视频预测综述：从确定性方法到生成方法

Ruibo Ming[1,2], Zhewei Huang[2], Zhuoxuan Ju[2,3], Jianming Hu[1],
Lihui Peng[1,*] and Shuchang Zhou[2,*]
[1]Tsinghua University
[2]Megvii Technology
[3]Peking University
mrb22@mails.tsinghua.edu.cn, hzwer@pku.edu.cn, nymphea@stu.pku.edu.cn,
{ hujm, lihuipeng } @mail.tsinghua.edu.cn, shuchang.zhou@gmail.com,

Abstract

视频预测是计算机视觉中的一项基本任务，旨在使模型能够根据现有视频内容生成未来帧序列。这项任务在各个领域得到了广泛的应用。在本文中，我们全面调查了该领域的历史和当代作品，包括使用最广泛的数据集和算法。我们的调查仔细研究了计算机视觉领域内视频预测的挑战和不断发展的前景。我们提出了一种新的分类法，以视频预测算法的随机性为中心。这种分类法强调了从确定性预测方法到生成性预测方法的逐步过渡，强调了方法的重大进步和转变。

1 介绍

视频预测任务的目标是在给定一系列历史帧的情况下预测未来的帧。它通常被归类为低级计算机视觉任务，因为它专注于像素值的直接处理。尽管有这种分类，但它与其他低级任务不同，它隐含地要求对场景动态和时间连贯性有更复杂的理解，这通常是高级视觉任务的特征。挑战在于设计能够有效实现这种平衡的模型，使用适量的参数来最小化推理延迟和资源消耗，从而使视频预测适用于实际应用。视频预测的这种独特地位表明，它在弥合计算机视觉中低级感知和高级理解之间的差距方面发挥着不可或缺的作用。目前的视频预测算法一般可以分为两类。一个类别需要扭曲参考帧（通常是最后一个观察到的帧）中的像素来构建未来的帧。然而，这组方法在模拟场景中物体的出现和消失（出生和死亡）时，本身就面临着困难。另一类包括从头开始生成新帧的方法。尽管这些方法有望捕捉物体动力学中的生死现象，但它们主要集中在像素级分布的建模上。因此，他们往往缺乏对潜在现实世界背景的综合理解，这对于创造性的预测能力至关重要。认识到这些局限性，我们提出了一种新的分类法，直观地总结在图 ?? 中，重点是算法的随机性。在 Section ?? 中，我们介绍了确定性算法，旨在基于确定性目标帧执行像素级拟合。然而，由于像素级指标鼓励模型对多个同样可能的结果进行平均，因此它们通常会产生模糊的输出。在 Section ?? 中，我们讨论了试图赋予模型在运动中进行随机预测的能力的算法。这包括将随机变量或分布引入确定性模型的方法，以及直接利用概率模型的方法。这种算法允许模型从运动分布中采样，鼓励生成明显偏离目标帧但仍然合理的预测。鉴于现有视频预测算法对高分辨率自然视频数据的创造性有限，难以预测包含许多生死现象的视频帧序列，我们在 Section ?? 中介绍了生成视频预测任务，该任务优先考虑在延长时间持续时间内生成合理的视频帧序列，而不是像素级精度。我们相信，长期视频合成的未来取决于视频预测和生成技术的协同整合。这种统一的方法将预测方法提供的上下文约束与对生成方法的增强理解相结合，在一个有凝聚力的框架内解决挑战。

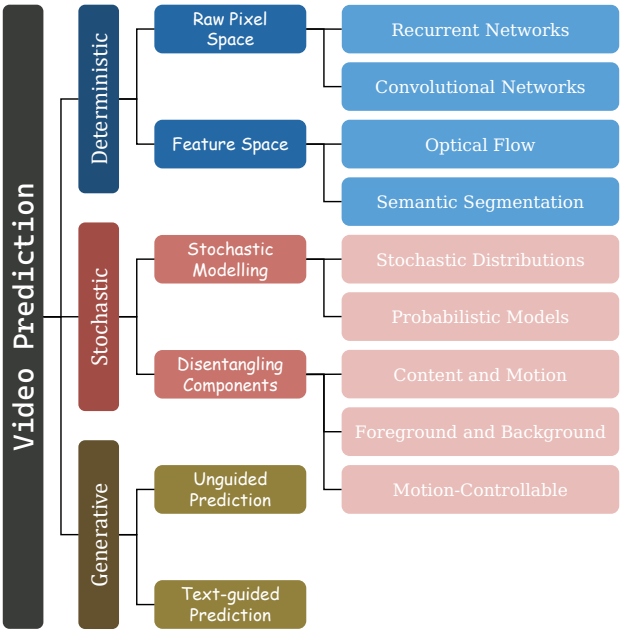

Figure 1: 视频预测算法分类概述

2 视频预测

2.1 问题定义

视频预测涉及根据对过去帧的分析来预测视频中的未来帧或序列。主要目标是开发能够准确预测视频序列中后续帧的视觉内容和可能运动的模型。这可以表述为条件生成建模问题，其中给定一系列观察到的帧 $X_{t_1:t_2}$ ，目

*Corresponding authors.

标是预测未来的帧 $Y_{t_2+1:t_3}$ 。

$$Y_{t_2+1:t_3} = X_{t_1:t_2} \cdot P(Y_{t_2+1:t_3}|X_{t_1:t_2}) \quad (1)$$

这里，t_1 表示初始时间步长，t_2 表示观察到的帧的最终时间步长，t_3 表示预测未来帧的最后一个时间步长。挑战在于学习一个映射函数，该函数可以捕获视频序列中复杂的时空依赖关系。

2.2 应用领域

视频预测的应用跨越了广泛的领域，展示了其在不同领域的重要性：这项任务在各个领域都至关重要，因为它使机器能够规划和响应动态环境。

自动驾驶。 视频预测对于自动驾驶汽车和无人机预测物体、行人和其他车辆的运动至关重要，从而确保安全导航。GAIA-1 [?] 采用统一的世界模型，结合多模态大语言模型和扩散过程来预测控制信号和未来帧。

机器人导航。 视频预测用于引导机器人通过动态环境，让它们有效地规划路径、移动物体和避开障碍物 [?] 。

电影业。 视频预测用于特效、动画和预可视化，帮助电影制作人创建逼真的场景并增强整体电影体验 [?] .

气象界。 视频预报在天气预报中起着至关重要的作用，帮助气象学家模拟和预测大气条件，有助于提高天气预报的准确性 [?]

2.3 挑战

在视频预测领域，存在长期存在的挑战，包括需要平衡低级像素处理和高层次场景动力学理解的算法、感知和随机评估指标的不足、实现长期预测的困难以及随机运动和生死现象的高分辨率数据集质量有限等。本节概述了视频预测领域的挑战，强调了持续进行研究和开发的必要性，以提高预测模型的准确性、可靠性和适应性。

高级指标。 广泛使用的低水平、确定性指标，如均方误差 (MSE)、峰值信噪比 (PSNR)、结构相似指数 (SSIM) 等，只能评估预测像素的像素精度，忽略了视频的视觉表现和不确定性。其直接结果是，模型倾向于对多个似是而非的结果进行平均，这通常会导致预测模糊，缺乏清晰度。研究人员越来越多地寻找替代方案，例如感知指标（DeePSiM [?] ，LPIPS [?] ）和随机指标（IS [?] ，FID [?] ），这些指标有望促进更具视觉吸引力和合理的预测。

长期预测。 虽然短期视频预测取得了重大进展，但由于动态场景中对象之间的长期依赖性和复杂的交互，在较长的时间范围内准确预测事件仍然具有挑战性。由于模型尺度不足以形成对现实世界的全面理解，现有的大多数视频预测模型主要对像素分布的变化进行建模。当面对长时间的自然视频时，这些模型很难在保持视觉质量的同时准确预测物体的运动。增强长期预测能力的一个有前途的方法是合并高级结构信息 [?] 。

普遍化。 数据量和模型复杂性之间的相互作用共同决定了算法性能的上限。尽管互联网上有大量的视频数据，但适合视频预测的高质量视频数据集的稀缺性仍然是一个限制因素，现有的数据集也带来了各种挑战，如数据分发简单、分辨率低、运动尺度小等。这些问题使得视频预测模型难以处理高分辨率内容和大型运动尺度，从而限制了其在多样化和看不见的场景中的实际实用性。实现高分辨率视频预测是一项复杂的任务，需要大量的计算资源 [?] 。由于计算负担过大，所涉及的复杂性使实时应用程序具有挑战性。

3 数据

为了评估视频预测模型的性能，我们不仅分析了它们的架构进步，还考虑了用于训练和测试的数据集的影响。视频预测模型的进步很大程度上取决于这些数据集的多样性、质量和特征。一般观察是数据集的适用性根据其维度和大小而变化，其中较低维度的数据集（通常具有较小的数据量）可能受到泛化性有限的影响。相比之下，高维数据集提供更广泛的数据，有助于增强模型的泛化能力。Table ?? 概述了视频预测中使用最广泛的数据集，突出了它们的数据量和其他监督模式，从而全面了解了该领域当前的数据集前景。如果原始论文或项目页面中没有报告特定的数据集详细信息，我们会计算平均值或中位数统计数据，以保持分析的一致性。

4 确定性预测

4.1 原始像素空间

视频预测的早期研究集中在原始像素空间建模上，这种方法简单明了，但计算要求很高。因此，它主要应用于小规模、低分辨率的数据集。

循环网络。 PredNet [?] 率先探索了视频预测中的循环神经网络，它从神经科学中的预测编码中汲取灵感，并采用循环卷积网络对视频特征进行有效处理。在此基础上，PredRNN [?] 通过修改具有双记忆结构的长短期记忆 (LSTM)，引入了显着的增强功能，旨在增强时空建模。尽管它取得了进步，但它在视频预测任务中遇到了梯度消失的挑战。为了解决这些局限性，ConvLSTM [?] 作为一个关键模型出现，巧妙地将 LSTM 与卷积神经网络 (CNN) 集成在一起，熟练地捕捉运动和时空动态，这一发展对后来的视频预测模型产生了重大影响。E3d-LSTM [?] 创新性地将 3D 卷积融入 RNN 中，并引入门控自注意力模块，从而显著提升了视频预测中的长期预测能力。

---

[1] https://www.csc.kth.se/cvap/actions
[2] https://data.caltech.edu/records/f6rph-90m20
[3] https://serre-lab.clps.brown.edu/resource/hmdb
[4] https://www.crcv.ucf.edu/data/UCF101.php
[5] http://jhmdb.is.tue.mpg.de
[6] http://www.cvlibs.net/datasets/kitti
[7] https://dreamdragon.github.io/PennAction
[8] http://medialab.sjtu.edu.cn/web4k/index.html
[9] https://cs.stanford.edu/people/karpathy/deepvideo
[10] https://www.cs.toronto.edu/ nitish/unsupervised_video
[11] https://www.cityscapes-dataset.com
[12] http://research.google.com/youtube8m
[13] https://sites.google.com/site/robotprediction
[14] https://davischallenge.org/davis2017/code.html
[15] https://developer.qualcomm.com/software/ai-datasets/sth-sth
[16] https://ogroth.github.io/shapestacks
[17] https://sites.google.com/view/svglp
[18] https://www.robonet.wiki
[19] http://toflow.csail.mit.edu
[20] https://maxbain.com/webvid-dataset
[21] https://github.com/JihyongOh/XVFI

Dataset	Year	Category	# Videos	# Clip Frames	Resolution	Extra Annotations
KTH Action[??]	2004	Human	2,391	95 *	160 × 120	Class
Caltech Pedestrian[??]	2009	Human	137	1,824	640 × 480	Bounding Box
HMDB51[??]	2011	Human	6,766	93 *	414 × 404 †	Class
UCF101[??]	2012	Human	13,320	187 *	320 × 240	Class
J-HMDB[??]	2013	Human	928	34 *	320 × 240	OF, Ins, HJ, Class
KITTI[??]	2013	Traffic	151	323 *	1242 × 375	OF, BBox, Sem, Ins, Depth
Penn Action[??]	2013	Human	2,326	70 *	480 × 270 †	Human Joint, Class
SJTU 4K[??]	2013	General	15	300	3840 × 2160	-
Sports-1M[??]	2014	Human	1,133,158	variable	variable	Class
Moving MNIST[??]	2015	Simulation	10,000	20	64 × 64	-
Cityscapes[??]	2016	Traffic	46	869 *	2048 × 1024	Semantic, Instance, Depth
YouTube-8M[??]	2016	General	8,200,000	variable	variable	Class
Robotic Pushing[??]	2016	Robot	59,000	25 *	640 × 512	Class
DAVIS17[??]	2017	General	150	73 *	3840 × 2026 †	Semantic
Something-Something[??]	2017	Object	220,847	45	427 × 240 †	Text
ShapeStacks[??]	2018	Simulation	36,000	16	224 × 224	Semantic
SM-MNIST[??]	2018	Simulation	customize	customize	64 × 64	-
RoboNet[??]	2019	Robot	161,000	93 *	64 × 48	-
Vimeo-90K[??]	2019	General	91,701	7	448 × 256	-
WebVid[??]	2021	General	10,732,607	449 *	596 × 336	Text
X4K1000FPS[??]	2021	General	4,408	65 *	4096 × 2160	-

* denotes the mean value. † denotes the median value.

Table 1: 使用最广泛的视频预测数据集的摘要，包括视频总数、每个视频的帧数、图像分辨率和其他注释等。(OF：光流，BBox：边界框，Sem：语义，Ins：实例，HJ：人体关节)

卷积网络。 卷积神经网络在视频预测技术的发展中发挥了重要作用。从 GDL [?] 开始，该领域取得了重大进展。在此之后，PredCNN [?] 在各种数据集上的表现优于其前身 [?]，从而建立了新的基准。SDC-Net [?] 引入了一种新颖的方法，即利用高分辨率视频帧预测技术，有效利用过去的帧和光流。在这些创新的基础上，最近推出的 SimVP [?] 标志着另一个里程碑。该方法重新审视了 ViT [?] 的进步，并引入了简化的 CNN 网络，证明了这种配置可以在视频预测中实现可比的性能。这一进展突显了卷积网络在视频预测中的应用的快速发展，凸显了向更高效和有效模型发展的持续趋势。

4.2 功能空间

在原始像素空间中进行预测通常会使模型负担过重，因为它们需要从头开始重建图像，这对于高分辨率视频数据集来说尤其具有挑战性。这种认识导致了研究人员关注点的转移。一些研究没有解决像素级预测的复杂性，而是转向了特征空间中的高级特征预测，例如光流和分割图。这些方法提供了一种更有效的方式来处理视频的复杂性 [?]。

光流预测。 视频预测领域通过创新方法取得了重大进展。FVS [?] 专注于提高预测质量，通过整合语义图、实例图和输入帧序列中的光流等补充信息，采用综合方法。这种方法虽然有效，但由于数据模态和计算需求的增加而带来了挑战。在相关的发展中，OPT [?] 的研究从照片级真实感视频插值的成功中汲取了灵感。它提出了一个专门为视频预测量身定制的优化框架，该框架基于高级插值算法的原理 [?]。基于这些概念，DMVFN [?] 代表了该领域的进一步发展。它扩展了密集体素流 [?] 的思想，并集成了可微路由模块。这一新增功能对于使模型能够更有效地捕捉和表示不同尺度的运动至关重要，展示了基于光流的视频预测方法的不断进步和改进。

未来的语义分割。 未来的语义分割代表了一种渐进的视频预测方法，主要侧重于预测即将到来的视频帧的语义图。该方法与传统的原始像素预测不同，转向语义图，以缩小预测范围并丰富场景理解。在这种情况下，S2S 模型 [?] 是一个开创性的端到端系统。它处理 RGB 帧及其语义图，作为输入和输出。这种集成不仅推动了未来的语义分割，而且提升了视频帧预测的任务，展示了语义级预测的独特优势。在此基础上，[?] 通过将光流与语义图合并来进一步创新。这种融合利用光流进行运动跟踪，利用语义图进行外观细节设计，利用前者对输入图像进行变形，后者对被遮挡的区域进行涂漆。

5 随机预测

在早期阶段，视频预测主要被视为低级计算机视觉任务，重点是采用确定性算法来增强像素级指标，如 MSE、PSNR 和 SSIM。然而，这种方法通过将可能的运动结果限制在单一的、固定的结果上，本质上限制了这些模型创造性输出的潜力 [?]。这通常会导致高像素级分数，但代价是产生模糊的图像，这大大降低了这些算法的实际适用性。认识到这个问题，视频预测领域已经发生了范式转变，从对短期确定性预测的依赖转向了长期随机预测。这种转变承认，虽然随机预测可能会产生与基本事实有很大差异的结果，但它在促进更全面的理解和增

强有关视频内容演变的创造性预测能力方面发挥着至关重要的作用。

5.1 随机性建模

对物体的不确定运动进行建模通常是通过将随机分布引入确定性模型或直接利用概率模型来实现的。

随机分布

在早期阶段，VPN [?] 使用 CNN 对基于像素分布的视频进行多重预测，而 SV2P [?] 通过对视频的随机分布估计来增强动作条件模型 [?] 。将焦点转移到视频元素的更整体视图上， [?] 提出了一种同时预测视频中语义分割、深度图和光流的概率方法。此外，SRVP [?] 使用常微分方程（ODE），而 PhyDNet [?] 使用偏微分方程（PDE）来计算随机分布。

概率模型

随着 [?] 奠定基础的开创性工作，对抗训练在预测不确定物体运动方面显着推进了视频预测任务。同样，vRNN [?] 和 GHVAE [?] 分别通过似然网络和层次结构增强 VAE，从而为随机预测方法的持续发展提供了另一个维度。认识到物体运动在很大程度上是确定性的，除非发生碰撞等不可预见的事件，SVG [?] 使用固定和可学习的先验对轨迹不确定性进行建模，这有效地融合了确定性和概率性方法。与此类似，但重点是增强时间方面， [?] 引入了一个序列鉴别器，旨在检测假帧。这种审查帧真实性的想法在 DIGAN [?] 中得到了进一步的扩展，其中重点转移到专注于识别非自然运动的运动鉴别器上。为了克服随机模型中的像素级预测挑战，一些工作引入了中间表示。S2S [?] 和 Vid2Vid [?] 将对抗性训练与未来的语义分割相结合。此外， [?] 利用 VAE 提取人类姿势信息，然后利用 GAN 来预测未来的姿势和帧。值得注意的是，直接对随机分布进行建模倾向于覆盖更广泛的预测分布，但视觉效果不佳。相比之下，概率模型能够产生更清晰的结果，但它们正在努力解决模式崩溃、训练困难和大量计算开销等问题。SAVP [?] 将这两种方法联系起来，将随机分布与对抗性训练相结合，旨在实现更广泛的预测分布和有希望的视觉质量。

5.2 解开组件

随机预测算法主要关注运动中的随机性。然而，这种方法往往忽略了视频中物体的生死现象。因此，许多研究将运动与其他视频元素隔离开来或人为地操纵其演变，从而更清楚地了解运动动力学，同时简化现实世界场景的复杂性。

内容和运动

视频预测算法通过强调复杂的图像细节来应对自然视频序列具有挑战性的复杂性。为此，他们专注于通过详细的本地信息有效地对外观进行建模，同时全面了解视频中固有的动态全局内容。然而，在机器人导航和自动驾驶等应用中，理解物体运动模式的重要性取代了对视觉美学的追求。这种优先级的转移鼓励了算法的发展，这些算法强调预测物体的运动，并将视频中的运动与外观区分开来。早期的工作 CDNA [?] 通过明确预测物体的运动开创了先例。它保持不变的外观特征，这有助于将模型应用于训练期间遇到的对象之外的对象。MoCoGAN [?] 自动学习以无监督的方式从内容中分离运动，而利用单独的内容和运动编码器路径的方法也已广泛应用于各种视频预测模型中。这种内容与运动分离的思想在 LMC [?] 中得到了进一步的探索，这使得运动编码器专注于基于残差帧的运动预测，而内容编码器则从输入帧序列中提取内容特征。MMVP [?] 采用不同的方法，仅使用一个图像编码器来提取信息，然后在图像解码器之前使用双流网络来分别处理运动预测和外观维护。AMC-GAN [?] 解决了运动的随机性，通过对抗性训练对多种合理的结果进行建模。过渡到不同的方法，SLAMP [?] 采用非对抗性方法，但专注于学习单独内容和运动的随机变量。为了进一步推进这一领域，LEO [?] 和 D-VDM [?] 利用扩散模型对内容和运动进行更逼真的解开，展示了这一方向的最新进展。

前景和背景

在视频预测领域，前景（物体）和背景（场景）的运动动态往往表现出鲜明的对比。前景对象通常表现出更强烈的运动，而场景往往保持相对静止。这种区别引导研究转向分别预测这些元素的运动，从而对视频动态提供更细致入微的理解。DrNet [?] 是这一领域的关键贡献，它专门解决了背景在视频帧中基本保持不变的场景。DrNet 专注于此类视频，无需学习复杂的场景动态，从而简化了预测过程。该模型巧妙地将图像分解为对象的内容和姿势。然后，它利用对抗性训练技术开发了一个场景鉴别器，用于评估两个姿势向量是否属于同一视频序列。

以人为本。 视频预测通常以前景运动为中心，尤其是当它涉及复杂的人体运动时。在这种情况下，各种专用数据集的一个常见假设是背景的相对静态性质，这是专注于详细人体运动的数据集的典型特征。这导致了对理解和预测人类姿势的研究的重点关注，以增强对前景运动的预测。这种方法的一个例子见于 [?] 的工作，它创新地预测骨架运动序列，然后使用骨架到图像转换器将这些序列转换为像素空间。他们的方法是一种有效的解决方案，将运动的抽象表示与视频预测的实际方面联系起来，展示了对以人为中心的视频预测任务所涉及的复杂性的细致入微的理解。

以对象为中心。 以对象为中心的视频预测（OCVP）领域专注于视频数据集，特别是旨在预测对象运动。这个概念最早是在 [?] 的著作中引入的，为视频预测这一专业领域奠定了基础。SlotFormer [?] 引入了基于 transformer 的自回归模型来学习视频序列中每个对象的表示。这项创新确保了对每个物体的一致和准确的跟踪。OKID 模型 [?] 代表了最近的进步。它采用 Koopman 算子，将视频独特地分解为不同的元素，特别是移动物体的属性和轨迹动力学。这种方法突出了一种更详细的分析视频序列中物体运动的方法，将其与以前的方法区分开来。

常规。 专注于人类姿势或物体的方法在特定的视频数据集中显示出相当大的前景，但由于它们依赖于预定义的结构并且难以适应可变背景，因此它们遇到了局限性，这阻碍了它们的泛化能力。这一挑战在性能上显而易见，虽然在某些条件下有效，但在面对动态背景变化时会步履蹒跚，这表明缺乏更广泛应用所需的多功能性。为了弥合这一差距，MOSO [?] 成为一种值得注意的方法，将运动、场景和物体确定为视频的关键元素。它通过区分场景和对象来更深入地研究内容分析。场景和对象被视为内容的进一步细分，其中场景表示背景，对象表示前景。MOSO 的创新贡献在于其为一般视频分析量身定制的两级网络。最初，MOSO-VQVAE 模型将视频帧分解为令牌级表示，通过视频重建任务磨练其功能。随后，该模型在第二阶段使用转换器，解决掩蔽令牌中的差异。

这种战略设计使模型能够在令牌级别处理各种任务，包括视频预测、插值、无条件视频生成等。

5.3 运动可控预测

在视频预测领域，出现了一个专注于运动显式控制的专业研究方向。这种方法的独特之处在于它强调根据用户定义的指令预测未来的物体位置，这与传统的依赖过去的运动趋势不同。该领域的核心挑战在于合成视频，使其遵守这些直接指令，同时保持自然和连贯的流程，这项任务需要对视频环境中的用户意图和运动动态进行细致入微的理解。这一挑战凸显了用户控制和自动预测之间错综复杂的平衡，标志着视频预测模型的概念化和实现方式发生了重大转变。

中风。 没有历史运动信息可用于从一张静止图像进行视频预测，因此出现了几种允许交互式用户控制的方法。iPOKE [?] 引入了一些技术，其中局部交互式笔触和戳使用户能够在一个静止图像中变形对象以生成一系列视频帧。这些笔触指示用户对对象的预期运动。遵循这一创新途径， [?] 介绍了一种交互式控制流体元素动画的方法。该领域的进步凸显了以用户为中心的方法在运动可控视频预测领域日益增长的重要性。

指示。 在旨在捕捉用户指定运动趋势的作品中，各种模态指令的整合，包括局部笔画、草图和文本，越来越普遍。VideoComposer [?] 通过组合文本描述、手绘笔触和草图来合成视频。这种方法尊重文本、空间和时间约束，利用视频潜在扩散模型和运动矢量进行明确的动态引导。从本质上讲，它可以生成与用户定义的运动笔触和形状草图一致的视频。同样，DragNUWA [?] 主要利用文本进行视频内容描述，并利用笔触进行未来的运动控制，从而实现可定制视频的预测。这些尝试通过扩大用户输入模式的范围来推进视频预测领域。

6 生成预测

在视频分析中，重点转向为生成视频预测而设计的算法，尤其是在处理表现出随机出生和死亡事件的视频时。这些事件在对象出现和消失时引入了不可预测性。这些算法需要对控制现实世界的基本物理原理有深刻的理解，以解决这种复杂性。他们不依赖于从历史框架推断的简单线性运动预测，而是通过复杂而富有想象力的建模技术来迎接挑战。因此，将单个静态图像转换为动态视频等任务（通常称为图像动画问题）成为应用生成视频预测技术的有希望的候选者。

6.1 非引导预测

在计算机视觉中，解决像素的非序列特性对 Transformer 架构提出了挑战。VQ-VAE [?] 和 VQGAN [?] 等创新技术结合了自回归模型和对抗性训练策略来解决图像量化问题。随后，VideoGPT [?] 和 LVM [?] 分别使用 VQ-VAE 和 VQGAN 构建自然视频预测模型。在大型语言模型（LLMs） [?] 兴起之前，转换器的变革性影响在时间序列建模中已经很明显。Video Transformer [?] 应用 Transformer 创建自回归模型，开创了 Transformer 架构在视频预测领域的应用。N Ü WA [?] 框架提出了一种通用的 3D Transformer 编码器-解码器架构，可适应各种数据模态和任务，进一步凸显了 transformer 在视频预测中的潜力。NUWA-Infinity [?] 通过创新的生成机制扩展了这一点，以实现无限高分辨率视频预测的雄心勃勃的目标。这一进展标志着我们不断努力统一不同模态的生成任务，展示了视频预测研究不断发展的前景。扩散模型 [?] 在图像生成领域占据主导地位。潜在扩散模型（LDM） [?] 探索了图像潜在空间的生成能力，显著提高了计算效率并降低了成本。LDM 为将扩散模型引入视频预测领域铺平了道路。LDM 在视频预测中的扩展表现出强大的生成能力 [?] 。此外，Video LDM [?] 利用预先训练的图像模型进行视频生成，提供多模态、高分辨率、长期的视频预测结果。同样，《SEINE [?] 》引入了一个通用的视频扩散模型，可以创建过渡序列，从较短的剪辑中生成更长的视频。

6.2 文本引导式预测

文本引导的生成式视频预测算法旨在通过集成上下文帧和文本引导来创建一系列帧。最近的例子包括 MMVG [?] 和 MAGVIT [?] 解决了新颖的文本引导视频完成（TVC）任务。该任务涉及根据各种条件完成视频，包括第一帧（视频预测任务）、最后一帧（视频倒带任务）或两者的组合（视频过渡任务），所有这些都由文本说明指导。这些模型利用自回归编码器-解码器架构，集成文本和帧特征，从而形成一个能够处理多个视频合成任务的统一框架。最近，旨在利用附加信息与 RGB 图像相结合来完成文本引导视频完成任务的研究工作激增。LFDM [?] 建立在潜在扩散模型的基础上，基于文本引导合成潜在空间中的光流序列。Emu Video [?] 采用了不同的方法，最初生成一个以文本指导为条件的图像，然后将其推断为视频。因此，该方法可以灵活地应用于基于具有不同文本输入的给定图像生成预测视频。DynamiCrafter [?] 将文本引导图像动画的应用范围扩大到开放域图像。SparseCtrl [?] 允许草图到视频生成、深度到视频生成和视频预测，并扩展输入范围。I2VGen-XL [?] 利用静态图像进行语义和定性指导，展示了文本引导视频预测研究的多样化方法。

7 未来研究

基于人工智能和计算机视觉领域的整体发展趋势，并考虑到第 ?? 节中讨论的视频预测领域的挑战，我们认为视频预测研究的发展轨迹应该与低级任务中数据简化和模型轻量化的传统趋势背道而驰。当务之急应该是利用大量的计算资源和高分辨率、长时间的视频数据。未来的研究应优先考虑制定评估指标，以激励随机预测的生成，从而扩大模型性能的潜力。最终目标是培养视频预测模型，深刻理解视频中的内在动态。这种模型可以在现实世界中以高随机复杂性在更长的时间范围内执行预测。

8 结论

在本次调查中，我们讨论了视频预测的各个方面，涵盖了主要数据集、不断发展的算法、当前挑战和未来趋势等关键主题。我们提出了一种基于算法随机性的新分类法，将视频预测分为确定性预测、随机预测和生成性预测，展示了从低级确定性方法向高级创意方法的重大转变。这项调查强调了在视频预测中平衡像素级精度与对复杂场景动态的深刻理解的重要性。此外，我们深入研究了随机生死现象的复杂性，倡导增强评估指标并利用大量计算资源和大规模视频数据集。这些见解旨在指导未来视频预测的研究方向。总之，视频预测研究至关重要，可能会对各种下游应用产生重大影响。随着该领域的发展，我们预计模型的发展将提供对现实世界中自然的更深刻和细致的把握。这种演变有望提高预测的准确

性、效率和创造力，为新的应用和研究铺平道路。这些进步将对计算机视觉和人工智能的更广泛领域做出重大贡献。

确认 这项工作得到了北京市自然科学基金 L231014 的支持。