# 综述报告——基于光流的视频插帧

姓名：张鹏杰 学号：3220231268

**摘要：**视频插帧是利用低帧率视频生成高帧率视频的技术，可以显著提高视频的流畅度和清晰度，光流法是一种常用的视频插帧方法，它通过计算相邻帧之间的像素运动来估计新的帧。本文分析总结了主要的光流插帧模型，展示了光流插帧技术的研究进展和现状，探讨了各种算法的原理，并在最后总结了光流插帧技术存在的主要矛盾，以及其未来的发展趋势。

**关键字：**视频插帧，光流，深度学习

# 引言

视频插帧（Video Interpolation）是一种[计算机视觉算法](https://so.csdn.net/so/search?q=%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%A7%86%E8%A7%89%E7%AE%97%E6%B3%95&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/q7w8e9r4/article/details/_blank)，用于在视频中插入额外的帧以提高视频的流畅度和观看体验。视频插帧技术可以应用于各种领域，比如电影特效、视频游戏、运动捕捉等，在近年来被广泛研究和应用。

基于光流的视频插帧通过计算视频相邻帧之间的双向光流，分析相邻帧之间的像素变化，从而推测出像素点在下一帧中的位置，然后使用合适的扭曲算法生成插值帧。传统的光流估计方法包括Lucas-Kanade[1][光流法](https://so.csdn.net/so/search?q=%E5%85%89%E6%B5%81%E6%B3%95&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/q7w8e9r4/article/details/_blank)、Horn-Schunck[2]光流法，基于深度学习的光流估计方法包括FlowNet[3]、PWC-Net[4]等。

近年来，深度学习方法在视频插帧领域取得了很大的突破，通过使用卷积神经网络（Convolutional Neural Networks，CNN）等深度学习框架，模型可以学习到视频中的时空特征，从而生成更真实的光流、更逼真的插值帧。由此，研究者们利用深度学习框架提出了许多基于光流的视频插帧模型，如SuperSloMo[5]、DAIN[6]、BMBC[7]、ABME[8]、RIFE[9]等。

本文共分为三章，第一章为引言，介绍了基于光流的视频插帧的基本概念，以及相关的模型方法，第二章为相关模型方法，以时间为线索对主要的模型方法进行了介绍，指明了各种方法主要的技术进展，第三章为总结，主要提出当前光流插帧技术存在的主要矛盾，以及可能的解决办法和研究方向。

# 相关模型方法

2.1 SuperSloMo

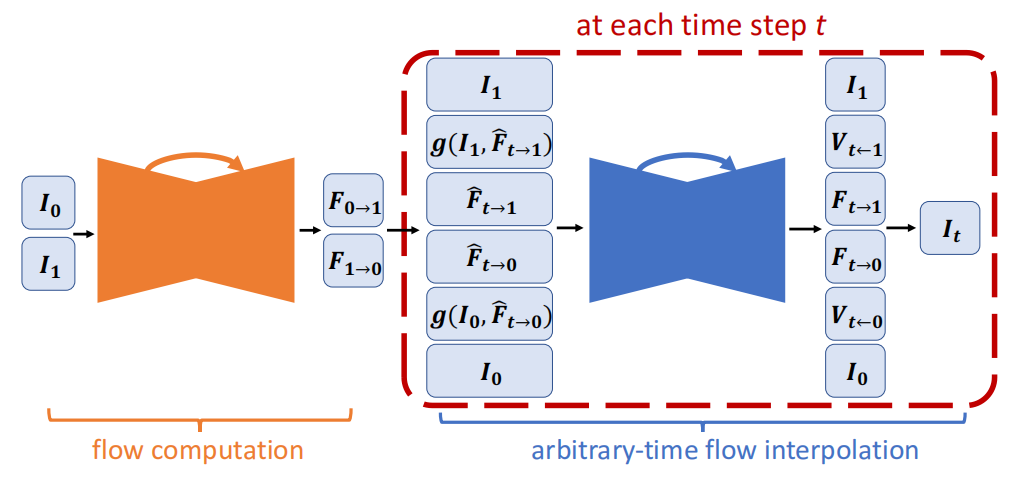


图1 SuperSloMo模型图[5]

2.1.1 基本原理

给定前后帧和以及时间，模型的目标是生成对应时刻的中间帧，SuperSlomMo通过融合利用光流扭曲的输入图像，生成目标插帧。

首先定义光流和，分别表征从到的光流，根据光流可以计算中间帧：

， （1）

其中表征反向扭曲函数，控制前后帧的权重，通过计算时序连续性和遮挡获得：对于时序性，若t趋近于0，则的贡献更高，若t趋近于1，则的贡献更高；对于遮挡，如果一个像素t在中间帧可见，那么它至少在前后帧当中的一帧中可见，因此引入可视映射和，其中，表征当从0运动到t时像素点p是否仍可见（0为被完全遮挡）。

由此有下式：

， （2）

其中是归一化系数。

2.1.2 光流插值

因为无法获取中间帧，直接计算光流和是困难的，为了解决这一问题，可以通过插值的方式从和中计算和。

将光流视为像素的位移，则可以近似表示为：

, （3）

同理可以近似表示为:

, （4）

同样利用时序加权平衡正反向光流的贡献：

（5）

2.1.3 网络模型

光流插值的方法在平滑区域表现很好，但在边界区域效果较差，会产生伪影，因此SuperSloMo设计一个UNet网络，对和进行细化，同时利用该网络生成和，模型架构如图1所示。

模型分为两部分，光流估计和光流插值，光流估计用于生成和的光流，光流插值用于细化和，同时生成和，最后利用公式（2）即可计算得出最终的插帧结果。

2.1.4 训练损失

重建损失：对中间帧的重建质量进行评估，在RGB空间域上对预测帧和真实帧做L1损失。

（6）

感知损失：对中间特征值做L2损失。

（7）

扭曲损失：主要用于对光流网络进行训练。

（8）

平滑损失：对光流进行平滑。

（9）

总的损失函数为：

（10）

2.2 DAIN

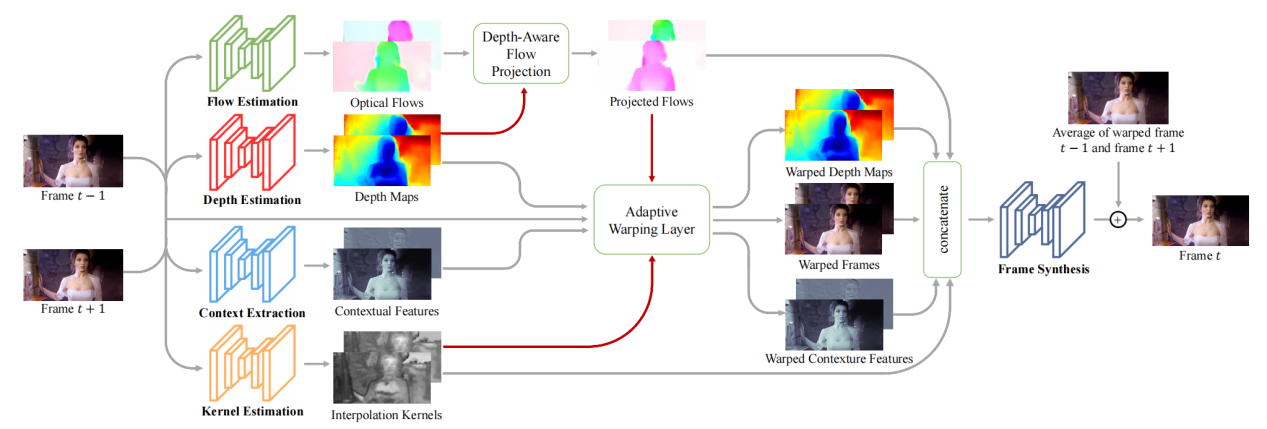


图2 DAIN模型图[6]

2.2.1 基本原理

DAIN采用与SuperSloMo相同的策略生成中间帧，即先生成光流，然后根据光流对前后帧进行扭曲。DAIN的优势在于引入深度作为光流插值的权值，将深度值较小的前景物体作为主要的光流源，缓解了光流插值中遮挡问题。

2.2.2 光流插值

不同于式（5），DAIN提出利用深度信息估计遮挡，以生成中间帧的光流。给定的深度图，表征在时刻t经过点x的像素，则映射后光流为：

（11）

其中权重是深度的倒数，表征为深度值较小的前景物体的光流遮挡了深度值较大的背景物体的光流，因此对该点的光流应该占有更高的权重：

（12）

类似地，可以根据深度图从中计算获得。

此外，对于没有光流经过的像素点，利用它邻域内的光流的均值填补空洞：

（13）

2.2.3 网络模型

模型分为以下模块：光流估计、深度估计、上下文提取、核估计、帧合成网络。DAIN利用深度感知的光流映射层获取中间光流，然后对输入帧、深度图、上下文特征进行自适应扭曲，最终利用帧合成网络生成中间帧。

光流估计。采用PWC-Net[4]的结构。

深度估计。采用MegaDepth[10]的结构。

上下文提取。构建残差卷积网络学习上下文特征。

核估计。局部插值核能被用于光流扭曲操作，基于光流确定窗口中心后，利用核对局部像素值（如4×4的区域内）进行聚合，以生成扭曲后的光流。DAIN利用Unet结构构造核的估计网络。

帧合成网络。为生成最终的插帧，DAIN将扭曲后的前后帧、深度图、上下文特征，以及映射后光流、插值核作为帧合成网络的输入，最终网络输出一个残差值，利用该值与前后帧的线性组合值相加，生成最终的合成帧。

2.3 BMBC

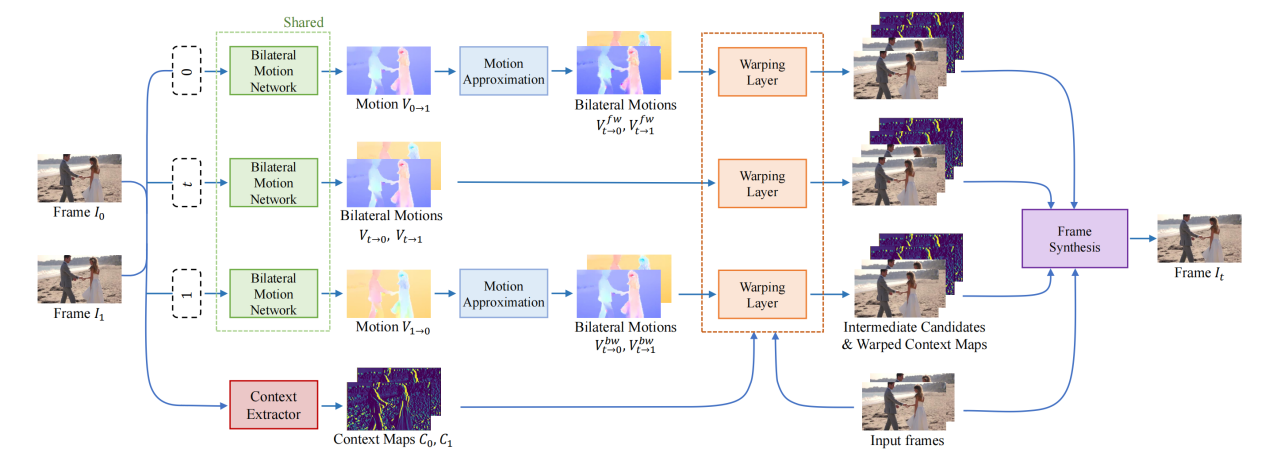


图3 BMBC模型图[7]

图3是BMBC模型的框架图，该模型设计双边光流估计的策略，利用光流网络直接生成中间帧到前后帧的光流。

给定输入和，首先模型估计双边光流和；其次估计双向光流和，根据这些运动近似四个前向双边运动；同时提取上下文特征和；然后使用六个双边光流扭曲输入帧、上下文特征，获得三组候选中间帧；最后利用帧合成网络将候选帧和输入帧融合，获得最终的插帧结果。

2.3.1 双边光流估计

为获得中间帧到、的光流值，BMBC假设相邻帧的运动是线性的，基于线性假设有，BMBC利用该线性关系将PWC-Net[4]光流网络进行修改，以预测双边光流，如图4所示。

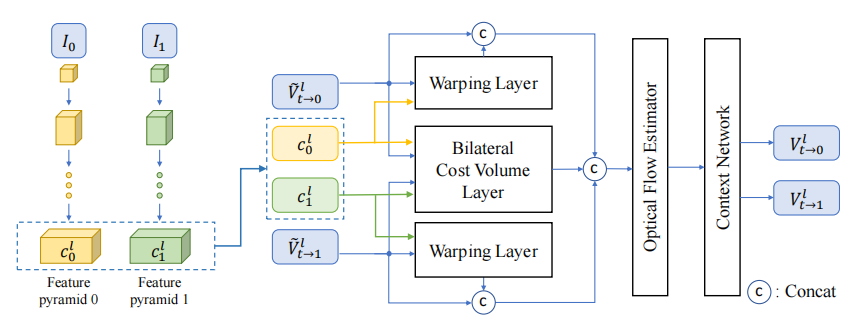


图4 BMBC双边光流估计模块图[7]

2.3.2 运动估计

尽管双边光流网络可以直接输出中间帧到前后帧的光流，但由于采用线性运动假设，和的精度可能不够高，尤其是对于存在遮挡的情况，与SuperSloMo类似，BMBC设计光流近似网络从和中近似得到和，详见公式（3）、公式（4）。

2.3.3 帧合成

如图3所示，结合三组光流获得三组候选中间帧，利用图5所示的结构，从特征中提取滤波卷积核，利用该滤波核将候选中间帧进行融合，生成最终的插帧结果。

（15）

（16）

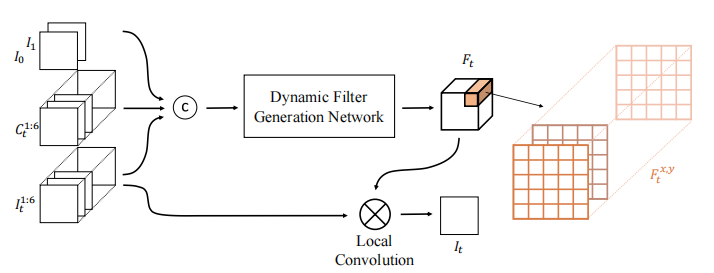


图5 BMBC帧合成模块图[7]

2.3.4 训练损失

BMBC将双边光流网络、动态滤波生成网络分别训练。

对于光流网络，设计基于Charbonnier的光度损失、平滑损失：

（17）

（18）

对于滤波生成网络，设计基于Charbonnier的损失：

（19）

2.4 ABME

BMBC建模了双边光流估计网络直接生成和，但是BMBC遵循线性运动的假设，对于非对称运动的光流估计仍有局限性，ABME在此基础上更进一步，提出非对称双边光流估计方法，一定程度上解决了图7所示的问题。

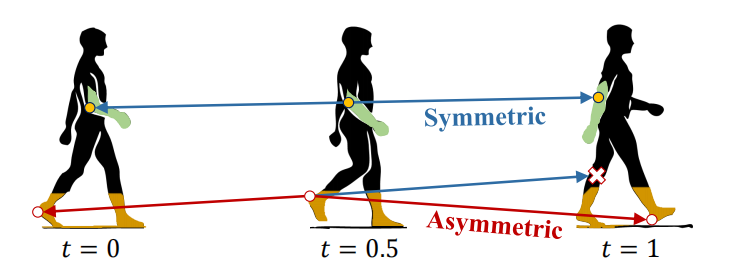


图6 对称、非对称双边光流示意图[8]

ABME的模型如图6所示，首先利用BMBC中提到的双边光流估计网络提取对称的双边光流和，然后基于此光流生成锚定帧，即插值帧的临时中间结果，然后利用ABMR-Net网络从前后帧、锚定帧中估计新的非对称的双边光流和，最终将结果输入帧合成网络，得到最终的输出结果。

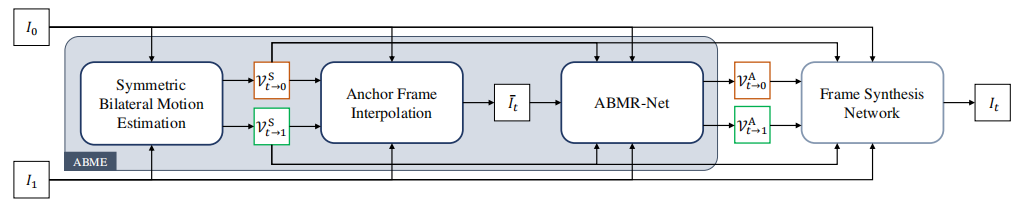


图7 ABME模型图[8]

2.4.1 锚定帧

获取对称双边光流和后，利用反向扭曲和时序权重，生成锚定帧：

（20）

为缓解遮挡产生的问题，尤其是相机平移时产生的边界遮挡，额外添加遮挡Mask如下：

（21）

（22）

2.4.2 非对称双边光流

基于锚定帧与输入帧、，利用图7所示的ABMR-Net模块生成非对称双边光流，、和、二者分别单独计算，得出光流和，此处的光流计算框架仍为PWC-Net[4]。

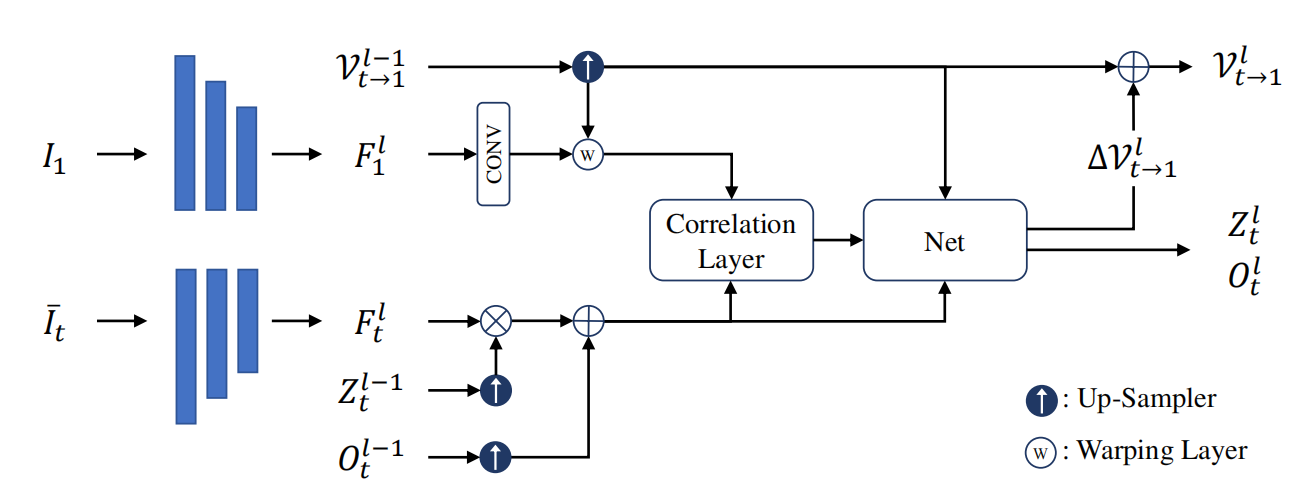


图8 ABMR-Net模块图[8]

由于锚定帧是对真实插帧的模拟，因此设计可靠性遮罩、遮挡遮罩以补偿其误差，其中初始化为0矩阵，基于双边光流的线性一致性初始化为：

（23）

2.4.3 帧合成

基于输入图像帧、对称双边光流、非对称双边光流，利用特征提取网络提取出的特征帧，ABME利用扭曲操作获得四个候选图像帧、十二个候选特征帧，并利用图8所示网络对其进行合成。

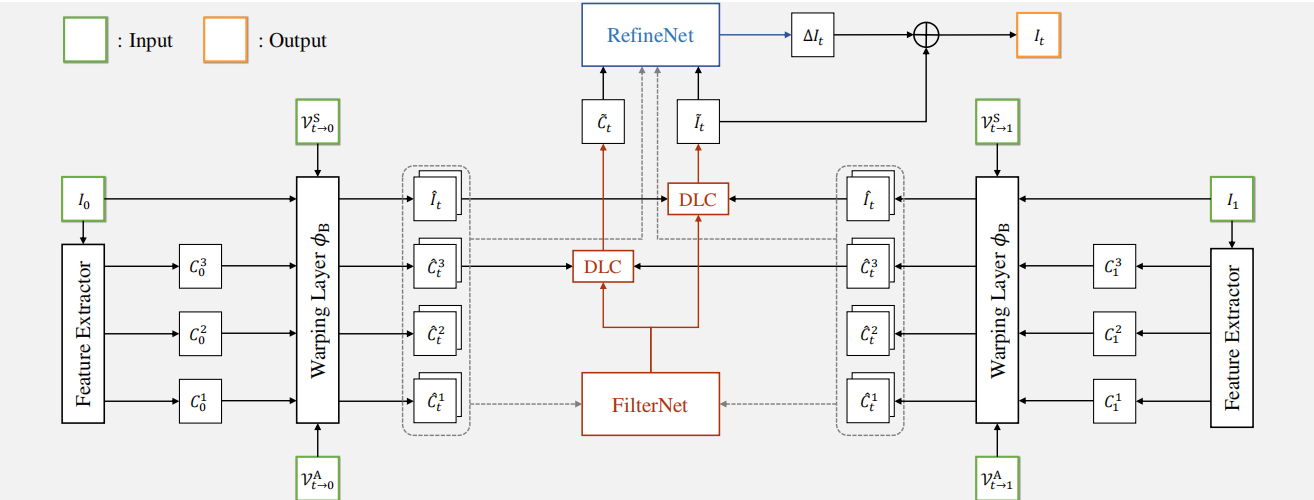


图9 ABME帧合成模块图[8]

给定多个候选帧，可以用权重对其进行加权融合，但是这种简单策略容易产生模糊伪影以及遮挡区域的重建错误，因此BMBC利用动态滤波对候选帧进行合成，而ABME提出了FilterNet和RefineNet两个网络实现该步骤。

FilterNet。用于生成卷积核，结构如图9所示，生成的卷积核被用于动态局部卷积（DLC），以生成滤波后的帧。

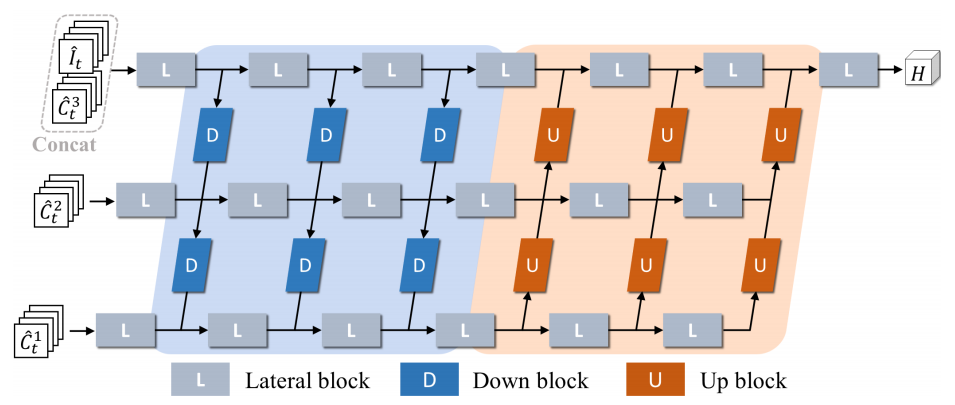


图10 FilterNet模块图[8]

RefineNet。动态滤波器只考虑局部邻域，如果邻域内由于运动误差、遮挡等问题没有包含某像素的信息，结果也会出现错误，因此为利用全局信息，RefineNet生成残差帧以细化滤波后的帧，它的结构与FilterNet相同，但其输入是滤波后的图像和特征以及原始金字塔特征。

2.4.4 模型训练

首先训练双边光流模块，然后训练AMBR-Net模块，最后以端到端的方式训练整个网络。

对光流网络，设置光度损失如下：

（24）

对于帧合成网络，设置损失如下：

（25）

其中为Charbonnier损失，为census损失。

2.5 RIFE

2.5.1 基本原理

视频插帧的关键是光流和的生成，前人的研究都是利用和对目标光流进行近似，即利用公式（3）、公式（4）揭示的线性关系进行估计，SuperSloMo、DAIN直接利用从、进行估计，而BMBC、ABME引入作为中间结果。

但是这种线性关系在面对复杂运动时总是存在局限性，因此RIFE提出直接利用端到端的框架设计新的模型，使模型拜托线性的约束，同时简化模型，提升模型的运行速度。

2.5.2 模型方法

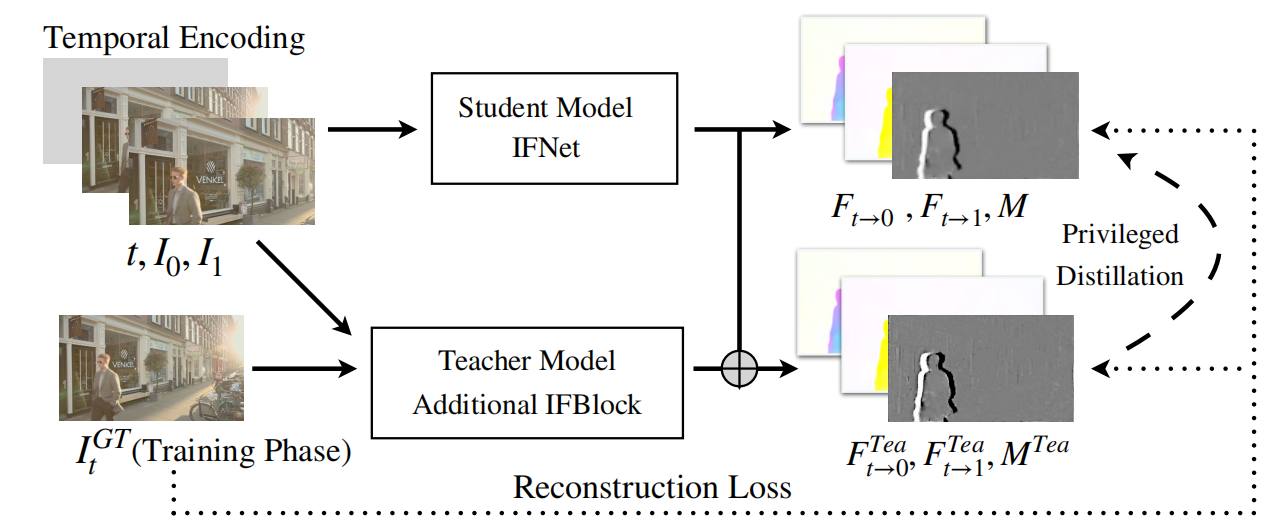


图11 RIFE架构图[9]

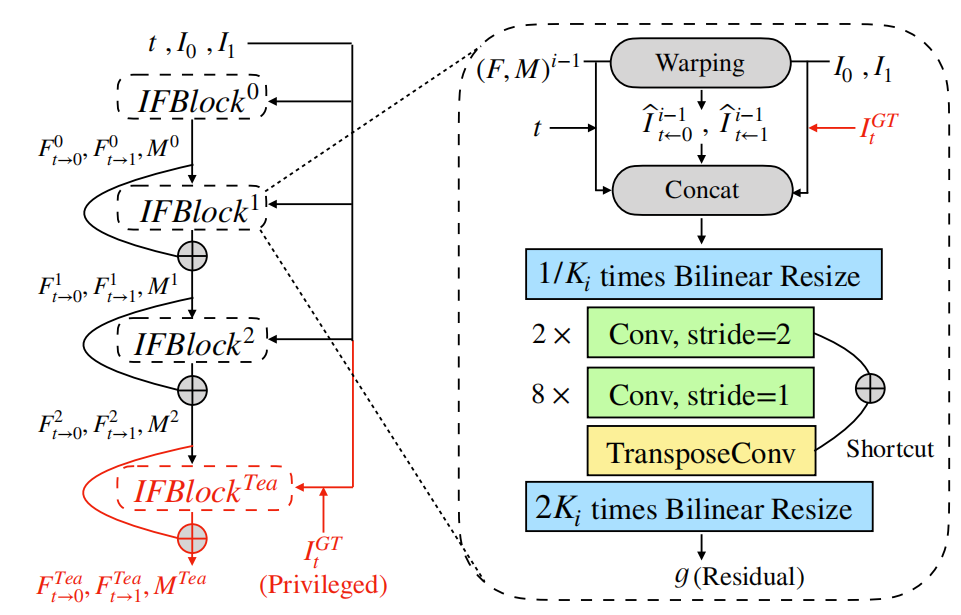


图12 FIBlock模块图[9]

如图11、图12所示，RIFE设计IFBlock模块，并将其进行堆叠形成端到端的结构，每层的输入为前后帧和、融合权重、光流和，输出为新的融合权重、光流和，有公式如下：

（26）

2.5.3 训练损失

由于端到端训练不容易收敛，在训练时作者额外设置最后一层IFBlock作为特权层，其可以接受真实的中间帧作为输入，由此设计蒸馏损失如下：

（27）

此外有重建损失如下：

（28）

# 总结

对基于光流的视频插帧而言，光流的准确性决定了最终插帧结果的质量，因此提取出更准确的中间帧到前后帧的光流是至关重要的，但是由于我们需要生成中间帧，而生成光流又需要前后帧，因此存在一个“先有鸡还是先有蛋”的问题，如何解决这一问题是进一步提升插帧模型性能的关键。

本文总结的若干方法中，前期方法如SuperSloMo、DAIN等利用前后帧间的光流对目标光流进行近似，中期方法如BMBC、ABME等将中间帧的思想引入光流网络，直接生成中间帧，但仍不能摆脱线性运动的假设，后期方法如RIFE提出利用端到端模型规避这一问题，总之都在探索提取出准确的光流。

因此，从实际的角度出发，可以进一步探索的方向包括：

1. 改进光流模型。当前的大部分插帧网络都忽视了对光流网络的改进，仍沿用Unet的原始结构，如果能参考最新光流估计的相关理论，一定能对最终的插帧结果有所帮助。
2. 引入多模态数据。对于遮挡问题，仅仅利用图像数据是很难有一个高效准确的解决方案，但如果能引入额外的信息如深度，基于多模态的输入构建光流插帧网络，可以对最终的结果产生有益影响。
3. 多任务模型。如DAIN中的深度估计一般，可以设计多任务模型，对光流网络、插帧网络产生促进作用。

# 参考文献

[1] LUCAS B D, KANADE T. An iterative image registration technique with an application to stereo vision; proceedings of the IJCAI'81: 7th international joint conference on Artificial intelligence, F, 1981 [C].

[2] HORN B K, SCHUNCK B G. Determining optical flow [J]. Artificial intelligence, 1981, 17(1-3): 185-203.

[3] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2015 [C].

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

[5] JIANG H, SUN D, JAMPANI V, et al. Super slomo: High quality estimation of multiple intermediate frames for video interpolation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[6] BAO W, LAI W-S, MA C, et al. Depth-aware video frame interpolation; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2019 [C].

[7] PARK J, KO K, LEE C, et al. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation; proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, F, 2020 [C]. Springer.

[8] PARK J, LEE C, KIM C-S. Asymmetric bilateral motion estimation for video frame interpolation; proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2021 [C].

[9] HUANG Z, ZHANG T, HENG W, et al. Real-time intermediate flow estimation for video frame interpolation; proceedings of the European Conference on Computer Vision, F, 2022 [C]. Springer.

[10] LI Z, SNAVELY N. Megadepth: Learning single-view depth prediction from internet photos; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].