第 54 卷第 1 期
2026 年 1 月

硅 酸 盐 学 报
JOURNAL OF THE CHINESE CERAMIC SOCIETY

Vol. 54，No. 1
January，2026

http://jccsoc.cbpt.cnki.net

DOI：10.14062/j.issn.0454-5648.20250794

基于多尺度特征增强和时序 Transformer 的 SiC 外延生长浓度预测模型

张忠义[1,2]，王 朗[1,2,3]，芦伟立[1,2,3]，李 帅[1,2,3]，杨建业[1,2,3]，高 楠[1,2,3]，王 波[1,2,3]，潘国平[1,2,3]，房玉龙[1,2,3]

(1. 中国电子科技集团公司第十三研究所，石家庄 050051；2. 河北省半导体智能制造创新中心，石家庄 050051；
3. 石家庄市半导体人工智能技术重点实验室，石家庄 050051)

摘 要：碳化硅(SiC)外延层掺杂浓度直接决定功率器件性能，现有调控依赖人工经验与离线测试，存在成本高、滞后性强的问题。为此，提出一种多尺度特征增强时序 Transformer(Multi-scale Feature-enhanced Temporal Transformer Network，MFT-Net)模型，整合多尺度卷积、压缩和激励(SE)模块、Transformer 及门控循环单元(GRU)模块，构建“当炉–下炉”双场景浓度预测体系。多尺度卷积捕捉毫秒至小时级参数动态，SE 强化核心特征，Transformer 建模全参数耦合，GRU 传递跨炉状态。基于 1200 炉数据实验表明，模型当炉预测相对误差低至 1.35%、决定系数 R^2 达 0.89，下炉预测相对误差为 1.66%，R^2 达到 0.87，显著优于传统模型。经统计，该模型预计可降低离线测试成本约 30%、提升工艺稳定性约 15%，为 SiC 外延智能化提供支撑。

关键词：碳化硅外延生长；核心层掺杂浓度；多尺度卷积；Transformer；工艺参数预测；深度学习
中图分类号：TM23　　**文献标志码**：A　　**文章编号**：0454–5648(2026)01–0035–14
网络出版时间：2026–01–05

碳化硅(SiC)作为第三代宽禁带半导体核心材料，其器件在多场景下性能显著优于传统硅基器件：相同电压下导通电阻仅为硅器件的 1/10~1/5，可大幅降低导通与开关损耗，适配高电压、高功率及高频应用场景，且能在潮湿、高温、强辐射等极端环境中长期稳定工作。目前，SiC 器件已在新能源汽车、光伏储能、高压输配电等领域规模化应用，成为高效能源转换的核心支撑技术[1–2]。

SiC 外延生长是器件制备的关键环节，需在衬底上生长晶体结构匹配、缺陷与杂质可控的 SiC 薄膜，核心目标是获得低缺陷密度、高均匀性、掺杂精准的外延层，以满足器件对击穿电压、导通损耗及可靠性的要求。过去数十年，高质量 SiC 衬底制备与外延工艺优化已取得商业化进展[3]，但外延层掺杂浓度的一致性调控仍依赖人工经验与离线测试，存在精度低、人力物力消耗大、无法预判工艺异常等问题，严重制约技术迭代效率。

近年来，机器学习与深度学习在半导体材料科学及工艺优化中展现出强大潜力，数据驱动方法逐渐成为传统物理仿真的有效补充甚至替代方案。在半导体材料领域，Sato 等[4]提出结合偏微分方程结构的神经网络架构，显著提升了半导体器件 TCAD 仿真效率；Wang 等[5]基于变分自编码器(VAE)构建 GaN 器件欧姆电阻自增强预测模型，实现了小样本数据的高效学习；Chen 等[6]将机器学习与第一性原理计算结合，为二维掺杂器件快速设计提供理论指导，大幅缩短了研发周期。深度学习的应用进一步拓展：Anger 等[7]利用卷积神经网络(CNN)对光致发光图谱进行分类，实现了晶圆缺陷的快速无损检测；Tang 等[8]基于深度等变神经网络学习框架，将混合密度泛函与材料结构建立直接关系，实现了半导体材料电子结构的快速预测。Kim 等[9]开发的深度学习晶圆缺陷检测算法准确率达 96.9%，展现出高维图像数据处理能力。

收稿日期：2025–11–02。　修订日期：2025–11–28。
基金项目：重点新材料研发及应用国家科技重大专项(2025ZD0619502)。
第一作者：张忠义(2000—)，男，助理工程师；
王 朗(1989—)，男，工程师。
通信作者：房玉龙(1986—)，男，博士，研究员。

Received date: 2025–11–02.　**Revised date:** 2025–11–28.
First author: ZHANG Zhongyi (2000–), male, Assistant Engineer.
WANG Lang (1989–), male, Engineer.
E-mail: 2841622099 @qq.com; wangl170326@163.com
Correspondent author: FANG Yulong (1986–), male, Ph.D., Professor.
E-mail: yvloong@163.com

在薄膜外延生长建模中，数据驱动方法也成为研究热点。例如，Wei 等[10]构建了 GaN 外延生长预测模型，通过正向模型预测 AlGaN 层浓度，基于此开发的逆向模型可在保证材料性能的前提下减少厚度设计量，并将设计时间缩短至秒级；Salimian 等[11]通过 ZnO 薄膜光谱数据预测方阻，经优化后模型性能优异，可减少生产测试成本，为外延过程在线监控提供了新思路。

然而，与广泛研究的 GaN、氧化物半导体等体系相比，针对 SiC 外延生长这一复杂过程，尤其是对其掺杂浓度进行精准预测的研究仍显不足。这一研究缺口主要源于 SiC 外延所特有的三重复杂特性：跨越毫秒至小时量级的多尺度参数动态、工艺参数间存在的全局强耦合关系，以及体现为跨炉次状态传递的长程记忆效应。这些特性对建模方法提出了严峻挑战：传统单尺度卷积神经网络难以捕捉全尺度特征，而长短期记忆网络(LSTM)在建模参数间复杂的全局非线性关联时存在局限，导致现有模型预测精度不足，难以满足高性能 SiC 外延工艺的控制需求。

近年来，一系列先进的时序建模框架被提出，但它们在应对 SiC 外延的上述挑战时，仍显现出一定的局限性。具体而言，Lim 等[12]提出的原应用于电力负荷预测(Temporal Fusion Transformers)虽能有效融合多变量时序注意力，但其特征提取前端并未针对多尺度参数动态进行定制化设计。Zhou 等[13]提出的 Informer 虽在长序列计算效率上取得突破，但其编码机制缺乏对局部多尺度特征的精细捕捉与对噪声参数的筛选能力。Shi 等[14]使用的 ConvLSTM 虽结合了卷积与循环单元，但其内在结构弱于刻画全工艺参数的全局耦合关系，且难以有效建模跨炉次的长程依赖。

为此，本工作从 SiC 外延工艺的物理机制出发，提出了多尺度特征增强-时序耦合 Transformer (MFT-NET)模型。该模型以多尺度卷积为核心特征提取基础，结合改进的 Transformer[15]与门控循环单元(GRU)时序记忆模块[16]，通过全连接层实现浓度映射。SiC 外延过程中的“多尺度特征”源于不同物理作用在时间尺度上的显著差异：在毫秒尺度上，主要表现为气体流量的瞬时波动，其快速变化会直接影响反应气体向衬底的输运过程，从而扰动掺杂原子的吸附速率；在分钟尺度上，体现为衬底温度的缓慢漂移，这种变化会显著改变原子的表面迁移与扩散行为，进而影响外延层中的浓度分布与梯度；在小时尺度上，则表现为总气体流量与生长时长等宏观参数的协同演化，这些共同决定了外延层的最终厚度与掺杂轮廓。

模型设计聚焦三大核心问题：首先，通过“多尺度卷积+SE”模块，利用多分支差异化卷积核捕捉从毫秒到小时的全尺度动态特征，并结合通道注意力机制强化对掺杂敏感的关键特征、抑制噪声干扰，以解决“多尺度参数关联与关键特征干扰”问题；其次，引入“改进 Transformer”模块，旨在捕捉工艺参数间复杂的全局非线性耦合关系，弥补传统模型在全局关联建模上的不足，解决“全参数耦合”问题；最后，通过“GRU 时序记忆”模块，定向建模上一工艺周期残留的物理状态对下一周期的影响，以解决“跨炉次状态传递”这一关键难题。

据调研，本工作是首个将多尺度卷积与改进 Transformer 协同应用于 SiC 外延浓度预测的模型，主要创新点为：1) 构建“多尺度卷积-Transformer”协同架构，填补 SiC 外延多尺度+全局耦合数据驱动建模的空白；2) 提出“多尺度特征提取-全局耦合建模”一体化策略，解决传统模型局部与全局特征割裂的问题；3) 在“当炉–下炉”双场景中实现高精度预测，性能优于传统 CNN、LSTM 及单一 Transformer 模型；4) 为 SiC 外延浓度调控提供“多尺度+全局协同”新范式，助力工艺智能化。

1 模型和方法

在本节中，首先介绍了模型的整体框架；随后，详细介绍了各个模块以及引入该模块的原因；接着，介绍了本工作中使用的数据预处理手段；最后，介绍了模型的损失函数。

1.1 MFT-NET 模型整体框架

MFT-NET 模型的整体架构如图 1 所示。模型的输入为工艺参数时序矩阵。首先，数据流入多尺度卷积模块[17]，该模块通过并行卷积分支提取参数在不同时间尺度上的局部特征，输出多尺度融合特征。紧接着，多尺度融合特征被送入压缩和激励(SE)模块，该模块通过通道注意力机制对特征进行自适应加权，强化关键参数通道并抑制噪声，输出增强后的特征。此后，为建模参数间的全局耦合关系，Transformer 模块对增强后的特征进行聚合，得到参数级全局特征，并通过自注意力机制计算其交互权重，输出蕴含复杂参数耦合关系的表征 X_{trans}。为捕获连续炉次间的状态依赖，X_{trans} 经池化后输入 GRU 模块，该模块结合上一炉的隐藏状态更新当前状态，并生成预测所用特征。最终，该特征通过全连接层映射，输出

最终的浓度预测值。该设计确保了从局部特征提取、全局关系建模到时序状态传递的端到端优化。

1.2　多尺度卷积模块

在 SiC 外延生长工艺中，工艺参数的变化表现出显著的多尺度特性：某些参数的瞬时波动(如气体流量、压力)会直接影响局部反应氛围，而一些参数的长程变化(如生长时长、总气体流量)则通过累积效应影响外延层厚度与掺杂分布。传统单尺度卷积模型因感受野固定，难以同时兼顾微观瞬时扰动与宏观趋势特征：小尺寸卷积核虽能捕捉毫秒至秒级波动，但无法覆盖长时累积效应；大尺寸卷积核虽可提取宏观特征，却会平滑掉关键微观信息。为有效捕捉不同时间尺度上的参数动态及其对浓度形成的协同作用机制，本工作设计了多尺度卷积模块，构建了一个三分支卷积结构，分别针对微观、介观和宏观尺度进行特征提取，如图 2 所示。

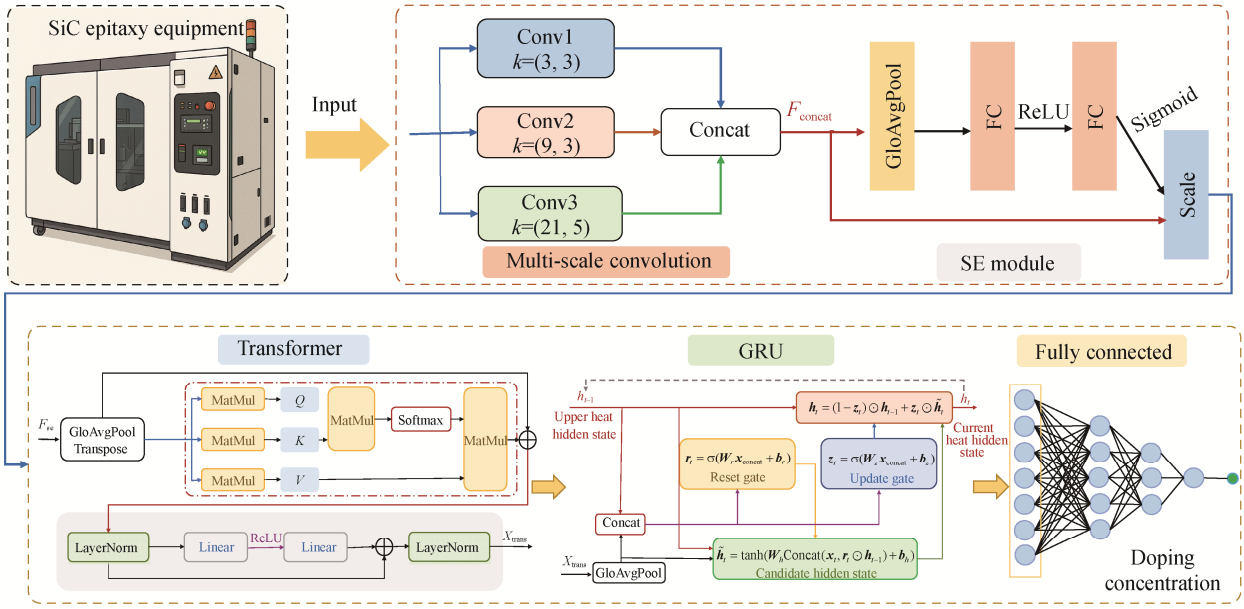


图 1　MFT-NET 模型框架图

Fig. 1　Overall framework of the MFT-NET model

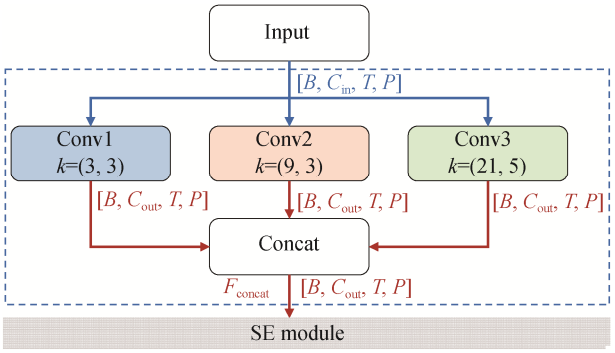


图 2　多尺度卷积模块示意图

Fig. 2　Schematic diagram of the multi-scale convolution module

设输入工艺参数矩阵为 $\boldsymbol{X}\in\mathbb{R}^{T\times P}$，其中 T 为时序采样点数，P 为工艺参数总数。为适配卷积操作，将其扩展为 $\boldsymbol{X}_{\text{exp}}\in\mathbb{R}^{C_{\text{in}}\times T\times P}$，其中输入通道数 $C_{\text{in}}=1$，各分支设计如下。

微观分支采用 3×3 卷积核，感受野覆盖 3 个连续时序点及 3 个相邻参数，专注于捕捉毫秒至秒级瞬时波动，其计算表示为：

$$\boldsymbol{F}_{\text{micro}}=\text{Conv2D}(\boldsymbol{X}_{\text{exp}},\boldsymbol{K}_{\text{micro}})+\boldsymbol{b}_{\text{micro}} \tag{1}$$

其中：$\boldsymbol{K}_{\text{micro}}$ 为卷积核；$\boldsymbol{b}_{\text{micro}}$ 为偏置。

介观分支采用 9×3 卷积核，感受野覆盖 9 个连续时序点及 3 个相邻参数，用于提取分钟级变化趋势，其计算表示为：

$$\boldsymbol{F}_{\text{meso}}=\text{Conv2D}(\boldsymbol{X}_{\text{exp}},\boldsymbol{K}_{\text{meso}})+\boldsymbol{b}_{\text{meso}} \tag{2}$$

宏观分支采用 21×5 卷积核，感受野覆盖 21 个连续时序点及 5 个相邻参数，为降低计算量并扩大感受野，设置步长 s=2，计算表示为：

$$\boldsymbol{F}_{\text{macro}}=\text{Conv2D}(\boldsymbol{X}_{\text{exp}},\boldsymbol{K}_{\text{macro}},s=2)+\boldsymbol{b}_{\text{macro}} \tag{3}$$

将三分支输出按通道维度拼接，得到多尺度特征

$$\boldsymbol{F}_{\text{concat}}=\text{Concat}(\boldsymbol{F}_{\text{micro}},\boldsymbol{F}_{\text{meso}},\boldsymbol{F}_{\text{macro}}) \tag{4}$$

其中：$\boldsymbol{F}_{\text{concat}}\in\mathbb{R}^{C_{\text{out}}\times T\times P}$；$C_{\text{out}}$ 为多尺度卷积模块的输出通道数。

1.3　SE 模块

在 SiC 外延生长过程中，不同工艺参数对核

心层掺杂浓度的影响具有显著差异性。Pearson 相关性分析[18]表明，在 75 项工艺参数中，仅少数参数与掺杂浓度的相关系数超过 0.3，这些参数直接参与掺杂原子的供给、扩散与沉积过程；其余多数参数与浓度的相关系数低于 0.1，其波动多为无关噪声，对浓度预测不构成显著影响。传统卷积模型对所有通道特征施加同等权重，容易导致关键特征被噪声淹没，甚至因参数间的耦合关系而产生错误归因。为强化对掺杂相关核心特征的选择性感知并抑制冗余干扰，本工作引入 SE 模块，通过通道注意力机制实现特征的自适应加权，如图 3 所示。

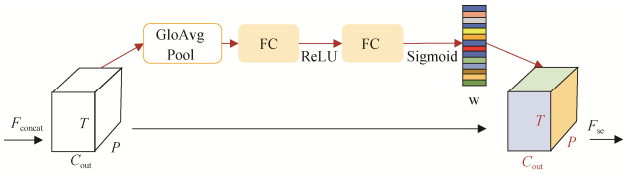


图 3　SE 模块示意图

Fig. 3　Schematic diagram of the squeeze-excitation (SE)

设多尺度卷积模块输出特征为 $\boldsymbol{F}_{\text{concat}} \in \mathbb{R}^{C_{\text{out}}\times T\times P}$。SE 模块通过压缩(Squeeze)和激励(Excitation)2 个操作实现通道权重的重新标定。

压缩操作通过对空间–参数维度的全局平均池化，将每个通道的特征压缩为一个标量，用于表征该通道参数的全局影响强度：

$$f_c = \frac{1}{T\times P}\sum_{t=1}^{T}\sum_{p=1}^{P}\boldsymbol{F}_{\text{concat}}(c,t,p) \tag{5}$$

得到通道统计向量 $\boldsymbol{f}=[f_1,f_2,\ldots,f_{C_{\text{out}}}]^T\in\mathbb{R}^{C_{\text{out}}}$，其数值大小反映了对应参数在全局过程中对浓度的影响权重。

激励操作通过 2 个全连接层实现通道注意力的非线性映射。首先将通道向量 $\boldsymbol{f}$ 压缩至 C/r 维(压缩比 r=4)，并经过 ReLU 激活函数：

$$\boldsymbol{f}_{\text{fc1}} = \text{ReLU}(\boldsymbol{W}_1\text{f}+\boldsymbol{b}_1) \tag{6}$$

其中：$\boldsymbol{W}_1\in\mathbb{R}^{(C/r)\times C}$ 为权重矩阵；$\boldsymbol{b}_1\in\mathbb{R}^{C/r}$ 为偏置。随后通过第二个全连接层恢复维度，并采用 Sigmoid 函数将输出约束在[0,1]区间，生成通道权重向量：

$$\boldsymbol{w} = \sigma(\boldsymbol{W}_2\boldsymbol{f}_{\text{fc1}}+\boldsymbol{b}_2) \tag{7}$$

其中：$\boldsymbol{W}_2\in\mathbb{R}^{C\times(C/r)}$ 为权重矩阵；$\boldsymbol{b}_2\in\mathbb{R}^{C}$ 为偏置；σ 表示 Sigmoid 激活函数。最终，通过元素级乘法对原始特征进行重新校准：

$$\boldsymbol{F}_{\text{se}}(c,t,p) = \boldsymbol{F}_{\text{concat}}(c,t,p)\cdot\boldsymbol{w}_c \tag{8}$$

输出增强后的特征 $\boldsymbol{F}_{\text{se}}\in\mathbb{R}^{C\times T\times P}$。该机制能够自适应地强化关键参数对应的特征通道，抑制冗余或噪声通道的干扰，从而提高模型对掺杂浓度预测的准确性与鲁棒性。

1.4　Transformer 模块

SiC 外延生长过程中，核心层掺杂浓度是由多参数间复杂的全局耦合作用共同决定的，这种耦合关系主要表现为直接耦合与间接耦合 2 种形式。传统卷积网络和循环神经网络难以有效刻画 75 项工艺参数间复杂的全局关联，尤其对跨参数的同步耦合作用建模能力有限。为从根本上解决全参数耦合关系的建模难题，本工作设计了参数级 Transformer 模块，通过基于自注意力的编码器结构显式地捕捉和量化任意两项参数之间的相互作用强度与模式，如图 4 所示。

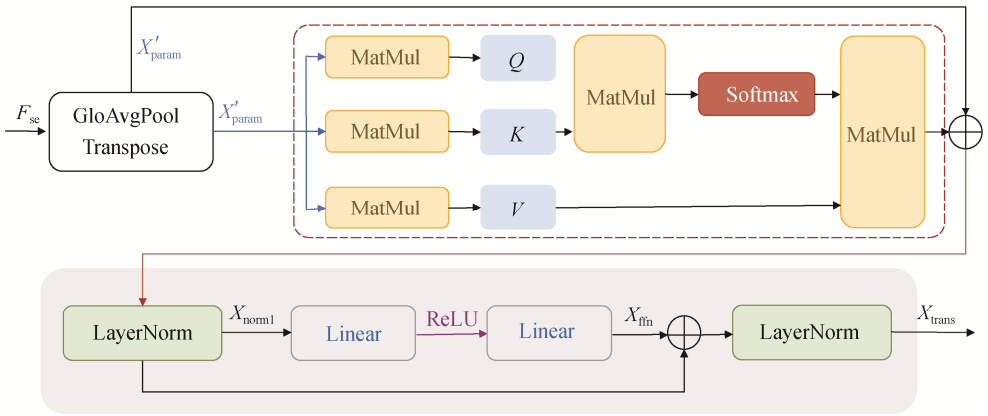


图 4　Transformer 模块示意图

Fig. 4　Schematic diagram of the Transformer module

首先，为消除时间动态对参数耦合关系建模的干扰，对前一模块输出的时序–参数特征 $\boldsymbol{F}_{\text{se}}\in\mathbb{R}^{C\times T\times P}$ 沿时间维度进行全局平均池化，将其压缩为参数级全局特征矩阵，并将该特征矩阵进行转置，以适应

后续的计算步骤：

$$\boldsymbol{X}_{\text{param}}(c,p)=\frac{1}{T}\sum_{t=1}^{T}F_{\text{se}}(c,t,p) \tag{9}$$

$$\boldsymbol{X}'_{\text{param}}=\boldsymbol{X}_{\text{param}}^{\text{T}} \tag{10}$$

其中：$\boldsymbol{X}'_{\text{param}}\in\mathbb{R}^{P\times C}$。该操作将每个参数在全部时间步上的动态行为聚合为一个全局表示，为后续参数间耦合关系的计算提供稳定特征表示。在得到参数级全局特征矩阵后，模型执行一次关键的张量转置操作。这一操作改变了特征的维度结构，使得建模的核心从时序依赖分析转向参数耦合分析。在传统 Transformer 架构中，自注意力机制计算的是序列内不同时间步之间的关联关系；而经过转置操作后，自注意力机制被重新导向计算所有工艺参数在整个工艺周期内的全局耦合关系。

随后，通过线性投影层将全局参数特征分别映射为查询(Query, $\boldsymbol{Q}$)、键(Key, $\boldsymbol{K}$)和值(Value, $\boldsymbol{V}$)3 个矩阵：

$$\boldsymbol{Q}=\boldsymbol{X}'_{\text{param}}\boldsymbol{W}^{Q},\quad \boldsymbol{K}=\boldsymbol{X}'_{\text{param}}\boldsymbol{W}^{K},\quad \boldsymbol{V}=\boldsymbol{X}'_{\text{param}}\boldsymbol{W}^{V} \tag{11}$$

其中：$\boldsymbol{W}^{Q},\boldsymbol{W}^{K},\boldsymbol{W}^{V}\in\mathbb{R}^{C\times d_k}$ 为可训练权重矩阵；$d_k=C/h=16$ 为投影维度；$h=8$ 为注意力头数量。通过缩放点积注意力机制计算参数间的耦合权重：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{12}$$

所得注意力权重矩阵 $\boldsymbol{A}=\text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{T}/\sqrt{d_k})\in\mathbb{R}^{P\times P}$ 具有明确的物理意义：其元素 $\boldsymbol{A}_{ij}$ 表示第 i 个参数与第 j 个参数之间的耦合强度。为捕捉不同类型的耦合关系，采用多头注意力机制并行地学习多种交互模式：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\boldsymbol{Head}_1,\ldots,\boldsymbol{Head}_h)\boldsymbol{W}^{O} \tag{13}$$

其中：每个注意力头 $\boldsymbol{Head}_i=\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)$，专注于捕捉一类特定的参数交互模式；$\boldsymbol{W}^{O}\in\mathbb{R}^{C\times C}$ 用于融合所有头的输出。

为增强模型表达能力并保障训练稳定性，引入残差连接与层归一化机制。将多头注意力的输出与原始参数特征相加并进行层归一化：

$$\boldsymbol{X}_{\text{norm1}}=\text{LayerNorm}(\boldsymbol{X}'_{\text{param}}+\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})) \tag{14}$$

该残差结构确保即使在深层网络中梯度也能有效回传，层归一化则稳定了特征的分布。其后通过一个两层的前馈神经网络进一步增强非线性拟合能力：

$$\boldsymbol{X}_{\text{ffn}}=\text{ReLU}(\boldsymbol{X}_{\text{norm1}}\boldsymbol{W}_{\text{ff1}}+\boldsymbol{b}_{\text{ff1}})\boldsymbol{W}_{\text{ff2}}+\boldsymbol{b}_{\text{ff2}} \tag{15}$$

其中：$\boldsymbol{W}_{\text{ff1}}\in\mathbb{R}^{C\times 4C}$。最后再次使用残差连接与层归一化：

$$\boldsymbol{X}_{\text{trans}}=\text{LayerNorm}(\boldsymbol{X}_{\text{norm1}}+\boldsymbol{X}_{\text{ffn}}) \tag{16}$$

输出特征 $\boldsymbol{X}_{\text{trans}}\in\mathbb{R}^{P\times C}$ 深度融合了所有参数间的耦合信息，其每一行对应一个参数的表征，该表征不仅包含参数自身的特性，还融合了其与其他所有参数交互所产生的修正信息。这种设计使得模型能够同时刻画如 N_2 流量与反应室总压间的竞争性耦合、衬底温度与 N_2 流量间的协同性耦合等多种作用类型，为最终浓度预测提供更加全面和准确的参数表征基础。

1.5　GRU 炉间状态传递模块

在 SiC 外延生长的工业化连续生产过程中，炉次间存在显著的状态传递效应，主要表现为掺杂原子残留、温度历史依赖及设备状态漂移等现象。传统单炉模型将每炉数据视为独立样本，无法记忆和利用历史状态信息，导致预测偏差超过 10%。为解决这一问题，本工作设计了 GRU 模块，通过其门控机制实现对关键状态信息的选择性记忆与传递，从而精准建模跨炉状态依赖关系，如图 5 所示。

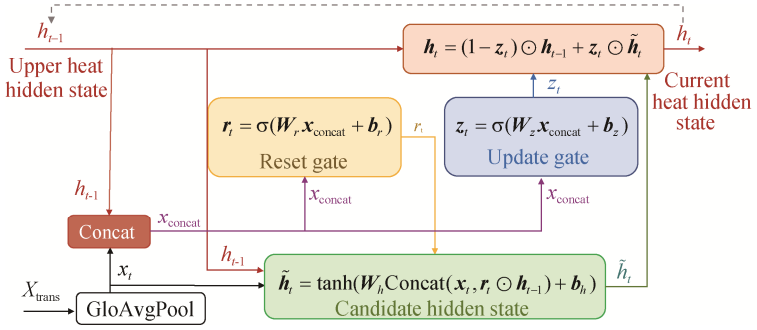


图 5　GRU 模块示意图

Fig. 5　Schematic diagram of the gated recurrent unit (GRU) inter-batch state transfer module

GRU 模块的隐藏状态在序列开始时初始化为零向量。在训练与推理过程中，处理完第 $t-1$ 炉数据后产生的隐藏状态 $\boldsymbol{h}_{t-1}$ 会被保留，并作为处理第 t 炉数据时的初始状态输入，从而实现反应室状态、

残留掺杂等跨炉次效应的建模。

GRU 模块的输入首先经过特征预处理。将 Transformer 模块输出的参数级特征矩阵 $\boldsymbol{X}_{\text{trans}} \in \mathbb{R}^{P\times C}$ 通过全局平均池化压缩为全局特征向量：

$$\boldsymbol{x}_t = \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{X}_{\text{trans}}(p,:) \tag{17}$$

其中：$\boldsymbol{x}_t \in \mathbb{R}^C$ 表征当前炉工艺参数的全局耦合特征。为实现跨炉状态传递，将该特征与上一炉的隐藏状态 $\boldsymbol{h}_{t-1}$ 进行拼接：$\boldsymbol{x}_{\text{concat}} = \text{Concat}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1})$，形成综合当前工艺参数与历史状态信息的输入特征。

GRU 通过更新门与重置门实现对状态信息的选择性控制。更新门控制历史状态的保留比例：

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \boldsymbol{x}_{\text{concat}} + \boldsymbol{b}_z) \tag{18}$$

重置门调节历史状态与当前特征的交互强度：

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \boldsymbol{x}_{\text{concat}} + \boldsymbol{b}_r) \tag{19}$$

候选隐藏状态的计算融合了筛选后的历史信息：

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_h \text{Concat}(\boldsymbol{x}_t, \boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h) \tag{20}$$

最终隐藏状态通过门控机制更新：

$$\boldsymbol{h}_t = (1-\boldsymbol{z}_t)\odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \tag{21}$$

该机制确保模型能够自适应地保留与浓度预测相关的历史状态，同时过滤无关噪声。

GRU 模块的输出包含 2 个部分：一是通过全连接层压缩得到的特征向量 $\boldsymbol{f}_{\text{gru}} \in \mathbb{R}^C$，用于当前炉的浓度预测；二是更新后的隐藏状态 $\boldsymbol{h}_t$，作为下一炉的状态输入，实现跨炉状态传递。

针对工业生产中的不同应用场景，模型提供 2 种预测模式。

1)当炉浓度预测：基于当前炉的工艺参数和上一炉隐藏状态，输出当前炉的浓度预测值，用于生长监测并节省测试成本。

2)下炉浓度预测：基于上一炉的工艺参数和上一炉隐藏状态，提前预测下一炉的浓度趋势，为实时工艺调控提供依据。

最终浓度预测通过 2 层全连接网络实现：

$$\boldsymbol{f}_{\text{fc1}} = \text{ReLU}(\boldsymbol{W}_{\text{fc1}} \boldsymbol{f}_{\text{gru}} + \boldsymbol{b}_{\text{fc1}}) \tag{22}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{W}_{\text{fc2}} \boldsymbol{f}_{\text{fc1}} + \boldsymbol{b}_{\text{fc2}} \tag{23}$$

其中输出层采用线性激活，确保预测值覆盖实际的浓度范围。

MFT-NET 模型中各个模块与 SiC 核心层浓度预测任务中要解决的问题对应关系以及后续实验安排如图 6 所示。

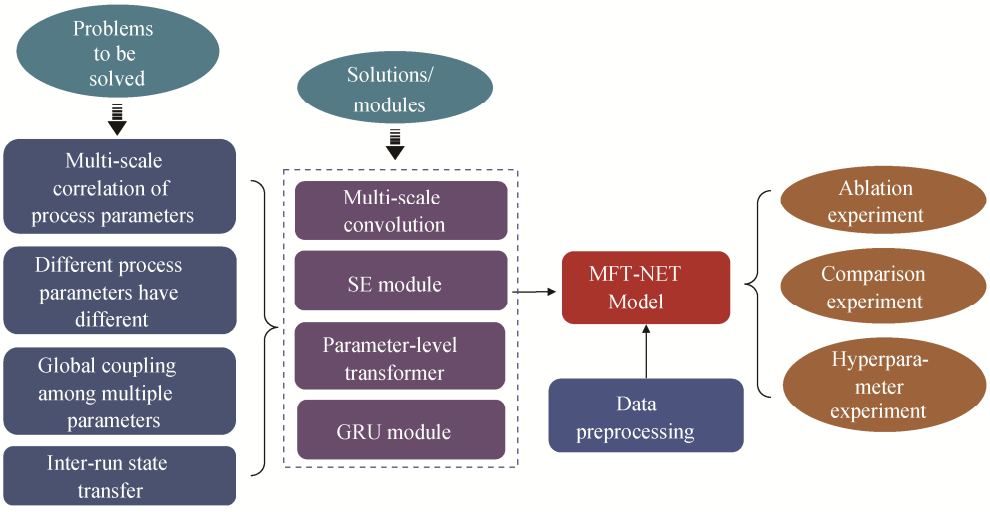


图 6 MFT-NET 模型中各个模块与其期望要解决的问题关系图

Fig. 6 Diagram of the relationship between each module in the MFT-NET model and the problems it intends to solve

1.6 损失函数

针对 SiC 外延生产数据中浓度分布存在波动的特性，传统均方误差(MSE)损失函数在训练过程中会平等对待所有样本误差。这导致其对少数偏离较大的异常样本赋予过高权重，迫使模型过度适应这些非典型工况，从而削弱了对主流稳定生产场景的预测准确性。为解决该问题，本工作设计了加权 MSE 损失函数，其核心思想是以常规工艺样本主导模型学习，同时对不同偏离程度的样本赋予差异化贡献。该函数通过动态调整样本权重，引导模型优先拟合占多数的稳定工艺规律，并对偏离样本进行适度而非过度的优化。后文将基于实际数据，对比分析传统 MSE 与加权 MSE 损失函数的预测结果，以验证所提方法的有效性。

加权 MSE 损失函数的具体定义如下：

$$L=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{w}_i(\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i)^2 \tag{24}$$

其中：N 为训练集样本数量；$\boldsymbol{y}_i$ 为第 i 炉核心层掺杂浓度实际值；$\hat{\boldsymbol{y}}_i$ 为第 i 炉浓度预测值；$\boldsymbol{w}_i$ 为第 i 炉的动态权重系数，其设计基于浓度分布的统计边界划分样本类型，对应权重规则如下：

$$\boldsymbol{w}_i=\begin{cases}1.0 & \left|\boldsymbol{y}_i-\mu_y\right|\le 2\sigma_y\\ \exp\left(-\dfrac{\left|\boldsymbol{y}_i-\mu_y\right|-2\sigma_y}{\sigma_y}\right) & 2\sigma_y<\left|\boldsymbol{y}_i-\mu_y\right|\le 3\sigma_y\\ 0.1 & \left|\boldsymbol{y}_i-\mu_y\right|>3\sigma_y\end{cases} \tag{25}$$

其中：μ_y 和 σ_y 分别为浓度均值和标准差动态调整。

2　结果与讨论

2.1　数据集和实验设置

实验数据基于某型号外延设备采集，数据来源于某 SiC 外延结构 2 年内的全工艺周期参数，共包含 1200 余炉有效生产数据。每炉数据包含 2 类核心信息。

1) 工艺参数数据：研究基于超过 1000 项原始监测参数，这些参数涵盖了气体流量(如 N_2、载气等)、温度(如衬底温度)、压力(反应室压力)、功率(射频功率)以及设备运行状态(如旋转转速、生长时间)等关键物理量。经 Shapiro-Wilk 正态性检验[19]证实，这些参数整体上符合近似正态分布特征，满足了 Pearson 相关性分析[18]对“连续变量”与“正态分布”的基本要求。相关研究[20]表明，在半导体制造系统中，Pearson 相关系数|r|对线性关系的量化精度显著优于其他方法，因此，本工作选用该方法进行相关性分析。在阈值设定上，本工作将|r| < 0.2 作为剔除弱相关参数的临界值。该标准在工业实践中得到验证：《Pearson 算法在半导体工艺上的应用》[21]在芯片成品率关联分析中同样采用此阈值，认为其能够有效过滤对核心工艺无实质影响的噪声参数。最终，依据相关性分析确定了 75 项关键参数作为模型的输入。

2) 浓度实测数据：每炉生长结束后，采用某型号高精度接触式浓度测试仪测量核心层掺杂浓度。为获得具有代表性的整片晶圆浓度值，并评估其均匀性，依据标准流程在每片晶圆的有效区域内选取 15 个均匀分布的测试点(包含中心及径向关键位置)。最终取 15 个测试点浓度的算术平均值作为该炉浓度的实际值，以消除局部波动并确保数据可靠性。核心层掺杂浓度是该模型的唯一输出变量。

需要说明的是，目标值 SIC 的核心层浓度信息均为 10^{16} 量级，对其进行了去量级操作；在后续的指标计算中，同样不考虑量级因素，而更关注数值本身的大小。另外，对数据进行了预处理操作，对于少量缺失值，采用基于缺失点前后时间步参数值的线性插值法进行填补。为消除参数间的量纲差异，对所有输入参数采用 Min-Max 标准化方法；针对原始工艺参数采样频率存在差异的问题，通过插值方式将所有参数统一为相同的采样频率，构建维度一致的时序矩阵，实现时序数据的规整化处理；数据集按时间顺序以 7.0:1.5:1.5 的比例划分为训练集、验证集和测试集，以模拟实际工业场景中的时序预测需求。

模型训练策略方面，优化器选用 AdamW[22]，该优化器在 Adam 动量更新基础上引入 L_2 正则化，有效抑制模型过拟合；学习率调度采用余弦退火策略，避免训练后期因学习率过高导致参数震荡；同时采用梯度裁剪避免梯度爆炸，确保模型在训练过程中稳定收敛。为实现模型性能最大化，构建了超参数优化体系，该体系融合了随机搜索与贝叶斯优化 2 种方法的优势：随机搜索通过在超参数空间进行广泛采样，快速定位有效区间，适用于对 Dropout 比率、Transformer 层数等关键参数进行初步筛选；贝叶斯优化则基于概率代理模型，建立超参数与模型性能的映射关系，实现精细化寻优，最终模型部分超参数设置如表 1 所示。除此之外，在讨论部分详细展示了不同超参数设置对模型性能的影响。

实验平台配置方面，CPU 采用 Intel Xeon Gold 6438，GPU 选用 NVIDIA RTX 4090；操作系统为 Ubuntu 22.04 LTS，深度学习框架采用 PyTorch 2.2.0(搭配 CUDA12.1)，数据处理依赖 Pandas 2.2.1、NumPy 1.26.4 与 Scikit-learn 1.4.1。

表 1　模型训练部分超参数设置

Table 1　Settings of partial hyperparameters for model training

Hyperparameter	Value
Dropout	0.2
Batch Size	64
Gradient Clipping Value	1.0
Output Channels (C_{out}) of Multi-scale Convolution	64
Number of Attention Heads	8
Batch Normalization	Yes
Number of Transformer Layers	3
Random Seed	42

2.2 评估方法

在 SiC 外延核心层掺杂浓度的回归预测任务中，为全面评估模型性能，从相对误差、绝对偏差及拟合优度多个维度选取了 4 种评估指标。

平均绝对百分比误差(MAPE)[23]通过计算预测值与实际值的相对误差，有效消除了浓度量级的影响，能够直观反映预测精度，其值越低代表模型性能越好，计算公式为：

$$\mathrm{MAPE}=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \tag{26}$$

其中：N 为测试集样本数量；y_i 为第 i 炉核心层掺杂浓度实际值；$\hat{y}_i$ 为对应预测值。

平均绝对误差(MAE)[24]衡量预测值与实际值的平均绝对偏差，对异常值不敏感，适用于评估模型在工业场景中的稳健性，计算公式为：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{27}$$

均方根误差(RMSE)[24]通过对误差进行平方放大较大偏差的影响，能反映预测值的波动情况，便于与工业控制标准进行对比，计算公式为：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{28}$$

决定系数(R^2)[25]用于量化模型对浓度变化的解释能力，其值越接近 1，说明模型对数据拟合效果越好，计算公式为：

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{29}$$

其中：$\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$ 为测试集实际浓度的均值。

综上，上述 4 类指标可从不同角度全面评估 MFT-NET 模型在当炉预测与跨炉预测双场景下的综合性能。

2.3 对比实验

为验证 MFT-NET 模型在 SiC 外延核心层掺杂浓度预测中的优越性，选取传统机器学习(Ridge 回归[26]、SVR[27])、CNN 架构(CNN、CNN + LSTM)、Transformer 类(Transformer、CNN + Transformer)共 7 类代表性模型，基于相同数据集、硬件平台与训练策略开展对比实验，测试集上“当炉预测”与“跨炉预测”双场景性能如表 2 所示，MFT-NET 模型预测值与真实值的对比如图 7 所示。

表 2 MFT-NET 模型与其他模型的对比实验结果

Table 2 Results of comparative experiments of MFT-NET model with other models

Prediction Scenario	Model	MAE/$\times10^{14}$ cm^{-3}	MAPE/%	RMSE/$\times10^{14}$ cm^{-3}	R^2
Current run prediction	Ridge Regression	4.01	2.97	5.94	0.54
	SVR	5.19	3.84	5.02	0.48
	CNN	3.12	2.31	3.99	0.61
	CNN+LSTM	3.11	2.30	3.56	0.67
	Transformer	2.74	2.03	3.54	0.68
	CNN+Transformer	2.05	1.52	2.98	0.76
	MFT-NET	1.83	1.35	2.81	0.89
Next run prediction	Ridge Regression	8.93	6.61	7.56	0.47
	SVR	6.77	5.01	6.81	0.43
	CNN	4.59	3.4	6.79	0.59
	CNN+LSTM	3.45	2.56	4.51	0.67
	Transformer	2.89	2.14	3.94	0.66
	CNN+Transformer	2.34	1.73	3.75	0.74
	MFT-NET	2.25	1.66	3.42	0.87

MAE is meanabsolute error, MAPE is meanabsolute percentage error, RMSE is root meansquared error, R^2 is coefficient of determination.

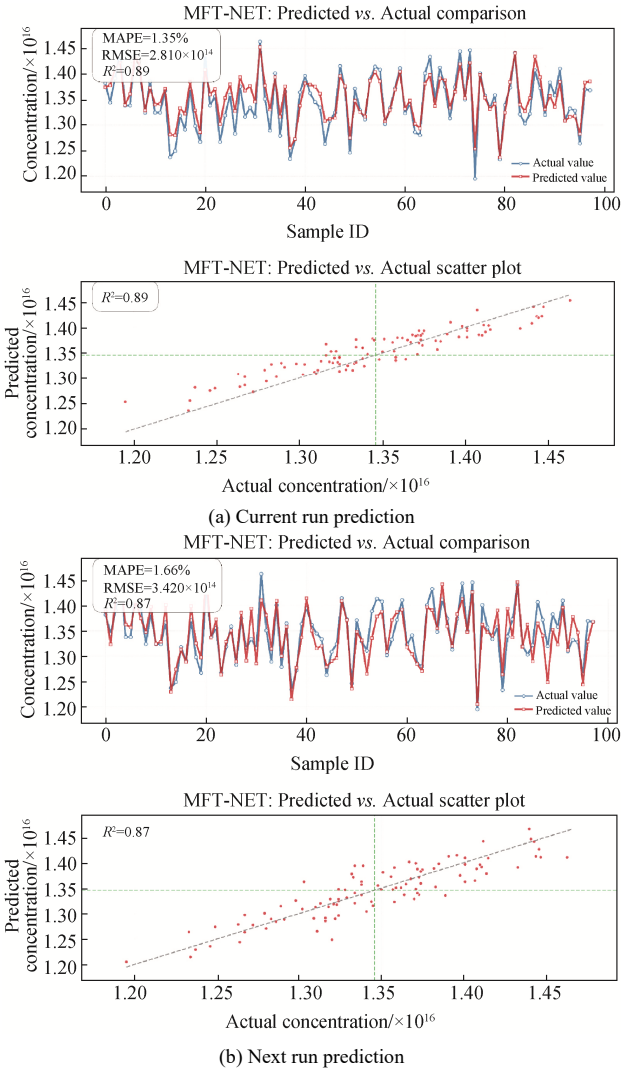


图 7 MFT-NET 模型预测值与真实值对比结果

Fig. 7 Results of comparison between predicted values and actual values of MFT-NET model

根据表 2 所示的对比实验结果，本工作从多个评估维度对不同类型模型的预测性能进行了系统分析。

传统机器学习模型在双场景下的综合表现欠佳。Ridge 回归在当炉预测中的 MAE 为 4.01×10^{14} cm^{-3}、MAPE 达 2.97%、RMSE 为 5.94×10^{14} cm^{-3}、R^2 仅为 0.54；跨炉预测时各项指标进一步恶化，MAE 升至 8.93×10^{14} cm^{-3}，MAPE 达 6.61%，RMSE 升至 7.56×10^{14} cm^{-3}，R^2 降至 0.47。这表明基于线性假设的传统回归方法难以适应 SiC 外延过程的复杂非线性特性。SVR 模型虽通过非线性核函数在一定程度上改善了拟合能力，当炉预测 MAPE 为 3.84%，R^2 为 0.48，但其特征工程依赖性强，且无法建模炉次间的状态传递，导致跨炉预测 MAPE 仍达 5.01%，R^2 仅 0.43，与实际工业应用要求存在显著差距。

CNN 架构模型在各项指标上均优于传统机器学习方法。基础 CNN 模型当炉预测的 MAE 为 3.12×10^{14} cm^{-3}，MAPE 为 2.31%，RMSE 为 3.99×10^{14} cm^{-3}，R^2 为 0.61；跨炉预测时 MAPE 升至 3.40%，R^2 降至 0.59，表明其固定尺度卷积核难以全面捕捉工艺参数的多尺度动态特征。CNN+LSTM 模型通过引入时序建模能力，当炉预测 MAE 降至 3.11×10^{14} cm^{-3}，MAPE 为 2.30%，RMSE 为 3.56×10^{14} cm^{-3}，R^2 提升至 0.67；但由于 LSTM 在处理长序列时的固有局限以及全局耦合建模能力的缺失，其跨炉预测 MAPE 仍达 2.56%，RMSE 为 4.51×10^{14} cm^{-3}，性能提升有限。

Transformer 类模型凭借其全局建模优势取得了进一步改进。基础 Transformer 模型在当炉预测中 MAE 为 2.74×10^{14} cm^{-3}，MAPE 为 2.03%，RMSE 为 3.54×10^{14} cm^{-3}，R^2 为 0.68；但因缺乏专门的时序处理机制，跨炉预测时 MAE 升至 2.89×10^{14} cm^{-3}，MAPE 为 2.14%，表现不尽理想。CNN+Transformer 混合模型在局部特征捕捉与全局建模间取得平衡，当炉预测 MAE 为 2.05×10^{14} cm^{-3}，MAPE 降至 1.52%，RMSE 为 2.98×10^{14} cm^{-3}，R^2 提升至 0.76；跨炉预测 MAPE 为 1.73%，R^2 为 0.74，但其特征选择机制的缺失限制了进一步的性能提升。

本工作提出的 MFT-NET 模型在 4 个评估指标上均展现出最优性能：当炉预测 MAE 低至 1.83×10^{14} cm^{-3}，MAPE 仅为 1.35%，RMSE 为 2.81×10^{14} cm^{-3}，R^2 达 0.89；跨炉预测 MAE 为 2.25×10^{14} cm^{-3}，MAPE 为 1.66%，RMSE 为 3.42×10^{14} cm^{-3}，R^2保持在 0.87 的高水平。与次优的 CNN+Transformer 模型相比，MFT-NET 在当炉预测中的 MAE 降低约 10.7%，MAPE 改善 11.2%，RMSE 降低 5.7%；在跨炉预测中 MAE 降低 3.8%，MAPE 改善 4.0%，RMSE 降低 8.8%。这一全面优势得益于模型的多模块协同设计：多尺度卷积确保全尺度特征提取，SE 注意力实现特征自适应筛选，Transformer 模块建立全局参数耦合关系，GRU 组件有效传递跨炉状态信息，从而完整契合了 SiC 外延工艺的复杂特性。

2.4　消融实验

为验证 MFT-NET 模型中不同组件的必要性及协同作用，以多尺度卷积作为基线模型(Model-A)，逐步引入核心模块以构建一系列对比模型，具体设置如下。

- Model-A (基线)：仅包含多尺度卷积模块。该模型用于捕捉工艺参数的多尺度动态特征，作为性能对比的基准。
- Model-B：在 Model-A 的基础上引入 SE 模块。该设计用于验证通道注意力机制对增强关键参数特征、抑制冗余噪声的有效性。
- Model-C：在 Model-A 的基础上引入 Transformer 模块。该设计用于测试自注意力机制在建模全局参数耦合关系中的作用。
- Model-D：同时集成多尺度卷积、SE 模块与 Transformer 模块。该模型用于检验特征提取、筛选与全局建模三者协同工作的效果。
- Model-E (MFT-NET)：在 Model-D 的基础上引入 GRU 模块，构成完整的最终模型。该设计旨在验证跨炉次状态传递机制对提升泛化能力的关键贡献。

消融实验结果如表 3 所示。

表 3　消融实验结果

Table 3　Results of ablation experiments

Prediction scenario	Model	MAE/ $\times10^{14}$ cm^{-3}	MAPE/%	RMSE/ $\times10^{14}$ cm^{-3}	R^2
Current run prediction	Model-A	3.04	2.25	3.84	0.63
	Model-B	3.02	2.24	4.98	0.68
	Model-C	2.14	1.59	4.55	0.79
	Model-D	1.94	1.44	3.47	0.84
	Model-E (MFT-NET)	1.83	1.35	2.81	0.89
Next run prediction	Model-A	4.33	3.21	6.54	0.61
	Model-B	4.87	3.61	5.88	0.71
	Model-C	3.66	2.71	5.62	0.75
	Model-D	3.12	2.31	4.32	0.76
	Model-E (MFT-NET)	2.25	1.66	3.42	0.87

基于表 3 所示的消融实验结果，MFT-NET 模型中各组件对模型性能的贡献得到了系统验证。通过逐步引入核心模块，各项评估指标均呈现出有规律的提升，充分证明了模型设计的有效性。在仅含多尺度卷积的基线模型(Model-A)基础上引入 SE 注意力模块(Model-B)后，当炉预测的 MAPE 由 2.25%微降至 2.24%，R^2 从 0.63 提升至 0.68，表明特征筛选机制有效强化了关键参数的表征；然而跨炉预测的 MAPE 从 3.21%升至 3.61%，这一现象揭示了 SE 注意力机制在应对跨炉次状态变化时的结构性局限：SE 模块通过全局池化生成的通道权重本质上是基于单炉次数据的静态特征重要性评估，其筛选标准偏向于优化当前炉次的瞬时拟合效果，可能过度抑制了那些在单炉内波动较小但对炉间差异具有指示意义的潜在状态信号，从而导致模型在面临设备状态累积变化时的泛化能力受损。当进一步加入 Transformer 模块构成 Model-C 后，模型性能实现显著跨越。当炉预测 MAE 从 3.02×10^{14} cm^{-3} 大幅降至 2.14×10^{14} cm^{-3}，MAPE 改善至 1.59%，R^2 跃升至 0.79；跨炉预测中 RMSE 从 5.88×10^{14} cm^{-3} 降至 5.62×10^{14} cm^{-3}，R^2 达到 0.75，证明全局耦合关系建模是性能提升的关键驱动力。Model-D 集成多尺度卷积、SE 注意力和 Transformer 模块后，各项指标得到进一步优化：当炉预测 RMSE 降至 3.47×10^{14} cm^{-3}，MAPE 改善至 1.44%，R^2 提升至 0.84；跨炉预测 MAE 降至 3.12×10^{14} cm^{-3}，MAPE 为 2.31%，显示出多模块协同的显著效果。最终引入 GRU 模块的完整模型(Model-E，MFT-NET)在所有指标上均达到最优：当炉预测 MAE 低至 1.83×10^{14} cm^{-3}，MAPE 仅为 1.35%，RMSE 为 2.81×10^{14} cm^{-3}，R^2 达 0.89；跨炉预测 MAE 为 2.25×10^{14} cm^{-3}，MAPE 为 1.66%，RMSE 为 3.42×10^{14} cm^{-3}，R^2 高达 0.87。特别值得注意的是，与 Model-D 相比，完整模型在跨炉预测中的 RMSE 降低了 20.8%，MAE 降低了 27.9%，这充分证明 GRU 模块通过门控机制实现了跨炉状态传递，是解决长程依赖问题的核心组件。

实验结果表明，MFT-NET 的完整架构通过多尺度特征提取、注意力筛选、全局耦合建模和状态传递 4 个模块的有机协同，实现了对 SiC 外延工艺中多尺度动态、参数耦合和跨炉依赖等复杂特性的全面适配，为工业级浓度预测提供了可靠解决方案。

2.5　讨论

本节针对模型性能优化与工业化适配展开分析。首先分析核心超参数(Transformer 层数、多尺度卷积核组合、Dropout 比率)对模型性能的内在影响；其次对比加权均方误差与传统均方误差对模型性能的影响差异。此外，采用 SHAP 值法解析关键特征重要性及作用机制，明晰算法工业化应用流程与应用价值。本节围绕上述内容深入讨论，为模型优化与工程化落地提供支撑。

2.5.1　Transformer 编码器层数影响

Transformer 层数决定了模型对复杂工艺参数间全局耦合关系的建模深度。测试了 1 至 4 层编码器对模型性能的影响，结果如图 8 所示。

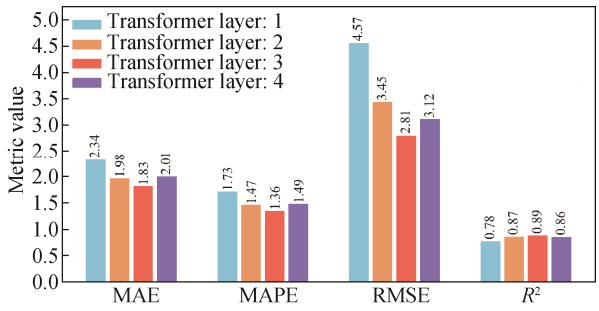


图 8　Transformer 层数对模型性能的影响

Fig. 8　Effect of Transformer encoder layer number on model performance

根据实验结果，随着 Transformer 层数的增加，模型性能呈现先上升后下降的趋势。当层数为 3 时，模型在当炉预测任务中达到最佳性能(MAPE=1.35%，R^2=0.89)。进一步增加至 4 层后，模型出现轻微的性能饱和现象，同时训练耗时显著上升，表明 3 层结构在模型复杂度与表达能力之间达到了最优平衡。更深层的网络可能因过拟合而导致泛化能力下降，因此后续实验均采用 3 层 Transformer 作为基准配置。

2.5.2　多尺度卷积核组合策略

多尺度卷积模块用于捕捉参数在不同时间尺度上的动态变化。为验证所设计的 ([3, 3]，[9, 3]，[21, 5]) 卷积组合的有效性，本工作将其与其余 2 种组合进行了对比，结果如图 9 所示。

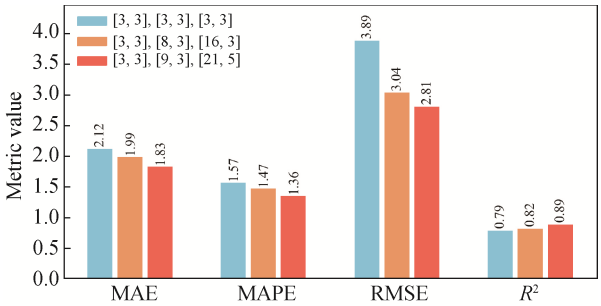


图 9　多尺度卷积核组合策略对模型性能的影响

Fig. 9　Effect of multi-scale convolution kernel combination strategy on model performance

实验数据显示，随着卷积核组合从单一尺度([3, 3]) 过渡到多尺度 ([3, 3], [8, 3], [16, 3])，再到提出的 ([3, 3], [9, 3], [21, 5]) 组合，模型性能呈现逐级提升态势。其中，最优组合在当炉预测中取得了 RMSE=2.81×10^{14} cm^{-3}、R^2=0.89 的显著优势，较单一尺度组合误差降低明显。结果表明，所设计的 ([3, 3], [9, 3], [21, 5]) 组合能够有效捕捉从秒级波动至小时级缓变的跨尺度工艺动态，为后续特征融合与状态传递提供了更具判别力的表征基础。

2.5.3　Dropout 比率影响　　Dropout 是控制模型过拟合的关键正则化超参数。测试了不同比率对模型泛化能力的影响，重点关注其在不同预测场景下的稳定性，结果如图 10 所示。

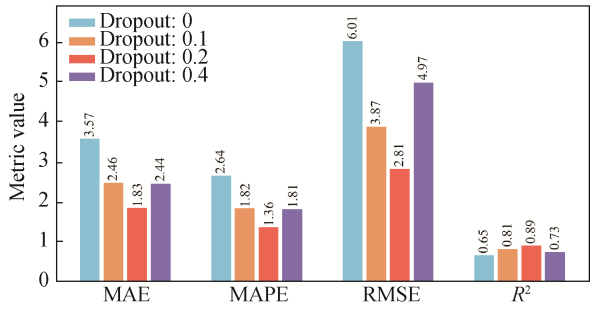


图 10　Dropout 对模型性能的影响

Fig. 10　Effect of Dropout ratio on model performance

实验发现，当 Dropout 比率为 0 时，模型在当炉预测场景上的性能急剧下降，表现出严重的过拟合。随着比率增加，模型在跨炉场景下的泛化能力持续提升，在比率为 0.2 时，模型在双场景下取得了最佳的综合性能。比率继续增加至 0.4 会导致模型学习能力不足，性能出现下降。因此，0.2 被确定为最优比率，能在保持强大拟合能力的同时，最大化模型的泛化性能。

2.5.4　MSE 损失和加权 MSE 结果对比　　相较于传统 MSE，加权 MSE 通过动态赋予偏离样本差异化权重，可有效抑制模型过拟合风险：一方面确保模型优先习得主流工艺规律，另一方面能够兼顾偏离样本的特征学习，实现对不同类型样本的均衡拟合。2 种损失函数的预测结果对比分析如图 11 所示。

与传统 MSE 损失相比，加权 MSE 在预测值与实际浓度的 MAPE 和 RMSE 均表现为更小，且对浓度变化规律的解释与拟合能力更优。鉴于 SiC 外延浓度分布具有波动性特征，加权 MSE 的动态权重策略可有效抑制对偏离样本的过拟合，更契合工业生产的实际需求。

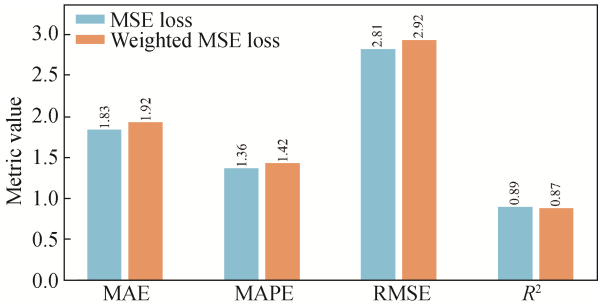


图 11　加权 MSE 损失函数和 MSE 损失函数对模型性能的影响

Fig. 11　Impact of weighted MSE loss function and MSE loss function on model performance

2.5.5　SHAP 分析　　基于博弈论 Shapley 值原理，SHAP [28]可赋予深度学习 “黑箱模型” 可解释性，其不仅能量化工艺参数贡献度，还能明晰参数数值变化与浓度预测结果的内在关联，为工艺参数调整提供明确方向；同时，SHAP 可解构单一炉次浓度预测结果的成因，并通过验证模型预测与 SiC 外延物理工艺常识的一致性，实现过拟合及数据异常的及时排查，避免模型预测局限于单纯数值输出。MFT-NET 在 SiC 外延掺杂浓度预测任务中的 SHAP 分析结果如图 12 所示。

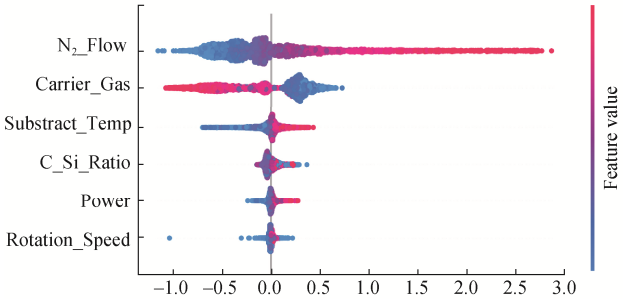


图 12　MFT-NET SHAP 分析结果

Fig. 12　SHAP analysis results of MFT-NET

SHAP 的分析结果与工程师认知高度吻合，进一步验证了模型预测的可靠性。关键参数的贡献方向符合预期：作为主要氮源的 N_2流量与掺杂浓度呈显著正相关；而作为稀释剂的载气流量则与浓度呈负相关。同时，衬底温度与射频功率作为影响反应动力学与活化效率的关键因素，亦显示出明确的正向贡献。值得注意的是，C/Si 比与转速的影响在模型中未呈现明确趋势，反映了实际工艺中多参数耦合的复杂性。

2.5.6　工业应用与价值　　基于 MFT-NET 模型，为 SiC 外延生长工业化生产构建 “数据驱动–智能

决策–精准调控”智能调控系统，形成全流程闭环方案。实际应用中，系统实时采集 SiC-CVD 设备工艺参数，模型据此输出核心层掺杂浓度预测值；若预测值超出预设区间，系统依托模型参数关联规律与耦合机制，定向生成多参数微调建议。工程师核验建议后执行调控，同时将调控后的数据反馈至模型训练库，驱动迭代，以持续优化系统性能。基于 MFT-NET 的智能调控系统计算流程如图 13 所示。

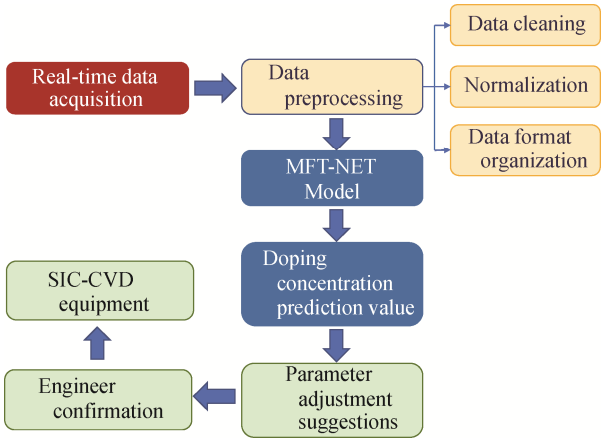


图 13　MFT-NET 赋能工业生产流程图

Fig. 13　Industrial production flow chart enabled by MFT-NET

从实际应用价值来看，该系统在成本控制与产品一致性管理上均体现出显著优势：在成本层面，该系统通过 MFT-NET 模型的当炉预测功能，可对约 30%的非关键炉次实现离线测试替代，结合传统测试中设备折旧、人工及产能损失等核心成本项，实现了离线测试成本的有效降低。在产品一致性层面，该系统通过模型驱动的精准调控，有效控制了 SiC 外延产品的浓度一致性(基于浓度标准差、超标炉次占比等关键指标加权核算，一致性较传统人工经验调控模式提升约 15%)，避免了因批次间差异导致的生产损耗与质量风险。

3　结论

针对 SiC 外延生长过程中核心层掺杂浓度调控所面临的响应滞后、经验依赖性强及测试成本高等挑战，本文提出了一种面向双预测场景的 MFT-NET 模型。通过集成多尺度卷积、SE 注意力机制、参数级 Transformer 与 GRU 状态传递模块，该模型实现了对局部动态、全局耦合与跨炉依赖的综合建模。实验表明，MFT-NET 在当炉与跨炉预测中均表现优异，当炉预测 MAPE 仅为 1.35%、R^2 达 0.89；跨炉预测 MAPE 为 1.66%、R^2 为 0.87，显著优于各类基线模型。本工作不仅为 SiC 外延掺杂浓度的精准预测提供了有效方法，也为工业环境下工艺智能调控与质量优化提供了可靠的技术路径。

参考文献:

[1] BUFFOLO M, FAVERO D, MARCUZZI A, et al. Review and outlook on GaN and SiC power devices: Industrial state-of-the-art, applications, and perspectives[J]. IEEE Trans Electron Devices, 2024, 71(3): 1344–1355.

[2] CHEN X F, YANG X L, XIE X J, et al. Research progress of large size SiC single crystal materials and devices[J]. Light Sci Appl, 2023, 12: 28.

[3] GHEZELLOU M, UL-HASSAN J. Influence of different hydrocarbons on chemical vapor deposition growth and surface morphological defects in 4H-SiC epitaxial layers[J]. Phys Status Solidi B: , 2024, 261(4): 2300535.

[4] SATO R, KUTSUKAKE K, HARADA S, et al. Machine learning for semiconductor process simulation described by coupled partial differential equations[J]. Adv Theory Simul, 2023, 6(9): 2300218.

[5] WANG Z H, LI L, LEON R C C, et al. Improving semiconductor device modeling for electronic design automation by machine learning techniques[J]. IEEE Trans Electron Devices, 2024, 71(1): 263–271.

[6] CHEN A, YE S M, WANG Z L, et al. Machine-learning-assisted rational design of 2D doped tellurene for fin field-effect transistor devices[J]. Patterns, 2023, 4(4): 100722.

[7] ANGER S, YAYAK A B, ROEDER G, et al. Combining full wafer inspection with deep learning to recognize wafers with critical defects[C]//2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). Saratoga Springs, NY, USA. IEEE, 2023: 1–6.

[8] TANG Z C, LI H, LIN P Z, et al. A deep equivariant neural network approach for efficient hybrid density functional calculations[J]. Nat Commun, 2024, 15(1): 8815.

[9] KIM T Y, PARK S, LIM C K, et al. Deep learning-based detection of defects in wafer buffer zone during semiconductor packaging process[J]. Multiscale Sci Eng, 2024, 6(1): 25–32.

[10] WEI S, HAO M L, YU L N, et al. Inverse design of GaN HEMT heterostructures *via* deep learning with enhanced auto-encoder framework[J]. Mater Today Commun, 2025, 46: 112629.

[11] SALIMIAN A, AMINISHAHSAVARANI A, UPADHYAYA H. Artificial neural networks to predict sheet resistance of indium-doped zinc oxide thin films deposited *via* plasma deposition[J]. Coatings, 2022, 12(2): 225.

[12] LIM B, ARıK S Ö, LOEFF N, et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting[J]. Int J Forecast, 2021, 37(4): 1748–1764.

[13] ZHOU H Y, ZHANG S H, PENG J Q, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[J]. Proc AAAI Conf Artif Intell, 2021, 35(12): 11106–11115.

[14] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. arXiv preprint arXiv: 1506.04214.

[15] DONG R, SONG Y, SIRIWARDANE E M D, et al. Discovery of 2D materials using Transformer network-based generative design[J]. Adv Intell Syst, 2023, 5(12): 2300141.

[16] DEY R, SALEM F M. Gate-variants of gated recurrent unit (GRU)

neural networks[C]//2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Boston, MA, USA. IEEE, 2017: 1597–1600.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 7132–7141.

[18] BENESTY J, CHEN J D, HUANG Y T, et al. Pearson correlation coefficient[M]//Noise Reduction in Speech Processing. Berlin, Heidelberg: Springer, 2009: 1–4.

[19] YU Q Y, LI L, ZHAO H, et al. Evaluation system and correlation analysis for determining the performance of a semiconductor manufacturing system[J]. Complex Syst Model Simul, 2021, 1(3): 218–231.

[20] SHAPIRO S S, WILK M B. An analysis of variance test for normality (complete samples)[J]. Biometrika, 1965, 52(3–4): 591–611.

[21] WANG Meng, LIANG Yuchao, GU Mei. The application of Pearson's algorithm in semiconductor processes[J]. J China Integrated Circuit, 2022, 31(5), 72–75, 81.

[22] LLUGSI R, EL YACOUBI S, FONTAINE A, et al. Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito[C]//2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). Cuenca, Ecuador. IEEE, 2021: 1–6.

[23] TAYMAN J, SWANSON D A. On the validity of MAPE as a measure of population forecast accuracy[J]. Popul Res Policy Rev, 1999, 18(4): 299–322.

[24] HODSON T O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not[J]. Geosci Model Dev, 2022, 15(14): 5481–5487.

[25] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Comput Sci, 2021, 7: e623.

[26] CAMERON A C, WINDMEIJER F A G. An R-squared measure of goodness of fit for some common nonlinear regression models[J]. J Econometrics, 1997, 77(2): 329–342.

[27] AWAD M, KHANNA R. Support vector regression[M]//Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley, CA: Apress, 2015: 67–80.

[28] MOSCA, E., SZIGETI, F., TRAGIANNA, S., et al. SHAP-based explanation methods: a review for NLP interpretability[C]//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea. 2022: 4593–4603.

A SiC Epitaxial Growth Doping Concentration Prediction Model Based on Multi-Scale Feature Enhancement and Temporal Transformer

ZHANG Zhongyi[1,2], *WANG Lang*[1,2,3], *LU Weili*[1,2,3], *LI Shuai*[1,2,3], *YANG Jianye*[1,2,3], *GAO Nan*[1,2,3], *WANG Bo*[1,2,3], *PAN Guoping*[1,2,3], *FANG Yulong*[1,2,3]

(1. The 13th Research Institute, CETC, Shijiazhuang 050051, China; 2. Hebei Province Semiconductor Intelligent Manufacturing Innovation Center, Shijiazhuang 050051, China; 3. Shijiazhuang Key Laboratory of Semiconductor Artificial Intelligence Technology, Shijiazhuang 050051, China)

Extended Abstract

Introduction Silicon carbide (SiC), as a representative third-generation wide-bandgap semiconductor material, exhibits superior performance over traditional silicon-based devices in high-voltage, high-power, and high-frequency applications. The doping concentration of the SiC epitaxial layer critically determines the performance of power devices. However, current doping control heavily relies on manual experience and offline testing, which incurs high costs and significant time delays. This study aims to address these limitations by proposing a data-driven model for accurate and real-time prediction of doping concentration in both *in-situ* and next-batch scenarios.

Methods This paper proposes a Multi-scale Feature-enhanced Temporal Transformer Network (MFT-NET) for predicting the core-layer doping concentration during SiC epitaxial growth. The model integrates four key modules:

Multi-scale Convolution Module: Employs parallel convolutional branches with kernel sizes of [3,3], [9,3], and [21,5] to capture parameter dynamics across millisecond-to-hour scales.

Squeeze-and-Excitation (SE) Module: Enhances feature channels correlated with doping concentration and suppresses noise *via* channel-wise attention.

Transformer Module: Models global coupling relationships among all 75 process parameters using multi-head self-attention.

Gated Recurrent Unit (GRU) Module: Captures cross-batch state dependencies to account for residual dopant and equipment state drift between consecutive epitaxial runs.

The model was trained and evaluated on a dataset comprising 1200 epitaxial runs from a commercial SiC epitaxial reactor. Data preprocessing included Pearson correlation-based feature selection and min-max normalization. A weighted mean squared error loss function was adopted to prioritize mainstream process samples.

Results and Discussion Experimental results demonstrate that MFT-NET significantly outperforms traditional models including Ridge Regression, Support Vector Regression (SVR), CNN, CNN+LSTM, Transformer, and CNN+Transformer. In the *in-situ* prediction scenario, MFT-NET achieved a mean absolute percentage error (MAPE) of 1.35%, a root mean square error (RMSE) of

2.81×10^{14} cm^{-3}, and a coefficient of determination (R^2) of 0.89. For next-batch prediction, it attained a MAPE of 1.66%, RMSE of $3.42\times10^{14}cm^{-3}$, and R^2 of 0.87.

Ablation studies confirmed the contribution of each module: the multi-scale convolution and SE modules improved feature extraction and selection; the Transformer enabled global parameter interaction modeling; and the GRU was crucial for capturing cross-batch state transfer, reducing next-batch prediction RMSE by 20.8% compared to the model without GRU.

Hyperparameter analysis revealed that a 3-layer Transformer, the proposed multi-scale kernel combination, and a Dropout ratio of 0.2 yielded optimal performance.

Conclusions The MFT-NET model effectively addresses the multi-scale dynamics, global parameter coupling, and cross-batch state transfer challenges in SiC epitaxial growth. It provides highly accurate doping concentration predictions for both *in-situ* monitoring and next-batch forecasting, with the potential to reduce offline testing costs by 30% and improve process stability by 15%. This work offers a robust data-driven framework for intelligent optimization of SiC epitaxial processes.

Keywords silicon carbide epitaxial growth; core-layer doping concentration; multi-scale convolution; Transformer; process parameter prediction; deep learning