融合 U-net 网络的纯卷积视频预测模型

谢玉枚[1]，蔡远利[2]，高海燕[3]，关翔锋[1]，唐伟强[4]

（1. 福建江夏学院电子信息科学学院，350108，福州；2. 西安交通大学电子与信息学部，710049，西安；
3. 厦门理工学院电气工程与自动化学院，361024，福建厦门；4. 兰州理工大学电气工程与
信息工程学院，730050，兰州）

摘要：为了解决基于深度学习视频预测中存在的时空特征提取不充分以及图像细节保留不足的问题，运用简单视频预测网络模型 SimVP 给出的 Inception 单元，提出了一种融合 U-net 网络的纯卷积视频预测模型(CUnet)。CUnet 模型由 3 个核心模块组成：首先，Cell 模块采用 2D 卷积层来提取空间特征，并将这些特征输入至多个 Inception 单元捕获时空特性；其次，DeCell 模块通过 Inception 单元捕获时空特征，并借助 2D 反卷积层进行上采样操作，恢复图像原始尺寸；最后，引入 U-net 作为主干网络，将 Cell 模块和 DeCell 模块有机整合，有效保留了图像的细节信息，实现了高质量的图像重建。实验结果表明：在 TaxiBJ 数据集上，与当前表现最佳的时间注意力单元网络模型 TAU 相比，CUnet 模型的预测精度提高了 5.23%；在 Human3.6M 数据集上，与当前表现最佳的快速傅里叶 Inception 网络模型 FFINet 相比，CUnet 模型的预测精度提高了 12.88%。CUnet 模型具有优秀的预测能力，可为纯卷积神经网络模型在视频预测领域的应用提供有益探索。

关键词：深度学习；视频预测；时空特征；U-net 网络；纯卷积神经网络

中图分类号：TP391.41 **文献标志码**：A

DOI：10.7652/xjtuxb202506012 **文章编号**：0253-987X(2025)06-0112-10

A Pure Convolutional Model Fused with U-net Network for Video Prediction

XIE Yumei[1], CAI Yuanli[2], GAO Haiyan[3], GUANG Xiangfeng[1], TANG Weiqiang[4]

(1. School of Electronic Information Science, Fujian Jiangxia University, Fuzhou 350108, China; 2. Faculty of Electrical and Information Engineering, Xi'an jiaotong University, Xi'an 710049, China; 3. School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen, Fujian 361024, China; 4. College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: To address the issues of insufficient spatiotemporal feature extraction and inadequate image detail preservation in deep learning-based video prediction, a pure convolutional video prediction model (CUnet) fused with the U-net network, using the Inception unit from the SimVP model, is proposed. CUnet model consists of 3 core modules. Firstly, the Cell module uses 2D convolutional layers to extract spatial features and feeds these features into multiple Inception units to capture spatiotemporal features. Secondly, the DeCell module captures spatiotemporal features through Inception units and performs upsampling operations using 2D deconvolutional layers to restore the original image size. Finally, U-net is introduced as the backbone network to organically integrate the Cell module and the DeCell module, effectively

收稿日期：2024-09-30。 作者简介：谢玉枚(1986—)，女，副教授；关翔锋(通信作者)，男，教授。 基金项目：国家自然科学基金资助项目(62463017)；福建省自然科学基金资助项目(2023J011099)。
网络出版时间：2025-02-26 网络出版地址：https://link.cnki.net/urlid/61.1069.t.20250226.0941.004

preserving the detailed information of the image and achieving high-quality image reconstruction. The experimental results showed that on the TaxiBJ dataset, compared with the currently best-performing TAU model, the prediction accuracy of the CUnet model had increased by 5.23%. On the Human3.6M dataset, compared with the currently best-performing FFINet model, the prediction accuracy of the CUnet model had increased by 12.88%. The CUnet model demonstrates exceptional predictive capabilities, offering valuable insights for the application of pure convolutional neural networks in the field of video prediction.

Keywords: deep learning; video prediction; spatiotemporal features; U-net; pure convolutional neural network

视频预测是通过对历史帧的学习，实现对未来帧的精准预测。视频预测技术已在交通流预测[1-4]、行为模式分析[5-7]等多个领域广泛应用。与静态图像相比，视频资料涵盖了更为复杂的动态信息，如运动轨迹、遮挡现象以及物体间的交互作用等，这使得视频预测成为视频表征领域一项充满挑战的研究课题。

随着人工智能技术的快速发展，近年来涌现了大量基于深度学习的视频预测方法[8]。凭借对时序数据强大的建模能力，循环神经网络模型在视频预测任务中得到了广泛的应用。Shi 等[9]提出了卷积长短期记忆网络模型(ConvLSTM)，该模型将全连接的长短期记忆网络(LSTM)扩展为具有卷积计算特性的 ConvLSTM 单元，并通过堆叠这些单元构建了 ConvLSTM 模型，提升了模型对空间特征的提取能力。为了有效地捕捉时空相关性，Wang 等[10]在设计时空单元的基础上提出了预测循环神经网络模型(PredRNN)，该模型可以同步提取并储存空间与时间信息，在时空序列预测数据集上取得了较好的预测效果。为缓解 PredRNN 模型梯度易消失问题及提高对短时非线性时空特征的提取能力，Wang 等[11]提出了 PredRNN＋＋模型，该模型能有效地抑制梯度消失。此外，为解决 PredRNN 中 LSTM 遗忘门饱和的问题，Wang 等[12]提出了记忆中记忆网络模型(MIM)，该模型将时空单元中的遗忘门替换为非平稳模块和平稳模块，从而提升了模型捕捉时空特征能力。

在视频预测中，纯循环神经网络模型在捕捉长距离依赖关系上存在挑战。为了解决这个问题，研究者将循环神经网络与卷积神经网络相结合，使用卷积神经网络来学习空间特征，使用循环神经网络来学习时间动态变化。Wang 等[13]提出了三维长短期记忆网络模型(E3D-LSTM)，该模型将三维卷积融入递归神经网络中，并引入自注意力模块，增强了模型捕捉长距离依赖关系的能力。Yu 等[14]提出了条件可逆网络模型(CrevNet)，该模型通过嵌套三维卷积模块的双向可逆自编码结构，能够保留更多的时空信息，进而提高了预测图像的清晰度。面对更复杂场景时，CrevNet 模型在预测运动趋势方面仍有局限，针对这一情况，Guen 等[15]提出了物理动态网络模型(PhyDNet)，该模型通过基于卷积的物理模块单元学习物理过程信息，并结合学习生成、消亡和发展等未知因素，提升了视频预测的精度。Chang 等[16]提出了运动感知单元网络模型(MAU)，该模型通过扩展预测单元的时间感受野，有效地捕捉了帧与帧之间的运动信息，提高了视频预测的准确性和连贯性。为了增强对时空特征的提取能力，从纯循环神经网络模型到循环神经网络与卷积神经网络融合的模型，网络模型结构日益复杂，但其预测性能的提升却并不显著。

传统的卷积神经网络主要用于捕捉空间特征，并不擅长处理时序信息，因而在视频预测任务中，采用纯卷积神经网络模型并不常见。2022 年，Gao 等[17]提出纯卷积神经网络模型——简单视频预测网络模型(SimVP)，该模型首先通过 2D 卷积提取空间特征，然后利用多个 Inception 单元捕捉时空特性，最后通过 2D 反卷积操作预测未来视频帧，具有很强的泛化能力和可扩展性。Li 等[18]在 SimVP 模型基础上提出了快速傅里叶 Inception 网络模型(FFINet)，该模型将 2D 卷积替换为快速傅里叶卷积，并且将 Inception 单元改为傅里叶变换 Inception 单元学习时间演变，提升了模型的预测能力。Tan 等[19]在 SimVP 模型基础上提出了时间注意力单元网络模型(TAU)，该模型将 Inception 单元替换为注意力模块，提升了并行处理能力和捕捉长距离时间演变的表现。然而，当前这些纯卷积神经网络模型在保留图像细节和图像重建方面仍存在一定的不足。

此外，Wen 等[20]基于生成对抗网络提出了端到端的视频生成方法，该方法通过两个级联的生成对抗网络捕捉运动和生成图像细节。为缓解预测步数增加导致的预测视频帧模糊问题，Kwon 等[21]提出了包含一个生成器和两个判别器的生成对抗网络模型，该模型能有效地保持过去帧和预测未来帧与原视频序列的一致性。为提高预测视频帧质量，Yu 等[22]提出了动态感知隐式生成对抗网络，该模型将连续信号编码到神经网络中，能有效提取视频序列的连续动态。然而，与循环神经网络和卷积神经网络相比，生成对抗网络模型的对抗训练过程存在不稳定性，同时生成对抗网络模型通常难以在对抗损失和重建损失之间保持平衡，这往往会造成预测结果出现模糊的情况[23]。

U-net 网络常应用于图像语义分割任务中，其在特征融合、图像细节捕获以及图像重建方面已获得广泛认可。近年来，将 U-net 网络应用于交通预测和视频异常检测等领域也取得了令人满意的结果[24-28]。纯卷积神经网络模型结构简单且预测效果出色，将 U-net 网络与纯卷积神经网络相结合有望进一步提高预测效果，但目前这方面的研究非常少。鉴于此，本文在 SimVP 模型给出的 Inception 单元基础上，提出一种融合 U-net 网络的纯卷积视频预测模型 CUnet。CUnet 模型包括 3 个部分：第 1 部分是 Cell 模块，视频数据先通过 2D 卷积提取空间特性，然后输入到多个 Inception 单元中捕捉时空特性，最后将这些单元的输出累加实现特征整合；第 2 部分是 DeCell 模块，视频数据先通过 Inception 单元捕捉时空特性，然后通过 2D 反卷积进行上采样恢复图像尺寸；第 3 部分引入 U-net 作为主干网络，利用其带有跳跃连接的编码解码结构，融合不同层级的时空特征，保留图像细节信息，实现图像重建。实验结果表明，本文 CUnet 模型在视频预测任务中表现出色，具有优秀的预测能力。

1 网络模型

1.1 CUnet 模型结构

本文 CUnet 模型由 Cell 模块、DeCell 模块和 U-net 网络 3 个模块组成，如图 1 所示。CUnet 模型引入 U-net 作为主干网络，将 Cell 模块作为 U-net 网络的编码器部分，将 Cell 模块的反模块 DeCell 作为 U-net 网络的解码器部分，并采用跳跃连接将编码器中的浅层特征与解码器中的深层特征相结合，以捕捉多尺度上下文信息。

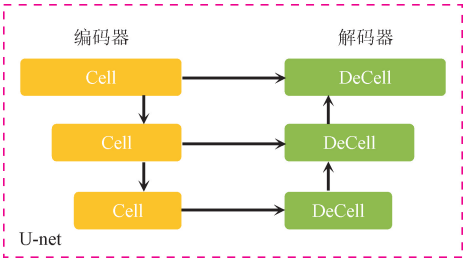


图 1 CUnet 模型结构

Fig. 1 Structure of CUnet model

1.2 Cell 模块

Cell 模块由 2D 卷积模块和 Inception 单元组成，Inception 单元结构沿用 SimVP 模型[17]的结构，Inception 单元与 GoogLeNet 网络模型中 Inception 单元类似。如图 2 所示，Inception 单元首先通过 1 个卷积层 Conv 进行初步特征提取，随后分别输入至 1 个 1×1 组卷积(GroupConv2D)、1 个 3×3 组卷积和 1 个 5×5 组卷积进行特征提取，最后将这些特征叠加输出。LSTM 是一种特殊的循环神经网络，设计用于解决长距离依赖问题，它通过引入 3 个门控机制来控制信息流动和更新记忆单元状态。Inception 单元与 LSTM 网络在处理时序信息方面表现出显著差异，LSTM 作为经典的时序模型，依靠门控机制来控制信息流动，能保证时序信息的连续性。然而，LSTM 模型结构复杂、计算量大、训练效率相对较低。相较而言，Inception 单元将时序维度映射至通道数维度，进而利用不同尺寸的卷积核提取时空特性，结构简洁、训练效率高，并且有效地扩展了网络的宽度，能够捕获不同尺度的特征。

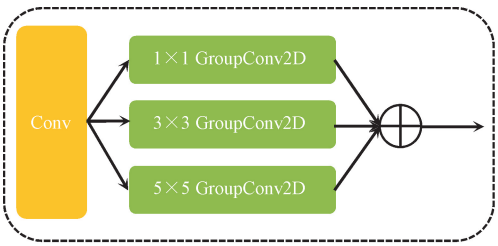


图 2 Inception 单元结构

Fig. 2 Structure of Inception

Cell 模块结构如图 3 所示，输入视频序列 X_0 的特征图大小为 $(B\times T, C_{\text{in}}, H, W)$，其中 B 为批量数、T 为序列个数、C_{in} 为输入通道数、H 为图像高度、W 为图像宽度。X_0 先通过 4 个 3×3 的 2D 卷积模块得到特征图大小为 $(B\times T, C_{\text{out}}, H/4, W/4)$ 的 X_4，其中 C_{out} 为输出通道数，接着将 X_4 的特征图大小变形为 $(B, T\times C_{\text{out}}, H/4, W/4)$，然后将 X_4 分别

输入至3 组Inception 单元块得到特征图大小为(B, $T\times C_{out}$, $H/4$, $W/4$) 的 X_5、X_6 和 X_7，接着将 $X4$、X_5、X_6 和 X_7 累加得到 X_8，最后将 X_8 的特征图大小变形为($B\times T$, C_{out}, $H/4$, $W/4$)。具体公式如下

$$\begin{cases} X_i = \sigma(N(\text{Conv2D}(X_{i-1}))), & 1\leqslant i\leqslant 4 \\ X_5 = \text{Inception}(\text{Inception}(X_4)) \\ X_6 = \text{Inception}(\text{Inception}(X_4)) \\ X_7 = \text{Inception}(\text{Inception}(X_4)) \\ X_8 = X_4 + X_5 + X_6 + X_7 \end{cases} \tag{1}$$

式中：N 为组规范化函数；σ 为 leakyReLU 激活函数；Conv2D 为 3×3 的 2D 卷积模块和 Inception 沿用 SimVP 的 Inception 单元。

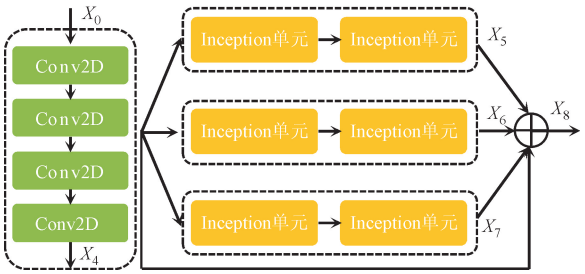


图 3　Cell 模块结构

Fig. 3　Cell module structure

1.3　DeCell 模块

DeCell 模块结构如图 4 所示，分为两个模块：2D 反卷积模块和 Inception 单元。将 Cell 模块输出的 X_8 的特征图大小变形为(B, $T\times C_{out}$, $H/4$, $W/4$)，然后通过 3 个 Inception 单元得到 X_{11}，再将 X_{11} 的特征图大小变形为($B\times T$, C_{out}, $H/4$, $W/4$)，最后通过 4 个 3×3 的 2D 反卷积模块得到特征图大小为($B\times T$, C_{out}, H, W) 的 X_{15}，具体公式如下

$$\begin{cases} X_i = \text{Inception}(X_{i-1}), & 9\leqslant i\leqslant 11 \\ X_i = \sigma(N(\text{DeConv2D}(X_{i-1}))), & 12\leqslant i\leqslant 15 \end{cases} \tag{2}$$

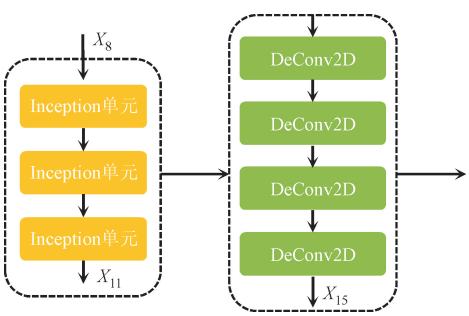


图 4　DeCell 模块结构

Fig. 4　DeCell module structure

式中：N 为组规范化函数；σ 为 leakyReLU 激活函数；DeConv2D 为 3×3 的反卷积模块。

2　实　验

2.1　实验设置

2.1.1　实验数据

选用视频预测任务中广泛认可的公共基准数据集 MovingMNIST、TaxiBJ 以及 Human3.6M 进行性能评估，数据集的具体信息见表 1。

表 1　各数据集的实验参数设置

Table 1　Experiment parameter settings for each dataset

数据集	训练集大小	测试集大小	特征图	T_{in}	T_{pre}	训练轮数
Moving-MNIST	10 000	10 000	(64,64,1)	10	10	2 000
TaxiBJ	19 627	1 334	(32,32,2)	4	4	80
Human-3.6M	73 404	8 582	(128,128,3)	4	4	100

注：T_{in} 为输入帧数量；T_{pre} 为预测帧数量。

(1) MovingMNIST 数据集[29]由一系列数字图像组成，每个图像有 2 个手写数字。这些数字图像会按照特定的轨迹进行移动，通过为每个数字分配不同的初始位置和速度，可以得到无限数量的序列。在本次实验中，利用过去的 10 帧视频数据来预测未来的 10 帧视频序列。

(2) TaxiBJ 数据集[29]是一个源自北京市出租车 GPS 轨迹的交通数据集。该数据集包含两个通道，它们分别表示在给定时间间隔内从其他区域进入某个区域的总交通量和离开某个区域前往其他区域的总交通量。在本实验中，利用过去 4 帧的数据来预测未来的 4 帧交通流量情况。

(3) Human3.6M 数据集[29]是一个用于 3D 人体姿态估计研究的大型公开数据集，它涵盖了 11 位表演者在多种场景和动作下的姿态数据。依照文献[13]的方法，针对行走场景选择了特定的数据子集进行实验。具体来说，使用子集{S_1, $S_5\sim S_8$}进行模型的训练，该子集包含了 2 624 个视频片段；而子集{S_9, S_{11}}则作为测试集，包含了 1 135 个视频片段。在本实验中，每个视频片段由 8 帧组成，通过观察前 4 帧来预测后续的 4 帧内容。

2.1.2　实验性能指标

为充分评估模型性能，采用均方误差、平均绝对误差以及结构相似性指数这 3 个指标来综合评价模

型在不同数据集上的性能表现。其中，均方误差是每个样本的预测值与真实值间像素级误差的平方的平均值，数值越小，说明模型的预测精度越高，一般可用于描述模型的精度，均方误差计算过程如下

$$\sigma=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(x(i,j)-y(i,j))^2 \tag{3}$$

式中：H 和 W 表示图像的高度和宽度；$x(i,j)$ 和 $y(i,j)$ 表示真实图像和预测图像在位置 (i,j) 处的像素值。

平均绝对误差 ρ 是每个样本的预测值与真实值间像素级误差的平均绝对值，数值越小，表示模型的预测能力越强，平均绝对误差计算过程如下

$$\rho=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}|x(i,j)-y(i,j)| \tag{4}$$

结构相似性指数 τ 的取值范围在 0 到 1 之间，数值越高，表明图像的质量越好，能够保留更多的结构信息。设真实图像为 x，预测图像为 y，则真实图像和预测图像的结构相似性的计算过程如下

$$\tau(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{5}$$

$$C_1=(k_1+L)^2 \tag{6}$$

$$C_2=(k_2+L)^2 \tag{7}$$

式中：μ_x 为真实图像的像素平均值；μ_y 为预测图像的像素平均值；σ_x^2 为真实图像的像素方差；σ_y^2 为预测图像的像素方差；σ_{xy} 为真实图像 x 和预测图像 y 的像素协方差；C_1 和 C_2 是常数，用于避免分母为 0；k_1 和 k_2 是较小的常数；L 为图像像素值的最大范围。

2.1.3 实现细节

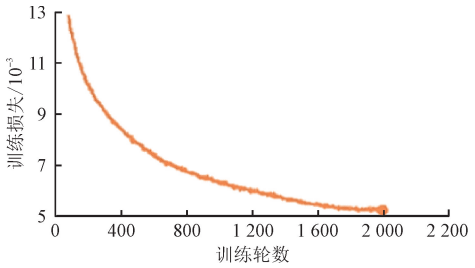
采用 PyTorch 框架实现提出的方法，并在 NVIDIA GeForce RTX 3080 Ti GPU 上进行了相应的实验。实验中，将批处理大小设定为 16，选用 Adam 作为优化算法，采用均方误差作为误差函数，并设置学习率为 0.001 以进行模型的训练。

2.2 实验结果与分析

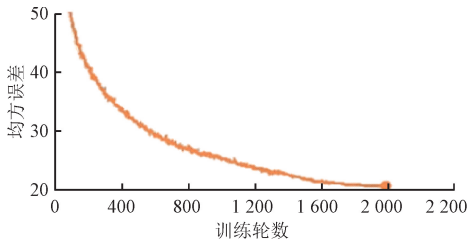
将本文提出的 CUnet 模型与视频预测领域的诸多主流模型进行对比，这些模型包括 ConvLSTM 模型[9]、PredRNN 模型[10]、PredRNN++模型[11]、MIM 模型[12]、E3D-LSTM 模型[13]、CrevNet 模型[14]、PhyDNet 模型[15]、SimVP 模型[17]、FFINet 模型[18]和 TAU 模型[19]。

2.2.1 MovingMNIST 数据集

本文 CUnet 模型在 MovingMNIST 数据集上训练时使用了 3 个 Cell 模块和 3 个 DeCell 模块，训练结果如图 5 所示。遵循文献[19]，模型总计进行了2 000轮训练。图 5(a)展示了模型在训练集上的损失(训练损失)随训练轮数增加而逐渐降低的过程；图 5(b)展示了模型在验证集上的均方误差随训练轮数增加而稳步下降的过程。



(a)CUnet 模型在训练集上的损失



(b)CUnet 模型在验证集上的均方误差

图 5 CUnet 模型在 MovingMNIST 数据集上的训练结果

Fig. 5 Training result of the CUnet model on the MovingMNIST dataset

对当前视频预测领域的多种主流模型在测试集上的实验结果进行了比较，结果如表 2 所示。

表 2 各模型在 MovingMNIST 数据集上的性能比较

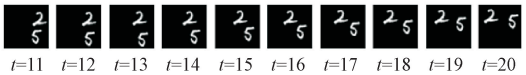
Table 2 Performance comparison of each model on MovingMNIST dataset

模型	均方误差	平均绝对误差	结构相似性指数
ConvLSTM[9]	103.3	182.9	0.707
PredRNN[10]	56.8	126.1	0.867
PredRNN++[11]	46.5	106.8	0.898
MIM[12]	44.2	101.1	0.910
E3D-LSTM[13]	41.3	86.4	0.910
PhyDNet[15]	24.4	70.3	0.947
CrevNet[14]	22.3		0.949
SimVP[17]	23.8	69.9	0.948
TAU[19]	19.8	60.3	0.957
FFINet[18]	19.2	60.4	0.958
CUnet	20.1	61.9	0.956

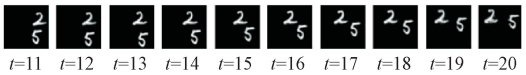
从表 2 的数据可以看出：在 MovingMNIST 数据集上，CUnet 模型在预测任务中展现了较低的均方误差和平均绝对误差；与纯卷积网络模型 SimVP 相比，CUnet 模型的均方误差和平均绝对误差分别降低了 15.55%和 11.44%；与最新的纯卷积网络模型 TAU 和 FFINet 相比，CUnet 模型的均方误差和平均绝对误差略有增加；在预测任务的结构相似性指数方面，与 SimVP 模型相比，其结构相似性指数提高了 0.84%；相较于最新的纯卷积网络模型 TAU 和 FFINet，CUnet 模型的结构相似性指数略显逊色。

图 6 展示了 CUnet 模型和 SimVP 模型在 MovingMNIST 数据集上的预测结果可视化。图 6(a)表示输入的 10 帧视频数据，图 6(b)表示真实的未来 10 帧视频数据，图 6(c)表示 CUnet 模型预测的 10 帧视频数据，图 6(d)表示 SimVP 模型预测的 10 帧视频数据。将真实目标视频数据与 CUnet 模型、SimVP 模型的预测视频数据相对比，可以观察到 CUnet 模型在图像细节保留方面有着明显的优势。尤其在数字“2”的呈现上，CUnet 模型准确地捕捉到了真实目标视频中数字“2”略微向上的折角特征，而 SimVP 模型则未能完全体现这一点。

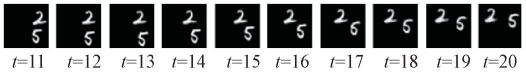
(a)输入数据



(b)真实数据



(c)CUnet 预测结果



(d)SimVP 预测结果

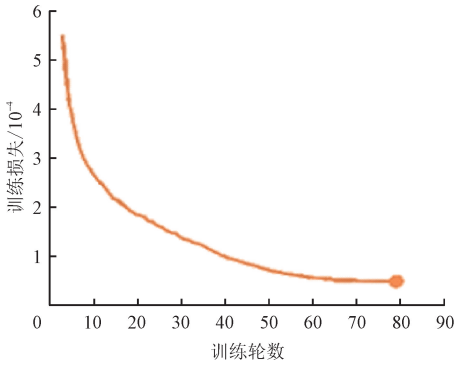
图 6　模型在 MovingMINST 数据集上的预测结果可视化

Fig. 6　Visualization of prediction results of the model on the MovingMNIST dataset

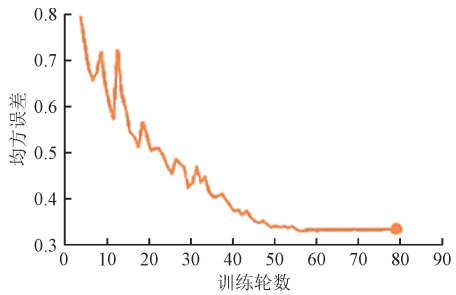
选取 MovingMNIST 数据集的视频序列，采用前 5 帧作为输入来预测后续的 15 帧，对 CUnet 模型和 SimVP 模型进行了 100 轮训练。在 MovingMNIST 测试集上，CUnet 模型的均方误差为 60.45，而 SimVP 模型的均方误差为 65.34。实验结果表明，在长时序预测任务中，CUnet 模型相较于 SimVP 模型展现出更高的预测精度。

2.2.2　TaxiBJ 数据集

CUnet 模型在 TaxiBJ 数据集上训练时使用了 2 个 Cell 模块和 2 个 DeCell 模块，训练结果如图 7 所示，遵循文献[19]，模型总计进行了 80 轮训练。图 7(a)描绘了模型在训练集上的损失随训练轮数增加而逐渐降低的过程；图 7(b)则反映了模型在验证集上的均方误差随训练轮数增加而稳步下降的现象。



(a)CUnet 模型在训练集上的损失



(b)CUnet 模型在验证集上的均方误差

图 7　CUnet 模型在 TaxiBJ 数据集上的训练结果

Fig. 7　Training result of the CUnet model on the TaxiBJ dataset

对比当前视频预测领域的多种主流模型在 TaxiBJ 测试集上的表现，结果如表 3 所示。

表 3　各模型在 TaxiBJ 数据集上的性能比较

Table 3　Performance comparison of each model on TaxiBJ dataset

模型	均方误差	平均绝对误差	结构相似性指数
ConvLSTM[9]	0.485	17.7	0.978
PredRNN[10]	0.464	17.1	0.971
PredRNN++[11]	0.442	101.1	0.910
MIM[12]	0.429	16.6	0.971
E3D-LSTM[13]	0.432	16.9	0.979
PhyDNet[15]	0.419	16.2	0.982
SimVP[17]	0.414	16.2	0.982
TAU[19]	0.344	15.6	0.983
FFINet[18]	0.412	16.2	0.982
CUnet	0.326	15.3	0.984

由表 3 可知，在 TaxiBJ 数据集上，无论是均方误差、平均绝对误差还是结构相似性指数，CUnet 模型均表现最优。与目前性能表现最好的纯卷积网络模型 TAU 相比，CUnet 模型的均方误差和平均绝对误差分别降低了 5.23%和 1.92%。同时，在预测任务的结构相似性指数表现上，CUnet 模型相较于 TAU 模型也有所提升，结构相似性指数提高了 0.1%。

图 8 展示了 CUnet 模型和 SimVP 模型在 TaxiBJ 数据集上的预测结果可视化。图 8(a)表示输入的 4 帧视频数据，图 8(b)表示真实的未来 4 帧视频数据，图 8(c)表示 CUnet 模型预测的 4 帧视频数据，图 8(d)表示 SimVP 模型预测的 4 帧视频数据。将真实目标视频数据与 CUnet 模型、SimVP 模型的预测视频数据相对比，可以看出 CUnet 模型的预测视频序列不仅图像质量更为清晰，而且在细节保留方面也优于 SimVP 模型。

(a)输入数据

(b)真实数据

(c)CUnet 预测结果

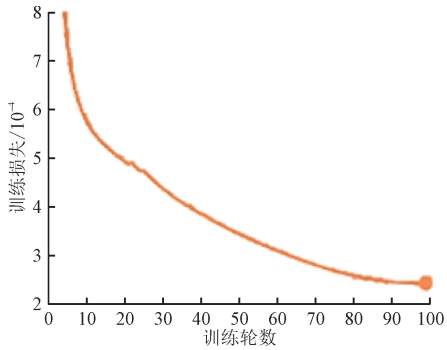
(d)SimVP 预测结果

图 8 模型在 TaxiBJ 数据集上的预测结果可视化

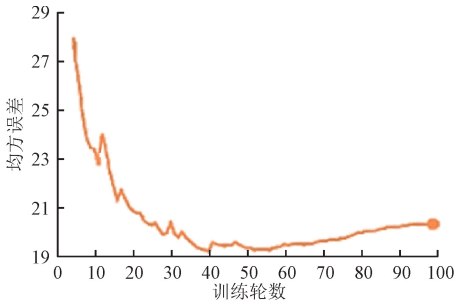
Fig. 8 Visualization of prediction results of the model on the TaxiBJ dataset

2.2.3 Human3.6M 数据集

CUnet 模型在 Human3.6M 数据集上训练时使用了 2 个 Cell 模块和 2 个 DeCell 模块，训练过程如图 9 所示。根据文献[17]，CUnet 模型总计进行了 100 轮训练。图 9(a)展示了模型在训练集上损失随训练轮数增加而逐渐降低的过程；图 9(b)则说明了模型在验证集上的均方误差随训练轮数增加而持续下降的情况。



(a)CUnet 模型在训练集上的损失



(b)CUnet 模型在验证集上的均方误差

图 9 CUnet 模型在 Human3.6M 数据集上的训练结果

Fig. 9 Training result of the CUnet model on the Human3.6M dataset

对当前视频预测领域的多种主流模型在测试集上的表现进行比较，结果如表 4 所示。

表 4 各模型在 Human3.6M 数据集上的性能比较

Table 4 Performance comparison of each model on Human3.6M dataset

模型	均方误差	平均绝对误差	结构相似性指数
ConvLSTM[9]	50.4	1 890	0.776
PredRNN[10]	48.4	1 890	0.781
PredRNN++[11]	45.8	1 720	0.851
MIM[12]	42.9	1 780	0.790
E3D-LSTM[13]	46.4	1 660	0.869
PhyDNet[15]	36.9	1 620	0.901
SimVP[17]	31.6	1 510	0.904
FFINet[18]	23.3	1 190	0.913
CUnet	20.3	325	0.987

由表 4 可知，在 Human3.6M 数据集上，无论是均方误差、平均绝对误差还是结构相似性指数，CUnet 模型均表现最佳。与目前表现最好的纯卷

积网络模型 FFINet 相比，CUnet 模型的均方误差和平均绝对误差分别下降了 12.88%和 72.69%。此外，在预测任务的结构相似性指数表现上，CUnet 模型相较于最佳的 FFINet 模型有了显著提升，结构相似性指数增加了 8.1%。图 10 展示了 CUnet 模型在 Human3.6M 数据集上的预测结果可视化。图 10(a)表示输入的 4 帧视频数据，图 10(b)表示真实的未来 4 帧视频数据，图 10(c)表示 CUnet 模型预测的 4 帧视频数据，可以观察到，CUnet 模型的预测视频序列能较大程度地复现真实目标视频序列的动态与关键特征。

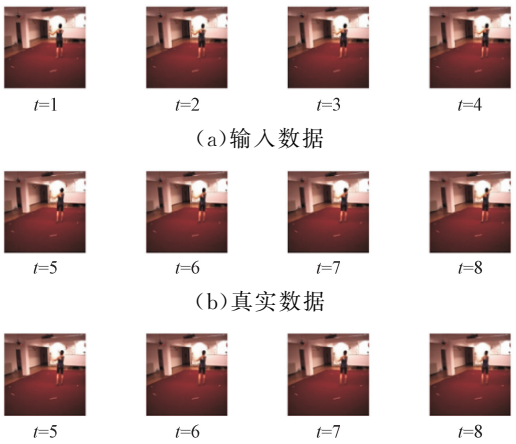


图 10　CUnet 模型在 Human3.6M 数据集上的预测结果可视化

Fig. 10　Visualization of prediction results of the CUnet model on the Human3.6M dataset

SimVP 模型是视频预测领域的一个标志性纯卷积神经网络基础模型，为后续研究提供了坚实基础。CUnet 模型、FFINet 模型和 TAU 模型均是在 SimVP 模型基础上进行改进的。通过 CUnet 模型在 MovingMNIST、TaxiBJ 以及 Human3.6M 数据集上实验表现看，相较于 SimVP 模型，CUnet 模型在均方误差、平均绝对误差以及结构相似性指数等关键指标上实现了显著的提升，这充分证明了 CUnet 模型在捕捉时空特征和保留图像细节方面的有效性。尽管在合成数据集 MovingMNIST 上，CUnet 模型的性能略低于表现最优异的纯卷积神经网络模型 TAU 和 FFINet，但在均方误差、平均绝对误差和结构相似性指数指标上仍表现出较高的竞争力。而在真实世界数据集 TaxiBJ 和 Human3.6M 的评测中，CUnet 模型则超越了其他所有竞争模型，展现出优秀的预测能力。这表明，TAU 模型和 FFINet 模型可能在诸如 Moving MNIST 这样的合成数据集上展现出较高的预测精度，而 CUnet 模型在处理 TaxiBJ 和 Human3.6M 这类复杂真实世界视频数据时展现出更强的鲁棒性和适应性。

3　消融实验

为验证 CUnet 模型的有效性，针对 Cell 和 DeCell 中的 Inception 的单元数以及 U-net 网络的层数做了消融实验，使用的数据集是 TaxiBJ，具体的实验结果如表 5 所示。

表 5　消融实验判断矩阵及模型总均方误差

Table 5　The judgment matrix of the ablation experiment and the total mean square error of the model

模型	Cell 模块				DeCell 模块				U-net 模块		均方误差
	Inc-3	Inc-6	Inc-并行	Inc-串行	Inc-3	Inc-6	Inc-并行	Inc-串行	Unet-1	Unet-2	
M1	有	无	有	无	有	无	有	无	有	无	36.7
M2	有	无	有	无	有	无	有	无	无	有	35.7
M3	有	无	有	无	有	无	无	有	无	有	35.1
M4	有	无	有	无	无	有	有	无	无	有	35.6
M5	有	无	有	无	无	有	无	有	无	有	34.2
M6	无	有	有	无	有	无	无	有	有	无	34.6
M7	无	有	有	无	有	无	有	无	无	有	35.1
M8	无	有	有	无	无	有	有	无	无	有	35.2
M9	无	有	无	有	无	有	有	无	无	有	33.8
M10	无	有	无	有	无	有	无	有	有	无	35.9
M11	无	有	无	有	无	有	无	有	无	有	34.6
M12	无	有	有	无	有	无	无	有	有	无	34.8
CUnet	无	有	有	无	有	无	无	有	无	有	32.6

注：Inc-3 表示有 3 个 Inception 单元；Inc-6 表示有 6 个 Inception 单元；Inc-并行表示 Inception 并行连接；Inc-串行表示 Inception 串行连接；Unet-1 表示 U-net 网络层数为 1 层；Unet-2 表示 U-net网络层数为 2 层；均方误差表示模型总均方误差；M1～M12 为本文构建的不同参数的消融实验模型。

由表 5 可知,随着 U-net 网络层数的增加,模型的均方误差表现有所提升,以 M12 模型和 CUnet 模型为例,M12 模型采用 1 层结构,CUnet 模型采用 2 层结构,在性能对比中,CUnet 模型展现出更低的均方误差。当 Cell 和 DeCell 模块采用了不对称的 Inception 连接方式时,模型的均方误差性能得到了进一步提升,如表 5 所示,对比 M11 模型和 CUnet 模型,M11 模型在 Cell 模块和 DeCell 模块中均以串行方式连接 Inception 单元,而 CUnet 模型在 Cell 模块中采用并行连接,在 DeCell 模块中采用串行连接。通过这种不对称的连接设计,CUnet 模型在均方误差表现上表现更为出色,其原因可能是不对称连接有助于模型在训练过程中捕获多样性特征,增强了模型的表达能力。

4 结 论

本文提出一种纯卷积视频预测模型——CUnet 模型,通过设计了一对不对称的 Cell 和 DeCell 模块,采用 U-net 作为主干网络,在 U-net 的编码端使用 Cell 模块捕捉视频数据的时空特性,在解码端利用 DeCell 模块进行上采样来实现图像尺寸恢复。通过这种模块化设计,CUnet 模型能够有效地融合不同层次的时空特征,进而实现对图像的高精度重建。采用本文 CUnet 模型在多个基准数据集上进行了广泛的实验,实验结果充分验证了该模型的有效性。CUnet 模型拓展了纯卷积神经网络模型在视频预测领域的应用范围,展示了纯卷积神经网络模型在时序预测任务中的潜力。

参考文献:

[1] 李善梅,宋思霓,王红勇,等. 基于多模态时空特征融合的终端区交通拥堵精细化预测[J/OL]. 北京航空航天大学学报. (2024-09-05)[2024-09-18]. https://doi.org/10.13700/j.bh.1001-5965.2024.0557.
LI Shanmei, SONG Sini, WANG Hongyong, et al. Refined prediction of terminal area traffic congestion based on multimodal spatiotemporal feature fusion [J/OL]. Journal of Beijing University of Aeronautics and Astronautics. (2024-09-05)[2024-09-18]. https://doi.org/10.13700/j.bh.1001-5965.2024.0557.

[2] NING Shuliang, LAN Mengcheng, LI Yanran, et al. MIMO is all you need: a strong multi-in-multi-out baseline for video prediction [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2023: 1975-1983.

[3] 吴宇轩,虞慧群,范贵生. 基于误差补偿的多模态协同交通流预测模型[J]. 电子学报, 2024, 52(8): 2878-2890.
WU Yuxuan, YU Huiqun, FAN Guisheng. Multimodal cooperative traffic flow prediction model based on error compensation [J]. Acta Electronica Sinica, 2024, 52(8): 2878-2890.

[4] CHENG Jinguo, LI Ke, LIANG Yuxuan, et al. Rethinking urban mobility prediction: a super-multivariate time series forecasting approach [EB/OL]. (2023-12-04)[2024-08-10]. https://arxiv.org/abs/2312.01699.

[5] 晏婕. 基于深度学习的视频帧序列预测算法研究[D]. 长春: 吉林大学, 2023.

[6] TANG Yujin, DONG Peijie, TANG Zhenheng, et al. VMRNN: Integrating vision mamba and LSTM for efficient and accurate spatiotemporal forecasting [EB/OL]. (2024-06-29)[2024-08-12]. https://arxiv.org/abs/2403.16536.

[7] 孔玮,刘云,李辉,等. 基于深度学习的行人轨迹预测方法综述[J]. 控制与决策, 2021, 36(12): 2841-2850.
KONG Wei, LIU Yun, LI Hui, et al. Survey of pedestrian trajectory prediction methods based on deep learning [J]. Control and Decision, 2021, 36(12): 2841-2850.

[8] 潘敏婷,王韫博,朱祥明,等. 基于无标签视频数据的深度预测学习方法综述[J]. 电子学报, 2022, 50(4): 869-886.
PAN Minting, WANG Yunbo, ZHU Xiangming, et al. A survey on deep predictive learning based on unlabeled videos [J]. Acta Electronica Sinica, 2022, 50(4): 869-886.

[9] SHI Xingjian, CHEN Zhourong, WANG Hao, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2015: 802-810.

[10] WANG Yunbo, LONG Mingsheng, WANG Jianmin, et al. PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 879-888.

[11] WANG Yunbo, GAO Zhifeng, LONG Mingsheng, et al. PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning [C]//Proceedings of the 35th International Conference

on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2018: 5123-5132.

[12] WANG Yunbo, ZHANG Ananjin, ZHU Hongyu, et al. Memory in memory: a predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 9146-9154.

[13] WANG Yunbo, JIANG Lu, YANG M H, et al. Eidetic 3D LSTM: a model for video prediction and beyond [C]//7th International Conference on Learning Representations. New York, USA: ICLR, 2019: 1-14.

[14] YU Wei, LU Yichao, EASTERBROOK S, et al. Efficient and information-preserving future frame prediction and beyond [C]//International Conference on Learning Representations. New York, USA: ICLR, 2020: 1-14.

[15] LE GUEN V, THOME N. Disentangling physical dynamics from unknown factors for unsupervised video prediction [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 11471-11481.

[16] CHANG Zheng, ZHANG Xinfeng, WANG Shanshe, et al. MAU: a motion-aware unit for video prediction and beyond [C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2021: 26950-26962.

[17] GAO Zhangyang, TAN Cheng, WU Lirong, et al. SimVP: simpler yet better video prediction [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 3160-3170.

[18] LI Ping, ZHANG Chenhan, XU Xianghua. Fast Fourier inception networks for occluded video prediction [J]. IEEE Transactions on Multimedia, 2024, 26: 3418-3429.

[19] TAN Cheng, GAO Zhangyang, WU Lirong, et al. Temporal attention unit: towards efficient spatiotemporal predictive learning [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2023: 18770-18782.

[20] WEN Shiping, LIU Weiwei, YANG Yin, et al. Generating realistic videos from keyframes with concatenated GANs [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(8): 2337-2348.

[21] KWON Y H, PARK M G. Predicting future frames using retrospective cycle GAN [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 1811-1820.

[22] YU S, TACK J, MO S, et al. Generating videos with dynamics-aware implicit generative adversarial networks [C]//10th International Conference on Learning Representations. [S. l.]: [s. n.], 2022: 1-14.

[23] OPREA S, MARTINEZ-GONZALEZ P, GARCIA-GARCIA A, et al. A review on deep learning techniques for video prediction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2806-2826.

[24] ZHANG Qianqian, FENG Guorui, WU Hanzhou. Surveillance video anomaly detection via non-local U-Net frame prediction [J]. Multimedia Tools and Applications, 2022, 81(19): 27073-27088.

[25] GAN K Y, CHENG Yutong, TAN H K, et al. Contrastive-regularized U-Net for video anomaly detection [J]. IEEE Access, 2023, 11: 36658-36671.

[26] 张杰，杨雪，龚智龙，等. 双向预测 BiP-GAN 的行人视频异常事件自动检测 [J/OL]. 武汉大学学报(信息科学版). (2025-02-11)[2025-02-18]. https://doi.org/10.13203/j.whugis20240259.
ZHANG Jie, YANG Xue, GONG Zhilong, et al. Bidirectional prediction BiP-GAN pedestrian video anomaly event automatic detection [J/OL]. Geomatics and Information Science of Wuhan University. (2025-02-11)[2025-02-18]. https://doi.org/10.13203/j.whugis20240259.

[27] WANG Zhaohua, LI Zhenyu, PAN Heng, et al. Large-scale measurements and prediction of DC-WAN traffic [J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(5): 1390-1405.

[28] LEI Shanzhong, SONG Junfang, WANG Tengjiao, et al. Attention U-Net based on multi-scale feature extraction and WSDAN data augmentation for video anomaly detection [J]. Multimedia Systems, 2024, 30(3): 118.

[29] TAN Cheng, LI Siyuan, GAO Zhangyang, et al. OpenSTL: a comprehensive benchmark of spatio-temporal predictive learning [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023: 69819-69831.

（编辑　刘杨　陶晴）