基于差分注意力的时空小波分析视频预测算法

金贝贝[1,2)], 胡瑜[1,2)*]

[1)] (中国科学院计算技术研究所智能计算机研究中心 北京 100190)
[2)] (中国科学院大学 北京 100149)
(huyu@ict.ac.cn)

摘 要: 针对视频预测中空间结构信息细节和时序运动依赖关系难以准确预测的问题, 受人类视觉过程的启发, 提出一种基于差分注意力机制的时空小波分析视频预测算法. 首先利用时空小波分析模块对视频内容进行多频分解, 增强模型对于高频细节信息以及过程性运动的理解能力; 然后利用差分注意力机制指导模型更高效、合理地分配注意力资源, 提升对瞬时运动特征的表达能力. 在 KTH, Cityscapes, BAIR, KITTI, Caltech Pedestrian 数据集上的实验结果表明, 所提算法在 PSNR, SSIM, LPIPS 评价指标上取得了比已有算法更优异的效果; 同时, 可视化的对比也表明所提算法的预测结果更加清晰.

关键词: 视频预测; 时序小波变换; 注意力机制; 空间小波分析
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2022.18895

Spatial-Temporal Wavelet Analysis Video Prediction Based on Differential Attention Mechanism

Jin Beibei[1,2)] and Hu Yu[1,2)*]

[1)] (*Research Center for Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)
[2)] (*University of Chinese Academy of Sciences, Beijing* 100149)

Abstract: Inspired by the visual process of human, a video prediction algorithm based on spatial-temporal wavelet analysis and differential attention is proposed to solve the problem that it is difficult to accurately predict the details of spatial structure information and the dependence of temporal motion. Firstly, the spatial-temporal wavelet analysis module is used to decompose the video in multiple frequencies, so as to enhance the model's ability to understand high-frequency details and procedural motion. Then, the differential attention mechanism guides the model to allocate attention resources more efficiently and reasonably, and improves the expression ability of instantaneous motion. Experimental results on the KTH, Cityscapes, BAIR, KITTI, Caltech Pedestrian datasets show the proposed algorithm achieves better results than the existing algorithms in the quantitative evaluation metrics of PSNR, SSIM and LPIPS. Meanwhile, the visualization results also show that the prediction of the proposed algorithm is clearer.

Key words: video prediction; temporal wavelet analysis; attention mechanism; spatial wavelet analysis

收稿日期: 2021-01-19; 修回日期: 2021-07-18. **基金项目**: 国家重点研发计划科技创新 2030——"新一代人工智能"重大项目(2018AAA0102701); 空间智能控制技术实验室开放基金(HTKJ2019KL502003). **金贝贝**(1992—), 女, 博士研究生, 主要研究方向为计算机视觉、视频预测; **胡瑜**(1975—), 女, 博士, 研究员, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为自动驾驶感知与决策、神经网络架构搜索、深度学习.

视频预测任务旨在根据已观测的视频帧预测未来的视频帧，与目标识别、分类和检测等任务不同，它无须对数据集进行人工标签化预处理. 互联网视频数据的爆炸式增长为其提供了极大的发展空间. 目前它已成为计算机视觉领域一个研究热点，并且在自动驾驶或机器人导航等领域拥有广泛的应用前景.

目前，视频预测仍然是一项非常具有挑战性的任务，它不仅需要理解空间维度中各个目标之间的相互关系，预测出清晰完整的外观特征，还需要掌握时序维度上各种动态运动的复杂演化. 基于深度学习的视频预测算法可以分为基于直接像素合成的视频预测算法[1-7]和基于变换的视频预测算法[8-12]. 基于直接像素合成的视频预测算法通常采用生成对抗网络(generative adversarial networks, GAN)的网络架构，利用生成器提取视频内容特征，直接输出预测的视频像素值；基于变换的视频预测算法则是间接地预测相邻帧之间的变换关系，之后将预测得到的变换关系应用到当前帧，得到要预测的下一帧内容.

由于视频数据的高维特性和各种时序运动演化的复杂性，现有的预测方法仍然存在空间上的细节丢失和时序上的运动预测不一致的问题，且生成的结果过于平滑，未能充分地保留高频细节信息. Jin 等[13]对此问题进行了详细分析，指出人类视觉系统对时空频率信息具有多通道特性，可以在对数尺度上将视网膜图像以近似相等带宽分解为不同频带进行处理，这些特性使人类视觉系统在处理视觉内容时能够更好地识别细节信息和运动信息. 小波分析是一种空间-尺度(时间-频率)分析方法，具有多分辨率(频率)分析的特点，能很好地表示空间(时间)频率信号的局部特征，这与人类视觉系统有很好的一致性.

因此文献[13]将小波分析的这种特性引入视频预测任务，在卷积神经网络(convolutional neural network, CNN)的基础上分别提出级联的空间小波分析模块(spatial wavelet analysis module, S-WAM)和时序小波分析模块(temporal wavelet analysis module, T-WAM)，来加强空间细节和时序过程性运动的预测. 在 KTH 和 BAIR 数据集上的实验结果表明，该方法取得了明显的性能提升. 然而，该方法中 T-WAM 模块针对时序信息的提取是建立在缓存的一段观测序列的基础上的，更多关注过程性的运动信息. 场景中不同目标之间的运动具有丰富的多样性，除了过程性运动之外，相邻帧间的瞬时运动对于理解场景动态信息和细节变化也同样重要. 此外，之前的工作[5-8]采取的是平均分配注意力资源的方式，在这种方式下，模型的注意力是基于静态信息获取的，与当前状态下场景中不同目标的运动变化程度无关. 实际预测过程中，运动较快的物体比运动较慢的物体有更大的预测难度，模型有必要根据物体的运动幅度有针对性地分配注意力. 视频的帧间差信息是通过对视频图像序列中的相邻 2 帧进行差分运算获得的，这些信息可以直接地反映运动目标的轮廓和运动的瞬时特征，是一种高效快速的运动检测方法. 本文提出差分注意力模块(differential attention module, DAM)，根据物体的帧间差信息生成注意力图，帮助模型在预测过程中更好地感知物体的动态运动特征，更加高效、合理地区分不同的运动目标. 这种建立在帧间变化信息上的注意力机制可以反映视频中的瞬时运动特征，与时序小波分析模块提供的过程性运动特征融合起来增强模型对于时序运动的表达能力.

本文还在文献[13]的基础上扩展了实验数据集，在多个视频数据集上的定量及定性的实验结果表明，本文算法在细节保留和时序运动一致性方面取得了更好的效果.

1　相关工作

目前，大多数的视频预测模型采取的都是直接的像素合成方法. 文献[2]提出一种多尺度预测架构，首次将对抗训练引入视频预测中用于提升预测结果的真实度，为了改善画面预测模糊的问题，提出图像梯度差损失(gradient difference loss, GDL)惩罚图像像素梯度预测的差异，之后其在视频预测模型中被广泛采用. 受到神经科学预测编码理论的启发，文献[3]提出一种视频预测架构 PredNet，利用循环卷积网络将实际信号和预期信号之间的差异视为误差信号以更新预测网络参数，然而该方法只适用于单帧预测，不适合进行长时多帧预测. 文献[4]提出一种新的时空长短期记忆(long short-term memory, LSTM)单元，可以利用 LSTM 同时提取和记忆时空表征，而不是将存储状态只局限于单个 LSTM 单元内，改善了预测时序一致性问题. 为了改进文献[4]中的梯度反向传播问题，文献[5]提出 PredRNN++，旨在提升模型的动态建模能力，利用梯度高速单元与级联的 LSTM 一起自适应地捕捉短期和长期视频依赖关系，让

模型做出更长期的预测. 为了获得更真实、合理的预测, 很多模型常常借助一些外部先验信息(如光流[6]、轨迹或动作标签[7-8])进行预测, 然而这些模型在很大程度上依赖于信息的准确性, 当这些信息无法准确获取时就不能很好地发挥作用, 同时外部信息的收集往往也费时、费力, 限制了它们的应用.

除了直接预测方法之外, 基于变换的方法通过预测帧间的变换关系间接地进行未来帧的预测. 文献[9]提出一个生成式模型, 将场景中的可变因素编码到要预测的变换核来生成预测帧, 帮助模型有效地区分场景中的变化量和不变量. 文献[10]提出一种在仿射变换空间中运行的基于变换的模型, 根据给定观测帧之间的历史仿射变换预测未来帧间的局部仿射变换, 并将其应用于前一观测帧的图像块上生成下一帧. 基于变换的预测模型不再需要存储低层细节内容, 只需要从观测序列提取足够的变换信息, 从而减少模型参数, 然而这些模型很容易受到噪声的影响, 性能不够稳定. 如在文献[12]中, 基于变换的方法在建模大幅度的运动时被证明是无效的.

本文在文献[13]的基础上加入帧间差注意力机制, 有效地提升了视频预测的质量. 小波分析被广泛应用于图像压缩、图像重建等领域, 小波变换可以将原始信号分解为不同分辨率(频率)的时频信号表示的集合, 在图像处理中, 经常使用的是离散小波变换(discrete wavelet transform, DWT). 文献[14]提出一种利用滤波器组的小波变换快速实现方法, 小波的滤波器组实现可以解释为计算给定母小波的离散子小波的小波系数.

参考文献[14], 空间 DWT 可以将原始图像分解为 1 幅低频分量图像和 3 幅高频分量图像(分别是水平方向、垂直方向和对角线方向). 时序 DWT 可以沿时间轴将一段视频序列分解为对应的低频分量和高频分量, 这些分解后的频域特征图是在不同频率下的原始图像的信息表达, 可以提供丰富的视觉信息. 多级离散小波分析可以通过在上一级得到的频率分量图像上重复类似的过程获得. 小波变换的多分辨率(频率)分析特性与人类视觉系统理解视觉信息的方式一致, 这为本文算法提供了生物学基础. DWT 技术的介绍详见文献[14].

2 本文算法

2.1 定 义

定义 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$ 表示长度为 m 的输入观测视频序列, $\boldsymbol{Y}=\{\boldsymbol{x}_{m+1},\boldsymbol{x}_{m+2},\cdots,\boldsymbol{x}_{m+n}\}$ 表示待预测的未来 n 帧视频序列的真实图像, $\hat{\boldsymbol{Y}}=\{\hat{\boldsymbol{x}}_{m+1},\hat{\boldsymbol{x}}_{m+2},\cdots,\hat{\boldsymbol{x}}_{m+n}\}$ 表示模型预测的未来 n 帧视频序列的预测图像. 模型训练的目的就是最小化 $\boldsymbol{Y}$ 和 $\hat{\boldsymbol{Y}}$ 之间的误差, 使预测更加清晰真实. 为了简化描述, 本文以 $n=1$ 为例进行算法描述.

2.2 网络结构

本文整体采用的是 GAN 架构[15], 包含一个生成器用来观测历史视频帧图像、提取特征, 并预测输出未来的视频帧序列, 以及一个判别器用来判断生成器的预测结果的真实性. 2 个模型交替训练, 不断对抗, 最终实现均衡状态. 本文的生成器是编码器-解码器结构, 生成器模型如图 1 所示, 在文献[13]的基础上增加了 DAM. 编码器需要根据输入的视频图像提取空间和时序上的特征, 从而将输入的时序序列编码为隐含层特征张量, 同时需要记忆历史信息, 提取视频中的时序运动特征; 解码器负责将这些编码器提取到的隐含层特征进行解码, 生成下一帧的预测, 预测结果可继续作为编码器的输入送到网络以进行后续帧的迭代预测.

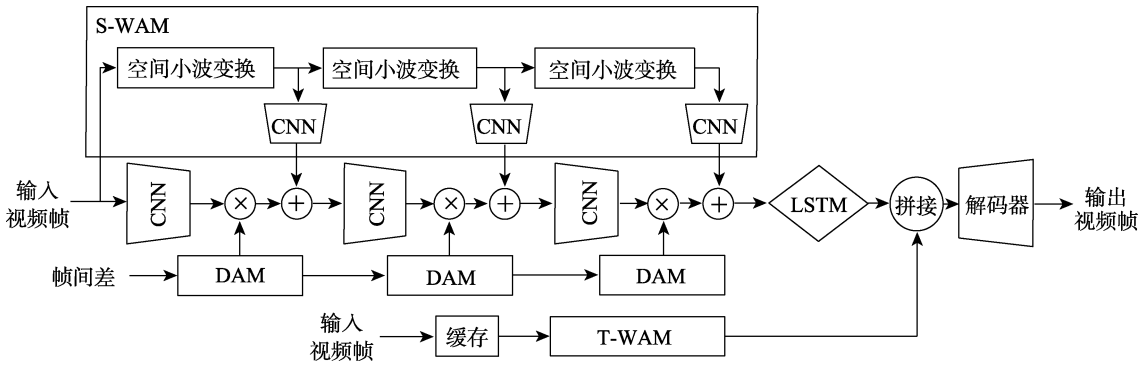

图 1 本文网络结构图

生成器的编码器包括主干网络、DAM、级联的 S-WAM 和 T-WAM 这 3 部分，解码器由反卷积层和上采样层组成. 由于视频预测是一种像素密集型的任务，预测输出和输入图像分辨率相同，因此主干网络既要充分地提取时空特征，又要尽可能避免网络中由于增大感受野引入的下采样层造成过多的纹理信息丢失. 本文采用超分辨率任务中广泛采用的 RRDB(residual-in-residual dense block)[16]实现主干网络的搭建，该模块利用多层的残差连接和密集连接有效地保留特征提取过程中的细节特征. 在每个时刻，输入视频帧到主干网络提取不同感受野下多尺度的空间信息，信息的提取过程可以表示为

$$\left[\boldsymbol{O}_t, \boldsymbol{h}_t\right] = f_{\mathrm{s}}\left(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}\right).$$

其中，f_{s} 表示主干网络；$\boldsymbol{O}_t$ 为时刻 t 编码器的输出；$\boldsymbol{h}_t$ 为 t 时刻 LSTM 的隐含层输出，保留网络的长短时记忆；$\boldsymbol{h}_{t-1}$ 为上一时刻 LSTM 的隐含层输出.

2.2.1　DAM

主干网络中 CNN 需要进行下采样增大感受野，会不可避免地引起细节丢失，影响了预测精度. 为了捕获视频中物体的瞬时运动，本文提出自适应 DAM. DAM 将帧差法得到的瞬时运动编码成注意力特征图，引导编码器主干结构根据不同位置的瞬时运动强度合理分配注意力，并级联了 3 个 DAM 将帧间的差分图编码为动态注意力图. 为了减少参数量，每级的注意力图在主干结构的特征图间进行了共享. 第 i 个 DAM 的处理过程为

$$\left[\boldsymbol{a}_t^i, \boldsymbol{G}_t^i\right] = f_{\mathrm{DAM}_i}\left(\boldsymbol{G}_t^{i-1}\right).$$

其中，f_{DAM_i} 表示第 i 个 DAM，由一个浅层 CNN 实现；$\boldsymbol{a}_t^i$ 表示在时刻 t 第 i 个 DAM 的注意力图输出，注意力图会分配到对应的主干网络层，与提取到的卷积特征融合；$\boldsymbol{G}_t^i$ 为 DAM 下采样层之后的输出. 特别地，第 1 级 DAM 的输入为 t 时刻相邻 2 帧的帧间差.

由于注意力特征是根据帧间差信息获得的，而不是单纯地由静态视频帧计算得到的，因此，注意力中引入基于相邻帧间物体运动信息得到的动态变化特征使模型可以分配更多的注意力到运动变化较大的目标上，实现更高效的特征学习.

2.2.2　S-WAM

为了在预测中保留更多的高频空间细节，考虑小波变换的多分辨率分析特性，S-WAM 负责增强高频信息的表示. S-WAM 包含 2 个阶段：首先通过 DWT 将视频帧分解为 1 个低频子带特征图和 3 个高频子带特征图，3 个高频子带特征图分别对应水平、垂直和对角线 3 个方向；然后这些不同频率下的子带图通过一个浅层的 CNN 进一步提取特征，最终获得与主干网络结构通道数一致的特征图. 本文级联了 3 个 S-WAM 进行金字塔形式的小波分析，每级的输出分别与主干网络对应的特征图结合，在多频率下对主干网络进行高频细节补偿，提升对于精细细节的预测效果. S-WAM 的处理过程可以描述为空间小波分解过程和特征提取过程 2 个阶段. 空间小波分解过程表示为

$$\left[\boldsymbol{A}_t^i, \boldsymbol{H}_t^i, \boldsymbol{V}_t^i, \boldsymbol{D}_t^i\right] = f_{\mathrm{DWT_s}}\left(\boldsymbol{I}_t^i\right).$$

其中，$f_{\mathrm{DWT_s}}$ 表示空间离散小波变换；$\boldsymbol{I}_t^i$ 表示时刻 t 第 i 个 S-WAM 的输入；$\boldsymbol{A}_t^i$ 表示小波分解得到的低频子带特征图；$\boldsymbol{H}_t^i, \boldsymbol{V}_t^i, \boldsymbol{D}_t^i$ 分别表示时刻 t 第 i 个 S-WAM 的空间小波分解得到的 3 个方向(水平、垂直、对角线)的高频子带特征图. 每级输出子带图的维度都是输入的一半. 特征提取过程为

$$\left[\boldsymbol{S}_t^i, \boldsymbol{A}_t^k\right] = f_{\text{S-WAM}}^i\left(\left[\boldsymbol{A}_t^i, \boldsymbol{H}_t^i, \boldsymbol{V}_t^i, \boldsymbol{D}_t^i\right]\right).$$

其中，$f_{\text{S-WAM}}^i$ 表示第 i 个 S-WAM 之后的 CNN；$\boldsymbol{S}_t^i$ 表示时刻 t 第 i 个 S-WAM 的输出.

2.2.3　T-WAM

为了建模视频数据中的时序上的多种频率下的运动特征，T-WAM 将一段视频序列分解为在时间轴上不同频率下的子带特征图，沿时间轴的一级离散小波分解可以将一段视频序列分解为长度各为原来一半的低频子带特征图和高频子带特征图. 本文对视频序列采取 3 级金字塔形式的时序离散小波分解，最后将每级的分解结果拼接起来输入到一个浅层 CNN 中进行特征提取. T-WAM 的处理过程为

$$\boldsymbol{O}_{\mathrm{T}} = f_{\text{T-WAM}}\left(\boldsymbol{X}\right).$$

其中，$f_{\text{T-WAM}}$ 表示时序离散小波分析；$\boldsymbol{O}_{\mathrm{T}}$ 表示 T-WAM 的输出. 将提取到的特征图和主干网络的输出进行融合作为解码器的输入，增强模型对于不同频率下的运动特征的分辨能力.

2.2.4　解码器和判别器

解码器的作用是完成从编码器提取的视频信息到下一帧视频图像的映射，需要根据编码器得到的融合特征图输出预测. 解码器的过程表示为

$$\hat{\boldsymbol{x}}_{t+1} = f_{\mathrm{dec}}\left(\left[\boldsymbol{S}_t, \boldsymbol{O}_{\mathrm{T}}\right]\right).$$

其中，$\boldsymbol{S}_t$ 表示主干网络的输出，$\boldsymbol{O}_{\mathrm{T}}$ 表示 T-WAM 的输出，二者拼接起来作为解码器整体的输入；$\hat{\boldsymbol{x}}_{t+1}$ 是模型预测的下一视频帧.

本文沿用文献[13]中解码器和判别器的网络结构，采用对抗训练的方式提高预测结果的真实性.

3 实验及结果分析

3.1 数据集及评价指标

本文采用多个视频预测任务中广泛采用的数据集来评估模型的性能.

KTH 数据集[17]是包含 25 组人的 6 种不同动作的数据集，采用第 1~16 组数据进行模型训练，第 17~25 组数据进行测试. 训练时，模型根据观测的 10 帧视频序列预测接下来的 10 帧视频帧；测试时，扩展到预测未来的 20 帧或 40 帧.

KITTI 数据集[18]是目前最受欢迎的国际上最大的自动驾驶场景下的计算机视觉算法评测数据集，用于评测立体图像、光流、视觉测距、三维物体检测和三维跟踪等计算机视觉技术在车载环境下的性能，包含市区、乡村和高速公路等场景采集的真实图像数据，每幅图像中最多有 15 辆车和 30 个行人，是一个非常全面的评测数据集.

BAIR 数据集[19]由一个随机移动的机械臂组成，该机械臂在桌子上推动物体. 由于手臂运动的高随机性和背景的多样性，这个数据集特别具有挑战性.

本文采用峰值信噪比(peak signal-to-noise ratio, PSNR)和结构相似性(structure similarity, SSIM)这 2 个指标进行预测结果的质量评估，数值越高表示性能越好；同时采用 LPIPS(learned perceptual image patch similarity)衡量预测结果的真实性，数值越低表示真实性越高；还通过实验结果的可视化对比来展示主观评价效果.

3.2 模型参数及损失函数设计

本文采用的软件运行环境为 Ubuntu16.04，深度学习平台配置为 Python3.6 和 PyTorch0.4.1. 硬件配置为 NVIDIA GTX 1080 Ti，采用 CUDA9.0 和 cuDNN v7，优化器为 Adam Optimizer.

本文中模型的训练融合了多种模态的损失函数，包括像素级损失函数和对抗训练损失函数. 像素级损失函数包括损失函数及梯度差损失函数[3]；对抗训练损失包括生成器损失和判别器损失. 计算损失函数时，将多帧的预测损失进行叠加来更新损失函数，这种整体性判断的设计有助于提升模型预测的一致性，获得更好的预测效果.

3.3 定量评估

本文在 KTH 数据集上分别进行观测 10 帧预测未来 20 帧和 40 帧的实验，每个指标值均是取所有帧预测结果的平均值. 由于训练时模型只预测未来 10 帧，扩展到未来 20 帧和 40 帧的预测在一定程度上可以反映模型的泛化性能.

从表 1 可以看出，本文算法在各个评价指标上都取得了较好的结果，尤其在 PSNR 和 SSIM 这 2 个指标上表现出了明显的优势.

表 1 不同算法在 KTH 数据集结果比较

算法	10→20			10→40		
	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS
MCNET[20]	25.95	0.804		23.89	0.730	
fRNN[21]	26.12	0.771		23.77	0.678	
PredRNN[4]	27.55	0.839		24.16	0.703	
PredRNN++[5]	28.47	0.865		25.21	0.741	
VarNet[22]	28.48	0.843		25.37	0.739	
E3D[23]	29.31	0.879		27.24	0.810	
MSNET[24]	27.08	0.876				
SAVP[25]	25.38	0.746	9.37	23.97	0.701	13.26
SV2P[26]	27.79	0.838	15.04	26.12	0.789	22.48
STMF[13]	29.85	0.893	11.81	27.56	0.851	14.13
本文(w/TW)	29.13	0.870	14.33	26.42	0.815	16.06
本文(w/SW)	28.37	0.839	16.16	25.98	0.762	19.45
本文(w/DAM)	29.41	0.872	13.09	26.77	0.833	16.40
本文	30.23	0.895	11.54	27.93	0.862	14.09

通过消融实验对比本文提出的各模块单独带来的性能提升. 表 1 中，“本文(w/TW)”表示只保留 T-WAM 所得的实验结果，“本文(w/SW)”表示只保留 S-WAM 所得的实验结果，“本文(w/DAW)”表示只保留 DAM 所得的实验结果. 可以看出，DAM 的效果比单独引入 2 个小波分析模块带来了更大的改善，T-WAM 的引入比 S-WAM 的引入性能提升更高，而将 3 个模块融合获得了最佳的性能提升，这充分地验证了视频数据比图像数据有更复杂的时空结构. 根据历史时空关系推断后续的时空关系，对时序特征的提取十分关键，本文应用的 DAM 和 T-WAM 分别关注视频的瞬时运动特征和过程性运动特征，为模型提供了丰富的时序信息，有助于模型做出更好的预测. 而 S-WAM 主要关注视频的空间细节信息，负责为模型提供更丰富的高频信息，使模型的预测更清晰.

表 2 和表 3 所示分别为不同算法在 KITTI 和 Cityscapes 数据集上进行的预测未来 1 帧、5 帧和 10 帧的实验对比，这 2 个数据集都是自动驾驶相关的数据集. 可以看出，本文算法在所有的实验设置下都达到了目前最先进的水平，尤其是对于更

长时的预测的优势，表明本文算法在时序特征提取方面的能力；还可以看出，随着预测时间的延长，模型的预测难度逐渐变大，造成这种现象的原因是视频预测中后面的视频帧预测是基于前面的视频帧预测结果而做出的，所以前面的预测误差会逐渐累积到后面的预测结果上，造成预测效果随时间逐渐变差，因此，减小时序上的预测误差对于做出更长时且准确的预测至关重要. 同样，本文在 KITTI 和 Cityscapes 数据集上进行了消融实验，结果与 KTH 数据集上的实验结果一致，融合 3 个模块带来最好的性能提升，DAM 效果比单独引入小波分析模块更好.

表 2　不同算法在 KITTI 数据集结果比较

算法	预测 1 帧			预测 5 帧			预测 10 帧		
	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS	PSNR/dB	SSIM	LPIPS
Copy-Last	15.6	0.472	65.7	9.4	0.250	70.9	6.1	0.210	89.3
DVF[27]	21.2	0.529	32.4	13.9	0.426	41.5			
PredNet[3]		0.562	55.3		0.475	62.9			
MCNET[20]	16.7	0.753	24.0	13.2	0.554	37.3	11.4	0.485	76.4
CtrlGen[28]	19.9	0.771		16.2	0.532		14.8	0.451	
VarNet[22]	21.8	0.764	35.4	16.9	0.574	35.9	14.7	0.526	61.1
STMF[13]	22.1	0.790	27.9	17.1	0.587	30.0	15.3	0.510	50.0
OMP[29]		0.792	18.4		0.607	30.4			
本文(w/TW)	21.5	0.749	23.9	16.4	0.605	30.9	14.8	0.528	53.7
本文(w/SW)	21.1	0.723	24.6	16.1	0.601	28.7	14.9	0.511	55.9
本文(w/DAM)	22.0	0.789	22.8	16.9	0.619	30.5	15.6	0.525	55.2
本文	22.3	0.795	21.2	17.4	0.622	30.2	15.6	0.541	54.3

表 3　不同算法在 Cityscapes 数据集结果比较

算法	预测 1 帧		预测 5 帧		预测 10 帧	
	SSIM	LPIPS	SSIM	LPIPS	SSIM	LPIPS
PredNet[3]	0.840	25.9	0.753	36.0	0.663	52.2
MCNET[20]	0.896	18.8	0.705	37.3	0.597	45.1
DVF[27]	0.838	17.3	0.711	28.7	0.634	36.5
Vidvid[30]	0.881	10.5	0.751	20.1	0.669	27.0
OMP-WC[29]	0.879	9.0	0.743	17.1	0.659	24.1
OMP-WM[29]	0.886	8.9	0.753	16.9	0.672	23.5
STMF[13]	0.891	8.5	0.756	16.5	0.674	23.2
本文(w/TW)	0.882	12.1	0.749	17.9	0.665	22.9
本文(w/SW)	0.875	12.2	0.738	19.2	0.659	25.4
本文(w/DAM)	0.893	10.9	0.754	17.6	0.677	23.3
本文	0.893	8.5	0.760	16.1	0.681	21.9

表 4 所示为 BAIR 数据集上的实验结果对比. BAIR 数据集有很高的随机性，在保证良好预测效果的同时，取得与真实值较高的时空一致性有非常大的挑战. 可以看出，本文算法在 PSNR 和 SSIM 这 2 个评价指标上都取得了最好的结果，说明该算法可以产生与真实样本更接近的预测结果；本文算法在 LPIPS 指标上也取得了与现有算法同等水平的预测效果，虽然基于随机变量的模型(如 SAVP, SV2P 等)在 LPIPS 上的表现更好一些.

为了进一步验证模型的泛化能力，本文将在 KITTI 数据集上训练得到的模型在 Caltech Pedestrian 数据集上进行测试. 实验设置与对比的模型相同，训练和测试时都是观测 10 帧，测试时采用的是预测未来 1 帧的数据指标进行对比. 实验结果如表 5 所示，可以看出，本文算法的泛化能力达到了最先进的水平.

表 4　不同算法在 BAIR 数据集结果比较

算法	PSNR/dB	SSIM	LPIPS
SAVP[25]	18.42	0.789	6.34
SAVP-V[25]	19.09	0.815	6.22
SV2P-I[26]	20.36	0.817	9.14
SVG-LP[31]	17.72	0.815	6.03
I-VRNN[32]		0.822	5.50
STMF[13]	21.02	0.844	9.36
本文	21.25	0.849	8.11

本文从时序运动的表达能力和空间细节的清晰度 2 个方面，通过对预测结果进行可视化来对模型进行定性的评估. 图 2 所示为在 KTH 数据集上进行的未来 40 帧的预测结果. 由于版面有限，每隔 1 帧输出 1 次，一共进行了 2 组不同行为(拳击及跑步)的实验，每组的第 1 行是真实图像序列，第 2 行是本文算法的预测结果，其余 2 行是其他算法的预测结果. 可以看出，本文算法的预测结果相比其他算法保留了更多的细节，也保持了更好的

表 5　不同算法在 Caltech Pedestrian 数据集结果比较

算法	PSNR/dB	SSIM	LPIPS
MCNET[20]		0.879	
PredNet[3]	27.6	0.905	7.47
ContextVP[33]	28.7	0.921	6.03
DVF[27]	26.2	0.897	5.57
VarNet[22]	27.9	0.912	
DPG[34]	28.2	0.923	5.04
C-GAN[35]	29.2	0.830	
PreCNet[36]	28.5	0.929	
CrevNet[37]	29.3	0.925	
STMF[13]	29.1	0.927	5.89
本文	29.2	0.931	5.92

运动时序一致性，如第 1 组中的胳膊和身体以及第 2 组中的身体和腿，都预测得更清楚；另外，2 组动作是不同的速度和形态下的过程化运动，本文算法的预测结果与真实值有更高的一致性，其他方法的预测结果中人过早跑出了画面，表明本文算法可以更好地处理各种不同的动态运动，具有较好的运动特征提取能力.

目前已有的神经网络模型中常常采用注意力机制使其具备专注于其输入(或特征)子集的能力，本文将注意力机制与帧间差算法相结合，将相邻帧之间的差异编码成注意力向量，用于指导模型对于特征的学习. 为了更直观地理解 DAM 在预测过程中所提取的特征，本文在 KTH 数据集上对注意力图进行了热力图形式的可视化. 从图 3 可以看出，注意力图可以充分反映画面中瞬时运动比较明显的区域，从而更好地指导模型预测时更合理地分配注意力，以获取所需关注的目标更丰富的信息，给予运动幅度较大的区域更高的关注，提升预测的效果.

图 4 所示为在 KITTI 数据集上预测未来 10 帧的可视化预测结果对比. 可以看出，本文算法的预测结果更为清晰，保留了较多的时序一致性特征，尤其是在车道线和标识牌的预测上效果更好. 因

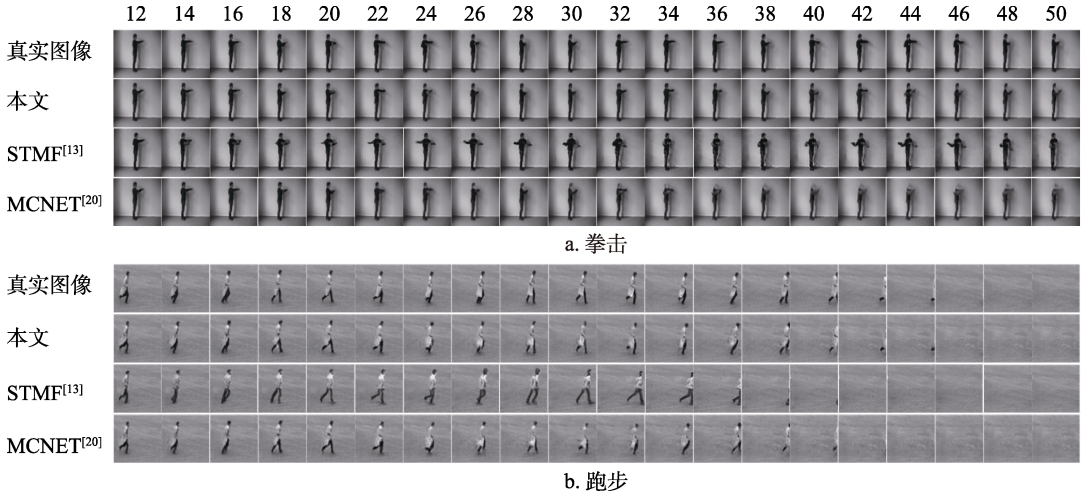

图 2　不同算法对 KTH 数据集行为预测结果可视化

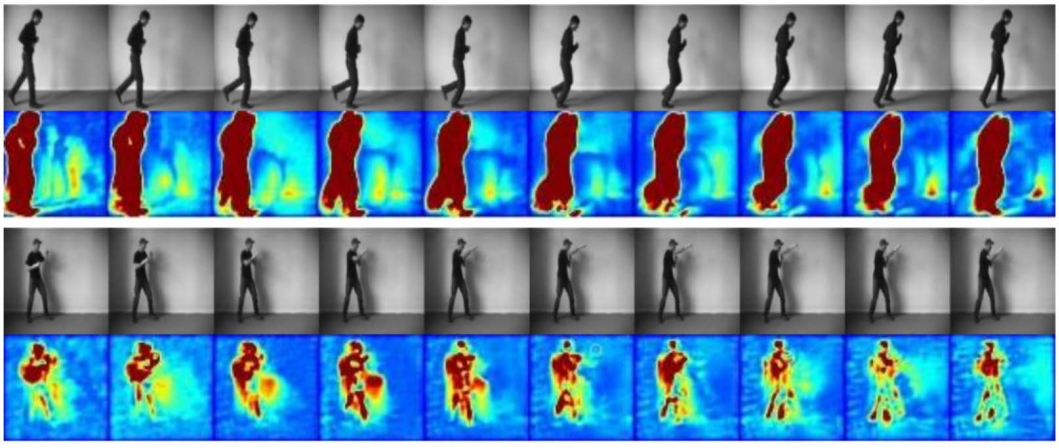

图 3　本文算法在 KTH 数据集上的注意力图可视化

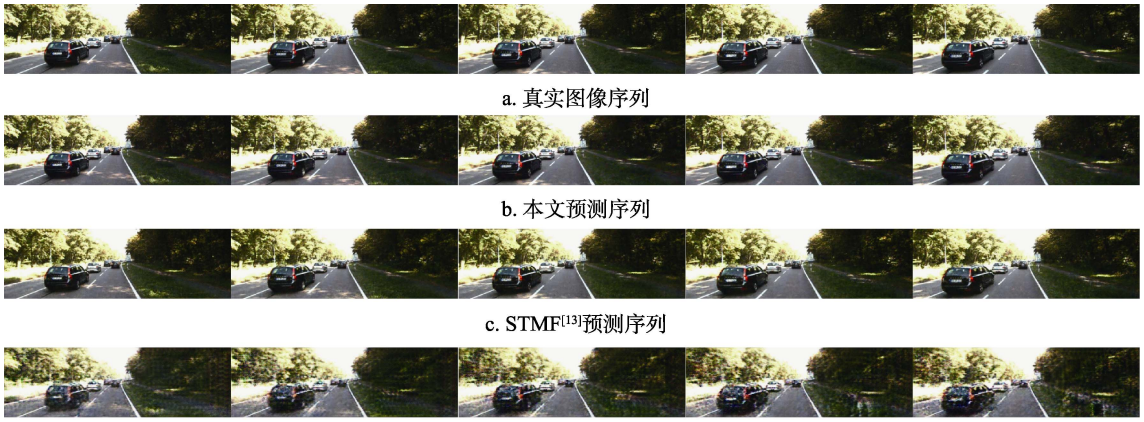

图 4　不同算法对 KITTI 数据集预测结果可视化

此，本文算法对于自动驾驶环境的预测训练可以提升模型对于环境的特征提取能力，有助于做出更安全可靠的驾驶决策.

4　结　语

针对当前视频预测模型中存在的细节缺失和多尺度时序运动预测不一致的问题，受人类视觉系统中处理视觉信息机制的启发，本文提出一种基于注意力机制的融合时空小波分析的视频预测算法，在文献[23]的基础上提出 DAM，将相邻帧之间的帧间变化信息编码为注意力向量来指导模型对瞬时运动的学习，提升文献[23]对于高价值信息的筛选能力，加强了模型对于环境的理解能力. 将改进后的模型扩展到更多的数据集上进行实验的结果表明，与最先进的方法相比，本文算法表现出了更优越的性能，在细节预测和时序一致性方面取得了更好的效果. 下一步会将本文算法与其他实际任务相结合，以展现模型的实际应用价值.

参考文献(References):

[1] Finn C, Levine S. Deep visual foresight for planning robot motion[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2017: 2786-2793

[2] Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error[OL]. [2021-01-19]. https://arxiv.org/abs/1511.05440

[3] Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning[OL]. [2021-01-19]. https://arxiv.org/abs/1605.08104

[4] Wang Y B, Long M S, Wang J M, *et al*. PredRNN: recurrent neural networks for predictive learning using spatial temporal LSTMs[C] //Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 879-888

[5] Wang Y B, Gao Z F, Long M S, *et al*. PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning[C] //Proceedings of the International Conference on Machine Learning. Lille: PMLR Press, 2018: 5123-5132

[6] Liang X D, Lee L, Dai W, *et al*. Dual motion GAN for future-flow embedded video prediction[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 1762-1770

[7] Oh J, Guo X X, Lee H, *et al*. Action conditional video prediction using deep networks in Atari games[C] //Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2863-2871

[8] Finn C, Goodfellow I, Levine S. Unsupervised learning for physical interaction through video prediction[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 64-72

[9] Chen B Y, Wang W M, Wang J Z. Video imagination from a single image with transformation generation[OL]. [2021-01-19]. https://arxiv.org/abs/1706.04124

[10] Vondrick C, Torralba A. Generating the future with adversarial transformers[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2992-3000

[11] van Amersfoort J, Kannan A, Ranzato M, *et al*. Transformation-based models of video sequences[OL]. [2021-01-19]. https://arxiv.org/abs/1701.08435

[12] Reda F A, Liu G L, Shih K J, *et al*. SDC-Net: video prediction using spatially-displaced convolution[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 747-763

[13] 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] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 4553-4562

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

[15] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative Adversarial Networks[J]. Communications of the ACM, 2020, 63(11): 139-144

[16] Wang X T, Yu K, Wu S X, *et al*. ESRGAN: enhanced super-resolution generative adversarial networks[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 63-79

[17] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C] //Proceedings of the 17th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2004: 32-36

[18] Geiger A, Lenz P, Stiller C, *et al*. Vision meets robotics: the KITTI dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237

[19] Ebert F, Finn C, Lee A X, *et al*. Self-supervised visual planning with temporal skip connections[OL]. [2021-01-19]. https://arxiv.org/abs/1710.05268

[20] Villegas R, Yang J M, Hong S, *et al*. Decomposing motion and content for natural video sequence prediction[OL]. [2021-01-19]. https://arxiv.org/abs/1706.08033

[21] Oliu M, Selva J, Escalera S. Folded recurrent neural networks for future video prediction[OL]. [2021-01-19]. https://arxiv.org/abs/1712.00311

[22] Jin B B, Hu Y, Zeng Y M, *et al*. VarNet: exploring variations for unsupervised video prediction[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2018: 5801-5806

[23] Wang Y B, Jiang L, Yang M, *et al*. Eidetic 3D LSTM: a model for video prediction and beyond[OL]. [2021-01-19]. https://openreview.net/pdf?id=B1lKS2AqtX

[24] Lee J, Lee J, Lee S, *et al*. Mutual suppression network for video prediction using disentangled features[OL]. [2021-01-19]. https://arxiv.org/abs/1804.04810

[25] Lee A X, Zhang R, Ebert F, *et al*. Stochastic adversarial video prediction[OL]. [2021-01-19]. https://arxiv.org/abs/1804.01523

[26] Babaeizadeh M, Finn C, Erhan D, *et al*. Stochastic variational video prediction[OL]. [2021-01-19]. https://arxiv.org/abs/1710.11252

[27] Liu Z W, Yeh R A, Tang X O, *et al*. Video frame synthesis using deep voxel flow[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4473-4481

[28] Hao Z K, Huang X, Belongie S. Controllable video generation with sparse trajectories[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7854-7863

[29] Wu Y, Gao R R, Park J, *et al*. Future video synthesis with object motion prediction[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 5538-5547

[30] Wang T C, Liu M Y, Zhu J Y, *et al*. Video-to-Video synthesis[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2018: 1152-1164

[31] Denton E, Fergus R. Stochastic video generation with a learned prior[C] //Proceedings of the International Conference on Machine Learning. Lille: PMLR Press, 2018: 1174-1183

[32] Castrejon L, Ballas N, Courville A. Improved conditional VRNNs for video prediction[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7607-7616

[33] Byeon W, Wang Q, Srivastava R K, *et al*. ContextVP: fully context-aware video prediction[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 781-797

[34] Gao H, Xu H Z, Cai Q Z, *et al*. Disentangling propagation and generation for video prediction[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9005-9014

[35] Kwon Y H, Park M G. Predicting future frames using retrospective cycle GAN[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1811-1820

[36] Straka Z, Svoboda T, Hoffmann M. PreCNet: next frame video prediction based on predictive coding[OL]. [2021-01-19]. https://arxiv.org/abs/2004.14878

[37] Yu W, Lu Y C, Easterbrook S, *et al*. Efficient and information-preserving future frame prediction and beyond[OL]. [2021-01-19]. https://openreview.net/forum?id=B1eY_pVYvB