Harmonious Feature Learning for Interactive Hand-Object Pose Estimation(2023 CVPR) 复现报告

罗志雄

2024 年 12 月 25 日

摘要

从单张图像中联合估计手和物体的 3D 姿态极具挑战性，因为当手与物体交互时，往往会发生严重的遮挡。现有的方法通常首先从单个骨干网络中提取粗略的手和物体特征，然后通过交互模块进一步相互增强。然而，这些工作通常忽略了手和物体在特征学习方面的竞争，因为骨干网络将它们都视为前景，并且它们通常相互遮挡。在该论文中，提出了一种新颖的和谐特征学习网络 (HFL-Net [15])。HFL-Net 引入了一种新的框架，该框架结合了单流和双流骨干网络的优势：它为手和物体共享一个公共 ResNet-50 模型的低级和高级卷积层的参数，而中级层则不共享。这种策略使手和物体能够被中级层作为唯一的目标提取，避免了它们在特征学习中的竞争。共享的高级层也迫使它们的特征保持和谐，从而促进它们之间的相互特征增强。特别是，提出通过将物体流中相同位置的特征与手特征连接来增强手的特征。随后采用自注意力层对连接的特征进行深度融合。实验结果表明，该方法在流行的 HO3D 和 Dex-YCB 数据库上始终优于最先进的方法。

关键词： 3D 手部姿态估计；物体位姿估计；特征提取

1 引言

1.1 选题背景

当人类与真实世界互动时，他们主要依靠双手。因此，准确理解手与物体之间的交互对于理解人类行为至关重要。它可以广泛应用于各种领域，包括虚拟现实的开发、增强现实和基于模仿的机器人学习等。近年来，基于单目 RGB 图像的手姿态估计和 6D 物体姿态估计取得了显著成果。然而，在交互情况下进行手-物体姿态估计的研究仍处于起步阶段。

如图 1所示，从单张图像中进行手物体姿态估计极具挑战性。主要原因在于，当手与物体相互作用时，会发生严重的遮挡；而遮挡反过来会导致信息丢失，从而增加了每个任务的难度。

现有的解决相互遮挡的方法主要分为两类：第一类方法是利用上下文信息。由于物理约束，交互中的手和物体在姿态上往往高度相关，这意味着其中一个的外观可以成为对另一个

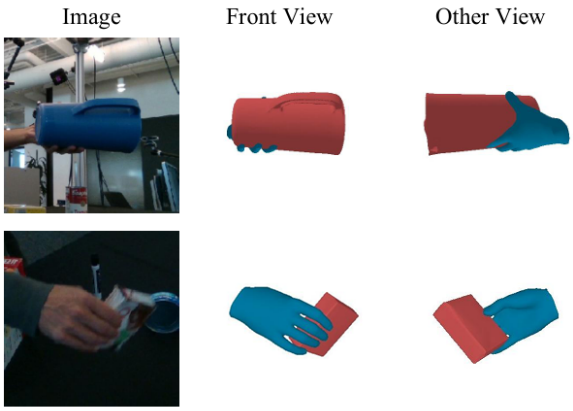

图 1. 手部遮挡情况

有用的上下文信息。采用这种方案的方法通常使用单个骨干网络分别提取手和物体的特征。这种统一的骨干网络模型确保手和物体特征位于相同的特征空间中，从而便于后续通过基于注意力的方法在手和物体之间进行相互特征增强。但是这种方法容易导致手和物体在特征学习上的竞争，从而降低估计精度。具体地说，当手和物体彼此靠近时，骨干网络模型会将它们都视为前景，因此可能无法区分手特征和物体特征。

另一种方法是使用两个独立的骨干网络分别提取手和物体的特征，从而避免竞争。当采用这种方法时，每个骨干网络只有一个目标作为前景。然而，这种策略的主要缺点包括模型尺寸较大，更重要的是，骨干网络之间存在不同的特征空间，这给手和物体之间的相互特征增强带来了困难。

1.2 选题依据

而这篇论文提出的和谐特征学习网络 (HFL-Net) 框架巧妙地解决了上述问题。采用基于 ResNet50 的 backbone，结合了单流和双流 backbone 的优点，通过共享低级和高级层参数，保留中级层独立，使手和物体特征在相同的特征空间中，避免了特征学习的竞争，并促进了相互增强。除此之外，该论文还提出了高效的手-物相互增强模块，构建了通过手来增强物体特征和通过物体来增强手部特征的注意力模块，巧妙的利用了手-物之间在物理学上形态高度相关的特性。

1.3 选题意义

HFL-Net 在交互式手部和物体姿态估计任务上取得了优异的性能，并且代码开源，方便进行复现和改进。通过复现 HFL-Net 的设计思路和实现细节，可以更深入地理解和谐特征学习在解决手物体交互遮挡问题上的优势，验证其在手部姿态估计领域的性能，并了解其是否适合解决手-物人机交互问题。论文提出的自注意力模块和交叉注意力模块，为手和物体特征交互提供了有效的机制，值得深入研究其原理和优化方法。除此之外，还能够了解该领域的评估指标和常用的公共数据集，了解该领域所用的最新技术，对以后的研究方向有初步的了解，打下扎实的基础。

2 相关工作

2.1 基于 RGB 图像的 3D 手部姿态估计

现有的基于 RGB 图像的 3D 手部姿态估计方法大致可以分为两类：无手部模型方法和基于手部模型的方法。

无手部模型方法 [3, 11, 19] 直接从单个 RGB 图像中预测 3D 手部关节坐标或 3D 手部网格顶点坐标。预测 3D 手部关节坐标更容易，但关节缺乏手部表面的几何信息。相比之下，3D 网格包含丰富的几何拓扑结构。为了获得网格顶点之间的合理拓扑结构，通常采用图卷积 [3, 19] 来细化顶点特征。然而，直接预测手部网格需要密集且准确的 3D 顶点标注，这些数据很难获得。

基于手部模型的方法 [12, 13, 17] 利用手部先验信息来简化网格预测任务。它们通常基于参数化手部模型，例如 MANO [18]，该模型已在大量手动扫描的各种姿态和形状的手部网格上进行预训练。有了这些先验信息，基于手部模型的方法只需要估计少量姿态和形状系数，就可以获得手部网格。在许多最新的工作 [12, 17] 中可以发现良好的结果。

2.2 基于 RGB 图像的 3d 手-物姿态估计

现有的基于 RGB 的 3D 手-物体姿态估计技术可以分为两大类：基于优化的方法 [1,5,9,20] 和基于学习的方法 [4, 7, 8, 10, 16]。

基于优化的方法根据手和物体的接触面以及物理约束（即吸引和排斥）来优化手和物体的位姿。估计手和物体之间的接触面通常很耗时 [5]。为了解决这个问题，Tse 等人 [20] 提出了一个基于图的网络来加速接触面的估计。

基于学习的方法设计了用于联合手和物体位姿估计的统一模型。它们通常采用现成的手部模型，例如 MANO [18]，并且还假设 3D 物体模型可用。因此，它们可以直接根据这些先验信息预测手和物体的位姿。早期工作 [4, 10] 采用双流骨干网络进行独立的手和物体位姿估计，这以更高的模型复杂度为代价。最近的工作 [7, 8, 16] 采用单流骨干网络提取手和物体特征。然而，它们忽略了如果采用单流骨干网络，手和物体特征学习是竞争性的。

在这篇论文中，提出了一种新的框架，它可以提取和谐的手和物体特征，这不仅缓解了手和物体姿态估计任务之间的竞争，而且还能够有效地增强手和物体特征之间的相互提升

3 本文方法

3.1 本文方法概述

该论文提出了一个和谐特征学习 HFL-Net 框架，如图 2所示，该框架包括一个经过精心设计的基于 ResNet-50 的骨干网络、手部和物体交互增强模块以及手部和物体解码器。当输入一张 RGB 图片时，经过 backbone 网络后，会分别提取手部和物体的特征，并经过交互模块进行特征增强，最后由解码器分别得到手部的 3D 网格、3D 关节坐标和物体的姿态。

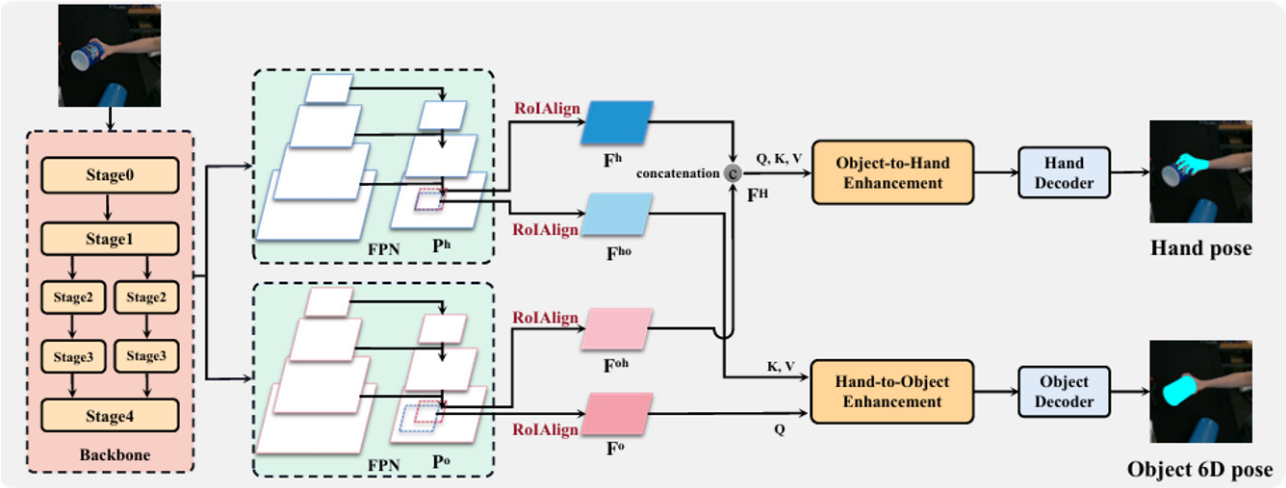

图 2. HFL-Net 网络框架图

3.2 特征提取骨架网络

当手部和物体发生交互时，容易产生遮挡，导致信息丢失，增加了手部和物体姿态估计的难度。现有的单流骨干网络将手部和物体都视为前景目标，容易导致特征学习上的竞争，难以区分手部和物体的特征。双流骨干网络虽然可以避免竞争，但会导致模型尺寸过大，且两个分支的特征空间不同，难以进行特征增强。

HFL-Net 骨干网络结合了单流和双流骨干网络的优点，将骨干网络中间层分为两个独立的分支，分别来提取手和物体的特征，避免在特征学习中竞争，然后通过共享最高层，迫使它们的特征保持和谐，从而促进它们之间的相互特征增强。具体来说，它保留了 ResNet-50 模型中 stage-0、stage-1 和 stage-4 层的结构，但为手和物体分别采用了独立的 stage-2 和 stage-3 层。stage-1 层输出的特征图分别输入到两组 stage-2 和 stage-3 层中。这样，每一组 stage-2 和 stage-3 层就只有一个前景目标，因此可以专注于手或物体的特征学习。

最后，stage-3 层输出的两组特征图被输入到相同的 stage-4 层中。由于 stage-4 层的参数共享，手和物体的特征空间被迫统一，从而便于后续它们特征之间的相互增强操作。之后通过采用 FPN 特征金字塔网络 [14] 来融合 stage-1 到 stage-4 的特征图，由于手和物体有独立的 stage-2 和 stage-3 层，所以它们采用了不同的 FPN。具体来说，在 FPN 特征金字塔中，从上到下逐一将各 stage 的特征图先经过 1x1 卷积层进行通道对齐，然后通过上采样（双线性插值）缩放到统一尺寸，最后各像素直接相加，得到融合增强后的特征 Ph (表示手的特征图) 和 Po (表示物体的特征图)。

3.3 物-手交互增强模块

手和物体在交互过程中，虽然会发生相互遮挡，但是由于手和物体在姿态上高度相关，意味着一个的外观可以提供对另一个有用的上下文信息。

所以当手被物体遮挡时，手握持的物体部分可以对手的特征进行增强，当物体被手遮挡时，用手的部分对物体进行增强。

以物体对手的特征进行增强为例。先是对 FPN 得到的手的特征图 Ph、物体的特征图 Po 进行 ROIAlign 操作，ROIAlign 是用于区域兴趣对齐（Region of Interest Align, ROI Align）

的操作，它的主要功能是从特征图中提取固定大小的特征区域，这些特征区域对应于输入的边界框（bounding boxes）。

具体来说，就是根据手部的边界框来提取 Ph 和 Po 中与手相关的特征区域，得到固定大小的特征图 Fh 和 Foh，其中，Fh 是手的特征图，Foh 是物体遮挡手的那部分特征。举个例子，在图 3中，红色框就是手的边界框，而蓝色框就是 Fh 注重的特征部分，绿色框就是 Foh 注重的物体遮挡手的那部分特征。

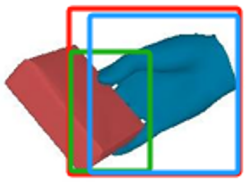

图 3. 手-物交互增强特征图

将 Fh 和 Foh 在通道维度上拼接后，生成特征图的 FH。再将 FH 输入到物-手交互增强模块中，如图 4所示。在物-手交互增强模块中，通过 1×1 卷积层实现像素内特征融合，然后通过多头注意力机制进行像素间的特征融合。在注意力机制中，Foh 会影响 Fh 中所有像素的特征。

最后输出特征图 Fhe，并将 Fhe 特征图输入到 Hand Decoder 解码器中得到手模型的参数，从而估计出手的姿态。

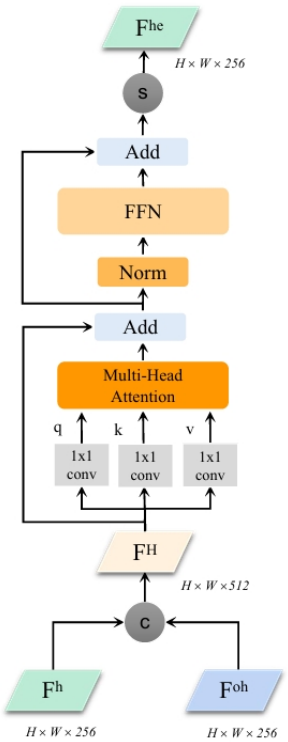

图 4. 物-手交互增强模块

4 复现细节

4.1 与已有开源代码对比

在此论文复现过程中，我参考了作者公布的源代码，该代码地址为 https://github.com/lzfff12/HFL-Net。在此代码的基础上，我重新训练了网络模型，在一个公共数据集 HO3d [6] 中进行训练并测试。

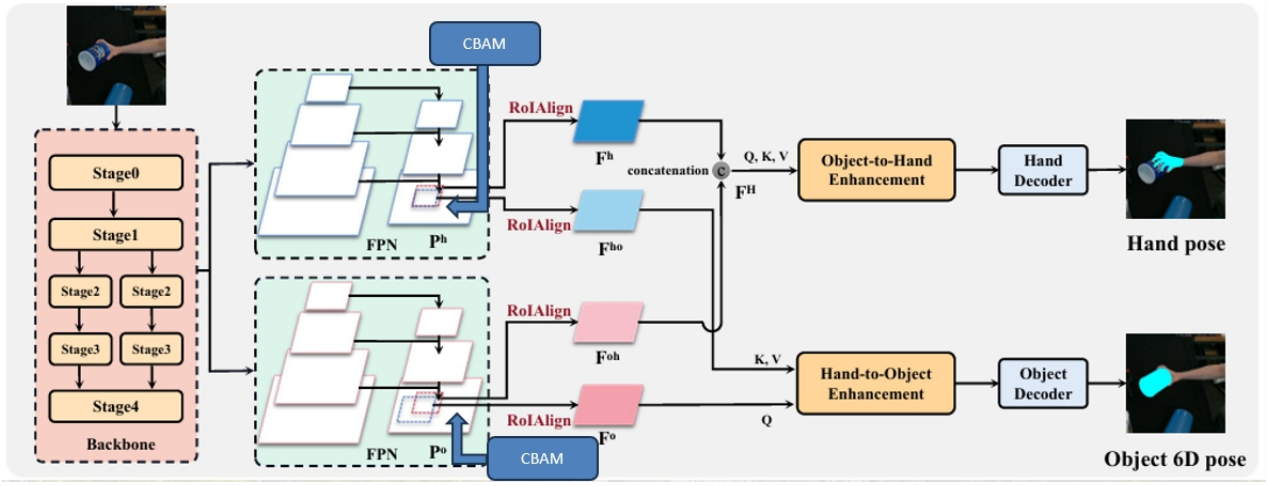

图 5. 加了 CBAM 块的 HFL-Net 网络框架图

此外，我还在该网络框架中，如图 5所示，添加了一个空间-通道维度的注意力机制模块——CBAM 模块，以此来弥补该网络模型在通道和空间上的特征信息关注的缺失。因为在此论文中，虽然关注了像素内的特征融合，但只是用了一个 1x1 的卷积核简单地来进行像素内特征融合。因此，我加入了一个通道注意力机制模块，生成一个权重序列，权衡不同通道的特征，更加合理的进行像素内的特征融合。同时，该论文使用了多头注意力机制来进行像素间的特征融合，我在此前先使用一个空间卷积注意力机制模块，生成一个权重矩阵，来平衡像素间的特征权重，来合理的突出重要的特征。

此改进在数据集 HO3d 中，经过实验发现，有一定的性能提升，能够体现出改进的效果。

4.2 实验环境搭建

软件环境的安装需要先根据项目给出的包环境需求文件 requirement.txt 来安装相对应的包，该代码使用的是 python 语言，并采用了 Pytorch 框架，可以使用 pip install 或者 conda install 命令来进行安装。然后下载对应的数据集 HO3d 或数据集 DexYCB [2]，HO3d 数据集下载地址在 此处和 DexYCB 数据集下载地址在 此处。之后，还需下载 MANO 手部模型文件，在 此处进行下载，下载完后存放到 assets/mano-models 文件夹中。

硬件环境需要一个有英伟达显卡（用于模型训练和推理）、linux 环境和存储空间至少 300G 的服务器，我已经在自己的设备，具有一张 24G 显存的 3090 显卡的服务器上进行运行和测试了。

4.3 界面分析与使用说明

下载完数据集后，记录数据集存放的文件夹的位置，并在项目中的名为 sh 文件夹里，修改 train-ho3d.sh 和 train-dex-ycb.sh 脚本文件，将脚本文件里的数据集存放路径参数改为自己数据集存放的文件夹路径。最后使用命令 sh 运行脚本文件，例如 sh train-ho3d.sh，就可以进行数据集的训练和测试了。同时也可以修改脚本文件里的保存模型参数、测试结果，训练结果的路径参数。

在该项目中，每经过 5 个 epoch 就会保存一次模型参数，并进行一次测试，由于 HO3d 数据集来源于一个公开的比赛，并没有公布测试集的 ground true 标签，因此测试生成的 json 文件需要放到 HO3d 比赛官网中进行测评，大概半小时会出结果。

4.4 创新点

我在此论文中的改进主要是，使用了 CBAM 卷积注意力机制模块，如图 6所示。为了强调空间和通道这两个维度上的有意义特征，CBAM 模块依次应用通道和空间注意模块，来分别在通道和空间维度上学习关注什么、在哪里关注。

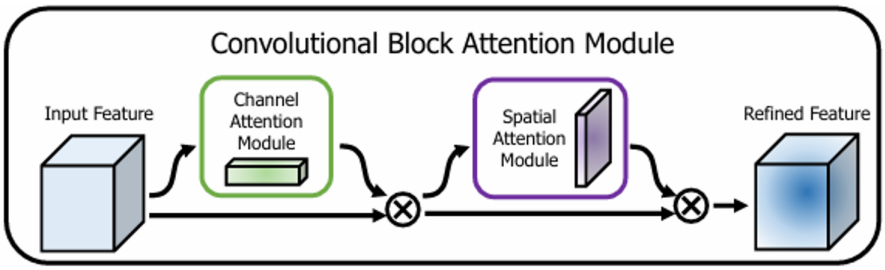

图 6. CBAM 卷积注意力机制模块

CBAM 的通道注意力模块如图 7所示，主要是通过对空间维度进行全局平均池化和全局最大池化，并经过两个全连接层和 Sigmoid 激活函数，生成一个通道注意力图，用于调节每个通道的权重。具体操作而言，就是先分别用 GAP 全局平均池化层和 GMP 全局最大池化得到两个 Cx1x1 的向量，然后再经过 conv，ReLU，conv，sigmoid 和进行归一化后，生成两个通道权重向量（Cx1x1）再相加，最后再将权重图与原特征图相乘。

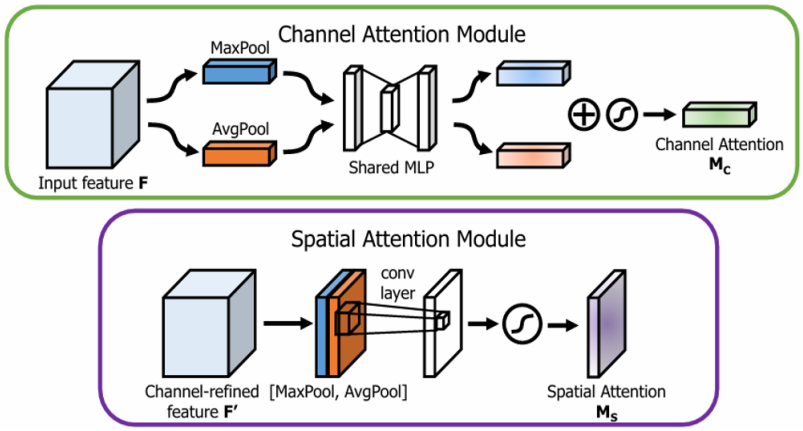

图 7. 通道和空间注意力机制模块

CBAM 的空间注意力模块如图 7所示，主要是通过对通道维度进行求平均和求最大值，合并得到一个通道数为 2 的卷积层，然后通过一个卷积，得到了一个通道数为 1 的空间像素权重图。具体操作而言，就是对特征图的每个位置的所有通道上先做最大池化和平均池化，来得到两个特征图后，在通道维度上拼接，再对这特征图进行 7x7 Conv，最后进行 BN 和 sigmoid 归一化，得到每个像素的权重图，再将权重图与原特征图相乘。

5 实验结果分析

此次论文复现，我主要针对 HO3d 这个数据集进行训练以及测试，HO3d [6] 一个具有手和对象的 3D 姿势注释的手对象交互数据集。该数据集包含来自总共 68 个序列的 66,034 个训练图像和 11,524 个测试图像。这些序列在多摄像头和单摄像头设置中捕获，包含 10 个不同的主体，从 YCB 数据集中操纵 10 个不同的对象。注释是使用优化算法自动获得的。

该实验主要包括 5 个指标，他们分别是

1. 平均关节误差 (PAMPVPE): 单位为毫米，估计手部 21 个关键点与真实值的平均误差，越低越好。
2. 平均网格误差 (PAMPVPE)：单位为毫米，手部网格重建的准确率，值越低越好。
3. F@5: 5 个最接近真实关键点的预测关键点中，至少有一个预测关键点与真实关键点之间的误差小于 5 毫米的概率，值越高越好。
4. F@15: 15 个最接近真实关键点的预测关键点中，至少有一个预测关键点与真实关键点之间的误差小于 15 毫米的概率，值越高越好

方法	Joint⬇(单位:mm)	Mesh⬇ (单位:mm)	F@5⬆	F@15⬆
原文	8.9	8.7	57.5	96.5
自己重新训练	9.6	9.6	53.6	**95.7**
改进后	**9.4**	**9.5**	**54.7**	95.4

图 8. 实验结果

具体的实验结果如图 8所示，所有结果都是在数据集 HO3d 中得到的。第一行是作者在论文中展示的。第二行是我将作者提供的代码，在数据集中保持原来的超参数进行训练且测试而来的，可以看见结果比作者公布的结果在 Joint 误差和 Mesh 误差上要高 0.7mm 和 0.9mm。第三行数据是我改进后的代码的实验结果，比自己重新训练的结果在 Joint 误差和 Mesh 误差上要低 0.2mm 和 0.1mm，并且在 F@5 中比自己训练的要高 1.1，可以看出改进还是有效果的。

6 总结与展望

本次计算机前沿技术复现，我选择的是和我研究领域相关的前沿论文《Harmonious Feature Learning for Interactive Hand-Object Pose Estimation》，这是一篇关于 3d 手部姿态估计的顶会论文，提出了一个和谐特征学习框架，解决了手-物之间特征学习存在相互竞争的问题。在数据集 HO3d 和 DexYCB 两个数据集上达到了 sota。通过复现这篇论文，我了解了手部姿态估计的前沿技术，对评估指标和常用的公共数据集有了一定的了解，并在此基础上对该论文提出的 HEF-Net 网络框架进行改进，添加了 CBAM 卷积注意力模块，有一定的提升效果。

但本次前沿论文复现还存在几点不足：

- 复现结果与原论文存在差距：可能是由于超参数设置、设备、数据集的差异或模型实现细节等原因导致自己重新训练并测试时，没有达到作者论文中所示的结果。
- CBAM 模块改进效果有限：虽然改进后在部分指标上有所提升，但提升幅度不大，同时也有指标没有提升，例如 F@15 指标就没有提升，需要进一步探索改进方法。
- 缺少对其他数据集的测试：HO3D 数据集可能存在局限性，需要测试更多数据集来验证改进后的模型的提升。

除此之外，未来还可进一步对模型结构进行研究，使用可以更有效的特征提取和交互增强方法，例如使用 Transformer 或图神经网络等。还可以利用手部模型的几何约束或物体模型的语义信息来提高估计精度。

参考文献

[1] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12397–12406, 2021.

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021.

[3] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10446–10455, 2021.

[4] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021.

[5] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[6] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020.

[7] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11090–11100, June 2022.

[8] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020.

[9] Yana Hasson, Gül Varol, Ivan Laptev, and Cordelia Schmid. Towards unconstrained joint hand-object reconstruction from rgb videos. *2021 International Conference on 3D Vision (3DV)*, pages 659–668, 2021.

[10] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808, 2019.

[11] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[12] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, June 2021.

[13] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12919–12928, 2021.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[15] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *2023 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 12989–12998, 2023.

[16] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, June 2021.

[17] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1495, 2022.

[18] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands. *ACM Trans. on Graphics*, 36:1 – 17, 2017.

[19] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11678–11687, 2021.

[20] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ale Leonardis, Feng Zheng, and Hyung Jin Chang. S2contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. *ArXiv*, abs/2208.00874, 2022.