DWAFM：融合注意力和频域多层感知机的动态加权图结构嵌入，用于交通预测

Sen Shi、IEEE 会员 Zhichao Zhang 和 Yangfan He

摘要——准确的交通预测是智能交通系统的一项关键任务。其核心难点在于准确建模交通数据中复杂的时空依赖。近年来，网络架构的改进未能带来显著的性能提升，而嵌入技术则展现出巨大的潜力。然而，现有嵌入方法往往忽略图结构信息，或仅依赖静态图结构，因此难以有效捕捉节点之间随时间变化的动态关联。为解决这一问题，本文提出一种新颖的动态加权图结构（DWGS）嵌入方法，该方法依托一种能够真实反映节点间动态关联强度随时间变化的图结构。我们首先将DWGS嵌入与空间—时间自适应嵌入、时间嵌入和特征嵌入相结合，再融入注意力和频域多层感知机（MLP），据此设计了一种新颖的交通预测模型，称为融合注意力和频域多层感知机的DWGS嵌入（DWAFM）。在五个真实世界交通数据集上的实验表明，DWAFM的预测性能优于一些当前最先进的方法。

索引词——注意力、频域多层感知机、图结构嵌入、交通预测。

arXiv:2603.00997v1 [cs.LG] 1 Mar 2026

一、引言

TRAFFIC 预测 [1] 是一项时空序列预测任务，根据历史时空观测数据推断未来的交通状态（如流量和速度）。近年来，深度学习在这一领域取得了显著进展。时空图神经网络（STGNN） [2]–[6] 以及基于注意力机制的模型 [7]–[10]凭借有效捕捉复杂时空依赖关系的卓越能力，受到了广泛关注。为提高预测准确率，研究重点转向设计更复杂的模型架构，例如新型图卷积 [11]–[16]和图结构学习 [17]–[23]。然而，研究人员逐渐发现，网络架构的改进往往无法显著提升模型性能。这一现象促使研究方向从复杂架构转向探索数据表示技术。

在此背景下，输入嵌入技术作为数据表示的关键组成部分，展现出巨大潜力，并已成为研究热点。具体而言，Shao等人提出了STID，通过引入独立的时空身份嵌入并将其与多层感知机（MLP）架构相结合，在保持高效率的同时取得了出色的性能 [24]。Liu等人提出了STAEformer，其设计的时空自适应嵌入组件显著提升了时空信息的表示能力 [25]。与此同时，Wang等人将预定义的在其 ST-MLP 研究中，将静态图结构信息融入空间表示，进一步提升了模型在时空预测任务中的性能 [26]。这些成功实践凸显了精心设计的嵌入技术在简化模型架构的同时提升性能的有效性。然而，这些方法要么完全避免使用预定义图结构，导致空间关系建模不足；要么引入预定义的静态图结构，无法准确反映早晚高峰等场景中节点关联强度的动态变化。

为克服这一关键局限，本文提出一种创新的动态加权图结构（DWGS）嵌入方法。在保留预定义图拓扑的同时，该方法以时空数据驱动的方式自适应学习图中每条边的时变权重，从而更准确地刻画时空数据集中隐含的动态空间依赖。此外，考虑到此类 DWGS 在表示节点间复杂空间关系方面可能存在局限，我们引入一种时空自适应嵌入，以捕捉数据中潜在的空间相关模式。总体而言，通过结合 DWGS嵌入与时空自适应嵌入，我们构建了一种统一且表达能力更强的空间表示，增强了对复杂时空依赖进行建模的能力。

最后，本文提出了一种新颖的交通预测模型，称为融合注意力和频域多层感知机的DWGS嵌入（DWAFM）。该模型结合时间嵌入和特征嵌入，整合注意力与频域多层感知机的协同工作机制，共同学习复杂的时空依赖关系，从而实现高性能交通预测。本研究的主要贡献总结如下：

- 本文提出了一种DWGS嵌入方法，通过数据驱动的方式自适应地学习动态加权邻接矩阵，以刻画节点间相关强度随时间的变化。
- 本文构建了一种新颖的交通预测模型DWAFM，有效结合了自注意力在空间关系建模方面的优势，以及频域多层感知机在时间模式提取方面的优势，从而高效建模复杂的时空依赖关系。

II. 预备知识

定义1. 交通网络：交通网络被建模为一个无向图 $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathrm{A})$，它将物理道路基础设施映射到拓扑空间中，其中顶点集 $\boldsymbol{V} = \{v_1, v_2, \ldots, v_N\}$表示 N 离散观测点，这些观测点对应于部署在道路路段或交叉路口处的物理传感器（例如线圈检测器或监控摄像头），用于记录连续的时间序列数据。边集 $\boldsymbol{E}$ 根据物理道路布局定义连通关系：当且仅当观测位置之间存在直接可通行的道路路段时，边 e_{ij} 才存在；邻接矩阵$\mathrm{A} \in \mathbb{R}^{N\times N}$则刻画网络内的空间相关性，其中每个元素$\mathrm{A}_{i,\ j}$量化了位置间的关联程度（例如空间邻近性或历史交通相关性）：若位置之间存在物理连接，则该值表示其关联程度；否则为0。

定义 2. 交通预测：交通预测的目标是利用过去 T 个时间段的交通数据 $\mathbf{X} \in \mathbb{R}^{T\times N\times D}$，预测未来 T_f 个时间段的数据 $\mathbf{Y} \in \mathbb{R}^{T_f\times N\times D}$。该交通预测任务定义为

$$\mathbf{X} \xrightarrow{F} \mathbf{Y},$$

其中 F 表示一个将历史数据映射到未来数据的学习函数， D 是交通数据特征的维度。

本研究部分得到了湖北大学湖北省应用数学重点实验室开放基金（项目编号：HBAM202404）的资助；部分得到了教育部系统控制与信息处理重点实验室基金（项目编号：Scip20240121）的资助；部分得到了南京工程学院人才引进启动基金（项目编号：YKJ202214）的资助。（通讯作者：Zhichao Zhang。）

Sen Shi 任职于中国南京南京信息工程大学数学与统计学院，邮编 210044（电子邮箱：202312380032@nuist.edu.cn）。

Zhichao Zhang 任职于中国南京南京信息工程大学数学与统计学院，邮编 210044；同时任职于中国武汉湖北大学湖北大学应用数学重点实验室，邮编 430062；并任职于中国上海上海交通大学教育部系统控制与信息处理重点实验室，邮编 200240（电子邮箱：zzc910731@163.com）。

Yangfan He 任职于中国南京市南京理工学院通信与人工智能学院、集成电路学院（邮编：211167），同时任职于中国南京市江苏省智能感知技术与系统工程研究中心（邮编：211167）（电子邮箱：Yangfan.He@njit.edu.cn）。

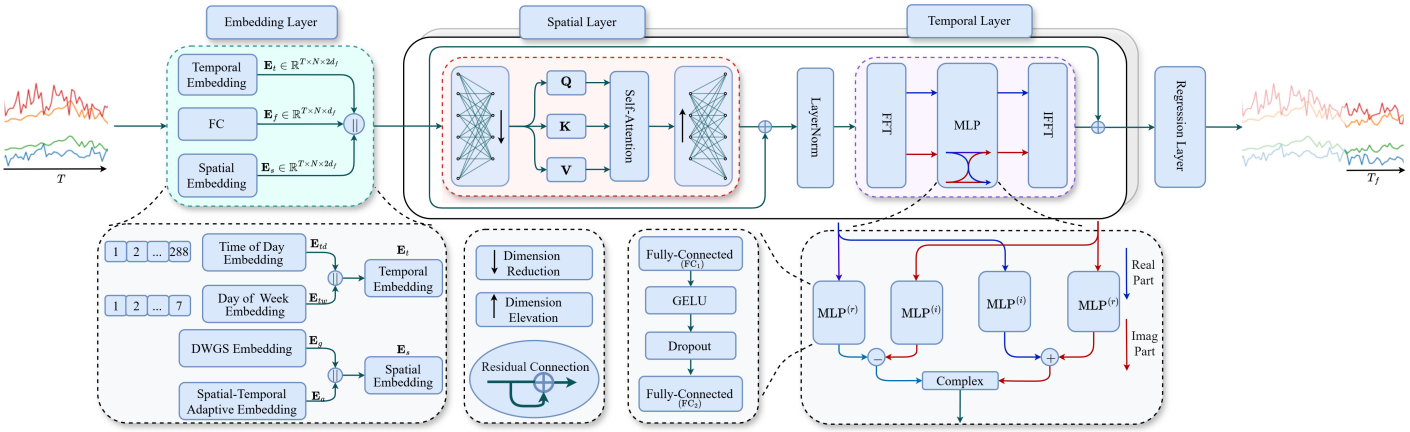


图 1：所提出的 DWAFM 总体框架。

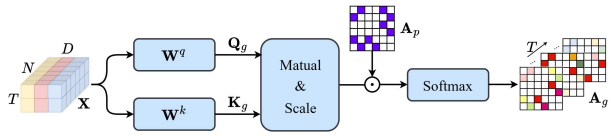


图 2：动态加权邻接矩阵的学习机制。

III. 方法

A. 总体框架

如图1所示，我们的模型由嵌入层、空间层、时间层和回归层组成。嵌入层融合特征嵌入、时间嵌入和空间嵌入以生成隐藏表示。空间层和时间层分别用于捕获交通序列之间的空间关系以及序列内部的时间依赖关系。回归层最终完成预测。

B. 嵌入层

嵌入层主要用于将原始输入数据映射到高维空间，以便更好地捕捉复杂的时空依赖。为保留原始数据 X 中的特征信息，首先将其转换为特征嵌入：

$$\mathbf{E}_f = \mathrm{FC}(\mathbf{X}) \in \mathbb{R}^{T\times N\times d_f},$$

其中，FC(·) 表示一个全连接层，d_f 表示嵌入维度。

对于时间嵌入，我们定义了两种可学习表示，用于捕捉交通数据中的周期性和趋势特征：$\mathrm{W}_d \in \mathbb{R}^{N_d\times d_f}$ 用于编码日内模式，$\mathrm{W}_w \in \mathbb{R}^{N_w\times d_f}$用于编码周内模式，其中，$N_d$ 和 N_w 分别表示一天中的时间间隔数和一周中的天数。对于每个时间步 t，根据其具体的时间位置，从这些表示中提取相应的嵌入向量。由于所有交通节点共享相同的时间特征，因此将提取的时刻和星期几嵌入沿节点维度复制并扩展，从而得到 $\mathrm{E}_{td} \in \mathbb{R}^{T\times N\times d_f}$ 和 $\mathrm{E}_{tw} \in \mathbb{R}^{T\times N\times d_f}$。随后，将 E_{td} 和 E_{tw} 拼接，形成时间嵌入

$$\mathbf{E}_t = \mathrm{Concat}\left[(\mathbf{E}_{td}, \mathbf{E}_{tw}), \mathrm{dim} = -1\right] \in \mathbb{R}^{T\times N\times 2d_f}.$$

对于空间嵌入，传统的图结构嵌入方法通常依赖预定义邻接矩阵，无法反映节点之间关联强度的动态变化。因此，我们提出了一种基于现有预定义邻接矩阵构建的动态加权邻接矩阵，用于捕捉节点间关联强度随时间的变化。如图 2 所示，我们对原始数据 X 使用自注意力，以学习连接权重的时间动态。该过程定义如下：

$$\mathbf{Q}_g = \mathbf{X}\mathbf{W}^q, \mathbf{K}_g = \mathbf{X}\mathbf{W}^k, \mathbf{A}_a = \mathrm{Softmax}\left(\frac{\mathbf{Q}_g\mathbf{K}_g^\top}{\sqrt{d_f}} \odot \mathbf{A}_p\right),$$

其中，W^q、$\mathrm{W}^k \in \mathbb{R}^{D\times d_f}$ 是可学习参数，A_p 是预定义邻接矩阵。为确保 A_p 中非连接位置的注意力分数为零，我们在应用 Softmax 之前，将注意力逻辑值中对应位置设为负无穷大。A_a 表示注意力分数，其为非对称矩阵。为构造有效的邻接矩阵，我们对 A_a 进行对称化，得到动态加权邻接矩阵 $\mathrm{A}_g = (\mathrm{A}_a + \mathrm{A}_a^\top)/2 \in \mathbb{R}^{T\times N\times N}$。随后，将可学习矩阵 $\mathrm{B} \in \mathbb{R}^{N\times d_f}$ 与 A_g 相乘，得到 DWGS嵌入

$$\mathbf{E}_g = \mathbf{A}_g\mathbf{B} \in \mathbb{R}^{T\times N\times d_f}.$$

考虑到仅依赖 E_g 可能不足以捕捉节点之间复杂的空间关系，我们引入空间-时间自适应嵌入 $\mathrm{E}_a \in \mathbb{R}^{T\times N\times d_f}$ 来捕捉潜在的空间关系，进一步丰富空间表示。随后，将 E_g 和E_a 拼接起来，形成空间嵌入

$$\mathbf{E}_s = \mathrm{Concat}\left[(\mathbf{E}_g, \mathbf{E}_a), \mathrm{dim} = -1\right] \in \mathbb{R}^{T\times N\times 2d_f}.$$

将上述所有嵌入向量拼接起来，即可得到最终的隐藏空间-时间表示

$$\mathbf{Z} = \mathrm{Concat}\left[(\mathbf{E}_f, \mathbf{E}_s, \mathbf{E}_t), \mathrm{dim} = -1\right] \in \mathbb{R}^{T\times N\times d_h}.$$

隐藏维度 d_h 设为 $5d_f$。

C. 空间层

受自注意力 [27], 的启发，我们引入了一种基于降维、自注意力和升维三阶段架构的高效空间依赖建模模块。该设计在保持计算效率的同时，有效捕捉序列间复杂的空间关系。给定隐藏的空间-时间表示 Z，我们首先使用两个一维卷积层（一维卷积神经网络）进行降维，将 $T \times d_h$维压缩为节点级表示

$$\mathbf{Z}_r = \mathrm{Conv1d}(\mathrm{ReLU}(\mathrm{Conv1d}(\mathbf{Z}))) \in \mathbb{R}^{N\times d_h}.$$

随后，我们在 Z_r上执行自注意力。缩放因子设为 $\frac{1}{\sqrt{d_h/2}}$，这一~~配置~~经超参数分析验证，可实现最佳性能。具体过程如下：

$$\mathbf{Q} = \mathbf{W}_q\mathbf{Z}_r, \mathbf{K} = \mathbf{W}_k\mathbf{Z}_r, \mathbf{V} = \mathbf{W}_v\mathbf{Z}_r,$$

$$\mathbf{Z}_s = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h/2}}\right)\mathbf{V},$$

其中，W_q、W_k、$\mathrm{W}_v \in \mathbb{R}^{d_h\times d_h}$ 为可学习参数。随后，通过另外两个一维卷积神经网络将得到的注意力输出 $\mathrm{Z}_s \in \mathbb{R}^{N\times d_h}$ 上采样回原始的时空维度，从而得到

$$\mathbf{Z}_e = \mathrm{Conv1d}(\mathrm{ReLU}(\mathrm{Conv1d}(\mathbf{Z}_s))) \in \mathbb{R}^{T\times N\times d_h}.$$

最后，我们引入残差连接，并进行层归一化，以生成时间层的输入表示 Z_t。即

$$\mathbf{Z}_t = \mathrm{LayerNorm}(\mathbf{Z}_e + \mathbf{Z}) \in \mathbb{R}^{T\times N\times d_h}.$$

这种结构化方法能够以较低的计算开销有效提取空间特征。

D. 时间层与回归层

在交通预测中，捕捉时间依赖关系通常通过循环神经网络 [2], [28], 时间卷积网络 [3], [4], 或注意力机制 [7], [29]来实现。近期研究 [30]表明，频域多层感知机在建模时间关系方面具有显著优势。受此启发，本文通过采用独立的多层感知机模块交叉处理复频谱的实部和虚部，改进了频域多层感知机架构，从而有效提取频域特征。具体而言，我们沿着 $\mathbf{Z}_t$ 的时间维度应用快速傅里叶变换（FFT，记为 $\mathcal{F}$），得到复数表示 $\mathbf{Z}_f = \mathcal{F}(\mathbf{Z}_t)$。随后，通过参数相互独立的多层感知机对实部 $\mathbf{Z}_f^r$ 和虚部 $\mathbf{Z}_f^i$进行交叉运算处理，得到

$$\hat{\mathbf{Z}}_f = \mathrm{MLP}^{(r)}(\mathbf{Z}_f^r) - \mathrm{MLP}^{(i)}(\mathbf{Z}_f^i) + j\left[\mathrm{MLP}^{(i)}(\mathbf{Z}_f^r) + \mathrm{MLP}^{(r)}(\mathbf{Z}_f^i)\right],$$

该结构能够更好地适应频域分量的复杂结构，并从全局视角有效捕捉关键时间模式。每个多层感知机的结构基于 MLP-Mixer[31] andis 定义如下：

$$\mathrm{MLP}(\mathbf{Z}_f^{r(i)}) = \mathrm{FC}_2\big(\mathrm{Dropout}(\mathrm{GELU}(\mathrm{FC}_1(\mathbf{Z}_f^{r(i)})))\big).$$

随后，对 $\hat{\mathbf{Z}}_f$ 应用逆快速傅里叶变换（IFFT，以 $\mathcal{F}^{-1}$ 表示），得到时间层的结果 $\hat{\mathbf{Z}}_t = \mathcal{F}^{-1}(\hat{\mathbf{Z}}_f) \in \mathbb{R}^{T\times N\times d_h}$。

最后，回归层生成预测结果由空间层和时间层处理后的最终输出 $\mathbf{Z}_{st} \in \mathbb{R}^{T\times N\times d_h}$ 得到。回归层定义如下：

$$\hat{\mathbf{Y}} = \mathrm{FC}(\mathbf{Z}_{st}).$$

此处，全连接层 $\mathrm{FC}(\cdot)$ 将 $\mathbf{Z}_{st}$ 中的维度 $T \times d_h$ 降至 $\hat{\mathbf{Y}}$ 中的维度 $T_f \times 1$。

IV. 实验结果与分析

A. 实验设置

数据集：我们在五个真实世界交通数据集上进行实验，包括三个高速公路交通流量数据集（PEMS03、PEMS04 和 PEMS08）以及两个交通速度数据集（PEMSD7(L) 和 PEMSD7(M)）。这些数据集的详细信息见补充材料的附录 A。

基线方法：本研究采用多种公开可用的基准方法进行比较，包括传统方法（HI [32]）、典型深度学习方法（STGCN [4], GWNet [3], StemGNN [17],STNorm [33], STID [24]），以及近两年提出的一些最新方法（DGCRN [21], MegaCRN [34], STWave [35], STAEformer [25], DFDGCN [36]）。在此，所有方法均在 BasicTS 平台 [37] 上实现，以确保公平且一致的评估。

评价指标：为评估模型的性能，我们采用了几种常用的回归评价指标，包括平均绝对误差（MAE）、均方根误差（RMSE）和平均绝对百分比误差（MAPE（%））。

模型设置：实验在配备一颗 Intel(R) Core(TM) i5-12400F CPU 和一张 NVIDIA RTX 4060Ti GPU 的计算机环境中进行。模型使用 Adam 优化器进行训练，初始学习率为 0.001，批次大小为 64。训练轮次为 80。五个数据集按照 6:2:2 的比例划分为训练集、验证集和测试集。进行预测时，我们将输入 T 和输出 T_f 的长度均设为 12 个时间步。详细的模型参数和代码已公开，见：https://github.com/ssnuist/DWAFM。

B. 预测结果

模型性能对比结果如表 I 所示，其中列出了各方法在连续 12 个时间间隔上的平均性能。最佳结果以粗体突出显示，次佳结果以下划线标出。DWAFM 在五个数据集的大多数评价指标上均取得了更优性能。尤其在 PEMS08 和 PEMSD7(M) 上，它在所有评估指标上均优于所有最先进的时空图神经网络和基于 Transformer 的方法。需要注意的是，尽管 DWAFM 在某些数据集上偶尔会产生略高的平均绝对百分比误差，但这源于平均绝对百分比误差对极小真实值（例如深夜低交通流量）固有的数学敏感性。由于我们基于平均绝对误差的训练目标优先准确拟合高峰时段的交通流量趋势，低流量时段的微小绝对偏差会被平均绝对百分比误差的计算不成比例地放大。此外，一些基线方法（例如 STAEformer）也表现出具有竞争力的性能。

C. 消融实验

为评估模型各个组件的性能，我们在 PEMS08 和 PEMSD7(M) 数据集上使用七种模型变体开展消融实验：

- 不使用 A_g。将动态加权邻接矩阵 A_g 替换为预定义邻接矩阵 A_p。
- 不使用 E_g。移除 DWGS嵌入 E_g。
- 不使用 E_s。移除空间嵌入 E_s。
- 不使用 E_t。移除时间嵌入 E_t。
- 不使用快速傅里叶变换。移除快速傅里叶变换和快速傅里叶逆变换模块，并将时间层替换为单个多层感知机架构。
- 不使用空间层。移除空间层。
- 不使用时间层。移除时间层。

如图 3 所示，模型的所有组件都有助于显著提升预测准确率。在嵌入方法中，移除 E_g 或 E_s 都会导致性能下降，这证实了通过自注意力学习 DWGS 并将其与丰富的空间特征（例如 E_a）相结合以提升性能的有效性。值得注意的是，移除快速傅里叶变换模块会导致性能下降，凸显了该模块在捕捉时间模式方面的重要作用。

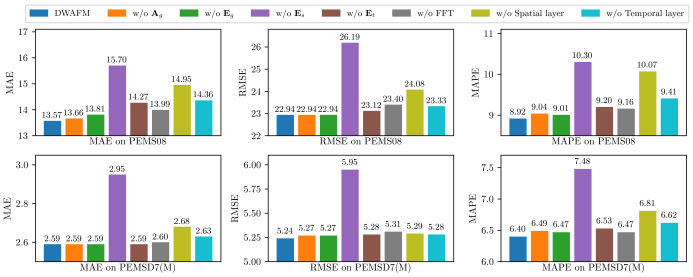


图 3：PEMS08 和 PEMSD7(M) 上的消融研究。

D. 参数敏感性分析

为探究核心超参数的影响，我们在 PEMS08 数据集上评估模型对嵌入维度 d_f 和空间自注意力缩放因子的敏感性。如图 4(a) 所示，平均绝对误差（MAE）通常会随着 d_f 增大而降低，并在 $d_f = 20$ 时达到最优。这表明，PEMS08 中复杂的交通流量模式需要更高维度的嵌入，才能捕获足够丰富的潜在特征。

对于缩放因子，我们没有采用标准 Transformer 缩放方式 $1/\sqrt{d_h}$，而是调整了该因子，以更好地适应交通数据固有的强局部相关性。实验结果表明，缩放因子设为 $1/\sqrt{d_h/2}$ 时，PEMS08 上的性能最佳。此调整能有效使注意力机制集中于高度相关的邻近节点，同时抑制来自距离较远、相关性较弱的传感器的噪声，从而增强模型对局部空间依赖关系的建模能力。此外，在 PEMSD7(M) 数据集上的参数敏感性评估揭示了最优设置存在数据集特异性差异。另外，我们还分析了模型对节点嵌入初始化的敏感性，详细内容和讨论见补充材料的附录 B。

E. 效率研究

在 PEMS08 上，以训练时间（每个 epoch 的秒数）、平均绝对误差和内存消耗（以气泡面积表示）为指标，将所提出的 DWAFM 的计算效率与最先进模型进行了比较。如图 4(b) 所示，与基于 Transformer 的架构（如 STAEformer）和主流图卷积方法（如 DGCRN）相比，DWAFM 实现了更优的性能—效率权衡。值得注意的是，尽管 DWAFM 的计算成本并非绝对最低，但其在不过度增加开销的情况下实现了显著的性能

表 I：性　　能比较：基线方法与所提出的 DWAFM 在五个常用真实世界交通数据集上的表现　　ets.

数据集	指标	HI	STGCN	GWNet	StemGNN	STNorm	STID	DGCRN	MegaCRN	STWave	STAEformer	DFDGCN	DWAFM
PEMS08	MAE	34.66	16.86	14.42	16.08	15.47	14.21	14.90	14.87	13.78	13.69	14.21	**13.57**
	RMSE	50.54	26.10	23.38	25.49	25.13	23.36	23.76	23.87	23.68	23.06	23.45	**22.94**
	MAPE	21.63	11.62	9.33	10.19	9.93	9.38	10.17	9.58	9.36	9.03	9.37	**8.92**
PEMS04	MAE	42.35	19.89	18.76	21.01	18.90	18.46	18.96	18.83	18.28	18.21	18.62	**18.14**
	RMSE	61.66	31.59	30.19	32.96	30.50	29.98	30.77	30.54	29.98	30.57	30.01	**29.72**
	MAPE	29.22	13.61	13.06	14.51	13.00	12.55	12.87	12.83	12.25	**12.21**	12.94	12.37
PEMS03	MAE	32.62	15.99	14.63	15.62	15.20	15.47	**14.53**	14.68	14.93	15.31	14.57	14.81
	RMSE	49.89	27.78	25.88	26.36	25.94	26.77	25.75	26.24	26.04	26.65	24.91	**23.96**
	MAPE	30.60	15.36	14.59	15.03	**14.55**	16.37	14.97	14.91	15.40	15.45	14.89	15.66
PEMSD7(M)	MAE	5.02	2.73	2.67	2.91	2.61	2.61	2.70	2.61	2.61	2.60	2.77	**2.59**
	RMSE	9.58	5.47	5.39	5.71	5.34	5.35	5.53	5.42	5.33	5.31	5.71	**5.24**
	MAPE	12.31	6.97	6.79	7.35	6.59	6.63	6.87	6.47	6.57	6.53	7.00	**6.40**
PEMSD7(L)	MAE	5.31	3.02	3.13	3.10	2.85	2.90	3.02	2.97	2.85	**2.79**	2.89	**2.79**
	RMSE	10.07	5.97	6.34	6.07	5.84	5.82	6.26	6.11	5.83	5.71	5.85	**5.70**
	MAPE	13.15	7.64	7.97	7.96	7.24	7.38	7.83	7.40	7.30	**7.05**	7.35	7.09

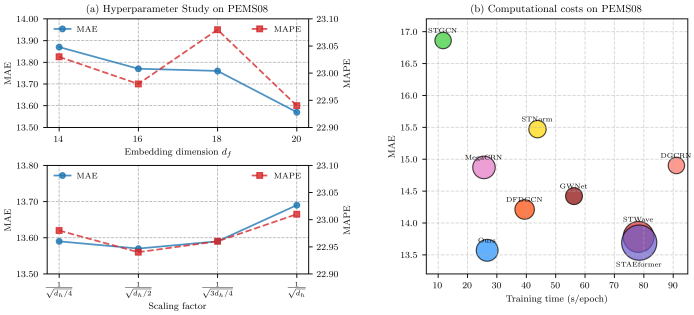


图 4：PEMS08 上的超参数研究与计算成本分析。

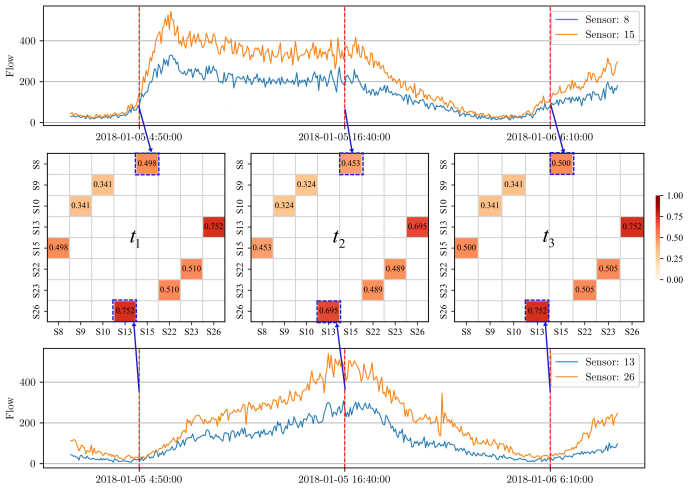


图 5：学习得到的动态加权邻接矩阵的可视化。

提升，展现出预测准确性与实际部署效率之间的出色平衡。此外，补充材料的附录 B 中提供了在 PEMS04 上进行的额外效率评估，进一步证实了 DWAFM 在性能与效率之间的良好平衡。

F. 时间建模模块的有效性

为评估所提出的频域多层感知机（Fre-MLPs），我们将其与卷积神经网络和多头自注意力机制进行了比较。如表 II 所示，在 PEMS08 和 PEMSD7(M) 数据集上，Fre-MLPs 的各项误差（平均绝对误差、均方根误差和平均绝对百分比误差）始终最低。借助频域变换，Fre-MLPs 能够有效捕获长程依赖和周期特征，克服卷积神经网络局部感受野的局限，以及自注意力机制中常见的冗余噪声问题。此外，Fre-MLPs 在表示精度与计算效率之间实现了更优的平衡。自注意力的计算复杂度较高（例如，在 PEMSD7(M) 上需要 23.40 秒），而 Fre-MLPs 的运行速度为 19.53 秒，与轻量级卷积神经网络（19.48 秒）基本相当，同时预测性能显著提升。

G. 可视化

为验证学习得到的 A_g 的有效性，我们在 PEMS04 上对其行为进行可视化和分析。我们选取两组传感器对（S8 和 S15，以及 S13 和 S26），以探究推断出的

表 II：不同时间建模模块的性能比较。

模型	指标 (PEMS08)				指标 (PEMSD7(M))			
	MAE	RMSE	MAPE	时间/秒	MAE	RMSE	MAPE	时间/秒
CNN	13.83	23.70	9.16	**26.63**	2.60	5.27	6.51	**19.48**
注意力	13.85	23.54	9.14	26.85	2.73	5.26	6.91	23.40
Fre-MLPs	**13.57**	**22.94**	**8.92**	26.71	**2.59**	**5.24**	**6.40**	19.53

相关强度与交通流实际变化之间的关系。如图 5 所示，S8 和 S15 之间的相关强度与其交通流相似性变化一致。在时间 t_1，高度相似的流量值意味着强相关性。到 t_2 时，流量出现分化，导致相关强度明显下降。在 t_3 时，流量重新趋于一致，相关性也随之恢复。传感器 S13 和 S26 也呈现出类似模式。这些结果表明，本文构建的 DWGS 能够有效捕捉传感器节点之间交通流相关性的动态演变。可视化结果验证了关联强度与观测数据之间的相关性假设，表明该模型对真实交通状况的变化具有良好的灵敏度和适应性。

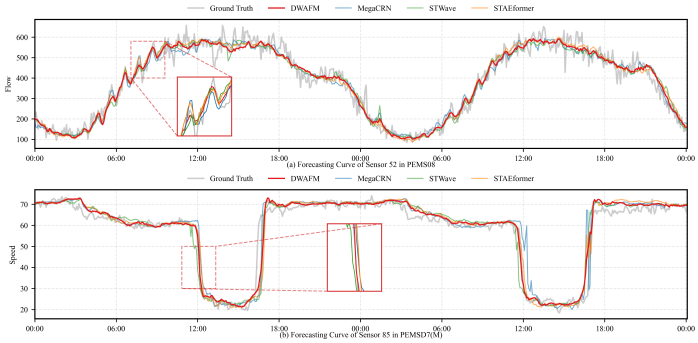


图6：真实交通预测结果的可视化。

在证明模型具备有效捕捉动态空间依赖的能力后，我们进一步通过直观地比较模型在 PEMS08 和 PEMSD7(M) 测试集上的预测值与真实值，评估其最终预测性能。具体而言，我们选取 PEMS08 的节点52和 PEMSD7(M) 的节点85作为代表，对 DWAFM 与若干先进模型（包括 MegaCRN、STWave 和 STAEformer）的预测曲线进行比较分析。

如图6所示，所提出的 DWAFM 在处理复杂交通动态方面表现出显著优势。在 PEMS08 流量上升且波动剧烈的时段，DWAFM 能够更准确地跟踪真实值的变化趋势。此外，在 PEMSD7(M) 的速度预测任务中，DWAFM 始终优于 STAEformer、STWave 和 MegaCRN 等竞争方法，尤其是在应对速度骤降（表明发生拥堵）以及速度稳定处于高速区间的时段时。

V. 结论

本研究聚焦于数据表示学习的一个核心方面，即输入嵌入技术。我们提出了一种新颖的DWGS嵌入方法，并介绍了DWAFM这一预测模型，该模型在五个交通基准上均取得了当前最先进的性能。实验分析证实，我们的嵌入方法所学习的图结构能够准确捕捉现实世界中随时间变化的节点间关系，而DWAFM能够有效捕捉序列内部及序列之间的依赖关系。未来的工作将致力于探索更加简洁高效的架构设计。

参考文献

[1] R. Jiang、D. Yin、Z. Wang、Y. Wang、J. Deng、H. Liu、Z. Cai、J. Deng、X. Song 和 R. Shibasaki，“DL-Traff：用于城市交通预测的深度学习模型综述与基准”，载于第30届 ACM 信息与知识管理国际会议（CIKM）论文集，2021年，第4515–4525页。[2] Y. Li、R. Yu、C. Shahabi 和 Y. Liu，“扩散卷积循环神经网络：数据驱动的交通预测”，载于学习表征国际会议（ICLR）论文集，2018年，第1–16页。[3] Z. Wu、S. Pan、G. Long、J. Jiang 和 C. Zhang，“用于深度时空图建模的图 WaveNet”，载于第28届人工智能国际联合会议（IJCAI）论文集，2019年，第1907–1913页。[4] B. Yu、H. Yin 和 Z. Zhu，“时空图卷积网络：用于交通预测的深度学习框架”，载于第27届人工智能国际联合会议（IJCAI）论文集，2018年，第3634–3640页。[5] Z. Shao、Z. Zhang、F. Wang 和 Y. Xu，“预训练增强的时空图神经网络用于多变量时间序列预测”，载于第28届 ACM SIGKDD 知识发现与数据挖掘国际会议（KDD）论文集，2022年，第1567–1577页。[6] J. Kong、X. Fan、M. Zuo、M. Deveci、X. Jin 和 K. Zhong，“ADCT-Net：通过双图交叉融合 Transformer 实现的自适应交通预测神经网络”，《信息融合》，第103卷，第102122页，2024年。[7] C. Zheng、X. Fan、C. Wang 和 J. Qi，“GMAN：用于交通预测的图多重注意力网络”，载于 AAAI 人工智能会议论文集，第34卷，第1期，2020年，第1234–1241页。[8] S. Guo、Y. Lin、H. Wan、X. Li 和 G. Cong，“学习时空图数据的动态性与异质性以进行交通预测”，《IEEE 知识与数据工程汇刊》（TKDE），第34卷，第11期，第5415–5428页，2021年。[9] W. Zheng、H. F. Yang、J. Cai、P. Wang、X. Jiang、S. S. Du、Y. Wang 和 Z. Wang，“将交通科学与表示学习相结合，实现全市路网拥堵预测”，《信息融合》，第99卷，101837，2023年，第。[10] J. Kong、X. Fan、M. Zuo、M. Deveci、X. Jin 和 K. Zhong，“ADCT-Net：通过双图交叉融合 Transformer 实现的自适应交通预测神经网络”，《信息融合》，第103卷，第102122页，2024年。[11] Z. Fang、Q. Long、G. Song 和 K. Xie，“用于交通流量预测的时空图常微分方程网络”，载于第27届 ACM SIGKDD 知识发现与数据挖掘会议（KDD）论文集，2021年，第364–373页。[12] W. Kong、Z. Guo 和 Y. Liu，“用于交通流量预测的时空枢纽图神经网络”，载于 AAAI 人工智能会议论文集，第38卷，第8期，2024年，第8627–8635页。[13] L. Zhao、Y. Song、C. Zhang、Y. Liu、P. Wang、T. Lin、M. Deng 和 H. Li，“T-GCN：用于交通预测的时间图卷积网络”，《IEEE 智能交通系统汇刊》（T-ITS），第21卷，第9期，第3848–3858页，2019年。[14] C. Song、Y. Lin、S. Guo 和 H. Wan，“时空同步图卷积网络：用于时空网络数据预测的新框架”，载于 AAAI 人工智能会议论文集，第34卷，第1期，2020年，第914–921页。[15] A. Liu 和 Y. Zhang，“通过交互式学习实现的用于交通预测的时空动态图卷积网络”，《IEEE 智能交通系统汇刊》（T-ITS），第25卷，7660，2024年7月，第7645–页。[16] M. Sabbaqi 和 E. Isufi，“图时间卷积神经网络：架构与理论分析”，《IEEE 模式分析与机器智能汇刊》（TPAMI），第45卷，第12期，第14 625–14 638页，2023年。[17] D. Cao、Y. Wang、J. Duan、C. Zhang、X. Zhu、C. Huang、Y. Tong、B. Xu、J. Bai、J. Tong 和 Q. Zhang，“用于多变量时间序列预测的谱时序图神经网络”，载于《神经信息处理系统进展》（NeurIPS），第33卷，2020年，第17 766–17 778页。[18] Z. Shao、Z. Zhang、W. Wei、F. Wang、Y. Xu、X. Cao 和 C. S. Jensen，“用于交通预测的解耦动态图时空神经网络”，《VLDB基金会会刊》（PVLDB），第15卷，2746，2022年11月，第2733–页。[19] C. Zheng、X. Fan、S. Pan、H. Jin、Z. Peng、Z. Wu、C. Wang 和 P. S. Yu，“用于交通预测的时空联合图卷积网络”，《IEEE 知识与数据工程汇刊》，第36卷，第1期，第372–385页，2023年。

[20] S. Lan、Y. Ma、W. Huang、W. Wang、H. Yang 和 P. Li，“DSTAGNN：用于交通流量预测的动态时空感知图神经网络”，载于《第39届国际机器学习会议（ICML）论文集》，2022年，第11 906–11 917页。[21] F. Li、J. Feng、H. Yan、G. Jin、F. Yang、F. Sun、D. Jin 和 Y. Li，“用于交通预测的动态图卷积循环网络：基准与解决方案”，《ACM知识发现与数据挖掘汇刊》（TKDD），第17卷，第1期，第1–21页，2023年。[22] L. Bai、L. Yao、C. Li、X. Wang 和 C. Wang，“用于交通预测的自适应图卷积循环网络”，载于《神经信息处理系统进展》（NeurIPS），第33卷，2020年，第17 804–17 815页。[23] Z. Wu、S. Pan、G. Long、J. Jiang、X. Chang 和 C. Zhang，“连接各个点：利用图神经网络进行多变量时间序列预测”，载于《第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（KDD），2020年，第753–763页。[24] Z. Shao、Z. Zhang、F. Wang、W. Wei 和 Y. Xu，“时空身份：一种简单但有效的多变量时间序列预测基线”，载于《第31届ACM信息与知识管理国际会议论文集》（CIKM），2022年，第4454–4458页。[25] H. Liu、Z. Dong、R. Jiang、J. Deng、J. Deng、Q. Chen 和 X. Song，“时空自适应嵌入使原始Transformer成为交通预测领域的先进方法”，载于《第32届ACM信息与知识管理国际会议论文集》（CIKM），2023年，第4125–4129页。[26] Z. Wang、Y. Nie、P. Sun、N. H. Nguyen、J. Mulvey 和 H. V. Poor，“ST-MLP：采用通道独立策略的级联时空线性框架，用于交通预测”，arXiv预印本 arXiv:2308.07496，2023年。[27] P. Veli˘ckovi´c、G. Cucurull、A. Casanova、A. Romero、P. Lio 和 Y. Bengio，“图注意力网络”，arXiv预印本 arXiv:1710.10903，2017年。[28] Z. Pan、Y. Liang、W. Wang、Y. Yu、Y. Zheng 和 J. Zhang，“利用深度元学习从时空数据进行城市交通预测”，载于《第25届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（KDD），2019年，第1720–1730页。[29] S. Guo、Y. Lin、H. Wan、X. Li 和 G. Cong，“学习时空图数据的动态性与异质性以进行交通预测”，《IEEE知识与数据工程汇刊》（TKDE），第34卷，第11期，第5415–5428页，2021年。[30] K. Yi、Q. Zhang、W. Fan、S. Wang、P. Wang、H. He、N. An、D. Lian、L. Cao 和 Z. Niu，“频域多层感知机是时间序列预测中更有效的学习器”，载于《神经信息处理系统进展》（NeurIPS），第36卷，2023年，第76 656–76 679页。[31] I. O. Tolstikhin、N. Houlsby、A. Kolesnikov、L. Beyer、X. Zhai、T. Un- terthiner、J. Yung、A. Steiner、D. Keysers、J. Uszkoreit、M. Lucic 和 A. Dosovitskiy，“MLP-Mixer：用于视觉的全多层感知机架构”，载于《神经信息处理系统进展》（NeurIPS），第34卷，2021年，第24 261 24 272页。[32] Y. Cui、J. Xie 和 K. Zheng，“历史惯性：一种被忽视但强大的长序列时间序列预测基线”，载于《第30届ACM信息与知识管理国际会议论文集》（CIKM），2021年，第2965–2969页。[33] J. Deng、X. Chen、R. Jiang、X. Song 和 I. W. Tsang，“ST-Norm：用于多变量时间序列预测的时空归一化”，载于《第27届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（KDD），2021年，第269–278页。[34] R. Jiang、Z. Wang、J. Yong、P. Jeph、Q. Chen、Y. Kobayashi、X. Song、S. Fukushima 和 T. Suzumura，“用于交通预测的时空元图学习”，载于《AAAI人工智能会议论文集》（AAAI），第37卷，第7期，2023年，第8078–8086页。[35] Y. Fang、Y. Qin、H. Luo、F. Zhao、B. Xu、L. Zeng 和 C. Wang，“当时空遇见小波：通过高效谱图注意力网络进行解耦交通预测”，载于《IEEE数据工程国际会议论文集》（ICDE），2023年，第517–529页。[36] Y. Li、Z. Shao、Y. Xu、Q. Qiu、Z. Cao 和 F. Wang，“用于交通预测的动态频域图卷积网络”，载于《IEEE声学、语音与信号处理国际会议论文集》（ICASSP），2024年，第5245–5249页。[37] Z. Shao、F. Wang、Y. Xu、W. Wei、C. Yu、Z. Zhang、D. Yao、T. Sun、G. Jin、X. Cao、G. Cong、C. S. Jensen 和 C. Xueqi，“探索多变量时间序列预测的进展：综合基准测试与异质性分析”，《IEEE知识与数据工程汇刊》（TKDE），第37卷，第1期，第291–305页，2024年。

DWAFM：融合注意力与频域多层感知机的动态加权图结构嵌入用于交通预测的补充材料

Sen Shi、Zhichao Zhang，IEEE会员，和 Yangfan He

附录 A 方法与实验补充细节

本附录提供有关实验设置和方法的补充信息。具体而言，表 I 总结了所评估数据集的统计特征。算法 1 概述了所提出 DWAFM 的完整训练过程和前向传播数据流。此外，表 II 对 DWAFM 与现有基线方法进行了系统的架构比较，阐明了动态图构建和时空建模方面的结构差异。

表 I：数据集统计信息。

数据集	节点	长度采样率		Type
PEMS08	170	17856	5min	交通流
PEMS04	307	16992	5min	交通流
PEMS03	358	26208	5min	交通流
PEMSD7(M)	228	12672	5min	交通速度
PEMSD7(L)	1026	12672	5min	交通速度

算法 1 DWAFM的学习算法

1: 输入：数据集 X $\in \mathbb{R}^{S\times N\times D}$，预定义图 $G=(V,E,A)$，初始化嵌入 $f_e(\cdot)$，空间 $f_s(\cdot)$、时间 $f_t(\cdot)$ 和回归 $f_r(\cdot)$ 层及其可学习参数 $\Theta_e,\Theta_s,\Theta_t,\Theta_r$，学习率 γ，批次大小 B，嵌入维度 d_f，模型层数 m。 2: 设置 $iter=1$ 3: 重复 4: 从 X 中随机选取一个批次（输入 $X\in\mathbb{R}^{B\times T\times N\times D}$，标签 Y $\in\mathbb{R}^{B\times T_f\times N}$）。 5: 计算 $Z=f_e(X;\Theta_e)\in\mathbb{R}^{B\times T\times N\times 4d_f}$ 6: 对于 $i=1$ 到 m 执行 7: 计算 $Z_s=f_s(Z;\Theta_s)$ 8: 计算 $Z_s=$ 层归一化(Z_s+Z) 9: 计算 $Z_t=f_t(Z_s;\Theta_t)+Z$ 10: 结束循环 11: 计算 $\hat{Y}=f_r(Z_t;\Theta_r)$ 12: 计算 $L=\text{loss}(\hat{Y},Y)$ 13: 根据 L计算 $\Theta_e,\Theta_s,\Theta_t,\Theta_r$ 的随机梯度。 14: 根据其梯度与学习率 γ 更新模型参数。 15: $iter=iter+1$。 16: 直至收敛

附录 B 补充实验结果

A. 参数敏感性与效率

为考察超参数敏感性，我们评估了嵌入维度 d_f 以及在 PEMSD7(M) 上的缩放因子。如图 1(a) 所示，模型在 $d_f=14$时表现最佳，因为维度过高可能导致模型在这个相对简单的数据集上过拟合。对于缩放因子，将其设为 $1/\sqrt{d_h/2}$ 而非标准值$1/\sqrt{d_h}$ 效果最佳。这一调整有效地强调强局部相关性并抑制远距离噪声，从而增强局部空间建模。

表 II：不同模型中图构建与空间-时间模块的比较。

模型	预定义图	图动态性	S-T 嵌入	核心架构 空间	核心架构 时间
STGCN	✓	静态	-	GCN	CNN
GWNet	✓	静态	-	GCN	CNN
StemGNN	-	静态	-	GCN	CNN
STID	-	-	✓	MLP	MLP
DGCRN	✓	动态	-	GCN	RNN
MegCRN	-	动态	-	GCN	RNN
STwave	✓	动态	-	注意力	卷积神经网络 + 注意力
STAEformer	-	-	✓	Transformer	Transformer
DFDGCN	✓	静态	✓	GCN	CNN
DWAFM	✓	动态	✓	注意力	Fre-MLPs

此外，我们在 PEMS04 上评估了计算效率（图 1(b)）。与 STAE-former 和 DGCRN 等基线方法相比，DWAFM 在训练时间和内存开销可接受的情况下，显著提升了准确率。这表明 DWAFM 在性能与效率之间实现了出色的平衡。

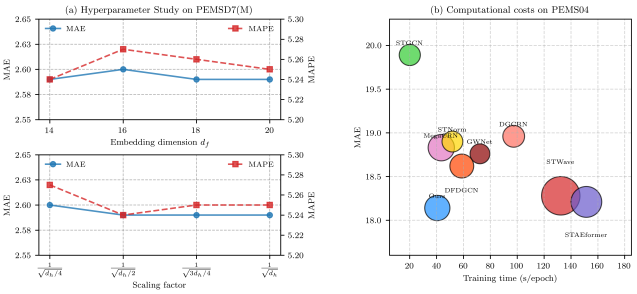


图 1：PEMSD7(M) 和 PEMS04 上的超参数研究与计算成本分析。

B. 节点嵌入初始化的敏感性分析

为严格评估模型对节点嵌入初始化的敏感性，我们采用四种标准方法进行了实验：Xavier均匀初始化、Kaiming均匀初始化、正交初始化和Xavier正态初始化。如表 III 所示，在两个数据集上，采用所有初始化方法时，评估指标均保持高度稳定。具体而言，在PEMSD7(M)上，平均绝对误差始终为2.59，而均方根误差和平均绝对百分比误差仅分别在5.24–5.25和6.40–6.43之间略有波动。PEMS08上也呈现出类似趋势，其平均绝对误差和均方根误差分别在13.54–13.59和22.94–23.00的狭窄范围内变化。这些一致的结果表明，所提出的模型对节点嵌入的初始化方案不敏感，从而验证了其强大的鲁棒性。

表 III：不同节点嵌入初始化方法下的性能比较。

方法	指标（PEMSD7(M)） MAE	RMSE	MAPE	指标（PEMS08） MAE	RMSE	MAPE
Xavier_均匀	2.59	5.24	6.40	13.57	22.94	8.92
Kaiming_均匀	2.59	5.25	6.40	13.59	22.98	8.94
正交初始化	2.59	5.25	6.43	13.59	23.00	8.92
Xavier_正态	2.59	5.24	6.42	13.54	22.95	8.90
均值 ± 标准差	**2.59 ± 0.00**	**5.25 ± 0.01**	**6.41 ± 0.02**	**13.57 ± 0.02**	**22.97 ± 0.03**	**8.92 ± 0.02**