中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

博 士 学 位 论 文

任务属性驱动的图像和视频生成

Task-driven image and video generation

论文作者 李震 指导教师 邵秀丽 程明明教授
申请学位 工学博士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 刘青山教授 评 阅 人 匿名评阅

南开大学研究生院

二〇二四年五月

摘要

在数字化时代，图像和视频生成技术在计算机视觉领域占据核心地位，其应用范围广泛，从娱乐到安全，从艺术创作到科学研究。随着互联网技术的发展和智能设备的普及，对高质量视觉内容的需求日益增长。近年来，基于数据驱动的自动生成技术，特别是生成对抗网络和扩散模型，为图像和视频生成领域带来了革命性的进步。然而，如何根据特定应用场景和任务需求精准控制生成内容的属性和特征，仍然是一个挑战。

本文旨在研究任务属性驱动的图像与视频生成技术，特别聚焦于视频补全、视频帧生成以及人像图像个性化生成任务。针对每个任务的特定问题，本文根据对应的任务属性设计了相应的研究方案，以提高生成内容的针对性和实用性。本文的主要贡献如下：

1）提出了一个端到端的时空联合一致性驱动的视频补全框架 E^2FGVI，该框架通过光流补全、特征传播和内容生成模块的定制化设计，优化了视频补全过程。与国际最先进的算法相比，本框架生成的视频片段展现了更好的时空一致性，显著提升了视频质量和处理速度。

2）在视频帧生成领域，本文提出了运动属性驱动的视频帧生成框架 AMT，通过采用全对多场变换和多场微调技术，显著提升了运动建模的鲁棒性和多样性。该框架能够精确且细致地建模全局运动属性，更好地处理大位移和运动边界区域的遮挡问题，使生成的视频帧视觉上更自然和连贯。

3）对于个性化图像生成任务，本文提出了身份属性驱动的人物图像个性化生成框架 PhotoMaker，通过编码任意数量的输入身份（Identity，ID）图像为堆叠的 ID嵌入，实现了高效的个性化文本到图像生成。为驱动 PhotoMaker训练，本文提出了面向 ID的数据构建流程，通过该流程可以收集包含大量人物 ID，每个 ID多张图的数据。相较于先前方法，PhotoMaker在提高生成图像的 ID保真度、速度改进和泛化能力方面都展现了其显著的优势，其为个性化内容生成开辟了新的可能性。

本文的研究不仅提供了满足特定应用需求的高质量视觉内容生成的新解决思路，也为相关研究领域带来了新的视角。随着技术的不断进步和研究的深入，

任务属性驱动的图像和视频生成技术预期将在未来展现出更广阔的潜力和价值。

关键词： 图像生成；视频生成；任务属性驱动；视频补全；视频帧生成；个性化定制；人像生成

Abstract

In the digital age, image and video generation technology occupies a core position in the field of computer vision, with a wide range of applications from entertainment to security, and from artistic creation to scientific research. With the development of internet technology and the proliferation of smart devices, the demand for high-quality visual content is growing. In recent years, data-driven automatic generation technologies, especially Generative Adversarial Networks and diffusion models, have brought revolutionary progress to the field of image and video generation. However, precisely controlling the attributes and characteristics of generated content according to specific application scenarios and task requirements remains a challenge.

This paper focuses on the study of task-driven image and video generation technology, with a special focus on video inpainting, video frame interpolation, and personalized portrait image generation tasks. For each task specific problems, this paper designs corresponding research schemes based on the task attributes to improve the pertinence and practicality of the generated content. The main contributions of this paper are as follows:

1. We propose an end-to-end spatio-temporal consistency-driven video inpainting framework, named E^2FGVI. This framework optimizes the video inpainting process through the customized design of optical flow completion, feature propagation, and content generation modules. Compared with the most advanced international algorithms, the video segments generated by this framework show better spatio-temporal consistency, significantly improving video quality and inference speed.

2. In the field of video frame interpolation, we propose the motion attribute-driven video frame interpolation framework, named AMT. AMT significantly enhances the robustness and diversity of motion modeling through the adoption of all-pairs correlation and multi-field refinement technologies. This framework can accurately and meticulously model global motions, better handling large displacements and occlusion issues in motion boundary areas, making the generated video frames visually more natural

and coherent.

3. For personalized image generation tasks, we propose the identity attribute-driven personalized human image generation framework, named PhotoMaker, which encodes any number of input ID images into stacked ID embeddings, achieving efficient personalized text-to-image generation. To drive the training of PhotoMaker, we propose an ID-oriented data construction process, through which data containing a large number of person IDs, with multiple images per ID, can be collected. Compared with previous methods, PhotoMaker shows significant advantages in improving the fidelity of generated images to ID, speed improvement, and generalization ability, opening up new possibilities for personalized content generation.

This research not only provides new solutions for the generation of high-quality visual content to meet specific application needs but also brings new perspectives to related research fields. With continuous technological progress and deeper research, task-driven image and video generation technology is expected to show broader potential and value in the future.

Key Words: Image Generation, Video Generation, Task-Driven, Video Inpainting, Video Frame Interpolation, Personalized Generation, Portrait Generation

目录

第一章 绪论

1.1 研究背景与意义

在数字化时代，图像和视频作为信息传递的主要媒介之一，其生成和处理技术在计算机视觉领域占据核心地位。该技术旨在通过计算机算法自动创建视觉内容，包括静态的图像和动态的视频。随着互联网技术的发展和智能设备的普及，人们对于图像和视频内容的需求日益增长，这不仅包括对高质量视觉内容的追求，还包括对内容创造和编辑的灵活性需求。图像和视频生成技术的应用范围极为广泛，涵盖了从娱乐到安全，从艺术创作到科学研究的多个领域。在娱乐产业中，高质量的图像和视频生成技术被用于电影和广告特效的制作、游戏场景的渲染以及虚拟现实内容的创建，为用户提供沉浸式的视觉体验。在社交媒体平台，个性化的图像和视频内容生成使得用户能够以更加丰富和创新的方式进行表达和交流。此外，图像和视频生成技术还在安全监控、自动驾驶模拟、医学影像分析等领域发挥着重要作用，通过生成高质量的模拟图像和视频来辅助决策和分析。

传统的图像和视频生成方法依赖于复杂的手工设计流程和专业的图像处理软件，这些方法不仅耗时耗力，而且在处理大规模数据和复杂场景时存在明显的局限性。随着人工智能技术，尤其是深度学习的快速发展，基于数据驱动的图像和视频自动生成技术成为可能。这些技术通过学习大量的图像和视频数据，掌握视觉内容的内在规律，从而能够自动生成具有高度真实感的图像和视频内容。

生成对抗网络 [1]（Generative Adversarial Network, GAN）由 Ian Goodfellow等人于2014年提出，开启了图像生成技术的新篇章。 GAN通过引入一个生成器（Generator）和一个判别器（Discriminator）的对抗训练机制，生成器负责生成尽可能真实的图像，而判别器则尝试区分真实图像与生成图像。这种对抗过程促进了生成器能力的不断提升，使得生成的图像质量大幅度提高。

GAN的提出引发了广泛的视觉生成研究热潮，衍生出了多种变体和改进模型，如条件 GAN [2]（cGAN）、循环 GAN [3]（CycleGAN）和渐进式增长的

GAN [4]（Progressive GANs）、StyleGAN [5–7]等，这些模型在图像及视频的风格转换、超分辨率、补全等多个领域展现了卓越的性能。然而，尽管 GAN在图像生成领域取得了巨大成功，它们仍存在一些挑战，如训练不稳定、模式坍塌问题以及生成结果的多样性不足等。

扩散模型 [8, 9]（Diffusion Models）的兴起为图像生成领域带来了新的变革。与 GAN直接从随机噪声生成图像的方式不同，扩散模型通过一系列渐进的步骤将噪声转化为图像，这一过程模仿了物理世界中的扩散过程。

扩散模型的核心思想是首先将数据添加噪声，通过一系列的反向步骤逐渐去除噪声，最终恢复出清晰的图像。扩散模型最初由 Sohl-Dickstein等人 [10]于2015年提出，但直到近几年才因其在图像生成质量上的显著提升 [11–13]而受到广泛关注。扩散模型在生成高质量图像方面展现出了巨大潜力，特别是在细节表现和图像多样性方面优于传统的 GAN模型。此外，扩散模型在文本到图像和视频生成、图像和视频编辑和超分辨率等任务中也表现出色，成为了图像生成领域的研究热点。

然而，尽管现有的图像和视频生成技术在视觉质量和生成效率上取得了显著进步，但如何根据特定的应用场景和任务需求，精准控制生成内容的属性和特征，仍然是一个挑战。不同的应用场景往往对图像和视频内容有着不同的要求，例如：在补全一段视频片段时，生成的片段需要与原始视频在待补全位置、姿态、背景等方面保持一致，同时也要确保补全后的视频在时间上的连续性。在人物图像生成中，则可能需要根据人物的身份特征来调整图像的细节。这些复杂的任务属性对图像和视频生成技术提出了更高的要求，也推动了任务属性驱动的图像与视频生成研究的发展。

在这样的背景下，如何将生成过程与特定的应用需求紧密结合，成为了研究和应用的新焦点。**任务属性驱动的图像与视频生成**，正是在这样的背景下应运而生的研究方向。任务属性驱动的图像视频生成技术通过引入明确的任务属性（如时空一致性、运动属性、身份属性等），来指导生成模型产生符合特定需求的内容。这种方法不仅能够提高生成内容的针对性和实用性，还能够在一定程度上解释和控制生成过程，使得生成的图像和视频更加符合实际应用的需求。它通过深入理解和利用任务属性，如场景布局、动作类型、特定语义信息等，来指导生成模型的学习和优化过程，从而实现更加精准和高效的定制化内容生成。这一方向的研究不仅能够推动图像和视频生成技术的进一步发展，还

能够为下游应用提供更加精准和高效的解决方案，实现图像和视频内容生成的实用化和个性化。

在任务属性驱动的视觉生成中，时空一致性、运动建模和身份一致性至关重要，因为它们共同决定了生成内容的真实性、连贯性和个性化程度，直接影响用户体验和内容的应用价值。这些属性确保视觉内容不仅在视觉上吸引人，而且在逻辑上合理，为用户提供沉浸式和连贯的观感体验。然而，确保时空一致性、良好的运动建模和优异的身份一致性在任务属性驱动的视觉生成中是极具挑战性的。具体来说，这些挑战有：

- 时空一致性方面的挑战：首先，时空一致性要求算法对时空关系可以高度的理解。在视频生成或图像序列生成中，算法需要根据时空上下文预测当前遮盖区域或者接下来的场景变化，这不仅包括物体的运动，还包括光照、阴影以及环境的变化。这种预测需要同时保持物理世界的逻辑和规律，而这对于算法来说是一个复杂的任务。其次，保持空间一致性要求算法能够理解和模拟物理世界的规则。例如，在视频中生成一个人去除后的背景时，算法需要确保时空上下文的一致性。此外，生成的背景能和时空上下文的背景自然连贯的融合，这需要算法具有高度的时空理解能力。最后，时空一致性的保持还需要大量的计算资源。为了生成高质量的视觉内容，算法需要处理和分析大量的数据，进行复杂的计算。这不仅增加了计算成本，也提高了对算法效率的要求。
- 运动建模方面的挑战：首先，运动的多样性和复杂性使得建模变得困难。现实世界中，不同物体的运动规律各不相同，在生成一段逼真且符合逻辑的视频时相机的位置和角度也应该有相应的运动。特别是在高速运动的生成中，算法需要能精确建模大范围小物体的快速运动，才能生成自然、真实的视频。其次，运动建模还需要考虑环境因素的影响。在现实世界中，运动中物体间的遮挡也增添了运动建模的难度。算法需要能够模拟这些外部因素对运动的影响，这不仅增加了建模的复杂度，也要求算法具有更高的适应性和灵活性。最后，运动建模还面临数据的挑战。为了训练算法准确模拟运动，需要大量的运动数据。然而，高质量的运动数据往往难以获取，特别是对于一些特殊的运动类型。此外，即使有足够的数据，数据的处理和分析也是一个挑战，需要算法能够从复杂的数据中学习到运动的本质规律。

- 身份一致性方面的挑战：首先，身份特征的复杂性使得保持一致性变得困难。每个人或物体都有其独特的特征，如面部特征、服装纹理等。在视觉生成过程中，算法需要准确捕捉这些特征，并在不同的场景和条件下保持这些特征的一致性。这不仅要求算法具有高度的识别能力，还要求算法能够在变化的环境中保持特征的稳定性。其次，身份一致性的保持还面临视角和光照变化的挑战。在不同的视角和光照条件下，同一物体或人物的外观可能会发生显著变化，这给保持身份一致性带来了难度。算法需要能够理解这些变化的本质，从而在生成的视觉内容中准确地保留身份特征。最后，保持身份一致性还需要算法具有高度的泛化能力。在实际应用中，算法可能会遇到前所未见的身份特征或变化情况。在这种情况下，算法仍然需要能够准确地识别和保持身份特征，这要求算法不仅要在特定的数据集上表现良好，还要能够适应新的情况和挑战。

面临时空一致性、运动建模和身份一致性等挑战，任务属性驱动的视觉生成技术的研究与开发显得尤为关键。这些挑战不仅突显了在实现高度真实感和个性化视觉内容生成过程中所需面对的技术难题，也为未来的研究指明了方向。随着技术的进步和研究的不断深入，有理由相信，这些挑战将会逐步被克服，从而显著提升视觉内容生成的质量和应用范围。

正是基于对这些挑战的深入理解，任务属性驱动的图像视频生成技术不仅为满足特定应用需求的高质量视觉内容自动生成提供了新的解决思路，同时也为计算机视觉和人工智能领域的研究带来了新的视角和挑战。研究者通过深入探索和有效利用任务属性，能够设计出更加智能和灵活的生成模型。这些模型根据不同应用场景的具体需求，能够自动调整生成策略，生成更加符合期望的图像和视频内容。此外，任务属性驱动的方法还促进了生成模型的理解能力和智能化水平的提升，推动了生成模型结构、学习策略等方面的技术创新，展现了在克服这些挑战过程中的积极进展和潜在价值。

总之，任务属性驱动的图像视频生成技术不仅能够提高生成内容的针对性和实用性，还促进了生成模型的智能化和技术的创新，为满足多样化应用需求提供了强有力的支持。随着研究的深入和技术的进步，这一领域有望在未来展现出更大的潜力和价值。

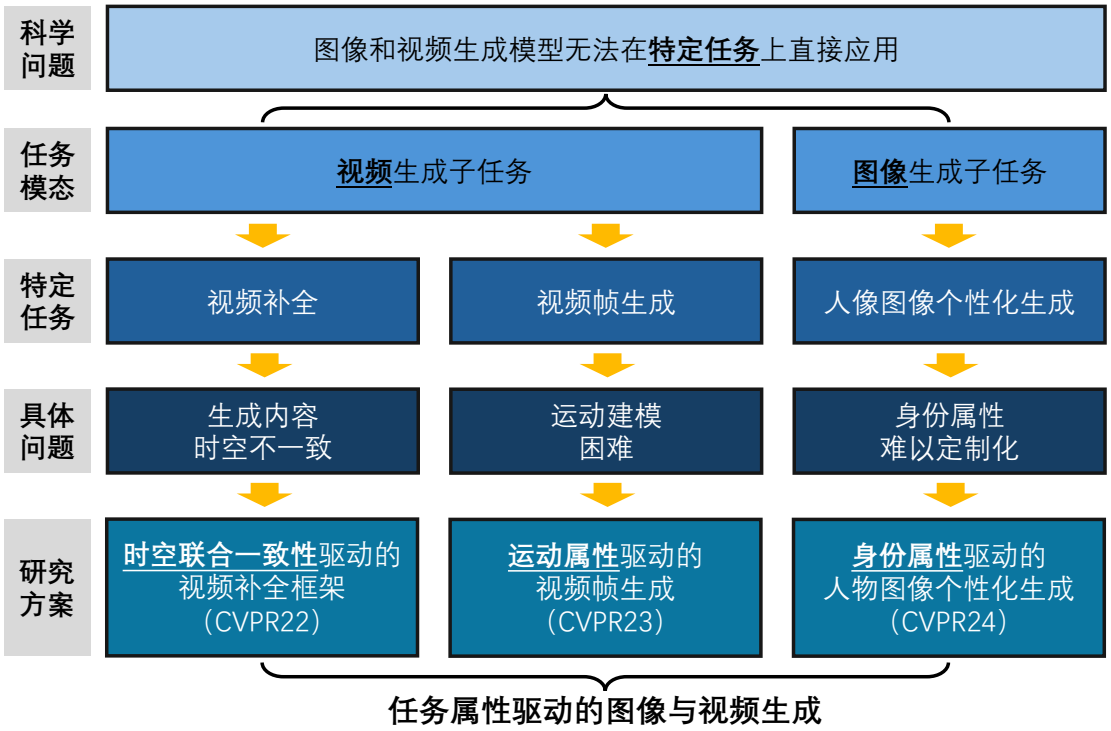

图 1.1: **任务属性驱动的图像与视频生成**的研究架构。

1.2 研究目标和主要贡献

本文旨在研究任务属性驱动的图像与视频生成技术，主要研究内容架构如图 1.1所示。如上一节所述图像和视频生成在面临特定任务时无法直接地进行应用。本文聚焦于三个图像和视频生成的子任务，分别是视频补全、视频帧生成以及人像图像个性化生成任务。每个任务都有具体的问题需要被解决，而这些问题都是任务本身所特有的，如视频补全生成中，生成内容容易出现时空不一致；在视频帧生成时会面临运动建模困难的问题；在进行个性化文生图生成时，会出现使用相应的身份属性进行个性化定制的问题。本文针对每个任务所面临的特定问题，根据对应的任务属性设计了相应的研究方案。这些研究方案又可以反哺图像和视频生成的发展。具体的研究方案如下：

1）视频补全是一项在视频编辑和后期制作中极为重要的技术，它要求生成的视频片段在视觉上不仅与原视频在内容上保持一致，而且在时间上也要连贯无缝。时空联合一致性驱动的视频补全框架正是基于这一需求设计的，它通过引入时空一致性作为任务属性，指导生成模型学习如何填补视频中的缺失部分，确保生成的视频片段在视觉和时间上都与原视频无缝对接。在最近的视频补全

方法中，常估计帧间传播像素的轨迹来捕捉时空一致性信息。然而，在这些方法中，估计轨迹的行为然后进行相应像素传播的行为往往是分阶段进行的。因此，这些方法不仅效率较低而且严重依赖早期阶段的中间结果。本文提出了一个端到端的时空联合一致性驱动的视频补全框架，称之为 E^2FGVI。该框架包含了三个定制化设计的可训练的模块，分别为光流补全、特征传播和内容生成模块。这三个模块与之前的基于光流的视频补全方法的三个阶段相对应，但不同点在于它们可以一同被优化。因此，本文提出的框架在处理视频补全任务时相对之前的方法会变得更加高效。实验结果表明，本文提出的方法在定性定量上都优于目前最先进方法的结果，并显示出较好的效率。

2）在视频生成和动画制作等领域，如何根据给定的运动轨迹或动作特征生成连贯的视频帧是一个关键问题。运动属性驱动的视频帧生成技术通过将运动属性作为任务指导，使得生成模型能够根据预定义的运动轨迹或动作特征生成相应的视频帧，从而实现更加自然和真实的动作表现。然而，现有的视频帧生成算法在进行运动建模时有两大缺陷：第一，受限于网络的表达能力，它们在遇到大位移时，无法可靠地建模出真正的帧间运动；第二，受限于建模方式单一，它们的建模的出的运动集合只能受限于狭小空间内，难以处理运动边界周围的遮挡和细节。因此，本文提出了一种新的运动属性驱动的视频帧生成框架，称为全对多场变换（All-pairs Multi-field Transforms，AMT）。该框架有两个基本设计：第一，本框架为所有像素对构建双向相关的匹配代价（correlation volume），该匹配代价可以很好的建模全局的运动属性，进而提高了预测光流的保真度。本框架还使用预测的双边光流来检索相关性匹配代价以更新光流和中间帧的内容特征；第二，本框架用一对已经更新过的较粗糙的光流生成多组细粒度的光流场，来扩大运动建模过程中的解集空间以保证在运动边界区域预测光流的多样性。结合以上两种设计，该框架能够生成更加真实的面向帧生成任务的光流，并减少在插帧期间建模大位移和处理遮挡区域的困难。这些优势使该框架在各种基准测试中以高效率实现了最先进的性能。此外，该框架在准确性和效率方面也优于基于 Transformer的模型。

3）身份属性驱动的人物图像个性化生成在个性化内容生成、虚拟角色设计等应用场景中，如何根据用户的特定需求生成具有特定身份属性的人物图像成为了一个热门话题。身份属性驱动的人物图像个性化生成技术通过引入简单的身份属性（如通过用户上传的自拍照进行解析后的特征），指导生成模型

产生符合用户个性化需求的人物图像，为用户提供更加丰富和个性化的视觉体验。近年来，文本到图像生成的技术取得了显著的进步，能够根据给定的文本提示生成逼真的人物照片。然而，现有的个性化生成方法无法同时满足高效率、不错的身份（Identity，ID）保真度和灵活的文本可控性的要求。本文中引入了 PhotoMaker，一种高效的个性化文本到图像生成方法，主要将任意数量的输入 ID图像编码成一个堆叠的 ID嵌入，以保留 ID信息。这样的嵌入作为一个统一的 ID表示，不仅可以全面地封装同一输入 ID的特性，而且还可以容纳不同ID的特性，以便后续的整合。这为更有趣和更有价值的实际应用铺平了道路。此外，为了驱动本文中 PhotoMaker的训练，本文提出了一个面向 ID的数据构建流程来组装训练数据。在通过所提出的流程构建的数据集的哺育下，本文的PhotoMaker比基于测试时间微调的方法展示出更好的 ID保留能力，同时提供了显著的速度改进、高质量的生成结果、强大的泛化能力和广泛的应用范围。

1.3 本文组织结构

本文的组织结构为：第一章介绍本文的课题背景并简述研究目标和主要贡献；第二章介绍了相关工作：包括视频补全生成技术、视频帧生成技术、人物图像个性化生成技术。第三章详细介绍了本文提出的时空联合一致性驱动的视频补全框架，并在两个公开数据集、多个时空一致性评价指标上证明了该网络架构的在目标去除和静态掩码遮挡两个典型任务上的优秀能力；第四章详细介绍了本文提出的运动属性驱动的视频帧生成，在多个公开数据集上证明了所设计算法在建模大运动进而进行视频帧生成的能力；第五章详细介绍了本文提出的身份属性驱动的人物个性化图像生成，其中还包括以人物身份为导向的数据集的构建以及利用这一数据集进行训练后的模型可以扩展应用在多个定制化生成的场景中，并证明了其进行身份一致性保持和身份信息表征的能力；第六章对本文的研究进行总结，并展望了基于现有成果的未来研究方向。

第二章 相关工作

本章将按照绪论部分章节 1.2中视频补全、视频帧生成、人物个性化定制三个方面，分别介绍相关的研究工作。

2.1 视频补全生成技术

2.1.1 图像补全

图像修复是一种有效的编辑工具，使用户能够屏蔽和编辑图像中的特定区域 [14–17]。随着深度学习的重大进展，一些工作通过利用生成对抗网络 (GAN) [1, 18, 19]获得了显着成就。这些方法经常随机屏蔽图像中的任何区域，并进行优化以恢复被屏蔽区域 [20–22]。通过这样的优化，这些模型能够填充与图像上下文一致的内容区域。然而，这些方法无法从图像上下文中推断出新对象，无法合成新的内容。

文本到图像扩散模型极大地促进了最近的进展 [11, 23–25]。具体来说，SD-Inpainting [13]和 ControlNet-Inpainting [26]都建立在大规模预训练的文本到图像模型之上，即 Stable diffusion [13]。他们微调了一个预训练的 T2I 模型，用于将随机掩码修复为修复掩码，将图像标题作为文本提示。尽管得到了不错的补全效果，但这些模型往往存在文本错位，即无法合成与文本提示一致的对象。Smartbrush [27]和 Imagen Editor [28]提出通过使用成对的对象描述数据进行训练来解决这个问题。然而，这些模型往往假设缺失区域总是存在目标，失去了执行上下文感知图像修复的能力。此外，如果图像补全的算法直接应用于视频补全生成会很容易造成补全后的视频中补全的内容不够可信且视频缺乏时间一致性。

2.1.2 视频补全

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 [29–31]、基于光流的方法 [32, 33]和基于注意力机制的方法 [34–37]. 一些采用三维卷积和注意力的方法 [29, 34, 38, 39]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 [39,40]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 [32,33]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。Kang等人 [41]提出了一个框架，通过引入新设计的流完成模块和一个利用错误指导图的错误补偿网络，增强了基于流的视频修复方法，旨在抵消传统基于流的方法的弱点。 Zhang等人 [42,43]在一个基于 Transformer的框架中，使用光流引导来增强视频修复。与之前工作不同的地方是，作者没有在框架中进行像素级传播，而是设计了一个端到端的可训练框架，在特征空间对可信内容进行传播。此外，作者的方法还受益于最近一些使用 Transformer来提升补全效果的工作 [36,37,44]。

作者的工作还启发了多个有影响力的应用。 Weder等人 [45]将本技术运用到3D重建中，来去除 NeRF [46]表征中不需要的物体。 Yang等人提出了 Track-Anything [47]，将作者提出的视频补全框架与 Segment-Anything [48]相结合，实现了用户可交互的视频补全生成。 Zhou等人 [49]将像素传播过程和特征传播过程融合在了一起提出了 ProPainter。

扩散模型 [8]的出现也为视频补全带来了新的生机， Ceylan等人 [50]完成的“Pix2video”工作引入了一种使用图像扩散模型的无需训练的、文本引导的视频编辑方法，该方法编辑一个锚帧，然后将更改传播到后续帧。 Hoppe等人 [51]引入了随机掩码视频扩散（RaMViD）方法，该方法使用3D卷积和一种新的条件技术将图像扩散模型扩展到视频，从而实现视频预测、填充和上采样。Cherel等人 [52]提出了一种使用内部扩散过程进行视频修复的技术，重点是提高修复视频内容的质量和一致性。 Voleti等人 [53]入了一种多功能的视频扩散模型，通过利用掩码条件方法，该模型在视频预测、生成和插帧方面表现出色。Zhang等人 [54]将文本引导引入视频补全中提出了 AVID框架。

2.1.3 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 [55,56], 视频分割 [57,58]，视频目标检测 [59]，深度估计 [60,61]，视频超分辨率 [62,63]，插帧 [64,65]等等。具体来说，许多视频修复和增强算法 [62,63,66–68] 依靠光流对齐来补偿帧之间的信息。最近的工作 [63,65,69–71] 利用可变卷积 [72] 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。作者的工作也与这些工作有相同的优点。

2.1.4 视觉 Transformer

最近，Transformer [73] 在视觉邻域获得了很多关注。视觉 Transformer [74] 以及它的跟进工作 [75–79] 在图像和视频表示学习方面取得了令人印象深刻的表现 [80–83]，如图像生成 [84]，目标检测 [85, 86]，和许多其他应用 [87–90]。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 [77, 78, 91]，以减少其计算复杂性，同时提升模型在有限感受野下的特征提取能力。 Swin Transformer [77] 通过转移本地窗口计算自注意力，加强了局部关系。 Focal Transformer [78] 引入焦点式的自注意力机制，增强了全局与局部的交互。

2.2 视频帧生成技术

2.2.1 视频插帧

随着深度学习的发展，视频插帧领域衍生出了大量基于深度学习的方法。现有的视频插帧方法通常可以分为不基于光流的方法和基于光流的方法。

不考虑光流的方法在模型进行视频帧生成时不明确表示中间运动。 基于相位的方法 [92, 93]直接预测中间帧的相位分解，但只能处理有限范围内的运动。 基于内核的方法是这个类别中的主流方法，通常旨在通过学习适应性内核来卷积输入帧，从而估计中间帧 [94, 95]。多年来，这个领域提出了许多改进，包括使用可变形卷积 [96, 96]，将插值运动估计形式化为分类 [97]，混合深度特征 [98]，引入双帧对抗损失 [65]，执行通道注意力 [99] 和利用3D时空卷积 [100]。最近，Shi等人 [101] 引入了一个基于 Transformer的框架，借助注意力机制来模拟长距离依赖性。通过直接产生像素值，这些方法往往会产生模糊的结果和人为的效果，特别是在快速移动的场景中 [102]。

基于光流的方法由于其的鲁棒性，这类方法已经成为视频插帧任务的主流。一般来说，基于光流的方法采用两阶段流程：(1) 光流估计和 (2) 帧合成。他们首先估计输入帧之间的光流，然后使用图像扭曲来合成中间帧 [103]。作为代表性的工作，Jiang等人 的 SuperSlomo [64] 采用了跳过连接的 U-Net来估计双向光流，假设运动是线性的。也有人做出了二次 [104] 和三次 [105, 106] 轨迹假设来近似中间运动。最近的工作探索了各种技术来提高中间光流估计和插值精度，

包括通过 softmax splatting进行前向扭曲 [107, 108]、体素流 [109]、周期一致性损失 [110, 111]、任务导向的流动性蒸馏损失 [112]、Gram矩阵损失 [113]、隐式神经函数 [114]、遮挡蒙版 [115]、锚点对齐 [116]、特权蒸馏 [117]和金字塔重复流动性估计 [118]。考虑额外的信息，如上下文图 [119]，深度图 [120] 和来自事件相机的辅助视觉信息 [121–123]，也可以进一步提高插值精度。 Park等人 采用了对称的双向运动场估计，并通过非对称的双向运动场进一步提高了中间运动估计的精度 [124]。 Lu等人 [102] 利用了 Transformer架构 [73] 来模拟长期依赖性。 Jin等人 [125] 提出了一个新的双向运动估计器，采用金字塔结构。 Zhang等人 [126] 提出了一种新的特征提取策略，通过混合 CNN和 Transformer架构来结合运动和外观信息。在网络结构选择上，类 UNet的网络结构是一种常见的选择 [107, 115, 119, 120] 去生成最终的合成帧，并且目前流行的 Transformer框架也被引进来 [101, 102, 127]去实现更好的生成。最近的一些工作 [112, 113] 结合效率考虑，舍弃掉了单独的生成网络。然而，这些方法仍然存在着不能很好地建模大位移和处理遮挡等问题。此外，AnimeInterp [128]和 [129]提出了专门针对动漫的插值方法，动漫通常展示最小的纹理和夸张的运动。

LDMVFI [130]是首个采用条件潜在扩散模型方法来解决视频帧生成任务的方案。为了利用潜在扩散模型进行视频插帧，这项工作引入了一系列开创性的概念。具体来说，该工作提出了一个专门针对视频帧插值的自动编码网络，该网络集成了高效的自注意力模块，并采用了可变形的基于内核的帧合成技术，大大提高了帧生成的性能。

此外，最近的视频生成工作 [131–135]都将插帧部分视为重要的模块应用在模型中。这些工作大多都是分阶段的进行视频生成，即先生成低分辨率低帧率的视频或视频编码再经由对应的插帧模块得到高帧率的视频或视频编码。

2.2.2 任务驱动的光流

最初，基于光流的视频处理方法是单独去进行光流估计和图像处理。但是，这种分两步进行的方法忽略了真实光流与特定任务所需目标之间的差别，进而导致在特定任务上产生次优解。 ToFlow [62]提出了面向任务的光流，显著促进了视频处理技术 [136–140]的发展。通常，面向视频插帧任务的光流与真实的光流基本一致，但是在局部细节上存在多样性的差异（比如，遮挡的区域）。Super Slomo [64]引入了一种掩码来精确的处理遮挡问题，并且提供了一种标准的范式来生成中间帧。目前已有多篇工作 [112, 116, 117, 141] 采用了这种方法。

IFRNet [112] 和 RIFE [117] 提出了面向任务的光流蒸馏损失来为训练过程中的中间光流提供先验。与它们不同，作者从架构设计的角度来考虑面向任务光流的估计。作者引入了全对关联性来加强运动建模的能力，保证了光流在粗粒度尺度上的一致性。在最细尺度上，作者采用多场细化,确保针对特定任务的光流区域的多样性。

2.2.3 匹配代价

匹配代价被应用在许多视觉任务中，用来表示匹配代价 [142–144]。在深度学习领域，匹配代价被证明了在光流估计领域的有效性 [145–149]。在这些相关工作中，最有影响力的是 PWC-Net [145]和 RAFT [147]。在视频插帧任务中，目前的方法 [124, 125, 150, 151]尝试遵循 PWC-Net的整体结构来引入匹配代价。然而，这些方法不仅只在局部区域搜索成本量，而且还依赖于从参考特征中扭曲获得的不准确特征，导致使用匹配代价性能增益有限。相反，作者提出的基于 RAFT的 AMT算法，可以通过迭代更新具有全对相关性的流场来扩大搜索空间，并且仅在可见帧之间构建匹配代价此外，作者还采用许多超越 RAFT 的 新颖和 特定于任务的设计。具体细节见章节 4.2。

2.3 人物图像个性化生成技术

个性化任务旨在从示例图像中捕获并利用作为生成条件的概念，这些概念通过文本不易描述。在本节中，作者提供了这些个性化条件的概述，将它们进行分类，以便更清楚地理解它们的多样化应用和功能。

2.3.1 文本到图像的扩散模型

扩散模型 [8, 23]在文本条件下的图像生成 [12, 13, 25, 152]方面取得了显著的进步，近年来引起了广泛的关注。这些模型的显著性能可以归因于高质量的大规模文本-图像数据集 [153–155]，基础模型的持续升级 [156, 157]，条件编码器 [158–160]，以及可控性的提高 [26, 161–163]。由于这些进步，Podell等人 [164]开发了目前最强大的开源生成模型——SDXL。鉴于其在生成人像肖像方面的卓越能力，作者基于这个模型构建了本论文的 PhotoMaker。然而，作者的方法也可以扩展到其他文本到图像的合成模型。

2.3.2 主题驱动的生成

主题驱动的文本到图像生成，使用特定主题的有限集合的图像基于文本描述生成定制的图像，已经取得了显著的进步。以前的主题驱动方法如DreamBooth [165]，Textual Inversion [166]，ELITE [167]，E4T [168]和 ProFusion [169]在微调过程中微调一个特殊的提示标记来描述目标概念。考虑到这两种开创性的工作都需要大量的时间进行微调，一些研究试图通过减少需要微调的参数数量 [170–173]或通过大数据集的预训练 [168, 174]来加速个性化定制的过程。尽管有了这些进步，但它们仍然需要对每个新概念进行大量的预训练模型微调，使得过程耗时，并限制了其应用。最近，一些研究 [167, 175–180]试图使用单个图像进行个性化生成，只需要一个前向传递，大大加速了个性化过程。这些方法要么利用个性化数据集 [177, 181]进行训练，要么在语义空间中编码要定制的图像 [167, 175, 176, 178, 182, 183]。这些方法在不需要额外微调的情况下实现了主题驱动的文本到图像生成。这些方法通常涉及训练额外的模块，同时保持核心预训练的文本到图像模型冻结。另外还有一种典型的方法是 IP-Adapter [163]，它的目标是通过为文本特征和图像特征分离交叉注意力层来解耦交叉注意力机制。同期工作 Anydoor [180]通过设计维持纹理细节同时允许多样化的局部变化的细节特征来补充常用的主题特征。

2.3.3 保持人物身份 ID的图像生成

保持人物身份 ID的图像生成是主题驱动生成的一个特例，但它专注于具有强语义的面部属性，并在现实世界的场景中得到了广泛的应用。现有的工作主要可以根据它们对测试时微调的依赖性分为两类。低秩适应 [184]（LoRA）是一种流行的轻量级训练技术，它在训练定制数据集之前将最小数量的新权重插入到模型中。然而，LoRA需要对每个新角色进行单独的训练，限制了其灵活性。相比之下，最近的发展引入了优化自由的方法，绕过额外的微调或反演过程。 Face0 [185]在 CLIP空间中用投影的面部嵌入覆盖最后三个文本标记，并使用联合嵌入作为条件来指导扩散过程。 FaceStudio [186]提出了一个混合引导的身份保持图像合成框架，其中面部嵌入被集成到 CLIP视觉嵌入和 CLIP文本嵌入中，通过线性投影，然后将合并的引导嵌入融入到 UNet中，以交叉注意力。 IP-Adapter-FaceID [163]使用来自面部识别模型的面部 ID嵌入，而不是 CLIP图像嵌入来保持 ID一致性。 InstantID [187]通过将人脸的五个特征点通过

类似 ControlNet [26]的方式注入网络来完成身份保持。作者的方法侧重于基于利用个性化数据集和在语义空间中编码要定制的图像两种技术方法的人像生成。具体来说，它不仅依赖于构建 ID导向的个性化数据集，还依赖于在语义空间中获取表示人的 ID的嵌入。与以前的基于嵌入的方法不同，作者的 PhotoMaker从多个 ID图像中提取堆叠的 ID嵌入。在提供更好的 ID表示的同时，所提出的方法可以保持与以前基于嵌入的方法相同的高效率。

第三章 时空联合一致性驱动的视频补全框架

本章主要研究时空联合一致性驱动的视频补全框架。章节 3.1中介绍研究背景、动机及解决方案概要； 章节 3.2介绍本章构建的时空联合一致性驱动的视频补全框架； 章节 3.3给出评测结果和结果分析； 章节 3.4对本章进行小结。

3.1 引言

3.1.1 研究背景

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生成指定的“遮挡”或“损坏”的区域。它被广泛地应用于现实世界的应用如物体和水印移除（object removal） [188]、视频修复（video restoration） [34]和视频外扩（video outpainting） [38, 189]。

如章节 2.1.2所述，尽管图像补全（image inpainting）已经取得了很多不错的进展 [21, 22, 190]，但由于复杂的视频场景和时间一致性的要求，视频补全仍然充满挑战。如果直接使用图像补全算法对每一帧进行补全，往往会使视频在时间上产生不一致性，并产生严重的伪影。而用户往往期待可以得到高质量的视频补全结果，这就需要算法同时需要兼顾**空间结构和时间一致性**。最近，深度学习的蓬勃发展，促使研究人员探索出许多相对传统算法更有效的解决方案 [29, 30, 32–34, 36–39, 191]。

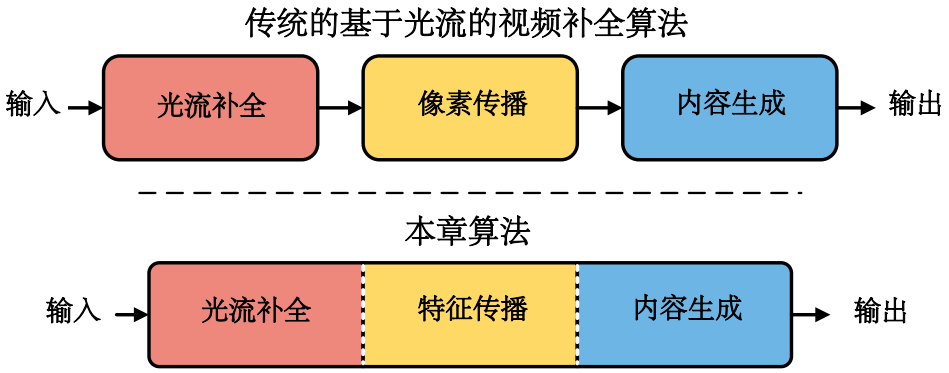

图 3.1: **传统的基于光流的方法 [32, 33]和本章框架**。 以前的基于光流的方法需要分别执行三个阶段，而本章框架的相应模块是以**端到端**的方式工作的。

在这些方法中，效果最好、最为典型的是基于光流的视频补全方法 [32,33]。这类方法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。与此同时，因为视频内容包含许多复杂的纹理结构信息，直接生成会很有难度。因此这类方法会优先去补全结构简单的光流，再根据补全的光流来进行像素传播。如图 3.1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*：由于在损坏的区域没有光流信息会直接影响后续的视频补全过程，这些方法首先通过现有的光流网络离线估计视频帧间的光流，然后再通过相应变化得到补全后的光流。（2） *像素传播*：通过在已补全的光流的引导下，通过可见区域间双向地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，剩余的缺失区域意味着在视频中一直存在遮挡，因此可以通过现有的预训练图像补全网络进行生成 [22, 190]。

即使这些方法可以获得不错的结果，但由于前两个阶段涉及许多手工操作（如：泊松混合、求解稀疏线性方程和索引每个像素的流动轨迹），整个基于光流的视频补全过程必须单独进行。这种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，这会大大影响最终的性能。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所依赖的上下文信息，最终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 [32]为例，补全一个大小为432×240，包含约70个帧的视频 [192]，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 3.2）。

3.1.2 研究动机与贡献

为了解决这些问题，本章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，名为 E^{2}FGVI。三个模块之间的密切协作，缓解了以前独立系统 [32, 33, 39, 40, 193]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，本章框架直接一步将其应用于待补全的视

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

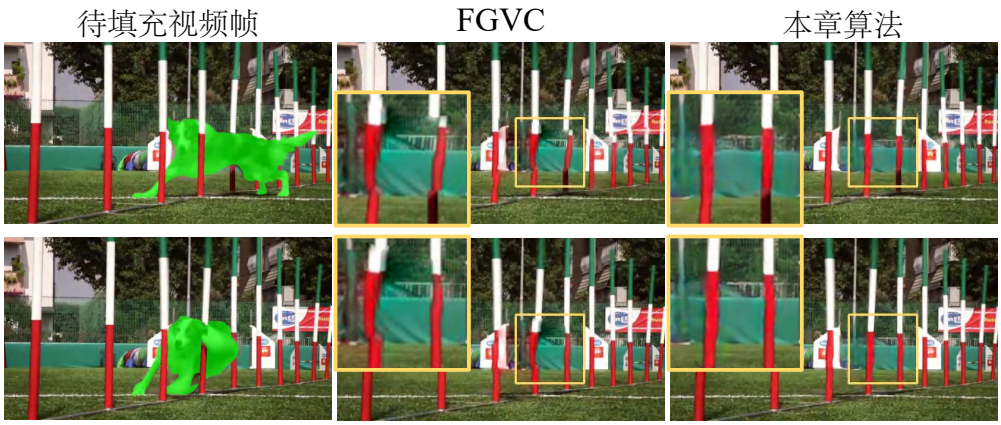

图 3.2: 本章框架与基于光流的目前最先进方法 FGVC [33]。与作者的方法相比，由于在内容生成过程中的错误积累和对时间信息的忽视，FGVC因此未能产生可靠的、具有时间一致性的结果。

频并直接输出补全后的光流，进而避免之前方法中多步复杂的操作（即先预测光流再补全光流）。对于特征传播模块，与像素级的传播不同，本章框架的光流引导传播过程是在可变形卷积的帮助下在特征空间进行的。因此可以直接在端到端地网络内部操作。此外传播模块通过更多可学习的采样偏移和特征级操作，其减轻了不准确光流对模型造成的影响。对于内容生成模块，本章框架提出了时空 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，局部和非局部的时空邻域都被考虑在内，从而能够产生更具有时间连贯性的补全结果。

实验结果表明本章提出的框架具有以下两个优势：

- 目前最优准确率：与目前最先进方法相比，该方法 E^{2}FGVI 在两个常用的面向失真度的指标（如，PSNR and SSIM [194]）、一个面向感知的指标（如 VFID [195]）和一个时间一致性衡量指标（如 E_{warp} [196]）方面取得了显著的改进。
- 高效: 作者的方法在 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，作者的方法显示出不错的推理速度。此外，在所有被比较的目前最先进方法中，作者的方法具有最低的计算复杂性（FLOPs）。

作者希望本章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，并为视频补全生成领域开拓一种新的技术路径。

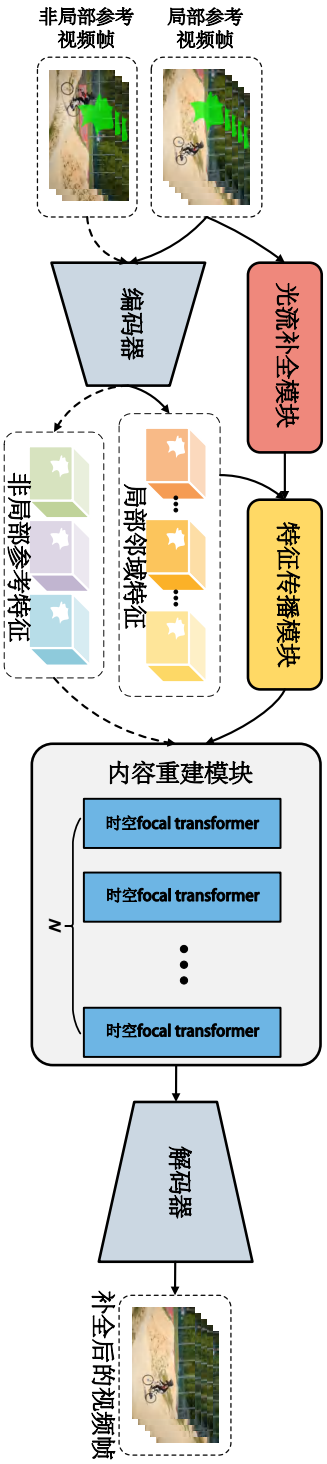

图 3.3: **本章算法 E^{2}FGVI的概览**。它包括：1）帧级内容编码器；2）光流补全模块；3）特征传播模块；4）由多个时空 Focal Transformer块组成的内容生成模块；5）帧级解码器。

3.2 方法

给出一个长度为 T 的包含待填充区域的视频序列 $\{X^t \in \mathbb{R}^{H\times W\times 3} \mid t = 1\ldots T\}$ 和相应的逐帧二进制掩码 $\{M^t \in \mathbb{R}^{H\times W\times 1} \mid t = 1\ldots T\}$，本章框架的目标是生成可靠的视频内容，使它在空间和时间维度上都与未遮掩的区域一致。在下文中，主要将讨论本章框架的主要组成部分。首先，该框架使用了一个上下文编码器，它将所有输入的包含待填充区域的视频帧编码为低分辨率的特征，以便在后续处理中提高计算效率；其次，作者通过一个光流补全模块提取并补全局部邻域之间的光流（见章节 3.2.1）；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（见章节 3.2.2）；第四，多层时空 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（见章节 3.2.3）。最后，一个解码器将生成补全后的视频特征放大，并将其重建为最终的视频序列 $\{\hat{Y}^t \in \mathbb{R}^{H\times W\times 3} \mid t = 1\ldots T\}$。

图 3.3 显示了本章节提出框架的整个流程（E^{2}FGVI）。值得注意的是，所有的模块都是可微的，它们一起构成了一个端到端的可训练架构。

在本节中，本章节将详细介绍框架中的详细操作。请注意，本框架中基于光流的模块只应用于从局部相邻帧中提取的特征，因为由于非局部帧中更容易出现过大的运动，因此这种运动建模会很困难，进而光流估计会退化甚至失败。此外，为了提高计算效率，本框架所有与光流相关的操作是在低分辨率空间进行的。

3.2.1 端到端的光流补全

在光流预测之前，本框架首先以1/4的分辨率对原始损坏的帧 X^t 进行下采样，这是为了与该框架编码器输出的低分辨率特征的空间分辨率相匹配。下采样的帧可以被表示为 $X^t_{\downarrow} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$。相邻帧 i 和 j 之间的光流是可以通过由预训练的光流估计网络计算得到。如果将光流估计网络表示为 $\mathcal{F}$，则有:

$$\hat{F}_{i\to j} = \mathcal{F}(X^i_{\downarrow}, X^j_{\downarrow}) \tag{3.1}$$

本框架使用来自轻量级光流估计网络的预训练权重来初始化光流补全模块，以利用其丰富的光流知识。

如大多数基于光流的视频补全方法 [32, 33]的做法一样，本框架通过公式 3.1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。由于视频中的缺失区域可以被视为光流估计的遮挡区域，因此会严重影响所估计光

流的质量，因此本框架需要在使用它们进行特征传播之前补全缺失区域的前向和后向光流。为了简单起见，作者使用 L1损失[2]来补全双向光流。

$$\mathcal{L}_{flow} = \sum_{t=1}^{T-1} \|\hat{F}_{t\to t+1} - F_{t\to t+1}\|_1 + \sum_{t=2}^{T} \|\hat{F}_{t\to t-1} - F_{t\to t-1}\|_1 \tag{3.2}$$

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，它们是从不包含缺失区域的原始视频中计算出来的。

本框架的光流补全模块与 DFVI [32]和 FGVC [33]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，作者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 [62]。（2） DFVI和 FGVC的光流补全效率较低（> 0.4秒/每光流）。这是因为它们需要先初始化光流，然后用多个阶段细化。而作者只用一个前馈传递来估计和补全光流，速度会快得多（< 0.01秒/每光流）。

3.2.2 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。以前向光流 $\hat{F}_{t\to t+1}$ 为例，它可以帮助算法捕捉到从第 t 帧到第 t+1帧的待填充区域的运动。一旦第 t 帧内容特征处的待填充区域的像素在第 t+1帧特征处的有效区域是已知的，在前向光流 $\hat{F}_{t\to t+1}$ 的帮助下，算法可以通过扭曲（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。经过扭曲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:

$$\hat{E}_b^t = \mathcal{P}_b(E^t, \mathcal{W}(\hat{E}_b^{t+1}, \hat{F}_{t\to t+1})) \tag{3.3}$$

其中 $\mathcal{W}(\cdot)$ 表示基于光流的空间扭曲操作，$\hat{E}_b^t$ 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 [197]激活的卷积层。

公式 3.3中的扭曲和合并操作近似于 DFVI和 FGVC中的整个传播过程，但本章算法在特征空间而不是图像空间中进行这些操作。传播特征 $\hat{E}_b^t$ 会将可信的内容信息逐渐传播到每个内容特征的待填充区域，这样的操作也有利于在光流引导的作用下，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光流估计质量的像素级传播不同，特征级传播利用了卷积层以在更大

[2]其他损失函数也可以用在公式 3.2中，但没有观察到它对最终的补全性能有明显改善。

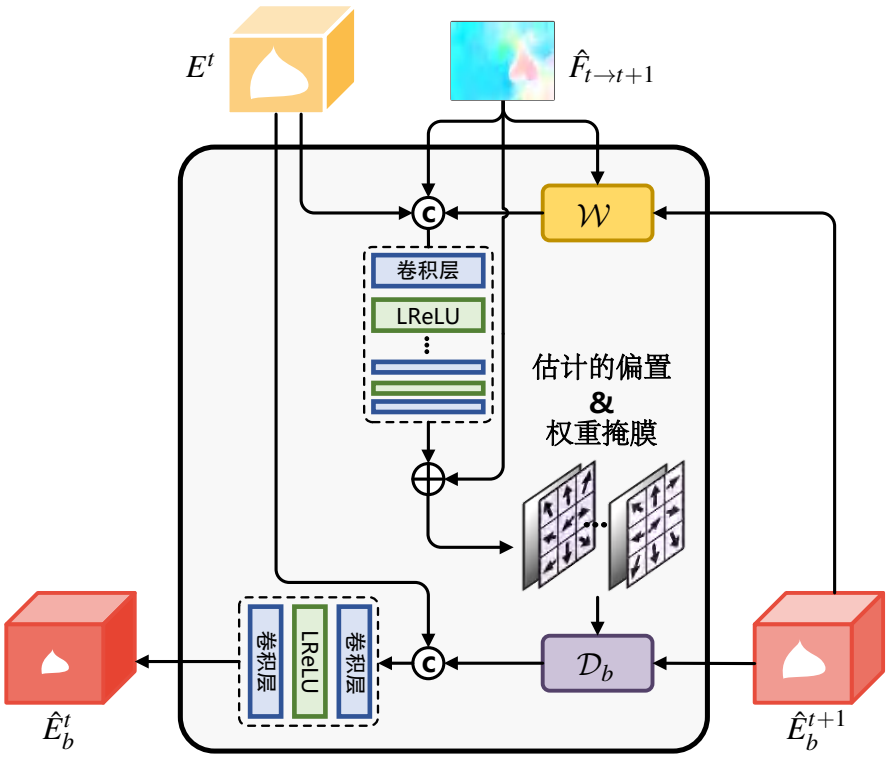

图 3.4: 使用已补全的前向流 $\hat{F}_{t\to t+1}$ 来指导特征后向传播的例子。其中，$\oplus$ 和 ©分别表示加法运算和连接运算。请注意，后向光流将以相反的方向进行。

的感受野的作用下自适应地将光流追踪到的信息进行合并，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，但它仍然需要面对公式 3.1中光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 [63, 69, 198, 199]的启发，本章算法采用了可调制的可变形卷积 [72]来进一步索引和加权候选特征点。如图 3.4所示，本章算法首先计算权重掩码 $W_{t\to t+1}$ 和与估计光流相关的偏移 $\Delta F_{t\to t+1}$。

$$[W_{t\to t+1}, \Delta F_{t\to t+1}] = \mathcal{C}_b(E^t, \mathcal{W}(\hat{E}_b^{t+1}, \hat{F}_{t\to t+1}), \hat{F}_{t\to t+1}) \tag{3.4}$$

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 分别是可变形卷积的核大小和群数。在该模块中可以通过将偏移量 $\Delta F_{t\to t+1}$ 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，本章算法使用可变形卷积层对后向特征 $\hat{E}_b^{t+1}$ 进行扭曲，而不使用公式 3.3中基于光流的扭曲，并进一步通过获得后向传播特征 $\hat{E}_b^t$:

$$\hat{E}_b^t = \mathcal{P}_b(E^t, \mathcal{D}_b(\hat{E}_b^{t+1}, W_{t\to t+1}, \hat{F}_{t\to t+1} + \Delta F_{t\to t+1})) \tag{3.5}$$

其中 $\mathcal{D}_b$ 表示可变形卷积层的操作。而权重掩码表示为 $W_{t\to t+1}$，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 [32, 33]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，本章算法使用一个可学习的 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 [32]提到的双向光流追踪的像素。

$$\hat{E}^t = \mathcal{I}(\hat{E}_f^t, \hat{E}_b^t) \tag{3.6}$$

其中 $\mathcal{I}$ 表示一个 1×1 大小的卷积层。

3.2.3 基于时空 Focal Transformer的内容补全

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 [33]中所讨论的，局部邻域的填充区域待生成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，本章算法将多个时空 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。 $\hat{\mathbf{E}}_l \in \mathbb{R}^{T_l\times\frac{H}{4}\times\frac{W}{4}\times C}$ 是所有非局部邻域的编码特征。本章算法使用软分割操作 [36]对级联的局部和非局部时间特征进行重叠块嵌入：

$$Z^0 = \mathrm{SS}([\hat{\mathbf{E}}_l, \mathbf{E}_{nl}]) \in \mathbb{R}^{(T_l+T_{nl})\times M\times N\times C_e} \tag{3.7}$$

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。

本章算法使用 Focal Transformer [78]从局部和非局部邻域中搜索以填补缺失的内容，而不是最近的工作中经常使用的原生的全局视觉 Transformer [74]。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 [77, 78]可以有效地降低计算和存储成本。（2）对于缺失区域的每个

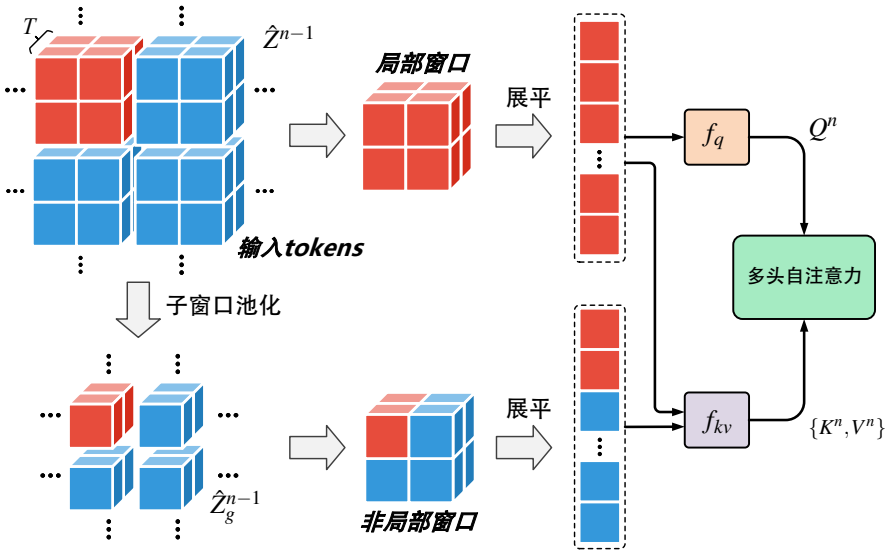

图 3.5: **时空 focal自注意力说明**。这里本章算法以窗口大小为 $2\times2\times2$ 为例。可以看到，keys 和 values $\{K^n, V^n\}$ 既包含细粒度的局部信息，又包含粗粒度的全局信息。

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对全局进行粗粒度注意力是合理的。

由于原来的 Focal Transformer无法处理序列数据，本章算法提出了一个时空 Focal Transformer，基本上将 focal窗口的大小从二维扩展到三维。具体来说，本章算法首先将输入 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力，本章算法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，本章算法通过两个线性投影层 f_q、f_{kv} 来计算 query， key和 value:

$$Q^n=f_q(\hat{Z}^{n-1}),\quad \{K_l^n,K_g^n,V_l^n,V_g^n\}=f_{kv}(\{\hat{Z}^{n-1},\hat{Z}_g^{n-1}\}) \tag{3.8}$$

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，本章算法从第 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。本章算法将相应的 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自注意力:

$$\text{Attention}(Q^n, K^n, V^n) = \text{Softmax}\left(\frac{Q^n (K^n)^T}{\sqrt{C_e}}\right) V^n \tag{3.9}$$

注意，注意力函数也可以以多头的方式进行，例子显示在图 3.5。

最后，在第 n 个 focal注意力中的整个过程被表述为：

$$Z'^n = \text{MFSA}(\text{LN}_1(Z^{n-1})) + Z^{n-1} \tag{3.10}$$

$$Z^n = \text{F3N}(\text{LN}_2(Z'^n)) + Z'^n \tag{3.11}$$

其中，MFSA和 LN分别表示多头 focal自注意力和层归一化 [200]。本章算法使用 F3N [36]来建立嵌入块间的关系。

3.2.4 训练目标

本章算法采用三个损失函数来优化网络模型。第一个是重建损失，通过L1距离测量合成视频 $\hat{\mathbf{Y}}$ 和原始视频 $\mathbf{Y}$ 之间的像素级差异:

$$\mathcal{L}_{rec} = \|\hat{\mathbf{Y}} - \mathbf{Y}\|_1 \tag{3.12}$$

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。本章算法采用了一个基于 T-PatchGAN [38]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。

对于视频补全生成器，对抗性损失为:

$$\mathcal{L}_{adv} = -E_{z \sim P_{\hat{\mathbf{Y}}}(z)}[D(z)] \tag{3.13}$$

第三种损失是光流一致性损失，如公式 3.2所示。

3.3 实验

3.3.1 设定

数据集： 为了验证所提方法的有效性，本节在两个最受学界认可的视频对象分割数据集上进行了评估，其中包括 YouTube-VOS [201]和 DAVIS [192]。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。作者遵循原始的分割模式，在 YouTube-VOS的测试集上展

示了实验指标的报告。 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。根据 FuseFormer [36]，测试集的50个视频片段被用来计算指标。作者在 YouTube-VOS数据集上训练本章算法的模型，并在 YouTube-VOS和DAVIS数据集上评估它。至于用于遮挡输入视频帧的掩码，在训练过程中，数据采样流程会为每一帧视频图像生成固定的或随机形状的掩码，以模拟静态掩码去除和物体去除的应用，具体如 [30, 34, 36, 37, 39]所使用的方式。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。

架构： 在作者的模型中，编码器和解码器使用与 FuseFormer [36]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，作者采用了一个轻量级模型 SPyNet [202]作为作者的光流补全模块。为了利用原始 SPyNet中学习到的光流先验，作者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 [36–38]相同。可变形卷积的核大小 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，作者使用软合成运算符 [36]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：对于训练损失函数，$\mathcal{L}_{rec}$， $\mathcal{L}_{adv}$ 和 $\mathcal{L}_{flow}$ 的权重分别为1， 10^{-2}，和 1。考虑到 GPU的内存限制，作者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 STTN [37] 和 FuseFormer [36]，在评估和测试过程中，本节使用一个大小为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。最终的模型使用 $\beta_1 = 0$ 和 $\beta_2 = 0.99$ 的 Adam优化器被训练了50万次。所有模块的初始学习率被设定为0.0001，并在40万次迭代时减少10倍。在本节的消融研究中，模型进行了25万次的迭代训练。所有实验都使用8个 NVIDIA Tesla V100 GPU进行训练，批量大小设置为8。

度量指标： 作者选择 PSNR、SSIM [194]、VFID [195]和光流的扭曲误差 E_{warp} [196]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 [36, 37]中得到了采用。在衡量时间上的一致性的指标中，采用了光流的扭曲误差 E_{warp}。

3.3.2 对比

量化结果： 本节展示了 YouTube-VOS [201]和 DAVIS [192]在固定掩码下的定量结果, 并将本章方法与以前的视频补全方法进行比较，包括 VINet [39]、DFVI [32]、LGTSM [30]、CAP [34]、STTN [37]、FGVC [33]和 Fuseformer [36]。正如在表 3.1中所示，本章方法在所有四个定量指标上都明显超过了以前的最先进算法。这一结果表明，本章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）以及更好的时空一致性（E_{warp}）的视频，验证了本章算法的优越性。

定性结果：作者选择了三种有代表性的方法，包括 CAP [34]、FGVC [33]和 Fuseformer [36]，来进行视觉比较。 图 3.6和 图 3.12 显示了视频补全和物体移除的结果。这些方法很难合理地恢复被遮挡区域的细节的同时，本文方法可以产生较真实的纹理和结构信息。这表明了所提方法的有效性。为了进一步的综合比较，作者对目标移除和消除静态掩码的应用都进行了用户研究。作者选择了五种方法，包括两种基于光流的方法（如，DFVI [32]和 FGVC [33]）、以及三种基于注意力的方法（如，CAP [34]、STTN [37]和 Fuseformer [36]）。作者邀请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、一个使用本文算法补全的视频和一个使用随机方法补全的视频组成。用户研究结果显示在图 3.7。从图中可以看到，与几乎所有方法的结果相比，志愿者显然更喜欢本文算法生成的结果。尽管在与 FGVC的比较中不存在这种明显的偏好，但本文算法仍然获得了大多数的投票。这表明，本文算法可以产生比其他方法更好的视觉效果。此外，本章还有对应的视频 demo可以在项目页面中查看。

效率比较： 作者使用 FLOPs和推理时间来衡量每种方法的效率。 FLOPs计算时时序窗口尺寸为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 3.1。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

表 3.1: 在 YouTube-VOS [201] 和 DAVIS [192]数据集上与目前最先进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 [36]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。

	准确性								高效性	
	YouTube-VOS				DAVIS				FLOPs	Runtime
Models	PSNR ↑	SSIM ↑	VFID ↓	$E_{warp}{}^{*}$ ↓	PSNR ↑	SSIM ↑	VFID ↓	$E_{warp}{}^{*}$ ↓		（s/frame）
VINet [39]	29.20	0.9434	0.072	0.1490	28.96	0.9411	0.199	0.1785	-	-
DFVI [32]	29.16	0.9429	0.066	0.1509	28.81	0.9404	0.187	0.1608	-	2.56
LGTSM [30]	29.74	0.9504	0.070	0.1859	28.57	0.9409	0.170	0.1640	1008G	0.23
CAP [34]	31.58	0.9607	0.071	0.1470	30.28	0.9521	0.182	0.1533	861G	0.40
FGVC [33]	29.67	0.9403	0.064	0.1022	30.80	0.9497	0.165	0.1586	-	2.44
STTN [37]	32.34	0.9655	0.053	0.0907	30.67	0.9560	0.149	0.1449	1032G	0.12
FuseFormer [36]	33.29	0.9681	0.053	0.0900	32.54	0.9700	0.138	0.1362	752G	0.20
E^{2}FGVI (本章算法)	**33.71**	**0.9700**	**0.046**	**0.0864**	**33.01**	**0.9721**	**0.116**	**0.1315**	682G	0.16

图 3.6: 和 CAP [34]、 FGVC [33]、 FuseFormer [36]以及本章算法 E²FGVI的定性结果比较。

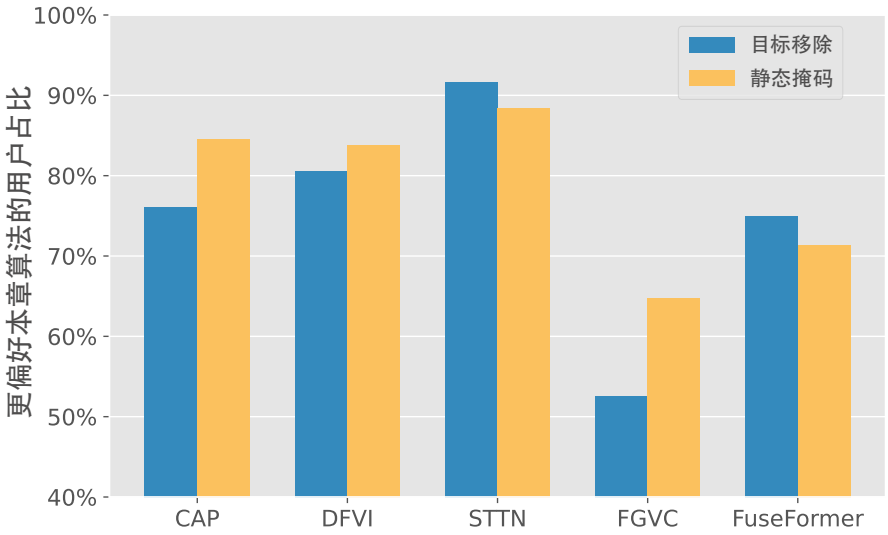

图 3.7: **用户研究结果**。 纵轴表示与其他方法相比，更喜欢本文算法的志愿者百分比。

表 3.2: 光流补全模块的消融实验。

模型变种	PSNR	SSIM
除去运动信息	32.08	0.9673
不补全光流	32.23	0.9682
补全光流	32.35	0.9688
使用 GT光流	32.54	0.9698

图 3.8: 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。

最低的 FLOPs。这表明，本文算法在视频补全应用上具有很高的效率。

表 3.3: 参数的比较。FuseFormer* 表示原始 FuseFormer的一个较大参数量版本。

	FuseFormer [36]	FuseFormer*	E^2FGVI
Params. (M)	36.6	41.6	41.8
PSNR/SSIM	31.74/0.9662	31.91/0.9669	32.35/0.9688

进一步的参数比较： 表 3.3中给出了参数量的增加是否给模型尽管作者的方法比目前最先进方法（如， FuseFormer [36]）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，作者在 FuseFormer中添加了残差块，以实现与作者相似的参数量。作者的方法仍然比更大参数量的 FuseFormer表现得更好。

3.3.3 消融实验

作者在光流补全、特征传播和注意力机制方面进行了三项消融实验，以验证本文算法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，作者调查了运动信息对视频补全的重要性。

表 3.4: 对特征传播模块的研究。“Flow”表示在公式 3.4中基于光流的扭曲函数 $\mathcal{W}$。“DCN”表示调制的可变形卷积 [72]。

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

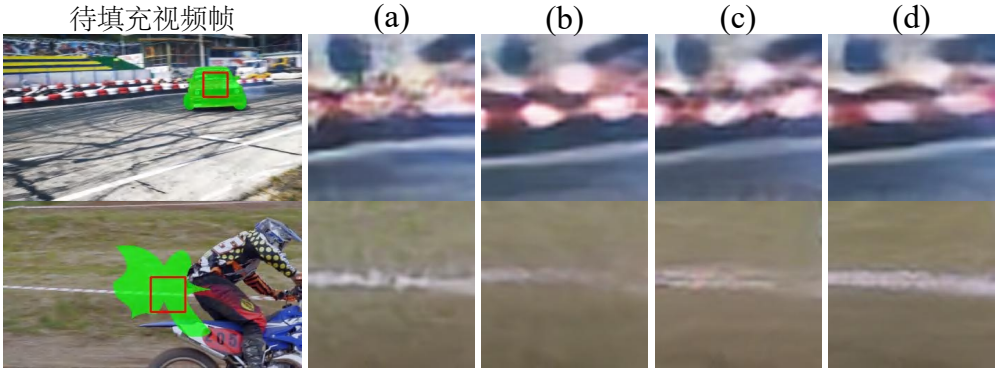

图 3.9: 特征传播模块的消融实验的定性结果。最后四列对应表 3.4四个例子。

通过只删除光流一致性损失 $\mathcal{L}_{flow}$，本文的光流补全模块不再提供关于对象运动的信息（见图 3.8），导致性能大幅下降，如表 3.2所示。其次，作者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 3.8）。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，本文算法获得了比以前更大的 PSNR和SSIM值。如图 3.8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 3.2图 3.8中，作者还展示了本文算法的潜在上界，该方法估计了不包含待填充区域的帧之间的光流。

对特征传播模块的实验： 在本文算法移除特征传播模块后（例子(a)在表 3.4），定量指标的数值急剧下降。从图 3.9 (a)中，作者可以看到，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扭曲和传播后（见公式 3.3）（例子(b)在表 3.4），由于本文算法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 3.9 (b)），PSNR值增加了很大幅度（0.42dB）。但是，基于光流的扭曲和传播很难恢复不能被光流追踪到的内容（图 3.9 (b)中的白线）。此外，对于仅涉及基

表 3.5: 对各种注意力机制的消融研究。FuseFormer [36] 是目前使用原有的全局注意力中最先进的方法。

注意力变体	PSNR	SSIM	FLOPs
全局注意力	31.74	0.9662	752G
局部注意力	31.57	0.9648	497G
Focal注意力	31.73	0.9653	560G

于可变形卷积的扭曲的特征传播模块（例子(c) 在 表 3.4），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扭曲相比，缺乏从相邻帧扭曲的有效信息，因此涉及更多伪影。通过将可变形卷积与流引导相结合例子在 表 3.4，PSNR和 SSIM值可以进一步提高。

在图 3.9 (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： 在本实验中作者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，包括原有的全局注意力（FuseFormer [36]），局部时空窗口注意力，和本文采用的时空 focal注意力。如表 3.5所示，全局注意力得到了最佳的量化性能，但也需要过大的计算量。局部注意力引入了局部时空窗口，就像 Video Swin Transformer [82]那样。虽然 FLOPs减少了34%，但由于注意力的计算被限制在局部窗口，因此导致性能不佳。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%。

3.3.4 以离线的方式补全光流

为了验证在线光流补全的有效性，作者使用 FGVC [196]光流补全模块以离线方式准备补全的光流。然后作者用 FGVC补全的光流重新训练一个模型。该模型的 PSNR值略高于作者的端到端设置（32.38 vs. 32.35（dB）），然而推理速度比作者的慢得多（1.21 vs. 0.16（秒/每帧））。

3.3.5 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，作者在进行内容生成之前，在图 3.10中可视化了时间窗口为5的平均局部邻域特征。图 3.10 中的四种情况对应于作者论文中 Tab. 3中的四种变体。对于没有特征传播的模型（图 3.10(a)），

图 3.10: 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。

显然，作者可以看到所有帧的待填充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扭曲（图 3.10(b)）或基于可变形卷积的扭曲（图 3.10(c)）的模型，损坏的区域被来自相邻帧的扭曲后的内容填充。由于有更多的采样特征点，基于可变形卷积的扭曲可以产生比基于光流的扭曲更平滑的内容。然而，特别是对于最后两个时间特征（ 图 3.10中最后两列），与基于光流的扭曲相比，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形卷积，最终的传播模块（图 3.10(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。

3.3.6 内容生成能力的研究

为了纯粹评估作者的方法的内容生成能力，作者首先预先填充了可以被流场追踪的像素 [33]。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，作者将预填充的视频分别送入一个图像补全模型 [22]和本文算法。作者的生成结果比 DAVIS数据集上的图像补全模型的 PSNR值大得多（31.74 vs. 30.80（dB））。

3.4 总结

本章提出了一个时空联合一致性驱动的端到端视频补全框架，名为 E^2FGVI。本章框架中设计的三个模块（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，本章框架在两个基准数据集上取得了目前最先进定量和定性性能，并且在推理时间和计算复杂度方面具有很好的优势。作者希望它可以成为未来工作的一个强有力的基线。

局限： 图 3.11 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，本文算法与 FGVC [33]和 FuseFormer [36]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。

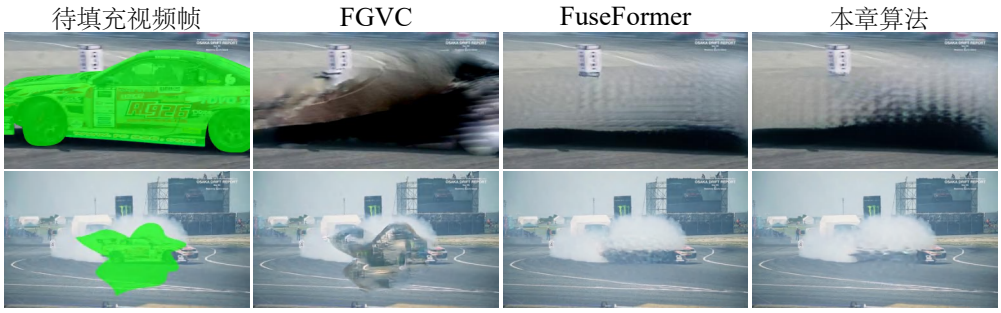

图 3.11: 两个失败案例（汽车漂移）。目前的视频补全方法不能处理大运动或大量的物体细节缺失，可能产生严重的伪影。

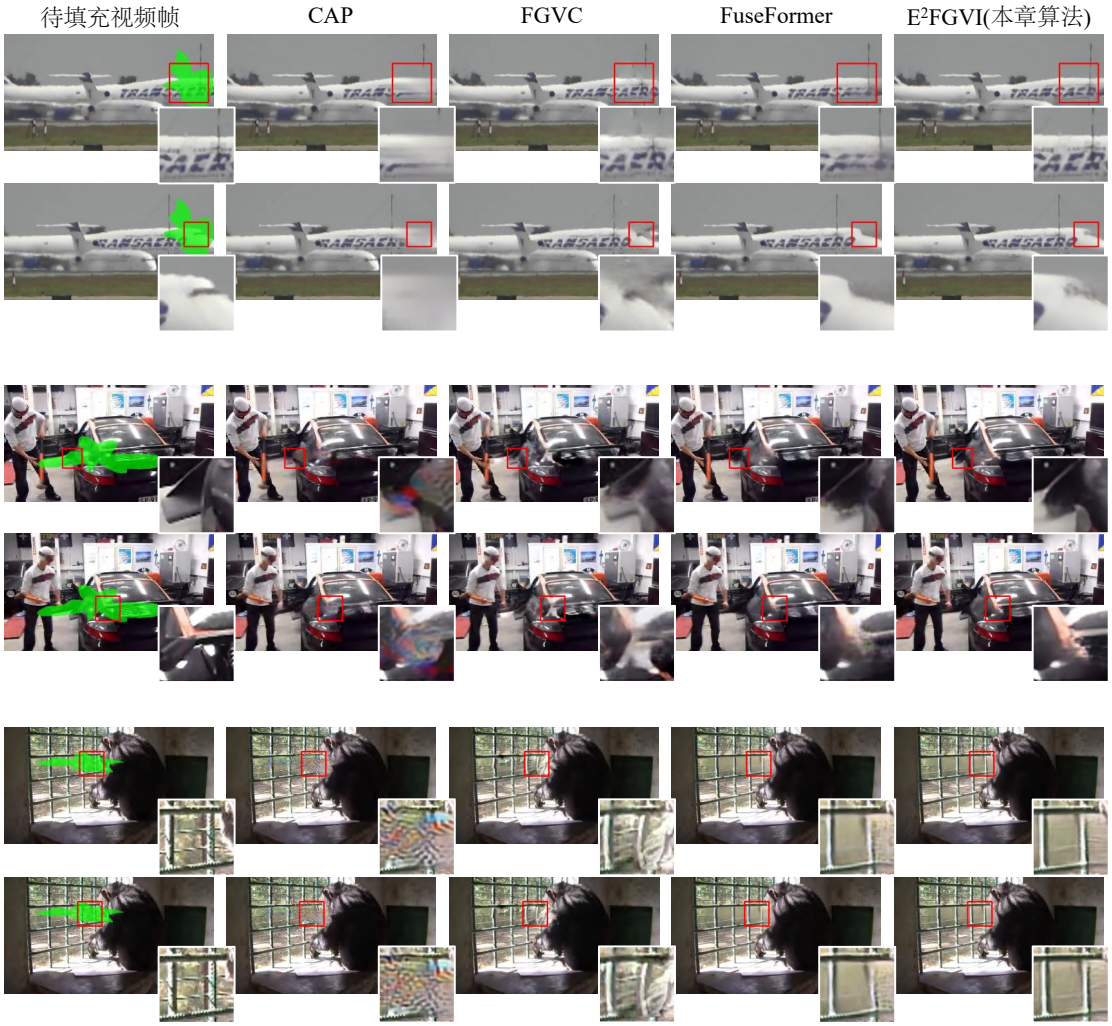

图 3.12: YouTube-VOS [201]数据集上的定性视频补全结果.

图 3.13: YouTube-VOS [201]数据集上的定性视频补全结果.

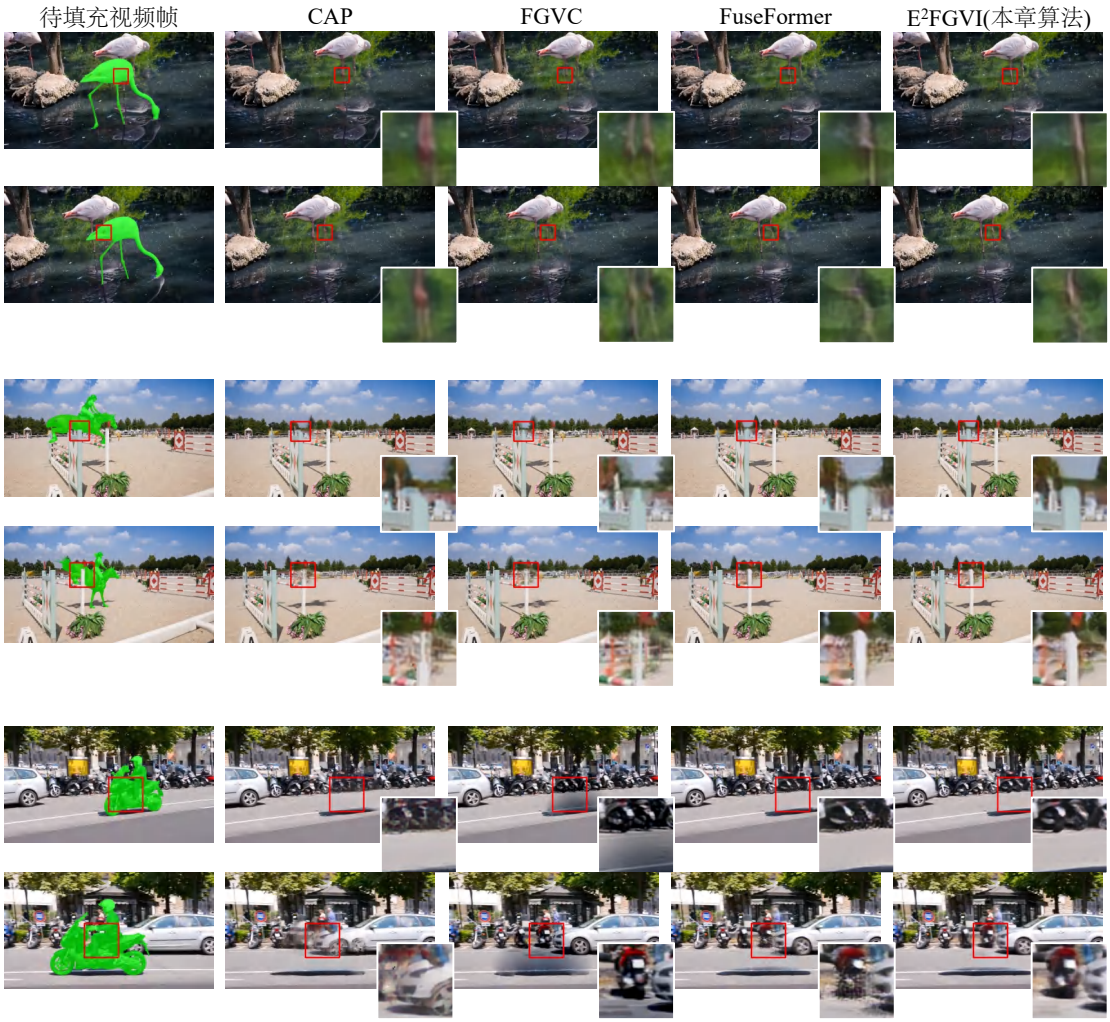

图 3.14: DAVIS [192]数据集上的定性物体移除结果.

图 3.15: DAVIS [192]数据集上的定性视频补全结果.

第四章 运动属性驱动的视频帧生成

本章主要研究运动属性驱动的视频帧生成。章节 4.1中介绍研究背景、动机及解决方案概要； 章节 4.2介绍本章构建的时空联合一致性驱动的视频补全框架； 章节 4.3给出评测结果和结果分析； 章节 4.4讨论了本章算法的核心设计与相关算法的区别。 章节 4.5对本章进行小结。

4.1 引言

4.1.1 研究背景

视频插帧（Video Frame Interpolation）是一种近年来备受关注的视频处理技术，旨在通过参考输入视频相关帧的信息生成中间帧，进而提高视频的时间分辨率。它已经被应用于各种下游任务，包括慢动作生成 [64, 70]、新视角合成 [203–205]、视频压缩 [206]、文本到视频的生成 [131]等。

最近，基于光流的视频插帧算法 [64, 102, 112, 117, 128, 207] 由于其有效性而在相关研究中占据主导地位。一种常见的基于光流的技术是通过给定的一对视频帧来估计其双边/双向光流，然后通过反向 [112, 115, 117] 或者正向 [107, 108, 119]扭曲（warping），将像素/特征传播到指定的时间步上。因此，合成帧的质量在很大程度上依赖于光流估计的结果。事实上，通过预训练的光流模型来近似中间光流是非常冗余的，并且这些光流在视频插帧任务中并不完全适用 [108, 117]。

解决这个问题的一个可行的方案是通过一个端到端的训练方式 [62, 64, 109, 112]去估计面向任务的光流。但是，像大位移和遮挡等困难问题仍亟待解决。这些困难主要来自光流估计自身的局限性。因此便引出了一个值得思考的问题：为什么之前采用的面向任务的光流仍然不能很好的解决视频插帧中的问题？

4.1.2 研究动机与贡献

最近的一些研究 [62, 112] 表明：面向任务的光流通常是和真实的光流一致的，但是在局部细节上存在差异。受此启发，作者尝试从以下两个方面来解决上述问题：

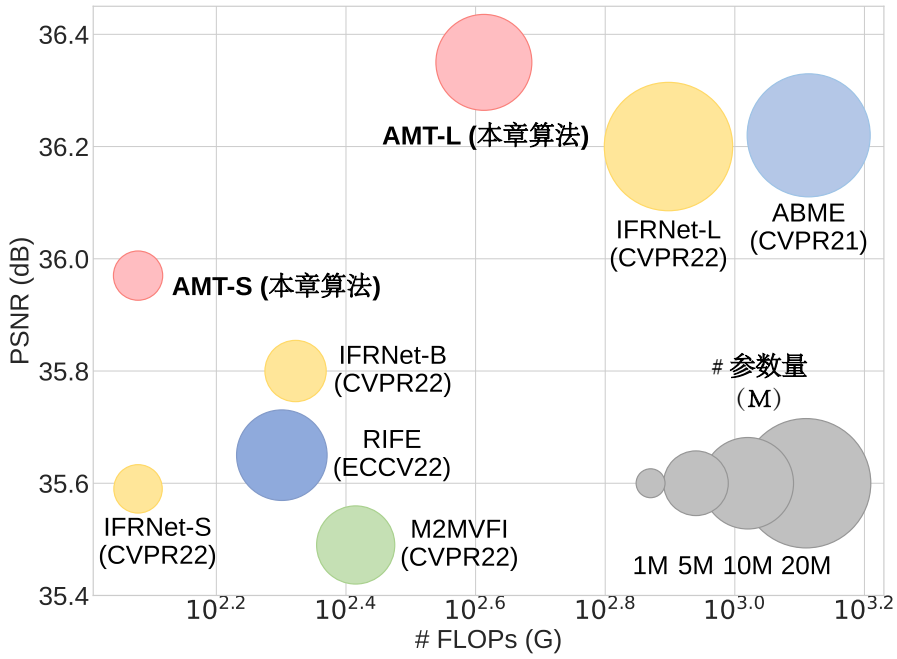

图 4.1: **性能与参数量和 FLOPs对比。** PSNR值是在 Vimeo90K [62]数据集中计算获得的。作者使用分辨率为720p的成对帧对来计算 FLOPs。可以看出作者的AMT方法优于目前最先进的方法，并且效率更高。

(i) 现有视频插帧方法预测的光流场与真实位移不够一致，尤其是遇到大位移时(见 图 4.2)。现有方法大多采用类似 UNet的架构 [208]和普通卷积来构建视频插帧模型。然而，这种类型的架构在对大位移运动进行建模 [147, 209–211]时容易在早期阶段累积错误。结果导致预测的光流场不准确。

(ii) 现有方法通过预测一对光流场，将解决方案的集合限制在一个狭小的空间内。这使他们难以处理运动边界周围的遮挡和细节，从而使最终结果恶化。(见图 4.2 和图 4.6).

在本文中，作者提出了一种新的视频插帧网络框架-**AMT**。针对先前工作的上述两个主要缺点，AMT 探索了两种新的设计来提高预测光流的保真度和多样性。

首先，作者的第一个设计是基于 RAFT [147]中的全对相关性，它充分模拟了帧之间的密集对应关系，特别是对于大位移。作者提出构建双向相关的匹配代价（cost volume）取代单向相关的匹配代价，并引入放缩查找策略来解决由于某一帧中存在的不可见信息引起的坐标不匹配问题。此外，通过该策略检索到的相关性有助于本章的模型通过跨尺度的方式联合更新双边光流和插帧内容

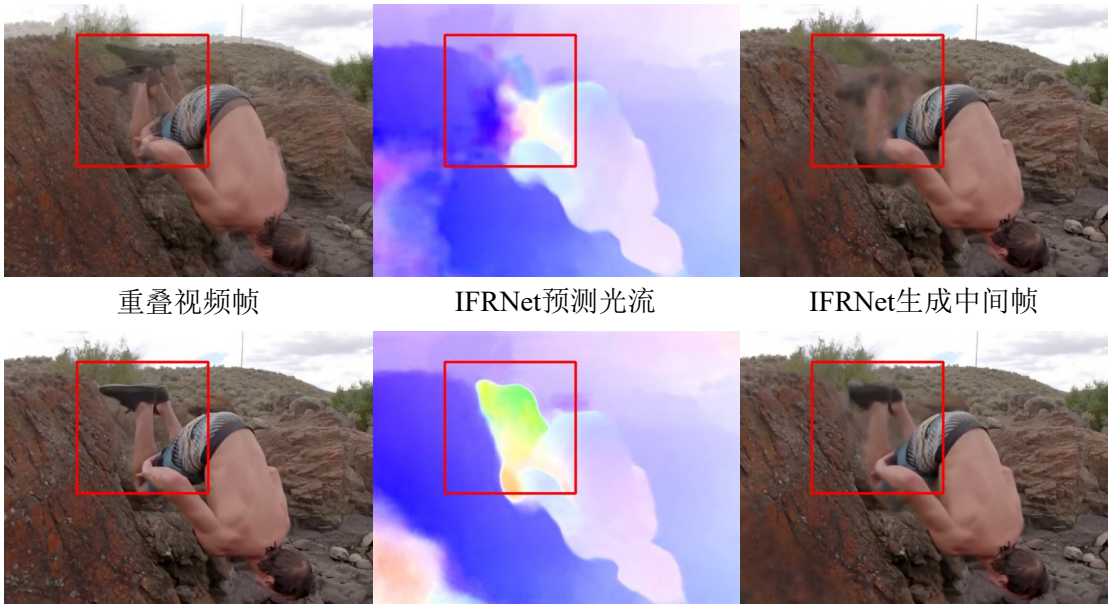

图 4.2: **估计光流和插帧画面的定性比较。** 本章提出的 AMT保证了中间光流的一致性，并精确地合成具有遮挡区域的快速移动对象，而以前最先进的 IFRNet [112]未能达到这样的效果。

特征。因此，本章提出的网络框架既保证了跨尺度光流的保真度，又为后续细化奠定了基础。

其次，考虑到预测一对流场难以应对场景中出现的遮挡情况，作者提出从一对更新过的粗双边光流中导出多组细粒度的光流。输入帧可以通过这些光流分别反向扭曲到目标时间步长。通过使用多个不同的流场为每个要插值的像素提供了足够的潜在值的来源，减轻了遮挡区域中的歧义性问题。

作者在公共基准测试中测试作者提出的 AMT方法，可以发现，与目前最先进的方法相比，作者的模型在不同的模型规模上均取得了较好的性能，并且更加高效（见图 4.2）。作者的轻量级模型在 Vimeo90K [62] 上的性能优于目前最先进的轻量级模型 IFRNet-B [112]，PSNR值提高了0.17dB，并且 FLOPs 和参数量仅为该方法的 60%。在更大尺度的模型上，作者的 AMT 在 Vimeo90K [62]上测试的 PSNR值超过之前的最先进的模型（即 IFRNet-L [112]）0.15dB，并且 FLOPs 只有其 75%，而参数量只有其的65%。此外，作者还提供了一个巨大尺度的模型来与基于 Transformer的最先进的方法 VFIFormer [102]进行比较。作者基于卷积的 AMT的方法与其性能相当，但是 VFIFormer的计算代价是作者提出的 AMT的23倍。由于 AMT良好的性能，作者希望本章算法能够为高效视频插

帧提供一个新的视角。

4.2 方法

给定一对输入帧 (I_0, I_1)，视频插帧任务的目标是生成特定时刻 t 的中间帧 I_t，其中 $0 < t < 1$。作者提出的 AMT是一种基于光流的单阶段方法，其中双边光流和插帧过程的中间特征联合起来进行更新和上采样。如图 4.3所示，它主要被分为三个主要组件： 1）一个同时提取特征和初始双边光流的的编码器； 2）一个多尺度的双向相关的匹配代价，用来在粗尺度下联合更新双边光流和中间特征； 3）一个多场细化操作，用来在最精细的尺度上使用多个光流组进行目标帧的插值。受益于此类设计，AMT可以在较粗尺度上估计出与真实位移量大致一致的运行矢量。同时，这些估计出的运动矢量在最细粒度上具备多样性，满足了以任务为导向的光流的需要。这些设计还使作者的 AMT 能够捕捉大位移运动并可以有效地处理遮挡区域。

4.2.1 初始光流和特征提取

作者采用了两个独立的特征提取器。它们均被应用于成对的输入帧 (I_0, I_1)，但它们的用途各不相同。第一个特征提取器是*相关性编码器*，它将输入帧映射到成对的密集特征以构建双向的相关性匹配代价。该特征提取器输出的特征 $\mathbf{g}_0, \mathbf{g}_1 \in \mathbb{R}^{H/8 \times W/8 \times D}$，空间分辨率为输入图像的1/8，通道数为 D。

第二个特征提取器是*内容编码器*，用来输出初始化的插值中间特征 x_t^1，并且预测初始的双边光流 $F_{t\to 0}^1$ 和 $F_{t\to 1}^1$。输出的初始化特征和预测的双边光流的空间分辨率与相关性编码器的输出相同。此外，内容编码器提取输入帧 I_0, I_1 中的金字塔特征 $\{X_0^l, X_1^l \mid l \in \{1,2,3\}\}$ 用以进一步的扭曲。它们的架构细节可以参考章节 4.3.3。

4.2.2 全对相关性

双向相关匹配代价：与 RAFT [147]类似，本章算法计算了所有特征向量对之间的点积相似性，以构建4D相关性匹配代价。给定一对特征 $\mathbf{g}_0, \mathbf{g}_1$，作者可以以下式计算其相关量 $\mathbf{C}$：

$$\mathbf{C}_{ijkl} = \sum_h \mathbf{g}_{0,ijh} \cdot \mathbf{g}_{1,klh}, \quad \mathbf{C} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{H}{8} \times \frac{W}{8}} \tag{4.1}$$

为了进一步测量跨尺度的相似性，相关性匹配代价的最后两个维度通过重

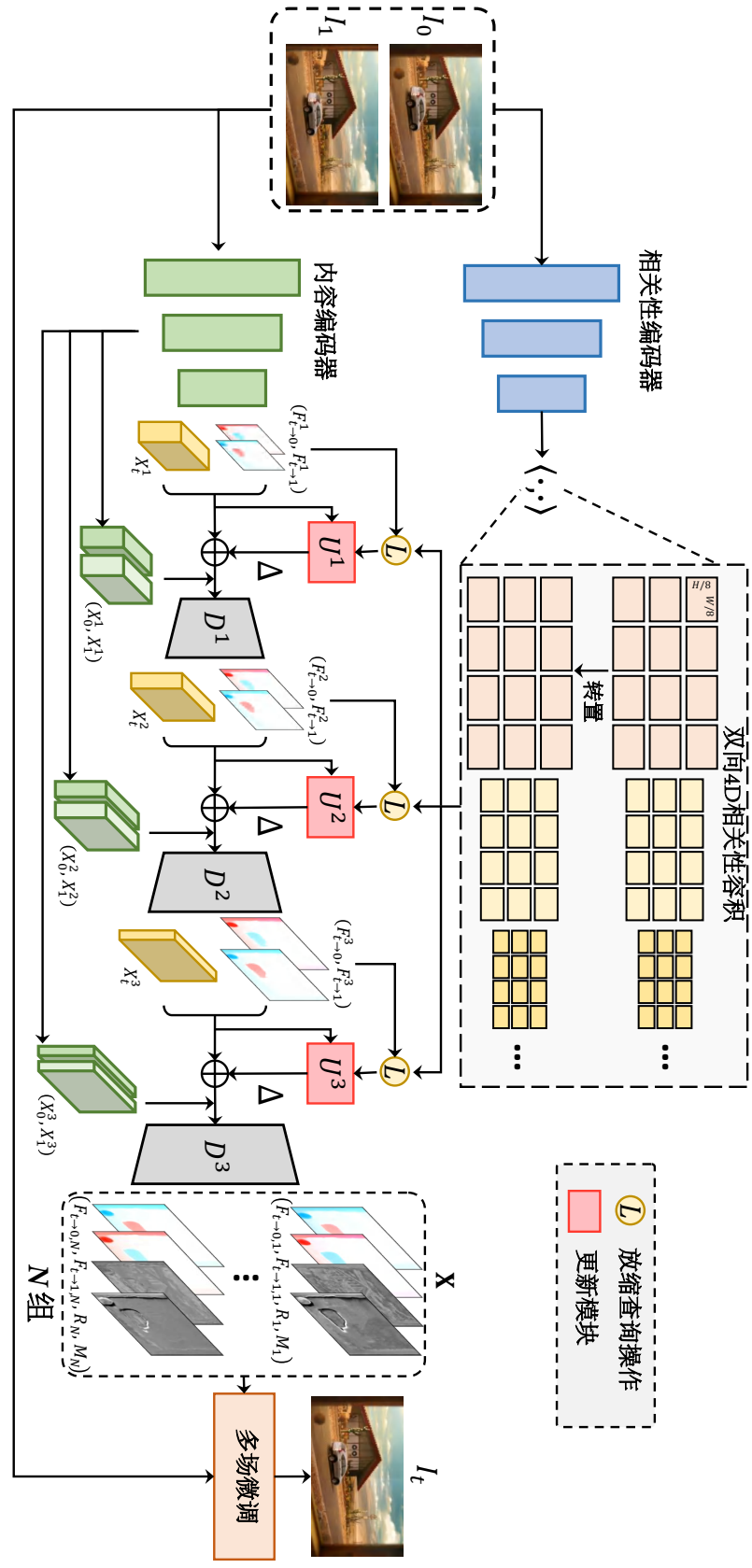

图 4.3: **AMT的概览**。首先，输入帧被送入到相关的编码器中提取特征，用于构建双向相关的匹配代价。然后，内容编码器提取可见帧的金字塔特征并生成初始双边流和插值中间特征。接下来，作者使用双向光流去检索双向的关联性来联合更新每一层中的光流场和中间特征。最后，作者生成了多组光流场，遮挡权重图和基于粗粒度的估计残差来生成中间帧。

复的二维平均池化层进行下采样，其中池化核大小为2，步长为 2。这样作者就得到了一个四层的相关金字塔 $\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4\}$。

但是，RAFT中的相关金字塔是单向的。它只反映从 I_0 到 I_1 之间的多尺度对应关系。因此作者将其称为前向相关性金字塔。这种单向的对应对于视频插帧任务来说是不够的，因为运动通常是不对称的 [104, 124]。为了在有限计算下得到双向的相关性金字塔而不重新计算矩阵乘法，作者直接转置相关量 $\mathbf{C}$ 来表示相反方向的对应关系。在得到转置后的相关量 $\mathbf{C}^T$，作者采用相同的池化操作来生成反向相关的金字塔 $\{\mathbf{C}_1^T, \mathbf{C}_2^T, \mathbf{C}_3^T, \mathbf{C}_4^T\}$。注意到，双向相关性匹配代价只需计算一次。这种紧凑的全局表征有利于本章提出的网络框架有效的感知大运动。

相关性放缩查找： 构建双向相关量后，作者使用估计的双边光流 $F_{t\to 0}^l$ 和 $F_{t\to 1}^l$ 来查询相关特征图。在 RAFT中，由于其估计光流和相关性匹配代价共享相同的坐标系，因此可以直接执行查找操作。例如，从第0帧到第1帧的运动 $F_{0\to 1}$ 以及与之对应的相关性匹配代价都是第0帧的坐标系。因此，相关特征图可以准确的从匹配的光流场中进行采样。然而，对于视频插帧任务，作者只能从可见帧(即 I_0, I_1)构建相关的容积（volume），但是估计光流(即 $F_{t\to 0}^l$ 和 $F_{t\to 1}^l$)需要使用不可见的中间帧 I_t。如果直接使用类似 RAFT的方式进行查询会导致坐标系之间的不匹配，进而使得相关性的查找不可信，并进一步影响光流的更新。一个直接的解决方法是将双边光流 $F_{t\to 0}^l$ 和 $F_{t\to 1}^l$ 转换到双向光流 $F_{0\to 1}^l$ 和 $F_{1\to 0}^l$。

为了实现这一目标，作者只需根据局部平滑运动假设来放缩估计的双边光流 [64, 124, 150]。具体来说，局部平滑运动假设认为移动物体在很小的时间间隔内部分重叠。因此，同一位置的双边光流和双向光流一般方向一致，但大小不同。因此，双向光流 $F_{0\to 1}^l$ 和 $F_{1\to 0}^l$ 可以通过以下方式进行估计：

$$F_{0\to 1}^l = \frac{1}{1-t} F_{t\to 1}^l, \quad F_{1\to 0}^l = \frac{1}{t} F_{t\to 0}^l \tag{4.2}$$

随后，作者使用估计的双边光流在双向相关性匹配代价中进行类似于RAFT中的查找操作。作者构建了两个以双向光流为中心且具有预定义半径的查找窗口。窗口中的查找操作是在双向相关金字塔中的所有级别上进行的。检索到的双向相关性被连接成一个特征图，以进一步联合后更新双边光流和插值的中间特征。

利用检索的相关性进行更新： 虽然 RAFT 以单一分辨率更新和生成预测的光流，但作者遵循大多数基于光流的视频插帧方法以采用从粗粒度到细粒度的

方式预测双边光流 [107, 112, 117, 119]。这是因为输入对的特征需要根据最新的光流预测逐渐进行扭曲，以生成更可信的中间特征。考虑视频插帧任务中双边流场与中间特征之间的相互关系 [109, 112, 117]，作者还对中间特征以及中间运动进行更新和上采样。

具体来说，在每一层 l 的更新阶段，作者采用了一个更新块依据检索到的双向相关性来联合预测双边光流区域 $F_{t\to0}^{l}, F_{t\to1}^{l}$ 和中间插值特征 X_t^l 的残差。在每个更新块中，双向相关特征和双边流首先通过两个卷积层。然后，它们与插值的中间特征连接并送入两个卷积层，进而取代 RAFT 中繁琐的 GRU 单元。最后，输出特征被发送到两个单独的输出头部，用于预测双边光流残差 $\Delta F_{t\to0}^{l}, \Delta F_{t\to1}^{l}$ 和一个插值特征残差 ΔX_t^l。每一个头部由两个卷积构成。

可以注意到的是检索到的相关特征的空间维度与相关性匹配代价的前两个维度(如：$\frac{H}{8}\times\frac{W}{8}$) 相同，但与上层的中间特征和运动的空间维度不同。因此，作者需要在将光流场和中间特征输入更新块之前相应地缩小光流场和中间特征的规模，并对预测残差进行上采样以进行更新。通过缩小尺寸，作者可以使更新块在低分辨率空间下工作，从而提高了效率。更新的中间特征 $\hat{X}_t^l$ 可以表述为 $\hat{X}_t^l = X_t^l + \Delta X_t^l$，其中 ΔX_t^l 为更新块输出的内容残差。需要更新的双边光流 $\hat{F}_{t\to0}^{l}, \hat{F}_{t\to1}^{l}$ 也通过相同的方法进行获取。

作者采用更新过的双边光流来扭曲输入帧的特征 X_0^l, X_1^l。本章用 X_0^l, X_1^l 以表示扭曲的特征。扭曲的特征、更新的双边光流和更新的双边光流被连接在一起，然后送入第 l 个解码器。第 l 个解码器 D^l 同时预测上采样的双边光流 $F_{t\to0}^{l+1}, F_{t\to1}^{l+1}$ 和中间特征 X_t^{l+1}，公式如下：

$$[F_{t\to0}^{l+1}, F_{t\to1}^{l+1}, X_t^{l+1}] = D^l([\hat{X}_0^l, \hat{X}_1^l, \hat{F}_{t\to0}^{l}, \hat{F}_{t\to1}^{l}, \hat{X}_t^l]) \tag{4.3}$$

特别的，公式 4.3没有考虑最后的一个解码器 D^3。这一解码器负责生成多个光流和视频插帧任务中特定的遮挡权重图。每个解码器的架构细节在章节 4.3.3中列出。

4.2.3 多场微调

在基于光流的视频插帧方法中，常用的插帧公式为：

$$I_t = M \odot \mathcal{W}(I_0, F_{t\to0}) + (1-M) \odot \mathcal{W}(I_1, F_{t\to1}) + R \tag{4.4}$$

其中，$\mathcal{W}$ 为反向扭曲的操作，$\odot$ 表示逐元素乘法。M 是一个估计的范

围0到1的遮挡权重图。 $F_{t\to 0}$ 和 $F_{t\to 1}$ 是最终预测的双边光流。 R 是估计的残差。上式既考虑了时间一致性又考虑了遮挡，因此可以有效地合成中间帧。

然而，由于遮挡区域中的每个位置都有许多潜在的候选像素，仅预测一对光流场会忽略掉这些候选像素，从而限制了在狭小空间中进行插值的解集。

基于先前预测的粗粒度的光流（大体与真实位移一致），多场微调模块派生出多对任务导向的细粒度光流（包含前向和后向光流）以供插帧任务使用。作者还针对每对光流联合估计了内容残差和遮挡权重图。处理公式如下:

$$\begin{aligned} \mathbf{X} &= D^3([\hat{X}_0^3, \hat{X}_1^3, \hat{F}_{t\to 0}^3, \hat{F}_{t\to 1}^3, \hat{X}_t^3]), \\ \mathbf{X} &= \{F_{t\to 0,n}, F_{t\to 1,n}, M_n, R_n | n \in \{1,2,...,N\}\} \end{aligned} \tag{4.5}$$

其中 N 为输出组的总数. $(F_{t\to 0,n}, F_{t\to 1,n})$、 M_n 和 R_n 分别是第 n 个估计的双边光流对、遮挡权重图和内容残差。值得注意的是, 公式 4.5 可以根据光流组的数量放大最后一个解码器的输出通道来轻松实现输出多组内容，保证了插帧过程的效率。最终的中间帧可以通过以下方式获得:

$$I_t = \mathcal{C}([I_t^1, ..., I_t^N]) \tag{4.6}$$

其中第 n 个插值帧 I_t^n 通过 公式 4.4 使用对应的一组双边光流对、遮挡权重图和内容残差进行计算。作者堆叠两个卷积层（表示为 $\mathcal{C}$），用于自适应合并候选帧并细化最终结果。更进一步的多对光流场的分析可以参见章节 4.3.7。

4.2.4 损失函数

本章的 AMT算法使用了三个损失。为了更好地预测面向任务的光流，作者使用了 IFRNet [112]中的光流蒸馏损失 $\mathcal{L}_{flow}$。它更集中于易于重建的光流区域，而不惩罚那些不太准确的区域。这种损失应用于更新多尺度的粗粒度光流，而在最后的多场微调阶段更新最精细的光流时并不引入该损失函数。这样设置的原因是因为希望最后一阶段基于之前的粗粒度光流派生出更符合任务特性的任务导向光流。 Charbonnier损失 [212] $\mathcal{L}_{char}$ 和 Census损失 [213] $\mathcal{L}_{css}$ 则用来监督中间帧的生成。前者测量真实插帧图像 I_t^{GT} 与生成图像 I_t 之间逐像素的差值，后者则计算 I_t^{GT} 和 I_t 的图像中经过 Census特征转换后块间的软汉明距离。最终目标函数定义如下:

$$\mathcal{L} = \lambda_{char}\mathcal{L}_{char} + \lambda_{css}\mathcal{L}_{css} + \lambda_{flow}\mathcal{L}_{flow}, \tag{4.7}$$

其中 λ_{char}， λ_{css}，和 λ_{flow} 是每个损失对应的权重。

表 4.1: **与目前最先进方法的定量比较**。根据计算复杂度，作者将现有的方法分为三组。对每一组来说，最好的方法会对应**加粗**，第二好的方法会对应 下划线。“OOM”表示在一张 RTX 3090 的 GPU上进行验证时，会出现内存溢出（out-of-memory）的问题。† 表示作者禁用该方法中的测试时间增强 [117]以便为了公平对比。

方法	Vimeo90K [62]	UCF101 [214]	SNU-FILM [99]				Xiph [215]		速度	参数量	FLOPs
			Easy	Medium	Hard	Extreme	2K	4K	(毫秒/帧)	(M)	(T)
AdaCoF [65]	34.38/0.972	35.20/0.970	39.85/0.991	35.08/0.976	29.47/0.925	24.31/0.844	34.86/0.928	31.68/0.870	52	21.8	0.36
M2M-VFI [108]	35.49/0.978	35.32/0.970	39.66/0.991	35.74/0.980	30.32/0.936	25.07/0.860	**36.44/0.943**	33.92/0.899	40	7.6	0.26
RIFE [117]	35.65/0.978	35.28/0.969	**40.06**/0.991	35.75/0.979	30.10/0.933	24.84/0.853	36.19/0.938	33.76/0.894	29	9.8	0.20
IFRNet-S [112]	35.59/0.979	35.28/0.969	39.96/0.991	35.92/0.979	30.36/0.936	25.05/0.858	35.87/0.936	33.80/0.891	25	2.8	0.12
IFRNet-B [112]	35.80/0.979	35.29/0.969	40.03/0.991	35.94/0.979	30.41/0.936	25.05/0.859	36.00/0.936	33.99/0.893	30	5.0	0.21
AMT-S(本章算法)	**35.97/0.983**	**35.35/0.971**	39.95/**0.994**	**35.98/0.983**	**30.60/0.940**	**25.30/0.865**	36.11/0.940	**34.29/0.901**	51	3.0	0.12
ToFlow [62]	33.73/0.968	34.58/0.967	39.08/0.989	34.39/0.974	28.44/0.918	23.39/0.831	33.93/0.922	30.74/0.856	88	1.4	0.62
DAIN [120]	34.71/0.976	34.99/0.968	39.73/0.990	35.46/0.978	30.17/0.934	25.09/0.858	35.95/0.940	33.49/0.895	664	24.0	5.51
CAIN [99]	34.78/0.974	35.00/0.969	39.95/0.990	35.66/0.978	29.93/0.930	24.80/0.851	35.21/0.937	32.56/0.901	71	42.8	1.29
BMBC [150]	35.01/0.976	35.15/0.969	39.90/0.990	35.31/0.977	29.33/0.927	23.92/0.843	32.82/0.928	31.19/0.880	2234	11.0	2.50
ABME [124]	36.22/0.981	35.41/0.970	39.59/0.990	35.77/0.979	30.58/0.937	**25.42/0.864**	**36.53/0.944**	33.73/0.901	560	18.1	1.30
IFRNet-L [112]	36.20/0.981	35.42/0.970	**40.10**/0.991	**36.12**/0.980	30.63/0.937	25.27/0.861	36.21/0.937	34.25/0.895	80	19.7	0.79
AMT-L(本章算法)	**36.35/0.982**	**35.42/0.970**	39.95/**0.991**	36.09/**0.981**	**30.75/0.938**	25.41/0.864	36.27/0.940	**34.49/0.903**	116	12.9	0.58
VFIFormer [102]	36.50/0.982	35.43/0.970	**40.13**/0.991	36.09/0.980	30.67/0.938	25.43/0.864	OOM	OOM	1293	24.1	47.71
EMA-VFI† [127]	36.50/0.980	35.42/0.970	39.58/0.989	35.86/0.979	**30.80**/0.938	**25.59**/0.864	**36.74/0.944**	34.55/**0.906**	211	66.0	0.91
AMT-G(本章算法)	**36.53/0.982**	**35.45/0.970**	39.88/**0.991**	**36.12/0.981**	30.78/ **0.939**	25.43 /**0.865**	36.38/0.941	**34.63**/0.904	250	30.6	2.07

4.3 实验

4.3.1 训练细节

作者在 Vimeo90K [62]训练集上使用 AdamW [216] 优化器，在 2 个 NVIDIA RTX 3090 GPUs 上训练 AMT模型 300 轮。总批量（batch size）大小为 24，学习率衰减遵循余弦衰减方案从 2×10^{-4} 到 2×10^{-5}。作者遵循 IFRNet [112]中采用的数据增强方法，具体包括随机翻转、随机旋转、反转序列顺序和随机裁剪大小为 224×224 的图像块。作者使用预训练好的 LiteFlowNet [217]模型预测的光流作为伪标签来监督中间生成的光流。 λ_{char}、λ_{css} 和 λ_{flow} 分别被设置为1、1和0.002。

4.3.2 基准测试

为了进行更加全面的比较，作者在包含不同运动场景的各种基准上评估作者的 AMT模型。作者选用了 PSNR 和 SSIM [194]来进行比较。本文中使用的主要基准测试数据如下：

Vimeo90K [62]: Vimeo90K 是最近视频插帧文献中最常使用的评估基准。其中包括了3782对分辨率为 448×256 的三元组图像。

UCF101 [214]: UCF101 数据集包含具有各种人类动作的视频。作者采用了 DVF [109]中的测试部分，其中包含379张尺寸为 256×256 的三元组图像。

SNU-FILM [99]: SNU-FILM数据集包含1,240对三元组帧，其宽度范围从368到720，高度范围从384到1280。按照运动幅度，它被分为四部分：简单、中等、困难和极难。

Xiph [215]: Xiph 数据集由 8 个 4K 分辨率的视频片段组成，最初由 Niklaus 等人提出 [107]。按照他们最初的评估设置，作者对该数据集进行了改造，包括通过缩小原始帧获得的“2K”版本，以及通过中心裁剪 2K 补丁生成的“4K”版本。

除了这些数据集之外，作者还提供了多帧插值的比较。比如使用了 Adobe240fps数据集 [218]，该数据集最初用于视频去模糊，现在也广泛用于视频插帧。它由高帧率视频（240 fps）组成，分辨率为1280 × 720，视频来自于118个视频剪辑中。

4.3.3 架构细节

作者构建了三种不同大小的模型，称为 AMT-S、AMT-L 和 AMT-G。为了

可以更好的重现，它们的架构细节分别展示在 图 4.9、图 4.10和图 4.11中。作者采用标准的残差模块 [219]和实例归一化（instance normalization） [220]应用于相关性编码器。查找半径设置为3。对于每个更新块，在较高级别（即 $l > 1$）的每个块后面都会跟随一个双线性上采样层。 IFRBlock表示 IFRNet [112]中提出的解码器，它共同估计双向光流和中间特征。为了进一步提高性能，在AMT-G中，作者对相关性特征进行上采样，以使其空间分辨率与当前插值特征对齐，从而便于在高分辨率空间中进行更新。

4.3.4 与先进方法的比较

本文将提出的 AMT方法与目前最先进（state-of-the-art，SOTA）的方法进行比较，包括 ToFlow [62]、DAIN [120]、CAIN [99]、AdaCoF [65]、BMBC [150]、RIFE [117]、ABME [124]、M2M-VFI [108]、IFRNet [112]、VFIFormer [102] 和 EMA-VFI [127]。作者利用 IFRNet [112]提供的代码进行基准测试。推理速度是指方法在 NVIDIA RTX 3090 GPU 上以 1280×720 分辨率进行 1000 次迭代的平均运行时间。为了确保公平比较，作者根据理论计算复杂度将 SOTA 方法分为三类。然后，作者为每组设计了一个对应复杂度的模型，称为 AMT-S、AMT-L 和 AMT-G。

定量比较： 如表 4.1所示，本文中的小模型 AMT-S在几乎所有基准测试中都取得了高效视频插帧方法中的最佳结果，特别是对于具有挑战性的设置。特别的，相比较之前先进的高效视频插帧的方法 IFRNet-B [112]，本文提出的 AMT-S在 Vimeo90K数据集上比其高0.17dB，并且参数量和 FLOPs仅有其60% 。这种差距在 SNU-FILM 的 Hard 和 Extreme 分区上变得更加明显，揭示了本文中 AMT模型在建模大运动方面的强大能力。对于大规模设置，与之前的 SOTA 方法 IFRNet-L [112]相比，本文的 AMT-L 显示出极具竞争力的结果，并且参数量为其65%，FLOPs为其75%。就推理速度而言，作者的方法与IFRNet相当。另外，与目前先进的基于 Transformer模型 (如 VFIFormer [102]和 EMA-VFI [127])相比，作者基于卷积的模型在准确率和效率方面也仍然具有优势。具体来说，作者的 AMT-G 在大多数情况下都优于它们，特别是在使用 SSIM 指标进行评估时。值得注意的是，本章模型的推理速度比 VFIFormer 快约 5 倍，而且参数量仅为 EMA-VFI 的一半。此外，VFIFormer 需要进行两阶段的分步训练和一共需要花费600次 epoch，而作者的模型只需要 300 个 epoch。对

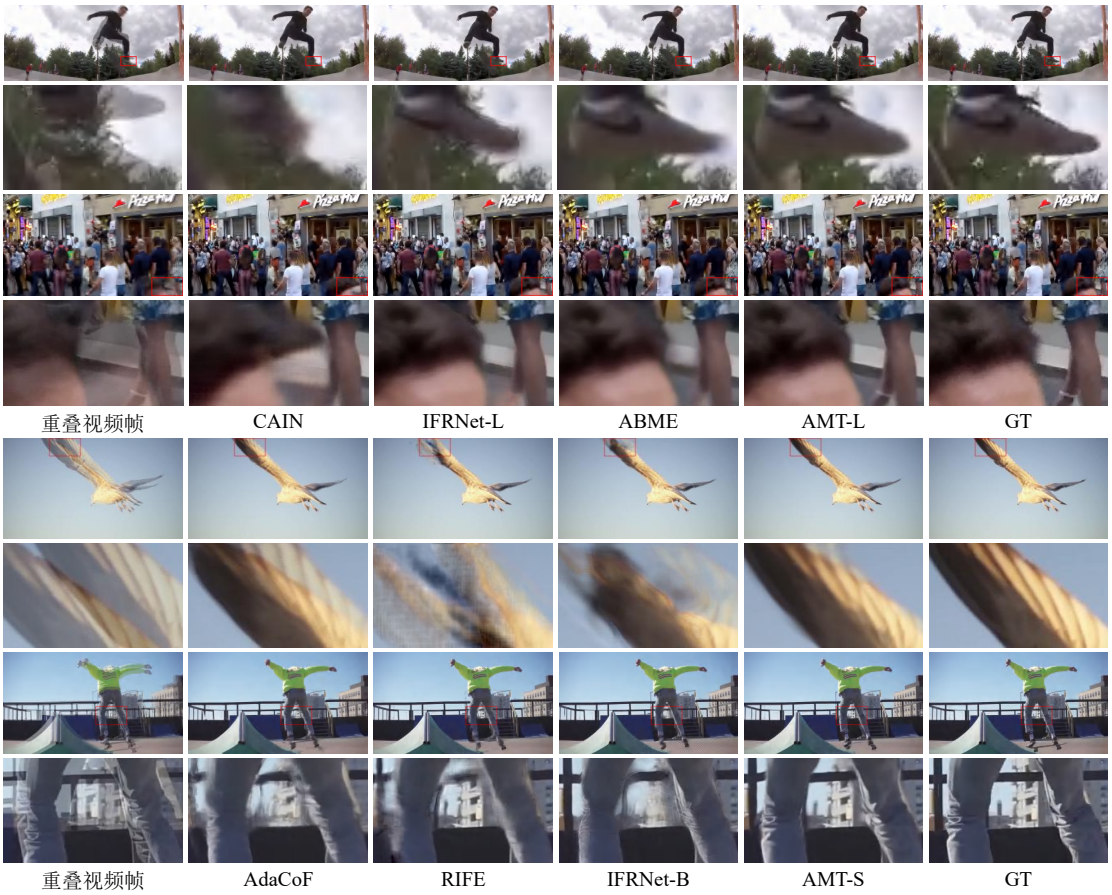

图 4.4: **不同视频插帧方法的定性结果**。作者根据计算成本将这些方法分为两组。作者的 AMT-S和 AMT-L可以精确合成大运动物体的边界，并可以高保真地重建遮挡区域。

于另一个 SOTA模型 EMA-VFI，其在训练期间引入了预热（warm-up）技术来提升性能，而作者的方法并未使用该技术。作者观察到，在将模型规模增加到巨大版本后，除了 Vimeo90K 数据集之外，本章算法的性能已饱和，这可能表明存在过拟合问题。

定性比较： 在图 4.4中，作者选择了多个具有代表性的方法，其中包括基于幻觉、基于内核和基于光流的方法包括：CAIN [99]、 AdaCoF [65]、ABME [124]、 RIFE [117]和 IFRNet(-S/-L) [112]。作者将这些对比算法与本章的AMT算法在 SNU-FILM [99]（Hard）数据集上的进行了视觉比较。为了公平比较，作者还根据计算成本将这些方法在视觉对比时分成两组。可以看出，之前的视频插帧方法在生成移动物体的边缘方面表现不佳。特别是当运动复杂时，容易生成模糊的边缘。由于作者充分考虑了视频插帧任务驱动的光流，作者的

图 4.5: **多帧生成效果对比**。作者的 AMT-S和 IFRNet-B [112]在 Adobe240 [218]数据集上的定性结果。从上到下的时间步长分别为 $1/4$ 和 $1/2$。

AMT算法能够真实地合成运动边界处的内容，并生成具有更少伪影的合理纹理。当背景物体被前景单方面地遮挡时，本文的 AMT算法仍然可以从另一个方向的参考帧中获得指导，而其他方法无法合成这些被遮挡的物体。本文还提供了更多视觉比较的结果，涵盖了两个基准数据集，包括 Vimeo90K [62]和 SNU-FILM [99]，以进一步展示所提出的 AMT的优越性。如图 4.8、图 4.13-图 4.18所示，作者的 AMT能够更忠实地合成具有大运动的对象，并生成更少伪影的逼真纹理。

多帧插帧：对于多帧插值设置，作者使用 GoPro数据集进行训练，并在 GoPro数据集的测试部分和 Adobe240数据集上进行模型评估。与之前的对比不一样的是，多帧插帧评测的目标是进行 $8\times$ 插值，即使用两个输入帧生成7个中间帧。其他训练设置和损失函数与作者主要论文中的设置一致。与最近的帧插值工作（如 [112, 117]）类似，作者为 $8\times$ 插值将一个时间嵌入向量注入到网络中。该向量中的元素都根据当前时间步骤设置为 t，其中 $t \in 1/8, 2/8, ..., 7/8$。作者将作者的 AMT-S与 DVF [109]、SuperSloMo [64]、DAIN [120]和 IFRNet-B [112]进行了比较。 $8\times$ 插值的结果如表 4.2所示。作者的方法在两个评估数据集上均获得了最佳的 PSNR和 SSIM结果，表明所提出的 AMT对多帧插值任务的有效性。

图 4.5 和 图 4.12 在 Adobe240数据集上对作者的方法和 IFRNet-B进行了可

表 4.2: **多帧插值（×8）的定量比较。**

方法	GoPro [221]		Adobe240 [218]	
	PSNR	SSIM	PSNR	SSIM
DVF [109]	21.94	0.776	28.23	0.896
SuperSloMo [64]	28.52	0.891	30.66	0.931
DAIN [120]	29.00	0.910	29.50	0.910
IFRNet-B [112]	29.97	0.922	31.93	0.936
AMT-S (本章方法)	**30.20**	**0.927**	**32.04**	**0.938**

视化比较。在这里，作者展示了时间步长为 1/4 和 1/2 的情况。可以看到，作者的方法可以生成更具时间一致性的结果，具有较少的伪影和更清晰的边缘。

表 4.3: **相关性匹配代价的设计。** 作者分别去除了相关性编码器, 建立一个单向的相关性匹配代价, 并建立了一个类 PWC-Net [222]的相关性匹配代价用于消融实验。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

模型变体	Vimeo	Hard	Extreme
去除相关性编码器	35.76	30.49	25.22
单向相关性匹配代价	35.93	30.34	25.18
类 PWC的局部相关性匹配代价	35.61	30.48	25.16
本章算法	**35.97**	**30.60**	**25.30**

4.3.5 消融实验

本文进行了消融实验，以验证本文提出的 AMT中两个关键组成部分（即全对相关性和多场特征细化）的有效性。所有的消融版本都基于 AMT-S进行，然后在 Vimeo90K [62]数据集和 SNU-FILM的 Hard和 Extreme分区 [99]上进行了评估。

4.3.6 关于全对相关性的消融实验

本节验证了本文提出的 AMT中全对相关性的有效性。所有的消融版本都基于 AMT-S进行，然后在 Vimeo90K [62]数据集和 SNU-FILM的 Hard和 Extreme分区 [99]上进行了评估。

体积设计： 正如表 4.3所示，本文的双向相关性匹配代价在视觉帧插值任务中充分考虑了输入帧之间的对应关系，从而比单向方法表现更好。此外，使用

表 4.4: **查找策略**。作者探究了初始化网格，类 RAFT [147]查找和本文提出的放缩查找变体。作者还探究了是否使用双边流来执行初始查找。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

查询策略	初始化策略	Vimeo	Hard	Extreme
初始二维网格		35.92	30.52	25.23
RAFT中查询策略	光流初始化	35.93	30.34	25.18
本章中放缩查询	零初始化	35.97	30.56	25.26
本章中放缩查询	光流初始化	**35.97**	**30.60**	**25.30**

表 4.5: **内容更新**。 本文分别通过使用可见帧的特征作为指导和丢弃内容更新来探究内容更新的影响。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

模型变体	Vimeo90K [62]	SNU-FILM [99]	
		Hard	Extreme
可见帧特征作为指导	35.95	30.53	25.21
移除内容更新	35.96	30.52	25.22
本章算法	**35.97**	**30.60**	**25.30**

独立的编码器（即相关性编码器）来构建相关性体积是必要的。作者可以观察到，当作者利用内容编码器的特征来构建相关性体积时，性能显著下降。作者还尝试了按照 PWC-Net [145]的方法构建相关性体积。这个变种的性能比其他任何方法都差，因为它的部分相关性体积限制了对运动信息进行充分建模的能力。

查找策略： 如表 4.4所示，作者可以观察到在使用 RAFT [147]中的基本查找策略时，性能明显下降。对于大的运动，其性能甚至比直接使用初始网格更差，这表明这种策略为光流更新提供了不准确的相关信息。在作者通过缩放对光流进行投影之后，相关性匹配代价和光流共享相同的坐标系，网络可以正确地充分利用查找过程。此外，来自内容编码器的初始流对为进一步查找提供了良好的初始点，这带来了性能增益。

内容更新： 在作者的 AMT 中，每个更新块都会接收中间内容特征作为上下文指导，并随着双边流进行更新。如果作者用可见帧中的特征替换上下文指导，则会引入不明确的信息，导致性能下降，如 表 4.5 所示。此外，作者在每个更新块中只保留一个头，仅更新流场而不更新中间特征，这导致大运动时

表 4.6: **跨尺度更新**。本文探究了不同尺度更新对结果的影响。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

第一层级	第二层级	第三层级	Vimeo90K [62]	SNU-FILM [99]	
				Hard	Extreme
			35.60	30.39	25.06
✔			35.84	30.55	25.19
✔	✔		35.92	30.58	25.28
✔	✔	✔	**35.97**	**30.60**	**25.30**
单一尺度			35.95	30.50	25.22

表 4.7: **光流对数目对模型性能影响**。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

光流对数量	Vimeo90K [62]	SNU-FILM [99]		FLOPs (G)
		Hard	Extreme	
1	35.84	30.52	25.25	116
3	35.97	30.60	25.30	121
5	36.00	**30.63**	**25.33**	127
7	**36.01**	30.57	25.25	135

PSNR 值下降。它表明全对相关性不仅有助于更新光流，而且有助于更新内容。

更新策略： 如 表 4.6 所示，所有跨层更新在作者的跨尺度更新策略中都是有效的。值得注意的是，如果作者丢弃所有更新，相当于没有全对相关性的模型，PSNR值将急剧下降。这证明了作者 AMT 中的 全对相关性的有效性。此外，仅在第一层进行 3× 迭代更新会降低性能。事实表明，跨尺度更新策略可以充分利用逐步细化的内容特征，从而实现更好的运动建模。

4.3.7 关于多场细化的消融实验

本节验证了本文提出的 AMT中多场特征细化的有效性。所有的消融版本都基于 AMT-S进行，然后在 Vimeo90K [62]数据集和 SNU-FILM的 Hard和 Extreme分区 [99]上进行了评估。

光流场的数量： 表 4.7 展示了流场数量对性能增益的影响。作者观察到，仅使用三对光流就可以带来显著的性能增益，这表明确保流场的多样性对于视

表 4.8: **多场融合**。 本文探究了残差模块(公式 4.4)和微调模块(公式 4.6)的作用。作者展示了这些变体的 PSNR 值，最佳的结果以**加粗**显示。默认设置以 灰色 标记。

模型变体	Vimeo90K [62]	SNU-FILM [99]	
		Hard	Extreme
除去残差	35.87	30.57	25.27
除去细化过程	35.89	30.51	25.19
本章算法	**35.97**	**30.60**	**25.30**

重叠视频帧　除去多场微调　加入多场微调　真实结果

图 4.6: **多场特征细化的效果**。 多场特征细化有助于网络更好地恢复遮挡区域。

频插帧任务的使用非常重要。PSNR 值随着场数的增加而上升，直到 7 对时出现饱和。作者在小型模型（即 AMT-S）中使用 3 对以提高效率，并在较大模型中使用 5 对以获得更好的性能。在图 4.6中，作者研究了多场细化对遮挡处理的影响。结果表明，采用多场细化后，作者的 AMT 可以合成被前景遮挡的背景，并具有更一致的纹理。

多场融合： 作者研究了一种变体，它删除了 公式 4.4 中每个候选帧的残差分量，但估计了最终插值结果中的残差部分。如 表 4.8 所示，该变体的结果低于原始设置，这表明作者需要分别补偿每个候选帧的细节。此外，如果作者用平均运算替换 公式 4.6 中的卷积运算符，性能将会下降（参见 表 4.8）。这表明作者的 AMT 进行自适应融合和细化非常重要。

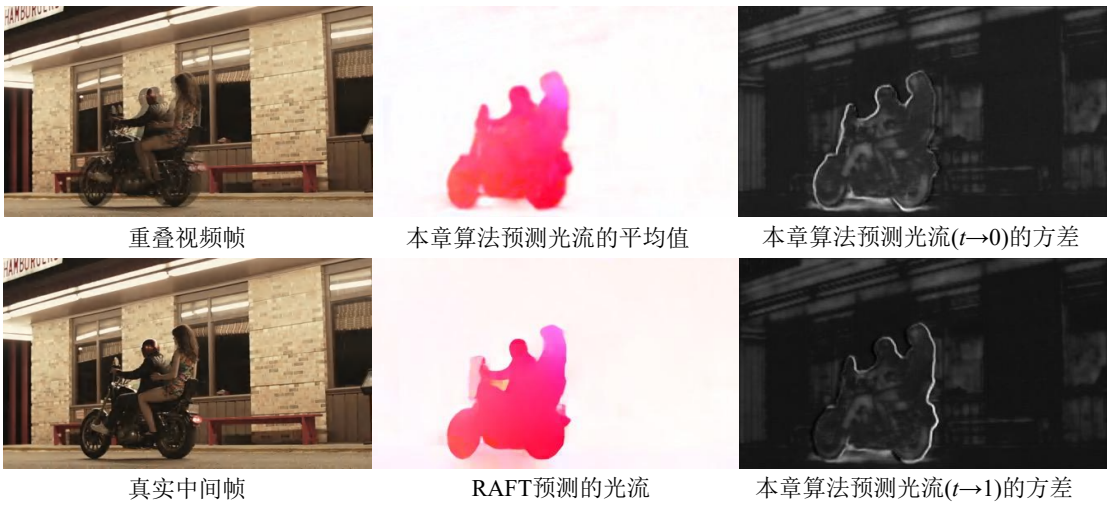

图 4.7: **多场微调模块估计出的光流对比**。多场微调模块使用三组光流对的平均图和方差图的可视化。作者提供 RAFT [147]的光流作为参考。

讨论：为了进一步讨论，作者将三个估计流对的平均值和偏差可视化。结果显示在图 4.7中。一方面，作者的平均流量与 RAFT [147] 估计的流量基本一致，近似于地面真实位移。另一方面，作者观察到流的多样性主要位于运动边界和纹理丰富的区域。这表明这些区域需要涉及更多潜在的候选像素来进行重建。通过这些可视化，可以看到本文的方法生成了非常有前景的面向任务的光流：通常与地面真实光流一致，但局部细节更加多样。

4.4 讨论

4.4.1 与 RAFT相比较进行讨论

Teed和 Deng在 [147]中提出了 RAFT, 该方法通过多尺度4D相关性体积进行迭代查找以更新光流场。鉴于其卓越的结果，当前最先进的光流估计方法 [148, 149, 209, 223, 224]都源自这种架构设计。此外，它还启发了立体匹配 [225]和场景流 [226]的发展。然而，RAFT类似的设计范式在帧插值领域并未得到充分研究。

为了更好地模拟帧插值中的大运动，作者基于 RAFT构建了 AMT。然而，与 RAFT不同，AMT涉及许多新颖的和针对特定任务的设计。为了更好地说明作者的模型，作者将从以下几个角度详细说明作者的 AMT与 RAFT之间的区别：

表 4.9: 对 RAFT样式 [147]设计进行的研究。本章算法最终的设置用 灰色 标记。

方案	Vimeo90K [62]	SNU-FILM [99]		FLOPs (G)
		Hard	Extreme	
单尺度预测	35.94	30.52	25.26	124
ConvGRU	35.99	30.58	25.27	132
绑定权重	35.93	30.56	25.22	121
凸上采样	35.99	30.56	25.28	123
原始模型	35.97	30.60	25.30	121

相关体积设计： RAFT构建了单向相关性体积，因为它只需要预测沿一个方向的光流。对于视频插帧任务，作者希望在两个方向上建模密集的对应关系，以更新双向光流。因此，作者构建了双向相关性体积。在论文的表 4.3中，作者验证了双向相关性体积的有效性。

内容编码器： 在 RAFT中，内容编码器仅从第一个输入帧中提取内容特征。由于视频插帧的特性，受到最近的单阶段视频插帧方法的启发，在作者的AMT中，内容编码器将图像对作为输入。它输出初始中间特征，初始双向光流以及来自输入图像对的金字塔特征。这个设计也受到最近的单阶段视频插帧方法的启发 [112, 113]。

相关性查找： 在 RAFT中，可以直接执行查找操作，因为相关性体积和预测的光流场具有相同的坐标系统。为了解决由于不可见帧引起的坐标不匹配问题，作者建议在查找操作之前对双向光流进行缩放。此外，作者检索双向相关性，而不是 RAFT中的单向相关性。作者使用初始双向光流 $(F^1_{t\to 0}, F^1_{t\to 1})$ 作为初始起点，而 RAFT使用零值。查找策略在表 4.4中进行了研究。

预测和更新方式： RAFT在单一分辨率上预测和更新流预测，而作者以粗到精的方式预测和更新双向光流。作者还提供了作者 AMT的一种变体来验证设计，该变体仅在输入最后一个解码器之前预测单一分辨率的流场。 表 4.9显示，这种变体的性能低于原始方法。这表明，预测多尺度光流对于帧插值非常重要。此外，作者还在表 4.6中研究了跨尺度更新的有效性。

更新块： 在更新块的设计中，作者的 AMT与 RAFT在五个方面有所不同：1）RAFT将从可见帧提取的特征视为内容指导，而作者使用表示不可见帧的插值中间特征； 2）RAFT在更新块中只有一个头部用于回归流残差，而作者有两

个头部用于联合预测内容和光流残差。前面两个方面已在主要论文的表2c中讨论过； 3）作者堆叠了两个卷积层，而不是 RAFT中繁琐的 ConvGRU单元，用于处理内容和动态特征。作者还调查了一种在每个更新块中配备 ConvGRU单元的变体。如表 4.9所示，与原始模型相比，这种变体表现出了较好的性能，但计算成本更高。因此，为了效率起见，作者选择堆叠两个卷积层； 4）在AMT中，更新块的权重在不同层级之间不共享。然而，权重绑定对 RAFT是有益的。 表 4.9表明，没有绑定权重的模型性能优于绑定权重的模型； 5）作者在 AMT中使用双线性上采样，而不是 RAFT中的凸起采样来放大光流场。如表 4.9所示，这两种上采样操作具有类似的性能，但是凸起采样会增加更多的计算成本。因此，作者选择在 AMT中使用双线性上采样。

最终目标： RAFT是为了光流估计而设计的，仅使用光流回归损失进行优化。然而，作者的 AMT是为帧插值而引入的，并受到面向任务的光流蒸馏损失和面向失真的内容损失的监督。作者需要考虑的不仅是估计的光流的保真度，还包括满足任务导向流的要求的多样性。因此，作者输出多个光流对，而不是RAFT中的单一光流场。此外，还需要考虑遮挡推理和残留幻觉，以生成更加真实的内容。

4.4.2 关于多场微调的讨论

一些研究也尝试预测多个光流对以准备中间内容候选项。具体来说，BMBC [150]通过双向运动网络和光流逼近预测了六个双边运动。ABME [124]基于非对称运动假设生成了四个双向光流场。在获取扭曲的候选帧和上下文特征后，这两个方法都依赖于动态滤波器甚至繁琐的综合网络来生成最终的中间帧。因此，它们在实际应用中效率较低。相比之下，作者的 AMT更加高效，如论文中的表1所示。作者在单次前向传递中生成多个光流场，而不是像 BMBC和ABME那样需要多个推理步骤。此外，作者仅在图像域中获得中间候选项，而不是在特征域中，并堆叠了两个轻量级卷积层来融合这些候选项。

M2M-VFI [108]与作者的多场微调最相关。它也可以在一步中生成多个光流并在图像域中准备扭曲的候选帧。然而，作者的多场景精炼与 M2M-VFI之间存在五个关键差异。首先，作者的方法通过反向扭曲生成候选帧，而 M2M-VFI中是通过前向扭曲生成的；其次，M2M-VFI预测多个光流以克服前向扭曲引起的重叠区域的空洞问题和伪影问题，但作者的目标是通过增强光流的多样性来减

轻遮挡区域和运动边界的歧义问题；第三，M2M-VFI需要首先通过一个现成的光流估计器估计双向光流，然后通过一个运动精化网络来预测多个双向光流。相反，作者直接在一个单阶段网络中估计多个双向光流。在这个网络中，作者首先在粗粒度上估计一对双向光流，然后从粗粒度光流对中导出多组细粒度的双向光流；第四，M2M-VFI一起估计了两个可靠性图和所有双向光流对，这些图可以进一步用于融合前向扭曲引起的重叠像素。如公式 4.5所示，作者不仅估计遮挡权重图，还估计了与每一对双向光流合作的残余内容。残余内容用于补偿扭曲后不可靠的细节。这一设计已在表 4.8中进行了研究；第五，作者堆叠了两个卷积层以自适应地合并候选帧，而 M2M-VFI则通过预先计算的权重图来规范化所有候选帧的总和。

4.5 总结

遵循面向任务光流的运动特征，我们引入了全对多场变换（名为 AMT）来实现高效的帧插值。它包含两个基本设计，即全对相关和多场细化。通过这两种设计，我们的方法在帧插值过程中能够有效地处理大运动和遮挡区域，并在多个基准测试中以高效率实现了最先进的性能。

局限性：尽管作者的方法在性能上表现出色，但由于从所有像素对计算的4D相关性体积，使其难以适应在资源受限的环境下进行非常高分辨率的输入。这是因为构建相关性体积的计算复杂度与图像分辨率的平方成正比。缓解这个问题的可能方法包括只在查找时计算每个相关性值 [147]或将4D相关性体积分解为两个3D相关性体积 [224]。

更广泛的影响：正如本文所呈现的，作者的 AMT可以合成两个可见帧之间较为真实的不存在的帧。鉴于其可靠的合成结果，作者的方法可能被滥用来伪造或篡改视频。

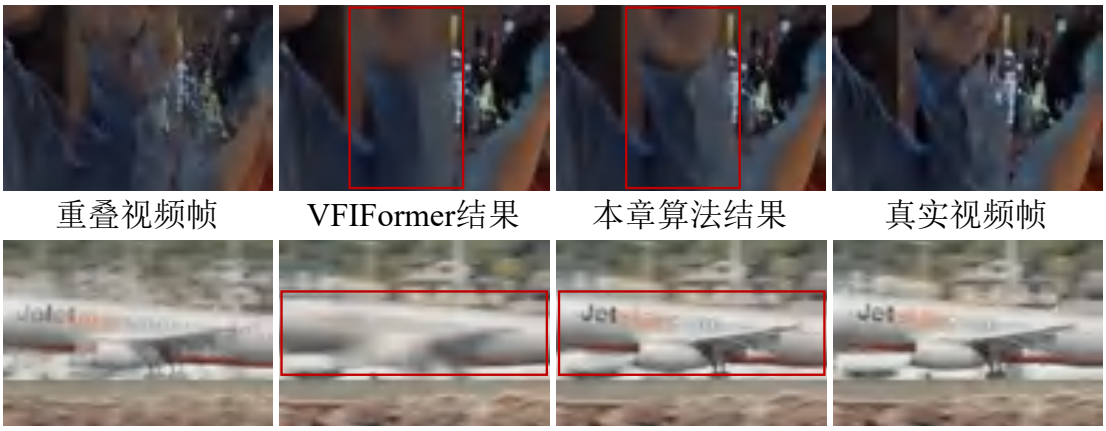

图 4.8: AMT-G与 VFIFormer的定性比较。作者的方法恢复了更清晰的结构和边缘。

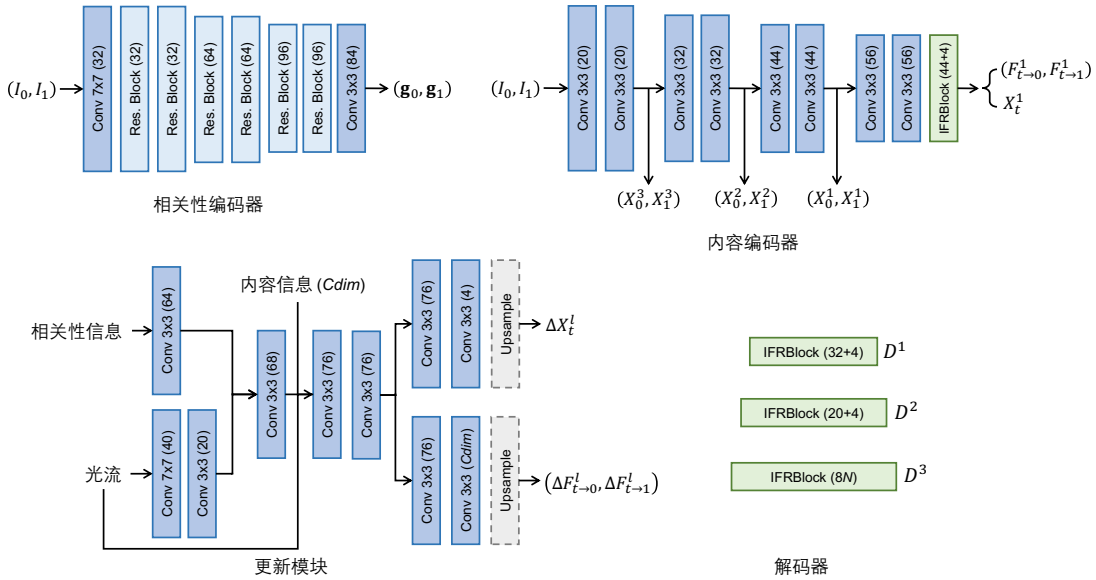

图 4.9: AMT-S的架构细节。括号中的数字表示输出通道数。 N 代表输出组的数量。 IFRBlock表示 IFRNet [112]中提出的解码器。Conv表示卷积层，Res. Block表示残差模块 [219]（Residual Block）。

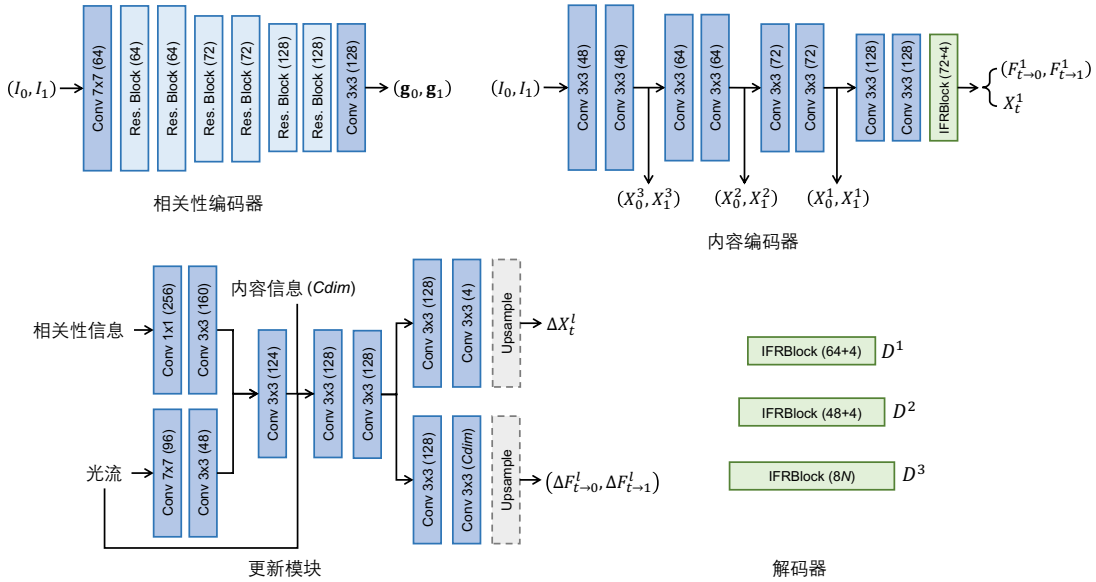

图 4.10: AMT-L的架构细节。括号中的数字表示输出通道数。 N 代表输出组的数量。 IFRBlock表示 IFRNet [112]中提出的解码器。Conv表示卷积层，Res. Block表示残差模块 [219]（Residual Block）。

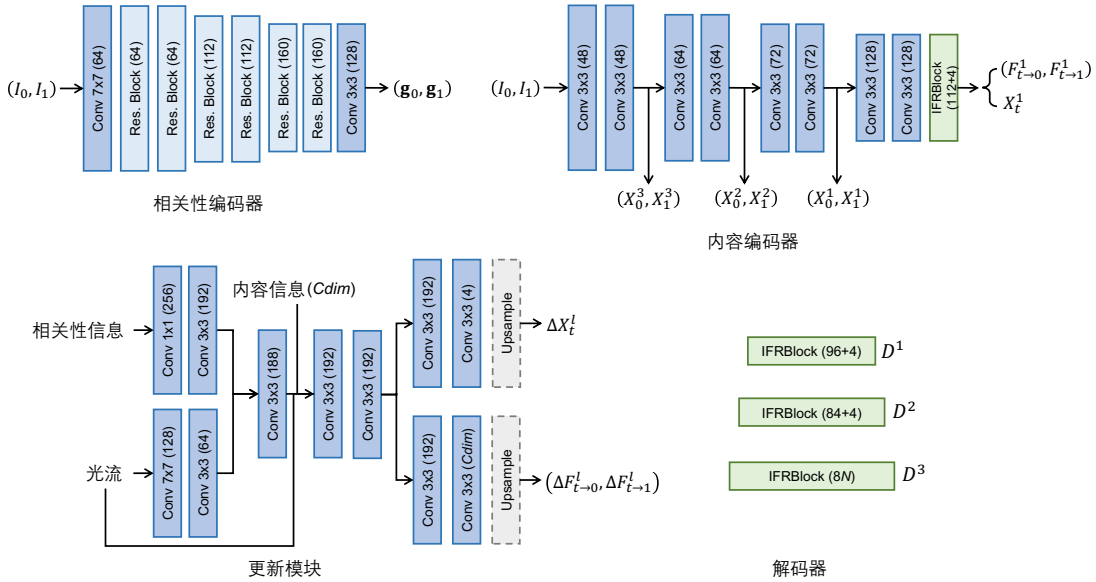

图 4.11: AMT-G的架构细节。括号中的数字表示输出通道数。 N 代表输出组的数量。 IFRBlock表示 IFRNet [112]中提出的解码器。Conv表示卷积层，Res. Block表示残差模块 [219]（Residual Block）。

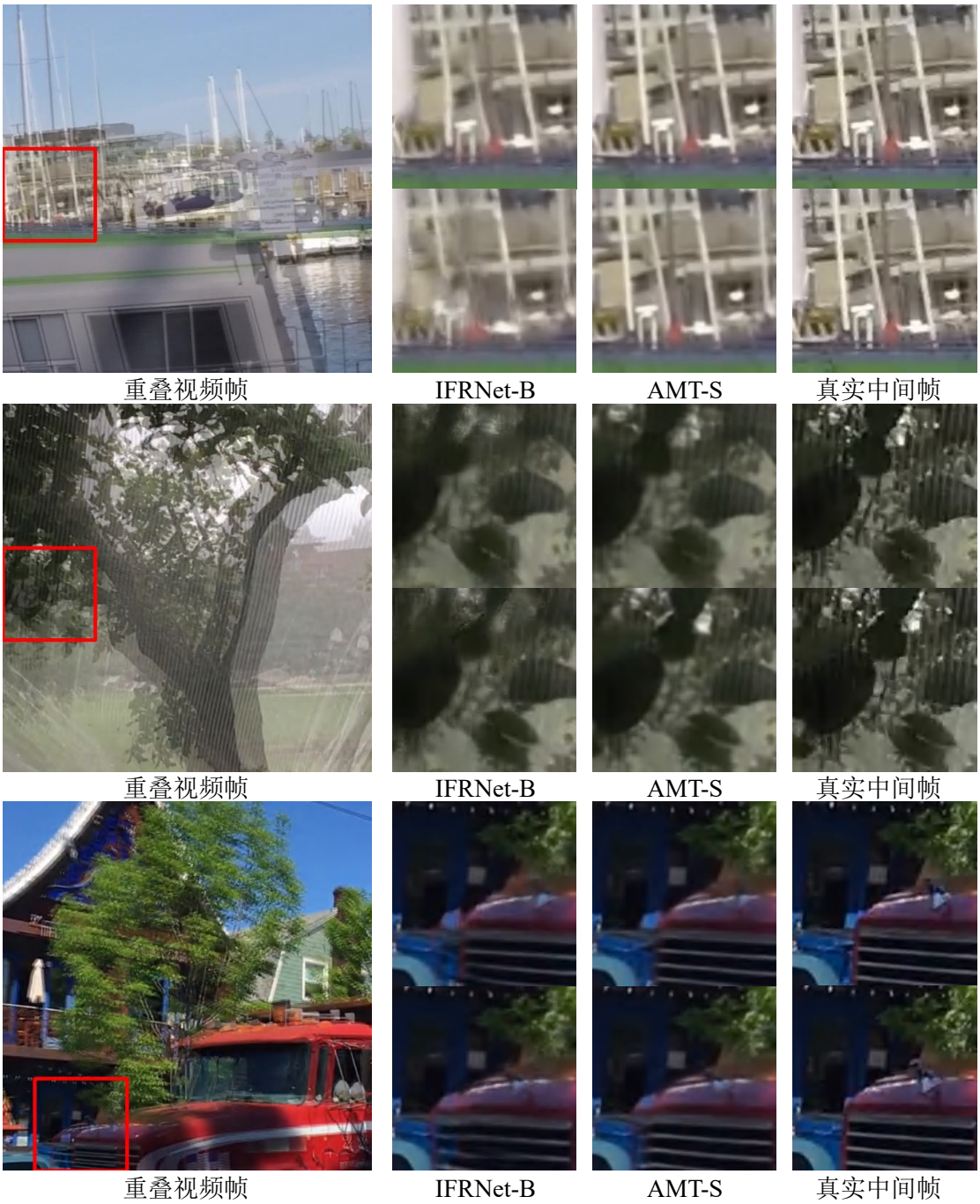

图 4.12: 在 Adobe240 [218]上的 AMT-S和 IFRNet-B [112]的定性结果。从上到下的时间步长分别是 1/4 和 1/2。

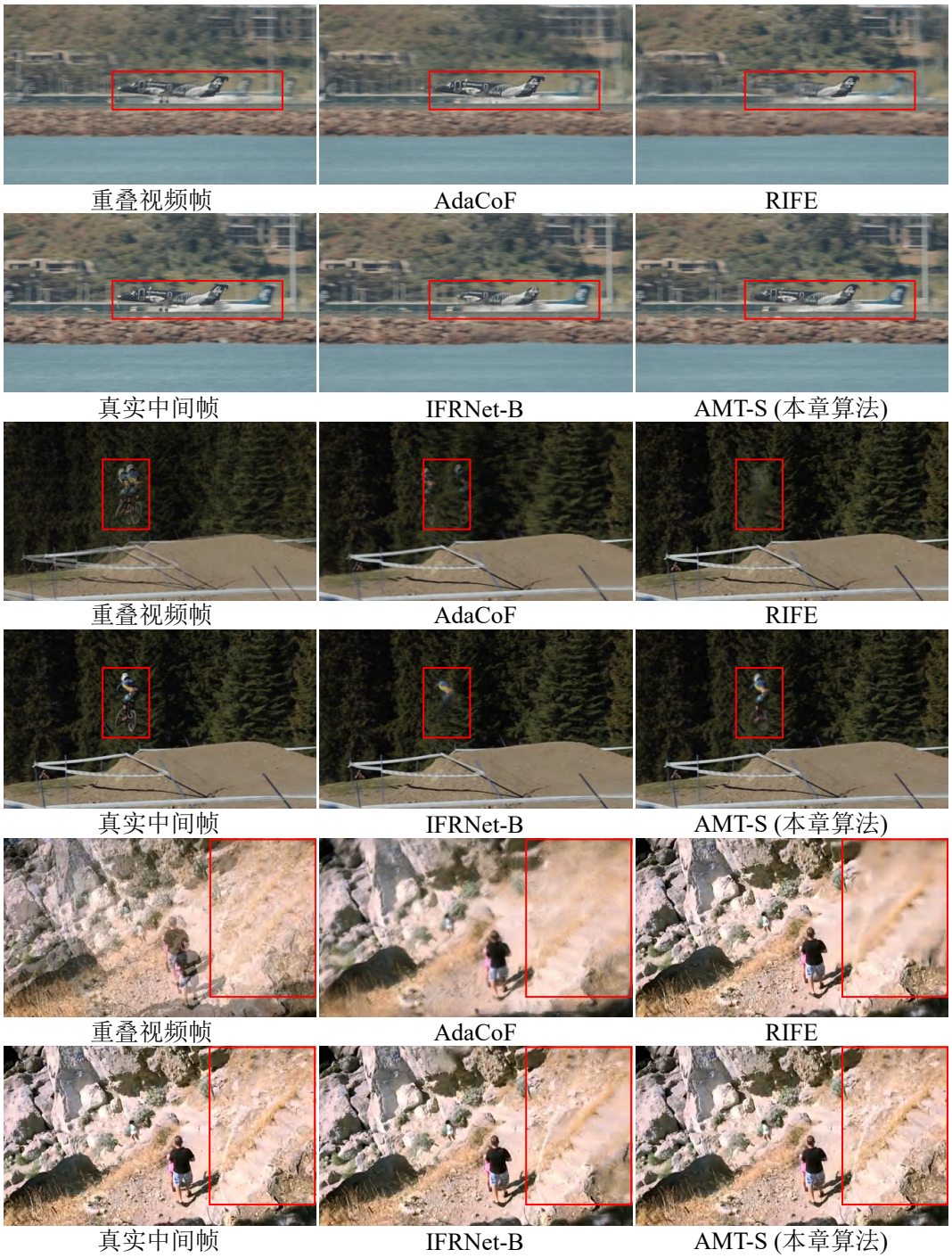

图 4.13: 在 Vimeo90K数据集 [62]上，对低计算复杂性方法进行的视觉比较。

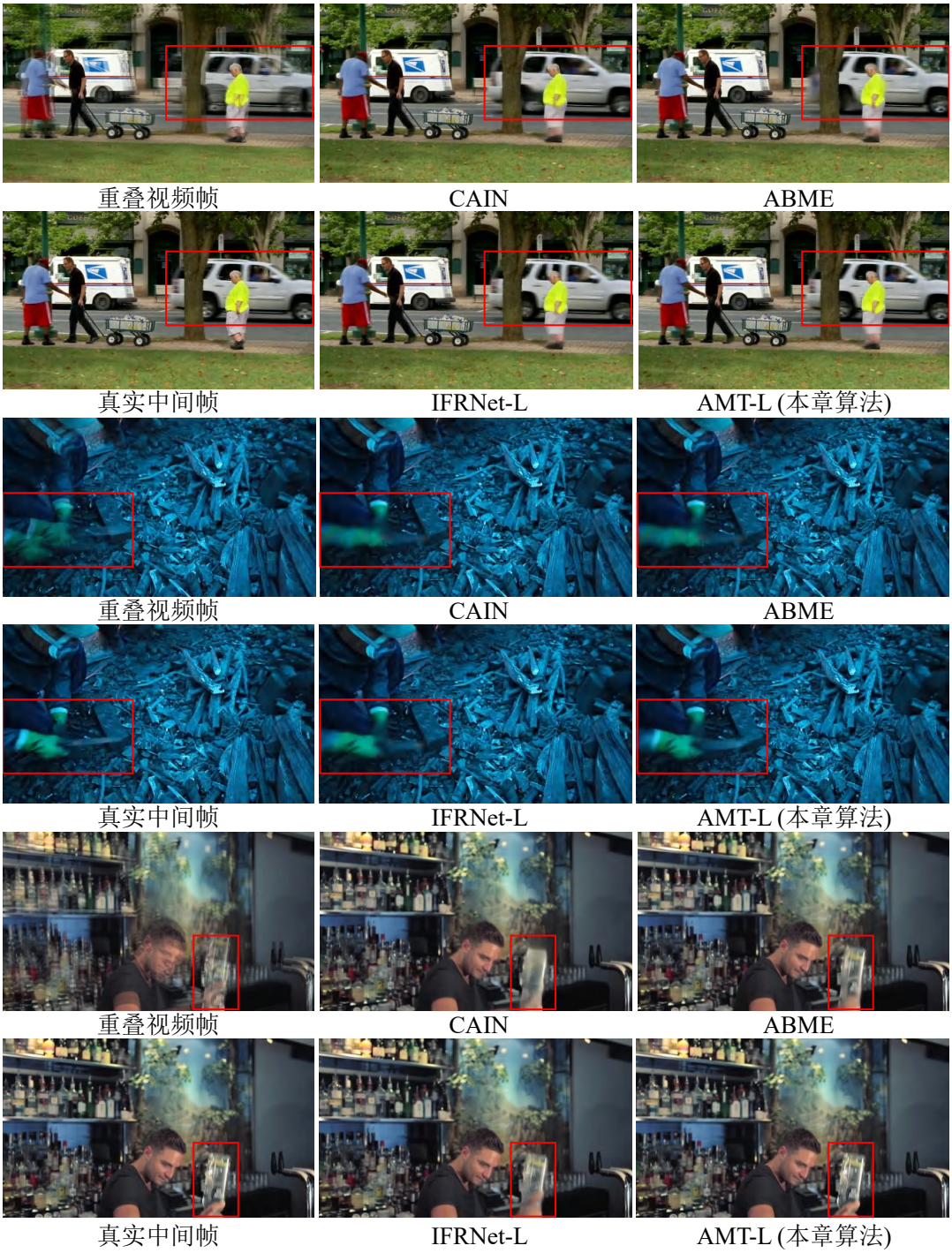

图 4.14: 在 Vimeo90K数据集 [62]上，对高计算复杂性方法进行的视觉比较。

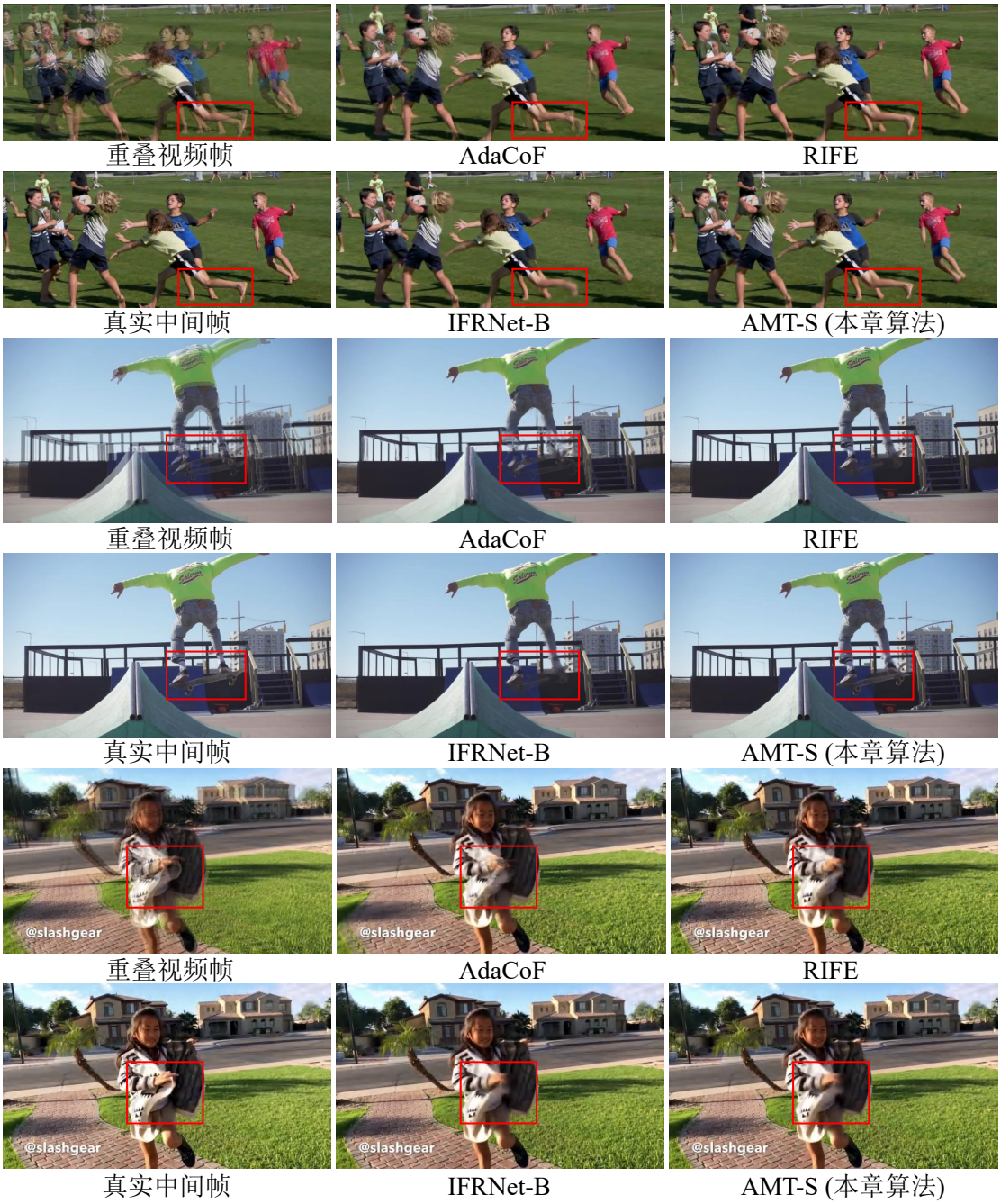

图 4.15: 在 SNU-FILM数据集 [99]的 Hard分区上，对低计算复杂性方法进行的视觉比较。

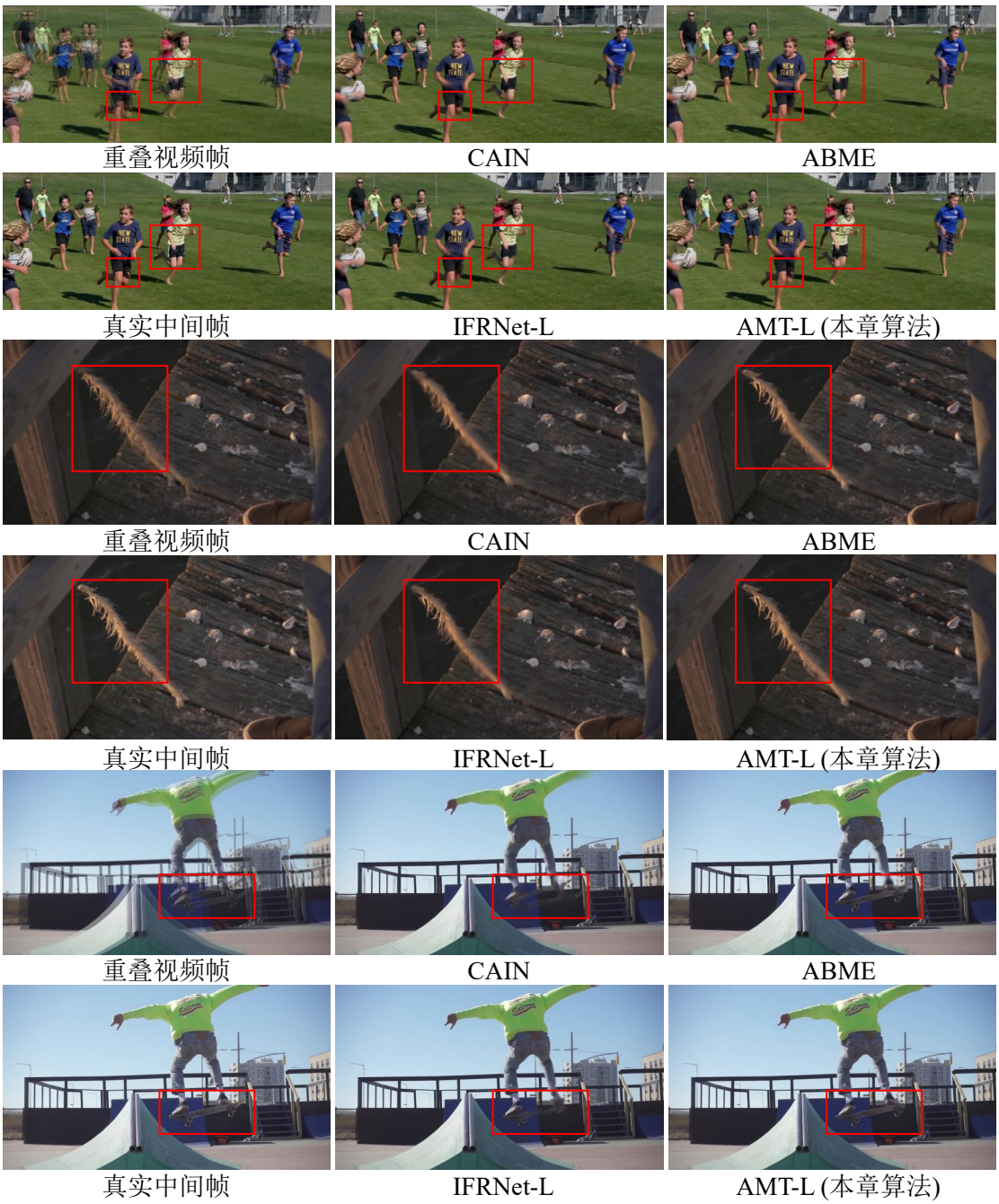

图 4.16: 在 SNU-FILM数据集 [99]的 Hard分区上，对高计算复杂性方法进行的视觉比较。

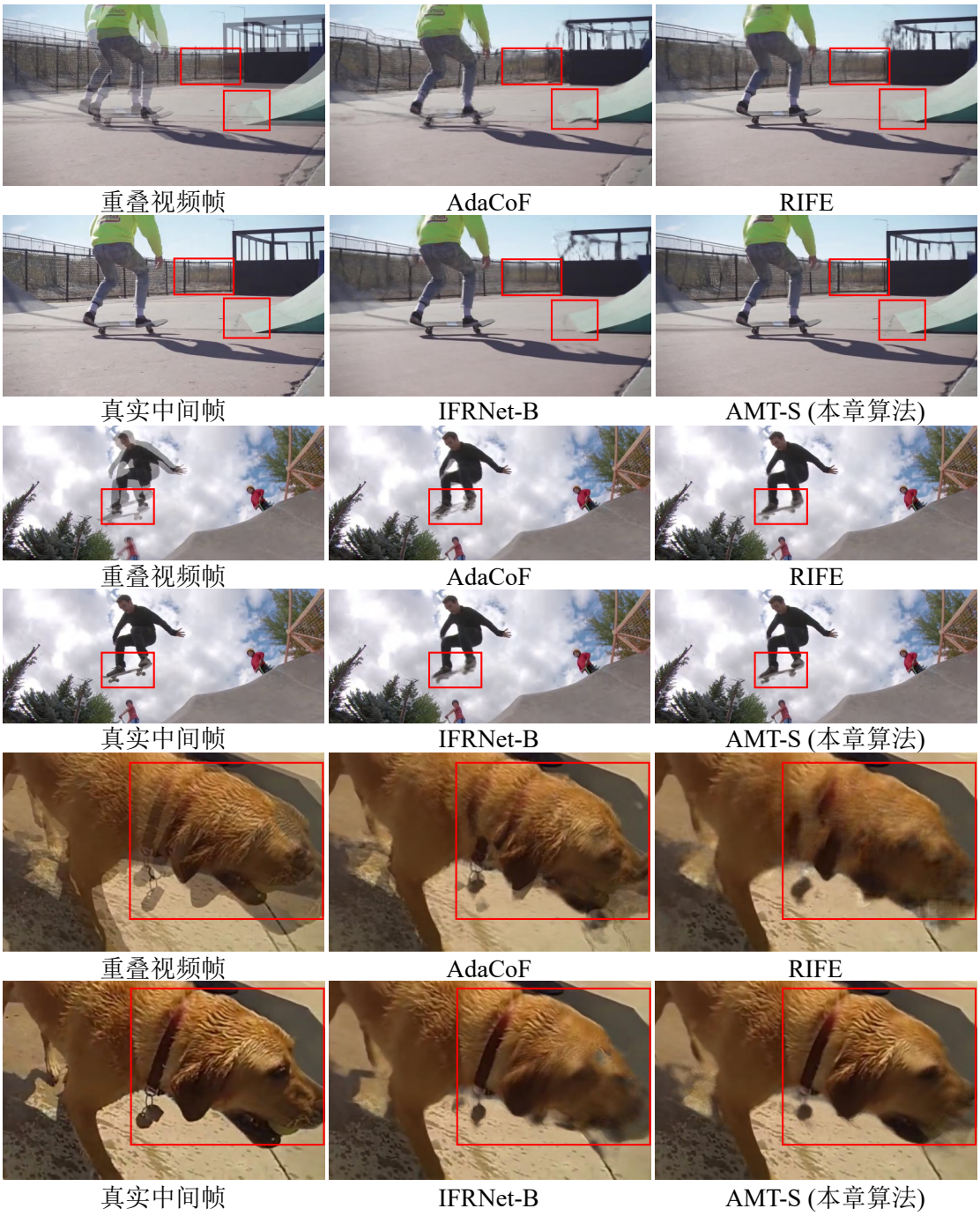

图 4.17: 在 SNU-FILM数据集 [99]的 Extreme分区上，对低计算复杂性方法进行的视觉比较。

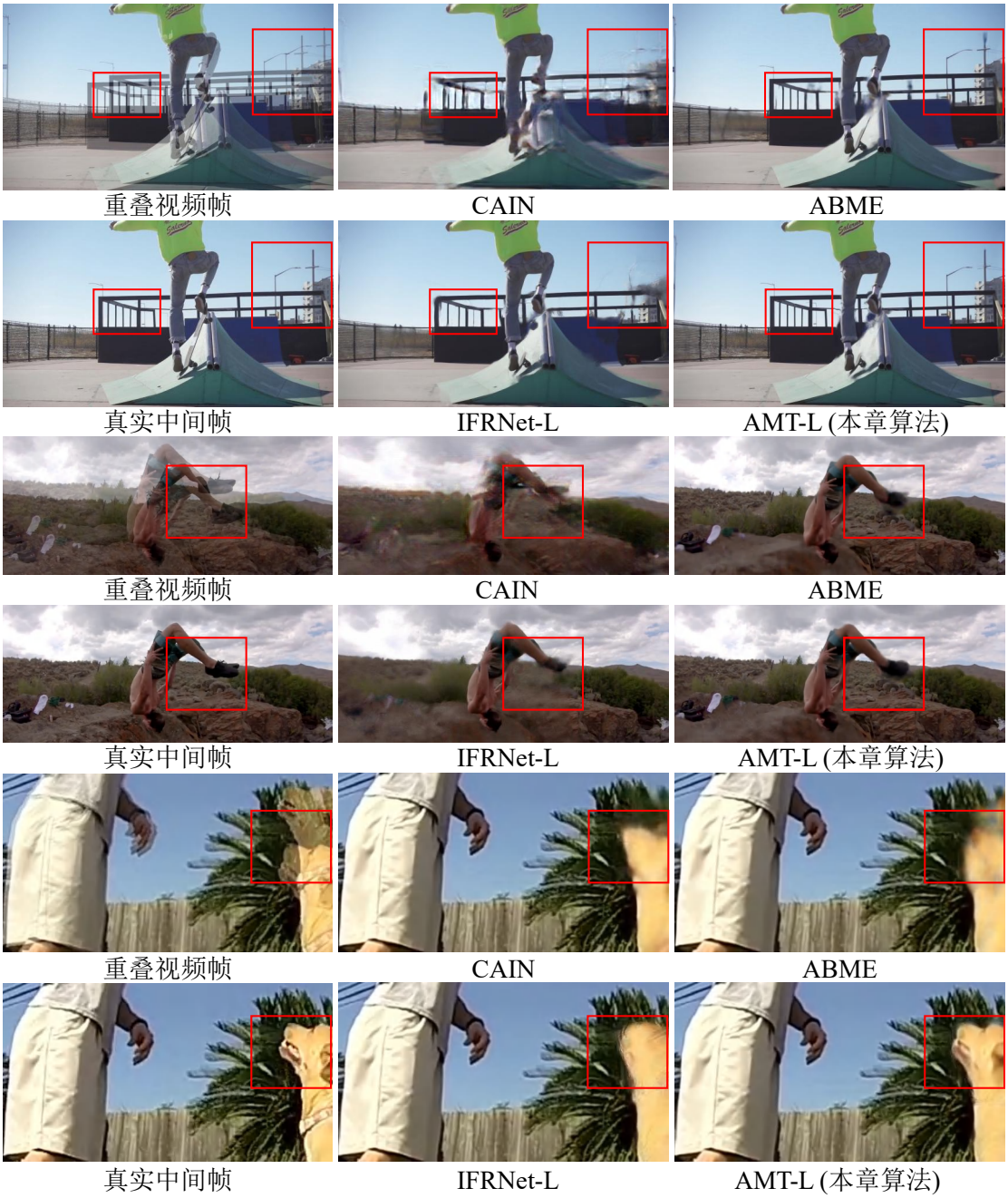

图 4.18: 在 SNU-FILM数据集 [99]的 Extreme分区上，对高计算复杂性方法进行的视觉比较。

第五章 身份属性驱动的人物图像个性化生成

本章主要研究身份属性驱动的人物图像个性化生成。章节 5.1中介绍研究背景、动机及解决方案概要； 章节 5.2介绍本章构建的身份属性驱动的人物图像个性化生成框架以及以 ID为导向的数据集收集流程； 章节 5.3给出评测结果和结果分析； 章节 5.4对本章进行小结。

5.1 引言

5.1.1 研究背景

人像相关的定制图像生成 [174, 227, 228]在近年受到了相当大的关注，与此同时也催生了许多应用，如个性化肖像照片 [229]、图像动画 [230]和虚拟试穿 [231]等。早期的方法 [232, 233]因受到生成模型（如生成对抗网络 [1, 5]）能力的限制，只能定制化地生成面部区域，进而导致生成的图像具有较低的生成多样性、场景丰富性和可控性。得益于更大规模的文本-图像配对训练数据集 [155]、更大的生成模型 [25, 164]以及可以提供更强语义嵌入的文本/视觉编码器 [158, 160]，基于扩散的文本到图像的生成模型最近一直在不断进化。这种进化使它们能够生成越来越真实的面部细节和丰富的场景。由于文本提示词和结构指导 [26, 162]的存在，可控性也得到了极大的提高。

同时，在愈发强大的基于扩散模型的文本到图像模型的滋养下，为满足用户对高质量定制结果的需求，许多基于扩散模型的定制化生成算法 [165, 166]应运而生。目前，在商业和社区中应用最为广泛使用的当属基于 DreamBooth的方法 [165, 171]。这种类型的应用需要用户上传几十张同一身份（Identity，ID）的图像来微调模型参数。尽管生成的结果具有高 ID保真度，但也有两个明显的缺点。第一是每次用于微调的定制数据都需要用户手动收集，因此非常耗时耗力；第二是定制每个 ID需要10-30分钟，且会消耗大量的计算资源，尤其是当生成模型的体积增长时，这种速度和资源的缺陷会更加明显。因此，为了简化和加速定制生成过程，受现有的以人为中心的数据集 [5, 234]的驱动，最近的工作尝试通过训练视觉编码器 [182, 183]或超网络 [174, 235]来将输入 ID图像表示为送入模型的嵌入或 LoRA [184]权重。经过训练，用户只需要提供一个待定制 ID的图

图 5.1: 给定一些输入 ID的图像，所提出的 PhotoMaker可以在一次前向传递中根据文本提示词生成多样化的个性化 ID图像。PhotoMaker在生成逼真的人像照片时，可以很好地保留输入图像池中的 ID信息。PhotoMaker还支持许多有趣的应用，如（a）改变属性、（b）将艺术作品或旧照片中的人物带入现实或（c）进行身份混合。

像，通过几十步的微调或甚至不需要任何微调过程就可以实现个性化生成。然而，这些方法定制的结果不能像 DreamBooth（见图 5.4）一样同时具备良好的 ID保真度和生成多样性。这是因为： 1）在这些方法的训练过程中，目标图像和输入 ID图像都从同一图像中采样。经此策略训练出的模很型容易记住图像中与 ID无关的特征，如表情和视角等信息，进而导致了可编辑性差； 2）仅依赖于一个要定制的 ID图像会使模型难以从其内部知识中辨别出要生成的 ID的特征，导致 ID保真度不佳。

5.1.2 研究动机与贡献

基于以上两点，受到 DreamBooth成功的启发，本章在设计算法时的目标是： 1）确保输入的 ID图像条件和目标图像在视角、面部表情和配件上具备变化，使模型避免记住与 ID无关的信息； 2）在训练过程中为模型提供多个不同

的图像，以更全面和准确地表征定制 ID的特征。

因此，作者提出了一个简单而有效的前馈定制人像生成框架，可以接收多个输入 ID图像，称为*PhotoMaker*。为了更好地表示每个输入图像的 ID信息，本章算法在语义级别堆叠多个输入 ID图像的编码，构造出一个堆叠的 ID嵌入。这个嵌入可以被视为要生成的 ID的统一表示，该嵌入中每个子部分对应一个输入 ID图像。为了更好地将这个 ID表示和文本嵌入集成到网络中，作者用堆叠的 ID嵌入替换了文本嵌入的类别词（例如，man和 woman）。得到的嵌入不仅表示了要定制的 ID信息还表示了要生成的上下文信息。通过这种设计，本章算法无需在网络中添加额外的模块，生成模型本身的交叉注意层就可以自适应地集成堆叠 ID嵌入中包含的 ID信息。

与此同时，堆叠的 ID嵌入允许作者在推理时接受任意数量的 ID图像作为输入，同时保持像其他无需微调的方法 [178, 183]那样的生成效率。具体来说，本章提出的方法在接收四个 ID图像时需要大约10秒来生成一个定制的人像照片，这比 DreamBooth类的需要微调的方法快大约130倍[1]。此外，由于本章提出的堆叠 ID嵌入可以更全面和准确地表示定制的 ID，本章提出的方法可以提供比最先进的无需微调的方法更好的 ID保真度和生成多样性。与以前的方法相比，本章提出的框架在可控性方面也有了很大的提高。它不仅可以进行常见的配合文本描述进行上下文生成、还可以改变输入人类图像的属性（例如，配件和表情）、生成与输入 ID完全不同视点的人像照片甚至修改输入 ID的性别和年龄（见图 5.1）。

值得注意的是，本章提出的 PhotoMaker也为用户生成定制的人像照片释放了很多可能性。具体来说，虽然在训练期间构建堆叠 ID嵌入的图像来自同一 ID，但 PhotoMaker可以在推理期间使用不同的 ID图像来形成堆叠 ID嵌入，以合并和创建一个新的定制 ID。合并的新 ID可以保留不同输入 ID的特征。例如，算法可以生成看起来像 Elun Musk的 Scarlett Johansson，或者一个混合了一个人和一个知名 IP角色的定制 ID（见图 5.1(c)）。同时，通过提示词权重 [236, 237]（Prompt Weighting）或改变输入图像池中不同 ID图像的比例，就可以简单地调整融合比例。这种性能展示了 PhotoMaker的灵活性。

本章提出的 PhotoMaker在训练过程中需要同时输入多个具有相同 ID的图像，因此需要 ID导向的人像数据集的支持。然而，现有的数据集要么不按 ID分

[1]在一台 NVIDIA Tesla V100上测试

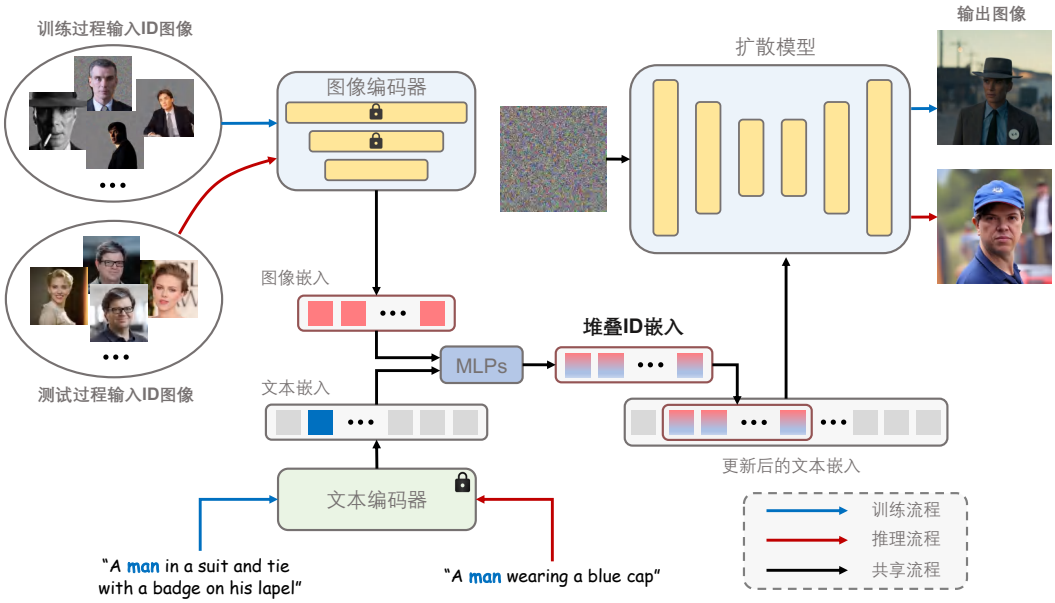

图 5.2: **本章提出的*PhotoMaker*流程的概览**。该算法首先从文本编码器和图像编码器中获取文本嵌入和图像嵌入。然后，该算法通过合并对应的类嵌入（例如，*man*和*woman*）和每个图像嵌入来提取融合嵌入。接下来，该算法沿着长度维度连接所有融合嵌入，形成一个堆叠的 ID嵌入。最后，该算法将堆叠的 ID嵌入输入到所有的交叉注意层，以自适应地合并生成模型中的 ID信息。注意，虽然该算法在训练过程中使用了具有二值分割图背景的同一 ID的图像，但该算法可以在推理过程中直接输入不同 ID的图像，无需将背景遮挡，就可以创建一个新的定制 ID。

类 [5, 155, 228, 238]，要么只关注脸部且不包括其他上下文信息 [232, 234, 239]。因此，作者设计了一个自动化的流程来构建一个与 ID相关的数据集，以便于本章提出的 PhotoMaker的训练。通过这个流程，作者可以构建一个包含许多 ID的数据集，每个 ID都有多个具有不同视角、属性和场景的图像。同时，这个流程可以为每个图像自动生成一个文本描述，标出对应的类别词 [165]，以更好地适应 PhotoMaker的训练需求。

5.2 方法

5.2.1 概述

给定几个要定制的 ID图像，本章提出的 PhotoMaker的目标是生成一个新的人像图像，保留输入 ID的特征，并在文本提示词的控制下改变生成 ID的内容或属性。尽管该方法像 DreamBooth一样输入多个 ID图像进行定制，但该方法

仍然享有与其他无需微调的方法相同的效率，通过单次前向传递完成定制，同时保持有希望的 ID保真度和文本可编辑性。此外，本章算法还可以混合多个输入 ID，生成的图像可以很好地保留不同 ID的特征，这为更多的应用提供了可能性。以上的能力主要来自本章提出的简单而有效的堆叠 *ID*嵌入，它可以提供输入 ID的统一表示。此外，为了方便训练本章提出的 PhotoMaker，本章还设计了一个数据构建流程来构建一个按 ID分类的人像数据集。 图 5.2展示了提出的PhotoMaker流程。 图 5.3展示了本章提出的数据构建流程。

5.2.2 堆叠 ID嵌入

编码器： 如最近的工作 [167, 176, 178]一样，本章使用 CLIP [158]图像编码器 $\mathcal{E}_{img}$ 来提取图像嵌入，因为它与扩散模型中的原始文本表示空间对齐。在将每个输入图像送入图像编码器之前，作者填充了除了特定 ID的身体部分以外的图像区域，以消除其他 ID和背景对训练过程的影响。由于用来训练原始 CLIP图像编码器的数据大部分是自然图像，为了使模型能够更好地从二值分割图图像中提取 ID相关的嵌入，作者在训练本章提出的 PhotoMaker时微调了图像编码器中的部分变换层。作者还引入了额外的可学习的投影层，将从图像编码器获取的嵌入注入到与文本嵌入相同的维度中。假设 $\{X^i \mid i = 1 \dots N\}$ 表示用户提供的 N 个输入 ID图像，作者因此得到了提取的嵌入 $\{e^i \in \mathbb{R}^D \mid i = 1 \dots N\}$，其中 D 表示投影维度。每个嵌入对应一个输入图像的 ID信息。对于给定的文本提示词 T，作者使用预训练的 CLIP文本编码器 $\mathcal{E}_{text}$ 提取文本嵌入 $t \in \mathbb{R}^{L \times D}$，其中 L 表示嵌入的长度。

堆叠： 最近的工作 [165, 166, 183]已经表明，在文本到图像的模型中，个性化的角色 ID信息可以由一些唯一的标记表示。本章提出的方法也有类似的设计，以更好地表示输入人像图像的 ID信息。具体来说，作者在输入标题中标记对应的类别词（例如，*man*和*woman*）（见章节 5.2.3）。然后，作者在文本嵌入中提取对应类别词位置的特征向量。这个特征向量 e^i 将与每个图像嵌入融合。作者使用两个 MLP层来执行这样的融合操作。融合的嵌入可以表示为 $\{\hat{e}^i \in \mathbb{R}^D \mid i = 1 \dots N\}$。通过结合类别词的特征向量，这个嵌入可以更全面地表示当前的输入 ID图像。此外，在推理阶段，这种融合操作也为定制生成过程提供了更强的语义可控性。例如，作者可以通过简单地替换类别词来定制人类 ID的年龄和性别（见章节 5.3.2）。

在获得融合嵌入后，作者沿着长度维度连接它们，形成堆叠的 *ID*嵌入：

$$s^* = \texttt{Concat}([\hat{e}^1, \ldots, \hat{e}^N]) \quad s^* \in \mathbb{R}^{N \times D} \tag{5.1}$$

这个堆叠的 ID嵌入在保留了每个输入 ID图像的原始表征同时，也可以作为多个 ID图像的统一表示。它可以接受任意数量的 ID图像编码嵌入，因此，它的长度 N 是可变的。与基于 DreamBooth的方法 [165, 171]相比，DreamBooth输入多个图像来微调模型进行个性化定制，而本章提出的方法本质上是同时向模型输入多个嵌入。在将同一 ID的多个图像打包成图像编码器的输入后，可以通过单次前向传递获得堆叠的 ID嵌入，这比基于微调的方法大大提高了效率。同时，与其他基于嵌入的方法 [167, 183]相比，这个统一的表示可以保持不错的 ID保真度和文本可控性，因为它包含了更全面的 ID信息。此外，值得注意的是，尽管作者在训练过程中只使用了同一 ID的多个图像来形成这个堆叠的 ID嵌入，但作者可以在推理阶段使用来自不同 ID的图像来构造它。这种灵活性为许多有趣的应用开启了可能性。例如，作者可以混合两个现实存在的人，或者混合一个人和一个知名的角色 IP（见章节 5.3.2）。

合并： 作者使用扩散模型中固有的交叉注意力机制来自适应地合并堆叠 ID嵌入中包含的 ID信息。作者首先用堆叠的 ID嵌入 s^* 替换原始文本嵌入 t 中对应类别词位置的特征向量，得到更新的文本嵌入 $t^* \in \mathbb{R}^{(L+N-1)\times D}$。然后，交叉注意力操作可以表示为：

$$\begin{cases} \mathbf{Q} = \mathbf{W}_Q \cdot \phi(z_t); \ \mathbf{K} = \mathbf{W}_K \cdot t^*; \ \mathbf{V} = \mathbf{W}_V \cdot t^* \\ \texttt{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}) \cdot \mathbf{V}, \end{cases} \tag{5.2}$$

其中，$\phi(\cdot)$ 是一个可以从输入潜在变量通过 UNet去噪器编码的嵌入。$\mathbf{W}_Q$、$\mathbf{W}_K$ 和 $\mathbf{W}_V$ 是投影矩阵。此外，作者可以通过提示词权重 [236, 237]调整一个输入 ID图像在生成新的定制 ID中的参与程度，这种方式也展示了本章提出的 PhotoMaker的灵活性。最近的工作 [170, 171]发现，通过简单地调整注意力层的权重，可以实现良好的 ID定制性能。为了使原始的扩散模型更好地感知堆叠 ID嵌入中包含的 ID信息，作者额外训练了注意力层矩阵的 LoRA [171, 184]残差。

5.2.3 ID导向的人像数据构建

由于本章提出的 PhotoMaker在训练过程中需要采样同一 ID的多个图像来

构建堆叠的 ID嵌入，作者需要使用一个按 ID分类的数据集来驱动本章提出的 PhotoMaker的训练过程。然而，现有的与人相关的数据集要么不注释 ID信息 [5, 155, 228, 238]，要么它们包含的场景丰富性非常有限 [232, 234, 239]（即，它们只关注脸部区域）。因此，在这一节，作者将介绍一个构建按 ID分类以人为中心的文本-图像数据集的流程。 图 5.3展示了提出的流程。通过这个流程，作者可以收集一个以 ID为导向的数据集，该数据集包含大量的 ID，每个 ID都有多个包含不同表情，属性，场景等的图像。这个数据集不仅便于本章提出的 PhotoMaker的训练过程，也可能激发未来的 ID驱动研究的潜力。经过一系列的过滤步骤，作者构建的数据集中的图像数量约为112K。它们按照约13,000个 ID名称进行分类。每个图像都附带一个对应 ID的二值分割图和一个带类别词位置信息的文本描述。

图像下载： 作者首先列出一个名人名单，可以从 VGGFace2 [240] 获取。作者根据列表在搜索引擎中搜索名字并爬取数据。每个名字作者下载大约100张图像。为了生成更高质量的肖像图像 [164]，作者在下载过程中过滤掉分辨率最短边小于512的图像。

人脸检测和过滤： 作者首先使用 RetinaNet [241]检测人脸边界框，并过滤掉小尺寸的检测结果（小于256 $\times$ 256）。如果一个图像不包含任何满足要求的边界框，该图像将被过滤掉。然后，作者对剩余的图像进行 ID验证。

ID验证： 由于一个图像可能包含多个脸，作者首先需要识别哪个脸属于当前的身份组。具体来说，作者将当前身份组的所有面部区域送入 ArcFace [242]中提取身份嵌入，并计算每对脸的 L2相似度。作者将每个身份嵌入与所有其他嵌入计算的相似度相加，得到每个边界框的得分。作者为每个包含多个脸的图像选择得分最高的边界框。在选择边界框后，作者重新计算每个剩余框的总得分。作者计算 ID组的总得分的标准差 δ。作者经验性地使用 8δ 作为阈值，过滤掉 ID不一致的图像。

裁剪和分割： 作者首先根据检测到的面部区域裁剪出一个更大的正方形框，同时确保裁剪后的面部区域可以占据图像的超过10%。由于作者需要在将输入 ID图像送入图像编码器之前去除与输入 ID无关的背景和 ID，作者需要生成指定 ID的二值分割图。具体来说，作者使用 Mask2Former [243]对“人物（person）”类进行全景分割。作者保留与 ID对应的面部边界框重叠最大的二值分割图。此

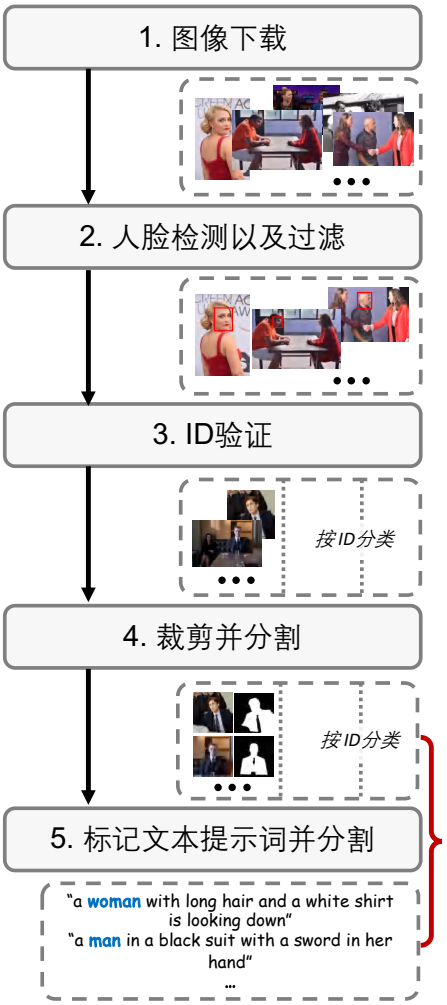

图 5.3: **本章提出的ID导向的数据组装流程概览**。红色括号表示最后 PhotoMaker算法训练所使用的数据对，即由第4步得到的图像和二值分割图和第5步得到的文本标记组成。

外，作者选择丢弃没有检测到二值分割图的图像，以及没有找到边界框和二值分割图区域重叠的图像。

字幕和标记： 作者使用 BLIP2 [244]为每个裁剪的图像生成一个文本描述。由于作者需要标记类别词（例如，*man*和*woman*）以便于文本和图像嵌入的融合。作者使用 BLIP2的随机模式重新生成不包含任何类别词的文本描述，直到出现一个类别词为止。在得到文本描述后，作者将文本描述中的类别词单数化，以便专注于单一的 ID生成。接下来，作者需要标记出对应于当前 ID的类别词的位置。只包含一个类别词的文本描述可以直接注释。对于包含多个类别词的文本描述，作者计算每个身份组的文本描述中包含的类别词。出现次数最多的

类别词将是当前身份组的类别词。然后，作者使用每个身份组的类别词来匹配和标记该身份组中的每个文本描述。对于不包含与对应身份组类别词匹配的类别词的文本描述，作者使用依赖解析（dependency parsing）模型 [245]根据不同的类别词分割文本描述。作者计算分割后的子文本描述与图像中特定 ID区域的 CLIP得分 [158]。此外，作者通过 SentenceFormer [246]计算当前段落的类别词与当前身份组的类别词之间的标签相似性。作者选择标记对应于 CLIP得分和标签相似性的乘积最大的类别词。

5.3 实验

5.3.1 设置

实现细节： 为了生成质量更高的人像图像，本章算法采用 SDXL模型 [164][2]作为本章提出的文本到图像合成模型。相应地，训练数据的分辨率调整为 1024×1024。作者使用 CLIP ViT-L/14 [158]和一个额外的投影层来获取初始图像嵌入 e^i。对于文本嵌入，作者保留 SDXL中的原始两个文本编码器进行提取。整个框架在8个 NVIDIA A100 GPU上使用 Adam [247]优化了两周，批大小为48。作者将学习率设置为 LoRA权重的 $1e-4$，其他可训练模块的 $1e-5$。在训练过程中，作者随机采样1-4个与当前目标 ID图像相同的 ID的图像来形成堆叠的 ID嵌入。此外，为了通过使用无分类器指导来提高生成性能，作者有10%的机会使用空文本嵌入来替换原始的更新文本嵌入 t^*。作者还使用了50%的概率使用二值分割图扩散损失 [248]，以鼓励模型生成更忠实于 ID相关区域的图像。在推理阶段，作者使用延迟主题条件 [183]来解决文本和 ID条件之间的冲突。作者使用50步的 DDIM采样器 [23]。无分类器指导（classifier-free guidance）的参数设置为5。

评价指标： 按照 DreamBooth [165]的做法，作者使用 DINO [249]和 CLIP-I [166]指标来衡量 ID保真度，并使用 CLIP-T [158]指标来衡量文本保真度。为了更全面的评价，作者还计算了生成图像和同一 ID的真实图像之间的面部相似度，通过检测和裁剪面部区域。作者使用 RetinaFace [241]作为检测模型。面部嵌入是通过 FaceNet [250]提取的。为了评价生成的质量，作者使用了 FID指标 [251, 252]。值得注意的是，由于大多数基于嵌入的方法倾向于将面部姿态和

[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

表 5.1: **用于测评的 ID名**。 对于每个人物，作者收集了四张与之对应的图像。

测试用人物 ID名	
① Alan Turing	⑭ Kamala Harris
② Albert Einstein	⑮ Marilyn Monroe
③ Anne Hathaway	⑯ Mark Zuckerberg
④ Audrey Hepburn	⑰ Michelle Obama
⑤ Barack Obama	⑱ Oprah Winfrey
⑥ Bill Gates	⑲ Rene Zellweger
⑦ Donald Trump	⑳ Scarlett Johansson
⑧ Dwayne Johnson	㉑ Taylor Swift
⑨ Elon Musk	㉒ Thomas Edison
⑩ Fei-Fei Li	㉓ Vladimir Putin
⑪ Geoffrey Hinton	㉔ Woody Allen
⑫ Jeff Bezos	㉕ Yann LeCun
⑬ Joe Biden	

表情融入到表示中，生成的图像往往在面部区域缺乏变化。因此，作者提出了一个指标，名为*生成人脸的多样性*，来衡量生成的面部区域的多样性。具体来说，作者首先检测并裁剪每个生成图像的面部区域。接下来，作者计算所有生成图像的所有面部区域之间的每对 LPIPS [253]得分，并取平均值。这个值越大，生成的面部区域的多样性越高。

评价数据集：本章提出的评价数据集包括25个 ID，包括来自 Mystyle [232]的9个 ID和作者自己收集的额外16个 ID。这些 ID与训练集中出现的 ID没有重叠，因此可以用来评价模型的泛化能力。为了进行更全面的评价，作者还准备了40个提示词，覆盖了各种表情、属性、装饰、动作和背景，这些提示词都列在列在表 5.2和表 5.3中。对于每个 ID的每个提示词，作者生成4个图像进行评价。用于评价的图像数据集包括手动选择的额外 ID和一部分 MyStyle [232]数据。对于每个 ID名称，作者有四个图像作为比较方法的输入数据，以及最终度量评价（即，DINO [249]、CLIP-I [166]和人脸相似度 [242]）。对于单嵌入方法（即，FastComposer [183]和 IPAdapter [163]），作者从每个 ID组中随机选择一张图像作为输入。作者在表 5.1中列出了用于评价的 ID名称。

5.3.2 应用

在本节中，作者将详细介绍本章提出的 PhotoMaker可以实现的应用。

表 5.2: **按照一般设置、服装、配饰、动作分类的评估文本提示词。** `class word`将被替换为`man`，`woman`，`boy`等。对于每个 ID和每个提示词，作者随机生成了四个图像进行评估。

类别	文本描述（Prompt）
通用	a photo of a `<class word>`
衣着	a `<class word>` wearing a Superman outfit
	a `<class word>` wearing a spacesuit
	a `<class word>` wearing a red sweater
	a `<class word>` wearing a purple wizard outfit
	a `<class word>` wearing a blue hoodie
配饰	a `<class word>` wearing headphones
	a `<class word>` with red hair
	a `<class word>` wearing headphones with red hair
	a `<class word>` wearing a Christmas hat
	a `<class word>` wearing sunglasses
	a `<class word>` wearing sunglasses and necklace
	a `<class word>` wearing a blue cap
	a `<class word>` wearing a doctoral cap
	a `<class word>` with white hair, wearing glasses
动作	a `<class word>` in a helmet and vest riding a motorcycle
	a `<class word>` holding a bottle of red wine
	a `<class word>` driving a bus in the desert
	a `<class word>` playing basketball
	a `<class word>` playing the violin
	a `<class word>` piloting a spaceship
	a `<class word>` riding a horse
	a `<class word>` coding in front of a computer
	a `<class word>` playing the guitar

表 5.3: **按照表情、视图和背景分类的评估文本提示词。** `class word`将被替换为`man`，`woman`，`boy`等。对于每个 ID和每个提示词，作者随机生成了四个图像进行评估。

类别	文本描述（Prompt）
表情	a <class word> laughing on the lawn
	a <class word> frowning at the camera
	a <class word> happily smiling, looking at the camera
	a <class word> crying disappointedly, with tears flowing
	a <class word> wearing sunglasses
视角变化	a <class word> playing the guitar in the view of left side
	a <class word> holding a bottle of red wine, upper body
	a <class word> wearing sunglasses and necklace, close-up, in the view of right side
	a <class word> riding a horse, in the view of the top
	a <class word> wearing a doctoral cap, upper body, with the left side of the face facing the camera
	a <class word> crying disappointedly, with tears flowing, with left side of the face facing the camera
背景	a <class word> sitting in front of the camera, with a beautiful purple sunset at the beach in the background
	a <class word> swimming in the pool
	a <class word> climbing a mountain
	a <class word> skiing on the snowy mountain
	a <class word> in the snow
	a <class word> in space wearing a spacesuit

图 5.4: **在基于文本的上下文生成应用上的定性比较**。作者将本章提出的方法与 DreamBooth [165]、Textual Inversion [166]、FastComposer [183]和 IPAdapter [163]进行比较，对于五个不同的身份和相应的提示词。作者观察到，本章提出的方法通常能够实现高质量的生成、不错的可编辑性和良好的身份保真度。

表 5.4: **在基于文本的上下文生成应用上的定量比较**。用于基准测试的指标包括保留 ID信息的能力（即，CLIP-I，DINO，和人脸相似度），文本一致性（即，CLIP-T），生成人脸的多样性，和生成质量（即，FID）。此外，作者定义个性化速度为在输入 ID条件后获取最终个性化图像所需的时间。作者在单个 NVIDIA Tesla V100 GPU上测量个性化时间。最好的结果以**粗体**显示，第二好的以下划线 标出。

	CLIP-T↑ (%)	CLIP-I↑ (%)	DINO↑ (%)	人脸相似性 ↑ (%)	生成人脸多样性 ↑ (%)	FID↓	速度 ↓ (s)
DreamBooth [165]	**29.8**	62.8	39.8	49.8	49.1	374.5	1284
Textual Inversion [166]	24.0	70.9	39.3	54.3	**59.3**	**363.5**	2400
FastComposer [183]	28.7	66.8	40.2	61.0	45.4	375.1	8
IPAdapter [163]	25.1	71.2	46.2	**67.1**	52.4	375.2	12
PhotoMaker (本章算法)	26.1	**73.6**	**51.5**	61.8	57.7	370.3	10

表 5.5: 非名人的客观指标比较。

方法	CLIP-T↑	DINO↑	Face Sim.↑	Face Div.↑
DreamBooth	**30.1**	44.4	37.7	47.5
FastComposer	25.9	**54.2**	**69.2**	39.1
IP-Adapter	23.3	47.5	61.4	37.7
PhotoMaker (本章算法)	29.5	50.5	66.7	**52.5**

对于每个应用，作者选择可能最适合相应应用的比较方法与本章提出的PhotoMaker进行对比。比较方法将从 DreamBooth [165]、Textual Inversion [166]、FastComposer [183]和 IPAdapter [163]中选择。对于每种方法，作者都优先使用每种方法提供的官方模型。对于 DreamBooth和 IPAdapter，作者使用它们的SDXL版本进行公平比较。对于所有应用，作者在本章提出的 PhotoMaker中选择了四个输入 ID图像来形成堆叠的 ID嵌入。作者也公平地使用四个图像来训练需要测试时间优化的方法。

简单上下文变化：作者首先展示了简单的上下文变化的结果，如修改头发颜色和衣服或根据基本提示词控制生成背景。由于所有的方法都可以适应这个应用，作者对生成的结果进行了定量和定性的比较（见表 5.4和图 5.4）。结果表明，本章提出的方法可以很好地满足生成高质量图像的能力，同时保证了高ID保真度（具有最大的 CLIP-T和 DINO得分，以及第二好的人脸相似度）。与大多数方法相比，本章提出的方法生成的图像质量更高，生成的面部区域展示了更大的多样性。同时，本章提出的方法可以保持与基于嵌入的方法一致的高效率。为了更全面的比较，作者在章节 5.3.3中展示了用户研究的结果。作者还在图 5.17中提供了一个更丰富的比较。

此外，本章还尝试通过 PhotoMaker生成 SDXL自己无法生成的 ID。这种情况可以被称为“非名人”案例。通过比较图 5.18和图 5.5，本章提出的方法可以成功地在输入“非名人”ID时也可以生成保真度高的输出图像。此外，作者还收集了12组非名人 ID以证明模型的泛化能力，其中包括作者身边同事的图像和由生成模型（即，GAN和扩散模型）生成的面部图像，每个 ID集大约有1-3个图像。作者选择了10个提示词进行评估。结果如图 5.6和表 5.5所示，作者的方法在文本一致性（CLIP-T）和 ID相似性（DINO和 Face Sim.）上排名第二，同时生成了最多样化的面部区域（Face Div.）。这证明了作者方法的全面性。

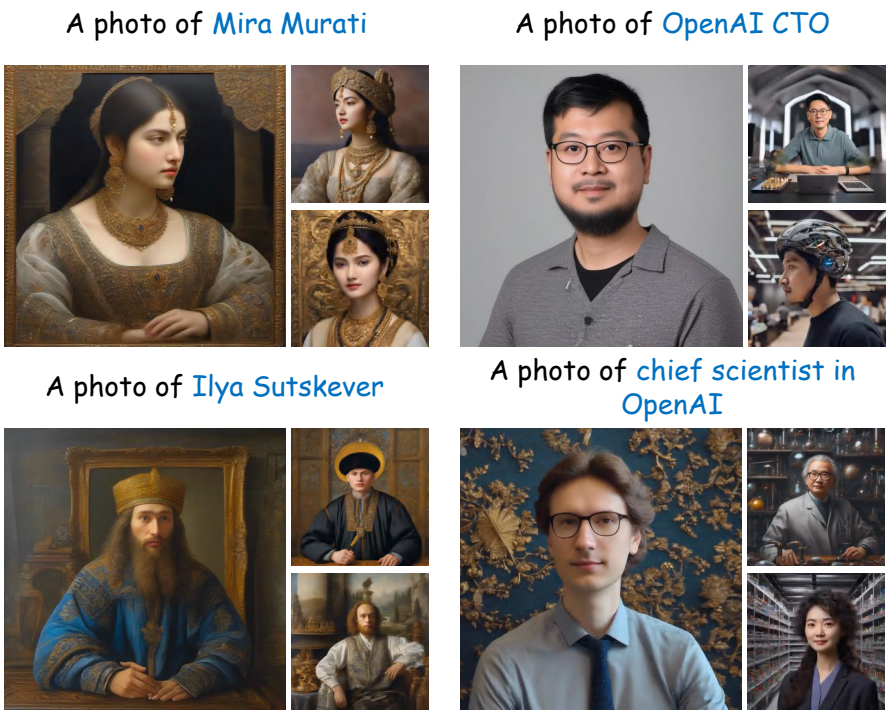

图 5.5: **两个 SDXL无法生成的人物的例子。** 作者替换了两种类型的文本提示词（如，姓名和职位）都无法使 SDXL生成 Mira Murati和 Ilya Sutskever。

图 5.6: **非名人的视觉比较。** 作者使用了由 GAN生成的面部图像 [254]作为参考图像。

将艺术品/旧照片中的人物带入现实： 通过将艺术画作、雕塑或一个人的旧照片作为输入，本章提出的 PhotoMaker可以将上个世纪甚至古代的人带到现在的世纪来为他们“拍照”。 图 5.7展示了结果。与本章提出的方法相比，Dreambooth和 SDXL都难以生成真实的人像图像，这些图像在真实照片中没有出现过。此外，由于 DreamBooth过于依赖定制图像的质量和分辨率，因此当使用旧照片进行定制生成时，DreamBooth难以生成高质量的结果。图 5.19和5.20中分别展示了更多的老照片和艺术作品回到现实生活的例子。对于作者比较的其他方法来说，实现这一点是相当具有挑战性的。

图 5.7: **PhotoMaker在艺术品和旧照片上的应用。** 本章算法能够将过去的人带回现实生活，或者改变输入 ID的年龄和性别。作者为 DreamBooth和 SDXL准备了一个提示词模板`A photo of <original prompt>, photo-realistic`。相应地，作者将原始提示词中的类别词改为名人的名字。

改变年龄或性别： 通过简单地替换类别词（例如 man和 woman），本章提出的方法可以实现性别和年龄的变化。 图 5.8展示了结果。尽管 SDXL和 DreamBooth也可以在提示词工程后实现相应的效果，但由于堆叠 ID嵌入的作用，本章提出的方法可以更容易地捕捉到角色的特征信息。因此，本章提出的结果显示了更高的 ID保真度。作者在图 5.21中提供了更多改变年龄或性别的视觉结果。

身份混合： 如果用户提供了不同 ID的图像作为输入，本章提出的 Pho-

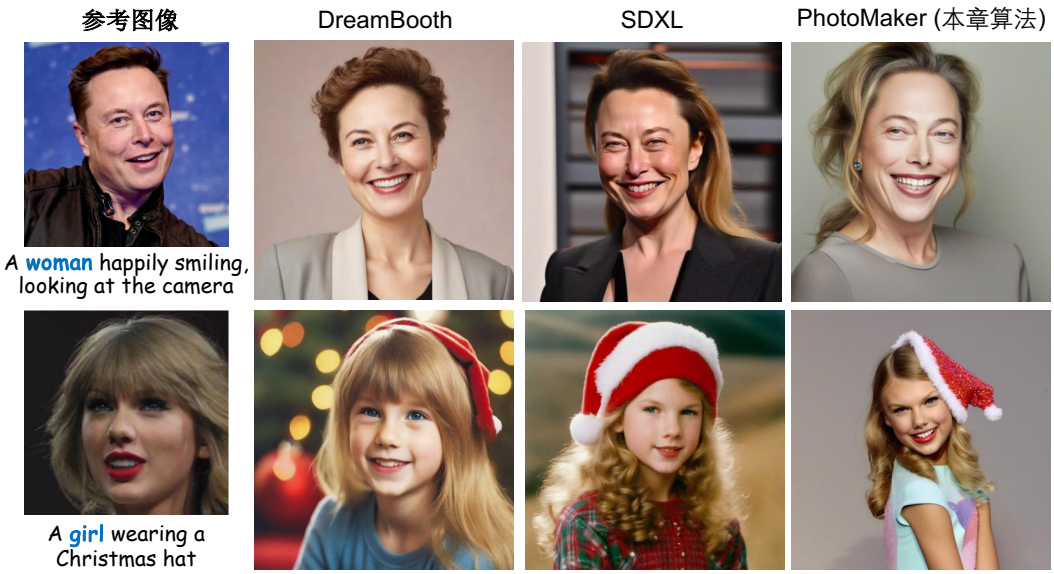

图 5.8: **PhotoMaker在改变年龄或性别上的应用**。 本章算法能够改变输入 ID的年龄和性别，作者为 DreamBooth和 SDXL准备了一个提示词模板`A photo of <original prompt>`。作者将类别词替换为`<name>, (at the age of 12)`。

toMaker可以很好地整合不同 ID的特征，形成一个新的 ID。从图 5.9中，作者可以看到，无论 DreamBooth还是 SDXL都无法实现身份混合。相比之下，无论输入是动漫 IP还是真人、无论性别差异，本章提出的方法可以在生成的新 ID上很好地保留输入的不同 ID的特征。作者在图 5.22中提供了更多的可视化效果展示。此外，作者可以通过控制相应 ID输入的数量或提示词权重来控制这个 ID在新生成的 ID中的比例。作者在图 5.14-5.15中展示了这种能力。展现了本方法相对其他方法更多的可能性。

风格化：在图 5.10中，作者展示了本章提出的方法的风格化能力。作者可以看到，在生成的图像中，本章提出的 PhotoMaker不仅保持了良好的 ID保真度，而且有效地展示了输入提示词的风格信息。这揭示了本章提出的方法驱动更多应用的潜力。 图 5.16展示了更多的结果。

5.3.3 用户研究

在这一节中，作者进行了一个用户研究，以进行更全面的比较。作者选择的比较方法包括 DreamBooth [165]、FastComposer [183]和 IPAdapter [163]。作者使用 SDXL [164]作为 DreamBooth和 IPAdapter的基础模型，因为它们的实现是开源的。作者为每个用户显示20个文本-图像对。每个对都包括输入 ID的参考图

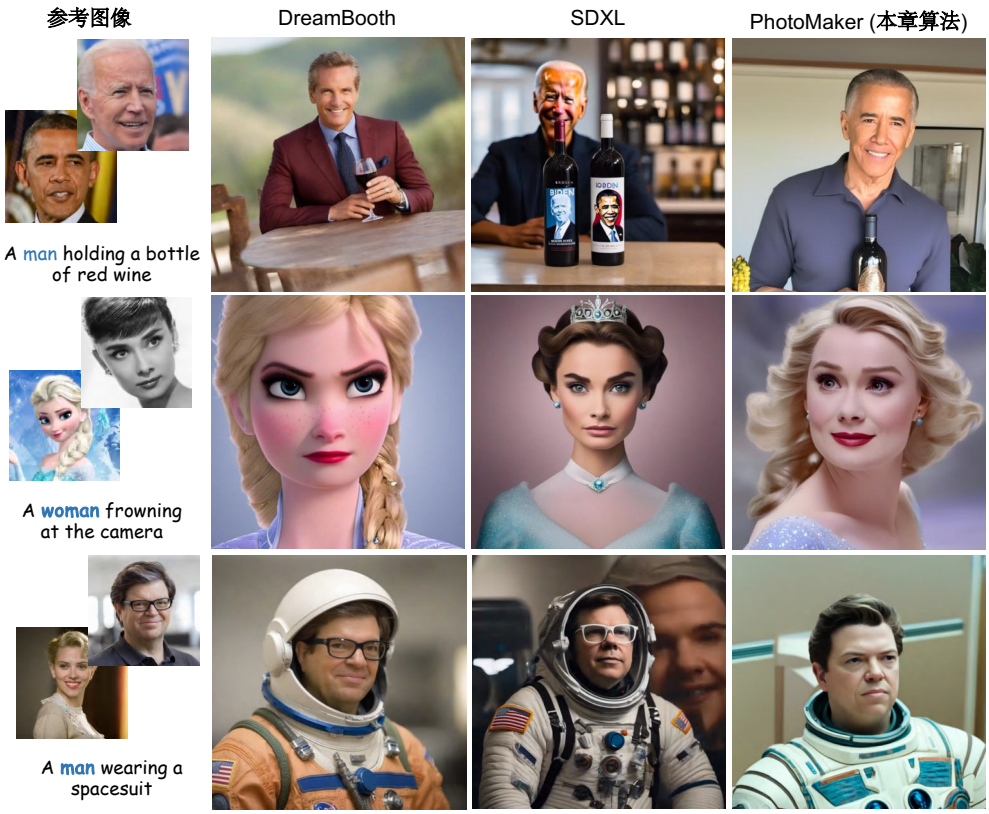

图 5.9: **身份混合**。本章提出的 PhotoMaker能够生成具有新 ID的图像，同时保留输入身份特征。作者为 SDXL准备了一个提示词模板`<original prompt>, with a face blended with <name:A> and <name:B>`。

像和相应的文本提示词。作者为每个文本-图像对的每种方法生成了四个随机生成的图像。作者要求每个用户回答这20组结果的四个问题： 1）哪种方法与输入人的身份最相似？ 2）哪种方法产生的生成图像的质量最高？ 3）哪种方法在图像中生成最多样化的面部区域？ 4）哪种方法生成的图像最能匹配输入的文本提示词？作者已经对所有方法的名称进行了匿名处理，并在每组回答中随机化了方法的顺序。共有40名候选人参加了本节的用户研究，共计收到了3,200张有效的投票。结果如图 5.11所示。

作者发现，本章提出的 PhotoMaker在 ID保真度、生成质量、多样性和文本保真度方面都有优势，特别是后三者。此外，作者发现 DreamBooth是在平衡这四个评价维度方面的第二好的算法，这可能解释了为什么它在过去比基于嵌入的方法更流行。同时，IPAdapter在生成图像质量和文本一致性方面表现出明显的劣势，因为它在训练阶段更关注图像嵌入。 FastComposer在生成面部区

图 5.10: **本章提出的 PhotoMaker的风格化结果。** 符号`<class>`表示它将被替换为`man`或`woman`。

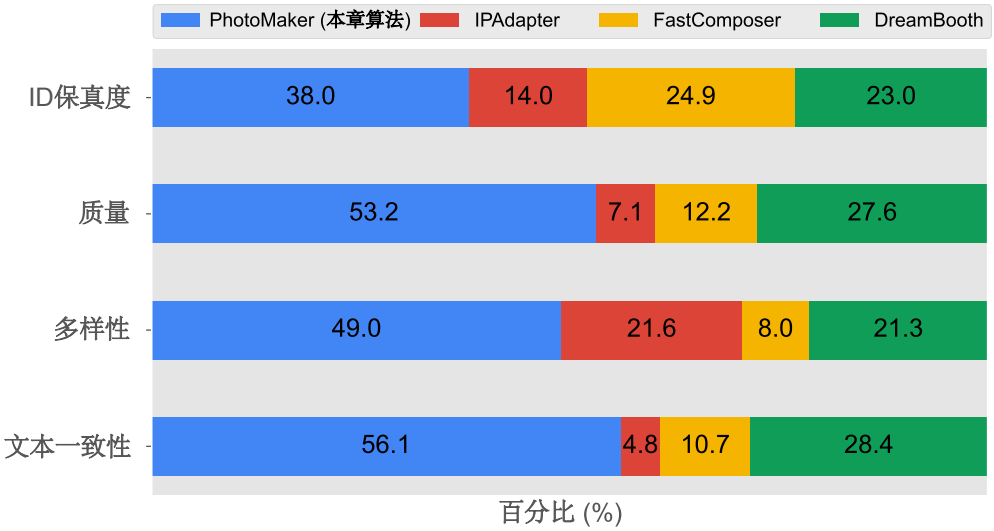

图 5.11: **用户对不同方法的 ID保真度、生成质量、面部多样性和文本保真度的偏好。** 为了便于说明，作者将每种方法收到的总票数的比例进行了可视化。本章提出的 PhotoMaker在这四个维度上占据了最大的比例。

域的多样性方面有明显的不足，因为他们的单嵌入训练流程。以上结果与主文中的表 5.4大致一致，除了 CLIP-T指标的差异。这可能是因为在手动选择最符合文本的图像时，人们更倾向于选择与文本中出现的对象协调的图像。相反，CLIP-T更倾向于关注对象是否出现。这可能表明了 CLIP-T的局限性。作者在图 5.17-5.20中提供了更多的可视化样本供参考。

图 5.12: **改变输入图像数量对生成结果的影响。** 可以观察到，随着输入图像数量的增加，ID的保真度也在增加。

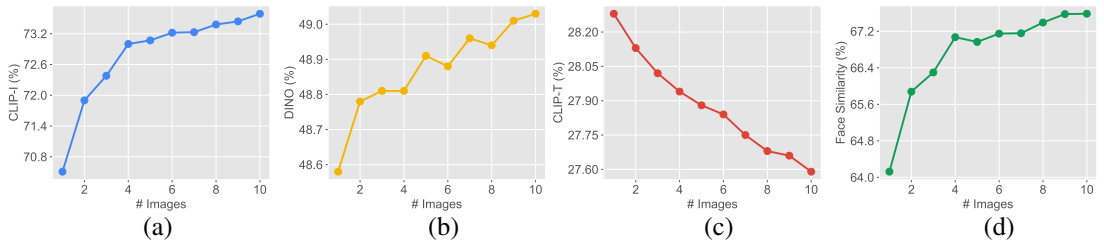

图 5.13: **输入 ID图像数量分别对 (a) CLIP-I、(b) DINO、(c) CLIP-T和 (d) 人脸相似度的影响。**

5.3.4 消融研究

作者将每个变体的总训练迭代次数缩短到八分之一，以进行消融研究。

输入 ID图像数量的影响： 作者探索了通过输入不同数量的 ID图像来形成提出的堆叠 ID嵌入的影响。在图 5.13中，作者在不同的指标上可视化了这种影响。作者得出的结论是，使用更多的图像来形成堆叠的 ID嵌入可以提高与 ID保真度相关的指标。这种改进在输入图像数量从一到两时尤其明显。随着输入图像数量的增加，ID相关指标的值的增长率显著减慢。此外，作者观察到 CLIP-T指标呈线性下降。这表明可能存在 ID保真度和文本可控性之间的权衡。从图 5.12中，作者可以看到增加输入图像的数量可以增加 ID的相似性。因此，更多的 ID图像形成堆叠的 ID嵌入可以帮助模型感知更全面的 ID信息，然后更准确地表示 ID来生成图像。此外，如 Dwayne Johnson的例子所示，性别编辑能力下降，模型更倾向于生成原始 ID的性别的图像。

表 5.6: **研究不同的嵌入组合方式**。 最好的结果以**加粗**显示。

	CLIP-T↑	DINO↑	人脸相似度↑	面部生成差异性↑
平均操作	**28.7**	47.0	48.8	**56.3**
线性层	28.6	47.3	48.1	54.6
堆叠操作	28.0	**49.5**	**53.6**	55.0

表 5.7: **训练数据采样策略**。 最好的结果以**加粗**显示。

	CLIP-T↑	DINO↑	人脸相似度↑	面部生成差异性↑
单个嵌入	27.9	50.3	50.5	**56.1**
单张图像	27.3	**50.3**	**60.4**	51.7
本章算法	**28.0**	49.5	53.6	55.0

组成多个嵌入的选择： 作者探索了三种组成 ID嵌入的方式，包括平均图像嵌入，通过线性层自适应地投影嵌入，以及本章提出的堆叠方式。从表 5.6中，作者可以看到堆叠方式在保证生成面部的多样性的同时，具有最高的 ID保真度，证明了其有效性。此外，这种方式比其他方式提供了更大的灵活性，包括接受任意数量的图像和更好地控制不同 ID的混合过程。

训练中多个嵌入的好处： 作者探索了两种其他的训练数据采样策略，以证明在训练过程中输入多个图像是必要的。第一种是只选择一个图像，这个图像可以与目标图像不同，来形成 ID嵌入（见表 5.7中的“单个嵌入”）。本章提出的多嵌入方式在 ID保真度上有优势。第二种采样策略是将目标图像视为输入

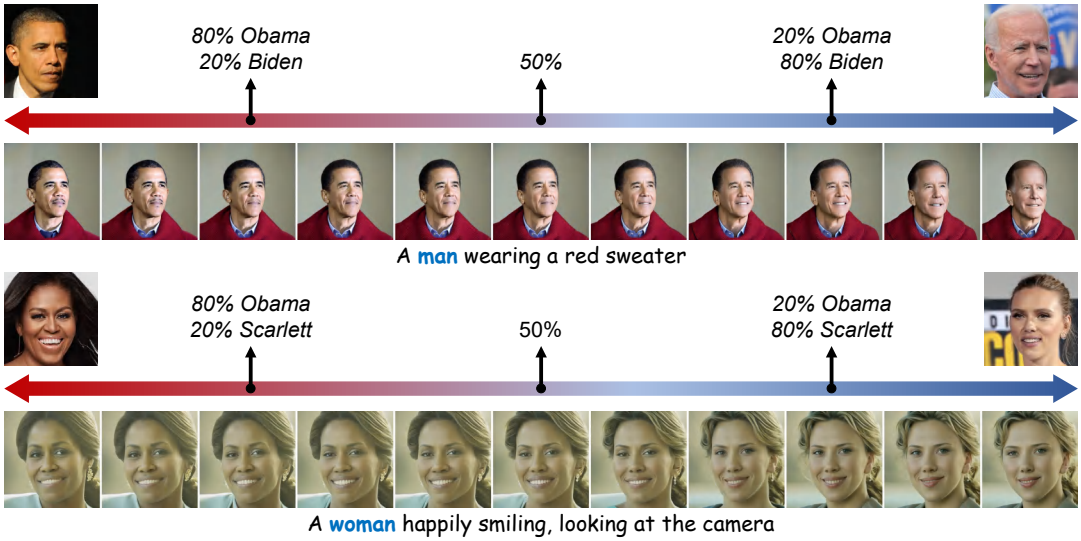

图 5.14: **输入样本池中不同 ID的图像比例对新 ID生成的影响。** 第一行描绘了从 Barack Obama到 Joe Biden的过渡。第二行描绘了从 Michelle Obama到 Scarlett Johansson的转变。为了提供更清晰的说明，图中使用百分比来表示输入图像池中每个 ID的比例。输入池中包含的图像总数为10。

ID图像（模拟大多数基于嵌入的方法的训练方式）。作者基于这个图像生成多个图像，使用不同的数据增强方法，并提取相应的多个嵌入。在表 5.7中可以看出，由于模型可以轻易地记住输入图像的其他无关特征，生成的面部区域缺乏足够的变化（低多样性）。

调整身份混合中的比例： 对于身份混合，本章提出的方法可以通过控制输入图像池中身份图像的百分比，或者通过提示词权重的方法 [236]来调整合并比例。这样，作者可以控制生成的新 ID的人要么更接近，要么更远离特定的输入 ID。 图 5.14展示了本章提出的方法如何通过控制输入图像池中不同 ID的比例来定制新的 ID。为了更好的描述，作者在这个实验中使用了总共10张图像作为输入。作者可以观察到两个 ID的图像之间的平滑过渡。这个平滑的过渡包括了肤色和年龄的变化。接下来，作者使用每个生成的 ID的四张图像进行提示词权重。结果如图 5.15所示。作者将与特定 ID相关的图像对应的嵌入乘以一个系数，以控制它在新 ID中的合并比例。与控制输入图像数量的方式相比，提示词权重需要更少的照片来调整不同 ID的合并比例，显示出其优越的可用性。此外，调整不同 ID混合比例的两种方式都展示了作者方法的灵活性。

身份混合： 作者在图 5.22中提供了更多身份混合应用的视觉结果。由于作

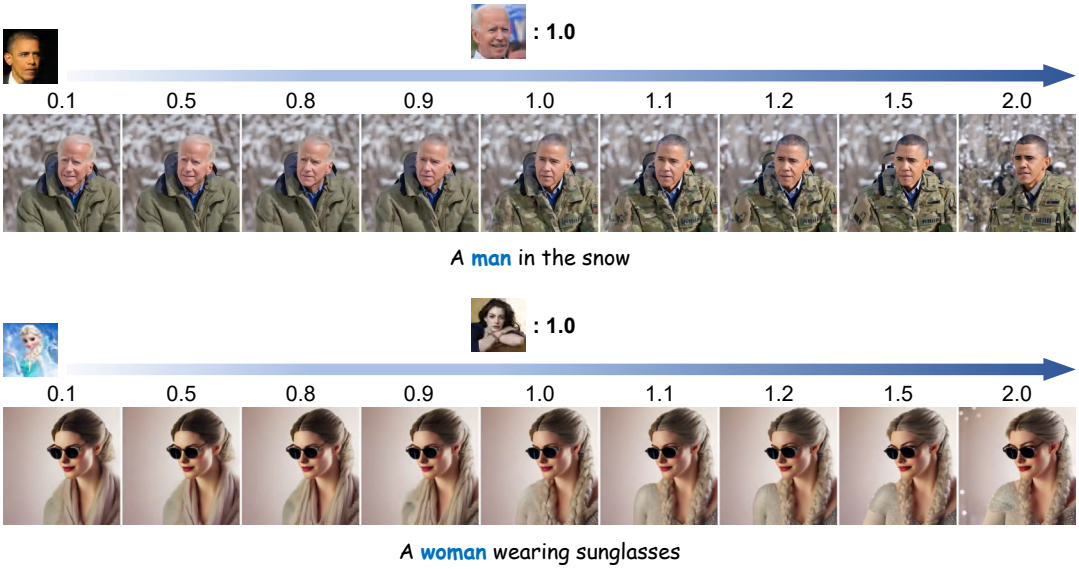

图 5.15: **提示词权重（Prompt Weighting）对新 ID生成的影响。** 第一行描绘了 Barack Obama和 Joe Biden的混合。从左到右的第一行表示图像中对应于 Barack Obama的 ID图像嵌入的权重逐渐增加。第二行描绘了 Elsa (Disney)和 Anne Hathaway的混合。 Elsa的权重逐渐增加。

者的堆叠 ID嵌入的优势，本章提出的方法可以有效地融合不同 ID的特征，形成一个新的 ID。然后，作者可以基于这个新的 ID生成文本控制。此外，本章提出的方法在身份混合过程中提供了很大的灵活性，如图 5.14-5.15所示。更重要的是，作者在主文中已经探讨了现有的方法在实现这个应用方面的困难。相反，本章提出的 PhotoMaker开启了大量的可能性。

5.4 总结

作者在本章中提出了*PhotoMaker*，这是一种高效的个性化文本到图像生成方法，专注于生成人像照片。 PhotoMaker利用了一个简单而有效的表示——堆叠的 ID嵌入，以更好地保留 ID信息。实验结果证明，与其他方法相比，本章提出的 PhotoMaker可以同时满足高质量和多样性的生成能力，良好的可编辑性，高推理效率和良好的 ID保真度。此外，作者还发现 PhotoMaker可以实现许多有趣的应用，这些应用是以前的方法难以实现的，如改变年龄或性别，将旧照片或艺术品中的人物带回现实，以及身份混合。

局限性： 首先，本章提出的方法只关注在图像中保持一个生成人的 ID信息，不能同时控制一张图像中多个生成人的 ID。其次，本章提出的方法擅长生

成半身肖像，但在生成全身肖像方面相对不太好。第三，本章提出的方法的年龄转换能力不如一些基于 GAN的方法 [255]精确。如果用户需要更精确的控制，可能需要对训练数据集的标题进行修改。最后，本章提出的方法基于 SDXL和作者构建的数据集，所以它也会继承来自模型和数据的偏好。

更广泛的影响：在这篇论文中，作者介绍了一种新的方法，能够生成高质量的人像图像，同时保持与输入身份的高度相似度。同时，本章提出的方法也可以满足高效率，适当的面部生成多样性和良好的可控性。

对于学术界，本章提出的方法为个性化生成提供了一个强大的基线。PhotoMaker数据创建流程使得可以创建更多具有不同姿势、动作和背景的多样化数据集，这对于开发更强大和更具泛化能力的计算机视觉模型是非常有用的。

在实际应用领域，PhotoMaker技术有可能革新娱乐业，可以在不需要大量CGI工作的情况下为电影或视频游戏创建逼真的角色。它也可以在虚拟现实中有所帮助，通过让用户在不同的场景中看到自己，提供更沉浸式和个性化的体验。值得注意的是，每个人都可以依赖本章提出的 PhotoMaker快速定制他们自己的数字肖像。

然而，作者承认，生成高保真度的人像图像的能力带来了伦理考虑。这种技术的普及可能导致生成肖像的不适当使用、恶意图像篡改和虚假信息的传播的激增。因此，作者强调开发和遵守伦理指南，并负责任地使用这项技术的重要性。作者希望 PhotoMaker贡献将进一步推动关于计算机视觉中人像生成的安全和伦理使用的讨论和研究。

图 5.16: **本章提出的 PhotoMaker的风格化结果，包括不同的输入 ID和不同的风格提示词。** 本章提出的方法可以无缝地转移到各种风格，同时防止生成逼真的结果。符号<class>表示它将被替换为man或woman。

图 5.17: **文本重定义上下文应用的更多视觉示例。** 本章提出的方法不仅提供了高 ID保真度，而且保留了文本编辑能力。作者为每个提示词随机采样三张图像。

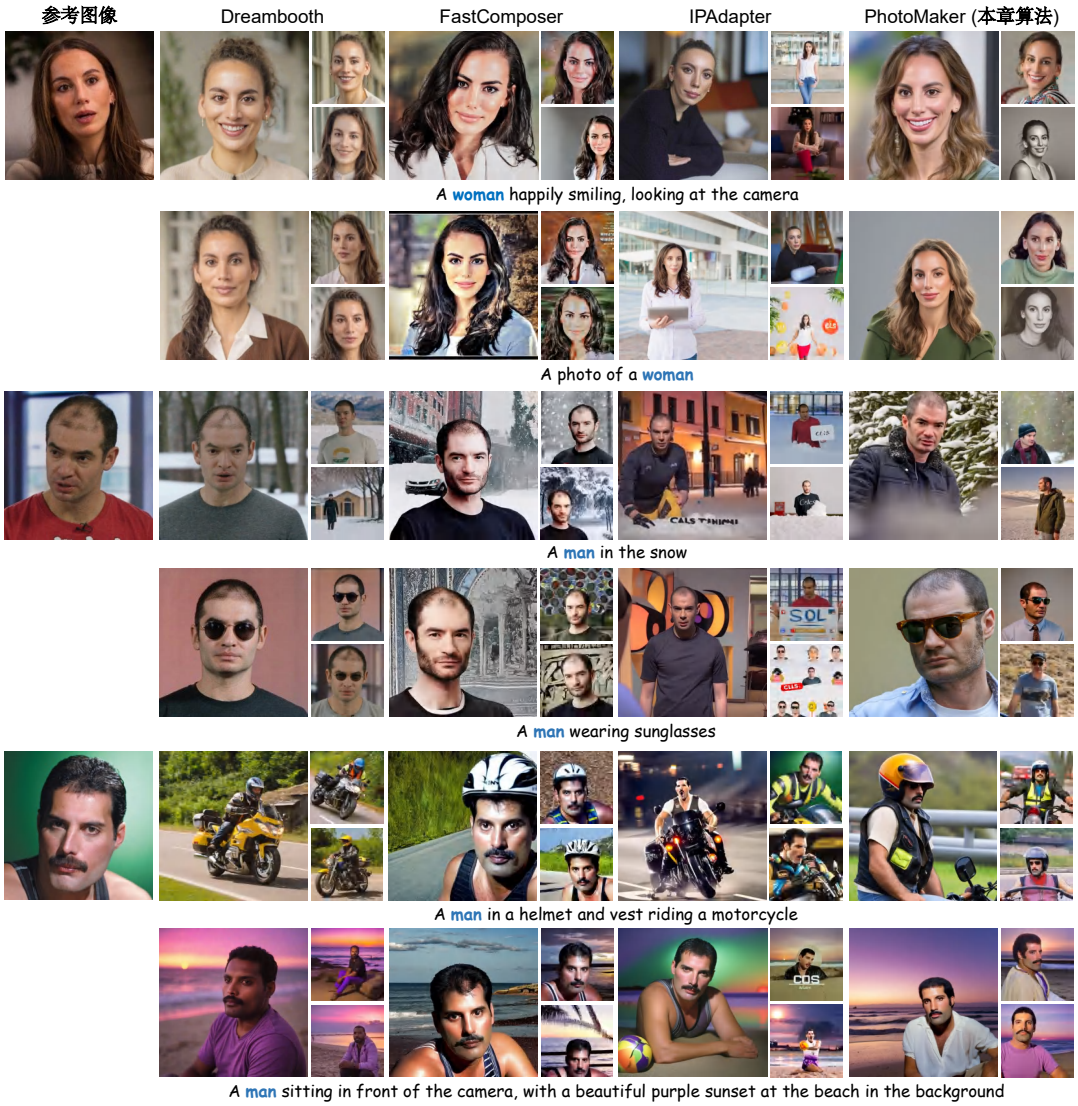

图 5.18: **文本重定义上下文应用的更多视觉示例**。本章提出的方法不仅提供了高 ID保真度，而且保留了文本编辑能力。作者为每个提示词随机采样三张图像。

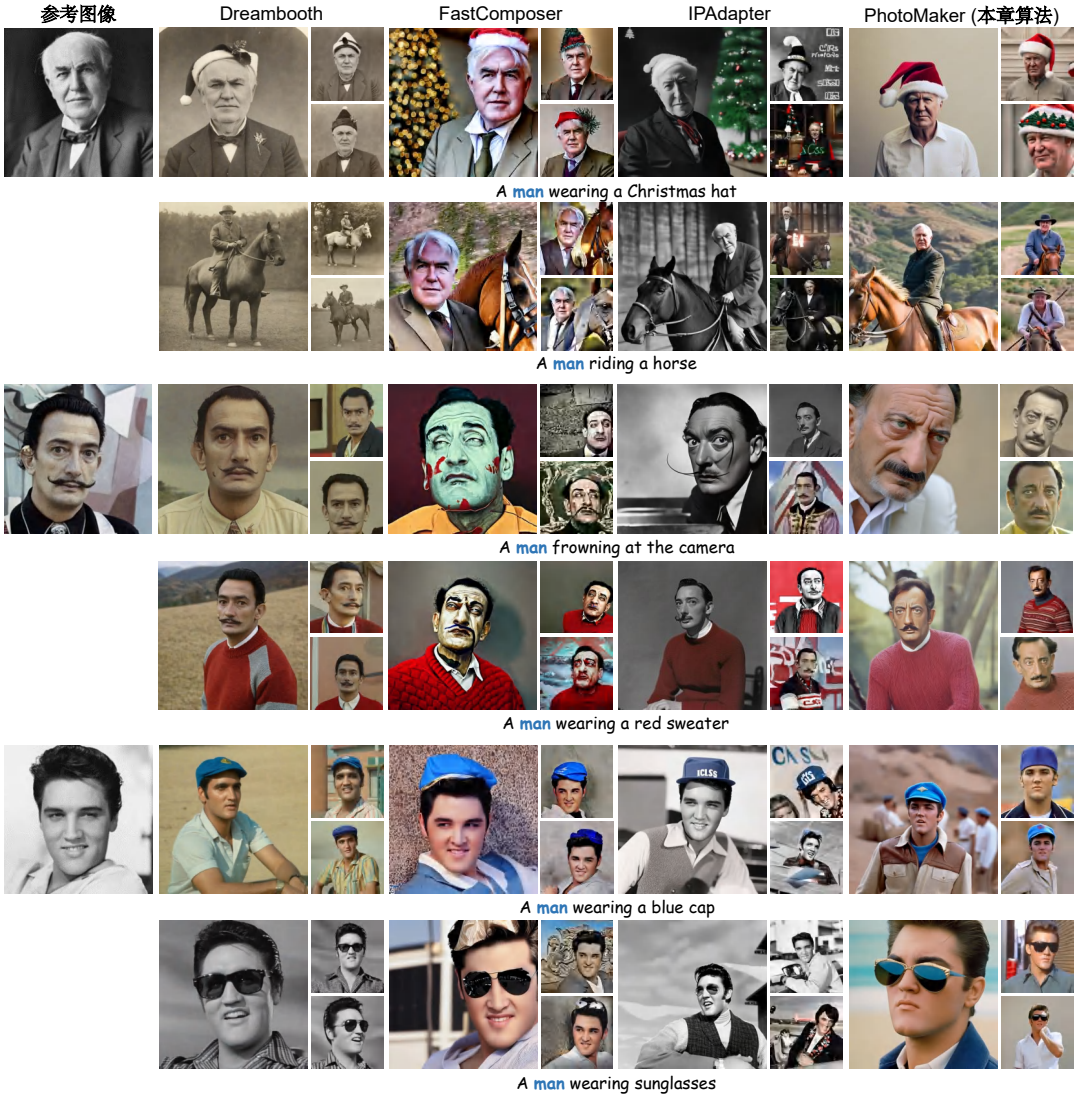

图 5.19: **将旧照片中的人物带回生活的更多视觉示例。** 本章提出的方法可以生成高质量的图像。作者为每个提示词随机采样三张图像。

图 5.20: **将艺术品中的人物带回生活的更多视觉示例。** 本章提出的 PhotoMaker可以生成照片般逼真的图像，而其他方法很难实现。作者为每个提示词随机采样三张图像。

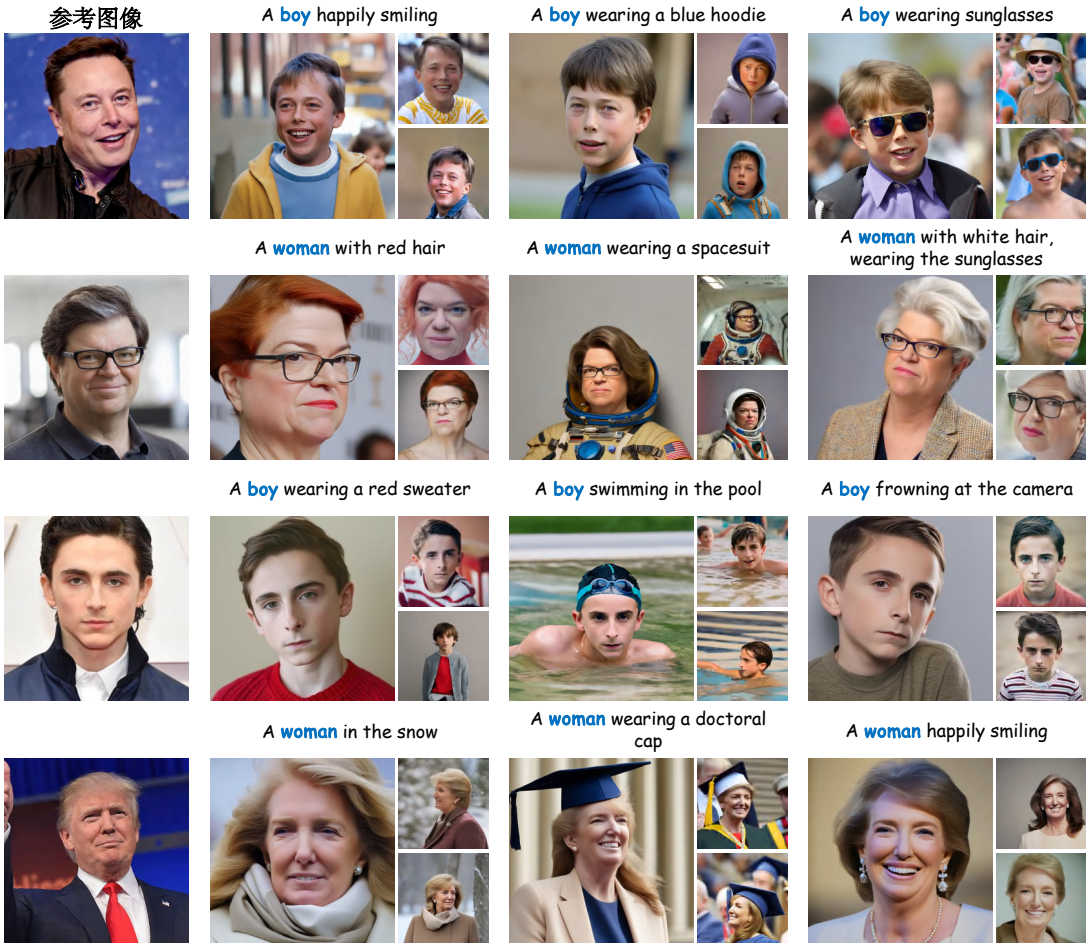

图 5.21: **改变每个 ID的年龄或性别的更多视觉示例。** 本章提出的 PhotoMaker在修改输入 ID的性别和年龄时，有效地保留了面部 ID的特征，并允许进行文本操作。作者为每个提示词随机采样三张图像。

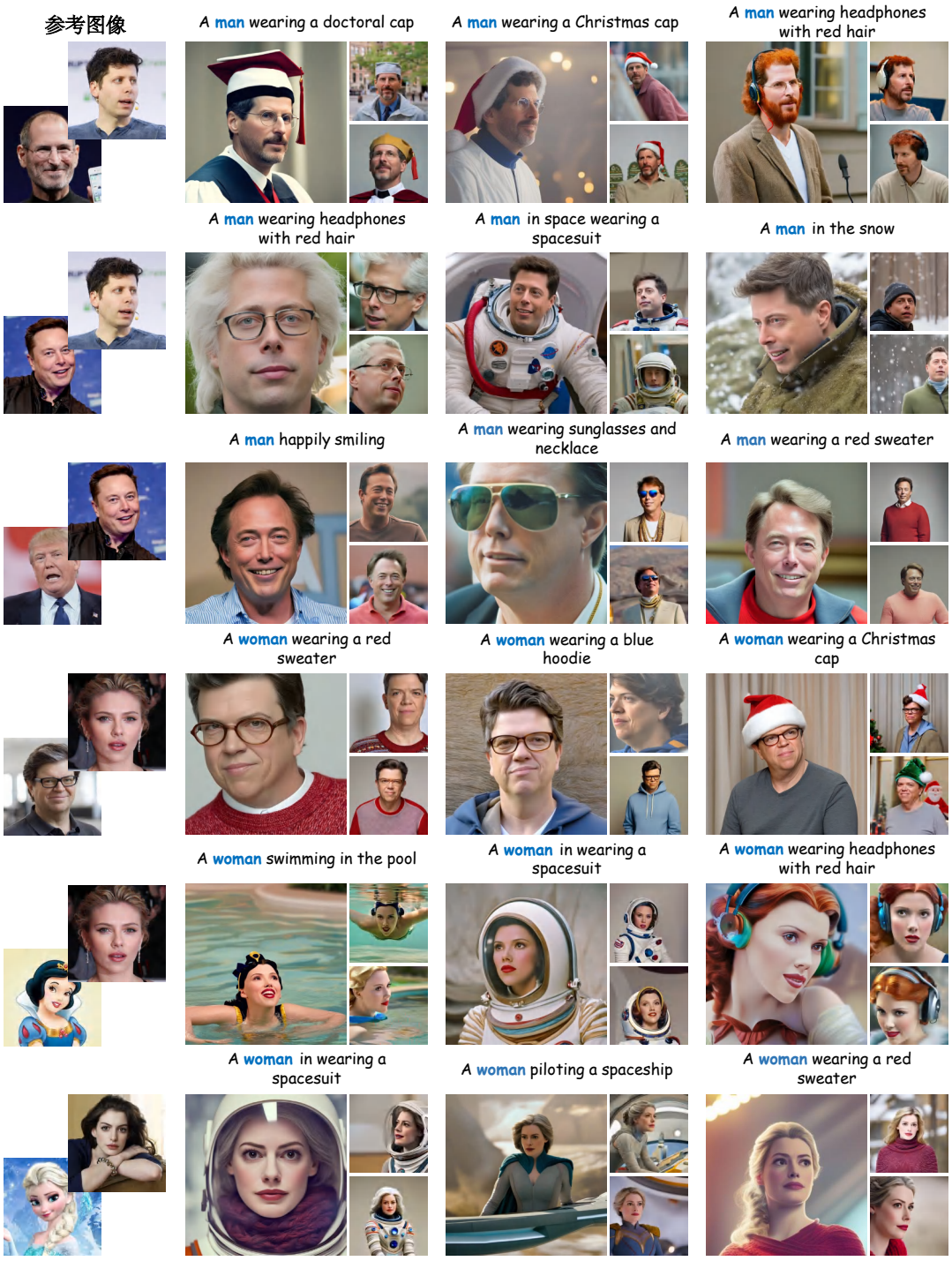

图 5.22: **身份混合应用的更多视觉结果。** 本章提出的 PhotoMaker可以在新生成的 ID图像中保持两个输入 ID的特征，同时提供高质量和文本兼容的生成结果。作者为每个提示词随机采样三张图像。

第六章 总结和展望

任务属性驱动的图像与视频生成技术，凭借其在满足特定需求和条件下生成高度定制化且真实度高的图像与视频内容的能力，对于内容创作、媒体娱乐、教育培训、安全监控等众多领域具有重大意义。这种方法不仅显著提升了生成内容的实用性和针对性，而且促进了计算机视觉和人工智能技术在图像与视频处理领域的应用与发展。本文围绕图像与视频生成的三个典型的子任务——视频补全、视频帧生成以及人物图像个性化定制——展开，针对每个任务的特定属性，即时空一致性、运动建模以及身份属性，分别提出了三种创新的研究方案。这些研究方案紧密依托于各自对应的任务属性，通过精心设计的方法不仅在性能上超越了当时世界上最先进的技术，同时也保持了良好的计算效率。通过这种任务属性驱动的方法论，本文的研究方案为图像与视频生成领域提供了新的视角和解决策略。这些方案的成功实施，不仅证明了任务属性驱动方法在提高生成内容质量和效率方面的有效性，也为未来在相关领域的研究和应用开辟了新的道路。

6.1 总结

本文首先介绍了任务属性驱动的图像与视频生成的研究背景和意义，并分析了图像与视频生成在特定任务上生成能力的局限性以及挑战。结合研究背景，本文确立了相应的研究目标并总结了主要的贡献。然后本文在第二章中结合本文的研究内容，对所涉及到的几个重要研究方向进行了介绍，并回顾了相关的研究工作。这些研究方向大致分为三个：视频补全技术、视频帧生成技术以及人物图像个性化技术。

在第三章，本文提出了一种时空联合一致性驱动的视频补全框架，名为E^2FGVI。该框架设计了三个可训练的模块：光流补全、特征传播和内容生成，这些模块模拟了基于光流的方法中的相应阶段，并通过它们之间的密切协作，减少了对中间结果的过度依赖，提高了生成效率。具体而言，光流补全模块直接输出补全后的光流，避免了多步复杂操作；特征传播模块在特征空间内通过可变形卷积进行，减轻了不准确光流的影响；内容生成模块引入了时空 Focal

Transformer，有效模拟了空间和时间维度上的长距离依赖关系，生成具有时间连贯性的补全结果。实验结果显示，E^{2}FGVI框架在多个评价指标上取得了显著改进，包括面向失真度的指标、面向感知的指标和时间一致性衡量指标。在效率上也表现出色，处理速度是之前基于光流方法的近15倍，同时具有最低的计算复杂性。这些优势使得 E^{2}FGVI框架成为视频补全领域的一个强有力的基线，为未来的研究开拓了新的技术路径。

在第四章，本文提出了一种运动属性驱动的视频帧生成框架，旨在克服先前工作在运动建模时的保真度和多样性方面的主要缺陷。该框架通过两种创新设计来提高运动建模的准确性和应对场景中的遮挡情况。首先，基于 RAFT中全对相关性的概念，该框架通过构建双向相关的匹配代价并引入放缩查找策略，有效模拟了帧间的密集对应关系，这种有效性尤其是在处理大位移时更为突出。这种方法不仅解决了坐标不匹配问题，还通过跨尺度方式联合更新双边光流和插帧内容特征，为光流的保真度提供了保障，并为后续的更加细化的运动建模奠定了基础。其次，为了应对场景中的遮挡问题，该框架从更新过的粗双边光流中导出多组细粒度光流，这些光流可以将输入帧反向扭曲到目标时间步长。通过为每个要插值的像素提供多个潜在值来源，该框架减轻了遮挡区域中的歧义性问题。在公共基准测试中，该框架与当前最先进的方法相比，不仅在性能上超越了现有的先进模型，同时在计算复杂度和参数量上也显示出显著的优势。因此，该框架为高效且准确的视频插帧技术提供了一个新的研究方向和基准。

在第五章，本文提出了一种身份属性驱动的人物图像个性化生成框架，名为 PhotoMaker。它能够接收多个输入人物相关图像并在语义级别上构造一个堆叠的 ID嵌入，作为生成定制人像的统一表示。这种设计允许模型在不需要额外训练的情况下，通过其交叉注意层自适应地集成 ID信息和文本嵌入，在保持输入身份（Identity，ID）信息的同时还提供了生成的多样性。 PhotoMaker展现了出色的生成效率，能够在大约10秒内生成定制人像，显著快于工业界中盛行的需要微调的方法。此外，它还提供了高度的可控性，允许用户改变输入人像的多种属性，甚至合并不同 ID的特征以创造全新的定制 ID。这一创新不仅展示了PhotoMaker的灵活性和高效性，也为定制人像生成领域提供了新的技术路径和研究可能性。为了支持 PhotoMaker的训练，作者还设计了一个自动化流程来构建一个与 ID相关的数据集，该数据集含有丰富的 ID人物、且每个人物都有多张对应的图像。每张图像都有与之对应的文本描述和分割掩码。这套数据流程可

以为未来的 ID驱动的相关研究作出一定铺垫。

6.2 展望

尽管本文在任务属性驱动的图像和视频生成方向取得了一定的研究成果，但仍然有许多方面值得进一步提升和研究。本节将详细介绍未来可能的研究以及发展方向：

- 第三章提出的时空一致性驱动的视频补全框架虽然在计算效率上相较于之前的方法有所提升，但在面对高分辨率和长视频时，其对计算资源的需求和处理速度的增加，限制了其在实时性任务或端侧部署的能力。因此，探索一种更高效的视频补全框架，以满足实时性需求，成为一个具有实践价值的研究方向。此外，图像补全算法作为视频补全的有效先验，其补全能力的转移对于视频补全来说是重要且有益的。一种可能的解决方案是利用图像补全结果作为初始化，通过设计网络模块来微调生成结果或特征的时空一致性。另一个研究方向是利用已训练的图像补全模型作为骨干，结合适当的模块（如可变形卷积或注意力机制）来聚合多帧特征。从数据角度来看，尽管本文仅使用了 Youtube-VOS [201]数据集进行模型训练，但目前趋势显示，使用更大规模的数据集可以获得更好的生成结果。因此，收集大规模高质量视频数据集，并设计适用于这些数据量的预训练方法，是未来研究的一个潜在方向。同时，SORA [256]的出现也预示着未来可能会有更鲁棒的大型视频补全算法被广泛应用。
- 第四章提出的运动属性驱动的视频帧生成框架有效地建模了广泛的运动范围，但随着输入帧分辨率的提高，其需要计算所有特征点的相似性，导致计算效率降低。因此，探索更高效的相关性计算方法成为一个重要的研究方向。尽管当前框架采用的是卷积架构，但鉴于 Transformer [73]在全局关系建模上的准确性，围绕 Transformer架构进行进一步的研究也具有潜力。此外，该框架在处理多帧生成时，需要对每一帧分别进行运动建模，而实际上待生成部分的运动具有一定的关联性。如何有效利用这种关联性进行更精确的多帧生成建模，也是一个值得探索的方向。同时，经过大规模训练的运动估计模型能为该任务提供有力的先验，帮助模型实现更鲁棒的帧生成。最近 SORA [256]的出现为帧生成任务展示了新的可能性，它不仅能够基于两个有关联的帧进行帧生成，还能够根据两个无关联的帧进行帧

生成，并确保所生成的帧在逻辑和物理上能够合理地串联成一个视频。这种能力为未来的帧生成任务提供了更广阔的探索空间和可能性。

- 第五章介绍的身份属性驱动的人物图像个性化生成框架虽然能够实现快速且高效的定制，但目前还需要用户上传多张图像以确保较高的 ID保真度。探索如何仅通过用户上传单张图像便实现高 ID保真度的定制化生成，代表了一个有潜力的研究方向。此外，尽管当前构建的 ID导向数据集规模相对有限，仅包含约一万个 ID和十万张图像，但扩大数据集规模以提升模型性能的可能性已在多项生成任务中得到验证。因此，利用该数据管线引入更多人物 ID和数据，以进一步提高模型在高 ID保真度方面的表现，是一个值得追求的目标。最近，DALLE3 [257]的研究表明，通过大型语言模型改写图像的文本描述能显著增强文本对生成图像的控制力。基于此，对数据集进行重构以扩展模型的文本控制能力也是一个可行的方案。此外，本文虽主要讨论了如何在生成图像时保持身份一致性，但将身份一致性的概念扩展到融合更多人物、生成更多物体类别、以及嵌入更多模态（如3D和视频）也是一个值得探索的领域。这一框架未来可以作为多模态生成和智能体控制生成中用户参与的重要组件，发挥更广泛的作用。

参考文献

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, ACM Communications (2020).

[2] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[3] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, 2017.

[4] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196 (2017).

[5] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: CVPR, 2019.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: CVPR, 2020.

[7] A. Sauer, K. Schwarz, A. Geiger, Stylegan-xl: Scaling stylegan to large diverse datasets, arXiv preprint arXiv:2201.00273 (2022).

[8] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: NeurIPS, 2020.

[9] A. Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: ICML, 2021.

[10] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning, PMLR, 2015.

[11] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, in: NeurIPS, 2021.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022.

[14] C. Barnes, E. Shechtman, A. Finkelstein, D. B. Goldman, Patchmatch, ACM TOG (2009).

[15] M. Bertalmio, G. Sapiro, Image inpainting, in: SIGGRAPH, 2000.

[16] M. Bertalmio, L. Vese, G. Sapiro, S. Osher, Simultaneous structure and texture image inpainting, in: CVPR, 2003.

[17] A. Criminisi, P. Perez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, TIP (2004).

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Generative image inpainting with contextual attention, in: CVPR, 2018.

[19] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. Huang, Free-form image inpainting with gated convolution, in: ICCV, 2019.

[20] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, M. Ebrahimi, Edgeconnect: Generative image inpainting with adversarial edge learning, in: ICCVW, 2019.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: CVPR, 2016.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Free-form image inpainting with gated convolution, in: ICCV, 2019.

[23] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: ICLR, 2021.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: ICML, 2021.

[25] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., Photorealistic text-to-image diffusion models with deep language understanding, in: NeurIPS, 2022.

[26] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: ICCV, 2023.

[27] S. Xie, Z. Zhang, Z. Lin, T. Hinz, K. Zhang, Smartbrush: Text and shape guided object inpainting with diffusion model, in: CVPR, 2023.

[28] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, et al., Imagen editor and editbench: Advancing and evaluating text-guided image inpainting, in: CVPR, 2023.

[29] C. Wang, H. Huang, X. Han, J. Wang, Video inpainting by jointly learning temporal structure and spatial details, in: AAAI, 2019.

[30] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, W. Hsu, Learnable gated temporal shift module for deep video inpainting, BMVC (2019).

[31] Y.-T. Hu, H. Wang, N. Ballas, K. Grauman, A. G. Schwing, Proposal-based video completion, in: ECCV, 2020.

[32] R. Xu, X. Li, B. Zhou, C. C. Loy, Deep flow-guided video inpainting, in: CVPR, 2019.

[33] C. Gao, A. Saraf, J.-B. Huang, J. Kopf, Flow-edge guided video completion, in: ECCV, 2020.

[34] S. Lee, S. W. Oh, D. Won, S. J. Kim, Copy-and-paste networks for deep video inpainting, in: ICCV, 2019.

[35] A. Li, S. Zhao, X. Ma, M. Gong, J. Qi, R. Zhang, D. Tao, R. Kotagiri, Short-term and long-term context aggregation network for video inpainting, in: ECCV, 2020.

[36] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, H. Li, Fuseformer: Fusing fine-grained information in transformers for video inpainting, in: ICCV, 2021.

[37] Y. Zeng, J. Fu, H. Chao, Learning joint spatial-temporal transformations for video inpainting, in: ECCV, 2020.

[38] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, W. Hsu, Free-form video inpainting with 3d gated convolution and temporal patchgan, ICCV (2019).

[39] D. Kim, S. Woo, J.-Y. Lee, I. S. Kweon, Deep video inpainting, in: CVPR, 2019.

[40] X. Zou, L. Yang, D. Liu, Y. J. Lee, Progressive temporal feature alignment network for video inpainting, in: CVPR, 2021.

[41] J. Kang, S. W. Oh, S. J. Kim, Error compensation framework for flow-guided video inpainting, in: ECCV, 2022.

[42] K. Zhang, J. Fu, D. Liu, Flow-guided transformer for video inpainting, in:

ECCV, 2022.

[43] K. Zhang, J. Peng, J. Fu, D. Liu, Exploiting optical flow guidance for transformer-based video inpainting, TPAMI (2024).

[44] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, H. Li, Decoupled spatial-temporal transformer for video inpainting, arXiv preprint arXiv:2104.06637 (2021).

[45] S. , G. Garcia-Hernando, A. Monszpart, M. Pollefeys, G. J. Brostow, M. Firman, S. Vicente, Removing objects from neural radiance fields, in: CVPR, 2023.

[46] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Communications of the ACM 65(1) 2021, 99–106.

[47] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, F. Zheng, Track anything: Segment anything meets videos, arXiv preprint arXiv:2304.11968 (2023).

[48] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: CVPR, 2023.

[49] S. Zhou, C. Li, K. C. Chan, C. C. Loy, Propainter: Improving propagation and transformer for video inpainting, in: ICCV, 2023.

[50] D. Ceylan, C.-H. P. Huang, N. J. Mitra, Pix2video: Video editing using image diffusion, in: ICCV, 2023.

[51] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, A. Dittadi, Diffusion models for video prediction and infilling, arXiv preprint arXiv:2206.07696 (2022).

[52] N. Cherel, A. Almansa, Y. Gousseau, A. Newson, Infusion: Internal diffusion for video inpainting, arXiv preprint arXiv:2311.01090 (2023).

[53] V. Voleti, A. Jolicoeur-Martineau, C. Pal, Mcvd-masked conditional video diffusion for prediction, generation, and interpolation, NeurIPS (2022).

[54] Z. Zhang, B. Wu, X. Wang, Y. Luo, L. Zhang, Y. Zhao, P. Vajda, D. Metaxas, L. Yu, Avid: Any-length video inpainting with diffusion model, arXiv preprint arXiv:2312.03816 (2023).

[55] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: CVPR, 2017.

[56] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video under-

standing, in: ICCV, 2019.

[57] Y.-H. Tsai, M.-H. Yang, M. J. Black, Video segmentation via object flow, in: CVPR, 2016.

[58] J. Cheng, Y.-H. Tsai, S. Wang, M.-H. Yang, Segflow: Joint learning for video object segmentation and optical flow, in: ICCV, 2017.

[59] X. Zhu, Y. Wang, J. Dai, L. Yuan, Y. Wei, Flow-guided feature aggregation for video object detection, in: ICCV, 2017.

[60] C. Godard, O. Mac Aodha, M. Firman, G. J. Brostow, Digging into self-supervised monocular depth estimation, in: ICCV, 2019.

[61] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, J. Kopf, Consistent video depth estimation, TOG (2020).

[62] T. Xue, B. Chen, J. Wu, D. Wei, W. T. Freeman, Video enhancement with task-oriented flow, IJCV (2019).

[63] K. C. Chan, X. Wang, K. Yu, C. Dong, C. C. Loy, Basicvsr: The search for essential components in video super-resolution and beyond, in: CVPR, 2021.

[64] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, J. Kautz, Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: CVPR, 2018.

[65] H. Lee, T. Kim, T.-y. Chung, D. Pak, Y. Ban, S. Lee, Adacof: Adaptive collaboration of flows for video frame interpolation, in: CVPR, 2020.

[66] T. H. Kim, M. S. Sajjadi, M. Hirsch, B. Scholkopf, Spatio-temporal transformer network for video restoration, in: ECCV, 2018.

[67] Y. Tian, Y. Zhang, Y. Fu, C. Xu, Tdan: Temporally-deformable alignment network for video super-resolution, in: CVPR, 2020.

[68] J. Pan, H. Bai, J. Tang, Cascaded deep video deblurring using temporal sharpness prior, in: CVPR, 2020.

[69] X. Wang, K. C. Chan, K. Yu, C. Dong, C. C. Loy, Edvr: Video restoration with enhanced deformable convolutional networks, in: CVPR Workshops, 2019.

[70] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, C. Xu, Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution, in: CVPR, 2020.

[71] G. Xu, J. Xu, Z. Li, L. Wang, X. Sun, M.-M. Cheng, Temporal modulation

network for controllable space-time video super-resolution, in: CVPR, 2021.

[72] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: CVPR, 2019.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[75] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: ICML, 2021.

[76] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, ICCV (2021).

[77] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, ICCV (2021).

[78] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal attention for long-range interactions in vision transformers, in: NeurIPS, 2021.

[79] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, arXiv preprint arXiv:2202.09741 (2022).

[80] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever, Generative pretraining from pixels, in: ICML, 2020.

[81] K. Desai, J. Johnson, Virtex: Learning visual representations from textual annotations, in: CVPR, 2021.

[82] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, arXiv preprint arXiv:2106.13230 (2021).

[83] M. Patrick, D. Campbell, Y. M. Asano, I. M. F. Metze, C. Feichtenhofer, A. Vedaldi, J. F. Henriques, Keeping your eye on the ball: Trajectory attention in video transformers, in: NeurIPS, 2021.

[84] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: ICML, 2018.

[85] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: ECCV, 2020.

[86] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, in: ICLR, 2021.

[87] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al., Rethinking attention with performers, ICLR (2021).

[88] B. Cheng, A. G. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, NeurIPS (2021).

[89] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: ICCV Workshops, 2021.

[90] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, S.-M. Hu, Attention mechanisms in computer vision: A survey, arXiv preprint arXiv:2111.07624 (2021).

[91] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, arXiv preprint arXiv:2107.00652 (2021).

[92] S. Meyer, O. Wang, H. Zimmer, M. Grosse, A. Sorkine-Hornung, Phase-based frame interpolation for video, in: CVPR, 2015.

[93] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, C. Schroers, Phasenet for video frame interpolation, in: CVPR, 2018.

[94] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive convolution, in: CVPR, 2017.

[95] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive separable convolution, in: ICCV, 2017.

[96] X. Cheng, Z. Chen, Video frame interpolation via deformable separable convolution, in: AAAI, 2020.

[97] T. Peleg, P. Szekely, D. Sabo, O. Sendik, Im-net for high resolution video frame interpolation, in: CVPR, 2019.

[98] S. Gui, C. Wang, Q. Chen, D. Tao, Featureflow: Robust video interpolation via structure-to-texture generation, in: CVPR, 2020.

[99] M. Choi, H. Kim, B. Han, N. Xu, K. M. Lee, Channel attention is all you need for video frame interpolation, in: AAAI, 2020.

[100] T. Kalluri, D. Pathak, M. Chandraker, D. Tran, Flavr: Flow-agnostic video representations for fast frame interpolation, in: WACV, 2023.

[101] Z. Shi, X. Xu, X. Liu, J. Chen, M.-H. Yang, Video frame interpolation transformer, in: CVPR, 2022.

[102] L. Lu, R. Wu, H. Lin, J. Lu, J. Jia, Video frame interpolation with transformer, in: CVPR, 2022.

[103] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, 2015.

[104] X. Xu, L. Siyao, W. Sun, Q. Yin, M.-H. Yang, Quadratic video interpolation, in: NeurIPS, 2019.

[105] Z. Chi, R. Mohammadi Nasiri, Z. Liu, J. Lu, J. Tang, K. N. Plataniotis, All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling, in: ECCV, 2020.

[106] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, D. Scaramuzza, Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion, in: CVPR, 2022.

[107] S. Niklaus, F. Liu, Softmax splatting for video frame interpolation, in: CVPR, 2020.

[108] P. Hu, S. Niklaus, S. Sclaroff, K. Saenko, Many-to-many splatting for efficient video frame interpolation, in: CVPR, 2022.

[109] Z. Liu, R. Yeh, X. Tang, Y. Liu, A. Agarwala, Video frame synthesis using deep voxel flow, in: ICCV, 2017.

[110] F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, B. Catanzaro, Unsupervised video interpolation using cycle consistency, in: ICCV, 2019.

[111] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, Y.-Y. Chuang, Deep video frame interpolation using cyclic frame generation, in: AAAI, 2019.

[112] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, J. Yang, Ifrnet: Intermediate feature refine network for efficient frame interpolation, in: CVPR,

2022.

[113] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, B. Curless, Film: Frame interpolation for large motion, in: ECCV, 2022.

[114] Z. Chen, Y. Chen, J. Liu, X. Xu, V. Goel, Z. Wang, H. Shi, X. Wang, Videoinr: Learning video implicit neural representation for continuous space-time super-resolution, in: CVPR, 2022.

[115] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, M.-H. Yang, Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement, TPAMI (2018).

[116] H. Sim, J. Oh, M. Kim, Xvfi: extreme video frame interpolation, in: ICCV, 2021.

[117] Z. Huang, T. Zhang, W. Heng, B. Shi, S. Zhou, Real-time intermediate flow estimation for video frame interpolation, in: ECCV, 2022.

[118] X. Jin, L. Wu, J. Chen, Y. Chen, J. Koo, C.-h. Hahm, A unified pyramid recurrent network for video frame interpolation, in: CVPR, 2023.

[119] S. Niklaus, F. Liu, Context-aware synthesis for video frame interpolation, in: CVPR, 2018.

[120] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, M.-H. Yang, Depth-aware video frame interpolation, in: CVPR, 2019.

[121] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, D. Scaramuzza, Time lens: Event-based video frame interpolation, in: CVPR, 2021.

[122] W. He, K. You, Z. Qiao, X. Jia, Z. Zhang, W. Wang, H. Lu, Y. Wang, J. Liao, Timereplayer: Unlocking the potential of event cameras for video interpolation, in: CVPR, 2022.

[123] J. Chen, Y. Zhu, D. Lian, J. Yang, Y. Wang, R. Zhang, X. Liu, S. Qian, L. Kneip, S. Gao, Revisiting event-based video frame interpolation, arXiv preprint arXiv:2307.12558 (2023).

[124] J. Park, C. Lee, C.-S. Kim, Asymmetric bilateral motion estimation for video frame interpolation, in: CVPR, 2021.

[125] J. Xin, W. Longhai, S. Guotao, C. Youxin, C. Jie, K. Jayoon, H. Cheul-hee, Enhanced bi-directional motion estimation for video frame interpolation, 2023.

[126] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, L. Wang, Extracting motion and

appearance via inter-frame attention for efficient video frame interpolation, in: CVPR, 2023.

[127] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, L. Wang, Extracting motion and appearance via inter-frame attention for efficient video frame interpolation, in: CVPR, 2023.

[128] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, Z. Liu, Deep animation video interpolation in the wild, in: CVPR, 2021.

[129] S. Chen, M. Zwicker, Improving the perceptual quality of 2d animation interpolation, in: ECCV, 2022.

[130] D. Danier, F. Zhang, D. Bull, Ldmvfi: Video frame interpolation with latent diffusion models, in: AAAI, 2024.

[131] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al., Make-a-video: Text-to-video generation without text-video data, arXiv preprint arXiv:2209.14792 (2022).

[132] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, S. Zhang, Modelscope text-to-video technical report, arXiv preprint arXiv:2308.06571 (2023).

[133] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, K. Kreis, Align your latents: High-resolution video synthesis with latent diffusion models, in: CVPR, 2023.

[134] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al., Lavie: High-quality video generation with cascaded latent diffusion models, arXiv preprint arXiv:2309.15103 (2023).

[135] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao, M. Z. Shou, Show-1: Marrying pixel and latent diffusion models for text-to-video generation, arXiv preprint arXiv:2309.15818 (2023).

[136] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, M.-M. Cheng, Towards an end-to-end framework for flow-guided video inpainting, in: CVPR, 2022.

[137] K. C. Chan, S. Zhou, X. Xu, C. C. Loy, BasicVSR++: Improving video super-resolution with enhanced propagation and alignment, in: CVPR, 2022.

[138] J. Lin, Y. Cai, X. Hu, H. Wang, Y. Yan, X. Zou, H. Ding, Y. Zhang, R. Timofte, L. Van Gool, Flow-guided sparse transformer for video deblurring, in: ICML,

2022.

[139] S. Yu, B. Park, J. Park, J. Jeong, Joint learning of blind video denoising and optical flow estimation, in: CVPR Workshops, 2020.

[140] X. Hu, Z. Huang, A. Huang, J. Xu, S. Zhou, A dynamic multi-scale voxel flow network for video prediction, in: CVPR, 2023.

[141] Z. Chi, R. M. Nasiri, Z. Liu, Y. Yu, J. Lu, J. Tang, K. N. Plataniotis, Error-aware spatial ensembles for video frame interpolation, arXiv preprint arXiv:2207.12305 (2022).

[142] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, M. Gelautz, Fast cost-volume filtering for visual correspondence and beyond, TPAMI (2012).

[143] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, A. Bry, End-to-end learning of geometry and context for deep stereo regression, in: ICCV, 2017.

[144] S. Jeon, S. Kim, D. Min, K. Sohn, Parn: Pyramidal affine regression networks for dense semantic correspondence, in: ECCV, 2018.

[145] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: CVPR, 2018.

[146] J. Hur, S. Roth, Iterative residual refinement for joint optical flow and occlusion estimation, in: CVPR, 2019.

[147] Z. Teed, J. Deng, Raft: Recurrent all-pairs field transforms for optical flow, in: ECCV, 2020.

[148] F. Zhang, O. J. Woodford, V. A. Prisacariu, P. H. Torr, Separable flow: Learning motion cost volumes for optical flow estimation, in: ICCV, 2021.

[149] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, H. Li, FlowFormer: A transformer architecture for optical flow, in: ECCV, 2022.

[150] J. Park, K. Ko, C. Lee, C.-S. Kim, Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation, in: ECCV, 2020.

[151] Z. Jia, Y. Lu, H. Li, Neighbor correspondence matching for flow-based video frame synthesis, in: ACM MM, 2022.

[152] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: Text-based real image editing with diffusion models, in: CVPR, 2023.

[153] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, in: CVPR, 2021.

[154] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, arXiv preprint arXiv:2111.02114 (2021).

[155] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, arXiv preprint arXiv:2210.08402 (2022).

[156] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al., Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis, arXiv preprint arXiv:2310.00426 (2023).

[157] W. Peebles, S. Xie, Scalable diffusion models with transformers, in: ICCV, 2023.

[158] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021.

[159] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, Openclip (2021).

[160] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, JMLR (2020).

[161] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, Y. J. Lee, Gligen: Open-set grounded text-to-image generation, in: CVPR, 2023.

[162] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, X. Qie, T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, arXiv preprint arXiv:2302.08453 (2023).

[163] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang, Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, arXiv preprint arXiv:2308.06721 (2023).

[164] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, Sdxl: Improving latent diffusion models for high-resolution image synthesis, arXiv preprint arXiv:2307.01952 (2023).

[165] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, in: CVPR, 2023.

[166] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion, in: ICLR, 2023.

[167] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, W. Zuo, Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation, in: ICCV, 2023.

[168] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, D. Cohen-Or, Designing an encoder for fast personalization of text-to-image models, arXiv preprint arXiv:2302.12228 (2023).

[169] Y. Zhou, R. Zhang, T. Sun, J. Xu, Enhancing detail preservation for customized text-to-image generation: A regularization-free approach, arXiv preprint arXiv:2305.13579 (2023).

[170] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, J.-Y. Zhu, Multi-concept customization of text-to-image diffusion, in: CVPR, 2023.

[171] S. Ryu, Low-rank adaptation for fast text-to-image diffusion fine-tuning, https://github.com/cloneofsimo/lora (2022).

[172] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, F. Yang, Svdiff: Compact parameter space for diffusion fine-tuning, in: ICCV, 2023.

[173] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, H. Zheng, Inserting anybody in diffusion models via celeb basis, in: NeurIPS, 2023.

[174] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, K. Aberman, Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, arXiv preprint arXiv:2307.06949 (2023).

[175] Y. Ma, H. Yang, W. Wang, J. Fu, J. Liu, Unified multi-modal latent diffusion for joint subject and text conditional image generation, arXiv preprint

arXiv:2303.09319 (2023).

[176] X. Jia, Y. Zhao, K. C. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, Y.-C. Su, Taming encoder for zero fine-tuning image customization with text-to-image diffusion models, arXiv preprint arXiv:2304.02642 (2023).

[177] W. Chen, H. Hu, Y. Li, N. Rui, X. Jia, M.-W. Chang, W. W. Cohen, Subject-driven text-to-image generation via apprenticeship learning, arXiv preprint arXiv:2304.00186 (2023).

[178] J. Shi, W. Xiong, Z. Lin, H. J. Jung, Instantbooth: Personalized text-to-image generation without test-time finetuning, arXiv preprint arXiv:2304.03411 (2023).

[179] J. Ma, J. Liang, C. Chen, H. Lu, Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning, arXiv preprint arXiv:2307.11410 (2023).

[180] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, H. Zhao, Anydoor: Zero-shot object-level image customization, arXiv preprint arXiv:2307.09481 (2023).

[181] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li, et al., Styledrop: Text-to-image generation in any style, in: NeurIPS, 2023.

[182] L. Chen, M. Zhao, Y. Liu, M. Ding, Y. Song, S. Wang, X. Wang, H. Yang, J. Liu, K. Du, et al., Photoverse: Tuning-free image customization with text-to-image diffusion models, arXiv preprint arXiv:2309.05793 (2023).

[183] G. Xiao, T. Yin, W. T. Freeman, F. Durand, S. Han, Fastcomposer: Tuning-free multi-subject image generation with localized attention, arXiv preprint arXiv:2305.10431 (2023).

[184] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: ICLR, 2022.

[185] D. Valevski, D. Wasserman, Y. Matias, Y. Leviathan, Face0: Instantaneously conditioning a text-to-image model on a face, arXiv preprint arXiv:2306.06638 (2023).

[186] Y. Yan, C. Zhang, R. Wang, Y. Zhou, G. Zhang, P. Cheng, G. Yu, B. Fu, Facestudio: Put your face everywhere in seconds, arXiv preprint arXiv:2312.02663

(2023).

[187] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, Instantid: Zero-shot identity-preserving generation in seconds, arXiv preprint arXiv:2401.07519 (2024).

[188] M. Ebdelli, O. Le Meur, C. Guillemot, Video inpainting with short-term windows: Application to object removal and error concealment, TIP (2015).

[189] S. W. Oh, S. Lee, J.-Y. Lee, S. J. Kim, Onion-peel networks for deep video completion, in: ICCV, 2019.

[190] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Generative image inpainting with contextual attention, in: CVPR, 2018.

[191] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, P. Pérez, Video inpainting of complex scenes, Siam journal on imaging sciences (2014).

[192] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: CVPR, 2016.

[193] D. Lao, P. Zhu, P. Wonka, G. Sundaramoorthi, Flow-guided video inpainting with scene templates, in: ICCV, 2021.

[194] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, TIP (2004).

[195] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-video synthesis, in: NeurIPS, 2018.

[196] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, M.-H. Yang, Learning blind video temporal consistency, in: ECCV, 2018.

[197] A. L. Maas, A. Y. Hannun, A. Y. Ng, et al., Rectifier nonlinearities improve neural network acoustic models, in: ICML, 2013.

[198] K. C. Chan, X. Wang, K. Yu, C. Dong, C. C. Loy, Understanding deformable alignment in video super-resolution, AAAI (2021).

[199] K. C. Chan, S. Zhou, X. Xu, C. C. Loy, Basicvsr++: Improving video super-resolution with enhanced propagation and alignment, CVPR (2022).

[200] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[201] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen,

T. Huang, Youtube-vos: Sequence-to-sequence video object segmentation, in: ECCV, 2018.

[202] A. Ranjan, M. J. Black, Optical flow estimation using a spatial pyramid network, in: CVPR, 2017.

[203] T. Zhou, S. Tulsiani, W. Sun, J. Malik, A. A. Efros, View synthesis by appearance flow, in: ECCV, 2016.

[204] J. Flynn, I. Neulander, J. Philbin, N. Snavely, Deepstereo: Learning to predict new views from the world's imagery, in: CVPR, 2016.

[205] Z. Li, S. Niklaus, N. Snavely, O. Wang, Neural scene flow fields for space-time view synthesis of dynamic scenes, in: CVPR, 2021.

[206] C.-Y. Wu, N. Singhal, P. Krahenbuhl, Video compression through image interpolation, in: ECCV, 2018.

[207] H. Zhang, Y. Zhao, R. Wang, A flexible recurrent residual pyramid network for video frame interpolation, in: ECCV, 2020.

[208] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015.

[209] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, Gmflow: Learning optical flow via global matching, in: CVPR, 2022.

[210] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, D. Metaxas, Global matching with overlapping attention for optical flow estimation, in: CVPR, 2022.

[211] F. L. Qiqi Hou, Abhijay Ghildyal, A perceptual quality metric for video frame interpolation, in: ECCV, 2022.

[212] P. Charbonnier, L. Blanc-Feraud, G. Aubert, M. Barlaud, Two deterministic half-quadratic regularization algorithms for computed imaging, in: ICIP, 1994.

[213] S. Meister, J. Hur, S. Roth, UnFlow: Unsupervised learning of optical flowwith a bidirectional census loss, in: AAAI, 2018.

[214] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402 (2012).

[215] C. Montgomery, Xiph.org video test media (derf's collection), in: Online,https://media.xiph.org/video/derf/, 1994.

[216] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2018.

[217] T.-W. Hui, X. Tang, C. C. Loy, LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation, in: CVPR, 2018.

[218] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, O. Wang, Deep video deblurring for hand-held cameras, in: CVPR, 2017.

[219] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.

[220] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[221] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: CVPR, 2017.

[222] Z. Ren, O. Gallo, D. Sun, M.-H. Yang, E. B. Sudderth, J. Kautz, A fusion approach for multi-frame optical flow estimation, 2019.

[223] S. Jiang, D. Campbell, Y. Lu, H. Li, R. Hartley, Learning to estimate hidden motions with global motion aggregation, in: ICCV, 2021.

[224] H. Xu, J. Yang, J. Cai, J. Zhang, X. Tong, High-resolution optical flow from 1d attention and correlation, in: ICCV, 2021.

[225] L. Lipson, Z. Teed, J. Deng, Raft-stereo: Multilevel recurrent field transforms for stereo matching, in: 3DV, 2021.

[226] Z. Teed, J. Deng, Raft-3d: Scene flow using rigid-motion embeddings, in: CVPR, 2021.

[227] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, Q. Xu, HumanSD: A native skeleton-guided diffusion model for human image generation, in: ICCV, 2023.

[228] X. Liu, J. Ren, A. Siarohin, I. Skorokhodov, Y. Li, D. Lin, X. Liu, Z. Liu, S. Tulyakov, Hyperhuman: Hyper-realistic human generation with latent structural diffusion, arXiv preprint arXiv:2310.08579 (2023).

[229] S. V. P. Ltd., Photo ai, https://photoai.com/, accessed: 2023-12-08 (2023).

[230] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, F. Wang, Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, in: CVPR, 2023.

[231] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, Toward characteristic-preserving image-based virtual try-on network, in: ECCV, 2018.

[232] Y. Nitzan, K. Aberman, Q. He, O. Liba, M. Yarom, Y. Gandelsman, I. Mosseri, Y. Pritch, D. Cohen-Or, Mystyle: A personalized generative prior, TOG (2022).

[233] A. Melnik, M. Miasayedzenkau, D. Makarovets, D. Pirshtuk, E. Akbulut, D. Holzmann, T. Renusch, G. Reichert, H. Ritter, Face generation and editing with stylegan: A survey, arXiv preprint arXiv:2212.09102 (2022).

[234] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: ICCV, 2015.

[235] M. Arar, R. Gal, Y. Atzmon, G. Chechik, D. Cohen-Or, A. Shamir, A. H. Bermano, Domain-agnostic tuning-encoder for fast personalization of text-to-image models, TOG (2023).

[236] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, D. Cohen-Or, Prompt-to-prompt image editing with cross attention control, in: ICLR, 2023.

[237] Huggingface, Prompt weighting, https://huggingface.co/docs/diffusers/using-diffusers/weighted_prompts (2023).

[238] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen, General facial representation learning in a visual-linguistic manner, in: CVPR, 2022.

[239] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, C. C. Loy, Mead: A large-scale audio-visual dataset for emotional talking-face generation, in: ECCV, 2020.

[240] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: FG, 2018.

[241] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: CVPR, 2020.

[242] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: CVPR, 2019.

[243] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: CVPR, 2022.

[244] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: ICML, 2023.

[245] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-

strength natural language processing in python (2020).

[246] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP, 2019.

[247] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.

[248] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, D. Lischinski, Break-a-scene: Extracting multiple concepts from a single image, in: SIGGRAPHAsia, 2023.

[249] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: ICCV, 2021.

[250] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: CVPR, 2015.

[251] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: NeurIPS, 2017.

[252] G. Parmar, R. Zhang, J.-Y. Zhu, On aliased resizing and surprising subtleties in gan evaluation, in: CVPR, 2022.

[253] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.

[254] D. Beniaguev, Synthetic faces high quality (sfhq) dataset (2022).

[255] Y. Alaluf, O. Patashnik, D. Cohen-Or, Only a matter of style: Age transformation using a style-based regression model, TOG (2021).

[256] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, A. Ramesh, Video generation models as world simulators, https://openai.com/research/video-generation-models-as-world-simulators (2024).

[257] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al., Improving image generation with better captions, Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf (2023).

致谢

在博士生涯结束之际，回顾过往几年，感慨良多。特此向各位致以最真诚的谢意。

本人衷心感激导师程明明教授以及邵秀丽教授在科研中和生活中对我的指导和帮助。程老师对我的信任和支持让我能探索更多科研的可能性，度过充实的博士时光。感谢实验室和实习期间同学们对我科研上的帮助和生活上的支持和陪伴。也感谢合作的郭春乐、侯淇滨老师以及王鑫涛给我的宝贵建议和帮助。特别感谢实验室的朋友们与我一起打桌游、打球、尝遍津南美食，让我的博士生活丰富多彩，充满欢声笑语。最后，感谢家人对我的无尽的爱护和无条件的支持，有你们是我莫大的幸运。

个人简历

李震，生于1994年10月31日。于2016年就读于四川大学电子信息学院，并于2019年获得电子与通信工程硕士学位。在2020年进入南开大学师从邵秀丽与程明明教授攻读计算机科学与技术博士学位，目前主要研究方向为图像以及视频的编辑和生成。

博士期间发表的学术论文：

1. **Zhen Li**, Mingdeng Cao, Xintao Wang, Zhongang Qi, , Ming-Ming Cheng, Ying Shan, PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding, In ***CVPR***, 2024. [Code]. (CCF-A类会议)
2. **Zhen Li***, Zuo-Liang Zhu*, Ling-Hao Han, Qibin Hou, Chun-Le Guo, Ming-Ming Cheng, AMT: All-pairs multi-field transforms for efficient frame interpolation, In ***CVPR***, 2023. [Code]. (CCF-A类会议)
3. **Zhen Li***, Cheng-Ze Lu*, Qianhua Qin, Chun-Le Guo, Ming-Ming Cheng, Towards an end-to-end framework for flow-guided video inpainting, In ***CVPR***, 2022. [Code]. (CCF-A类会议)
4. **Zhen Li***, Zeng-Sheng Kuang*, Zuo-Liang Zhu, Hong-Peng Wang, Xiu-Li Shao, Wavelet-based texture reformation network for image super-resolution, ***IEEE TIP***, 2022. [Code]. (SCI一区，CCF-A类期刊，影响因子11.041)
5. Zuo-Liang Zhu*, **Zhen Li***, Ruixun Zhang, Chun-Le Guo, Ming-Ming Cheng, Designing an illumination-aware network for deep image relighting, ***IEEE TIP***, 2022. [Code]. (SCI一区，CCF-A类期刊，影响因子11.041)
6. Yupeng Zhou, **Zhen Li**, Chun-Le Guo, Song Bai, Ming-Ming Cheng, Qibin Hou, SRFormer: Permuted self-attention for single image super-resolution, In ***ICCV***, 2023. [Code]. (CCF-A类会议)
7. Xin Jin*, Ling-Hao Han*, **Zhen Li**, Zhi Chai, Chunle Guo, Chongyi Li, DNF: Decouple and feedback network for seeing in the dark, In ***CVPR***, 2023. [Code]. (CCF-A类会议)

8. Gang Xu, Jun Xu, **Zhen Li**, Liang Wang, Xing Sun, Ming-Ming Cheng, Temporal modulation network for controllable space-time video super-resolution, In ***CVPR***, 2021. [Code]. (CCF-A类会议)
9. Chang-Bin Zhang*, Peng-Tao Jiang*, Qibin Hou, Yunchao Wei, Qi Han, **Zhen Li**, Ming-Ming Cheng, Delving deep into label smoothing, In ***IEEE TIP***, 2022. [Code]. (SCI一区，CCF-A类期刊，影响因子11.041)