基于膨胀卷积与注意力机制的多维时序预测方法

刘尚东[1,2]，袁帅祥[1]，徐鹤[1,2]，刘思行[3]，季一木[1,2]
（1. 南京邮电大学计算机学院、软件学院、网络空间安全学院，南京 210023；）
（2. 江苏省高性能计算与智能处理工程研究中心，南京 210023；）
（3. 江苏鱼跃医疗设备股份有限公司，镇江 212300）

摘要

针对多元时间序列预测中通道间复杂非线性依赖与时序偏移现象导致模型预测精度不足的问题，提出了一种基于膨胀卷积与通道注意力机制的多维时序预测模型（Dilated Convolutional Time Series Transformer, DCTST）。首先，通过倒置嵌入模块将每个变量的整个历史序列独立编码为高维令牌，保留全局时序特征；其次，设计通道注意力模块，结合自注意力机制与可学习路由机制，并引入膨胀卷积模块在变量维度上扩展感受野，捕获多尺度局部依赖；最后，通过跨层残差连接整合原始输入与深层特征，增强模型稳定性。理论分析和实验表明，与主流通道独立性模型（如 PatchTST）和通道相关性模型（如 iTransformer）相比，DCTST 在多个真实数据集和合成数据集上的预测均方误差（MSE）和平均绝对误差（MAE）均有显著降低。实验结果表明，DCTST 能够有效建模通道间依赖，提高多元时间序列的预测精度与鲁棒性，适用于智能电网、交通流量等实际工程场景。

关键词：时间序列预测；膨胀卷积；Transformer；注意力机制
中图分类号： TP391

引言

随着数据存储与传感器技术的快速发展，海量多维时间序列数据在诸多领域呈现爆发式增长。这类数据包含多个变量，每个变量表示一个单变量时间序列，对其的预测目标是通过分析历史观测值，提取规律与趋势，从而为未来的决策提供依据。这类数据不仅在时间维度上有着复杂的动态变化模式，而且在不同变量（通道）之间存在潜在的非线性关联与多尺度特征互动。传统统计学分析方法[1,2]在捕捉这类多维时间序列的复杂关联时，往往面临较大的挑战，导致预测精度和泛化能力受限。因此，深度学习模型[3-7]逐渐崭露头角。尤其是基于注意力机制的 Transformer 架构[8]，在自然语言处理和时间序列分析等领域均取得了显著成果。

基金项目：江苏省前沿技术研发计划资助（基金号: BF2025617）

此外，一些具备线性化处理特征的多层感知机（Multi-Layer Perceptron, MLP）方法[9,10]在时间序列预测中也表现出色。

此前，基于 Transformer 的模型为自然语言处理而设计，在捕获时间序列长期依赖关系方面取得了成功，但早期在时间序列预测上的应用中，将同一时间步的不同变量数据视为单个标记的策略面临几个问题：1、没有考虑变量之间的时序偏移，可能对 token 的加权表示混入噪声，削弱相关 token 的关注。2、每个 token 的单步表示无法充分包含历史信息。3、不同物理量被嵌入为同一 token，物理意义的边界模糊，生成混合特征而非保持独立性。结果这些模型[3,4,11,12]受到了基于 MLP 方法[13]的挑战，这些方法为多变量时间序列建模提供了创新的视角。

在这两大方法的竞争下，研究者们不断融合彼此的优点以改进模型性能。PatchTST[14]使用分块嵌入策略，通过处理较小的时间片段改善单通道的自相关性表示，这种方式类似于自然语言处理中子词级标记化的思想。然而，PatchTST 仅关注通道内关系，忽略了多维时序数据通道间的潜在依赖。同时，Crossformer[15]在分块嵌入的策略上通过可学习的路由机制整合通道间交互，显著提升预测精度，但以高昂的计算复杂性为代价。相比之下，iTransformer[16]通过将每个变量的整个序列视为一个嵌入进行跨通道注意力建模，极大提高了计算效率，但由于粗粒度的嵌入方式，其在捕捉时间细节依赖时表现不足。

此外，线性模型也对通道相关性建模进行了探索。这些方法[9,17]通常通过显式通道融合与归纳偏置来增强预测能力，但在复杂数据场景下仍受限于精度和计算效率之间的权衡。

为应对此挑战，我们提出了一种高效的膨胀卷积时间序列预测模型（DCTST）。与传统依赖注意力机制的模型不同，DCTST 将膨胀卷积结构引入多元时间序列建模框架中，既保持对全局通道关系的整体把握，又精准捕捉通道间复杂的细微共性，显著提升预测性能并降低对计算与存储资源的过度需求。我们的贡献如下：

(1)深入探索当前通道相关性模型中注意力机制在多维时间序列预测中的应用，系统分析变量嵌入与维度分段嵌入方式下的缺陷，明确通道相关性建模中的关键痛点与未来改进方向。

(2)提出了 DCTST，将膨胀卷积结构融入多维时间序列预测模型中，实现对通道间潜在关联的高效建模，克服现有变量嵌入与维度分段嵌入方式的不足。

(3)在多个真实世界数据集以及自建多维相关性数据集上，DCTST 显著提升预测精度与泛化表现，并通过有效降低计算开销与存储压力，提高了模型在大规模场景下的适用性。

1. 相关理论

1.1. 时间序列预测研究现状

在传统统计方法阶段，自回归整合移动平均（Autoregressive integrated moving aver age, ARIMA）模型和向量自回归模型（Vector Autoregression, VAR）等模型通过线性依赖关系建模时间序列的趋势与自回归特性，这些模型适用于单变量或多变量时间序列的线性特性，但难以处理非线性和长时间依赖，且对多维特征之间复杂动态关系的建模能力有限。随着数据规模和复杂性的增加，深度学习方法逐步成为主流。以循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）[18]和门控循环单元（Gated Recurrent Unit, GRU）[19]为代表的递归神经网络通过递归结构和门控机制捕捉长序列中的时间依赖关系，显著缓解了传统模型的局限性。然而，递归结构的计算效率较低，且在面对超长序列时依然存在梯度消失等问题。为解决这些问题，卷积方法应运而生，其中时域卷积网络（Temporal Convolutional Network, TCN）[20]通过扩张卷积设计，显著扩大了感受野，在捕获长时间依赖的同时避免了梯度消失。然而，这一阶段的模型大多聚焦于时间维度的建模，未能充分考虑多维特征间的动态关系。进入 Transformer 及其变体阶段后，Transformer 通过注意力机制，灵活捕捉长序列中的依赖关系，并支持大规模并行计算，成为长序列建模的强有力工具。随后，为适应时间序列预测任务，诸如 Informer[3]、Autoformer[4]和 PatchTST[14]等变体模型相继提出，通过稀疏注意力和序列分块等优化策略，提升了长序列的建模效率与预测精度。在此过程中，研究者逐步认识到多变量时间序列中通道之间的潜在关联性对模型性能的重要性，开始从通道独立建模向通道相关性建模递进[21-24]。

在 Transformer 类模型的嵌入策略上，可以从注意力机制的应用角度将其分为整体嵌入和分段嵌入两类。这一区分的核心在于建模序列依赖关系的粒度不同：整体嵌入通过序列级的嵌入来捕获各维度序列的相互依赖关系。而分段嵌入则将序列划分为多个补丁，强调补丁级的依赖关系，补丁级的依赖一般作用在捕获时间维度相关性上。多维时间序列预测目前逐步形成两大类别。其中通道独立性模型的设计逻辑是针对每个通道独立建模，将时间序列的复杂性拆解为单一维度的问题。这样的模型避免了多通道间复杂交互的建模需求。DLinear[13]是这一类别的典型代表，通过时间序列分解，将序列拆分为趋势部分和残差部分，分别建模以提升精度。TimeMixer[10]则进一步通过不同尺度的分解与建模，增强了对单通道时间依赖的多层次捕捉能力。PatchTST 在单通道内部引入分块嵌入，捕捉块间的相关性，并最终通过线性层进行预测归纳。然而，通道独立性模型也存在显著局限性：它们仅在时间维度分析单个通道，而忽略了跨通道的动态关联。在现实世界的多变量时序预测任务中，不同通道之间可能存在复杂的相互依赖关系，相比之下，通道相关性模型通过显式建模特征之间的交互关系，进一步挖掘了多变量的潜在动态关系，是对通道独立模型的扩展与深化。这类模型强调不仅要捕捉单通道内的表现，还需考虑跨通道间的关联性。Crossformer 是通道相关性模型的典型代表之一，通过分段嵌入的注意力机制捕捉时间依赖，并引入可学习的路由策略增强通道间的交互，降低分块后跨维度捕获相关性的计算成本。TimeXer[25]通过不重合分段嵌入降低了显存消耗，并通过可学习路由策略有效捕捉全局特征，以协变量预测的视角强调跨维度依赖的重要性。此外，TSMixer[9]通过交替在时间与通道维度上进行 MLP 结构混合，进一步提升了对复杂关系的建模能力。iTransformer 则对全局序列进行注意力计算，以高效方式处理通道依赖关系。SOFTS[17]在 iTransformer 的视角上加以改进，提出了替代注意力机制的模块 STAR,采用基于 MLP 的通道交互方法，通过将序列聚合为全局核心表示，再分配给单个序列进行融合，进一步提高了计算效率，但在时间维度的局部表示上表现较差。

在本文中，我们介绍了基于膨胀卷积与通道注意力机制的多维时序预测模型(Dilated Convolutional Time Series Transformer, DCTST)，它在通道维度上使用扩

张卷积来增强序列表示和注意力机制。它在捕捉通道相关性方面以高计算效率实现了最先进的性能。

2. 模型架构

多元时间序列预测的核心挑战在于有效建模变量间复杂的相互依赖关系及其随时间演变的动态模式。为了应对这一挑战，我们提出了 DCTST 模型，其核心思想是通过一种新颖的 DCAttention 模块，协同利用注意力机制和膨胀卷积，分别从全局和局部视角捕捉跨通道的时空特征。模型采用仅编码器架构，主要由嵌入层、堆叠的编码器层和投影层构成，其整体结构如图 1 所示。

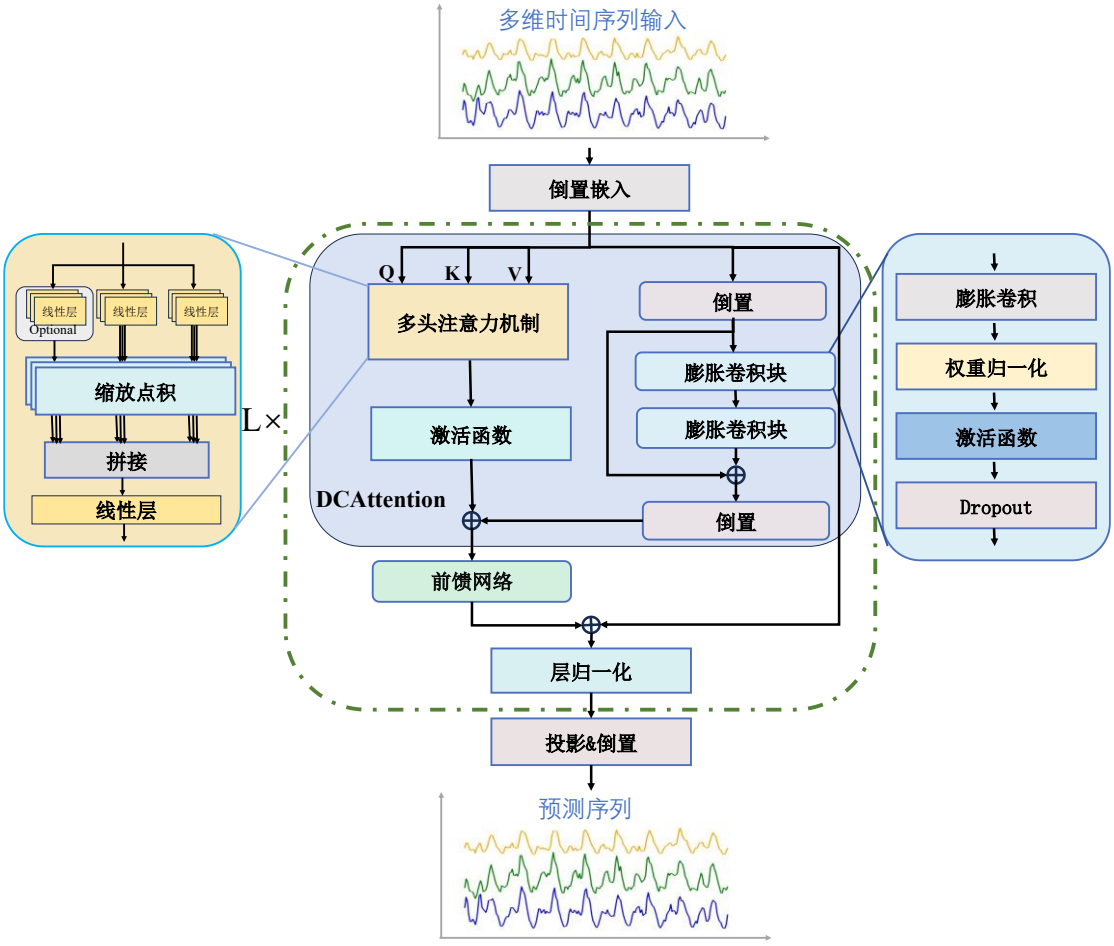


图 1 模型结构

Figure 1 Model architecture

给定一个包含 N 个变量、回看窗口长度为 T 的多元时间序列输入$X \in R^{N \times T}$，模型的目标是预测未来 S 个时间步的序列$\hat{Y} \in R^{N \times S}$。考虑到在实际传感器监测场景中，多元时间序列是以$X \in R^{T \times N}$的形式逐点采集生成的。为了构建变量间的全

局依赖，本模型通过倒置将数据重构为$N \times T$布局。接着嵌入层通过一个多层感知机将每个变量的整个时间序列映射到一个高维特征空间，生成初始的隐藏表示$H^0 \in R^{N \times D}$，其中 D 为嵌入维度。这一过程可以表述为：

$$H^0 = MLP(X^T) \in R^{N \times D} \tag{1}$$

其中，X^T代表对原始输入矩阵的转置操作，MLP 负责将每个变量的整段历史序列投影为高维表示。随后，H^0被送入由 L 个相同结构编码器层堆叠而成的编码器。每个编码器层对输入表示进行渐进式处理，其输出作为下一层的输入，即

$$H^{i+1} = EncoderLayer_i(H^i), i = 0,1, \dots, L-1 \tag{2}$$

其中每个编码层的内部计算流程由**公式 3、4、5** 定义：

$$H^i_{fused} = DCAttention(H^i) \tag{3}$$

$$H^i_{ff} = FeedForward(H^i_{fused}) \tag{4}$$

$$H^{i+1} = LayerNorm(H^i_{ff} + H^i) \tag{5}$$

每个编码器层的内部计算包含三个核心步骤：首先，隐藏表示H^i经由核心的 DCAttention 模块进行通道间的信息交互与特征提取，得到融合表示H^i_{fused}；接着，H^i_{fused}通过一个标准的前馈神经网络进行非线性变换，得到H^i_{ff}；最后，通过层归一化处理每层初始输入H^i与前馈网络输出H^i_{ff}的残差和，得到该层的输出H^{i+1}。经过 L 层编码后，最终的隐藏表示H^L通过投影层（一个 MLP）映射为预测结果$\hat{Y}^T$,公式表示为：

$$\hat{Y}^T = Projection(H^L) \tag{6}$$

最后经过转置得到预测结果。

2.1. DCAttention 模块

DCAttention 模块是 DCTST 模型的创新核心，其设计旨在解决传统方法在平衡计算效率与捕获复杂通道相关性方面面临的困境。该模块由两个并行且互补的组件构成：注意力层和膨胀卷积层。具体而言，输入特征首先进入全局注意力分支，该分支引入了可选的（Optional）查询投影层，使模型能够根据特征维度

灵活调整全局核心表示的构建方式，感知跨通道的全局依赖和共同模式。同时，输入特征被送入膨胀卷积分支，利用沿变量维度的扩张卷积操作，以在局部范围内挖掘同一时间不同通道间协同变化的模式。

2.1.1 注意力层

在注意力层中，我们将每个变量的整个嵌入序列视为一个独立的令牌（Token）。这种处理方式得益于 Transformer 架构在序列建模中的成功经验，特别是其通过查询（Query）、键（Key）、值（Value）机制计算相关性权重的能力。自注意力机制的计算遵循标准公式：

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

其中$Q = HW_Q$,$K = HW_K$,$V = HW_V$分别由输入嵌入H经过线性变换得到。$W_Q, W_K, W_V \in R^{D\times d_k}$是可学习的权重矩阵，其中$d_k$是查询和键的维度。Softmax 函数对相似性分数进行归一化。

受前人工作启发[26]，为了在捕获全局通道交互的同时维持线性计算复杂度，我们提出了一种非对称的全局注意力机制。该机制的核心是一个可学习的全局特征提取器$Extr \in R^{h\times D}$，其被定义为一个可学习的全局变量矩阵，其中 h 是分组因子。$Extr$通过均值池化生成全局查询向量Q_{global}，公式表示为：

$$Q_{global} = \frac{1}{h}\sum_{i=1}^{h} Extr[i] \tag{8}$$

随后以$Q_{global} \in R^{1\times D}$作为查询向量，以隐藏表示$H$所映射的 K、V 作为键值对，通过缩放点积注意力计算,公式表示为：

$$GlobalAttention\left(Q_{global}, K, V\right) = Softmax\left(\frac{Q_{global}K^T}{\sqrt{d_k}}\right)V \tag{9}$$

最终得到融合全局模式的注意力输出H_{global}，它聚合了所有通道的全局信息。这种设计避免了计算变量间两两交互的N^2复杂度，使模型能够高效地处理变量数众多的多元时间序列。图 2 展示了两种注意力对输入嵌入的处理方式。

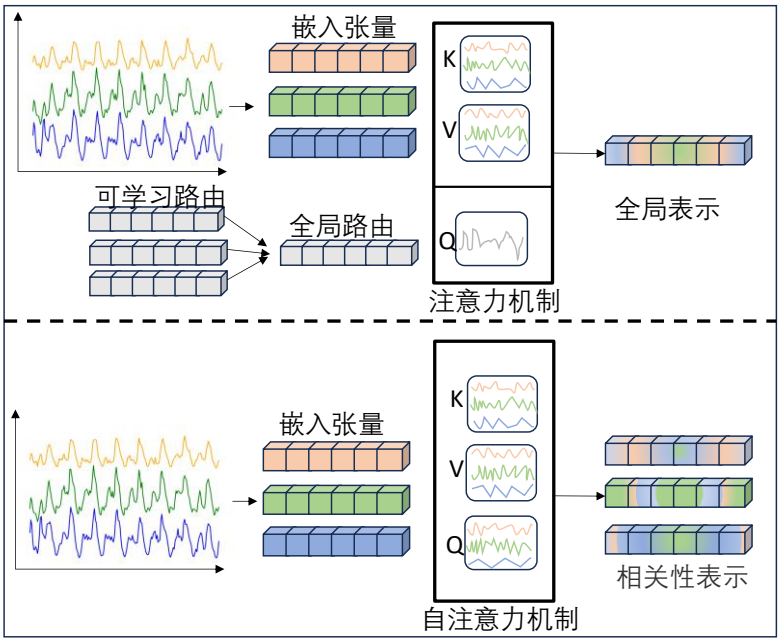


图 2 注意力机制的两种实现方案(上图为全局注意力机制，下图为自注意力机制)

Figure 2 Two implementation schemes of attention mechanism (upper for global attention, lower for self-attention)

2.1.2 膨胀卷积层

尽管注意力机制擅长捕获全局依赖，但对于序列中局部范围内变量间存在的短程、高强度相关性，其捕捉能力可能不足。为此，我们设计了膨胀卷积层，专门用于在通道维度上提取局部时空特征。

与通常在时间维度上进行滑动的传统时序卷积不同，我们的膨胀卷积操作沿着变量维度进行。具体而言，给定输入嵌入矩阵$H \in R^{N\times D}$，我们首先将其转置为$H^T \in R^{D\times N}$。这一转置操作使得卷积核可以在变量索引维度上滑动，从而直接感知不同变量在同一隐含时间点上的关系。对于每个变量 c，在其局部感受野内的卷积操作定义为：

$$\tau_{i,k',t'} = \sum_{k'=0}^{K-1} W_{i,k',t'} \cdot H^T[t', c - d \cdot k'] \tag{10}$$

其中，K 是卷积核长度，d 是扩张率，i 是输出通道索引，$W_{i,k',t'}$是卷积核权重。膨胀卷积允许模型在不增加参数量的情况下，以指数级速度扩大感受野，从而捕获更长范围的通道依赖。随后，我们将所有隐含时间步t'上的卷积结果进行

累加，并加上偏置项，得到每个输出通道 i 对于变量 c 的卷积结果$Z_{i,c} = \sum_{t'=0}^{D-1} \tau_{i,c,t'} + bias_i$。将所有输出通道的结果堆叠，得到矩阵$Z \in R^{D\times N}$，再经过转置操作后，最终得到与输入$H$维度一致的膨胀卷积输出$Z^T \in R^{N\times D}$。图 3 描述了膨胀卷积的处理流程。

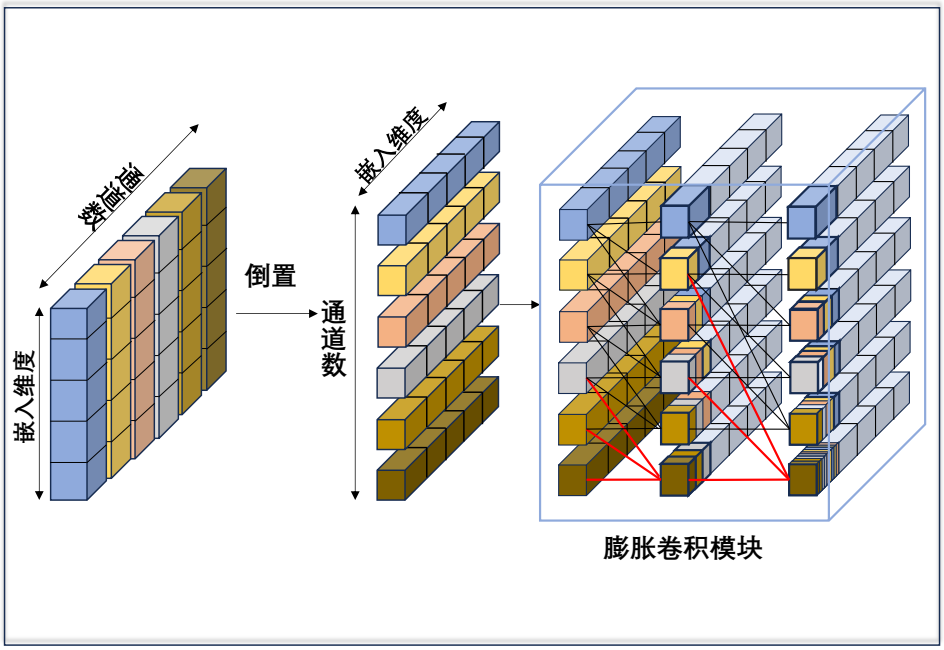


图 3 膨胀卷积层处理流程

Figure 3 Processing flow of dilated convolution layer

2.1.3 输出融合

注意力层和膨胀卷积层的输出以互补的方式融合，形成 DCAttention 模块的最终输出。融合公式为：

$$H_{fused} = ReLU\big(Attention(H)\big) + Z^T \tag{11}$$

其中，$Attention(H)$是注意力机制提供的富含全局通道交互信息的表示，将每个通道用相似度高的通道以不同的权重进行累加来表示。而Z^T是膨胀卷积捕捉的局部通道特征。ReLU 激活函数引入了非线性。这种融合策略确保了模型能够同时利用全局上下文和局部细节，从而构建出更具表现力的特征表示。

2.2 膨胀卷积的梯度分析

为了深入理解膨胀卷积层如何促进模型学习，我们对其反向传播过程进行了分析。核心在于考察输出Z^T中每个元素对卷积核权重$W_{i,k',t'}$的偏导数$\frac{\partial Z^T[c,t]}{\partial W_{i,k',t'}}$。这里，$Z \in R^{D\times N}$是一个由 D 个卷积核的结构$Z_{i,:} \in R^{1\times N}$堆叠而成的矩阵，通过将 Z 转置得到膨胀卷积的输出$Z^T \in R^{N\times D}$，其中$Z^T[c,t] = Z[i,c]$，c 代表变量索引，t 代表隐含时间步,i 则为卷积核的编号。

根据前向传播公式，我们有：

$$\frac{\partial Z^T[c,t]}{\partial W_{i,k',t'}} = \frac{\partial Z_{i,c}}{\partial W_{i,k',t'}} = \frac{\partial\left(\sum_{t'=0}^{D-1}\sum_{k''=0}^{K-1} W_{i,k'',t'} \cdot H^T[t', c-d\cdot k''] + bias_i\right)}{\partial W_{i,k',t'}} = H^T[t', c-d\cdot k'] \tag{12}$$

这一结果表明，在梯度计算中，每个卷积核参数$W_{i,k',t'}$的更新仅依赖于输入特征$H^T[t', c-d\cdot k']$本身。至关重要的是，这个梯度与时间步 t 无关，尽管$Z^T[c,t]$是跨所有隐含时间步累加的结果。这意味着，在权重更新时，每个时间步 t 的梯度是独立计算的，不会因为其他时间步的信息而产生混淆。这一特性确保了膨胀卷积在整合所有隐含时间步信息以构建长期时间依赖的同时，不会引入不必要的时间混叠噪声，从而很好地保持了时间维度上的独立性。这对于时间序列预测任务至关重要，因为保持清晰的时间因果关系有助于提高预测的准确性。

由于卷积核权重的梯度更新在形式上独立于其他时间步，其学习过程本质上是在挖掘同一隐含时间步t'下，不同通道间的协同变化模式。具体来说，权重$W_{i,k',t'}$直接作用于局部感受野内的一组通道特征$\{H^T[t', c-d\cdot k'] | k' \in [0, K-1]\}$。在训练过程中，通过反向传播，梯度$\frac{\partial Z^T[c,t]}{\partial W_{i,k',t'}} = H^T[t', c-d\cdot k']$指示了参数更新的方向。如果某一时刻，某个通道$c-d\cdot k$的特征值$H^T[t', c-d\cdot k']$与目标输出存在强烈的相关性，那么对应于该通道的卷积核权重$W_{i,k',t'}$就会获得一个幅度较大的更新。

相比之下，若不进行转置而直接在原始时间维度上应用膨胀卷积，则梯度$\frac{\partial Z[c,t]}{\partial W_{i,k',t'}}$将变为$H[c', t-d\cdot k']$。在这种情况下，局部时间步$\{t-d\cdot k'\}$的信息会被混合到同一个输出时间步 t 中，这很可能在数据中引入因果噪声，破坏模型对时

间先后顺序的建模能力。因此，我们采用的沿通道维度的膨胀卷积策略在理论上是更优的选择。在图 4 中展示了两种膨胀卷积滑动方向的差异。

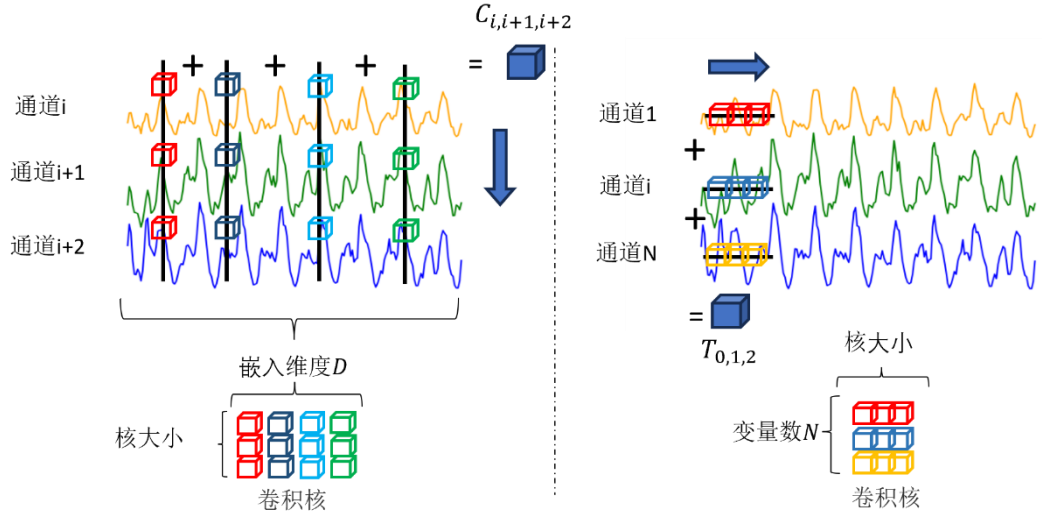


图 4 膨胀卷积的方向(左图为沿通道维度滑动，右图为沿时间维度滑动)
Figure 4 Directions of dilated convolution (left for sliding along channel dimension, right for sliding along temporal dimension)

2.3 DCTST 模型

尽管基于 Transformer 的模型在多元时间序列预测领域已取得显著进展，能够通过自注意力机制灵活捕捉长期时间依赖性和复杂模式，然而，在面对高维、噪声显著且变量间存在复杂动态关联的实时预测任务时，此类模型常面临计算开销庞大、对局部通道相关性捕捉不足以及抗噪能力较弱等挑战，从而限制了预测精度与泛化性能。

因此，本研究提出了 DCTST 模型，旨在通过引入膨胀卷积结构增强特征表示与通道交互效率，平衡模型复杂度与预测准确性。在 DCTST 中，我们设计了 DCAttention 核心模块，将膨胀卷积与注意力机制并行融合，以同时从全局和局部视角提取多变量时间序列中的时空特征。该模型通过整体序列嵌入、多层编码器堆叠和线性投影，实现端到端的高效预测。

DCTST 模型首先利用多层感知机将每个变量的完整时间序列嵌入为高维表示，随后通过堆叠的编码器层进行渐进式特征提炼。每个编码器层核心的 DCAttention 模块分别执行全局注意力计算与沿通道维度的膨胀卷积操作：注意力机制通过可学习的全局查询向量以线性复杂度捕获跨变量依赖，而膨胀卷积则通过扩张滑动感知局部通道间的时间共性模式；二者输出经 ReLU 激活与残差连

接融合，再通过前馈网络与层归一化完成特征增强。最终，投影层将编码器输出的高级表示映射为未来时间步的预测值。

基于 DCTST 的多变量时间序列预测流程如下：

(1) 数据输入与嵌入：获取历史多元时间序列数据，通过嵌入层将每个变量的序列映射为高维特征向量，生成初始隐藏表示。

(2) 特征编码与交互：将嵌入表示输入多层编码器，每层通过 DCAttention 模块并行计算全局注意力权重与膨胀卷积特征，二者融合后经前馈网络非线性变换，通过残差连接与层归一化优化梯度流动。

(3) 输出预测：将最终编码层的输出特征通过投影层（MLP）映射至目标预测时间步长，生成未来序列的预测结果。

(4) 模型输出与评估：输出预测值，并采用均方误差（MSE）和平均绝对误差（MAE）等指标评估预测性能，验证模型有效性。

3 实验与分析

3.1 实验数据

为全面且严谨地评估 DCTST 模型在多元时间序列预测任务上的泛化能力与鲁棒性，本研究在涵盖了能源、交通、气象、工业运维等多个关键领域的共 13 个公开数据集及 4 个自建的合成数据集上进行了广泛的实验。这些数据集在变量规模、序列长度、采样频率以及变量间依赖关系的复杂性上均具有显著的差异性。真实世界数据集包括 ETT 系列(电变压器每小时/每 15 分钟数据)[3]、weather（气候数据）[3]、ECL（每小时电力负荷数据）[3]、Traffic（每小时高速公路占用率数据）[27]、Solar-Energy（每 10 分钟太阳能发电数据）[27]、Flight（每小时航班数据）[28]以及 PEMS 系列（每 5 分钟交通速度数据）[27]。此外，为专门检验模型捕捉复杂通道间依赖关系的能力，我们设计并构建了一个新的合成数据集 MUL，包含 XBL、Waveform、Waveform2、Multi-correlation 四个子数据集。该数据集通过引入随机事件驱动时序滞后、波形组合相关性以及多通道触发事件等显式生成模式，增强了变量间的关联性。所有数据均经过严格的预处理，包括数据清洗和标准化。本研究使用的数据统计信息详见表 1。

表 1 实验数据集描述

Table 1 Description of experimental datasets

数据集	变量数	回看窗口长度	预测长度	训练/验证/测试集大小	采样频率	领域
ETT-h1	7	96	{96,192,336,720}	{8545,2881,2881}	1 小时	能源
ETT-h2	7	96	{96,192,336,720}	{8545,2881,2881}	1 小时	能源
ETT-m1	7	96	{96,192,336,720}	{34465,11521,11521}	15 分	能源
ETT-m2	7	96	{96,192,336,720}	{34465,11521,11521}	15 分	能源
ECL	321	96	{96,192,336,720}	{18317,2633,5261}	1 小时	能源
Traffic	862	96	{96,192,336,720}	{12185,1757,3509}	1 小时	交通
Weather	21	96	{96,192,336,720}	{36792,5271,10540}	10 分	气象
Solar	137	96	{96,192,336,720}	{36601,5161,10417}	10 分	能源
Flight	7	96	{96,192,336,720}	{18317,2633,5261}	1 小时	交通
PEMS03	358	96	{12,24,48,96}	{15617,5135,5135}	5 分	交通
PEMS04	307	96	{12,24,48,96}	{10172,3375,3375}	5 分	交通
PEMS07	883	96	{12,24,48,96}	{16911,5622,5622}	5 分	交通
PEMS08	170	96	{12,24,48,96}	{10690,3548,3548}	5 分	交通
XBL	8	96	96	{20809,2905,2905}	模拟	合成
Waveform	199	96	{96,192,336,720}	{13809,1905,3905}	模拟	合成
Waveform2	32	96	{96,192,336,720}	{13809,1905,3905}	模拟	合成
Multi-correlation	105	96	{96,192,336,720}	{13809,1905,3905}	模拟	合成

3.2 实验配置

所有实验均在一台配备 NVIDIA GeForce RTX 4090 GPU（24GB 显存）的工作站上进行。实验程序使用 Python 3.8 编程语言编写，集成开发环境为 PyCharm。模型构建基于 PyTorch 深度学习框架，并利用了 NumPy、Pandas 等库进行数据预处理，Matplotlib 和 Seaborn 库进行结果可视化。

DCTST 模型的训练采用 Adam 优化器，初始学习率从集合{1×10^{-3}, 3×10^{-4}, 7×10^{-4}}中通过网格搜索确定。训练周期数固定为 10，并采用学习率衰减策略，每完成一个训练周期后将学习率减半。为确保结果的统计可靠性，每个实验均使用 5 个不同的随机种子运行，并报告其平均性能与标准差。模型的关键结构与超参数设置如表 2 所示，这些参数通过在一组保留的验证集上进行超参数搜索确定。

表 2 DCTST 模型主要参数设置

Table 2 Main parameter settings of DCTST model

所属模块	参数名	参数设置
嵌入层	嵌入维度（D）	{128,256,512,1024}
	实现方式	单层 MLP
编码器	编码器层数（L）	{1,2,3,4,5}
	层归一化	启用
	残差连接	启用
DCAttention	注意力类型	全局注意力/自注意力
	膨胀卷积层数	{1,2,3,4,5,6,7}
	卷积核大小	3
	膨胀率(d)	2
训练超参数	优化器	Adam
	损失函数	均方误差（MSE）

在模型训练过程中，我们采用早停法来防止过拟合，监控验证集上的损失，若其连续 5 个周期未提升则提前终止训练。同时，使用模型检查点回调功能，保存验证集上性能最佳的模型参数用于最终测试。

3.3 实验评价指标

为客观、定量地评估模型预测结果与真实值之间的偏差，本研究采用均方误差和平均绝对误差作为核心评价指标。

均方误差计算公式如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

平均绝对误差计算公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

其中，n 为样本数量，y_i为第 i 个样本实际值，$\hat{y}_i$为第 i 个样本预测值。在所有实验结果中，MSE 和 MAE 的值越低，表明模型的预测精度越高。此外，为了更全面地评估模型，我们还进行了深入的消融实验以验证各模块贡献，模型效率分析（包括内存使用和训练时间）以衡量其实际应用潜力，以及在数据缺失场景下的鲁棒性测试。这些综合评估共同确保了对 DCTST 模型性能多维度、多视角的严谨考察。

3.4 实验结果分析

为了全面评估 DCTST 模型在多元时间序列预测任务中的性能，我们在多个真实世界数据集和合成数据集上进行了广泛的实验，并与 7 个主流基线模型进行了比较。固定回看窗口长度 T=96，预测长度 S 根据不同数据集设置为{96,192,336,720}或{12,24,48,96}。评估指标采用均方误差（MSE）和平均绝对误差（MAE），结果取所有预测长度的平均值，系列数据集如（ETT 与 PEMS）则对子数据集的平均值进一步取平均。

表 3 多维时间序列数据集预测结果

Table 3 Prediction results on multidimensional time series datasets

数据集	指标	DCTST (自注意力)	DCTST (全局注意力)	SOFTS	TimeXer	TimeMixer	iTransformer	Crossformer	PatchTST	TSMixer
ETT(Avg)	MSE	0.373	0.371	0.375	**0.365**	0.367	0.383	0.685	0.381	0.388
	MAE	0.393	0.392	0.394	**0.388**	**0.388**	0.399	0.578	0.397	0.398
Weather	MSE	0.250	0.248	0.255	0.241	**0.240**	0.258	0.264	0.265	0.256
	MAE	0.277	0.274	0.278	**0.271**	**0.271**	0.278	0.320	0.285	0.279
ECL	MSE	**0.161**	0.163	0.174	0.171	0.182	0.178	0.244	0.190	0.186
	MAE	**0.256**	0.258	0.264	0.270	0.273	0.270	0.334	0.276	0.287
Traffic	MSE	0.410	**0.396**	0.409	0.466	0.485	0.428	0.550	0.453	0.522
	MAE	0.278	**0.266**	0.267	0.287	0.298	0.282	0.304	0.286	0.357
Solar	MSE	**0.197**	0.202	0.229	0.209	0.216	0.233	0.641	0.236	0.261
	MAE	**0.244**	0.251	0.256	0.287	0.280	0.261	0.639	0.266	0.297
PEMS	MSE	**0.101**	**0.101**	0.108	0.130	0.159	0.119	0.220	0.217	0.125

Flight	MAE	0.200	**0.199**	0.206	0.226	0.262	0.218	0.304	0.306	0.231
	MSE	0.158	0.159	0.160	0.164	0.162	**0.156**	0.380	0.168	0.182
XBL	MAE	**0.264**	0.265	0.271	0.278	0.271	0.265	0.417	0.282	0.290
	MSE	**0.011**	0.011	0.017	0.074	0.581	0.088	0.014	0.782	0.050
Waveform	MAE	**0.064**	0.064	0.065	0.181	0.472	0.200	0.074	0.886	0.157
	MSE	0.014	**0.013**	0.494	0.065	0.602	0.571	0.017	0.799	0.035
	MAE	0.048	**0.045**	0.504	0.152	0.811	0.568	0.072	0.979	0.099
Waveform 2	MSE	**0.005**	**0.005**	0.187	0.039	0.421	0.238	0.007	0.673	0.091
	MAE	0.030	**0.028**	0.229	0.106	0.571	0.294	0.051	0.862	0.186
Multi-correlation	MSE	**0.015**	0.020	0.030	0.189	0.473	0.072	0.021	0.651	0.201
	MAE	0.047	**0.044**	0.076	0.200	0.762	0.144	0.062	0.774	0.288

实验结果表明，DCTST 在绝大多数数据集上均取得了最优或接近最优的性能。在真实世界数据集方面，模型在 ECL 电力负荷数据集上 MSE 达到 0.161，在 Traffic 交通流量数据集上 MSE 为 0.396，在 Solar-Energy 太阳能发电数据集上 MSE 为 0.197，均显著优于对比基线。特别是在变量数众多的 PEMS 交通数据集上，DCTST 的 MSE 为 0.101，展现出处理高维数据的强大能力。

为了量化评估模型的计算效率，我们在回顾窗口与预测窗口均为 96 步的实验设置下，对比了各模型在 Traffic 和 ECL 数据集上的性能表现与资源消耗，实验设置批次大小为 16，记录单次迭代的平均训练时间与显存消耗。详细结果如表 4 所示。

表 4 模型计算效率与资源消耗对比表

Table 4 Comparison of model computational efficiency and resource consumption

模型	Traffic(862)			ECL(321)		
	速度（ms/iter）	显存消耗（MB）	MSE	速度（ms/iter）	显存消耗（MB）	MSE
DCTST(global)	49	3588	0.373	37	1918	0.137
DCTST(self)	109	7993	0.383	66	2807	0.136
SOFTS	58	1893	0.376	38	1341	0.143
TimeMixer	177	12579	0.462	40	2863	0.153
TimeXer	210	12335	0.428	100	5423	0.140
iTransformer	76	6003	0.395	40	1613	0.148
PatchTST	188	10045	0.427	134	6337	0.164
TSMixer	50	783	0.493	41	573	0.157

Crossformer	400	22970	0.522	110	7549	0.219

由表 4 可见，DCTST(Global)选用的非对称注意力机制保持极高预测精度，Traffic 上 MSE 最低，ECL 上 MSE 仅次于 Self 版本 0.001 的同时，展现了较高的计算效率。在 Traffic 数据集上，其训练速度相比 PatchTST 提升了约 3.8 倍，显存占用仅为 Crossformer 的约 15%。即便与以高效著称的线性模型（如 SOFTS、TSMixer）相比，DCTST(Global)在速度相当的情况下，预测精度有显著优势。证明本文提出的非对称全局注意力机制结合膨胀卷积架构，能够有效平衡模型复杂度与预测性能，避免大规模数据的注意力平方级复杂度的计算，适合大规模多维时序数据的实际部署。

在专门设计的合成数据集 MUL 上，DCTST 的优势更加明显。在 XBL 子集上 MSE 低至 0.011，在 Waveform 和 Waveform2 子集上 MSE 分别为 0.014 和 0.005，在多变量关联子集上 MSE 为 0.015，全面超越了其他对比模型。图 5 与图 6 展示了 DCTST 与各模型预测结果的可视化情况。

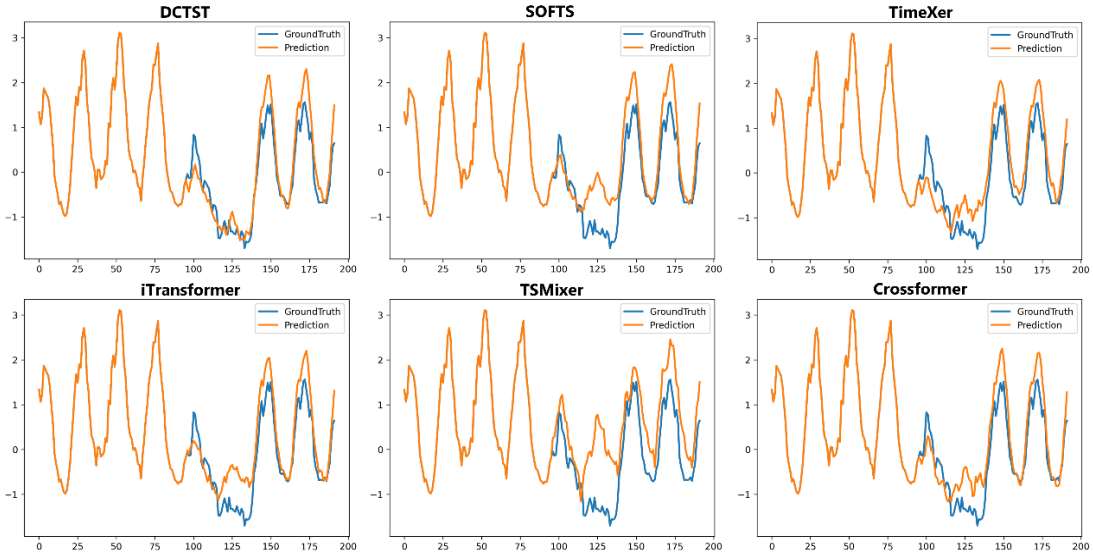


图 5 DCTST 与不同模型对 ECL 数据集预测曲线

Figure 5 Prediction curves of DCTST and different models on ECL dataset

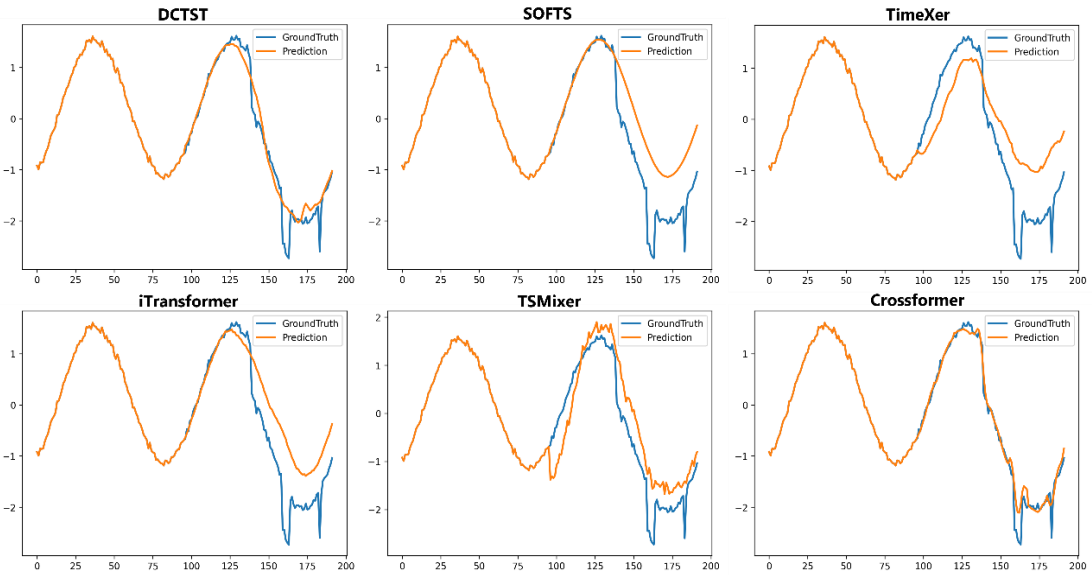


图 6 DCTST 与不同模型对 XBL 数据集预测曲线

Figure 6 Prediction curves of DCTST and different models on XBL dataset

图 7 和图 8 以箱线图的形式，分别展示了各模型在 ECL 和 XBL 数据集上各通道绝对误差和（SAE）的分布情况。图中，中位线位置越低，表明模型在大部分通道上的平均预测误差越小；箱体（IQR）越短，表示模型对不同通道的预测性能越稳定；离群点越少，说明模型在极端困难的通道上仍能保持较好的鲁棒性。

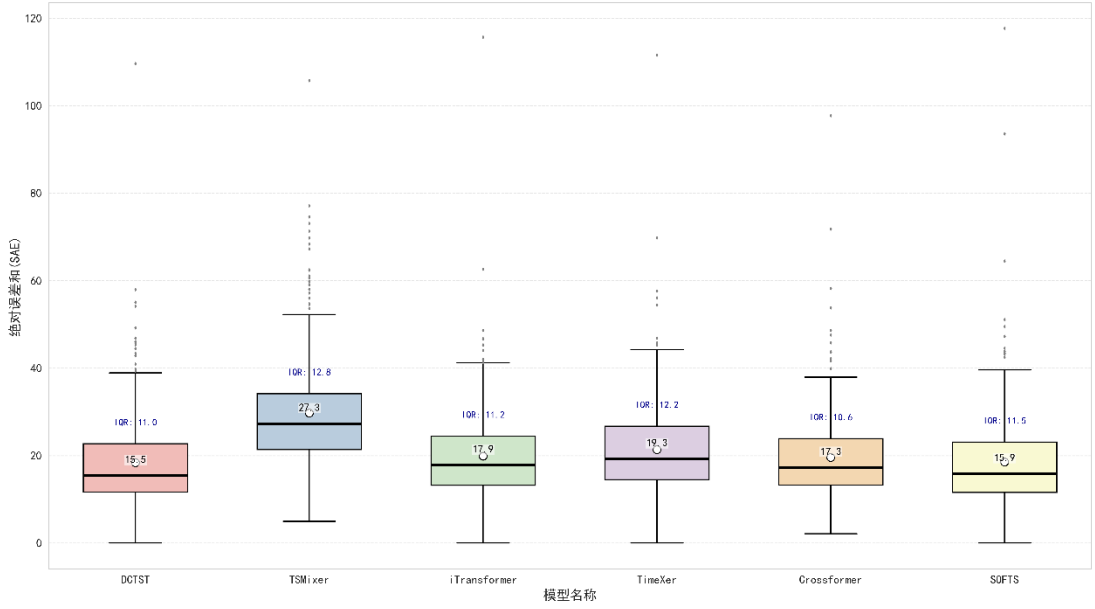


图 7 ECL 数据集各模型通道误差分布箱线图

Figure 7 Box plot of channel error distribution of each model on ECL dataset

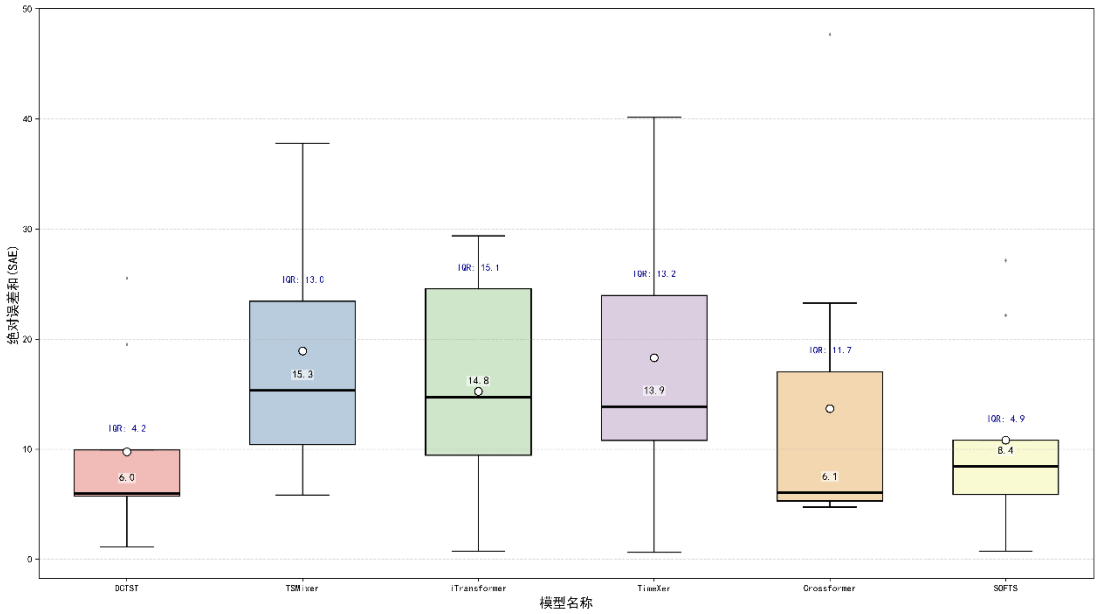


图 8 XBL 数据集各模型通道误差分布箱线图

Figure 8 Box plot of channel error distribution of each model on XBL dataset

从图 7 可以看出，DCTST 在 ECL 数据集上的中位数（15.5）显著低于 TSMixer（27.3）和 TimeXer（19.3），且四分位距（IQR=11.0）较窄，说明 DCTST 不仅整体精度高，且在不同特征的通道间表现一致性强。从图 8 的 XBL 数据集结果来看，DCTST 的优势更为明显，其中位数（6.0）和 IQR（4.2）均为所有模型中最小，且离群点极少，证明了模型在处理复杂多变量相关性时的精度与稳定性。

为了深入验证 DCTST 模型中各核心组件的有效性，我们设计了系统的消融实验。实验通过控制变量法，逐步分析和评估了 DCAttention 模块的不同设计策略、前馈网络（FFN）的作用以及残差连接方式对模型性能的影响。实验在 ECL、Traffic 和 Solar-Energy 三个具有代表性的真实世界数据集上进行，评估指标选取 MSE，“Attention + Dilated”表示并行融合；“Dilated(Attention)”表示先膨胀卷积后注意力；“Attention(Dilated)”表示先注意力后膨胀卷积。最佳结果以加粗显示。结果如表 5 所示，最好结果加粗显示。

表 5 DCAttention、FFN 和残差连接的消融实验结果（平均 MSE）

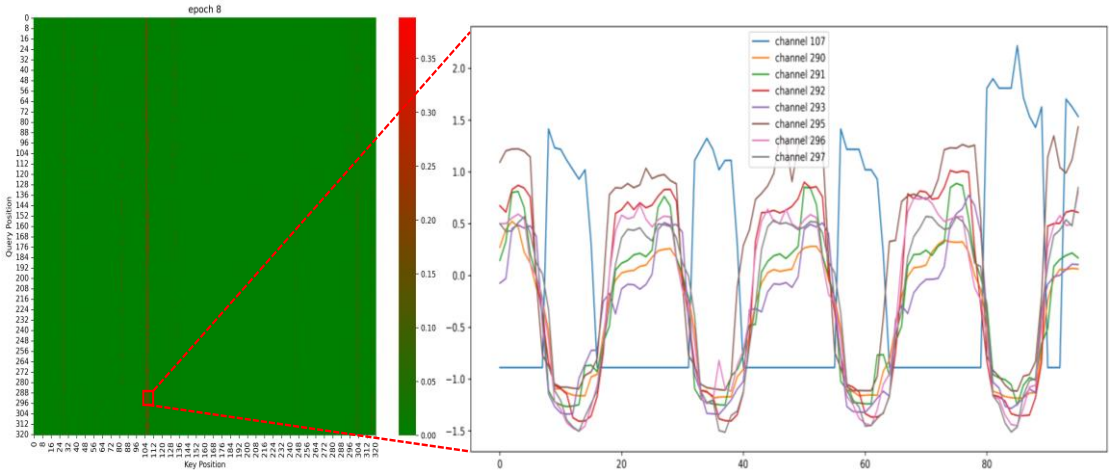
Table 5 Ablation results of DCAttention, FFN and residual connection (average MSE)

案例编号	DCAttention 配置	FFN 是否使用	残差连接方式	ECL	Traffic	Solar
1	Attention+Dilated	是	跨层	**0.161**	**0.410**	**0.197**

2	Attention+Dilated	否	跨层	0.165	0.428	0.207
3	Attention+Dilated	是	子层	0.165	**0.410**	**0.197**
4	Attention+Dilated	否	子层	0.165	0.419	0.214
5	Attention	是	子层	0.178	0.428	0.233
6	Attention	是	子层	0.184	0.436	0.209
7	Attention	否	跨层	0.186	0.436	0.207
8	Dilated	是	子层	0.169	0.442	0.208
9	Dilated	否	子层	0.169	0.453	0.199
10	Dilated	是	跨层	0.168	0.447	0.204
11	Dilated(Attention)	否	子层	0.164	0.426	0.206
12	Attention(Dilated)	否	跨层	0.165	0.413	0.198
13	Dilated(Attention)	否	子层	0.165	0.432	0.209
14	Attention(Dilated)	否	跨层	0.170	0.421	0.215
15	Dilated(Attention)	是	跨层	0.165	0.411	0.199
16	Attention(Dilated)	是	子层	0.164	0.433	0.201
17	Dilated(Attention)	是	跨层	0.168	0.420	0.202
18	Attention(Dilated)	是	子层	0.164	0.412	0.212

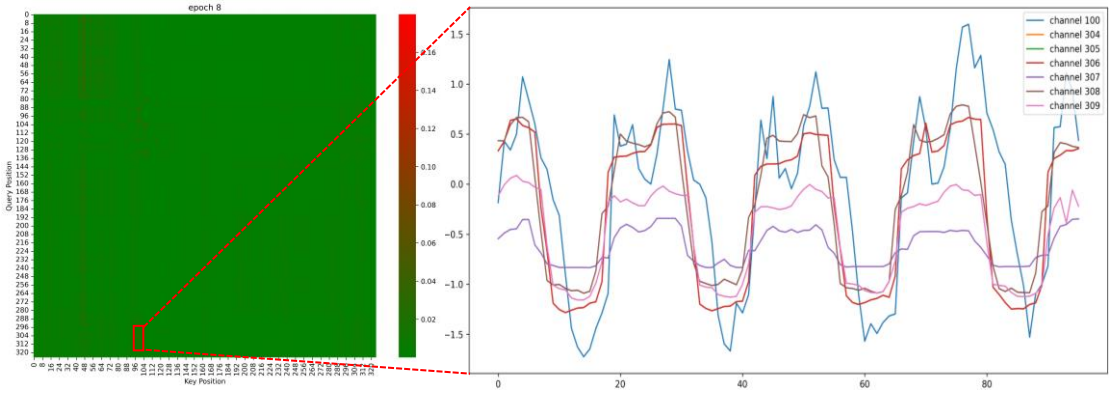
从表 5 的消融实验结果可以得出，DCTST 模型采用的注意力机制与膨胀卷积并行融合结构具有显著优势，该结构在三个基准数据集上均取得了最优性能。具体而言，完整配置在 ECL、Traffic 和 Solar-Energy 数据集上的 MSE 分别达到 0.161、0.410 和 0.197，明显优于仅使用单一组件的变体模型。

为了深入探究 DCTST 模型在通道相关性建模方面的机理，我们对其注意力图进行了可视化分析。具体而言，我们将 DCTST 中多个注意力头的注意力权重进行平均处理，以观察模型在不同变量间建立的注意力分布模式。在图 9 展示了 DCTST 与 iTransformer 在相同训练数据上的注意力图。



(a)DCTST 的注意力权重

(a) Attention weights of DCTST



(b)iTransformer 的注意力权重

(b) Attention weights of iTransformer

图 9 DCTST 与 iTransformer 的注意力图

Figure 9 Attention maps of DCTST and iTransformer

DCTST 学习到的通道间注意力权重显著高于仅使用整体序列嵌入结合注意力机制的对比模型 iTransformer。这一现象表明，通过引入膨胀卷积模块，DCTST 能够更有效地捕获变量间的深层依赖关系。从图 9 可以看出，DCTST 的注意力图呈现出更加集中和显著的模式，相似度数值也更高，而 iTransformer 的注意力分布相对分散。进一步分析发现，DCTST 对于具有周期同步性但峰谷相位相反的序列仍然保持了较高的注意力相似度。这意味着 DCTST 能够识别并利用这种特殊的同步特征，即使变量间的极值点出现时间存在偏移，但只要它们遵循相同的周期规律，模型就能建立有效的关联。相比之下，iTransformer 的注意力机制主要聚焦于峰谷完全同步的变量组合，对相位相反的同步模式捕捉能力较弱。这种差异说明膨胀卷积模块通过扩大变量之间的感受野，使模型能够识别出更复杂

的同步关系。这不仅包括简单的峰值对齐，还涵盖相位差、周期同步等更丰富的时序特征。这种能力使得 DCTST 能够从多元时间序列中提取更有意义的通道交互信息。

我们在 ECL 数据集上进行了控制变量实验，通过随机掩码的方式模拟不同严重程度的数据缺失情况。具体而言，我们在测试集上分别设置了 6.25%、12.5%、25%三种遮挡率，随机将相应比例的各个通道时间步数据置为零，然后观察各模型在残缺数据上的预测性能变化。所有模型均使用完整数据训练，仅在测试阶段引入缺失值，以模拟现实中最常见的场景。图 10 展示了实验的结果。

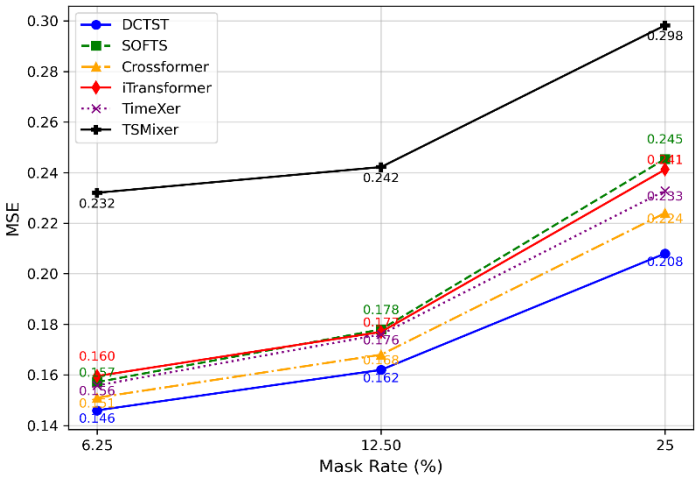


图 10 不同掩码率下的模型表现

Figure 10 Model performance under different mask rates

如图 10 所示，随着遮挡率从 6.25%增加到 25%，DCTST 的 MSE 增长曲线最为平缓，性能衰减明显小于对比模型。这种强鲁棒性主要得益于 DCTST 独特的结构设计。其膨胀卷积模块能够通过滑动窗口在局部区域内聚合特征信息，即使部分时间点数据缺失，模型仍能从相邻通道和时间的有效数据中恢复关键模式。相比之下，严重依赖全局注意力机制的模型如 iTransformer，当数据出现大规模缺失时，其注意力权重的计算会受到较大干扰，导致性能急剧下降。而 DCTST 在不同缺失率下均表现出卓越的稳定性。

4 结语

本研究针对多元时间序列预测中复杂通道相关性建模的挑战，提出了一种融合膨胀卷积与注意力机制的 DCTST 模型。通过创新性地设计并行处理的 DCAttention 模块，模型在保持线性计算复杂度的同时，有效平衡了全局依赖捕捉与局部特征提取。实验结果表明，该模型在多个基准数据集上均取得了最优性能，在预测精度、抗缺失鲁棒性和计算效率方面显著优于主流基线方法。这些优势使其在资源受限的工业场景中具备良好的应用前景。未来的研究工作将聚焦于模型在非均匀采样序列和在线学习场景中的扩展应用。

参考文献：

[1] 杨海民，潘志松，白玮. 时间序列预测方法综述[J]. 计算机科学, 2019, 46(1): 21-28.

[2] 张美英，何杰. 时间序列预测模型研究综述[J]. 数学的实践与认识, 2011, 41(18): 189-195.

[3] Haoyi Zhou,Zhang Shanghang,Peng Jieqi, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting[C]. Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 11106-11114.

[4] Haixu Wu,Xu Jiehui,Wang Jianmin, et al. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting[C]. Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021: 22419-22430.

[5] Minhao Liu,Zeng Ailing,Chen Muxi, et al. Scinet: Time Series Modeling and Forecasting with Sample Convolution and Interaction[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 5816-5828.

[6] 陈加毫，谢良，廖思灏. 基于多粒度多尺度深度时空模型的长期序列预测方法[J]. 计算机科学, 2024, 51(04): 102-113.

[7] 詹熙，黎维，潘志松. Multi-Shapelet:一种基于Shapelet的多变量时间序列分类方法[J]. 数据采集与处理, 2023, 38(02): 386-400.

[8] Ashish Vaswani,Shazeer Noam,Parmar Niki, et al. Attention is All You Need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.

[9] Chen Sian,Chunliang Li,Yoder Nathanael-C. Tsmixer: An All-Mlp Architecture for Time Series Fore- Casting[DB/OL]. https://openreview.net/forum?id=wbpxTuXgm0.

[10] Shiyu Wang,Wu Haixu,Shi Xiaoming, et al. Timemixer: Decomposable Multiscale Mixing for Time Series Forecasting[C]. Proceedings of the 12th International Conference on Learning Representations, 2024: 1-27.

[11] YanJun Zhao,Ma Ziqing,Zhou Tian, et al. Gcformer: An Efficient Framework for Accurate and Scalable Long-Term Multivariate Time Series Forecasting[C]. CIKM, 2023: 3464-3473.

[12] Tian Zhou,Ma Ziqing,Wen Qingsong, et al. Fedformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting[C]. ICML, 2022: 27257-27275.

[13] Ailing Zeng,Chen Muxi,Zhang Lei, et al. Are Transformers Effective for Time Series

Forecasting[C]. Proceedings of the 37th AAAI Conference on Artificial Intelligence, 2023: 11121-11128.

[14] Yuqi Nie,Nguyen Nam-H,Sinthong Phanwadee, et al. A Time Series is Worth 64 Words: Long-Term Forecasting with Transformers[C]. Proceedings of the 11th International Conference on Learning Representations, 2023: 1-24.

[15] Yunhao Zhang,Yan Junchi,Han Lu, et al. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting[C]. Proceedings of the 11th International Conference on Learning Representations, 2023: 1-21.

[16] Yong Liu,Hu Tengge,Zhang Haoran, et al. Itransformer: Inverted Transformers are Effective for Time Series Forecasting[C]. Proceedings of the 12th International Conference on Learning Representations, 2024: 1-25.

[17] Lu Han,Chen Xu-Yang,Ye Han-Jia, et al. Softs: Efficient Multivariate Time Series Forecasting with Series-Core Fusion[C]. Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024: 64145-64175.

[18] Sepp Hochreiter,Schmidhuber Jürgen. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[19] Kyunghyun Cho,van Merriënboer Bart,Gulcehre Caglar, et al. Learning Phrase Representations Using Rnn Encoder-Decoder for Statistical Machine Translation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1724-1734.

[20] Shaojie Bai,Kolter J-Zico,Koltun Vladlen. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling[DB/OL]. https://arxiv.org/abs/1803.01271.

[21] Xiangfei Qiu,Wu Xingjian,Lin Yan, et al. Duet: Dual Clustering Enhanced Multivariate Time Series Forecasting. in SIGKDD, 2025.

[22] Xingjian Wu,Qiu Xiangfei,Li Zhengyu, et al. Catch: Channel-Aware Multivariate Time Series Anomaly Detection Via Frequency Patching. in ICLR, 2025.

[23] Yuxuan Shu,Lampos Vasileios. Deformtime: Capturing Variable Dependencies with Deformable Attention for Time Series Forecasting[J]. Transactions on Machine Learning Research, 2025.

[24] 杨永鹏，杨震，杨真真. 基于时序分解和注意力图神经网络的交通预测[J]. 数据采集与处理, 2025, 40(02): 417-430.

[25] Yuxuan Wang,Wu Haixu,Dong Jiaxiang, et al. Timexer: Empowering Transformers for Time Series Forecasting with Exogenous Variables[C]. Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024: 469-498.

[26] Andrew Jaegle,Gimeno Felix,Brock Andy, et al. Perceiver: General Perception with Iterative Attention[C]. Proceedings of the 38th International Conference on Machine Learning, 2021: 4651-4664.

[27] Guokun Lai,Chang Wei-Cheng,Yang Yiming, et al. Modeling Long-and Short-Term Temporal Patterns with Deep Neural Networks[C]. The 41st international ACM SIGIR conference on research & development in information retrieval, 2018: 95-104.

[28] Wanlin Cai,Liang Yuxuan,Liu Xianggen, et al. Msgnet: Learning Multi-Scale Inter-Series Correlations for Multivariate Time Series Forecasting[C]. Proceedings of the 38th AAAI Conference on Artificial Intelligence, 2024: 11141-11149.

基金项目：江苏省前沿技术研发计划资助（基金号: BF2025617）
This work was funded by Frontier Technologies R&D Program of Jiangsu（No.: BF2025617).
通信作者：徐鹤（E-mail：xuhe@njupt.edu.cn）

刘尚东(1979-)，男，博士，副教授．研究方向：人工智能、网络空间安全、大数据及云计算技术．Email: lsd@njupt.edu.cn

袁帅祥（2001-），男，硕士研究生．研究方向：人工智能和大数据、深度学习．Email: 1023040816@njupt.edu.cn

徐鹤（1985-），男，博士，教授，CCF 高级会员．研究方向：大数据和人工智能、物联网技术、知识表示与推理等. Email: xuhe@njupt.edu.cn

刘思行（1991-），男，硕士，软件工程师．研究方向：人工智能、生物电信号，生物化学信号分析与处理。Email: liu.sx@yuyue.com.cn

季一木（1978-），男，博士，教授，博士生导师，CCF 会员．研究方向：高性能计算、大数据和人工智能．Email: jiym@njupt.edu.cn

Multivariate Time Series Forecasting Method Based on Dilated Convolution and Attention Mechanism

LIU Shangdong，YUAN Shuaixiang[1]，XU He[1, 2]，LIU Sixing[3]，JI Yimu[1, 2]
（1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023;）
（2. Jiangsu HPC and Intelligent Processing Engineer Research Center, Nanjing 210023）
（3. Jiangsu Yuyue Medical Equipment and Supply Co., Ltd. Zhenjiang 212300）

Abstract

Aiming at the problem of insufficient prediction accuracy of models caused by complex nonlinear dependencies and temporal shifts between channels in multivariate time series forecasting, this paper proposes a multivariate time series forecasting model based on dilated convolution and channel attention mechanism (Dilated Convolutional Time Series Transformer, DCTST). Firstly, an inverted embedding module independently encodes the entire historical sequence of each variable into high-dimensional tokens, preserving global temporal features. Secondly, a channel attention module is designed, combining a self-attention mechanism with a learnable routing strategy, and a dilated convolution module is introduced to expand the receptive field along the variable dimension to capture multi-scale local dependencies. Finally, cross-layer residual connections integrate the original input with deep-level features,

enhancing model stability. Theoretical analysis and experiments show that compared to mainstream channel-independent models (e.g., PatchTST) and channel-dependent models (e.g., iTransformer), DCTST achieves a significant reduction in both Mean Squared Error (MSE) and Mean Absolute Error (MAE) on multiple real-world and synthetic datasets. The experimental results indicate that DCTST can effectively model inter-channel dependencies, improving the prediction accuracy and robustness of multivariate time series, making it suitable for practical engineering scenarios such as smart grids and traffic flow forecasting.

Highlights:

1. Proposes DCTST, a model integrating dilated convolution and channel attention to capture complex inter-channel dependencies in multivariate time series forecasting.
2. Introduces a learnable channel attention module and a dilated convolution module, enabling effective modeling of both global temporal features and multi-scale local dependencies across variables.
3. Achieves state-of-the-art performance, significantly reducing prediction errors (MSE/MAE) over strong baselines on multiple datasets, demonstrating high robustness for practical applications like smart grids.

Key Words: Time Series Forecasting; Dilated Convolution; Transformer; Attention Mechanism