结构与纹理分解的多尺度3D解耦卷积视频预测

郑明魁 吴孔贤 邱鑫涛 郑海峰 赵铁松
（福州大学物理与信息工程学院 福州 350108）

摘 要 视频预测旨在利用历史帧预测未来图像帧，是一项逐像素的密集预测任务。目前的非自回归模型采用多帧输入多帧输出的架构，有效避免了误差累积。针对现有方法在对视频数据降维处理时使用跨步卷积进行下采样而导致局部细节丢失的问题，本文采用了特征域结构与纹理分离学习的思路，去除细节后的低频结构信息具有更强的时间相关性，有利于局部区域结构像素时空相关性的预测，而高频细节特征则采用一个独立的增强模块进行学习。在此基础上，本文设计了一种多尺度的3D解耦卷积模块，将3D卷积解耦为2D卷积和1D卷积来专注学习低频结构的空间和时间特性，这种解耦方式在提高对象形态预测性能的同时还减少了模型参数和内存消耗。最后采用一种高频细节小尺度增强模块，用来学习分解后的高频信息并预测图像的纹理，提升视频预测的细节质量。在合成数据以及真实场景数据集上的实验结果表明，本文所设计的算法兼顾了时空一致性和细节表现力，在视频中运动物体的整体结构与局部细节预测方面展现出更高的准确性，其中在Moving MNIST数据集上的MSE为15.7，分别比现有算法如SimVP、TAU、SwinLSTM、VMRNN等降低了34.0%、20.7%、11.3%、4.8%，在其他数据集上的实验结果也表现出一定的优越性。

关键词 视频预测；多帧输入多帧输出；结构与纹理分离；3D解耦卷积
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.01832

Multi-Scale 3D Decoupled Convolutional Video Prediction Method Based on Structure and Texture Decomposition

ZHENG Ming-Kui WU Kong-Xian QIU Xin-Tao ZHENG Hai-Feng ZHAO Tie-Song
（School of Physics and Information Engineering, Fuzhou University, Fuzhou 350108）

Abstract Video prediction is a fundamental task in computer vision, aiming to predict future frames based on a series of historical frames. It is a dense pixel-level prediction task with broad application value in fields such as autonomous driving, traffic flow prediction, and weather forecasting. Traditional video prediction methods typically rely on autoregressive model architectures, which use a cyclical strategy, taking the output of the previous frame as the input for the next frame, and recursively predicting in a loop. However, current models still face unresolved challenges. In particular, many existing approaches perform down sampling through strided convolution when reducing the dimensionality of video data, which inevitably leads to pixel loss and neglect of local details, thereby compromising the clarity of the predicted results. To mitigate this issue, non-autoregressive models have been proposed, featuring a multi-frame input and multi-frame output architecture that generates future frames in parallel, breaking away from the cyclic framework and effectively avoiding the accumulation of prediction errors. However,

收稿日期：2024-10-29；在线发布日期：2025-05-20。本课题得到国家自然科学基金项目（62171134）、福建省科技重大专项专题项目（2022HZ026007）资助。**郑明魁**，博士，副教授，中国计算机学会（CCF）会员，主要研究领域为多模态信息智能编码、计算机视觉与触觉。E-mail：zhengmk@fzu.edu.cn。 **吴孔贤**，硕士研究生，主要研究领域为视频预测、视频编码。**邱鑫涛**，硕士研究生，主要研究领域为视频超分辨率。 **郑海峰**，博士，教授，主要研究领域为智能车联网、6G网络/通感一体化、具身智能感知与决策。**赵铁松**（通信作者），博士，教授，主要研究领域为图像处理与计算机视觉、智能视频编码与通信、触感信息与虚拟现实。

existing models still face pressing issues that need to be addressed. Objects in videos often exhibit irregular motion, and the variability in video content along with multiple possible motion trajectories make it challenging for network models to predict image motion accurately, resulting in blurred image details in the predicted frames. To tackle this challenge, this paper introduces a novel research approach that leverages the characteristics of wavelet transforms through the separated learning of feature domain structure and texture to enhance the quality of video prediction. Under this separated structure, the low-frequency structural information, after detail removal, has stronger temporal correlation, which aids in more accurate spatiotemporal prediction of image regions. High-frequency detail features are learned through an independent enhancement module to improve the local quality of video prediction. Additionally, by using a two-level wavelet transform, down sampling operations can be performed to reduce image resolution without losing pixel information, and corresponding up sampling operations can be achieved through inverse wavelet transform. This symmetrical structure maximizes the retention of image information and allows for more accurate prediction of subsequent images. Furthermore, this paper designs a multi-scale 3D decoupled convolution module that uses convolutional kernels of different sizes to learn regional features at various scales. This module decouples traditional 3D convolution into 2D and 1D convolutions. This decoupling method focuses on learning the spatial and temporal characteristics of low-frequency structures, which not only improves predictive performance but also reduces the model′s parameters and memory consumption. This design enables the model to more effectively capture both short-term and long-term temporal dependencies, thereby enhancing the accuracy and coherence of video prediction. Finally, a high-frequency detail enhancement module on a small scale is designed to learn the decomposed high-frequency information and predict image details and textures, enhancing the local quality of video prediction. The experimental results on synthetic data and real-world datasets show that the algorithm designed in this paper has higher prediction accuracy than existing algorithms. It has more accurate prediction performance in local details and overall prediction morphology. Among them, the MSE on the Moving MNIST dataset is 15.7, which is 34%, 20.7%, 11.3%, and 4.8% lower than the existing advanced algorithms SimVP, TAU, SwinLSTM, and VMRNN respectively.

Keywords video prediction; multi-frame input and multi-frame output; structure and texture separation; decoupled 3D convolution

1 引 言

视频预测在给定若干连续帧的条件下预测未来图像帧，不需要人工标记[1]。其因在自动驾驶[2]、交通流量预测[3]、天气预测[4]、异常检测[5-6]等领域有广泛的应用价值而受到越来越多的关注。视频由不同帧图像构成，具备一定的时间相关性和空间相关性，视频预测需要在空间域准确地描述视觉内容，并在时间域合理地对运动做出预测。然而视频中复杂多样的背景和具有多种可能的运动轨迹使得视频预测成为一项极具挑战性的任务。

2015年Srivastava[7]等人首次将LSTM(Long Short-Term Memory)应用于视频预测任务以来，视频预测技术在近些年来不断蓬勃发展。早期的方法主要通过混合卷积神经网络和循环神经网络来分别学习视频的空间相关性和时间依赖性。ConvLSTM[4](Convolutional LSTM network)方法将LSTM的内部计算方式由全连接改为卷积计算，使其具有用于捕获时空相关性的结构。该方法是自回归模型的代表，自回归模型基于循环的策略，使用上一帧的输出作为下一帧的输入。之后的很多研究成果[8-15]都在此基础上做出改进，例如通过改变ConvLSTM的内部结构或是堆叠之后的输出连接

走向来更新视频预测模型。事实证明，循环神经网络的结构对于学习时间序列有独特的优势，但是这种根据当前帧预测下一帧的自回归模型训练起来速度较慢，并且随着预测的帧数变长，前期预测帧中的小错误容易堆叠和累积，导致后期预测出现偏移，存在一定的局限性。

相比于容易造成误差累积的自回归结构，以多帧输入来预测多帧输出的非自回归架构模型，可以有效解决这种局限性。该模型架构将数据一次性输入编码器并从解码器输出若干帧。与单帧输入单帧输出的循环结构相比，该方法不依赖于前一帧的预测结果，误差不会积累。如基于CNN构建的SimVP[16]模型（Simpler yet Better Video Prediction），没有依靠循环神经网络来进行推理，便实现了比PredRNN[8]、E3D-LSTM[10]、MIM[12]这些依赖循环架构的模型都要好的效果。后续推出的改进算法SimVP. v2[17]和TAU[18]在模型中间翻译器部分采用新的卷积模块建模时间特征，更简单高效。此外，Ning等人[19]也充分利用了多输入多输出的非自回归结构，结合Transformer设计一种MIMO-VP模型，取得了较好的性能。

鉴于扩散模型（Diffusion Models）在图像领域的成功，Ho等人[20]通过将图像扩散模型中的2D U-Net[21]更改为3D U-Net[22]，初步引入了视频扩散模型的概念。在此基础上，Voleti等人[23]设计了一个掩蔽条件视频扩散（MCVD）通用框架，能够灵活处理视频预测、无条件生成和插值等多种任务。Höppe等人[24]也通过引入3D卷积，将图像扩散模型扩展到视频领域，提出了随机掩码视频扩散（RaMViD）模型，用于实现视频预测。这些基于扩散模型的视频预测方法为该领域的研究提供了新的思路和启发。然而，值得注意的是，这类生成式模型在追求视频的连贯性和真实性方面表现出色，但在预测精度上还有待进一步提高。

现有视频预测框架首先对视频数据进行降维处理，将数据映射到一个低维潜在空间，再对视频数据进行时空相关性学习，最后通过解码过程将视频数据还原回原始空间大小。然而在对视频数据进行降维处理的过程中，现有算法仅利用不同步长设置的卷积操作来实现下采样，如图1所示，这样会导致图像像素的损失，从而影响到模型对图像细节的预测能力。因此，模型框架设计应充分考虑视频图像的整体信息与细节部分的平衡，确保既能保持整体结构特征，又能有效捕捉局部细节。视频图像主要包含低频结构与高频细节信息，网络模型对局部区域低频结构预测偏差是造成视频中对象形态变化的主要原因，而高频细节变化则可能造成预测图像模糊。对此本文采用特征域结构与纹理分解的方法，使用离散小波变换将特征信息的低频结构与高频细节信息分离学习，去除细节后的低频结构信息具有更强的时间相关性与运动一致性，有利于区域像素的时空相关性预测，而高频细节特征则采用一个独立的增强模块进行学习。这种结构与细节的分离学习，能够使得模型既关注到物体的运动，也注重局部细节的信息，有效预测视频中运动对象形态变化的规律，并增强局部细节。

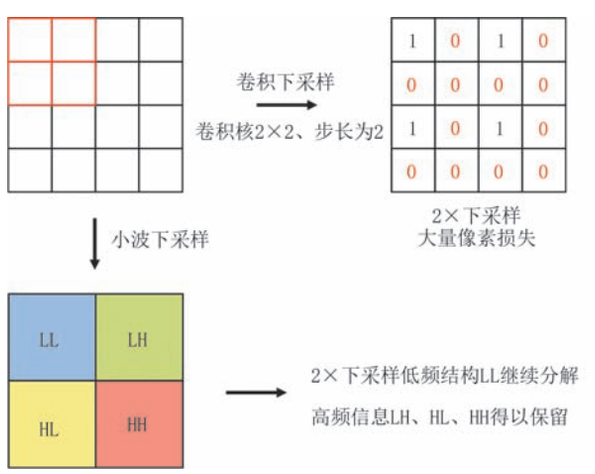

图1 卷积和小波下采样2倍后的对比示意图（原始特征图为4×4，卷积下采样后保留的像素标记为1，损失的像素标记为0）

视频和图像的本质区别就在于时间维度的扩展，因此模型对时间依赖性的捕捉能力便显得尤为重要。非自回归架构在一定程度上解决了循环中误差积累的问题，但是同时也弱化了获取帧与帧之间强相关性的能力。如何捕获视频帧中的时空相关性也是值得思考的问题，3D卷积能够同时考虑空间和时间维度的信息，从而更好地捕获数据的时空特征和时间相关性，减少信息的丢失，对视频数据有较好的处理优势。但是利用3D卷积的堆叠来建模时间特征会导致昂贵的计算成本和内存需求。为了降低计算次数，本文将3D卷积解耦成一个空间域上的卷积串联一个时间域上的卷积。空间卷积可以捕捉空间上的相关性，而时间卷积则捕捉时间上的相关性，比如运动信息。这种解耦方式有助于网络更专注于学习某一特定类型的模式，并且能够大大减小模型的参数量和计算量。

本文的主要贡献如下：

(1)设计了一种结构与纹理特征分离的学习架构，将特征划分为低频结构和高频细节并分别学习。该分离机制有助于模型更有效地捕捉运动趋势与细节变化。

(2)设计了一种多尺度的3D解耦卷积模块，对特征分解后的低频结构信息时空相关性进行建模。解耦后的多尺度3D卷积网络既提升了时空特征提取能力，增强了结构预测性能，与标准的3D卷积相比还减少了模型的参数和内存消耗。

(3)设计了一种高频细节小尺度增强模块，用来学习分离出的高频信息并预测图像的细节和纹理信息，提升视频预测的局部细节质量。

2　相关工作

2.1　视频预测

视频预测即根据过去的图像帧预测未来帧的过程。近年来，视频预测领域蓬勃发展，取得了一定成就，目前主流的方法分为两大类，自回归模型和非自回归模型。分类的依据是模型是否使用上一帧的预测的结果作为预测下一帧的输入。

自回归模型遵循循环的策略来生成下一帧，单帧输入单帧输出，循环直至输出所有预测的帧。ConvLSTM[4]是视频预测领域中自回归模型的代表，后来的许多工作都是在其之上的变体。PredNet[9]基于ConvLSTM[4]做预测，再将其与真实帧做残差，并将获得的残差编码传递到下一层，以此实现误差的前向传递。PredRNN[8]增加时空记忆单元，提出了一个新的时空LSTM单元，能够垂直和水平地传递记忆状态。PredRNN++[11]提出了一个梯度高速公路单元来缓解梯度消失，并提出了一个Casual-LSTM模块来级联空间和时间记忆。MIM[12]改进了LSTM的忘记门，使模型有了捕捉时空序列中非平稳特征的能力。MAU[25]提出了一个捕捉运动信息的动作感知单元来拓宽时间感受野。SwinLSTM[15]将Swin Transformer融合进简化的LSTM，这是一种用自注意力机制取代ConvLSTM中的卷积结构的扩展。VMRNN[26]则将Vision Mamba与LSTM结合，提出了一种新的循环单元。这种自回归模型有一定的优点，它可以循环输出无限长时间范围的长期预测，并且由于推理是按顺序处理的，内存需求不会随着预测长度的增加而增加。但与此同时也容易造成误差积累的问题。模型的误差在每个时间步传播，导致后期的预测越来越偏离真实值。

自回归模型容易产生误差积累，导致学习到的特征逐渐偏离原本轨迹。非自回归模型采用多帧输入多帧输出的结构，能够一次性预测所有的帧。并且输出的帧与帧之间是独立的，避免了误差积累的问题。SimVP[16]的出现，激发了众多学者对非自回归架构的研究兴趣。其仅以简单的2D卷积网络建模空间和时间特征，并实现了高效的性能。后续同一作者团队又推出SimVP.v2[17]和TAU[18]，在中间翻译器部分采用新的卷积模块建模时间特征，设计出更简单更高效的模型，并实现了更好的性能。Ning等人[19]充分利用了多输入多输出的非自回归结构，提出了一种新的MIMO-VP模型，取得了很好的性能，也再次证实了这种架构的有效性。最近，越来越多非自回归的工作涌现，都取得了不错的性能。例如PastNet[27]和IAM4VP[28]，是最近的两个没有使用循环的模型，具有很好的性能。而FFINet[29]和PLA-SM[30]在设计中间翻译器的同时通过引入额外的训练手段，加强了模型的空间学习能力，提升了预测效果。

除了上述两大类自回归和非自回归的预测模型外，近年来扩散模型作为一种新的生成模型，也逐渐在视频预测领域引起了广泛的关注。Ho等人[20]首次展示了将扩散模型运用到视频任务上的可行性，Voleti等人[23]利用简单的非递归2D卷积架构构建了一个通用的MCVD框架实现视频预测。Ye等人[31]采用神经元随机微分方程预测时间运动信息，并利用预测的运动特征和前一帧的图像扩散模型自回归生成视频帧。Zhang等人[32]提出了一种基于扩散模型的框架，通过沿时间维度外推分布进行视频预测。这些方法有一定的发展前景，但目前在预测精度上还存在局限性。

视频内容的快速变化以及物体运动的多样性带来了较大的预测精度挑战，特别是在动态场景中，物体的多种运动模式和复杂时间依赖性使得传统方法难以有效提升预测精度。上述算法在预测视频对象的整体动作精度以及图像局部细节的清晰度方面还有待提高，对此本文在非自回归预测架构的基础上，通过特征域结构与纹理分离学习的思路进行预测，分解后的低频结构信息具有更强的时间相关性与运动一致性，有利于区域像素的时空相关性预测，而高频细节特征则采用独立的小尺度分组卷积增强模块进行学习，使得所设计的算法既关注到物体的结构

运动信息，也学习了视频图像的局部纹理细节，从而提高视频预测的质量。

2.2 小波变换

近年来，频域信息在深度学习的应用引起了广泛关注。传统卷积神经网络主要在空间域进行特征提取，然而诸多研究表明，频域同样包含大量有助于理解图像和时序数据的结构化信息。例如，在图像预测[33-36]任务中，FreqFusion[34]利用频域技术解决特征融合中的类内不一致和边界位移问题。FADC[35]将频率引入膨胀卷积中，FDConv[36]则将其融入动态卷积中实现了频率多样性权重的构建。

而小波变换作为一种典型的时频分析工具，也在多个视觉任务中展现出优越性。例如，WavResNet[37]提出了用小波变换来降低输入和标签流形的拓扑复杂度，提升了深度网络在图像去噪中的性能。Wave-ViT[38]将小波与Transformer结合，通过对键/值进行无损下采样来降低自注意力的计算成本，提高图像识别的精度。MWCNN[39]采用小波变换来替代传统的池化操作完成特征映射的下采样过程，从而降低其分辨率，完成图像恢复任务。然而，这些工作都是针对单张图像进行操作，难以直接应用于视频预测这种需要处理时序信息的任务。于是本文在现有研究的基础上，将CNN与小波变换结合，分离多个视频帧的高低频信息，并分别学习各自的特征，最后融合学习到的高低频信息进行多个视频帧的预测。这一创新性的方法避免了图像在下采样过程中的信息损失，有效保留了图像的细节信息，为视频预测提供了新思路。

3 方 法

3.1 任务定义

视频预测任务利用过去的视频图像预测生成未来的图像帧。其问题定义如下所示：给定过去时刻的T帧视频序列$\mathcal{X}_T=\{x_i\}_{t-T+1}^{t}$，预测出未来时刻的$T'$帧视频序列$\mathcal{P}_{T'}=\{x_i\}_{t+1}^{t+T'}$。这里的$x_i\in\mathbb{R}^{C\times H\times W}$是通道$C$，高度为$H$，宽度为$W$的图像。在模型训练中，本文把视频序列表示为4维张量$\mathcal{X}_T\in\mathbb{R}^{T\times C\times H\times W}$。模型的本质是在学习一个带有可学习参数的映射函数$\mathcal{F}_\theta:\mathcal{X}_T\mapsto\mathcal{P}_{T'}$，并最小化损失函数的过程。最优参数为$\Theta^*=\arg\min_{\Theta}\mathcal{L}(\mathcal{F}_\Theta(\mathcal{X}_T),\mathcal{Y}_{T'})$，其中$\mathcal{F}_\Theta(\mathcal{X}_T)$为预测值，$\mathcal{Y}_{T'}$为真值，$\mathcal{L}$为损失函数。

3.2 总体框架

由于视频运动的不规则性，特别是网络对局部区域低频结构预测偏差是造成视频对象形态变化的主要原因。本文使用离散小波变换将特征域的低频结构与高频细节信息分解后，低频结构信息在时间和空间上具有更强的相关性，有利于网络模型在总体上估计和推测视频运动变化，而采用一个细节增强模块独立估计高频信息，这种结构与细节分离学习的方法可以提高视频预测的质量。解耦后的低频结构信息在时间和空间上的相关性存在大量冗余，捕捉时空相关性的能力显得尤为重要，这对模型的时空建模能力提出了极高的要求。

本文提出了如图2所示的整体框架，模型遵循空间编码器-时空学习器-空间解码器的整体架构。空间编码器主要关注帧内的信息，通过二维卷积对视频帧的空间信息进行特征提取，以此来捕获每一帧图像的空间相关性，在此基础上引入两级离散小波变换[40]对特征图前后进行两次下采样，得到高低频分离的特征图。将高频信息送入高频细节小尺度增强模块，来增强对图像中细节、纹理的学习。在时空学习器部分，将两次小波变换后的低频结构信息送入低频结构时空学习模块，通过将不同尺度的3D解耦卷积按U-Net的架构堆叠来建模视频的时空依赖性，以此来充分学习视频帧的时间和空间变化。最后，空间解码器将处理过的高频信息与低频信息按通道维度拼接后采用逆小波变换实现两次上采样操作，并通过二维卷积生成对未来帧的预测。

3.3 基于结构和纹理细节的特征域小波分解

3.3.1 离散小波变换原理

小波变换[40]是一种非常有效的信号分析与处理工具，它能够对图像进行多尺度分解，从而在保留图像主要特征的同时，对细节信息进行更精细的分解。离散小波变换和其逆变换常用于信号的多分辨率处理。以下是它们的原理：

(1)离散小波变换

离散小波变换的基本思想是通过小波函数对信号进行分解，将信号的高低频部分分离，从而获得信号在不同尺度下的表示。主要分解过程是：通过一对滤波器(低通滤波器和高通滤波器)对信号进行卷积。低通滤波器提取信号的低频信息，高通滤波器提取信号的高频信息。然后对低频部分继续进行滤波(低通和高通)，形成一个递归分解过程。其过程通过公式表示为

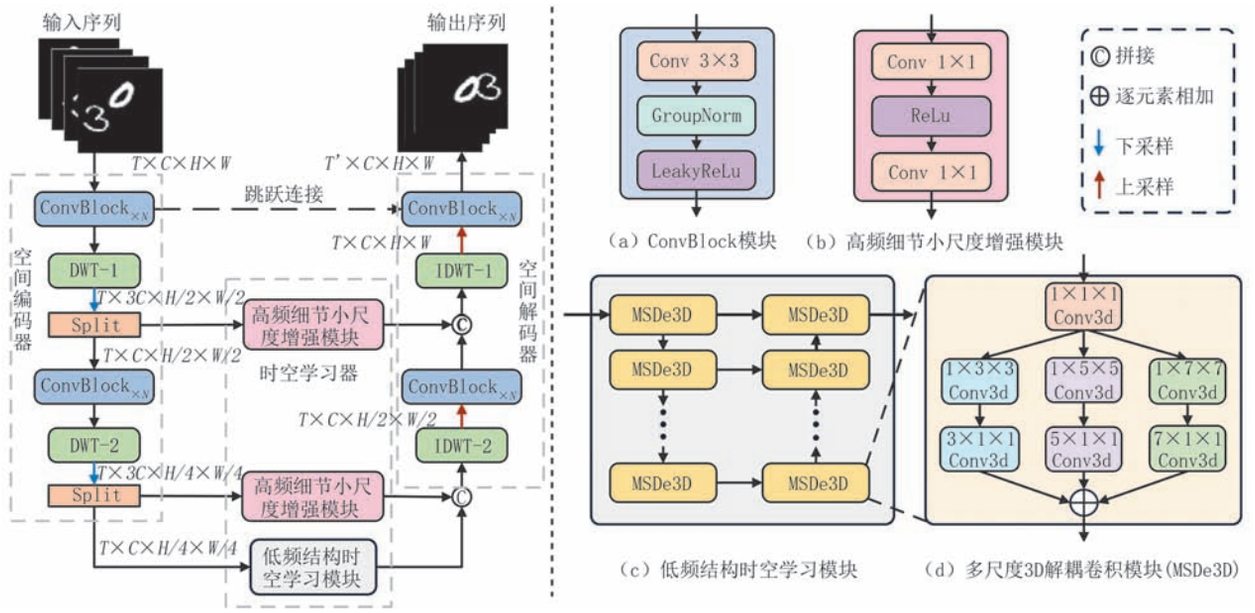

图2　整体结构框图（N为超参数，本文统一设置为2）

$$c_a=\sum_k h_k\cdot x_{2k},\ c_d=\sum_k g_k\cdot x_{2k} \tag{1}$$

其中，h_k和g_k分别是低通和高通滤波器，x是信号，c_a是低频部分，c_d是高频部分。

(2)逆离散小波变换

逆离散小波变换的过程是对小波分解后的系数进行重构，使其恢复原始信号。逆变换的步骤与正变换相反，即将低频部分和高频部分重新构造出原始信号。其重构过程为：逆变换通过对每一个级别的低频和高频部分，使用适当的滤波器进行重建。通过对低频和高频信号进行加权合成，能够将多次分解得到的细节和结构信息逐步合成回原始信号$x(t)$。其过程通过公式表示为

$$x(t)=\sum_k h'_k\cdot c_{a,2k}+\sum_k g'_k\cdot c_{d,2k} \tag{2}$$

其中，k'和g'_k是逆滤波器，$c_{a,2k}$是低频部分，$c_{d,2k}$是高频部分。

3.3.2　模块具体实现

以往的视频预测模型都是采用跨步卷积进行下采样操作来减少特征图的空间维度。此操作在降低分辨率的同时也容易造成图像像素的缺失，导致图像的细节信息不能完整地保留下来。而小波变换可以将信号分解为不同尺度下的低频结构信息和高频细节信息，这种分解过程降低了输入信号的分辨率，但并不会导致信息的丢失。在此基础上，本文将小波变换与卷积相结合，在特征域使用了二级哈尔小波变换，降低特征分辨率的同时解耦图像特征域的高低频信息。通过两次小波分解，可以更精细地提取低频结构特征与高频细节特征，为后续的低频结构信息时空相关性学习以及细节增强提供更丰富的多尺度信息。

图3示意了特征域小波变换的分解过程。从图中可以看出，图像首先经过卷积处理，提取出特征域信息。接着，利用离散小波变换对图像特征域进行分解，将其拆解为低频结构信息和高频纹理细节分量。在第一次小波变换后，对低频分量再进行一次卷积处理，随后进行二次小波变换，得到更加精细的结构和细节信息。通过两级离散小波变换，图像的特征域分辨率实现了四倍下采样。每次小波变换都在卷积处理之后进行。最后，通过应用逆离散小波变换，可以恢复特征图的原始分辨率。

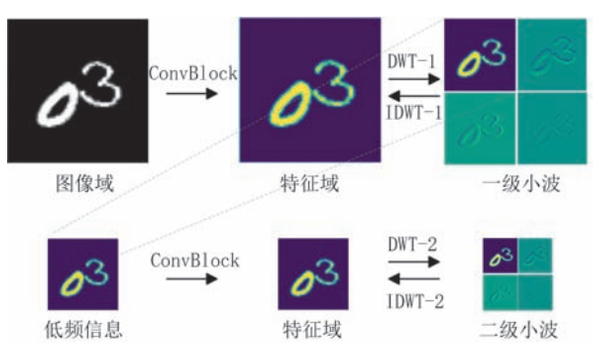

图3　特征域小波分解示意图

将两级离散小波应用至本文的模型中，模块的具体实现如下：首先对经过ConvBlock处理的特征进行一级小波分解，得到具有两倍下采样的高频特征和低频特征。然后将获得的低频特征再次输入

ConvBlock处理并进行二级小波分解，获得具有四倍下采样的高频特征和低频特征。整体过程如式(3)-(5)所示。

$$F_k = ConvBlock_{\times N}(f_k) \tag{3}$$

$$F_{dwt-k} = DWT_k(F_k) \tag{4}$$

$$F_{LL_k}, F_{(LH_k, HL_k, HH_k)} = Split(F_{dwt-k}), k = 1, 2 \tag{5}$$

需要注意的是，当k取1时，式(1)中的f_1为原始数据集的多帧输入，k取2时，f_2为第一次小波变换后分离的低频信息F_{LL_1}。其中F_{dwt-k}为经过小波变换后得到的特征图。F_{LL_k}为频率分离后的低频特征，$F_{(LH_k, HL_k, HH_k)}$为频率分离后的高频特征。

因为小波变换具有可逆性，在解码器部分采用逆小波变换，在此起到上采样的作用，其可逆性能够充分保留高频信息，很好地将分解的图像信息还原，此过程避免了细节的丢失，能够帮助获得更高质量的预测效果。具体实现如式(6)～(9)所示。

$$Z_{idwt-2} = IDWT_2\left(Cat\left(Z_2, Z^{j+1}\right)\right) \tag{6}$$

$$Z_{out-2} = ConvBlock_{\times N}(Z_{idwt-2}) \tag{7}$$

$$Z_{idwt-1} = IDWT_1\left(Cat\left(Z_1, Z_{out-2}\right)\right) \tag{8}$$

$$Z_{out-1} = ConvBlock_{\times N}\left(Cat\left(Z_{idwt-1}, Z_{enc1}\right)\right) \tag{9}$$

式中的Z_1，Z_2分别是第一次和第二次经过学习后的高频特征，Z^{j+1}是经过低频时空学习模块后的特征，Z_{idwt-1}，Z_{idwt-2}分别为第一次与第二次小波变换相对应的逆变换。Z_{enc1}是来自编码器部分的跳跃连接。Cat是按通道维度进行Concat拼接。

3.4 低频结构信息多尺度3D解耦卷积模块

为了从视频数据中学习低频结构时空动态变化，加强学习帧与帧之间的相关性，本文的思想主要是采用3D卷积进行时空依赖性的学习。3D卷积能从时间、高度、宽度三个维度对特征进行学习，对视频数据有天然的处理优势。但以此为代价的是参数量和计算内存的消耗。为了降低计算复杂度而又不损失预测效果，本文设计了多尺度3D解耦卷积模块。如图2(d)所示，使用不同尺度的卷积核来学习局部和全局的特征。3D解耦卷积指的是将原本需要在三维空间内同时进行的卷积操作，通过解耦的方法，分解成两个相对独立的操作步骤(如3×3×3的卷积分解成1×3×3和3×1×1)：先进行2D卷积以处理空间维度，再进行1D卷积以处理时间维度。在此之前特征先经过一个1×1×1的卷积进行处理，此操作是为了方便计算通道数。该模块具体可以被表示为：

$$\hat{z}^j = Conv3d_{1\times1\times1}(z^j) \tag{10}$$

$$z^{j+1} = \sum_{k\in\{3,5,7\}} Conv3d_{k\times1\times1}\left(Conv3d_{1\times k\times k}(\hat{z}^j)\right) \tag{11}$$

式中，z^j为整个3D解耦卷积模块的输入，$\hat{z}^j$为经过1×1×1卷积后的特征图，z^{j+1}为经过整个3D解耦卷积后得到的特征图。

为了比较3D解耦卷积与标准3D卷积的计算成本，本文将这两种类型的卷积所涉及的FLOPs次数(忽略加法计算)进行了对比：

两种方法的计算示意图如图4所示，设输入特征图的尺寸为$X_{in}\in\mathbb{R}^{C_{in}\times T\times H\times W}$，标准3D卷积使用大小为$K_t\times K_h\times K_w$的卷积核在时间、高度、宽度三个方向上移动，生成的特征图尺寸为$X_{out}\in\mathbb{R}^{C_{out}\times T'\times H'\times W'}$，则标准3D卷积的$Flops$次数为

$$FLOPs_{3D} = C_{in}\times K_t\times K_h\times K_w\times C_{out}\times T'\times H'\times W' \tag{12}$$

3D解耦卷积将标准3D卷积分解成空间2D卷积和时间1D卷积，即$1\times K_h\times K_w$和$K_t\times1\times1$，计算量也分为两部分，中间产生的特征图尺寸为$X_{middle}\in\mathbb{R}^{C_{out}\times T'\times H'\times W'}$，最终生成的特征图尺寸为$X_{out}\in\mathbb{R}^{C_{out}\times T'\times H'\times W'}$，则3D解耦卷积的$Flops$次数为

$$FLOPs_{De3D} = C_{in}\times K_h\times K_w\times C_{out}\times T'\times H'\times W' + C_{out}\times K_t\times C_{out}\times T'\times H'\times W' \tag{13}$$

这里和实验设置中一致，前后两次卷积(先2D后1D)后的特征图大小不变，中间和最后生成的特征图大小都为$T'\times H'\times W'$，并且为了方便比较，假设输入通道和输出通道一致，则最终解耦3D与标准3D的FLOPs比例为$\frac{1}{K_t}+\frac{1}{K_h\times K_w}$，实验设置中$K_t$、$K_h$、$K_w$大小一致，都取$K$，化简后最终比例为$\frac{K+1}{K^2}$，本文的模型结构中$K$分别取3、5、7。则计算得到二者比例为4/9、6/25、8/49。将结果进行倒数，发现3D卷积的计算次数/3D解耦卷积的计算次数分别为2.25、4.167、6.125(值越来越大)。由此得出结论，卷积核越大时，3D解耦卷积替代3D卷积的这一过程能节约的计算成本是越高的。

这种解耦策略的优势在于它能够更精确地控制模型在不同维度上的特征学习能力。2D卷积专注于提取每一帧内的空间特征，如边缘、纹理和形状，而1D卷积则专注于捕捉帧与帧之间的时间动态，如运动和变化趋势。这种分步解耦处理的方法使得模型能够更加灵活地调整对局部和全局特征的敏感度，从而在不同的预测任务中实现更优的性能。

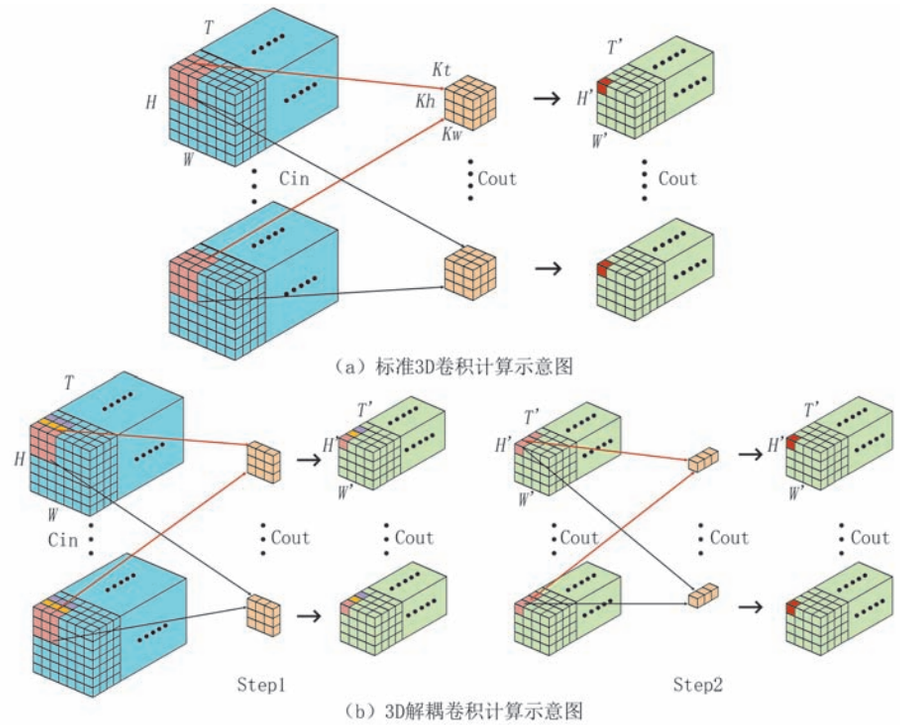

图4　标准3D卷积计算与3D解耦卷积计算

此外，解耦3D卷积模块的设计还允许灵活地调整卷积核的大小和数量。这种灵活性不仅提高了模型的适应性，也有助于处理不同分辨率的视频数据。通过多尺度卷积核的运用，模型能够同时捕捉到视频数据中的微观细节和宏观模式，这在复杂的视频内容理解和预测中尤为重要。

3.5　高频细节小尺度增强模块

在视频预测中，高频特征包括图像的细节和纹理信息，高频特征的提取和增强是提升视频帧质量的重要步骤。在视频序列中，这些特征有助于捕捉场景中的动态变化，如物体的运动和形状变化，这对于视频内容的理解和预测至关重要。为了解决视频帧中由于局部运动不一致和内容变化等造成细节模糊的问题，引入了高频细节小尺度增强模块，用来提升视频帧中的细节信息。

高频特征增强模块的核心是1×1小核卷积，这种卷积操作因其感受野较小，能够更精确地捕捉图像中的高频细节，如边缘和纹理。这种卷积操作常用于特征融合和通道降维，它在保持图像分辨率不变的情况下，可以有效地提取局部特征。通过这种方式，模型能够更准确地捕捉到图像中的微小细节，从而提高视频帧的清晰度和视觉质量。

为了进一步优化模型的性能，论文采用了分组卷积技术。分组卷积通过将输入特征分成多个组，每组独立进行卷积操作，可以显著减少模型的参数量和计算复杂度。这种方法不仅减少了模型的存储需求，还提高了训练和推理的速度，使得模型更适合在资源受限的环境中部署。分组卷积的并行计算特性，也使得模型能够更高效地处理大量数据，这对于实时视频处理应用尤为重要。

在算法的模型中分别将两次小波处理后的高频特征送入高频细节小尺度卷积增强模块，用来对高频信息进行加强学习。增强后的高频特征可以表示为

$$Z_k = Conv_{1\times1}\left(ReLu\left(Conv_{1\times1}\left(F_{LH_k,HL_k,HH_k}\right)\right)\right),\quad k=1,2 \tag{14}$$

其中，当k取1时，式中F_{LH_k,HL_k,HH_k}为第一次小波变换后分离的高频特征，当k取2时，F_{LH_k,HL_k,HH_k}为第二次小波变换后分离的高频特征。

4　实验结果与分析

4.1　实验环境及超参数设置

本文在Pytorch 1.12.1框架下实现了该方法，在NVIDIA GeForce RTX 4090进行了实验，并基于CUDA 10.2平台进行加速计算。使用的操作系统为Ubuntu 18.04，开发语言为Python 3.8.18。

对于超参数设置，在Moving MNIST数据集

上，batch_size为16，学习率(lr)为0.001，空间编码器隐藏层数量(hid_S)为64，时空学习器隐藏层数量(hid_T)为512，小波下采样次数为2，时空学习器中MSDe3D的块数(N_T)为8。在TaXIBJ数据集上，batch_size为16，lr为0.0005，hid_S为32，hid_T为256，小波下采样次数为1，N_T为10。在KTH数据集上，batch_size为4，lr为0.001，hid_S为64，hid_T为256，小波下采样次数为2，N_T为6。三个数据集都使用Adam优化器进行优化，并使用onecycle进行学习率动态调整。

4.2 评估指标

本文遵循当前的研究[16]，对不同的数据集采用不同的评估指标。对于Moving MNIST，TaXIBJ数据集，采用MSE、MAE、SSIM来评估预测质量。对于KTH数据集，采用SSIM、PSNR来评估预测质量。

4.3 数据集

本文通过合成场景和真实场景数据集定量评估本文的模型，三个数据集的详细信息如表1所示。

表1 数据集介绍

数据集	训练集	测试集	图像尺寸	输入帧数	输出帧数	训练轮次
MMNIST	10 000	10 000	(1,64,64)	10	10	2000
TaXIBJ	19627	1334	(2,32,32)	4	4	50
KTH	5200	3167	(1,128,128)	10	20	100

Moving MNIST[7]是一个广泛使用的合成数据集，由0-9中的两位手写体数字组成，在64×64网格内独立移动并从边界反弹，它是时空预测学习的标准基准。其中训练集有10 000个视频序列，测试集有10 000个视频序列，每个视频序列由20帧图像组成，前10帧作为输入，后10帧作为目标值。

TaXIBJ[41]包含从出租车GPS监视器收集的真实北京出租车轨迹数据。TaXIBJ的每一帧都是一个32×32×2的热图，其中最后一个维度表示进出同一区域的交通流量。本文遵循文献[42]中的实验设置，并使用最后四周的数据进行测试，其余的用于训练。本文使用4个观测值来预测接下来的4个连续帧，并将数据归一化为[0,1]。本文采用每帧MSE作为度量。

KTH[43]包含25名受试者在四种不同场景中多次执行的六种人类动作(散步，慢跑，跑步，拳击，挥手，拍手)的灰度视频，将160×120的大小调整为128×128分辨率。使用1-16号人员进行训练，17-25号进行测试。训练时10帧预测后面10帧，测试时10帧预测20帧。

4.4 实验定量分析

4.4.1 Moving MNIST数据集实验

本文在Moving MNIST数据集上根据前十帧来对后十帧进行预测，并对比了近十年来主要模型算法的结果，其中包括自回归的循环模型以及非自回归的多进多出模型，实验结果如表2所示。

表2 在Moving MNIST上的实验结果

方法	来源	MSE↓	MAE↓	SSIM↑
ConvLSTM[4]	(NIPS 2015)	103.3	182.9	0.707
PredRNN[8]	(NIPS 2017)	56.8	126.1	0.867
PredRNN++[11]	(PMLR 2018)	46.5	106.8	0.898
MIM[12]	(CVPR 2019)	44.2	101.1	0.910
LMC[42]	(CVPR 2021)	41.5	-	0.924
E3D-LSTM[10]	(ICLR 2018)	41.3	87.2	0.910
MAU[25]	(NeurIPS 2021)	27.6	-	0.937
PhyDNet[44]	(CVPR 2020)	24.4	70.3	0.947
SimVP[16]	(CVPR 2022)	23.8	68.9	0.948
TAU[18]	(CVPR 2023)	19.8	60.3	0.957
MIMO-VP[19]	(AAAI 2023)	17.7	51.6	0.964
SwinLSTM[15]	(ICCV 2023)	17.7	-	0.962
VMRNN[26]	(CVPR 2024)	16.5	-	0.965
TSIF[45]	(ESWA 2025)	18.3	55.9	0.961
Ours	-	**15.7**	**50.7**	**0.966**

从表2中可以看到本文的模型取得了最有效的结果。与同类型的SimVP模型相比，MSE和MAE分别降低了34.0%、26.4%。与MIMO-VP[19]、SwinLSTM[15]的结果相比，本文模型的MSE降低了11.3%。与最近的工作TSIF[45]相比，MSE降低了14.2%。

本文将模型的预测效果进行了可视化，结果如图5所示。从图中可以看出，本文的预测结果和真实帧的运动轨迹几乎完全一致，数字的形态清晰可见，尤其是边缘的部分，最接近真实帧，如最后一帧数字的整体轮廓非常清晰，数字“3”的勾也没有模糊重影。预测结果证明了本文的小波下采样模块在保留高频细节方面的有效性。通过分解输入图像的特征域，将其分为高频结构和低频细节信息。在此过程中，高频细节得到了有效保留，使得在后续的预测阶段可以利用这些细节来生成更加清晰的图像。高频细节小尺度增强模块进一步处理这些高频部分，提升了图像的细节表现，使得边缘和细节更加清

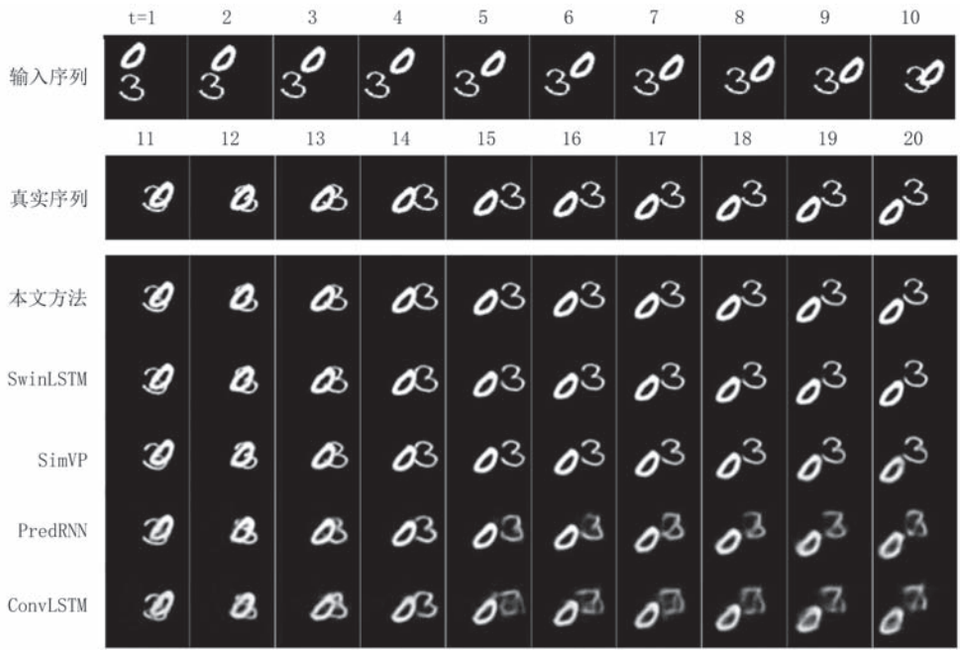

图5　Moving MNIST可视化结果

晰。此外，预测帧整体形态与真实帧保持一致，这得益于3D解耦卷积在时空特征提取方面的作用。通过分别处理空间和时间特征，提高了模型对时空依赖关系的学习能力。解耦3D卷积能够更准确地捕捉时空动态特征，确保模型在进行长时间帧预测时，能够生成与真实运动轨迹一致的结果。

总结来说，本文的方法通过特征域小波下采样和3D解耦卷积的时空特征提取能力，以及高频细节小尺度增强模块的结合，使得模型在Moving MNIST数据集上的预测效果在定量指标和视觉效果上均达到了较好的表现。

4.4.2　TaXIBJ数据集实验

本文在TaXIBJ数据集上采用前四帧的交通流量图来预测后四帧，并与经典的模型以及最新的预测方法进行对比来评估本文模型的性能，如表3所示。

从表3可知，除了MAE略微高于TAU模型，本文的方法在MSE和SSIM评估指标上取得了最优的性能。与SimVP相比，本文的MSE降低了19.6%，MAE降低了1.9%。为了直观感受性能之间的差异，本文还将数据结果进行了可视化，如图6所示。由于图像之间的差异不明显，本文特地将预测帧与真实帧的帧差结果也进行了可视化，亮度越高，代表差值越大。可以看出，本文方法的像素值差异比同类型的SimVP模型更小，体现出更优的性能。

表3　在TaXIBJ上的实验结果

方法	来源	MSE × 100↓	MAE↓	SSIM↑
ConvLSTM[4]	(NIPS 2015)	48.5	17.7	0.978
PredRNN[8]	(NIPS 2017)	46.4	17.1	0.971
PredRNN++[11]	(PMLR 2018)	44.8	16.9	0.977
E3D-LSTM[10]	(ICLR 2018)	43.2	16.9	0.979
MIM[12]	(CVPR 2019)	42.9	16.6	0.971
PhyDNet[44]	(CVPR 2020)	41.9	16.2	0.982
SimVP[16]	(CVPR 2022)	41.4	16.2	0.982
TAU[18]	(CVPR 2023)	34.3	**15.6**	0.983
TSIF[45]	(ESWA 2025)	40.1	16.0	0.982
Ours	-	**33.3**	15.9	**0.987**

在交通流量预测中，高频细节往往代表突发的交通拥堵或变化，这些细节对预测准确性至关重要。通过小波下采样，本文能够将图像特征分解为高频细节和低频结构部分，确保在下采样过程中尽量保留高频细节。这种方法有效地避免了重要信息的丢失，使得模型能够在预测阶段生成更加清晰和准确的图像。交通流量图是不断随空间和时间变化的，具有复杂而又动态变化的时空依赖性，解耦3D卷积通过分别处理空间和时间特征，有效捕捉和处理了这些复杂的时空依赖性，从而提高了预测的准确性。

4.4.3　KTH数据集实验

本文在KTH数据集上的实验结果如表4所示。实验中，训练时采用10帧预测10帧的模式，测试时

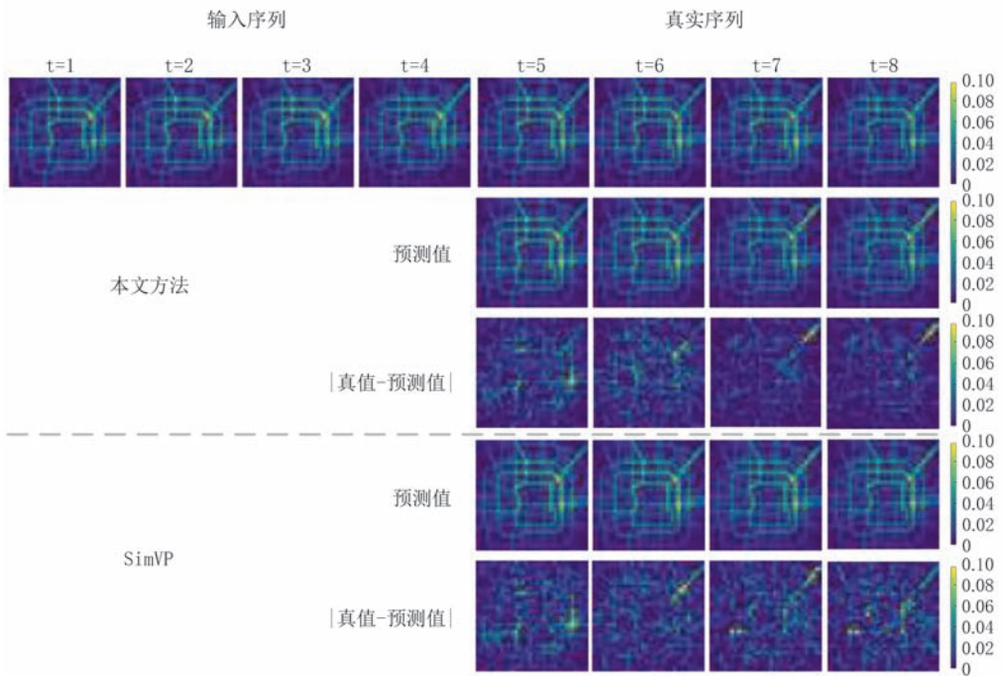

图6 TaXIBJ可视化结果

将预测得到的10帧再次输入网络，预测接下来的10帧，从而总共得到20帧的预测结果。这种预测长度的增加，进一步考验了模型的时空学习能力以及预测长序列的能力。本文与自回归和非自回归的模型都进行了比较，另外为了使得论述更加全面，本文也与使用Diffusion Model的方法进行了对比。在表格的前两行展示了MCVD[23]、ExtDM[32]方法的结果，可以发现这类模型的PSNR较低，精度不如主流的一些视频预测算法，这是因为这类生成式的模型在动作预测的准确度上效果有待提高。从表4中可以看出，与其他模型相比本文的模型取得了最优的性能。

可视化结果如图7所示。在长序列预测中，本文的模型与真实视频序列保持了最相似的形态，动作的走向以及脚步的姿态与真实帧最为接近。与同类型非自回归模型SimVP相比，在预测的11到20帧中，本文的方法表现出更高的准确性，人物的整体跑步姿势以及头部、手部、脚部的细节预测都更加贴近真实帧。如图8所示，特别将11至13帧展示出来，本文方法在红框部分细节预测都是与真实帧最相似的，预测是最精准的。在后面的21到25帧，尽管本文的预测结果在后期逐渐变得模糊，但人物的整体形态、跑步姿势仍然更像真实序列，尤其是脚部的动作。再观察26到30帧图像，本文方法则逐渐模糊，导致最后几帧清晰度下降。因此对于长序列的学习能力还有待进一步提高。而相比自回归模型代表PredRNN，前几帧预测较为准确，但随着预测长度的增加，预测的人物姿态逐渐偏离真实值，这是由于循环误差累积造成的。这也证明了非自回归结构在长预测任务中的优势。

表4 在KTH上的实验结果

方法	来源	KTH(10→20)	
		SSIM↑	PSNR↑
MCVD[23]	(NIPS 2022)	-	23.84
ExtDM[32]	(CVPR 2024)	-	24.75
ConvLSTM[4]	(NIPS 2015)	0.712	23.58
MCnet[46]	(ICLR 2017)	0.804	25.95
PredRNN[8]	(NIPS 2017)	0.839	27.55
PredRNN++[11]	(PMLR 2018)	0.865	28.47
E3D-LSTM[10]	(ICLR 2018)	0.879	29.31
Znet[13]	(ICME 2019)	0.817	27.58
SimVP[16]	(CVPR 2022)	0.905	33.72
TAU[18]	(CVPR 2023)	0.911	34.13
VMRNN[26]	(CVPR 2024)	0.907	34.06
TSIF[45]	(ESWA 2025)	0.907	33.87
Ours	-	**0.912**	**34.15**

本文的模型在KTH数据集上的表现证明了其在处理长预测任务中前期预测的有效性。通过小波下采样和3D解耦卷积的结合，本文的模型能够更好地捕捉和处理时空依赖性，提高了视频前期预测的准确性。但对长序列的后期学习能力有待进一步提高。这些结果展示了本文模型在动作预测任务中的

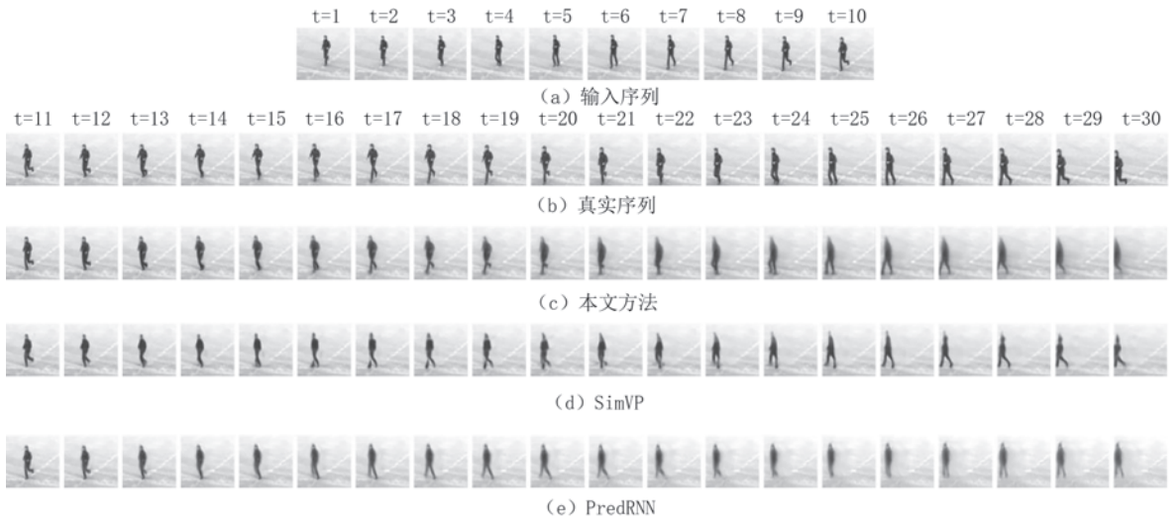

图7　KTH可视化结果

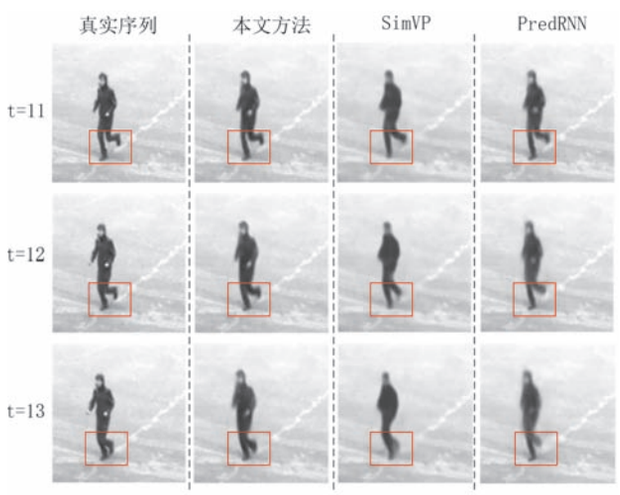

图8　KTH局部细节图

潜力，为未来研究的改进方向提供了重要参考。

4.5　消融实验

为验证本文方法中提出的模块的有效性，本文在Moving MNIST数据集上进行了以下消融实验。实验中，空间编码器部分包括小波下采样、卷积下采样、高频细节小尺度增强模块。时空学习器部分则包括2D（使用纯2D卷积进行时空学习，采用SimVP中的Inception模块，并去掉尺度为11的大核卷积）、3D（在相同设置下使用3D卷积替代2D卷积）和De3D（3D解耦卷积）。

整个实验采用相同的超参数设置：epoch为200，学习率为0.001。

4.5.1　模块有效性分析

在进行消融实验的过程中，论文逐步分析了不同模型组件对视频预测性能的影响。通过对比模型1和模型2，能够验证小波下采样技术的有效性。小波下采样通过将视频特征分解为高频结构与低频纹理细节分量后分别处理，仍然保留了重要的细节特征信息，从而在模型2中实现了MSE的改善。

进一步地，模型2与模型3的比较揭示了高频特征增强模块的显著贡献。MSE的降低0.45表明，通过专门设计的模块来增强高频细节特征，可以有效地提升模型对细节和纹理等关键信息的捕捉能力，从而进一步提高预测的准确性。

在模型3和模型4的对比中，观察到3D卷积在处理时空特征方面的强大能力，这反映在MSE的显著下降上。3D卷积通过在三个维度上捕捉特征，增强了模型对视频序列中时空动态的理解。然而，这种性能的提升并非没有代价，3D卷积显著增加了模型的参数量和计算需求，这在资源受限的环境中可能成为一个限制因素。

最后，模型4与模型5的比较中，看到了3D解耦卷积模块的显著优势。MSE的进一步降低1.83，结合参数量的减少和浮点计算次数的大幅度降低，表明了解耦3D卷积在保持预测性能的同时，显著提高了计算效率。这种设计通过分步骤处理时空特征，减少了不必要的计算，同时保持了对关键信息的捕捉能力，实现了性能与资源消耗之间的良好平衡。

综合这些消融实验的结果，可以得出结论，小波下采样、高频细节小尺度增强模块以及解耦3D卷积模块都是提升视频预测模型性能的关键步骤。这些改进不仅各自对模型性能有积极影响，而且它们之间的协同作用进一步优化了模型的整体表现。通过

这些技术的融合，能够在保持高效计算的同时，实现对视频内容的准确预测。

4.5.2 参数量与计算复杂度分析

为了更全面地评估各模型的性能，我们对消融实验中的不同模块组合从参数量(Params)和计算复杂度(FLOPs)的角度进行了对比分析。根据表5中的结果可以观察到：

(1)参数量(Params)分析

模型1到模型3中，随着空间编码器部分的优化(引入小波下采样和高频细节小尺度增强模块)，参数量略有增加，从25.521 M提升至25.563 M再到25.633 M，但变化幅度并不大，表明空间编码器优化在参数量上的代价较低。

模型3到模型5中，时空学习器部分从2D卷积到3D卷积再到De3D模块，参数量大幅下降，此处3D卷积的参数量要比2D卷积更小的原因是在实验设计中，3D卷积模块中去掉了2D卷积中的大核卷积(11×11)。而De3D模块比3D卷积模块参数量小的原因正是由于将3D卷积解耦成2D和1D的结果。可以看见最终优化后模型5的参数量为18.015 M，比模型1减少了29.4%，展现了De3D模块的高效性。

(2)计算复杂度(FLOPs)分析

在计算复杂度方面，加入小波下采样(模型2)和高频细节小尺度增强模块(模型3)使FLOPs略微增加，但仍保持在可接受范围内(从11.267 G增加到13.722 G)。

与此相对，时空学习器从2D卷积替换为3D卷积后(模型4)，计算复杂度显著提升至336.896 G，而在采用De3D替代后(模型5)，FLOPs又降低至52.497 G，仅为模型4的15.6%。这进一步表明，De3D模块在性能与效率之间取得了良好的平衡。

表5 消融实验(模块消融)

		模型1	模型2	模型3	模型4	模型5
空间编码器	小波下采样		√	√	√	√
	卷积下采样	√				
	高频细节小尺度增强模块			√	√	√
时空学习器	2D	√	√	√		
	3D				√	
	De3D					√
评估指标	MSE↓	34.26	34.16	33.71	26.36	24.53
	Params(M)	25.521	25.563	25.633	19.024	18.015
	FLOPs(G)	11.267	12.778	13.722	336.896	52.497

4.5.3 小波下采样次数消融实验

为了验证本文的设计思路，对小波下采样的次数单独进行消融实验。实验结果如表6所示。

表6 小波下采样次数消融实验

数据集(评估指标)	分辨率	一次小波	二次小波
MMNIST(MSE↓)	64×64	26.83	**24.53**
TaXIBJ(MSE/100↓)	32×32	**33.3**	37.4
KTH(SSIM↑)	128×128	0.897	**0.912**

其中MMNIST数据集训练200轮，TaXIBJ训练50轮，KTH训练100轮。每一个数据集的消融实验采用相同的超参数设置以及相同的环境。唯一不同的只有小波下采样的次数(一次小波为两倍下采样，二次小波为四倍下采样。这里一次小波分解后经过一次高频细节小尺度增强模块，二次小波分解后则分别经过两次高频细节小尺度增强模块)。

从表格中可以看出，MMNIST和KTH数据集采用二次小波的效果是更好的，而TaXIBJ数据集则采用一次小波效果更好。TaXIBJ数据集的分辨率较小(32×32)，因此在进行二次小波下采样时，下采样的倍数较高，特征图尺寸过小，可能导致低频结构的丢失，使得模型无法有效学习到数据中的重要特征。相较之下，一次小波只进行一次下采样，能够较好地保留足够的低频信息，使得模型能在较小的特征图上学习到更多重要的低频结构和纹理，从而导致一次小波的表现更优。MMNIST和KTH这两个数据集的分辨率较大(MMNIST 64×64，KTH 128×128)，因此在采用二次小波时，由于高频细节小尺度增强模块的加入，能够更好地捕捉到高频细节和小尺度特征，这对于高分辨率图像尤其

重要。二次小波通过两次高频细节小尺度增强模块，有助于提高高频特征的表示，从而改善图像的质量，因此这两个数据集在二次小波下表现更好。

综上所述，数据集的分辨率与小波下采样的次数密切相关。在分辨率较大的数据集（如MMNIST和KTH），二次小波能够更好地保留和增强细节信息，因此在这些数据集上表现更好；而对于分辨率较小的数据集（如TaXIBJ），一次小波因其较少的下采样次数，更适合学习低频结构和较为简单的特征，导致一次小波的效果更优。

5 总结与讨论

本文因视频内容变化以及物体多种可能的运动等带来的预测难题，提出了一种基于特征域结构与纹理分解的多尺度3D解耦卷积视频预测方法，旨在提升视频预测的精度与质量。通过将特征信息分解为低频结构与高频细节，模型能够分别学习并处理不同频率的信息。低频结构信息的时间相关性和运动一致性通过3D解耦卷积模块建模，而高频细节则通过小尺度分组卷积增强模块进行细化处理，从而提高了局部细节的清晰度。相比传统的3D卷积方法，本文提出的3D解耦卷积模块不仅能够有效捕获视频的低频结构时空相关性，还大幅减少了计算成本和内存需求。此外，高频细节增强模块专注于提升图像的细节质量。总体而言，本文设计的模型既关注物体的运动信息，又兼顾图像的局部细节，实现了更精确的时空特征提取与细节还原，为视频预测任务提供了新的参考方案。

但是，本文工作仍存在一些局限性。例如，对于长序列的时空学习能力，会随着预测帧的增加而逐渐衰退，由于计算资源的限制，本文方法选择的是一些小型的数据集进行训练，对于大型复杂的数据集的学习能力还有待探究。未来，还需要继续优化模型，尤其是在长期帧预测以及如何处理高分辨率视频方面，这些领域仍然是需要重点攻克的难点。

参 考 文 献

[1] Pan Min-Ting, Wang Yun-Bo, Zhu Xiang-Ming, et al. A review of deep prediction learning methods based on unlabeled video data. Acta Electronica Sinica, 2022, 50 (04): 869-886(in Chinese)
(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. 电子学报，2022，50(04):869-886)

[2] Castrejon L, Ballas N, Courville A. Improved conditional vrnns for video prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7608-7617

[3] Chandra R, Bhattacharya U, Bera A, et al. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8483-8492

[4] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015, 28: 802-810

[5] Cao C, Lu Y, Zhang Y. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. IEEE Transactions on Image Processing, 2024, 33: 1810-1825

[6] Cao C, Zhang H, Lu Y, et al. Scene-dependent prediction in latent space for video anomaly detection and anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(1): 224-239

[7] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 843-852

[8] Wang Y, Long M, Wang J, et al. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 879-888

[9] Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104, 2016

[10] Wang Y, Jiang L, Yang M H, et al. Eidetic 3D LSTM: A model for video prediction and beyond//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

[11] Wang Y, Gao Z, Long M, et al. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 5123-5132

[12] Wang Y, Zhang J, Zhu H, et al. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9154-9162

[13] Zhang J, Wang Y, Long M, et al. Z-order recurrent neural networks for video prediction// Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 230-235

[14] Wang Y, Wu H, Zhang J, et al. Predrnn: A recurrent neural network for spatiotemporal predictive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 2208-

2225

[15] Tang S, Li C, Zhang P, et al. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 13470-13479

[16] Gao Z, Tan C, Wu L, et al. Simvp: Simpler yet better video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3170-3180

[17] Tan C, Gao Z, Li S, et al. Simvp: Towards simple yet powerful spatiotemporal predictive learning. arXiv preprint arXiv:2211.12509, 2022

[18] Tan C, Gao Z, Wu L, et al. Temporal attention unit: Towards efficient spatiotemporal predictive learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18770-18782

[19] Ning S, Lan M, Li Y, et al. MIMO is all you need: A strong multi-in-multi-out baseline for video prediction//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(2): 1975-1983

[20] Ho J, Salimans T, Gritsenko A, et al. Video diffusion models. //Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 8633-8646

[21] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015: 234-241

[22] Çiçek Ö, Abdulkadir A, Lienkamp S S, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation// Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 2016: 424-432

[23] Voleti V, Jolicoeur-Martineau A, Pal C. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation//Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 23371-23385

[24] Höppe T, Mehrjou A, Bauer S, et al. Diffusion models for video prediction and infilling. arXiv preprint arXiv:2206.07696, 2022

[25] Chang Z, Zhang X, Wang S, et al. Mau: A motion-aware unit for video prediction and beyond//Proceedings of the 35th Conference on Neural Information Processing Systems, 2021: 26950-26962

[26] Tang Y, Dong P, Tang Z, et al. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 5663-5673

[27] Wu H, Xu F, Chen C, et al. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction// Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 2917-2926

[28] Seo M, Lee H, Kim D, et al. Implicit stacked autoregressive model for video prediction. arXiv preprint arXiv: 2303.07849, 2023

[29] Li P, Zhang C, Xu X. Fast Fourier inception networks for occluded video prediction. IEEE Transactions on Multimedia, 2024, 26: 3418-3429

[30] Li P, Zhang C, Yang Z, et al. Pair-wise layer attention with spatial masking for video prediction. arXiv preprint arXiv: 2311.11289, 2023

[31] Ye X, Bilodeau G A. STDiff: Spatio-Temporal Diffusion for Continuous Stochastic Video Prediction//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 6666-6674

[32] Zhang Z, Hu J, Cheng W, et al. Extdm: Distribution extrapolation diffusion model for video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 19310-19320

[33] Chen L, Fu Y, Wei K, et al. Instance segmentation in the dark. International Journal of Computer Vision, 2023, 131(8): 2198-2218

[34] Chen L, Fu Y, Gu L, et al. Frequency-aware feature fusion for dense image prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10763-10780

[35] Chen L, Gu L, Zheng D, et al. Frequency-adaptive dilated convolution for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 3414-3425

[36] Chen L, Gu L, Li L, et al. Frequency dynamic convolution for dense image prediction. arXiv preprint arXiv:2503.18783, 2025

[37] Bae W, Yoo J, Chul Ye J. Beyond deep residual learning for image restoration: persistent homology-guided manifold simplification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 145-153

[38] Yao T, Pan Y, Li Y, et al. Wave-vit: Unifying wavelet and transformers for visual representation learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel, 2022: 328-345

[39] Liu P, Zhang H, Lian W, et al. Multi-level wavelet convolutional neural networks. IEEE Access, 2019, 7: 74973-74985

[40] Mallat S G. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7): 674-693

[41] Zhang J, Zheng Y, Qi D. Deep spatio-temporal residual networks for citywide crowd flows prediction//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 1655-1661

[42] Lee S, Kim H G, Choi D H, et al. Video prediction recalling long-term motion context via memory alignment learning// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 3054-3063

[43] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach//Proceedings of the 17th International

Conference on Pattern Recognition, Cambridge, UK, 2004: 32-36

[44] Guen V L, Thome N. Disentangling physical dynamics from unknown factors for unsupervised video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11474-11484

[45] Yuan J, Wu F, Zhao L, et al. A dual-stage spatiotemporal information fusion network for video prediction. Expert Systems with Applications, 2025, 276: 127189

[46] Villegas R, Yang J, Hong S, et al. Decomposing motion and content for natural video sequence prediction. arXiv preprint arXiv:1706.08033, 2017

ZHENG Ming-Kui, Ph. D., associate professor. His research interests include multi-modal information intelligent coding, computer vision and haptics.

WU Kong-Xian, M. S. candidate. Her current research interests include video prediction and video encoding.

QIU Xin-Tao, M. S. candidate. His current research interest is video super resolution.

ZHENG Hai-Feng, Ph. D., professor. His research interests include intelligent vehicle networking, 6G/synaesthesia integration, embodied intelligent perception and decision making.

ZHAO Tie-Song, Ph. D., professor. His research interests include image processing and computer vision, intelligent video coding and communication, tactile information and virtual reality.

Background

Nowadays, people have increasingly high demands for video services, with a growing demand for ultra high definition, high frame rate, and immersive videos. Along with this comes the huge challenge of storing and transmitting massive video data. Although network transmission rates are faster in the 5G era, improving the performance of video encoding technology and compressing massive video data more efficiently is the fundamental solution to storage and transmission problems. Due to temporal correlation being the most important characteristic of video signals, inter frame predictive coding has become the core of video encoding. High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC) both use block based motion estimation and motion compensation techniques to fully utilize the temporal features in the video to generate the predicted result of the current frame to be encoded. Then, only the residual between the current frame and its predicted frame needs to be encoded to reconstruct the current frame at the decoding end. Obviously, the higher the inter frame prediction accuracy, the less data needs to be encoded and transmitted in the end. However, the above inter frame prediction techniques have the following problems: (1) the accuracy of block based motion estimation is poor; (2) Motion compensation uses simple motion models, which are difficult to fit complex movements; (3) The types of reference frames used for inter frame prediction are not diverse enough. Due to the above reasons, the inter frame prediction accuracy in mainstream encoding standards is not ideal and urgently needs to be further improved to enhance video encoding efficiency.

Video prediction algorithms have rapidly developed in recent years due to their ability to utilize massive amounts of unlabeled natural data to learn the intrinsic representations of videos, saving a lot of manual labeling time. They also have broad application value in fields such as weather prediction, traffic flow prediction, robot decision-making, and limited driving. By utilizing video prediction technology, the reconstructed frames already encoded in the video encoder are used to generate reference frames that are more similar to the current frame through deep learning methods, and integrated into the video encoder to achieve more efficient video compression efficiency.

This article has made certain contributions to improving the accuracy of video prediction. Experimental results on synthetic data and real scene datasets show that the algorithm designed in this article has higher prediction accuracy than existing algorithms. It has more accurate prediction performance in local details and overall prediction morphology. Among them, the MSE on the Moving MNIST dataset is 15.7, which is 34%, 20.7%, 11.3%, and 4.8% lower than the existing advanced algorithms Simvp, TAU, SwinLSTM, and VMRNN, respectively.

This paper is supported by the National Natural Science Foundation of China: Research on Observable Compression Effect and Its Application in Ultra High Definition Video Encoding (62171134), and the Fujian Provincial Science and Technology Major Special Project: Research and Industrial Application of High end Intelligent Controllers for Autonomous Unmanned Systems (2022HZ026007).