基于无标签视频数据的深度预测学习方法综述

潘敏婷[1],王韫博[1],朱祥明[1],高思宇[1],龙明盛[2],杨小康[1]
(1. 上海交通大学人工智能研究院、人工智能教育部重点实验室,上海 201109; 2. 清华大学软件学院,北京 100084)

摘 要: 基于视频数据的深度预测学习(以下简称“深度预测学习”)属于深度学习、计算机视觉和强化学习的交叉融合研究方向,是气象预报、自动驾驶、机器人视觉控制等场景下智能预测与决策系统的关键组成部分,在近年来成为机器学习的热点研究领域. 深度预测学习遵从自监督学习范式,从无标签的视频数据中挖掘自身的监督信息,学习其潜在的时空模式表达. 本文对基于深度学习的视频预测现有研究成果进行了详细综述. 首先,归纳了深度预测学习的研究范畴和交叉应用领域. 其次,总结了视频预测研究中常用的数据集和评价指标. 而后,从基于观测空间的视频预测、基于状态空间的视频预测、有模型的视觉决策三个角度,分类对比了当前主流的深度预测学习模型. 最后,本文分析了深度预测学习领域的热点问题,并对研究趋势进行了展望.

关键词: 深度学习; 自监督学习; 计算机视觉; 视频预测; 有模型的视觉决策
中图分类号: TP389.1 **文献标识码:** A **文章编号:** 0372-2112(2022)04-0869-18
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211209

A Survey on Deep Predictive Learning Based on Unlabeled Videos

PAN Min-ting[1], WANG Yun-bo[1], ZHU Xiang-ming[1], GAO Si-yu[1], LONG Ming-sheng[2], YANG Xiao-kang[1]
(1. *MoE Key laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai* 201109, *China*;
2. *School of Software, Tsinghua University, Beijing* 100084, *China*)

Abstract: Deep predictive learning based on video data (hereinafter referred to as "deep predictive learning") is a research direction of deep learning, being interacted with computer vision and reinforcement learning. It is a key part of intelligent prediction and decision-making systems in weather forecasting, autonomous driving, robotics, and other scenarios, and has become a hot research field of machine learning in recent years. Deep predictive learning follows the self-supervised learning paradigm, using internal constraints from unlabeled video data to learn the underlying spatiotemporal patterns. In this paper, we review the existing deep learning techniques for predictive learning in detail. First, we summarize the research scope and application fields of deep predictive learning. Second, we present the datasets and evaluation metrics commonly used in this research field. Third, we summarize current mainstream deep prediction learning models from three perspectives: predictive models based on observation space, predictive models based on state space, and visual planning methods based on the predictive models. Finally, we discuss the hot issues and future research directions in the field of deep predictive learning.

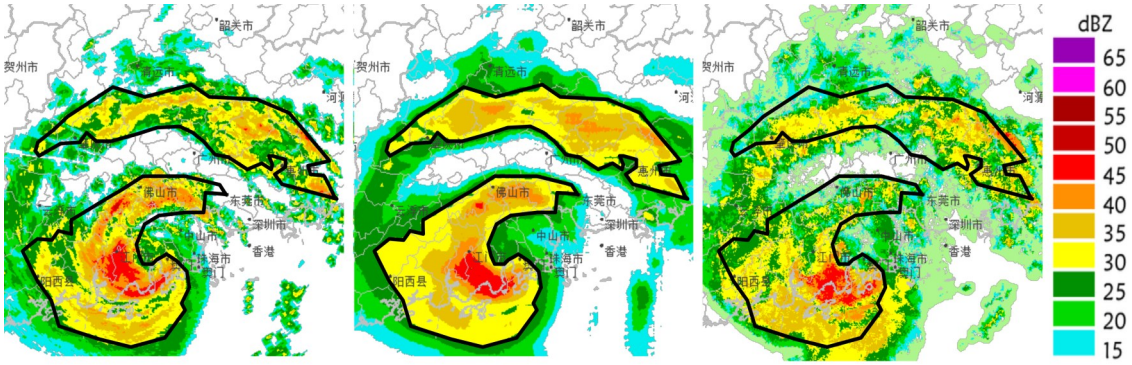
Key words: deep learning; self-supervised learning; computer vision; video prediction; model-based visual planning

1 引言

近年来,随着移动互联网、智能安防监控、时空数据采集与传感器网络等技术的迅猛发展,各行业中的视频数据体量呈指数级增长. 运用深度学习方法对海量视频数据进行建模,在无须额外人工标注的情况下理解其时空结构特性,对气象预报、自动驾驶、机器人视觉控制等若干场景下智能预测与决策系统具有重要意义,这使得基于无标签视频数据的深度预测学习(以下简称“深度预测学习”)成了近年来一个备受关注的研究领域. 预测学习的交叉应用场景众多,本文依照近年来国际学术界的主流研究成果,重点讨论其在计算机视觉和视觉决策场景下的具体内涵.

收稿日期:2021-09-01;修回日期:2022-02-17;责任编辑:王天慧
基金项目:国家自然科学基金(No.62106144, No.U19B2035);上海市科技重大专项(No.2021SHZDZX0102);上海市青年科技英才扬帆计划(No.21Z510202133)

首先，在计算机视觉的应用范畴下，预测学习的核心任务是指，基于一段连续的视频历史观测，预测其在未来一段时间范围内的变化．给定一个 n 帧视频序列 $(X_{t-n},\cdots,X_t)$，预测随后一段 m 帧视频序列 $(\hat{X}_{t+1},\cdots,\hat{X}_{t+m})$．利用深度学习模型，刻画观测空间中历史数据与未来数据之间确定性的映射关系，从而实现对未来时空变化趋势的高质量、精细化预测，已被成功应用于多种时空大数据平台中，其中包含短时临近强对流天气预报[1]、城市交通状况预测[2~4]等典型交叉应用场景．例如在气象短临预报中，需要根据前一时段内的雷达回波影像序列预测出未来 0~2 h 内每间隔 6 min 的雷达回波影像．在图 1 所展示的例子中，由清华大学团队主导研发的“新一代灾害性天气短时临近预报业务平台”首次将深度预测学习方法应用于中央气象台天气预报业务系统，表现出了超越传统数值模型与光流外插模型的预报水平，大幅提升了我国短临灾害性天气精细化预报能力，证明了深度预测学习具有广阔的交叉领域应用前景与重要的科学研究价值．



强台风天气在未来一小时后的雷达回波影像（真实实况）　深度预测学习天气预报产品（基于论文[31]的预测模型）　中央气象台原SWAN系统预报产品（基于非机器学习的光流外插算法）

图 1　深度预测学习在气象预报领域的应用示例，图中所示为从过去一小时雷达回波历史观测中预测到的未来一小时雷达回波．

此外，在许多基于时空观测信号的智慧决策系统中，视频预测模型也有着广阔的应用前景和商用价值．此类深度预测学习方法主要应用于机器人视觉决策任务[5~8]，其具体问题定义为给定 n 帧视频序列 $(X_{t-n},\cdots,X_t)$，以及机器人未来可能选取的动作序列 $(a_t,\cdots,a_{t+m-1})$，以视频帧 $(\hat{X}_{t+1},\cdots,\hat{X}_{t+m})$ 的形式，预测在相应未来时刻执行对应动作所可能产生的后果．此类预测模型的一种典型的应用场景是部分可见的马尔可夫决策过程（Partially Observable Markov Decision Process，POMDP）．在该问题中，场景的状态信息是不完全可知的，即视觉观测数据无法准确反应全部的物理机理，一些近期研究工作利用深度预测学习方法，在隐状态空间中融合机器人的动作信息与隐状态时空深度表征，刻画动作、状态、环境三者之间的动态关系．显然，提升预测模型的精度，可以有效改善下游视觉控制与决策任务的执行效果．

从交叉应用场景看预测学习的本质，视频数据作为一种典型的具有网格化空间结构的高维时间序列，其最大特点是在时间上具有长时非平稳趋势与非确定性趋势，同时在单一时刻又具有高维空间相关性（例如图 1 中的雷达回波影像）．传统的机器学习方法大多将时空数据当作多组单变量时间序列进行独立建模，其最大问题是特征学习能力不足，难以捕获空间相关性与非线性时空动态，故而难以形成长时、精细化的预测．深度预测学习遵循自监督学习的训练范式，不需要额外的标注信息，利用上述时空数据特性实现自监督训练，在无标签情况下建模数据中紧耦合的时间与空间相关性，从复杂、海量、高维、非线性的时空数据中挖掘重要的空间结构，并刻画其随时间的动态变化．预测学习模型与面向视频数据的生成模型不同．后者更关注生成数据的分布与真实数据分布的统计差异，而不需要严格保证生成结果相对观测数据的合理性；而前者相当于集成了因果推断模型和条件生成模型，不仅需要关注于观测空间中的生成质量，而且要尽可能地从历史观测中推断时空状态信息，因此需要更强的特征提取能力．在本文的后续讨论中，我们据此将主流的视频预测网络按照在观测空间或状态空间中的建模时空动态进行归纳对比．具体分类方式如图 2 所示．

本文第 2 节将归纳观测空间中的视频预测模型，主要包含基于卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Networks，RNN）的若干神经网络架构．第 3 节将总结基于语义状态空间或隐状态空间的深度预测网络，探究低维状态空间中的时空特征表达与解耦方法，以及基于此的长时预测方法和不确定性预测方法．第 4 节将归纳基于深度预测模型的视觉决策前沿方法，讨论如何结合预测学习提高交互环境中控制和决策水平．第 5 节将介绍该研究领域内的典型数据集和模型评价指

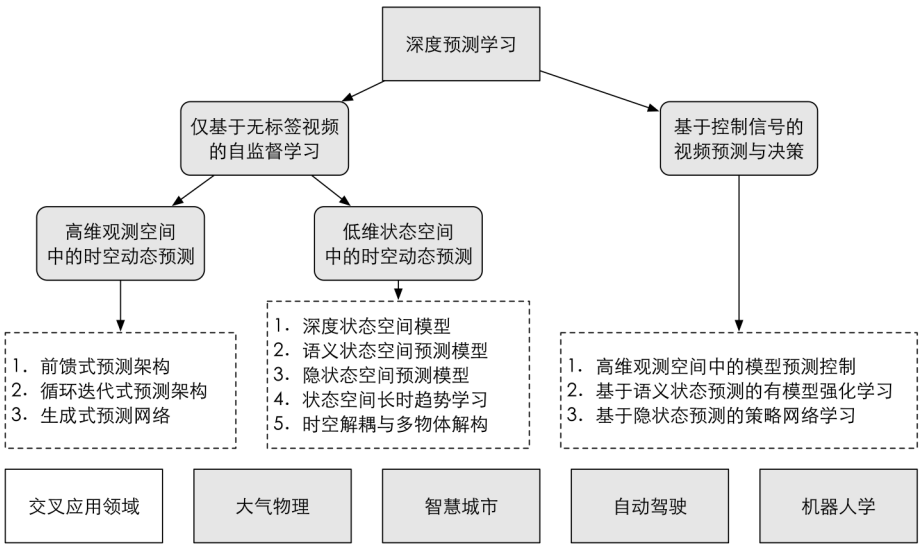


图2　深度预测学习主流方法的谱系图

标. 最后,第6节将讨论深度预测学习的开放问题与未来发展趋势.

2　高维观测空间中的预测学习方法

一类典型的视频预测模型是直接在观测到的像素空间中预测视频的未来序列,而没有将空间特征提取到低维度的状态空间并在低维空间中建模动态变化. 此类视频预测模型的优点是极大地保存了空间细节特征,有助于高精度的预测. 现有工作主要贡献是有效的高维时空特征表达,既要表示出帧内的空间信息,又要表示出帧间的时间信息,通过同时学习时空相关性,实现高质量的视频预测. 目前,观测空间中的时空预测模型可分为基于卷积神经网络的前馈式预测模型、基于循环神经网络的迭代式预测模型和生成式深度预测网络,表1基于这一分类对相关的模型进行了对比. 本节详细介绍这3类模型.

2.1　基于卷积神经网络的前馈式预测模型

卷积神经网络作为目前最主流的特征提取网络,被广泛应用于图像分类、语义分割、目标检测等计算机视觉领域. 卷积操作通过卷积核可以提取到图片中的

表1　观测空间中的视频预测模型对比(ED表示图像编码-解码架构,Adv表示对抗损失函数)

预测方法		年份	网络架构	损失函数	数据集	评价指标
基于卷积神经网络的前馈式预测模型	BeyondMSE[11]	2016	多尺度CNN	L1, Adv, GDL	Sports1M, UCF101	PSNR, SSIM
	Vukotic等[13]	2016	ED	L2	KTH	MSE
	DFN[14]	2016	ED	交叉熵	Moving MNIST	二值交叉熵
	Jin等[18]	2020	ED, GAN	L2, Adv, GDL	KTH, KITTI, BAIR, Caltech Pedestrian	PSNR, SSIM, LPIPS, FVD
	FutureGAN[22]	2018	3D-CNN, GAN	WGAN-GP[64]	Moving MNIST, KTH	MSE, SSIM
基于循环神经网络的循环迭代式预测模型	Ranzato等[28]	2014	rCNN	交叉熵	UCF101	MSE
	Srivastava等[30]	2015	LSTM	交叉熵	UCF101, Moving MNIST	交叉熵, 平方损失
	Shi等[31]	2015	ConvLSTM	交叉熵	Moving MNIST, Radar Echo	交叉熵, MSE
	TrajGRU[32]	2017	Trajectory GRU	有权重的L2	MovingMNIST++, HKO-7	CSI, HSS, B-MSE
	PredRNN[34]	2017	ST-LSTM	L2	Moving MNIST, KTH, Radar echo	MSE, PSNR, SSIM
	fRNN[35]	2018	bGRU	L1	Moving MNIST, KTH, UCF101	MSE, PSNR, SSIM
	E3D-LSTM[23]	2019	E3D-LSTM	L1, L2	Moving MNIST, KTH, TaxiBJ	MSE, PSNR, SSIM
	MotionRNN[36]	2021	ConvGRU	L1, L2, GDL	Human3.6M, Moving MNIST	MSE, SSIM, PSNR
	PredRNN-V2[63]	2021	ST-LSTM	L2	Moving MNIST, KTH, Radar echo, Traffic4Cast, BAIR Pushing	MSE, PSNR, SSIM, LPIPS
生成式深度预测网络	VideoGAN[25]	2016	3D-CNN,GAN	Adv	Two million videos from Flickr	Amazon Mechanical Turk
	TGAN[40]	2017	GAN	Adv	Moving MNIST, UCF101	IS (Inception Score)
	VideoFlow[45]	2020	Glow[46]	对数似然	BAIR Pushing	FVD, PSNR, SSIM

空间结构信息[9],但因为卷积核有限的感受视野,无法构建出远距离的空间依赖关系,有效的解决方法是堆叠更多的卷积层[10]或增大卷积核的尺寸. Mathieu 等人[11]提出多尺度建模的方式保持远程依赖关系,将不同分辨率的图片通过上采样对齐后输入到大尺度网络中. 为提取视频帧之间的时间依赖关系,Oh 等人[12]将连续的历史观测在通道维度上进行拼接后输入卷积层,经过编码和解码操作后直接得到预测的视频图像. Vukotic 等人[13]在编码网络和解码网络中加入时间变量,直接生成任意指定时刻的视频图像,不需要使用逐步迭代的方式进行多帧预测.

卷积网络建模时空特征的一大挑战是在视频预测的测试阶段,卷积核的参数往往是固定不变的,而不同的视频片段具有不同的运动模式,只使用单一变换的方法会造成预测失准. 为此,Brabandere 等人[14]、Xue 等人[15]和 Xu 等人[16]分别提出使用动态卷积网络,根据输入视频图片动态地改变卷积核的参数,由此预测出来的视频图片更加清晰、准确.

卷积网络建模时空特征的另一大挑战是其自身对时间动态信息的刻画能力有限,预测误差在长期范围内往往呈指数级增长. 为此,Jaderberg 等人[17]提出了空间转换模块,通过对卷积网络的特征图进行平移、旋转和缩放等空间变换建模物体运动. Jin 等人[18]在时间维度和空间维度上分别采用离散小波变换分解出不同频率的特征信息,解决视频预测时外观细节缺失和动作模糊的问题. Du 等人[19]提出用 3D 卷积网络处理连续的序列信息,在一定程度上解决了 2D 卷积网络无法表达时间维度的问题,可以在保留空间信息的同时有效提取到视频中的时间信息,有效提升了视频分类准确性[20,21]. 此后,3D 卷积网络也被使用到视频预测任务中[22~26].

因观测空间维度较高,观测空间中的深度预测网络需要特别考虑计算效率和预测质量. 与下文介绍的循环网络模型相比,大多数前馈式预测模型在大规模 GPU 上表现出更高的并行计算效率,但不善于构建远距离视频观测之间的长期依赖关系,在较长预测时效下的模型效果有待进一步提升.

2.2 基于循环神经网络的循环迭代式预测模型

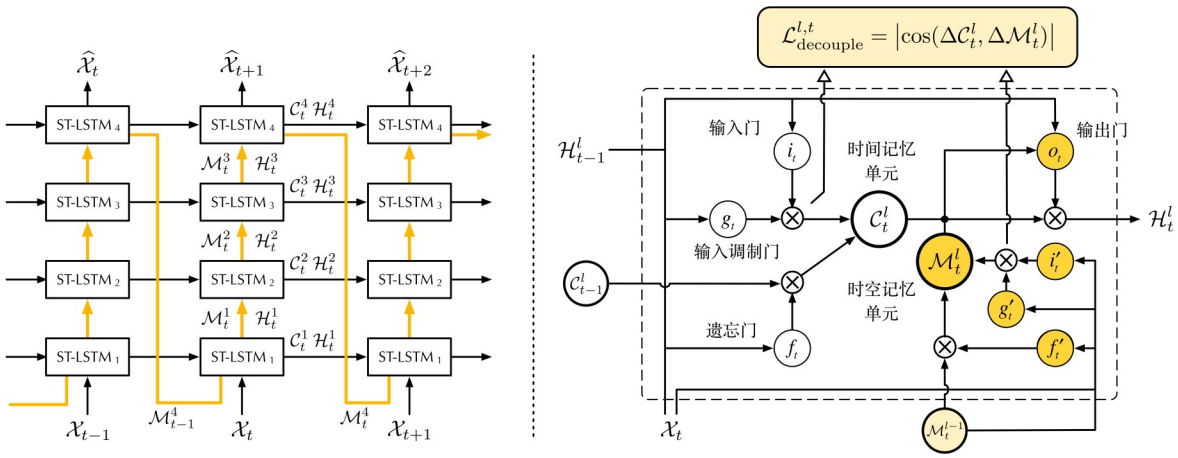
循环神经网络是为了处理序列数据而专门设计的,是自然语言处理等领域的关键技术. 循环神经网络的输出不仅依赖当前时刻的输入,还与历史时刻的网络状态有关,实现了信息记忆的功能. Graves[27]根据这一特性,结合长短时记忆单元(Long Short-Term Memory, LSTM)进行序列生成,提升了时序预测的预测长度. Ranzato 等人[28]首次将循环网络应用在视频数据上,实现了单帧预测. Sutskever 等人[29]开创性地提出了 LSTM 编码器-解码器框架,为序列到序列的学习任务提供了一个通用框架. Srivastava 等人[30]借鉴这一框架,将视频中的历史观测序列编码为一个固定长度的特征向量,并传递给 LSTM 解码器,进行未来视频多帧迭代预测. 上述模型所采用的 LSTM 层基于全连接算子,并没有考虑到对空间结构信息的学习. 为此,Shi 等人[31]提出了 ConvLSTM 网络,将 2D 的输入图片转换为 3D 张量,在 LSTM 内部的状态转移函数中采用卷积结构,通过堆叠多个 ConvLSTM 层形成最终的预测模型. 网络中某个单元格的未来状态是由其附近的输入和过去状态决定的,公式如下:

$$\begin{aligned} i_t &= \sigma(W_{xt} * X_t + W_{ht} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot C_{t-1} + b_o) \\ H_t &= o_t \odot \tanh(C_t) \end{aligned} \tag{1}$$

其中,X_t表示输入数据;C_t表示存储单元状态;H_t表示输出的隐状态;i_t, f_t, o_t分别表示输入门、遗忘门和输出门;*表示卷积操作;⊙表示哈达玛乘积.

ConvLSTM 模型中的卷积循环结构具有位置不变性,而自然界中的运动和转换通常是位置变化的. 因此 Shi 等人[32]提出了 Trajectory GRU(TrajGRU)模型,它可以学习到神经元动态的连接方式,比固定连接的 Convolutional GRU(ConvGRU)[33]更加灵活. Wang 等人[34]提出 PredRNN 模型,用时空长短时记忆单元(ST-LSTM)代替传统的 LSTM 单元,该单元所特有的记忆状态以“之”字形进行更新,如图 3(a)所示,信息首先跨层向上传递,然后随时间向前传递. 图 3(b)给出了 ST-LSTM 的内部详细结构,相比传统的 LSTM 单元,增加了一个输入调制门g_t和一个时空记忆状态M_t^l,在同一时间步长中将信息从l-1 层垂直传输到当前节点. 在 PredRNN-V2[63]中,两种记忆状态的增量信息经由损失函数被显式分离.

进一步地,Wang 等人[23]提出了 E3D-LSTM 模型,将 3D 卷积整合到 LSTM 网络中,可以让模型存储更有用的视频短期特征和建立局部依赖关系,而对于长期关联,通过一个门控制的自注意力模块,让当前的存储状态与其历史信息相互作用,由此同时提高了短期依赖和长期关联的特征表达. Oliu 等人[35]引入双向映射门控循环单元(bijective Gated Recurrent Units, bGRU),把 GRU 堆叠起来进行双向映射,在编解码的过程中参数是共享的,与标准的 GRU 网络不同,网络的输入是上一层的状态值,并使用一组额外的逻辑门来更新其输出,从而在多帧预测时避免了重新编码的过程,减小了误差的传播. 上述模型的相关信息在表 1 中列出. Wu 等人[36]则关注运动本身复杂的时空变化,物理世界的运动可以自然地分解为瞬态变化和运动趋势,使用 MotionRNN 结构分别对视频中物体的瞬时变化和长时运



(a) PredRNN[34]与PredRNN-V2[63]的网络框架　　(b) 核心卷积循环网络单元ST-LSTM的内部详细结构

图3 PredRNN[34]与PredRNN-V2[63]的整体网络结构,及其核心时空长短时记忆单元ST-LSTM的内部结构

动趋势进行建模.

2.3 生成式深度预测网络

在许多真实场景下,基于给定的历史观测,未来的视频序列具有多种可能性,使用均方误差损失函数(l_2 loss)训练的确定性模型[28,27],会回归每个像素点上所有可能像素强度的平均值,造成预测图片的模糊. 鉴于生成式对抗网络(Generative Adversarial Network,GAN)[37]的成功,对抗训练也被用来消除视频预测的歧义性. Mathieu 等人[11]提出基于条件生成式对抗网络(Conditional GAN,CGAN)[38]的 BeyondMSE 模型,验证了使用对抗损失函数可以得到比使用l_2损失函数更加清晰的视频预测结果. CGAN 在视频预测中的网络结构如图 4 所示. 同时训练相互对抗的预测网络和判别网络. 预测网络的训练目标包含两部分,一部分是尽量逼近预测目标的真实值,另一部分是尽量"欺骗"判别网络,使其输出错误的分类结果;而判别网络的训练目标是区分真实数据和预测数据. 通过对抗训练,可以有效剔除掉那些明显错误的预测结果.

近五年来,更多的预测学习模型采用对抗训练方式[25,39~43]提高生成图像序列的清晰度. 其中 Vondrick 等人[25]提出 VideoGAN,用两路卷积网络分别对视频的前景和背景进行生成. TGAN[40]使用相似的神经网络架构,将上述CGAN的预测网络分解为时间生成器和图像生成器. 采用对抗损失函数的一大挑战是训练的稳定性和模式崩塌问题,即生成器无法覆盖多种可能的时空特征模式,在训练过程中会收敛到单一的模式状态[43]. 现有的对抗预测网络的另一个尚未完全解决的问题是,对气象应用中的雷达回波序列、城市计算应用中的交通热力图序列的时空变化的捕捉能力不足. 这些场景要求精测的精细化程度高,即看重像素级预测结果的绝对准确度,对抗损失更关注图像全局质量,往往在MSE等像素级评价指标上表现不佳. 另一种典型的生成模型是变分自编码器(Variational Autoencoders,VAE). Denton 等人[44]提出了一种改进的 VAE 的模型,可以在更长的时间视野中合成未来视频帧. 虽然

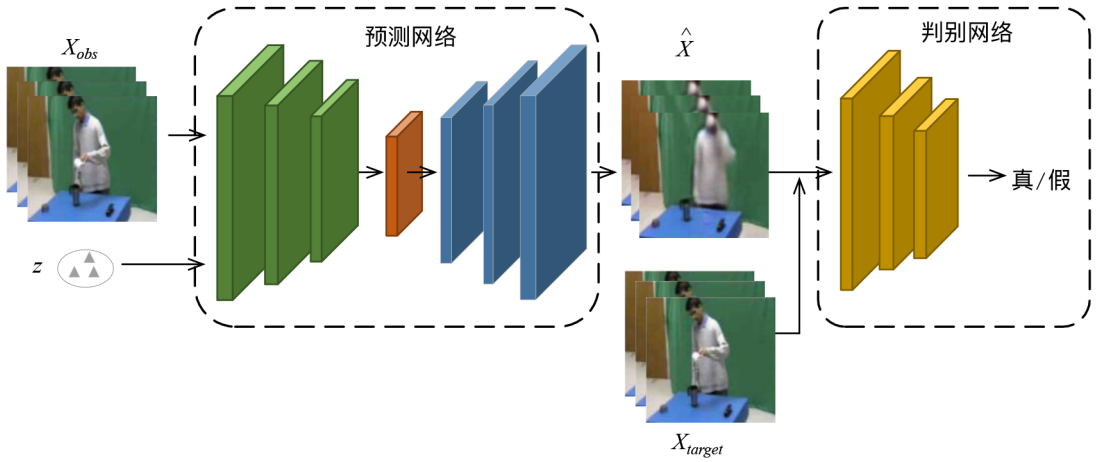


图4 CGAN应用于视频预测问题的通用网络结构

这些模型可以对未来分布进行建模,但预测分布仍然是在像素空间上,这使得模型倾向于产生模糊的预测.在生成式对抗网络和变分自编码器之外,Kumar等人[45]提出了VideoFlow,拓展了基于流的生成模型(Flow-Based Generative Model)[46,47],直接优化数据似然(data likelihood),生成具有随机性的高质量未来视频序列.

3 低维状态空间中的预测学习方法

视频图像中的数据特征具有高维度、高复杂的性质,直接在观测空间中对未来分布进行建模是一项具有挑战性的任务,且观测空间中的特征虽然可以保留很多的细节,但存在大量的冗余信息,增加了计算的复杂度.一个可行且有效的办法是,将低维表示的时空动态特征从视频序列中抽离出来,在有外部语义的状态空间或隐空间中进行状态前向传递的学习.对低维动态变量进行建模,一方面能够消除图像解码与时间动态之间的强耦合,从而更灵活地学习长序列相关性和时空动态随机性;另一方面通过有效降低特征维度,去除了冗余信息,提升了计算效率.表2总结了基于状态空间的视频预测模型.

3.1 状态空间模型

状态空间模型(State Space Model,SSM)是描述动态系统的完整模型,根据输入改变系统内部状态,并由此改变系统的输出,常被用于序列建模[4,48-53].对于观测到的n个序列数据$(X_{t-n},\cdots,X_t)$,状态空间模型假设每一个数据X_t都是依据隐藏状态H_t生成的,这些连续或离散的隐藏状态形成了一个马尔可夫过程.SSM的通用表达式为[54]

$$\begin{aligned} H_t&=g(H_{t-1},\epsilon_t)\\ X_t&=f(H_t,\sigma_t)\end{aligned} \tag{2}$$

其中,g表示状态转换模型,f表示观测模型,ϵ_t和σ_t分别表示状态转换模型和观测模型中的噪声.很多经典的时间序列模型,如马尔可夫模型、隐马尔可夫模型、深度隐马尔可夫模型和向量自回归模型等,都可以表示为状态空间模型.给定潜在的隐藏状态H_t,观察值X_t与历史信息F_t是独立的.写出后L步序列的后验分布如下:

$$p(X_{t+1:t+L}\mid F_t)=\int p(H_t\mid F_t)\prod_{i=t+1}^{t+L}p(X_i\mid H_i)p(H_i\mid H_{i-1})\mathrm{d}H_{t:t+L} \tag{3}$$

由上式可知,在解决时空序列预测问题上,SSM有其自身优势,贝叶斯公式使其可以自然地模拟动力系统的不确定性,编码不同未来结果的概率.

3.2 基于语义状态空间的深度预测模型

为了应对特征维度过高的问题,在语义分割、实例分割和人体姿态等计算机视觉任务中,常将观测空间降维为高级特征表示.在预测学习中,同样可以对具有特定语义信息进行表达和学习,基于外部语义先验来提高预测图片的质量.语义分割[55]和实例分割[56]将动态视觉场景分解为语义实体,如行人、车辆和障碍物等,场景动态由像素级建模变为语义实体建模,对动态场景理解具有重要意义.其中一个具有代表性的方法是Jin等人[57]提出的PEARL框架.该框架执行两个互补的预测学习任务:其一是使用单帧预测网络从输入数据中捕获时间上下文;其二是将时间上下文特征输入到帧解析网络,通过变换层生成未来的像素级分割.Wu等人[58]使用语义分割将前景和背景分离,同时使用实例分割将前景中的不同运动物体分离,因不同物体具有不同的运动方式,识别每一个运动物体并预测它们的移动路径和尺度变化,通过背景的非刚性变形和运动物体的仿射变换来预测未来场景.Bei等人[59]等人同样使用分割网络描述场景布局,使用光流定义物体运动,先将预测的光流图作用于当前帧做运动变换,然后根据预测的语义图进行图像渲染[60,61].Wu等人[62]试图分析视频中每个运动物体对应的物理状态,将不同物体分割后通过物理引擎模块预测其物理状态的变化.上述方法主要考虑2D场景,Henderson等人[65]在3D场景上对多个3D物体进行建模,每个物体都有一个外观表示和一个随时间变化的3D定位,并基于此预测场景的未来变化趋势,可建模由视角变化而引起的外观变化.

除了场景分割图,人体姿态与物体关键点也可以作为重要的语义状态约束.一种常用方法是通过有监督地训练人体姿态预测器,来学习人体关键点的动态变化,然后结合预测的姿态生成未来视频序列[66,67].此类方法虽然在具有静态背景的长期预测任务中表现出一定的优势,但因在预测中没有考虑全局信息变化,无法处理气象预报、自动驾驶等场景下的预测问题.此外,训练关键点检测器通常需要额外的监督信息,不适用于无标签视频数据场景.为了解决后一个问题,Minderer等人[68]通过添加物体关键点特征约束,对关键点坐标空间中的动态建模,在视频预测任务中实现了无监督的关键点提取.该模型则直接从视频中学习基于关键点的表示,不需要像素数据之外的任何监督.为了进一步建模动态环境中的物体间交互信息,Bodla等人[69]采用分级预测的方式,先学习布局特征,即物体位置和人体关键点,对视频中人与物体之间的关系进行建模,然后再进行视频预测.

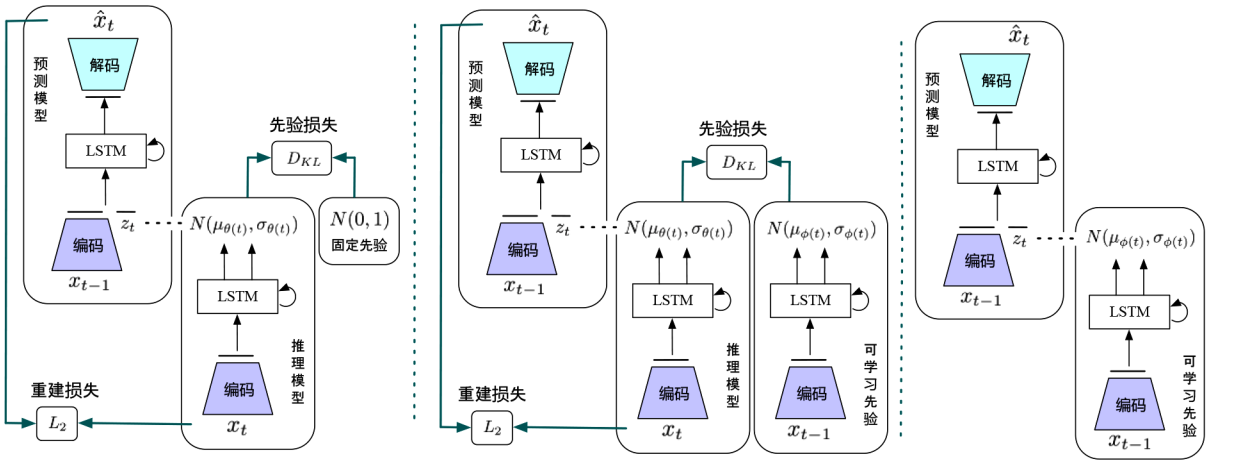
3.3 基于隐状态空间的深度预测模型

对未来视频中固有的不确定性进行概率建模是视频预测的关键技术之一.面向图像数据和时序数据的

表2　基于深度网络状态空间的视频预测模型对比(ED表示图像编码-解码架构,Adv表示对抗损失函数)

预测方法		年份	网络架构	损失函数	数据集	评价指标
基于语义状态空间的深度预测模型	Villegas等[66]	2017	LSTM, ED	L2, Adv	Penn Action, Human3.6M	PSNR
	VDA[62]	2017	CNN	L2	Block towers	MSE
	Struct-VRNN[68]	2019	RNN	L2, KL	Basketball, Human3.6M	FVD, SSIM, PSNR
	Bodla等[69]	2021	RNN, GAN	L1, Adv	UMD-HOI[90], Bimanual[91]	PSNR, SSIM, LPIPS
	Bei等[59]	2021	MLP, RNN	L1, 交叉熵, KL	Cityscapes, KITTI	PSNR, SSIM, LPIPS
隐状态空间学习与时空概率预测	SV2P[73]	2018	CDNA[89]	L1, L2, KL	BAIR robot pushing, Human3.6M, Robotic pushing	PSNR, SSIM
	SVG[44]	2018	LSTM, ED	L2, KL	KTH, BAIR robot pushing	PSNR, SSIM
	SAVP[74]	2018	GAN, VAE	L1, KL	KTH, BAIR robot pushing	PSNR, SSIM
	Gur等[76]	2020	GAN, VAE	L2, KL, Adv	UCF101, YouTube 8M	FID
	GHVAE[75]	2021	VAE	ELBO, KL	RoboNet, Human3.6M, KITTI, Cityscapes	FVD, SSIM, LPIPS
基于状态空间的长时趋势建模	EPVA[41]	2018	LSTM, ED	L2, Adv	Toy Dataset, Human3.6M	SSIM
	GCP[79]	2020	LSTM, ED	交叉熵	Human3.6M	PSNR, SSIM
	Lee等[82]	2021	ConvLSTM, ED	L2, L1	Moving MNIST, KTH, Human3.6M	MSE, PSNR, SSIM, LPIPS
预测学习中的时空表征解耦	DrNet[83]	2017	LSTM, ED	L2, CE, Adv	MNIST, KTH	IS, PSNR, SSIM
	DDPAE[85]	2018	VAE	ELBO	Moving MNIST	BCE, MSE
	RNN-EM[86]	2018	RNN-ED	KL	Bouncing balls	BCE
	PhyDNet[84]	2020	ConvLSTM, ED	L2	Moving MNIST, Traffic BJ, Human 3.6	MSE, MAE, SSIM

概率式隐变量模型包括变分自编码器(Variational Auto-Encoder, VAE)[70,71]和变分循环神经网络(Variational Recurrent Neural Network, VRNN)[72]等. 基于此, Babaeizadeh等人[73]提出了一种随机变分视频预测框架(Stochastic Variational Video Prediction, SV2P),使用隐变量和变分推断对未来视频序列的概率分布建模. Denton等人[44]提出随机视频生成模型(Stochastic Video Generation, SVG),其变分循环神经网络的隐状态信息从可学习的先验分布中采样得到,即$z_t \sim p_\psi(z_t|x_{1:t-1})$,其模型结构如图5所示.



(*a*) 带有固定隐状态先验的SVG模型的训练过程　(*b*) 带有可学习先验的SVG模型的训练过程　(*c*) 带有可学习先验的SVG模型的预测过程

图5　SVG[44]模型架构图

遵循类似的思路,SAVP模型[74]将GAN和VAE进行结合,在有效增强不确定性建模能力的同时,显著提高了视频序列的生成质量. 此外,一些现有工作[75~77]通过构建层次化的隐变量网络结构,也实现了提升视频预测质量的效果. 其中,Gur等人[76]提出层级块状VAE-GAN结构,使用patch-VAE生成粗糙的图片,增强

预测结果的多样性，并使用patch-GAN补充空间细节，提升图像质量. Wu等人[75]提出了多层变分预测网络的分层训练策略，逐层叠加地训练整个预测模型的隐状态子模块，有效优化了模型的内存开销. Wang等人[78]提出基于粒子滤波算法的贝叶斯预测网络(BP-Net)，同时考虑视频观测空间到隐空间的感知不确定性和从历史隐状态到未来隐状态的动态不确定性.

基于低维隐状态空间的预测模型，如上述典型的SVG，在BAIR数据集上SSIM，PSNR等评价指标下的预测结果不如基于高维状态空间的PredRNN-V2，因为在低维隐状态空间中建模会丢失高维空间中较多的细节信息，造成生成数据的模糊和失真. 但它的优势是可以通过加入随机变量建模数据的多样性分布，输出差异化的预测结果，易于与下游控制决策任务进行结合.

3.4 基于状态空间的长时趋势建模

目前大多数预测学习方法只关注较短时效(十帧左右)下的视频预测质量，但在预测时效延长后，预测图像的质量会随着视频上下文的消失而急剧下降. 为了实现可靠的长序列视频预测，使用层级预测是一个有效的解决方法，能减小长期预测造成的累加误差. Villegas等人[66]先估计出输入视频中的高级语义状态(如人体姿态)，通过预测该结构在未来如何演变辅助未来视频序列的生成. Wichers等人[41]也采用类似的方式，但无需额外的标注信息，并在特征空间中使用对抗训练来改善特征表达. Pertsch等人[79]和Kim等人[80]采用由粗到细的优化策略或构建变分隐状态的层次化结构，在视频的多个时间尺度上进行预测，即先将预测过程分为多个子段，然后再优化每个子段，从而有效缓解逐步预测带来的误差累积，实现远距离的时空状态转移. 利用类似神经图灵机(Neural Turing Machine)[81]的外部存储网络是强化长时趋势建模的另一个思路. Lee等人[82]提出长时运动上下文记忆单元(LMC-Memory)，首先从长序列中提取并在外部记忆单元中存储视频的长时上下文信息，而后用已存储的上下文信息来匹配短期序列，从而有效延长了深度网络的预测时效.

3.5 预测学习中的时空表征解耦

以视频数据为代表的高维时空序列数据往往具有复杂的时空耦合性，该特性在某种程度上增大了预测学习的难度. 因此，许多针对时空表征的自监督解耦方法应运而生，其主流思想是将数据分解为“内容”和“运动”两部分，并独立地进行特征学习和预测.

Denton等人[83]提出了一种基于GAN的预测式时空解耦模型DrNet，利用时间相关性和对抗损失函数将视频数据的每一帧分解为静止和可随时间变化的两部分，并分别在合成和真实数据集的长时预测任务上验证了模型的特征解耦能力，但其缺点是需要为对抗训练构造额外的训练数据. Minderer等人[68]提出了Struct-VRNN模型，结合关键点信息进行图像表示，并在关键点坐标空间中对其进行了动力学建模，也可视为带有语义约束的时空特征解耦. Guen等人[84]提出PhyDNet模型，在ConvLSTM网络的基础上加入偏微分方程约束，有效提取了数据中的物体先验知识，提升了模型的解耦和预测能力，但该方法在处理复杂的真实时空序列数据时表现欠佳.

考虑以物体为中心的时空特征表达，Hsieh等人[85]提出了DDPAE模型，将高维的时空数据表示为多个低维的时不变状态分量(表示空间解构得到的多个物体的结构信息)和时变分量(表示不同物体各自的动态信息)的组合，在人工构造的数据集上具有较好的解耦能力. 但该方法假设时空动态仅包含简单刚性物体的位移，忽略了物体间可能发生的遮挡、物体的形变等，因而实用性较为有限. 此外，van Steenkiste等人[86]提出的RNN-EM模型，用RNN串联多个物体解构后的特征，并利用图网络对物体间的交互建模，有效刻画了不同物体的时间动态特征，然而该方法对遮挡、形变等复杂时空动态场景的刻画能力依然有限.

Zablotskaia等人[87]拓展了Greff等人[88]提出的无监督静态多物体解构方法IODINE，有效学习了动态场景下的多物体状态信息与物体间的相关性.

4 基于预测模型的视觉决策方法

预测学习与强化学习和视觉决策算法关系密切. 智能体学习与世界交互时，一个核心挑战是预测动作对环境产生的影响. 目前学习物理交互的许多方法都需要环境状态的标注信息进行训练，然而，要将真实世界的交互学习扩展到多种应用场景时，获取有标注的数据是不切实际的. 为了在环境状态无标注的情况下学习物理对象的运动，可以通过前文总结的视频预测方法来估计环境在给定动作序列的条件下的在观测空间中的反馈，即一种以动作为条件的视频预测模型. 此类模型能明确地对场景中与动作序列相对应的运动物体进行建模，实现更优的视觉控制与决策. 表3总结了基于预测模型的视觉决策方法.

4.1 基于动作序列的视频预测模型

Oh等人[12]提出首个以动作为条件的视频预测模型，在编码器和解码器之间加入一个动作转换机制，研究控制输入条件下高维图像的长期预测，并在游戏仿真环境Atari上完成了模型验证. Schmidhuber等人[92]提出的世界模型(World Model)，成为首个基于预测学习时空特征表达的视觉强化学习方法. Chiappa等人[93]提出循环环境模拟器模型(Recurrent Environment Simulator)，能够做出未来数百个时间步长的时空连贯预测，

表3　基于预测模型的视觉决策方法对比(ED表示图像编码-解码架构)

预测方法		年份	网络架构	视频预测(世界模型)损失函数	数据集
高维观测空间中的有模型视觉决策	Visual MPC[95]	2018	ConvLSTM	L2	\
	PoliNet[96]	2019	ED	L1, L2	Stanford 2D-3D-S[122] and Matterport3D[123]
低维语义空间中的有模型视觉决策	O2P2[98]	2019	CNN	L2, perceptual loss	\
	SMORL[99]	2020	CNN	KL	MuJoCo, Multiworld
低维隐空间中的有模型视觉决策	PlaNet[51]	2019	RSSM	KL	DeepMind Control Suite
	Dreamer[7]	2019	RSSM	KL, contrastive loss	DeepMind Control Suite

模拟环境对给定动作序列的响应,从而进行有效的提前规划和决策. Wang等人[63]提出了PredRNN-V2模型,在PredRNN的卷积循环网络单元的基础上,实现了高维时空状态与低维动作信息的有效融合,其图模型如图6所示.

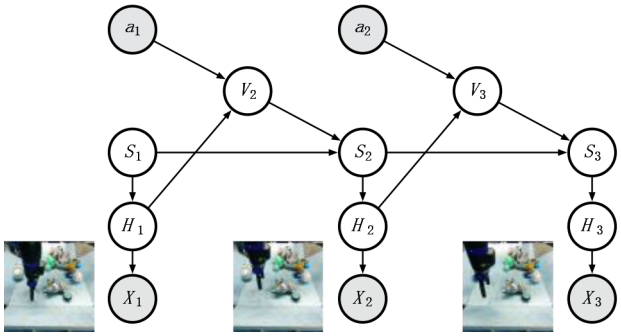


图6　动作条件下的PredRNN-V2模型[63]

上述方法为有模型的强化学习(Model-Based Reinforcement Learning,MBRL)提供了研究基础. 智能体通过预测采取不同动作后的未来结果(表示为未来图像序列),最终选择能够产生最优反馈的动作. 这使智能体能够在许多任务中高效执行,减少与环境交互产生的代价. 近几年,基于视频预测模型的强化学习方法被运用到到高维的图像环境中,这些方法表现出良好的性能,同时所需要的数据远少于无模型强化学习方法(Model-Free Reinforcement Learning).

4.2　高维观测空间中的有模型视觉决策

在基于预测模型的视觉控制中,观察信号往往存在于高维的像素空间,通常不足以揭示环境的确切内在状态. 该问题可被定义为部分可见的马尔可夫决策过程(POMDP),具有离散的时间步长为 $t \in [1, T]$,隐状态为 s_t,观察到的图像为 o_t,连续行动为 a_t,即时奖励为 r_t. Babaeizadeh等人[94]总结并总体评估了有模型的视觉强化学习方法,如图7所示,深度预测网络包含高维观测空间或低维隐空间中的状态转移函数、观察函数和奖励函数. 此处以基于低维隐状态的方法为例,分别有状态转移函数 $s_t \sim p(s_t \mid s_{t-1}, a_{t-1})$,观察函数 $o_t \sim p(o_t \mid s_t)$ 和奖励函数 $r_t \sim p(r_t \mid s_t)$. 也即不仅要预测环境未来的时空动态,也要预测环境未来的奖励反馈. 决策模型整体的训练目标是找到策略 $p(a_t \mid o \leqslant t, r \leqslant t, a \leqslant t)$,使未来奖励的预期总和 $E_p \sum_{t=1}^{T} \gamma^t r_t$ 最大,其中是 γ 贴现因子,T 是视野长度,根据策略采样得到的动作计算期望. 在MBRL中,我们根据先前的观察和未来的动作假设,通过预测其分布来近似预期奖励 $p(r_t \mid o \leqslant t, a \geqslant t)$,然后通过策略优化方法,寻找高回报的动作序列.

如上所述,在有模型的视觉决策中,深度预测网络按照学习表征的不同可以分为五大类. 第一类近似于直接根据未来的动作和以前的观察估计预期奖励,并不明确地预测图像. 另外四类在未来预期奖励 $p(r_t \mid o \leqslant t, a \geqslant t)$ 之外,还预测下一步(或多步)观察值 o_{t+1}. 在状态转移函数上其具体可分为在观测空间中对环境的转移函数进行建模 $\hat{o}_{t+1} \sim p(o_t \mid o \leqslant t, a \geqslant t)$ 或直接在隐空间中建模 $h_t \sim p(h_t \mid h_{t-1}, a_t)$,其中 h_t 是模型在时间步骤 t 的隐状态. 在奖励函数上其具体可分为使用学习的隐空间预测未来的奖励 $r_t \sim p(r_t \mid h_t)$ 或从预测的未来观测中进一步预测未来的奖励 $r_t \sim p(r_t \mid o_{t+1})$.

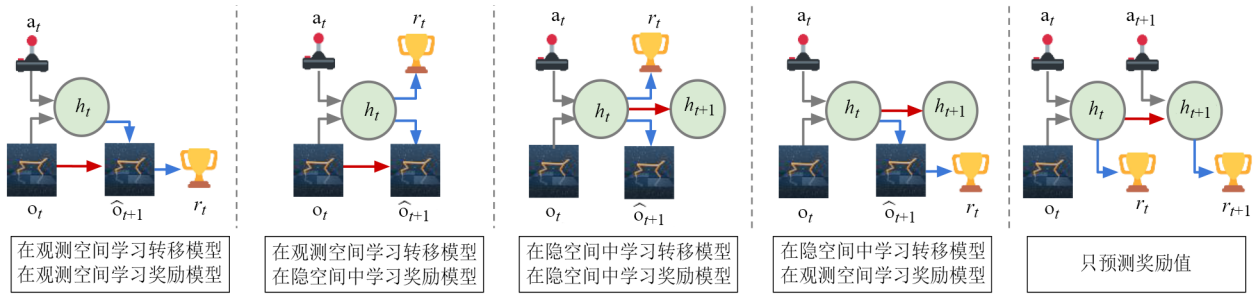


图7　有模型的视觉决策的多种方法对比[94]

在观测空间上学习预测模型的典型案例是 Ebert 等人[95]提出的 Visual MPC 算法，利用深度强化学习和模型预测控制(Model Predictive Control，MPC)从原始的视觉输入中学习复杂的机器人技能，并将其用于现实世界的机器人任务. 在测试时，该方法需要指定目标图像，即理想的目标状态是以图像形式表现的，并需要用户指定历史观测中的特定像素(表示某个物体)在目标图像上的目标位置(即表示该物体的移动目标). Hirose 等人[96]提出一种新的神经网络架构 PoliNet，基于视觉输入信号进行机器人导航，同样使用基于深度预测网络的模型预测控制算法，而无需对环境进行精确地图测绘.

4.3 低维语义空间中的有模型视觉决策

鉴于直接在图像空间上进行规划控制可解释性较差，且不能对特定的语义信息进行建模，基于语义空间上的有模型视觉决策方法使用了更具结构化的时空特征表达来进一步提高智能体学习的效率. 具体来说，此类方法可以细分为基于粒子表征的预测学习和基于物体表征的预测学习.

求解 POMDP 问题的一个主要困难是在感知不确定的情况下推断未观测到的真实状态的多模态分布，并使决策算法依赖感知到的状态的信念分布(即各种可能的状态具有不同的置信度). Wang 等人[97]提出 DualSMC 模型，在有明确语义定义的状态空间中，使用深度学习技术，连通了用于状态估计的粒子滤波算法、用于动态建模的深度预测网络和基于状态信念分布的 MBRL 算法. 该模型能够从基于粒子的信念状态中学习到显著降低感知不确定性的决策策略，是深度预测学习在视觉规划控制领域的典型应用.

对于任何智能体来说，仅使用高维、非结构化和未标记的视频观测数据来学习感知、预测与决策技能是一个棘手的问题. Janner 等人[98]提出了 O2P2(Object-Oriented Prediction and Planning)模型，利用预测学习构建显式的多物体表征，共同学习从视频观测映射到物体表征的感知函数、预测多物体之间两两物理交互函数，以及将物体表征映射回像素的渲染函数. 该方法基于预测学习提取的物体表征，运用 CEM(Cross-Entropy Method)算法实现下游控制任务. Zadaianchuk 等人[99]提出了 SMORL(Self-supervised Multi-Object RL)模型，通过预测未来时刻的观测图像，以无监督的方式提取场景中潜在的结构化向量，将观测图像和决策目标图像分别表示为一组以物体为中心的状态表征(包括物体类别和位置等)，并将其与强化学习相结合，使智能体能够分而治之地处理环境中的多个动作对象.

4.4 低维隐空间中的有模型视觉决策

很多视觉控制与决策场景不易人为界定有语义的状态空间，需要智能体具有自监督提取隐状态表征，并基于此隐状态空间执行决策算法的能力. 为了能使智能体执行多步的视觉预测和决策，Hafner 等人[51]提出 PlaNet 模型，在隐空间中训练转移预测模型和奖励预测模型. 相较于在高维观测空间进行状态转移函数的建模，隐状态预测的存储与计算开销小，使得智能体可以并行地预测大量基于动作假设的轨迹，继而运用 CEM 算法，实现基于预测状态轨迹和预测回报函数的控制决策. PlaNet 与无模型强化学习方法相比，显著减少了智能体和环境交互的次数，在 DeepMind Control Suite 视觉控制仿真环境中取得了相似或更优的表现. PlaNet 的一大特色是在训练预测模型时，区别于以往单独使用确定性模型或随机性模型，使用循环状态空间模型(Recurrent State Space Model，RSSM)来同时建模环境中的确定性部分和随机部分.

传统的序列决策算法受到固定视野长度的限制，使得智能体可能产生相对“短视”的行为. 在 PlaNet 的后续工作 Dreamer 中[7]，首先从与环境的交互中学习未来观测与未来奖励函数的预测网络. 在第二部分中，智能体借助学习得到的环境模型“假想”出大量的隐状态轨迹，并在这些轨迹上利用强化学习算法学习值函数，通过隐状态状态转移模型回传梯度，找到使值函数最大化的策略. 在“假想”的隐空间中学习策略模型进一步提升了强化学习的训练效率，智能体可以根据学习到的策略在环境中行动.

5 数据集和评价指标

深度预测学习方法弱化了对视频标注的依赖，因而多数深度学习与计算机视觉研究中常用的视频数据集都可以作为深度预测模型的训练集和测试集. 本节根据数据内容或应用场景的不同，归纳总结该领域中常用的视频数据集，并介绍这些数据集上相应的模型评价指标. 各数据集具体特性如表 4 所示.

5.1 预测学习专用的合成数据集

Bouncing balls 数据集[100]是训练高维序列生成模型的常用数据集. 该数据集模拟 3 个球在盒子里弹跳的过程，包含 4000 段训练视频、200 段验证视频和 200 段测试视频.

Moving MNIST 数据集[30]中的视频是由两个(或多个)从静态 MNIST 数据集中抽取的数字，通过移动产生的 20 帧(或更多)视频序列组成，数字的移动速度和方向任意. 该数据集有两种典型设置，训练集的视频数量分别是固定的和无限的.

Block towers 数据集[101]是一个由合成数据和真实数据组成的小积木数据集，合成数据是使用 3D 游戏引

擎创建,把几个不同颜色的积木堆叠起来,并随机进行推倒.

5.2 人体动作视频数据集

KTH数据集[102]于2004年发布,包含2391个视频,每个视频平均时长为4秒,视频中25个人在4个不同场景下做出6类不同动作,如行走、跑步、挥手等. 这一数据集是当时拍摄的最大人体动作数据集,包括尺度变化、衣着变化和光照变化,但背景比较单一.

Weizmann数据集[103]仅有90个视频,包含10个动作,每个动作有9个不同样本,每段视频只有一个人在做单一动作. 因为相机固定,所以背景也是单一的.

HMDB-51数据集[104]中的视频多数来源于电影和网络视频,包含6766段视频,平均时间为3.15秒,人在视频中执行51类动作,每类至少有101段视频. 该数据集提供了拍摄视角和相机移动等标注信息.

UCF101数据集[105]是UCF50的扩展,从YouTube网站上收集,包含101个动作类别,13320个视频,所有视频的帧率都为25 fps. 该数据集是预测模型中最常用的数据集.

Penn Action数据集[106]是一个来自宾夕法尼亚大学的动作和人类姿态识别数据集. 它包含了15个不同动作的2326个视频序列,提供人的关节点和视角标注.

Human3. 6M数据集[107]是一个人体姿态数据集,记录了11个志愿者执行的15种不同类型的动作. 它标注了所有志愿者的深度图、姿态、2D框和3D扫描掩膜等. 此外,该数据集通过在真实视频中插入高质量的3D人体模型来扩展,以创建一个真实而复杂的背景.

Sports1M数据集[20]是由已经标注的YouTube视频组成. 它包含487类运动,每个视频的分辨率、时长和帧率都不同. 它的规模比UCF101数据集大,视频超过100万个,视频中的动作也更频繁.

5.3 城市交通热力图与车辆驾驶视频数据集

Caltech Pedestrian数据集[108]是一个专注于检测行人的数据集,因为它标注有行人边界框. 在137个视频片段中,总共有25万个已标注的视频帧,行人有2300个,行人边界框有350000个. 该数据集还提供了边界框和遮挡标签之间的时间对应关系.

Kitti数据集[109]是移动机器人和自动驾驶最流行的数据集之一,也是计算机视觉算法的基准. 它用各种传感器记录的数小时的交通场景,包括高分辨率的RGB相机、灰度立体声相机和3D激光扫描仪. 原始的数据集并不包含标注信息,而在2015年进行了语义分割和实例分割标注.

Cityscape数据集[110]是一个用于城市街道场景语义理解的大型数据库,记录了50个城市的街景,共30个类别,提供语义、实例和密集标注,含大约5000张精细标注的图像和20 000张粗糙标注的图像.

TaxiBJ数据集[111]是在混乱场景中收集得到的交通流量数据集,不会随时间均匀变化,不同时间的交通流量是不同的,每一帧都是一个32×32×2大小的图像,其

表4　常用的视频预测学习数据集

数据集		年份	视频数量	真实/合成	分辨率	标注信息	现有工作示例
预测学习专用的合成数据集	Bouncing balls[100]	2008	4000	合成	/	/	PGN[120], PGP[121]
	MovingMNIST[30]	2015	/	真实+合成	64×64	/	DFN[14], FutureGAN[22], ConvLSTM[31], PredRNN[34]
	Block towers[101]	2016	12 781	真实+合成	/	/	PhysNet[101],VDA[62]
人体动作视频数据集	KTH[102]	2004	2391	真实	160×120	动作	PredRNN[34], SVG[44], SAVP[74]
	Weizmann[103]	2005	90	真实	180×144	动作	MCnet[39]
	HMDB-51[104]	2011	6766	真实	/	动作, 视角	Behrmann 等[24], Srivastava 等[30],
	UCF101[105]	2012	13 320	真实	320×240	动作	BeyondMSE[11], TGAN[40]
	Penn Action[106]	2013	2326	真实	480×270	动作, 姿态, 视角	Villegas 等[66]
	Human3.6M[107]	2014	/	真实+合成	1000×1000	动作, 深度图, 姿态, 2D框, 3D扫描掩膜	MIM[111], Villegas 等[66], Struct-VRNN[68], SV2P[73]
	Sports1M[20]	2014	1 133 158	真实	/	运动	BeyondMSE[11], Srivastava 等[30]
城市交通热力图与车辆驾驶视频数据集	Caltech Pedest.[108]	2009	137	真实	640×480	行人边界框	Jin 等[18]
	Kitti[109]	2013	151	真实	1392×512	里程	Jin 等[18], Wu 等[58], GHVAE[75]
	Cityscapes[110]	2016	50	真实	2048×1024	语义, 实例	PEARL[57], Bei 等[59], Wu 等[58]
	TaxiBJ[111]	2019	/	真实	32×32	/	MIM[111]
	Traffic4Cast[112]	2019	/	真实	495×436	/	PredRNN-V2[63]
机器人视觉预测数据集	RobotPushing[89]	2016	57 000	真实	640×512	机械臂姿态	SV2P[73]
	BAIR Pushing[113]	2017	45 000	真实	64×64	机械臂姿态	PredRNN-V2[63], SV2P[73]
	RoboNet[114]	2019	161 000	真实	/	机械臂姿态	GHVAE[75]

中两通道表示进出同一区域的车流量.

Traffic4Cast数据集[112]于2019年以视频帧的形式记录了柏林、莫斯科和伊斯坦布尔连续交通流量的GPS轨迹. 每个帧的大小是495×436×3. 每个像素的值对应于5分钟内获取100 m×100 m区域范围的交通信息,包括平均速度、流量和主要交通方向.

5.4 机器人视觉预测数据集

Robotic Pushing数据集[89]是为学习物体的运动而创建的,它包含10个不同的机械臂与现实世界中的物体相互作用的过程,机械臂具有7个自由度.

BAIR Robot Pushing数据集[113]是BAIR实验室研究机器人无监督学习训练采集而来的,机器人学习环境中的物理学,并预测其行动对环境产生的影响. 数据集是由机械臂数小时自监督学习产生.

RoboNet数据集[114]是由4个不同实验室的7个机械臂做各种自监督训练而组成的,BAIR实验室就是其中之一. 该数据集的目标是成为一个与ImageNet图像数据集一样的通用标准.

5.5 评价指标

视频预测模型最直接的评估指标是计算像素级的均方误差(Mean Square Error, MSE)[115~117],另一个与MSE相关且更流行的评价指标是峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)[34,66,89,118]. PSNR是最大像素值的比值,例如,在8比特图像中使用255除以MSE来对比两张图片. 两张图像之间的差异越小,PSNR的值就越高,则生成图像的质量也越高.

虽然像素级别的MSE和PSNR很容易计算,但它们不能测量生成图像的整体视觉质量. 为解决这个问题,现有研究通常采用结构相似性(Structural Similarity Index Measure, SSIM)[70]作为评价指标. 为了捕捉两幅图像之间的结构差异,SSIM将光照信息与物体的结构进行区分,通过对比图片之间的亮度、对比度和结构来计算图片的相似度.

上述度量之外,参考主流的生成模型评价指标,现有视频预测方法亦采用基于预训练深度网络的图像感知相似度(Learned Perceptual Image Patch Similarity, LPIPS)[36]. LPIPS通过比较神经网络的内部激活值来衡量两幅图像的相似性,即在归一化的深度特征上计算L2距离. 类似地,近期工作引入Fréchet Video Distance (FVD)[119],衡量预测视频与真实视频之间序列级别的质量差异.

6 存在的问题和未来研究方向

6.1 视频预测领域的开放问题

虽然目前视频预测领域的研究取得了突破性的进展,但仍存在很大的进步空间. 为了提高视频预测的能力,需要对该领域现存的问题进行归纳和分析.

首先,现在大多数的预测方法局限于短期范围,从长期来看,模型的预测能力远没有达到期望的效果. 对于长时间范围的预测,通用的方法是使用迭代式的连续帧预测,这种预测方式会造成误差的累积,使结果一步步远离目标值. 目前循环神经网络在该领域仍被广泛用于时间依赖关系的建立,但在同时构建时空强耦合关系上还有所欠缺. 此外,当前研究普遍使用的视频预测评价指标主要来源于图像相似度的比较,在时间维度上没有得到有效度量,因此迫切需要一个更加客观的、准确的评价指标.

其次,预测的视频帧存在分辨率过低的问题. 损失函数对预测质量有直接的影响:均方误差损失函数会造成预测图片的模糊;对抗损失函数虽然可一定程度上消除视频预测的歧义性,但训练不稳定且易出现模式崩塌问题;利用KL散度直接约束由视频编码成的隐状态虽然可以对概率分布直接建模,但在真实应用场景下的视频生成质量还有待提高.

最后,目前的视频预测方法,无论是确定性的模型,还是概率式的预测模型,其建模过程完全是以数据为驱动的. 然而一方面,因为现实世界是复杂多变的,观测数据往往充满噪声的;另一方面,在数据噪声的背后,视频数据往往可以反映出真实的物理规律,其隐藏的状态信息是有迹可循的. 现有的大部分模型在预测未来时空状态的时候,并没有充分考虑对观测序列背后的本质物理过程的推断.

综上所述,未来的视频预测研究方向主要包括:第一,寻找循环神经网络的一种可替代模型,直接建立当前帧与目标帧的时空关联性,减少迭代造成的误差累积;第二,借鉴生成对抗网络中的对抗损失函数,设计一种合理的损失函数,避免当前常用的均方误差损失函数所造成的图像模糊问题;第三,进一步探索基于物理过程的时空动态建模方法,以刻画真实场景的本质属性.

6.2 基于预测的决策领域的开放问题

对于基于视频预测模型的智能决策算法,尤其是有模型强化学习决策方法,目前的主要挑战是:第一,对环境的建模存在误差,而且随着智能体与环境的迭代交互,累积误差越来越大,使得算法难以收敛到最优解;第二,模型的泛化性较差,对于各种复杂的现实环境,希望决策可以在各种环境中发挥作用. 但目前的多种主流方法都是使用针对特定任务经过调整的超参数进行训练的,只能处理仿真场景下(例如DeepMind Control Suite)相对单一的运动预测和视觉决策问题,它们经常因新颖的任务或环境而失效. 研究该问题的核心是提升预测模型(也称世界模型)对于新场景的适配与

迁移能力.

如何将深度强化学习算法应用于现实环境的实际应用中,是该领域研究最大的动因.有模型强化学习的首要研究方向是提升模型的泛化能力.人类之所以能够在新任务上进行快速学习,是因为我们会重复利用过去的知识技能,因此有模型强化学习也可以利用人类这一特性,借鉴迁移学习方式,提升其领域适配性和通用性.另外,将以动作序列为条件的视频预测模型应用于下游视觉控制与决策任务中也是未来研究的重要思路,让结合动作信息的视频预测模型引导决策任务的执行,提升世界模型在真实场景中的泛化能力,从而提升有模型决策方法的样本效率和在真实场景下的模型适配能力.

参考文献

[1] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2015: 802-810.

[2] CHANDRA R, BHATTACHARYA U, BERA A, et al. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8483-8492.

[3] CASTREJON L, BALLAS N, COURVILLE A. Improved conditional vrnns for video prediction[C]//Proceedings of The IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 7608-7617.

[4] ZHANG J, ZHENG Y, QI D, et al. DNN-based prediction model for spatio-temporal data[C]//Proceedings of The 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Burlingame, California: Association for Computing Machinery, 2016: 1-4.

[5] EBERT F, FINN C, LEE A X, et al. Self-supervised visual planning with temporal skip connections[C]//Proceedings of The 1st Annual Conference on Robot Learning. California: PMLR, 2017: 344-356.

[6] HA D, SCHMIDHUBER J. World models[EB/OL]. (2018-05-27)[2021-09-01]. https://arxiv.org/abs/1803.10122.

[7] HAFNER D, LILLICRAP T, BA J, et al. Dream to control: Learning behaviors by latent imagination[EB/OL]. (2019-11-03)[2021-09-01]. https://arxiv.org/abs/1912.01603.

[8] WANG Y, LIU B, WU J, et al. DualSMC: Tunneling differentiable filtering and planningunder continuous POMDPs[C]//Proceedings of The TwentyNinth International Joint Conference on Artificial Intelligence. Yokohama: arXiv, 2020: 4190-4198.

[9] LECUN Y, BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of The IEEE, 1998, 86(11): 2278-2324.

[10] JAIN V, MURRAY J F, ROTH F, et al. Super-vised learning of image restoration with convolu-tional networks[C]//Proceedings of The 11th International Conference on Computer Vision. Rio de Janeiro: IEEE, 2007: 1-8.

[11] MATHIEU M, COUPRIE C, LECUN Y. Deep multiscale video prediction beyond mean square error[EB/OL]. (2015-11-17)[2021-09-01]. https://arxiv.org/abs/1511.05440.

[12] OH J, GUO Xiao-xiao, LEE H, et al. Action-conditional video prediction using deep networks in Atari games[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2015: 2863-2871.

[13] VUKOTIC V, PINTEA S L, RAYMOND C, et al. One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network[C]//International Conference on Image Analysis and Processing. Cham: Springer, 2017: 140-151.

[14] JIA X, DE BRABANDERE B, TUYTELAARS T, et al. Dyna-mic filter networks[C]//Proceed ings of The Advances in Neural Information Systems Processing. Barcelona: arXiv, 2016: 667-675.

[15] XUE T, WU J, BOUMAN K L, et al. Visual dynamics: Stochastic future generation via layered cross convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2236-2250.

[16] XU J, NI B, LI Z, et al. Structure preserving video prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1460-1469.

[17] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[C]//Proceedings of the Advances in Neural Information Process Systems ing. Montreal: MIT Press, 2015: 2017-2025.

[18] JIN B, HU Y, TANG Q, et al. Exploring spatial-temporal multi-frequency analysis for high-fide lity and temporal-consistency video prediction [C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4554-4563.

[19] TRAN D , BOURDEV L , FERGUS R , et al. Learning spatiotemporal features with 3D con volutional networks

[C]//Proceedings of The IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4489-4497.

[20] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolu tional neural networks[C]//Proceedings of The IEEE Conference on Computer Vision and Recognition Pattern. Columbus: IEEE, 2014: 1725-1732.

[21] XU H, DAS A, SAENKO K. R-c3d: Region con volutional 3d network for temporal activity detection[C]//Proceedings of The IEEE Inter national Conference on Computer Vision. Venice: IEEE, 2017: 5783-5792.

[22] AIGNER S, KORNER M. Futuregan: Antici pating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans[EB/OL]. (2018-10-02)[2021-09-01]. https://arxiv.org/abs/1810.01325.

[23] WANG Y, JIANG L, YANG M H, et al. Eidetic 3d lstm: A model for video prediction and beyond[C]//International Conference on Learning Representations. New Orleans: OpenReview, 2018: 1-14.

[24] BRHRMANN N, GALL J, NOROOZI M. Unsupervised video representation learning by bidirectional feature prediction[C]//Proceedings of The IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 1670-1679.

[25] VONDRICK C, PIRSIAVASH H, TORRALBA A. Generating videos with scene dynamics[C]//Proceedings of The Advances in Neural Informa tion Processing Systems. Barcelona: arXiv, 2016: 613-621.

[26] HAN T, XIE W, ZISSERMAN A. Video representation learning by dense predictive coding[C]//Proceedings of The IEEE/CVF International Conference on Computer Vision Workshops. Seoul: IEEE, 2019: 1483-1492.

[27] GRAVES A. Generating sequences with recurrent neural networks[J]. arXiv preprint arXiv, 2013, 1308.0850.

[28] RANZATO M A, SZLAM A, BRUNA J, et al. Video (language) modeling: a baseline for generative models of natural videos[EB/OL]. (2014-12-20)[2021-09-01]. https://arxiv.org/abs/1412.6604.

[29] SUTSKEVER I, VINYALS O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2014: 3104-3112.

[30] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using lstms [C]//Proceedings of The 32nd International Conference on Machine Learning. Lille: JMLR, 2015: 843-852.

[31] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Proceedings of the Advances in Neural Information Processing Systems. Montreal: MIT Press, 2015: 802-810.

[32] SHI X J, GAO Z H, LAUSEN L, et al. Deep learningfor precipitation nowcasting: A benchmark and a new model [EB/OL]. (2017-06-12) [2021-09-01]. https://arxiv. org/abs/1706.03458.

[33] BALLAS N, YAO L, PAL C, et al. Delving deeper into convolutional networks for learningvideo representations [EB/OL]. (2015-11-19) [2021-09-01]. https://arxiv. org/abs/1511.06432.

[34] WANG Y, LONG M, WANG J, et al. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms[C]//Proceedings of The Advances in Neural Information Processing Systems. Long Beach: MIT Press, 2017: 879-888.

[35] OLIU M, SELVA J, ESCALERA S. Folded recurrent neural networks for future video prediction[C]//Proceedings of The European Conference on Computer Vision. Munich: Springer, 2018: 716-731.

[36] WU H X, YAO Z Y, LONG M S, et al. MotionRNN: A flexible model for video prediction with space time-varying motions[EB/OL]. (2018-03-03) [2021-09-01]. https://arxiv. org/abs/2103.02243v2.

[37] GOODFELLOW I, POUGET-ABADIEb J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2014: 2672-2680.

[38] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2021-09-01]. https://arxiv.org/abs/1411.1784.

[39] VILLEGAS R, YANG J, HONG S, et al. Decom-posing motion and content for natural video sequence prediction[EB/OL]. (2017-06-25) [2021-09-01]. arXiv preprint arXiv, 2017, 1706.08033.

[40] SAITO M, MATSUMOTO E, SAITO S. Tem-poral generative adversarial nets with singular value clipping[C]//Proceedings of The IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2830-2839.

[41] VILLEGAS R, ERHAN D, LEE H. Hierarchicallong-term video prediction without super-vision[C]//Proceedings of the International Conference on Machine Learning. Stockholm: PMLR, 2018: 6038-6046.

[42] JIN B, HU Y, TANG Q, et al. Exploring spatial-temporal

multi-frequency analysis for high-fidelity and temporal-consistency video prediction[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4554-4563.

[43] HENAFF M, ZHAO J, LECUN Y. Prediction under uncertainty with error-encoding networks[EB/OL]. (2017-11-24) [2021-09-01]. https://arxiv.org/abs/1711.04994.

[44] DENTON E, FERGUS R. Stochastic videogeneration with a learned prior[C]//International Conference on Machine Learning. Stockholm: PMLR, 2018: 1174-1183.

[45] KUMAR M, BABAEIZADEH M, ERHAN D, et al. Videoflow: A conditional flow-based model for stochastic video generation[EB/OL]. (2017-11-24) [2021-09-01]. https://arxiv.org/abs/1903.01434.

[46] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1x1 convolutions[EB/OL]. (2018-07-09)[2021-09-01]. https://arxiv.org/abs/1807.03039.

[47] PRENGER R, VALLE R, CATANZARO B. Waveglow: A flow-based generative network forspeech syn-thesis[C]//Proceedings of The ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE. 2019: 3617-3621.

[48] BAYER J, OSENDORFER C. Learning stochastic recurrent networks[EB/OL]. (2014-11-27)[2021-09-01]. https://arxiv.org/abs/1411.7610.

[49] FRACCARO M, SONDERBY S K, PAQUET U, et al. Sequential neural models with stochasticlayers[EB/OL]. (2016-05-24)[2021-09-01]. https://arxiv.org/abs/1605.07571.

[50] KRISHNAN R, SHALIT U, SONTAG D. Structured inference networks for nonlinear state space models[C]//Proceedings of The AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 2101-2109.

[51] HAFNER D, LILLICRAP T, FISCHER I, et al. Learning latent dynamics for planning from pixels[C]//International Conference on Machine Learning. Long Beach: JMLR, 2019: 2555-2565.

[52] ASAHARA A, MARUYAMA K, SATO A, et al. Pedestrian-movement prediction based on mixed Markov-chain model [C]//Proceedings of The 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Chicago: ACM, 2011: 25-33.

[53] MATHEW W, RAPOSO R, MARTINS B. Predicting future locations with hidden Markov models[C]//Proceedings of The 2012 ACM Conference on Ubiquitous Computing. Pittsburgh: ACM, 2012: 911-918.

[54] MURPHY K P. Machine Learning: A Probabilistic Perspective[M]. Cambridge: MIT press, 2012.

[55] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440.

[56] WANG W, YU R, HUANG Q, et al. Sgpn: Simi larity group proposal network for 3d pointcloud instance segmentation [C]//Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2569-2578.

[57] JIN X, LI X, XIAO H, et al. Video scene parsing with predictive feature learning[C]//Proceedings of The IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5580-5588.

[58] WU Y, GAO R, PARK J, et al. Future video synthesis with object motion prediction[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 5539-5548.

[59] BEI X, YANG Y, SOATTO S. Learning seman tic-aware dynamics for video prediction[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 902-912.

[60] PATWARDHAN K A, SAPIRO G, BERTALMI M. Video inpainting under constrained camera motion[J]. IEEE Transactions on Image Processing, 2007, 16(2): 545-553.

[61] XU R, LI X, ZHOU B, et al. Deep flow-guided video inpainting[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3723-3732.

[62] WU J, LU E, KOHLI P, et al. Learning to see physics via visual de-animation[C]//Proceedings of The Advances in Neural Information Processing Systems. Long Beach: MIT, 2017: 153-164.

[63] WANG Y, WU H, ZHANG J, et al. PredRNN: A recurrent neural network for spatiotemporal predictive learning [EB/OL]. (2021-03-17)[2021-09-01]. https://arxiv. org/abs/2103.09504.

[64] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[EB/OL]. (2017-03-31)[2021-09-01]. https://arxiv.org/abs/1704.00028v2.

[65] HENDERSON P, LAMPERT C H. Unsupervised object-centric video generation and decomposition in 3D [EB/OL]. (2020-07-07)[2021-09-01]. https://arxiv. org/abs/2007.06705.

[66] VILLEGAS R, YANG J, ZOU Y, et al. Learning to generate long-term future via hierarchical prediction[C]//Proceedings

of The 34th International Conference on Machine Learning. Sydney: JMLR, 2017: 3560-3569.

[67] WALKER J, MARINO K, GUPTA A, et al. The pose knows: Video forecasting by generating pose futures[C]//Proceedings of The IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3332-3341.

[68] MINDERER M, SUN C, VILLEGAS R, et al. Unsupervised learning of object structure and dynamics from videos [EB/OL]. (2019-06-19) [2021-09-01]. https://arxiv. org/abs/1906.07889.

[69] BODLA N, SHRIVASTAVAh G, CHELLAPPA R, et al. Hierarchical video prediction using relational layouts for human-object interact ions[C]//Proceedings of The IEEE/CVF Con ference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 12146-12155.

[70] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2013-12-20)[2021-09-01]. https://arxiv.org/abs/1312.6114v5.

[71] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[C]//International Conference on Machine Learning. Beijing: PMLR, 2014: 1278-1286.

[72] CHUNG J, KASTNER K, DINH L, et al. A recurrent latent variable model for sequential data[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2015: 2980-2988.

[73] BABAEIZADEH M, FINN C, ERHAN D, et al. Stochastic variational video prediction[EB/OL]. (2017-10-30)[2021-09-01]. https://arxiv.org/abs/1710.11252.

[74] LEE A X, ZHANG R, EBERT F, et al. Stochastic adversarial video prediction[EB/OL]. (2018-04-04)[2021-09-01]. https://arxiv.org/abs/1804.01523.

[75] WU B, NAIR S, MARTIN-MARTIN R, et al. Greedy hierarchical variational autoencoders for large-scale video prediction[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 2318-2328.

[76] GUR S, BENAIM S, WOLF L. Hierarchical patch vae-gan: Generating diverse videos from a single sample [EB/OL]. (2020-06-22) [2021-09-01]. https://arxiv. org/abs/2006.12226.

[77] SONDERBY C K, RAIKO T, MAALOE L, et al. Ladder variational autoencoders[C]//Proceedings of The Advances in Neural Information Processing Systems. Barcelona: IEEE, 2016: 3738-3746.

[78] WANG Y, WU J, LONG M, et al. Probabilistic video prediction from noisy data with a posterior confidence[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 10830-10839.

[79] PERTSCH K, RYBKIN O, EBERT F, et al. Long-horizon visual planning with goal-conditioned hierarchical predictors [C]//Proceedings of The Advances in Neural Information Processing Systems. Virtual Conference: Curran Associates, 2020: 17321-17333.

[80] KIM T, AHN S, BENGIO Y. Variational temporal abstraction[C]//Proceedings of the Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2019: 11570-11579.

[81] GRAVES A, WAYNE G, DANIHELKA I. Neural turing machines[EB/OL]. (2014-10-20)[2021-09-01]. https://arxiv.org/abs/1410.5401.

[82] LEE S, KIM H G, CHOI D H, et al. Video prediction recalling long-term motion context via memory alignment learning [C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3054-3063.

[83] DENTON E, BIRODKAR V. Unsupervised learning of disentangled representations from video[EB/OL]. (2017-05-31) [2021-09-01]. https://arxiv.org/abs/1705.10915.

[84] GUEN V L, THOME N. Disentangling physical dynamics from unknown factors for unsupervised video prediction[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pat tern Recognition. Seattle: IEEE, 2020: 11474-11484.

[85] HSIEH J T, LIU B, Huang D A, et al. Learning to decompose and disentangle representations for video prediction [EB/OL]. (2018-06-11) [2021-09-01]. https://arxiv. org/abs/1806.04166v1.

[86] VAN STEENKISTE S, CHANG M, GREFF K, et al. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions[EB/OL]. (2018-02-28)[2021-09-01]. https://arxiv.org/abs/1802.10353.

[87] ZABLOTSKAIA P, DOMINICI E A, SIGAL L, et al. Unsupervised video decomposition using spatio-temporal iterative inference[EB/OL]. (2020-06-25) [2021-09-01]. https://arxiv.org/abs/2006.14727.

[88] GREFF K, KAUFMAN R L, KABRA R, et al. Multi-object representation learning with itera tive variational inference [C]//Proceedings of the International Conference on Machine Learning. Long Beach: PMLR, 2019: 2424-2433.

[89] FINN C, GOODFELLOW I, LEVINE S. Un supervised

learning for physical interaction through video prediction [C]//Proceedings of the Advances in Neural Information Processing Systems. Barcelona: MIT Press, 2016: 64-72.

[90] GUPTA A, KEMBHAVI A, DAVIS L S. Observing human-object interactions: Using spatial and functional compatibility for recogni tion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1775-1789.

[91] DREHER C R G, WACHTER M, ASFOUR T. Learning object-action relations from bimanual human demonstration using graph networks[J]. IEEE Robotics and Automation Letters, 2019, 5(1): 187-194.

[92] SCHMIDHUBER J, HUBER R. Learning to generate artificial fovea trajectories for target detection[J]. International Journal of Neural Systems, 1991, 2: 125-134.

[93] CHIAPPA S, RACANIERE S, WIERSTRA D, et al. Recurrent environment simulators[EB/OL]. (2017-04-07) [2021-09-01]. https://arxiv.org/abs/1704.02254.

[94] BABAEIZADEH M, SAFFAR M T, HAFNERD, et al. Models, pixels, and rewards: Evaluating design trade-offs in visual model-based reinforcement learning[EB/OL]. (2020-12-08)[2021-09-01]. https://arxiv.org/abs/2012.04603.

[95] EBERT F, FINN C, DASARI S, et al. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control[EB/OL]. (2018-12-03)[2021-09-01]. https://arxiv.org/abs/1812.00568.

[96] HIROSE N, XIA F, MARTIN-MARTIN R, et al. Deep visual mpc-policy learning for naviga tion[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 3184-3191.

[97] WANG Y, LIU B, WU J, et al. DualSMC: Tunneling differentiable filtering and planning under continuous POMDPs [C]//Proceedings of The Twenty-Ninth International Joint Conference on Artificial Intelligence. Virtual Conference: AAAI 2020: 4190-4198.

[98] JANNER M, LEVINE S, FREEMAN W T, et al. Reasoning about physical interactions with object-oriented prediction and planning[EB/OL]. (2018-12-28) [2021-09-01]. https://arxiv.org/abs/1812.10972v1.

[99] ZADAIANCHUK A, SEITZER M, MARTIUS G. Self-supervised visual reinforcement learning with object-centric representations[EB/OL]. (2020-11-29)[2021-09-01]. https://arxiv.org/abs/2011.14381.

[100] SUTSKEVER I, HINTON G E, TAYLOR G W. The recurrent temporal restricted boltzmann machine[C]//Proceedings of The Advances in Neural Information Processing Systems. Vancouver, British Columbia: MIT Press, 2009: 1601-1608.

[101] LERER A, GROSS S, FERGUS R. Learning physical intuition of block towers by exam ple[C]//Proceedings of the 32nd International Conference on Machine Learning. New York: JMLR, 2016: 430-438.

[102] SCHULDT C, LAPTEV I, CAPUTO B. Recog nizing human actions: a local SVM approa ch[C]//Proceedings of The 17th International Conference on Pattern Recognition. Cambridge: IEEE, 2004: 32-36.

[103] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.

[104] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition[C]//Proceedings of The International Conference on Computer Vision. Barcelona: IEEE, 2011: 2556-2563.

[105] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012-12-03) [2021-09-01]. https://arxiv. org/abs/1212.0402v1.

[106] ZHANG W, ZHU M, DERPAMIS K G. From actemes to action: A strongly-supervised representation for detailed action understand ing[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 2248-2255.

[107] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6m: Large scale datasets and predic tive methods for 3d human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325-1339.

[108] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 34(4): 743-761.

[109] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[110] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3213-3223.

[111] WANG Y, ZHANG J, ZHU H, et al. Memory in memory: A predictive neural network for learn ing higher-order non-stationarity from spatio temporal dynamics[C]//Proceedings of The IEEE/CVF Conference on Computer Vision

and Pattern Recognition. Long Beach: IEEE, 2019: 9154-9162.

[112] MARTIN H, HONG Y, BUCHERD, et al. Traffic4cast-Traffic Map Movie Forecasting-Team MIE-Lab[EB/OL]. (2019-10-27) [2021-09-01]. https://arxiv. org/abs/1910.13824.

[113] EBERT F, FINN C, LEE A X, et al. Self-supervised visual planning with temporal skip connections[C]//Proceedings of the 1st Annual Conference on Robot Learning. Mountain View, California: PMLR, 2017: 344-356.

[114] DASARI S, EBERT F, TIAN S, et al. Robonet: Large-scale multi-robot learning[EB/OL]. (2019-10-24)[2021-09-01]. https://arxiv.org/abs/1910.11215v2.

[115] KWON Y H, PARK M G. Predicting future frames using retrospective cycle gan[C]//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1811-1820.

[116] HO Y H, CHO C Y, PENG W H, et al. Sme-net: Sparse motion estimation for parametric video prediction through reinforcement learning[C]//Proceedings of The IEEE/CVF International Conference on Computer Vision. Long Bea ch: IEEE, 2019: 10462-10470.

[117] LIANG X, LEE L, DAI W, et al. Dual motion GAN for future-flow embedded video prediction[C]//Proceedings of The IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1744-1752.

[118] XU J, NI B, YANG X. Video prediction via selective sampling[C]//Proceedings of The Advances in Neural Information Processing Systems. Montreal: MIT Press, 2018: 1712-1722.

[119] VILLEGAS R, PATHAK A, KANNAN H, et al. High fidelity video prediction with large stochastic recurrent neural networks[C]//Proceedings of the Advances in Neural Infor mation Processing Systems. Vancouver: MIT Press, 2019: 81-91.

[120] LOTTER W, KREIMAN G, Cox D. Unsuper vised learning of visual structure using predictive generative networks [EB/OL]. (2015-11-19)[2021-09-01]. https://arxiv.org/abs/1511.06380.

[121] MICHALSKI V, MEMISEVIC R, KONDA K. Modeling deep temporal dependencies with recurrent grammar cells [C]//Proceedings of The Advances in Neural Information Processing Systems. Quebec: MIT Press, 2014: 1925-1933.

[122] ARMENI I, SAX S, ZAMIR A R, et al. Joint 2d-3d-semantic data for indoor scene understanding[EB/OL]. (2017-02-03) [2021-09-01]. https://arxiv. org/abs/1702.01105.

[123] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3d: Learning from rgb-d data in indoor environments [EB/OL]. (2017-09-18)[2021-09-01]. https://arxiv.org/abs/1709.06158v1.

作者简介

潘敏婷 女,1996年生,广西贵港人. 现为上海交通大学博士研究生. 主要研究方向为视频数据预测模型.

E-mail: panmt53@sjtu.edu.cn

王韫博(通讯作者) 男,1989年生,吉林长春人. 现为上海交通大学人工智能研究院助理教授. 主要研究方向为深度学习,尤其是预测学习、时空动态系统建模、有模型的强化学习决策.

E-mail: yunbow@sjtu.edu.cn

朱祥明 男,2000年生,上海人. 现为上海交通大学本科生. 研究方向为计算机视觉、强化学习.

E-mail: xmzhu76@sjtu.edu.cn

高思宇 女,1999年生,山东青岛人. 现为上海交通大学硕士研究生. 主要研究方向为时空序列数据预测.

E-mail: siyu.gao@sjtu.edu.cn

龙明盛 男,1985年生,广西河池人. 现为清华大学副教授,国家优秀青年科学基金获得者. 主要研究方向为机器学习的理论和算法,尤其是迁移学习、深度学习和面向科学的机器学习方法.

E-mail: mingsheng@tsinghua.edu.cn

杨小康 男,1972年生,浙江东阳人. 现为上海交通大学教授,教育部长江学者特聘教授,国家杰出青年科学基金获得者,国家万人计划创新领军人才. 主要研究方向为计算机视觉和机器学习.

E-mail: xkyang@sjtu.edu.cn