EvHand-FPV: 基于事件的高效 3D 手部追踪从第一人称视角出发

*Zhen Xu, *Guorui Lu, Chang Gao, Qinyu Chen

Abstract— **手部追踪为直观交互范式带来了巨大的潜力，但基于帧的方法往往难以满足准确性、低延迟和能效的要求，特别是在资源受限的环境中，如扩展现实（XR）设备。**

事件相机通过异步感知亮度变化，在毫瓦级功耗下提供 μ 秒级的时间分辨率。

在这项工作中，我们提出了 EvHand-FPV，这是一个轻量级框架，用于从单个事件相机进行以自我为中心（First-Person-View）的三维手部追踪。我们构建了一个基于事件的第一人称视角数据集，该数据集结合了带有三维标签的合成训练数据和带有二维标签的真实事件数据用于评估，以解决缺乏自我为中心基准的问题。EvHand-FPV **还引入了一个基于手腕的兴趣区域（ROI），通过几何线索定位手部区域，并结合了一种端到端映射策略，将 ROI 偏移嵌入网络中以减少计算而无需显式重建，以及一种带有辅助几何特征头部的多任务学习策略，在测试时不会增加开销的情况下改进表示。在我们的实际第一人称视角测试集上，**EvHand-FPV **将 2D-AUCp 从 0.77 提高到 0.85，同时参数量从 11.2 M 减少至 1.2 M 减少了 89%，每次推理的 FLOPs 从 1.648 G 降低到了 0.185 G，降低了 89%。它在合成数据上的 3D-AUCp 也保持了 0.84 的竞争水平。这些结果展示了准确且高效的基于事件的第一人称视角手部追踪，适合于设备端 XR 应用。该数据集和代码可在 https://github.com/zen5x5/EvHand-FPV 获得。**

I. 介绍

手部追踪已成为自然人机交互的核心技术，改变了用户与数字环境互动的方式。在扩展现实（XR）应用中，准确的手部追踪消除了对物理控制器的需求，通过自然手势实现虚拟对象的直观操作。这项能力已集成到商用 AR/VR 设备如 Meta Quest [1] 中，并且也正在向机器人遥操作系统 [2]、手语识别 [3] 和医疗康复 [4] 等领域扩展。这些应用不仅需要对具有高度自由度的手部姿势进行精确估计，还需要实现实时追踪并满足极低延迟的要求，在快速手部动作的情况下这一点尤为挑战，此时低延迟和高时间保真度至关重要。

当前的手部跟踪解决方案主要依赖于传统的基于帧的相机 [5]–[8]，这些相机以固定的帧率捕获整个图像，而不考虑场景动态。这种方法对于资源受限设备来说存在局限性。典型的 2D RGB 传感器在 25 – 30 FPS 时消耗约 200 毫瓦，而 3D 深度系统则需要在类似帧率下消耗 3 – 5 瓦 [9]。尝试通过降低帧率来减少功率消耗 [10] 严重损害了跟踪响应性，在快速手部移动时导致运动模糊和时间信息丢失，而这正是准确跟踪最为关键的时刻。

事件相机通过异步捕获亮度变化，以 μs 级的时间分辨率和 mW 级的功耗来缓解这些限制。与传统相机不同，这些受生物启发的传感器异步仅捕获像素级别的亮度变化，实现了微秒级的时间分辨率，同时消耗低至 10 mW [11]–[14]。这种稀疏、事件驱动的感知机制消除了运动模糊，并为追踪快速移动提供了卓越的时间保真度。最近基于事件的手部跟踪系统展示了有希望的结果 [15]–[18]，验证了这项技术的潜力。

然而，现有的基于事件的方法未能充分利用事件相机的效率优势，采用了参数超过 10M 的深度学习模型，这带来了显著的计算负担。这种不匹配阻碍了在 AR/VR 头戴设备和眼镜上的部署，在这些设备中，计算能力和电源预算是高度受限的。此外，这些方法仅关注第三人称视角，忽视了对于虚拟键盘输入、AR 界面操作和机器人远程操控等交互应用至关重要的第一人称（以自我为中心）手部追踪。[19]–[21]

为了解决这些挑战，我们提出了**手部 EV-第一人称视角**：一个 Efficient event-based 3D手 跟踪框架来自 First-PersonView。我们的方法重新思考了基于事件的手部跟踪，以实现实用部署，在减少模型参数和计算负载达 89%的同时，实现了优于现有最先进技术（SOTA）的精度。我们证明了通过仔细设计数据表示、网络架构和训练策略，可以实现高效且准确的自我中心手部跟踪。

本工作的主要贡献如下：

- 我们提出了一种轻量且准确的基于事件的框架，用于从第一人称视角进行三维手部追踪。
- 我们构建了一个基于事件的第一人称视角手部追踪数据集，其中包括带有 3D 标签的合成训练数据以及首次出现的带 2D 标签的真实测试事件数据，解决了自中心视角的关键数据缺口问题 [22]。
- 我们介绍了一种基于手腕的 ROI 方法，采用端到端

*Zhen Xu and Guorui Lu contribute equally. Zhen Xu, Guorui Lu, Qinyu Chen are with the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands. {z.xu.11, g.lu, q.chen}@liacs.leidenuniv.nl

Chang Gao is with Department of Microelectronics, Delft University of Technology, The Netherlands.

映射策略，将 ROI 偏移嵌入网络中，从而减少输入大小并实现无需显式重建的有效对齐。

- 我们设计了一种多任务学习策略，该策略包含一个辅助的几何特征预测任务，引导网络朝向更具判别性的表示方向发展，在不增加推理阶段任何额外开销的情况下提高准确性。

II. 相关工作

A. 三维手部重建

从视觉数据中重建 3D 手部已经得到了广泛的研究，方法大致可以分为两类范式。第一类采用端到端的深度学习直接将图像空间映射到手部姿态空间 [5]–[7], [23]。这些方法利用深度神经网络的强大能力进行高效的推理，并展示了有希望的表现。然而，它们通常缺乏几何约束，并且需要大规模标注数据集进行训练，限制了其泛化能力。第二类则结合预先定义的参数化手部模型作为先验知识，以确保生理上合理的重建结果。如 MANO [24] 和 SMPL [25] 这样的模型提供了强有力的结构约束，改善了重建的手部的真实性，特别是在训练数据有限或存在遮挡的场景中。

当前的方法主要依赖于传统的成像模态，包括 RGB 摄像头 [5]–[7], [23], [26]–[31] 和深度传感器 [32]–[37]。虽然这些技术得益于成熟的硬件和大量的数据集，但它们面临着根本性的限制：低时间分辨率（通常为 30-60 FPS），无法捕捉快速的手部运动，快速移动时显著的运动模糊，以及随着帧率增加而大幅提高的高能耗，使得它们不适合资源受限的边缘设备和高速跟踪应用。我们的任务专注于手部追踪，即实时连续估算手部姿态。由于我们采用 MANO 模型来表示手部，因此我们的方法也与 3D 手部重建方法相关。

B. 基于事件的手部跟踪

事件相机代表了视觉传感的一种范式转变，它们以异步方式捕捉像素级别的亮度变化，而不是完整的帧 [14]。这种仿生方法提供了微秒级的时间分辨率、固有的运动模糊减少、高动态范围（超过 140 dB）以及超低功耗（毫瓦级别），使其特别适合高速运动捕捉和边缘计算应用。

1) *第三人称视角方法*：最近的进展成功地将事件相机应用于控制下的第三人称视角中的 3D 手部追踪。EventHands [16] 通过直接从事件流中回归 MANO 参数，开创了纯基于事件的手部追踪，使用局部归一化事件表面（LNES）表示。该方法在 1 kHz 下实现了实时性能，并展示了出色的合成到现实的泛化能力，尽管仅在合成数据上进行训练。然而，它需要超过 11 M 参数并且只能重建单个手部。

EvHandPose [18] 通过引入手部流表示和弱监督框架解决了领域差距挑战，并附带了一个大规模的真实世界事件数据集。虽然这减少了对合成数据的依赖，但计算需求仍然很大。Ev2Hands [15] 扩展了双手机器重建的能力，使用点云表示，但同样需要大量的计算资源，模型参数超过 10M。

薛等人提出了一个基于几何的替代方案 [15]，提供了一个优化框架，消除了对大规模标注的需求，并提供了强大的可解释性。然而，其迭代优化过程限制了实时性能，制约了其在在线跟踪场景中的适用性。

2) *第一人称视角方法*：尽管取得了这些进展，现有的基于事件的方法仅限于静态的第三人称摄像机设置，忽视了第一人称视角的研究。一方面，当前的数据集缺乏足够的自我中心数据；另一方面，这类视角通常出现在资源受限的平台如 VR 头显上，在这种平台上模型大小和计算成本是关键考虑因素。然而，先前的基于事件相机的手部追踪研究大多忽略了这些问题。

EventEgoHands [22] 近期开创了基于事件的 3D 手部网格重建，引入了一个手部分割模块，该模块采用 U-Net 从 LNES 表示中提取手部区域，过滤掉由相机运动引起的背景事件。他们还创建了 N-HOT3D，这是一个大型合成数据集，包含 447K 样本，专门设计用于第一人称视角，该数据集通过使用 v2e 模拟器 [38] 从 HOT3D 数据集中生成。尽管 EventEgoHands 成功展示了基于事件的手部追踪的可行性，但它保留了现有方法所具有的计算复杂度特征。我们的 EvHand-FPV 框架在第一人称视角挑战的见解基础上构建，特别针对资源受限部署进行优化，引入了轻量级替代方案，在大幅降低计算需求的同时保持准确性。

III. 方法

A. 数据集

1) *合成数据*：我们使用了事件模拟 [16]，它基于 MANO 手部模型 [24]，来生成 3D 右手序列。从每个配置好的视点，模拟器产生同步的 RGB 帧、事件流、3D 关节坐标以及它们的 12 维主成分分析（PCA）转换标签，所有这些都以 μs 精度的时间戳存储。为了增加数据多样性，我们改变了相机视点并应用了随机的手势变换、平移和旋转。总共，我们生成了 720,000 ms 的合成数据用于训练，以及 60,000 ms 用于测试。如图 1(a) 左列所示，该数据集包括带有时间戳的 RGB 帧、事件流、关节坐标以及相应的 PCA 标签。

2) *真实世界数据*：我们收集最多 60 秒的真实数据，使用与 30 FPS 的 RGB 同步的 DAVIS346 事件相机。我

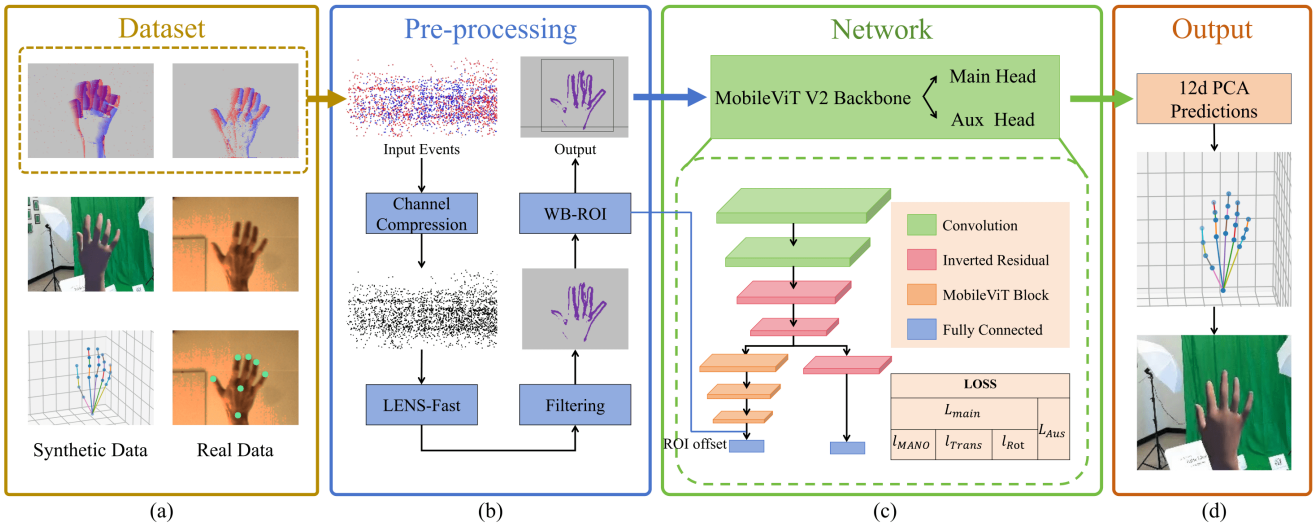

Fig. 1. **概述**: (a) **数据集**: 合成数据由 MANO 手部模型使用仿真生成，以及使用事件相机收集的真实世界数据。(b) **预处理**: 事件流由通道压缩、LNES-Fast 事件表示和噪声滤波处理。然后应用基于手腕的 ROI 方法来定位手部区域并裁剪输入。(c) **轻量级多任务网络架构**: 裁剪后的 ROI 及其偏移量被馈入 MobileViT V2 主干网络。该网络遵循多任务设计，具有一个主预测头（手部姿势）和一个辅助头（事件分布的几何统计）。(d) **输出**: 12 维预测包括 MANO PCA 分量、平移和旋转，这些被重建为 3D 手部关节并渲染成网格用于可视化。

们的真实数据持续时间达到 60,000 毫秒，允许测试实际性能稳定性。为了确保方法的通用性，从具有不同手形和尺寸的受试者中收集数据，并且记录的手部运动涵盖了目标场景中的常见运动模式，如平面平移、深度移动、手腕旋转和手势变化。我们使用 MediaPipe [39] 初始标注 RGB 帧中手关节的 2D 坐标，并手动校准标签。图 1(a) 的右栏显示了真实世界的手部数据，包括 RGB 数据、事件数据以及手关节的标注坐标。

B. 事件表示

事件相机生成的数据格式为 $e(t,x,y,p)$，其中 t 表示时间戳，x 和 y 代表空间坐标，而 p 表示事件极性。$p=0$ 表示一个负面事件，表明该时刻像素变暗，而 $p=1$ 表示一个正面事件，表明像素变亮。我们通过三个主要步骤处理事件，如算法 1 所总结的：通道压缩、事件积累和噪声过滤。

1) *信道压缩*: 在采用多通道表示 [16] 时，为了泛化性需要进行极性增强，因为事件极性受到亮度 [14] 的影响，即背景变化可能导致相同运动下的不同事件极性。在这个项目中，为了降低复杂度，我们将极性通道压缩成单个通道，在此通道中，像素处出现任意一种极性即可标记为一个事件。如果在同一个 bin 内的同一像素上发生两种极性，则强度不会加倍。我们的实验确认这种简化不会影响性能。

2) *事件积累*: 为了将事件数据准备为神经网络的输入，事件被划分为时间窗并累积成表示形式。在我们的 EvHand-FPV 框架中，我们使用固定时间窗口的方法来保证最小频率。已经提出了各种方法来将时间窗中的事件累积成表示形式，例如 EventHands [16] 中使用的 LNES。LNES 在一个固定的 100 毫秒时间窗口（99 毫秒重叠）内聚合事件，并将其归一化为一个 2 通道图像，每个通道对应一种极性，实现了 1 千赫的高时间分辨率。然而，固定的时间窗大小可能导致信息冗余或过度稀疏。在快速手部运动期间，在短时间内生成了大量的事件，足以形成丰富的图像。在这种情况下，继续添加早期事件只会模糊表示形式。相反，当手部缓慢移动并产生少量事件时，缩短时间窗口可能会导致表示过于稀疏。

为了解决这个问题，我们引入 LNES-Fast，在保持原始时间窗口长度约束的同时，为合并事件的数量添加一个上限。在进行事件堆叠时，我们不再遍历完整的时间窗口，而是同时监控经过的时间和累积事件计数，并在计数超过预定义阈值时提前终止。该策略避免了在快速运动中过度模糊，同时仍然遍历足够的时间以防止稀疏性，从而实现缓慢运动。因此，LNES-Fast 在不牺牲表示质量的情况下减少了冗余和计算开销。

3) *噪声过滤*: 为了减轻事件堆叠过程中积累的噪声，我们利用了虚假事件通常作为孤立异常点出现这一观察结果；因此，我们应用高斯模糊来抑制孤立的虚假事件并平滑表示。这种方法仅作用于当前帧，消除了调用前一帧和后一帧信息进行过滤的需要，从而降低了时间维度上的计算复杂度。

Algorithm 1 LNES -快速通道压缩和噪声过滤

Input: $E = \{e_i \mid i = 1 \cdots n\}$: 事件 bin;
t_0: 当前时间戳;
L: 窗口大小;
θ: 计数限制;
k, σ: 高斯模糊核和方差。
Output: 去噪单通道表示 $\tilde{\texttt{img}}$

信道压缩：将同一像素的正/负事件合并为一个事件
初始化 $\texttt{events\_cnt} \leftarrow 0$
初始化 $\texttt{img} \leftarrow 0$
for e_i in E do
  if $events\_cnt \geq \theta$ then
    中断
    #↑ 快速早期停止
  end if
  for each $e_i \in E$ do
    $\texttt{img}[y,x] \leftarrow \max(\texttt{img}[y,x], \frac{L-(t_0-t_i)}{L})$
    # ↑ 窗口归一化权重
    $\texttt{events\_cnt} \leftarrow \texttt{events\_cnt} + 1$
  end for
end for
$\tilde{\texttt{img}} \leftarrow \text{GaussianBlur}(\texttt{img}, \texttt{kernel} = k, \sigma)$
#↑ 噪声过滤
返回 $\tilde{\texttt{img}}$

C. 基于手腕的 ROI 和轻量级端到端映射

事件数据本质上包含可以利用的空间坐标信息，以高效地估计目标的近似位置。这使得可以使用感兴趣区域 (ROI)，通过仅关注裁剪后的区域来减少计算负载。一种常见的策略是预先定义多个候选区域，并在输入 [40] 时选择事件密度最高的一个。

1) 脱部定位：我们的方法是基于解剖学观察，手腕是手臂和手之间最狭窄的连接处。如图 2 所示，我们通过确定其垂直坐标 Y 以及左右水平边界 X_L 和 X_R 来定位手腕。这些变量 Y,X_L 和 X_R 初始化为 $Y = H-1$,$X_L = 0$ 和 $X_R = W-1$，其中 H 和 W 表示事件帧的高度和宽度。随着 Y 的减小，宽度 $X_R - X_L$ 的减少表明收敛于手腕。一旦宽度开始再次扩大，即 X_L 减少而 X_R 增加，这表明我们远离了手腕，并终止搜索过程。为了对抗噪声的鲁棒性，在确定每行中最左和最右事件坐标时应用了一个阈值，防止孤立事件引起的误检测。

2) ROI 构建：定位手腕后，我们构建一个以手腕中点 (X_c, X_c) 为中心的 ROI 框，其中 $X_c = (X_L + X_R)/2$ 和 $Y_c = Y$。给定预定义的 ROI 大小 $h \times w$，垂直范围设置为 $[Y_c - (h-10), Y_c + 10]$，保留手腕下方 10 个像素，并将剩余像素向上分配。水平范围定义为 $[Y_c - w/2, Y_c + w/2]$。

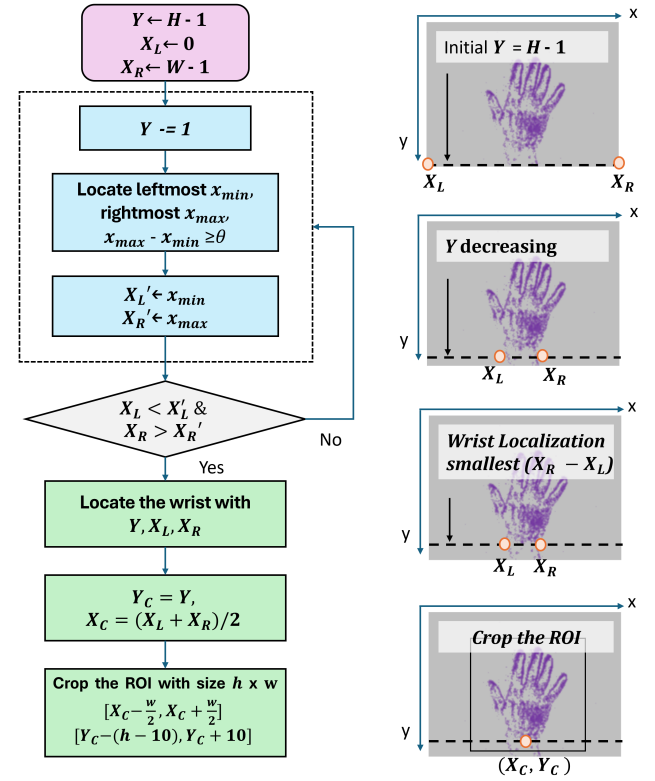

Fig. 2. **基于手腕的** ROI: 流程图和插图。

3) 轻量级端到端映射：虽然 ROI 减少了计算复杂性，但将其与基于 PCA 的输出集成起来却是一个挑战：明确地将 ROI 预测重建到原始坐标空间中是代价高昂的。为了解决这个问题，我们引入了一种轻量级的端到端映射策略。将 ROI 偏移（即 ROI 框左上角的坐标）与展平后的特征连接，并传递给全连接层，使模型能够隐式学习这种变换。这样的设计避免了显式的重建过程，并为 ROI 到原始空间的对齐提供了一种新颖且高效的解决方案。

D. 轻量级多任务学习网络架构

整个网络的架构如图 1(c) 所示。选定的 ROI 被作为第一层输入送入网络，而原始图像中的 ROI 偏移量则被送入主任务的最后一全连接层，以恢复在原始空间系统中的预测结果。ROI 特征图首先通过一个基本特征提取层，该层包含两个 2D 卷积层和 MobileViT V2 [41] 中的两个倒残差层。随后，它经过两条并行分支：一条辅

助（AUX）任务分支，由一个倒残差层、全局平均池化和一个线性层组成；另一条主任务分支，由 MobileViT V2 块、全局平均池化和两个线性层组成。

1) *轻量级骨干选择*：先前的研究采用了大量参数的模型，例如拥有超过 11M 个参数 [42] 作为其骨干网络 [16], [18] 。然而，我们的目标是设计一个轻量级系统，这些模型不符合我们的需求。因此，我们仅保留 ResNet-18 作为基准，并专注于为资源受限设备专门设计的轻量级模型，包括 ShuffleNet V2 [43] 、MobileNet V3 [44] 和 MobileViT V2 [41] 。最终，我们选择 MobileViT V2 作为我们的模型骨干，因为它在可比参数规模下实现了最快的收敛速度和最佳性能，并避免了与其它 ViT 模型相关的过度计算开销。

为了进一步降低计算复杂度并使模型更易于硬件部署，我们简化了模型内的两个组件。首先，我们将 SiLU 激活函数替换为简化的 ReLU，这在计算上对硬件更为友好。其次，我们在线性自注意力模块中用泰勒级数展开近似替换了 Softmax 函数。实验验证表明，这两种优化措施不会导致准确率下降。

骨干网络的输出是 12 维的，由三个部分组成：前 6 个维度表示 MANO 模型的 PCA 成分，第 7 到第 9 个维度对应手部的三维平移，最后 3 个维度编码手部旋转。

2) *多任务学习*：为了提高准确性，我们设计了一个多任务学习架构。我们发现像手心点、标准差和 ROI 的协方差矩阵等参数与手的 3D 空间位置密切相关。基于这一观察，我们在主干模型的中间层引入了额外的输出头，使用这些参数作为标签。这个辅助任务鼓励模型的较低层和中间层学习更有意义的特征，这有助于提高性能。此外，由于在推理过程中可以移除辅助分支，因此在部署时不会增加额外的计算开销。

辅助任务的标签和输出具有七个维度，包括以下信息的归一化值：事件坐标在 x 和 y 轴上的均值和标准差、协方差矩阵的两个特征值以及与最大特征值对应的特征向量的方向角。

E. 损失

总损失由主任务和辅助任务的损失组成。总损失函数如下所示：

$$\text{Loss}_{\text{Total}} = \text{Loss}_{\text{Main}} + w_{\text{Aux}} \times \text{Loss}_{\text{Aux}} \tag{1}$$

其中，辅助任务损失 w_{Aux} 的权重设置为 0.5，以防止辅助任务干扰主任务的优化方向。

1) *主要任务损失*：我们分别为主任务中的三个组件计算单独的损失项，即 MANO、平移和旋转，并在聚合整体主任务损失时为它们分配不同的权重。在确定权重时，我们既考虑保持这三个损失组件之间的量级相当，又考虑到它们对整体性能的不同贡献。根据实验结果，我们采用以下加权方案：

$$\begin{aligned}\text{Loss}_{\text{Main}} = (&6 \times w_{\text{MANO}} \times l_{\text{MANO}} \\ &+3 \times w_{\text{Trans}} \times l_{\text{Trans}} \\ &+3 \times w_{\text{Rot}} \times l_{\text{Rot}})/12\end{aligned} \tag{2}$$

其中 $w_{\text{MANO}} = 10, w_{\text{Trans}} = 10000, w_{\text{Rot}} = 20$。

2) *辅助任务损失*：对于辅助任务，我们计算了所有七个组件的总均方误差、事件坐标在 x 轴和 y 轴上的平均值和标准差、两个特征值以及事件分布特征椭圆的方向角，且未对每个组件分配不同的权重。

IV. 实验

我们在真实世界基于事件的手部数据上评估了我们的 EvHand-FPV 方法，包括与 SOTA 工作的比较和消融研究以评估各个组件的贡献。我们使用 PyTorch Lightning [45] 在单个 NVIDIA RTX 3090 GPU 上训练模型。我们将批处理大小设置为 32，训练周期设为 20。优化器是 Adam，学习率为 1×10^{-4}。

A. 评估指标

1) *二维度量*：PCK 指标，指的是根对齐的关键点正确率百分比，在手部追踪工作中通常采用此指标 [28] 。为了考虑尺度变化，我们采用了手掌归一化的 2D-PCK（2D-PCKp）及其对应的曲线下的面积（2D-AUCp）[16] 。神经网络模型的预测结果被输入到事件模拟中，基于 MANO 模型计算 3D 手部关节位置 [24] ，然后将这些关节投影到 2D 图像平面上以计算 2D-PCKp 和 2D-AUCp。

2) *3D 度量*：通过仿真.ev 软件转换的预测输出的 3D 坐标可以直接与 3D 真实标签进行比较，以计算 3D-PCK 和 3D-AUC。然而，由于获取现实世界数据的 3D 标注固有的困难，我们对该指标的评估仅限于合成数据。

B. 与先前工作比较

图 3 和表 I 展示了我们的方法与第三人称事件驱动手部追踪领域的 SOTA 方法 EventHands [16] 的比较。EventHands 在其原始的第三人称设定中，在真实数据上达到了 0.77 2D-AUCp，而在合成数据上达到了 0.85 3D-AUC。然而，当应用于我们的第一人称视角（自中心）任务时，其性能急剧下降到在真实数据上的 2D-AUCp 仅为 0.12，在合成数据上的 3D-AUC 为 0.17，这显示了视点之间的显著领域差异以及 EvHand-FPV 的必要性。为了公平起见，我们也重新训练了 EventHands 中使用的

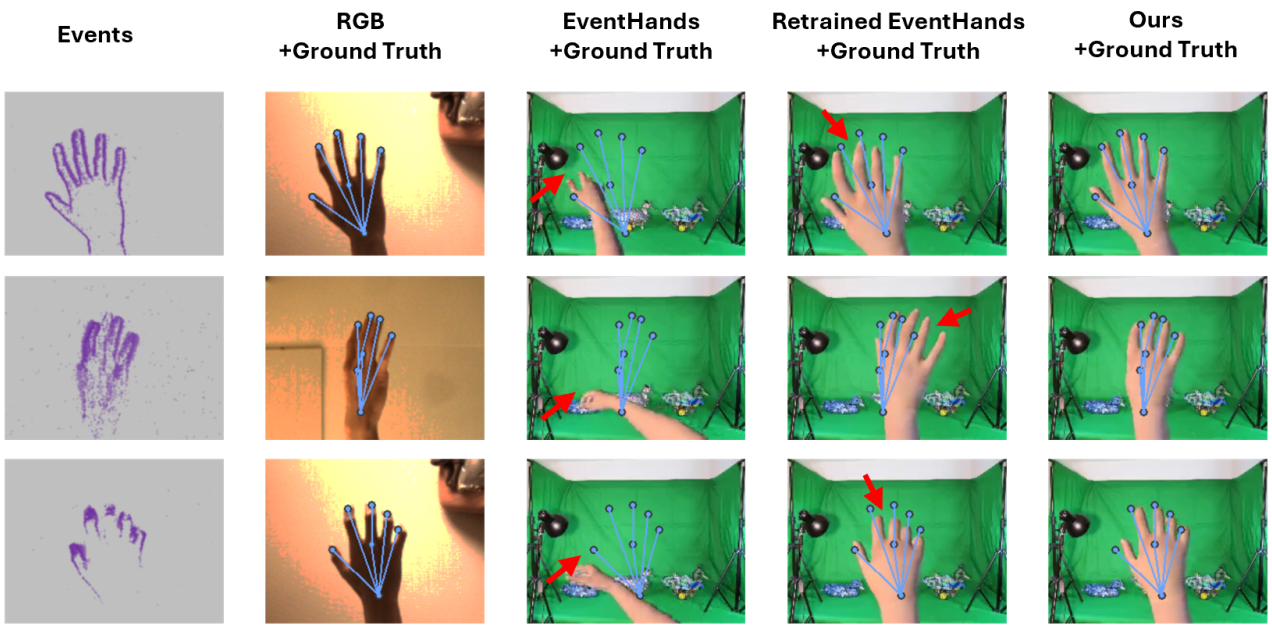

Fig. 3. **定性评估**。我们将我们的方法与 EventHands [16] 在我们的实际 FPV 测试集上进行了比较。预测的手部姿态以渲染的网格形式显示，而真实标注则用蓝色线条和关节表示。红色箭头指示明显的失败部分。RGB 图像未作为输入使用，此处仅作为参考。

TABLE I

与 EventHands 的比较 [16]

方法	**数据集**	2D-AUCp (Real Data) ↑	3D-AUCp (Synthetic Data) ↑	**参数** ↓	**浮点运算数** ↓
EventHands	EventHands	0.77	0.85	11.2 M	1.648 G
EventHands (w/o retrain)	EvHand-FPV	0.11	0.17	11.2 M	1.648 G
EventHands (retrained)	EvHand-FPV	0.77	0.82	11.2 M	1.648 G
EvHand-FPV (Ours)	EvHand-FPV	0.85	0.84	1.2 M	0.185 G

ResNet-18 骨干网络在我们的数据集上，结果达到了在真实数据上的 2D-AUCp 为 0.77，在合成数据上的 3D-AUC 为 0.82，这证实了我们数据集的难度与先前的工作相当。

我们的 EvHand-FPV 在真实数据上实现了 0.85 的 2D-AUCp，在合成数据上实现了 0.84 的 3D-AUC，同时将模型大小从 11.2 M 参数减少到 1.2 M 参数，并将计算成本从 1.648 G FLOPs 减少到 0.185 G FLOPs，这两个指标都减少了 89%，这在轻量级约束下的事件驱动的第一人称手部追踪中达到了 SOTA 水平。

C. 消融研究

1) *骨干结构*: 我们首先评估不同的轻量级骨干网络，以确定最适合我们的框架的架构。具体来说，我们将 ShuffleNet V2 [43], MobileNet V3 [44], 和 MobileViT V2 [41] 与在 EventHands 中使用的 ResNet-18 进行了比较。它们的性能、参数数量和计算负载详见表 II。这些结果是基于原始图像输入且没有辅助任务分支获得的。ShuffleNet V2 和 MobileNet V3 非常紧凑，但显示出有限的准确性（分别为 0.78 和 0.76 的 2D-AUCp）。MobileViT V2 在相同的参数预算下实现了最佳准确性（0.80 2D-AUCp），得益于其基于变换器的设计及线性注意力机制。最终，我们选择了 MobileViT V2 作为提出方法的骨干网络。为了公平比较，在这些实验中，我们将这些轻量级模型修剪到相同数量的参数（0.1M）。然而，后续的实验使用官方发布的 MobileViT V2（1.2M）来实现更好的性能。

2) *架构修改*: 虽然 MobileViT V2 中的 SiLU 激活函数具有平滑性和非单调性等优点，但在硬件（例如 FPGA 或 ASIC）上的实现带来了显著的挑战 [46], [47]。考虑到这一点，并且承认事件图像本质上比 RGB 图像更简单，我们将 MobileViT V2 内的所有激活函数替换为相对简单的 ReLU。实验表明这种替换不会导致性能损失。另一

TABLE II

骨干网络之间的比较

模型	2D-AUCp ↑	参数 ↓	浮点运算次数 ↓
ResNet-18	0.77	11.2 M	1.648 G
ShuffleNet V2 (Pruned)	0.78	0.1 M	0.025 G
MobileNet V3 (Pruned)	0.76	0.1 M	0.011 G
MobileViT V2 (Pruned)	0.80	0.1 M	0.023 G

TABLE III

基于手腕的 ROI 和多任务学习的消融研究。

投资回报率	多任务	2D-AUCp↑	浮点运算次数 ↓
No	No	0.85	0.322 G
Yes	No	0.84	0.185 G
No	Yes	0.87	0.322 G
Yes	Yes	0.85	0.185 G

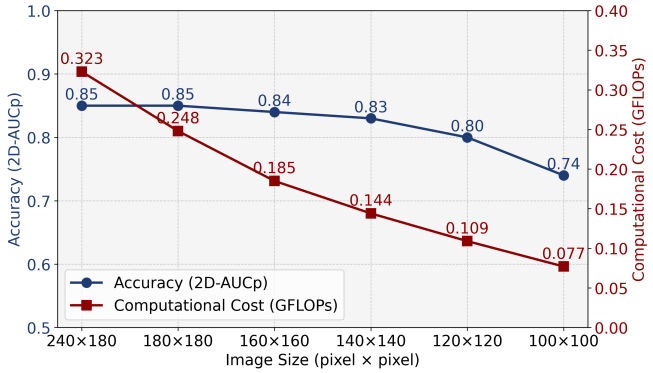

Fig. 4. ROI 大小对准确性（2D-AUCp）和计算成本（GFLOPs）的影响。大幅降低分辨率显著减少了 FLOPs，同时保持了直到 160×160 的竞争力准确性。

个计算密集型组件是 softmax 函数。为了适应轻量级系统，我们用泰勒级数展开近似替换了原始的 softmax 函数。实验表明这种替换也不会导致性能损失。这些修改使得 MobileViT V2 更适合轻量级部署而不牺牲准确性。

3) ROI 和多任务学习：图 4 说明了在不同感兴趣区域大小下，我们的方法性能和计算成本的变化。仅以精度下降 0.01（从 0.85 降到 0.84）为代价，使用 160×160 的感兴趣区域大小可将计算成本降低 42.72%。因此，我们采用了这个尺寸作为我们的方法。

我们在 MobileViT-V2 0.5× 主干上进行比较实验，以考察 ROI 和我们的辅助任务的影响。如表 III 所示，添加辅助任务在有无 ROI 的情况下都提高了性能，而 ROI 显著降低了计算成本，但准确率略有下降。在推理过程中会剪枝辅助分支，因此其参数和计算不计入部署成本中。ROI 与辅助任务的结合对于实现高准确率和低计算成本都是必不可少的。

V. 结论

在这项工作中，我们提出了 EvHand-FPV，一个用于从第一人称视角使用事件相机进行 3D 手部跟踪的高效轻量级框架。通过引入基于手腕的 ROI 检测、LNES-Fast 表示和优化后的 MobileViT V2 多任务学习网络架构，我们的方法在真实 FPV 测试集上实现了 0.85 的 2D-AUCp，并且与 EventHands 相比，参数数量和计算成本减少了 89%。这些结果表明，在严格的资源限制下实现准确的实时自我中心手部跟踪是可行的，这使得我们的框架非常适合 AR/VR 头戴设备和其他可穿戴平台。

References

[1] Meta Platforms, Inc., “Meta quest 3,” https://www.meta.com/quest/quest-3/, 2023, accessed: 2025-08-22.

[2] R. Li, H. Wang, and Z. Liu, “Survey on mapping human hand motion to robotic hands for teleoperation,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2647–2665, 2022.

[3] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign language transformers: Joint end-to-end sign language recognition and translation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[4] M. Bressler, J. Merk, T. Gohlke, F. Kayali, A. Daigeler, J. Kolbenschlag, and C. Prahm, “A virtual reality serious game for the rehabilitation of hand and finger function: Iterative development and suitability study,” JMIR Serious Games, vol. 12, p. e54193, Aug 2024.

[5] C. Zimmermann and T. Brox, “Learning to estimate 3d hand pose from single rgb images,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 10 2017.

[6] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, “Hand pose estimation via latent 2.5d heatmap regression,” in Proceedings of the European Conference on Computer Vision (ECCV), 9 2018.

[7] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, “3d hand shape and pose estimation from a single rgb image,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6 2019.

[8] J. Liebers, S. Brockel, U. Gruenefeld, and S. S. and, “Identifying users by their hand tracking data in augmented and virtual reality,” International Journal of Human – Computer Interaction, vol. 40, no. 2, pp. 409–424, 2024.

[9] A. Colaco, A. Kirmani, N.-W. Gong, T. Mcgarry, L. Watkins, and V. K. Goyal, “3dim: Compact and low power time-of-flight sensor for 3d capture using parametric signal processing,” in Proc. Int. Image Sensor Workshop. Citeseer, 2013, pp. 349–352.

[10] D. Rotman, O. Cohen, and G. Gilboa, “Frame rate reduction of depth cameras by rgb-based depth prediction,” in 2016

IEEE International Conference on the Science of Electrical Engineering (ICSEE), 2016, pp. 1–5.

[11] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," IEEE Journal of Solid-State Circuits, vol. 43, no. 2, pp. 566–576, 2008.

[12] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds," IEEE Journal of Solid-State Circuits, vol. 46, no. 1, pp. 259–275, 2011.

[13] B. Son, Y. Suh, S. Kim, H. Jung, J.-S. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo, Y. Roh, H. Lee, Y. Wang, I. Ovsiannikov, and H. Ryu, "4.1 a 640×480 dynamic vision sensor with a 9µm pixel and 300meps address-event representation," in 2017 IEEE International Solid-State Circuits Conference (ISSCC), 2017, pp. 66–67.

[14] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 154–180, 2022.

[15] Y. Xue, H. Li, S. Leutenegger, and J. Stückler, "Event-based non-rigid reconstruction from contours," 2022. [Online]. Available: https://arxiv.org/abs/2210.06270

[16] V. Rudnev, V. Golyanik, J. Wang, H.-P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt, "Eventhands: Real-time neural 3d hand pose estimation from an event stream," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10 2021, pp. 12 385–12 395.

[17] C. Millerdurai, D. Luvizon, V. Rudnev, A. Jonas, J. Wang, C. Theobalt, and V. Golyanik, "3d pose estimation of two interacting hands from a monocular event camera," in 2024 International Conference on 3D Vision (3DV), 2024, pp. 291–301.

[18] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi, "Evhandpose: Event-based 3d hand pose estimation with sparse supervision," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 9, pp. 6416–6430, 2024.

[19] J. Grubert, L. Witzani, E. Ofek, M. Pahud, M. Kranz, and P. O. Kristensson, "Effects of hand representations for typing in virtual reality," in 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2018, pp. 151–158.

[20] W. Lin, L. Du, C. Harris-Adamson, A. Barr, and D. Rempel, "Design of hand gestures for manipulating objects in virtual reality," in Human-Computer Interaction. User Interface Design, Development and Multimodality, M. Kurosu, Ed. Cham: Springer International Publishing, 2017, pp. 584–592.

[21] Y. Qin, W. Yang, B. Huang, K. V. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," 2024. [Online]. Available: https://arxiv.org/abs/2307.04577

[22] R. Hara, W. Ikeda, M. Hatano, and M. Isogawa, "Eventego-hands: Event-based egocentric 3d hand mesh reconstruction," in 2025 IEEE International Conference on Image Processing (ICIP), 2025, pp. 1199–1204.

[23] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular rgb image," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10 2019.

[24] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," ACM Transactions on Graphics, vol. 36, no. 6, p. 1 – 17, Nov. 2017. [Online]. Available: http://dx.doi.org/10.1145/3130800.3130883

[25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: a skinned multi-person linear model," ACM Trans. Graph., vol. 34, no. 6, Oct. 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[26] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 666–682.

[27] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 89–98.

[28] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2018.

[29] L. Yang, S. Li, D. Lee, and A. Yao, "Aligning latent spaces for 3d hand pose estimation," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2335–2343.

[30] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly supervised 3d hand pose estimation via biomechanical constraints," in Proceedings of the European conference on computer vision (ECCV). Springer, 2020, pp. 211–228.

[31] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo, "Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20 544–20 554.

[32] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 680–689.

[33] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7113–7122.

[34] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge et al., "Depth-based 3d hand pose estimation: From current achievements to future goals," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2636–2645.

[35] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8417–8426.

[36] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-

voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in Proceedings of the IEEE conference on computer vision and pattern Recognition, 2018, pp. 5079–5088.

[37] L. Fang, X. Liu, L. Liu, H. Xu, and W. Kang, "Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image," in Proceedings of the European conference on computer vision (ECCV). Springer, 2020, pp. 120–137.

[38] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2021, pp. 1312–1321.

[39] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," 2020. [Online]. Available: https://arxiv.org/abs/2006.10214

[40] S. Tan, J. Yang, J. Huang, Z. Yang, Q. Chen, L. Zheng, and Z. Zou, "Toward efficient eye tracking in ar/vr devices: A near-eye dvs-based processor for real-time gaze estimation," IEEE Transactions on Circuits and Systems I: Regular Papers, pp. 1–13, 2025.

[41] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," 2022. [Online]. Available: https://arxiv.org/abs/2206.02680

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[43] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," 2018. [Online]. Available: https://arxiv.org/abs/1807.11164

[44] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. [Online]. Available: https://arxiv.org/abs/1905.02244

[45] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019, version 1.4, Apache-2.0 License. [Online]. Available: https://github.com/Lightning-AI/lightning

[46] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.03118

[47] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017. [Online]. Available: https://arxiv.org/abs/1710.05941