Modeling and Simulation 建模与仿真, 2025, 14(3), 426-434
Published Online March 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.143235

Hans汉斯

基于多尺度空洞可分离卷积的视觉Transformer的端到端可训练头部姿态估计

尧京京

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年2月24日；录用日期：2025年3月17日；发布日期：2025年3月26日

摘　要

在本文中，我们基于Hopenet网络和视觉Transformer提出了一种用于RGB图像头部姿势估计的新方法，并设计了一种新颖的架构，由以下三个关键组件组成：(1) 骨干网络，(2) 视觉Transformer，(3) 预测头。我们还对骨干网络进行了改进，采用多尺度空洞可分离卷积以增强特征提取能力。相比于传统卷积神经网络和视觉Transformer提取特征的方式，我们的骨干网络在降低图像分辨率的同时，能够更有效地保留关键信息。通过消融实验，我们验证了基于多尺度空洞可分离卷积的骨干网络在特征保留能力上优于传统的深度卷积网络和视觉Transformer架构。我们在300W-LP和AFLW2000数据集上进行了全面的实验与消融研究。实验结果表明，所提出的方法在头部姿势估计任务上，相较于Hopenet及部分基于Transformer编码器的方法(如HeadPosr)，在准确性和鲁棒性方面均实现了显著提升。

关键词

姿势估计，多尺度空洞可分离卷积，视觉Transformer，Transformer编码器

End-to-End Trainable Head Pose Estimation with Vision Transformer Based on Multi-Scale Dilated Separable Convolution

Jingjing Yao

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science & Technology, Shanghai

Received: Feb. 24th, 2025; accepted: Mar. 17th, 2025; published: Mar. 26th, 2025

Abstract

In this paper, we propose a novel approach for head pose estimation from RGB images, leveraging

文章引用: 尧京京. 基于多尺度空洞可分离卷积的视觉 Transformer 的端到端可训练头部姿态估计[J]. 建模与仿真, 2025, 14(3): 426-434. DOI: 10.12677/mos.2025.143235

the Hopenet network and Vision Transformer. Our method introduces an innovative architecture comprising three key components: (1) a backbone network, (2) a Vision Transformer, and (3) a prediction head. To enhance feature extraction capabilities, we further improve the backbone network by incorporating multi-scale dilated separable convolutions. Compared to traditional convolutional neural networks and Vision Transformers for feature extraction, our backbone network effectively preserves critical information while reducing image resolution. Through ablation studies, we validate that the proposed backbone network, equipped with multi-scale dilated separable convolutions, outperforms conventional deep convolutional networks and Vision Transformer-based architectures in terms of feature retention. We conduct extensive experiments and ablation studies on the 300W-LP and AFLW2000 datasets. Experimental results demonstrate that our approach significantly improves both accuracy and robustness in head pose estimation, outperforming Hopenet and certain Transformer-based encoder methods, such as HeadPose.

Keywords

Head Pose Estimation, Multi-Scale Dilated Separable Convolutions, Vision Transformer, Transformer Encoder

1. 引言

面部表情检测[1]是近年来广泛研究的领域，其中头部姿势估计作为面部表情检测中的一个重要子任务，受到了极大的关注。头部姿势估计在多个领域中具有重要应用，如人机交互[2]、自动驾驶[3]、表情驱动与注意力检测[4]等。当前，卷积神经网络已成为头部姿势估计研究中的主流方法。最近，[5]提出了一种基于多损失训练的无关键点头部姿态估计方法，该方法通过结合分类和回归来进行角度预测。尽管有些方法[6] [7]利用面部关键点来提升头部姿势估计的精度，但在关键点检测困难的情况下，依然面临显著挑战，且对关键点检测的精度要求较高。此外，也有方法[8]采用 RGB-D 图像进行估计，虽然 RGB-D 图像能够提供高精度的头部姿态估计，但其在数据存储需求高、功耗大和受环境影响等方面存在一定局限性。

受到 Transformer 网络[9]的启发，Transformer 最初应用于自然语言处理领域，凭借其自注意力机制和捕获远距离特征的能力，逐渐被扩展应用于计算机视觉领域[10]。因此，方法[11]基于 Transformer 网络提出了头部姿势估计的架构。与基于卷积神经网络的方法相比，基于视觉 Transformer [10]的头部姿势估计在效果上表现优越，但在骨干网络的优化上仍有较大的提升空间。目前，骨干网络多采用卷积神经网络和视觉 Transformer 进行特征提取。由于 Transformer 的注意力机制及其在捕获远距离特征方面的优势，基于视觉 Transformer 的骨干网络在特征提取能力上往往优于传统的卷积神经网络。然而，基于 Transformer 的网络通常采用补丁下采样方法，如补丁合并。在使用切片合并时，合并操作是基于相邻的补丁进行的，而不是基于全局图像结构进行的。这意味着，尽管每个补丁内部的信息能够得到保留，但在合并过程中，补丁之间的空间关系可能会被忽略，从而导致模型无法获得足够的空间结构信息。例如，合并后的补丁表示可能无法有效地区分不同区域之间的细节差异(如物体的边缘和纹理)。

近年来，深度可分离卷积因其能够在减少卷积计算成本的同时保持与标准卷积相近的性能而广泛受到关注。[12]提出了一种深度可分离卷积注意力模块，旨在通过关注重要信息并捕捉通道和空间位置之间

的关系来提升模型性能。基于此，本文提出了一种多尺度空洞可分离卷积，用于骨干网络中，以增强特征信息的保留。结合 Hopenet 网络和视觉 Transformer 的设计理念，本文提出了一种基于 Transformer 编码器的头部姿势估计模型架构，旨在进一步提升头部姿势估计的精度和效率。

本文的主要贡献如下：

(1) 提出了一种从 RGB 图像中预测头部姿势的方法。本文架构由三个主要模块组成：① 骨干网络：用于提取空间图像特征；② 视觉 Transformer：用于进一步处理和捕捉图像中的长程依赖关系；③ 预测头：设计了一个全连接层，对 Transformer 编码器的输出进行处理，得到偏航角、俯仰角和滚转角预测结果。

(2) 我们提出了一种多尺度空洞可分离卷积方法，应用于骨干网络中，旨在有效捕捉下采样输出中的空间和通道特征。该方法通过从多个尺度提取特征，能够全面捕捉局部和全局信息，精准地提取图像中的关键特征，从而提升特征表达能力。

(3) 我们进行了广泛的消融实验，以验证所提出的多尺度空洞可分离卷积方法在头部姿势估计任务中的优势和有效性。

(4) 在使用 300W-LP 数据集进行训练，并在 AFLW2000 数据集上进行测试时，基于单张图像进行头部姿势估计的性能超过了 Hopenet 和 HeadPosr 方法，显示了我们方法的优越性。

2. 国内外研究现状

头部姿态估计(Head Pose Estimation, HPE)是计算机视觉领域的核心研究方向，旨在从图像或视频数据中推断人类头部在三维空间中的朝向。随着研究的不断深入，国内外学者提出了多种方法，以提升头部姿态估计的准确性和鲁棒性。

近期，[5]提出了一种基于 ResNet 的无关键点端到端头部姿态估计方法，该方法摒弃了传统基于关键点检测的流程，直接通过深度残差网络进行姿态回归，提高了模型的稳定性和计算效率。此外，Cao 等人[13]提出了一种基于向量表示的姿态估计方法，该方法采用旋转矩阵的三个向量作为姿态参数，并利用深度神经网络进行回归预测，有效克服了欧拉角表示中存在的不连续性问题，提高了大角度姿态估计的准确性和稳定性。

与此同时，国内学者也对头部姿态估计进行了深入研究。例如，[14]提出了一种多尺度卷积神经网络(MSCNN)框架，该模型利用不同尺度的卷积核对输入的头部图像进行特征提取，并引入 1 × 1 卷积以减少计算开销，同时保留关键信息。

3. 整体设计

3.1. 问题表述

基于 RGB 图像的头部姿势估计问题可以表述为：给定一组 RGB 图像 $I=\{i_1,i_2,\cdots,i_n\}$，这些图像对应一组姿势地面真值 $V=\{v_1,v_2,\cdots,v_n\}$，$v_i=[y,p,r]$，y,p,r 分别代表偏航角、俯仰角和滚转角。我们就是要让模型找到一个关系 R，$\overline{v}=R(i)$，使得 $\overline{v}$ 无限接近 v。关系 R 是使用姿势地面真值和预测姿势之间的平均绝对误差(MAE)的优化来计算的：

$$L=\sum_{i=1}^{n}\frac{\left|\overline{v}_i-v_i\right|}{n} \tag{1}$$

3.2. 模型架构

本文架构由 Hopenet 的初始层和多尺度空洞可分离卷积构成的骨干网络、视觉 Transformer 和预测头

组成。网络的整体架构如图 1 所示，具体的子结构描述如下：

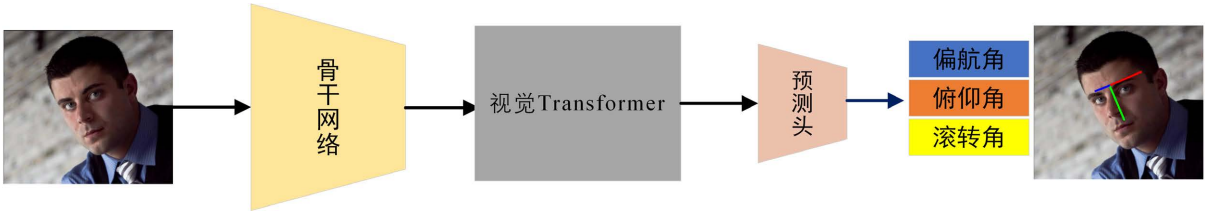

Figure 1. The model architecture proposed in this paper
图 1. 本文模型架构

1) 骨干网络：架构中的主干部分由 ResNet50 的前置初始层去除最大池化操作后，结合本文提出的多尺度空洞可分离卷积构成。

给定一组图像，$I \in R^{B\times C\times H\times W}$，B 代表批量大小，C 代表图像中的通道，H 代表高度，W 代表图像的宽度。经过 ResNet50 的前置初始层处理后，得到特征图 $F \in R^{B\times 64\times 112\times 112}$，随后对每个通道的特征进行切片操作，得到特征 $F' \in R^{B\times d\times P\times P}$，其中 $d = 64\times 112/P\times 112/P$，$P\times P$ 为切片后的每个补丁的尺寸大小，切片操作后的特征图被输入到本文提出的多尺度空洞可分离卷积中。

多尺度空洞可分离卷积分为两个主要部分：多尺度空洞卷积和逐点卷积。首先，输入特征图经过多尺度空洞卷积处理，对每个通道分别应用不同的空洞卷积。多尺度空洞卷积使用多个不同的空洞率(例如{5, 3, 1})，用于从不同感受野范围内提取高层次语义特征，并将这些不同尺度的信息融合。此操作能够有效提取全局和局部特征，捕捉空间信息，同时避免不同通道之间信息的混合。多尺度空洞卷积的网络结构如图 2 所示。在第二步操作中，使用逐点卷积(1×1 卷积核)对深度卷积输出的各通道特征进行组合，从而捕捉通道之间的相关性。

在经过多尺度空洞卷积和逐点卷积后，特征图都将通过批归一化层和 GELU 激活函数进行处理，确保网络训练的稳定性和非线性表达。通过多尺度空洞卷积处理后的特征图大小保持不变，而经过逐点卷积处理后，特征图维度调整为所需的输出维度 $F' \in R^{B\times C'\times P\times P}$，并通过形状重塑操作进行适配，最终为后续的 Transformer 编码器提供输入 $F' \in R^{B\times A\times C'}$，$A = P\times P$。

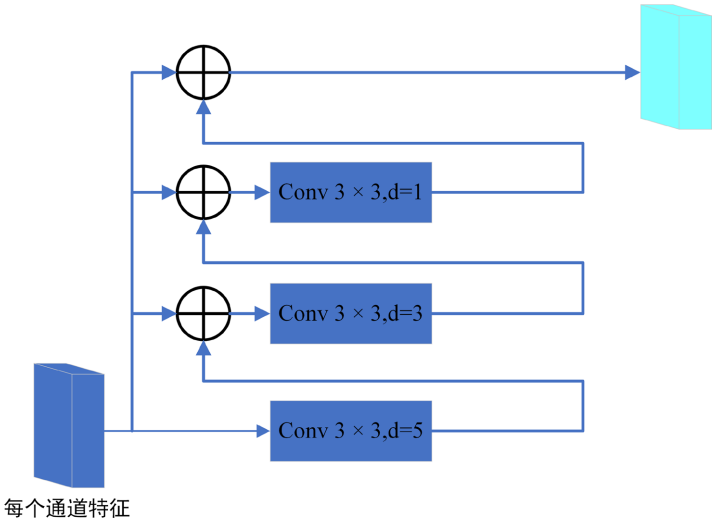

Figure 2. Multi-scale dilated convolutional network architecture
图 2. 多尺度空洞卷积网络结构

2) 视觉 Transformer：视觉 Transformer 原理如图 3 所示，在此过程中，我们将一个可学习的嵌入向量沿着维度为1拼接到骨干网络输出特征的前端，作为分类token，其形状为 $R^{B\times1\times C'}$，该token在Transformer编码器输出时的状态，作为图像的最终表示。为了保留位置信息，位置嵌入被加到特征中。我们采用标准的可学习 1 维位置嵌入，形状为 $R^{B\times(A+1)\times C'}$。最终得到的嵌入向量序列作为输入传递给 Transformer 编码器进行处理。

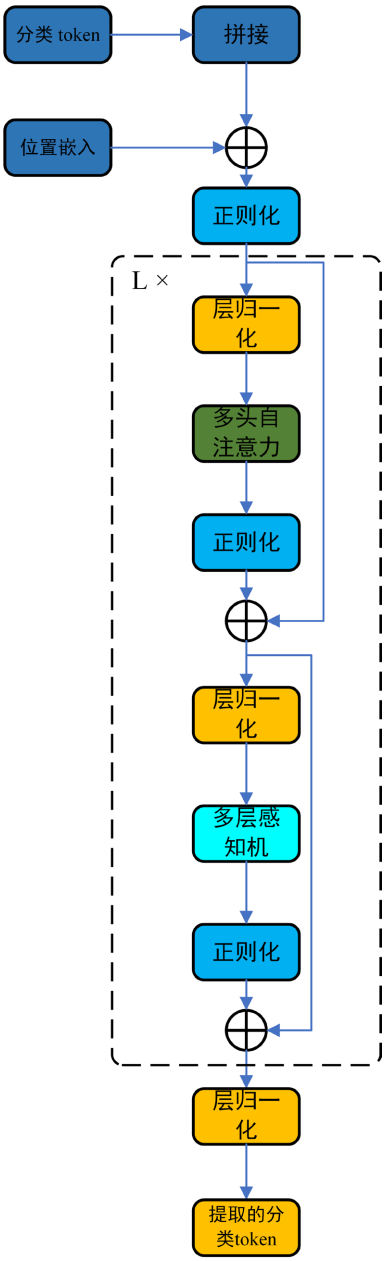

Figure 3. Vision Transformer
图 3. 视觉 Transformer

Transformer 编码器是通过堆叠 L 个相同的单元块来构建，每个单元块由多头自注意力(MSA)模块和多层感知器(MLP)模块组成。Transformer 编码器自注意力的计算过程如下：

$$Q = \hat{F}W_Q \tag{2}$$

$$K = \hat{F}W_K \tag{3}$$

$$V = \hat{F}W_V \tag{4}$$

$$SA\left(\hat{F}\right) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_t}}\right)V \tag{5}$$

$$d_t = C'/h \tag{6}$$

$\hat{F} \in R^{B\times(A+1)\times C'}$ 是每个 Transformer 层的输出，W_Q、W_K 和 $W_V \in R^{C'\times C'}$ 分别代表查询矩阵、关键矩阵和值矩阵。h 为多头自注意模块的注意力头数。

Transformer 模块表示如下：

$$F^{t-1} = \text{MSA}\left(\text{LN}\left(\hat{F}^{t-1}\right)\right) + \hat{F}^{t-1} \tag{7}$$

$$\hat{F}^{t} = \text{MLP}\left(\text{LN}\left(F^{t-1}\right)\right) + F^{t-1} \tag{8}$$

3) 预测头：这里我们使用一个全连接层，经过视觉 Transformer 后分类 token $\in R^{B\times 1\times C'}$，分类 token 经过预测头后变为 $\text{R}^{\text{B}\times 1\times 3}$，即偏航角、俯仰角和滚转角。

4. 实验设计与验证

4.1. 数据集

实验在两个公开开源数据集上进行验证，分别是 300W-LP [15]和 AFLW2000 [16]数据集。

1) 300W-LP 数据集：作为 300W 数据集的扩展版本，300W-LP 包含了 XM2VTS、IBUG、LFPW、AFW 和 HELEN 等多个子数据集，共计 61,225 张图像样本。

2) AFLW2000 数据集：该数据集包含 2000 张图像，涵盖了丰富的头部姿势变化。图像中的人脸呈现多种照明、表情及姿势变化。

在本文中，采用 300W-LP 数据集进行训练，并使用 AFLW2000 数据集进行测试。实验结果报告了偏航角、俯仰角和滚转角的误差值，以及这三个角度的平均误差。

4.2. 实施细节

本文使用 Pytorch 库实现模型，并在训练过程中应用了随机裁剪和缩放数据增强技术。网络训练采用 Adam 优化器，进行了 120 个迭代周期，初始学习率设为 0.01。此外，学习率在每 40 个周期后衰减为原值的 0.1 倍。训练过程中，使用 300W-LP 数据集，批量大小设为 32。模型中的超参数设置为：P = 14，d = 64，L = 12，h = 8。所有实验均在 NVIDIA GeForce GTX 3090 GPU 上执行。

4.3. 实验结果与分析

表 1 展示了本文方法与 Hopenet [5]、HeadPosr [11]的定量比较。Hopenet 是一种基于卷积神经网络的头部姿势估计方法，通过多任务学习结合分类和回归技术，进行偏航角、俯仰角和滚转角的估计。HeadPosr 则基于 Transformer 编码器进行头部姿势预测，其骨干网络采用 ResNet 的初始层。通过这些对比，我们验证了所提方法在准确性和鲁棒性方面的优势。

表 1 显示，在 300W-LP 数据集上训练的情况下，HeadPosr 和我们的模型在头部姿势估计任务中的表现均优于 Hopenet，这表明结合 ResNet 架构与 Transformer 编码器的模型相较于仅使用 ResNet 架构的模

型具有更好的性能，进一步验证了 Transformer 编码器对头部姿势估计任务的积极作用。与 Hopenet 相比，我们的方法在偏航角、俯仰角和滚转角的表现分别提高了 31.95%、23.22%和 34.48%。在平均值方面，我们的方法相较于 Hopenet 提高了 29.99%。与 HeadPosr 相比，我们的方法在偏航角、俯仰角和滚转角的表现分别提升了 4.53%、3.77%和 2.79%。在平均值方面，我们的方法相对于 HeadPosr 提高了 3.66%。我们的方法优于 HeadPosr，可能得益于我们提出的多尺度空洞可分离卷积在特征提取方面的优势。在下一小节中，我们将通过消融实验进一步验证多尺度空洞可分离卷积对特征提取的提升作用。

Table 1. Comparison of our proposed method with Hopenet and HeadPosr on the AFLW2000 dataset, trained on the 300W-LP dataset (EH64 denotes the use of 6 encoders and 4 attention heads in the Transformer architecture)
表 1. 本文方法与 Hopenet、HeadPosr 在 AFLW2000 数据集上的比较，都在 300W-LP 数据集上训练(EH64 表示在 Transformer 中使用 6 个编码器和 4 个注意力头)

方法	滚转角误差	偏航角误差	俯仰角误差	平均误差
Hopenet [5]	6.38	6.51	7.32	6.77
HeadPosr EH64 [11]	4.30	4.64	5.84	4.92
Ours	4.18	4.43	5.62	4.74

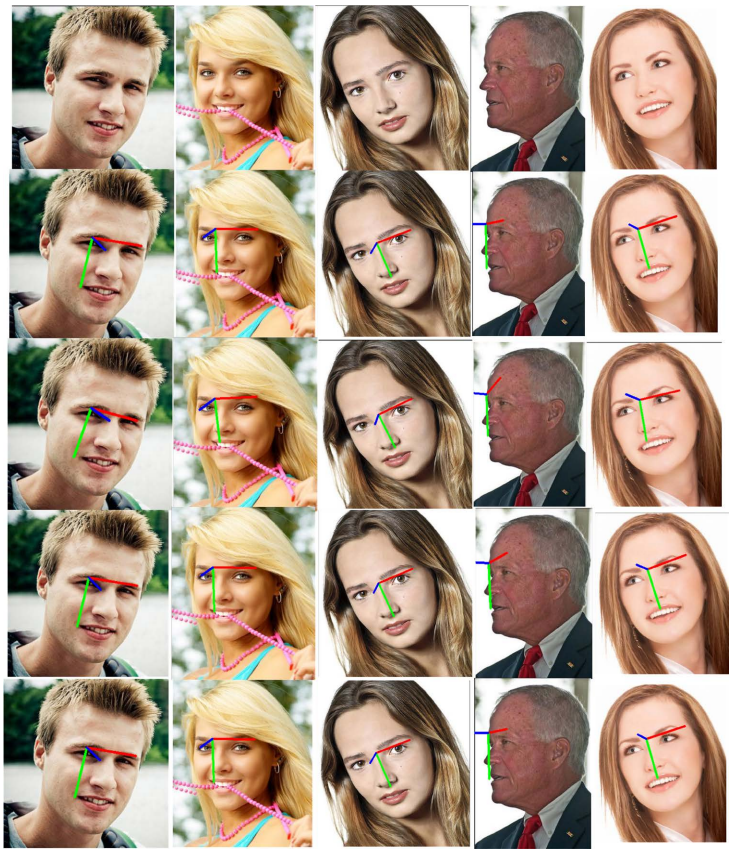

Figure 4. Comparison of qualitative results of our method with Hopenet and HeadPosr. The first row presents image samples from the test set, while the second row shows the corresponding ground-truth head pose values. The third row displays the output results of Hopenet, the fourth row presents the output results of HeadPosr, and the fifth row shows the output results of our proposed method
图 4. 本文方法与 Hopenet 和 HeadPosr 的定性结果的比较。第一行展示了来自测试集的图像样本；第二行为相应测试图像的姿势地面真实值；第三行展示了 Hopenet 的输出结果；第四行展示了 HeadPosr 的输出结果；第五行展示了本文方法的输出结果

图 4 展示了本文方法与 Hopenet 和 HeadPosr 方法测试的一些图像样本。可以看出，本文方法表现优于 Hopenet 和 HeadPosr 方法，我们方法结果与测试图像的姿势地面真实值最相近。

4.4. 消融实验

本节评估了三种不同骨干网络架构对头部姿势估计任务的影响：传统的基于卷积神经网络的骨干网络、基于 Vision Transformer 的骨干网络以及基于多尺度空洞可分离卷积的骨干网络。对于传统的基于卷积神经网络的骨干网络，我们使用卷积核大小为 8、填充为 0、步长为 8 的标准卷积层替代切片操作和多尺度空洞可分离卷积，该卷积层的输入维度为 64，输出维度为 64，目的是确保提取的特征能够符合输入到 Vision Transformer 编码器的要求。对于基于 Vision Transformer 的骨干网络，我们直接使用 Vision Transformer 作为骨干网络，骨干网络的 Vision Transformer 的编码器层数设为 3 层，其他设置保持不变。所有三种架构均在 300W-LP 数据集上进行训练，并在 AFLW2000 数据集上进行测试。实验结果如表 2 所示。

Table 2. Comparison of three different backbone networks (compared on the AFLW2000 dataset, trained on the 300W-LP dataset.)
表 2. 三种不同骨干网络结果的比较(在 AFLW2000 数据集上的比较，在 300W-LP 数据集上训练)

方法	滚转角误差	偏航角误差	俯仰角误差	平均误差
基于(8 × 8)卷积层的骨干网络	4.43	4.79	5.74	4.98
基于 ViT 的骨干网络	4.32	4.75	5.58	4.88
基于多尺度空洞可分离卷积的骨干网络(本文)	4.18	4.43	5.62	4.74

从表 2 可知，基于 Vision Transformer 的骨干网络模型在性能上优于传统的基于卷积神经网络的骨干网络模型。与此同时，我们的模型在这两者之上，表现出更优的结果。这表明，所提出的多尺度空洞可分离卷积在骨干网络中的应用，有效提升了特征提取能力，并在头部姿势估计任务中发挥了重要作用。

5. 结束语

本文基于 Hopenet 网络和视觉 Transformer 设计了一种新颖的架构，用于头部姿势估计。我们提出了一种多尺度空洞可分离卷积，有效地提取全局和局部特征，捕捉来自不同感受野范围内的空间信息，同时避免了通道间的信息混合。通过逐点卷积进一步建模通道间的相关性，显著增强了骨干网络的特征提取能力。最终，在两个公共数据集上验证了所提模型的性能以及多尺度空洞可分离卷积对骨干网络的积极作用。实验结果表明，本文提出的模型在性能上优于其他最新方法，且相比于传统基于 ResNet 初始层的深度卷积网络或视觉 Transformer，应用多尺度空洞可分离卷积显著提升了特征提取的效果和头部姿势估计的精度。

本文设计的模型仍有不足之处，虽然多尺度空洞可分离卷积有效地捕捉和融合不同感受野的空间信息，但由于空洞卷积涉及较大的步幅和不同尺度的处理，可能导致计算开销增大，训练缓慢，在未来的工作中我们尝试采用更高效的空洞卷积方法或将卷积操作进行并行化来减少计算开销。

参考文献

[1] 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). 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1021-1030. https://doi.org/10.1109/iccv.2017.116

[2] Bisogni, C., Nappi, M., Pero, C. and Ricciardi, S. (2021) FASHE: A Fractal Based Strategy for Head Pose Estimation. *IEEE Transactions on Image Processing*, **30**, 3192-3203. https://doi.org/10.1109/tip.2021.3059409

[3] Murphy-Chutorian, E., Doshi, A. and Trivedi, M.M. (2007) Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation. 2007 *IEEE Intelligent Transportation Systems Conference*, Bellevue, 30 September-3 October 2007, 709-714. https://doi.org/10.1109/itsc.2007.4357803

[4] Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A. and Rehg, J.M. (2018) Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018., Springer, 397-412. https://doi.org/10.1007/978-3-030-01228-1_24

[5] Ruiz, N., Chong, E. and Rehg, J.M. (2018) Fine-grained Head Pose Estimation without Keypoints. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 2074-2083. https://doi.org/10.1109/cvprw.2018.00281

[6] Kazemi, V. and Sullivan, J. (2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1867-1874. https://doi.org/10.1109/cvpr.2014.241

[7] Kumar, A., Alavi, A. and Chellappa, R. (2017) KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors. 2017 12*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2017), Washington, 30 May-3 June 2017, 258-265. https://doi.org/10.1109/fg.2017.149

[8] Fanelli, G., Weise, T., Gall, J. and Van Gool, L. (2011) Real Time Head Pose Estimation from Consumer Depth Cameras. In: Mester, R. and Felsberg, M., Eds., *Pattern Recognition*, Springer, 101-110. https://doi.org/10.1007/978-3-642-23123-0_11

[9] Meyer, G.P., Gupta, S., Frosio, I., Reddy, D. and Kautz, J. (2015) Robust Model-Based 3D Head Pose Estimation. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 3649-3657. https://doi.org/10.1109/iccv.2015.416

[10] Vaswani, A. (2017) Attention Is All You Need. arXiv: 1706.03762.

[11] Dosovitskiy, A. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[12] Dhingra, N. (2021) HeadPosr: End-To-End Trainable Head Pose Estimation Using Transformer Encoders. 2021 16*th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG* 2021), Jodhpur, 15-18 December 2021, 1-8. https://doi.org/10.1109/fg52635.2021.9667080

[13] Liu, F., Xu, H., Qi, M., Liu, D., Wang, J. and Kong, J. (2022) Depth-Wise Separable Convolution Attention Module for Garbage Image Classification. *Sustainability*, **14**, Article 3099. https://doi.org/10.3390/su14053099

[14] Cao, Z., Chu, Z., Liu, D. and Chen, Y. (2021) A Vector-Based Representation to Enhance Head Pose Estimation. 2021 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2021, 1187-1196. https://doi.org/10.1109/wacv48630.2021.00123

[15] 梁令羽, 张天天, 何为. 多尺度卷积神经网络的头部姿态估计[J]. 激光与光电子学进展, 2019, 56(13): 79-86.

[16] Zhu, X., Lei, Z., Liu, X., Shi, H. and Li, S.Z. (2016) Face Alignment across Large Poses: A 3D Solution. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 146-155. https://doi.org/10.1109/cvpr.2016.23