【ICCV-2023 Best Paper 详解】Tracking Everything Everywhere All at Once

转载自 https://zhuanlan.zhihu.com/p/661157000

- 原文：https://arxiv.org/pdf/2306.05422.pdf
- 官网：http://omnimotion.github.io/

本文提出的OmniMotion可以对视频中的每个像素进行准确的长距离 运动估计 。OmniMotion使用准三维规范体积来表示视频，并通过局部空间和规范空间之间的双射（bijections）进行逐像素跟踪。这种表示方法能够确保全局一致性，跟踪遮挡目标，并对相机和物体运动的任意组合进行建模。

一句话总结方法：本文利用了单个canonical volume的bijections、光照一致性、以及坐标相关网络

和

提供的时空平滑性，来调和不一致的pairwise flow，并填充对应graph中缺失的内容。

1. Introduction

运动估计主要有两种方法：sparse feature tracking ，dense optical flow
这两种表示方法都无法完全模拟视频的运动：成对的光流无法捕捉较长时间窗口内的运动轨迹，而稀疏跟踪则无法对所有像素的运动进行建模。

为了弥补这一差距，已经有许多方法尝试估计视频中 dense 和long-range像素轨迹。这些方法包括简单地将两帧光流场串联在一起，以及直接预测多帧中每个像素的轨迹。然而，这些方法在运动估计时都使用了有限的上下文信息，忽略了时间上或空间上相距较远的信息。这种局部性可能导致长轨迹中累积的误差和运动估计中的时空不一致性。即使先前的方法考虑了long-range context，它们仍然在2D领域进行操作，发生遮挡时会丢失跟踪。
总的来说，同时产生dense和long-range轨迹仍然是该领域的一个未解决问题，其中存在三个关键挑战：（1）在长序列中保持准确的跟踪，（2）通过遮挡跟踪点，（3）在时空上保持一致性。

在这项工作中，我们提出了一种全面的视频运动估计方法OmniMotion，利用视频中的所有信息来共同估计每个像素的完整运动轨迹。它使用准三维表示，其中一个规范的三维体被映射到每帧的局部体积中，通过一组local-canonical bijections。这些bijections作为动态多视角几何的灵活松弛，模拟了相机和运动场景的组合。这种表示保证了循环一致性，并且可以跟踪所有像素，即使在遮挡时，因此称为"Everything , Everywhere"。我们通过对整个视频的运动进行联合求解来优化表示。一旦优化完成，该表示方法可以在视频的任何坐标上查询，得到跨越整个视频的运动轨迹。

总结一下，我们提出了一种方法，具有以下特点：1）为整个视频中的所有点生成全局一致的完整运动轨迹；2）能够在遮挡情况下跟踪点；3）能够处理具有任意相机和场景运动组合的真实场景视频。我们在TAP视频跟踪基准测试[15]上定量地展示了这些优势，取得了领先的 性能 ，远远超过了所有先前的方法。

2. Related Work

Sparse feature tracking. 多帧图像中的特征跟踪[4, 42]应用非常广泛，例如Structure from Motion（SfM）[1, 56, 59]和SLAM[17]。**尽管稀疏特征跟踪[13, 43, 57, 67]可以建立long-range对应关系，但这种对应关系仅限于一组有区别的兴趣点，并且通常只限于刚性场景。**因此，下面我们将重点关注能够为一般视频生成密集像素运动的工作。

Optical flow. 一般来说，光流被表述为一个优化问题[6, 7, 24, 75]。然而最近一些方法使用神经网络 直接预测光流，具有更好的质量和效率[20, 25, 26, 61]。其中一种领先的方法是RAFT[66]，它通过基于4D相关体的迭代更新来估计光流场。**虽然光流方法可以在相邻帧之间进行精确的运动估计，但它们不适用于长程运动估计：将成对的光流连接成更长的轨迹会导致漂移并无法处理遮挡，而直接计算远距离帧之间的光流（即更大的位移）通常会导致时间上的不一致性[8,75]。**多帧光流估计方法[27, 29, 52, 70]可以解决两帧光流的一些局限性，但仍然难以处理长程运动。

Feature matching. 虽然光流方法通常用于处理相邻帧，但其他技术可以估计远距离视频帧之间的密集对应关系[41]。一些方法通过自监督或弱监督的方式学习这样的对应关系[5, 10, 13, 37, 53, 71, 73, 78]，利用循环一致性等线索，而其他方法[18,30,62,68,69]则使用更强的监督信号，例如从三维重建流程中生成的几何对应关系[39, 56]。然而，成对匹配方法通常不考虑时间上下文，这可能导致在长视频中出现不一致的跟踪和较差的遮挡处理。相比之下，我们的方法能够通过遮挡产生平滑的轨迹。

Pixel-level long-range tracking. 最近的一个显著方法是PIPs[23]，它利用一个较小的时间窗口（8帧）内的上下文信息，来估计遮挡中的多帧轨迹。然而，为了对超过这个时间窗口长度的视频生成运动轨迹，PIPs仍然需要链式对应，这个过程容易出现漂移，并且会在超过8帧窗口的范围内无法继续跟踪被遮挡的点。与我们的工作同时进行的是，一些学习方法用于以前馈方式预测像素级的长程轨迹。MFT[46]学习选择最可靠的光流序列来进行长程跟踪。TAPIR[16]通过采用TAP-Net[15]启发的匹配阶段和PIPs[23]启发的细化阶段来跟踪点。CoTracker[31]提出了一种灵活而强大的跟踪算法，采用Transformer 架构在整个视频中跟踪点。**我们的贡献与这些工作是互补的：任何这些方法的输出都可以作为全局运动优化时的输入监督。**

Video-based motion optimization. 在概念上与我们的方法最相关的是经典方法，它们在整个视频上全局优化运动[2,12,36,54,55,60,63,72]。例如，粒子视频从初始光流场中生成一组半密集的长程轨迹（称为粒子）[55]。然而，它不能在遮挡情况下进行跟踪；当一个被遮挡的实体重新出现时，它将被视为不同的粒子。Rubinstein等人[54]进一步提出了一种组合分配方法，可以在遮挡情况下进行跟踪并生成更长的轨迹。然而，该方法只针对具有简单运动的视频生成半密集轨迹，而我们的方法能够估计一般视频中所有像素的长程运动。还有一个相关的方法是ParticleSfM[82]，它通过成对光流进行长程对应关系的优化。与我们的方法不同，ParticleSfM侧重于在SfM框架内进行相机姿态估计，只优化来自静态区域的对应关系，而将动态对象视为异常值。

Neural video representations. 虽然我们的方法与最近使用基于坐标的多层感知器（MLP）[44,58,65]对视频建模的方法存在相似之处，但先前的工作主要集中在诸如新视角合成[38, 40, 47, 48, 77]和视频分解[32, 81]等问题上。相比之下，我们的工作针对的是密集的、长程的运动估计挑战。虽然一些动态新视角合成的方法会产生2D运动作为副产品，但这些系统需要已知的相机姿态，而且生成的运动通常是错误的[21]。一些动态重建方法[9,76,79,80]也可以产生2D运动，但这些方法通常以物体为中心，关注关节物体。另外，基于视频分解的表示方法，如分层神经图集[32]和可变形精灵[81]，通过求解每个帧与全局纹理图集之间的映射来进行。可以通过反转此映射来得到帧与帧之间的对应关系，但这个过程既昂贵又不可靠。此外，这些方法只能使用有限数量的固定排序层来表示视频，限制了它们对于建模复杂真实世界视频的能力。

4. OmniMotion representation

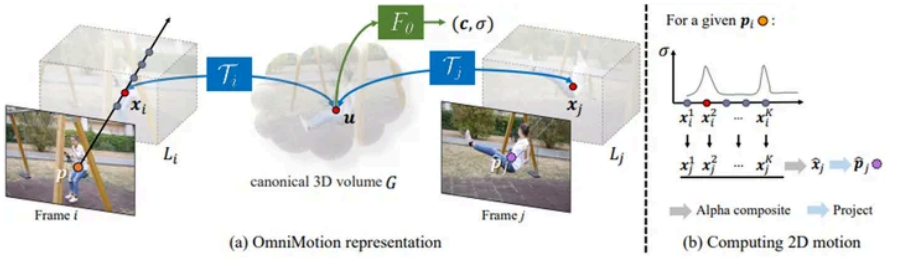

Figure 2: *Method overview*. (a) Our OmniMotion representation is comprised of a canonical 3D volume G and a set of bijections $\mathcal{T}_i$ that map between each frame's local volume L_i and the canonical volume G. Any local 3D location x_i in frame i can be mapped to its corresponding canonical location u through $\mathcal{T}_i$, and then mapped back to another frame j as x_j through the inverse mapping $\mathcal{T}_j^{-1}$. Each location u in G is associated with a color c and density σ, computed using a coordinate-based MLP F_θ. (b) To compute the corresponding 2D location for a given query point p_i mapped from frame i to j, we shoot a ray into L_i and sample a set of points $\{x_i^k\}_{k=1}^K$, which are then mapped first to the canonical space to obtain their densities, and then to frame j to compute their corresponding 3D locations $\{x_j^k\}_{k=1}^K$. These points $\{x_j^k\}_{k=1}^K$ are then alpha-composited and projected to obtain the 2D corresponding location $\hat{p}_j$.

由于OmniMotion并没有明确地将相机运动和场景运动解耦，因此得到的表示并不是一个物理上准确的3D场景重建，**我们称之为准3D表示。这种对动态多视角几何的松弛能够避免动态3D重建中的不确定性。**
我们保留了在遮挡情况下进行一致且准确的长期跟踪所需的属性：

1. 首先，通过在每个局部帧和一个规范帧之间建立双射关系，OmniMotion保证了在所有局部帧上全局循环一致的3D映射，模拟了现实世界中度量3D参考帧之间的一对一对应关系。
2. 其次，OmniMotion保留了关于投影到每个像素上的所有场景点以及它们的相对深度排序的信息，**使得即使这些点暂时被遮挡，也能够对其进行跟踪。**

4.1. Canonical 3D volume

我们使用一个canonical volume

代表视频内容，该体积充当了观察场景的三维地图集。与NeRF [44]类似，我们在上定义了一个基于坐标的网络，将每个canonical 3D坐标映射到密度和颜色。 密度很重要，它告诉我们canonical space中的surfaces位置。结合3D bijections，我们能够跟踪surfaces在多帧中的运动，并推断出遮挡情况下的关系。颜色，允许我们在优化过程中计算光照损失。

4.2. 3D bijections

我们定义了一个连续bijective mapping，将每帧的局部坐标系中的3D点映射到canonical 3D坐标系中，即，其中是帧索引。注意，canonical坐标与时间无关，可以视为特定场景点或3D轨迹在时间上的全局一致的“索引”。通过组合这些bijective mappings及其逆映射，我们可以将一个局部3D坐标系中的3D点映射到另一个中 。 Bijective mappings可以确保3D点在每帧中的对应关系是循环一致的，因为它们来自相同的canonical point。

为了能够捕捉真实世界运动的expressive maps，我们将这些bijections进行参数化为invertible neural networks（INNs）。受到最近homeomorphic shape modeling [35, 49]的启发，我们使用Real-NVP [14]，因为它具有简单的表达式和解析可逆性。Real-NVP通过组合简单bijective transformations（称为affine coupling layers）来构建bijective mappings。一个affine coupling layer将输入分成两部分：第一部分保持不变，但用于参数化施加于第二部分的仿射变换。

我们修改了这个架构，还引入了每帧的latent code[35, 49]作为condition。然后，所有的可逆映射都由相同的可逆网络进行参数化，但是使用不同的latent code:

从5.4. Network中可以看出，latent code是关于时间 的。
从附录中可以看出 负责转换canonical空间 的3D坐标 和对应射线上坐标的。

4.3. Computing frame-to-frame motion

有了上述表示后，那么如何得到第帧中任一像素的2D运动呢？直觉上，我们可以通过在光线方向上采样多个点将被查询像素“提升”到3D空间，然后使用bijections将这些3D点“映射”到目标帧，通过alpha合成“渲染”出3D点，最后“投影”回2D空间以获得一个对应关系。

5. Optimization

优化过程的输入包括一个视频序列和一组运动预测（现有方法的输出，可能带有噪声）作为指导。优化过程为整个视频输出完整的、全局一致的运动估计。

5.1. Collecting input motion data

在我们的大部分实验中，我们使用RAFT [66]来计算input pairwise correspondence。我们还尝试了另一种dense correspondence方法，TAP-Net [15]。评估结果展示了我们的方法在不同类型的输入对应关系下具有良好的一致性。以RAFT为例，我们首先通过穷举计算所有成对的光流。由于光流方法在大位移下可能产生明显错误，我们采取循环一致性和外观一致性检查来过滤掉虚假的对应关系。当认为结果可靠时，我们还可以选择性地利用chaining来augment the flows。更多详细信息可以在补充材料中找到。

过滤之后，不完整的光流场仍然存在噪声和不一致性。下面介绍我们的优化方法，将这些嘈杂、不完整的pairwise motion整合成完整的、准确的long-range motion。

5.2. Loss functions

主要损失函数是flow loss。目标是最小化预测flow与监督flow之间的MAE：

最后，为了确保估计的3D运动在时域上的平滑性，我们加入了正则化项去惩罚较大的加速度。例如，对于第帧中的3D位置，我们使用公式1将其分别映射到第和帧，得到和

这种优化思想利用了到单个canonical volume的bijections、光照一致性、以及基于坐标的网络

和提供的时空平滑性，来调和不一致的pairwise flow，并填充对应graph中缺失的内容。

5.3. Balancing supervision via hard mining

穷举的成对流输入使优化阶段可用的有用运动信息最大化。然而，这种方法，特别是当与流过滤过程相结合时，可能导致动态区域中运动样本的不平衡收集。刚性背景区域通常具有许多可靠的成对对应关系，而快速移动和变形的前景对象在过滤后通常具有少得多的可靠对应关系，尤其是在遥远的帧之间。这种不平衡可能导致网络完全关注主要（简单）背景运动，而忽略代表监控信号一小部分的具有挑战性的运动对象。

为了解决这个问题，我们提出了一个简单的策略，用于在训练中挖掘困难的例子。具体来说，我们周期性地缓存流预测，并通过计算预测流和输入流之间的欧几里得距离来计算误差图。在优化过程中，我们指导采样，以便更频繁地对具有高误差的区域进行采样。我们在连续帧上计算这些误差图，我们认为我们的监控光流是最可靠的。有关更多详细信息，请参阅补充资料。

5.4. Implementation details

Network. 主要有以下3个网络：详见附录。

- 输出latent code的网络，使用了一个2层MLP（每层256通道）作为GaborNet [19]来计算每帧的latent code。这个MLP的输入是时间，输出的维度为128。
- canonical representation
- ，也是一个GaborNet（3层512个通道），其输入以上两个网络的输出，输出是预测的颜色和密度，公式为
  Representation. 我们将所有像素坐标

  归一化到[-1, 1]，并将深度范围设置为[0, 2]，并为每帧定义一个局部的3D空间。虽然这个方法可以将内容放置在canonical space 的任意位置，但我们仍然用初始化得到大致位于单位球内的canonical locations，这样确保输入进是well-conditioned。
  为了提高训练过程中的数值稳定性，我们将canonical 3D 坐标进行了缩放后再传递给
  ，这与mip-NeRF 360 [3]中的操作类似。

6. Evaluation

评测结果就不说了，感兴趣的去看论文吧（个人觉得方法更重要）。

Supplement

A. Preparing pairwise correspondences

我们直接使用现有方法（如RAFT[66]和TAP-Net[15]）的pairwise correspondences，并将它们合并为覆盖整个视频的密集、全局一致和准确的对应关系。作为预处理阶段，我们详尽地计算所有成对的对应关系，并使用循环一致性和外观一致性检查对其进行过滤。

当计算基础帧和目标帧之间的flow field 时，如果可能，我们总是用前一个预测作初始化。虽然这改善了远距离帧之间的预测效果，但是远距离帧之间的流预测仍可能包含显著误差，因此我们过滤掉那些周期一致性误差（i.e., forward-backward flow consistency error）大于3个像素的flow vector estimates。

Figure 5: *Erroneous correspondences after cycle consistency check*. The red bounding box highlights a common type of incorrect correspondences from flow networks like RAFT [66] that remains undetected by cycle consistency check. The left images are query frames with query points and the right images are target frames with the corresponding predictions. Only correspondences on the foreground object are shown for better clarity.

经过上述过滤之后，我们仍然经常观察到一种持续的错误（无法通过循环一致性检查来检测）。**图5展示了这类spurious correspondence，我们推测这是因为flow networks难以估计两帧之间显著变形区域的运动，于是转向对周围区域的运动进行插值。例如在图5的例子中，前景人员的flow被“锁定”在背景层上。由于这些不正确的flows与第二层运动（例如，背景运动）一致，于是通过了循环一致性检查。**

为了解决这个问题，我们增加了外观检查：用DINO[10]为每个像素提取密集特征，并过滤掉特征余弦相似度＜0.5的对应关系。在实践中，我们对所有pairwise flows采取循环一致性检查，当两帧相距超过3帧时增加外观检查。我们发现，这种过滤过程能够在不针对每个序列进行调整的情况下，很好消除了不同序列中流场中的主要错误。图6中展示了经过循环一致性和外观一致性检查后的结果。

Figure 6: *Correspondences from RAFT [66] after both cycle and appearance checks*. The left column shows a single query frame, and the right column displays target frames with increasing frame distances to the query frame from top to bottom. The filtered correspondences are reliable without significant errors.

这种方法的一个缺点是，当一些区域在目标帧中突然被遮挡时，这些正确的flows也会被过滤掉。对于某些correspondence methods（如RAFT），在遮挡期间保留这些运动信号可以得到更好的运动估计结果。

因此，我们设计了一种简单的策略来检测遮挡区域的可靠流。对于每个像素，我们计算其到目标帧的正向流a、循环流b（从目标像素返回到源帧的流）、以及第二个正向流c。这个过程有两次循环一致性检查：a和b之间的一致性形成标准的循环一致性检查，而b和c之间的一致性形成一个次要的、补充性的一致性检查。**我们找出a和b不一致但b和c一致的像素，并将其视为遮挡像素。**我们发现这种方法在识别遮挡区域的可靠流方面很有效，特别是当两帧之间的时间距离较近时。因此，如果这些对应关系跨越小于3帧的时间距离，我们允许它们绕过循环一致性检查。我们的实验在使用RAFT流的方法变体中使用了这个额外的信号，但对于TAP-Net变体则没有使用，因为我们发现后者在遮挡事件附近预测的对应关系不太可靠。

C. Additional implementation details

Network architecture 图7中展示了将local坐标系和canonical坐标系之间映射的可逆网络的架构。由六个affine coupling layers组成（图7中仅显示了第一层），这六个layers带有alternating split patterns。每个layer的可学习组件是一个3层MLP（每层具有256个通道），输入是一帧的latent code 和输入坐标，输出是一个尺度s和一个平移t，该尺度和平移被应用于输入坐标上得到。所有输入坐标都经过相同处理。我们发现，对输入坐标加上位置编码[44]可以提高MLP的拟合能力，我们将频率数量设置为4。

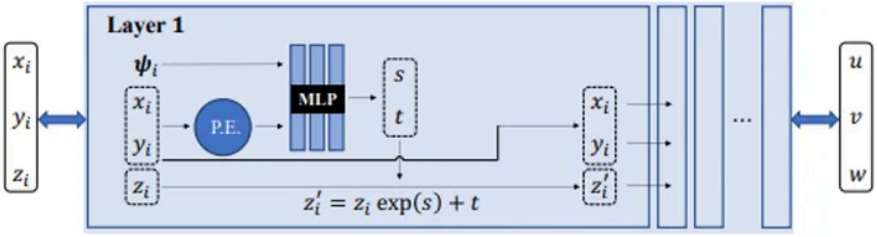

Figure 7: *Network architecture for the mapping network* M_{θ}. We show the first affine coupling layer, which is representative of the subsequent layers, except for the different splitting patterns used. As mentioned in the main paper, this architecture is fully invertible, i.e., it can be queried in either direction, from (u, v, w) to (x, y, z) and vice-versa.

感觉还是采用的Nerf那一套，和CVPR Best Paper一样，把工程量堆上去了，学术上创新欠缺了些。