ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 4,Jul. 2026,pp. 1103－1117
DOI:10. 16337/j. 1004-9037. 2026. 04. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于扩散模型改进的数字孪生网络流量生成技术

赵晓红[1]，陈　浩[1]，施冠炬[2]，管　荑[3]，丁旭辉[2]，朱　超[2]

(1. 国网山东省电力公司电力科学研究院科技信息中心，济南 250003；2. 北京理工大学网络空间安全学院，北京 100081；3. 国网山东省电力公司电力调度控制中心，济南 250001)

摘　要：在工业4.0、智慧城市等领域，数字孪生技术正在被广泛应用，其通过构建与物理实体一一对应的虚拟模型，能够实现对设备、产品和系统的实时监控、预测性维护及优化管理。然而，为数字孪生高效生成准确的网络流量数据包仍然是一项极具挑战的任务。该任务具有时序依赖性，且在复杂网络行为和多样化场景的影响下，流量生成过程存在高度的不确定性和复杂性。为了解决这一问题，本文提出了基于扩散模型的大批量网络流量数据包生成算法——FlowDiff和时序扩散生成模型(Temporal diffusion generative model，TDGM)。FlowDiff算法将流量数据包的生成作为反向扩散过程，根据流量的时序特性和外部场景条件，通过逐步去噪生成符合实际网络流量特征的数据包。TDGM是为了适应网络流量中的时序依赖和周期性变化特性设计的模型。该模型引入时序感知特征嵌入层，将流量的时序信息与重要特征进行融合，从而增强了流量特征之间的时空关联。模型还结合卷积神经网络和Transformer模块，提取流量的局部特征和全局特征，进行有效融合。最后，将历史流量数据和周期性特征作为扩散模型的条件输入，利用交叉注意力机制进一步优化生成过程。实验结果表明，FlowDiff在真实网络流量数据集上的表现优异，在与真实数据流量的比较中，充分验证了生成的准确性和有效性，展示了其在数字孪生系统中应用的有效性和优势。

关键词：网络流量生成；扩散模型；数字孪生；深度学习；时序依赖

中图分类号：TP393　　**文献标志码：**A

引用格式：赵晓红，陈浩，施冠炬，等. 基于扩散模型改进的数字孪生网络流量生成技术[J]. 数据采集与处理，2026，41(4)：1103-1117. ZHAO Xiaohong，CHEN Hao，SHI Guanju，et al. Improved network traffic generation technology for digital twins based on diffusion model[J]. Journal of Data Acquisition and Processing，2026，41(4)：1103-1117.

引　　言

数字孪生技术[1]是一种将物理实体与其虚拟副本实时映射的新型智能技术，正在逐步成为工业4.0、智慧城市和智能制造等领域的关键支撑工具。通过构建与现实对象一一对应的虚拟模型，数字孪生不仅能够实现设备运行状态的实时监控，还可支持预测性维护与优化管理等功能。然而，构建高精度、功能完善的数字孪生系统依赖于大量高质量的数据支持。这些数据涵盖了设备运行参数、生产过程状态、环境变量、系统健康指标以及用户反馈等多个维度，对模型的实时性与准确性至关重要。数字孪生的高质量数据和信息需要强大的网络支撑，尤其是在安全防护、数据传输和隐私保护等方面，同样需要能够精确模拟网络行为和流量特征的模型，这些模型通常依赖于真实的网络流量数据。然而，获取大量真实的网络流量数据往往受到隐私、法律和技术限制[2]。因此，如何在不侵犯隐私和不违反法律的前提下生成大量逼真的网络流量，成为数字孪生研究中亟需解决的关键问题[3-4]。

基金项目：国网山东省电力公司科技项目(52062624000K)。
收稿日期：2025-07-17；**修订日期：**2025-09-14

在网络流量生成领域,已有大量研究致力于解决如何从真实流量中学习并生成逼真的网络数据。早期生成研究主要基于统计建模方法,如马尔可夫链和自回归模型,文献[5]提出了一种基于数据包级别的源基础流量建模方法,使用隐马尔可夫模型刻画流量的时间依赖性与分布特性,具备良好的灵活性,适用于多种应用层协议。文献[6]则探讨了自回归(Autoregressive ,AR)模型在网络流量时间序列建模中的优势,并回顾了包括自回归移动平均(Autoregressive moving average,ARMA)模型、自回归积分移动平均(Autoregressive integrated moving average,ARIMA)模型在内的多种自回归变体在流量预测中的应用。上述方法通过构建概率分布模型,能够在一定程度上生成或预测符合特定模式的网络流量。然而,这类传统方法在应对大规模或动态变化的网络环境时,常面临模型精度有限、计算复杂度高等问题,难以有效捕捉复杂的时序依赖、非平稳性与突发流量变化,从而限制了其在实际场景中的适用性。

近年来,随着深度学习和生成对抗网络(Generative adversarial network,GAN)的兴起,基于深度学习的网络流量生成方法逐渐成为研究的热点。例如,文献[7]提出的基于生成对抗网络和自编码器(Autoencoder,AE)的异常检测模型,通过数据预处理、重采样和重构等多个环节,有效地提升了对异常流量的识别能力。然而,该方法仍存在一定的局限性,尤其是在处理大规模复杂网络流量时,模型的训练效率较低,且在面对高度不平衡的流量数据时,生成效果稳定性和准确性难以保证。此外,文献[8]分别采用条件表格生成对抗网络(Conditional tabular generative adversarial network,CTGAN)[9]和三元组变分自编码器(Triplet-based variational autoencoder,TVAE)[10]生成5G网络表格数据,并通过统计分布拟合与皮尔逊相关性等方法对生成质量进行评估。尽管该方法在表格数据层面取得了一定效果,但其在依赖精确网络拓扑信息的同时,难以有效建模网络流量中的时序依赖与周期性波动,尤其在面对动态变化的流量环境时,生成结果在多样性与准确性方面仍有较大改进空间。

扩散模型(Diffusion model,DM)作为一种强大的生成模型,在生成图像等领域表现出了卓越的能力[11]。与图像数据相比,网络流量具有更高的信息密度,且通常包含更短且高度结构化的片段,扩散模型则能够通过精细化的去噪过程,捕捉和生成复杂的时序依赖和周期性变化,弥补传统生成模型在面对复杂流量模式时的不足。尤其在动态和多变的网络环境中,扩散模型的灵活性和强大的生成能力使其成为解决流量生成问题的理想选择[12]。

针对网络流量数据具有特征稀疏、结构紧凑等特点,本文在模型结构上对文字编码器进行轻量化设计,以提升在流量数据场景下的处理效率与建模精度。同时,考虑到网络流量的时序性与数据包间的相关性,提出一种结合扩散模型与时间序列分析的新型生成方法——时序扩散生成模型(Temporal diffusion generative model,TDGM)。该模型通过引入傅里叶变换构建周期性特征,使生成过程能够更好地捕捉网络流量中的周期波动与长短期依赖关系。实验表明,TDGM不仅生成样本更贴近真实流量模式,也更适应复杂多变的网络环境,有助于提升数字孪生系统的实用性。

1 扩散模型

扩散模型作为一种生成模型,近年来在机器学习领域取得了显著进展,尤其在生成高质量图像和文本等任务中展现出了卓越的性能。扩散模型的理论基础源于去噪扩散概率模型(Denoising diffusion probabilistic model, DDPM)[13],该模型通过正向和反向的马尔可夫链过程对数据进行噪声的添加与去除。在正向过程(噪声添加过程)中,数据逐渐被转化为简单的噪声分布(如高斯噪声);而在反向过程(去噪过程)中,模型利用深度神经网络逐步恢复数据,逆转噪声的添加过程。通过这一过程,扩散模型能够模拟复杂数据分布的演化,进而实现对数据分布的有效学习与生成。这一机制使得扩散模型在生成任务中具有较强的表现力和灵活性,特别是在高维数据生成方面取得了突破性进展[14]。

1.1 前向扩散

前向扩散过程[15]旨在将原始数据通过多步噪声注入过程转化为纯噪声。设数据点x_0服从真实数据分布$q(x)$,在每个时间步$t\in 1,2,\cdots,T$中,通过马尔可夫链递归添加高斯噪声,得到

$$x_t=\sqrt{1-\beta_t}\,x_{t-1}+\sqrt{\beta_t}\,\epsilon \quad \epsilon\sim\mathcal{N}(0,I) \tag{1}$$

式中:I为高斯噪声的方差;β_t为在时间步t的噪声强度参数,其控制了噪声的量。β_t是在(0,1)范围内递增的小值,使得数据逐渐被噪声覆盖,x_t采样自真实数据分布$q(x)$,是时间t对应的扩散状态。

通过不断地应用式(1)递归公式,可以从原始数据x_0开始,经过T次迭代,生成接近高斯噪声的x_T。这个过程从根本上保证了数据逐步丧失其原有特征,最终在T步后接近一个已知的标准正态分布$\mathcal{N}(0,I)$。为了更加直观地看到这一过程,可以使用累积分布的表示形式,将每个时间步数据x_t直接表示为x_0加上独立的高斯噪声叠加,即

$$x_t=\sqrt{\bar{\alpha}_t}\,x_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon \tag{2}$$

式中:$\alpha_t=1-\beta_t$,那么由于每一步添加的噪声独立,且服从正态分布,$\bar{\alpha}_t=\prod_{s=1}^{t}(1-\beta_s)$,可以表示为从0到$t$时间步的所有$(1-\beta_s)$的乘积;$\epsilon$为标准正态分布的噪声,即$\mathcal{N}(0,I)$。

通过式(2),可以看到在早期时间步,数据x_t仍然保持了相对较多的原始信息,但随着t的增加,$\sqrt{\bar{\alpha}_t}$的值逐渐减小,噪声的影响逐渐占主导地位。鉴于数据包的可视性较低,故对数据包原始数值进行了加噪处理,并采用统计量进行表征,得到如图1所示的结果。图1中展示了一个完整pcap包在分别添加噪声100次、500次、1 000次、2 000次、3 000次和4 000次后的情况。通过这些噪声水平的对比,可以直观地观察到不同噪声强度下数据包整体表现的变化。

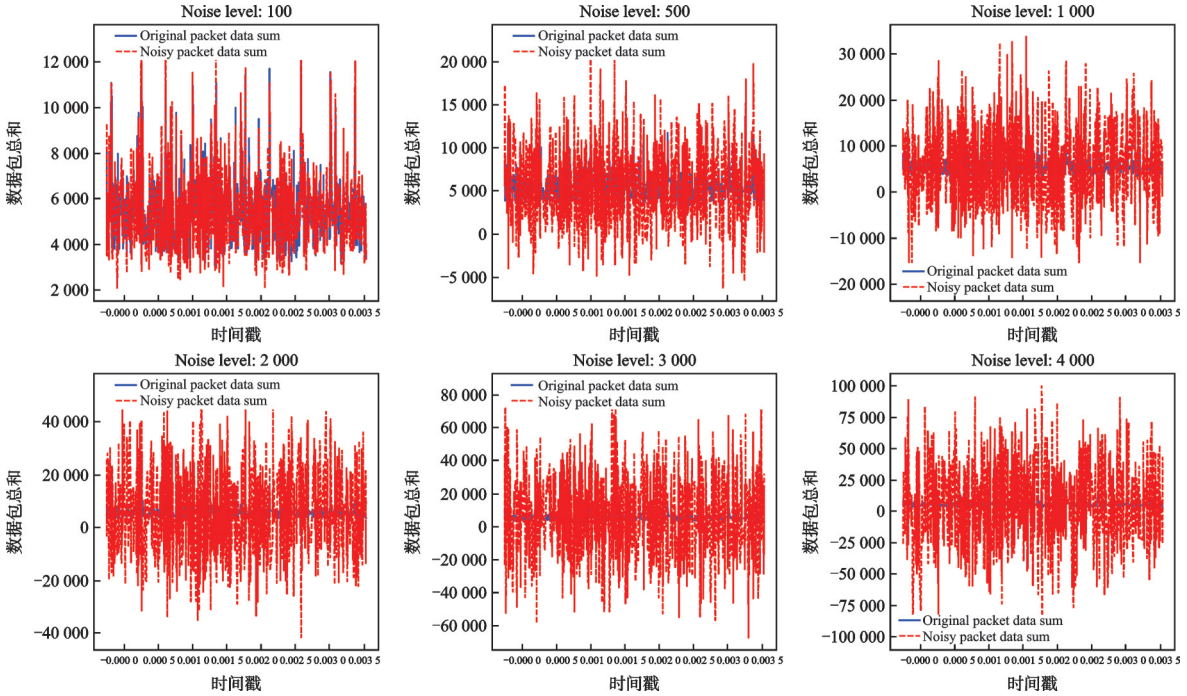

图1 逐步加噪过程

Fig.1 Step-by-step noise addition process

1.2 反向扩散

反向扩散过程[15]可视为前向扩散的逆过程，目标是从x_T(纯噪声)开始，逐步去噪，生成更加接近原始数据x_0的样本。反向扩散即从一个简单的初始状态(如白噪声)逐步去噪，直至还原出一个复杂的目标分布(如图像、文本或其他高维数据)。这个过程可以被形式化为一个多步骤的马尔可夫链，学习一个网络模型p_θ去近似这个条件概率方便反向扩散过程，使数据逐渐从噪声分布演化为目标分布，即

$$p_\theta(x_{0:T})=p(\boldsymbol{x}_T)\prod_{t=1}^{T}p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{3}$$

则每一个反向时间步t的去噪操作可以表示为$p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t))$。$\mu_\theta(x_t,t)$为由神经网络学习到的均值函数，表示从时间步$t$的样本$x_t$生成$t-1$时刻样本$x_{t-1}$的均值。$\Sigma_\theta(x_t,t)$为生成样本的方差，通常也由神经网络来参数化。显然，在已知初始状态x_0下，更容易求解，均值函数与方差范围被确定，故逆向的条件概率可以表示为

$$q(\boldsymbol{x}_{t-1}|(\boldsymbol{x}_t,\boldsymbol{x}_0))=\mathcal{N}(\boldsymbol{x}_{t-1};\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t,\boldsymbol{x}_0),\tilde{\beta}_t\boldsymbol{I}) \tag{4}$$

经过数学推导得到下面公式

$$q(\boldsymbol{x}_{t-1}|(\boldsymbol{x}_t,\boldsymbol{x}_0))=\exp\left(-\frac{1}{2}\left(\frac{\boldsymbol{x}_t^2-2\sqrt{\alpha_t}\,\boldsymbol{x}_t\boldsymbol{x}_{t-1}+\alpha_t\boldsymbol{x}_{t-1}^2}{\beta_t}+\frac{\boldsymbol{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\,\boldsymbol{x}_0\boldsymbol{x}_{t-1}+\bar{\alpha}_{t-1}\boldsymbol{x}_0^2}{1-\bar{\alpha}_{t-1}}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right)\right) \tag{5}$$

式中：$\tilde{\beta}_t=\dfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$；$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)=\dfrac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}+\dfrac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$；$x_t=\sqrt{\bar{\alpha}_t}\,x_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon$。故$\tilde{\boldsymbol{\mu}}_t=\dfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\dfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)$。

训练扩散模型的目标是通过最大化似然估计，使得模型能够准确地从噪声中生成数据。

2 TDGM模型

TDGM旨在融合扩散模型与时间序列分析方法，以生成更真实、复杂的网络流量数据。传统流量生成模型多依赖于统计方法或GAN[16]的框架，难以有效应对具有强时序依赖和复杂时空特性的网络流量，存在建模能力弱、生成结果不稳定等问题。

TDGM在设计上引入了两方面的改进：一是通过集成潜在空间生成模块与时序捕捉模块[17]，显式地对流量的周期性与渐进性特征进行建模，从而更精确地还原网络行为模式[18]；二是赋予模型条件控制能力，使其能够根据不同的网络场景、时间段或语义描述，灵活生成多样化的流量数据。与传统模型相比，TDGM展现出更强的泛化能力和场景适应性。

2.1 模型架构

实验模型架构参考潜在空间生成模型，由Encoder文字编码器、变分自编码器(Variational autoencoder, VAE)和U型卷积网络(U-shaped convolutional network, UNet)3个模块组成。

(1)Encoder：文字编码器由嵌入层、归一化层、自注意力层和前馈神经网络层组成[19-20]，如图2所示，用于对输入文本进行特征提取与表示学习。该编码器参考了CLIP架构，并在注意力模块和前馈网络后引入双重残差连接，以增强模型的数值稳定性和梯度传播能力。相较于原始设计，残差结构保留了层间信息流，有助于在与非线性特征融合时保持语义一致性，尤其在前馈网络后配合自定义激活函数的非线性变换，显著提升了高维语义建模能力及训练过程中的泛化表现。

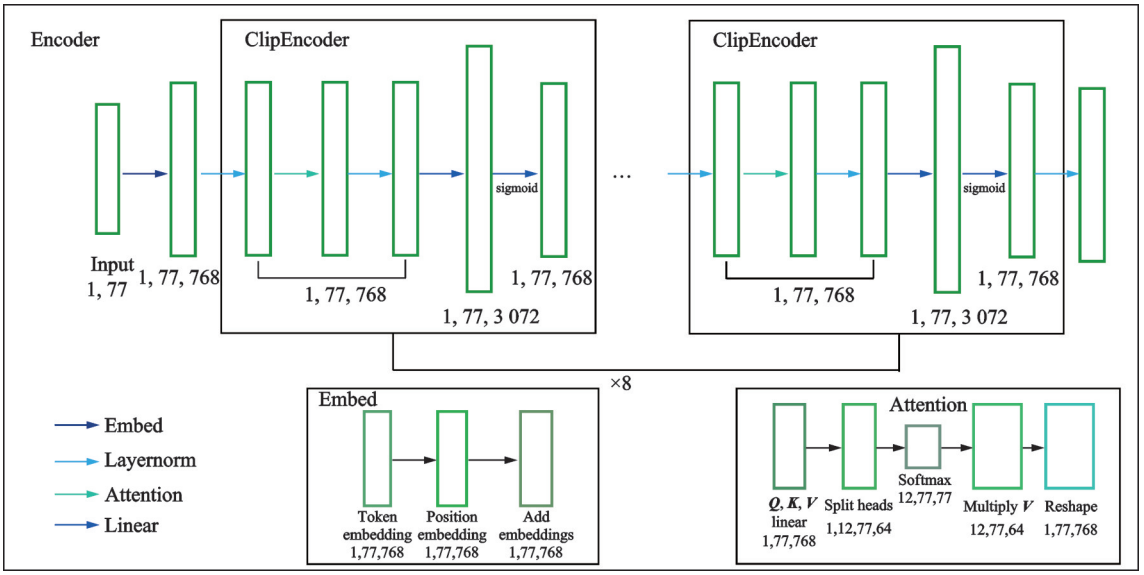

图2 Encoder模型架构

Fig.2 Encoder model architecture

(2)VAE:VAE模块由编码器、潜在变量采样器和解码器组成,如图3所示。编码器包含多层卷积、残差块、归一化与自注意力结构,负责将输入数据压缩为潜在空间的均值与方差分布[21]。采样器利用重参数化技巧从该分布中采样潜在变量[22],确保训练过程的可微性。解码器则由反卷积、残差块、归一化层和上采样结构构成,用于逐步恢复特征图的空间分辨率,并生成与输入一致尺寸的网络流量样本。该潜在空间压缩机制不仅能有效减少冗余信息,还可增强对流量包序列、协议分布等高维流量特征的建模能力,结合注意力机制,进一步提升了对时序依赖和全局语义的捕捉效果。

(3)UNet:UNet架构由下采样层、中间处理层和上采样层组成,如图4所示。下采样部分通过卷积、残差结构与自注意力模块逐层提取深层特征,并缩减空间维度;中间部分引入残差网络与Trans-

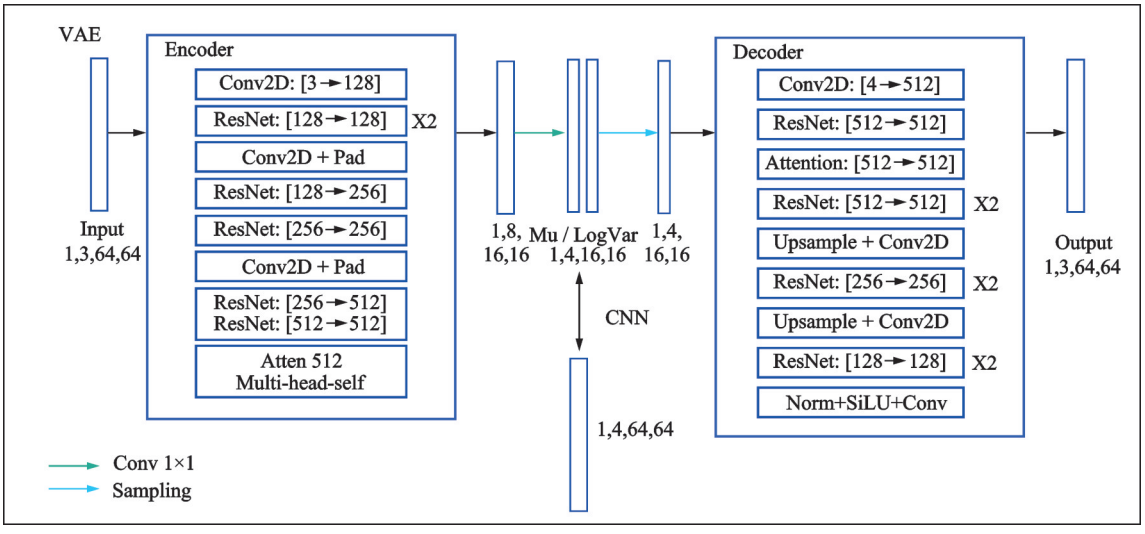

图3 VAE模型架构

Fig.3 VAE model architecture

former结构,用于捕捉全局上下文及长程依赖;上采样部分结合反卷积与跳跃连接逐步恢复空间维度,同时保留底层的细粒度信息。通过引入时间步嵌入,该架构能够在特征提取过程中考虑时间位置编码,为后续扩散过程提供时序感知能力。

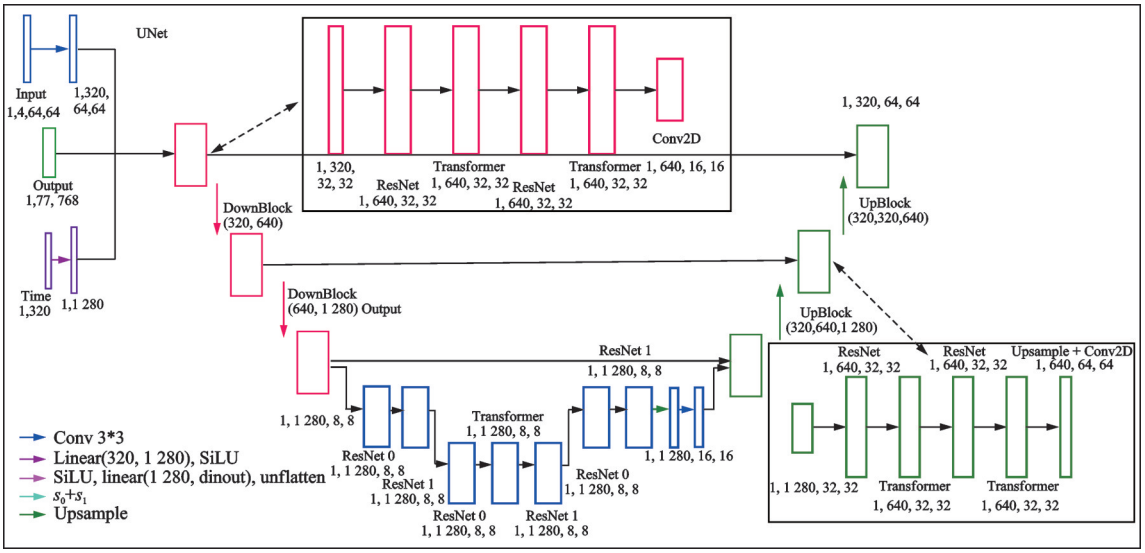

图4 UNet模型架构

Fig.4 UNet model architecture

考虑到网络流量在时间维度上的周期性波动,TDGM引入傅里叶变换、正弦波函数精确建模这些周期性特征。网络流量通常呈现与日夜节律、节假日等因素相关的周期性变化,为此,模型采用如下形式表达流量在时间t时刻的周期性行为

$$P(t)=\sum_{i=1}^{n}A_i\sin(\omega_i t+\phi_i) \tag{6}$$

式中:A_i为周期性波动的幅度;ω_i为周期性波动的频率,通常与日周期或周周期相关;ϕ_i为周期的相位。这些周期性成分通过傅里叶分析提取后,与原始流量特征一同输入模型,为生成过程提供辅助引导。

在潜在空间建模部分,TDGM将周期性特征与经过VAE编码器压缩后的网络流量特征进行拼接,构造融合的潜在表示。假设x为PCAP张量,$p(t)$是周期性特征,那么模型的输入构造如下

$$Z=f_{\text{encode}}([x,p(t)]) \tag{7}$$

式中:f_{encode}为编码器网络,$[x,p(t)]$表示拼接后的PCAP数据和周期性特征。通过这种方式,周期性特征与网络流量数据共同作用于潜在空间,使得生成模型不仅能够学习到网络流量的时序依赖关系,还能够生成具有周期性波动的网络流量。

在解码与生成阶段,融合后的潜在表示通过解码器映射回原始维度,生成的网络流量样本不仅保留了长期趋势,还重构了典型的周期性特征。这一机制显著提升了生成流量的真实性与时间一致性,使模型在复杂网络环境中具有更强的适应性与泛化能力。最终,TDGM能够输出结构合理、变化规律贴近实际网络状态的合成流量样本,为数字孪生系统提供高质量的数据支撑。

2.2 算法实现

为实现网络流量数据的时序特征建模与噪声学习,设计了FlowDiff算法,其具体流程如算法1所示。首先,根据每条流量记录在时序序列中的位置构造时间步索引T_{step},以增强模型对流量前后依赖关

系的感知能力。随后,利用预训练的VAE编码器将输入特征映射至潜在空间,获得紧凑的潜在特征F_{VAE}。在此基础上,从标准正态分布中采样随机噪声z,并通过噪声缩放因子α将其注入潜在特征,得到加噪特征F'。同时,根据文本描述d_{text}提取目标噪声信息T_{noise},并以加噪特征与目标噪声之间的均方误差L_{MSE}作为优化目标。通过最小化该损失,模型能够学习网络流量潜在特征与目标噪声之间的映射关系,从而为后续扩散过程中的噪声预测与流量数据生成提供基础。

算法1 FlowDiff算法

Input: X原始网络流量数据集,T_{step}用于标识每条记录在时序序列中的位置的时间步数索引,F_{VAE}预训练的VAE编码器,α噪声缩放因子,d_{text}用于提取目标噪声信息的文本描述。

Output: L_{MSE}均方误差损失。

begin

(1) 时序感知特征嵌入:为每条流量记录添加时间步索引,并按时间顺序对数据进行排序,确保生成模型能够学习时序依赖关系。通过时间步索引构建时序特征向量:$X_{\mathrm{temporal}}=[X,T_{\mathrm{step}}]$,并嵌入特征图。其中,$X$表示流量数据,$T_{\mathrm{step}}$为时间步索引,表示每条记录在时间序列中的位置

(2) 使用VAE提取特征图:$F_{\mathrm{VAE}}=\mathrm{VAE}(F)$

(3) 生成随机噪声:$z\sim N(0,1)$

(4) 将噪声添加到特征图:$F'=F_{\mathrm{VAE}}+\alpha\cdot z$,其中,$\alpha$为噪声缩放因子

(5) 从文本中计算目标噪声:$T_{\mathrm{noise}}=\mathrm{ExtractNoise}(T)$基于文本描述提取噪声信息

(6) 计算均方误差损失:$L_{\mathrm{MSE}}=\frac{1}{N}\sum_{i=1}^{N}(F_i'-T_{\mathrm{noise}})^2$

(7) 返回损失: return L_{MSE}

end

3 实验分析

3.1 实验环境配置与模型参数设置

实验的设备环境配置为32 GB内存、Intel(R) Core(TM) i9-10900F CPU@2.80 GHz中央处理器、Windows10 22H2专业版操作系统、NVIDIA GeForce RTX 4090 24 GB显卡。编程环境为Python = 3.9,PyTorch = 1.12.1,CUDA = 12.3,Transformers = 4.26.1,diffusers = 0.12.1。

潜在空间生成模型分为编码器、VAE和UNet 3部分,在编码器部分,权重被冻结,输入形状为[1, 77]的文字编码,输出为[1, 77, 768]的特征表示。编码器采用预训练模型,不参与反向传播,仅用于特征提取。在VAE模块中,输入大小为[1, 3, 64, 64],通过Encoder提取特征图[1, 4, 16, 16]。VAE中的均值层和方差层分别提取均值与方差,VAE生成的特征图会进行噪声调整,缩放系数设置为0.182 15。训练批次大小设置为1,训练轮数设置为50,学习率设置为2e−4,优化器设置为Adam,损失函数设置为均方误差(MSE)与KL散度组合,在UNet模块中,输入为[1, 4, 16, 16]的特征图,UNet网络的权重参与训练,输出与输入形状一致[1, 4, 16, 16]。通过1 000个时间步的随机选择,添加标准正态噪声以模拟扩散过程,并通过UNet去噪。时间步数设置为1 000,训练批次大小设置为4,训练轮数设置为200,学习率设置为2e−5,优化器设置为AdamW,学习率减少因子设为线性降低至0。

3.2 数据集介绍

实验使用由MAWI工作组维护的MAWI网络流量数据集,该数据集是WIDE项目的一部分,旨在

捕获和记录互联网骨干网络的流量数据。表1汇集展示了所使用的网络流量数据集的流量类型。MAWI数据集为网络流量特征分析、异常检测及网络性能评估等研究提供了丰富的真实流量数据支持。

表 1 网络流量数据的流量类型

Table 1 Traffic types of network traffic data

流量类型	描述	示例
HTTP流量	基于HTTP协议的网页访问数据	GET/POST请求
DNS流量	域名解析系统流量，用于解析域名到IP地址	DNS请求与响应
TCP流量	基于TCP协议的通信数据，包含连接、传输等状态	3次握手、数据传输
UDP流量	无连接的数据传输协议流量，通常用于实时应用	流媒体传输
ICMP流量	用于网络设备间的诊断和通信，例如Ping操作	Ping请求与响应

流量数据主要来自以下采样点。

Samplepoint-F：Samplepoint-F 自 2006 年 7 月 1 日开始运行，负责记录 WIDE 项目与上游 ISP 的传输链路流量。该采样点的流量数据包括。

日常流量记录：自 2006 年至 2024 年的每日流量记录。

(1)长时间流量记录。例如 2007 年 1 月的 48 h 流量、2008 年 3 月的 72 h 流量、以及 2024 年 6 月的 24 h 流量等；

(2)链路升级。链路于 2007 年从 100 Mb/s 升级至 1 Gb/s；

(3)在 2016 年去除了 150 Mb/s 的承诺接入速率(CAR)；

(4)2023 年 7 月，WIDE 项目在 BBIX 开始运行，扩展了采样点 F 的流量覆盖范围。

该采样点的流量数据有效反映了长期网络传输中的变化，特别是在链路升级、网络结构调整等场景下，为网络性能和流量行为研究提供了重要支持。此外，自 2013 年以来，采样点还包含 USCANT 项目对全球 IPv4 地址空间的探测性 ICMP 流量。

3.3 评价方法

为了全面评估生成模型在网络流量生成任务中的效果，设计了一套综合的检测方法，所生成结果的评价流程如图 5 所示，分为 3 个主要步骤。首先，使用 KL 散度量化生成流量与真实流量的概率分布差异。KL 散度通过衡量两者概率分布的不一致性，提供了对生成流量真实性的初步评估。该方法揭示了生成流量与真实流量在概率空间中的显著差异，随后采用统计特征比较与 FID(Frechet inception distance)进一步评估生成效果。统计特征比较选取了如平均包大小、方差和峰值等关键流量特征，旨在评估生成流量与真实流量在基础统计量上的相似度。该过程能够快速验证生成流量是否具备与真实流量一致的基本特征，为生成模型提供了重要的初步验证手段。FID 方法能够提取流量的高阶特征，分析生成流量与真实流量在特征空间中的相似性。在该流程中，UNet 的训练分为预训练阶段与正式训练阶段：预训练侧重于利用大规模、多样化的网络流量数据学习通用特征提取能力，为后续任务提供稳定

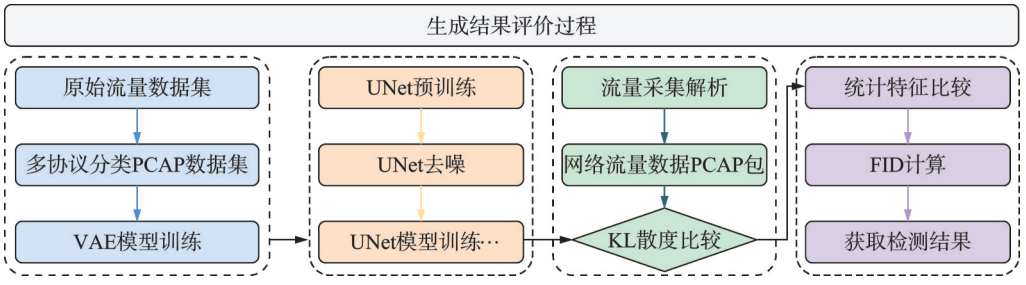

图 5 生成结果评价流程

Fig.5 Evaluation process for generated results

的初始参数;正式训练则基于目标任务数据(如本文筛选并处理后的MAWI数据集)进行定制化去噪建模,以优化特定场景下的生成质量和泛化性能。两阶段在训练目标、数据来源及优化策略上存在差异,从而确保模型既具备通用特征学习能力,又能精准适配目标任务。

评估方法中,设:x_{real}代表真实流量的统计特征;x_{gen}代表生成流量的统计特征;w_i为每个统计特征的权重(可根据实际情况设置,默认等权)。

通过加权欧氏距离计算统计特征之间的差异,总的差异度量D表示为

$$D=\sqrt{\sum_{i=1}^{n} w_i\left(\frac{x_{\text{real},i}-x_{\text{gen},i}}{x_{\text{real},i}}\right)^2} \tag{8}$$

使用FID度量生成流量与真实流量的高阶特征差异,其计算公式为

$$\text{FID}(\boldsymbol{x}_{\text{r}},\boldsymbol{x}_{\text{g}})=\|\boldsymbol{\mu}_{\text{r}}-\boldsymbol{\mu}_{\text{g}}\|_2^2+\text{tr}(\boldsymbol{\Sigma}_{\text{r}}+\boldsymbol{\Sigma}_{\text{g}}-2\cdot(\boldsymbol{\Sigma}_{\text{r}}\boldsymbol{\Sigma}_{\text{g}})^{1/2}) \tag{9}$$

式中:$\boldsymbol{\mu}_{\text{r}}$和$\boldsymbol{\mu}_{\text{g}}$分别代表从真实流量和生成流量中提取的特征均值向量;$\boldsymbol{\Sigma}_{\text{r}}$和$\boldsymbol{\Sigma}_{\text{g}}$分别为协方差矩阵;$\|\boldsymbol{\mu}_{\text{r}}-\boldsymbol{\mu}_{\text{g}}\|_2^2$为均值向量之间的欧氏距离的平方;tr表示矩阵的迹,即矩阵对角元素之和;$(\boldsymbol{\Sigma}_{\text{r}}\boldsymbol{\Sigma}_{\text{g}})^{1/2}$为协方差矩阵乘积的平方根矩阵。

综合KL散度、统计特征距离和FID 3种指标,从分布相似性、基础统计一致性和高阶语义匹配3个层面对生成流量进行评估,确保其在实际应用中的可用性与可靠性,并为模型优化提供有效反馈。

3.4 数据预处理

在MAWI网络流量数据集的基础上,进行了数据预处理和特征提取,构建了适合生成模型输入的64×64张量。表2展示了所使用的数据预处理和特征提取的步骤,具体步骤如下:

表2 数据预处理与特征提取步骤

Table 2 Data preprocessing and feature extraction steps

步骤	描述	输出形状
数据筛选	筛选关键特征(时间戳、IP、协议、包大小等)	流量记录列表
数据分块	按时间顺序划分64条记录为一组	64×64张量
特征提取	提取包大小、协议类型、IP地址等特征	64维特征向量
特征处理	归一化与缺失值处理	完整的64维特征
输入构建	构建64×64张量作为模型输入	64×64张量

(1)关键特征筛选。对MAWI数据集中原始网络流量数据进行了筛选,保留了时间戳、源/目的IP、协议类型和数据包大小等关键特征。随后,按时间顺序对流量数据进行排序,以确保生成模型能够学习到流量的时序特性。

(2)数据分块与处理。为了适应生成模型的输入格式,将流量数据划分为大小为64×64的张量,其中64行表示连续的流量记录,64列代表提取的流量特征。对于无法直接提取到完整64维特征的流量记录,采用插值和统计分布补齐等方法进行处理,保证数据的完整性。

(3)特征归一化与补全。在特征提取过程中,对包大小、协议类型、源/目的IP地址及端口号等信息进行归一化处理,以消除不同特征之间的量纲差异。对于缺失或不完整的特征,通过基于相似流量统计特征的补全方法进行修复,以确保每条流量记录形成完整的64维特征向量。

(4)时序信息构建融入。为每条流量记录添加时间戳并排序,同时引入时间步索引,帮助模型学习时序依赖和变化趋势。此外,针对周期性波动(如日周期、周周期),通过傅里叶变换或正弦波函数建模周期性特征,叠入特征向量,增强对流量周期变化的捕捉能力。

通过上述预处理步骤,将数据投入训练,学习MAWI数据集中网络流量的分布模式,进而生成符合实际流量特征的样本。

3.5 实验结果与分析

对生成模型在网络流量生成任务中的表现进行评估，分析了模型生成不同类型流量时的适应性和准确性。实验结果揭示了不同流量类别下，生成模型在保持原始数据特性上的效果。

为了保证实验的代表性，实验设计中包含了多种流量类型。表3显示了实验中使用的各类网络流量样本数量。这些数据涵盖了常见的网络协议，如IPv4、ICMP、TCP等。不同流量类型在实验中的数量分布较为均匀，以确保生成模型能够充分学习和捕捉不同流量的特性。

表3 网络流量样本数量分布

Table 3 Distribution of network traffic sample quantities

流量类型	ARP	IPv4	IPv4-TCP	IPv4-UDP	IPv4-ICMP	IPv4-ICMP-DNS	IPv4-CMP-DNS	IPv4-UDP-Raw
样本数量	30	250	621	104	308	21	22	53

模型评价过程中涉及到以下一系列评价指标。

(1) MSE：用于衡量生成数据与原始数据之间的均方误差，数值越小，表示生成数据与原始数据越接近。

(2) KL散度：用于衡量生成数据与真实数据分布之间的差异，数值越小，表示生成数据的分布越接近真实数据。

(3) FID：主要衡量生成数据与原始数据在特征空间的分布差异，数值越小则表示生成的数据质量越高。

不同模型生成的流量分布与原始流量分布的对比，如图6所示，图中数据1表示TDGM生成结果，数据2表示原始数据结果。从图中可以看出，生成模型在大多数组别的MSE和KL散度上都表现出较低的误差，表明生成数据与原始数据在数值和分布上的相似度较高。特别地，IPv4的KL散度和FID最低，说明在此组数据中，生成模型能够更好地捕捉到原始数据的特征分布。

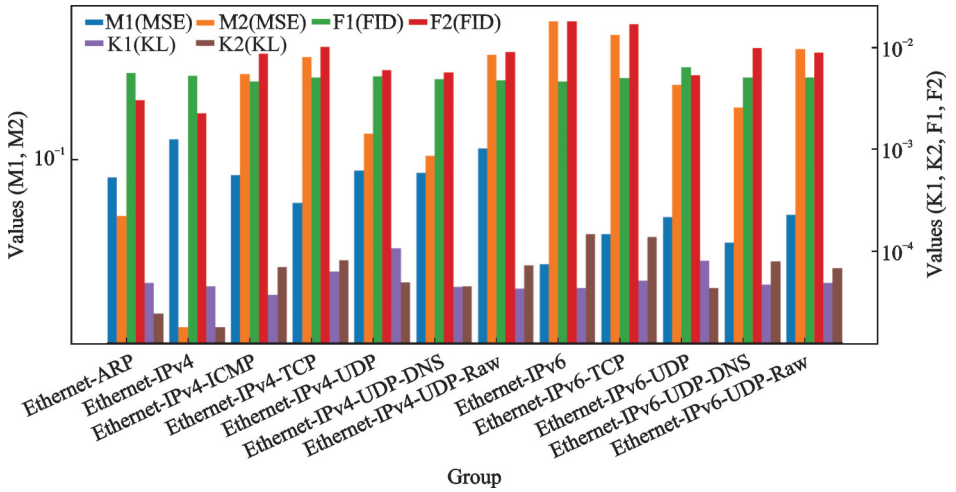

图6 MSE、FID和KL差异图

Fig.6 Divergence diagram of MSE, FID and KL

为进一步分析,图7展示了每个流量组的五星雷达图,这些图对比了每个组中的原始数据和生成数据在最大值(Max)、95th百分位(95th Percentile)、偏度(Skewness)、峰度(Kurtosis)和加权欧氏距离(Weighted Euclidean distance)5个维度上的差异。

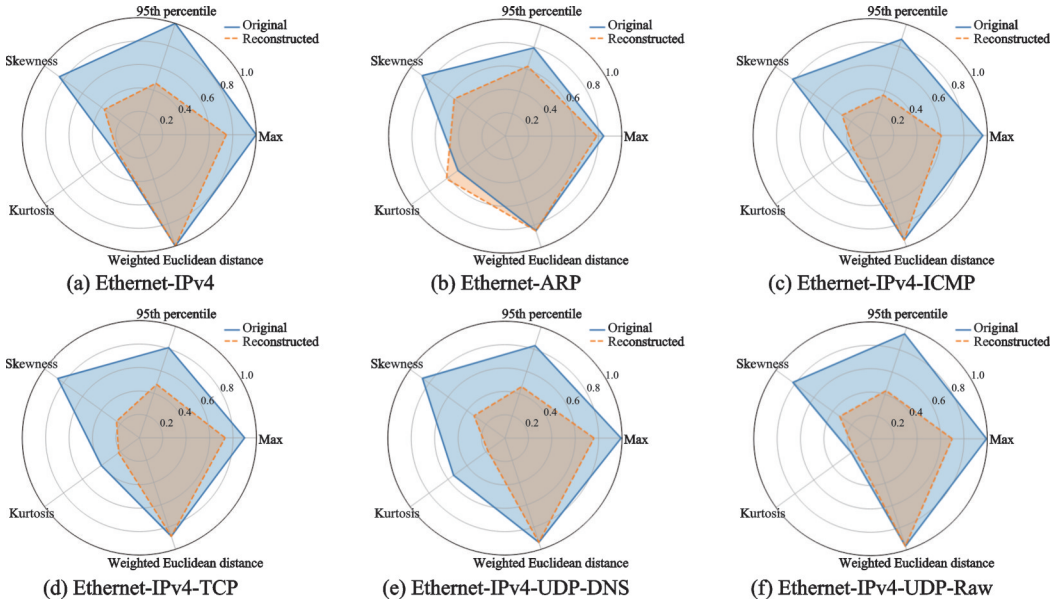

图7 不同Ethernet协议的雷达图

Fig.7 Radar chart of different Ethernet protocols

从图7可以看出,对于大多数流量类型,生成数据在各个维度上与原始数据较为一致,尤其是在最大值和95th百分位上,生成模型展现了较强的重现能力。然而,在偏度和峰度这两个衡量分布形状的指标上,生成数据与原始数据存在一定的差异。

从实验结果可以看出,生成模型在多数流量类型上的性能表现出色,尤其是对于大规模样本的流量类型(如IPv4、TCP流量),生成数据与原始数据在多个评估指标上相近,展现了生成模型强大的泛化能力。实验结果表明,生成模型在处理大样本流量(如IPv4和TCP流量)时表现出较好的泛化能力,但对于小样本流量(如ARP、DNS流量)的生成效果还有待提升。

为进一步验证基于扩散模型方法在网络流量生成任务中的优势,选取典型的生成对抗网络GAN作为对照模型,并在相同数据集与特征预处理条件下进行对比实验。评价指标依旧采用最大值、95th百分位、偏度、峰度及加权欧氏距离5个维度,并通过雷达图直观对比不同模型生成结果与真实流量的接近程度。图8为从左到右Ethernet-IPv4-TCP、Ethernet-IPv4-UDP-Raw、Ethernet-IPv6-TCP、Ethernet-IPv6-UDP-Raw 4类流量下的实验结果。

从4类典型流量(Ethernet-IPv4-TCP、Ethernet-IPv4-UDP-Raw、Ethernet-IPv6-TCP、Ethernet-IPv6-UDP-Raw)的对比结果可以看出,GAN在极值拟合(最大值、95th百分位)上部分场景占优,尤其在流量峰值较为明显的数据集中表现较好;然而,其在高阶统计特征(偏度、峰度)和整体分布一致性(加权欧氏距离)方面存在明显不足,易出现模式崩溃导致特征缺失。相比之下,扩散模型在多数场景下更接近真实流量的分布形态,尤其在偏度、峰度等衡量分布形状的指标上表现稳定,能够较好地保留周期性波动、长尾分布与异常尖峰特征;但在个别场景(如IPv4-UDP-Raw的加权欧氏距离、IPv6-TCP的极值指标)上仍有优化空间。整体而言,扩散模型在生成质量与分布拟合度方面优势更为显著,适用于对分布细节要求较高的数字孪生与网络仿真任务。

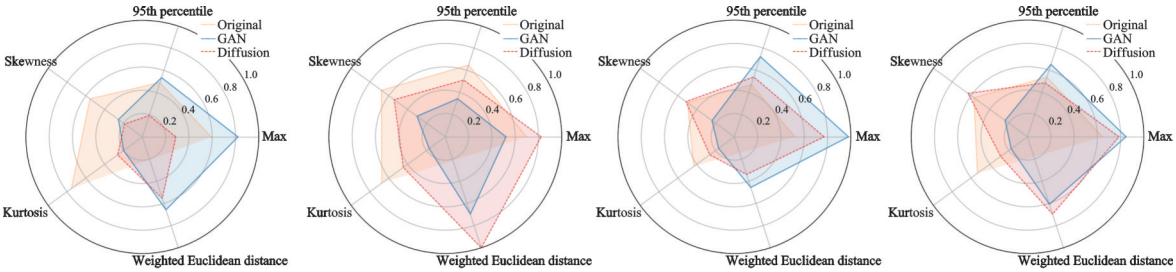

图8 Diffusion与GAN模型的对比分析

Fig.8 Comparative analysis of Diffusion and GAN models

综合分析可知，GAN在极值（最大值、95th百分位）拟合上有一定优势，尤其是在流量峰值较为突出的场景中表现较好，但在高阶统计特征（偏度、峰度）和整体分布一致性（加权欧氏距离）方面稳定性较差，容易出现模式崩溃（Mode collapse）导致的特征丢失。而扩散模型凭借逐步去噪的生成机制，能够更好地捕捉流量的分布形态与高阶统计特征，在多数场景下表现出更高的生成质量与分布拟合度。这一特性对于需要保留周期性波动、长尾分布以及异常尖峰的数字孪生与网络仿真应用尤为关键。

4 结束语

提出了基于扩散模型的网络流量生成方法FlowDiff，使用时序扩散生成模型TDGM，利用随机加噪与逐步去噪机制，实现了符合真实特征的大批量流量数据生成。实验结果表明，FlowDiff在建模时序依赖与周期性波动方面表现优异，能够准确还原IPv4和TCP等流量的真实分布。通过MSE、KL散度与FID等指标验证，该方法有效缩小了生成数据与真实数据之间的差异，具有良好的通用性与生成质量，适用于多种网络环境。进一步分析可见，在最大值和95th百分位上，生成模型展现了较强的重现能力，说明模型成功再现了细粒度的突发特性；而在包含24～72 h连续片段的评估中，综合KL散度与FID等指标仍保持低位，表明模型对昼夜周期等长期波动也具备良好刻画能力。由于合成流量同时保留了尖峰与周期两种关键形态，其可以在不泄露隐私的前提下安全注入数字孪生或离线仿真平台，用于链路扩容、拥塞控制、QoS调度与灰度发布等网络控制任务，并为安全压测与异常检测提供可靠、可重复的测试数据支撑。综上所述，FlowDiff与TDGM不仅在定量评估中表现突出，也可以为实际网络运维决策提供有力的数据基础。

参考文献：

[1] 沈高青，蔡圣所，雷磊，等. 基于数字孪生和强化学习的低空智联网协同认知干扰[J]. 数据采集与处理，2024，39(1)：15-30.
SHEN Gaoqing，CAI Shengsuo，LEI Lei，et al. Cooperative cognitive jamming in low-altitude intelligent network based on digital twin and reinforcement learning[J]. Journal of Data Acquisition and Processing，2024，39(1)：15-30.

[2] DONG X，YU J，LUO Y，et al. Achieving an effective，scalable and privacy-preserving data sharing service in cloud computing[J]. Computers & Security，2014，42：151-164.

[3] 陶飞，刘蔚然，刘检华，等. 数字孪生及其应用探索[J]. 计算机集成制造系统，2018，24(1)：1-18.
TAO Fei，LIU Weiran，LIU Jianhua，et al. Digital twin and its potential application exploration[J]. Computer Integrated Manufacturing Systems，2018，24(1)：1-18.

[4] 陶飞，马昕，戚庆林，等. 数字孪生连接交互理论与关键技术[J]. 计算机集成制造系统，2023，29(1)：1-10.
TAO Fei，MA Xing，QI Qinglin，et al. Theory and key technologies of digital twin connection and interaction[J]. Computer In-

tegrated Manufacturing Systems, 2023, 29(1): 1-10.

[5] DAINOTTI A, PESCAP A, ROSSI P S, et al. Internet traffic modeling by means of hidden Markov models[J]. Computer Networks, 2008, 52(14): 2645-2662.

[6] ERGEN D, ONUR E. On network traffic forecasting using autoregressive models[EB/OL]. (2019-12-27). https://doi.org/10.48550/arXiv.1912.12220.

[7] 郭森森,王同力,慕德俊. 基于生成对抗网络与自编码器的网络流量异常检测模型[J]. 信息网络安全, 2022, 22(12): 7-15.
GUO Sensen, WANG Tongli, MU Dejun. Anomaly detection model based on generative adversarial network and autoencoder [J]. Netinfo Security, 2022, 22(12): 7-15.

[8] KHATIMAN M N A, ABU-SAMAH A, AZMAN M A, et al. Generation of synthetic 5G network dataset using generative adversarial network (GAN)[C]//Proceedings of 2023 IEEE 16th Malaysia International Conference on Communication. [S.l.]: IEEE, 2023: 141-145.

[9] XU L, SKOULARIDOU M, CUESTA-INFANTE A, et al. Modeling tabular data using conditional GAN[C]//Proceedings of Conference on Neural Information Processing Systems. [S.l.]: [s.n.]: 2020.

[10] ISHFAQ H, HOOGI A, RUBIN D. TVAE: Triplet-based variational autoencoder using metric learning[J]. arXiv preprint arXiv:1802.04403, 2018.

[11] 高欣宇,杜方,宋丽娟. 基于扩散模型的文本图像生成对比研究综述[J]. 计算机工程与应用, 2024, 60(24): 44-64.
GAO Xinyu, DU Fang, SONG Lijuan. A comparative review of text-to-image generation based on diffusion models[J]. Computer Engineering and Applications, 2024, 60(24): 44-64.

[12] 王子昂,汤艳君,王子晨,等. 基于去噪扩散概率模型的网络流量入侵检测方法研究[J]. 信息安全研究, 2024, 10(5): 421-430.
WANG Ziang, TANG Yanjun, WANG Zichen, et al. Research on a network traffic intrusion detection method based on denoising diffusion probabilistic models[J]. Journal of Information Security Research, 2024, 10(5): 421-430.

[13] 岳忠牧,张喆,吕武,等. De-DDPM: 可控、可迁移的缺陷图像生成方法[J]. 自动化学报, 2024, 50(8): 1539-1549.
YUE Zhongmu, ZHANG Zhe, LÜ Wu, et al. De-DDPM: A controllable and transferable defect image generation method[J]. Acta Automatica Sinica, 2024, 50(8): 1539-1549.

[14] 陈家林,周永霞. 基于扩散模型的手部网格重建算法[J]. 计算机应用研究, 2025, 42(5): 1564-1569.
CHEN Jialin, ZHOU Yongxia. Hand mesh reconstruction algorithm based on diffusion models[J]. Application Research of Computers, 2025, 42(5): 1564-1569.

[15] 刘雨生,肖学中. 基于扩散模型微调的高保真图像编辑[J]. 计算机应用, 2024, 44(11): 3574-3580.
LIU Yusheng, XIAO Xuezhong. High-fidelity image editing based on diffusion model fine-tuning[J]. Journal of Computer Applications, 2024, 44(11): 3574-3580.

[16] 刘万军,程裕茜,曲海成. 基于生成对抗网络的图像自增强去雾算法[J]. 系统仿真学报, 2024, 36(5): 1093-1106.
LIU Wanjun, CHENG Yuqian, QU Haicheng. Image self-enhancement dehazing algorithm based on generative adversarial networks[J]. Journal of System Simulation, 2024, 36(5): 1093-1106.

[17] 王猛,杨观赐. 融合时序关系和上下文信息的时间动作检测方法[J]. 贵州大学学报(自然科学版), 2024, 41(6): 78-84, 90.
WANG Meng, YANG Guanci. Temporal action detection method integrating temporal relationships and contextual information [J]. Journal of Guizhou University (Natural Sciences), 2024, 41(6): 78-84, 90.

[18] 李睿,贾悠,焦哲,等. 基于时间序列的网络行为异常检测[J]. 通信技术, 2020, 53(10): 2550-2554.
LI Rui, JIA You, JIAO Zhe, et al. Network behavior anomaly detection based on time series[J]. Communications Technology, 2020, 53(10): 2550-2554.

[19] 王勇,周慧怡,俸皓,等. 基于深度卷积神经网络的网络流量分类方法[J].通信学报, 2018, 39(1): 14-23.
WANG Yong, ZHOU Huiyi, FENG Hao, et al. Network traffic classification method based on deep convolutional neural networks[J]. Journal on Communications, 2018, 39(1): 14-23.

[20] 王永生,陈振,刘利民,等. 基于GAT-Transformer时间序列模型的SOWQP[J]. 计算机仿真, 2024, 41(3): 321-326, 358.

WANG Yongsheng, CHEN Zhen, LIU Limin, et al. SOWQP based on the GAT-Transformer time-series model[J]. Computer Simulation, 2024, 41(3): 321-326, 358.

[21] 刘云祥,王一宾.基于潜在表示的自适应权重多视图子空间聚类算法[J].电脑知识与技术, 2023, 19(17): 10-15.
LIU Yunxiang, WANG Yibin. Adaptive weighted multi-view subspace clustering algorithm based on latent representations[J]. Computer Knowledge and Technology, 2023, 19(17): 10-15.

[22] 张宇,黎靖,马铭,等. YOLOLW: 一个新的轻量级目标检测模型[J]. 计算机与现代化, 2024(11): 91-98.
ZHANG Yu, LI Jing, MA Ming, et al. YOLOLW: A novel lightweight object detection model[J]. Computer and Modernization, 2024, (11): 91-98.

[23] 贾军营,杨芯茹,杨海波,等.改进CLIP-ReID的跨模态行人重识别[J].计算机系统应用, 2025, 34(1): 153-160.
JIA Junying, YANG Xinru, YANG Haibo, et al. Cross-modal person re-identification based on improved CLIP-ReID[J]. Computer Systems & Applications, 2025, 34(1): 153-160.

作者简介:

赵晓红(1990-),女,工程师,研究方向:计算机网络、人工智能,E-mail:xiaohongzhao1990@163.com。

陈浩(1998-),男,工程师,研究方向:强化学习、信息通信工程。

施冠炬(2000-),男,硕士研究生,研究方向:深度学习、计算机网络。

管黉(1979-),女,高级工程师,研究方向:通信网络、电力系统自动化。

丁旭辉(1991-),男,特别副研究员,博士生导师,研究方向:卫星通信、物理层迭代接收技术。

朱超(1989-),**通信作者**,男,副教授,博士生导师,研究方向:边缘计算、云计算,E-mail:chao@bit.edu.cn。

(编辑:夏道家)

Improved Network Traffic Generation Technology for Digital Twins Based on Diffusion Model

ZHAO Xiaohong[1], CHEN Hao[1], SHI Guanju[2], GUAN Ti[3], DING Xuhui[2], ZHU Chao[2*]

(1. Science and Technology Information Center, Electric Power Research Institute of State Grid Shandong Electric Power Company, Jinan 250003, China; 2. Schoole of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China; 3. Power Dispatch Control Center, State Grid Shandong Electric Power Company, Jinan 250001, China)

Abstract: Generating realistic network traffic for digital twin systems is difficult because traffic packets exhibit temporal dependence, periodic variation, and scenario-driven uncertainty. To address this problem, this paper proposes FlowDiff, a diffusion-based packet generation algorithm, and TDGM, a temporal diffusion generation model. FlowDiff formulates packet generation as a conditioned reverse-diffusion process that progressively denoises latent traffic features. TDGM integrates temporal-aware embedding, periodic features, convolutional neural network (CNN) and Transformer modules, and cross-attention to capture local, global, and long-range traffic patterns. Experiments on real network traffic datasets show that the generated samples are close to real traffic in MSE, KL divergence, FID, and statistical features, demonstrating the effectiveness of the method for digital twin network simulation.

Highlights:

1. A large-scale network traffic packet generation algorithm, FlowDiff, is proposed for digital twin scenarios. It formulates packet generation as a reverse-diffusion process conditioned on temporal characteristics and external network scenarios, thereby generating traffic samples closer to real network behavior.

2. A temporal diffusion generation model (TDGM) is developed by introducing temporal-aware feature embedding and Fourier-based periodic feature modeling, enabling the model to capture short-term dependence, long-term trends, and periodic fluctuations in network traffic.

3. A hybrid generative architecture is designed by integrating latent-space compression, local feature extraction, and global context modeling. The combination of VAE, UNet/CNN, Transformer, and cross-attention improves feature representation and conditional generation capability for complex traffic data.

4. Experiments on the real MAWI network traffic dataset evaluate the proposed method using MSE, KL divergence, FID, and statistical feature distance, together with comparison.

Key words: network traffic generation; diffusion model (DM); digital twin; deep learning; temporal dependence

Foundation item: Science and Technology Project of State Grid Shandong Electric Power Company (No.52062624000K).
Received: 2025-07-17; **Revised:** 2025-09-14
***Corresponding author, E-mail**: chao@bit.edu.cn.