

Tracking Everything Everywhere All at Once



1. Related Work

Optical flow: 光流传统上被表述为一个优化问题，最近的进展使利用神经网络直接预测光流成为可能，提高了质量和效率。虽然光流方法允许在连续帧之间进行精确的运动估计，但它们不适合进行远程运动估计：

2. OmniMotion representation

OmniMotion 是一个新的测试时间优化方法，用于从视频序列中估计密集和长距离运动。； OmniMotion 将视频用准三维规范体表示，并通过局部和规范空间之间的双射（满足单射和满射）映射到每一帧的局部 volume。local-canonical bijection 被参数化为神经网络。并捕捉摄像机和场景的运动，不解开两者。可以被认为是从一个固定的，静态的照相机产生的local volume 的渲染。

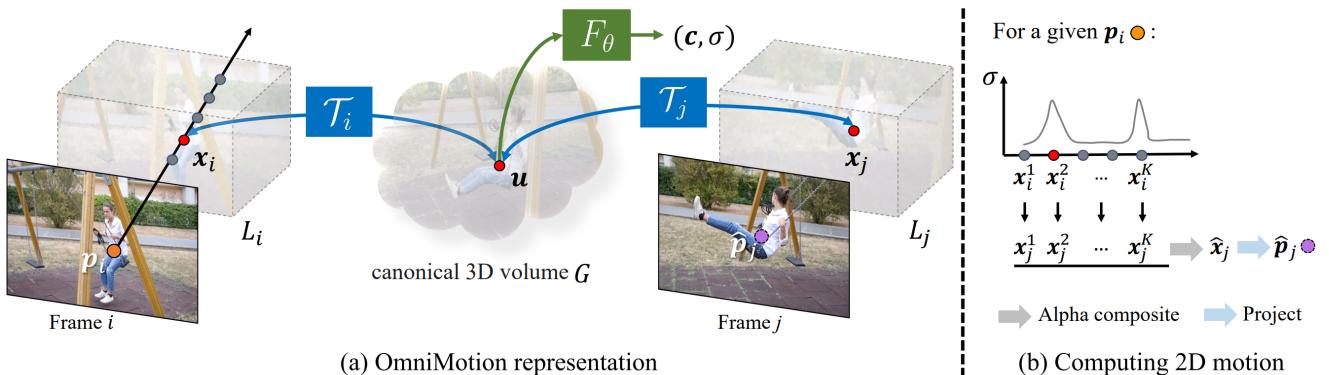


Figure 2: Method overview. (a) Our OmniMotion representation is comprised of a canonical 3D volume G and a set of bijections \mathcal{T}_i that map between each frame's local volume L_i and the canonical volume G . Any local 3D location x_i in frame i can be mapped to its corresponding canonical location 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 local 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 .

Canonical 3D volume

G 作为观测场景的三维图集，在 G 上定义一个基于坐标的网络 F_θ 将每个规范坐标 $u \in G$ 映射到一个密度 σ 和颜色 c 。存储在 G 中的密度是告诉曲面在规范空间中的位置。结合3D bijections，允许我们在多个帧上跟踪 surfaces，以及关于遮挡关系的原因；存储在 G 中的颜色允许在优化过程中计算photometric loss。

3D bijections

连续的双射映射 \mathcal{T}_i ，将来自每个局部坐标系 L_i 的三维点 x_i 映射到规范的三维坐标系 $u = \mathcal{T}_i(x_i)$ ，其中 i 是帧索引。可以从一个局部的三维坐标中映射一个frame (L_i) to another (L_i):

$$\mathbf{x}_j = \mathcal{T}_j^{-1} \circ \mathcal{T}_i(\mathbf{x}_i) \quad (1)$$

为了允许能够捕捉真实世界运动的表达性地图，将双射参数化为可逆神经网络（INNs）。使用Real-NVP (Real-NVP组合被称为仿射耦合层的简单双射变换来构建双射映射。仿射耦合层将输入分割成两部分；第一部分保持不变，但用于参数化，应用于第二部分的仿射变换。)。修改这个架构，condition on a per-frame latent code ψ_i , 然后所有的可逆映射 T_i 都由相同的可逆网络 M_θ 参数化，but with different latent codes: $T_i(\cdot) = M_\theta(\cdot; \psi_i)$.

Computing frame-to-frame motion

通过在射线上的采样点将查询像素“提升”到3D，使用双射 \mathcal{T}_i 和 \mathcal{T}_j “映射”这些3D points 到一个目标帧 j ，通过 alpha compositing “渲染”这些来自不同样本的3D points，最后“project”回2D以获得假定的对应关系。

假设相机运动包含了 local-canonical bijections \mathcal{T}_i 并且使用fixed, orthographic camera。在 \mathbf{p}_i (query pixel) 处的射线定义为 $\mathbf{r}_i(z) = \mathbf{o}_i + z\mathbf{d}$ ，在射线上采集 K 个样本 \mathbf{x}_i^k ，这相当于在 \mathbf{p}_i 上附加一组深度值 $\{z_i^k\}_{k=1}^K$ 。将这些样本映射到规范空间，然后查询密度网络 F_θ ，来获得这些样本的密度和颜色。以第 k 个样本 \mathbf{x}_i^k 为例，它的密度和颜色可以写成 $(\sigma_k, \mathbf{c}_k) = F_\theta(M_\theta(\mathbf{x}_i^k; \psi_i))$ 。我们还可以沿着每个样本的射线映射到帧 j 中相应的三维位置 $\hat{\mathbf{x}}_j^k$ (Eq. 1)。

聚合所有样本的对应样本 $\hat{\mathbf{x}}_j^k$ ，以生成一个单一的对应 $\hat{\mathbf{x}}_j$ 。这种聚合类似于NeRF中样本点的颜色的聚合方式：使用alpha合成，第 k 个样本的alpha值为 $\alpha_k = 1 - \exp(-\sigma_k)$ 。 $\hat{\mathbf{x}}_j$ 为：

$$\hat{\mathbf{x}}_j = \sum_{k=1}^K T_k \alpha_k \mathbf{x}_j^k, \quad \text{where} \quad T_k = \prod_{l=1}^{k-1} (1 - \alpha_l)$$

用类似的过程合成 \mathbf{c}^k ，得到 \mathbf{p}_i 的图像空间颜色 \hat{C}^i 。

3. Loss functions

$$\mathcal{L}_{flo} = \sum_{f_{i \rightarrow j} \in \Omega_f} \|\hat{f}_{i \rightarrow j} - f_{i \rightarrow j}\|_1 \quad (3)$$

$\hat{f}_{i \rightarrow j} = \hat{\mathbf{p}}_j - \mathbf{p}_i$ ($\hat{\mathbf{p}}_j$ 是使用固定正交相机模型对 $\hat{\mathbf{x}}_j$ 进行投影，得到查询位置 \mathbf{p}_i 的预测的二维相应位置 $\hat{\mathbf{p}}_j$)； Ω_f 是 the set of all the filtered pairwise flows

$$\mathcal{L}_{pho} = \sum_{(i, \mathbf{p}) \in \Omega_p} \|\hat{C}^i(\mathbf{p}) - C_i(\mathbf{p})\|_2^2 \quad (4)$$

Ω_p 是所有帧上的所有像素位置的集合。

为了保证 M_θ 估计的三维运动的时间平滑性，我们应用了一个正则化项来惩罚大的加速度。给定帧 i 中的采样三维位置 \mathbf{x}_i ，我们使用等式 1 将其映射到帧 $i-1$ 和帧 $i+1$ ，分别产生3D点 \mathbf{x}_{i-1} 和 \mathbf{x}_{i+1} ，然后最小化3D加速度：

$$\mathcal{L}_{reg} = \sum_{(i, \mathbf{x}) \in \Omega_x} \|\mathbf{x}_{i+1} + \mathbf{x}_{i-1} - 2\mathbf{x}_i\|_1 \quad (5)$$

Ω_x 是所有帧的局部三维空间的并集

分别最小化公式 3、4、5，最终 loss function 为

$$\mathcal{L} = \mathcal{L}_{flo} + \lambda_{pho} \mathcal{L}_{pho} + \lambda_{reg} \mathcal{L}_{reg}$$

权重 λ 控制了每个项的相对重要性。