文章编号:1671-6833(2025)05-0051-09

基于动态记忆与运动信息的目标中心视频预测算法

韩晨晨, 卢宪凯, 王志成, 熊筱舟

(山东大学 软件学院,山东 济南 250101)

摘 要: 针对在视频预测任务中需要维持视频帧间目标空间和时间一致性的问题,提出了基于动态记忆与运动信息的目标中心视频预测算法。首先,引入目标中心模型解耦场景中的目标,确保视频目标在长期动态预测中的一致性和稳定性,有效维持目标的空间一致性;其次,设计目标动态记忆模块,用于捕捉视频的长期依赖并对目标动态进行精确建模,克服现有视频预测方法在预测目标间动态交互上的不足,提升预测目标的时间一致性;再次,利用相邻帧的特征相似性矩阵捕捉帧间运动信息,构建视频序列的时空关系,强化帧间的时间一致性;最后,利用交叉注意力机制融合视频目标的时序和结构信息来提升视频预测效果。通过在具有复杂目标交互的 Obj3D 和 CLEVRER 数据集上进行视频预测实验,结果表明:相较于较先进的基于目标中心的视频预测算法,所提算法在 *PSNR*、*SSIM* 两个指标上性能分别提升了 4.5%,1.4%,并在 *LPIPS* 指标上降低了 20%。

关键词:视频预测; 目标中心学习; 场景解析; 无监督学习; 时空预测

中图分类号:TP391.4;TP183 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.02.011

受人类预见未来能力的启发,研究者们亦致力于探索和开发视觉感知模型预测未来的潜能。其中,视频预测任务即利用有限的历史视频帧预测未来帧,已成为学术界和工业界关注的热点[1]。视频预测任务在人体动作预测[2]、自动驾驶[3]及气象预警[4]等多个领域展现出巨大的应用前景。然而,在处理多目标动态交互的复杂场景时,由于预测结果常受到目标运动信息丢失(时间不一致)和目标结构失真(空间不一致)的影响,给下游应用带来了极大挑战。因此,如何对目标间的交互进行精准建模并保持时空一致性,仍是一个亟待解决的问题。解决视频预测的时空不一致问题,确保视频帧间目标的时空一致性,成为本领域研究的重点。

为了解决视频预测的时间不一致问题,准确地进行运动估计尤为关键。Sun 等[5]通过卷积神经网络来学习帧间的空间表示,但缺乏明确的时间状态转移机制,导致时间感知范围有限,运动信息容易丢失。Pan 等[6]采用循环神经网络的显式结构对时间状态进行编码,并通过卷积操作表征空间结构信息。然而,这些研究在复杂的时空特征提取和状态转换过程中,不可避免地牺牲了图像的外观细节,尽管运动预测较为准确,但未来帧的图像质量较差。此外,考虑到视频中的不同目标具有各自的几何形态和运动特性,Lee 等[7]尝试解耦视频中的目标,通过语义分割或实例分割信息在复杂场景中实现语义一致的视频预测,以解决长期预测中的空间不一致问题。然而,在实际应用中,语义或实例信息并非总是可用的,这限制了此类视频预测方法的适用范围。

为克服现有视频预测方法在维持时空一致性方面的局限,本文首先通过运动矩阵捕捉视频目标的运动信息,从而更准确地预测视频目标的运动趋势,保持视频序列的连贯性和一致性,显著改善时间不一致问题。此外,本文引入了无监督的目标中心学习方法[8],用于对视频场景进行解耦。通过该方法,视频目标被有效解构为一组目标中心特征(slot 特征),并通过解码这些 slot 特征生成未来的视频帧,实现在无密集标签的情况下维持视频帧间目标的空间一致性,有效解决了空间不一致问题。

收稿日期:2024-10-25;**修订日期**:2024-12-10

基金项目:山东省自然科学优秀青年基金资助项目(ZR2024YQ006);山东省高等学校青创团队计划(2023KJ027)

通信作者:卢宪凯(1990—),男,山东济南人,山东大学研究员,博士,主要从事计算机视觉、视频目标分割、目标跟踪、机器学习、遥感图像分析等研究,E-mail:luxiankai@sdu.edu.cn。

引用本文:韩晨晨,卢宪凯,王志成,等. 基于动态记忆与运动信息的目标中心视频预测算法[J]. 郑州大学学报(工学版),2025,46(5):51-59. (HAN C C, LU X K, WANG Z C, et al. Object-centric video prediction algorithm based on dynamic memory and motion information[J]. Journal of Zhengzhou University (Engineering Science), 2025, 46(5):51-59.)

综上所述,本文的贡献如下:

(1)本文提出了一种目标中心的视频预测方法,利用目标中心模型解耦场景目标的能力,有效维持了未来视频中目标的空间一致性。

(2)通过建立运动矩阵捕捉视频目标的运动信息,并结合slot动态记忆模块对目标动态进行建模,有效维持了未来视频帧的时间一致性。

1 相关工作

1.1 视频预测

近年来,视频预测任务在计算机视觉领域引起了广泛关注,针对这一挑战,众多方法相继被提出。早期的研究主要关注从图像序列中提取时空信息来维持视频的时间一致性。Lin等[9]通过卷积神经网络对空间数据进行编码,并结合LSTM模型捕捉时间依赖。Wang等[10]提出了一种基于ConvLSTM的循环节点间的记忆状态转换方法,更好地传递输入视频的外观特征来提升预测效果。Villegas等[11]利用帧间差分表示运动信息,分别对运动和外观特征进行编码,并将两者整合到一个端到端框架中。Voleti等[12]基于去噪扩散模型学习视频帧的时空关系,以历史帧为条件来实现视频预测任务。然而,这些研究虽然能在一定程度上预测时间一致的后续视频,但由于无法有效解耦视频中的目标,常导致预测结果出现模糊或扭曲的视觉外观,无法维持视频目标的空间一致性。为了解决这一问题,SLAMP[13]和video-to-video[14]方法利用光流对视频的外观和运动进行分解,以辅助视频预测。Bei等[15]设计了一种语义感知的动态模型,该模型预测并融合了未来视频帧的光流图和语义图。除了光流图和语义图,Wu等[16]进一步利用实例图来区分目标和背景。然而,这些方法通常需要额外的数据或模块,在处理多目标、复杂动作和高分辨率视频时,其效率难以保证。

为此,本文采用目标中心学习的方法解耦视频中的目标,实现目标中心的视频预测。该方法不仅能够更准确地预测目标之间的动态交互,还有效保留了目标的结构化信息,从而避免生成模糊或扭曲的视觉外观。同时,本文采用运动矩阵的方式实现时空预测,通过捕捉目标运动信息,实现了运动与外观的有效解耦,确保了运动预测的准确性以及高质量的图像外观。

1.2 目标中心学习

无监督的目标中心学习旨在将场景的模块化、组合性和因果结构表示为一组目标特征,而不需要额外的监督。这通常通过在Slot Attention[8]架构上引入归纳偏差来实现,这种偏差迫使模型将输入数据编码到一个集合结构的瓶颈中。在瓶颈中,目标的特征表示发生竞争或表现为排他性绑定,从而将场景分解为多个目标slot特征。目标中心学习的方法最初应用于合成图像数据,随后通过调整重建目标、3D场景分解和合成视频生成等技术手段,并结合多模态数据与先验知识,逐渐扩展到更复杂的图像处理领域。

目前,已有一些方法使用目标中心学习的方法解决视频预测任务。例如,OCVT[17]通过Transformer对多帧的目标表示进行处理,但OCVT依赖人工解缠的目标特征,且在训练过程中需要利用匈牙利匹配进行对齐,这限制了其性能。而SlotFormer[18]和OCVP[19]则利用视频目标中心模型,将视频场景分解为时序对齐的目标slot特征,并使用Transformer对交互目标进行建模以预测未来slot特征,然而,该方法未能充分利用目标运动信息,导致对目标运动的预测不够准确,影响了视频的时间一致性。为此,本文模型使用运动矩阵捕捉运动信息,并结合目标slot记忆模块预测目标间的动态交互,以得到更准确的未来slot特征。

2 目标中心的视频预测方法

2.1 问题及框架描述

给定一个包含 T 帧的视频序列 $X_{1,T}=\{x_t\in\mathbf{R}^{C\times H\times W}\}_{t=1}^{T}$,其中,$x_t$ 表示第 t 个视频帧,其宽度为 W,高度为 H,通道数为 C。视频预测任务旨在学习一个映射函数 $\mathcal{F}_\theta: X_{1,T}\rightarrow Y_{T+1,T+T'}$,其以 $X_{1,T}$ 作为输入,未来 T' 帧的预测视频序列 $Y_{T+1,T+T'}=\{y_t\in\mathbf{R}^{C\times W\times H}\}_{t=T+1}^{T+T'}$ 作为输出,其中,θ 表示模型的参数;y_t 为预测的第 t 个视频帧。

为解决上述任务,目标中心的视频预测模型网络结构如图1所示。模型首先利用CNN卷积网络提取视频序列 $X_{1,T}$ 的图像特征 $\{f_t\}_{t=1}^{T}$,基于运动矩阵的未来帧预测模块计算相邻帧的特征相似度矩阵 $\{\hat{\boldsymbol{M}}_{t,t+1}\}_{t=1}^{T-1}$,用其表示视频目标的运动趋势;其次,将其作为输入,利用3D卷积模型预测视频目标的未来运动趋势 $\{\hat{\boldsymbol{M}}_{T,T+t}\}_{t=1}^{T'}$;再次,通过特征聚合得到未来帧特征 $\{\hat{f}_{T+t}\}_{t=1}^{T'}$;最后,利用目标中心的场景分解模块,将未来帧特征解耦为目标slot特征 $\{\bar{S}_t\}_{t=T+1}^{T+T'}$。目标slot动态记忆模块首先利用目标中心的场景分解模块将图像特征 $\{f_t\}_{t=1}^{T}$ 解耦为一组目标slot特征 $\{S_t\}_{t=1}^{T}$,并输入记忆缓冲区;其次,利

用自注意力机制在记忆缓冲区对目标的动态交互进行建模,预测未来帧的 slot 特征 $\{\tilde{S}_t\}_{t=T+1}^{T+T'}$;再次,目标 slot 融合模块通过交叉注意力机制融合 $\{\bar{S}_t\}_{t=T+1}^{T+T'}$ 和 $\{\tilde{S}_t\}_{t=T+1}^{T+T'}$,得到融合视频目标时序和结构信息的 slot 特征 $\{\hat{S}_t\}_{t=T+1}^{T+T'}$;最后,模型利用目标中心的视频场景分解模块解码未来帧的 slot 特征 $\{\hat{S}_t\}_{t=T+1}^{T+T'}$,预测出未来帧图像 $\{\hat{y}_t\}_{t=T+1}^{T+T'}$。接下来将具体介绍模型的各个模块。

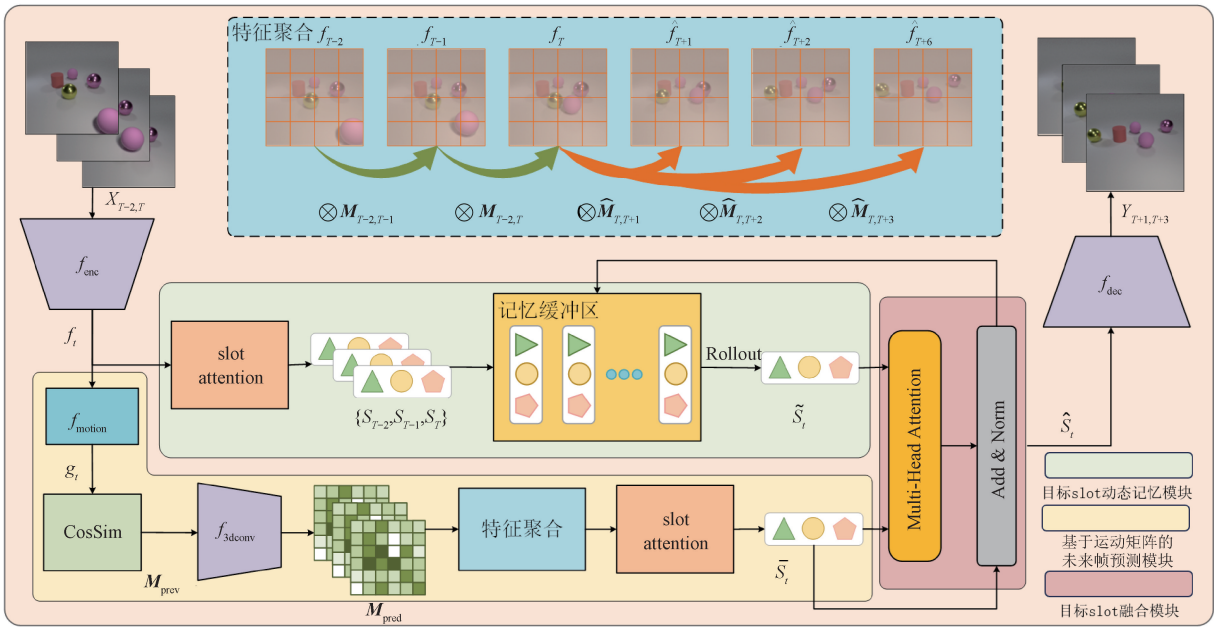

图 1　目标中心的视频预测模型网络结构

Figure 1　Network architecture of object-centric video prediction model

2.2　目标中心场景分解模块

本模型利用 SAVi[20] 模型解耦视频帧生成一组目标 slot 特征,实现对视频场景的有效分解,目标中心场景分解模块网络结构如图 2 所示。

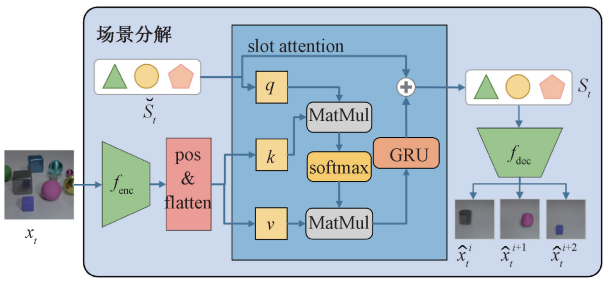

图 2　目标中心场景分解模块网络结构

Figure 2　Network architecture of object-centric scene decomposition module

对输入帧数为 T 的视频序列 $\{x_t\}_{t=1}^{T}$,目标中心模型计算生成 T 组目标 slot 特征 $\{S_t\}_{t=1}^{T}$,其中,$S_t \in \mathbf{R}^{N\times D_{\text{slot}}}$,由 N 个维度为 D_{slot} 的 slot 向量组成,N 为模型期望从视频中解耦出的目标数目。首先,模型通过卷积神经网络(CNN)编码器 f_{enc} 提取视频图像 x_t 的特征 f_t,并对特征添加位置编码和进行平坦化处理,得到一组特征 p_t,具体公式如下:

$$f_t = f_{\text{enc}}(x_t); \tag{1}$$

$$p_t = \text{flatten}(f_t + \text{pos}(f_t))。 \tag{2}$$

式中:$f_t \in \mathbf{R}^{h\times w\times D_{\text{enc}}}$;$p_t \in \mathbf{R}^{M\times D_{\text{enc}}}$,$M$ 为平坦化后特征图的尺寸,D_{enc} 为特征维度;flatten 为平坦化操作;pos 为位置编码操作。

接下来,模型构建一组初始的 slot 特征 $\breve{S}_t$。为实现 slot 向量在视频帧之间的时序对齐,即确保相同下标的 slot 向量在不同帧中表示相同的目标,模型对初始帧 $t=1$ 的 $\breve{S}_t$ 使用高斯采样进行构建,而对后续帧 $t \geqslant 2$ 的 $\breve{S}_t$,模型利用前一帧的 slot 特征 S_{t-1} 进行构建,通过这种逐帧传播的方式保证每帧的初始化 slot 特征是由前一帧的 slot 特征利用式(3)构建的,以实现视频目标在时序上的对齐,如式(3)所示:

$$\breve{S}_t = S_{t-1} + \text{Multi-HeadSelfattn}(S_{t-1})。 \tag{3}$$

式中:$\breve{S}_t$ 为第 t 帧的初始 slot 特征,$\breve{S}_t \in \mathbf{R}^{N\times D_{\text{slot}}}$;$S_{t-1}$ 为 $t-1$ 帧的 slot 特征;Multi-HeadSelfattn 表示多头自注意力函数。

为将 slot 向量与图像中的目标进行绑定以实现场景分解,模型对初始 slot 特征 $\breve{S}_t$ 执行 slot attention 操作,更新 slot 特征为 S_t:

$$S_t = f_{\text{SA}}(\breve{S}_t, p_t); \tag{4}$$

$$f_{\text{SA}}(\breve{S}_t, p_t) = \breve{S}_t + \text{GRU}(\breve{S}_t, f_{\text{dot}}(\breve{S}_t, p_t)); \tag{5}$$

$$f_{\text{dot}}(\breve{S}_t, p_t) = \text{softmax}\left(\frac{k(p_t) \cdot q(\breve{S}_t^T)}{\sqrt{D_{\text{slot}}}}\right) v(p_t)。 \quad (6)$$

式中:f_{SA} 表示 slot attention 机制;f_{dot} 表示点积注意力机制,其将 slot 向量与图像目标进行绑定,实现有效的场景分解;k、q、v 表示线性层,D_{slot} 表示变换特征维度;softmax 表示函数;GRU 为门控递归单元。

模型采用 slot 空间广播解码器 f_{dec}[21] 对目标 slot 特征进行解码以重构图像,公式如下:

$$\bar{m}_t^n, \bar{x}_t^n = f_{\text{dec}}(S_t^n); \quad (7)$$

$$\hat{m}_t = \text{softmax}(\bar{m}_t); \quad (8)$$

$$\hat{x}_t = \sum_{n=1}^{N} \bar{m}_t^n \odot \bar{x}_t^n。 \quad (9)$$

式中:S_t^n 为目标 n 在第 t 帧的 slot 表示;$\bar{x}_t^n$ 为目标 n 在第 t 帧的 RGB 图像预测,$\bar{x}_t^n \in \mathbf{R}^{C\times H\times W}$;$\bar{m}_t^n$ 为目标 n 在第 t 帧的掩码预测,$\bar{m}_t^n \in \mathbf{R}^{1\times H\times W}$;$\hat{m}_t$ 为 $\bar{m}_t^n$ 通过 softmax 函数归一化得到的加权掩码;$\hat{m}_t$ 与 $\bar{x}_t$ 进行加权求和得到合成重构帧 $\hat{x}_t$,$\hat{x}_t \in \mathbf{R}^{C\times H\times W}$;$\odot$ 表示哈达玛积。

目标中心模型采用无监督的方式,通过最小化重建帧损失来训练模型,重建损失 $\mathcal{L}_{\text{rec}}$ 公式如下:

$$\mathcal{L}_{\text{rec}} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2。 \quad (10)$$

2.3 基于运动矩阵的未来帧预测模块

本模型通过构建运动矩阵[22]的方式捕捉目标运动信息,预测未来视频帧特征,维护视频目标的时间一致性。

模型首先通过由若干卷积层构成运动捕捉模块 f_{motion},对编码器 f_{enc} 提取的图像特征 f_t 进行运动增强处理,过滤掉与运动无关的特征,得到增强后的特征 $g_t = f_{\text{motion}}(f_t) \in \mathbf{R}^{h\times w\times c}$,这使构建的运动矩阵更加聚焦于与运动紧密相关的特征。其次,模型利用余弦相似度来计算运动矩阵,由相邻帧特征 $\{g_t, g_{t+1}\}$ 得到运动矩阵 $\boldsymbol{M}_{t,t+1} \in \mathbf{R}^{h\times w\times h\times w}$。具体来说,对于运动矩阵中位置为 $(h_t, w_t, h_{t+1}, w_{t+1})$ 的元素 $M_{t,t+1}^{h_t,w_t,h_{t+1},w_{t+1}}$ 的计算公式如下:

$$M_{t,t+1}^{h_t,w_t,h_{t+1},w_{t+1}} = \text{CosSim}(g_t^{h_t,w_t}, g_{t+1}^{h_{t+1},w_{t+1}})。 \quad (11)$$

式中:CosSim 表示余弦相似度的计算函数;$g_t^{h_t,w_t}$ 和 $g_{t+1}^{h_{t+1},w_{t+1}}$ 分别表示相邻帧特征 g_t 和 g_{t+1} 在位置为 (h_t, w_t) 和 (h_{t+1}, w_{t+1}) 的特征。

对输入的 T 帧视频 $X_{1,T}$,首先,根据上述方法可获得由 $T-1$ 个运动矩阵组成的集合 $\boldsymbol{M}_{\text{prev}} = \{\boldsymbol{M}_{1,2}, \boldsymbol{M}_{2,3}, \cdots, \boldsymbol{M}_{T-1,T}\}$。为预测未来视频 $Y_{T+1,T+T'}$,模型使用 3D 卷积神经网络 f_{3Dconv} 对运动矩阵进行时序预测,预测未来的运动矩阵集合 $\boldsymbol{M}_{\text{pred}} = \{\hat{\boldsymbol{M}}_{T,T+1}, \hat{\boldsymbol{M}}_{T,T+2}, \cdots, \hat{\boldsymbol{M}}_{T,T+T'}\}$,以推测视频目标的未来运动趋势。其次,模型采用矩阵乘法的方式聚合未来帧特征。具体来说,这一过程将预测的运动矩阵与相应的特征图进行逐元素相乘,以合成未来帧的特征表示,具体公式如下:

$$\hat{f}_{T+t} = \sum_{i=1}^{T} [f_i \cdot (\prod_{n=i}^{T-1} \boldsymbol{M}_{n,n+1}) \cdot \hat{\boldsymbol{M}}_{T,T+t}]。 \quad (12)$$

式中:$\hat{f}_{T+t}$ 表示预测的第 $T+t$ 帧的特征,$\hat{f}_{T+t} \in \mathbf{R}^{h\times w\times D_{\text{enc}}}$。最后,模型通过 2.2 节所述的目标中心模块解耦 $\hat{f}_{T+t}$,得到一组 slot 特征 $\bar{S}_{T+t} \in \mathbf{R}^{N\times D_{\text{slot}}}$,用来表示未来视频帧 y_{T+t} 的目标 slot 特征。

通过上述方式,模型利用运动矩阵中包含的运动信息,将不同时间步的特征图按照预测的运动趋势进行融合,合成未来帧的特征表示,并解耦目标 slot 特征。这种方法成功解耦了运动和外观预测,既捕捉了运动信息,又保留了特征图中的空间信息,有助于生成更准确、更自然的未来帧预测结果,维护视频预测过程中视频目标的时间一致性。

2.4 目标 slot 动态记忆模块

为更准确地预测目标间的动态交互,提升预测视频的时间一致性,本模型引入了 slot 动态记忆模块。此模块将历史帧的目标 slot 特征输入到记忆缓冲区中,并通过前向滚动机制捕捉历史 slot 特征的长期依赖,对目标间的动态交互建模。

模型通过目标中心模块对视频帧特征 $\{f_t\}_{t=1}^{T}$ 进行场景分解,得到一组 slot 特征序列 $\{S_t\}_{t=1}^{T}$,并将 $\{S_t\}_{t=1}^{T}$ 输入记忆缓冲区 $\mathcal{M} \in \mathbf{R}^{L\times N\times D_{\text{slot}}}$,其中 L 为记忆缓冲区的长度,N 为每帧分解的 slot 向量的数量。记忆缓冲区采用先进先出的数据结构实现,确保为每个目标保留最多 L 个时间步的 slot 特征。在时间步 t,记忆缓冲区会向前滚动,整合多帧的目标特征得到 $\tilde{S}_t \in \mathbf{R}^{N\times D_{\text{slot}}}$。具体公式如下:

$$\tilde{S}_t = \text{Rollout}(\mathcal{M}_{<t}); \quad (13)$$

$$\text{Rollout}(\mathcal{M}_{<t}) = \tilde{S}_{t-1} + \text{Linear}(\text{Multi-HeadSelfattn}(\mathcal{M}_{<t}))。 \quad (14)$$

式中:$\mathcal{M}_{<t}$ 表示记忆缓冲区中时间步数小于 t 的 slot 特征;Rollout 表示记忆向前滚动过程;Linear 表示线性层。

在 Rollout 过程中,模型首先利用自注意机制擅长捕捉长期依赖的优势对记忆缓存 $\mathcal{M}_{<t}$ 进行时序

建模,以捕捉视频目标间的长期动态交互关系;其次,利用线性层进行维度变换;最后,通过残差连接保留原始 slot 特征的信息,得到时间步 t 的目标 slot 特征 $\widetilde{S}_t$。

2.5 目标 slot 融合模块

在 2.3 节和 2.4 节,模型分别得到了两组 slot 特征:$\bar{S}_t$ 和 $\widetilde{S}_t$。其中,$\bar{S}_t$ 是通过运动矩阵预测的未来帧特征经过目标中心模块解耦得到的,因此其对目标运动的预测更为精确;$\widetilde{S}_t$ 是通过 slot 记忆模块捕捉视频的长期依赖关系生成的,其更好地预测了目标间的动态交互,并蕴含丰富的时序信息。为结合这两组 slot 特征的优势,模型引入了 slot 融合模块,其利用交叉注意力机制实现了 $\bar{S}_t$ 和 $\widetilde{S}_t$ 的有效融合,公式如下:

$$\hat{S}_t = \bar{S}_t + \text{Multi-HeadSelfattn}(q = \bar{S}_t, k = \widetilde{S}_t, v = \widetilde{S}_t)。\tag{15}$$

式中:$\hat{S}_t$ 表示融合 slot 特征,$\hat{S}_t \in \mathbf{R}^{N\times D_{\text{slot}}}$。

通过 slot 解码器 f_{dec} 解码融合的 slot 特征 $\hat{S}_t$,即得到预测的未来帧 $\hat{y}_t$,此外,$\hat{S}_t$ 也被插入到记忆缓冲区中,用以更新和扩展记忆内容。模型通过最小化未来图像帧和未来 slot 特征的均方差(*MSE*)损失对未来帧预测模块进行监督训练,公式如下:

$$\mathcal{L}_{\text{I}} = \frac{1}{T'}\sum_{t=T+1}^{T+T'} \| \hat{y}_t - y_t \|^2;\tag{16}$$

$$\mathcal{L}_{\text{S}} = \frac{1}{T'\cdot N}\sum_{t=T+1}^{T+T'}\sum_{n=1}^{N} \| \hat{S}_t^n - S_t^n \|^2;\tag{17}$$

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{I}} + \mathcal{L}_{\text{S}}。\tag{18}$$

式中:$\mathcal{L}_{\text{I}}$ 表示图像 *MSE* 损失;$\mathcal{L}_{\text{S}}$ 表示 slot 特征的 *MSE* 损失;$\mathcal{L}_{\text{pred}}$ 表示总的视频预测损失。

3 实验结果与分析

3.1 数据集

遵循以往目标中心的视频预测研究[18-19],并考虑到本文使用的目标中心模型仅能分解合成视频数据的局限性,本文选取 Obj3D[23] 和 CLEVRER[24] 数据集进行算法性能的测试。但目前目标中心模型发展迅速,已有研究[25]应用于自然视频数据,本文模型在未来也将扩展到自然视频预测。

Obj3D 数据集包含训练视频 2 920 个,测试视频 200 个,视频分辨率为 64×64。数据通过在场景中放置 3 至 5 个静态物体,然后从场景前方发射一个球体与这些物体碰撞而生成。

CLEVRER 数据集包含 10 000 个训练视频和 5 000 个测试视频,视频分辨率为 64×64。该数据集有着更多的物体数目和更多样化的目标交互,比 Obj3D 数据集更具挑战性。

3.2 实验环境与模型参数

本文采用的软件运行平台为 Ubuntu20.04 版 64 位,深度学习环境软件配置为 Python3.9 和 PyTorch1.11。硬件配置为 NVIDIA3090 显卡,采用 CUDA11.2,使用 Adam 优化器、OneCycle 学习率调整策略来训练模型。

模型的超参数主要包括学习率、训练轮数、批处理大小、输入特征图尺寸、slot 特征维度、slot 特征数量等。在 Obj3D 数据集和 CLEVRER 数据集上,学习率分别设置为 0.001 0 和 0.000 5,由于 CLEVRER 数据集相较于 Obj3D 数据集目标数量更多且交互更复杂,故在 CLEVRER 数据集上设置更小的学习率,以保证模型更稳定的训练;根据模型的收敛情况,训练轮数分别设置为 200 和 50;根据服务器性能,批处理大小统一设置为 16;过大的下采样比例会影响目标中心模型的性能,过小的下采样比例会导致计算运动矩阵时消耗更多资源,故统一设置编码器下采样比例为 8,并设置特征图维度 D_{enc} 为 128;为平衡计算效率和模型性能,统一设置 3D 卷积层数为 3,记忆缓冲区长度 L 为 6;根据 bj3D 和 CLEVRER 两个数据集视频中的目标数量,分别设置 slot 特征数量为 6 和 7,并统一设置 slot 特征维度 D_{slot} 为 128;输入帧数 T 统一设置为 5,并在预测帧数 T' 为 15 和 25 的情况下进行训练和实验。

模型首先在数据集上用重建损失 $\mathcal{L}_{\text{rec}}$ 对目标中心模块进行预训练,以确保场景分解得到的目标 slot 特征能够准确地表示场景中的目标,随后使用预测损失 $\mathcal{L}_{\text{pred}}$ 训练 2.3~2.5 节所述未来帧预测模块。

3.3 测试指标

本文遵循前人的研究,使用峰值信噪比 *PSNR*[26]、结构相似指数 *SSIM*[26] 和感知图像相似性 *LPIPS*[27] 3 个指标来评估视频预测性能。

PSNR 通过比较原始图像与处理后图像之间的像素差异来量化图像质量;*SSIM* 则是一种更为全面的图像相似度评估方法,它不仅考虑了亮度、对比度和结构等关键信息,而且更符合人类的主观感知;*LPIPS* 是一种基于深度学习的感知图像相似性评估指标,它通过模拟人类视觉系统的感知特性,更准确地衡量了两个图像之间的感知差异,而不仅仅是简单的像素级比较。

此外,本文使用调整兰德指数 *ARI* 和平均交并比 *mIoU* 对预测的实例掩码和边界框进行评测,以评估模型在目标动态建模方面的能力。

3.4 视频预测实验和结果

表1为视频预测定量结果,图3展示了本文方法与其他方法在Obj3D数据集上视频预测定性结果。从表1数据中看出,本文方法在 *PSNR* 和 *LPIPS* 指标上优于所有基准模型,在 *SSIM* 指标上也展现出竞争力。

与基于目标中心的方法相比,本文方法在各指标上均有显著提升。在Obj3D数据集上,在预测帧数为15时,本文方法与OCVP-Seq相比,在 *PSNR*、*SSIM* 指标分别提升了4.5%和1.4%,并在 *LPIPS* 指标上降低了20%,证明了本文方法在性能上超越了仅依赖Transformer的OCVP-Seq预测方法。相比于Obj3D数据集,模型在CLEVRER数据集上随着预测帧数的增加有着更严重的性能损失。这首先是因为CLEVRER数据集有着更复杂的运动交互,导致了更多的误差累积;其次,CLEVRER数据集在视频中会出现新的目标,模型无法预测输入帧中不存在的目标,且新出现的目标也会影响目标中心模型的时序对齐,这也是目标中心模型待解决的问题。

实验结果显示,本文方法与非目标中心的视频预测方法相比,在性能上也展现出了明显优势,特别是在更能体现人类感知的 *LPIPS* 指标上,在Obj3D数据集上,在预测帧数为15时,相比于ConvLSTM下降了56.5%。结合图3的可视化结果,验证了本文方法通过目标中心模块解耦视频目标,更好地维持了视频目标的结构化信息,有效避免了目标结构的失真。在 *SSIM* 指标上,本文方法和其他目标中心方法一样,与非目标中心的方法相比较差。这主要因为 *SSIM* 指标考虑到了图像的亮度、对比度和结构,并重点关注图像局部细节。而目标中心模型在解耦目标的过程中会对目标亮度、对比度和

表1 视频预测定量结果

Table 1 Quantitative results of video prediction

方法类型	方法	Obj3D						CLEVRER					
		预测帧数=15			预测帧数=25			预测帧数=15			预测帧数=25		
		PSNR	*SSIM*	*LPIPS*	*PSNR*	*SSIM*	*LPIPS*	*PSNR*	*SSIM*	*LPIPS*	*PSNR*	*SSIM*	*LPIPS*
非目标中心方法	ConvLSTM[9]	34.32	0.929	0.046	28.95	0.849	0.112	27.31	0.892	0.232	26.86	0.891	0.367
	PhyDNet[10]	30.08	0.921	0.034	28.21	0.892	0.053	27.32	0.891	0.187	26.70	0.880	0.202
	STMFANet[28]	31.37	0.934	0.032	29.35	0.921	0.047	28.21	0.901	0.163	27.23	0.899	0.213
	SimVP[29]	33.78	**0.963**	0.024	32.61	0.942	0.039	29.27	0.919	0.118	28.39	0.918	0.154
	MMVP[22]	33.41	0.960	0.027	32.45	**0.946**	0.043	29.39	**0.923**	0.109	28.53	**0.927**	0.177
目标中心方法	SlotFormer[18]	32.89	0.931	0.025	30.87	0.892	0.041	30.13	0.895	0.063	28.24	0.858	0.103
	OCVP-Seq[19]	33.10	0.932	0.025	30.93	0.891	0.041	30.74	0.905	0.055	28.69	0.869	0.099
	OCVP-Par[19]	32.99	0.931	0.025	30.85	0.890	0.043	30.63	0.899	0.057	28.24	0.864	0.101
	本文方法	**34.58**	0.945	**0.020**	**32.87**	0.919	**0.029**	**31.53**	0.914	**0.041**	**28.86**	0.881	**0.093**

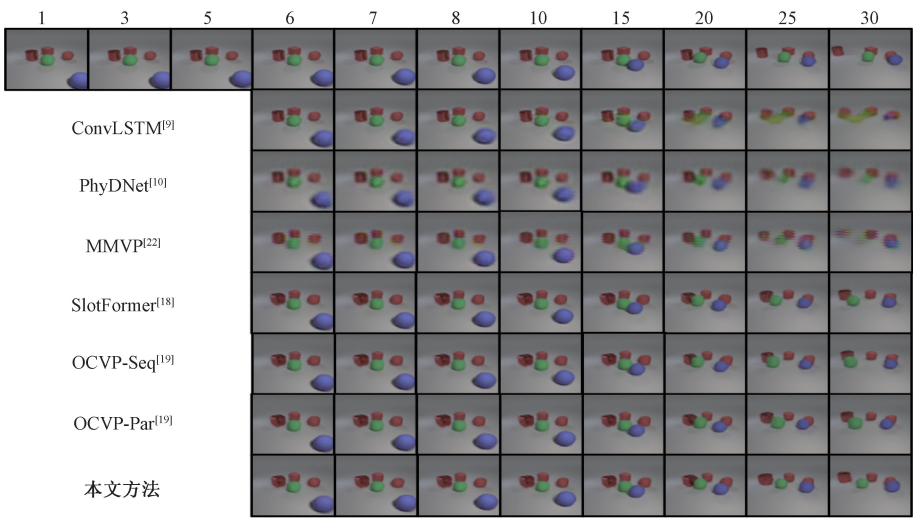

图3 视频预测定性结果

Figure 3 Qualitative results of video prediction

结构产生误差,这些局部误差对 *SSIM* 有着显著影响。

由图 3 可以看出,与非目标中心的方法相比,本文方法预测的早期图像质量较差,这主要因为通过目标中心模块解耦视频目标导致的外观损失。由于非目标中心的方法缺乏对目标的明确认知,在几个时间步后预测质量就会显著降低,无法维持目标的空间一致性,而本文方法更好地维持了目标结构。相比于其他基于目标中心的方法,本文方法更准确地预测了目标的运动趋势和碰撞交互,更好地维持了视频目标的时间一致性。

3.5　目标动态预测实验和结果

本节通过评估目标中心场景分解模块生成的目标边界框和分割掩码的质量,来衡量模型对视频目标动态建模的能力。具体而言,通过解码未来帧的目标 slot 特征,生成每个目标的掩码权重,使用 argmax 函数获得各目标的分割掩码,并与真实掩码进行对比,计算 *ARI* 和 *mIOU* 指标。在 CLEVRER 数据集上,本文方法与其他方法目标动态定量结果如表 2 所示。

表 2　目标动态定量结果

Table 2　Quantitative results of object dynamics

方法	*ARI*	*mIOU*
SlotFormer[18]	0.609	0.568
OCVP-Seq[19]	0.623	0.575
OCVP-Par[19]	0.631	0.583
本文方法	**0.638**	**0.586**

相比于其他目标中心的方法,本文方法在 *ARI* 和 *mIOU* 指标上取得了最佳性能,说明本文方法实现了更好的目标动态预测,能够精确地预测目标的运动轨迹和目标间的碰撞交互,并有效维持目标的结构完整性。

图 4 为目标动态预测定性结果,以可视化方式展示了本文方法预测的未来帧图像和分割掩码,相比于非目标中心的方法以视频帧为单位的预测,本文方法可预测视频中每个目标的运动轨迹,这给视频预测带来了更广阔的应用前景,在未来可扩展至多人动作预测和自动驾驶等任务。

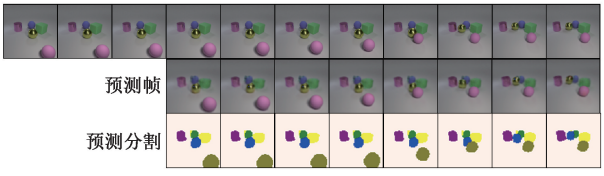

图 4　目标动态预测定性结果

Figure 4　Qualitative results of object dynamics prediction

3.6　消融实验

本节在 Obj3D 数据集上进行消融实验,以验证本文方法中各个模块的作用。消融实验结果如表 3 所示。由表 3 可知,相比于 No-Motion 项,加入基于运动矩阵的视频预测模块后,基于本文方法的 *PSNR* 和 *SSIM* 指标分别提高了 5.2%和 3.5%,说明该模块对提升视频预测准确性具有显著作用。相比于 No-Memory 项,加入目标 slot 动态记忆模块后,基于本文方法的 *PSNR* 和 *SSIM* 指标分别提高了 2.6%和 0.9%,表明该模块有助于视频预测性能的进一步增强。分别在记忆缓冲区长度 L 为 2、4 的情况下进行视频预测实验,实验结果表明,在记忆容量大的情况下,本文模型可以捕捉更多的长期依赖,用以预测目标间的交互,使得预测结果更准确。分别在 3D 卷积层数 Conv3Dlayers 为 1、2 的情况下进行视频预测实验,实验结果表明,更多的 3D 卷积层可获取更好的预测结果,主要原因是更多的 3D 卷积层可更准确地预测视频目标的运动趋势。

表 3　消融实验结果

Table 3　Ablation experiment results

消融项	*PSNR*	*SSIM*	*LPIPS*
No-Motion	32.86	0.913	0.045
No-Memory	33.72	0.937	0.029
L-4	34.23	0.940	0.023
L-2	33.93	0.943	0.028
Conv3Dlayers-2	34.45	0.939	0.024
Conv3Dlayers-1	34.34	0.934	0.027
本文方法	**34.58**	**0.945**	**0.020**

4　结论

本文提出了一种视频预测算法,该算法通过目标中心模型将场景分解为独立的视频目标,为每个目标提取出精确的 slot 特征,并有效融合运动信息对目标动态进行建模。通过这种目标中心的方法,算法能够确保视频目标长期动态预测的一致性与稳定性,维持目标的空间一致性,避免生成模糊或扭曲的视觉效果。此外,算法引入了运动矩阵来精准捕捉运动信息,并通过动态记忆模块对目标的动态变化进行建模。这使算法能够学习目标间复杂的时空交互,实现对目标动态的精确预测,确保视频序列的时间一致性。实验结果表明,本文算法在多个复杂交互的数据集上表现优异,证明结合运动信息与动态记忆的目标中心时空预测方法在处理复杂动态场景时具有巨大潜力。

参考文献:

[1] 李卫军, 张新勇, 高庚潇, 等. 基于门控时空注意力的视频帧预测模型[J]. 郑州大学学报(工学版), 2024, 45(1): 70-77, 121.
LI W J, ZHANG X Y, GAO Y X, et al. Video frame prediction model based on gated spatio-temporal attention [J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(1): 70-77, 121.

[2] MARTINEZ J, BLACK M J, ROMERO J. On human motion prediction using recurrent neural networks[C]// 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2891-2900.

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

[4] DAI K, LI X T, YE Y M, et al. MSTCGAN: multiscale time conditional generative adversarial network for long-term satellite image sequence prediction [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[5] SUN F, BAI C, SONG Y, et al. MMINR: multi-frame-to-multi-frame inference with noise resistance for precipitation nowcasting with radar[C]//The 26th International Conference on Pattern Recognition. Piscataway: IEEE, 2022: 97-103.

[6] PAN T, JIANG Z Q, HAN J N, et al. Taylor saves for-later: disentanglement for video prediction using Taylor representation [J]. Neurocomputing, 2022, 472: 166-174.

[7] LEE W, JUNG W, ZHANG H, et al. Revisiting hierarchical approach for persistent long-term video prediction[EB/OL]. (2021-04-14)[2024-08-10]. https://doi.org/10.48550/arXiv.2104.06697.

[8] LOCATELLO F, WEISSENBORN D, UNTERTHINER T, et al. Object-centric learning with slot attention[J]. Advances in Neural Information Processing Systems, 2020, 33: 11525-11538.

[9] LIN Z H, LI M M, ZHENG Z B, et al. Self-attention ConvLSTM for spatiotemporal prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11531-11538.

[10] WANG Y B, LONG M S, WANG J M, et al. PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs[J]. Advances in Neural Information Processing Systems, 2017, 30: 879-888.

[11] VILLEGAS R, YANG J M, HONG S, et al. Decomposing motion and content for natural video sequence prediction[EB/OL]. (2017-07-25)[2024-08-10]. https://doi.org/10.48550/arXiv.1706.08033.

[12] VOLETI V S, JOLICOEUR-MARTINEAU A, PAL C. MCVD: masked conditional video diffusion for prediction, generation, and interpolation[J]. Advances in Neural Information Processing Systems, 2022, 36: 23371-23385.

[13] AKAN A K, ERDEM E, ERDEM A, et al. SLAMP: stochastic latent appearance and motion prediction[C]// 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 14708-14717.

[14] WANG T C, LIU M Y, ZHU J Y, et al. Video-to-video synthesis[EB/OL]. (2018-08-20)[2024-08-10]. https://doi.org/10.48550/arXiv.1808.06601.

[15] BEI X Z, YANG Y C, SOATTO S. Learning semantic-aware dynamics for video prediction[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 902-912.

[16] WU Y, GAO R R, PARK J, et al. Future video synthesis with object motion prediction[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5538-5547.

[17] WU Y F, YOON J, AHN S. Generative video transformer: can objects be the words? [EB/OL]. (2021-07-20)[2024-08-10]. https://doi.org/10.48550/arXiv.2107.09240

[18] WU Z Y, DVORNIK N, GREFF K, et al. SlotFormer: unsupervised visual dynamics simulation with object-centric models[EB/OL]. (2022-10-12)[2024-08-10]. https://doi.org/10.48550/arXiv.2210.05861.

[19] VILLAR-CORRALES A, WAHDAN I, BEHNKE S. Object-centric video prediction via decoupling of object dynamics and interactions[C]//2023 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2023: 570-574.

[20] ELSAYED G F, MAHENDRAN A, VAN STEENKISTE S, et al. SAVi++: towards end-to-end object-centric learning from real-world videos[EB/OL]. (2022-06-15)[2024-08-10]. https://doi.org/10.48550/arXiv.2206.07764.

[21] WATTERS N, MATTHEY L, BURGESS C P, et al. Spatial broadcast decoder: a simple architecture for learning disentangled representations in VAEs[EB/OL]. (2019-06-21)[2024-08-10]. https://doi.org/10.48550/arXiv.1901.07017.

[22] ZHONG Y Q, LIANG L M, ZHARKOV I, et al. MMVP: motion-matrix-based video prediction[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4250-4260.

[23] LIN Z X, WU Y F, PERI S, et al. Improving generative

imagination in object-centric world models [EB/OL]. (2020-10-05) [2024-08-10]. https://doi.org/10.48550/arXiv.2010.02054.

[24] YI K X, GAN C, LI Y Z, et al. CLEVRER: CoLlision events for video REpresentation and reasoning[EB/OL]. (2019-10-03) [2024-08-10]. https://doi.org/10.48550/arXiv.1910.01442.

[25] ZADAIANCHUK A, SEITZER M, MARTIUS G. Object-centric learning for real-world videos by predicting temporal feature similarities[J]. Advances in Neural Information Processing Systems, 2023, 36: 61514-61545.

[26] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[27] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 586-595.

[28] JIN B B, HU Y, TANG Q K, et al. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4553-4562.

[29] GAO Z Y, TAN C, WU L R, et al. SimVP: simpler yet better video prediction[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3160-3170.

Object-centric Video Prediction Algorithm Based on Dynamic Memory and Motion Information

HAN Chenchen, LU Xiankai, WANG Zhicheng, XIONG Xiaozhou

(School of Software, Shandong University, Jinan 250101, China)

Abstract: In response to the challenges of maintaining structural and temporal consistency between video frames in video prediction tasks, an object-centric video prediction algorithm based on dynamic memory and motion information was proposed. Firstly, by introducing an object-centric model, the objects in the scene were decoupled to ensure the consistency and stability of long-term dynamic modeling of video objects, to effectively maintain the structural consistency of video objects. Secondly, an object dynamic memory module was designed to capture the long-term dependencies of videos and model object dynamics, to overcome the shortcomings of existing video prediction methods in predicting dynamic interactions between objects and enhancing the temporal consistency of video objects. Thirdly, the feature similarity matrix of adjacent frames was used to capture the motion information between frames and model the spatiotemporal relationships of the video sequence, further strengthened the temporal consistency of video objects. Finally, a cross-attention mechanism was utilized to integrate the temporal and structural information of video objects, further improved the video prediction performance. Experiments on video prediction were conducted on the Obj3D and CLEVRER datasets with complex object interactions. The results showed that compared to the state-of-the-art object-centric video prediction algorithms, the proposed algorithm increased performance on the *PSNR* and *SSIM* metrics by 4.5% and 1.4%, respectively, and also achieved a 20% reduction in the *LPIPS* metric.

Keywords: video prediction; object-centric learning; scene parsing; unsupervised learning; spatiotemporal prediction