RAFT：递归全对域变换方法计算光流

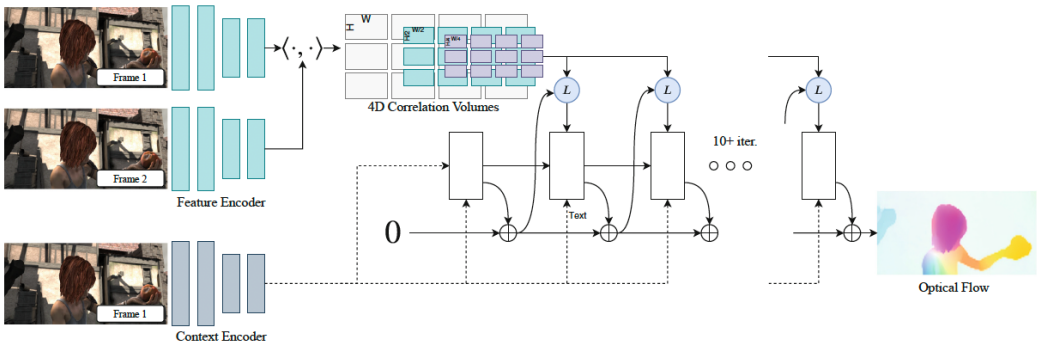
1. RAFT 概述

光流预测是估计视频帧之间每像素运动的任务。传统上，光流被视为一对图像之间，稠密位移场空间上的手工优化问题，往往需要在相似图像区域匹配和运动合理性先验之间做出权衡，以达到优化目标。然而，传统光流预测受限于快速移动物体、遮挡、运动模糊和无纹理表面等问题，同时在复杂角情况下难以保持鲁棒性。

另外，传统光流预测同其他传统视觉任务相同，都存在需要人工优化具体模型以适配具体任务的问题，即传统光流预测对图像的特征提取、运动先验等任务都需要人工优化而非自主学习。因此，具有自主迭代优化的基于深度学习的计算机视觉任务逐渐流行，并涌现了大量将传统视觉思想与深度学习方法相结合的工作。

RAFT（Recurrent All-Pairs FieldTransforms for Optical Flow）是以传统光流预测为启发，用卷积神经网络提取图像特征，对相邻帧的特征进行不同尺度下的全对相关性计算得到相关性金字塔，最后用 GRU(Gated Recurrent Unit）更新并预测光流。

2 技术细节



图表 1 RAFT 模型架构图

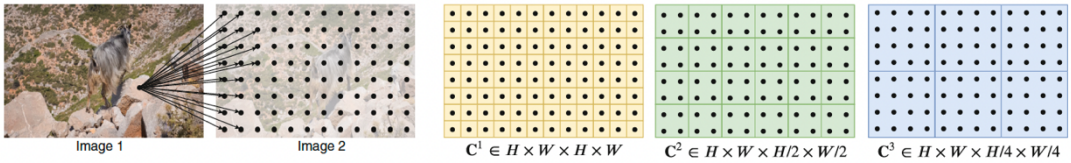
2.1 特征提取

使用卷积网络从输入图像中提取特征。特征编码器网络同时应用于 Image1 和 Image2，并以较低的分辨率将输入图像映射到密集的特征图。特征编码器g_θ以

1/8 分辨率输出特征：$\mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H/8\times W/8\times D}$，其中 D = 256。特征编码器由 6 个残差块组成，2 个为 1/2 分辨率，2 个为 1/4 分辨率，2 个为 1/8 分辨率。

此外，该方法还使用了上下文网络。上下文网络仅从第一个输入图像 Image1 中提取特征。该特征提取器h_θ与g_θ相同。特征网络g_θ和上下文网络h_θ共同构成了该方法的第一阶段.

2.2 计算视觉相似度



图表 2 相关体计算与相关性金字塔构建

RAFT 通过构建所有特征向量对之间的完整相关体来计算视觉相似性。给定图像特征$g_\theta(I1) \in \mathbb{R}^{H\times W\times D}$和$g_\theta(I2) \in \mathbb{R}^{H\times W\times D}$，通过取所有特征向量对之间的点积形成相关体。相关体的维度及其计算方式如下:

$$C\big(g_\theta(I_1), g_\theta(I_1)\big) \in \mathbb{R}^{H\times W\times H\times W}$$

$$C_{ijkl} = \sum_h g_\theta(I_1)_{ijh} \cdot g_\theta(I_2)_{klh}$$

RAFT 通过将相关体的最后两个维度与核大小 1、2、4 和 8 和等效步长池化来构建一个 4 层金字塔$\{C1, C2, C3, C4\}$（图 2）。因此，体积C_k的尺寸为$H\times W\times H/2^k\times W/2^k$。相关性金字塔给出了大位移和小位移的信息；然而，通过保持前两个维度（$I1$维度）保持了高分辨率信息，允许恢复快速移动的小物体的运动。

2.3 迭代更新

RAFT 的更新算子从初始起点$f_1 = 0$开始，估计出一系列光流估计值$\{f_1, \ldots, f_N\}$。在每次迭代中，它都会生成一个更新方向Δf，并将其应用于当前估计值:$f_{k+1} = \Delta f + f_k$。

更新操作符的一个核心组件是基于 GRU 单元的门控激活单元，其中全连接层被卷积层所取代。GRU 输出的隐藏状态通过卷积层来预测流量更新量 Δf。输

出的流量分辨率是输入图像的 1/8。在训练和评估期间，将预测的流量场进行上采样，以匹配真实值的分辨率。

3 模型表现

RAFT 的设计借鉴了许多现有作品，但具有显著的创新性。首先，RAFT 维护并更新一个单一的高分辨率流场。这与流行的由粗到细的设计不同，在这些工作中，流场首先在低分辨率下估计，然后在高分辨率下进行上采样和细化。通过在单一高分辨率流场上操作，RAFT 克服了由粗到细级联的几个局限性：从低分辨率错误中恢复的困难、容易遗漏小的快速移动物体以及训练多阶段级联通常所需的大量训练迭代（通常超过 100 万次）。

其次，更新操作符具有新颖的设计，它由一个卷积门控循环单元（GRU）组成，该单元在 4D 多尺度相关体积上执行查找操作；相比之下，先前工作中的细化模块通常仅使用普通卷积层或相关层。

在 KITTI 数据集上，RAFT 实现了 5.10% 的 F1-all 错误率，比已发表的最佳结果（6.10%）降低了 16%。在 Sintel（最终帧）数据集上，RAFT 获得了 2.855 像素的端点误差，比已发表的最佳结果（4.098 像素）降低了 30%。此外，RAFT 具有强大的跨数据集泛化能力，并且在推理时间、训练速度和参数数量方面都表现出高效性。