文章编号：2095-2163(2020)06-0303-07 中图分类号：TP391.4 文献标志码：A

面向三维人脸重建的自编码体素网络研究

董俊呈，左旺孟
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：在过去的几十年里，单幅人脸图像三维重建技术在计算机视觉和图形学领域中获得了极大的关注。自编码体素网络可以通过一张照片来估计一张人脸的高质量的体素模型，拥有不错的效果。本文从自编码体素网络的模型结构，引导项和损失函数三个方面对其进行了改进，给出了改进方案和测试结果，证明改进是有效的。

关键词：三维重建；自编码体素网络；体素模型

Research on self-coding volumetric network for 3D face reconstruction

DONG Juncheng, ZUO Wangmeng
(College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 In the past few decades, the three-dimensional reconstruction technology ofsingle face images has gained great attention in the field of computer vision and graphics. The self-encoding volumetric network can estimate the size of a face through a photo, and it has a good effect. Here the three aspects of the self-encoding voxel network model structure, the leading term and the loss function are improved, and an improvement scheme and test results are proposed to prove that the improvement is effective.

【Key words】 3d face reconstruction; self-encoding volumetric network; volumetric model

0 引 言

本文主要研究单幅人脸图像的三维重建问题，基于 VRN 论文的相关方法和技术，完成面部照片三维重建任务的端到端的神经网络。本文首先验证了现有各种三维重建方案的效果、性能和可行性，同时对 3DMM 和 VRN 进行复现并验证效果；其次，验证基本无误，并且复现效果达到 baseline 水平后对 VRN 的模型结构，损失函数和引导项这三个方向进行了改进。

1 对现有工作的复现和验证

1.1 三维可变形模板(3DMM)

本文实现了传统的 3DMM 重建方法，用蒙特卡洛法对输入进行拟合，在适当的初始化条件下可以得到不错的效果。

代码实现的操作大体如下：

a.读取 BFM 数据集，经 PCA 后构建特征值和特征向量，目标是计算拟合所对应的的各个特征值系数。

b. 对于任意一个要拟合的人脸，检测 36，39，42，45，31，33，35，48，54，51，57 号特征点，计算在齐次坐标系下经过平移，水平拉伸和竖直拉伸后得到的与原图对应特征点的 MSE 距离最小的情况作为初始化，如图 1 所示。

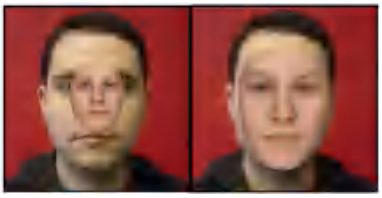


图 1 通过人脸特征点进行初始化

Fig. 1 Initialization by face landmark

c.如图 2 所示，调用蒙特卡洛算法，以颜色直方图的 MSE 距离作为优化目标，对三维人 ß 脸的特征向量系数进行优化。如果拟合中误差小于设定的最小阈值，则可以提前结束；如果误差大于设定的最大阈值，则认为模型已经偏离梯度下降方向，结束拟合过程，返回-1；否则，算法进行 2 000 次后停止，返回当前的最好结果。

如果初始化得当，最终可以取得较好的拟合结果，如图 3 所示。

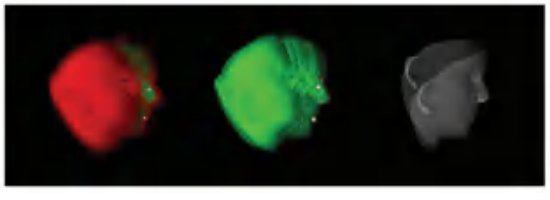


图 2 蒙特卡洛算法进行拟合过程

Fig. 2 The fitting process with Monte Carlo

作者简介：董俊呈(1996-)，男，硕士研究生，主要研究方向：计算机视觉、三维重建；左旺孟(1977-)，男，博士，教授，教师，主要研究方向：计算机视觉、机器学习与生物特征识别等方面的研究。
通讯作者：董俊呈 Email：syao_ran@ qq.com

收稿日期：2020-03-24

图 3 传统 3DMM 拟合结果

Fig. 3 The result of the traditional 3DMM fitting

利用蒙特卡洛方法对三维人脸进行拟合伪代码如下：

算法 1 利用蒙特卡洛方法对三维人脸进行拟合

输入：待拟合三维人脸特征向量系数矩阵 G，输入 RGB 图片 I，蒙特卡洛步长 l

输出：拟合结果人脸特征向量系数矩阵

```
1.function MontFit(G, I, l):
2.   for i in range(2000):
3.        if MSE(Z(P(G)), Z(I)) > ThresholdMax:
4.             return -1
5.        end if
6.        if MSE(Z(P(G)), Z(I))<ThresholdMin:
7.             return G
8.        end if
9.        L← {for i in range(20), MontStep(G, l)} + {G}
10.       temp L← {for i in range(20), MSE(Z(P(L[i])), Z(I))}
11.       G← L[minIndex(tempL)]
12.  end for
13.return G
```

1.2 自编码体素网络(VRN)

VRN 是一个端到端的神经网络，输入是一张三通道 RGB 或灰度的任意姿态，任意光照，任意表情，允许遮挡的人脸照片，输出是一个三维人脸的体素表示[1]，即一个 192×192×200 的三维矩阵，其中数字“1”代表该位置有一个体素立方体，“0”则代表没有，这个三维人脸向 Z 轴的垂直投影应该与输入人脸对齐。需要注意的是，由于姿态变化，人脸（尤其是鼻子导致的）会有自遮挡问题，因此这个体元表示与简单输出一张深度图是有区别的。

本文将 VRN release 的 MATLAB 代码重写成了 pytorch 代码，完成了 training 和 testing 的工作，并用原文所列出的训练集对模型进行了训练并达到了 baseline，在原文中提供的测试集 Florence 和 AFLW2000-3D 上均达到了原文的水平，同时对文中用于比较 VRN 性能的重建方法 EOS 和 3DDFA 在对应数据集上进行了验证，与 VRN 提供的数据基本一致，本文复现 VRN 的可视化结果如图 4 所示。

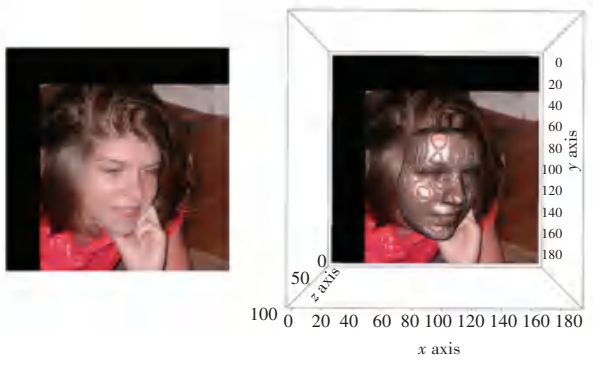
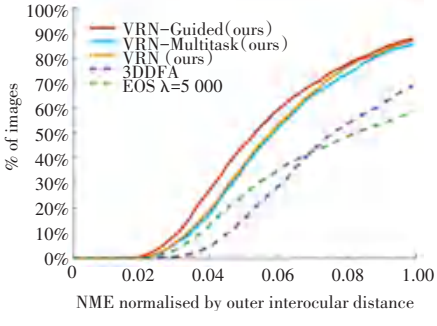


图 4 VRN 复现的可视化结果

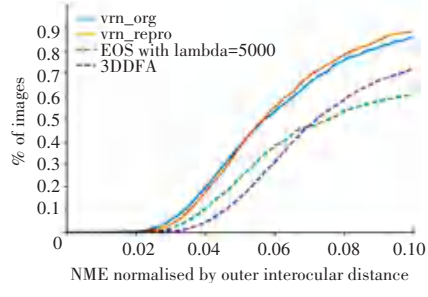
Fig. 4 Visualization results of VRN reproduction

同时测试了文中用于比较效果的 3DDFA 和 EOS，证明 VRN 的方法是可行的。图 5 是在 AFLW2000-3D 上比较 VRN，复现 VRN（VRN-repro），EOS 和 3DDFA 的 NME 损失，图 5（a）是 VRN 论文中的结果，图 5（b）是复现的结果；图 6 是在 Florence 上比较 VRN，复现 VRN（VRN-repro），EOS 和 3DDFA 的 NME 损失，图 6（a）是 VRN 论文中的结果，图 6（b）是复现的结果。本文在各数据集上各个方法的平均 NME 损失值如表 1 所示。



（a）原文中的结果

（a）Result in the original paper

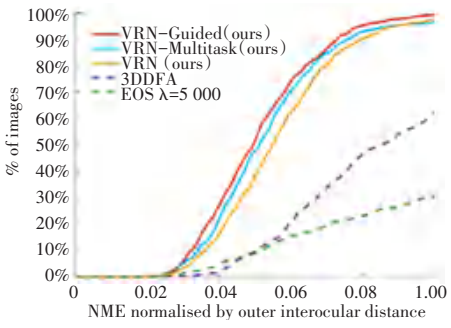


（b）复现结果

（b）Result of the reproduction

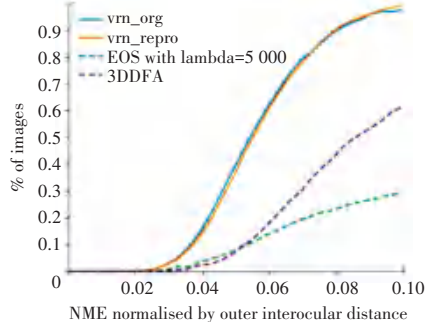
图 5 AFLW2000 数据集上的结果

Fig. 5 The result on AFLW2000



（a）原文中的结果

（a）Result in the original paper



（b）复现结果

（b）Result of the reproduction

图 6　Florence 数据集上的结果

Fig. 6　The result on Florence

表 1　在各数据集上各个方法的平均 NME 损失

Tab. 1　The average NME loss of each method on each data set

Method	AFLW2000	Florence
VRN	0.068 3	0.057 0
VRN-repro	0.068 1	0.057 1
3DDFA	0.100 6	0.097 4
EOS	0.097 1	0.125 1

2　对自编码体素网络的改进

VRN 网络是一个端到端的，简洁轻量的模型，但是模型的表达效果仍然没有达到理想的效果。因此，本文又训练了 vrn-multitask，来提取人脸特征点的热度图，把热度图信息和原图一起输入到 vrn-guided 中来优化输出，确实得到了提升。但是本文认为 VRN 采用的 U-Net 结构是可以改进的，尝试如 Fish-Net 这些被证明相同结构下效果更好的网络[2]。另外，只有二维的特征点信息并不能最好的起到引导的作用，希望加入 pose 等更多的信息来对 VRN 进行引导，试着得到更好一些的效果。VRN 采用的全局的损失本文认为也是有一定不足的，显然人脸内部的体素权重应当小于靠近边缘和表面的体素。

2.1　对自编码体素网络结构的改进

在 VRN 中，本文使用两个串联的 UNET 端到端训练了一个输出体元人脸的网络，U-Net 使用的“上/下采样+跳跃连接”的结构，使得其构成的神经网络具有易收敛、轻量级，深层网络容易更快的获取浅层网络梯度，保留了图片各个像素的位置信息的优点。但也存在当多个 U-Net 共同工作于同一个模型时，各个 U-Net 直接配合较差的问题，据此 UNET 被提出后，已经产生了很多基于 UNET 结构的其他模型结构，如 FishNET 等。

Fish-Net 是对 U-Net 的一种改进。Fish-Net 认为，当多个 U-Net 串联时，单个 U-Net 内的对应上采样和下采样之间有跳跃连接，但两个相邻的 U-Net 之间的下采样和上采样之间没有跳跃连接，因此两个 U-Net 之间的通路可能会成为梯度传播的瓶颈；同时 Fish-Net 的作者提取了相邻两个 U-Net 对应的下采样层和上采样层，发现从语义信息的角度这两个特征也处于不同的域。因此 Fish-Net 除了将下采样层和自身对应的上采样层进行连接，还将每个 U-Net 的上采样层和后面相邻的一个 U-Net 的下采样层做了跳跃连接，使得后面的 U-Net 可以更容易的感受到前面 U-Net 的梯度。

在 Fish-Net 中，有两种用于上采样和下采样的卷积块，分别是上采样-重制块（UR-block）和下采样-重制块（DR-block）。通过在 FishNet 中设计的身体和头部，将尾部和身体各个阶段的特征连接到头部。Fish-Net 精心设计了头部中的各层，以使其中没有 I-conv。头部中的层是由串联，具有特征的卷积和池化层组成。因此，Fish-Net 解决了尾部在躯干网络前获得梯度传播的问题，用到的两种方法分别是：1）排除头部的 I-conv 和 2）在身体和头部使用串联。为了避免像素之间重叠，对于跨度为 2 的下采样 Fish-Net，将卷积核大小设置为 2×2，消融实验显示了网络中不同种类的内核大小对实验效果的影响。为了避免 I-conv 问题，应避免采用上采样方法中的加权反卷积，为简单起见，Fish-Net 选择最近邻插值进行上采样，由于上采样操作将以较低的分辨率稀释输入特征，Fish-Net 在重制模块中还应用了膨胀卷积，该方法被证明是可行并且确实可以提高 UNET 效果的，本文将 UNET 替换成 FishNET，并对数据结构进行相应的更改并重新训练，实验证明在相同参数和模型规模下，不论是 AFLW2000 数据集上，表 2 所示，还是 Florence 数据集上，表 3 所示，FishNET 的表现都要优于 UNET（图 7）。

另外，本文提出了 MR-UNET，如图 8 所示，来对原 UNET 进行多尺度条件下的改进，实验结果表明，在相同的网络规模和参数量下，Stacked UNET 表现不如原 UNET，但随着网络规模的增加，其准确

度依然有很高的上限,且其网络结构和输出的特征与 FishNET 和 UNET 有着较好的契合度。因此,本文在后面的实验中也使用该网络来产生用于引导原网络的 pose 信息。

表 2 AFLW2000 上各个模型的参数规模和对应的 NME-LOSS

Tab. 2 The parameter scale and corresponding NME-LOSS of each model on AFLW2000

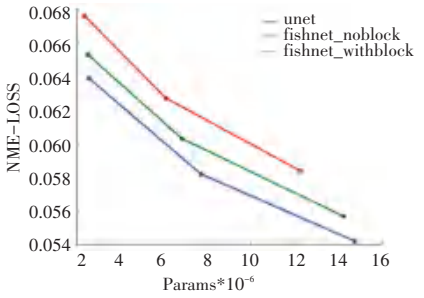
Method	Params	NME-LOSS
UNET-2	2.365M	0.067 6
UNET-5	6.109M	0.062 7
UNET-10	12.251M	0.058 4
FISHNET_NoBlcok-2	2.511M	0.065 3
FISHNET_NoBlcok-5	6.843M	0.060 3
FISHNET_NoBlcok-10	14.236M	0.055 7
FISHNET_Blcok-2	2.561M	0.063 9
FISHNET_Blcok-5	7.709M	0.058 2
FISHNET_Blcok-10	14.790M	0.054 2
3DDFA	2.874M	0.101 2
EOS	3.163M	0.097 1

表 3 Florence 上各个模型的参数规模和对应的 NME-LOSS

Tab. 3 The parameter scale and corresponding NME-LOSS of each model on Florence

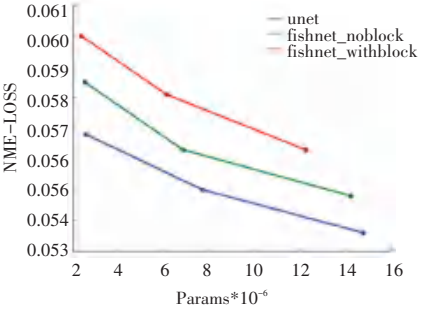
Method	Params	NME-LOSS
UNET-2	2.365M	0.060 0
UNET-5	6.109M	0.058 1
UNET-10	12.251M	0.056 3
FISHNET_NoBlcok-2	2.511M	0.058 5
FISHNET_NoBlcok-5	6.843M	0.056 3
FISHNET_NoBlcok-10	14.236M	0.054 8
FISHNET_Blcok-2	2.561M	0.056 8
FISHNET_Blcok-5	7.709M	0.055 0
FISHNET_Blcok-10	14.790M	0.053 6
3DDFA	2.874M	0.097 5
EOS	3.163M	0.125 3

MR-Net 全程端到端训练模型,使用 RMSProp 方式。首先关闭所有上下采样通路,使模型中只有主干网络(第一行)处于工作状态,初始化学习率,每 40 个 epoch 后学习率衰减为之前的 0.1。在训练中对数据进行一系列增强操作:输入图片被施加一个 XOY 平面的旋转,旋转处于{-45,…,45}之间的整数,然后被施加一个随机的平移操作,平移距离是{-15,…,15}之间的整数像素,然后被施加一个缩放,由于尽量不使面不变形过于明显,以及方便 groundtruth 的 z 方向,可以根据输入图片的变化产生对应变化,本文中的缩放均使用等比例缩放,这样 groundtruth 的三维数据可以直接按照相同的比例进行缩放,随机缩放比例处于 1-{-0.15,…,0.15}之间,随机选取 20%的样本做水平翻转,最后输入数图片在 RGB 三个通道分别做等比例的随机亮度调整,调整范围在{0.6,…,1.4}之间。同时,作为对应的三维人脸也要做同样的变换,与输入的 RGB 图片保持对齐。



(a) AFLW2000 下的结果

(a) Result on AFLW2000



(b) Florence 下的结果

(b) Result on Florence

图 7 FishNET 与 UNET 的参数规模和 NME-LOSS 关系的比较

Fig. 7 Comparison of FishNET and UNET parameter scale and NME-LOSS relationship

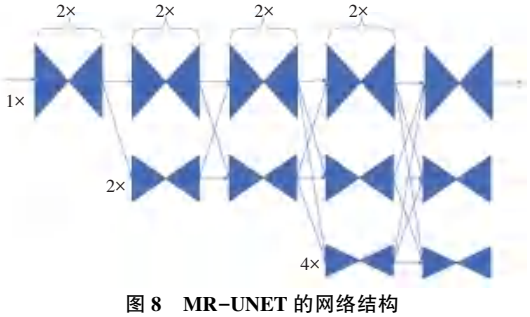


图 8 MR-UNET 的网络结构

Fig. 8 MR-UNET network structure

在主网络训练至 LOSS 不再下降,打开对应通道,使第二行的网络加入训练,训练参数相较于第一行训练参数均减少为原先的一半,训练至 LOSS 不再下降;同样在模型的 LOSS 稳定且不再下降后,打开对应通路将第三行的网络加入模型,训练方式仿照第一二行的情况,同样需注意第三行和前两行的数据应保持等比例情况下的一致,且 groundtruth 的三维人脸应做对应的变换来与输入图像保持对齐。

同样在模型的 LOSS 稳定且不再下降后,打开对应通路将第三行的网络加入模型,训练方式仿照第二行的情况,同样需注意第三行和前两行的数据应保持等比例情况下的一致,且 groundtruth 的三维人脸应做对应的变换来与输入图像保持对齐。

本文对 MR-UNET 与 UNET 的参数规模和

NME-LOSS 关系的比较，结果如表 4 所示。可见 MR-UNET 在单幅人脸图像三维重建任务上达到了最低的 NME-LOSS。

表 4　MR-UNET 与 UNET 的参数规模和 NME-LOSS 关系的比较

Tab. 4　Comparison of the parameter scale of MR-UNET and UNET and the relationship between NME-LOSS

Method	Params	NME-LOSS
UNET-2	2.365M	0.067 6
UNET-5	6.109M	0.062 7
UNET-10	12.251M	0.058 4
MR-Net	23.277M	**0.051 7**
3DDFA	2.874M	0.101 2
EOS	3.163M	0.097 1

2.3　对自编码体素网络引导项的研究

简单的两个串联的 UNET 模型表达能力有限，因此又训练了一个 vrn-multitask 用于输出人脸特征点的热度图，模型结构如图 9 所示。将这个热度图与原输入连接到一起，输入网络进行重建，让这个特征点的热度图对原模型进行引导，称为 vrn-guided，网络结构如图 10 所示。

图 9　VRN-multitask 的网络结构

Fig. 9　VRN-multitask network structure

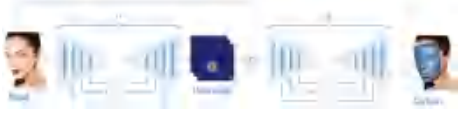


图 10　VRN-guided 的网络结构

Fig. 10　VRN-guided network structure

在 vrn-duided 中，首先训练了一个叉状网，如图 9 所示。输入图片进入一个 U-Net 后，输出的特征被分为两份，分别输入到两个单独的 U-Net 中，上半部分用于预测输入人脸的热度图，下半部分用于预测三维重建结果，其中面部特征点热度图和三维体素人脸的损失同时能影响到左边第一个 U-Net 的参数学习，vrn 原文中称这个网络为 vrn-multitask，这个模型可以同时预测输入图片中人脸的特征点概率分布热度图和重建体素三维模型。从模型角度来看，vrn-multitask 的左下半部分（去除第二列最上面的一个 U-Net）与 vrn-unguided 模型结构一致。

提取 vrn-multitask 的左上半部分的热度图提取网络。首先将 RGB 人脸图片输入该网络，得到 192 ×192 × 68 的面部特征点热度图矩阵，将其和输入图片的 192 × 192 × 3 的矩阵连接，这一步要确定两者维度的对齐，一起输入到重建网络中进行重建，这个流程的模型就是 vrn-guided，结构如图 10 所示。

本文认为二维的特征点的热度图并不能最好的对模型进行引导，原图中很多信息并没有被包含进去：如姿态、光照等信息。因此，希望能训练一个网络对姿态等信息进行预测，并与特征点信息一起对原模型进行引导，尝试达到比 VRN 更好的效果。

2.3.1　面部特征点信息用于引导

在 VRN 原文中，本文使用了一个另外的网络用于面部特征点的检测，将检测结果转化为 192x192x68 的热度图与 vrn-unguided 连接后再输入到 UNET 中，用于引导三维重建过程，本文首先复现了该工作并达到了 baseline，复现结果的 NME-LOSS，如表 5 所示。

表 5　VRN-guided 复现结果的 NME-LOSS

Tab. 5　NME-LOSS of the VRN-guided reproduction

Method	AFLW2000	Florence
VRN-guided	0.637	0.509
VRN-guided-repro	0.639	0.509

本文认为就人脸特征点的表达来说，使用热度图并不是唯一且最好的方法。通过面部特征点提取的神经网络获得人能理解的面部特征点的热度图，再从热度图转化为机器能理解的神经网络特征，经历了两次不同 domain 的翻译过程，这个翻译的过程可能导致一些信息的损失和网络训练难度增加。因此，本文在 LFPW，HELEN，AFW，AFLW 等数据集上训练了一个以 UNET 为基本结构的面部特征点检测网络，在 IBUG 和 MUG 数据集上测试达到 dlib 的标准化 MSE 误差，将倒数第二层的特征提取代替原先的特征点热度图进行引导。

希望倒数第二层的特征更好地起到引导重建的作用，本文首先将预测特征点网络和三维重建网络分别训练作为初始化，打开连接两个网络的通道一起训练，重复上面的两个步骤几次以后，得到最终结果，模型结构如图 11 所示。

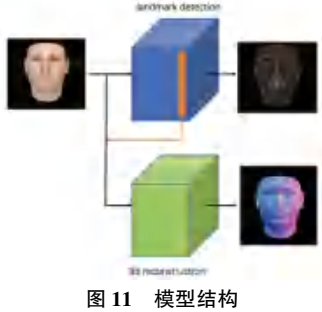
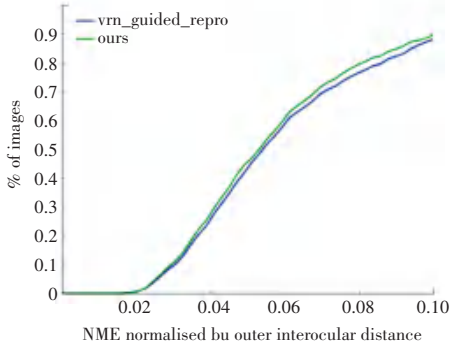


图 11　模型结构

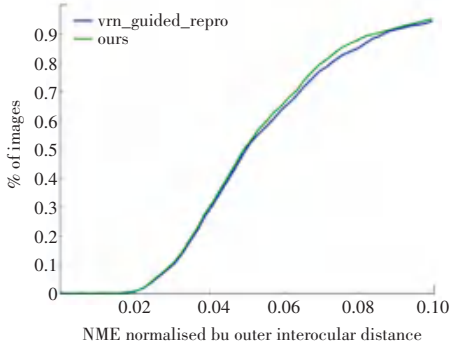
Fig. 11　The model structure

最终保持了面部特征点检测的准确性,达到了 dlib 相当的 baseline,同时得到了比使用特征点热度图更高的结果,如图 12 所示。图 12(a)是在 AFLW2000 上的结果,图 12(b)是在 Florence 上的结果。



(a) AFLW2000 下的结果

(a) Result on AFLW2000



(b) Florence 下的结果

(b) Result on Florence

图 12 特征点信息引导方法与 VRN 的比较

Fig. 12 Comparison of vrn and method with feature point information guidance

2.3.2 面部姿态信息用于引导

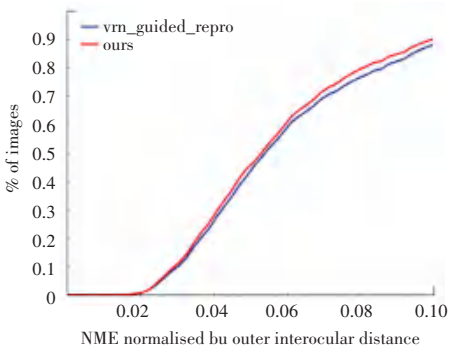
同时本文发现 MR-UNET 在面部姿态预测有着很好的表现,因此本文参考面部特征点信息用于引导的方法,将 MR-UNET 的倒数第二层特征用于补充引导 VRN-guided,最终在原基础上得到了更好的效果。图 13(a)是在 AFLW2000 上的结果,图 13(b)是在 Florence 上的结果。

首先单独训练 MR-UNET 和 VRN-guided 作为初始化,然后将两个网络连接起来同时训练,重复这两个步骤若干次直到重建损失不再下降。

2.4 对自编码体素网络损失函数的研究

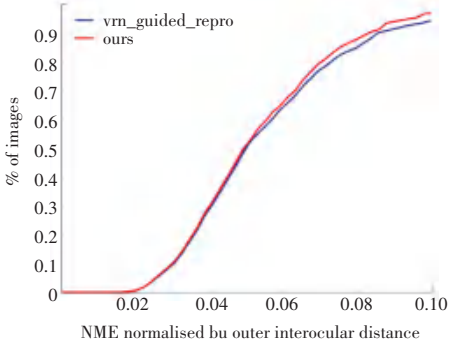
在 VRN 的原文中,使用了一个全局的交叉熵损失函数作为网络的 LOSS 进行训练,式(1):

$$L = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D}\left[V_{uhd}\log\hat{V}_{uhd} + (1-V_{uhd})\log(1-\hat{V}_{uhd})\right]. \quad (1)$$



(a) AFLW2000 下的结果

(a) Result on AFLW2000



(b) Florence 下的结果

(b) Result on Florence

图 13 姿态信息引导的方法与 vrn 的比较

Fig. 13 Comparison of vrn and method with pose information guidance

近期在目标检测领域 Focal-Loss 被提出,用于优化交叉熵损失函数[3]。目标检测通常被分成两阶段和一阶段两种算法,前者的代表是 Faster RCNN,这类算法准确率高但执行效率低,虽然可以通过减少 proposal 的数量或者降低输入图像的分辨率等方式来进行提速,但实际上治标不治本,速度并没有质的提升;后者的代表是 yolo,这种直接回归的检测算法效率高,但准确度低。经过实验研究表明单阶段的算法不如两阶段的算法准确度高是因为样本类别不均匀,在目标检测中,成千上万个候选位置中只有少部分是正样本,导致样本不均衡,这使负样本占据了总 LOSS 的大部分,而且大多数都是简单样本,导致了模型优化偏离了预期,之前的 OHEM 方法也试图解决样本不均匀的情况,但是它虽然增加了分错的样本的权重,却忽略了容易分类的样本。针对这个问题,本文提出了 focal loss,通过减少易分类样本的权重使得模型在训练时能够更加专注于难分类的样本,同时在原文中还训练了一个 retinaNet 来证明 focal loss 是有效的。实验结果表明 retinaNet 即具有单阶段检测器的速度,又拥有两阶段检测器的准确度。