Software Engineering and Applications 软件工程与应用, 2025, 14(2), 201-216
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.142019

Hans汉斯

基于Transformer和运动边界掩码关注变化的视频插帧方法

石明光，王晓红，马春运

上海理工大学出版学院，上海

收稿日期：2025年2月23日；录用日期：2025年3月27日；发布日期：2025年4月8日

摘　要

为了提升基于光流的视频插帧方法在变化区域的生成质量，我们提出了一种新颖的两阶段视频插帧框架，该框架在光流运动信息的约束下指导中间帧的细化。捕捉长距离相关性信息能够提高光流估计的准确性，因此，我们提出了一种BWT-FlowNet用于光流估计，该网络通过集成双级窗口Transformer和内容感知机制来捕捉视频序列中的长距离时空交互。随后，利用光流中的运动信息预测运动边界掩模(MB mask)，以帮助网络在中间帧细化过程中聚焦于内容变化区域。我们还开发了一种运动边界感知细化网络(MBAR Net)用于中间帧的细化过程。在MBAR Net的子层中使用金字塔MB mask以突出运动区域。此外，引入掩模感知损失函数(Mask Perceptual Loss)以有效约束内容变化区域，从而提高预测帧的质量。实验表明，我们提出的方法在多个公共基准测试中均取得了优异的性能。

关键词

视频插帧，光流估计，掩码，Transformer

Transformer-Based Video Frame Interpolation with MB Mask Guidance

Mingguang Shi, Xiaohong Wang, Chunyun Ma

College of Publishing, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 23rd, 2025; accepted: Mar. 27th, 2025; published: Apr. 8th, 2025

Abstract

To enhance the generation quality of flow-based video frame interpolation methods in changing regions, we propose a novel two-stage video frame interpolation framework that guides the

文章引用: 石明光, 王晓红, 马春运. 基于 Transformer 和运动边界掩码关注变化的视频插帧方法[J]. 软件工程与应用, 2025, 14(2): 201-216. DOI: 10.12677/sea.2025.142019

refinement of intermediate frames under the constraint of optical flow motion information. Capturing long-range relevant information can enhance the accuracy of optical flow estimation. Therefore, we propose a BWT-FlowNet for optical flow estimation, which integrates a bi-level window Transformer with content awareness to capture long-range spatial-temporal interactions in video sequences. Then, a Motion Boundary Mask (MB Mask) is predicted by leveraging the motion information from optical flow, which is used to help the network focus on content-changing areas during the refinement of intermediate frames. We also develop a Motion Boundary-Aware Refinement Net (MBAR Net) to refine the process of intermediate frames. Pyramid MB Masks are utilized in sublayers of the MBAR Net to highlight motion regions. In addition, the Mask Perceptual Loss function is introduced to constrain content-changing areas effectively, improving the quality of predicted frames. Experiments demonstrate that our proposed method achieves excellent performance on several public benchmarks.

Keywords

Video Frame Interpolation, Optical Flow Estimation, Mask, Transformer

1. 引言

视频插帧(VFI)是视频处理中的一项关键任务，其目标是在连续帧之间生成中间帧。该技术已广泛应用于新视角合成[1]、视频压缩[2]、慢动作视频生成[3]-[5]以及动画合成[6]，以提高视频制作的效率并降低成本。

深度学习已被广泛应用于视频插帧，主要方法包括基于核的方法[3] [4] [7]-[9]、基于相位的方法[10] [11]以及基于光流的方法[12]-[23]。基于相位的方法通过调整像素相位巧妙地捕捉运动信息，但由于假设运动的周期性，它们仅适用于有限范围内的运动，并且容易产生模糊的结果。基于核的方法利用核函数捕捉视频中的时空信息。然而，这些方法通常将运动估计与图像合成合为一步，从而限制了模型生成高质量中间帧的能力。此外，由于固定核函数本身存在局限性，这些方法在处理复杂运动模式及大运动时表现出局限性。

近年来，基于光流的方法取得了显著进展。光流可在像素级别更精确地捕捉物体运动，从而提供更准确的运动信息。这些方法大多分为两个阶段：第一阶段预测光流，并利用光流对原始输入帧或上下文特征进行变形；第二阶段通过带有各种特征的细化网络增强粗糙变形帧的细节。

现有的基于光流的方法未能综合整合这两个阶段，因为光流估计和随后在这些运动区域的细化或细节生成是分开处理的。通常，视频中的许多物体和背景要么是静止的，要么是线性运动、视角变换、缩放等，其运动相对容易预测。另一方面，运动边界周围的区域通常表现出更复杂的运动模式、大位移和内容变化，这导致在中间帧合成过程中出现模糊和伪影。为了解决这一问题，我们提出了一种新颖的两阶段视频插帧框架，该框架通过光流运动信息约束中间帧的细化。在我们的方法中，光流中的运动信息用于预测运动边界掩模(MB mask)，该掩模在帧细化阶段用于突出显示内容变化区域。

光流估计是基于光流方法的核心，因为它提供了视频帧之间的像素级运动信息，并有效地对输入帧进行变形和对齐。以往的研究[24]-[28]提出了各种用于光流估计的网络，但这些方法通常依赖于卷积神经网络(CNN)，它们主要关注局部区域，难以有效捕捉长距离相关性信息。这一局限性在涉及大位移或复杂

运动模式的场景中面临挑战。为此，我们引入了 Transformer [29]用于光流估计，因为它们能够更好地捕捉长距离时空依赖特征。最近的研究[30]-[33]表明，将 CNN 与 Transformer 结合可以显著提高网络性能。我们设计了一种 BWT-FlowNet，该网络集成了 CNN 与双级路由注意力[34]，以增强光流估计。在 Transformer 中采用了具有内容感知能力的稀疏注意力机制，并通过过滤策略高效捕捉视频序列中的时空相关性。

以往用于生成中间帧的方法在运动边界区域往往效果不佳，因为这些区域的物体表现出更复杂的运动模式和大位移。因此我们引入了运动边界感知细化网络(MBAR Net)用于细化中间帧，该网络在金字塔结构的 MB mask 的引导下专注于内容变化和运动区域。

为了更好地优化中间帧的细粒度质量，我们引入了一种掩模感知损失(Mask Perceptual Loss)。该损失函数利用 MB mask 对预测帧和目标帧进行预处理，然后使用预训练的 VGG 网络计算它们的感知损失。

总的来说，我们的研究贡献包括：

1) 提出了一种两阶段视频插帧框架，该框架在光流运动估计的约束下对中间帧进行细化。第一阶段生成的光流用于预测运动边界掩模(Motion Boundary mask)，该掩模在细化阶段被用于增强网络对内容变化显著区域的聚焦能力。

2) BWT-FlowNet 的设计结合了卷积神经网络(CNN)和双级窗口 Transformer，以实现从粗到细的光流估计。为了高效捕捉视频序列中的时空长距离依赖关系，Transformer 中采用了具有内容感知能力的注意力机制。

3) 我们开发了运动边界感知细化网络(Motion Boundary Aware Refinement Net，简称 MBAR Net)用于中间帧的细化。MB mask 被处理成金字塔结构，使网络在中间帧细化过程中能够更多地关注运动边界区域。

4) 为了有效约束预测帧中的运动相关区域，我们引入了基于 MB mask 的掩模感知损失函数(Mask Perceptual Loss)，以提升图像质量，尤其是在内容变化区域。

2. 网络模型设计

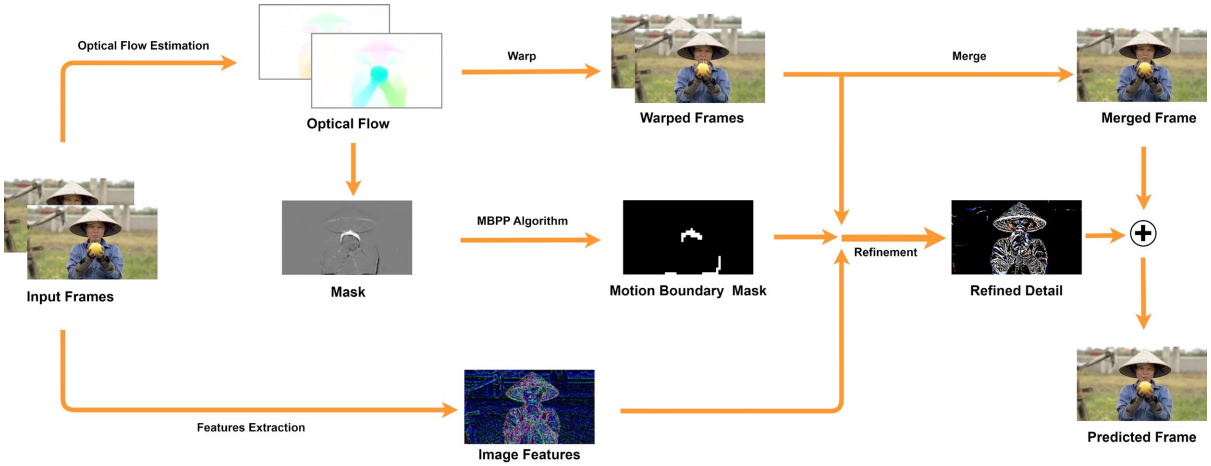

Figure 1. The main idea of our proposed method
图 1. 方法思想图

我们提出了一种视频插帧框架，该框架在光流运动信息的约束下对中间帧进行细化，如图 1 所示，本文使用一个网络估计光流，并基于这些光流导出的运动信息预测运动边界掩模(Motion Boundary mask，

简称 MB mask)。通过光流对两帧连续的图像进行变形，随后将变形后的图像处理成一个融合帧。接着，将输入帧的特征、光流以及 MB mask 输入到一个细化网络中，该网络在 MB mask 的引导下专注于内容变化和运动区域。细化网络的输出与融合帧结合，生成最终的细化中间帧。

2.1. 网络框架

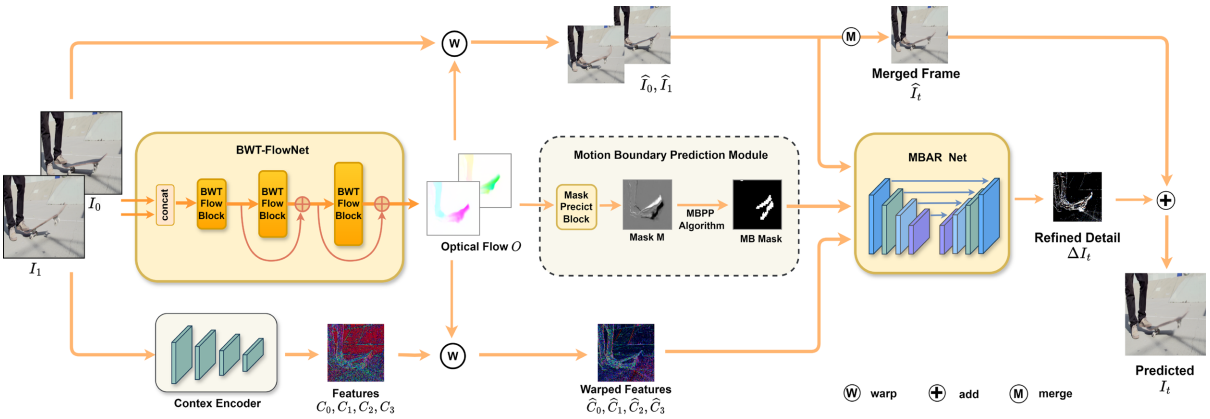

Figure 2. Overview of our proposed framework
图 2. 方法的总框架图

基于在光流运动信息约束下对中间帧进行细化的概念，我们设计了一种用于视频插帧(VFI)的框架，如图 2 所示。给定输入帧 I_0 和 I_1，采用双级窗口 Transformer 的 BWT-FlowNet 被设计用于估计更精确的光流，这些光流随后被用于通过反向变形获得变形帧 $\hat{I}_0, \hat{I}_1$。在光流所获取的运动信息的引导下，掩模预测模块计算掩模 M，用于指示连续帧上的运动内容。通过变形帧和掩模 M，可以得到粗糙的中间帧 $\hat{I}_t$，其表达式为：

$$\hat{I}_t = M \odot \hat{I}_0 + (1 - M) \odot \hat{I}_1. \tag{1}$$

为在光流运动信息约束下指导中间帧的细化，本文提出一种运动边界块预测(MBPP)算法，用于将掩模处理成运动边界掩模(MB mask)，以明确限定运动及内容变化区域。随后，MB mask 被下采样为金字塔结构 M_{prd}，供细化网络使用。

引入上下文编码器可以从输入图像中提取详细的高维特征 C_0, C_1, C_2, C_3，这些特征通过光流 O 进行反向变形，得到变形特征 $\hat{C}_0, \hat{C}_1, \hat{C}_2, \hat{C}_3$，其表达式为：

$$\hat{C}_0, \hat{C}_1, \hat{C}_2, \hat{C}_3 \triangleq Warp\left(O, Encoder\left(I_0, I_1\right)\right). \tag{2}$$

最终，将输入帧、光流、变形帧、多尺度变形特征以及 M_{prd} 输入到运动边界感知细化(MBAR)网络中，该网络预测中间帧的细化细节 ΔI_t。通过将细化细节 ΔI_t 与融合帧相结合，得到最终结果 I_t，其表达式为：

$$I_t = \hat{I}_t + \Delta I_t. \tag{3}$$

2.2. BWT-FlowNet

本文设计了 BWT-FlowNet 以实现更精确的光流估计，如图 3 所示。光流能够有效地指导输入帧和图像特征的变形，而精确的运动信息则增强了对运动区域的识别能力。

(1) BWT-FlowNet：与传统的卷积神经网络(CNN)相比，Transformer 能够更有效地捕捉长距离依赖关

系和全局上下文信息，这在光流估计中尤为重要，尤其是在处理大运动以及提高估计精度方面。因此，我们引入了一种新的视觉 Transformer，它在 BWT Flow Block 中采用了双级路由注意力(Bi-Level Routing Attention) [34]，以捕捉视频序列中的长距离时空相关性。

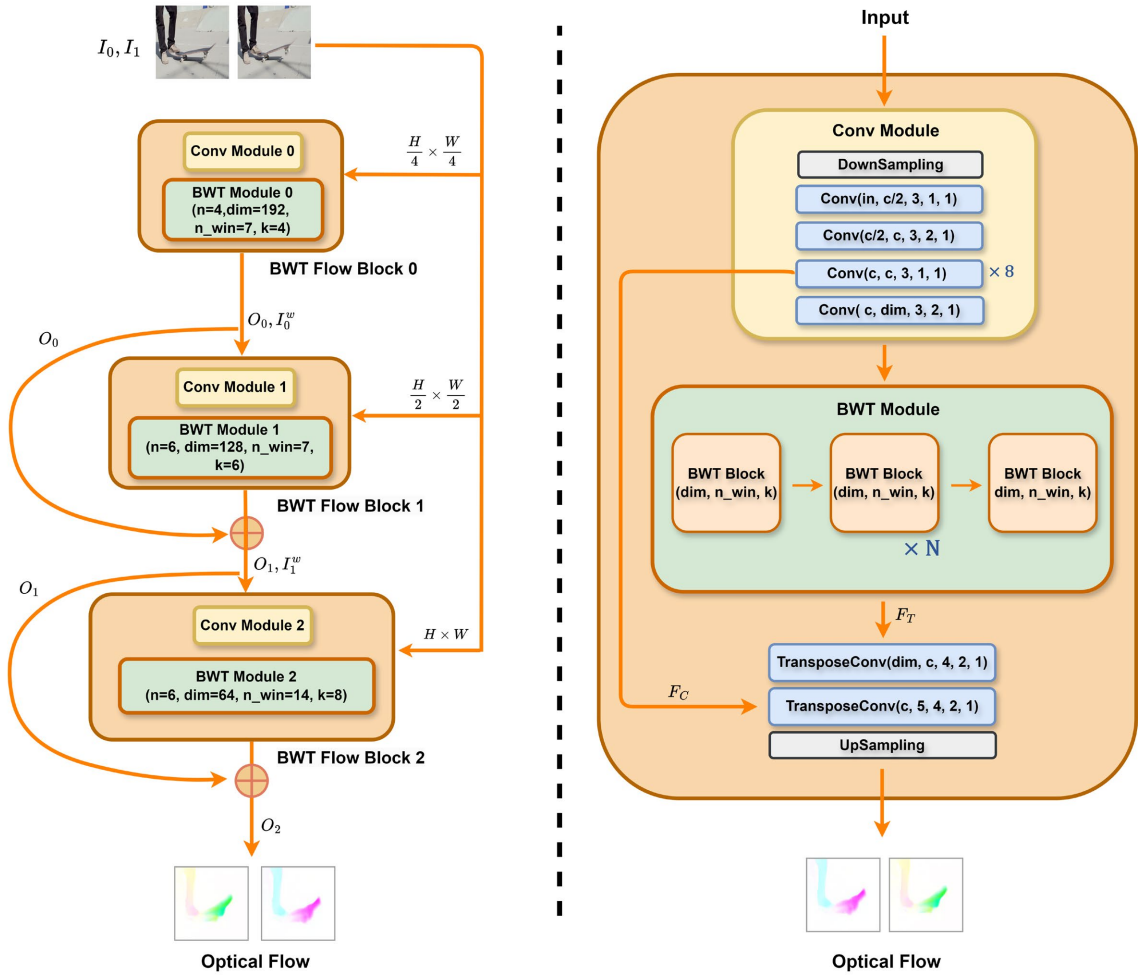

Figure 3. Design of BWT-FlowNet
图 3. BWT-FlowNet 的具体设计

如图 3 左侧所示，BWT-FlowNet 中的 BWT Flow Block 采用了三级金字塔结构，用于从粗到细地预测光流。光流 O_i 的计算公式为：

$$O_i = O_{i-1} + \text{BWT Flow Block}_i\left(O_{i-1}, I_{i-1}^w\right), \tag{4}$$

其中，O_i 表示由第 $i-1$ 个 BWT Flow Block 预测的光流，I_{i-1}^w 表示通过 O_{i-1} 对输入帧进行反向变形得到的变形帧。参数 k 随着特征尺寸的增加而持续增大，使得 BWT Flow Block 能够从最相关的 top-k 区域捕获更多特征。

(2) BWT Flow Block：将 CNN 与 Transformer 结合可以提升网络性能，因为 Transformer 能够捕捉长距离相关性和全局结构，而 CNN 则用于提取细粒度的局部特征。如图 3 右侧所示，我们的 BWT Flow Block 同时包含一个卷积模块(Conv Module)和一个双级窗口 Transformer 模块(BWT Module)，以实现更全面的特征提取和运动估计。

在 BWT Flow Block 中，通过两个连续的卷积块将特征通道数 c 增加，再使用 8 个连续的卷积块获取高维特征。随后，在 BWT 模块中使用 BWT 块处理特征，其通过双级路由注意力(Bi-Level Routing Attention, BRA)从最重要的区域捕捉长距离时空相关性特征。最后，将输出特征输入到转置卷积块中以计算光流，公式为：

$$O = \mathrm{TransConv}\left(F_C\right) + \mathrm{TransConv}\left(F_T\right), \tag{5}$$

其中，F_C 表示通过深度卷积获得的特征，F_T 表示由 BWT 块捕捉到的长距离相关性特征。

(3) BWT Block：双级路由注意力(BRA)是一种具有内容感知能力的稀疏注意力机制，它在 BWT 块中通过两阶段的过滤过程捕捉长距离依赖关系。首先，在区域级别构建区域亲和图，并剪枝掉与查询无关的区域。然后，在剩余的相关区域内执行细粒度的 token-to-token 注意力。

给定一个特征图 X，BWT 块中的 BRA 将其划分为 $S\times$S 个不重叠的区域，每个区域包含 HW/S^2 个特征向量。然后通过线性投影和权重矩阵 W^q, W^k, W^v 分别得到查询、键、值张量 Q, K, V，公式为：

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v, \tag{6}$$

其中，$X^r \in \mathcal{R}^{S^2 \times \frac{\mathrm{HW}}{S^2} \times \mathrm{C}}$ 是从 $X \in \mathcal{R}^{H \times \mathrm{W} \times \mathrm{C}}$ 重新排列得到的。通过对 Q 和 K 分别进行区域级平均，得到区域级查询和键 Q^r，$K^r \in \mathcal{R}^{S^2 C}$。随后，通过 $A^r = Q^r \left(K^r\right)^T$ 得到区域间亲和图 A^r，它表示两个区域在语义上的相关性。接着，BRA 选择相关性最高的 k 个索引，并将它们存储在索引矩阵 I^r 中，其中 $I^r = topkIndex\left(A^r\right)$。$I^r$ 的第 i 行包含与第 i 区域最相关的 k 个区域的索引。最后，得到输出 L，公式为：

$$L = Attention\left(Q, K^g, V^g\right) + LCE\left(V\right). \tag{7}$$

其中，K^g, V^g 是从最相关的 k 个区域中收集的键和值张量，而 LCE(V)表示局部上下文增强，如文献[35]所述。特别地，在最后一个 BWT Flow 块中，BRA 的局部窗口数量 S 被设置为 14，k 被设置为 8，从而使 BRA 能够专注于更细粒度的区域。注意力窗口大小的调整策略如图 4 所示。

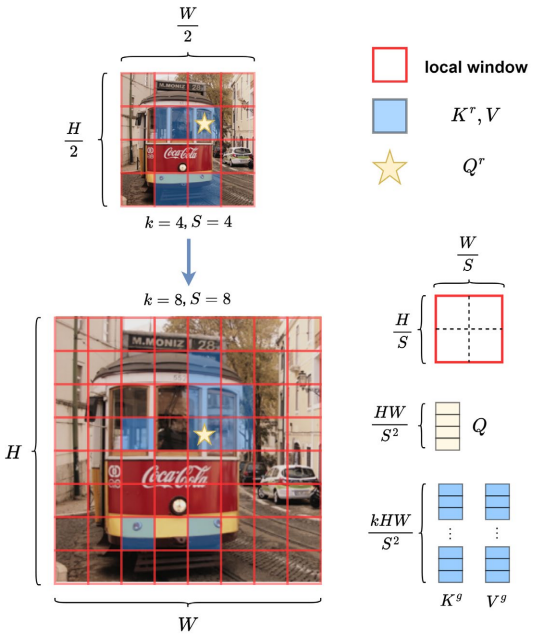

Figure 4. The change of attention window for BRA
图 4. BRA 注意力窗口的变化

2.3. Motion Boundary Mask

在视频序列中，两帧之间显著变化的区域往往只有全部内容的一小部分，而其他部分保持静止或仅简单运动，如图 5 所示。为此，需要区分出明显运动的区域，并在合成中间帧时关注这些部分。

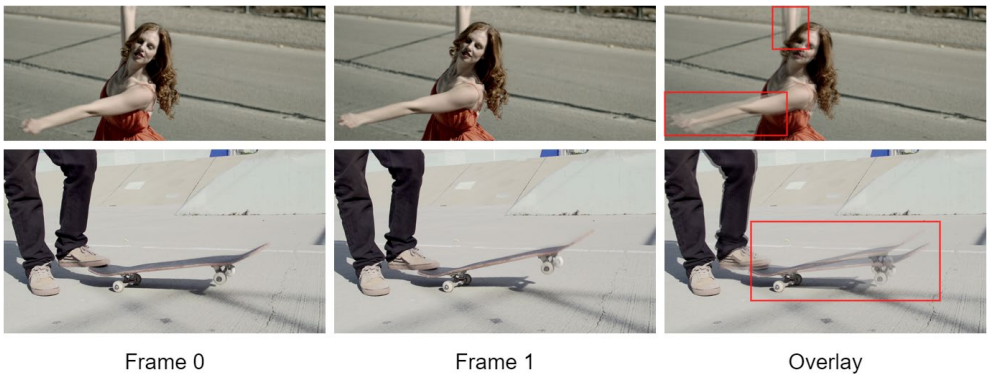

Figure 5. Two frames and their overlay
图 5. MB mask 的处理过程

借助 BWT-FlowNet 所获得的精确光流信息，掩模预测模块(Mask Predict Block)能够预测出限制运动区域的掩模 M。为帮助细化网络聚焦于内容变化区域，本文设计了一种运动边界块预测算法来处理掩模 M。首先，设定两个阈值 T_{low} 和 T_{high}，用于区分具有显著运动的区域。随后，将掩模值低于 T_{low} 或高于 T_{high} 的部分设置为 1，其余所有值设置为 0，从而将 M 转换为二值掩模 M_{binary}，其计算公式为：

$$M_{binary} = Threshold\left(M, T_{low}, T_{high}\right). \tag{8}$$

合成的中间帧通常在内容变化显著的运动边界附近存在模糊和伪影问题，而像素级掩模无法充分指示与运动相关的区域。为了适当扩展掩模区域以获取更多上下文特征，将 M_{binary} 的掩模区域转换为大小为 L × L 的块：

$$M_{patch} = Patchif\left(M_{binary}, L, N\right). \tag{9}$$

在每个块中，若有 N 个或更多像素的值为 1，则该块中的所有元素均被设为 1。最终获得运动边界掩模(MB mask)，记作 M_{patch}。整个过程如图 6 所示。其中，T_{low}、T_{high}、L 和 N 均为超参数，根据实验结果进行调整。

Figure 6. Processing of MB mask
图 6. MB mask 的处理过程

随后，对 M_{patch} 进行下采样，以获得金字塔结构的 MB mask，这些掩模被输入到 MBAR Net 的子编码器中，使其能够更加关注运动和内容变化区域。

具体来说，依据两个连续帧的运动信息，可预测出一个对应的 MB mask，如图 7 所示。该 MB mask

被送入到 MBAR Net 的子编码器，并与编码器的输入特征进行拼接，从而对特征中的部分区域进行强调。这种特征处理方法，能够使 MBAR Net 在细化中间帧时，关注运动边界相关的特征，最终提高这些部分的图像质量。

Figure 7. Overlay of two frames and the MB mask
图 7. 两帧的重叠和它们的 MB mask

2.4. Motion Boundary Aware Refinement Net

为了在中间帧的细化过程中专注于内容变化区域并提升这些区域的图像质量，本文开发了一种运动边界感知细化网络(Motion Boundary Aware Refinement Net，简称 MBAR Net)，该网络在不同的子编码器层中采用了金字塔结构的 MB mask，如图 8 所示。

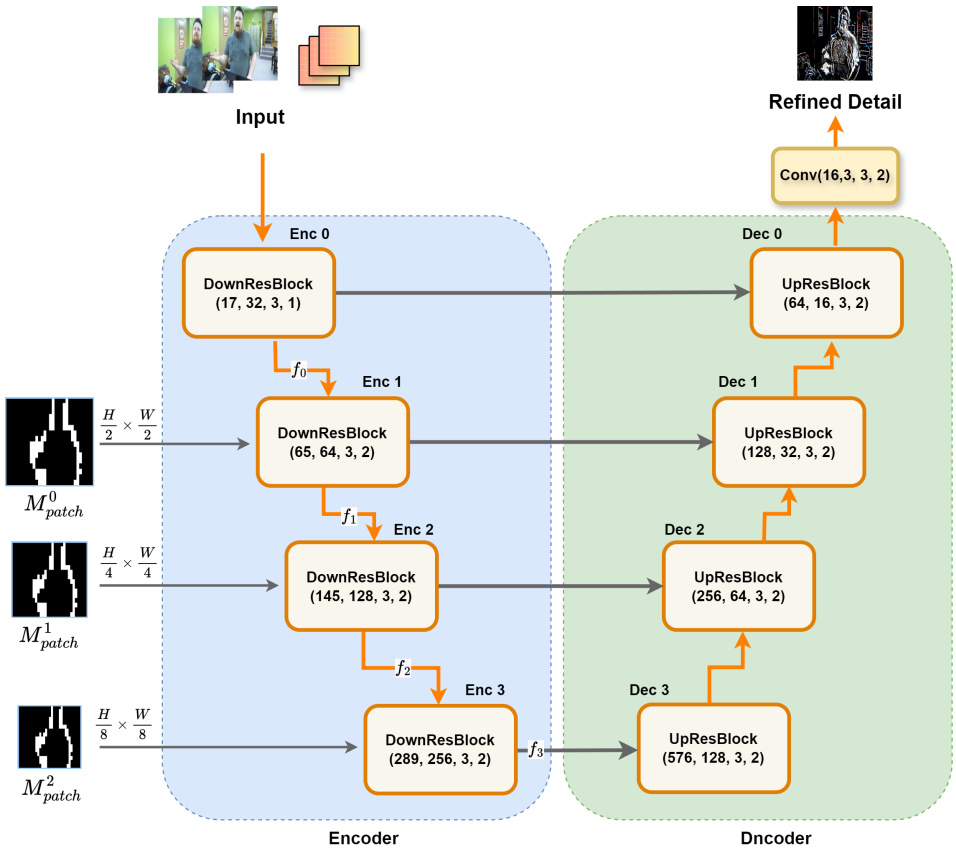

Figure 8. Architecture of Motion Boundary Aware Refinement Net
图 8. 运动边界感知细化网络的架构

为了与不同子编码器的输入特征尺寸相匹配，MB mask 被下采样为不同尺寸的掩模，记作 M^i_{patch}。子编码器的输入特征与 M^i_{patch} 进行拼接，然后通过 DownResBlock 进行处理。$Encoder_i$ 的输出特征 f_i 的计算公式为：

$$f_i = CBAM\left(Enc_i\left(M^i{}_{patch}, \hat{C}_i, f_{i-1}\right)\right). \tag{10}$$

在 MB mask 的引导下，MBAR Net 能够在编码和解码过程中更加关注运动区域。每个子编码器的输出特征通过 DownResBlock 中的 CBAM [36]进行处理，通过空间和通道注意力优化中间特征。此外，残差块[37]也被应用于 DownResBlock 和 UpResBlock 中，以有效学习特征，增强网络的泛化性能及稳定性。为在下采样过程中更好地保留特征信息，DownResBlock 中采用了平均池化，而非基于卷积的下采样方法。

2.5. 损失函数

本文采用 L_1 损失作为重建损失，用于建模中间帧的重建质量，公式为：

$$L_{rec} = I_t^{GT} - I_{t1}. \tag{11}$$

此外，为了衡量光照变化，本文使用了 Census 损失[38] [39] L_{cen}，其定义为 I_t^{GT} 和 I_t 的 Census 变换图像块之间的软汉明距离。

掩模感知损失(Mask Perceptual Loss)。运动边界周围的区域表现出更复杂、更大的运动。然而，一些损失函数会在整个内容上计算预测值与目标值之间的差异，未能更多地关注图像中的运动区域。为此，我们引入了基于提出的 MB mask 的掩模感知损失(MP 损失)，以解决这一问题。首先，使用 MB mask 分别与预测帧 I_t 和目标帧 I_t^{GT} 相乘，以保留内容变化的区域。随后，利用预训练的 VGG 网络提取丰富的感知特征，计算处理后的预测帧与目标帧之间的感知损失，公式为：

$$L_{mp} = VGG\left(M_{patch} \times I_t^{GT}, M_{patch} \times I_t\right). \tag{12}$$

蒸馏损失(Distillation Loss)。在光流估计中，本文采用了蒸馏损失 L_{dis} 来提高预测质量。在 BWT-FlowNet 中增加了一个额外的 BWT Flow Block 作为引导块，从目标帧接收信息，从而在训练过程中更准确地预测光流。蒸馏损失的计算公式为：

$$L_{dis} = O^{Teacher}_{t\to 0} - O_{t\to 01} + O^{Teacher}_{t\to 1} - O_{t\to 11}. \tag{13}$$

完整目标函数。我们的完整目标函数定义为：

$$L = L_{rec} + \lambda_{cen} L_{cen} + \lambda_{mp} L_{mp} + \lambda_{dis} L_{dis}, \tag{14}$$

其中，λ_{cen}，λ_{mp} 和 λ_{dis} 分别是 L_{cen}，L_{mp} 和 L_{dis} 的权重。

3. 实验结果与分析

3.1. 实验数据集

我们的模型在以下多个数据集上进行了评估：

Vimeo90K [12]：包含两个子集，即三元组(Triplet)和七元组(Septuplet)数据集，分辨率为 448 × 256。

UCF101 [40]：与人类行为相关，包含 379 个三元组，分辨率为 256 × 256。

Middlebury [41]：我们在 Middlebury 的 OTHER 子集上进行了测试，其中包含分辨率为约 640 × 480 的图像。

SNU-FILM [42]：包含 1240 个三元组，分辨率为 1280 × 720，根据难度分为四个子集：简单(Easy)、

中等(Medium)、困难(Hard)和极限(Extreme)。

本文在 Vimeo-90k 数据集上训练模型，并在其他数据集上评估其性能指标。

3.2. 实验细节

网络架构。在 BWT-FlowNet 中，本文为每个 BWT Flow Block 配置了不同的嵌入维度。每个 Flow Block 包含 6 个 BWT 块，用于捕捉相关窗口的数量和参数 k 的设置如图 3 所示。卷积部分(Conv Module 中的参数 c)的通道数分别设置为 360、225、135 和 135。在 MB mask 处理的实现中，T_{low} 和 T_{high} 分别设置为 0.2 和 0.8，而 L 和 N 分别设置为 8 和 2。对于本文的小型模型，第一个 BWT Flow Block 仅包含 4 个 BWT 块，且在经过 Conv Module 和 MBAR Net 后，特征尺寸被下采样为原始尺寸的一半。

训练细节。本文模型使用 AdamW 优化器进行训练，学习率设置为 1×10^{-4}。训练和测试均在 RTX 3090 GPU 上进行，批量大小设置为 8，模型总共训练 250 个周期。本文从训练样本中随机裁剪出 224 × 224 的图像块，并通过随机翻转和时间反转对其进行数据增强。

3.3. 与其他方法对比

Table 1. Quantitative comparison of VFI results on several datasets
表 1. 不同视频插帧方法在多个数据集上的定量指标对比

方法	Vimeo90K	UCF101	M.B.	SNU-FILM			
				Easy	Medium	Hard	Extreme
	PSNR/SSIM	PSNR/SSIM	IE.	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
ToFlow	33.73/0.968	34.58/0.967	2.15	39.08/0.989	35.31/0.977	28.44/0.918	23.39/0.831
SepConv	33.79/0.970	34.78/0.967	2.27	39.41/0.990	34.97/0.976	29.36/0.925	24.31/0.845
DAIN	34.71/0.976	34.99/0.968	2.04	39.73/0.990	35.46/0.978	30.17/0.934	25.09/0.858
CAIN	34.65/0.973	34.91/0.969	2.28	39.89/0.990	35.61/0.978	29.90/0.929	24.78/0.851
BMBC	35.01/0.976	35.15/0.969	2.04	39.90/0.990	35.31/0.977	29.33/0.927	23.92/0.843
RIFE	35.61/0.978	35.28/0.969	1.96	39.80/0.990	35.76/0.979	30.36/0.935	25.27/0.860
RIFE-Large	36.10/0.980	35.29/0.969	1.94	40.02/0.991	35.92/0.979	30.49/0.936	25.24/0.862
AdaCoF	34.47/0.973	34.90/0.968	2.24	39.80/0.990	35.05/0.975	29.46/0.924	24.31/0.844
M2M	35.47/0.978	35.28/0.969	2.09	39.66/0.990	35.74/0.979	30.30/0.936	25.08/0.860
MFRNet	35.96/.978	35.35/.970	1.90	—	—	—	—
ABME	**36.18/0.981**	35.38/0.970	2.01	39.59/0.990	35.77/0.979	**30.58/0.936**	**25.42/0.864**
Ours	36.01/0.980	**35.43/0.970**	**1.87**	**40.23/0.991**	**36.03/0.979**	30.35/0.934	25.06/0.855

本文将所提方法与最先进的视频插帧(VFI)方法进行了比较，这些方法包括基于核的方法 SepConv [4]、AdaCoF [43]，基于光流的方法 ToFlow [12]、DAIN [5]、SoftSplat [18]、CAIN [42]、BMBC [44]、RIFE [13] 和 ABME [45]，以及 MFRNet [23]和 M2M [46]。对于本文方法的结果，我们运行了 RIFE 和 IFRNet [21]的公开代码，并参考了 IFRNet 和[47]中的其他结果。本文采用常见的指标(如 PSNR 和 SSIM)进行定量评估，结果如表 1 所示。此外，本文还对提出方法与几种最先进的方法进行了视觉比较，如图 9 所示。在预测结果的第一行中，本文的模型能够清晰地渲染出运动幅度较大的手指，优于其他方法。其他视觉比较也证明了本文的模型具有卓越的视觉性能。

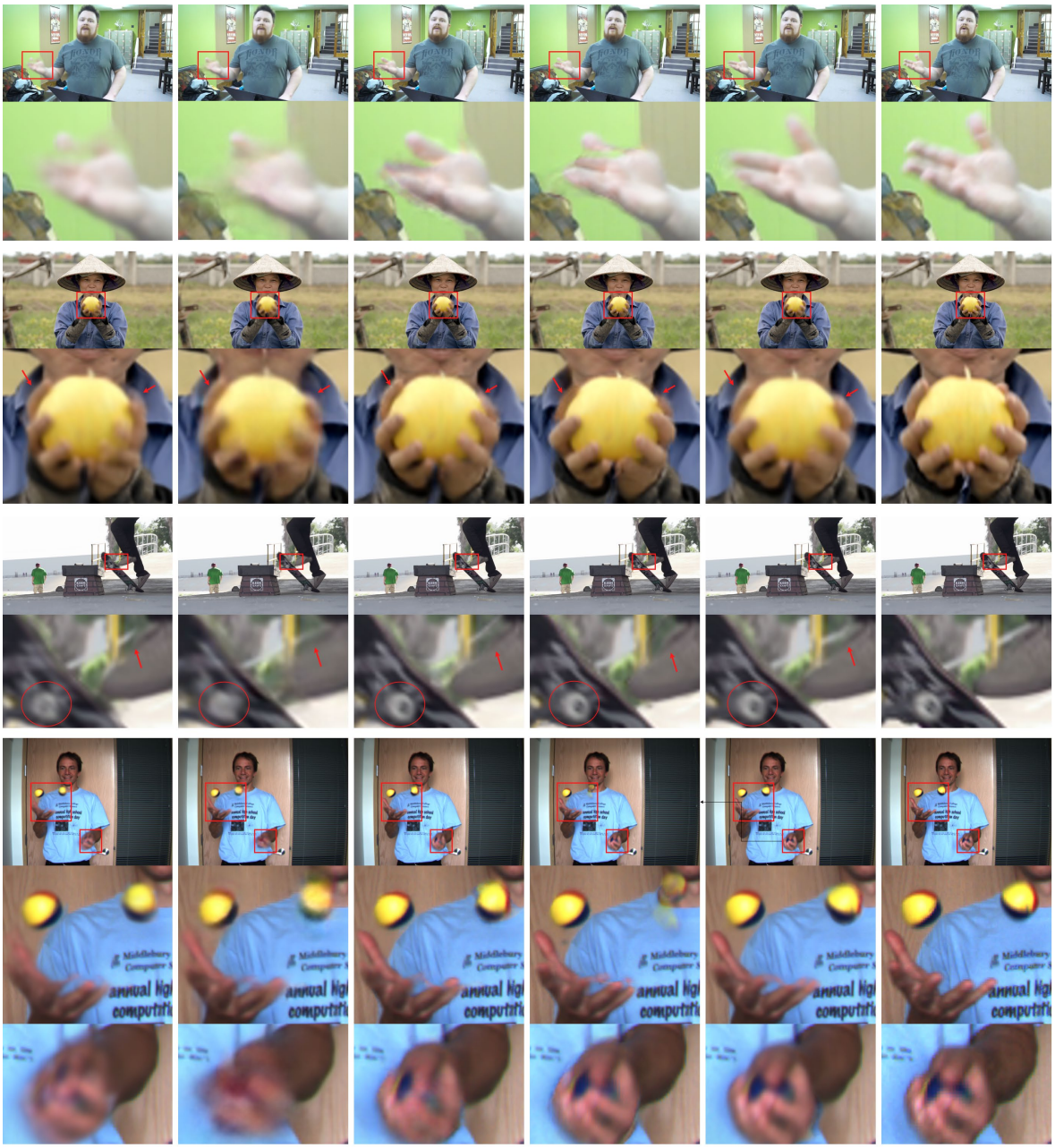

Figure 9. Visual comparison between different VFI algorithms
图 9. 不同视频插帧算法的可视化对比

3.4. 消融实验

在本节中，本文进行了几项消融研究，以探讨本文提出的方法。消融实验使用了几个参数较少的模型，并在 Vimeo90K 和 SNU-FILM 数据集上进行训练和测试。小型模型采用了较少的 BWT 块和缩减的卷积通道。

MB mask 的效果：在细化阶段使用多尺度 MB mask 作为引导以增强预测精度是本文设计的一种新方法。为了验证我们提出的两阶段框架的有效性，本文在 MB mask 上进行了消融实验。通过表 2 所示的

实验发现，采用 MB mask 的模型在 Vimeo90K 数据集上将 PSNR 提高了 0.108 dB，本文的模型也在 SNU-FILM 数据集上提升了性能。

Table 2. Ablation on MB mask
表 2. MB mask 的消融实验

模型	Vimeo90K	SNU-FILM	
		Medium	Hard
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
wo. M	35.668/.978	35.88/.979	30.23/.934
w. M	**35.776/.979**	**35.91/.980**	**30.29/.934**

BWT Block 效果：BWT 模块是捕捉长距离依赖关系的关键组件。本文通过改变模型中 BWT 块的数量来研究其有效性。表 3 比较了光流预测网络中不同数量 BWT 块的性能指标，包括无 BWT 块的版本、包含 4 个 BWT 块的版本以及包含 6 个 BWT 块的版本。可以观察到，在 Vimeo-90k 数据集上，模型 3 相较于模型 1 实现了 0.12 dB 更高的 PSNR；在 SNU-FILM 数据集的中等难度和高难度子集上，模型 3 也分别实现了 0.14 dB 更高的 PSNR。对比模型 2 和模型 3 可以明显看出，适当增加 BWT 块的数量能够提升模型预测的准确性。

Table 3. Ablation on BWT Block
表 3. BWT Block 的消融实验

模型	Vimeo90K	SNU-FILM	
		Medium	Hard
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
1) *wo.* BWT Block	35.61/.977	35.77/.979	30.17/.933
2) 4 BWT Blocks	35.65/.978	35.86/.979	30.28/.934
3) 6 BWT Blocks	**35.73/.978**	**35.91/.979**	**30.31/.934**

多尺度上下文编码器和 CBAM 的效果：编码器提取的多尺度图像特征用于中间帧合成，其中更准确的特征能够带来更精确的预测。本文比较了在细化网络中使用多尺度上下文编码器与普通卷积模块的性能。如表 4 所示，使用上下文编码器的模型在 Vimeo90K 数据集上相比仅使用普通卷积的模型，PSNR 提升了 0.13 dB。

此外，该模型还在 SNU-FILM 数据集的中等难度和高难度子集上显示出性能提升。进一步地，本文研究了在编码器中使用 CBAM 的效果。包含 CBAM 的模型在 Vimeo90K 数据集上比第二好的模型提升了 0.05 dB，在 SNU-FILM 数据集的中等难度和高难度子集上分别提升了 0.04 dB 和 0.03 dB。

Table 4. Ablation on Context Encoder and CBAM in refinement network
表 4. Context Encoder 和细化网络中 CBAM 的消融实验

模型	Vimeo90K	SNU-FILM	
		Medium	Hard
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
wo. Enc	35.65/.978	35.86/.979	30.28/.934
w. Enc	35.78/.979	35.90/.979	30.31/.934
w. Enc + CBAM	**35.83/.979**	**35.94/.979**	**30.34/.934**

掩模感知损失的效果：此外，本文还对掩模感知损失进行了消融实验，结果如表 5 所示。可以观察到，在引入掩模感知损失后，模型在 Vimeo90K 数据集上实现了 0.031 dB 的 PSNR 提升，在 SNU-FILM 数据集的中等难度和高难度子集上分别实现了 0.12 dB 和 0.06 dB 的提升。

Table 5. Ablation on Mask Perceptual loss
表 5. 掩码感知损失函数的消融实验

模型	Vimeo90K	SNU-FILM	
		Medium	Hard
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
wo. MP loss	35.776/.979	35.91/.905	30.29/.934
w. MP loss	**35.807/.979**	**36.03/.980**	**30.35/.934**

此外，本文还评估了不同消融模型的推理结果，如图 10 所示。从图中对比可以证明，BWT-FlowNet、上下文编码器、MB mask 和 MP 损失的整合显著提升了我们模型生成更准确中间帧的能力。

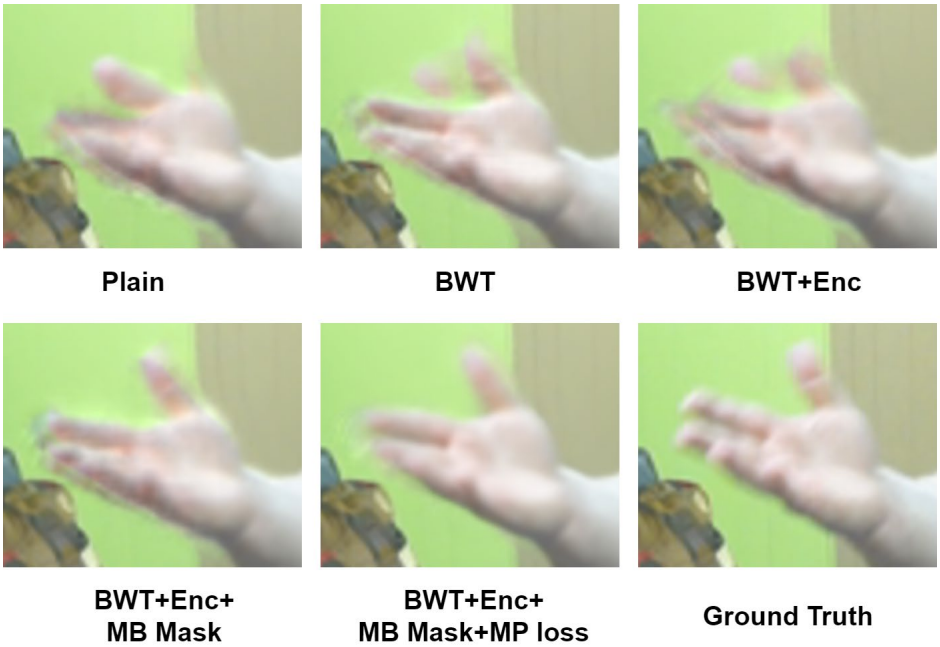

Figure 10. Visual comparison of inference results from different ablation models
图 10. 不同消融实验模型的推理结果对比

4. 结论

在本研究中，我们提出的视频插帧框架在光流运动信息的约束下有效地细化了中间帧。运动边界掩模(MB mask)被用于指示内容变化区域，从而指导中间帧的细化。BWT-FlowNet 在光流估计方面表现出色，能够有效处理大运动和遮挡问题。此外，本文开发了 MBAR Net 以充分利用 MB mask 和其他特征来细化中间帧，并引入了掩模感知损失函数，以有效关注合成帧中的内容变化区域。这种基于掩模的损失函数与视频插帧(VFI)任务的需求高度契合，也可被其他 VFI 方法采用。定量和定性评估表明，本文提出的方法具有卓越的性能和泛化能力。在未来的工作中，我们将探索新的方法以提升光流估计和中间帧细化的效果，从而进一步优化视频插帧技术。

参考文献

[1] Flynn, J., Neulander, I., Philbin, J. and Snavely, N. (2016) Deep Stereo: Learning to Predict New Views from the World's Imagery. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5515-5524. https://doi.org/10.1109/cvpr.2016.595

[2] Wu, C.Y., Singhal, N. and Krahenbuhl, P. (2018) Video Compression through Image Interpolation. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Tel Aviv, 23-27 October 2018, 416-431.

[3] Liu, Z., Yeh, R.A., Tang, X., Liu, Y. and Agarwala, A. (2017) Video Frame Synthesis Using Deep Voxel Flow. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4473-4481. https://doi.org/10.1109/iccv.2017.478

[4] Niklaus, S., Mai, L. and Liu, F. (2017) Video Frame Interpolation via Adaptive Separable Convolution. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 261-270. https://doi.org/10.1109/iccv.2017.37

[5] Bao, W., Lai, W., Ma, C., Zhang, X., Gao, Z. and Yang, M. (2019) Depth-Aware Video Frame Interpolation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 3698-3707. https://doi.org/10.1109/cvpr.2019.00382

[6] Siyao, L., Zhao, S., Yu, W., Sun, W., Metaxas, D., Loy, C.C., *et al*. (2021) Deep Animation Video Interpolation in the Wild. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 6583-6591. https://doi.org/10.1109/cvpr46437.2021.00652

[7] Niklaus, S., Mai, L. and Liu, F. (2017) Video Frame Interpolation via Adaptive Convolution. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2270-2279. https://doi.org/10.1109/cvpr.2017.244

[8] Peleg, T., Szekely, P., Sabo, D. and Sendik, O. (2019) IM-Net for High Resolution Video Frame Interpolation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 2393-2402. https://doi.org/10.1109/cvpr.2019.00250

[9] Shi, Z., Liu, X., Shi, K., Dai, L. and Chen, J. (2020) Video Interpolation via Generalized Deformable Convolution. arXiv: 2008.10680.

[10] Meyer, S., Wang, O., Zimmer, H., Grosse, M. and Sorkine-Hornung, A. (2015) Phase-Based Frame Interpolation for Video. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1410-1418. https://doi.org/10.1109/cvpr.2015.7298747

[11] Meyer, S., Djelouah, A., McWilliams, B., Sorkine-Hornung, A., Gross, M. and Schroers, C. (2018) PhaseNet for Video Frame Interpolation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 498-507. https://doi.org/10.1109/cvpr.2018.00059

[12] Xue, T., Chen, B., Wu, J., Wei, D. and Freeman, W.T. (2019) Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision*, **127**, 1106-1125. https://doi.org/10.1007/s11263-018-01144-2

[13] Huang, Z., Zhang, T., Heng, W., Shi, B. and Zhou, S. (2022) Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 624-642. https://doi.org/10.1007/978-3-031-19781-9_36

[14] Chi, Z., Mohammadi Nasiri, R., Liu, Z., Lu, J., Tang, J. and Plataniotis, K.N. (2020) All at Once: Temporally Adaptive Multi-Frame Interpolation with Advanced Motion Modeling. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 107-123. https://doi.org/10.1007/978-3-030-58583-9_7

[15] Jiang, H., Sun, D., Jampani, V., Yang, M., Learned-Miller, E. and Kautz, J. (2018) Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 9000-9008. https://doi.org/10.1109/cvpr.2018.00938

[16] Liu, Y., Liao, Y., Lin, Y. and Chuang, Y. (2019) Deep Video Frame Interpolation Using Cyclic Frame Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8794-8802. https://doi.org/10.1609/aaai.v33i01.33018794

[17] Niklaus, S. and Liu, F. (2018) Context-Aware Synthesis for Video Frame Interpolation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1701-1710. https://doi.org/10.1109/cvpr.2018.00183

[18] Niklaus, S. and Liu, F. (2020) Softmax Splatting for Video Frame Interpolation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5436-5445. https://doi.org/10.1109/cvpr42600.2020.00548

[19] Xu, X., Siyao, L., Sun, W., Yin, Q. and Yang, M.H. (2019) Quadratic Video Interpolation. arXiv: 1911.00627.

[20] Zhang, H., Zhao, Y. and Wang, R. (2019) A Flexible Recurrent Residual Pyramid Network for Video Frame Interpolation. ICCV.

[21] Kong, L., Jiang, B., Luo, D., Chu, W., Huang, X., Tai, Y., *et al*. (2022) IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 1959-1968. https://doi.org/10.1109/cvpr52688.2022.00201

[22] Jin, X., Wu, L., Chen, J., Chen, Y., Koo, J. and Hahm, C. (2023) A Unified Pyramid Recurrent Network for Video Frame Interpolation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 1578-1587. https://doi.org/10.1109/cvpr52729.2023.00158

[23] Zhu, G., Qin, Z., Ding, Y., Liu, Y. and Qin, Z. (2024) MFNet: Real-Time Motion Focus Network for Video Frame Interpolation. *IEEE Transactions on Multimedia*, **26**, 3251-3262. https://doi.org/10.1109/tmm.2023.3308442

[24] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., *et al*. (2015) FlowNet: Learning Optical Flow with Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2758-2766. https://doi.org/10.1109/iccv.2015.316

[25] Sun, D., Yang, X., Liu, M. and Kautz, J. (2018) PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8934-8943. https://doi.org/10.1109/cvpr.2018.00931

[26] Teed, Z. and Deng, J. (2020) RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 402-419. https://doi.org/10.1007/978-3-030-58536-5_24

[27] Liu, P., King, I., Lyu, M.R. and Xu, J. (2019) DDFlow: Learning Optical Flow with Unlabeled Data Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8770-8777. https://doi.org/10.1609/aaai.v33i01.33018770

[28] Liu, P., Lyu, M., King, I. and Xu, J. (2019) Selflow: Self-Supervised Learning of Optical Flow. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4566-4575. https://doi.org/10.1109/cvpr.2019.00470

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. 2017 *Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5998-6008.

[30] d'Ascoli, S., Touvron, H., Leavitt, M.L., *et al*. (2021) Convit: Improving Vision Transformers with Soft Convolutional Inductive Biases. *Proceedings of the 38th International Conference on Machine Learning*, 18-24 July 2021, 2286-2296.

[31] Li, Y., Zhang, K., Cao, J., *et al*. (2021) Localvit: Bringing Locality to Vision Transformers. arXiv: 2104.05707.

[32] Wang, C., Xu, H., Zhang, X., Wang, L., Zheng, Z. and Liu, H. (2022) Convolutional Embedding Makes Hierarchical Vision Transformer Stronger. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 739-756. https://doi.org/10.1007/978-3-031-20044-1_42

[33] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., *et al*. (2021) CvT: Introducing Convolutions to Vision Transformers. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 22-31. https://doi.org/10.1109/iccv48922.2021.00009

[34] Zhu, L., Wang, X., Ke, Z., Zhang, W. and Lau, R. (2023) BiFormer: Vision Transformer with Bi-Level Routing Attention. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 10323-10333. https://doi.org/10.1109/cvpr52729.2023.00995

[35] Ren, S., Zhou, D., He, S., Feng, J. and Wang, X. (2022) Shunted Self-Attention via Multi-Scale Token Aggregation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10843-10852. https://doi.org/10.1109/cvpr52688.2022.01058

[36] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018., Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[37] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[38] Meister, S., Hur, J. and Roth, S. (2018) UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 7251-7259. https://doi.org/10.1609/aaai.v32i1.12276

[39] Zhong, Y., Ji, P., Wang, J., Dai, Y. and Li, H. (2019) Unsupervised Deep Epipolar Flow for Stationary or Dynamic Scenes. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 12087-12096. https://doi.org/10.1109/cvpr.2019.01237

[40] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. arXiv: 1212.0402.

[41] Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J. and Szeliski, R. (2010) A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, **92**, 1-31. https://doi.org/10.1007/s11263-010-0390-2

[42] Choi, M., Kim, H., Han, B., Xu, N. and Lee, K.M. (2020) Channel Attention Is All You Need for Video Frame Interpolation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 10663-10671. https://doi.org/10.1609/aaai.v34i07.6693

[43] Lee, H., Kim, T., Chung, T., Pak, D., Ban, Y. and Lee, S. (2020) AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5315-5324. https://doi.org/10.1109/cvpr42600.2020.00536

[44] Park, J., Ko, K., Lee, C. and Kim, C. (2020) BMBC: Bilateral Motion Estimation with Bilateral Cost Volume for Video Interpolation. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 109-125. https://doi.org/10.1007/978-3-030-58568-6_7

[45] Park, J., Lee, C. and Kim, C. (2021) Asymmetric Bilateral Motion Estimation for Video Frame Interpolation. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 14519-14528. https://doi.org/10.1109/iccv48922.2021.01427

[46] Hu, P., Niklaus, S., Sclaroff, S. and Saenko, K. (2022) Many-to-Many Splatting for Efficient Video Frame Interpolation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 3543-3552. https://doi.org/10.1109/cvpr52688.2022.00354

[47] Lu, L., Wu, R., Lin, H., Lu, J. and Jia, J. (2022) Video Frame Interpolation with Transformer. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 3522-3532. https://doi.org/10.1109/cvpr52688.2022.00352