DOI:10.11918/202503001

多尺度特征建模的图像时间序列预测网络

沈 瑜[1],马煜堃[1],赵永刚[2],魏子易[1],李江柽[1],王若暄[1],刘佳英[1],闫佳荣[1]

(1.兰州交通大学 电子与信息工程学院,兰州 730070;2.兰州萃英信息科技有限公司,兰州 730070)

摘 要:为提高图像时间序列预测的精度,本研究提出了一种基于长短期记忆网络(long short-term memory,LSTM)与注意力机制的时间序列预测网络:MA-LSTM。该网络整体由多尺度注意力模块(multi-scale attention block,MAB)、多尺度注意力层(multi-scale attention layer,MALayer)和超分辨率重建模块(super resolution reconstruction module,SRRM)组成,以多尺度特征建模为核心,着重提升时空特征表达能力与长程依赖建模能力。首先,MA-LSTM 设计了 MAB 模块,通过时空特征增强层提升模型的细节建模能力,并利用通道特征增强层加强了特征图的跨维度信息交互,解决了 SwinLSTM 对于细粒度特征捕捉不足的问题。其次,MA-LSTM 引入了简化的 LSTM 结构,与 MAB 结合构建了 MALayer,增强模型对时序信息的建模能力。最后,在特征图重建时设计了 SRRM 模块,有效增强模型预测输出的细节表达能力。研究表明,MA-LSTM 在 MovingMNIST 和 KTH 两个不同领域的数据集上,结构相似性指数分别达到 0.960 2 和 0.924 3,与 SwinLSTM、PhyDNet、PredRNN、ConvLSTM 网络进行的对比试验结果表明,结构相似性指数最高提升了 0.337 和 0.212,展现了其在时序预测任务中的高效性和适用性,且具备跨领域的推广潜力。此外,消融实验进一步证明了本文所提出模块的有效性。

关键词:图像时间序列;预测网络;LSTM;移位窗口注意力;多注意力融合

中图分类号:TP183 **文献标志码**:A **文章编号**:0367-6234(2026)01-0119-12

Multi-scale feature modeling for image time-series prediction network

SHEN Yu[1], MA Yukun[1], ZHAO Yonggang[2], WEI Ziyi[1], LI Jiangcheng[1], WANG Ruoxuan[1], LIU Jiaying[1], YAN Jiarong[1]

(1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China; 2. Lanzhou Trying Information Technology Co., Ltd., Lanzhou 730070, China)

Abstract: To improve the accuracy of image time-series prediction, a time-series prediction network of MA-LSTM is proposed based on LSTM and attention mechanism. This model is consist of multi-scale attention module (MAB), multi-scale attention layer (MALayer) and super-resolution reconstruction module (SRRM), it could improve the express spatiotemporal features and long-range dependencies. Firstly, MAB module is designed, and detail modeling is improved through the spatiotemporal feature enhancement layer (GSTA), then the channel feature enhancement layer (GCA), overcoming SwinLSTM′s insufficient capture of fine-grained features, is used to enhance the cross-dimensional information interactions of the feature map. Secondly, a simplified LSTM structure is employed, and MALayer is constructed in combination with MAB to improve modeling of time series information. Finally, the SRRM module is designed during feature map reconstruction to improve the prediction output. Experimental results show that MA-LSTM achieves a structural similarity index(SSIM) of 0.960 2 and 0.924 3 on two datasets in different fields: MovingMNIST and KTH. Compared with SwinLSTM, PhyDNET, PredRNN, and ConvLSTM networks, the highest accuracy improvement of 0.337 and 0.212, respectively. This model demonstrates the higher efficiency and applicability in time series prediction tasks and the well potential for cross-domain promotion, and the ablation experiments also show the effectiveness of the proposed module.

Keywords: image time series data; prediction network; LSTM; shifted window attention; multi-attention fusion

时序预测在降雨量预测、天气变化预测、车流量预测、经济学等多个领域应用广泛[1];然而时序数据通常情况下具有高度的非线性、复杂的时间依赖性和干扰数据,传统方法难以有效建模其时空演化

收稿日期:2025-03-01;**录用日期**:2025-04-28;**网络首发日期**:2025-08-29
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20250829.0905.002
基金项目:国家自然科学基金青年科学基金A类(42325502);甘肃省重点研发计划(甘科计[2024]10号-24YFGA037);国家自然科学基金(62241106,61861025);甘肃省科技专员专项(甘科计[2023]18号-23CXGA0008);“智慧天路”建设重大专项-QZzhtlzx(2023QZzhtl1102);兰州局集团公司科技研究开发计划LZJKY2024079-1;中国国家铁路集团有限公司重点课题(N2023X050);兰州交通大学重点研发项目(LZJTU-ZDYF2305)
作者简介:沈 瑜(1982—),女,教授,博士生导师
通信作者:马煜堃,3379038113@qq.com

规律,为时序数据预测带来了巨大的挑战,近年来使用数据驱动的深度学习展现出强大的建模能力,使得更加精确的时序预测成为可能[2]。

现有预测方法主要基于卷积神经网络[3](convolution neural networks, CNNs)、多层感知机[4](multilayer perception, MLP)与循环神经网络[5](recurrent neural networks, RNNs)展开。1990 年,循环神经网络首次成为处理序列数据的重要工具。1997 年, Hochreiter 等[6]提出了(long short-term memory,LSTM)网络,显著提升了在时间序列预测和语音识别等任务中的性能表现。2015 年,Shi 等[7]提出了 ConvLSTM,通过引入卷积结构增强了模型建模能力。2017 年,Wang 等[8]提出了 PredRNN,以及 2020 年,Li 等[9]提出了 MIM,Guen 等[10]提出了 PhyDNet,在捕捉复杂时空依赖关系方面进一步提升了性能,但总体仍未脱离卷积神经网络,卷积操作本身聚焦于捕捉局部特征与联系[11],基于 CNN 的网络缺乏对于全局特征的高效捕捉,限制了其在时序预测任务中的表现。

在基于 CNNs 网络发展的同时,2017 年 Vaswani 等[12]提出了用于自然语言处理任务的 Transformer 架构,其采用全新的自注意力机制,实现了对数据的全局建模。2020 年,Google Research 团队提出了 Vision Transformer 结构,将 Transformer 架构成功应用于视觉任务中[13]。2021 年,Microsoft Research Asia 团队提出了 SwinTransformer[14]结构,引入了滑动窗口机制,在大幅度降低计算开销的同时提升了特征提取的效率与精度。2023 年,Tang 等[15]提出了 SwinLSTM 结构,通过集成 SwinTransformer 的移动窗口注意力机制代替卷积结构至 LSTM 中,实现了更优秀的时间序列图片预测能力,然而移位窗口注意力机制更关注对数据的全局建模能力,导致其对于细粒度信息的捕捉存在一定的局限,在小目标特征表达上易出现信息丢失的问题[16]。

为了解决现有方法在时序预测中的局限性,本文提出了一种基于 LSTM 结构,融合了注意力与卷积的图像时间序列预测网络——MA-LSTM。该模型设计了用于高效提取时间信息、空间信息的多尺度注意力模块(MAB)、用于增强特征图重建时特征表达能力的超分辨率重建模块(SRRM),经过简化的 LSTM 与 MAB 模块形成纵横交错的特征提取机制,能够充分提取不同尺度的时空特征。MA-LSTM 着重提升了时间建模、空间建模以及长程依赖的建模能力。

1 MA-LSTM 模型框架

本文所提出的 MA-LSTM 模型,是一种基于简化 LSTM 结构和 SwinTransformer 改进的图像时间序列预测网络,集成了多重卷积注意力机制,主要由多尺度注意力层(multi-scale attention layer, MALayer)模块和超分辨率增强重建模块构成,MALayer 内部包含了简化的 LSTM 网络以及多个多尺度注意力模块(multi-scale attention block, MAB),网络的总体结构图如图 1 所示。

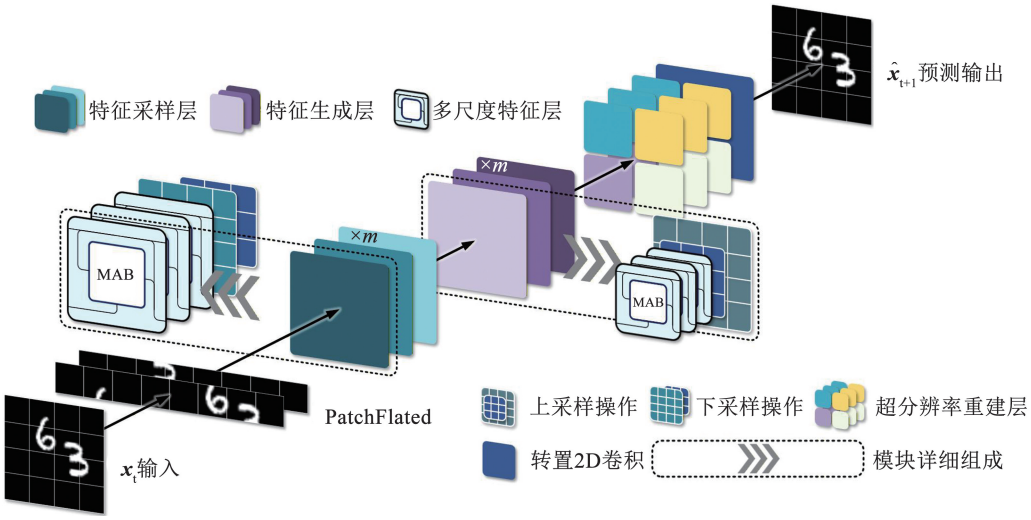


图 1 MA-LSTM 网络总体结构图

Fig. 1 Overall structure of MA-LSTM network

MA-LSTM 首先会将输入划分为互不重叠的补丁块,经过 PatchFlated 层后,送入特征采样层。每一个特征采样层由多个 MALayer 和一个下采样模块组成,其中 MALayer 负责提取多尺度特征,下采样模块则降低数据维度以减小计算压力,同时帮助 MALayer 在不同尺度提取特征。随后数据通过特征生成层逐步恢复维度,特征生成层的结构对称地包含了数个 MALayer 和一个上采样模块,与特征采样

层总体形成了 U 型结构。经过特征生成层后，特征图进入超分辨率重建模块，进一步增强特征信息并还原特征图，最终得到预测输出。

特征生成层和特征采样层分别接受来自前序状态对应层级的隐藏状态信息 $\boldsymbol{H}_t$ 和细胞状态信息 $\boldsymbol{C}_t$，在每个特征生成层和特征采样层中有数个 MALayer，MA-LSTM 的网络总体计算方法如下：

$$\begin{aligned}\boldsymbol{F}^{(0)} &= f_{\text{Flatten}}(\boldsymbol{x}_t) \in \mathbb{R}^{N\times P^2}\\ \boldsymbol{F}^{(m)} &= \prod_{n=1}^{m} f_{\text{Down}}\Big(\prod_{k=0}^{i-1} f_{\text{MALayer}}(\boldsymbol{F}^{(k)})\Big)\\ \boldsymbol{F}^{(2m)} &= \prod_{n=1}^{m} f_{\text{Up}}\Big(\prod_{k=0}^{j-1} f_{\text{MALayer}}(\boldsymbol{F}^{(m+k)})\Big)\\ \hat{\boldsymbol{x}}_{t+1} &= f_{\text{SRRM}}(\boldsymbol{F}^{(2m)})\end{aligned} \tag{1}$$

式中：$\boldsymbol{F}^{(i)}$ 为 MA-LSTM 网络中第 i 层输出的特征图，N 为 patch 的数量，P 为每个 patch 展开后的维度，f_{Flatten} 为特征图展平操作，f_{Down} 与 f_{Up} 分别为下采样与上采样，f_{MALayer} 为 MALayer 模块，f_{SRRM} 为特征图重建操作。

时序预测任务对模型的长程依赖能力提出要求，MA-LSTM 简化了传统 LSTM 门结构以匹配引入的 Transformer 结构，修改的门结构简化了数据流动阻力，能够更高效地捕捉数据长程依赖关系。

全局信息的建模能力对于时序预测网络的精准预测同样重要，针对常用 CNN 时序预测网络全局建模能力较弱的问题，MA-LSTM 引入了 SwinTransformer 的移位窗口注意力机制，能够从全局进行特征建模，更好地处理时序数据。同时为了解决 SwinTransformer 在细粒度细节捕捉上的不足，MALayer 集成的多尺度注意力模块，在集成了 SwinTransformer 的基础上，引入了时空特征增强层（global spatial-temporal attention，GSTA）与通道特征增强层（global channel attention，GCA），两种机制融合了通道注意力、空间注意力与卷积，能够有效补充 SwinTransformer 小细节捕捉能力缺失的问题，同时加强了特征图跨维度交互能力，提升了对复杂场景的理解能力，提高模型的整体性能。

此外，MA-LSTM 在特征生成层后额外设计了一个超分辨率重建模块（super resolution reconstruction module，SRRM）用于增强特征图的特征表达，SRRM 通过多尺度特征金字塔结构[17]强化特征信息表示，使得模型在重建输出图像时拥有更优秀的精确度表现。

1.1　多尺度注意力层（MALayer）

MALayer 的主要组成部分是简化的 LSTM 结构与 MAB 模块。二者分别负责横向的长短期记忆捕捉和纵向的全局特征提取，最终形成高效的时空交互机制，MALayer 的网络结构如图 2 所示。

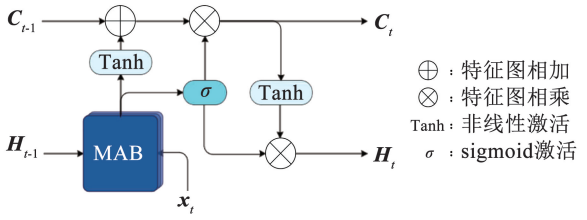


图 2　MALayer 网络结构图

Fig. 2　Structure of MALayer

以 ConvLSTM 的计算公式为例简化门控结构，ConvLSTM 的关键方程为：

$$\begin{aligned}\boldsymbol{I}_t &= \sigma(\boldsymbol{W}_{\text{xi}} * \boldsymbol{x}_t + \boldsymbol{W}_{\text{hi}} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_i)\\ \boldsymbol{F}_t &= \sigma(\boldsymbol{W}_{\text{xf}} * \boldsymbol{x}_t + \boldsymbol{W}_{\text{hf}} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_f)\\ \boldsymbol{O}_t &= \sigma(\boldsymbol{W}_{\text{xo}} * \boldsymbol{x}_t + \boldsymbol{W}_{\text{ho}} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_o)\\ \boldsymbol{C}_t &= \boldsymbol{F}_t \cdot \boldsymbol{C}_{t-1} + \boldsymbol{I}_t \cdot \tanh(\boldsymbol{W}_{\text{xc}} * \boldsymbol{x}_t + \boldsymbol{W}_{\text{hc}} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_c)\\ \boldsymbol{H}_t &= \boldsymbol{O}_t \cdot \tanh(\boldsymbol{C}_t)\end{aligned} \tag{2}$$

式中：σ 为 sigmoid 激活函数，“tanh”为非线性激活函数，$\boldsymbol{I}_t$ 为输入门输出，$\boldsymbol{F}_t$ 为遗忘门输出，$\boldsymbol{O}_t$ 为输出门输出，$\boldsymbol{H}_t$ 为隐藏状态，$\boldsymbol{C}_t$ 为细胞状态，“·”为特征图点对点相乘，“$*$”为卷积操作。每个门控结构都由两组完全不同的权重 $\boldsymbol{W}_{\text{xx}}$ 与偏置 $\boldsymbol{b}$ 与之对应，不同的权重和偏置决定了门控的不同用途，但是结构本身却完全相同。

MALayer 通过注意力直接捕捉并计算全局特征，从而消去所有权重，输入门、遗忘门与输出门的公式完全一致，输出并为 $\boldsymbol{F}_t$，细胞状态和隐藏状态的计算公式也相应得到更新，在 MALayer 中，MAB 模块直接嵌套在 $\boldsymbol{F}_t$ 计算式中，简化后的计算公式如下：

$$\begin{aligned}\boldsymbol{F}_t &= \sigma(f_{\text{MAB}}(\boldsymbol{x}_t, \boldsymbol{H}_{t-1}))\\ \boldsymbol{C}_t &= (\tanh(f_{\text{MAB}}(\boldsymbol{x}_t, \boldsymbol{H}_{t-1})) + \boldsymbol{C}_{t-1}) \cdot \boldsymbol{F}_t\\ \boldsymbol{H}_t &= \boldsymbol{F}_t \cdot \tanh(\boldsymbol{C}_t)\end{aligned} \tag{3}$$

式中 f_{MAB} 为多尺度注意力模块。

1.2　多尺度注意力模块（MAB）

SwinLSTM 通过将移位窗口注意力结构引入 LSTM 实现了对特征的全局建模，但窗口注意力机制对于细粒度细节捕捉以及跨通道信息建模存在不足。MAB 针对上述问题做出了改进，引入了两种通道空间复合的注意力机制，即时空特征增强层和通道特征增强层，MAB 的整体结构见见图 3。

MAB 在总体结构上呈现顺序结构，在首层 MAB 接受当前时间步 $\boldsymbol{x}_t$ 的特征图后，与前序时间步的 $\boldsymbol{H}_{t-1}$ 逐元素相加，通过移位窗口注意力机制，初次进行全局特征提取，随后与 $\boldsymbol{x}_t$ 做跳跃连接后输出。第 2 层特征图直接通过全局时空特征注意力机制进行更细节的时空特征建模，随后输出与 $\boldsymbol{x}_t$ 做跳跃连接，最后经过 MLP 层输出。第 3 层特征图与 $\boldsymbol{H}_{t-1}$ 输入进行逐元素相加后，通过全局通道特征注意力机制进行跨维度特征建模提取，经过 MLP 以后加权输出。

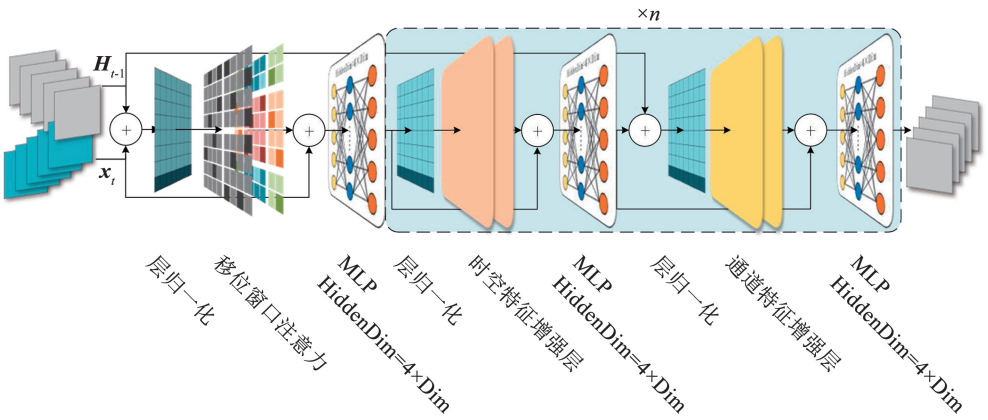


图 3　MAB 网络结构图

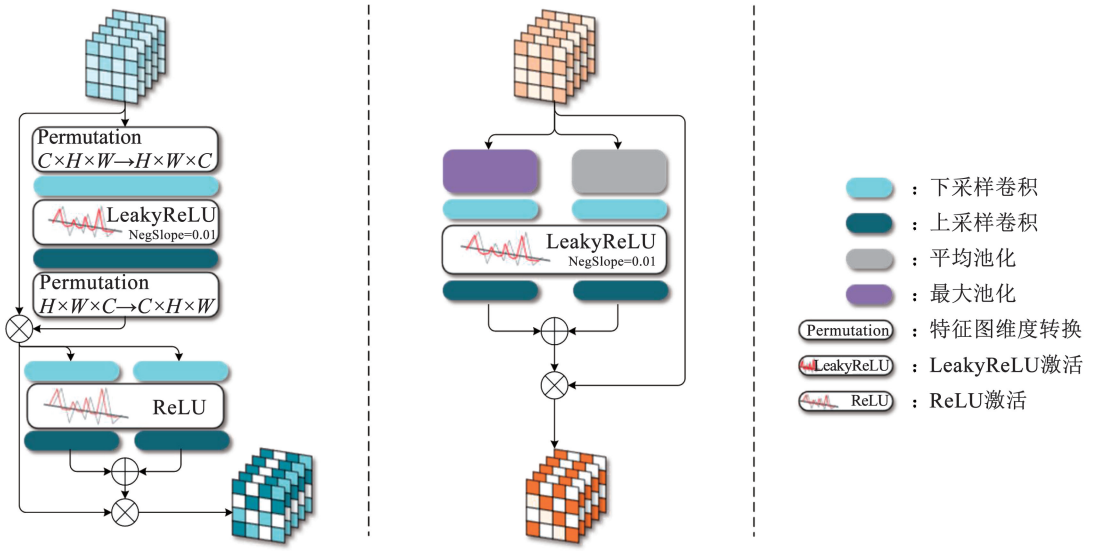
Fig. 3　Structure of the MAB

如果需要更深层次的网络,MAB 模块会根据所需网络深度层数的奇偶性进行堆叠,多层堆叠的 MAB 由式(4)表示。

$$\boldsymbol{F}^{(l)}=\begin{cases}f_{\mathrm{SWA}}(\boldsymbol{x}_t+\boldsymbol{H}_{t-1})+\boldsymbol{x}_t, & l=1\\ f_{\mathrm{GSTA}}(\boldsymbol{F}^{(l-1)})+\boldsymbol{F}^{(l-1)}, & l=2i\\ f_{\mathrm{GCA}}(\boldsymbol{F}^{(l-1)}+\boldsymbol{H}_{t-1})+\boldsymbol{F}^{(l-1)}, & l=2i+1\end{cases} \tag{4}$$

式中:$\boldsymbol{F}^{(l)}$ 为MAB在第l层的输出,f_{SWA} 为移位窗口注意力机制,f_{GSTA} 为全局时空特征注意力机制,f_{GCA} 为全局通道特征注意力机制,$i\in(0,+\infty)$用于选择在不同深度下经过的模块。

图 4 展示了 MAB 中时空特征增强层(global spatial-temporal attention,GSTA)与通道特征增强层(global channel attention,GCA)的结构设计。GSTA 进一步放大全局特征的交互效应,并保留细粒度的特征细节;GCA 通过两种不同的池化方式结合以增强每个通道上的特征表达。



(a) GSTA模块结构图　　(b) GCA模块结构图

图 4　模块结构图

Fig. 4　Module structure diagram

在 GSTA 中,所有输入特征首先经过跨维度特征层,提取出的特征与原输入逐元素相乘;后经过时空特征层,输出特征再次与主元素相乘,最后输出增强特征图。跨维度特征层中包含一个降维卷积,将特征图的维度压缩为 c/r(r 为缩减倍率,在本文中 $r=4$),而后通过 LeakyReLU 激活函数引入非线性变换,最后通过升维卷积还原特征图并输出维度增强特征图。LeakyReLU 相较于 ReLU 可有效避免“神经元死亡”,通过保留负值梯度提升模型的鲁棒性。维度注意力模块的计算过程如下:

$$\begin{aligned}&\boldsymbol{F}^{(l-1)}\in\mathbb{R}^{C\times H\times W}\xrightarrow{\text{permutation}}\boldsymbol{F}'^{(l-1)}\in\mathbb{R}^{H\times W\times C}\\&\boldsymbol{F}_1=\boldsymbol{W}_{3\times3}^{\mathrm{Up}}*(f_{\mathrm{LeakyReLU}}(\boldsymbol{W}_{3\times3}^{\mathrm{Down}}*\boldsymbol{F}'^{(l-1)}))\\&\boldsymbol{F}_1\in\mathbb{R}^{H\times W\times C}\xrightarrow{\text{permutation}}\boldsymbol{F}'_1\in\mathbb{R}^{C\times H\times W}\\&\boldsymbol{F}_{\mathrm{dim}}=\boldsymbol{F}^{(l-1)}\odot\boldsymbol{F}'_1\end{aligned} \tag{5}$$

式中：W 为降维升维时卷积的权重，上标注明了权重所属卷积的功能，C、H、W 分别表示特征图的通道、高、宽的尺寸。

时空特征层中包含一个浅层的特征 U 型池，输入数据会分别经过卷积核为 7×7 和 3×3 的两个降维卷积，经过卷积后特征图维度压缩为 c/r，经过 ReLU 函数激活后再次通过对应卷积核一致的两个反向卷积，sigmoid 归一化后与主元素相乘，最终输出同时施加维度注意力和时空注意力的特征图。时空特征层的计算公式为：

$$\boldsymbol{F}_1 = f_{\text{LeakyReLU}}(\boldsymbol{W}_{7\times7}^{\text{Down}} * \boldsymbol{F}_{\text{dim}}), \boldsymbol{F}_2 = f_{\text{LeakyReLU}}(\boldsymbol{W}_{3\times3}^{\text{Down}} * \boldsymbol{F}_{\text{dim}})$$
$$\boldsymbol{F}_{\text{spatial}} = \boldsymbol{W}_{7\times7}^{\text{Up}} * \boldsymbol{F}_1 + \boldsymbol{W}_{3\times3}^{\text{Up}} * \boldsymbol{F}_2$$
$$\boldsymbol{F}^{(l)} = \boldsymbol{F}_{\text{dim}} \odot \boldsymbol{F}_{\text{spatial}} \tag{6}$$

式中 $\boldsymbol{F}^{(l)}$ 为 GSTA 的最终输出，$\boldsymbol{W}$ 为卷积层的权重，其中下标表明该权重所属卷积的卷积核尺寸，上标表明该权重所属卷积的功能，$f_{\text{LeakyReLU}}$ 为 LeakyReLU 激活函数。

不同尺度的卷积层能够有效提取多尺度的特征，有效弥补移动窗口注意力机制对于细粒度细节捕捉不足和窗口间信息交互受阻的问题。U 型池结构通过上下采样，能够聚合更大的特征感受野，增强模型的长程依赖能力，增强特征表达的多样性和鲁棒性。

GCA 通过两种不同的池化方式结合，以增强每个通道上的特征表达。所输入的特征图分别在空间维度上进行全局最大池化和全局维度池化，生成两个向量，特征图通过共享的降维卷积层将维度压缩至 $c/r(r=4)$，并使用 LeakyReLU 激活函数进行非线性变换。再通过共享的升维卷积层将通道数还原回原尺寸。两个经过卷积增强的特征向量相加，生成融合后的通道特征向量，融合后的特征向量通过 Sigmoid 激活函数进行归一化，生成最终的通道权重。输入特征图的每个通道与对应的通道权重逐元素相乘，完成对通道特征的增强，GCA 的计算式如下：

$$\boldsymbol{F}_1 = f_{\text{MaxPool}}(\boldsymbol{F}^{(l-1)}),\ \boldsymbol{F}_2 = f_{\text{AvgPool}}(\boldsymbol{F}^{(l-1)})$$
$$\boldsymbol{F}_{\text{channel\_i}} = f_{\text{LeakyReLU}}(\boldsymbol{W}_{3\times3}^{\text{Down}} * \boldsymbol{F}_i)), i = 1,2$$
$$\boldsymbol{F}^{(l)} = \boldsymbol{F}^{(l-1)} \odot \sum_{i=1}^{2} \boldsymbol{W}_{3\times3}^{\text{Up}} * \boldsymbol{F}_{\text{channel\_i}} \tag{7}$$

通过全局池化与共享卷积操作，GCA 能够有效建模不同通道之间的依赖关系，提升通道间的信息交互，保留特征间的关键关系，同时缓解“神经元死亡”问题。

1.3　超分辨率重建模块

超分辨率重建模块（SRRM）通过多尺度特征增强和融合机制，为特征图还原阶段提供更丰富的特征信息，改善 MA-LSTM 的细节表达能力。超分辨率重建模块采用特征金字塔生成调制后的特征图，将输入特征图平均分割为 4 个部分，其中的 3 个部分会经历自适应下采样，最终形成 4 个不同尺度的特征图：分别为 $\frac{1}{4}C\times H\times W$，$\frac{1}{4}C\times\frac{1}{2}H\times\frac{1}{2}W$，$\frac{1}{4}C\times\frac{1}{4}H\times\frac{1}{4}W$，$\frac{1}{4}C\times\frac{1}{8}H\times\frac{1}{8}W$。所有特征图随后会通过一个多尺度深度卷积加权层进行特征增强。该层由 4 层具有 3×3 卷积核的卷积层组成，每层卷积通过不同尺度的卷积核进行特征加权，从而有效提取和保留重要的局部和全局信息，SRRM 的网络结构如图 5 所示。

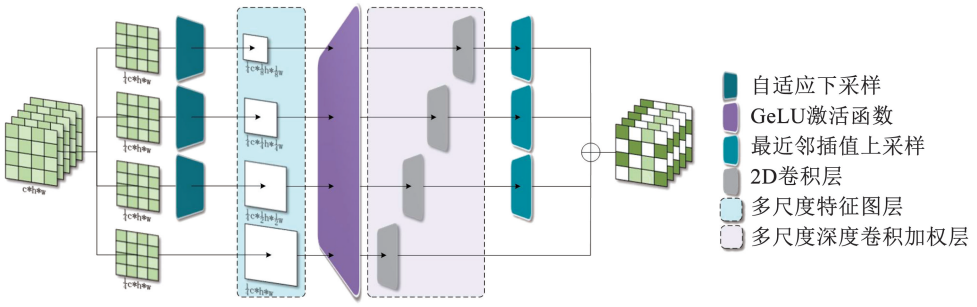


图 5　SRRM 网络结构图

Fig. 5　Structure of the SRRM

增强后的特征图通过最近邻插值上采样至原始大小，4 个尺度的特征图在通道维度上拼接，形成多尺度特征图。为了进一步整合多尺度特征，模块引入了一个融合卷积层（卷积核为 3×3）对拼接后的特征图进行整合。随后，使用 GeLU 激活函数引入非线性映射，增强特征图的表达能力。

最后，经过非线性映射的调制特征图与原始输入特征图逐元素相乘完成特征调制，调制在保留原始输入中的关键信息的同时强化由金字塔生成的多尺度特征，确保重建后的特征图具有更高的细节表达能力，SRRM 的计算式见式（8）。

$$\{\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3,\boldsymbol{F}_4\}=S(\boldsymbol{F}^{(l-1)})$$
$$\boldsymbol{F}'_i=\boldsymbol{W}^{\text{Down}}_{3\times3}*\boldsymbol{F}_i\quad \boldsymbol{F}'_i\in\mathbb{R}^{\frac{1}{2^{i-1}}C\times\frac{H}{2}\times\frac{W}{2}}\quad \forall i\in\{2,3,4\}$$
$$\boldsymbol{F}''_i=\boldsymbol{W}^{\text{Enhance}}_{3\times3}*\boldsymbol{F}'_i\quad \forall i\in\{1,2,3,4\}$$
$$\boldsymbol{F}'''_i=\boldsymbol{W}^{\text{Up}}_{3\times3}*\boldsymbol{F}''_i\quad \forall i\in\{2,3,4\}$$
$$\boldsymbol{F}_{\text{activated}}=f_{\text{GeLU}}(\boldsymbol{W}^{\text{Fusion}}_{3\times3}(\text{Concat}(\boldsymbol{F}''_1,\boldsymbol{F}'''_2,\boldsymbol{F}'''_3,\boldsymbol{F}'''_4)))$$
$$\boldsymbol{F}^{(l)}=\boldsymbol{F}_{\text{activated}}\odot\boldsymbol{F}^{(l-1)}\tag{8}$$

式中：S 为特征图分割操作，Concat 为特征图拼接。

超分辨率重建模块的多尺度特征增强设计能够在多个尺度上增强特征图的特征表达，通过融合卷积层和 GELU 激活函数整合，实现了局部特征和全局信息的高效融合。SRRM 弥补了在特征图还原时存在的特征缺失问题，可为输出预测结果提供更清晰的数据支持。

2 实验结果及分析

2.1 损失函数及评价指标

MA-LSTM 采用均方误差（mean square error，MSE）作为模型的损失函数，MSE 能够有效降低模型在训练时可能出现的梯度爆炸、消失现象，促进模型优化[18]，计算方法由式(9)表示。

$$S_{\text{MSE}}(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i)^2\tag{9}$$

式中：n 为用于训练的所有样本总数，$\boldsymbol{y}_i$ 为真实值的第 i 个样本值。$\hat{\boldsymbol{y}}_i$ 为预测值的第 i 个样本值。对于每个样本计算预测值与真实值之间的误差，MSE 越小，则表示预测值越接近真实值。

为了更全面地评估模型的预测性能，本文采用结构相似性指数（SSIM）[19]和峰值信噪比（PSNR）作为补充对比评价指标。相对于 MSE，SSIM 更加关注结构信息的保持情况，PSNR 更侧重整体图像质量的评价，通过结合 MSE、SSIM 与 PSNR，可以更全面地评估模型的总体能力，SSIM 与 PSNR 的计算方法可由式(10)表示：

$$\text{SSIM}(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)=\frac{(2\boldsymbol{\mu}_{y}\boldsymbol{\mu}_{\hat{y}_i}+C_1)(2\boldsymbol{\sigma}_{y\hat{y}_i}+C_2)}{(\boldsymbol{\mu}^2_{y_i}+\boldsymbol{\mu}^2_{\hat{y}_i}+C_1)(\boldsymbol{\sigma}^2_{y_i}+\boldsymbol{\sigma}^2_{\hat{y}_i}+C_2)}$$
$$\text{PSNR}(\hat{\boldsymbol{y}}_i)=10\log_{10}\left(\frac{S_{\text{MAX}}(\hat{\boldsymbol{y}}_i)_I^2}{S_{\text{MSE}}(\hat{\boldsymbol{y}}_i)}\right)\tag{10}$$

式中：$\boldsymbol{\mu}_{y_i}$ 与 $\boldsymbol{\mu}_{\hat{y}_i}$ 分别是真实图像和预测图像的均值，$\boldsymbol{\sigma}^2_{y_i}$ 与 $\boldsymbol{\sigma}^2_{\hat{y}_i}$ 分别是真实图像和预测图像的方差，$\boldsymbol{\sigma}_{y\hat{y}_i}$ 则是其二者的协方差，C_1 与 C_2 是稳定常数，$S_{\text{MAX}}(\hat{\boldsymbol{y}}_i)_I$ 为图像的最大像素值。其中，SSIM 越接近 1，表示预测结果与真实图像越相似，PSNR 值越高，表示图像质量越好。

2.2 数据集

MovingMNIST 数据集常用来进行视频预测模型性能评估。基于经典的 MNIST 手写数据集，生成动态的手写数字序列，每一幅图像的大小为 64 * 64 像素，每个序列包含 10 ~ 20 帧，根据参数设置，图像中会有一个或多个数字同时在图像中移动，可能会发生碰撞、反弹和穿过边界等[20]。MovingMNIST 数据集为动态数据集，无固定的训练数量，每次根据模型的参数动态确定总生成数量，在本模型中采用 10 000 个序列进行训练，每个序列长度为 20。

KTH 数据集由罗马尼亚科学院开发，是一个大规模的 3D 人体姿态数据集，包含了 11 名不同受试者在不同角度和动作下采集的数据，包含了 360 万种人体姿势以及其对应的图像[21]。在测试中使用步行姿势用于训练，将所有的图像调整大小至 64 * 64 * 1。

2.3 实验结果

本文选用了具有代表性的图像时间序列预测网络以及近期所发布的网络进行对比：ConvLSTM 是率先结合卷积操作与 LSTM 的网络结构；SwinLSTM 是首个将 SwinTransformer 的滑动窗口注意力机制与 LSTM 相结合的网络；PhyDNet 在物理建模能力上存在优势；PredRNN 通过引入时空记忆单元增强了长程时序依赖能力。所有模型均在 Windows 11 系统下运行，CPU 使用 Intel Core i5 13600KF，GPU 使用 Nvidia RTX 3090 24G，环境主体为 python 3.9、pytorch 2.1.0、CUDA 11.8。所有模型训练批次大小设置为 8，MA-LSTM 迭代了 1 000 轮，对比模型均使用推荐的训练参数，并按照原文推荐迭代轮数进行训练，如果模型较早出现了饱和现象（连续 n/10 轮未出现评价指标得分提升，n 为模型推荐的迭代次数），则提前终止迭代。

表 1 为 MA-LSTM 与对比模型在两个数据集（MovingMNIST 和 KTH）上的测试指标。针对 MovingMNIST 数据集，MA-LSTM 在 MSE、SSIM 和 PSNR 的 3 个指标上均表现优异。具体而言，MA-LSTM在 MovingMNIST 数据集上达到了最佳 MSE 值为 18.4，SSIM 为 0.9602，PSNR 为 35.16。与 SwinLSTM 相比，MA-LSTM 在 SSIM 上提升了 0.01，MSE 上提升了 0.9，PSNR 上提升了 0.82。此外，相较于其他对比模型，MA-LSTM 的 MSE 提升数值最大达到了 84.9，SSIM 最大提升数值为 0.337，PSNR 最大提升数值为 11.58，显示出其在图像时间序列预测中的显著优势。

在 KTH 数据集上的测试结果同样表明，MA-LSTM 在 MSE、SSIM 和 PSNR 三个指标上均取得优异的表现。MA-LSTM 在 KTH 数据集上的最佳 MSE

值为 21.5，SSIM 值为 0.924 3，PSNR 值为 37.23。相较于 SwinLSTM，MA-LSTM 的 SSIM 值提升了 0.02，MSE 值提升了 3.1，PSNR 值提升了 0.39。与其他对比模型相比，MA-LSTM 的 SSIM 值最大提升幅度为 0.2，PSNR 值提升幅度为 2.89。

表 1　MA-LSTM 与对比模型在不同数据集上的指标

Tab. 1　MA-LSTM and comparative models′ performance metrics on different datasets

对比模型	MovingMNIST 数据集			KTH 数据集		
	MSE 值	SSIM 值	PSNR 值	MSE 值	SSIM 值	PSNR 值
MA-LSTM	**18.4**	**0.960 2**	**35.16**	**21.50**	**0.924 3**	**34.73**
ConvLSTM	103.3	0.623 0	23.58	38.13	0.712 0	21.59
SwinLSTM	19.3	0.950 0	34.34	24.60	0.903 0	34.34
PredRNN	56.8	0.867 0	27.55	37.90	0.852 0	26.52
PhyDNet	24.4	0.942 0	32.13	45.00	0.829 0	29.04

在主要评价指标外，针对 MovingMNIST 数据集和 KTH 数据集，本文还提出了部分补充评价指标，如图 6 所示。模型收敛速度得分考察所有模型在相同环境变量下，模型完全收敛的时长，此项评分依据所有模型在相同环境下模型收敛的时长。细节维持能力得分考察模型在预测时，对于目标的细节的保持能力，此项评分依据模型的 PSNR 得分并参考所有模型生成的最后一张图片比对真实值主体细节的损失程度。结构维持能力得分考察模型在预测时，能否保持主体目标的结构性完整、不损坏目标结构，此项评分依据所有模型生成的最后一张图片，与真实值差分得出偏差，并计算偏差面积，其计算公式为

$$S = \sum_{i,j} \{ | \hat{y}_{20}(i,j) - y_{20}(i,j) | > T \} \quad (11)$$

式中：$\hat{y}_{20}$ 为网络预测输出的最后一幅图像，y_{20} 为对应的真实值图像，T 为计算阈值，S 为面积（越小越好）。

轨迹预测能力得分考察模型在预测时，能否有效预测目标的运动轨迹，不发生明显的偏离，此项评分依据所有模型的 SSIM 得分与左右模型生成的最后一张图片主体坐标比对真实值主体坐标的情况。通过主要评价指标以及补充指标，可以更全面地评价 MA-LSTM 的总体性能，图 6 展示了各项指标的具体表现，在下文的细节对比中亦有体现。

取所有模型最佳迭代权重，固定输入序列，使所有模型预测相同步数并对比结果。图 7 展示了在 MovingMNIST 数据集上的目标序列。

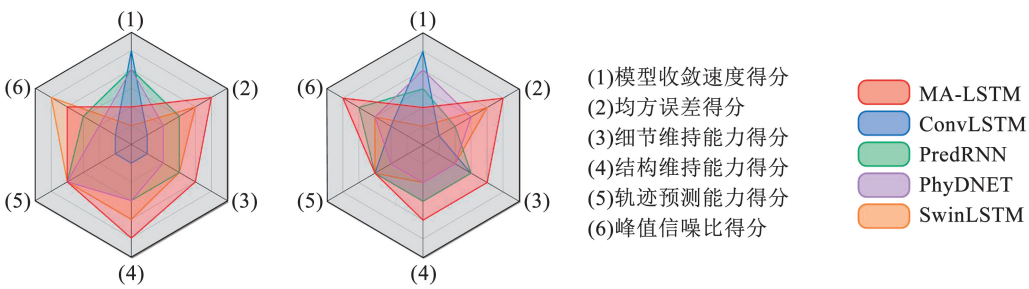


图 6　MA-LSTM 与对比模型在两个数据集上的表现雷达图

Fig. 6　Radar chart of the performance of MA-LSTM and the comparison model on two datasets

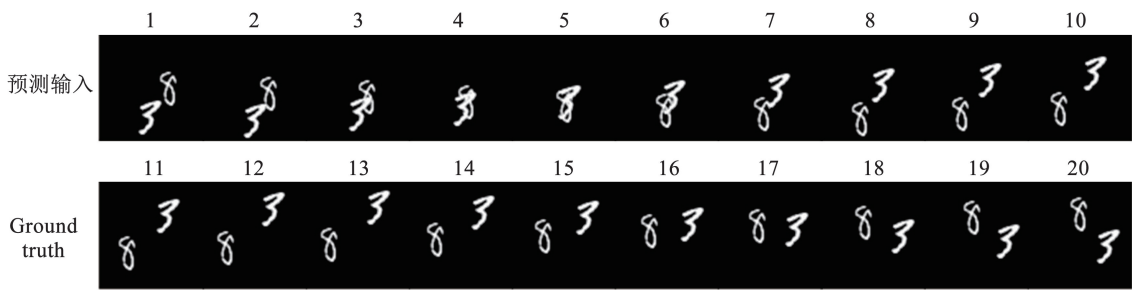


图 7　MovingMNIST 数据集目标序列

Fig. 7　Dataset target sequence of MovingMNIST

图 8 为所有模型向后预测 10 步的结果。MA-LSTM 在向后预测的 10 步中,所有数字未出现结构性损坏,所有数字运动方向与真实值高度吻合,在最后一步出现了一定的目标背景交界模糊。ConvLSTM 在第 1 步即出现了严重的数字结构性损坏,第 2 步数字结构性完全损坏,所预测的运动基本符合真实值,表明 ConvLSTM 在捕捉空间特征方面能力较差。SwinLSTM 直至第 4 步才出现了数字“8”的结构变形,但随着步数增高,结构性变形没有明显加剧,所预测的数字运动方向与真实值高度吻合,具有较强的空间信息捕捉能力以及长程依赖能力。PredRNN 在第 3 步时,数字 8 与数字 3 均出现了结构性变形,随着步数增高,结构性变形程度明显加剧,所预测的数字运动方向与真实值高度吻合,在捕捉空间特征方面的能力存在一定欠缺。PhyDNet 在第 3 步时,数字 8 出现了结构性变形,但随着步数增高,结构性变形没有明显加剧,所预测的数字运动方向与真实值高度吻合,在捕捉空间特征方面的同样存在一定欠缺。

图 8 所有模型在 MovingMNIST 数据集上的预测结果

Fig. 8 Prediction results of all models on the MovingMNIST dataset

图 9 展示了经过截取并放大的最后一幅图像,并标明了其出现的结构变形,以对比不同模型的结构维持能力。SwinLSTM 在所有预测序列中,保留了完整的数字结构 ,对比模型则出现了不同程度的结构变形。

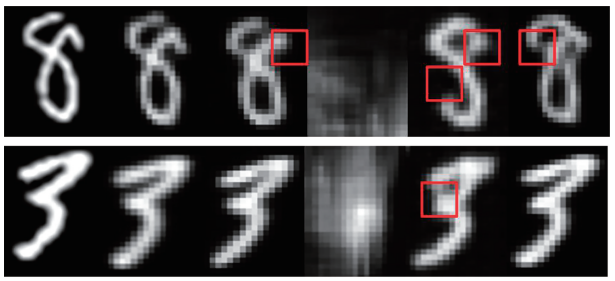


图 9 目标结构维持能力对比

Fig. 9 Comparison of target structure maintenance capabilities

在 KTH 数据集上的测试,选取了一组完整的任务挥拳动作,图 10 展示了在 KTH 数据集上的目标序列。

图 11 为所有模型向后预测 10 步的预测结果,可以看到,MA-LSTM 在向后预测的 10 步中,左臂的挥拳动作出现了部分变形,与真实值存在一定偏差,但整体而言,人物的主体结构得以保持,且人物主体细节仍然保留了较高的完整性,展现出其较强的局部细节捕捉能力。ConvLSTM 的表现相对于 MovingMNIST 数据集有所进步,能够在一定程度上保留挥拳的手臂,人物主体结构出现了比较明显的缺失。SwinLSTM 无法有效预测挥拳动作,人物主体的手臂结构完全丢失,细节缺失较为严重,证明其未能有效捕捉动作的全貌,建立准确的长程依赖数据。PredRNN 能够在一定程度上预测挥拳动作,但同样出现了人物主体手臂结构丢失的现象,同时人物主体细节缺失明显。PhyDNET 完全无法预测挥拳动作,人物主体手臂结构完全丢失,同时人物主体细节缺失明显。

如图 12 所示,将模型在两个序列上所输出的最后一幅图像与真实值做绝对值差分,并统计所得到的图像中非 0 像素的面积,面积越小,证明其结构维持能力越强,其中若出现多个重复轮廓(图 12 中的红框),则证明其轨迹预测出现偏差。

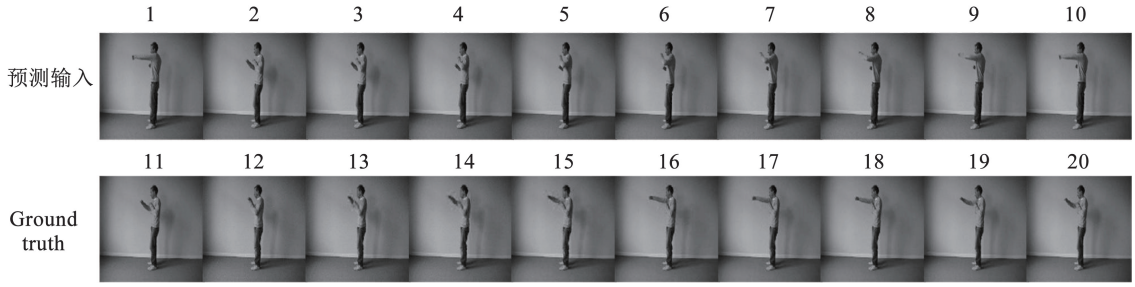


图 10　KTH 数据集目标序列

Fig. 10　Dataset target sequence of KTH

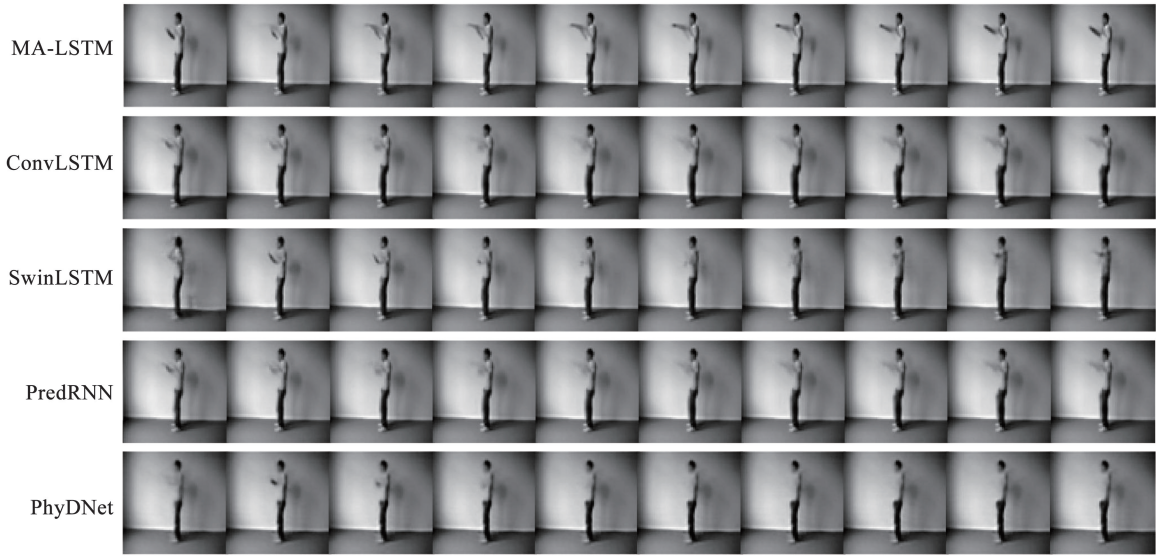


图 11　所有模型在 KTH 数据集上的预测结果

Fig. 11　Prediction results of all models on the KTH dataset

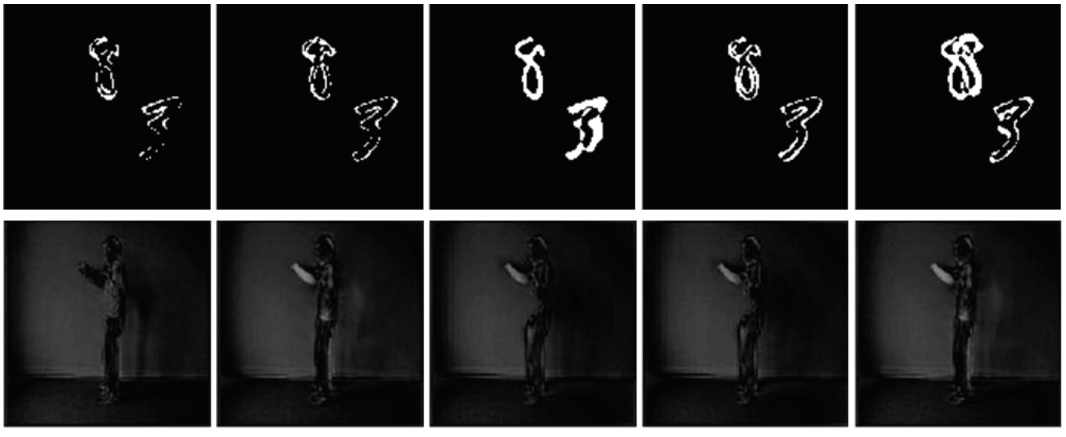


图 12　模型输出差分对比图

Fig. 12　Model output difference comparison chart

在 MovingMNIST 数据上的表现不难发现，MA-LSTM 与 SwinLSTM 均展现出优异的预测结果，说明 Transformer 结构对于运动目标轨迹的捕捉更为擅长，拥有出色的轨迹预测能力；而在 KTH 数据集上，MA-LSTM、ConvLSTM 与 PredRNN 均能在不同程度反应小区域内的挥拳动作，证实了卷积操作对于小区域的信息提取更加充分。MA-LSTM 中的 MAB 设计，充分结合了 Transformer 结构与卷积结构的优势，因此在两个数据集上均获得了更加优秀的表现。

同时，为了探究 MA-LSTM 的长程依赖能力以及可视化表达，图 13 展示了 MA-LSTM 在特征提取层以及特征生成层的特征图，并将其转置为热力图的形式表现。在图 13 中，行 2 为特征提取层热力图，行 3 为特征生成层热力图。特征提取层能够有

效提取当前层网络的输入数据,并保留一定步长的前序隐藏状态数据;如行 3 所示,从 $t=4$ 时刻开始,特征生成层已经能够生成比较明显的预测轨迹,直至最后时间步,特征生成层都能够有效地生成轨迹预测。

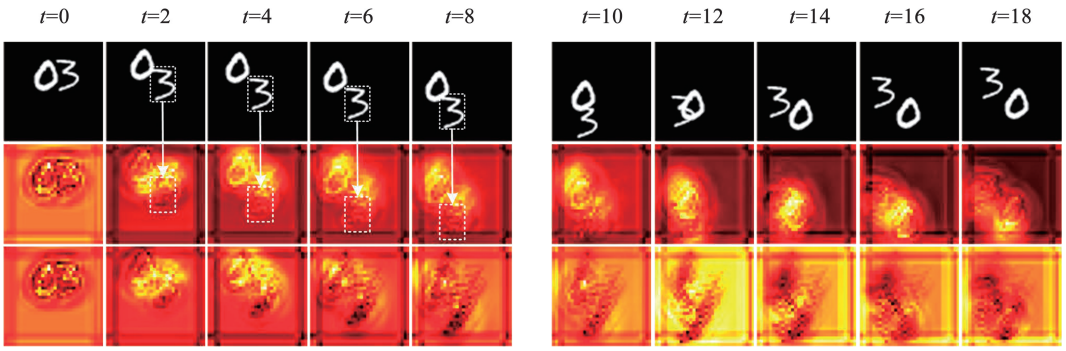


图 13 MA-LSTM 特征提取层与特征生成层转置热力图

Fig. 13 Visualization of MA-LSTM long-range dependence capability

2.4 消融实验

本文基于 MovingMNIST 数据集以及 KTH 数据集进行了消融实验,主要针对 MAB 有效性以及 SRRM 模块有效性两个方面展开了测试,MAB 的测试结果如表 2 所示。

表 2 MAB 消融实验对比指标

Tab. 2 Comparison between the model and the mainstream algorithms

模型深度	最低 MSE 值	最高 SSIM 值	平均训练时长/s
1-1-1-1	51.822 10	0.839 6	**629**
2-2-2-2	39.579 03	0.897 9	868
3-3-3-3	38.408 67	0.905 1	1 095
4-4-4-4	38.235 27	0.902 2	1 591
2-4-4-2	37.718 17	0.909 2	927
2-6-6-2	**32.056 03**	**0.909 9**	1 106

MAB 深度的测试在 MovingMNIST 数据集展开,分别进行了 6 种不同 MAB 深度的模型,验证模型均采用 2 层特征采样层和 2 层特征生成层,通过对称改变每一层的 MAB 数量测试不同网络深度下 MA-LSTM 的表现。图 14 为不同 MAB 深度下网络的验证结果。其中,1-1-1-1 代表特征采样层与特征生成层的每一层都设计有 1 个 MAB 模块;依此类推 2-4-4-2 则代表特征采样层的第 1 层设计有 2 个 MAB 模块以及第 2 层设计有 4 个 MAB 模块,特征生成层的第 1 层设计有 4 个 MAB 以及第 2 层设计有 2 个 MAB 模块。当平等地增加每个层的 MAB 数量时,在 4-4-4-4 时网络出现了退化现象,在训练时长显著增加的情况下出现了指标退化。因此考虑非平等增加层数量,在 2-6-6-2 时网络指标增幅明显,超越 4-4-4-4深度的同时训练时长显著低于 4-4-4-4 深度,与 3-3-3-3 深度的训练时长接近。

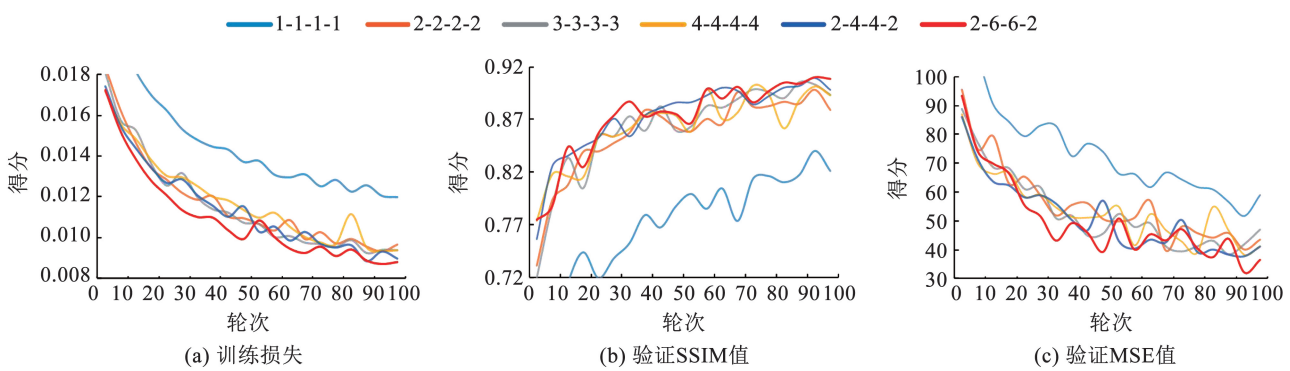


图 14 MAB 深度消融实验验证结果

Fig. 14 Ablation experimental training results about MAB

对 MAB 进行消融试验得到的结果,结合不同深度网络下 MAB 的表现,给出一个用于计算不同任务下的模型深度估计经验公式

$$D=\frac{\alpha(T_{\text{in}}\times T_{\text{out}})+\beta(H\times W)}{\gamma} \tag{12}$$

式中:D 为模型中 MAB 的累计深度,α、β、γ 为超参数(以本文所示训练环境确立,$\alpha=0.005$、$\beta=0.5$、$\gamma=4.5$),T_{in}、T_{out} 分别为输入序列长度、输出序列长度,H、W 分别为图像的长宽。

针对 SRRM 的测试在 KTH 数据集展开,则分别进行仅 TransConv2D 模块、TransConv2D&SRRM 双模块、仅 BilinearInterpolation 模块以及 Bilinear-

Interpolation&SRRM 双模块 4 种方案进行对比，所有对比模型均迭代相同代数，对比不同模型的最优 MSE 与 SSIM 得分。验证结果如图 15 所示，可以看到，TransConv2D&SRRM 的组合，在收敛速度上体现出明显优势，拥有更低的 MSE 和更高的 SSIM 得分，在趋势上相对于其他模型也有明显的优势。重新分组，将未添加 SRRM 的模型与添加了 SRRM 的结果进行对比，SRRM 也表现出明显的特征增强优势，相对于另外两个模型拥有更低的 MSE 和更高的 SSIM 得分，增加了模型收敛速度。

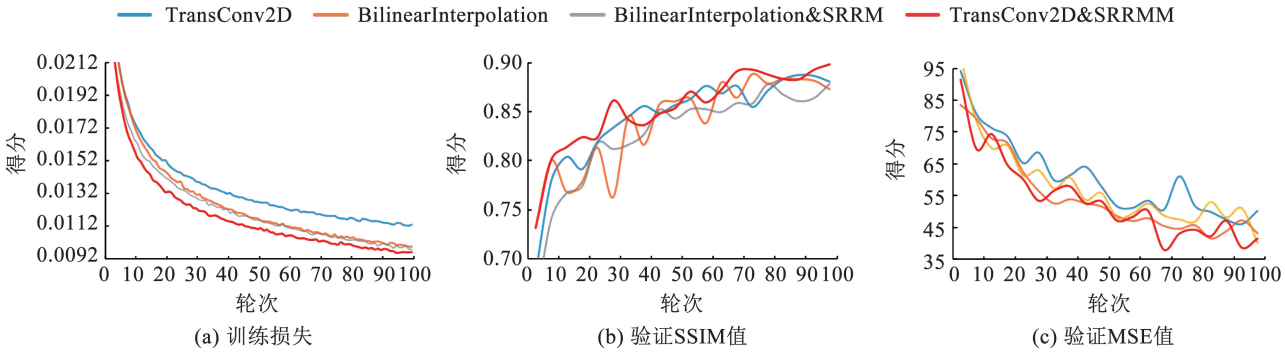


图 15　SRRM 有效性消融实验验证结果

Fig. 15　Ablation experimental training results about SRRM

3　结　论

本文提出了一种基于多尺度特征建模的图像时间序列预测网络 MA-LSTM，并从结构设计、性能评估和模块有效性验证等方面展开了系统研究，主要结论如下。

1）提出了多尺度注意力模块（MAB），融合了移位窗口注意力机制与卷积结构，弥补了传统 Transformer 在细粒度建模和局部特征提取方面的不足。模块内部的时空特征增强层（GSTA）通过多尺度卷积与 U 型池化结构提升了空间建模能力，通道特征增强层（GCA）利用不同尺度的通道池化加强了跨通道的信息交互。

2）通过简化 LSTM 与 MAB 构建了多尺度注意力层（MALayer），形成了横向与纵向交错的信息提取机制，降低了模型复杂度，显著增强了对长程时序依赖的建模能力。

3）设计了超分辨率重建模块（SRRM），采用特征图拆分与多尺度金字塔卷积策略，显著增强了网络在重建阶段的细节恢复能力，从而提升了最终预测图像的质量。

4）在 MovingMNIST 和 KTH 两个具有代表性的数据集上进行广泛实验验证，MA-LSTM 在多个主流性能指标上均优于现有模型（SwinLSTM、ConvLSTM、PredRNN、PhyDNet），在结构维持能力、目标轨迹预测能力和细节保留能力方面表现出显著优势。通过消融实验，进一步验证了 MAB 和 SRRM 模块的有效性，并在此基础上提出了基于输入长度、图像尺寸与输出长度的网络深度估算经验公式，为不同任务下的模型配置提供参考。

参考文献

[1] QIU Xiangfei, WU Xingjian, LIN Yan, et al. DUET: Dual clustering enhanced multivariate time series forecasting [J/OL]. [2025-12-20]. DOI:10.48550/arXiv.2412.10859

[2] LIM B, ZOHREN S. Time-series forecasting with deep learning: A survey[J]. Philos Trans A Math Phys Eng Sci, 2021, 379(2194): 20200209. DOI: 10.1098/rsta.2020.0209

[3] LI Zewen, LIU Fan, YANG Wenjie, et al. A survey of convolutional neural networks: Analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999. DOI: 10.1109/TNNLS.2021.3084827

[4] LIN Shengsheng, LIN Weiwei, HU Xinyi, et al. Cyclenet: Enhancing time series forecasting through modeling periodic patterns[J]. Advances in Neural Information Processing Systems, 2024, 37: 106315. DOI: 10.48550/arXiv.2409.18479

[5] SEO J H, KIM K D. An RNN-based adaptive hybrid time series forecasting model for driving data prediction [J]. IEEE Access, 2025, 13: 54177. DOI: 10.1109/ACCESS.2025.3554803

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation MIT-Press, 1997, 9(8): 1735

[7] SHI Xinjian, CHEN Zhourong, WANG Hao, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. Advances in Neural Information Processing Systems, 2015, 28: 802. DOI: 10.1145/3620679.3620688

[8] WANG Yunbo, WU Haixu, ZHANG Jianjin, et al. Predrnn: A recurrent neural network for spatiotemporal predictive learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 2208. DOI: 10.1109/TPAMI.2022.3165153

[9] YI Kun, GE Yixiao, YANG Shusheng, et al. Masked image

modeling with denoising contrast[C]// 11th International Conference on Learning Representations, ICLR 2023. Kigali, Rwanda: OpenReview. net, 2023: 1. DOI: 10.48550/arXiv.2205.09616

[10] GUEN V L, THOME N. Disentangling physical dynamics from unknown factors for unsupervised video prediction [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 11471. DOI: 10.48550/arXiv.2003.01460

[11] CHEN Peng, ZHANG Yingying, CHENG Yunyao, et al. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting[C]// 12th International Conference on Learning Representations. Hybrid, Vienna, Austria: ICLR, 2024. DOI: 10.48550/arXiv.2402.05956

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998. DOI: 10.48550/arXiv:1706.03762

[13] WU Chuang, HE Tingqin. A survey of applications of vision transformer and its variants[C]//2024 IEEE 10th International Conference on Intelligent Data and Security, IDS 2024. New York City, NY, USA: IEEE, 2024: 21. DOI: 10.1109/IDS62738.2024.00011

[14] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//18th IEEE/CVF International Conference on Computer Vision, ICCV 2021. Canada: IEEE, 2021: 9992. DOI: 10.48550/arXiv.2103.14030

[15] TANG Song, LI Chuang, ZHANG Pu, et al. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm [C]// Proceedings 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 13424. DOI: 10.48550/arXiv.2308.09891

[16] REKAVANDI A M, RASHIDI S, BOUSSAID F, et al. Transformers in small object detection: A benchmark and survey of state-of-the-art[J]. ACM Computing Surveys, 2025, 58(3): 1. DOI:10.1145/3758090

[17] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 2117. DOI: 10.48550/arXiv.1612.03144

[18] ZHOU Jinxin, LI Xiao, DING Tianyu, et al. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features[C]//Proceedings of the 39th International Conference on Machine Learning, ICML 2022. Vienna, Austria: PMLR, 2022: 27179. DOI: 10.48550/arXiv.2203.01238

[19] WANG Zhou, BOVIK A C, HAMID R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600. DOI: 10.1109/TIP.2003.819861

[20] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R, et al. Unsupervised learning of video representations using lstms[C]// 32nd International Conference on Machine Learning, ICML 2015. Lille, France: PMLR, 2015: 843. DOI: 10.48550/arXiv.1502.04681

[21] KANG S M, WILDES R P. Review of action recognition and detection methods[J]. arXiv: 1610.06906. DOI: 10.48550/arXiv.1610.06906

(编辑 吕雪梅)