动态相机姿态及其寻找方法

Chris Rockwell[1,2] Joseph Tung[3] Tsung-Yi Lin[1]
Ming-Yu Liu[1] David F. Fouhey[3] Chen-Hsuan Lin[1]

NVIDIA[1] University of Michigan[2] New York University[3]

https://research.nvidia.com/labs/dir/dynpose-100k

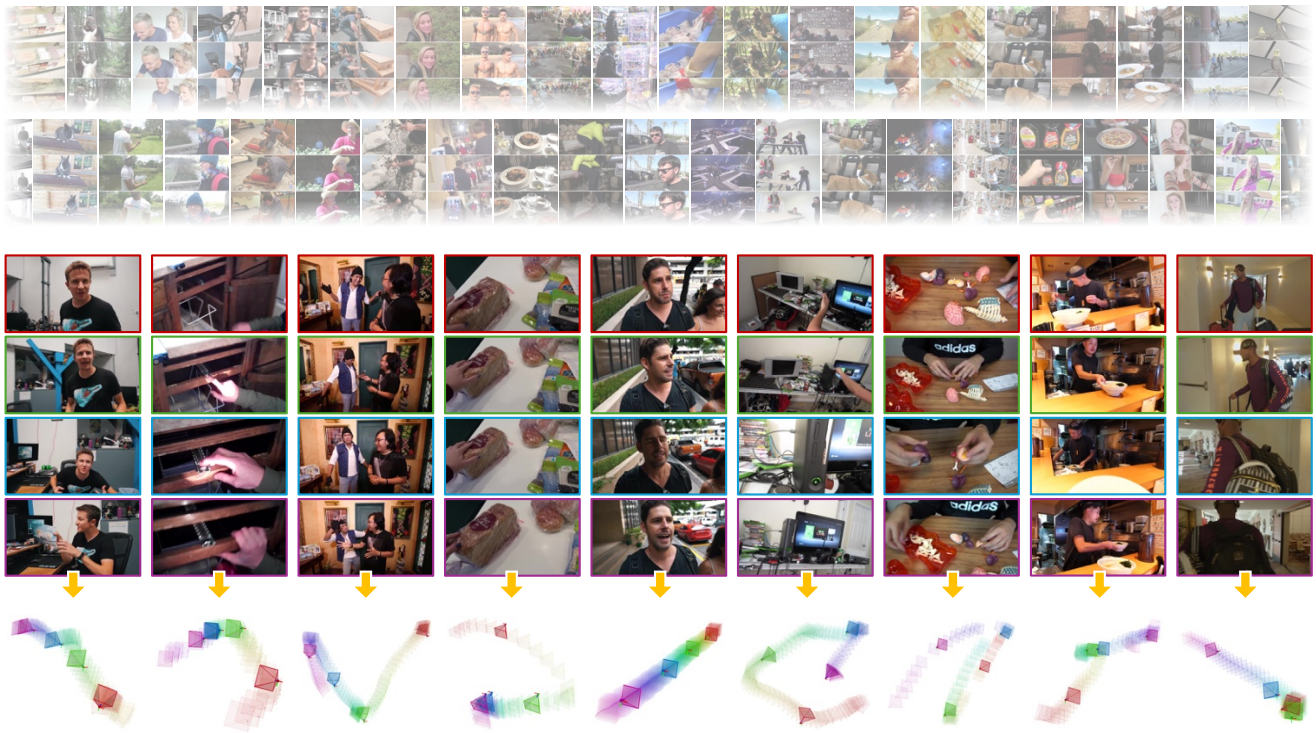

Figure 1. We introduce DynPose-100K, a large-scale video dataset of dynamic content with camera annotations. DynPose-100K consists of 100,131 Internet videos that span diverse settings. We curate DynPose-100K such that videos contain dynamic content while ensuring the cameras are able to be estimated (including intrinsics and poses). Towards this end, we address two challenging problems: (a) identifying the videos suitable for camera estimation, and (b) improving the camera estimation algorithm for dynamic videos.

Abstract

在大规模动态互联网视频中标注相机姿态对推动如逼真视频生成和模拟等领域的发展至关重要。然而，收集这样一个数据集是困难的，因为大多数互联网视频不适合姿态估计。此外，对动态互联网视频进行标注即使对于最先进的方法来说也是一个重大挑战。本文中，我们介绍了一个大规模动态互联网视频数据集，该数据集标注了相机姿态。我们的收集流程通过精心结合的任务特定和通用模型集解决了过滤问题。对于姿态估计，我们结合了最新的点跟踪、动态掩膜和由运动恢复结构技术，以实现对当前最新方法的改进。我们的分析和实验表明，*DynPose-100K* 数据集在多个关键属性上既具有大规模性又具有多样性，为多种下游应用的进步开辟了途径。

1. 介绍

使用摄像机信息标注大规模动态互联网视频有可能推动计算机视觉和机器人领域的许多问题的发展。这种

www.xueshuxiangzi.com

数据集可以为生成模型提供动力，创建摄像机控制的动态视频和四维场景。它也可以用于大规模视图合成模型的训练，以应用于扩展现实。此外，这些数据在机器人领域可能成为变革性的，适用于诸如模仿学习或在逼真的模拟环境中进行训练等任务。

然而，构建一个包含摄像机标注的动态视频数据集是具有挑战性的。首先，大多数互联网视频不适合用于摄像机姿态估计。它们可能是卡通、包含大量后期处理或缺乏清晰的参考框架。接下来，可估计的视频通常质量较低或者不包含场景动态。此外，即使可以识别出合适的目标视频，估计动态摄像机姿态的过程仍然存在挑战。在静态环境中效果良好的技术，例如结构从运动中提取，在面对移动的物体、变化的外观以及在互联网视频中常见的其他复杂动态时表现不佳。

鉴于这些挑战，大多数现有的数据集采取了不同的方法。一种工作思路是使用合成数据 [9, 28, 51, 52, 62, 105]，其中可以获得真实的相机数据。然而，由于 3D 素材的高成本，这通常导致较小规模的数据集（通常为 <500 个视频），并且方法面临实际模拟的差距。最近的研究在受限领域中从真实视频中获取相机数据，例如自我中心的视频 [81]、自动驾驶视频 [75]、手工收集的宠物视频 [72] 或特定动作 [27] 的视频。这些设置通过密集视点 [81]、LiDAR 传感数据 [75]、多摄像机 [27, 37] 和转盘式视频 [72] 使结构光形态建模更容易。然而，这些设置也对收集的数据施加了重要的限制。

我们介绍了 DynPose-100K，这是一大集合动态互联网视频，带有摄像机注释。示例视频和注释在图 1 中出现。该数据集包含多样的内容，各种动态对象的显著大小和针对的视频长度；并在 § 4 中进行了分析。我们解决了为适合的动态视频进行筛选以及用精心设计的管道对相应摄像机位姿进行注释的核心挑战（§ 3）。我们的筛选管道结合了一系列专门的专家，处理视频通常不合适的常见原因，以及一个可以检测并消除各种问题的通用 VLM。我们使用一种新的方法产生准确的位姿，该方法融合了运动掩蔽、对应物跟踪和 SfM 的最先进技术。我们将这种方法应用于 Panda-70M [12, 94]，筛选出 3.2 百万视频以产生 10 万段具有高质量摄像机信息的视频。

我们的贡献总结如下：

1. 我们引入了 DynPose-100K，这是一个大型动态互联网视频数据集，并附有相机信息标注。分析表明，该数据集在内容和动态方面具有多样性，并有针对性的视频长度。
2. 我们提出了一种使用专业模型和 VLM 的过滤流程，灵感来源于互联网视频分析。
3. 我们提出了一种动态姿态估计流程，将最先进的跟踪和遮罩组件整合在一起。
4. 实验表明，过滤选择的视频的精确度远高于其他方法，而姿势估计在各项指标和设置中最大可减少误差达 90 % 。

2. 相关工作

这项工作旨在收集和注释带有摄像机姿态的动态图像互联网视频。为此，我们在数据整理和姿态估计中面临挑战。

Dynamic camera pose datasets. 由于预测动态相机姿态的挑战，数据集通常采用策略来替代或辅助标准的 SfM。合成数据集很有吸引力，因为可以获得真实的姿态 [9, 28, 51, 52, 62, 105, 105]。然而，对工程化动态资产的要求意味着数据集规模较小，且通常包含有限的动态，比如 *e.g.* 物体飞行 [51] 或坠落 [28]。在真实世界的视频中，通过 *e.g.* 转盘式拍摄 [72]、密集的视点 [81]、多摄像机 [27]、鱼眼拍摄 [37] 和 LiDAR [75]，SfM 变得更容易。我们利用了 Panda-70M [12] 的多样化互联网视频数据集并展示了我们的方法可以在不依赖于先前工作的约束下标注高质量的姿态。同一时间进行的工作 CamCo [93] 和 B-Timer [45] 分别收集了 12K 和 40K 多样化视频 [3]；我们收集了 100K 并公开发布数据。

给定对一个静态场景的多个重叠视图

Camera pose estimation.，经典的 SfM [66, 74, 76] 和 SLAM [54] 是精准的黄金标准，而在给定大量数据集 [64, 87, 97, 106] 的情况下，学习的方法最近变得具有竞争力 [21, 41, 73, 77, 80, 82, 85, 88, 98, 101]。动态的互联网视频更具挑战性，因为动态对应不能用于束调整，也限制了静态对应。最近的研究 [26, 43, 50, 103] 利用学习预测 [13, 69, 103] 或运动和语义线索 [31, 33, 78] 来遮罩动态物体。我们采用在语义、交互、运动和跟踪 [15, 31, 36, 63, 78] 方面的最新技术来升级遮罩方法。

由于光线和外观的变化，互联网视频中的对应关系估计也是一项挑战 [43]。因此，我们展示了掩蔽 + 经典 SfM [66] 在多样化的互联网视频中失败。ParticleSfM [103] 通过传播的光流 [78] 引入密集对应关系，并使用更稳健的全局捆绑调整 Theia-SfM [76]。我们采用了 Theia-SfM 并通过 BootsTAP [20] 利用长期跟踪 [19, 30, 68, 84] 的进展来升级对应关系估计。并行研究 TracksTo4D [39] 在给定二维轨迹 [38] 的情况下预测姿态和动态掩码，但其专注于小动态和循环摄像机运动 [72]；我们基于更强的二维跟踪 BootsTAP 来解决多样化的互联网视频。并行研究 DATAP-SfM [96] 在合成的 FlyingThings3D [51] 上训练联合跟踪和运动预测，我们则利用由大量真实数据 [16, 20, 46, 63] 训练的跟踪和掩码方法并收集一个大规模数据集。并行研究 MonST3R [100] 和 MegaSaM [44] 分别将学习到的 DUSt3R [86] 和 DROID-SLAM [79] 泛化到动态场景；我们围绕经典 SfM 进行姿态标注构建，因为它在较长视频中具有精确性。

3. DynPose-100K： 数据集策划

DynPose-100K 是一个大规模的视频数据集，包括了带有相机注释的多样化动态互联网内容。组装这样一个数据集具有挑战性，主要因为与相机姿态估计相关的

www.xueshuxiangzi.com

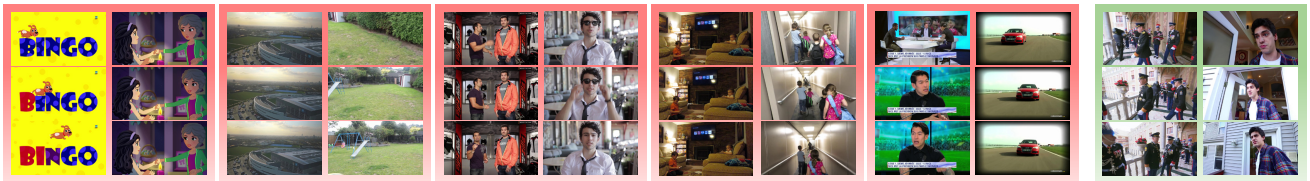

Figure 2. Panda-Test 数据集统计。统计反映在保留的 1K 视频 Panda-Test 集上的人工标注，详见 § 4.1 。由于各种问题，只有 9 % 是目标动态相机姿态估计视频，*e.g.* 静态场景、低质量或非真实内容，以及参考帧模糊或不明确。我们专注于移动相机以促进下游任务，如 *e.g.* 相机控制的视频生成和学习姿态估计。我们使用专业化模型和通用 VLM 的结合来删除不合适的视频。

两个问题：(a) 绝大多数互联网视频不适合成功的相机姿态估计，以及 (b) 动态视频上的相机姿态估计本质上是一个具有挑战性的问题。我们将在本节中解决这些挑战。

3.1. 候选人选择标准

代表性互联网视频数据集中的大多数视频（例如 Panda-70M [12] ）都不适合用来估计相机姿态：图 2 显示，来自 Panda-70M 数据集的随机选择的 1,000 个视频中，只有 9 % 的视频满足动态环境下相机姿态估计的标准。我们首先定义特定的标准来识别适合相机姿态估计的视频，然后通过定义的数据过滤步骤检测这些标准。

我们认为视频应该满足三个标准：

C1. 真实世界和高质量视频。视频必须来自真实世界 *i.e.* ，而不是卡通、计算机屏幕录制、并排/合成视频以及经过大量编辑和处理的内容。视频还应具备足够的质量、分辨率、帧率，并且不应有非透视失真。

C2. 姿势预测的可行性。视频不应包含严重的放大或缩小效果、突然的镜头切换或模糊的参考框架（*e.g.* 在移动的汽车内拍摄的场景）。没有静态对应关系的视频，*e.g.* ，当背景完全模糊或被遮挡时应予以排除。

C3. 动态相机和场景。非静态相机允许进行非平凡的姿势预测训练或相机控制的视频生成方法。具有移动摄像机和动态场景的视频通常也提供更丰富的数据，*e.g.* ，涉及人的互动，支持更广泛的下游应用。

3.2. 候选视频选择

过滤过程旨在自动选择符合这三个标准的视频。我们的过滤管道使用专门的模型（*e.g.* 训练用于检测失真或估算相机焦距）识别常见的不适合原因。另一方面，各种其他问题，如 *e.g.* ，后期编辑的文本，这些专门的模型无法有效处理。为此，我们使用近期的视觉-语言模型（VLM）[60] 并提示它们检测各种问题。正如将在 § 4.1 中显示的那样，虽然单个过滤器（包括 VLM）自身是不足够的，但它们的组合使用是有效的。

Filtering using task-specific models. 专用模型旨在解决视频被拒绝的具体且经常发生的原因。这些模型在时间上进行了帧子采样，以实现高效处理。

1. 动画片和演示。这使用分类器 [16] 来去除（C1）*e.g.* 动画或屏幕录像的标准；以及（C3）*e.g.* 静态场景，比如坐在摄像机前的人或不太可能互动的场景。
2. 非透视失真。具有较高预测失真的视频会被删除。失真会使需要可靠对应的应用程序变得复杂，因此我们认为它们的质量较低（C1）。
3. 焦距。去除焦距预测方差高的视频，因为它们通常包含变焦效果或镜头切换（C2）。我们还排除长焦距的视频，因为它们通常包括背景模糊过度，不适合可靠姿态估计（C2），或者是使用静止摄像机的运动镜头（C3）。
4. 动态对象遮罩。对于预测遮罩尺寸过大的视频，被认为不适合姿态估计（C2）并被移除，因为可靠的姿态估计需要足够的静态对应关系。
5. 光流。高峰值的连续光流通常是由镜头切换 (C2) 引起的；这样的视频通过预测流被去除。低平均连续流通常表示静态视频 (C3)；这些也被去除。
6. 点跟踪。预测的点跟踪点突然消失可能表示镜头转换、背景模糊或剧烈的放大/缩小，使得视频不适合姿态估计（C2）。相反，极其稳定的跟踪则表明摄像机或场景都没有显著移动（C3）。这两种情况都被去除。

Filtering using general VLM. 通用 VLM 处理多种视频拒绝的可能性。我们提示 GPT-4o mini [59, 60] 一系列涵盖所有三类标准的八个问题。它们涵盖了：视频是否来自一个模糊的参考，如果背景太模糊以至于无法匹配对应关系，如果帧被扭曲或具有很长的焦距，如果场景是静态的，如果视频是卡通，是否经过后期处理，以及是否有儿童出现在视频中。我们发现与专业过滤器有重叠之处，*e.g.* 识别长焦距，是有帮助的，因为 VLM 可以被提示考虑多样的线索进行移除。

从 3.2M Panda-70M 视频开始，我们应用过滤（§ ~ 3.2 ）剩下 137K 动态视频。我们在 107K 上估计姿态，并删除轨迹帧注册数少于 80 % 的，剩下 100K 视频。我们依次应用过滤以提高效率。首先在整个集合上应用轻量级的 Hands23、光流和焦距过滤。我们删除低评分的视频，剩下 1.63M 视频。接下来运行失真过滤并删除低评分视频，剩下 1.53M；然后跟踪，剩下 679K；再

www.xueshuxiangzi.com

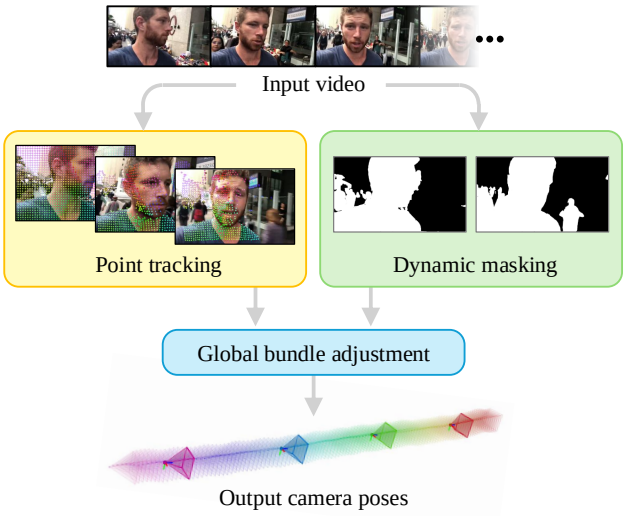

Figure 3. 姿态估计方法。我们在滑动窗口中应用最先进的点跟踪方法，以产生密集的、长期的对应关系。使用补充的动态掩码来去除非静态轨迹。剩余的静态轨迹被作为输入提供给全局束调整。

然后是遮罩，剩下 462K；最后是 VLM，之后我们应用严格的最终过滤得到 137K 用于姿态估计。

3.3. 动态相机姿势估计

在收集到合适的视频后，我们的目标是获得高质量的相机姿态。在动态互联网视频上进行姿态估计具有挑战性：不仅因为动态物体遮挡了基础静态场景，而且静态场景可能发生外观变化，导致对应估计困难。我们的方法解决了这两个挑战（图 3）。对于遮挡，我们使用了语义、交互、运动和跟踪方面的最新方法，每个组件都提升了整体性能。对于对应，我们使用了最新的点跟踪方法 [20]，在视频中以滑动窗口的方式应用。最后，我们应用了全局捆绑调整 [76]，该方法已被证明能够有效处理互联网视频所带来的挑战 [103]。

我们首先目标是在输入视频中分割 *动态* 区域。我们的方法利用了几个领域的进展来生成精确的掩膜。

1. 语义分割。我们使用 OneFormer 对常见的动态类别（*e.g.* 人类、车辆、动物和运动装备）进行分割。
2. 对象交互分割。这种方法会对持有的物体进行遮罩，这些物体可能是动态的但超出了语义类别 *e.g.* 餐具。我们使用 Hands23 [16] 手-物体交互分割模型。
3. 运动分割。这处理了不属于常见类别或涉及人类操作的动态物体，*e.g.* 树叶沙沙作响或流水。我们使用 RoDynRF [50] 的运动掩模来移除基于光流 [78] 具有高 Sampson 误差 [31] 的区域。
4. 掩码传播。传播在帧之间提供平滑的掩码和精确的对象边界。我们使用 SAM2 [63] 来传播掩码。

Point tracking. 我们使用点跟踪来估计对应关系，该方法利用视频中的时间信息来改善估计。与成对估计不同，跟踪利用了点在连续帧之间通常只移动少量的事实。此外，在多个帧中跟踪同一点能够在每个帧组合中快速增加配对对应关系。为此，我们使用 BootsTAP [20] 来跟踪一个点网格向前数个帧，向前移动，然后以滑动窗口方式重复。基于网格的点跟踪促进了更密集的对应关系，而扩展的跟踪持续时间支持长期对应关系，减少漂移并有助于闭环。点网格的滑动窗口跟踪确保每个帧在前面的窗口被遮挡或发生漂移的情况下维护足够数量的对应关系。

我们使用全局捆绑调整方法 Theia-SfM [76]，使用来自 tracklets 的对应作为输入。从单个 tracklet 中，我们提取所有对的对应，排除包含动态掩膜中 tracklet 的帧的对。

4. DynPose-100K： 数据集分析

在这项工作中，由于动态互联网视频的规模和多样性，我们重点关注这种视频。我们评估了过滤流程在识别适合姿势估计（§ 4.1）的视频方面的有效性，并对生成的数据集（§ 4.2）进行统计分析。

4.1. 在 Panda-Test 上的过滤评估

我们的过滤目标是从各种互联网视频中识别出适合姿态估计的视频。

Dataset. 我们在从 Panda-70M 中随机选择的 1,000 个保留视频上评估过滤性能，这些视频我们称之为 Panda-Test。每个视频都经过手动分类，以判断是否适合姿态估计，结果产生了 90 个（9 %）适合的视频。这些 90 个视频在 DynPose-100K 、*e.g.* 内容、长度和动态大小方面表现出类似的多样性。

Metrics. 我们的主要评估指标是准确率和召回率。我们优先考虑准确率，旨在尽量减少 DynPose-100K 中包含的不准确视频数量。然而，召回率也很重要，因为低召回率意味着需要处理更多数量的视频来实现同等规模的数据集。

Baselines and ablations. 我们将我们的方法与现有和替代的过滤方法进行比较。我们使用我们的 SfM 重建点（CamCo [93]）和重投影误差（受 B-Timer [45] 启发）进行过滤，同时结合视频适用性和交互分类器 Hands23 [16]。此外，我们以类似于我们的过滤管道（§ 3.2）的方式提示 GPT-4o mini [60]。我们实现了这次比较的两个版本：一个产生二值输出，另一个为每个示例分配分数。最后，我们对 § 3.2 的过滤方法的每个步骤进行了消融研究。

Results. 图 5 显示我们的筛选以高精度和召回率选择视频，超过基线方法。实际上，在用于收集 DynPose-100K 的阈值下，筛选的测试精度为 0.78，这是除重投影误差在 0.02 召回率外，没有任何基线可在任何召回率下达到的水平。我们方法中的每个组件都提高了性能，其中 VLM 提供了很大的提升，尽管作为一种单独的方法它相对较弱。

www.xueshuxiangzi.com

Dataset	Real/Syn.	Num. vids.	Num. frames	Domain	Access
T.Air Shibuya [62]	Syn.	7	0.7K	Street	Public
MPI Sintel [9]	Syn.	14	0.7K	Scripted	Public
PointOdyssey [105]	Syn.	131	200K	Walking	Public
FlyingThings3D [51]	Syn.	220	2K	Objects	Public
Kubric Movi-E [28]	Syn.	400	10K	Objects	Public
EpicFields [81]	Real	671	19,000K	Kitchens	Public
Waymo [75]	Real	1,150	200K	Driving	Public
CoP3D [72]	Real	4,200	600K	Pets	Public
Ego-Exo4D [27]	Real	5,035	23,000K	Set tasks	Public
Stereo4D [37]	Real	110,000	10,000K	S. fisheye	Public
CamCo [93]	Real	12,000	385K	Diverse	Private
B-Timer [45]	Real	40,000	19,000K	Diverse	Private
DynPose-100K	Real	100,131	6,806K	Diverse	Public

Table 1. Dynamic camera pose datasets. DynPose-100K has the most videos of diverse Internet video datasets. Datasets with more frames are private or more uniform; *e.g.* stereo fisheye is typically outdoor PoV walking. We use short videos, yielding fewer frames but high dynamics (Figure 6).

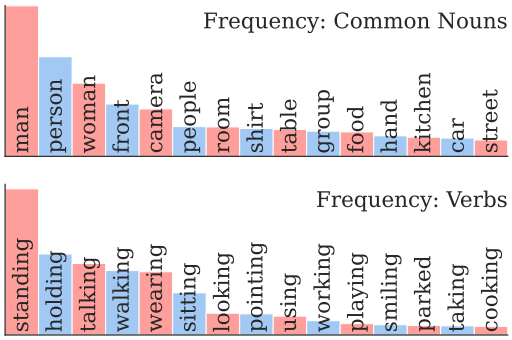

Figure 4. Diverse content. Frequent nouns cover varied subjects: person, hand, car; objects: shirt, table, food; and settings: room, kitchen, street. Verbs span diverse actions: using, working, playing.

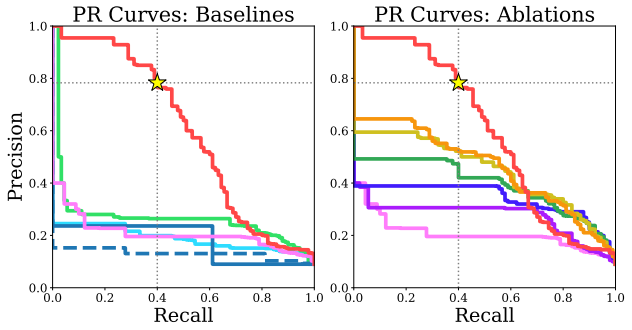

Figure 5. 熊猫测试的动态视频过滤。我们展示了基线和消融的 PR 曲线。我们的过滤明显超越了所有基线和消融。☆代表 DynPose-100K 的操作阈值。对于基线，我们展示：■ 重建点（CamCo [93]），■ 重投影误差，（固体 ■ ）GPT-4o mini [60]：二元，（虚线 ■ ）GPT-4o mini [60]：得分，■ Hands23 [16]，和 ■ 我们的。对于消融，我们从 ■ Hands23 开始并添加组件直到恢复到 ■ 我们的。具体来说，我们描绘：■ Hands23，■ + 流动，■ + 跟踪，■ + 掩膜，■ + 焦点，■ + 失真，■ +VLM（我们的）。

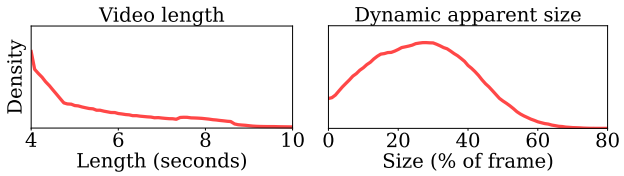

Figure 6. Left: Targeted video length. DynPose-100K videos are primarily 4-10s, ideal for dynamic pose: shorter videos contain little ego-motion, longer videos have less dense dynamics and ego-motion. Right: Diverse dynamic apparent size. Mean size in % across video. Large dynamic objects occlude static correspondences, making pose estimation challenging. Videos may average small size in the case of only a few dynamic frames.

4.2. 数据集统计

在收集了大量目标视频后，我们评估了所产生的数据集的特征。除了评估数据集的大小之外，我们还检查结果视频是否展示了期望的属性：*e.g.* 多样化的内容、适当的视频长度和多样的动态物体大小。

Dataset size. 表 1 显示 DynPose-100K 包含的视视频比现有的多样化数据集要多得多。最大的替代方案通常局限于特定场景，比如厨房、驾驶、步行或宠物转台视频 [27, 37, 72, 75, 81]。同时进行的工作 CamCo [93] 和 B-Timer [45] 也有多样化内容，但视频数量较少且是私有的。

Video content. 图 4 展示了在 Panda-70M 的 [12] 标注中与 DynPose-100K 视频相关的常见名词和动词。常见名词涵盖了广泛的主题、对象和环境；而动词反映了各种动作。这种语言多样性表明 DynPose-100K 包含了内容多样的视频。

Video length. 图 6 左侧显示了 DynPose-100K 中视频长度的分布。大多数视频的长度在 4 到 10 秒之间。这些短视频通常动态丰富，并且时间足够长，可以体现显著的相机运动。

Dynamic object size. 图 6，右图显示了在 DynPose-100K 中动态物体表观大小的分布，从小到大不等。大的动态物体会遮挡静态对应物，使得姿态估计具有挑战性。较小的表观动态物体通常对应于更远的物体，这些物体可以快速移动，使得精确遮蔽具有挑战性。视频中动态物体几乎占据整个画面的情况会被过滤掉，因为它们使得姿态估计变得不可行。

5. 相机姿态估计的评估

我们通过控制实验来评估整理后的数据集 DynPose-100K 在姿态估计方面的效果。由于缺乏真实的相机位姿数据，直接在动态的互联网视频上进行评估是具有挑战性的。因此，我们的评估方法是双管齐下的：（1）

www.xueshuxiangzi.com

Method	All 36 videos (Iden. Rot. + Rand. Tr. if fail)				8 videos: all succeed only		
	% Vids. reg. ↑	ATE (m) ↓	RPE Tr. (m) ↓	RPE Rot. (°) ↓	ATE (m) ↓	RPE Tr. (m) ↓	RPE Rot. (°) ↓
Iden. Rot. + Rand. Tr.	100.	0.652	0.139	1.60	0.390	0.080	1.00
DROID-SLAM [79]	100.	0.198	0.046	1.75	0.048	0.017	0.82
DUSt3R [86]	97.2	0.412	0.177	20.1	0.256	0.124	18.5
MonST3R [100]	100.	0.149	0.046	1.21	0.036	0.011	0.46
LEAP-VO [13]	100.	0.206	0.049	1.70	0.037	0.011	0.73
COLMAP [66]	44.4	0.388	0.082	2.03	0.122	0.026	1.91
COLMAP+Mask [66]	38.9	0.323	0.085	1.64	0.089	0.017	1.36
ParticleSfM [103]	97.2	0.185	0.075	2.99	0.051	0.047	2.91
Ours	100.	0.072	0.033	1.31	0.003	0.002	0.30

Table 2. Lightspeed 上的相机姿态估计。我们的姿态估计算法对所有序列进行配准，并将所有序列的轨迹误差减少了 50 %（左图），在简单序列中减少了 90 %（右图），*vs.* 所有其他方法。

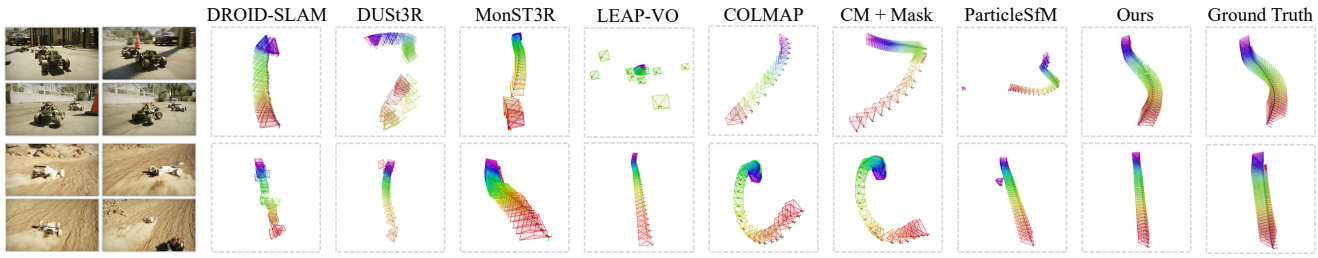

Figure 7. 在 Lightspeed 上的预测轨迹。姿态随时间的序列：R O Y G B V 。我们在左侧可视化 Lightspeed 的照片级真实渲染。在此动态场景中，静态方法 DROID-SLAM、COLMAP 和 DUSt3R 难以成功登记一致的序列，或者造成曲率过多或过少。上图：动态方法 MonST3R、LEAP-VO 和 ParticleSfM 无法产生平滑的序列。下图：MonST3R、LEAP-VO 和 COLMAP+Mask 增加了曲率。我们的方法生成了平滑且准确的轨迹。

我们使用提供真实姿态以供直接比较的真实感合成渲染数据集 Lightspeed；（2）我们通过在 Panda-Test 上标注 1 万个精确对应点，直接在互联网视频上进行评估，从而可以使用 Sampson 误差 [31] 来进行姿态评估。

我们主要的比较对象是 ParticleSfM [103]，其流程与我们的相似，包括静态跟踪、动态屏蔽和全局 SfM。我们还与标准的 SfM 方法 COLMAP [66] 进行比较，既在其原始形式下，也在使用动态屏蔽过滤匹配点的情况下进行比较。最后，我们将我们的方法与最先进的基于学习的静态方法 DROID-SLAM [79] 和 DUSt3R [86]，以及动态方法 LEAP-VO [13] 和 MonST3R [100] 进行对比评估。

5.1. 对 **Lightspeed** 的姿态评估

我们介绍了 Lightspeed，这是一个具有挑战性的照片级真实动态姿态估计基准，包含真实的相机姿态。这个数据集是姿态估计的良好基准，因为它共享 DynPose-100K 的几个重要特征：多样化的环境，大型动态物体和几秒钟的不同剪辑长度。

Dataset. 我们使用 NVIDIA Racer RTX [57]，如图 7 所示，该图展示了几辆遥控车在具有挑战性的室内和室外场景中快速移动的情景，特征包括日夜光照和第一人称及第三人称视角。我们将视频拆分为多个片段，去除了那些具有非透视失真、背景模糊、场景静止或摄像机处于静止或简单线性轨迹（参考 [92]）的片段，最终得到 36 个序列。我们从相应的 3D 资产中提取出了真实的相机位姿。

根据之前的研究 [50, 69, 103]，我们报告了平均轨迹误差（ATE）以及旋转和平移的相对位姿误差（RPE）。有关详细信息，请参阅之前的研究。简而言之，这些指标是在轨迹对齐和缩放后计算的，相对误差是在顺序帧上计算的。为了计算所有视频的得分，我们将无法收敛的预测轨迹替换为随机均匀平移和单位旋转。我们还分别报告了所有方法都收敛的视频子集上的得分。

Results. 表 2 显示 COLMAP 和 COLMAP+Mask 在 Lightspeed 中难以处理许多具有挑战性的序列。DROID-SLAM、DUSt3R、LEAP-VO 和 ParticleSfM 提供了注册，但不够准确。MonST3R 提供了较好的轨迹误差，而我们的方法显然更为优越，将轨迹误差减少了全视频的 50 % 和成功案例中的 90 % 。所有替代方法至少在图 7 的一个序列中遇到困难，而我们的方法则处理了两个序列。

5.2. **Panda-Test** 上的姿势评估

接下来，我们在动态互联网视频上进行评估。我们使用来自 Panda-Test 的 90 个视频子集（§ 至 4.1），其中包含各种具有可估计摄像机姿态的挑战性对象。

Metrics. 无标签动态互联网视频中地面真相相机姿态不可用，因此我们选择标注对应点并从预测姿态中计算 Sampson 误差。我们从横跨 90 个视频的不同图像

www.xueshuxiangzi.com

Baselines	% Vids. registered ↑	All 90 estimable videos (Identity if fail) Mean per-video reprojection error, 720p % < 5 Pix ↑	% < 10 Pix ↑	% < 30 Pix ↑	Mean ↓	52 videos: all succeed only Mean per-video reprojection error, 720p % < 5 Pix ↑	% < 10 Pix ↑	% < 30 Pix ↑	Mean ↓
Identity	100.	0.0	0.0	2.2	151	0.0	0.0	0.0	133
DROID-SLAM [79]	100.	57.8	77.8	94.4	11.0	59.6	84.6	96.2	6.78
DUSt3R [86]	96.7	0.0	6.7	48.9	43.0	0.0	9.6	57.7	30.3
MonST3R [100]	100.	55.6	78.9	90.0	9.86	63.5	84.6	90.4	9.71
LEAP-VO [13]	100.	64.4	76.7	96.7	7.59	75.0	84.6	98.1	6.03
COLMAP [66]	82.2	51.1	62.2	73.3	27.5	71.2	82.7	92.3	9.03
COLMAP+Mask [66]	67.8	47.8	58.9	75.6	30.1	69.2	82.7	96.2	6.10
ParticleSfM [103]	92.2	70.0	76.7	88.9	12.5	80.8	86.5	96.2	6.77
Ours	95.6	72.2	84.4	98.9	5.76	82.7	94.2	100.	3.75

Table 3. 熊猫测试上的相机姿态估计。在 10K 图像对上的重投影误差，按视频进行归一化到 720p。静态方法 DUSt3R 和 COLMAP 在面对动态环境时表现不佳，而 DROID-SLAM 缺乏精度。ParticleSfM 注册的视频比 COLMAP+Mask 更多，但两者在注册和精度上都不及我们的方法。MonST3R 和 LEAP-VO 注册了所有帧，导致中等错误，但我们的方法在所有指标上表现更好。

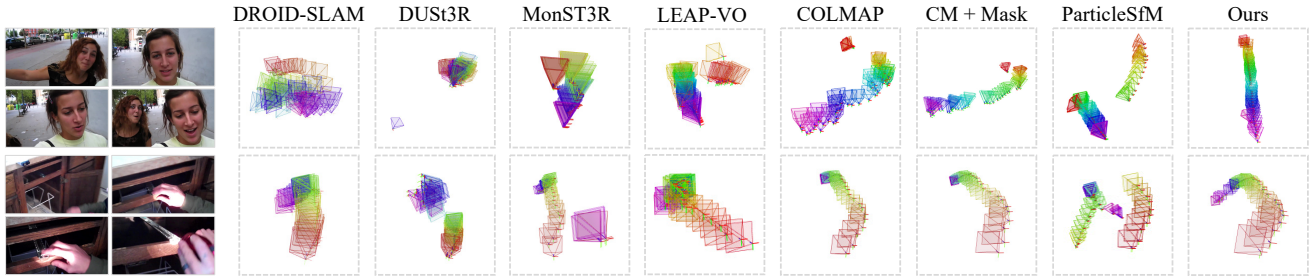

Figure 8. 在 Panda-Test 上的预测轨迹。随着时间变化的姿态序列: R O Y G 乙 五 。像 DROID-SLAM、DUSt3R 和 COLMAP 这样的静态方法在面对动态场景时会遇到困难。MonST3R、LEAP-VO、COLMAP+Mask 和 ParticleSfM 可能在处理大规模动态区域（顶部）和跨越不同外观和光照进行跟踪（底部）时遇到困难，而我们的方法可以处理这些情况。

对中策划了 1 万个对应点用于评估。图像对由专家标注员手动标注以达到粗略准确性。这些匹配通过使用 SuperPoint [18] 和 LightGlue [47] 进行细化，具体方法为在 720p 图像上选择与人工标注点相距 10 个像素以内的 LightGlue 匹配。如果找不到这样匹配，则对应点会被丢弃。帧对在彼此间隔 2.5 秒内随机选择。

每对的重投影误差是 Sampson 误差的平方根。Sampson 误差计算的是从标注的对应点到使用基础矩阵（基于预测的相对相机姿态和内参）计算出的极线的平方距离。度量被对每个视频中的所有对进行平均，从而支持逐视频分析。我们的主要度量指标是跨视频的平均误差，以及平均重投影误差在阈值范围内的视频百分比；我们使用归一化的 720p 分辨率下的 5、10 和 30 像素。未注册的帧使用最近邻的姿态填充。如果整个序列未注册，则使用单位矩阵。

表 3 显示了与 Lightspeed 相似的趋势：静态方法 DROID-SLAM 和 DUSt3R 不准确，而即使加上动态掩码，COLMAP 仍然难以处理很多视频。ParticleSfM 有所改进，但仍有很高的平均误差。MonST3R 和 LEAP-VO 可以处理所有的帧，但我们的方法降低了各项指标上的误差。在所有方法都成功的 52 个序列中，我们的方法比所有其他方法在平均误差上减少了超过 35%。图 8 展示了来自两个具有挑战性的序列的定性结果。定性结果与定量结果一致：DROID-SLAM、DUSt3R 和 COLMAP 表现不佳，而 MonST3R、LEAP-VO、COLMAP+Mask

Method	% Vids. reg. ↑	All 90 estimable videos (Identity if fail) Mean per-video reprojection error, 720p % < 5 Px ↑	% < 10 Px ↑	% < 30 Px ↑	Mean ↓
DUSt3R	96.7	0.0	6.7	48.9	43.0
+ Synthetic (MonST3R [100])	100.	55.6	78.9	90.0	9.86
+ DynPose-100K (Ours)	100.	54.4	82.2	92.2	8.78

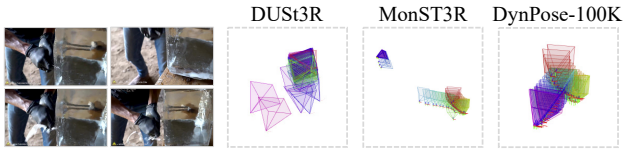

Table 4. Camera estimation on Panda-Test. DynPose-100K fine-tuning of DUSt3R has lower mean error and similar to or better than accuracy compared to synthetic data (MonST3R). We train with only 2K videos / 140K frames, smaller than the 1.3M frames used to train MonST3R; demonstrating efficient supervision.

和 ParticleSfM 面临失败案例。我们的方法在处理涉及较大动态或外观变化的困难情况下表现更佳。显然，结合使用的用于掩码和跟踪的最先进组件（§ 3.3 ）在每个相应任务中都取得了高效的流水线效果。

6. 结论

在本文中，我们介绍了 DynPose-100K ，一个大型动态互联网视频数据集，附有摄像机姿态标注。该数据集通

www.xueshuxiangzi.com

过精心设计的视频过滤和摄像机姿态估计流程进行策划，并通过实验验证。我们希望 DynPose-100K 可以开启令人兴奋的新可能性。例如，表 4 显示该数据集可以微调 DUSt3R [86]，以比 MonST3R [100] 的合成数据产生更低的平均误差。为了补充 DynPose-100K，我们收集了高质量的合成数据集 Lightspeed，以支持真实姿态基准测试。

Acknowledgments. DF 得到了 NSF IIS 2437330 的支持。我们感谢 Gabriele Leone 和 NVIDIA Lightspeed 内容技术团队分享了原始 3D 资产和场景数据以创建 Lightspeed 基准测试。我们感谢 Yunhao Ge, Zekun Hao, Yin Cui, Xiaohui Zeng, Zhaoshuo Li, Hanzi Mao, Jiahui Huang, Justin Johnson, JJ Park 和 Andrew Owens 在这个项目中给予的宝贵启发、讨论和反馈。

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In RSS , 2022.

[2] Sherwin Bahmani, Xian Liu, Yifan Wang, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In ECCV , 2024.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In ICCV , 2021. 3

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In NeurIPS , 2021.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv , 2023.

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In CVPR , 2023.

[7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In ICLR , 2025.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

[9] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In ECCV , 2012. 3, 6

[10] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In CVPR , 2022.

[11] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In 3DV , 2017.

[12] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In CVPR , 2024. 3, 4, 6

[13] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In CVPR , 2024. 3, 7, 8

[14] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. arXiv , 2024.

[15] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In CVPR , 2024. 3

[16] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. NeurIPS , 2023. 3, 4, 5, 6

[17] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. In NeurIPS , 2024.

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In CVPRW , 2018. 8

[19] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In ICCV , 2023. 3

[20] Carl Doersch, Yi Yang, Dilara Gokay, Pauline Luc, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ross Goroshin, João Carreira, and Andrew Zisserman. Bootstap: Bootstrapped training for tracking-any-point. In ACCV , 2024. 3, 5

[21] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. arXiv , 2024. 3

[22] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. IJCV , 2010.

[23] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In CoRL , 2024.

[24] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In NeurIPS , 2022.

[25] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In ICLR , 2024.

www.xueshuxiangzi.com

[26] Lily Goli, Sara Sabour, Mark Matthews, Marcus Brubaker, Dmitry Lagun, Alec Jacobson, David J Fleet, Saurabh Saxena, and Andrea Tagliasacchi. Romo: Robust motion segmentation improves structure from motion. arXiv , 2024. 3

[27] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In CVPR , 2024. 3, 6

[28] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In CVPR , 2022. 3, 6

[29] Annika Hagemann, Moritz Knorr, and Christoph Stiller. Deep geometry-aware camera self-calibration from video. In ICCV , 2023.

[30] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In ECCV , 2022. 3

[31] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision . Cambridge university press, 2003. 3, 5, 7

[32] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. arXiv , 2024.

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV , 2017. 3

[34] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv , 2022.

[35] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. arXiv , 2023.

[36] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In CVPR , 2023. 3

[37] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. In CVPR , 2025. 3, 6

[38] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In ECCV , 2024. 3

[39] Yoni Kasten, Wuyue Lu, and Haggai Maron. Fast encoder-based 3d from casual videos via point track processing. In NeurIPS , 2024. 3

[40] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. In SIGGRAPH , 2017.

[41] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In CVPR , 2021. 3

[42] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In NeurIPS , 2024.

[43] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In CVPR , 2021. 3

[44] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In CVPR , 2025. 3

[45] Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, and Jiahui Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. arXiv , 2024. 3, 5, 6

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV , 2014. 3

[47] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In ICCV , 2023. 8

[48] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In CVPR , 2024.

[49] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In ICCV , 2023.

[50] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In CVPR , 2023. 3, 5, 7

[51] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In CVPR , 2016. 3, 6

[52] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In CVPR , 2023. 3

[53] Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. In ICRA , 2007.

[54] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. T-RO , 2015. 3

[55] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In CVPR , 2023.

[56] NVIDIA. Cosmos world foundation model platform for physical ai. arXiv , 2025.

[57] NVIDIA GeForce. NVIDIA Racer RTX | The future of graphics powered by GeForce RTX 40 Series, 2022. [Online; accessed Access Date: 2024-11-9]. 7

www.xueshuxiangzi.com

[58] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In CoRL , 2023.

[59] OpenAI. Gpt-4 technical report, 2023. 4

[60] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. 4, 5, 6

[61] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In ECCV , 2024.

[62] Yuheng Qiu, Chen Wang, Wenshan Wang, Mina Henein, and Sebastian Scherer. Airdos: Dynamic slam benefits from articulated objects. In ICRA , 2022. 3, 6

[63] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. arXiv , 2024. 3, 5

[64] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In ICCV , 2021. 3

[65] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas MÃijller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In CVPR , 2025.

[66] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In CVPR , 2016. 3, 7, 8

[67] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In CVPR , 2017.

[68] Jenny Seidenschwarz, Qunjie Zhou, Bardienus Duisterhof, Deva Ramanan, and Laura Leal-Taixé. Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. arXiv , 2024. 3

[69] Shihao Shen, Yilin Cai, Wenshan Wang, and Sebastian Scherer. Dytanvo: Joint refinement of visual odometry and motion segmentation in dynamic environments. In ICRA , 2023. 3, 7

[70] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv , 2022.

[71] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. arXiv , 2023.

[72] Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In CVPR , 2023. 3, 6

[73] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In 3DV , 2025. 3

[74] Noah Snavely. Scene reconstruction and visualization from internet photo collections. PhD Thesis , 2008. 3

[75] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In CVPR , 2020. 3, 6

[76] Chris Sweeney. Theia multiview geometry library: Tutorial & reference. http://theia-sfm.org. 3, 5

[77] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. arXiv , 2018. 3

[78] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In ECCV , 2020. 3, 5

[79] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. NeurIPS , 2021. 3, 7, 8

[80] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In NeurIPS , 2023. 3

[81] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. NeurIPS , 2024. 3, 6

[82] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In CVPR , 2024. 3

[83] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In CVPR , 2021.

[84] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In ICCV , 2023. 3

[85] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In CVPR , 2025. 3

[86] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In CVPR , 2024. 3, 7, 8, 9

[87] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In IROS , 2020. 3

[88] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. TartanVO: A generalizable learning-based VO. In CoRL , 2021. 3

[89] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. In ECCV , 2024.

[90] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A

www.xueshuxiangzi.com

unified and flexible motion controller for video generation. In SIGGRAPH , 2024.

[91] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In CVPR , 2020.

[92] J. Wulff, D. J. Butler, G. B. Stanley, and M. J. Black. Lessons and insights from creating a synthetic optical flow benchmark. In ECCVW , 2012. 7

[93] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv , 2024. 3, 5, 6

[94] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In CVPR , 2022. 3

[95] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In SIGGRAPH Asia , 2023.

[96] Weicai Ye, Xinyu Chen, Ruohao Zhan, Di Huang, Xiaoshui Huang, Haoyi Zhu, Hujun Bao, Wanli Ouyang, Tong He, and Guofeng Zhang. Datap-sfm: Dynamic-aware tracking any point for robust dense structure from motion in the wild. arxiv , 2024. 3

[97] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In ICCV , 2023. 3

[98] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In CVPR , 2018. 3

[99] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In CVPR , 2021.

[100] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In ICLR , 2025. 3, 7, 8, 9

[101] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In ECCV , 2022. 3

[102] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. arXiv , 2024.

[103] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In ECCV , 2022. 3, 5, 7, 8

[104] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. In ICLR , 2025.

[105] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In ICCV , 2023. 3, 6

[106] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In SIGGRAPH , 2018. 3

[107] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In NeurIPS , 2023.

www.xueshuxiangzi.com