而人脸图像搜索技术作为图片搜索的一个分支，因其需求广泛，开启了飞速发展的时代。目前，视频、直播等媒体形式已经成为了生活中的主流，然而，对于视频、直播中人或物的搜索，仍停留在截取图像后需自行搜索的阶段，而主流搜索引擎的图片搜索功能，使用起来又非常繁琐。同时，视频实时搜索技术正处于发展阶段，现有技术面临诸多问题，如程序鲁棒性普遍较低，难以同时保证准确性与实时性都达到高度可用。为了解决这些问题，本文从实际应用角度出发，提出了一种基于特征混合聚类和关键点检测的智能人脸搜索框架：将实时人脸检测与人脸识别搜索分割开来，先将人脸粗过滤后再进行人脸精细搜索。此方案同时保证了人脸检测的实时性与人脸搜索的准确性，此外，由系统能效实验可知，本框架在多种状况下都高度可用。

本文的具体贡献如下：

第一，提出了一种新的“由粗到细”的基于特征混合聚类和关键点检测的智能人脸搜索框架，该框架采用了逐级筛选的策略。在人脸实时检测端，使用了基于多尺度深度特征混合聚类的人脸检测算法，将人脸图像进行快速划分。而在人脸数据搜索端，基于人脸检测端的筛选结果，使用基于脸部关键点检测算法[2]提取人脸特征，由于人脸关键点检测的方法具有高精度的特性，所以人脸识别的准确率达到了 99.8%，实现了对互联网视频数据的实时检测与高鲁棒的视频人脸数据智能搜索。

第二，提出了一种基于多尺度深度特征混合聚类的人脸检测算法，通过 *Softmax*[3]与 *center loss*[4]两种聚类算法进行联合训练，其中，*Softmax* 函数进行数据分类，*center loss* 函数生成聚类中心。同时，又加入了中心损失函数的校正回归，提升了网络对人脸的泛化能力，增加了算法效率。

第三，构造了两个公开的影视类人脸数据集，一个是影视剧剧集截图所构造的人脸数据集，适用于各种深度人脸特征网络模型的训练；另一个是高清明星人脸数据集(含合照)，适用于人脸面部特征精确划分算法的训练。

2 相关工作

随着人工智能技术的发展，人脸识别技术取得了巨大突破，其凭借着非接触和几乎无感的优良特性，在人类社会中发挥了重要的作用，全方位地渗透到了生活中。人脸识别技术经过几十年研究与发展，形成了多种人脸识别的解决方案，用来满足不同场景的应用需求。面对更加复杂的应用要求，人脸识别的研究也从最单一简单的背景转变到各种复杂场景之下，对于姿态、光照、表情、年龄、妆容、遮挡、噪声、种族、性别差异等影响因素都需要有良好的解决办法[5]。

目前，大部分人脸数据集都是在较为理想的环境下采集的，由于几乎没有其他因素的干扰，大多数的人脸识别算法对处于该环境下的人脸具有良好的识别率。但是，在现实生活中，除了在较为理想环境下的人脸识别外，还有对于实时记录的公共或私人监控系统进行人脸识别的需要。然而，由于拍摄条件参差不齐，如有些监控本身分辨率就较低，再加上环境复杂、距离较远、角度较偏等因素，就会导致图像的分辨率极低，普通的人脸识别技术难以对其进行有效的识别。所以，对于复杂场景下的低分辨率人脸图像的识别研究具有很大的潜在价值与行业需求，此时，低分辨率人脸识别应运而生[6]。

随着人脸识别技术作为公安、刑侦等领域的技术手段，被大量运用在保卫国家安全，保护人民群众及其财产安全，其相关领域的研究、专利数量开始井喷式增长。2014 年前后，人们开始将深度学习与人脸识别相结合，deep face[7]横空出世，此后，类似于 deep face、DeepIDs[8]等基

于深度学习的人脸识别技术如雨后春笋般涌出。2015 年起，我国提出了一系列政策与发展目标，其中，《中华人民共和国国民经济和社会发展第十三个五年规划纲要》中的第六篇“强化信息安全保障”，极大地促进了我国人脸识别领域的发展，到“十三五”规划结束之时，我国众多知名企业，都拥有了自己的一套人脸识别体系，其算法的完成程度，几乎都处于国际领先地位[9]。但是，美国、日本等国家由于人工智能行业起步较早，仍有大量技术领先。人脸识别技术在发展的同时，其应用领域也得到了拓宽，对此，研究人员提出了根据任务而定的高效的网络集成方法[10]。而在一些场景下，人们除了需要人脸识别技术准确地识别人脸外，对其识别速度也有了严格要求，为了解决这个问题，DFSD[11]、S^3FD[12]、MTCNN[13]、CenterFace[14]等优秀算法相继被提出[15]。本文提出的基于特征混合聚类和关键点检测的智能人脸搜索框架，就参考了 CenterFace 中关于中心损失函数的相关思想。

以往的人脸检测方法继承了基于锚点的通用目标检测框架，可以细分为两类：两步法（Faster R-CNN[16]）和一步法（SSD[17]）。与两步法相比，一步法的效率较高，召回率也较高，但是会导致很高的假阳性率，且降低了人脸的定位精度。之后，区域生成网络（Region Proposal Network，RPN[18]）的两段法直接开始应用于人脸检测，SSH[19]与 S^3FD 在一个单一网络中开发了一个不变尺度的网络，用来检测来自不同层的多尺度人脸。但是，基于锚点的方法不能很好地兼顾召回率与鲁棒性。

目前，较先进的人脸检测网络通过使用网络预训练模型 VGGNet[20]和 ResNet[21]，在 WIDER FACE[22]上取得了较高的准确率。但是，这些人脸检测技术因为其庞大而复杂的神经网络耗时较长，且模型的规模也非常大，很难应用于实际。其次，基于卷积神经网络的 VGGNet 等网络预训练模型，没有对人脸特征进行标记，不利于人脸特征对齐匹配的应用。因此，将人脸的检测与对齐算法进行一体化设计，并达到较好的实时性与准确率，成为了实际应用中至关重要的环节。

受到无锚点通用目标检测框架的启发，Xu 等[14]提出了一种轻量化、高效率的人脸检测和对齐方法 CenterFace，其网络可以进行端到端训练。该算法使用脸边界框的中心点来表示脸部位置，面部框的大小和坐标直接被回归到中心位置的图像特征，从而使人脸检测和对齐问题转化为人脸的关键点估计问题。热图中的峰值对应面部的中心，每个峰值的图像特征可以预测脸的大小和人脸关键点的大小。经相关数据集评估，结果表明，该网络对于人脸图像实现了较好的分辨。

但 *Softmax* 损失函数只保证了特征的可分性，并不要求类内紧凑和类间分离，因此并不适用于人脸识别。CenterFace 尽管基于中心损失函数，使人脸特征辨识度显著增高，但是在 WIDER FACE 高难测试场景下仅达到 78.2% 的精度，在 Wildest Faces[23]数据集中的测试结果也并不理想，模型的训练方法还有待改进。

实时应用场景下的人脸识别算法要求轻量化、高效率。而在应用更加广泛的人物身份识别场景中，人脸识别算法则需要更高的准确性，各大互联网公司在此方面几乎做到了极致。2021 年，美国国家标准技术研究所发布的人脸识别供应商测试结果中，商汤科技与依图、百度等公司均在前列，对于人脸图像的查准率均已超过 99.99%（数据来自 Face Recognition Vendor Test（FRVT）| NIST）。但由于这些公司的网络模型都过于庞大，而且大多并未公开，所以进行实验时一般选择开源、轻量化的人脸识别模块。

Face_recognition 是一个极为简洁的人脸特征识别库，可以使用 Python 和命令行工具进行提取、识别、操作人脸。该项目基于行业内领先的

C＋＋开源库 Dlib 中的深度学习模型，并在由美国麻省大学阿莫斯特分校提供的 Labeled Faces in the Wild[24]人脸数据集上进行测试，测试结果准确率高达 99.38%，已经超过了人类肉眼识别的平均水平。Dlib 提供了两个人脸检测方法，使用方向梯度直方图(Histogram of Oriented Gradient，HOG)特征[25]进行回归或者基于卷积神经网络进行识别，其中，卷积神经网络方法需要使用 GPU 加速。

与其他人脸识别算法的步骤一致，Face_recognition 首先进行人脸检测：输入图像经过降维处理后，生成方向梯度直方图，系统通过匹配，找到 HOG 图案中最像人脸梯度特征的那一部分，从而完成人脸分离。得到人脸图像，然后进行人脸对齐，使用面部特征点估计算法找到图像中对应的特征点，Dlib 中的模型拥有 68 个人脸特征定位点。人脸特征经仿射变换后映射到标准人脸模型上，最终经过具有 29 个转换层的深度残差网络 ResNet[21]生成 128 位特征码，随后只需进行简单的向量间距计算(主要采用欧氏距离)，就可以得到输入人脸图像与其余人脸图像的人脸间距绝对值，值越小越可能是同一张人脸。这种特性使得 Dlib 在人脸编码搜索方面有着得天独厚的优势，兼顾了人脸识别的准确性与人脸编码搜索的实时性。

3 算法框架

3.1 总体设计

首先，算法实时截取视频、直播等流媒体图像。然后，将图像通过轻量化、高效率的人脸检测算法(基于多尺度深度特征混合聚类的人脸检测算法)进行人脸图像的初步筛选，过滤掉背景干扰。最后，经过粗筛选后的人脸图像再经使用者挑选，进入基于脸部关键点检测算法(Dlib 人脸识别)，由于面部特征规范化的特性，使得识别准确率得到极大提升。通过以上粗筛选、细搜索的应用策略，前者为后者搜索扫除障碍，争取时间，后者将前者的成果进行更加精细的呈现，这种逐级筛选所带来的合力，使得该框架的实时性与精准度都能得到极大的保障。

算法按照处理阶段可分为预处理与在线处理两部分，框架如图 1 所示。其中，在线处理部分又被细分为实时人脸检测部分(粗筛选)与人脸在线搜索部分(细搜索)。

3.1.1 预处理部分

预处理部分分为两个步骤。首先，通过网络爬虫爬取海量目标数据图像，此步骤为预处理阶段，故无需对效率作太多要求，此时可以引入针对性的基于深度学习的人脸特征识别算法，将人们更有可能搜索的结果图片提取到数据库中(如：明星、公众人物等)，达到计算机系统快表(Cache)的效果，大幅度提升系统的搜索效率。也可制定人脸图像排序的规则，将所需要的人脸图像入库，提升库中图像被检索的概率。此时，入库的不仅有图片，还包括相应页面的统一资源定位符等信息。

在获取完海量目标图片后，将目标图片导入库处理程序，再通过基于脸部关键点检测算法，对目标图片进行人脸存在性判断，若存在人脸图像，则建立人脸空间，并遍历图片找到全部人脸，存入对应图像的空间。之后，通过脸部关键点检测算法中深度残差神经网络提取人脸特征并完成编码，当全部图像中的所有人脸进行编码后，将人脸特征码与对应图片的统一资源定位符一同存入数据库中，等待搜索。

3.1.2 实时在线处理与搜索部分

在进行视频实时人脸数据搜索时，首先需要导入当前播放的视频流，若无接口，可直接使用系统屏幕录制功能，程序通过一定算法截取视频帧；若存在可导入视频源，系统就对视频进行提前处理，加入关键帧算法，提升系统的可用性，

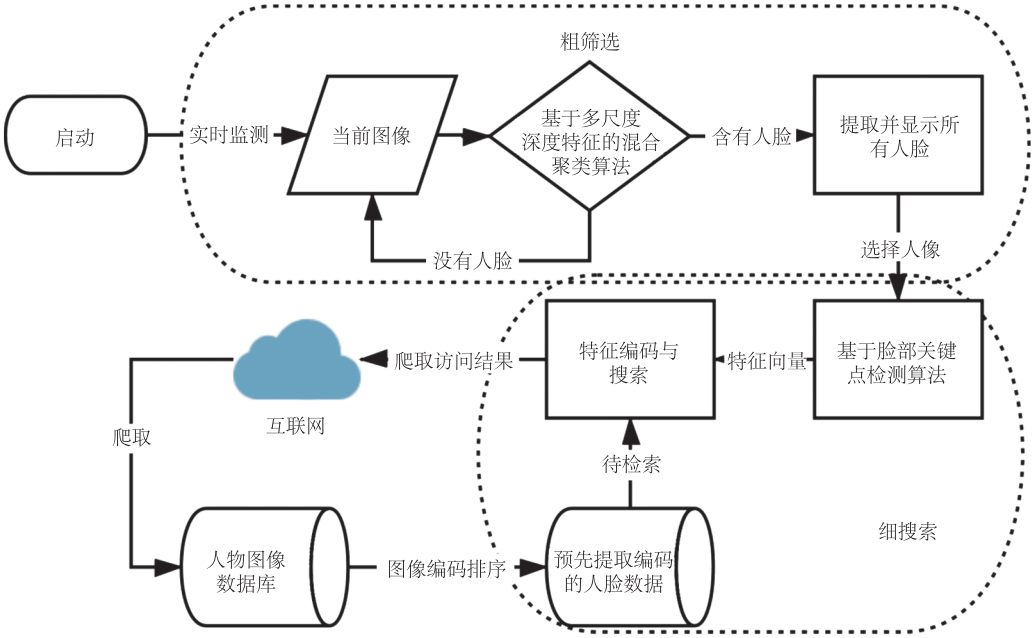


图 1　实时人脸数据搜索算法框架

Fig. 1　Real time face data search algorithm framework

减少用于处理无效帧所消耗的计算资源。在获取到视频帧后，为了减少人脸提取所消耗的计算资源，提升系统实时性，此时应采用轻量化、高效率的基于多尺度深度特征混合聚类的人脸识别算法，算法在轻量化结构的辅助下，能够迅速将视频帧中人脸全部找出。该步骤只进行人脸检测而不进入搜索，所以系统能轻松处理每秒几十帧的数据，这样做极大程度地节省了计算资源，提升了系统的可用性。

在实时人脸检测部分结束后，经过粗筛选的去除背景干扰的人脸图像出现在使用者眼前，再进行人脸在线搜索，系统将通过基于脸部关键点检测算法对人脸特征进行提取，生成特征码。由于人脸关键点特征具有自身高命中率的特性，可将人脸搜索精度提升至 99.8%，同时，基于脸部关键点检测算法采用规范的特征编码，所以系统能在可接受时间内找到若干匹配的人脸，并基于人脸特征间距进行排序。然后生成人脸所在图片的统一资源定位符序列进行网络爬虫，此时，采用的爬虫算法应是简单高效的深度网络爬虫，从而保证使用者所需数据能够即时地下载返回并呈现在使用者眼前。该系统框架借助准确且高效的基于脸部关键点检测算法，能够做到轻松处理数万级的数据。

在线处理部分，粗筛选阶段为细搜索阶段扫除障碍、排除背景等因素干扰的同时，也因其算法轻量化的特点，能够实现高帧率人脸图像的实时处理；细搜索部分在继承粗筛选阶段的结果后，以极高的命中率，出色地解决了粗筛选阶段所做不到的高准确率问题。两个阶段相辅相成，共同行使职能，最终使该系统框架既兼顾了实时性，也达到了高鲁棒性。

3.1.3　系统目标

为了进一步提升系统鲁棒性，使框架高度可用，本文提出并解决了以下几个问题。

第一，在许多视频或直播的场景中，人物并不是正襟危坐在镜头前，尤其是在电视剧与电影中。影视作品为了体现镜头感，往往会出现人脸只有侧面或存在遮挡的情况，在某些大场面中，单个人脸会变得很小，且演员本身又带妆，

在多种因素的影响下，人脸识别的错误率会极大增加，特别是图像中有多张人脸时更加容易发生错误。此外，直播时的突发场景，尤其是在户外直播时，过高或过低的场景亮度以及手持摄影设备直播时的抖动，都极大程度地增加了图像的噪声。因此需要构建一个能在复杂场景下识别人脸的算法——不仅要满足框架中提到的轻量化、高效率等优点，还需要对低分辨率下的人脸具有足够的识别率。

第二，人脸图像的在线搜索模块，与实时人脸检测模块不同，该模块对于人脸处理与搜索的时间需求相对来说比较宽松，但对人脸识别精度有较高的要求。因此，需要一个以提高人脸识别准确性为主，同时又具有一定即时性的算法，并且为了之后搜索时方便编码，该算法提取出来的特征，需要有极高的规范性。

第三，由于项目基于单个系统架构，并非分布式架构，因此，在海量数据中及时地搜索和返回目标，成为了搜索模块的首要任务。而通过分析传统的图像搜索引擎结构可知，建立搜索的第一步是完成所搜图片的编码，在这里需要非常规范化的人脸编码，其既能最大限度地保留人脸信息，又能尽可能地保持相同的码位以便搜索，在编码完成后，建立高效的索引进一步提升搜索速度。

第四，对于返回的搜索结果，应当是最符合搜索者期待的结果，但是，在海量图片下人脸识别的结果有可能非常多，本文借助人脸识别等技术，将人们最可能期望得到的搜索结果筛选出来，优化了搜索者的搜索体验。同时，减轻了系统下行的压力。

3.2 基于多尺度深度特征混合聚类的人脸检测算法

对于人脸识别的实际应用来说，通过深度神经网络所获得的特征，除了需要实现人脸可分，还需要有良好的辨别率。*Softmax* 损失函数确保了提取的特征可分，但过于模糊的分类方法，会使其人脸识别的能力减弱，在加入 *center loss* 中心损失函数后，人脸识别的准确性得到提高，每个类别的中心和特征向量都拥有相同的维度。通过训练的不断深入，分类中心不断迭代，最小化人脸特征与中心距离，深度特征高度分离。通过 *Softmax* 和 *center loss* 的联合训练，最终使类间间距增大，类内间距减小。达到了较高的识别率，同时因为聚类的原因，使得原本存在的图像噪点被舍去，降低了系统资源占用，提升了系统性能。

3.2.1 训练数据集

为了使模型有效且具有一定健壮性，本文选择自制数据集，数据来自视频、直播中各种场景下的人脸画面，数据集采用程序对电影、电视剧、直播等视频数据随机截取，筛选掉无人脸的图像，确保数据集中的图像接近真实使用场景。具体图像示例如图 2 所示。

图 2 影视人脸数据库例图

Fig. 2 Example of video face database

本数据集中包含 218 位人物以不同姿势、表情、妆容呈现的 15 844 张人脸图像，并且含有不同的锐度、亮度等背景信息。另外，由于视频中镜头的移动，会出现运动模糊等复杂情况，同时还伴有着其他的干扰因素。其中，人脸图像多于 1 张的人物有 157 位(占比 72.02%)。选用数据集中 11 091 张(占比 70%)图像用作训练数据，其余图像用作测试数据(测试数据只包含拥有两张图片以上的人物)，具体划分如表 1 所示。

表 1　数据集划分

Table 1　Data set partition

划分	全部图像	训练	测试	多张
人物（人）	218	203	30	157
图像（张）	15 844	11 091	4 753	15 583

3.2.2 基于多尺度深度特征的混合聚类

本文采用的 CenterFace 神经网络框架的工作原理如图 3 所示，中心脸的框架与常用的图像识别神经网络近似，主要区分点为损失函数的差异。

对于低分辨率人脸的识别，CenterFace 还存在缺陷，当使用 *Softmax* 与中心损失函数联合进行特征训练时，虽然达到了类内间距不断减小的目的，但类中心与原点之间的距离也变小了，这不仅缩小了类内间距，还缩小了类间间距，始终无法达到良好的分类效果。为了解决此问题，本文提出了一种新的解决方案，基于向量间欧氏距离，将中心损失函数聚类所导致的类间间距缩小的问题作修正回补，基于模损失函数，得到回归函数：

$$L_R=\sum_{n=1}^{N}\left(\log\frac{\mathrm{e}^{W_{y_n}^{T}\boldsymbol{x}_n+d_{y_n}+\frac{1}{2}\left(\boldsymbol{x}_n-c_y\right)^2}}{\sum_{m=1}^{M}\mathrm{e}^{W_m^{T}\boldsymbol{x}_n+d_m}}\right)$$

其中，L_R 为回归函数；N 为每个分批的大小；$\boldsymbol{x}_n$ 表示第 n 个特征向量；y_n 为 $\boldsymbol{x}_n$ 的标签；M 为训练集的分类数；W_m 为 CenterFace 网络全连接层中权重第 m 列；W_{y_n} 为 CenterFace 网络全连接层中权重第 y_n 列；d 为误差值；c_y 为聚类中心，T 为训练迭代次数。

经回归函数矫正后，所得结果在缩小类内间距的同时，增大了类间间距，这种做法较大程度上增加了特征的泛化能力。

基于特征向量空间欧氏距离进行分类后，很容易发现，在 *Softmax* 与中心损失函数进行联合训练的同时，进行回归矫正，该结果比之前单一的损失函数的结果有更强的区分能力和更准确的识别能力，如图 4 所示。这种高区分度将会带来搜索效率的提升：此前被误识别的‘假人脸’此时都不参与识别，系统能空出更多的额外资源用来处理真的人脸，处理速度加快的同时，系统占用也在降低。此外，CenterFace 中的训练步骤也存在缺陷，通过颠倒卷积层中参数和类

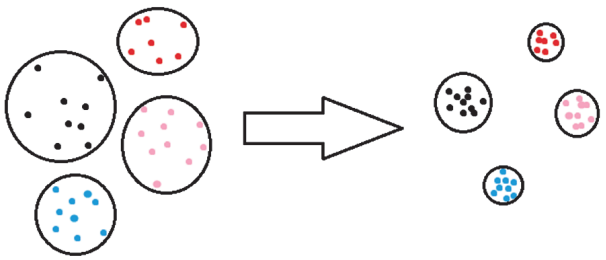


图 4　单一损失函数与联合损失函数的比较

Fig. 4　Comparison between single loss function and joint loss function

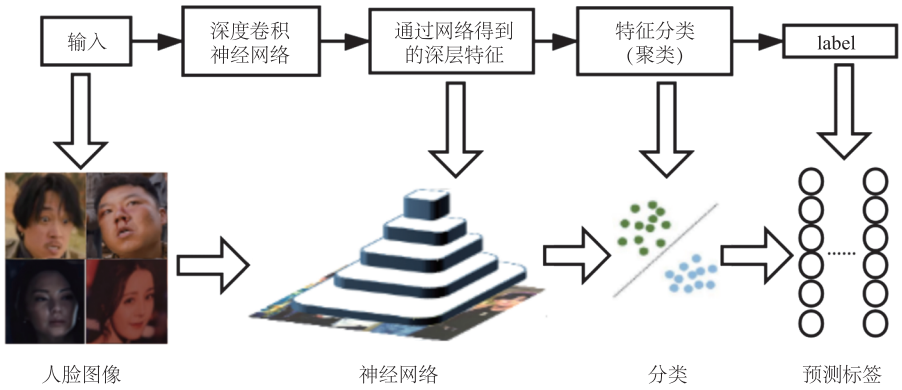


图 3　CenterFace 工作原理

Fig. 3　How CenterFace works

中心参数的更新顺序，可以进一步提升函数的收敛能力。

3.3 基于脸部关键点检测算法的网络爬虫筛选器

本文在进行基础人脸数据库的构建时，用爬虫技术对网络人物的图像进行爬取，可能会使库中存在许多几乎不可能被访问到的图像，这不仅增加了存储成本，而且降低了系统搜索效率。如果在爬取图像数据时，就对图像进行筛选，将高颜值、高清晰度、高访问量的人脸图像排在搜索队列头部，并增加搜索与访问次数等属性，建立反馈机制，提升更有可能被人们搜索的图像的搜索优先级，建立一套由人脸识别与前馈神经网络相结合的网络爬虫结果处理机制，在搜索时不断优化搜索体验，最终框架见图 5。

在线搜索时，图像特征码的生成，采用的是基于脸部关键点检测算法提取特征，而离线数据库的建立也需要以同样的方法生成特征码，所以，本文在爬虫的反馈端也使用基于脸部关键点检测算法，便于数据统一。当图像被搜索或访问后，将该图像传入基于脸部关键点检测算法的网络进行分析，然后输出人脸特征标签信息，如性别、年龄、颜值(基于脸部关键点检测算法内置属性)等信息，将这些信息反馈至爬虫模块，通过自学习网络修改参数，提升经常被搜索图片的优先级，从而提升搜索者的搜索体验。

3.4 多人脸检测

人脸识别技术的本质，就是对一张人脸图像进行量化分析后，与其他经过量化分析后的人脸图像作比较。但在实际应用过程中，数据库中很少存储只含有单个面部的照片，即使是单人照、证件照等照片，在进行人脸识别时仍需要进行人脸切割、人脸对齐等必要步骤。在互联网中存在海量的人脸图片，大部分图片含有多张人脸，而且同一人的单人脸图片中的面部图像，并不一定比含有多张人脸的图片中的面部图像更具有代表性。传统的人脸识别技术采用的是逐一对比的方法，将该技术应用到含有多张人脸图像的图片中，能够扩大人脸识别技术的使用范围。

对于搜索端，当截取到的图片含有多张人脸图像时，先用算法对该图片进行判断，将所有人脸图像进行定位后，基于人脸中心位置与人脸特征位置进行切割，再通过人工选择其中一张人脸图像，并将其传入搜索网络。而对于人脸数据库端，在获得一张网络图片后，建立与图片相对应的人脸图像空间，对图片中每个人脸都进行检测提取，并存储到对应图片空间中的每张人脸图像都对应着人脸空间的一个元素。将传入搜索网络的人脸图像与数据库中人脸图像逐一比对，比对完成后，通过人脸图像所在的人脸空间找到对应图片，爬取图片的统一资源定位符进行结果返

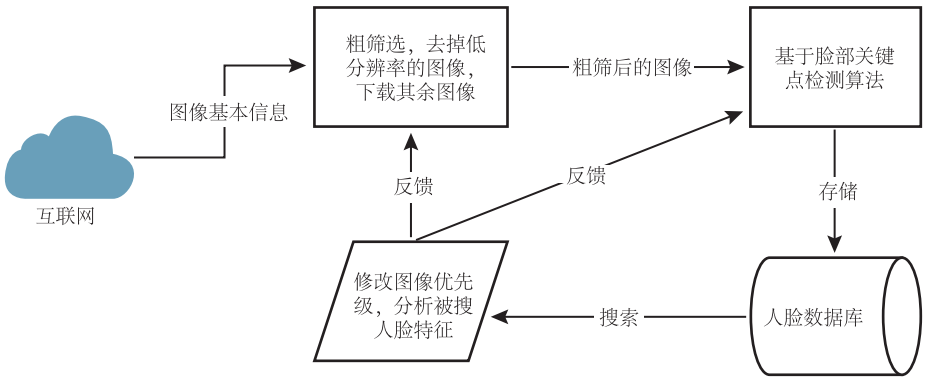


图 5 基于脸部特征关键点检测算法的网络爬虫筛选器框架

Fig. 5 Crawler filter framework based on facial feature key point detection algorithm

回。通过这种方法，无论匹配图片或者搜索图片有一张还是多张人脸图像，系统都能够给出有效的结果返回。

4 实　验

为了验证优化过后系统框架的运行效率，本节将利用一些数据集对框架中的算法进行评估。

4.1 基于多尺度深度特征混合聚类的人脸检测算法运行效率

现有的人脸检测器大多数基于卷积神经网络（Convolutional Neural Networks，CNN），在进行大量的人脸识别时，往往需要 GPU 进行加速，基于 CPU 运行时很难保证实时采样识别。本文训练所得到的人脸检测模型的大小仅有 6.8 MB，比如今动辄数百兆的 CNN 模型小了数十倍，系统效率得到很大提升。本测试基于英特尔 Z170 芯片组，CPU 为 Intel Core i7-7700，GPU 为 NVIDIA GTX1060，测试对象为一张 314 人的合照，测试方法包含两个开源人脸识别算法：基于 OpenCV 的 haar cascades[26]与精细搜索中使用到的基于脸部关键点检测算法，4 个低分辨率下的快速人脸检测算法：DSFD、S^3FD、基于 PyramidBox[27]方法的 PyramidBox＋＋[28]、LFFD[29]，基于 GPU 的运行效率测试结果如表 2 所示。由表 2 可知，本文所使用的方法，除了对基于脸部关键点检测算法有巨大的效率优势外，在相同数据集下，与腾讯旗下优图团队的 DFSD 方法和在 WIDER FACE 下取得了较好成绩的 S^3FD、LFFD 方法，以及 PyramidBox＋＋方法相比，本文算法的单张图片人脸识别效率是最高的，即便是高分辨率的多人脸图像，平均一秒钟也能处理 30 张左右。此外，本实验还对一些含有较少人脸的图像进行了测试，本文算法的效率均排在首位，由于篇幅限制，本文只选取了最能验证算法效率的含有 314 张人脸图像的图片测试结果进行展示。

对于 CPU，采集网络摄像头视频流进行测试，针对上述人脸识别算法过于缓慢以至于难以使用，本文采用 MNN（Mobile Neural Network）[30]、NCNN（Tencent/ncnn: ncnn is a high-performance neural network inference framework optimized for the mobile platform（github.com））、ONNX（Open Neural Network Exchange）[31]等方法生成模型，和原生模型进行比对，对比结果如表 3 所示。由表 3 可知，本文算法具有最高的效率，较原本算法提升了 31.2%。CPU 下运行帧数可达 30 帧，具有较好的实时性。对于多人脸任务的处理（测试环境为 10～30 人的各种复杂场景），CPU 下的 6 帧显然已经无法满足要求。在实际应用中，采用 GPU 并行处理数据，在大多数场景

表 2　基于 GTX 1060 的运行效率

Table 2　Running efficiency based on GTX 1060

算法	运行效率（ms）		
	分辨率：640×480	分辨率：1 280×720	分辨率：1 920×1 080
OpenCV haar cascades	553.7	1 238	2 937
基于脸部关键点检测算法	680.5	1 536	3 091
DSFD	140.1	301.5	665.4
PyramidBox＋＋	108.5	278.0	497.9
S^3FD	57.60	138.6	255.4
LFFD	19.92	78.76	156.8
CenterFace	13.57	29.41	53.40
本文算法	10.34	18.15	34.01

下能够稳定处理 30 帧的画面，具体数据及内容见第 4.3 节。

表 3 CPU 下的算法表现

Table 3 Algorithm performance under CPU

模型	单次平均处理时间（ms）	帧数	多人脸帧数
CenterFace_MNN	51.05	18	3
CenterFace_NCNN	63.33	15	3
CenterFace_ONNX	87.13	12	2
CenterFace_tensorrt	99.91	10	2
CenterFace	41.57	24	5
本文算法	32.11	30	6

4.2 基于多尺度深度特征混合聚类的人脸检测算法识别率

通用影视人脸数据集中，共包含 15 844 张图像，218 位人物。本实验选取其中 30 位人物的 1 000 张人脸图像作为测试，仍然使用 GPU 测试中的 7 种算法，并引入 MTCNN 作为比较。此外，还在测试中加入 200 张假人脸，测试结果如表 4 所示。由表 4 可知，本文算法对于人脸数据有着较好的辨别效果。结合第 4.1 节表 2 的实验结果可知，本文算法在对比 DSFD、S^3FD 算法几十分之一的处理时间中，达到了 97.8% 的识别率，兼具了实时性与准确性。对于假阳性样本的抗干扰性，与 MTCNN 等基于深度神经网络方法相比，本文算法仍具有较好的抗干扰能力，但是仍然难以超越基于脸部关键点检测方法。此外，通过对表 4 分析可得，与未被改进的算法相比，基于深度特征混合聚类的人脸识别算法有更高的人脸识别率以及更低的假阳性率。

表 4 各算法在自制数据集下识别率与假阳性率

Table 4 Recognition rate and false positive rate of each algorithm under self-made data set

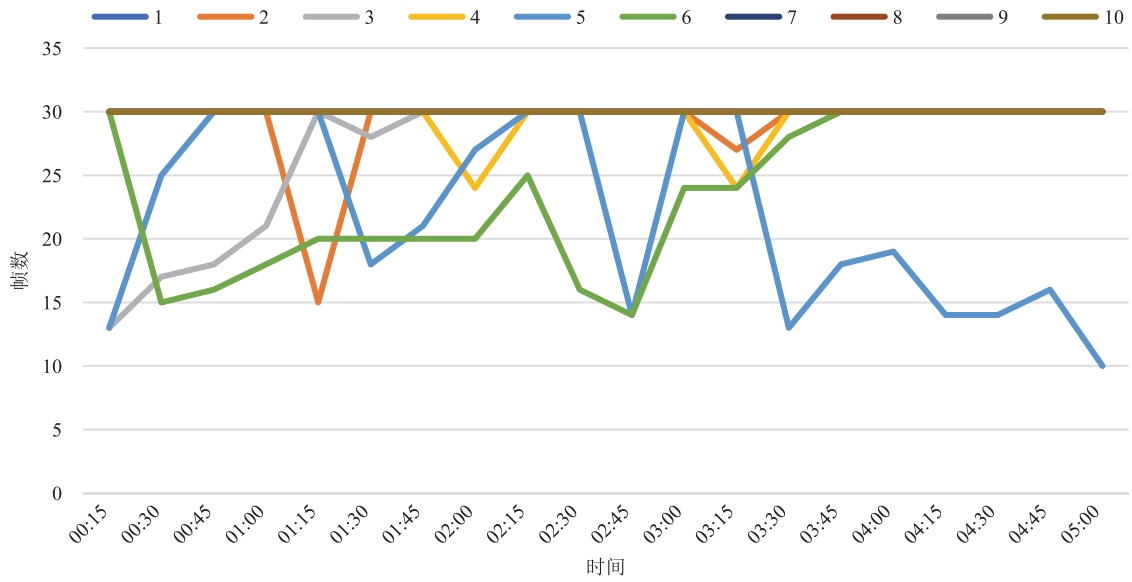
算法	识别率(%)(真实人脸)	假阳性率(%)(人造脸)
OpenCV haar cascades	98.9	23.5
基于脸部关键点检测算法	99.8	1.50
PyramidBox＋＋	98.0	82.5
LFFD	92.1	95.0
DSFD	98.0	78.0
S^3FD	95.8	82.0
MTCNN	89.7	91.5
CenterFace	96.1	96.5
本文算法	97.8	34.5

4.3 系统能效

人脸关键点检测算法与爬虫算法，非本文研究核心重点，本文仅将其技术运用于系统之中，故不进行单独的算法性能测试，两算法的具体原理见第 3 节。

本文选取 10 个影视、直播的视频流，将每个视频播放 5 min，并对系统的综合效能进行了测试。为了保持系统稳定性，将其处理帧数限制在 30 帧/s，每 15 s 进行一次帧率采样，采样结果如图 6 所示。由图 6 可知，除视频 5、视频 6 帧率严重波动，视频 3 开始时出现帧率波动外，其余视频仅有小幅波动，且视频 7～10 没有出现任何波动。通过对视频内容分析可知，视频 5 与视频 6 都是大型战争题材影视剧，含有近千人冲锋陷阵的场面，视频 3 则是在视频开始时有数百人上朝的多人场景，这表明系统在大多数情况下能够每秒稳定处理 30 张画面，即使出现人数较多的场景，系统最低帧率仍然不小于 10 帧。但此时 CPU 负载达到了 80%，为了降低系统资源的占用，在实际使用过程中将帧数调至 5(此处的帧数代表着系统每秒中处理图像的数量，并不影响视频的播放帧数，直播、电视剧播放帧数仍维持在 60 帧/s，电影一般为 24 帧/s，运行系统时并不会产生播放卡顿的现象)，此帧率在满足人们对于实时搜索需求的同时，对系统资源的消耗较低。

在此基础上，本文还对大数据下人脸数据搜索能效进行了测试，在依靠基于脸部关键点检测算法的网络爬虫筛选器所生成的 54 000 张名人图片的数据集中(数据集采用 MySQL 进行存储)，搜索结果的平均用时为 283 ms(此处并没有统计



注：视频 7~10 完全重合

图 6　系统综合性能测试表

Fig. 6　System comprehensive performance test table

爬虫下载时间，由于网络延时等因素，在 50 M 下载带宽下，总使用时间在 0.5～1.5 s 范围内不等）。该实验结果充分表明了本框架的可用性，且具有一定的研究价值。

5　结　语

本文基于现有视频、直播中人物信息获取难的问题，针对现有技术使用范围受限、信息有限、难以兼顾实时性与准确性等问题，提出了一种全新的基于特征混合聚类和关键点检测的智能人脸搜索框架。该框架的关键在于将人脸实时检测与大数据下人脸的精准搜索分割开来，使用更轻量化、高效率的基于多尺度深度特征混合聚类的人脸检测算法进行实时人脸检测，极大程度上减少了系统负担。在人脸搜索阶段，使用实时性与准确性并重的基于脸部关键点检测算法，使系统总体具有良好的实时性与精准性，且对系统资源的占用较低。同时，本文还提出了基于多尺度深度特征混合聚类的人脸检测算法，方法运用 *Softmax* 与 *center loss* 函数联合进行训练，并加入回归修正函数，使得最终结果既减少了类内间距，又增加了类间间距。本文算法不仅提高了人脸识别的准确率，还将许多非人脸的噪声干扰剥离了聚类中心，一定程度上减轻了系统的负担。通过基于多尺度深度特征混合聚类的人脸检测算法的实验结果可以发现，与同类型前沿的方法相比，本文算法具有更高的效率，同时也达到了同类型算法中较高的识别率及对假阳性样本较高的抵抗性。本文还将人脸识别技术应用在了网络爬虫中，使得搜索结果高度相关。

此外，本文提供了两个可供研究的数据集，一个主要包含视频、直播等场景下的视频截图中的人脸，可以通用于各种基于深度学习的人脸算法的模型进行训练；另一个为较高清晰度的明星、网络名人人脸数据集，其包含合照、独照、剧照等，可用于人脸精细化特征提取。

本文基于深度神经网络实现了视频、直播等媒体数据下的实时人脸搜索框架，并经过充分的数据测试，结果表明，本系统在兼顾实时性与准

确性的同时，还具有良好的环境兼容性与稳定性，为大数据下实时人脸搜索系统的应用打下了坚实的基础。

参 考 文 献

[1] Julesz B. Visual pattern discrimination [J]. IRE Transactions on Information Theory, 1962, 8(2): 84-92.

[2] 刘兆丰. Dlib 在人脸识别技术中的运用 [J]. 电子制作, 2020, (21): 39-41＋7.
Liu ZF. Application of Dlib in face recognition technology [J]. Practical Electronics, 2020, (21): 39-41＋7.

[3] He YL, Zhang XL, Ao W, et al. Determining the optimal temperature parameter for Softmax function in reinforcement learning [J]. Applied Soft Computing, 2018, 70: 80-85.

[4] Shi ZL, Wang H, Leung CS. Constrained center loss for convolutional neural networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[5] Yang GZ, Huang TS. Human face detection in a complex background [J]. Pattern Recognition, 1994, 27(1) : 53-63.

[6] 黄兴晗, 杜小甫, 刘沂杰. 人脸识别技术分类比较 [J]. 电子测试, 2021, (17): 96-97＋29.
Huang XH, Du XF, Liu XJ. Classification and comparison of face recognition technologies [J]. Electronic Test, 2021, (17): 96-97＋29.

[7] Huang PL, Han JW, Zhang DW, et al. CLRNet: component-level refinement network for deep face parsing [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[8] Li SZ, Jain AK. Handbook of face recognition [M]. New York: Springer, 2011.

[9] GrotherPJ, Ngan ML, Hanaoka KK. Ongoing Face Recognition Vendor Test (FRVT) part 2: Identification [Z/OL]. https://pages.nist.gov/frvt/html/frvt11.html#_FRVT_Participation_Statistics_.

[10] Siqueira H, Magg S, Wermter S. Efficient facial feature learning with wide ensemble-based convolutional neural networks [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5800-5809.

[11] Scagliarini A, Bogner S, Harting J. Editorial for the Special Issue “DSFD 2017” [J]. Computers and Fluids, 2018, 179: 670-671.

[12] Zhang S, Zhu X, Lei Z, et al. S^3FD: Single Shot Scale-invariant Face Detector [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 192-201.

[13] Zhang KP, Zhang ZP, Li ZF, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[14] Xu YY, Yan W, Yang GK, et al. CenterFace: joint face detection and alignment using face as point [J]. Scientific Programming, 2020.

[15] 余璀璨, 李慧斌. 基于深度学习的人脸识别方法综述 [J]. 工程数学学报, 2021, 38(4): 451-469.
Yu CC, Li HB. Overview of face recognition methods based on deep learning [J]. Chinese Journal of Engineering Mathematics, 2021, 38(4): 451-469.

[16] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 201.

[17] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector [C] // European Conference on Computer Vision, 2016: 21-37.

[18] Hu P, Ramanan D. Finding tiny faces [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 951-959.

[19] Najibi M, Samangouei P, Chellappa R, et al. SSH: Single Stage Headless face detector [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 4875-4884.

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

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C] // Proceedings of IEEE

Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[22] Yang S, Luo P, Loy CC, et al. WIDER FACE: a face detection benchmark [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5525-5533.

[23] Bilge YC, Yucel MK, Cinbis RG, et al. Red Carpet to Fight Club: Partially-supervised domain transfer for face recognition in violent videos [C] // 2021 IEEE Winter Conference on Applications of Computer Vision, 2021: 3358-3369.

[24] Huang GB, Learned-Miller E. Labeled Faces in the Wild: updates and new reporting procedures [Z]. https://people.cs.umass.edu/~elm/papers/lfw_update.pdf.

[25] Déniz O, Bueno G, Salido J, et al. Face recognition using Histograms of Oriented Gradients [J]. Pattern Recognition Letters, 2011, 32(12): 1598-1603.

[26] Hapsari DTP, Berliana CG, Winda P, et al. Face detection using haar cascade in difference illumination [C] // 2018 International Seminar on Application for Technology of Information and Communication, 2018.

[27] Tang, X, Du DK, He ZQ, et al. PyramidBox: a context-assisted single shot face detector [C] // Proceedings of the European Conference on Computer Vision, 2018: 797-813.

[28] Li, ZH, Tang X, Han JY, et al. PyramidBox＋＋: high performance detector for finding tiny face [J]. arXiv Preprint, arXiv: 1904.00386, 2019.

[29] He Y, Xu D, Wu L, et al. LFFD: a light and fast face detector for edge devices [J]. arXiv Preprint, arXiv: 1904.10633, 2019.

[30] Jiang X, Wang H, Chen Y, et al. MNN: a universal and efficient inference engine [J]. airXiv Preprint, arXiv: 2002.12418, 2020.

[31] Ye XY, Chen XT, Chen HH, et al. Deep learning network for face detection [C] // 2015 IEEE 16th International Conference on Communication Technology, 2015: 504-509.