# 基于深度值前向投影的视频帧插值模型

陈祖儿[1] 郑明魁[1] 张承琰[1] 易天儒[1]
CHEN Zu'er ZHENG Mingkui ZHANG Chengyan YI Tianru

**摘 要** 视频帧插值技术应用广泛，其目的是在给定两个连续的视频帧条件下，生成中间帧。针对向投影过程中经常出现的多个像素投影到同一个位置的像素重叠问题，提出了一种基于深度值前向投影的视频帧插值模型。根据提出的深度估计模块的深度值对前向投影过程进行线性加权，并具有深度平移不变性，对重叠像素区域的前景物体边界和背景像素的像素重建有一定的效果提升。实验结果表明，所提出的算法在公开的视频帧内插数据集 Vimeo-90k 上测试结果良好，与其他算法相比，在 PSNR、SSIM 和 LPIPS 性能评价指标上均能达到较为优秀的性能指标，验证了算法的优越性。

**关键词** 视频帧插值；视频帧预测；前向投影；深度估计；图像合成

doi：10.3969/j.issn.1672-9528.2024.04.001

## 0 引言

视频帧插值（video frame interpolation ，VFI）指在合成两个连续视频帧之间的中间帧，是计算机视觉领域的一个经典任务。目前的视频帧插值方法根据其基本框架的不同，大致可 以分为基于核的方法和基于流的方法两类。

（1）基于核的方法。具体而言，这种方法是将视频帧输入进神经网络，通过训练学习卷积核，分别对前后两帧的图像块进行局部卷积来预测中间待插值帧的像素值。根据输入信息的不同，还可以继续细分为只输入视频帧信息的方法。Long 等人 [1] 将两帧图像直接输入到网络，得到一个输出图像，通过不断迭代调整网络参数，使得网络输出帧和真实中间帧的差值达到最小，以实现最优的插帧效果。Cheng 等人 [2] 使用可变形可分离的卷积核。Choi 等人 [3] 使用 PixelShuffle 算子和通道注意力机制来隐私地捕获运动信息。Zhang 等人 [4] 提出了 CNN 与 transformer 的混合架构，利用帧间注意力机制分别提取图像的运动信息和外观特征。Niklaus 等人 [5] 融合了视频上下文相关信息，提出了对视频内容敏感的中间帧生成的方法。Cheng 等人 [6] 结合了多尺度的纹理特征、位置特征和变换域特征。Bao 等人 [7] 结合不同尺度上的光流信息和深度信息对中间视频帧进行插值。

（2）基于流的方法。首先，计算输入帧和目标帧之间的运动估计或光流，来确定目标像素点的位置信息，流估计的准确程度直接决定了后续目标帧插值的质量。其次，选取合适的插值算法进行前向扭曲或后向扭曲，以此合成目标帧。其中对光流估计的改进方法还包括直接估计中间帧光流以及双向流估计中间流。Liu 等人 [8] 将体素引入到光流估计中将像素抽象为体素，进而在体素层面上做光流预测，以计算出更精确的光流。Huang 等人 [9] 提出了一个名为 IFNet 的神经网络，以更快的速度从粗到细直接估计相邻帧到中间帧的端到端的中间流。Xue 等人 [10] 提出了面向任务的光流估计算法，针对不同的任务设计不同的网络结构来估计光流。Van 等人 [11] 利用对抗训练的方式，在多尺度上估计光流，按从粗粒度到细粒度的方式来合成中间帧，使得生成帧的主观质量很好。Bao 等人 [12] 利用双向流的加权组合来估计中间流。除此之外，Simon 等人 [13] 提出了一种插值优化算法，将光流估计结果作为前向扭曲的依据。

近几年，视频帧插值技术在不断进步，上述的两类范式方法已经取得了较为不错的实验结果。但是由于现实世界中相机和物体的运动带来了复杂的像素移动和亮度变化，这对视频帧插值算法而言仍然是一个巨大的挑战，会造成插值帧出现伪影、空洞以及多像素重叠等一系列问题。目前的视频帧插值算法在针对运动物体边界以及像素遮挡的问题上，仍然没有显著性的效果提升。

为了有效解决上述提出的问题，本文提出了一个基于深度值前向投影的视频帧插值模型。主要贡献如下：（1） 提出了一种像素级的高分辨率深度图估计算法，该算法结合了来自 PWC-Net 网络输出的光流信息以及来自数据集的相机位姿信息， 能够得到更高精度的深度图；（2）提出了一种基 于深度值的前向投影插值算法，解决了像素重叠（多个像素 映射到同一个位置）问题，对运动物体边界和像素遮挡问题 的插值结果有一定的效果提升。

1. 福州大学 福建福州 350108

[ 基金项目 ] 国家自然科学基金项目（61902071）；福建省自然科学基金计划资助项目（2020J01466）；2020 年福建省高等学校科技创新团队（产业化专项）

## 1 模型方法

模型的总体框架图如图 1 所示。给定输入视频帧 $I_0$、$I_1$，经过光流预测网络后，可以得到双向流 $F_{0\to1}$ 和 $F_{1\to0}$ 并估计中间流 $F_{0\to t}$ 和 $F_{1\to t}$，同时可以得到特征金字塔。利用双向流估计深度图 $D_0$、$D_1$，并依据此结果对原图以及特征空间进行前向投影。最后将扭曲结果输入图像合成网络，得到合成的中间帧 $I_t$。

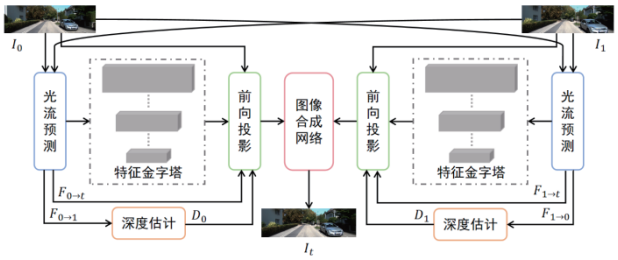


图 1 模型总体框架图

### 1.1 光流预测网络

在基于流的方法框架下，从待插值帧 $I_t$ 的角度，希望知道中间流 $F_{0\to t}$ 和 $F_{1\to t}$，从而从视频帧 $I_0$、$I_1$ 进行扭曲得到插值结果。然而对于视频帧插值任务而言，估计中间流和合成插值帧之间存在着“先有鸡还是先有蛋”的问题，因为都属于事先不存在的对象，所以它们的估计是一个难点，现成的光流估计模型也不能直接应用于视频帧插值框架中。给定输入帧 $I_0$、$I_1$，可以利用现成的光流估计网络，如 PWC-Net[14]，估计双向流 $F_{0\to1}$ 和 $F_{1\to0}$，并在接下来的实验测试中证明，该模型效果表现良好。得知双向流后，根据时间步长估计中间流 $F_{0\to t}$ 和 $F_{1\to t}$，其中 $t\in(0,1)$ 定义了期望的时间位置。

$$F_{0\to t}=t\cdot F_{0\to1} \tag{1}$$

$$F_{1\to t}=(1-t)\cdot F_{1\to0} \tag{2}$$

同时，利用 PWC-Net 网络中现成的金字塔结构对视频帧 $I_0$、$I_1$ 进行多尺度的特征提取，生成特征表示的 6 层金字塔。不同的特征尺度，具有不同的感受野，能够提供更多的特征信息。从第一层到第六层，特征通道的数量分别为 16、32、64、96、128、196。

### 1.2 深度估计模块

在视频帧的预测过程中，深度误差可能导致前向投影过程中出现裂纹、孔洞、背景像素穿透等不良现象，因此深度估计的精度非常重要。使用 RGB-D 传感器的主动式估计方法存在精度低的问题，因此提出利用即时定位与地图构建（simultaneous localization and mapping，SLAM）中的对极几何方法生成像素级别的高精度深度图，即在已知两帧图像对应点的坐标关系和相机的位姿信息的情况下，通过对极约束估计该点的深度值。

两个像素点 $p_0$、$p_1$ 的像素位置为：

$$s_0p_0=KP,s_1p_1=K(RP+t) \tag{3}$$

式中：$s_0$、$s_1$ 分别表示像素点 $p_0$、$p_1$ 对应的深度值，$\boldsymbol{K}$ 为相机内参矩阵，$P$ 为空间中的点 $[X,Y,Z]^{\mathrm{T}}$，$\boldsymbol{R}$、$\boldsymbol{t}$ 分别为相机运动的旋转矩阵和平移矩阵。

将 $x_0$、$x_1$：

$$x_0=K^{-1}p_0,x_1=K^{-1}p_1 \tag{4}$$

为两个像素点的归一化平面上的坐标代入公式（2）得：

$$s_1x_1=s_0Rx_0+t \tag{5}$$

两边同时左乘$x_0^{\wedge}$，得：

$$s_0x_0^{\wedge}x_0=s_1x_0^{\wedge}Rx_1+x_0^{\wedge}t=0 \tag{6}$$

式（6）结果为 0。因此，只要知道两个匹配点 $p_0$、$p_1$ 的像素坐标关系，就可以将上式看成 $s_1$ 的方程，求出 $I_1$ 时刻帧的每个像素点对应的深度值 $s_1$。同理，也可以得到 $I_0$ 时刻帧的每个像素点对应的深度值 $s_0$。

空间点 $P$ 在不同时刻帧的像素坐标是不同的，对于求解两帧之间像素点 $p_0$、$p_1$ 的对应关系问题，迄今为止有许多研究者对此进行了深入的研究。其中针对图像的特征提取与匹配问题，研究者们提出过许多经典的传统算法。SIFT 算法充分考虑了相机的运动以及光照条件的变化，但随之而来的是巨大的计算开销。ORB 算法在计算速度方面有了明显改进，但只能匹配相对稀疏的局部特征点。近几年，基于深度学习的光流估计网络取得了很大的进展，因此本文采用 PWC-Net 光流估计网络（目前性能最好的光流估计网络之一）来获得高精度、高效率的全局光流。PWC-Net 网络输入视频帧 $I_0$、$I_1$，输出两帧之间的光流 $(u_1{-}u_0, v_1{-}v_0)$，其中 $(u_0, v_0)$ 和 $(u_1, v_1)$ 分别表示匹配点 $p_0$、$p_1$ 相对应的像素坐标。根据公式（4），已知匹配点 $p_0$、$p_1$ 像素坐标的对应关系，即可求出该点的深度值 $s_0$、$s_1$，从而得到 $I_0$、$I_1$ 帧图像对应的深度图 $D_0$、$D_1$。

### 1.3 前向投影模块

在图像变换的算法里（即移动图像中的像素），分为前向扭曲和后向扭曲。前向扭曲是指像素从原始图像的坐标移动到目标图像的坐标。该算法的优点是计算速度快，但与此同时带来的缺点是会导致部分信息的丢失或者混叠，比如空洞和像素重叠。后向扭曲是指像素从目标图像的坐标向原 始图像的坐标进行映射。该算法的优点是能够保持较好的 图像质量，但缺点是计算过于复杂、计算速度慢。

后向扭曲是一种常见的技术，在无监督深度估计以及光流估计等领域中广泛使用，应用于许多深度学习网络框架中。然而通过后向扭曲算法来合成视频中间帧需要知道 $F_{t\to0}$ 和 $F_{t\to1}$，这在视频帧插值领域中，以时刻为 $t$ 的视频帧 $I_t$ 的角度而言，计算此中间流是非常复杂且困难的。相比之下，前向扭曲更适合视频帧插值任务。但目前还没有提出非常合适的前向扭曲算法，来解决多个像素映射到同一个位置的像素重叠问题，因此前向扭曲会导致重建后的图像模糊问题。目前

处理这种映射模糊性的常用方法是叠加算法：

$$u = p - (q + F_{0\to t}[q]) \tag{7}$$

$$b(u) = \max(0,1-|u_x|)\cdot\max(0,1-|u_y|) \tag{8}$$

$$I_t[p] = \sum_{\forall q\in I_0} b(u) I_0[q] \tag{9}$$

式中：$p$、$q$ 分别表示 $I_0$ 帧与 $I_t$ 帧上的像素点；$F_{0\to t}[q]$ 表示 $q$ 像素位置上的光流；$b(u)$ 表示双线性核权值。但这种简单的叠加操作会导致像素重叠区域的亮度不一致问题。因此，本文提出了一种基于深度值的前向投影插值算法来解决这些固有的限制，其中 $D[q]$ 表示像素点 $q$ 处的深度值。

$$I_t[p] = \frac{\sum_{\forall q\in I_0}\exp(D[q])\cdot b(u)\cdot I_0[q]}{\sum_{\forall q\in I_0}\exp(D[q])\cdot b(u)} \tag{10}$$

使用深度值 $D[q]$ 对 $I_0[q]$ 进行线性加权，这种方法可以根据像素点处的深度值，更好地将重叠像素区域的前景物体和背景像素有效分开。除此之外，这种算法还具有深度平移不变性，它对于相对于深度 $D[q]$ 的平移 $\varphi$ 是不变的，当将多个像素映射到同一位置时，这是一个特别重要的属性。例如，当前景物体位于 $D[q]$=1，背景位于 $D[q]$=10，或者是前景物体位于 $D[q]$=101，背景位于 $D[q]$=110 时，都将对前景物体和背景进行同等处理。

$$\begin{aligned} I_t[p] &= \frac{\sum_{\forall q\in I_0}\exp(D[q]+\varphi)\cdot b(u)\cdot I_0[q]}{\sum_{\forall q\in I_0}\exp(D[q]+\varphi)\cdot b(u)} \\ &= \frac{\sum_{\forall q\in I_0}\exp(D[q])\cdot\exp(\varphi)\cdot b(u)\cdot I_0[q]}{\sum_{\forall q\in I_0}\exp(D[q])\cdot\exp(\varphi)\cdot b(u)} \\ &= \frac{\sum_{\forall q\in I_0}\exp(D[q])\cdot b(u)\cdot I_0[q]}{\sum_{\forall q\in I_0}\exp(D[q])\cdot b(u)} \end{aligned} \tag{11}$$

### 1.4 图像合成网络

图像合成网络根据扭曲后的输入图像及其对应的特征金字塔生成插值结果。本文的图像合成网络不仅输入扭曲后的视频帧，而且输入多个分辨率上扭曲后的特征空间，这使得合成网络能够做出更好的预测。借鉴三行六列的 GridNet[15] 网络结构来完成这项任务。

## 2 实验与结果

### 2.1 实验环境与数据集

本实验基于 Ubuntu18.04 系统上，主要配置为：Intel Xeon(R) Silver 4210R CPU@2.40 GHz×40，NVIDIA GeForce RTX 3090 GPUs。数据集选用公共可用的 Vimeo-90k[16] 数据集进行测试，输入图片的分辨率为 448×256。模型在训练时 Batch_size 设置为 4，Epoch 设置为 80，学习率设置为 0.000 2。

### 2.2 评价指标

为了验证本文提出的模型结构的优越性，需要对其进行性能评估。为此，分别从主观方面和客观方面进行评价分析。其中，主观方面主要是依靠人眼从图像的色彩、细节等方面对图像的整体效果进行评价；客观方面，本文选用了最常用的图像质量评价指标，包括：峰值信噪比（peak signal to noise ratio，PSNR），通常用于评估一幅图像与原始图像之间的相似度，PSNR 的值越高，表示两幅图像之间的相似度越高，质量越好；结构相似性（structural similarity index measure，SSIM），是一种考虑了亮度、对比度、结构差异的用于衡量两幅图像之间相似度的指标，SSIM 的值越高，表示图像的质量越好；感知损失（learned perceptual image patch similarity，LPIPS），用于度量两张图像之间更符合人类感知情况的差别，LPIPS 的值越低，表示两张图像越相似。

### 2.3 实验结果与分析

选择了最具代表性的几个图像场景作为结果展示，其中包括了单一物体以及复杂的室内外场景。可以从图 2 看出，主观上人眼基本不能区分出合成帧与真实值的差别，合成帧在整体场景结构、亮度、色彩以及细节上都拥有很好的合成效果。并且用红色矩形框特别圈出了图片中的细节部分，可以看到在马腿毛发处的细节以及前景汽车与背景的边界处的预测效果都很好。

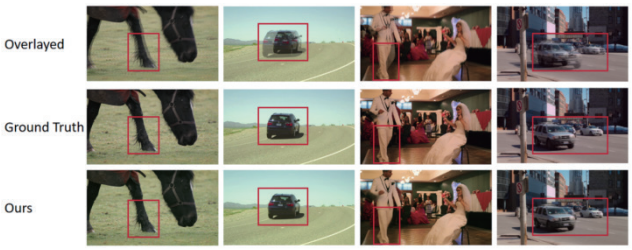


图 2 实验结果

同时，为了更好地验证本文提出的模型结构的性能，选取了近几年提出的 DAIN、SepConv、SoftSplat、CDFI[17]、M2M[18]、IFRNet[19] 视频插帧算法作为对比。由表 1 可知，本文算法分别在 PSNR、SSIM、LPIPS 三个性能指标上都能取得最优秀的性能指标。与 SoftSplat 算法对比，能取得 0.02 dB 的 PSNR 提升和 0.009 dB 的 SSIM 提升；与 IFRNet 算法对比，在 SSIM 性能持平的情况下，能取得 0.32 dB 的 PSNR 提升。这验证了本文算法的优越性。

表 1 实验客观评价指标对比

| | Ours | DAIN | SepConv | SoftSplat | CDFI | M2M | IFRNet |
|---|---|---|---|---|---|---|---|
| PSNR↑ | 36.12 | 34.70 | 33.80 | 36.10 | 35.17 | 35.47 | 35.80 |
| SSIM↑ | 0.979 | 0.964 | 0.956 | 0.970 | 0.977 | 0.978 | 0.979 |
| LPIPS↓ | 0.020 | 0.022 | 0.027 | 0.021 | — | — | — |

## 3 结语

为了解决视频帧内插过程中前向投影带来的图像模糊性问题，本文提出了一个基于深度值前向投影的视频帧插值模型。该模型框架在传统的基于光流的方法上，加入了深度估计模块，并改进了前向投影算法，能够估计一个高精度的深度图，并解决多个像素映射到同一个位置的问题。经过实验测试证明，在 Vimeo-90k 数据集上的视觉和客观评价方面都取得了较好的插值帧实验结果。与目前几种先进的插值方法

进行比较，本文所提出的算法在 PSNR 和 SSIM 评价指标上均有一定的提高，验证了本文算法模型的有效性。本文算法还有值得进一步改进的地方，后续将考虑对视频帧进行前景物体和背景分离操作，希望在更大程度上提高合成帧的图像质量。

**参考文献：**

[1]LONG G, KNEIP L, ALVAREZ J M, et al. Learning image matching by simply watching video[J]. European conference on computer vision, 2016, 9910: 434-450.

[2]CHENG X H, ZHEN Z C. Video frame interpolation via deformable separable convolution[C]//Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI). Palo Alto:Association for the Advancement of Artificial Intelligence,2020:10607-10614.

[3]CHOI M, KIM H, HAN B, et al. Channel attention is all you need for video frame interpolation[C]//Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI).Palo Alto:Association for the Advancement of Artificial Intelligence, 2020:10663-10671.

[4]ZHANG G, ZHU Y, WANG H, et al. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2023: 5682-5692.

[5]NIKLAUS S, LIU F. Context-aware synthesis for video frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2018: 1701-1710.

[6]CHENG X, CHEN Z. A multi-scale position feature transform network for video frame interpolation[J]. IEEE transactions on circuits and systems for video technology, 2019, 30(11): 3968-3981.

[7]BAO W, LAI W S, ZHANG X, et al. MEMC-Net: motion estimation and motion compensation driven neural network for video interpolation and enhancement[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(3): 933-948.

[8]LIU Z, YEH R A, TANG X, et al. Video frame synthesis using deep voxel flow[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway:IEEE, 2017: 4463-4471.

[9]HUANG Z, ZHANG T, HENG W, et al. RIFE:real-time intermediate flow estimation for video frame interpolation[J]. European conference on computer vision, 2022, 13674: 624-642.

[10]XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow[J]. International journal of computer vision, 2019, 127(8): 1106-1125.

[11]VAN A J, SHI W, ACOSTA A, et al. Frame interpolation with multi-scale deep loss functions and generative adversarial networks[EB/OL].(2019-02-26).https://arxiv.org/pdf/1711.06045.pdf.

[12]BAO W, LAI W S, MA C, et al. Depth-aware video frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2019: 3703-3712.

[13]NIKLAUS S, LIU F. Softmax splatting for video frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2020:5436-5445.

[14]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 Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2018: 8934-8943.

[15]DAMIEN F, REMI E, ELISA F, et al. Residual conv-deconv grid network for semantic segmentation[C]//Proceedings of the British Machine Vision Conference (BMVC). New York:Springer-Verlag New York, Inc., 2018: 8981-8989.

[16]XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow[J]. International journal of computer vision, 2019, 127(8): 1106-1125.

[17]DING T, LIANG L, ZHU Z, et al. CDFI: compression-driven network design for frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2021: 8001-8011.

[18]HU P, SIMON N, STAN S, et al. Many-to-many splatting for efficient video frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2022: 3553-3562.

[19]KONG L, JIANG B, LUO D, et al. Ifrnet:intermediate feature refine network for efficient frame interpolation[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2022: 1969-1978.

**【作者简介】**

陈祖儿（1999—），女，福建福州人，硕士研究生，研究方向：视频帧间预测编码。

郑明魁（1976—），通信作者（email：zhengmk@fzu.edu.cn），男，福建福州人，副教授，博士，研究方向：多媒体通信与编码。

（收稿日期：2024-01-14）