GMFlow: Learning Optical Flow via Global Matching

Haofei Xu[1*] Jing Zhang[2] Jianfei Cai[1] Hamid Rezatofighi[1] Dacheng Tao[3,2]

摘要

基于学习的光流估计一直被带有流回归的卷积的框架所主导。这种框架本质上局限于局部相关性，因此很难应对光流领域长期存在的大位移这一挑战。为缓解这一问题，当前的代表性算法 RAFT 通过大量的迭代改进来逐步提升其性能，但这种序列化的处理方式使得推理时间增大了很多。为了实现高效高精度的光流估计，我们将光流重新定义为一个全局匹配问题。我们提出了一个 GMFlow 框架，它由三个主要部分组成：用于特征增强的 Transformer，用于全局匹配的相关性和 softmax 层，以及一个用于光流传播的自注意力层。此外，我们进一步使用了一个细化步骤，在高分辨率复用 GMFlow 进行残差流估计，取得了很好的性能。我们使用一次细化步骤的网络的性能超过了细化 31 次的 RAFT，并且我们的运行速度更快。

关键词：Transformer；光流估计；全局匹配

1 引言

自从开创性的基于学习的 FlowNet[1]以来，卷积回归已经应用在光流上很久了。为了将匹配信息编码到网络中，成本量（相关性）被视为一个重要因素应用于流行的框架。但基于回归的方法中，搜索空间被视为后续卷积回归的通道维度，导致成本量是一个预定义的大小，使得搜索空间被限制在一个局部范围内，因此难以处理大位移。

为了解决大位移问题，RAFT[2]提出了一种具有大量迭代改进的框架，在不同迭代阶段将卷积应用于不同的局部变量，从而逐渐实现全局搜索，但大量的迭代使得推理时间线性的增加，限制了它的实际应用。

受图像对之间的稀疏匹配启发，我们不再在预定义的局部成本量上运行额外的卷积层，而是将光流作为一个全局匹配问题。图 1提供了这两种流量估计方法的概念比较。Transformer 提供了更具有辨别力的特征表示，通过比较特征相似性，我们的流量预测可以通过可区分的匹配层（相关性层和 softmax 层）获得。

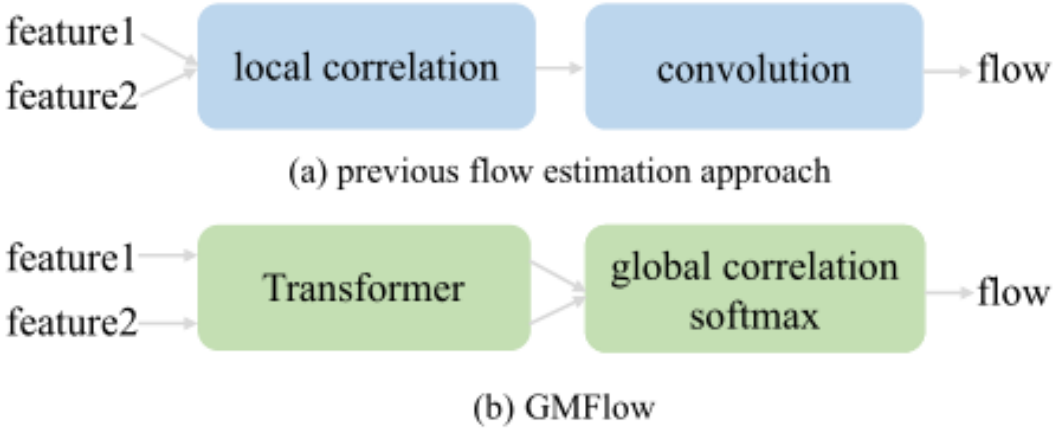

图 1: 流量估计方法的概念比较

在本文中，我们提出来一个 GMFlow 框架来实现光流的全局匹配公式。具体来说，我们先从卷积骨干网络中提取出密集特征输入到由自注意力、交叉注意力和前馈网络组成的 Transformer 中来实现特征增强。然后，我们通过关联所有成对特征来比较特征相似性。之后，我们使用可微分的 softmax 匹配层获得流估计。为了解决被遮挡和边界外的像素问题，我们加入一个额外的自我注意层，通过利用特征自相似性将高质量的流预测从匹配的像素传播到不匹配的像素。我们进一步介绍了一个细化步骤，在更高的特征分辨率复用 GMFlow 进行残差流预测，取得很好的性能。我们的贡献：

● 我们将光流估计重新表达为全局匹配问题，彻底改变了主流的流回归模型，从而有效的解决了大位移问题。

● 我们提出了一个 GMFlow 框架来实现全局匹配公式。

● 我们进一步提出了一个细化步骤来利用更高分辨率的特征，使得我们能够使用相同的 GMFlow 框架来进行残差流估计。

●GMFlow 在 Sintel 测试中的表现优于 31 次迭代的 RAFT，且运行速度更快。

2 相关工作

2.1 流估计方法

流估计方法是现有流行光流框架的基础，尤其是由粗到细的方法 PWC-Net[3]和迭代细化方法 RAFT[2]。它们都执行某种多阶段的改进，无论在多个尺度还是在单一分辨率上。对于每个阶段的流量预测，它们在流程上的概念是相似的，即从具有卷积的局部成本量回归光流，但这种方法难以处理大位移。但也有两个例外，DICL[4]使用卷积执行局部匹配，GLU-Net[5]使用卷积从全局相关性中回归流，但这也使其自身受限于固定的图像分辨率，因此需要额外的子网络来处理这个问题。与这些方法不同的是，我们使用 Transformer 执行全局匹配，并且我们表明确实有可能在不依赖大量改进的情况下获得高度准确的结果。

2.2 大位移

大位移一直是光流估计长期面临的问题。为了缓解这个问题，一种方法是用从粗到细的方法来逐步估计大位移。但如果分辨率太粗糙，从粗到细的方法往往会错过快速移动的小物体。RAFT[2]则保持单一的高分辨率并通过大量迭代改进逐渐改建初始预测，但使得推理时间大量增加。相比之下，我们的新框架简化了光流估计流程并以高精度和高效率估计大位移，这是通过重新制定光流问题和强大的 Transformer 实现的。

2.3 transformer

我们使用更简单的 softmax 操作和简单的流传播层来处理遮挡。我们的目标是学习强大的特征表示（特别是交叉注意力）以进行匹配。

3 本文方法

光流估计本质上是一个匹配问题，旨在找出对应的像素点。为此，我们可以比较每一个像素的特征相似度，并确定给出最高相似度的对应像素，这样的过程要求特征的辨识度很高。聚合图像本身的空间上下文和来自另一幅图像的信息可以直观地减少歧义并提高它们的独特性。这种设计理念使稀疏

特征匹配框架取得了巨大成就。稀疏匹配的成功带来了较大的观点变化，促使我们将光流表述为显式全局匹配问题，以应对大位移的挑战。

3.1 全局匹配公式

给定两个连续的视频帧 I_1 和 I_2，我们首先使用权重共享卷积网络提取下采样的密集特征 $F_1, F_2 \in R^{H*W*D}$，其中 H、W 和 D 分别表示高度、宽度和特征维度。由于两帧中的对应关系应该具有很高的相似性，我们首先通过计算它们的相关性来比较 F_1 中每个像素与 F_2 中所有像素的特征相似性，可以通过矩阵乘法来实现：

$$C = \frac{F_1 F_2^T}{\sqrt{D}} \in R^{H*W*H*W} \tag{1}$$

其中，相关矩阵 C 中的每个元素表示 F_1 中坐标 $p_1(i,j)$ 与 F_2 中 $p_2(k,l)$ 之间的相关值，$\frac{1}{\sqrt{D}}$ 是归一化因子，避免点积后的值过大。

为了识别对应关系，最常用的方法是直接获取相关性最高的位置，但这个操作是不可微分的，这阻碍了端到端的训练。为了解决这一问题，我们使用了一个可区分的匹配层，我们用 softmax 操作对 C 的最后两个维度进行归一化，得到一个匹配分布：

$$M = softmax(C) \in R^{H*W*H*W} \tag{2}$$

我们可以通过对具有匹配分布 M 的像素网络 $G \in R^{H*W*2}$ 的二维坐标进行加权平均来获得对应关系 $\hat{G}$:

$$\hat{G} = MG \in R^{H*W*2} \tag{3}$$

最后计算相应坐标之间的差值就可以得到光流 V:

$$V = \hat{G} - G \in R^{H*W*2} \tag{4}$$

这种基于 softmax 的方法不仅可以实现端到端训练，还可以提供亚像素精度。

3.2 特征增强

由于 F_1 和 F_2 只有两组特征，没有空间位置的概念，因此我们首先向特征添加固定的 2D 正弦和余弦位置编码。添加位置信息使得匹配过程不仅考虑特征相似性，还考虑他们的空间距离，有助于解决歧义和提高性能。

添加位置信息后，我们使用 6 个堆叠的自注意力，交叉注意力和前馈网络（FFN）块来提高初始特征的质量。对于自注意力，注意力机制中的 query、key 和 value 是相同的特征。对于交叉注意力，key 和 value 相同，但与 query 不同以引入它们的相互依赖关系。这个过程对 F_1 和 F_2 对称地执行，即：

$$\hat{F_1} = \tau(F_1 + P, F_2 + P), \hat{F_2} = \tau(F_2 + P, F_1 + P) \tag{5}$$

其中 τ 是一个 Transformer，P 是位置编码，τ 的第一个输入是 query，第二个是 key 和 value。

为了提高效率，我们采用了 Swim Transformer[6]的转移局部窗口注意力策略。与使用固定窗口大小的 Swim 不同，我们将特征拆分为固定数量的局部窗口，以使窗口大小与特征大小自适应。具体来说，我们将大小为 $H*W$ 的输入特征拆分为 $K*K$ 个窗口，每个窗口的大小为 $\frac{H}{K}*\frac{W}{K}$，并在每个局部窗口内独立执行自注意力和交叉注意力。对于每两个连续的局部窗口，我们将窗口分区移动 $(\frac{H}{2K}, \frac{W}{2K})$，以引入跨窗口连接。在我们的工作中，我们拆分为 $2*2$ 个窗口，每个窗口大小为 $\frac{H}{2}*\frac{W}{2}$，体现了速

度与精度的权衡。

3.3 流传播策略

我们基于 softmax 的流估计方法隐含地假设相应的像素在两个图像中都是可见的，因此可以通过比较它们的相似性来匹配它们。然而，这个假设对于被遮挡和边界外的像素是无效的。为了解决这个问题，如图 2所示，通过观察光流场和图像本身具有很高的结构相似性，我们建议通过测量特征自相似性将匹配像素中的高质量流预测传播到不匹配的像素。这个操作可以通过一个简单的自注意力层有效地实现：

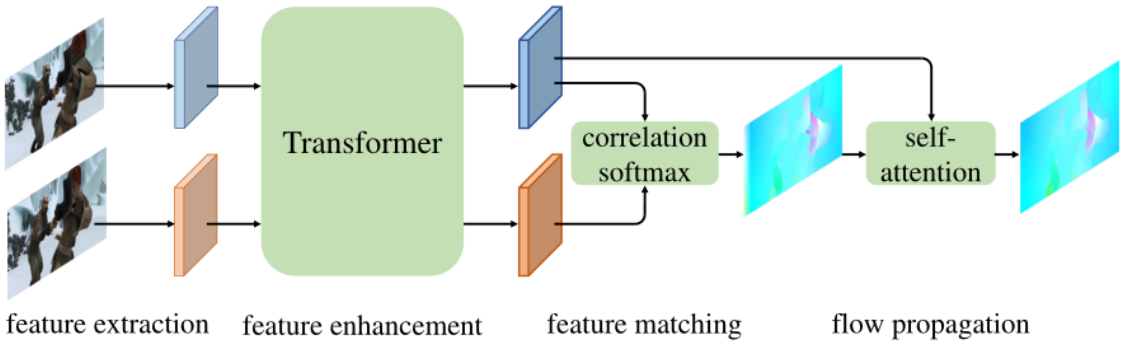

图 2: GMFlow 框架总览

$$\tilde{V} = softmax(\frac{\hat{F}_1\hat{F}_1^{\ T}}{\sqrt{D}})\hat{V} \in R^{H*W*2} \tag{6}$$

其中

$$\hat{V} = softmax(\frac{\hat{F}_1\hat{F}_2^{\ T}}{\sqrt{D}})G - G \tag{7}$$

3.4 细化

到目前为止提出的框架（基于 1/8 的特征）已经可以实现具有竞争力的性能。它可以通过引入额外的更高分辨率 (1/4) 特征进行细化来进一步改进。具体来说，我们首先将之前的 1/8 流量预测上采样到 1/4 分辨率，并用当前流量预测扭曲第二个特征。然后将细化任务简化为残差流学习，其中可以使用图 2 中描述的相同 GMFlow 框架，但在局部范围内。具体来说，我们在 Transformer 中分割成 8×8 个局部窗口（每个具有原始图像分辨率的 1/32），并对每个像素执行 9×9 局部窗口匹配。从 softmax 层获得流预测后，我们对流传播执行 3×3 局部窗口自注意力操作。

请注意，这里我们与全局匹配阶段共享细化步骤中的 Transformer 和自注意力权重，这不仅减少了参数，还提高了泛化能力。为了生成 1/4 和 1/8 特征，我们还共享主干特征。

具体来说，我们采用与 TridentNet[7]类似的方法，但分别使用步幅为 1 和 2 的权重共享卷积。这种权重共享设计也导致比特征金字塔网络更好的性能。

3.5 训练损失

$$L = \sum_{i=1}^{N} \gamma^{N-i} \|V_i - V_{gt}\|_1 \tag{8}$$

其中 N 是包括中间和最终预测在内的流量预测的数量，$\gamma = 0.9$ 是呈指数增长的权重，以便为后期预测提供更高的权重。

4 复现细节

4.1 与已有开源代码对比

GMFlow 已在 github 上开源，你可以通过这个链接来获取代码：https://github.com/haofeixu/gmflow。

4.2 实验环境搭建

代码使用 python3.8 和 pytorch1.9 实现，具体实验环境可以参照 enviroment.yml。

4.3 数据集

数据集共有五个：

- **FlyingChairs_release**：用于训练集
- **FlyingThings3D**: 用于训练集
- **HD1K**: 用于测试集
- **KITTI**: 用于测试集
- **Sintel**: 用于基准测试

4.4 创新点

GMFlow 打破了传统的光流估计流程，其创新点如下：

- 将光流估计重新表达为全局匹配问题。
- 很好的处理了光流估计中的“大位移”问题。

5 实验结果分析

GMFlow 可以实现双向光流估计，并且能够现实遮挡检测。输入的测试图像如图 3所示。我们对输入的测试图像做双向光流估计，得到的结果如图 4

图 3: 输入的测试图像

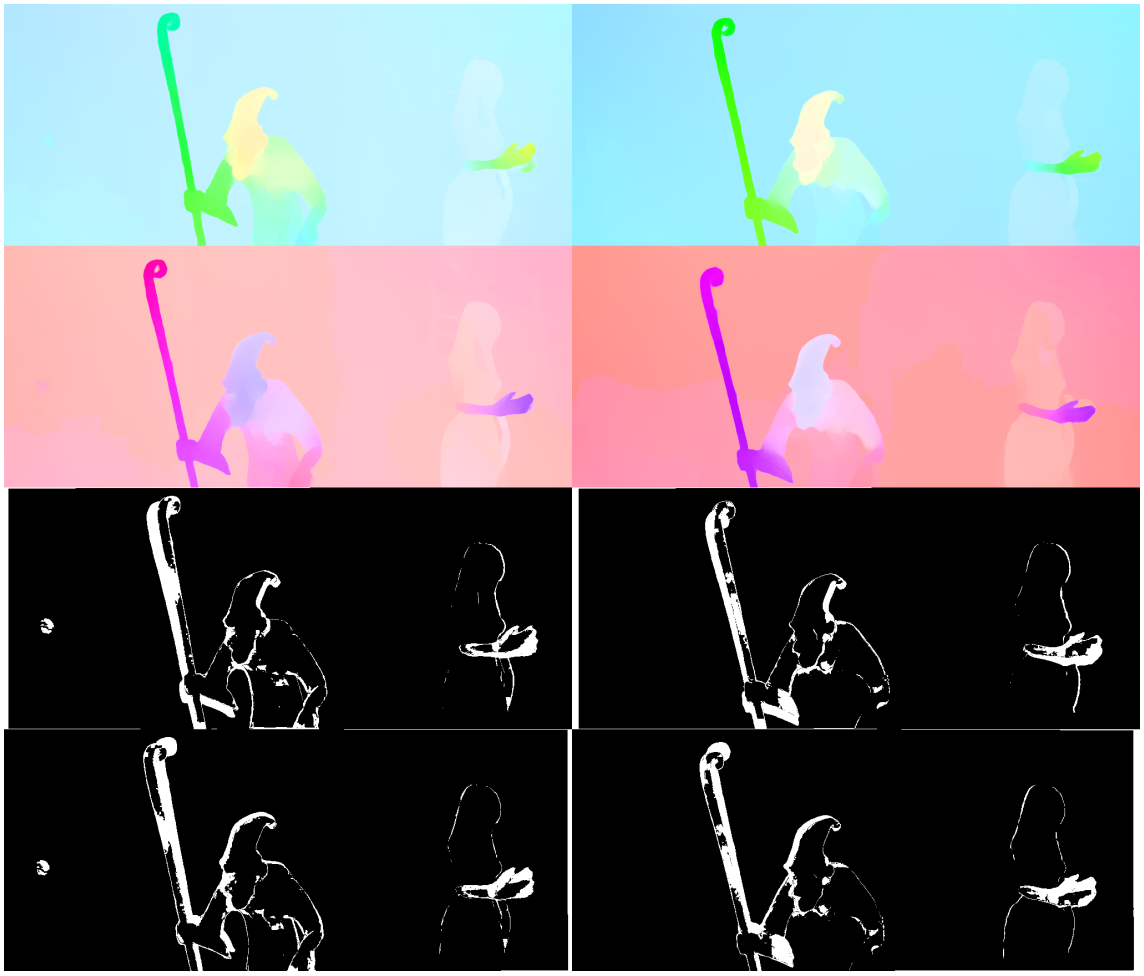

图 4: 双向光流估计与遮挡检测

6 总结与展望

我们的框架在遮挡区域仍有未来改进空间，此外，当训练集和测试集的数据有很大的区别时，我们的模型不能表现得很好。但幸运的是，目前有许多可用的大规模的数据集可以用来增强 Transformer 的泛化能力。

参考文献

[1] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2758-2766.

[2] TEED Z, DENG J. Raft: Recurrent all-pairs field transforms for optical flow[C]//European conference on computer vision. 2020: 402-419.

[3] SUN D, YANG X, LIU M Y, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8934-8943.

[4] WANG J, ZHONG Y, DAI Y, et al. Displacement-invariant matching cost learning for accurate optical flow estimation[J]. Advances in Neural Information Processing Systems, 2020, 33: 15220-15231.

[5] TRUONG P, DANELLJAN M, TIMOFTE R. GLU-Net: Global-local universal network for dense flow and correspondences[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern

recognition. 2020: 6258-6268.

[6] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[7] LI Y, CHEN Y, WANG N, et al. Scale-aware trident networks for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6054-6063.