文章编号:1671-4598(2024)05-0282-08 DOI:10.16526/j.cnki.11-4762/tp.2024.05.040 中图分类号:TP391.4 文献标识码:A

基于合成数据的刚体姿态实时估计网络

刘千山,林雪剑,朱 枫,李佩东
(国能宝日希勒能源有限公司,内蒙古 呼伦贝尔 021500)

摘要:现有刚体姿态估计存在数据稀缺、复杂场景下的低鲁棒性及低实时性等问题,为此提出一种基于合成数据的刚体目标位姿追踪网络结构;采用时空间特征融合技术,捕捉时间与空间特征信息,生成具有时空敏感的特征图;利用残差连接学习更为丰富和抽象的优质特征,改善追踪目标的准确性;对稀缺数据进行数据增强,生成符合现实物理特性的复杂合成数据,以此训练深度学习模型,提高模型的泛化性;在YCB-Video数据集中选取7个物体进行实时姿态追踪实验,结果表明,提出的方法相较于同类相关方法,在复杂场景下对刚体姿态估计的更为准确,在实时估计效率上表现最优。

关键词:姿态追踪;数据稀缺;时空特征融合;残差连接;数据增强

Real Time Estimation Network for Rigid Body Posture Based on Synthetic Data

LIU Qianshan, LIN Xuejian, ZHU Feng, LI Peidong
(Guoneng Baorixile Energy Co., Ltd., Inner Mongolia, Hulunbuir 021500, China)

Abstract: There are the characteristics of scarce data, low robustness in complex scenes, and poor real-time for existing rigid body pose estimation, for this reason, a rigid object pose tracking network based on synthetic data is proposed. Temporal and spatial feature fusion techniques are used to capture temporal and spatial feature information, generating spatiotemporal sensitive feature maps. Residual connectivity is utilized to learn more diverse and abstract high-quality features, improving the accuracy of tracking the target. Data augmentation is performed on scarce data to generate complex synthetic data that conforms to the real physical characteristics, which is used to train the deep learning model and improve the generalization of the model. Seven objects are selected on the YCB-Video dataset for real-time pose tracking experiments, the results show that compared with similar related methods, the proposed method is more accurate in estimating the poses of rigid bodies in complex scenarios, and it has an optimal performance in real-time estimation efficiency.

Keywords: pose tracking; data scarcity; spatiotemporal feature fusion; residual connectivity; data augmentation

0 引言

刚体姿态估计是计算机视觉领域的重要分支,其难点是解算出目标刚体和相机之间的空间多自由度姿态变换,从而实现由平面维度到立体维度的结构转化。在智能驾驶领域,由激光雷达、超声波雷达和高清摄像头组成的基于视觉算法的智能领航技术能够让汽车实现自主导航,并及时规避潜在的道路风险,保护乘客安全。在工业控制领域中,姿态估计常用于工业机械臂精准抓取、分拣、装配和放置操作,从而提高生产效率和产品质量,同时降低生产成本和风险。因此,一个性能优秀、通用性强的姿态估计算法对于实现各类视觉任务至关重要。

姿态估计方法依据特征提取方式的不同可分为传统手工设计阶段和深度学习阶段。在传统手工设计阶段中按表征空间划分为基于特征点的姿态估计方法[1-6]、基于模板的姿态估计方法[7-10]和基于3D坐标的姿态估计方法[11-12]。传统手工设计阶段的方法缺点在于面对复杂场景环境时鲁棒性较差,当场景或目标物体发生改变时需要手动更新网络中的大量超参数,人力成本和计算成本较高。

随着深度学习和物体CAD模型的出现,计算机视觉领域迎来加速发展,AlexNet深度卷积神经网络[13]的提出标志着目标检测步入深度学习阶段。卷积神经网络(CNN)以其多层抽象表示图像特征的能力,在计算机视觉任务中相较于传统手工设计阶段的姿态估计方法,表现出更好的姿态估计效果,逐渐成为深度学习阶段的主流目标检测方法。大量优秀网络的出现标志着目标姿态估计同样步入深度学习阶段,如Faster RCNN[14]、YOLO[15]、SSD[16]等,姿态估计的精度和效能不断提升。深度传感器的引入为计算机视觉领域提供了新的可能,深度数据的出

收稿日期:2023-03-15; **修回日期**:2023-04-01。
基金项目:国家自然科学基金(61601213)。
作者简介:刘千山(1986-),男,大学本科。
引用格式:刘千山,林雪剑,朱 枫,等.基于合成数据的刚体姿态实时估计网络[J].计算机测量与控制,2024,32(5):282-289.

现极大改善了目标姿态估计的准确性，如 PWP3D[17]、DenseFusion[18]、RGF[19]等。追踪视频序列中连续变化的目标姿态相较于估计单张快照中的物体姿态，更具挑战性和复杂性，难点包括网络实时追踪效率、追踪目标平滑连贯的结果、面对遮挡场景和多目标场景的鲁棒性等。近年来，互联网上出现的大量 3D 模型加快了深度学习在追踪目标姿态领域的发展，显著提高了追踪视频序列中物体姿态的准确性和鲁棒性[20-22]。一个性能优异的姿态追踪网络训练需要大量手工标记的真实数据驱动，高昂的数据标注成本和复杂的位置信息成为制约网络发展的条件，降低了这类方法的实用性。

为实现在复杂的挑战性场景中实时、精确估计目标刚体的姿态，本文提出一种基于卷积神经网络的深度神经网络，即使在少量合成数据驱动下网络仍然能做到高精准度和高追踪效率。本文提出的时空间特征融合技术能够提取并学习目标物体在连续帧中的相似特征，充分捕捉特征的时间信息和空间信息，生成具有时空感知性的特征图。残差连接提取特征图中更丰富和抽象的优质边缘特征，加强网络对于遮挡挑战下的鲁棒性。在数据稀缺的情况下，图像增强通道在模拟目标的真实碰撞和重力属性[23-24]的基础上，实现域随机化，随机生成具有不同纹理、光照和干扰物体的增强图像用于训练，降低数据收集成本同时弥补合成数据与真实数据的差距。通过在数据集 YCB-Video 上进行实验，并与近年来的相关方法进行定性、定量对比表明，本文方法在收敛速度、精度及追踪性能上均获得了最优结果，能够满足实际应用的需求。

1　系统结构及原理

当目标姿态估计网络应用于实际场景中，通常会遇到图像质量和目标特性带来的干扰导致目标姿态估计结果出现错误，成为姿态估计任务中的挑战性问题。其中，图像质量干扰包括背景环境噪声、目标物体遮挡、场景光照干扰和图像模糊噪声，目标特性包括目标物体纹理复杂或缺失、类间可分性低和类间多样性大。本文提出方法首先使用数据增强通道对输入图像进行处理，提高网络对图像质量和目标特性干扰的鲁棒性，并设计时空间特征融合技术和残差连接提高网络追踪性能。算法系统流程如图 1 所示。

2　合成数据增强通道

数据增强是计算机视觉中数据预处理阶段的重要部分，通过引入数据多样性、减少过拟合风险和提高模型性能，为稀缺数据下的模型训练提供坚实基础。真实数据从真实世界中通过传感器、摄像头和其他数据采集设备获得，具有反映真实场景多样性和复杂性的特点，包括真实场景中的噪声、干扰和变化，可以更好地训练模型以应对真实使用场景，具有合成数据无法比拟的优势，但获取大规模真实数据成本高昂且存在数据不平衡问题。合成数据通过模拟或生成算法收集，可以根据需求生成大量样本，数据收集成本低，但网络在训练时可能过分适应合成数据的特点，在真实场景中表现不佳。本算法在模拟目标真实碰撞和重力属性[23-24]的基础上，引入深度信息生成合成数据，其中目标姿态从预置姿态中采样，干扰物体、纹理、光照、角度随机采样，弥补合成数据相较真实数据的差距，使网络具有泛化到不同使用场景的能力，减轻数据稀缺带来的影响。合成数据的深度信息相较于真实数据有一定偏差，我们对生成的合成数据深度信息进行双边滤波处理，平滑轮廓噪声并弥补与真实数据的不足。合成数据增强通道包含数据生成通道和数据增强通道，数据生成通道用于生成合成数据，其目的是通过模拟各种场景和变化来增加数据多样性。

在数据生成后，数据进入数据增强通道，这一阶段目的是进一步丰富和多样化数据，以更好地训练深度学习模型。数据增强通道包括以下操作。

图像缩放和旋转：对图像进行缩放和旋转操作，引入不同角度和视角的变化，帮助模型提取不同距离和观察角度下的目标特征；

HSV 变换：通过调整图像的色相、饱和度和明度，增强网络模型对于不同环境的鲁棒性；

目标物体或场景增加高斯噪声和模糊：模拟实际场景中因相机抖动或失焦出现的噪声或模糊问题；

去除目标物体部分深度信息：在实际使用场景中模型可能需要处理模糊的深度信息，帮助模型适应实际使用场

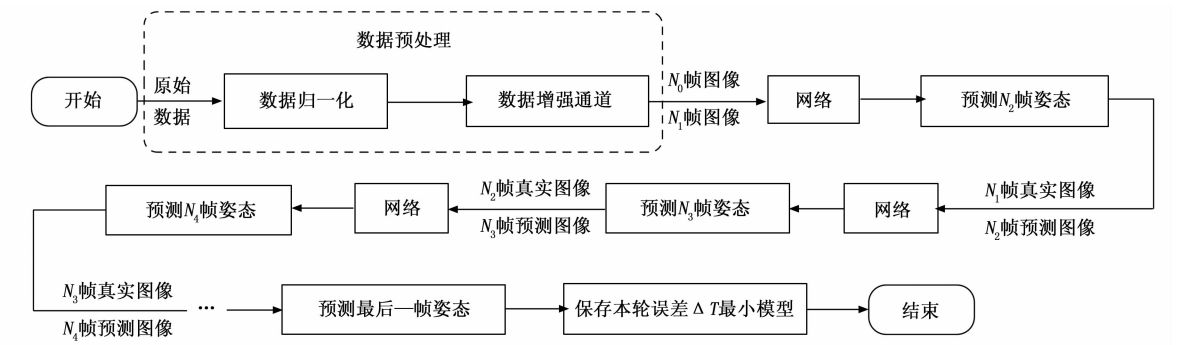

图 1　算法流程图

景中的深度变化；

遮挡图像部分信息操作：模拟图像中的遮挡可以帮助模型理解部分可见目标，增强模型面对遮挡场景时的鲁棒性。

数据生成通道可视化结果如图 2 所示，数据增强通道可视化结果如图 3 所示。本文所采用的姿态追踪方法将在下一章具体阐述。

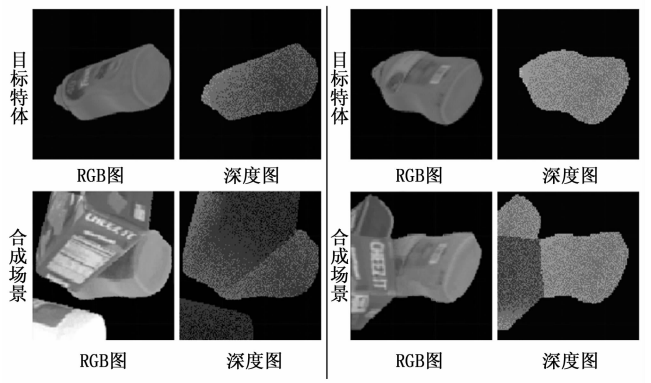

图 2 数据生成通道可视化结果

3 姿态追踪网络

3.1 李群和李代数

在同时定位与地图构建（SLAM，simultaneous localization and mapping）中，常使用李群中的特殊欧式群 SE（3）表示物体在三维空间中的运动，表示为：

$$SE(3)=\left\{\boldsymbol{T}=\begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\\ 0^T & 1\end{bmatrix}\in\mathbb{R}^{4\times4}\,|\,\boldsymbol{R}\in SO(3),\boldsymbol{t}\in\mathbb{R}^3\right\} \tag{1}$$

式中，$\boldsymbol{t}$ 为平移向量，$\boldsymbol{T}$ 为变换矩阵，$\boldsymbol{R}$ 为旋转矩阵，其作用是共同表示一个坐标系到另一个坐标系转移的过程，变换矩阵 $\boldsymbol{T}$ 为当前相机位置进行的移动和旋转。特殊正交群 SO（3）代表旋转矩阵的合集，表示为：

$$SO(3)=\{R\in\mathbb{R}^{3\times3}\,|\,\boldsymbol{R}\boldsymbol{R}^T=I,\det(\boldsymbol{R})=1\} \tag{2}$$

对于任意旋转矩阵 $\boldsymbol{R}$，满足：

$$\boldsymbol{R}\boldsymbol{R}^T=I \tag{3}$$

对公式（3）两端时间 t 求导，得：

$$\boldsymbol{R}'(t)\boldsymbol{R}^T(t)+\boldsymbol{R}(t)\boldsymbol{R}'^T(t)=0 \tag{4}$$

化简公式（4）得到：

$$\boldsymbol{R}'(t)\boldsymbol{R}^T(t)=-(\boldsymbol{R}'(t)\boldsymbol{R}'^T(t))^T \tag{5}$$

由公式（5）可知 $\boldsymbol{R}'(t)\boldsymbol{R}^T(t)$ 为反对称矩阵，记作：

$$\boldsymbol{R}'(t)\boldsymbol{R}^T(t)=\varphi^\wedge(t) \tag{6}$$

对公式（6）两端右乘 $R(t)$，得到：

$$\boldsymbol{R}'(t)=\varphi^\wedge(t)\boldsymbol{R}(t) \tag{7}$$

由公式（7）可得：

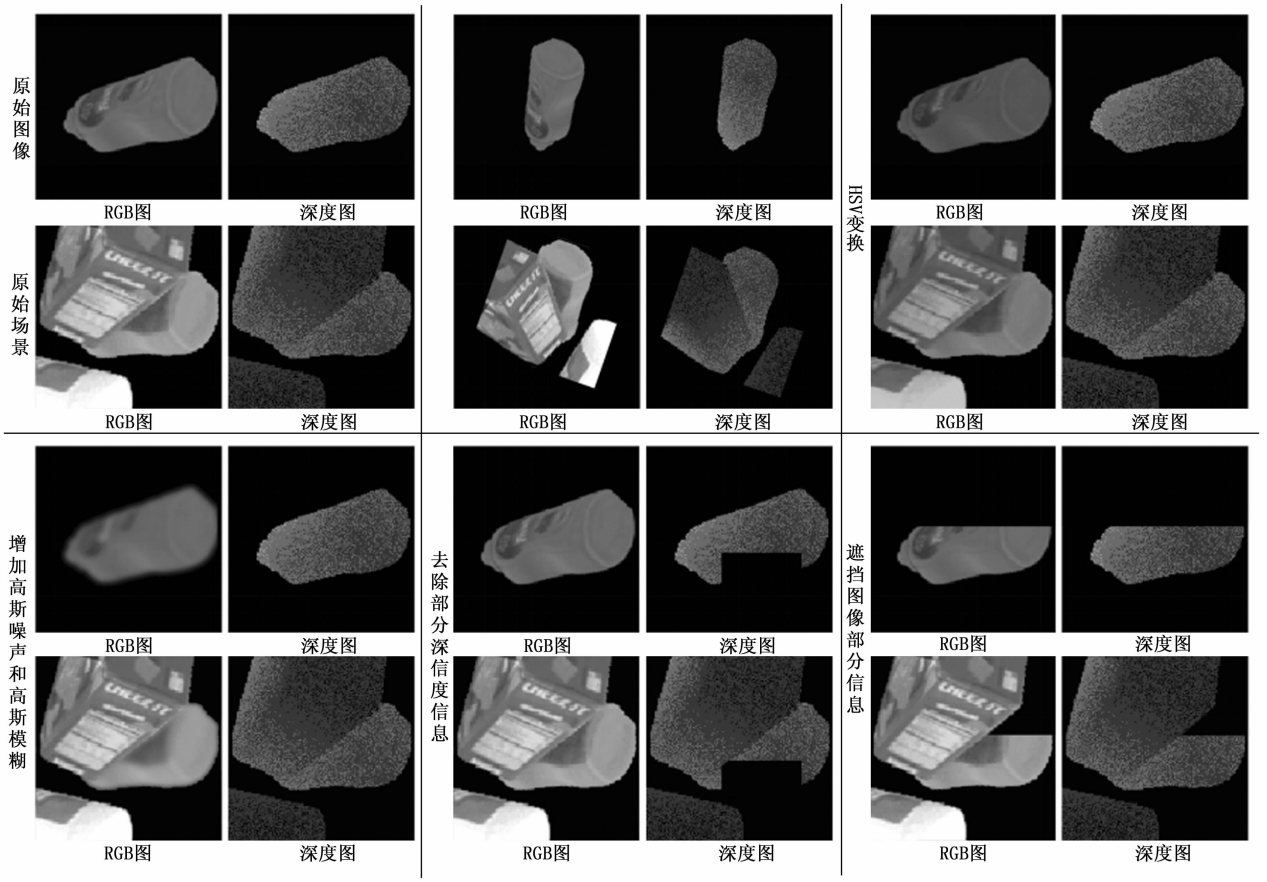

图 3 数据增强通道可视化结果

$$\boldsymbol{R}'(t)=\begin{bmatrix}0 & -\varphi_3 & \varphi_2\\ \varphi_3 & 0 & -\varphi_1\\ -\varphi_2 & \varphi 1 & 0\end{bmatrix}\boldsymbol{R}(t) \tag{8}$$

由公式（8）可知，对旋转矩阵 $\boldsymbol{R}$ 左乘 $\varphi^{\wedge}(t)$ 即为求导。在初始条件 $\boldsymbol{R}(0)=I$ 的情况下求解微分方程公式（8），得到：

$$\boldsymbol{R}(t)=\exp(\varphi^{\wedge}(t)) \tag{9}$$

公式（9）又称为李代数 $so(3)$ 和特殊正交群 $SO(3)$ 的指数映射关系。同理，可得李代数 $se(3)$ 和特殊欧式群 $SE(3)$ 指数映射关系如下：

$$\exp(\xi^{\wedge})=\begin{bmatrix}\sum_{n=0}^{\infty}\frac{1}{n!}(\varphi^{\wedge})^n & \sum_{n=0}^{\infty}\frac{1}{(n+1)!}(\varphi^{\wedge})^n\rho\\ 0^T & 1\end{bmatrix}\triangleq \begin{bmatrix}R & J\rho\\ 0^T & 1\end{bmatrix}=T \tag{10}$$

式中 $\boldsymbol{J}$ 为雅可比矩阵，表示为：

$$\boldsymbol{J}=\sum_{n=0}^{\infty}\frac{1}{(n+1)!}(\varphi^{\wedge})^n=\sum_{n=0}^{\infty}\frac{1}{(n+1)!}(\theta a^{\wedge})^n \tag{11}$$

对公式（11）中雅可比矩阵 $\boldsymbol{J}$ 进行泰勒展开，整理可得雅可比矩阵 $\boldsymbol{J}$ 的表达式为：

$$\boldsymbol{J}=\frac{\sin\theta}{\theta}I+(1-\frac{\sin\theta}{\theta})aa^T+\frac{1-\cos\theta}{\theta}a^{\wedge} \tag{12}$$

由公式（1）和（10）可得李代数 $se(3)$ 表达式为：

$$se(3)=\left\{\xi=\begin{bmatrix}\rho\\ \varphi\end{bmatrix}\in\mathbb{R}^6,\rho\in\mathbb{R}^3,\varphi\in\mathbb{R}^3\right\} \tag{13}$$

式中，ξ 表示李代数 $se(3)$ 中的元素，包含物体的物体一组三维移动 ρ 和三维旋转 φ。

3.2　网络结构

本文设计的网络结构如图 4 所示。网络输入数据为 176×176 大小的连续 3 帧 RGB-D 图像，图像包含 3 个 RGB 通道以及 1 个深度信息通道。在计算特征分支差异时采用 $\bar{\xi}$ 表示目标物体预测姿态相较真实姿态的移动误差和旋转误差。

对第 0 帧图像 N_0 和第 1 帧图像 N_1 提取时间特征得到第 0 帧特征 P_{T0} 和第 1 帧特征 P_{T1}，第 0 帧图像和第 1 帧图像的特征差 ΔP_1 计算公式为：

$$\Delta P_1=\alpha(P_{T0}-P_{T1}) \tag{14}$$

式中，α 为预定义的鲁棒损失函数，P_{T0} 和 P_{T1} 为经过特征提取后的像素强度值[25]。将特征差 ΔP_1 和第 1 帧图像 N_1 的空间特征 P_{S1} 输入时空间特征融合模块，用于预测第 2 帧图像 N_2 的物体姿态 P_2，公式为：

$$P_2=MLP(P_{S1}\oplus\Delta P_1) \tag{15}$$

式中，MLP 为多层感知机，$\oplus$ 用于对第 1 帧图像 N_1 的空间特征 P_{S1} 和特征差 ΔP_1 进行串联，二者经过卷积操作后维度相同。预测姿态 P_2 用于计算第 2 帧图像 N_2 帧的预测姿态相较于真实姿态 $\overline{P_2}$ 的旋转误差 ρ 和平移误差 θ。损失函数 L 基于均方误差（MSE，mean-square error）进行计算，表示每一批中不同预测姿态相较于真实姿态的平移和旋转差异大小，公式为：

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{16}$$

式中，N 为待测样本总数，y_i 为样本标签，$\hat{y}_i$ 表示样本预测概率。根据本文方法需求，对 MSE 公式（16）进行修改，将预测姿态的旋转损失和平移损失作为目标，损失函数 L 的定义如下：

$$L=M_1(\rho-\bar{\rho})^2+M_2(\varphi-\bar{\varphi})^2 \tag{17}$$

式中，M_1 和 M_2 分别表示网络模型移动误差和旋转误差的权重，需要手动设定，在本文方法中设置移动误差和旋转误差权重取值均为 1，表示网络训练时移动和旋转同样重要。$\boldsymbol{\rho}$ 和 $\bar{\boldsymbol{\rho}}$ 表示预测三维移动矩阵，$\boldsymbol{\varphi}$ 和 $\bar{\boldsymbol{\varphi}}$ 表示预测三维旋转矩阵。由于元素 ξ 表示每一个预测姿态 P_2 相较于真实姿态 $\overline{P_2}$ 的误差损失，最优预测姿态 $P_2{}^*$ 表示为：

$$\Delta P_t{}^*=\underset{\xi}{\operatorname{argmin}}\{(\rho-\bar{\rho})^2+(\varphi-\bar{\varphi})^2\} \tag{18}$$

3.3　残差连接

传统的卷积神经网络由多个隐藏层堆叠而成，这些隐藏层通过非线性激活函数相互连接，通过拟合最优解 $H(x)=x$，在网络中逐层提取目标物体特征。随着卷积神经网络隐藏层堆叠层数的逐渐增加，网络性能也愈发优异，但

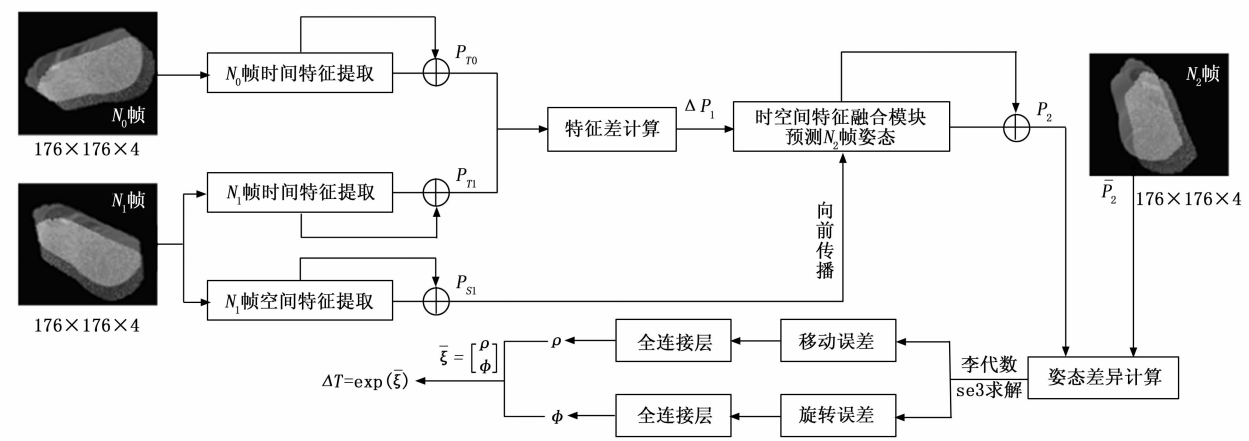

图 4　网络结构

是梯度消失和梯度爆炸的出现阻碍了卷积神经网络的发展。因为在深层网络中隐藏层堆叠过多，导致图像原始特征在逐层提取过程中逐渐消失或对同一特征反复提取，梯度在反向传播过程中逐渐减小或增大，导致训练和优化深层网络成为问题。

ResNet[26]网络提出了一种包含深度残差的网络模型，在网络层中加入图像原始特征跳跃连接，使每一个网络层都有原始特征的输入，解决了深层卷积神经网络出现的梯度消失和梯度爆炸问题，深度残差结构如图 5 所示。ResNet 网络由多个残差结构堆叠组成，每个残差结构包含一个恒等映射和跳跃连接，帮助网络将输入的特征映射添加到非线性变换的输出中，残差结构通过求解原始特征跳跃连接后的残差映射来间接提取物体特征：

$$H(x) = F(x) + x \tag{19}$$

由于残差结构的引入，ResNet 网络可以构建深层神经网络而不需要考虑梯度爆炸或梯度消失，这使得它成为深度学习中的一个重要突破，许多以残差结构为基础的深度神经网络在不同计算机视觉任务中出现并表现优异。

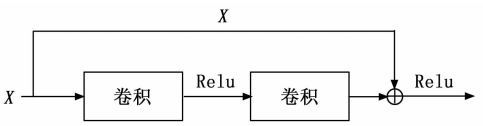

图 5 残差模块

本文方法在 ResNet 网络残差模块的基础上，将激活函数 ReLU 替换为激活函数 Swish。激活函数 ReLU 因其简单的计算方式和稀疏性成为神经网络选择的主流，缺点是当输入值小于 0 时，神经元没有响应能力，降低网络模型的表达能力。相较于激活函数 ReLU，激活函数 Swish 在输入小于 0 时同样有输出存在，且导数图像更加平滑，在深层网络中相较于激活函数 ReLU 具有更好的适用性[27]。图 6 为 Swish 函数及导数图像。在 Swish 函数图像中，当 $x > 0$ 时，函数仍然有输出，且当 $x \to +\infty, f(x) \to x$；当 $x \to -\infty, f(x) \to 0$。Swish 激活函数表示为：

$$f(x) = x \cdot \mathrm{sigmoid}(x) = \frac{x}{1+e^{-x}} \tag{19}$$

Swish 激活函数导数 $f'(x)$ 为：

$$f'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \tag{20}$$

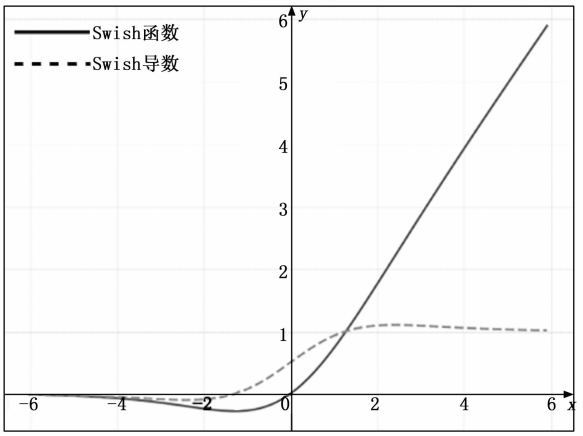

图 6 Swish 函数及其导数图像

3.4 时空间特征融合

本文方法提出了一种新的时空间特征融合模块，通过提取连续帧的时空间特征差异，获取目标物体在连续变化时出现的特征信息，模块结构如图 7 所示，图中⊕表示特征拼接操作。将输入大小为 $h \times w$ 的特征图通过自适应平均池化后分解为水平和垂直方向的特征分量，分别从时间和空间分量以不同的空间感受野和时域感受野进行特征捕获，强化网络对物体连续变化特征的提取能力，同时减少网络模型收敛所需轮次和时间，加快网络模型收敛速度。输出特征图 X_c 表示为：

$$X_c = f[Conv(t^w \otimes t^h)] \tag{21}$$

式中，t^w 和 t^h 为特征图二次分割后的垂直和水平特征分量，⊗用于对二者进行特征乘操作，$Conv$ 为卷积操作，保证输出特征图和输入特征图的大小、维度相同。

4 实验结果与分析

实验均在以下配置中完成：硬件环境：cpu 为英特尔 i7-8700 处理器，显卡为英伟达 RTX 3060，内存 16 GB；软件环境：系统为 Ubuntu，Python 版本为 3.6.9、OpenCV 版本为 4.0.0.21。

实验采用两种指标进行评估本文方法和同类网络模型

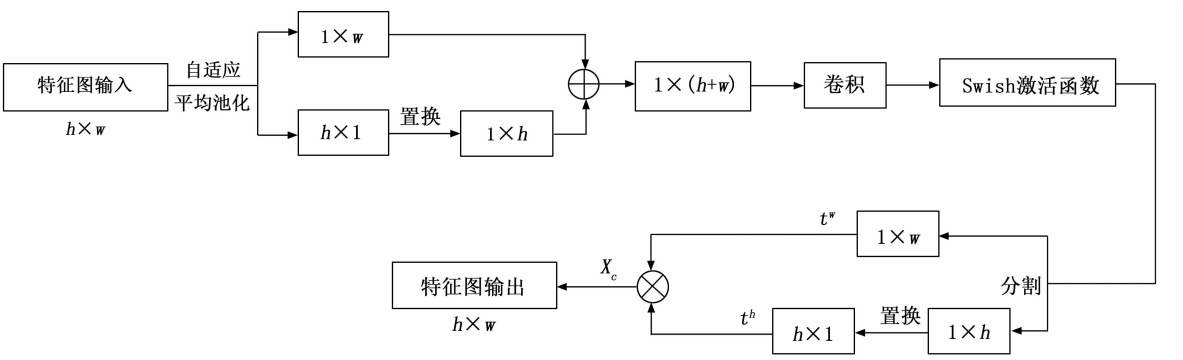

图 7 时空间特征融合模块结构

追踪目标物体的准确率，分别是 ADD（Average Distance of Descriptors）和 ADD-S（Average Distance of Descriptors for Symmetry）。其中，ADD 指标用于评估网络模型追踪非对称刚体的准确率，通过收集所有预测点到真实点的欧几里德距离，并计算所有点对距离平均值作为输出，得到该网络模型追踪目标物体的准确率；ADD-S 指标用于评估网络模型追踪对称刚体的准确率，通过将目标物体的预测点和真实点分别投影到同一面上，计算投影点对之间的平均距离得出网络模型追踪非对称刚体的准确率。两种评估指标均基于 ROC 曲线下与坐标轴围成的面积（AUC，area under curve）进行度量，从而衡量网络模型追踪目标物体姿态的准确率[27]。ADD 和 ADD-S 评估指标计算公式如下：

$$ADD = \frac{1}{m}\sum_{x\in M}\left\|\boldsymbol{R}x+\boldsymbol{T}-(\hat{\boldsymbol{R}}x+\hat{\boldsymbol{T}})\right\| \tag{22}$$

$$ADD-S = \frac{1}{m}\sum_{x_1\in M}\min_{x_2\in M}\left\|\boldsymbol{R}x_1+\boldsymbol{T}-(\hat{\boldsymbol{R}}x_2+\hat{\boldsymbol{T}})\right\| \tag{23}$$

式中，m 为网络模型预测目标物体总点数，$\boldsymbol{R}$ 和 $\boldsymbol{T}$ 是网络模型追踪目标物体预测姿态的旋转矩阵和平移矩阵，$\hat{\boldsymbol{R}}$ 和 $\hat{\boldsymbol{T}}$ 是目标物体真实姿态的旋转矩阵和平移矩阵，设置预测点与真实点误差最大阈值为 0.1 m，如果大于最大阈值，则表示该网络模型中预测点偏离真实点。

网络模型采用自适应运动估计算法（Adam，adaptive moment estimation）动态调整学习率，根据网络训练时的历史梯度不断自适应更新调整学习率，从而使损失函数 L 不断减小，加快网络模型收敛速度。在网络模型训练初期，初始学习率设置为 0.01，较大的学习率使网络模型在初期快速收敛；随着训练轮次的增加，Adam 不断动态降低网络模型学习率，更加精准地找到损失最小的网络模型，即在评估过程中表现最佳的网络模型。

4.1　YCB-Video 数据集与实验

本文方法和同类相关方法在训练和评估过程中均在 YCB-Video 数据集上进行，该数据集包含了多个复杂、遮挡的挑战性场景，选取 7 个 YCB 物体和 12 段用于评估网络模型追踪性能的视频序列，每个 YCB 物体均包含约 200 000 组手工标注的训练图像。此外，在实验中设计消融实验，以评估本文方法中提出的不同模块对网络性能的影响。

本文方法使用合成数据生成通道生成的合成数据进行训练，通过约 20 000 组经过数据增强的合成数据训练得到最优网络模型，为同类相关方法训练数据的十分之一。表 1 在 YCB-Video 数据集上本文方法及同类相关方法的对比评估结果，其中大夹钳和超大夹钳为对称刚体，以 * 标出。在本文方法和同类相关方法的追踪过程中只允许初始化一次，在追踪过程中不允许重新初始化姿态。

实验结果表明，本文方法与同类相关方法相比，在 ADD 和 ADD-S 指标中分别达到了 92.56 和 94.22，表现最佳，追踪效果最稳定。在网络模型追踪目标姿态过程中，实时追踪频率也至关重要，其决定了网络模型追踪目标物体时的连贯程度和流畅度。本文方法的实时追踪频率达到 79 Hz，与同类相关方法相比，本文方法实时追踪频率最高，在复杂场景下网络模型追踪过程仍较为流畅。追踪目标姿态过程可视化实验结果如图 8 所示。图 8 中 A 表示缺乏纹理的对称刚体大夹钳，B 表示富有纹理的非对称刚体厨艺罐头，C 表示无纹理的非对称刚体马克杯，同一组图像从左到右按照时间序列排序，其中物体 A 和物体 B 均遇到了遮挡的挑战场景。由图 8 可知，在存在遮挡和多目标物体的复杂场景中，本文方法预测姿态与真实姿态接近，在部分场景下预测姿态和真实姿态相同，验证了本文方法能够有效、稳定地追踪目标物体，且在复杂场景下具有高鲁棒性，追踪效果平滑连贯，可以投入到真实场景中进行使用。

4.2　消融实验

通过在 YCB-Video 数据集上进行不同模块的消融实验，

表 1　不同方法在 YCB-Video 数据集上追踪评分比较

方法	评估指标	物体							
		厨艺罐头	番茄罐头	芥末瓶	肉罐头	马克杯	大夹钳 *	超大夹钳 *	平均
DenseFusion[28]	ADD	—	—	—	—	—	—	—	—
	ADD-S	96.40	94.60	97.20	91.30	97.50	72.90	69.80	88.53
PoseCnn＋ICP＋DeepIM[29]	ADD	78.00	90.30	**97.10**	82.20	88.90	54.20	36.50	75.31
	ADD-S	96.30	94.80	**98.00**	90.30	**98.20**	77.90	77.80	90.47
DeepIM[30]	ADD	89.00	89.10	92.00	78.00	80.40	73.90	49.30	78.81
	ADD-S	93.80	93.20	95.10	88.90	91.20	84.10	90.30	90.94
VIPose[31]	ADD	—	71.54	70.87	—	—	—	—	71.21
	ADD-S	—	82.79	82.12	—	—	—	—	82.46
HFF6D[32]	ADD	91.80	**94.22**	96.89	50.77	90.73	—	—	38.22
	ADD-S	95.79	**97.25**	**97.76**	**76.83**	**96.38**	**—**	**—**	**39.00**
本文方法	ADD	**92.33**	92.26	96.32	**87.67**	**92.17**	**95.52**	**91.65**	92.56
	ADD-S	95.21	96.10	94.97	**92.13**	94.32	92.15	94.67	94.22

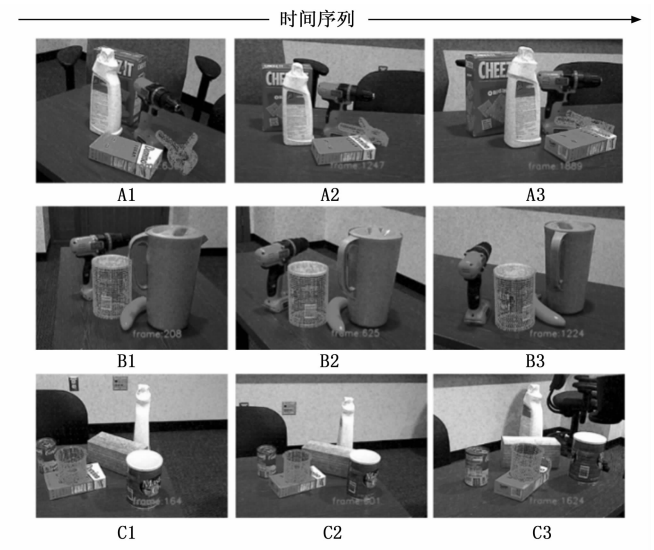

图 8 本文方法在 YCB-Video 数据集中的可视化姿态追踪结果

验证本文方法所提出的不同模块的有效性，实验结果如表 2。采用 3 个 YCB 目标物的 20 000 组数据训练作为基准从而验证不同模块的有效性。

表 2 消融实验

消融模块	评估标准	物体			
		厨艺罐头	马克杯	大夹钳 *	平均
基准	ADD	92.33	92.17	95.52	93.34
	ADD-S	95.21	94.32	92.15	93.89
数据增强	ADD	52.37	45.62	62.47	53.49
	ADD-S	61.15	47.85	56.65	55.22
残差连接	ADD	68.97	62.25	75.56	68.93
	ADD-S	72.55	65.64	68.67	68.95
时空间特征融合	ADD	64.56	41.65	46.77	50.99
	ADD-S	68.96	45.84	50.98	55.26

1）数据增强。不使用数据增强通道生成的数据进行训练，采用 20 000 组合成数据驱动网络来验证数据增强通道的有效性。未使用数据增强通道的评估指标 ADD 和 ADD-S 分别降低 39.85 和 38.67，证明数据增强通道能够加强训练样本的多样性，增强网络在复杂场景下的鲁棒性。

2）残差连接。从网络中去除残差连接进行训练和评估。未采用残差采样连接的评估指标 ADD 和 ADD-S 降低了和 24.41 和 24.94。由于残差连接能够较好保留物体的原始特征，加强网络提取物体的边缘特征信息，同时在追踪过程中减小因拍摄移动产生的目标物体抖动。

3）时空间特征融合。将时空间特征融合模块替换为传统的卷积模块和 ReLU 激活函数，未采用时空间特征融合的网络评估得分最低，分别降低 42.35 和 38.63。证明时空间特征融合能够有效提取连续帧中物体的姿态差异，增强网络对于目标姿态追踪的准确性。

消融实验结果表明，在上述模块的共同作用下，网络可以长期稳定地实时追踪目标物体的 6D 姿态，实验结果验证了不同模块对网络的有效性。

5 结束语

在这项工作中，本文设计了一种刚体姿态实时估计网络结构，该网络通过提取并学习连续帧中目标物体姿态的差异特征，从而做到实时、稳定的目标姿态追踪效果。通过在复杂和遮挡挑战性数据集 YCB-Video 上评估实验，证明本文提出的网络可以在数据稀缺的条件下进行训练并快速收敛，在复杂场景中进行追踪能力评估，本文方法相较于同类相关方法做到了最佳的实时目标追踪效果。与其他方法相比，本方法追踪精度更高，能够更精确地实时追踪目标物体，在遮挡场景下仍能做到较好的追踪效果。本文方法不足是网络结构存在一定冗余，当场景中存在相似的目标物体时，追踪速度可能无法维持稳定。未来的研究将致力于简化网络结构并实现多目标姿态追踪，以提高在复杂场景条件下的实时追踪目标物体姿态性能。

参考文献：

[1] LOWE D G. Object recognition from local scale-invariant features [C] //Proceedings of the Seventh IEEE International Conference on Computer Vision, IEEE, 1999, 2: 1150 - 1157.

[2] SCHMIDT T, NEWCOMBE R A, FOX D. DART: dense articulated real-time tracking [C] //Robotics: Science and Systems, 2014, 2 (1): 1 - 9.

[3] SAVARESE S, FEI-FEI L. 3D generic object categorization, localization and pose estimation [C] //2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007: 1 - 8.

[4] UMEYAMA S. Least-squares estimation of transformation parameters between two point patterns [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991, 13 (04): 376 - 380.

[5] HINTERSTOISSER S, LEPETIT V, RAJKUMAR N, et al. Going further with point pair features [C] //Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part III 14. Springer International Publishing, 2016: 834 - 848.

[6] VIDAL J, LIN C Y, MARTí R. 6D pose estimation using an improved method based on point pair features [C] //2018 4th International Conference on Control, Automation and Robotics (iccar), IEEE, 2018: 405 - 409.

[7] HINTERSTOISSER S, LEPETIT V, LLIC S, et al. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes [C] //Computer Vision-ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Revised Selected Papers, Part I 11. Springer Berlin Heidelberg, 2013: 548 - 562.

[8] RIOS-CABRERA R, TUYTELAARS T. Discriminatively

trained templates for 3d object detection: A real time scalable approach [C] //Proceedings of the IEEE International Conference on Computer Vision, 2013: 2048 - 2055.

[9] KEHL W, TOMBARI F, NAVAB N, et al. Hashmod: A hashing method for scalable 3D object detection [J]. ArXiv Preprint ArXiv: 1607.06062, 2016.

[10] AUBRY M, MATURANA D, EFROS A A, et al. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 3762 - 3769.

[11] BRACHMANN E, KRULL A, MICHEL F, et al. Learning 6d object pose estimation using 3d object coordinates [C] // Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6 - 12, 2014, Proceedings, Part II 13, Springer International Publishing, 2014: 536 - 551.

[12] GALL J, YAO A, RAZAVI N, et al. Hough forests for object detection, tracking, and action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33 (11): 2188 - 2202.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25.

[14] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 28.

[15] REDMON J, DIWVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779 - 788.

[16] LIU W, AAGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector [C] //Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21 - 37.

[17] PRISACARIU V A, REID I D. PWP3D: Real-time segmentation and tracking of 3D objects [J]. International Journal of Computer Vision, 2012, 98: 335 - 354.

[18] WANG C, XU D, ZHU Y, et al. Densefusion: 6d object pose estimation by iterative dense fusion [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3343 - 3352.

[19] WANG H, SRIDHAR S, HUANG J, et al. Normalized object coordinate space for category-level 6d object pose and size estimation [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2642 - 2651.

[20] WEN B, MITASH C, REN B, et al. se (3) -tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains [C] //2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020: 10367 - 10373.

[21] LI Y, WANG G, JI X, et al. Deepim: Deep iterative matching for 6d pose estimation [C] //Proceedings of the European Conference on Computer Vision (ECCV), 2018: 683 - 698.

[22] LIU Y, PENG J, DAI W, et al. Joint spatial and scale attention network for multi-view facial expression recognition [J]. Pattern Recognition, 2023, 139: 109496.

[23] MITASH C, BEKRIS K E, BOULARIAS A. A self-supervised learning system for object detection using physics simulation and multi-view pose estimation [C] //2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017: 545 - 551.

[24] TREMBLAY J, TO T, BIRCHFIELD S. Falling things: A synthetic dataset for 3d object detection and pose estimation [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 2038 - 2041.

[25] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40 (3): 611 - 625.

[25] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.

[26] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions [C] //Proceedings of the 6th International Conference on Learning Representations, 2017.

[27] XIANG Y, SCHMIDT T, NARAYANAN V, et al. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes [J]. ArXiv Preprint ArXiv: 1711.00199, 2017.

[28] HE Y, SUN W, HUANG H, et al. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11632 - 11641.

[29] LI Z, WANG G, JI X. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7678 - 7687.

[30] LI Y, WANG G, JI X, et al. Deepim: Deep iterative matching for 6d pose estimation [C] //Proceedings of the European Conference on Computer Vision (ECCV), 2018: 683 - 698.

[31] GE R, LOIANNO G. Vipose: Real-time visual-inertial 6d object pose tracking [C] //2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021: 4597 - 4603.

[32] LIU J, SUN W, LIU C, et al. Hff6d: Hierarchical feature fusion network for robust 6d object pose tracking [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32 (11): 7719 - 7731.