基于光流引导的端到端视频补全框架*

Zhen Li[1†] Cheng-Ze Lu[1†] Jianhua Qin[2] Chun-Le Guo[1‡] Ming-Ming Cheng[1]
[1]TKLNDST, CS, Nankai University [2]Hisilicon Technologies Co. Ltd.
zhenli1031@gmail.com, czlu919@outlook.com, qinjianhua@hisilicon.com
{guochunle, cmm}@nankai.edu.cn

摘要

在最近的视频补全方法中，常用到利用传播像素的轨迹来捕捉跨帧运动信息的光流。然而，这些方法中，基于光流的手工处理过程是单独应用在视频补全流程的。因此，这些方法的效率较低，而且严重依赖早期阶段的中间结果。在本文中，我们通过精心设计的三个可训练的模块（光流补全、特征传播和内容生成），提出了一个端到端的视频补全框架（E^2FGVI）。这三个模块与之前基于光流的方法的三个阶段相对应，但可以共同优化，从而使得视频补全过程更加高效。实验结果表明，我们提出的方法在定性定量上都优于 SOTA 方法，并显示出较好的效率。我们的代码开源在:https://github.com/MCG-NKU/E2FGVI.

1. 引言

视频补全 (Video inpainting) 的目的是在整个视频片段中用合理和连贯的内容来填补“损坏的”区域。它被广泛地应用于现实世界的应用，如物体移除 (object removal) [16]，视频修复 (video restoration) [28] 和视频补全 (video completion) [7, 40]。

尽管图像补全 (image inpainting) 已经取得了重大进展 [43, 61, 62]，但由于复杂的视频场景和损坏的视频帧，视频补全 (video inpainting) 仍然充满挑战。如果直接对每一帧进行图像补全，往往会使视频在时间上产生不一致性，并产生严重的伪影。而在高质量的视频补全中，空间结构和时间一致性都需要被考虑。深度学习的最新进展促使研究人员探索出更有效的解决方案 [7, 8, 17, 23, 28, 34, 39, 51, 58, 64]。

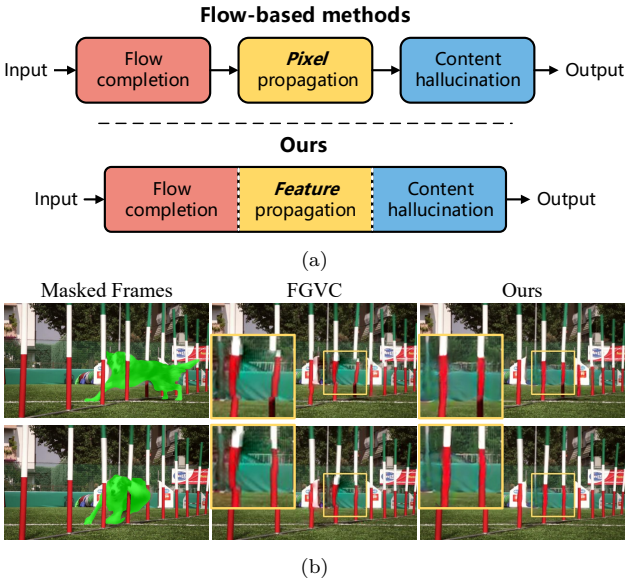

图 1. (a) 传统的基于光流的方法 [17, 58] 和我们的方法. 以前的基于光流的方法分别进行这三个阶段，而我们的相应模块是以端到端的方式工作。(b) 我们的方法与基于光流的 SOTA 方法 FGVC [17]. 由于在内容生成过程中的错误积累和对时间信息的忽视，与我们的方法相比,FGVC 未能产生可靠的、具有时间一致性的结果。

在这些方法中，典型的基于光流的方法 [17, 58] 将视频补全视为一个像素传播问题，以保持时间上的一致性。如图1 (a) 所示，这些方法可以分解为三个相互关联的阶段。(1) 光流补全: 首先应该估计光流，因为在损坏的区域没有光流场会影响后面的过程。(2) 像素传播: 他们通过在已补全的光流的引导下，用可见区域双向传播像素来填补损坏的视频中的空缺。(3) 内容生成: 在传播之后，剩余的缺失区域可以通过预先训练的图像补全网络进行填充 [61, 62]。

*本文为 CVPR'22 论文 [30] 的中文翻译版。
†具有相同贡献
‡C.L. Guo 是通讯作者.

不幸的是, 即使可以获得令人印象深刻的结果, 但由于前两个阶段涉及许多手工操作 (e.g. 泊松混合、求解稀疏线性方程和索引每个像素的流动轨迹), 整个基于光流的补全过程必须单独进行。独立的过程引起了两个主要问题。一个是在早期阶段发生的错误会在后续阶段积累和放大, 这进一步大大影响了最终的性能。具体来说, 不准确的光流估计会误导像素的传播, 并进一步混淆内容生成阶段, 产生不准确的补全结果。其次, 这些复杂的手工设计的操作只有在没有 GPU 加速的情况下才能处理。因此, 推理视频序列的整个过程是非常耗时的。以 DFVI [58] 为例, 补全一个大小为 432×240, 包含约 70 个帧的视频 [45], 需要约 4 分钟[1], 这在大多数实际应用中是不可接受的。此外, 除了上述缺点外, 在内容生成阶段只使用预先训练好的图像补全网络, 忽略了跨时空邻域的内容关系, 导致视频中生成的内容不一致 (见图1 (b))。

为了解决这些问题, 在本文中, 我们精心设计了三个可训练的模块, 包括 (1) 光流补全、(2) 特征传播和 (3) 内容生成模块, 这些模块模拟了基于光流的方法中的相应阶段, 并进一步构成了光流引导的视频补全的端到端框架 (E^{2}FGVI)。三个模块之间的密切协作, 缓解了以前独立系统 [17, 23, 26, 58, 68] 对中间结果的过度依赖, 并能以更有效的方式工作。

具体来说, 对于光流补全模块, 我们直接一步将其应用于待补全的视频, 避免多步复杂的操作。对于特征传播模块, 与像素级的传播不同, 我们的光流引导传播过程是在可变卷积的帮助下在特征空间进行的。通过更多可学习的采样偏移和特征级操作, 传播模块减轻了不准确光流对模型造成的影响。对于内容生成模块, 我们提出了一个 temporal focal transformer, 以有效地模拟空间和时间维度上的长距离依赖关系。在这个模块中, 局部和非局部的时间邻域都被考虑在内, 从而能够产生更具有时间连贯性的补全结果。

实验结果表明我们的框架具有以下两个优势:

- 目前最优准确率: 与 SOTA 方法相比, 该方法 E^{2}FGVI 在两个常用的面向失真度的指标 (i.e., PSNR and SSIM [54])、一个面向感知的指标 (i.e., VFID [52]) 和一个时间一致性衡量指标 (i.e., E_{warp} [25]) 方面取得了显著的改进。
- 高效: 我们的方法在 Titan Xp GPU 上以每帧 0.12 秒的速度处理 432×240 的视频, 这比以前基于光流的方法快了近 15 倍。与同样可以端到端部署的方法相比, 我们的方法显示出不错的推理速度。此外, 在所有被比较的 SOTA 方法中, 我们的方法具有最低的计算复杂性 (FLOPs)。

希望我们提出的具有上述优势的端到端框架可以作为视频补全邻域的一个强有力的基线。

2. 相关工作

Video inpainting. 在深度学习发展的基础上, 视频补全已经取得了巨大的进展。这些方法可以大致分为三类: 基于三维卷积方法 [8, 21, 51], 基于光流的方法 [17, 58], 和基于注意力机制的方法 [28, 29, 34, 64]. 一些采用三维卷积和注意力的方法 [7, 23, 28, 51] 通常会产生时间上不一致的结果, 这是因为时间上的接受区域有限。为了产生更多的时间连贯性结果, 许多工作 [23, 68] 将光流视为视频补全的强大先验因素, 并将其纳入网络。然而, 直接计算无效区域内的图像之间的光流是非常困难的, 因为这些区域本身就成为阻碍因素, 限制了性能。最近, 基于光流 [17, 58] 的方法首先进行光流的补全, 并使用被补全的光流沿其轨迹传播索引的像素。不过我们没有手工进行像素级传播, 而是设计了一个端到端的可训练框架, 在特征空间进行传播。此外, 我们的方法还受益于最近一些使用 transformers 来提升补全效果的工作 [33, 34, 64]。

基于光流的视频处理. 跨帧的运动信息可以很好地帮助处理许多与视频相关的任务, 如视频理解 [3, 32], 视频分割 [11, 49], 视频目标检测 [67], 深度估计 [18, 37], 视频超分辨率 [4, 59], 插帧 [22, 27] 等等。具体来说, 许多视频修复和增强算法 [4, 24, 41, 47, 59] 依靠光流对齐来补偿帧之间的信息。最近的工作 [4, 27, 53, 55, 56] 利用可变卷积 [65] 来模拟光流的行为, 但它具有更多可学习的偏置, 以实现更有效的对齐。我

[1]我们在 Intel(R) Core(TM) i7-6700K CPU,NVIDIA Titan Xp GPU 上测试

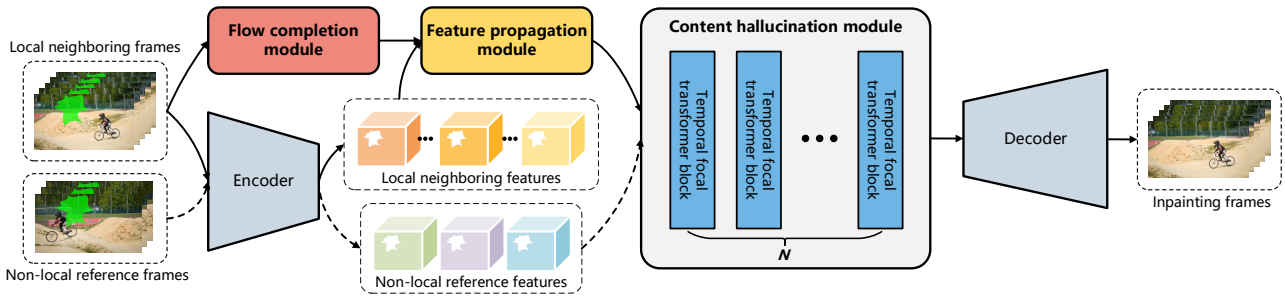

图 2. E^{2}FGVI 的概览. 它包括: 1) 帧级内容编码器,2) 光流补全模块,3) 特征传播模块,4) 由多个 temporal focal transformer 块组成的内容生成模块,5) 帧级解码器。

们的工作也与这些工作有相同的优点。

Vision transformer. 最近, Transformer [50] 在视觉邻域获得了很多关注。Vision Transformer [15] 以及它的跟进工作 [19,35,48,60,63] 在图像和视频表示学习方面取得了令人印象深刻的表现 [9,13,36,44], 如图像生成 [42], 目标检测 [2,66], 和许多其他应用 [10,12,20,31]。由于自注意力机制的二次复杂性, 许多工作部署了有效的基于窗口的注意力机制 [14,35,60], 以减少其计算复杂性, 同时提高模型在有限感受野的能力。Swin Transformer [35] 通过转移本地窗口计算自注意力, 加强了局部关系。Focal Transformer [60] 引入焦点式的自注意力机制, 增强了全局与局部的交互。

3. 方法

给出一个长度为 T 的有损视频序列 $\{X^t \in \mathbb{R}^{H\times W\times 3} \mid t = 1\dots T\}$ 和相应的逐帧二进制掩码 $\{M^t \in \mathbb{R}^{H\times W\times 1} \mid t = 1\dots T\}$, 我们的目标是合成可靠的视频内容, 使它在空间和时间维度上都与被破坏 (遮掩) 的区域一致。在下文中, 我们将讨论我们方法的主要组成部分。首先, 我们使用了一个上下文编码器, 它将所有被破坏的帧编码为较低分辨率的特征, 以便在后续处理中提高计算效率。其次, 我们通过一个光流补全模块提取并补全局部邻域之间的光流 (第3.1节)。第三, 补全的光流协助从局部邻域中提取的特征来完成特征对齐和双向传播 (第3.1节)。第四, 多层 temporal focal transformers 通过将传播的局部邻域特征与非局部参考特征相结合来进行内容生成 (第3.2节)。最后, 一个解码器将填充的特征放大, 并将其重建为最终的视频序列 $\{\hat{Y}^t \in \mathbb{R}^{H\times W\times 3} \mid t = 1\dots T\}$。

图2 显示了我们提出模型的整个流程 (E^{2}FGVI)。值得注意的是, 所有的模块都是可微的, 并构成了一个端到端的可训练架构。

3.1. 光流补全和特征传播

在本节中, 我们将详细介绍我们方法中与光流有关的操作。请注意, 我们只将基于光流的模块应用于从局部相邻帧中提取的特征, 因为由于经常发生在非局部帧中的大运动的存在, 光流估计会退化甚至失败。此外, 为了提高计算效率, 与光流相关的操作是在低分辨率空间进行的。

端到端的光流补全. 在光流预测之前, 我们首先以 1/4 的分辨率对原始损坏的帧 X^t 进行下采样, 这与编码的低分辨率特征的空间分辨率相匹配。下采样的帧被表示为 $X^t_{\downarrow} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$. 相邻帧 i 和 j 之间的光流预测是由光流估计网络计算出来的 $\mathcal{F}$:

$$\hat{F}_{i\to j} = \mathcal{F}(X^i_{\downarrow}, X^j_{\downarrow}). \tag{1}$$

我们使用来自轻量级光流估计网络的预训练权重来初始化该网络, 以利用其丰富的光流知识。

继大多数基于光流的视频补全方法 [17,58] 之后, 我们通过 Eq. (1)估计前向光流 $\hat{F}_{t\to t+1}$ 和后向光流 $\hat{F}_{t\to t-1}$, 用于光流引导的双向传播。由于损坏的视频中的缺失区域成为光流估计的遮挡因素, 严重影响了估计光流的质量, 因此我们需要在使用它们进行特征传播之前恢复前向和后向光流。为了简单

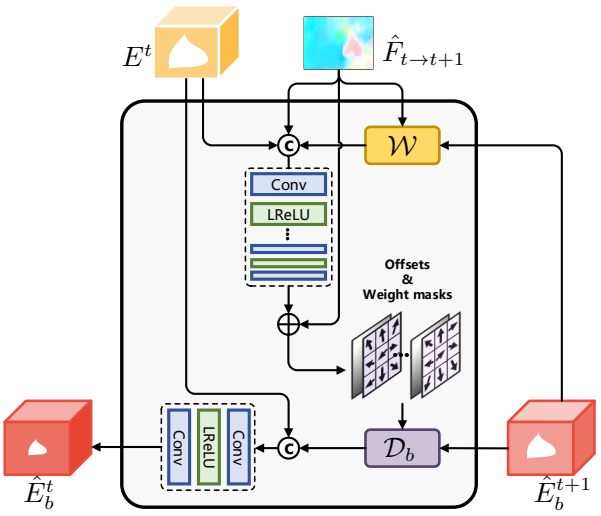

图 3. 使用已补全的前向流 $\hat{F}_{t\to t+1}$ 来指导特征后向传播的例子. 其中,⊕ 和 © 分别表示加法运算和连接运算。请注意, 后向流将以相反的方向进行。

起见, 我们使用 L1 损失[2]来补全双向光流。

$$\mathcal{L}_{flow} = \sum_{t=1}^{T-1} \|\hat{F}_{t\to t+1} - F_{t\to t+1}\|_1 + \sum_{t=2}^{T} \|\hat{F}_{t\to t-1} - F_{t\to t-1}\|_1, \tag{2}$$

其中 $F_{t\to t+1}$ 和 $F_{t\to t-1}$ 分别是真实的前向和后向光流, 它们是从未被破坏的原始视频中计算出来的。

我们的光流补全模块与 DFVI [58] 和 FGVC [17] 主要有两个方面的差异: (1) DFVI 和 FGVC 是分别部署了光流补全网络和传播算法。相比之下, 我们的光流补全模块可以以端到端的方式与其他网络组件一起训练, 这有利于该模块生成面向任务的光流 [59]。(2) DFVI 和 FGVC 的光流补全效率较低 ($>$ 0.4s/flow), 因为它们需要先初始化光流, 然后用多个阶段细化。而我们只用一个前馈传递来估计和补全光流, 速度会快得多 ($<$ 0.01s/flow)。

光流引导的特征传播. 假设 $\{E^t \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C} \mid t = 1\dots T_l\}$ 是从上下文编码器中提取的局部时间邻域特征, 其中 T_l 表示局部相邻帧的长度。以前向流 $\hat{F}_{t\to t+1}$ 为例, 它可以帮助我们捕捉到从第 t 帧到第 t 帧的损坏区域的运动。一旦第 t 帧内容特征处的损坏区域的像素在第 t+1 帧特征处的有效区域是已知的, 在前向流 $\hat{F}_{t\to t+1}$ 的帮助下, 我们可以通过 warping 第 t+1 帧后向传播特征 $\hat{E}_b^{t+1}$ 到当前时间步来直观地利用这一有效信息。warping 的特征可以进一步与当前内容特征 E^t 合并, 并通过后向传播函数 $\mathcal{P}_b(\cdot)$ 进行更新:

$$\hat{E}_b^t = \mathcal{P}_b(E^t, \mathcal{W}(\hat{E}_b^{t+1}, \hat{F}_{t\to t+1})), \tag{3}$$

其中 $\mathcal{W}(\cdot)$ 表示基于光流的空间 warping 操作,$\hat{E}_b^t$ 是第 t 时间步长的反向传播特征, 传播函数 $\mathcal{P}_b(\cdot)$ 代表两个具有 LeakyReLU [38] 激活的卷积层。

Eq. (3)中的 warping 和合并操作近似于 DFVI 和 FGVC 中的整个传播过程, 但我们在特征空间而不是图像空间中进行这些操作。传播特征 $\hat{E}_b^t$ 随着合理的内容逐渐涉及到每个内容特征的损坏区域而逐步更新, 这也有利于在流引导的作用下, 关联所有跨局部邻域特征。与基于光流的方法中非常耗时且严重依赖流估计质量的手工制作像素级传播不同, 特征级传播利用卷积层在更大的感受野的作用下自适应地将光流追踪信息合并, 并可以通过 GPU 加速。

尽管特征级传播可以比 FGVC 和 DFVI 更快、更有效, 但它仍然需要面对 Eq. (1)中流估计结果不准确造成的问题, 这将在传播过程中带来不相关的信息, 进一步阻碍最终的性能。为了缓解这个问题, 受 [4–6, 53] 的启发, 我们采用了调制的可变卷积 [65] 来进一步索引和加权候选特征点。如图3所示, 我们首先计算权重掩码 $W_{t\to t+1}$ 和与估计光流相关的偏移 $\Delta F_{t\to t+1}$。

$$[W_{t\to t+1}, \Delta F_{t\to t+1}] = \mathcal{C}_b(E^t, \mathcal{W}(\hat{E}_b^{t+1}, \hat{F}_{t\to t+1}), \hat{F}_{t\to t+1}), \tag{4}$$

其中 $\mathcal{C}_b(\cdot)$ 表示多个级联卷积层。计算出的权重掩码 $M_{t\to t+1}$ 和偏移量 $\Delta F_{t\to t+1}$ 的大小都是 $\frac{H}{4} \times \frac{W}{4} \times K^2 \times G$, 其中 K 和 G 分别是可变卷积的核大小和群数。我们可以通过将偏移量 $\Delta F_{t\to t+1}$ 加入到已补全的光流 $\hat{F}_{t\to t+1}$ 中, 进一步生成每个空间位置的 $K^2 \times G$ 候选特征点。偏移量 $\Delta F_{t\to t+1}$ 和补全的光流 $\hat{F}_{t\to t+1}$ 之间的关系是互利的。一方面, 更灵活的采样位置可以很好地弥补光流补全的不准确。另一方面, 补全的光流提供了较好的初始采样位置, 这使得它很容易在其周围环境中找到更有意义的内容。然后, 我们使用可变卷积层对后向特征 $\hat{E}_b^{t+1}$ 进行 warping, 而不使用 Eq. (3)中基于光流的 warping,

[2]其他损失函数也可以用在 Eq. (2)中, 但我们没有观察到它对最终的补全性能有明显改善。

并进一步通过获得后向传播特征 $\hat{E}_b^t$:

$$\hat{E}_b^t = \mathcal{P}_b(E^t, \mathcal{D}_b(\hat{E}_b^{t+1}, W_{t\to t+1}, \hat{F}_{t\to t+1} + \Delta F_{t\to t+1})), \quad (5)$$

其中 $\mathcal{D}_b$ 表示可变形卷积层的操作。权重掩码 $W_{t\to t+1}$, 其值通过 sigmoid 函数归一化, 可以应用于每个采样像素, 以衡量其有效性。上述操作是参考 [17, 58] 双向使用的, 不过前向传播特征 $\hat{E}_f^t$ 可以用同样的方式反向获得。最后, 我们使用一个可学习的 1×1 大小的卷积层来自适应地融合前向和后向传播特征, 而不是使用预先定义的规则来结合 [58] 提到的双向光流追踪的像素。

$$\hat{E}^t = \mathcal{I}(\hat{E}_f^t, \hat{E}_b^t), \quad (6)$$

其中 $\mathcal{I}$ 表示一个 1×1 大小的卷积层。

3.2. Temporal focal transformer

仅使用局部时间邻域像素提供的信息对视频补全来说是不够的。正如 [17] 中所讨论的, 局部邻域的损坏内容可能出现在非局部邻域中。因此, 非局部时空邻域中的信息可以被视为局部邻域中这些缺失区域的一个较好的参考。在这里, 我们将多个 temporal focal transformer 块堆叠起来, 有效地结合局部和非局部时间邻域的信息, 以进行内容生成。

假设 T_{nl} 是选定的非局部帧的数量。$\widehat{\mathrm{E}}_l \in \mathbb{R}^{T_l \times \frac{H}{4} \times \frac{W}{4} \times C}$ 是所有非局部邻域的编码特征。我们使用软分割操作 [34] 对级联的局部和非局部时间特征进行重叠块嵌入。

$$Z^0 = \mathrm{SS}([\widehat{\mathrm{E}}_l, \mathrm{E}_{nl}]) \in \mathbb{R}^{(T_l+T_{nl}) \times M \times N \times C_e}, \quad (7)$$

其中 SS 表示软分割的操作, Z^0 是包含局部和非局部时间信息的嵌入 token, $M \times N$ 是嵌入空间维度,C_e 是特征维度。

我们使用 focal transformer [60] 而不是最近的工作中经常使用的 vision transformer [15] 从局部和非局部邻域中搜索以填补缺失的内容。原因列举如下:(1) 与执行细粒度的全局注意力机制相比, 通过基于窗口的注意力机制 [35, 60] 可以有效地降低计算和存储成本。(2) 对于缺失区域的每个 token, 由于图像的局部自相似性, 只在局部区域进行细粒度的自注意力是合理的, 而粗粒度的关注是全局性的。

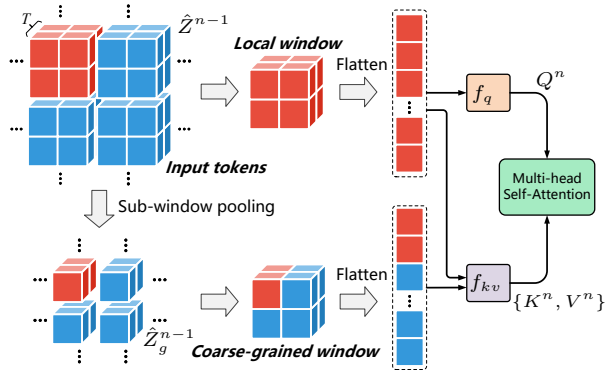

图 4. temporal focal self-attention 说明。这里我们以窗口大小为 $2 \times 2 \times 2$ 为例. 我们可以看到,keys 和 values $\{K^n, V^n\}$ 既包含细粒度的局部信息, 又包含粗粒度的全局信息。

由于原来的 focal transformer 无法处理序列数据, 我们提出了一个 temporal focal transformer, 基本上将 focal 窗口的大小从二维扩展到三维。具体来说, 我们首先将输入 token Z^{n-1}(其中 $n \in [1, N]$ 和 N 是 focal transformer 块的堆叠数), 分割成大小为 $s_t \times s_h \times s_w$ 的子窗口网格。被分割的 token $\hat{Z}^{n-1} \in \mathbb{R}^{(\frac{(T_l+T_{nl})}{s_t} \times \frac{M}{s_h} \times \frac{N}{s_w} \times C_e) \times (s_t \times s_h \times s_w)}$ 可以直接用于计算细粒度的局部注意力。为了在粗粒度上进行全局注意, 我们使用了一个线性嵌入层 f_p, 通过以下方式在空间上池化子窗口 $\hat{Z}_g^{n-1} = f_p(\hat{Z}^{n-1}) \in \mathbb{R}^{(\frac{(T_l+T_{nl})}{s_t} \times \frac{M}{s_h} \times \frac{N}{s_w} \times C_e) \times s_t}$。然后, 我们通过两个线性投影层 f_q、f_{kv} 来计算 query, key 和 value:

$$Q^n = f_q(\hat{Z}^{n-1}), \quad \{K_l^n, K_g^n, V_l^n, V_g^n\} = f_{kv}(\{\hat{Z}^{n-1}, \hat{Z}_g^{n-1}\}). \quad (8)$$

为了使用局部-全局交互计算注意力, 对于第 i 个子窗口 $Q_i^n \in \mathbb{R}^{s_t \times s_h \times s_w \times C_e}$ 内的 queries, 我们从第 i 个局部窗口 $K_{l,i}^n \in \mathbb{R}^{s_t \times s_h \times s_w \times C_e}$ 和第 i 个展开的粗粒度窗口 $K_{g,i}^n \in \mathbb{R}^{s_t \times s_h \times s_w \times C_e}$ 获取 keys。这种操作可以并行处理。我们将相应的 keys 和 values 分别用 $K^n = \{K_l^n, K_g^n\}$ 和 $V^n = \{V_l^n, V_g^n\}$ 连接起来, 然后计算出 Q_i^l 的 focal self-attention:

$$\mathrm{Attention}\,(Q^n, K^n, V^n) = \mathrm{Softmax}\left(\frac{Q^n (K^n)^T}{\sqrt{C_e}}\right) V^n. \quad (9)$$

注意,attention 函数也可以以多头的方式进行, 例子显示在图4。

最后, 在第 n 个 focal attention 中的整个过程

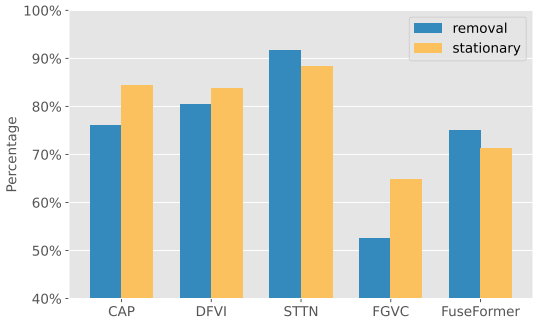

图 5. 用户研究结果. 纵轴表示与其他方法相比, 更喜欢我们的方法的志愿者百分比。

被表述为:

$$Z'^{n} = \mathrm{MFSA}(\mathrm{LN}_1(Z^{n-1})) + Z^{n-1}, \tag{10}$$

$$Z^{n} = \mathrm{F3N}(\mathrm{LN}_2(Z'^{n})) + Z'^{n}, \tag{11}$$

其中,MFSA 和 LN 分别表示多头 focal self-attention 和层归一化 [1]。我们使用 F3N [34] 来建立嵌入块间的关系。

3.3. 训练目标

我们采用三个损失函数来优化我们的模型。第一个是重建损失, 通过 L1 距离测量合成视频 $\widehat{\mathrm{Y}}$ 和原始视频 Y 之间的像素级差异:

$$\mathcal{L}_{rec} = \|\widehat{\mathrm{Y}} - \mathrm{Y}\|_1. \tag{12}$$

第二种是对抗性损失, 它已被证明在生成高质量和真实的内容方面很有效。我们采用了一个基于 T-PatchGAN [7] 的判别器, 使模型同时关注所有跨时空邻域的全局和局部特征。这个判别器 D 的训练目标是:

$$\begin{aligned}\mathcal{L}_D = E_{x\sim P_{\mathrm{Y}}(x)}[\mathrm{ReLU}(1 - D(x))] + \\ E_{z\sim P_{\widehat{\mathrm{Y}}}(z)}[\mathrm{ReLU}(1 + D(z))],\end{aligned} \tag{13}$$

对于视频补全生成器, 对抗性损失为:

$$\mathcal{L}_{adv} = -E_{z\sim P_{\widehat{\mathrm{Y}}}(z)}[D(z)], \tag{14}$$

第三种损失是光流一致性损失, 如公式 Eq. (2)所示。训练细节可在补充材料中找到。

4. 实验

4.1. 设定

数据集. 为了验证所提方法的有效性, 我们在两个热门的视频对象分割数据集上进行了评估, 即 YouTube-VOS [57] 和 DAVIS [45]。YouTube-VOS 具有不同的场景, 包括 3471、474 和 508 个视频片段, 分别用于训练、验证和测试。我们遵循原始的分割模式, 在 YouTube-VOS 的测试集上贴出了实验指标的报告。DAVIS 由 60 个用于训练的视频片段,90 个用于测试的视频片段组成。按照 FuseFormer [34], 测试集的 50 个视频片段被用来计算指标。我们在 YouTube-VOS 数据集上训练我们的模型, 并在 YouTube-VOS 和 DAVIS 数据集上评估它。至于掩码, 在训练过程中, 我们生成固定的和与物体接近的掩码, 以模拟视频补全和物体去除的应用, 具体如下 [8, 23, 28, 34, 64] 在评估中, 固定的掩码被用来计算客观指标, 由于缺乏参照物, 采用与物体接近的掩码进行定性比较。

度量指标. 我们选择 PSNR、SSIM [54]、VFID [52] 和光流的 warping 误差 E_{warp} [25] 来评估最近的视频补全方法的性能。具体来说,PSNR 和 SSIM 是经常使用的指标, 用于面向失真的图像和视频评估。VFID 测量两个输入视频之间感知上的相似性, 并在最近的视频补全工作 [34, 64] 中得到了采用。采用光流的 warping 误差 E_{warp} 来衡量时间上的一致性。

4.2. 对比

量化结果. 我们贴出了 YouTube-VOS [57] 和 DAVIS [45] 在固定掩码下的定量结果, 并将我们的方法与以前的视频补全方法进行比较, 包括 VINet [23], DFVI [58], LGTSM [8], CAP [28], STTN [64], FGVC [17] 和 Fuseformer [34]。正如在表1中所示, 我们的方法在所有四个定量指标上都大大超过了以前的 SOTA 算法。优异的结果表明, 我们的方法可以生成失真度更低 (PSNR 和 SSIM)、内容上更有视觉可信性 (VFID) 以及更好的空间和时间一致性 (E_{warp}) 的视频, 这验证了我们提出的方法的优越性。

定性结果. 我们选择了三种有代表性的方法, 包括 CAP [28]、FGVC [17] 和 Fuseformer [34], 来进行视觉比较。图6显示了视频补全和物体移除的结果。那些方法很难合理地恢复被遮挡区域的细节的同时, 我们提出的方法可以产生较真实的纹理和结构信息。

表 1. 在 YouTube-VOS [57] 和 DAVIS [45] 数据集上与 SOTA 视频补全模型进行定量比较。↑ 表示越高越好。↓ 表示越低越好。$E_{warp}{}^*$ 指 $E_{warp} \times 10^{-2}$。每个方法都是按照 FuseFormer [34] 中的程序进行评估的。VINet、DFVI 和 FGVC 都不是端到端的训练方法, 因此它们的 FLOPs 是不可预测的。

	Accuracy								Efficiency	
	YouTube-VOS				DAVIS				FLOPs	Runtime
Models	PSNR ↑	SSIM ↑	VFID ↓	$E_{warp}{}^*$ ↓	PSNR ↑	SSIM ↑	VFID ↓	$E_{warp}{}^*$ ↓		(s/frame)
VINet [23]	29.20	0.9434	0.072	0.1490	28.96	0.9411	0.199	0.1785	-	-
DFVI [58]	29.16	0.9429	0.066	0.1509	28.81	0.9404	0.187	0.1608	-	2.56
LGTSM [8]	29.74	0.9504	0.070	0.1859	28.57	0.9409	0.170	0.1640	1008G	0.23
CAP [28]	31.58	0.9607	0.071	0.1470	30.28	0.9521	0.182	0.1533	861G	0.40
FGVC [17]	29.67	0.9403	0.064	0.1022	30.80	0.9497	0.165	0.1586	-	2.44
STTN [64]	32.34	0.9655	0.053	0.0907	30.67	0.9560	0.149	0.1449	1032G	0.12
FuseFormer [34]	33.29	0.9681	0.053	0.0900	32.54	0.9700	0.138	0.1362	752G	0.20
E^{2}FGVI (Ours)	33.71	0.9700	0.046	0.0864	33.01	0.9721	0.116	0.1315	682G	0.16

图 6. 和 CAP [28], FGVC [17], FuseFormer [34] 的定性结果比较。

这表明了所提方法的有效性。为了进一步的综合比较, 我们对物体移除和视频补全的应用都进行了用户研究。我们选择了五种方法, 包括两种基于光流的方法 (i.e., DFVI [58] and FGVC [17]), 以及三种基于注意力的方法 (i.e., CAP [28], STTN [64], and Fuseformer [34])。我们邀请 20 人参加用户研究。每个志愿者都会看到随机抽样的 40 个视频三元组, 并被要求选择一个视觉效果更好的补全视频。每个三元组由一个原始视频、一个使用我们的方法补全的视频、一个使用随机方法补全的视频组成。用户研

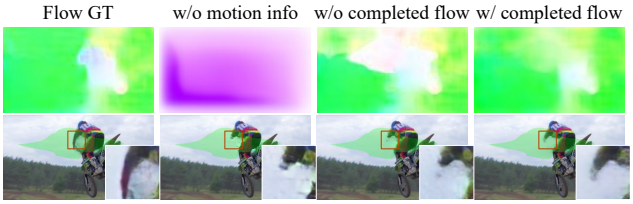

图 7. 光流补全模块的消融研究。第一行显示了不同情况下光流补全模块产生的结果。第二行可视化了相应的补全帧。

表 2. 光流补全模块的消融实验.

Case	PSNR	SSIM
w/o motion information	32.08	0.9673
w/o completed flow	32.23	0.9682
w/ completed flow	32.35	0.9688
Flow GT	32.54	0.9698

究结果显示在图5。我们可以看到, 与几乎所有方法的结果相比, 志愿者显然更喜欢我们的结果。尽管在与 FGVC 的比较中不存在这种明显的偏好, 但我们提出的方法仍然获得了大多数的投票。这表明, 我们提出的方法可以产生比其他方法更好的视觉效果。

效率比较. 我们使用 FLOPs 和推理时间来衡量每种方法的效率。FLOPs 计算时 temporal size 为 8, 运行时间是在单个 Titan Xp GPU 上使用 DAVIS 数据集测量的。

比较结果显示在表1。所提出的方法与基于 transformer 的方法运行时间相当, 但比基于光流的方法快了近 15 倍。此外, 与所有其他方法相比, 它拥有最低的 FLOPs。这表明, 我们提出的方法在视频补全方面具有很高的效率。

4.3. 消融实验

我们在光流补全、特征传播和注意力机制方面进行了三项消融实验, 以验证我们框架所提出的模块的有效性。所有的消融研究都是在 DAVIS 数据集上进行的。

对光流补全模块的实验.

首先, 我们调查了运动信息对视频补全的重要性。通过只删除光流一致性损失 $\mathcal{L}_{flow}$, 我们的光流补全模块不再提供关于对象运动的信息 (见图7), 导致性能大幅下降, 如表2所示。其次, 我们研究了通过固定光流补全模块中的预训练权重来研究补全光流的必要性。有了关于光流的初步知识, 光流补全模块将被遮挡的区域视为遮挡因素, 并为可见区域提供初始光流估计 (见图7)。与没有运动信息的模型相比, 性能有明显的提高。然而, 这种模型忽略了损失区域的运动信息。在我们通过训练光流补全模块以使光流一致性损失最小化来补全光流后, 我们获得了比以前更大的 PSNR 和 SSIM 值。如图7所示, 有完整光流的模型可以恢复更多关于人类手臂的更真实内容。此外, 在表2图7中, 我们还展示了我们的方法的潜在上界, 该方法估计了未损坏的帧之间的光流。

表 3. 对特征传播模块的研究.'Flow' 表示在 Eq. (4)中基于光流的 warping 函数 $\mathcal{W}$。'DCN' 表示调制的可变卷积 [65]。

	(a)	(b)	(c)	(d)
Flow	✗	✓	✗	✓
DCN	✗	✗	✓	✓
PSNR	31.73/0.9653	32.15/.9677	32.17/0.9676	32.35/.9688

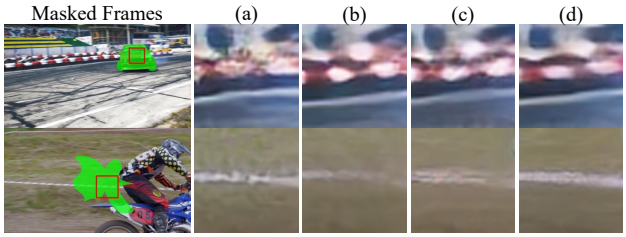

图 8. 特征传播模块的消融实验的定性结果. 最后四列对应表3四个例子。

对特征传播模块的实验。在我们从模型中移除特征传播模块后 (例子 (a) 在表3), 定量指标的数值急剧下降。从图8 (a) 中, 我们可以看到, 这个模型产生的结果存在严重的伪影和不连续的内容。在这个模型加入基于光流的 warping 和传播后 (见 Eq. (3))(例子 (b) 在表3), 由于我们可以在光流的帮助下将相邻帧的有效像素带到不可见的区域, 生成的内容变得更加有效 (见图8 (b)),PSNR 值增加了很大幅度 (0.42dB)。但是, 基于光流的 warping 和传播很难恢复不能被光流追踪到的内容 (图8 (b) 中的白线)。此外, 对于仅涉及基于可变卷积的 warping 的特征传播模块 (例子 (c) 在 表3), 在更多可学习的偏移的帮助下, 可以更清楚地恢复结构细节, 但由于与基于光流的 warping 相比, 缺乏从相邻帧 warping 的有效信息, 因此涉及更多伪影。通过将可变卷积与流引导

表 4. 对各种注意机制的消融研究。FuseFormer [34] 是目前使用 vanilla global attention 的 SOTA 方法。

Case	PSNR	SSIM	FLOPs
FuseFormer	31.74	0.9662	752G
Local attention	31.57	0.9648	497G
Focal attention	31.73	0.9653	560G

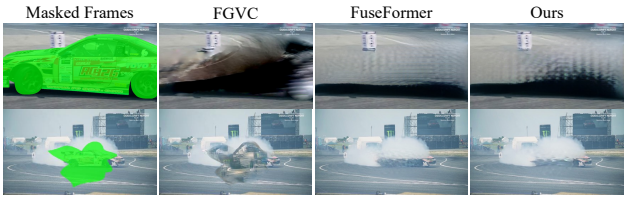

图 9. 两个失败案例（汽车漂移）. 目前的视频补全方法不能处理大运动或大量的物体细节缺失, 可能产生严重的伪影。

相结合例子在 表3）,PSNR 和 SSIM 值可以进一步提高。

在图8 (d) 中, 这个模型在所有变体中取得了视觉上的最佳效果, 同时保留了较潜力的结构细节。这证明了特征传播模块的有效性。

对注意力机制的研究实验. 我们删除了光流补全和特征传播模块, 纯粹比较不同的注意机制的差异, 包括 vanilla global attention (FuseFormer [34]), local window attention, and focal attention。如表4所示,vanilla global attention 实现了最好的量化性能, 但也需要过大的计算量。Local attention 引入了局部窗口, 就像 Video Swin Transformer [36] 那样。虽然 FLOPs 减少了 34%, 但注意力的计算被限制在局部窗口, 导致性能不佳。Focal attention 显示了性能和计算之间的良好权衡。它的 PSNR 和 SSIM 值与 FuseFormer 相当, 与 Local attention 相比, 计算成本只增加了 12%。

4.4. 局限

图9 显示了两种失败情况。当遇到大的运动或跨帧的大量物体细节缺失时, 我们的方法与 FGVC [17] 和 FuseFormer [34] 一样, 在缺失区域产生了难以置信的内容和许多伪影。这表明这些情况对视频补全来说仍然具有挑战性。

5. 总结

我们提出了一个端到端的基于光流的可训练视频补全模型, 名为 E^{2}FGVI。我们精心设计的三个模块（即光流补全、特征传播和内容生成模块）相互协作, 解决了以往方法的许多瓶颈问题。实验结果表明, 我们的方法在两个基准数据集上取得了 SOTA 定量和定性性能, 并且在推理时间和计算复杂度方面具有很好的优势。我们希望它可以成为未来工作的一个强有力的基线。

致谢. 这项工作得到了中国国家重点研发计划的资助 (NO. 2018AAA0100400),NSFC (NO. 61922046), 中国教育部 S&T 创新项目和中国博士后科学基金 (NO.2021M701780) 的资助。我们也感谢用于本研究的 MindSpore、CANN 和 Ascend AI Processor 的支持。

参考文献

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020. 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017. 2

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In CVPR, 2021. 2, 4

[5] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. AAAI, 2021. 4

[6] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. CVPR, 2022. 4

[7] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. ICCV, 2019. 1, 2, 6, 13

[8] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting. BMVC, 2019. 1, 2, 6, 7

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In ICML, 2020. 3

[10] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. NeurIPS, 2021. 3

[11] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In ICCV, 2017. 2

[12] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. ICLR, 2021. 3

[13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In CVPR, 2021. 3

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. arXiv preprint arXiv:2107.00652, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021. 3, 5

[16] Mounira Ebdelli, Olivier Le Meur, and Christine Guillemot. Video inpainting with short-term windows: Application to object removal and error concealment. TIP, 2015. 1

[17] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In ECCV, 2020. 1, 2, 3, 4, 5, 6, 7, 9, 13, 14

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In ICCV, 2019. 2

[19] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. arXiv preprint arXiv:2202.09741, 2022. 3

[20] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624, 2021. 3

[21] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. In ECCV, 2020. 2

[22] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In CVPR, 2018. 2

[23] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In CVPR, 2019. 1, 2, 6, 7

[24] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In ECCV, 2018. 2

[25] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In ECCV, 2018. 2, 6, 13

[26] Dong Lao, Peihao Zhu, Peter Wonka, and Ganesh Sundaramoorthi. Flow-guided video inpainting with scene templates. In ICCV, 2021. 2

[27] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In CVPR, 2020. 2

[28] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In ICCV, 2019. 1, 2, 6, 7, 14

[29] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In ECCV, 2020. 2

[30] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17562–17571, June 2022. 1

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image

restoration using swin transformer. In ICCV Workshops, pages 1833–1844, 2021. 3

[32] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In ICCV, 2019. 2

[33] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. arXiv preprint arXiv:2104.06637, 2021. 2

[34] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In ICCV, 2021. 1, 2, 5, 6, 7, 9, 13, 14

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV, 2021. 3, 5

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. arXiv preprint arXiv:2106.13230, 2021. 3, 9

[37] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. TOG, 2020. 2

[38] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In ICML, 2013. 4

[39] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. Siam journal on imaging sciences, 2014. 1

[40] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In ICCV, 2019. 1

[41] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In CVPR, 2020. 2

[42] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In ICML, 2018. 3

[43] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In CVPR, 2016. 1

[44] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In NeurIPS, 2021. 3

[45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In CVPR, 2016. 2, 6, 7, 14, 17, 18

[46] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In CVPR, 2017. 13

[47] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In CVPR, 2020. 2

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In ICML, 2021. 3

[49] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In CVPR, 2016. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 3

[51] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In AAAI, 2019. 1, 2

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In NeurIPS, 2018. 2, 6

[53] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In CVPR Workshops, 2019. 2, 4

[54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. TIP, 2004. 2, 6

[55] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In CVPR, 2020. 2

[56] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In CVPR, 2021. 2

[57] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In ECCV, 2018. 6, 7, 14, 15, 16

[58] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In CVPR, 2019. 1, 2, 3, 4, 5, 6, 7

[59] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. IJCV, 2019. 2, 4

[60] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In NeurIPS, 2021. 3, 5

[61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In CVPR, 2018. 1

[62] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In ICCV, 2019. 1, 13

[63] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. ICCV, 2021. 3

[64] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In ECCV, 2020. 1, 2, 6, 7, 13

[65] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In CVPR, 2019. 2, 4, 8

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In ICLR, 2021. 3

[67] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In ICCV, 2017. 2

[68] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In CVPR, 2021. 2

附录

A. 架构和训练细节.

架构. 在我们的模型中, 编码器和解码器使用与 FuseFormer [34] 相同的架构。编码器和译码器的通道维度 C 被设定为 128。为了提高计算效率, 我们采用了一个轻量级模型 SPyNet [46] 作为我们的光流补全模块。为了利用原始 SPyNet 中学习到的光流先验, 我们使用预训练的权重来初始化这个模块。T-PatchGAN 的架构细节与以前的工作 [7,34,64] 相同。可变卷积的核大小 K 和组数 G 分别被设定为 3 和 16。focal transformer 块的数量 N 被设定为 8,token 的嵌入维度 C_e 被设定为 512。嵌入空间维度 $M \times N$ 为 20×36。分区子窗口的大小 $s_t \times s_h \times s_w$ 设置为 $(T_l + T_{nl}) \times 5 \times 9$。在内容生成模块结束时, 我们使用软合成运算符将嵌入 tokens 合成为特征, 这些特征与原始 tokens 具有相同的空间大小。

训练细节. 对于训练目标,$\mathcal{L}_{rec}$, $\mathcal{L}_{adv}$ 和 $\mathcal{L}_{flow}$ 的权重分别为 1, 10^{-2}, 和 1。考虑到 GPU 的内存限制, 我们将视频中的所有帧调整为 432×240, 用于训练、评估和测试。在训练过程中, 局部 (T_l) 和非局部帧 (T_{nl}) 的数量分别为 5 和 3。局部帧是连续的片段, 而非局部帧是从视频中随机抽出的, 用于训练。按照 [64] 和 FuseFormer [34], 在评估和测试过程中, 我们使用一个大小为 10 的滑动窗口来获取局部相邻帧, 并以 10 的采样率对非局部相邻帧进行均匀采样。我们使用 $\beta_1 = 0$ and $\beta_2 = 0.99$ 的 Adam 优化器。最终的模型被训练了 500K 次, 所有模块的初始学习率被设定为 0.0001, 并在 400K 次迭代时减少 10 倍。在我们的消融研究中, 我们对模型进行了 25 万次的迭代训练。我们使用 8 个 NVIDIA Tesla V100 GPU 进行训练, 批量大小设置为 8。为了方便同行复现, 我们的代码已开源 [3]。

[3] https://github.com/MCG-NKU/E2FGVI

B. 更多实验

B.1. 以离线的方式补全光流

为了验证在线光流补全的有效性, 我们使用 FGVC [25] 光流补全模块以离线方式准备补全的光流。然后我们用 FGVC 补全的光流重新训练一个模型。该模型的 PSNR 值略高于我们的端到端设置 (32.38 vs. 32.35 (dB)), 然而推理速度比我们的慢得多 (1.21 vs. 0.16 (s/frame))。

B.2. 深入了解光流引导的特征传播模块

为了进一步研究特征传播模块的有效性, 我们在进行内容生成之前, 在图10中可视化了 temporal size 为 5 的平均局部邻域特征。图10 中的四种情况对应于我们论文中 Tab. 3 中的四种变体。对于没有特征传播的模型 (图10(a)), 显然, 我们可以看到所有帧的损坏区域仍然存在于这些特征中, 进一步限制了内容生成的表现。对于只使用基于光流的 warping(图10(b)) 或基于可变卷积的 warping (图10(c)) 的模型, 损坏的区域被来自相邻帧的 warped 内容填充。由于有更多的采样特征点, 基于可变卷积的 warping 可以产生比基于光流的 warping 更平滑的内容。然而, 特别是对于最后两个时间特征 (图10中最后两列), 与基于光流的 warping 相比, 由没有光流引导的模型填充的区域有更明显的边界, 这意味着在没有运动信息的情况下传播的有效内容较少。通过采用带有光流引导的可变卷积, 最终的传播模块 (图10(d)) 在所有情况中以最合理、最自然的内容填充了缺失部分。这是一个很好的证明, 说明了可变偏移和补全的光流场之间的互利关系。

B.3. 内容生成能力的研究

为了纯粹评估我们的方法的内容生成能力, 我们首先预先填充了可以被流场追踪的像素 [17]。因此, 剩余的未填充像素很可能在其他视频帧中不可见。然后, 我们将预填充的视频分别送入一个图像补全模型 [62] 和我们的模型。我们的生成结果比 DAVIS 数据集上的图像补全模型的 PSNR 值大得多 (31.74 vs. 30.80 (dB))。

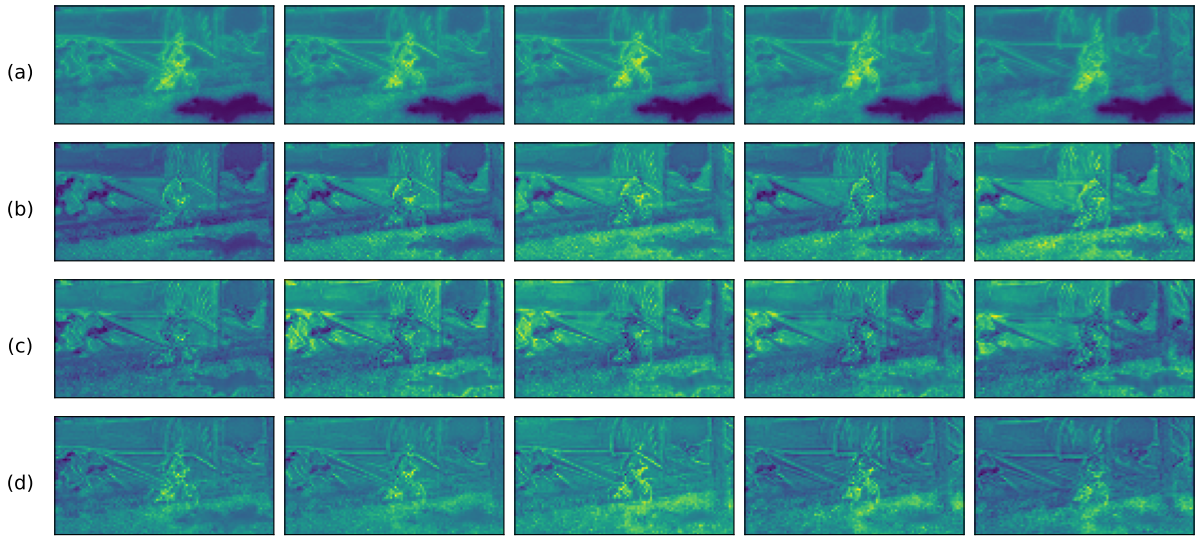

图 10. 在不同的实验设置下, 进入内容生成阶段前的帧平均特征的可视化. (a) 没有光流引导的特征传播。(b) 无变形卷积的光流导特征传播 (主论文中的公式 3)。(c) 没有光流引导的特征传播。(d) 在流场和可变形卷积的帮助下, 最终的光流引导特征传播模块。(**放大看效果更好**)

表 5. 参数的比较. FuseFormer* 表示原始 FuseFormer 的一个较大参数量版本。

	FuseFormer [34]	FuseFormer*	E^{2}FGVI
Params. (M)	36.6	41.6	41.8
PSNR/SSIM	31.74/0.9662	31.91/0.9669	32.35/0.9688

B.4. 参数的比较

我们在表格. 5中给出了这些参数尽管我们的方法比 SOTA 方法 (i.e., FuseFormer [34]) 多消耗了 ∼14% 的参数, 但与其他方法相比, 它实现了性能和计算复杂性之间的较好的权衡。为了进一步比较, 我们在 FuseFormer 中添加了残差块, 以实现与我们相似的参数量。我们的方法仍然比更大参数量的 FuseFormer 表现得更好。

B.5. 更多定性的结果

在本节中, 我们在两个基准数据集上提供了额外的可视结果, 包括 YouTube-VOS [57] 和 DAVIS [45], 以进一步显示所提出的 E^{2}FGVI 的优越性。现将 CAP [28]、FGVC [17] 和 FuseFormer [34] 的重建结果进行比较。如图. 11-14所示, 我们的 E^{2}FGVI 比其他方法能产生更有效的纹理和结构信息, 在损失的区域中也能产生更连贯的内容。**我们的** demo **在我们的项目页面中显示**。

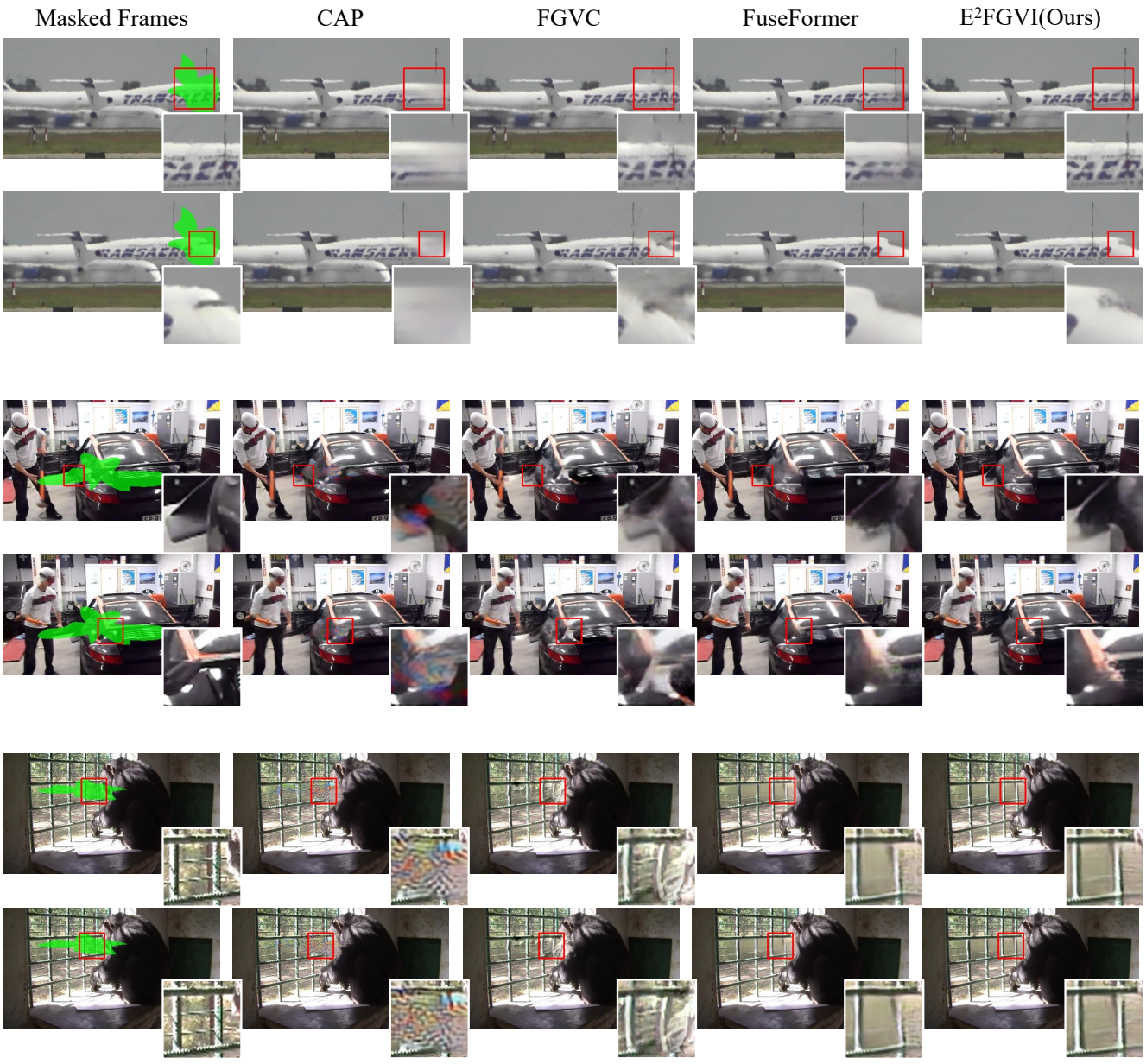

图 11. YouTube-VOS [57] 数据集上的定性视频补全结果.

图 12. YouTube-VOS [57] 数据集上的定性视频补全结果.

图 13. DAVIS [45] 数据集上的定性物体移除结果.

图 14. DAVIS [45] 数据集上的定性视频补全结果.