Computer Science and Application 计算机科学与应用, 2023, 13(12), 2500-2508
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312248

Hans汉斯

基于DECA的单目人脸三维重建研究

王承伟[1]，张翠军[1,2*]，王振凯[1]，李昊渊[1]，张一帆[1]

[1]河北地质大学信息工程学院，河北 石家庄
[2]河北地质大学人工智能与机器学习研究室，河北 石家庄

收稿日期：2023年11月26日；录用日期：2023年12月21日；发布日期：2023年12月29日

摘 要

针对基于单目图像的DECA模型在人脸三维重建时精度不高，且容易出现过拟合的问题，提出用Vision Transformer (ViT)改进DECA模型的特征提取器部分，增强模型的局部和全局理解能力，提取更高维的特征，以提高人脸特征点的检测精度和人脸重建的精确性。进一步，引入DropKey策略，将ViT中的Key作为Drop对象，惩罚注意力峰值，以改善训练过程中的过拟合问题。实验结果表明，在引入ViT和DropKey策略后，人脸三维重建的效果有明显的提升。

关键词

人脸重建，深度学习，Vision Transformer，DropKey

Research on Three-Dimensional Reconstruction of Monocular Face Based on DECA

Chengwei Wang[1], Cuijun Zhang[1,2*], Zhenkai Wang[1], Haoyuan Li[1], Yifan Zhang[1]

[1]College of Information Engineering, Hebei GEO University, Shiajizhuang Hebei
[2]Laboratory of Artificial Intelligence and Machine Learning, Hebei GEO University, Shiajizhuang Hebei

Received: Nov. 26th, 2023; accepted: Dec. 21st, 2023; published: Dec. 29th, 2023

Abstract

In view of the low accuracy of the monocular image-based DECA model in face 3D reconstruction and the problem of overfitting, Vision Transformer (ViT) is proposed to replace the feature ex-

*通讯作者。

文章引用: 王承伟, 张翠军, 王振凯, 李昊渊, 张一帆. 基于 DECA 的单目人脸三维重建研究[J]. 计算机科学与应用, 2023, 13(12): 2500-2508. DOI: 10.12677/csa.2023.1312248

tractor part of the DECA model to enhance the local and global understanding ability of the model and extract higher-dimensional features. To improve the accuracy of face feature point detection and face reconstruction. Further, the DropKey strategy is introduced, and the Key in ViT is used as Drop object to punish the attention peak, so as to improve the overfitting problem in the training process. The experimental results show that after the introduction of ViT and DropKey strategies, the effect of face 3D reconstruction has been significantly improved.

Keywords

Face Reconstruction, Deep Learning, Vision Transformer, DropKey

1. 引言

人脸三维重建，作为计算机视觉领域的一个关键分支，致力于从单张或多张人脸图像中还原人脸的三维几何结构。单目人脸重建方法，相较于多目方法，具备更低的图像获取成本和更快的数字人物头部网格生成效率。

传统的人脸三维重建方法主要依赖几何学和计算机视觉技术，如多视角几何、结构光、稠密重建等。这些方法通过特征提取和模型匹配等步骤，以图像本身表达的信息，如视差和相对高度，完成三维模型的还原。林琴等人[1]结合局部立体匹配算法，对初步估计的脸部稠密视差值进行平滑处理，重建人脸的点云信息，在人脸数据库 Bosphorus 上获得了更加精确的重建结果。Castelan 等人[2]引入 SFS (Shape From Shading)方法，利用成像表面亮度的变化，解析出人脸表面的矢量信息，从而重建出人脸深度信息。叶于平等[3]提出了基于 3D 优化的标定方法优化结构光系统标定参数来提高重建精度，通过基于 GPU 的非刚性配准算法和纹理融合等算法重建出高精度高保真度的人脸动画表情。Blanz 等人[4]提出了基于变形模型的方法，该方法通过调整 3DMM 模型(3D Morphable Model)中描述不同形状和纹理的 PCA 系数拟合三维人脸模型，利用自适应对齐算法将目标图像进行对齐，使得生成模型更好的匹配目标图像；Cao 等人[5]又在 3DMM 的基础上增加了人脸表情。

随着深度学习方法在单目人脸三维重建中的迅速发展，人脸重建质量得到了显著提升。一些研究探索了卷积神经网络(CNN)的应用，以解决在人脸重建领域的困难。Tuan Tran 等人[6]提出 3DMM CNN 方法使用卷积神经网络 ResNet101 [7]对 3DMM 模型的形状系数和纹理系数直接进行了回归。Zhu 等人[8]针对 3DMM 的输入只有一张图像的问题，将 RGB 图像和 PNCC (Projected Normalized Corrdinate Code)特征合并输入，通过权重调整的方式优先拟合关键形状参数，提高了模型的精度。Feng 等人[9]提出的 PRNet 模型利用 UV 位置图描述三维形状并在计算损失函数时对不同区域的顶点加权，以更精准的预测坐标，实现了以端到端的方式实现人脸三维重建。

经典的基于 3DMM 模型的研究[10]都会面临着数据采集和处理方面的严重困难，而且难以精确捕捉人脸形状和纹理的复杂变化。为了解决这一问题，近年来提出了 FLAME 头部模型[11]，它通过整合多源异构数据集构建了更为精确的模型，可同时描述形状和纹理。Detailed Expression Capture and Animation (DECA)模型[12]则进一步引入深度学习技术，以生成 UV 图和细节、形状和表情等参数，从而更为鲁棒地重建人脸的形状和表情。

DECA 作为单目人脸重建领域的深度学习模型，在提高人脸几何结构和还原面部纹理上发挥着重要作用。尽管它成功结合了的 FLAME 模型，但在实验中，DECA 模型在重建精度方面表现不佳，且容易受到过拟合问题的影响。本文引入了 Vision Transformer (ViT)作为特征提取器[13]，并采用 DropKey 技术[14]改善训练问题。ViT 是一种强大的深度学习架构，因其卓越的特征提取能力而有望提高人脸几何结构的精度和纹理的还原质量，从而提供更精准的面部信息。同时，DropKey 技术的引入有助于减轻过拟合问题的影响，增强 DECA 模型的泛化性能，提高模型在不同数据集和场景中的表现。

2. 改进 DECA 模型

2.1. 整体框架

DECA 模型是基于 FLAME 模型的进一步发展，它使用深度学习技术从一个低维的潜在表征中生成 UV 位置图，并通过训练回归器来预测细节、形状和表情等参数。改进的 DECA 模型以二维人脸图像为输入，经过 Encoder 层进行特征编码，在 Encoder 层中采用了 ViT 模型和 DropKey 相结合的方法来提取图像的特征。随后，这些特征通过一个全连接层被映射为一个低维的潜在编码，包括相机编码(camera code)、反射率编码(albedo code)、光照编码(light code)、形状编码(shape code)、姿势编码(pose code)和表情编码(expression code)。这个潜在编码进一步被用于解码，其中反射率编码通过生成网络 D_A 输出反射率贴图，用于体现模型的纹理和颜色；FLAME 模型通过形状编码，姿势编码和表情编码对 3D 模型进行形变调整；最后利用可微渲染器(Differentiable Renderer)渲染最终的二维人脸图像以及通过最小化输入图像和输出图像之间的差异来进行模型的优化。整体框架如图 1 所示。

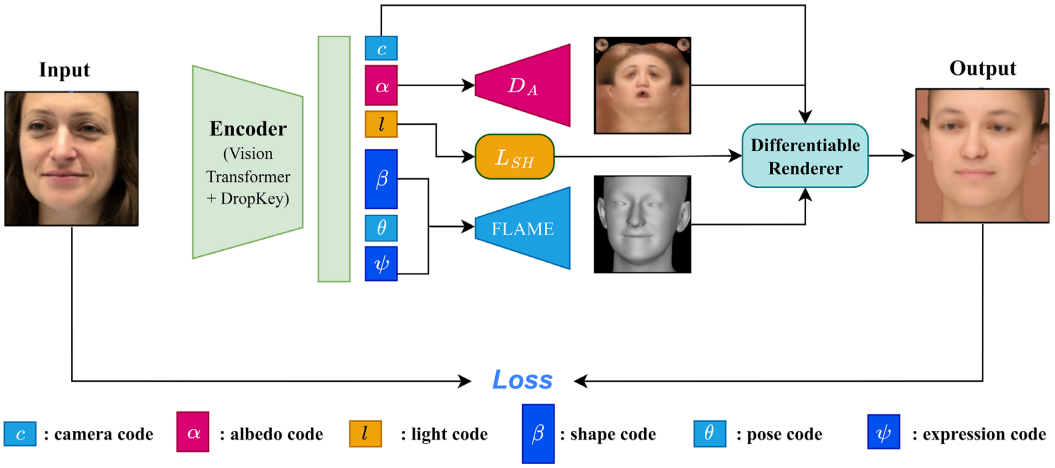


Figure 1. Improve the overall DECA frame diagram
图 1. 改进 DECA 整体框架图

2.2. Vision Transformer

ViT 作为深度学习模型架构，旨在将自然语言处理领域的 Transformer 模型中的注意力机制引入到计算机视觉任务中，在图像分类，目标检测和图像特征提取等多项视觉任务上获得了与 CNN 相媲美甚至更出色的表现。

与传统的 CNN 不同，ViT 无需手动设计卷积核或池化层来提取特征，它将输入图像分割成小图像块(patch)，然后通过 Embedding 层将这些图像块转换为 token 序列并将这些 token 作为模型的输入。在 token 输入 Transformer Encoder 之前额外加入[class]token 和位置编码，位置编码使得模型可以理解输入图像块

之间的相对位置关系，有助于模型对图像全局结构的理解。Transformer Encoder 层从输入的 token 中提取关键特征，其中自注意力机制用于理解输入图像块之间的关系，多头注意力机制使得模型能够在不同空间位置上关注不同特征，有助于捕捉图像的全局和局部信息。最后，MLP Head 层对从 Transformer Encoder 层得到的特征进行进一步的处理和输出。MLP Head 由一个或多个全连接层组成，其目的是将高维的特征向量转换成适合于不同具体任务的结果，例如生成用于图像分类的类别标签、用于目标检测的边界框和类别信息，以及用于语义分割的像素级别掩模。ViT 的网络结构如图 2 所示。

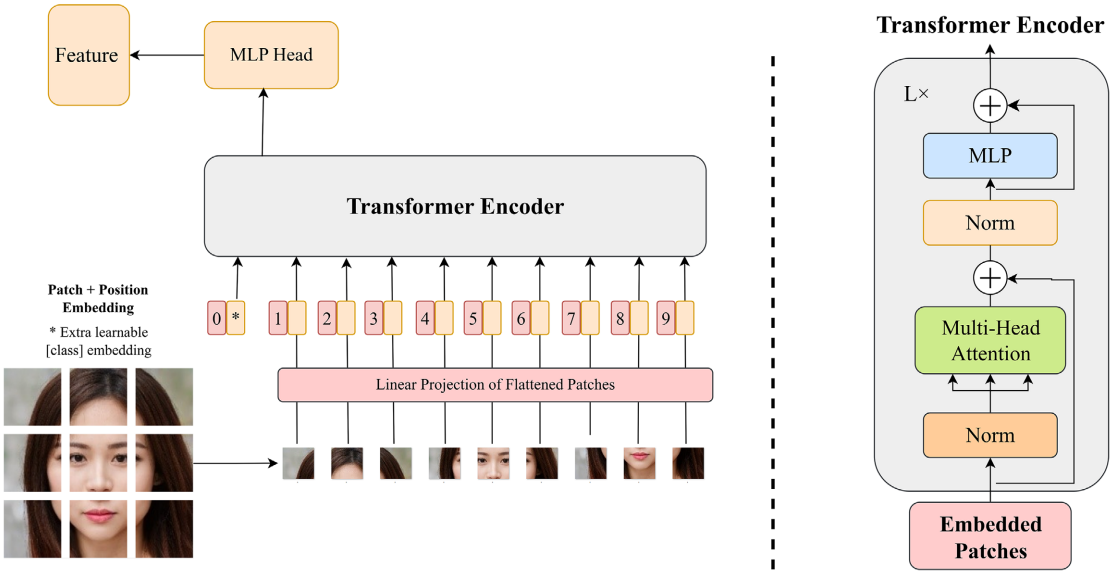


Figure 2. Vision Transformer network structure diagram
图 2. Vision Transformer 网络结构图

2.3. DropKey 策略

当面临相对较小的数据集时，基于 Transformer 的算法容易受到过拟合问题的困扰，当前的 ViT 模型通常采用 CNN 中常见的 Dropout 正则化策略，即在注意力权重图上进行随机 Drop 并为不同深度的注意力层设置统一的 Drop 概率。然而，在 Softmax 归一化后进行随机 Drop 可能会破坏注意力权重的概率分布，且无法对权重峰值进行惩罚，导致 ViT 过拟合局部特定信息，如图 3(b)所示。另一方面，不同深度的注意力层需要不同的 Drop 概率，恒定的 Drop 概率会导致训练不稳定，使得模型高维语义信息缺失或者低维细节特征过拟合。

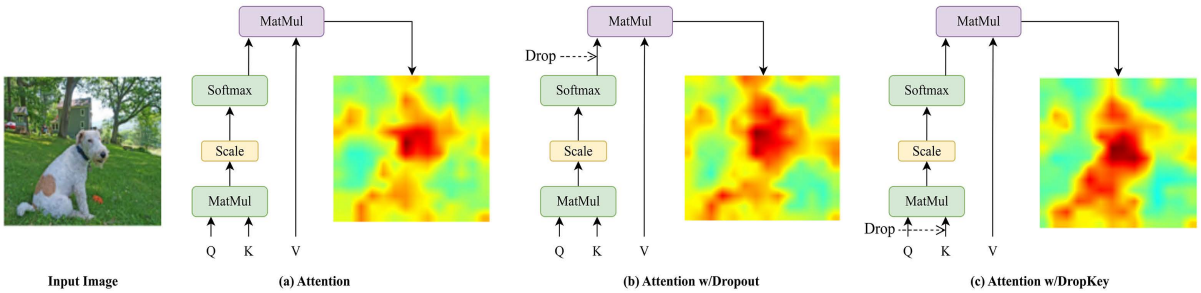


Figure 3. Attention-layer comparison map based on DropKey and Dropout
图 3. 基于 DropKey 和 Dropout 的注意力层对比图

为了克服 CNN 采用的 Dropout 策略在 ViT 上缺乏有效性的问题，DropKey 将 Key 设置为 Drop 对象，能够对注意力峰值进行惩罚，使得 ViT 模型可以更关注与目标有关的其他图像块，有助于捕捉全局特征，如图 3(c)所示。与 Dropout 不同，DropKey 为不断加深的注意力层设置递减的 Drop 概率策略，防止 ViT 模型过拟合并且有足够的高维特征进行稳定的训练，DropKey 的具体操作如表 1 所示。

Table 1. Attention with DropKey code
表 1. 使用 DropKey 的注意力机制代码

算法 1：Attention with DropKey code

```
# N: token number, D: token dim
# Q: query (N, D), K: key (N, D), V: value (N, D)
# use_DropKey: whether use DropKey
# mask_ratio: ratio to mask
def Attention(Q, K, V, use_DropKey, mask_ratio)
    attn = (Q * (Q.shape[1] ** -0.5)) @ K.transpose(-2, -1)
    # use DropKey as regularizer
    if use_DropKey == True:
        m_r = torch.ones_like(attn) * mask_ratio
        attn = attn + torch.bernoulli(m_r) * -1e-12
    attn = attn.softmax(dim=-1)
    x = attn @ V
return x
```

2.4. 损失函数

改进 DECA 模型的损失函数由多个损失组成，定义如下：

$$L_{reco} = L_{lmk} + L_{eye} + L_{pho} + L_{id} + L_{sc} + L_{reg} \tag{1}$$

其中，L_{lmk} 为人脸特征点重投影损失，L_{eye} 为眼睛闭合损失，L_{pho} 为光度损失，L_{id} 为身份损失，L_{sc} 为形状一致性损失，L_{reg} 为正则化损失。

人脸特征点重投影损失用于度量人脸图像中 2D 人脸特征点真实值 k_i 与 FLAME 模型表面上对应的特征点 $M_i \in R^3$，经由估计的相机模型投影到图像中的误差，定义如下：

$$L_{lmk} = \sum_{i=1}^{68} \left\| k_i - s\Pi\left(M_i\right) \right\|_1 \tag{2}$$

眼部闭合损失表示上眼睑和下眼睑对应的特征点 k_i 和 k_j 的相对偏移，并度量其与 FLAME 模型表面上的对应特征点 M_i 和 M_j 投影到二维图像中的偏移之间的误差，定义如下：

$$L = \sum_{(i,j)\in E} \left\| k_i - k_j - s\Pi\left(M_i - M_j\right) \right\|_1 \tag{3}$$

其中，E 是上眼睑和下眼睑特征点对的集合。

光度一致性损失表示输入图像和输出的渲染图像之间的误差，定义如下：

$$L_{pho} = \left\| V_I \odot \left(I - I_r\right) \right\|_{1,1} \tag{4}$$

其中，V_I 为值为 1 的人脸遮罩，$\odot$ 为逐元素乘积。

身份损失函数表示将人脸编码成低维嵌入向量之间的误差，定义如下：

$$L_{id} = 1 - \frac{f\left(I\right) f\left(I_r\right)}{\left\| f\left(I\right) \right\|_2 \cdot \left\| f\left(I_r\right) \right\|_2} \tag{5}$$

形状一致性损失表示给定同一个人物的两个不同的图像 I_i 和 I_j，编码器 Encoder 应该输出相同的形状参数即 $\beta_i = \beta_j$。利用这一特点，用 β_j 替换 β_i，同时保持所有其他参数不变，这组新参数可以很好地重建图像 I_i，定义如下：

$$L_{id} = L(I_i, R\left(M\left(\beta_j, \theta_i, \psi_i\right), B\left(\alpha_i, l_i, N_{u,v,i}\right), c_j\right) \tag{6}$$

L_{reg} 表示对形状、光照和反照率的正则化损失，定义如下：

$$E_\beta = \|\beta\|_2^2, E_\psi = \|\psi\|_2^2, E_\alpha = \|\alpha\|_2^2 \tag{7}$$

其中，E_β 表示形状的正则化损失，E_ψ 表示光照的正则化损失，E_α 表示反照率的正则化损失。

2.5. 评价指标

实验采用 NoW 基准测试[15]，该基准测试为从单目人脸图片中进行三维重建的项目提供评估数据集，包括 100 个受试者的 2054 张人脸图像以及每个受试者的一个 3D 人脸扫描网格。评价指标为已重建 3D 网格进行刚性对齐(旋转，平移，缩放)到扫描 3D 网格中，计算预测和扫描网格中一系列相对应的特征点数值之间的误差 E_{dis}，定义如下：

$$E_{dis} = \frac{1}{N}\sum_{i=1}^{N} \rho\left(P_i^{gt}, P_i^{pd}\right) \tag{8}$$

其中，N 为采样点个数，ρ 为对应点计算误差的函数，P_i^{gt} 为扫描网格采样点的真实值，P_i^{pd} 重建网格采样点的预测值。

3. 实验和结果

3.1. 定量评估

实验训练所采用的数据集包括 VGGFace2 数据集[16]、BUPT-Balancedface 数据集[17]和通过 StyleGan2 模型[18]生成的黄种人数据集，这些数据集需要经过二维人脸关键点检测算法[19]进行人脸 68 个特征点标注以及使用人脸分割模型[20]进行人脸遮罩的标注。设置四种算法模型(1) DECA/ResNet50；(2) DECA/ViT；(3) DECA/ViT + DropKey (ratio = 0.01)；(4) DECA/ViT + DropKey (ratio = 0.05)，分别计算 NoW 评估数据集的评价指标 E_{dis} 的中位数，平均数和标准差误差，对比结果如表 2 所示。

Table 2. Comparison of the results of the four algorithms in the NoW dataset
表 2. 四种算法在 NoW 数据集的结果对比

算法模型	Media (mm)	Mean (mm)	Std (mm)
DECA/ResNet50	1.29	1.62	1.39
DECA/ViT	1.29	1.63	1.38
DECA/ViT + DropKey (ratio = 0.01)	1.26	1.60	1.35
DECA/ViT + DropKey (ratio = 0.05)	1.25	1.58	1.33

表 2 数据表明，模型(2)将特征提取器替换成 ViT 模型后，可以达到和使用卷积神经网络 ResNet50 的模型(1)重建三维人脸相对等的结果；然而，可以从图 4 看出，引入 ViT 模型后，训练会出现强烈的损失曲线震荡且容易过拟合。模型(3)中的 DropKey 策略(ratio = 0.01)可以激励模型更多关注与目标有关的其他图像块，有助于捕捉全局鲁棒特征，避免了模型过度拟合低维特征。模型(4)将 Drop 概率设置为 0.05，提高了模型对高注意力值部分的惩罚程度，保证了模型有充足的高维特征进行稳定的训练，提升了模型的性能。

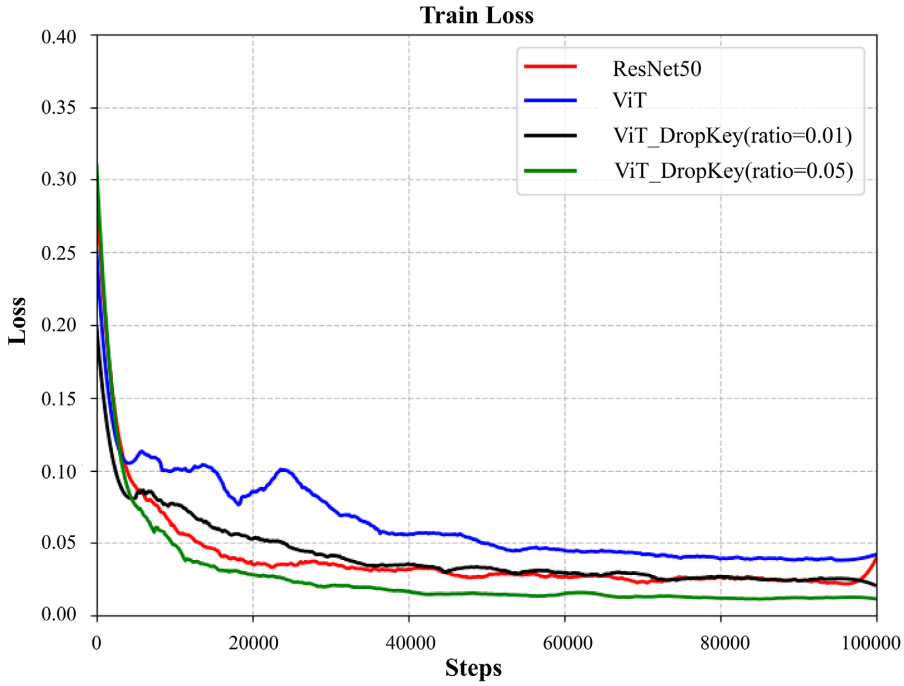
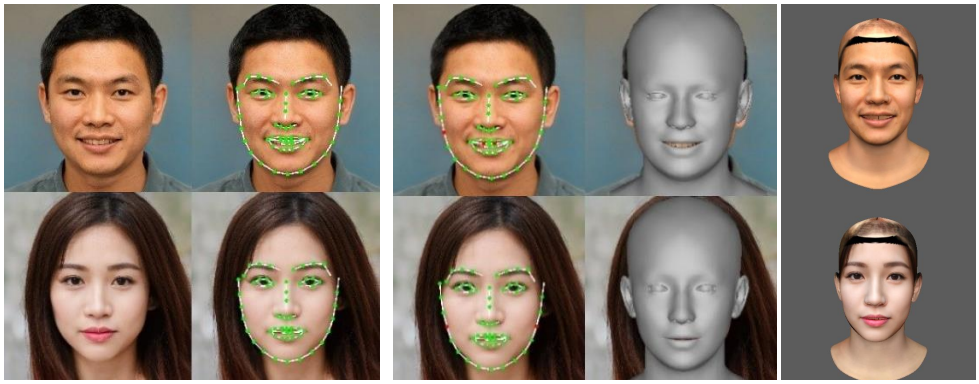


Figure 4. Training loss curve
图 4. 训练损失曲线图

3.2. 定性分析

图 5 中，(a)为 StyleGan2 数据集上的人脸图像和标签图片；(b)为(a)通过使用 ViT 模型和 DropKey 策略的 DECA 模型预测的人脸特征点和 FLAME 头部模型；(c)为使用 Albedo 贴图渲染后的三维人脸模型。可以看出，改进 DECA 模型对人脸特征点预测准确以及对人脸图片进行三维重建的效果较好。



(a) 原始和标签图像　(b) 模型重建效果　(c) 渲染效果

Figure 5. Face 3D reconstruction effect chart
图 5. 人脸重建效果图

4. 结论

本文通过引入 Vision Transformer 模型和 DropKey 策略对 DECA 模型进行改进，并将改进后的模型用于单目人脸三维重建。改进后的 DECA 模型对于二维人脸图像重建三维人脸有着良好的效果：一方面，

Vision Transformer 模型中多头注意力使得模型能够在不同的位置关注不同特征，有利于捕捉图像的全局和局部信息；另一方面，DropKey 的注意力峰值惩罚和随加深的注意力层递减 Drop 概率的策略，防止了 ViT 模型过拟合和有足够多的高维特征进行稳定训练。实验结果表明，改进后的 DECA 模型对人脸特征点预测准确，纹理逼近真实图像，模型重建精度得到提高。

基金项目

省部级社科类重点项目(项目名称：弘扬地质精神，传承优良学风；项目编号：XFCC2023ZZ028)。

参考文献

[1] 林琴, 李卫军, 董肖莉, 宁欣, 陈鹏. 基于双目视觉的人脸三维重建[J]. 智能系统学报, 2018, 13(4): 534-542.

[2] Castelán, M. and Hancock, E.R. (2004) Acquiring Height Maps of Faces from a Single Image. *Proceedings* 2*nd International Symposium on* 3*D Data Processing, Visualization and Transmission*, Thessaloniki, 6-9 September 2004, 183-190.

[3] 叶于平. 高真实感人脸表情动态三维重建及迁移方法研究[D]: [博士学位论文]. 深圳: 中国科学院大学(中国科学院深圳先进技术研究院), 2022.

[4] Blanz, V. and Vetter, T. (1999) A Morphable Model for the Synthesis of 3D Faces. *Proceedings of the* 26*th Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, 8-13 August 1999, 187-194. https://doi.org/10.1145/311535.311556

[5] Cao, C., Weng, Y., Zhou, S., *et al*. (2013) FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, **20**, 413-425. https://doi.org/10.1109/TVCG.2013.249

[6] Tuan Tran, A., Hassner, T., Masi, I., *et al*. (2017) Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5163-5172. https://doi.org/10.1109/CVPR.2017.163

[7] He, K.M., *et al*. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[8] Zhu, X.Y., *et al*. (2016) Face Alignment across Large Poses: A 3D Solution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 146-155. https://doi.org/10.1109/CVPR.2016.23

[9] Feng, Y., Wu, F., Shao, X., *et al*. (2018) Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Net-Work. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 534-551. https://doi.org/10.1007/978-3-030-01264-9_33

[10] Guo, J.Z., *et al*. (2020) Towards Fast, Accurate and Stable 3D Dense Face Alignment. *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 152-168. https://doi.org/10.1007/978-3-030-58529-7_10

[11] Li, T., Bolkart, T., Black, M.J., *et al*. (2017) Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, **36**, Article No. 194. https://doi.org/10.1145/3130800.3130813

[12] Feng, Y., *et al*. (2021) Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Transactions on Graphics*, **40**, Article No. 88. https://doi.org/10.1145/3450626.3459936

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2010) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[14] Li, B., Hu, Y., Nie, X., *et al*. (2023) DropKey for Vision Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 18-22 June 2023, 22700-22709. https://doi.org/10.1109/CVPR52729.2023.02174

[15] NoW Challenge (2019) https://now.is.tue.mpg.de/

[16] Cao, Q., Shen, L., *et al*. (2018) Vggface2: A Dataset for Recognizing Faces across Pose and Age. 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, 15-19 May 2018, 67-74. https://doi.org/10.1109/FG.2018.00020

[17] Wang, M., Deng, W., Hu, J., *et al*. (2019) Racial Faces in the Wild: Reducing racial Bias by Information Maximization Adaptation Network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 692-702. https://doi.org/10.1109/ICCV.2019.00078

[18] Karras, T., Laine, S., Aittala, M., *et al*. (2020) Analyzing and Improving the Image Quality of StyleGAN. *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 14-19 June 2020, 8110-8119. https://doi.org/10.1109/CVPR42600.2020.00813

[19] Bulat, A. and Tzimiropoulos, G. (2017) How Far Are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1021-1030. https://doi.org/10.1109/ICCV.2017.116

[20] Nirkin, Y., Masi, I., Tuan, A.T., *et al*. (2018) On Face Segmentation, Face Swapping, and Face Perception. 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, 15-19 May 2018, 98-105. https://doi.org/10.1109/FG.2018.00024