E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS


**中图法分类号**:TP319.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)05-1434-11

**论文引用格式**:Sang G L, Xiao S D and Zhao Q J. 2023. Soft threshold denoising and video data fusion-relevant low-quality 3D face recognition. Journal of Image and Graphics, 28(05):1434-1444(桑高丽, 肖述笛, 赵启军. 2023. 联合软阈值去噪和视频数据融合的低质量3维人脸识别. 中国图象图形学报, 28(05):1434-1444)[DOI:10.11834/jig.220695]

# 联合软阈值去噪和视频数据融合的低质量3维人脸识别

桑高丽[1],肖述笛[2],赵启军[2*]

1. 嘉兴学院信息科学与工程学院,嘉兴　314001; 2. 四川大学计算机学院,成都　610065

**摘　要**:**目的**　低质量3维人脸识别是近年来模式识别领域的热点问题;区别于传统高质量3维人脸识别,低质量、高噪声是低质量3维人脸识别面对的主要问题。围绕低质量3维人脸数据噪声大、依赖单张有限深度数据提取有效特征困难的问题,提出了一种联合软阈值去噪和视频数据融合的低质量3维人脸识别方法。**方法**　首先,针对低质量3维人脸中存在的噪声问题,提出了一个即插即用的软阈值去噪模块,在网络提取特征的过程中对特征进行去噪处理。为了使网络提取的特征更具有判别性,结合softmax和Arcface(additive angular margin loss for deep face recognition)提出的联合渐变损失函数使网络提取更具有判别性特征。为了更好地利用多帧低质量视频数据实现人脸数据质量提升,提出了基于门控循环单元的视频数据融合模块,实现了视频帧数据间互补信息的有效融合,进一步提高了低质量3维人脸识别准确率。**结果**　实验在两个公开数据集上与较新方法进行比较,在Lock3DFace(low-cost kinect 3D faces)开、闭集评估协议上,相比于性能第2的方法,平均识别率分别提高了0.28%和3.13%;在Extended-Multi-Dim开集评估协议上,相比于性能第2的方法,平均识别率提高了1.03%。**结论**　提出的低质量3维人脸识别方法,不仅能有效缓解低质量噪声带来的影响,还有效融合了多帧视频数据的互补信息,大幅提高了低质量3维人脸识别准确率。

**关键词**:3维人脸识别;低质量3维人脸;软阈值去噪;联合渐变损失函数;视频数据融合

# Soft threshold denoising and video data fusion-relevant low-quality 3D face recognition

Sang Gaoli[1], Xiao Shudi[2], Zhao Qijun[2*]

*1. College of Information and Engineering, Jiaxing University, Jiaxing　314001, China;*

*2. College of Computer Science, Sichuan University, Chengdu　610065, China*

**Abstract**: **Objective** 3D sensors-portable are developed and focused on user-friendly 3D facial data. Its low-quality 3D face recognition is concerned about more in the context of pattern recognition in recent years. Low quality 3D face recognition is challenged of the problem of low quality and high noise. To suppress high noise in low-quality 3D face data and alleviate the difficulty of extracting effective features in terms of limited single-depth data, we develop a novel low-quality 3D face recognition method on the basis of soft threshold denoising and video data fusion. **Method** First, a trainable soft thresh-

**收稿日期**:2022-07-07;**修回日期**:2022-11-28;**预印本日期**:2022-12-05

*通信作者:赵启军　qjzhao@scu.edu.cn

**基金项目**:国家自然科学基金项目(61773270);嘉兴学院"百青计划"(CD70621004);浙江省教育厅科研项目(Y202249424)

**Supported by**: National Natural Science Foundation of China (61773270); Hundred Youth Program of Jiaxing University (CD70621004); Scientific Research Fund of Zhejiang Provincial Education Department (Y202249424)

old denoising module is developed to denoise the features in the process of feature extraction. To denoise the features in the process of network feature extraction, deep learning method is melted into the soft threshold denoising module designed using the neural network model beyond threshold-manual method. Then, to make the features extracted more distinctive, a joint gradient loss function is fed into softmax and Arcface(additive angular margin loss for deep face recognition) to extract more effective features. Finally, to make use of multiple frames of low-quality video data, a recurrent unit-gated video data fusion module is proposed to improve the quality of face-related data, which can optimize the mutual-benefited information between video frame data. **Result** To verify the effectiveness, comparative analysis is carried out in respect of two popular low-quality 3D face datasets, called the Lock3DFace(low-cost kinect 3D faces) and the Extended-Multi-Dim dataset. To be clarified, the experiments are followed by the prior training and testing protocol. Specifically, each of three protocols mentioned below are in comparison with the method of second-highest performance. For the Lock3DFace closed-set protocol, the average recognition rate is increased by 3.13%; For the Lock3DFace open-set protocol, the average recognition rate is optimized by 0.28%; For the Extended-Multi-Dim open-set protocol, the average recognition rate is improved by 1.03%. Furthermore, the ablation study demonstrates that the effectiveness and the feasibility of soft threshold denoising and video data fusion as well. **Conclusion** A trainable soft threshold denoising module is developed to denoise the low-quality 3D faces. The joint gradient loss function can be used to extract more distinctive features in relevant to softmax and Arcface. Furthermore, a video-based data fusion module is used to fuse information-added between video frames and the accuracy of low-quality 3D face recognition can be improved further. This low-quality 3D face recognition method can alleviate the degree of noise and integrate more effective information in terms of multiple frames of video data, which is potential to optimize low-quality 3D face recognition.

**Key words**: 3D face recognition; low-quality 3D face; soft threshold denoising; joint gradient loss function; video data fusion

## 0 引　言

近年来,受益于便携式3维传感技术的发展,基于低质量3维人脸的识别研究受到越来越多的关注。区别于传统高质量3维人脸数据(徐成华 等,2004),基于便携式3维传感器采集的3维人脸数据存在严重的质量差、噪声大和精度低等问题。图1展示了传统高质量和低质量3维人脸数据对比图。可以看出,低质量3维人脸数据表面存在大量毛刺,数据采集精度较低,给基于低质量3维人脸识别的研究带来很大困难。目前基于低质量3维人脸的识别精度很难令人满意(He 等,2016;Mu 等,2019;Liu 等,2019;龚勋和周炀,2021),基于低质量3维人脸识别方法的研究非常有限且面临诸多挑战。

现有基于低质量3维人脸识别方法主要围绕低质量3维数据质量提升、有效特征提取等方面开展研究,存在以下困难:1)在低质量3维人脸数据提升方面,现有方法大都基于单张深度数据优化或基于单张深度数据的重建进行低质量3维人脸识别研究。基于单张深度数据所能获取形状信息有限,如何利用现有多帧视频数据之间的互补信息进行低质量3维数据质量提升亟待解决。2)在有效特征提取方面,低质量3维人脸受噪声影响较大,导致其形状信息存在较大误差,增加了有效特征提取难度。


(a) 高质量3维图像

(b) 低质量3维图像

图1　低质量和高质量3维人脸数据对比图

Fig. 1　Low-quality and high-quality 3D face data comparison diagram ((a) high-quality 3D faces; (b) low-quality 3D faces)

针对上述存在问题,本文主要贡献如下:1)针对低质量3维人脸中存在的噪声影响,本文提出了一个即插即用的软阈值去噪模块。不同于传统的阈值去噪方法严重依赖于大量经验,本文结合深度学习方法,利用神经网络模型自动学习软阈值,在网络提取特征的过程中对特征进行去噪处理。2)为了实现低质量3维人脸多帧视频数据的融合,提出基于门

控循环单元的低质量3维人脸视频数据融合模块，自动提取低质量3维人脸视频帧数据间的依赖关系，实现视频帧数据间互补信息的有效融合。3）在有效特征提取方面，结合softmax和Arcface（additive angular margin loss for deep face recognition）提出了联合渐变损失函数，使网络提取更具有判别性特征，进一步提高了低质量3维人脸识别准确率。

## 1 相关工作

对低质量3维人脸的研究始于2010年以后，随着便携式3维采集设备Kinect v1的出现，3维人脸数据的获取变得更方便，也更能满足实际应用的需求。由于这些3维人脸数据质量较差，早期关于低质量3维人脸识别的研究主要是基于传统人脸识别方法，通常将这些低质量3维数据与2维RGB图像结合来进行人脸识别，以减轻RGB图像在识别中遇到的姿态、遮挡和光照等因素影响。例如，Li等人（2013）提出一套首先利用深度数据同时将RGB图像和深度图像归一化到正面姿态的预处理方法，然后通过稀疏表示分别对纹理和深度图进行相似度计算，再对相似度简单融合进行识别。在数据规模为52人的CurtinFace低质量3维人脸数据库（Li等，2013）中的不同姿态、表情、光照和遮挡等图像上都取得了较好效果。Hsu等人（2014）为了应对姿态变化，同时针对低质量3维人脸数据噪声大的问题，提出3D表面重建技术，利用特征点对人脸对齐，然后提取图像的局部二值模式（local binary patterns，LBP）特征，并使用稀疏表示分类进行识别。

随着Kinect v2和RealSense等更多便携式3维采集设备的相继出现和大型低质量3维人脸数据集Lock3DFace（low-cost kinect 3D faces）（Zhang等，2016）和Extended-Multi-Dim（Hu等，2019）的发布，一方面，使用这些设备获取的低质量3维人脸数据质量相比之前有了一定程度的改善；另一方面，由于低质量3维人脸数据库规模的扩大，逐渐出现了一些基于深度学习的方法来解决低质量3维人脸识别问题。Cui等人（2018）提出了第1个基于深度学习的低质量3维人脸模型，证明了利用深度学习方法对低质量3维人脸识别的可能性。在低质量3维人脸数据提升方面，为了减轻噪声、遮挡、姿态和表情等因素的影响，Hu等人（2019）提出了基于深度图像归一化为正面姿态和中性表情的低质量3维人脸识别算法。Mu等人（2019）提出了一个轻量化的深度学习模型和数据预处理方法。数据处理流程包括点云恢复、表面细化和数据增强等，轻量化的深度学习模型则由5层卷积神经网络结构块组成，并在其中使用4个跳跃连接来结合不同语义层面的信息，以生成更具有鉴别性的特征。为了减弱低质量3维人脸识别中噪声的影响，Zhang等人（2021）认为采集低质量3维人脸数据噪声服从一种分布，从而导致相应特征存在扰动，因此受扰动的特征也服从一种潜在分布（即给定3维人脸的后验分布），并提出基于低质量3维人脸识别的分布表示方法，在Lock3DFace数据集上取得了很好的识别结果，但该算法网络结构复杂且参数量较大，时间复杂度较高，训练难度大。

在特征提取方面，区别于早期的传统方法，Hu等人（2019）提出以高质量3维人脸数据为引导，提出3种使用高质量3维人脸引导低质量3维人脸识别模型训练的策略，减轻了低质量3维人脸特征提取难度。然而该算法需要同时使用高质量和低质量数据进行训练，数据获取难度较大，且目前缺少其他包含高低质量3维人脸的数据集。龚勋和周炀（2021）针对低质量3维人脸难以提取有效特征的问题，提出了基于dropout的空间注意力机制和类间正则化损失，有效提高了低质量3维人脸识别准确率。

## 2 网络模型介绍

### 2.1 网络结构

图2为本文提出的低质量3维人脸识别模型的整体结构。首先，将视频帧数据$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)\in\mathbf{R}^{n\times 128\times 128}$经过预处理（Yang等，2015）得到3D人脸的法线贴图输入到软阈值（soft thresholding，STD）-Led3D网络，在STD-Led3D网络的特征提取过程中，插入软阈值去噪模块（soft threshold denoising module，STDM）对数据噪声进行过滤，得到视频帧的特征表示$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)\in\mathbf{R}^{n\times 960}$，然后将这些特征输入门控循环单元融合模块对特征向量进行融合，得到视频级特征表示$\boldsymbol{r}^v\in\mathbf{R}^{n\times 960}$。图2中，MSFF（multi-scale-feature fusion）为多尺度特征融合模块，SAV（spatial attention ectorization）为空间注意矢量化模块。

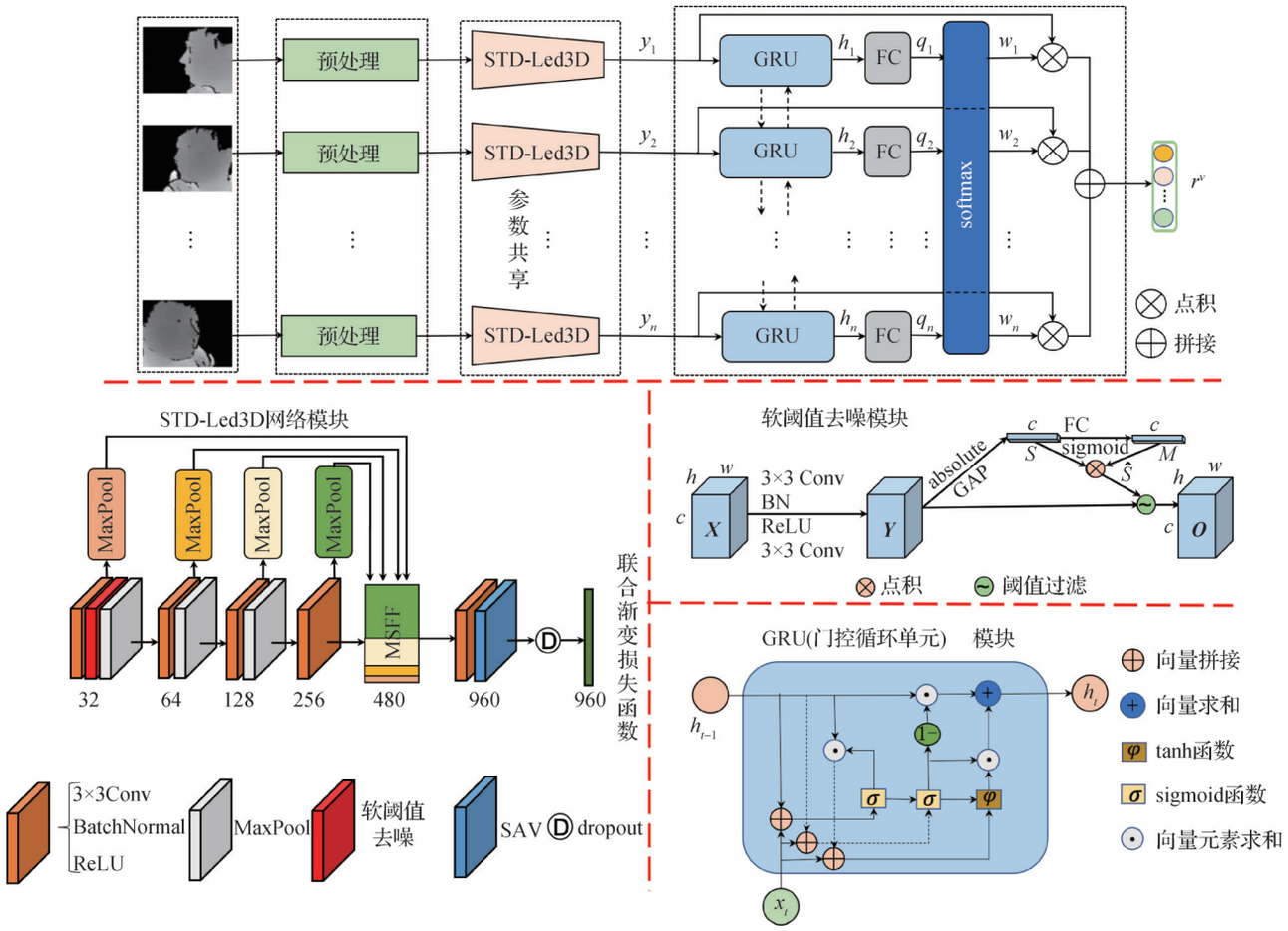


图 2　联合软阈值去噪和视频数据融合的低质量 3 维人脸识别模型示意图

Fig. 2　Diagram of soft threshold denoising and video data fusion-relevant low-quality 3D face recognition

## 2.2　软阈值去噪模块

传统阈值化方法通常利用人工设计的滤波器将有用信息转化为积极或消极的特征，并将噪声信息转化为接近零的特征。然而，设计这样的过滤器需要大量的经验。而深度学习中的梯度下降算法可以自动对滤波器的阈值进行学习，避免了阈值设定的开销。本文设计了一个即插即用软阈值去噪模块（STDM）以减轻噪声对网络提取特征的影响，提高模型对噪声的鲁棒性，其结构如图 2 所示。

软阈值去噪模块的输入为 $c \times h \times w$ 大小的特征图 $\boldsymbol{X}$，其中 $c$ 为通道数，$h$ 和 $w$ 分别为特征图的高度和宽度。输入特征图 $\boldsymbol{X}$ 首先通过由 3×3 卷积层、批归一化层、ReLU 激活层和 3×3 卷积层组成的连通结构进行特征变换，得到特征图 $\boldsymbol{Y}$。然后对变换后的特征图 $\boldsymbol{Y}$ 取绝对值和全局池化来获取软阈值模块的初始阈值 $\boldsymbol{S}$。为了使 $\boldsymbol{S}$ 不会过大，将 $\boldsymbol{S}$ 通过一层全连接层和 sigmoid 层，得到范围为 0～1 的缩放向量 $\boldsymbol{M}$。将 $\boldsymbol{M}$ 作用于向量 $\boldsymbol{S}$ 得到每个通道最终阈值 $\hat{\boldsymbol{S}}$，最后利用该阈值对特征图 $\boldsymbol{Y}$ 中的噪声进行过滤得到软阈值去噪模块输出 $\boldsymbol{O}$。

考虑到软阈值去噪模块主要作用是在网络提取特征的过程中进行特征去噪，而随着网络层靠后，噪声特征和有用特征将会混合到一起。因此，为了保证更好的去噪效果，本文将软阈值去噪模块插入到 Led3D（Mu 等，2019）网络中的第 1 个结构块后。Led3D 网络是第 1 个专门设计用来提高低质量 3 维人识别准确性和效率的卷积神经网络。

## 2.3　联合渐变损失函数

损失函数是特征提取的关键部分，特征提取过程也是使损失函数最小化的过程。损失函数越小，说明网络对当前训练数据的拟合能力越好，特征判别性越高。由于低质量 3 维人脸识别是细粒度识别问题，人脸之间相似性很强，如何设计损失函数来优化网络，使同类特征靠近、不同类特征尽量远离变得尤为重要。过去一段时间，低质量 3 维人脸识别领域的研究大多使用 softmax 损失来优化模型。但

softmax仅保证类别是可分的，并不要求同类特征紧凑、异类特征分离，使得最后识别准确率较低。而Arcface（Deng等，2019）损失函数可以使类内特征更加紧凑，同时类间特征产生明显的距离。

为了利用两个损失函数的优点，进一步提高网络的特征提取能力，使网络提取的特征同类更近、不同类更远，本文将softmax损失函数与Arcface损失函数相结合，提出了一种联合渐变损失函数，计算为

$$L = \lambda L_{\mathrm{s}} + (1 - \lambda) L_{\mathrm{a}} \tag{1}$$

$$\lambda = \begin{cases} 1 - \left(\dfrac{i}{T}\right)^{3} & i < T \\ 0 & i \geqslant T \end{cases} \tag{2}$$

式中，$\lambda$为权重参数，$i$表示迭代次数，$L_{\mathrm{s}}$和$L_{\mathrm{a}}$分别为softmax和Arcface损失函数。$\lambda$的值会随着训练次数不同而改变。具体来说，在训练的最初始阶段，$\lambda$为1，损失函数完全由softmax决定。随着迭代次数增加，当迭代次数达到$T$（根据网络的实际收敛情况，本文选用$T$为1 500）时，$\lambda$变为0，损失函数完全由Arcface决定。

直观上，本文提出的联合渐变损失函数会首先利用softmax优化网络，使网络迅速收敛。在训练过程中，逐渐增加Arcface的权重，慢慢提升模型训练难度，逐渐使同类特征距离更近、不同类间特征距离更远，从而使模型收敛到一个更好的特征空间。

### 2.4 门控循环单元数据融合模块

门控循环单元（gated recurrent unit，GRU）（Dey和Salem，2017）能够很好地对序列数据之间的相关信息进行建模，并已广泛用于各类时序任务中。本文使用门控循环单元来建模低质量3维人脸视频数据之间的相关性，提出基于门控循环单元数据融合模块对每帧低质量3维人脸视频数据进行融合，通过对每帧视频的所有特征表示来预测每个特征表示中每个维度的向量权值，然后加权和得到整个视频序列融合后的低质量3维人脸特征表示。

GRU的具体结构如图2所示。设当前节点输入为$x_t$，上一节点传送的包含先前节点相关信息的隐藏状态为$h_{t-1}$，利用两者，GRU会得到当前时间步的隐状态输出$h_t$，并将其传递到下一时间步，其过程可表示为

$$z_t = \sigma(h_{t-1}\boldsymbol{W}_z \oplus x_t \boldsymbol{U}_z) \tag{3}$$

$$r_t = \sigma(h_{t-1}\boldsymbol{W}_r \oplus x_t \boldsymbol{U}_r) \tag{4}$$

$$\tilde{h}_t = \tau(r_t \odot h_{t-1}\boldsymbol{W}_h \oplus x_t \boldsymbol{U}_h) \tag{5}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{6}$$

式中，$\sigma$表示sigmoid函数，$\tau$表示tanh函数，$\oplus$表示向量拼接，$\odot$表示向量元素相乘，$\boldsymbol{W}_z$, $\boldsymbol{U}_z$, $\boldsymbol{W}_r$, $\boldsymbol{U}_r$, $\boldsymbol{W}_h$, $\boldsymbol{U}_h$分别表示可学习的权重矩阵。

GRU首先利用当前时间步输入$x_t$和上一时间步输出的隐藏状态$h_{t-1}$来获取更新门控信息$z_t$和重置门控信息$r_t$。值得注意的是，在获取$z_t$和$r_t$前会将对应数据通过sigmoid函数，该函数会将数据范围变换为[0, 1]，转换之后的值越接近于1，代表记忆下来的信息越多，而越接近于0代表遗忘的信息越多。求得门控信息之后，GRU首先会使用$r_t$对输入的隐状态信息$h_{t-1}$进行重置，并与$x_t$进行拼接，再利用tanh激活函数功能将数据范围缩放为[-1, 1]，由此得到中间状态信息$\tilde{h}_t$。最后GRU使用更新门$z_t$来完成对记忆的遗忘和选择，$(1 - z_t) \odot h_{t-1}$表示对上一时间步隐藏状态中不重要的信息进行遗忘，$z_t \odot \tilde{h}_t$表示选择性记忆当前时间步$\tilde{h}_t$中的信息，通过两者可以得到当前时间步的输出$h_t$。

本文提出的门控循环单元数据融合模块由一个双向门控循环单元、一个全连接层和一个softmax归一化层构成，如图2所示。STD-Led3D网络输出的视频帧特征表示$\boldsymbol{Y}$会首先输入双向门控循环单元完成视频帧之间的依赖关系建模，即双向门控循环单元会对视频帧特征表示分别进行正方向和反方向的处理，得到每帧数据与其前后视频帧之间的关系表示$\boldsymbol{H}_b \in \mathbf{R}^{n \times 960}$和$\boldsymbol{H}_a \in \mathbf{R}^{n \times 960}$，这两个特征向量随后被拼接为$\boldsymbol{H} \in \mathbf{R}^{n \times 960}$，$\boldsymbol{H}$被送入全连接层预测视频帧特征的初始权值$\boldsymbol{Q} \in \mathbf{R}^{n \times 960}$。在特征融合前，利用softmax操作对所有特征表示在同一维度进行归一化。具体来说，给定初始权重集合$\boldsymbol{Q} = \{q_1, q_2, \cdots, q_n\}$，第$t$个视频帧特征向量的第$j$个成分归一化计算过程为

$$w_{ij} = \frac{\exp(q_{ij})}{\sum_{j=1}^{n} \exp(q_{ij})} \tag{7}$$

式中，$q_{ij}$表示第$i$个视频的第$j$个分量。在获得每帧数据每个维度的权重之后，即可将其与STD-Led3D输出特征加权求和，获得最终视频级特征表示，计算为

$$r^{v} = \sum_{i=1}^{n} y_{ij} \odot w_{ij} \tag{8}$$

式中，$n$表示视频帧数量，$\odot$表示向量元素相乘。

## 3 实验结果与分析

### 3.1 数据库及评价协议

3.1.1 数据库情况

为了评估本文方法的有效性，在 Lock3DFace（Zhang 等，2016）和 Extended-Multi-Dim（Hu 等，2019）两个低质量3维人脸数据集上进行验证。

Lock3DFace数据集采集自 Kinect v2，包含509人的5 711个视频样本，并伴随有表情、姿态、遮挡和时间流逝等方面的变化。数据集包括两个独立的部分，两部分采集时间间隔最长达7个月。所有的509人参加了第1阶段的数据采集，169人参加了第2阶段的数据采集。Lock3DFace数据集样例如图3所示。

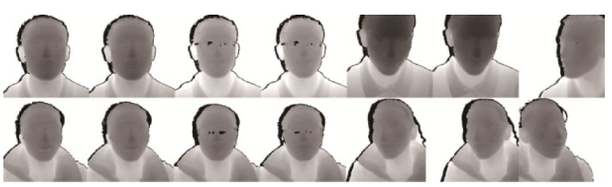

图3 Lock3DFace 数据集样例

Fig. 3 The samples of Lock3DFace dataset

Extended-Multi-Dim 是第1个包含高、低质量的3维人脸数据集，其中低质量深度图和彩色图像使用 RealSense 设备采集，高质量3维人脸使用 SCU（Sichuan University）高精3维扫描仪采集。该数据集共包含902个不同样本，是最大的多模态人脸数据集。每个采集样本伴随3种表情、水平方向[+90°，−90°]和俯仰角方向[+15°，−15°]连续姿态变化的数据。Extended-Multi-Dim数据集样例如图4所示。

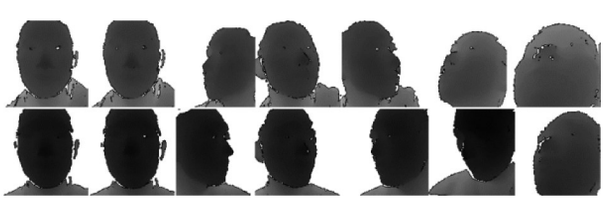

图4 Extended-Multi-Dim数据集样例

Fig. 4 The samples of Extended-Multi-Dim dataset

3.1.2 评估协议

1）Lock3DFace闭集评估协议。本文采用与 Mu 等人（2019）相同的实验设置进行训练和测试。具体地说，将每个身份的第1个自然表情视频数据作为训练数据，剩余视频划分为表情、遮挡、姿态和时间4个测试子集。其中，时间子集只使用自然和表情数据。由于Lock3DFace数据集中每个视频内数据相似性过大，因此所有视频都以相等的间隔选出6帧作为代表。最后生成6 617个训练样本，1 283个表情子集测试样本，1 004个遮挡子集测试样本，1 010个姿态子集测试样本，676个时间子集测试样本。与 Mu 等人（2019）方法一致，在 Lock3DFace 数据集上训练的模型都使用高质量3维人脸数据集 FRGC v2 和 Bosphorus（Savran 等，2008）合并之后的数据集进行预训练，预训练学习率为0.005，其他与正式训练保持一致。

2）Lock3DFac开集协议。随机从509个身份中选取300个身份的所有视频数据作为训练数据，剩余209个身份的视频数据作为测试集。并对训练视频数据中每个人的第1个自然表情视频数据进行数据增强，其余使用原始数据，共生成3 272个训练样本。测试集中每个人的第1个自然无表情视频作为图库样本，剩余视频作为测试样本，共包含自然、表情、遮挡、姿态和时间5个测试子集，分别包括205，520，407，417，256个样本。与Lock3DFace闭集协议一致，所有视频都以相等的间隔选出6帧作为代表。

3）Extended-Multi-Dim开集评估协议。采用与 Hu 等人（2019）相同的实验设置进行训练和测试，训练集包括430人，约5 082组训练样本。测试时，以每个身份的第1个自然无表情视频中的第1个样本作为图库数据，其余视频中所有样本作为测试数据，共包含自然（中性表情）、表情（张嘴、皱鼻、闭眼等）、姿态1（水平旋转头部）和姿态2（顺时针旋转头部）4个子集，分别包括2 184，1 356，857，870个样本。与Lock3DFace协议一致，所有视频都以相等的间隔选出6帧作为代表。

### 3.2 消融实验

本文所有消融实验均采用Lock3DFace闭集评估协议，并以Led3D网络为基准网络模型。

为了验证本文提出的软阈值去噪模块、渐变损失函数模块和门控循环单元模块的有效性，分别在基准模型上应用相应的模块以及同时叠加所有模块进行消融实验。

首先，为了验证软阈值去噪模块的有效性，在基准模型的不同位置添加软阈值去噪模块（STDM），结

果如表1所示。模型1表示在基准模型中的第1个卷积神经网络(convolutional neural networks, CNN)模块前加入STDM;模型2表示在第1个CNN模块后添加STDM;模型3表示在第2个CNN模块后添加STDM;模型4表示在第3个CNN模块后添加STDM。所有模型都使用softmax损失函数进行训练。

从表1可以看出,与基准模型相比,随着在基准模型中加入STDM的位置逐渐靠后,Rank-1识别准确率先上升后下降。其中,模型2的Rank-1识别准确率最高,相比基准模型高了2.49%,而模型4的准确率最低,相比基准模型低了约2.27%。由于网络利用不同层进行特征变换,随着层次深入,有效特征和噪声特征会逐渐混合到一起难以分离,因此插入位置越深,效果越差。根据表1的结果,本文使用模型在第1个CNN模块后添加软阈值去噪模块。

**表1　不同位置添加软阈值去噪模块的Rank-1识别率**

**Table 1　Rank-1 recognition rate of adding soft threshold denoising module in different locations**

/%

| 方法 | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|
| 基准 | 76.65 | 34.86 | 32.84 | 23.52 | 45.92 |
| 模型1 | 77.67 | **37.95** | 30.77 | 22.34 | 46.29 |
| 模型2 | **81.01** | 33.86 | **37.08** | **26.18** | **48.61** |
| 模型3 | 78.91 | 33.57 | 35.21 | 22.63 | 46.77 |
| 模型4 | 74.79 | 31.67 | 31.76 | 20.12 | 43.65 |

注:加粗字体表示各列最优结果。

为了直观展示软阈值去噪模块在特征提取过程中的有效性,选取1幅高质量3维人脸数据,向其添加不同强度的高斯噪声,然后将经过软阈值去噪模块前后的特征可视化,观察模块的去噪效果。如图5所示,第1列为加入高斯噪声,第2～7列为经过软阈值去噪模块前的特征,第8～13列为经过软阈值去噪模块后的特征。不难看出,随着噪声强度的增加,输出特征图中噪声响应越多,人脸判别性区域特征越来越不明显。而经过软阈值去噪模块后的特征包含噪声响应较少,有效特征也更加明显。尽管在此添加的是高斯噪声,不难得出,本文提出的软阈值去噪模块不仅在直观上确实减弱了特征中的噪声,在性能上也提高了低质量3维人脸识别准确率(表1)。

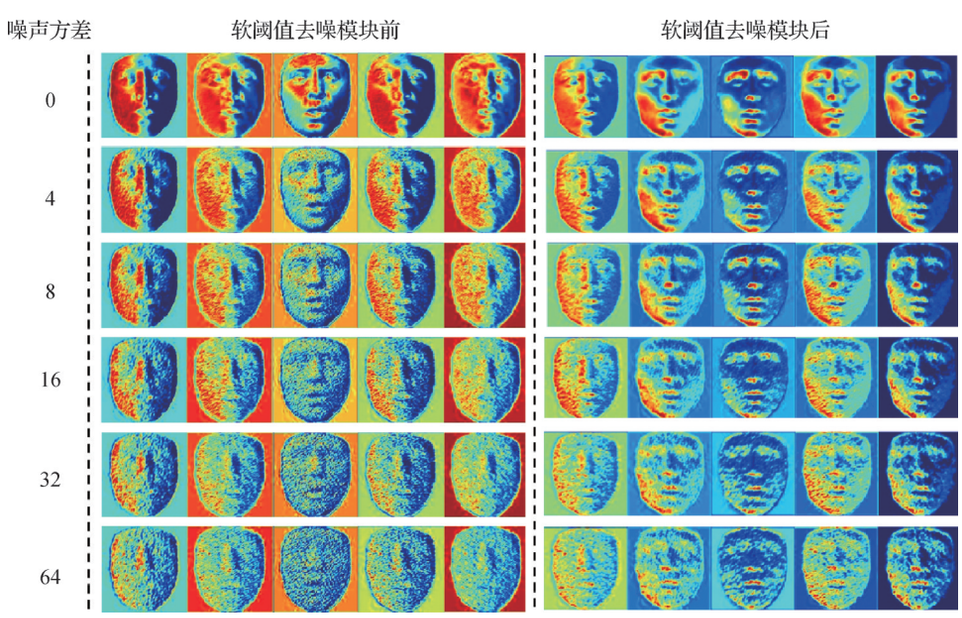


图5　软阈值去噪模块前后特征可视化

Fig. 5　Feature visualization begore and after the soft threshold denoising module

其次,为了验证联合渐变损失函数的有效性,分别使用softmax、Arcface和联合渐变损失3种损失函数对基准模型进行训练。其中,Arcface和联合渐变损失函数中的超参$s$和$m$使用了4组不同的设置。如表2所示,在不同超参设置下,使用softmax损失函数的模型结果远不如使用Arcface和联合渐变损失的模型结果,Rank-1准确率最多相差12.06%。另外,随着角边距惩罚项$m$逐渐增大,使用Arcface损失函数训练的模型平均准确率先是增大后又减小,这是由于随着角边距惩罚项的增大,同类特征距离

更紧凑，不同类特征距离变大，特征判别性高，识别效果好。而当角边距惩罚项超过一定值之后，其准确率下降。这是由于角边距惩罚项过大导致模型学习难度增加，而无法学习到一个很好的特征空间，因此识别结果大幅下降。与此相反的是，使用联合渐变损失函数的模型不仅在同样角边距惩罚项设置下比使用 Arcface 损失函数的模型取得的 Rank-1 识别准确率都高，同时，两者准确率差距随着角边距惩罚项的增大先是逐渐减小后又逐渐增大，在 $m = 0.7$ 时达到了约 10.20% 的差距。以上结果说明，本文提出的联合渐变损失函数在不同参数设置下都有助于基准模型收敛到一个更好的特征空间，提升了模型识别准确率。

**表 2　不同损失函数的 Rank-1 识别率**

**Table 2　Rank-1 recognition rate of different loss functions**

/%

| 损失函数 | $s$ | $m$ | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|---|---|
| softmax | – | – | 76.65 | 34.86 | 32.84 | 23.52 | 45.92 |
| Arcface | 32 | 0.1 | 84.75 | 38.15 | 39.74 | 28.11 | 51.90 |
| Arcface | 32 | 0.3 | 87.86 | 42.43 | 43.39 | 30.62 | 55.34 |
| Arcface | 32 | 0.5 | 90.66 | 45.92 | 44.97 | 28.70 | 57.20 |
| Arcface | 32 | 0.7 | 80.86 | 38.45 | 34.91 | 18.05 | 47.78 |
| 联合渐变损失 | 32 | 0.1 | 84.98 | 40.74 | 41.03 | 30.33 | 53.33 |
| 联合渐变损失 | 32 | 0.3 | 88.64 | 44.42 | 44.47 | **30.77** | 56.40 |
| 联合渐变损失 | 32 | 0.5 | 91.05 | 45.72 | 45.86 | 29.44 | 57.63 |
| 联合渐变损失 | 32 | 0.7 | **91.60** | **48.00** | **46.25** | 26.48 | **57.98** |

注：加粗字体表示各列最优结果，“–”表示该参数无效。

为了进一步直观展示联合渐变损失函数的有效性，分析各参数对方法性能的影响，将联合渐变损失函数（以 $s = 32$，$m = 0.7$ 为例）的训练损失曲线图可视化。作为对比，softmax 和 Arcface 损失函数也一并可视化，如图 6 所示。可以看出，softmax 损失函数的损失值起始值比较低，且很快收敛至 0 附近；Arcface 损失函数的起始值很高，收敛至 15 附近就趋于平缓；而联合渐变损失函数的起始值与 softmax 一致，训练中经历了先降低后增加再降低的过程，这是由于联合渐变损失函数中 Arcface 所占权重逐渐增大的缘故。结合表 2 的结果，说明联合渐变损失函数能在加快模型训练收敛速度的同时，使模型学习到一个更好的特征空间，从而提高低质量 3 维人脸识别准确率。

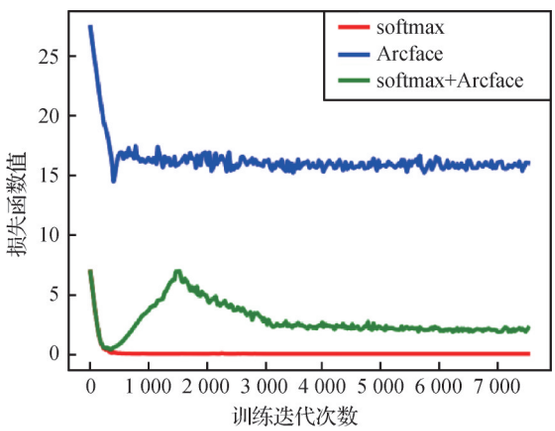


图 6　损失函数对比

Fig. 6　Comparison of different loss functions

再次，为了验证本文提出的门控循环单元数据融合模块的有效性，设计了以下 3 种基准数据融合模型。1）投票法（Vote）。该方法在获取每个视频帧的身份之后，采用投票的方式确定整个视频的身份；2）最大池化法（Maxpool）。该方法对所有视频帧使用最大池化获取同一维中的最大响应值构成视频特征表示；3）单向门控循环单元（单向 GRU）。该方法使用一层单向 GRU 网络对输入视频帧特征进行融合，由于最后一个节点的输出包含前面所有帧的相关信息，因此直接将其作为整个视频的特征表示。本文提出的视频数据融合模型则是在软阈值去噪和联合渐变损失函数的基础上，添加双向门控循环单元模块，结果如表 3 所示。

**表 3　不同融合模型的 Rank-1 识别结果**

**Table 3　Rank-1 recognition rate of different fusion models**

/%

| 融合模型 | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|
| Vote | 92.67 | 46.51 | 47.43 | 29.73 | 58.80 |
| MaxPool | 92.83 | **47.11** | 47.23 | 30.62 | 59.10 |
| 单向 GRU | 92.60 | 46.22 | 46.73 | 30.62 | 58.67 |
| 视频数据融合 | **93.06** | 46.31 | **48.12** | **39.95** | **59.43** |

注：加粗字体表示各列最优结果。

从表 3 可以看出，投票法 Vote 和最大池化法 MaxPool 没有对视频帧之间的互补特征进行学习，所以 Rank-1 识别率较差。此外，由于前置 CNN 网络已经提供了可识别特征，而单向 GRU 融合模型中最后

一个节点输出的特征进行加权的表示可能会与原始CNN输出特征有较大不同，在数据量有限的情况下，进行新的特征学习可能会导致过拟合，所以单向GRU识别性能比视频数据融合方法差。本文提出的视频数据融合模型在大部分测试子集中都实现了最高的识别准确率，在最后平均识别率上都高于其他方法，表明了本文提出的视频数据融合方法的有效性。

最后，为了验证本文提出软阈值去噪模块、联合渐变损失函数和门控循环单元模块叠加之后的有效性，在表1取得最优软阈值去噪模块位置的模型2上分别叠加联合渐变损失函数和门控循环单元模块，结果如表4所示。从表4可以看出，相比基准模型，本文提出的任一模块都对最终的识别准确率有益，且叠加之后的模型取得了最佳识别性能。

**表4　叠加模块在Lock3DFace闭集协议上的对比结果**

**Table 4　Comparison results of superposition of modules on Lock3DFace close-set protocol**

/%

| 方法 | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|
| 基准 | 76.65 | 34.86 | 32.84 | 23.52 | 45.92 |
| 软阈值去噪模块 | 81.01 | 33.86 | 37.08 | 26.18 | 48.61 |
| 软阈值去噪模块+联合渐变损失 | 92.23 | 45.92 | 47.34 | 30.77 | 58.98 |
| 软阈值去噪模块+联合渐变损失+GRU模块 | **93.06** | **46.31** | **48.12** | **39.95** | **59.43** |

注：加粗字体表示各列最优结果。

### 3.3　与现有方法比较

#### 3.3.1　Lock3DFac闭集协议实验结果

为验证本文方法的性能，与现有的低质量3维人脸识别方法VGG16（Visual Geometry Group network）（Simonyan和Zisserman，2014）、ResNet34（residual network）（He等，2016）、Inception-V2（Ioffe和Szegedy，2015）、MobilNet-V2（Sandler等，2018）、Led3D（lightweight and efficient deep approach for 3D faces）（Mu等，2019）、SAD（龚勋和周炀，2021）、NAN（neural aggregation network）（Yang等，2017）和MAA（meta attention-based aggregation）（Liu等，2019）方法进行对比。

表5展示了上述方法在Lock3DFace闭集上的实验结果。可以看出，本文方法实现了最好的性能，相比其他最好的结果，准确率提升了3.13%。相比其他方法，本文方法的识别结果总体上有所提升，这是由于本文方法提出的软阈值去噪模块在特征提取过程中对噪声进行过滤，减轻了噪声的影响；而本文提出的联合渐变损失函数有效利用了softmax和Arcface损失函数各自的优点，有效降低了模型的训练难度，使网络收敛到一个判别性更好的特征空间；另外，本文提出的视频数据融合模块可以融合视频帧之间的互补特征，其对视频帧之间序列进行建模，因此效果最好。

**表5　不同算法在Lock3DFace闭集协议上的对比结果**

**Table 5　Comparison results of different algorithms on Lock3DFace close-set protocol**

/%

| 方法 | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|
| VGG16 | 79.63 | 36.95 | 21.70 | 12.84 | 42.80 |
| ResNet34 | 62.83 | 20.32 | 22.56 | 20.70 | 32.23 |
| Inception-V2 | 80.48 | 32.17 | 33.23 | 12.54 | 44.77 |
| MobilNet-V2 | 85.38 | 32.77 | 28.30 | 10.60 | 44.92 |
| Led3D | 86.94 | 48.01 | 37.67 | 26.12 | 54.28 |
| SAD | 87.02 | **65.47** | 45.17 | 23.52 | 54.83 |
| NAN | 92.99 | 45.02 | 45.64 | 29.29 | 57.99 |
| MAA | 92.99 | 46.31 | 48.12 | 31.95 | 59.40 |
| REC | 74.12 | 18.63 | 28.57 | 13.17 | 34.53 |
| 本文 | **93.06** | 46.31 | **48.12** | **39.95** | **59.43** |

注：加粗字体表示各列最优结果。

#### 3.3.2　Lock3DFace开集协议实验结果

表6展示了不同方法在Lock3DFace开集上的实验结果。可以看出，本文方法在平均识别率上依然高于其他所有对比模型，再次说明了本文提出方法的有效性。同时，可以发现本文方法在姿态子集中的结果远高于其他对比模型，说明本文方法由于学习了视频帧之间的互补特征，从而提高了识别准确

率。另外，与表5相比，表6中相同测试子集的识别结果更好，这是因为在开集上本文使用了多种类型的训练数据进行训练，使模型泛化能力得到了增强。

**表6　不同算法在Lock3DFace开集协议上的对比结果**

**Table 6　Comparison results of different algorithms on Lock3DFace open-set protocol**

/%

| 方法 | 自然 | 表情 | 遮挡 | 姿态 | 时间 | 平均 |
|---|---|---|---|---|---|---|
| NAN | 98.05 | 96.73 | 60.20 | 60.67 | 62.50 | 75.46 |
| MAA | 99.02 | 98.46 | 74.94 | 69.78 | 72.27 | 82.88 |
| 本文 | **99.02** | **98.65** | **75.68** | **69.54** | **73.05** | **83.16** |

注：加粗字体表示各列最优结果。

3.3.3　Extended-Multi-Dim开集协议实验结果

表7展示了不同方法在Extended-Multi-Dim开集上的实验结果。可以看出，本文方法实现了最好的性能，相比其他最好的MAA方法平均有1.03%的准确率提升。在姿态1和姿态2测试子集中，本文方法的识别结果比MAA方法分别高出了0.58%和2.87%。Extended-Multi-Dim中的视频数据相比Lock3DFace数据集中的视频数据，人脸在姿态子集中有较大的姿态变化，因此未考虑视频帧相关性的其他方法在两个姿态测试子集中的效果较差。而本文方法使用门控循环单元获取不同帧之间的互补信息，通过互补信息预测特征表示每个维度的权重，然后对特征表示加权求和获取最后的视频级特征表示，有效提高了低质量3维人脸识别准确率。同时，也说明了本文方法在应对较大姿态变化时具有良好性能。

**表7　不同算法在Extended-Multi-Dim开集协议上的对比结果**

**Table 7　Comparison results of different algorithms on Extended-Multi-Dim open-set protocol**

/%

| 方法 | 自然 | 表情 | 姿态1 | 姿态2 | 平均 |
|---|---|---|---|---|---|
| ResNet34 | 92.68 | 87.24 | 55.99 | 41.54 | 75.31 |
| Inception-V2 | 86.73 | 79.47 | 52.56 | 37.36 | 69.67 |
| MobilNet-V2 | 90.10 | 85.67 | 55.39 | 38.87 | 73.33 |
| BDB | 94.20 | 90.70 | 60.80 | 42.30 | 80.00 |
| Led3D | 96.20 | 94.80 | 67.90 | 47.20 | 81.80 |
| NAN | 97.85 | 98.01 | 96.50 | 84.02 | 95.39 |
| MAA | 98.99 | 98.16 | 98.25 | 87.93 | 96.83 |
| 本文 | **99.51** | **98.75** | **98.83** | **90.80** | **97.86** |

注：加粗字体表示各列最优结果。

## 4　结　论

本文围绕低质量3维人脸数据噪声大、依赖单幅有限深度数据提取有效特征困难的问题，提出了一种联合软阈值去噪和视频数据融合的低质量3维人脸识别方法。本文提出的软阈值去噪模块，将去噪过程直接融入深度学习网络模型，避免了传统阈值设置严重依赖人工经验的缺陷；在有效特征提取方面，本文结合softmax和Arcface提出的联合渐变损失函数使网络提取更具有判别性特征；另外，本文提出的视频数据融合模块，利用门控循环单元对低质量3维人脸视频帧特征数据间的依赖关系建模，实现视频帧数据间互补信息的有效融合，进一步提高了低质量3维人脸识别准确率。大量的对比实验证明了本文网络模型的有效性。

## 参考文献（References）

Cui J Y, Zhang H, Han H, Shan S G and Chen X L. 2018. Improving 2D face recognition via discriminative face depth estimation//Proceedings of 2018 International Conference on Biometrics. Gold Coast, Australia: IEEE: 140-147 [DOI: 10.1109/ICB2018.2018.00031]

Deng J, Guo J, Xue N and Zafeiriou S. 2019. Arcface: additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4690-4699 [Doi: 10.1109/CVPR.2019.00482]

Dey R and Salem F M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks//Proceedings of the 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Boston, USA: IEEE: 1597-1600 [DOI: 10.1109/MWSCAS.2017.8053243]

Gong X and Zhou Y. 2021. 3D face recognition for low quality data. Journal of University of Electronic Science and Technology of China, 50(1): 43-51 (龚勋, 周炀. 2021. 面向低质量数据的3D人脸识别. 电子科技大学学报, 50(1): 43-51) [DOI: 10.12178/1001-0548.2020321]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on

Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hsu G S J, Liu Y L, Peng H C and Wu P X. 2014. RGB-D-based face reconstruction and recognition. IEEE Transactions on Information Forensics and Security, 9(12): 2110-2118 [DOI: 10.1109/TIFS.2014.2361028]

Hu Z G, Gui P H, Feng Z Q, Zhao Q J, Fu K R, Liu F and Liu Z X. 2019. Boosting depth-based face recognition from a quality perspective. Sensors, 19(19): #4124 [DOI: 10.3390/s19194124]

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR.org: 448-456

Li B Y L, Mian A S, Liu W Q and Krishna A. 2013. Using kinect for face recognition under varying poses, expressions, illumination and disguise//Proceedings of 2013 IEEE Workshop on Applications of Computer Vision (WACV). Clearwater Beach, USA: IEEE: 186-192 [DOI: 10.1109/WACV.2013.6475017]

Liu Z X, Hu H, Bai J Q, Li S H and Lian S G. 2019. Feature aggregation network for video face recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 990-998 [DOI: 10.1109/ICCVW.2019.00128]

Mu G D, Huang D, Hu G S, Sun J and Wang Y H. 2019. Led3D: a lightweight and efficient deep approach to recognizing low-quality 3D faces//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5766-5775 [DOI: 10.1109/CVPR.2019.00592]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B and Akarun L. 2008. Bosphorus database for 3D face analysis//Proceedings of the 1st European Workshop on Biometrics and Identity Management. Roskilde, Denmark: Springer: 47-56 [DOI: 10.1007/978-3-540-89991-4_6]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: #1556 [DOI: 10.48550/arXiv.1409.1556]

Xu C H, Wang Y H and Tan T N. 2004. Overview of research on 3D face modeling. Journal of Image and Graphics, 9(8): 893-903 (徐成华, 王蕴红, 谭铁牛. 2004. 3维人脸建模与应用. 中国图象图形学报, 9(8): 893-903) [DOI: 10.3969/j.issn.1006-8961.2004.08.001]

Yang J L, Ren P R, Zhang D Q, Chen D, Wen F, Li H D and Hua G. 2017. Neural aggregation network for video face recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5216-5225 [DOI: 10.1109/CVPR.2017.554]

Yang X D, Huang D, Wang Y H and Chen L M. 2015. Automatic 3D facial expression recognition using geometric scattering representation//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Ljubljana, Slovenia: IEEE: #7163090 [DOI: 10.1109/FG.2015.7163090]

Zhang J J, Huang D, Wang Y H and Sun J. 2016. Lock3DFace: a large-scale database of low-cost kinect 3D faces//Proceedings of 2016 International Conference on Biometrics. Halmstad, Sweden: IEEE: #7550062 [DOI: 10.1109/ICB.2016.7550062]

Zhang Z H, Yu C C, Xu S and Li H B. 2021. Learning flexibly distributional representation for low-quality 3d face recognition//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI: 3465-3473 [DOI: 10.1609/aaai.v35i4.16460]

## 作者简介

桑高丽,女,副教授,主要研究方向为模式识别。
E-mail: glsang@zjxu.edu.cn
赵启军,通信作者,男,教授,博士生导师,主要研究方向为模式识别和机器视觉。E-mail: qjzhao@scu.edu.cn
肖述笛,男,硕士研究生,主要研究方向为图像处理。
E-mail: xiaoshudi@stu.scu.edu.cn