DOI: 10.16452/j.cnki.sdkjzk.2026.01.010 文章编号:1672-3767(2026)01-0112-10

基于变分自动编码器与扩散模型的时序推荐模型

赵铁柱[1],黄咏健[1],杨秋鸿[2],陈奕霖[1]
(1. 东莞理工学院 计算机科学与技术学院,广东 东莞 523808;
2. 东莞城市学院 人工智能学院,广东 东莞 523419)

摘 要:在推荐系统中,变分自动编码器和扩散模型能够有效针对"用户-物品"的交互生成过程进行建模,但受限于变分自动编码器表征能力和扩散模型易受噪声的影响,模型性能得不到进一步提升。针对此问题,本研究提出一种结合变分自动编码器和扩散模型并融入时序信息的推荐模型(TRMVADM)。TRMVADM 首先根据用户-物品交互次序对交互数据进行赋权,然后利用编码器对时序数据进行维度压缩,并在潜在空间中进行前向与反向扩散,最后通过解码器推断出用户物品交互的概率。TRMVADM 有效利用了编码器维度压缩的特点和扩散模型挖掘潜在用户-物品交互信息的能力,提高了模型的推荐性能。在 MovieLens 和 Yelp 数据集上的实验结果表明,TRMVADM 在评价指标召回率和归一化累计折损增益上,相比于次优基线模型分别平均提升 5.4% 和 2.04%。

关键词:变分自动编码器;扩散模型;时序信息;用户-物品交互;推荐系统
中图分类号:TP181 **文献标志码**:A

Temporal recommendation model based on variational autoencoder and diffusion model

ZHAO Tiezhu[1], HUANG Yongjian[1], YANG Qiuhong[2], CHEN Yilin[1]
(1. School of Computer Science and Technology, Dongguan University of Technology, Dongguan 523808, China; 2. School of Artificial Intelligence, Dongguan City University, Dongguan 523419, China)

Abstract: In recommendation systems, variational autoencoders and diffusion models can effectively model the generation process of "user-item" interactions. However, due to the limited representation capability of variational autoencoders and the susceptibility of diffusion models to noise, the performance of the models cannot be further improved. To address this issue, this paper proposed a temporal recommendation model based on variational autoencoder and diffusion model (TRMVADM). TRMVADM first assigned weights to interaction data based on the order of user-item interactions. It then used an encoder to compress temporal data dimensions and performs forward and backward diffusion in the latent space. Finally, it inferred the probability of user-item interactions through an decoder. TRMVADM effectively utilizes the dimension compression feature of the encoder and the ability of the diffusion model to explore latent user-item interaction information, thereby improving the recommendation performance of the model. Experimental results on the MovieLens and Yelp datasets demonstrate that TRMVADM can achieve an average improvement of 5.4% in recall rate and 2.04% in normalized discounted cumulative gain compared to the suboptimal baseline model.

Key words: variational autoencoder; diffusion model; temporal information; user-item interaction; recommendation system

收稿日期:2024-05-24
基金项目:广东省普通高校重点领域专项项目(2021ZDZX3007);东莞市社会发展科技项目(20231800936732);东莞城市青年教师发展基金项目(2022QJY005Z);杨振宁创新班导师制科研项目(2022463030408)
作者简介:赵铁柱(1983—),男,湖南双峰人,副教授,博士,主要研究方向为数据挖掘、大数据、分布式计算/存储.
黄咏健(1998—),男,广东东莞人,硕士研究生,主要从事推荐系统、大数据研究,本文通信作者.
E-mail:huangyongjian@dgut.edu.cn

随着电子商务的蓬勃发展,海量的商品信息与交互数据造成了信息过载。推荐系统是解决信息过载的有效手段,根据用户的兴趣与行为习惯为用户推荐可能感兴趣的商品[1]。然而,传统的推荐算法无法挖掘到"用户-物品"数据中复杂的非线性交互信息,导致模型泛化性能和精度较差。随着生成式模型的发展,许多研究人员开始把生成式模型运用到推荐系统中,提出了生成式推荐模型[2-4]。此类模型普遍假设用户-物品交互行为是由对应的潜在因子决定的,例如用户兴趣等,这和真实世界的用户-物品交互的产生原理基本吻合,故能有效针对交互生成过程进行建模。

现有的生成推荐模型主要分为三类:基于生成对抗网络(generative adversarial networks,GAN)[2]的推荐模型、基于自动编码器(autoencoders,AE)[3]的推荐模型和基于扩散模型(diffusion model,DM)[4]的推荐模型,这些模型能够克服传统推荐模型的固有缺陷,但仍然存在一些问题阻碍了其推荐性能的进一步提升。面对交互数据稀疏问题和噪声问题,Wen 等[5]提出带有条件 GAN 的个性化推荐条件生成对抗网络 (personalized recommendation with conditional generative adversarial networks,PRGAN),该网络将条件评分向量生成形式化为用户物品匹配问题,利用对抗训练的模式优化网络参数,缓解了条件评分向量的稀疏问题,提高了判别器的学习效率。然而,对抗训练过程通常是不稳定的,并存在梯度消失的问题,导致了模型性能不理想。基于 AE 的推荐模型利用编码器近似潜在因子的分布,训练过程稳定,生成的数据较符合用户兴趣。Zhang 等[6]将注意力机制运用在 AE 中,提出 A-SAERec(attentive stacked sparse autoencoder)模型,利用矩阵分解神经网络,增强训练过程的稳定性,提高了推荐性能。然而,复杂模型的后验分布通常难以处理,这使基于 AE 的推荐模型表征能力受限,易处理性较弱。针对 GAN 训练不稳定和 AE 表征能力弱问题,Wang 等[7]将 DM 运用在推荐系统中,提出扩散推荐模型 (diffusion recommender model,DiffRec),该模型以去噪的方式学习生成过程,通过构造扩散过程的多个时间步,简化了复杂的后验分布,并利用去噪过程提升了训练的稳定性。然而,基于 DM 的推荐模型利用去噪神经网络模型的迭代过程,从噪声数据中恢复成原来的数据,这意味着基于 DM 的推荐模型易受到噪声的干扰,在实际应用时推荐结果不稳定。

针对 GAN、AE 和 DM 的不足,本研究提出一种基于变分自动编码器(variational autoencoders,VAE)[8]与扩散模型的时序推荐模型(temporal recommendation model based on variational autoencoder and diffusion model,TRMVADM)。TRMVADM 利用 VAE 在用户交互向量和潜在向量中进行转换,缓解 DM 中噪声对模型的影响,并利用 VAE 对时序数据的维度进行压缩,充分提取用户-物品交互的时序特征,提升扩散过程的效率。TRMVADM 中的 DM 通过构造正向和反向扩散过程的多个时间步,形成一个易于处理的后验分布,有效解决了 AE 中对复杂后验分布建模困难的问题,并且 DM 通过去噪神经网络模型的迭代过程有效提高了模型训练的稳定性。

1 相关工作

AE 是一种可以进行无监督学习的神经网络模型,其拥有编码器和解码器成对结构。VAE 是一种改进的 AE 模型,引入 Kullback-Leibler(KL)离散度强制潜在表示服从预先定义的高斯分布,其主要优点是可以从潜在空间中进行采样生成新的数据。Liang 等[8]将 VAE 引入到协同过滤任务中,认为这种非线性概率模型可以超越线性因素模型有限的建模能力,能有效捕获到用户-物品数据之间的协同特征。VAE 能有效提取用户物品之间的协同信号,但其性能受到训练样本数量的严重限制。为解决此问题,Wang 等[9]提出自监督 VAE 模型,该模型首先通过数据增强为每个用户构建多个视图,然后设计一个预设任务对齐从每个用户的不同视图中学到的表示,从而提高在稀疏交互数据集上的泛化能力。基于 AE 的推荐模型在处理稀疏数据方面取得较好的效果,但不能在细粒度层面区分正样本和负样本。针对此问题,Liu 等[10]提出一种简单的线性结构 VAE 框架,该框架结合正样本和关键负样本的重构损失函数,有效挖掘出用户潜在的真实偏好。这些基于 AE 的推荐模型可以深入提取稀疏用户数据中的非线性特征,但在易处理性和表征能力方面还有所欠缺,简单的 AE 结构不能有效捕获异构的用户偏好。

相比于 AE,基于 DM 的推荐模型通过扩散过程的多个时间步,能更深入挖掘交互数据的潜在信息,其主要包括前向扩散过程和反向扩散过程,前向扩散过程首先逐步向用户数据注入噪声,然后利用反向扩

散的去噪过程推断出用户-物品的交互概率。Walker 等[11]首次将 DM 运用在协同过滤当中,提出协同扩散生成式模型 (collaborative diffusion generative model,CODIGEM),通过获取复杂的非线性模式,对用户-物品交互数据进行有效建模,生成强协同信号和鲁棒的潜在表示,以提高模型的泛化能力和推荐性能。然而,CODIGEM 在每一个时间步利用不同的 AE 进行预测,在推理过程中只使用第一个 AE 估计交互概率,导致该模型性能受限且不适用于交互序列建模。针对 CODIGEM 的问题,DiffRec[7]使用共享的多层感知机(multilayer perceptron,MLP)进行多步预测,并利用去噪过程进行推断,有效提高了模型性能。Li 等[12]在序列推荐任务中把 DM 用于物品表示的构建和不确定性的注入,提出面向序列的推荐模型,将物品的潜在方面和用户的不同意图建模为分布,很好地契合了序列推荐的范式。这些基于 DM 的推荐模型通过增噪和去噪的过程有效提升了模型的泛化性能,但易受到噪声规模的影响,从而导致推荐结果不稳定。当前,生成推荐模型仍有许多不足,主要体现在表征能力受限和易受噪声影响[13]等问题。针对上述问题,本研究提出结合 VAE、DM 和时序信息的 TRMVADM 模型。

2 TRMVADM 模型

2.1 模型概述

TRMVADM 模型主要包括 VAE 和 DM,其结构如图 1 所示。在真实空间中,TRMVADM 的 VAE 结构主要由参数为 ϕ 的编码器 f_ϕ 和参数为 ψ 的解码器 g_ψ 构成。用户 u 交互向量 $\boldsymbol{x}_u$ 先经时序重加权得到 $\boldsymbol{x}_0$,$\boldsymbol{x}_0$ 再通过 f_ϕ 编码变为扩散过程的初始向量 $\boldsymbol{z}_0$,$\boldsymbol{z}'_0$ 是与 $\boldsymbol{z}_0$ 对应的最终向量,$\boldsymbol{z}'_0$ 经过 g_ψ 的解码恢复为 $\boldsymbol{x}'_0$,最后根据 $\boldsymbol{x}'_0$ 的排序结果对用户进行推荐。扩散过程在潜在空间中进行,时间步数为 T,$\boldsymbol{z}_0$ 经前向扩散过程得到 $\boldsymbol{z}_T$,反向扩散过程的初始向量 $\boldsymbol{z}'_T$ 由 $\boldsymbol{z}_T$ 初始化,用参数为 θ 的 MLP 去噪(记为 MLP(θ)),$\boldsymbol{z}'_T$ 通过 MLP(θ)的 T 次去噪得到 $\boldsymbol{z}'_0$,去噪过程的中间向量为 $\{\boldsymbol{z}'_1,\boldsymbol{z}'_2,\cdots,\boldsymbol{z}'_{T-1}\}$。综上,TRMVADM 模型的工作流程主要分为时序重加权、编码、前向扩散过程、反向扩散过程和解码。

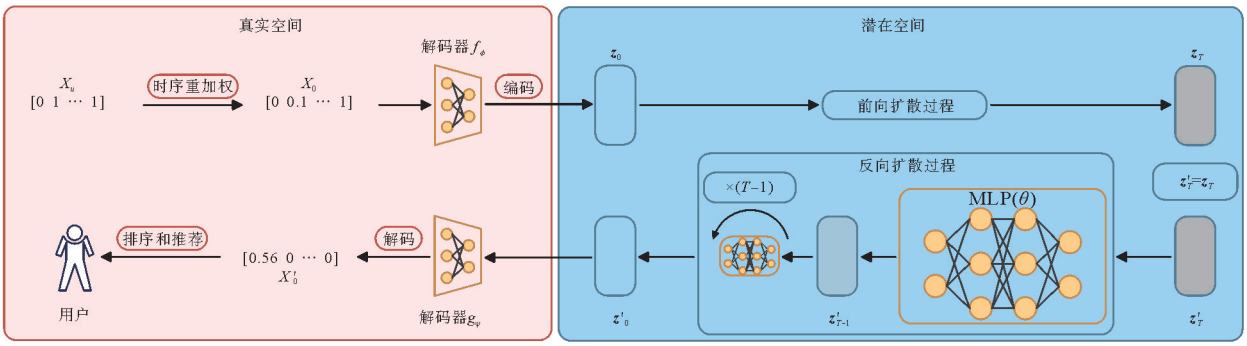


图 1 TRMVADM 结构

Fig. 1 Structure of TRMVADM

2.2 时序重加权

一般而言,用户的偏好会随着时间的推移而发生改变,对此本研究设计了一种基于时序的重加权策略,为用户后期交互的数据分配更大的权重。假设在一个商务系统中,物品集合为 I,其中物品 $i\in I$,用户 u 的物品交互向量为 $\boldsymbol{x}_u=[x_u^1,x_u^2,\cdots,x_u^{|I|}]$。交互向量按照用户-物品交互次序先后进行排列,当 $x_u^i=1$ 时表示用户 u 和物品 i 有过交互行为,$x_u^i=0$ 时则相反。假设用户 u 和 M 个物品交互过,其按交互次序的交互物品序列为 $S=\{i_1,i_2,\cdots,i_M\}$,其中 i_m 表示第 m 个交互过的物品的 ID。此处利用基于时间感知的线性方案:$w_m=w_{\min}\cdot\left(\frac{w_{\max}}{w_{\min}}\right)^{\frac{m-1}{M-1}}$,定义交互物品的权重向量 $\boldsymbol{w}=[w_1,w_2,\cdots,w_M]$,其中 $w_{\min}$ 和 $w_{\max}$ 是交互权重的下界和上界。由于模型最终输出的是物品的交互概率,所以 $w_{\min}$ 和 $w_{\max}$ 的范围是 $w_{\min}<w_{\max}\in(0,1]$。根据上述定义,$\boldsymbol{x}_u$ 通过重赋权可变为时序交互向量 $\boldsymbol{x}_0$,具体过程如式(1)、式(2)所示:

$$\boldsymbol{x}_0=\boldsymbol{x}_u\odot\bar{\boldsymbol{w}}, \tag{1}$$

$$\bar{w}[i]=\begin{cases} w[I_{dx}(i)], & i\in S; \\ 0, & i\notin S。 \end{cases} \tag{2}$$

式中:⊙表示点乘,$I_{dx}(i)$表示用户 u 交互物品序列 S 中物品 i 的索引,$w[I_{dx}(i)]$为向量 $\boldsymbol{w}$ 里索引 $I_{dx}(i)$对应的值,$\bar{w}[i]$为向量$\bar{\boldsymbol{w}}$ 里索引 i 对应的值。

2.3 编码与解码

TRMVADM 中 f_ϕ 的作用是将真实的数据空间映射到潜在变量空间,g_ψ 的作用则相反。经时序重加权后,向量 $\boldsymbol{x}_0$ 会输进 f_ϕ,f_ϕ 不直接输出一个潜在向量 $\boldsymbol{z}_0$,而是先输出一个期望为 μ 和方差为 σ^2 的多维高斯分布,并首先从该高斯分布中采样出 $\boldsymbol{z}_0$,然后通过扩散过程得到 $\boldsymbol{z}'_0$,最后利用 g_ψ 输出用户-物品交互概率向量 $\boldsymbol{x}'_0$。但在 f_ϕ 的输出中进行采样的过程是不可微的,所以该过程采用了重参数化方法[8]。重参数化过程和数据 $\boldsymbol{x}'_0$生成过程为:

$$\boldsymbol{z}_0=r(\boldsymbol{x}_0)=\mu_\phi(\boldsymbol{x}_0)+\sigma_\phi(\boldsymbol{x}_0)\odot\boldsymbol{\epsilon},\boldsymbol{\epsilon}\sim N(\boldsymbol{0},\boldsymbol{I}), \tag{3}$$

$$x'_0=g_\psi(z'_0)。 \tag{4}$$

式中:$\mu_\phi(\boldsymbol{x}_0)$和 $\sigma_\phi(\boldsymbol{x}_0)$分别表示 f_ϕ 输出分布的期望和方差,r 表示重参数化过程,N 表示高斯分布,$\boldsymbol{I}$ 表示单位矩阵。

2.4 前向扩散过程与反向扩散过程

经重参数化过程后,得到潜在向量 $\boldsymbol{z}_0$,假设 $\boldsymbol{z}_0\sim q(\boldsymbol{z}_0)$,TRMVADM 中的 DM 通过在 T 个时间步中添加高斯噪声构造出潜在变量$\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_T\}$,$\alpha_t$ 用于控制每一个时间步的噪声规模,其中 $t\in\{1,2,\cdots,T\}$,$\alpha_t\in(0,1)$。在前向扩散过程中,根据重参数化方法和两个独立高斯噪声的可加性[7],$\boldsymbol{z}_t$ 可直接由 $\boldsymbol{z}_0$ 得出:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0)=N\left(\boldsymbol{z}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{z}_0,(1-\bar{\alpha}_t)\boldsymbol{I}\right)。 \tag{5}$$

式中:$\bar{\alpha}_t=\prod_{t'=1}^{t}\alpha_{t'}$,利用重参数化过程可得$\boldsymbol{z}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{z}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$,其中 $\boldsymbol{\epsilon}\sim N(\boldsymbol{0},\boldsymbol{I})$。根据生成个性化的推荐的要求[11],总时间步 T 不宜设置过大。为了调节每一时间步的噪声规模,此处设计一种线性噪声方案,如式(6)所示:

$$\alpha_t=1-s\cdot\left[\alpha_{\min}+\frac{t-1}{T-1}(\alpha_{\max}-\alpha_{\min})\right],s\in[0,1]。 \tag{6}$$

式中:s 用于控制总噪声规模,$\alpha_{\min}$、$\alpha_{\max}$ 分别为添加噪声的下、上界,$\alpha_{\min}<\alpha_{\max}\in(0,1)$。

在反向扩散过程中,利用 MLP(θ)对 $\boldsymbol{z}'_T$进行迭代去噪,去噪过程 $\boldsymbol{z}'_t\to\boldsymbol{z}'_{t-1}$如式(7)所示:

$$p_\theta(\boldsymbol{z}'_{t-1}|\boldsymbol{z}'_t)=N(\boldsymbol{z}'_{t-1};\mu_\theta(\boldsymbol{z}'_t,t),\sigma^2(t)\boldsymbol{I})。 \tag{7}$$

式中:$\mu_\theta(\boldsymbol{z}'_t,t)$是 MLP($\theta$)预测上一个时间步数据分布的期望,$\theta$ 是 MLP(θ)的参数,方差 $\sigma^2(t)$在每一个时间步都是固定的,分布 p_θ 用于近似分布 q。

2.5 模型训练与推理

TRMVADM 模型需要对 VAE 和 DM 一起进行优化,故优化过程包含了 VAE 和 DM 的优化项。

2.5.1 VAE 优化

假设 $\boldsymbol{z}_0\sim q(\boldsymbol{z}_0|\boldsymbol{x}_0)$,$q(\boldsymbol{z}_0|\boldsymbol{x}_0)$的真实分布是未知的,因此可通过 f_ϕ 近似该分布,令 $\boldsymbol{z}_0\sim p_\phi(\boldsymbol{z}_0|\boldsymbol{x}_0)$,VAE 可通过最小化 KL 离散度 $D_{KL}(p_\phi(\boldsymbol{z}_0|\boldsymbol{x}_0)||q(\boldsymbol{z}_0|\boldsymbol{x}_0))$进行优化,其等价于最大化证据下界(evidence lower bound,ELBO)[3]。但在 TRMVADM 的 VAE 中未采用传统的 ELBO,而是采用了类似多项式变分自编码器 (multinomial variational autoencoders,MultiVAE)[8]的目标函数,具体为:

$$L_{VAE}(\boldsymbol{x}_0,\phi,\psi)=-\mathrm{E}_{p_\phi(\boldsymbol{z}_0|\boldsymbol{x}_0)}[\ln q_\psi(\boldsymbol{x}_0|\boldsymbol{z}_0)]+\beta\cdot D_{KL}(p_\phi(\boldsymbol{z}_0|\boldsymbol{x}_0)||q(\boldsymbol{z}_0)),\beta\in(0,1)。 \tag{8}$$

式中:等号右边第一项为重构项,第二项为先验匹配项,β 用于控制先验匹配项的强度,在优化过程中利用模拟退火方法[8]进行迭代,TRMVADM 中的 VAE 的优化通过最小化式(8)实现。

2.5.2 DM 优化

TRMVADM 模型里的 DM 优化过程也通过最大化 ELBO 实现,其中先验匹配项位无可训练参数,在优化过程中可直接忽略,新的 ELBO 为:

$$L_{\mathrm{DM}}(\boldsymbol{z}_0,\theta)=\mathrm{E}_{q(\boldsymbol{z}_1|\boldsymbol{z}_0)}[\ln p_\theta(\boldsymbol{z}'_0|\boldsymbol{z}'_1)]-\sum_{t=2}^{T}\mathrm{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}[D_{\mathrm{KL}}(q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t,\boldsymbol{z}_0)||p_\theta(\boldsymbol{z}'_{t-1}|\boldsymbol{z}'_t))]。\tag{9}$$

式中,等号右边第一项为重构项,第二项为去噪匹配项。在去噪匹配项中,$q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t,\boldsymbol{z}_0)$可通过贝叶斯法则变为:

$$q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t,\boldsymbol{z}_0)\propto N(\boldsymbol{z}_{t-1};\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t),\sigma^2(t)\boldsymbol{I}),\tag{10}$$

$$\begin{cases}\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t)=\dfrac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{z}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{z}_0}{1-\bar{\alpha}_t},\\ \sigma^2(t)=\dfrac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}。\end{cases}\tag{11}$$

式中:$\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t)$和$\sigma^2(t)$分别是$q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t,\boldsymbol{z}_0)$的均值和方差。根据式(7)和式(10),第 t 步去噪匹配项可由式(12)计算得到,

$$L_{\mathrm{DM}}^{t}=\mathrm{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\left[\frac{1}{2\sigma^2(t)}\|\mu_\theta(\boldsymbol{z}'_t,t)-\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t)\|_2^2\right]。\tag{12}$$

式中:$\mu_\theta(\boldsymbol{z}'_t,t)$用于逼近$\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t)$,$\boldsymbol{z}'_t=\boldsymbol{z}_t$。根据式(11),利用同样的方法对 $\mu_\theta(\boldsymbol{z}'_t,t)$进行分解,分解后为:

$$\mu_\theta(\boldsymbol{z}'_t,t)=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{z}'_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)}{1-\bar{\alpha}_t}。\tag{13}$$

式中:$\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)$是基于 $\boldsymbol{z}'_t$和 t 对 $\boldsymbol{z}_0$ 进行预测。将式(11)中的 $\mu'(\boldsymbol{z}_t,\boldsymbol{z}_0,t)$和式(13)中的 $\mu_\theta(\boldsymbol{z}'_t,t)$代入式(12)中,可得出新的 L_{DM}^{t}:

$$L_{\mathrm{DM}}^{t}=\mathrm{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}\left[\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}}-\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)\|\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)-\boldsymbol{z}_0\|_2^2\right]。\tag{14}$$

由于 TRMVADM 对相邻时间步的信噪比差不敏感,所以将$\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}}-\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)$置为 1,得到:

$$L_{\mathrm{DM}}^{t}=E_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}[\|\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)-\boldsymbol{z}_0\|_2^2]。\tag{15}$$

通过实验结果可知,利用式(15)计算 L_{DM}^{t} 效果更好,故采用此式。

在重构项中,此处定义 L_{DM}^{1} 作为式(9)中重构项的负值形式,具体形式为:

$$L_{\mathrm{DM}}^{1}=E_{q(\boldsymbol{z}_1|\boldsymbol{z}_0)}[\|\boldsymbol{z}'_\theta(\boldsymbol{z}'_1,1)-\boldsymbol{z}_0\|_2^2]。\tag{16}$$

式(16)也采用了式(15)的形式估计高斯对数似然。根据式(15)和式(16),DM 中的 ELBO 可表示为 $-\sum_{t=1}^{T}L_{\mathrm{DM}}^{t}$,因此最大化 $ELBO$ 等价于最小化$\sum_{t=1}^{T}L_{\mathrm{DM}}^{t}$,目标函数为:

$$L_{\mathrm{DM}}(\boldsymbol{z}_0,\theta)=\sum_{t=1}^{T}\mathrm{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_0)}[\|\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)-\boldsymbol{z}_0\|_2^2]。\tag{17}$$

故可通过优化 $\boldsymbol{z}'_\theta(\boldsymbol{z}'_t,t)$中的$\theta$ 最小化 $L_{\mathrm{DM}}(z_0,\theta)$。

2.5.3 训练和推理过程

TRMVADM 的目标函数包含了 VAE 和 DM 的目标函数,如式(18)所示:

$$L(\boldsymbol{x}_0,\phi,\psi,\theta)=\gamma\cdot L_{\mathrm{VAE}}(\boldsymbol{x}_0,\phi,\psi)+L_{\mathrm{DM}}(r(\boldsymbol{x}_0),\theta),\gamma\in(0,1)。\tag{18}$$

式中:γ 用于控制 VAE 目标函数的作用强度,与式(8)类似,γ 也使用模拟退火方法[8]进行迭代优化。TRMVADM 整个训练过程如算法 1 所示。由于在不同的时间步中,优化难度一般不一样,所以本研究采用重要性采样[7]强调学习具有较大损失值 L_{DM}^{t} 的时间步,p_t 的分布如式(19)所示:

$$p_t\propto\frac{\sqrt{\mathrm{E}[(L_{\mathrm{DM}}^{t})^2]}}{\sqrt{\sum_{t'=1}^{T}\mathrm{E}[(L_{\mathrm{DM}}^{t'})^2]}},\tag{19}$$

式中 $\sum_{t=1}^{T} p_t = 1$,在获取足够多的 L_{DM}^t 之前,先采用均匀采样 $U(1,T)$。

在 TRMVADM 推理过程中,模型首先利用式(3)对 $\boldsymbol{x}_0$ 进行维度压缩得到 $\boldsymbol{z}_0$,随后在前向扩散过程 $\boldsymbol{z}_0 \to \boldsymbol{z}_1 \to \cdots \to \boldsymbol{z}_T$ 中对 $\boldsymbol{z}_0$ 逐步加噪,然后设置 $\boldsymbol{z}'_T = \boldsymbol{z}_T$ 进行反向扩散过程 $\boldsymbol{z}'_T \to \boldsymbol{z}'_{T-1} \to \cdots \to \boldsymbol{z}'_0$,最后利用 g_ψ 对 $\boldsymbol{z}'_0$ 进行解码得到 $\boldsymbol{x}'_0$。根据 TRMVADM 推理出的用户-物品交互概率向量 $\boldsymbol{x}'_0$ 对物品进行排序,利用排序结果可得到物品推荐列表,TRMVADM 的推理过程如算法 2 所示。在反向扩散的过程中,TRMVADM 忽略了方差并利用式(13)计算 $\boldsymbol{z}'_{t-1} = \mu_\theta(\boldsymbol{z}'_t, t)$,从而得到上一个去噪转换时间步的数据。

在时间复杂度方面,VAE 的训练和推理过程的复杂度均为 $O(P_1)$,DM 的训练和推理过程的复杂度均为 $O(TP_2)$,其中 P_1 表示 VAE 的参数量,P_2 表示 DM 的参数量。由于 TRMVADM 采用了时间步采样方法,其训练过程的复杂度为 $O(P_1+P_2)$,而推理过程包含逐步去噪,其复杂度为 $O(P_1+TP_2)$。在空间复杂度方面,VAE、DM 和 TRMVADM 没有占用额外的存储空间,复杂度都为 $O(1)$。

算法 1　TRMVADM 训练

输入:全部用户-物品交互数据 X' 和随机初始化的 ϕ、ψ、θ
输出:已经优化过的 ϕ、ψ、θ
1) 使用式(1)和式(2)对 X' 进行赋权,得到 $\bar{X}$
2) repeat
3)　　采样一批用户时序数据 $X \subset \bar{X}$
4)　　for 每个用户时序数据 $\boldsymbol{x}_0 \in X$ do
5)　　使用式(3)计算 $\boldsymbol{z}_0$
6)　　采样 $t \sim \mathrm{U}(1,T)$ 或 $t \sim p_t$,$\boldsymbol{\epsilon} \sim N(\boldsymbol{0},\boldsymbol{I})$
7)　　使用式(5)计算 $\boldsymbol{z}_t$,并设置 $\boldsymbol{z}'_t = \boldsymbol{z}_t$
8)　　使用式(18)计算 $L(\boldsymbol{x}_0,\phi,\psi,\theta)$
9)　　采用梯度下降法优化 ϕ、ψ、θ
10)　　end for
11) until 算法收敛
12) 结束

算法 2　TRMVADM 推理

输入:用户 u 的物品交互数据 $\boldsymbol{x}_u$ 和训练过的 ϕ、ψ、θ
输出:用户 u 的物品交互概率 $\boldsymbol{x}'_0$
1) 采样 $\boldsymbol{\epsilon} \sim N(\boldsymbol{0},\boldsymbol{I})$
2) 使用式(1)和式(2)计算 $\boldsymbol{x}_0$
3) 使用式(3)计算 $\boldsymbol{z}_0$
4) 使用式(5)计算 $\boldsymbol{z}_T$,并设置 $\boldsymbol{z}'_T = \boldsymbol{z}_T$
5) for $t = T, T-1, \cdots, 1$ do
6)　　根据式(13)计算 $\boldsymbol{z}'_{t-1} = \mu_\theta(\boldsymbol{z}'_t, t)$
7) end for
8) 使用式(4)计算 $\boldsymbol{x}'_0$
9) 结束

3　实验

为验证 TRMVADM 模型的实际效果,本研究在 3 个不同数据集上将 TRMVADM 模型进行性能对比实验、消融实验、不同目标函数对比实验、推荐效率分析和参数敏感性分析。

3.1　实验设置

3.1.1　数据集和评价指标

实验采用的 3 个数据集分别是 MovieLens 10M 数据集、MovieLens 20M 数据集和 Yelp 数据集。MovieLens 数据集是一个电影用户评分基准数据集,Yelp 数据集是一个包含不同商店用户评分的商业数据集,数据集的基本信息如表 1 所示。本研究采取标准数据预处理方法,首先对于所有数据集,采用用户评分 3 分及以上的交互数据,并且只保留至少 10 次交互的用户和至少被交互过 10 次的物品;然后将所有评分设置为 1,并按照用户物品交互时间戳的先后顺序对所有交互数据进行排序;最后按照 7∶2∶1 的比例划分训练集、验证集和测试集。

表 1　数据集基本信息

Table 1　Basic information of datasets

数据集	类型	用户数	物品数	总交互数	稀疏百分比/%
MovieLens 10M	电影	69 584	9 175	8 233 567	98.71
MovieLens 20M	电影	137 523	14 258	16 447 894	99.16
Yelp	商业	72 487	43 748	204 3402	99.94

实验使用 TopN 推荐任务中常用的召回率(recall rate,Recall)和归一化折损累计增益(normalized discounted cumulative gain,NDCG)两种指标评估推荐模型的性能。其中,Recall 用于表示有多少比例的用户-物

品交互记录包含在最终的预测推荐列表里,NDCG 用于评估预测的推荐列表和用户真实交互列表的差距。假设推荐列表长度为 K,Recall 和 NDCG 分别对应 R@K 和 N@K,根据文献[11]中的实验设定,K 设为 20 或 50。

3.1.2 实验环境和参数设置

实验所使用的软件环境为 Python 3.11.5,Pytorch 2.1.0,CUDA 12.4,硬件环境为 NVIDIA RTX 4070;训练初始学习率为 0.000 1,每迭代 40 次衰减一半,批大小为 128,优化器采用 AdamW,训练过程中使用早停策略,如迭代 40 次后验证集的性能没有增长就停止训练;在 TRMVADM 模型中,MLP(θ)隐层大小范围是{[256],[256,512],[1 024]},T 的范围是{2,5,10,20,40,80,120},s 的范围是{0,0.000 01,0.000 1,0.005,0.01,0.1,1},$\alpha_{\min}$ 和 $\alpha_{\max}$ 的范围分别是{0.000 1,0.001,0.005}和{0.005,0.01,0.1},γ 的范围是{0.01,0.03,0.05,0.07,0.09},$w_{\min}$ 的范围是{0.1,0.3,0.5,0.7,0.9},$w_{\max}$ 固定为 1。

3.1.3 基线模型

本实验将 TRMVADM 和其他有竞争力的基线模型进行比较,分别是 MF-BPR[14]、ENMF[15]、LightGCN[16]、MultiVAE[8]、CODIGEM[11]和 DiffRec[7],其中前 3 个为非生成式模型,后 3 个为生成式模型。MF-BPR 是一种经典的基于矩阵分解的协同过滤方法,ENMF 是一种基于神经网络的矩阵分解模型,LightGCN 是一个基于图卷积网络的推荐模型。

3.2 模型性能对比实验

在模型性能对比实验中,TRMVADM 模型和其他 6 种基线模型在 3 个数据集上对比 R@20、R@50、N@20 和 N@50 等 4 个指标,对比结果如表 2 所示,其中最优结果以加黑表示,次优结果以下划线表示。由表 2 可以看出,TRMVADM 模型除了在 MovieLens 10M 数据集的 N@20 指标上没有取得最优,在其他指标上均取得最优性能,其中 MovieLens 20M 数据集与 Yelp 数据集的指标相对于次优结果提升最低为 1.6%,最高为 9.68%,表明 TRMVADM 能够充分利用 VAE 和 DM 的优点对时序数据进行建模,有效提升模型的推荐性能,验证了将两者结合的优越性。

表 2 模型性能对比

Table 2 Comparison of model performance

模型	MovieLens 10M				MovieLens 20M				Yelp			
	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
MF-BPR	0.066 2	0.127 1	0.047 5	0.066 9	0.067 5	0.126 9	0.047 9	0.066 5	0.002 7	0.004 8	0.001 4	0.001 9
ENMF	0.108 4	0.232 0	0.070 8	0.110 0	0.106 8	0.223 6	0.071 4	0.108 5	0.053 1	0.099 7	0.026 3	0.037 9
LightGCN	0.103 6	0.215 7	0.069 8	0.105 5	0.100 0	0.204 6	0.069 2	0.102 5	0.066 1	0.125 1	0.032 0	0.046 4
MultiVAE	0.106 8	0.239 5	0.068 7	0.110 5	0.107 0	0.232 5	0.070 7	0.110 4	0.057 8	0.112 3	0.027 1	0.040 4
CODIGEM	0.1103	0.224 4	0.073 1	0.109 5	0.109 4	0.219 5	0.073 7	0.109 0	0.052 5	0.099 6	0.025 9	0.037 7
DiffRec	0.126 6	0.249 7	**0.087 4**	0.126 2	0.125 7	0.239 7	0.088 2	0.124 1	0.062 1	0.112 3	0.030 8	0.043 3
TRMVADM	**0.132 0**	**0.264 0**	0.084 0	**0.127 0**	**0.136 3**	**0.262 9**	**0.090 6**	**0.131 6**	**0.067 9**	**0.127 1**	**0.033 3**	**0.047 7**

3.3 消融实验

为了评估 TRMVADM 模型里各个组成部分对整体性能的贡献,分别在 TRMVADM 模型中分别去除时序重加权策略和 VAE、VAE、时序重加权策略得到对应模型 RMDM、TRMDM、RMVADM,然后在 3 个数据集上进行实验,并加入 MultiVAE 进行对比,实验结果如表 3 所示。由表 3 可以看出,RMVADM 的指标相对于 MultiVAE 提升最低为 8.24%、最高为 24.47%,相对于 RMDM 提升最低为 10.68%、最高为 66.69%,表明结合了 VAE 和 DM 的方法相比于单纯基于 VAE 或 DM 的方法,更能深入挖掘用户-物品潜在的交互信息。而融合了时序信息的 TRMVADM 的指标相对于 RMVADM 提升最低为 1.81%、最高为 10.96%,表明 TRMVADM 中的 VAE 能够有效捕获时序偏移特征,并利用 DM 对时序数据进行扩散,从而提升了模型的性能。TRMDM 的性能表现最差,是由于 DM 的前向加噪过程破坏了原有的用户时序信息,导致 DM 无法对用户-物品的交互生成过程进行正确建模,时序重加权策略在此处起到负面效果。

表 3 消融实验结果

Table 3 Results of ablation experiments

模型	MovieLens 10M				MovieLens 20M				Yelp			
	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
MultiVAE	0.106 8	0.239 5	0.068 7	0.110 5	0.107 0	0.232 5	0.070 7	0.110 4	0.057 8	0.112 3	0.027 1	0.040 4
RMDM	0.100 7	0.183 1	0.068 3	0.095 0	0.085 4	0.153 1	0.058 9	0.080 8	0.057 2	0.104 3	0.028 1	0.039 9
TRMDM	0.085 6	0.153 2	0.058 4	0.080 6	0.074 7	0.129 6	0.051 6	0.069 4	0.054 5	0.099 3	0.026 8	0.038 1
RMVADM	0.122 9	0.259 3	0.075 7	0.119 6	0.129 2	0.255 2	0.088 0	0.128 2	0.064 2	0.122 0	0.031 1	0.045 1
TRMVADM	**0.132 0**	**0.264 0**	**0.084 0**	**0.127 0**	**0.136 3**	**0.262 9**	**0.090 6**	**0.131 6**	**0.067 9**	**0.127 1**	**0.033 3**	**0.047 7**

3.4 不同目标函数对比实验

为了探究不同目标函数对模型性能的影响,实验中设计了 TRMVADM 对应的变体 TRMVADM(w)和 TRMVADM(ε),其中 TRMVADM(w)的 L_{DM}^{t} 使用式(14)计算,TRMVADM(ε)的 L_{DM}^{t} 使用文献[17]中预测噪声 ε 计算。TRMVADM 和对应变体在 3 个数据集上进行了对比实验,实验结果如表 4 所示。由表 4 可以看出,TRMVADM(w)和 TRMVADM(ε)的性能表现均不及 TRMVADM,表明使用式(15)的去噪匹配项的 TRMVADM 更适用于推荐任务。式(14)中二级范数的权重是通过调节 $\boldsymbol{z}_{\theta}'(\boldsymbol{z}_{t}',t)$ 精确预测 $\boldsymbol{z}_{0}$,由于 TRMVADM 对相邻时间步的信噪比差不敏感,使调节机制呈现负面效果,并且预测噪声 ε 对推荐结果作用较小,故两种变体都不合适。

表 4 不同目标函数的模型对比

Table 4 Comparison of models with different objective functions

模型	MovieLens 10M				MovieLens 20M				Yelp			
	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
TRMVADM(w)	0.081 3	0.148 3	0.061 8	0.083 3	0.080 2	0.149 1	0.060 7	0.082 7	0.012 3	0.023 7	0.005 8	0.00 86
TRMVADM(ε)	0.106 3	0.190 9	0.078 3	0.105 5	0.101 5	0.180 1	0.076 5	0.101 7	0.018 8	0.035 8	0.010 0	0.014 5
TRMVADM	**0.132 0**	**0.264 0**	**0.084 0**	**0.127 0**	**0.136 3**	**0.262 9**	**0.090 6**	**0.131 6**	**0.067 9**	**0.127 1**	**0.033 3**	**0.047 7**

3.5 模型推荐效率分析

表 5 分别展示了 MultiVAE、DiffRec 和 TRMVADM 在 3 个数据集上的计算开销,其中参数量表示模型需要训练的参数量,平均推理时间表示训练时每个 epoch 的平均推理时间。由表 5 可见,相比于基于 VAE 的 MultiVAE 和基于 DM 的 DiffRec,由于 TRMVADM 包含了 VAE 和 DM 结构,其参数量多于前两者。但得益于 VAE 数据降维特性和 DM 中多概率分布的易处理性,在 MovieLens 10M 数据集上,TRMVADM 的平均推理时间与其他模型差异不大,在数据维度较高的 MovieLens 20M 和 Yelp 数据集上相比于次优结果分别降低了 16.4 和 2.19 s,证明 TRMVADM 在训练参数量较大的情况下,能够满足推荐任务的效率需求。

表 5 参数量与推理时间对比

Table 5 Comparison of parameter numbers and inference time

模型	MovieLens 10M		MovieLens 20M		Yelp	
	参数量/百万个	平均推理时间/(s/epoch)	参数量/百万个	平均推理时间/(s/epoch)	参数量/百万个	平均推理时间/(s/epoch)
MultiVAE	11.38	8.69	17.49	39.14	52.90	12.97
DiffRec	18.81	9.07	29.23	34.57	89.65	16.02
TRMVADM	30.31	8.84	45.93	18.17	136.55	10.78

3.6 参数敏感性分析

为了研究 T、s、γ 和 $w_{\min}$ 对 TRMVADM 的影响,本研究根据 3.1.2 节提到的参数范围在 MovieLens

20M 数据集上进行了不同参数的性能对比实验,实验结果如图 2 所示。由图 2 可以看出:①对于 T,由于正向扩散过程的噪声相对较小,增加扩散过程的 T 对模型性能影响较小,而且随着 T 的增加,模型的推理时间会逐步上升,因此考虑到时间步过大会造成较高的计算负担,把 T 设置在 10 以内较为合适;②对于 s,由 0 到 0.000 01,模型性能得到初步提升,表明 DM 中去噪训练的有效性。但是随着 s 继续增大,模型性能变化较小,说明 VAE 能够缓解 DM 中噪声对模型的影响,提升模型推理的稳定性;③对于 γ,随着 γ 的加大,R@50 和 N@50 均先上升后下降,其峰值在 0.03 附近,表明 TRMVADM 的核心在于 DM 中的建模过程,VAE 起到辅助作用;④对于 $w_{\min}$,R@50、N@50 的峰值分别在 0.3、0.5 附近,由于 $w_{\min}$ 过于逼近 $w_{\max}$ 会导致模型性能明显下降,故 $w_{\min}$ 设置在[0.1,0.5]区间较为恰当。

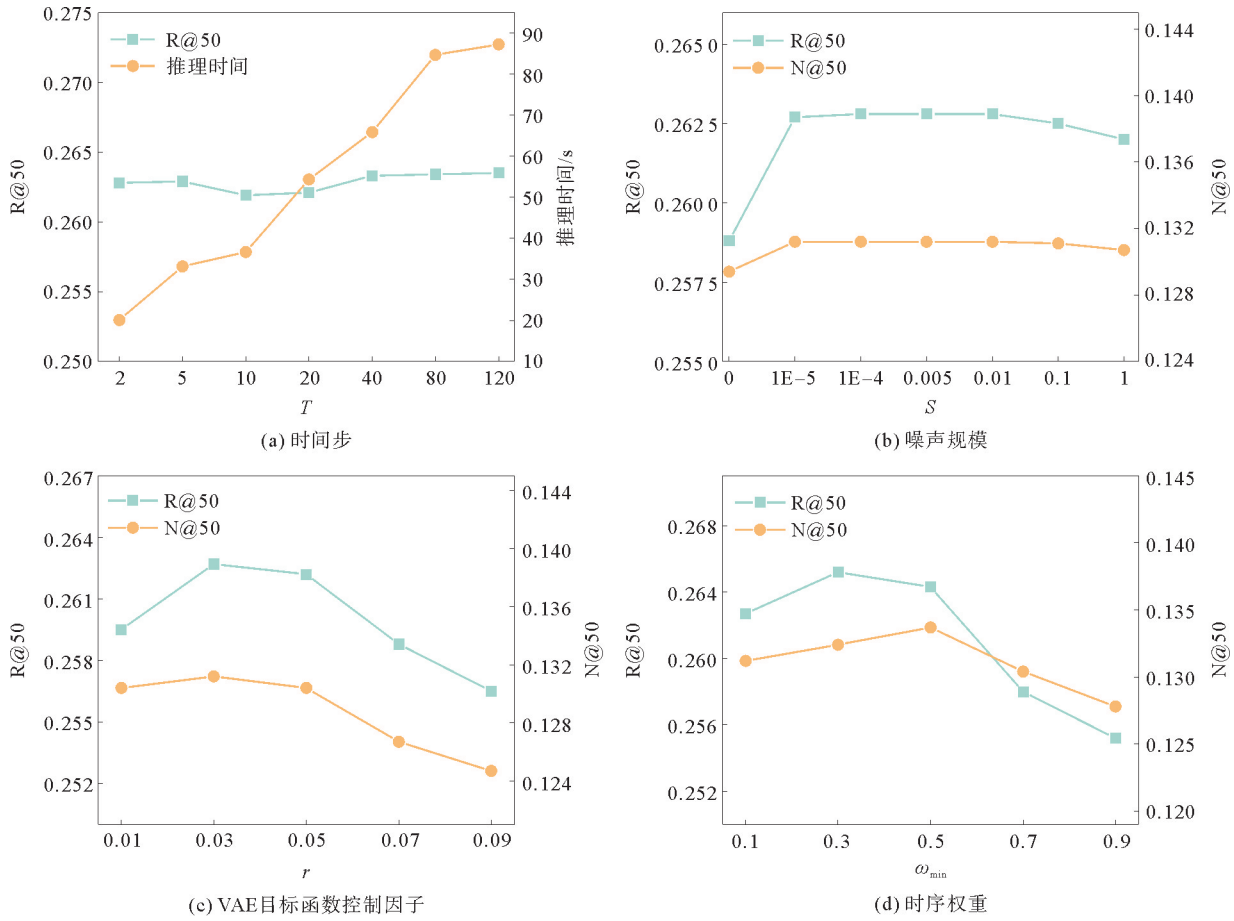


图 2 不同参数的性能对比

Fig. 2 Performance comparison of different parameters

4 结论

针对基于 DM 或 VAE 的推荐模型的缺点,本研究提出通过融合 VAE 和 DM,并结合时序信息对用户-物品的交互生成过程进行建模的生成推荐模型。该模型利用时序重加权策略对用户-物品交互数据赋予时序信息,通过 VAE 进行数据维度压缩和提取交互数据的时序特征,并利用 DM 的扩散过程深入挖掘交互数据的潜在信息,提升模型推荐性能;同时 VAE 有效缓解了噪声对模型的影响,DM 简化了复杂后验分布的处理,提升模型的稳定性和表征能力。所提出的模型有效性在 MovieLens 10M、MovieLens 20M 和 Yelp 数据集上得到了验证。下一步的工作将探索基于条件引导的生成推荐范式,使推荐结果更符合用户的偏好。

参考文献:

[1] 赵建立,姚彬,陈建建,等.结合隐式反馈与相关性建模的概率流式张量分解推荐模型[J].山东科技大学学报(自然科学版),2023,42(3):76-84.
ZHAO Jianli,YAO Bin,CHEN Jianjian,et al. Probabilistic streaming tensor decomposition recommendation model with implicit feedback and correlation modeling[J]. Journal of Shandong University of Science and Technology(Natural Science),2023,42(3):76-84.

[2] MANAA N,SERIDI H,MENDJEL M S M. Advancements in recommender systems through the integration of generative adversarial networks[J]. International Journal of Informatics and Applied Mathematics,2023,6(2):35-45.

[3] HUANG T,LIANG C,WU D,et al. A debiasing autoencoder for recommender system[J]. IEEE Transactions on Consumer Electronics,2023,70(1):3603-3613.

[4] QIN Y F,WU H J,JU W,et al. A diffusion model for POI recommendation[J]. ACM Transactions on Information Systems,2023,42(2):1-27.

[5] WEN J,CHEN B Y,WANG C D,et al. PRGAN:Personalized recommendation with conditional generative adversarial networks[C]//2021 IEEE International Conference on Data Mining (ICDM). IEEE,2021:729-738.

[6] ZHANG Y H,ZHAO C,YUAN M,et al. Unifying attentive sparse autoencoder with neural collaborative filtering for recommendation[J]. Intelligent Data Analysis,2022,26(4):841-857.

[7] WANG W J,XU Y Y,FENG F L,et al. Diffusion recommender model[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:Association for Computing Machinery,2023:832-841.

[8] LIANG D W,KRISHNAN R G,HOFFMAN M D,et al. Variational autoencoders for collaborative filtering[C]//Proceedings of the 2018 World Wide Web Conference. 2018:689-698.

[9] WANG J,LIU G D,WU J,et al. Self-supervised variational autoencoder for recommender systems[C]//2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE,2021:831-835.

[10] LIU W,HOU U L,LIANG S,et al. Revisiting positive and negative samples in variational autoencoders for top-N recommendation[C]//28th International Conference on Database Systems for Advanced Applications. Cham:Springer Nature Switzerland. Tianjin,Apr. 17-20,2023:563-573.

[11] WALKER J,ZHONG T,ZHANG F L,et al. Recommendation via collaborative diffusion generative model[C]//International Conference on Knowledge Science,Engineering and Management. Cham:Springer International Publishing,2022:593-605.

[12] LI Z H,SUN A,LI C L. DiffuRec:A diffusion model for sequential recommendation[J]. ACM Transactions on Information Systems,2023,42(3):1-28.

[13] 赵铁柱,梁校伦,杨秋鸿,等.基于异质信息对齐和重排序的跨模态行人重识别方法[J].山东科技大学学报(自然科学版),2024,43(2):79-89.
ZHAO Tiezhu,LIANG Xiaolun,YANG Qiuhong,et al. Cross-modal person re-identification method based on heterogeneous information alignment and reranking[J]. Journal of Shandong University of Science and Technology(Natural Science),2024,43(2):79-89.

[14] YOU Y L,WANG Z Y. Preference-aware Bayesian personalized ranking for point-of-interest recommendation[J]. Journal of Intelligent and Fuzzy Systems,2023,44(5):7113-7119.

[15] CHEN C,ZHANG M,ZHANG Y F,et al. Efficient neural matrix factorization without sampling for recommendation [J]. ACM Transactions on Information Systems (TOIS),2020,38(2):1-28.

[16] HE X N,DENG K,WANG X,et al. LightGCN:Simplifying and powering graph convolution network for recommendation[C/OL]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020. DOI: 10.1145/3397271.3401063.

[17] ROMBACH R,BLATTMANN A,LORENZ D,et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:10684-10695.

(责任编辑:齐敏华)

DOI: 10.16452/j.cnki.sdkjzk.2026.01.011　　文章编号:1672-3767(2026)01-0122-09

基于多维空间视觉感知的仓库盘点方法

钟　诚[1],田殿雄[1],赵文炎[1],卢泽钰[1],张良安[2],李　勇[2]
(1.国网冀北电力有限公司唐山供电公司,河北 唐山 063000;2.上海洲固电力科技有限公司,上海 200040)

摘　要:传统人工仓库物资盘点方式效率低下,难以满足现代智慧仓储对物资精准管理的需求。现有的自动识别技术存在易受环境干扰、成本高昂以及在复杂场景下稳定性不足等局限,尤其在动态仓库环境中,物资盘点本质上是一个多维时空信息处理问题,需要有效融合时间、空间与距离等不局限于三维物理结构的高维度抽象特征,同时还需应对遮挡、光线变化和小目标检测等挑战。针对上述问题,本研究提出一种基于多维空间视觉感知的仓库盘点方法(MDSVP-WIM)。该方法通过混合高斯模型动态更新背景来抑制环境干扰,结合二维卷积与图像分割技术提升小目标识别精度,并引入跨帧追踪-联合投票机制,增强系统在动态场景中的稳定性与鲁棒性。在 WareSegNet 与 SpatioTrack-360 数据集上的实验表明,本研究方法的平均精度均值(mAP)分别达到 0.97 和 0.96,F1 值为 0.94 和 0.93,性能显著优于 FCN、PSPNet 等基准模型,为复杂仓储环境下的物资自动化盘点提供了一种高效、稳定的方案。

关键词:智慧仓储;物资盘点;视觉感知;混合高斯模型;跨帧追踪
中图分类号:TB3; U411　　**文献标志码**:A

A warehouse inventory method based on multidimensional spatial visual perception

ZHONG Cheng[1], TIAN Dianxiong[1], ZHAO Wenyan[1], LU Zeyu[1], ZHANG Liang'an[2], LI Yong[2]
(1. Tangshan Power Supply Company of State Grid Jibei Electric Power Co., Ltd., Tangshan 063000, China;
2. Shanghai Zhougu Power Technology Co., Ltd., Shanghai 200040, China)

Abstract: The inefficiency of the traditional manual material inventory method makes it difficult to meet the precise management needs of modern intelligent warehousing. However, the existing automatic recognition technologies have limitations such as susceptibility to environmental interference, high costs, and insufficient stability in complex scenarios. Especially in dynamic warehouse environments, material inventory is essentially a multidimensional spatiotemporal information processing problem, which requires effective integration of high-dimensional abstract features, such as time, space, and distance that are not limited to three-dimensional physical structures, and addresses challenges such as occlusion, light changes, and small object detection. In response to the above issues, this study proposes a multidimensional spatial visual perception based warehouse inventory method (MDSVP-WIM). This method suppresses environmental interference by dynamically updating the background through a Gaussian mixture model, improves the accuracy of small object recognition by combining the techniques of two-dimensional convolution and image segmentation, and enhances the stability and robustness of the system in dynamic scenes by introducing cross-frame tracking joint voting mechanism. The experiments on the WareSegNet dataset and SpatioTrack-360 dataset show that the proposed method, with the mean average precision (mAP) reaching 0.97 and 0.96 respectively and the F1-Score reaching 0.94 and 0.93 respectively, has a significantly better performance than benchmark models such as FCN and PSPNet. This study provides an efficient

收稿日期:2024-12-26
基金项目:山东省自然科学基金项目(ZR2024MF142)
作者简介:钟　诚(1978—),男,吉林长春人,硕士,高级经济师,主要研究方向为物资管理.
赵文炎(1985—),男,湖北仙桃人,硕士,高级工程师,主要研究方向为物资管理、轨迹预测,本文通信作者.
Email:15175579560@139.com

and stable visual perception solution for automated material inventory in complex warehousing environments.

Key words: intelligent warehousing; material inventory; visual perception; Gaussian mixture model; cross-frame tracking

随着物流行业的快速发展,仓储向智能化、自动化方向的深度转型,传统的人工盘点方式已难以满足高效、精准的管理需求[1-2]。尤其是在复杂仓库场景中,高密度货架导致的视觉遮挡、多品类货物混放带来的特征干扰,以及低光照或强反光等环境因素,进一步放大了人工操作的局限性。因此,开发高效的自动仓储物资盘点方法已成为物流仓储行业的迫切需求[3]。

目前主流的自动识别技术中,条码识别和磁卡识别等技术存在纸质标签易损坏、易遮挡以及电子标签成本较高等问题。在此背景下,基于计算机视觉的自动盘点技术正逐渐成为行业研究热点。该技术通过深度学习模型,对视频中货物视觉特征进行自动识别与提取,有效避免了传统标签技术带来的局限性。

基于卷积神经网络(convolutional neural network,CNN)的目标检测算法以物体及边界框为核心表征方式,但存在小目标像素占比低导致特征提取不充分、漏检率高达 30%,以及近距离密集目标边界模糊引发锚框回归重叠、误检率超过 25%等问题[4]。这类传统模型侧重于单帧图像处理,缺乏对跨帧时序信息的有效利用,难以应对货物移动导致的连续帧间特征变化。此外,基于自注意力机制的模型虽然在长距离依赖建模中性能优异,但在仓库场景中对局部细节的捕捉能力通常弱于 CNN 模型[5]。

针对上述问题,本研究提出一种基于多维空间视觉感知的仓库盘点方法(multidimensional spatial visual perception based warehouse inventory method, MDSVP-WIM),其主要创新有:

1) 多维空间去噪模块利用混合高斯(mixture of Gaussian, MoG)模型动态更新背景,能够处理时间、距离等抽象特征,并通过马氏距离计算像素与背景分布的差异,实现前景目标的快速分离,减少动态背景干扰;

2) 多维空间视觉感知模块通过多维卷积并行提取多尺度特征,模块中的多个分支分别从不同角度对图像进行处理,每个分支提取的特征包含了不同的空间信息,从而增强对高维抽象特征的感知能力;

3) 多维空间跨帧跟踪-联合投票模块以边界框几何中心点为核心跟踪锚点,结合视频流连续性实现动态目标的稳定定位,并通过多帧结果融合提升语义判别准确性,有效增强系统在动态场景下对移动货物的跟踪与识别的鲁棒性;

4) 实验结果表明,MDSVP-WIM 在 WareSegNet 和 SpatioTrack-360 数据集上的检测平均精度均值(mean average precision, mAP)分别达到 0.97 和 0.96,F1 值分别为 0.94 和 0.93,性能显著优于全卷积网络(fully convolutional networks,FCN)、金字塔场景解析网络(pyramid scene parsing network, PSP-Net)等基准模型,为复杂仓储环境下实现高效、智能的物资盘点提供了有效的解决方案。

1　相关工作

在仓库自动化管理领域,目标识别与盘点技术的研究主要围绕计算机视觉与深度学习展开。现有方法可归纳为传统特征工程方法、基于 CNN 的目标检测方法和语义分割方法三大类。

早期研究多依赖于人工设计的特征算子提取目标信息。例如,王紫芮等[6]采用尺度不变特征变换方法生成 128 维特征向量,但该方法计算复杂度高,对边缘光滑目标特征响应弱,在动态仓储场景中漏检率高达 40%;Dalal 等[7]提出的定向梯度直方图方法需手动设置滑动窗口参数,在处理密集遮挡目标时易出现特征混淆,误检率达 35%,且难以有效应对多尺度目标共存场景。因此,传统方法因依赖人工特征设计,泛化能力有限,难以适应仓储场景的复杂性与动态性。

近年来,基于 CNN 的目标检测算法逐渐成为主流,其通过卷积层自动学习特征,实现了从人工特征到数据驱动特征的跨越,显著提升了检测精度与泛化能力。Faster R-CNN[8]作为两阶段算法的代表,能够端到端地完成目标定位与分类,在标准尺寸目标上检测精度较高,但对小目标的漏检率也偏高。YOLO 系列[9]采用单阶段架构,检测速度较快,如 YOLOv3 引入多尺度预测机制,提升了对不同尺寸目标的检测能力,但在仓储复杂环境中仍存在因干扰导致的误检问题[10]。

针对上述目标检测的粗粒度缺陷,语义分割技术通过像素级分类提供了更精细化的特征。Shelhamer 等[11]提出的 FCN 以卷积层替代全连接层,支持任意尺寸输入,为语义分割任务奠定了基础,但 FCN 对边界复杂、形状不规则的目标分割效果较差,在仓储异形货物分割中精度仅 60%。Zhao 等[12]提出的 PSP-Net 通过金字塔池化模块融合多尺度上下文信息,但在仓储密集遮挡环境中仍存在特征混淆和误检率高的问题。全景分割网络 UPerNet[13]作为面向全景分割的统一框架,借助编码器-解码器结构与特征金字塔,将仓储场景分割精度提升至 70%,但小目标漏检率仍达 30%。双重注意力网络 DANet[14]通过空间与通道双重注意力机制增强长程依赖建模,但局部细节捕捉能力较弱,计算复杂度较高,相比 UPerNet 精度提高约 3%。专为遥感图像分割设计的多尺度动态图卷积神经网络 DMNet[15]引入动态尺度选择机制,但对小目标特征提取仍不充分,在密集目标场景下误检率超过 25%。

为突破上述瓶颈,研究者进一步提出了通分万物模型 v2(segment anything model v2, SAM v2)[16],该模型是首个将图像分割与视频分割统一到端到端框架中的通用大模型,基于时空记忆注意力机制缓存历史帧键值,能以 44 帧/s 的速度进行零样本掩码输出。动态上下窗口方法(dynamic context-swin v2, DC-Swin v2)[17]通过嵌入动态上下文窗口,可根据目标密度自适应调整大小与步长,在仓储任务中分割精度达到 75%。

2 MDSVP-WIM 方法设计

针对小目标特征提取不充分、复杂边界分割困难等问题,本研究提出 MDSVP-WIM,其整体架构如图 1 所示,由多维空间去噪模块、多维空间视觉感知模块以及跨帧跟踪-联合投票模块构成,输入为多物资监控的视频流,输出识别物资的类别、位置及置信度等信息。

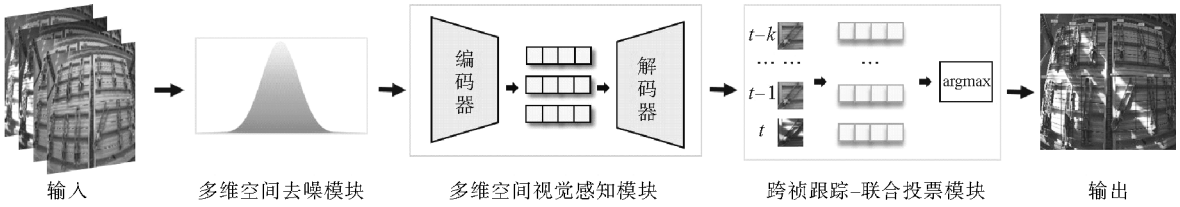


图 1 MDSVP-WIM 架构

Fig. 1 MDSVP-WIM architecture

2.1 多维空间去噪模块

多维空间去噪模块旨在减弱视频流中外部环境因素对检测目标的影响,其核心在于整合时间变化、空间关系和多通道特征等高维度信息,实现对视频流的多维特征建模。已有研究多采用背景减除方法(background subtraction methods, BSM),其核心思想是通过从当前帧中减去背景模型,从而识别出前景目标所对应的像素区域。

与传统方法不同,MDSVP-WIM 的多维空间去噪模块将输入的 RGB 像素数据扩展到长度为 5 的向量数据,该向量的 5 个特征值分别为 RGB 像素数据、像素点的空间距离和像素点在连续两帧(t 和 $t+1$ 时刻)图像中的空间距离变化率。如式(1)所示,$\boldsymbol{L}_{(x,y)}$ 不再是传统的 RGB 像素值,而是当前像素点(x,y)所对应长度为 5 的向量,其中 $\boldsymbol{L}_{(x,y)}$ 中的前 3 个元素分别对应像素点的三通道颜色值,$\boldsymbol{L}_{(x,y)}$ 中的第 4 个元素表示深度传感器获取像素点的空间距离,$\boldsymbol{L}_{(x,y)}$ 中的第 5 个元素表示像素点在连续两帧图像中的空间距离变化率。$\boldsymbol{L}^{\mathrm{bg}}_{(x,y)}$ 为估计的背景模型在 5 个维度上的特征值,$\boldsymbol{F}_{(x,y)}=(F^1_{(x,y)},F^2_{(x,y)},\cdots,F^5_{(x,y)})$用于衡量像素点偏离背景的程度。通过应用阈值函数进行前景分割,得到二值分割结果 $\boldsymbol{B}_{(x,y)}=(B^1_{(x,y)},B^2_{(x,y)},\cdots,B^5_{(x,y)})$,判断依据为 $\boldsymbol{F}_{(x,y)}$ 在 5 个维度上的特征值与预设阈值 T 的大小关系,如式(2)所示。

$$\boldsymbol{F}_{(x,y)}=\boldsymbol{L}_{(x,y)}-\boldsymbol{L}^{\mathrm{bg}}_{(x,y)}。\tag{1}$$

$$B^i_{x,y}=\begin{cases}255, & |F^i_{(x,y)}|>T;\\ 0, & |F^i_{(x,y)}|\leqslant T。\end{cases}\tag{2}$$

式中,$i=1,2,\cdots,5$。

MDSVP-WIM 中的多维空间去噪模块通过融合连续帧序列特征和空间距离维度,减弱背景干扰和对象噪点,增强主体边缘特征。为了便于数据训练,本研究使空间距离变化率服从多元高斯分布来完成背景减除,将每个像素的历史空间距离变化率视为高斯分布的混合值。当新帧在 $t+1$ 时刻出现时,通过计算每个像素与高斯分布的马氏距离来比较背景差异,并据此判断是否更新 $t+1$ 时刻分布的权重。

2.2　多维空间视觉感知模块

多维空间视觉感知模块以去噪后的单帧图像为输入,采用编码器-解码器结构提取多尺度空间特征,并完成像素级语义分割,输出与输入尺寸相同的类别掩码,为后续跨帧跟踪与联合投票提供精确的目标区域依据。

MDSVP-WIM 的空间视觉感知模块采用编码-解码器结构,编码卷积层的内部结构如图 2 所示,其设计遵循"归一化→特征增强→再归一化→特征投影"的流程。首先对输入特征进行批量归一化,然后引入注意力机制模块以增强特征的判别性。该模块采用轻量化深度可分离卷积网络架构,包含 1×1 卷积层、激活函数以及 MDSVP-EM 模块。特征先经过注意力模块处理后,再通过批量归一化层以保持稳定的数值分布,最后输入到前馈神经网络进行多阶特征投影和表达增强。前馈神经网络由 1×1 卷积层、d,3×3 卷积与 MDSVP-EM 构成。其中,1×1 卷积层用于通道调整与特征映射;深度可分离 d,3×3 卷积通过标准 3×3 卷积的中尺度感受野强化局部空间特征提取,d 表示卷积层的输入通道数,深度可分离卷积通过解耦空间与通道进行卷积;MDSVP-EM 则在空间中对特征进行非线性投影与增强调制,结构如图 3 所示。

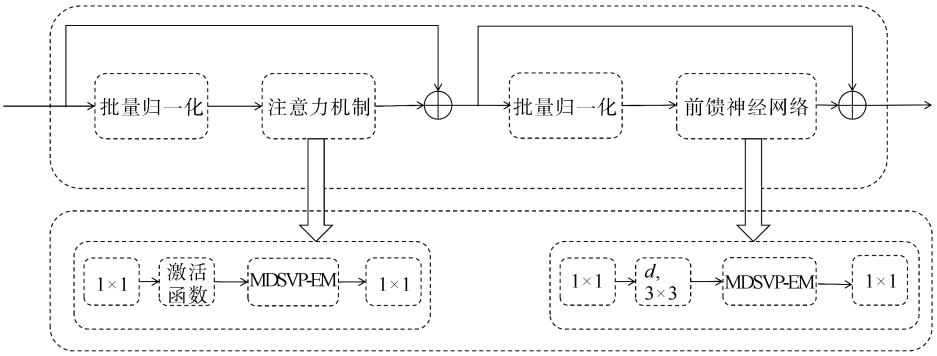


图 2　编码卷积层结构图

Fig. 2　The structure of encoder convolutional layer

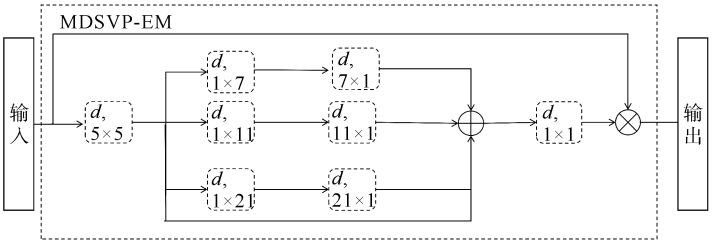


图 3　MDSVP-EM 结构图

Fig. 3　Structure of MDSVP-EM

MDSVP-EM 模块采用卷积并行提取策略,并对关键区域进行显式增强。首先,输入特征通过 5×5 的深度可分离卷积进行初步编码,然后并行输入至不同尺度的条状深度可分离卷积分支中。接着,对多尺度特征进行通道级的加权融合,融合后的特征再通过一个 1×1 的可分离卷积计算每个像素的注意力权重,以评估其在空间表达中的重要性。最后将该注意力图与原始输入特征进行逐元素乘法操作,得到增强后的空间特征表示。其中,条状深度可分离卷积分支的卷积核大小分别为 7、11 和 21。卷积核的参数量和计算量随卷积核的增长呈指数增长,为实现轻量化和高效率,如图 3 所示,采用 7×1 和 1×7 的卷积操作来模拟 7×7 标准二维卷积,同时结合激活函数和深度可分离卷积进行特征融合。通过带有激活函数的

逐通道深度可分离卷积,融合来自不同感受野的通道特征,获得目标各区域像素的注意力分数,最后通过加权计算强化目标主体特征表示,如式(3)和式(4)所示。

$$\boldsymbol{A}_{\mathrm{tt}}=\mathrm{Conv}_{1\times1}\left(\sum_{i=0}^{2}\mathrm{scale}_i\left(\mathrm{DeepConv}(\boldsymbol{X})\right)\right), \tag{3}$$

$$\boldsymbol{O}_{\mathrm{ut}}=\boldsymbol{A}_{\mathrm{tt}}\otimes\boldsymbol{X}。 \tag{4}$$

式中,$\boldsymbol{X}$ 表示 MDSVP-EM 模块的输入特征,即当前层当前时刻每个像素经过逐通道卷积处理后的历史信息;$\boldsymbol{A}_{\mathrm{tt}}$ 和 $\boldsymbol{O}_{\mathrm{ut}}$ 分别表示注意力图和输出特征;$\otimes$表示元素矩阵乘法运算;DeepConv 表示深度卷积;scale_i 表示图 3 中的第 i 个分支。前馈神经网络将这些增强后的特征表示编码为低维特征,同时保留有价值的特征。

为捕捉多层次语义信息,通过跳跃连接融合每一层编码卷积层输出的主体特征表示,并采用上采样策略恢复分辨率。本研究使用结构简单且拟合能力强的多层感知机(multilayer perceptron,MLP)来提取高级语义特征。首先,用一维卷积将编码器最后三层通道的特征全部对齐到相同维度;然后,对每层特征依次采用最近邻上采样进行放大,并与同尺度的跳跃连接特征拼接;拼接后的特征通过深度可分离卷积融合细节与语义信息,再经过一维卷积逐步压缩通道数,在保留空间细节特征的同时减少冗余通道,降低计算量并突出关键语义;最后,使用与两层 1×1 卷积操作等效的 MLP 实现逐像素分类。

2.3 跨帧跟踪-联合投票模块

为提高模型稳定性与鲁棒性,本研究提出目标跨帧跟踪-联合投票模块。该模块利用视频流的连续性,通过跨帧跟踪实时定位目标,并采用联合投票机制融合历史帧信息。具体地说,首先以边界框中心点作为跟踪初始点,确定初始框的中心点坐标,然后在后续帧中根据中心点的位移调整边界框位置。对于一个边界框,其左上角坐标为$(x_{\min},y_{\min})$,右下角坐标为$(x_{\max},y_{\max})$,中心点$(x_{\mathrm{center}},y_{\mathrm{center}})$的计算方法为:

$$(x_{\mathrm{center}},y_{\mathrm{center}})=\left(\frac{x_{\min}+x_{\max}}{2},\frac{y_{\min}+y_{\max}}{2}\right)。 \tag{5}$$

跨帧追踪算法基于目标在连续帧间的位置连续性假设,通过中心点的位移调整边界框位置。在后续帧中,算法根据当前边界框的几何坐标重新计算中心点,并据此调整边界框位置。

为进一步提升类别判断的准确性,采用联合投票机制对多帧中同一跟踪区域的主体特征进行语义判断。该方法采用特征记忆池存储同一跟踪对象在多帧中的区域像素信息,将多次识别结果融合为一个更精确的单一结果。联合投票机制可形式化表示为:

$$\hat{S}_{i,j}=\arg\max_{C=1}^{k}N_{i,j}(C)。 \tag{6}$$

式中:$\hat{S}_{i,j}$ 表示像素点(i,j)的联合投票结果,k 为当前分割检测目标的数量;$N_{i,j}(C)$表示在 k 次分割结果中将类别 C 分配给像素(i,j)的次数,argmax 表示选取使 $N_{i,j}(C)$最大的类别 C。

3 实验

3.1 实验数据集

为评估模型性能,本研究采用无人仓库自行采集的 WareSegNet 和 SpatioTrack-360 数据集,主要包含电网物资及零部件图像。数据集总涵盖 120 h 视频、10 368 000 张图像,涉及 10 类物资,场景覆盖多样化的仓储环境。数据集按 75%∶20%∶5%的比例划分为训练集、测试集和验证集。

3.2 评估指标

本研究采用精确度(Precision)、召回率(Recall)、平均精度(mean average precision,mAP)和 F1 值作为定量评估指标,分别记为 P、R、m_{AP} 和 F_1,如式(7)~(11)所示。

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}, \tag{7}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}, \tag{8}$$

$$A_{\mathrm{P}}=\int_0^1 P(R)\,\mathrm{d}R, \tag{9}$$

$$m_{\mathrm{AP}}=\sum_{i=1}^{C}\frac{A_{\mathrm{P}}(i)}{C}, \tag{10}$$

$$F_1=2\times\frac{P\times R}{P+R} \tag{11}$$

式中:T_{P} 表示真阳性的数量,F_{P} 表示假阳性的数量,F_{N} 表示假阴性的数量,C 为类别总数。

3.3　对比实验

3.3.1　基准方法介绍

在对比实验中,选取以下 7 种主流方法作为基准,与 MDSVP-WIM 进行性能比较。

1) FCN[11]:全卷积神经网络,采用反卷积层实现上采样,能够以端到端方式进行图像语义分割。

2) PSPNet[12]:采用金字塔池化模块聚合多尺度上下文信息,有效融合全局与局部特征,提升场景解析能力。

3) UPerNet[13]:全景分割网络,通过共享的特征提取器和多个任务特定分支实现语义分割与实例分割的融合。

4) DANet[14]:双重注意力网络,基于双重注意力机制,分别在空间和通道维度上建模上下文依赖,增强模型对复杂场景的理解能力。

5) DMNet[15]:多尺度动态图卷积神经网络,为图像分割设计,引入动态尺度选择机制,根据图像特征自适应选择感受野大小,以更好地捕捉不同尺度的目标特征。

6) SAM v2[16]:通分万物模型 v2,在支持零样本迁移的基础上,引入可提示记忆解码器,提升小目标分割与边缘一致性,并通过缓存历史帧键值对,增强跨帧间目标表示的稳定性。

7) DC-Swin v2[17]:基于 Swin Transformer 结构,嵌入动态上下文窗口机制,可根据目标密度自适应调整窗口大小与步长,在保持线性计算复杂度的同时扩大感受野,并结合局部-全局双路径编码提升特征表达能力。

3.3.2　定量分析

本实验围绕检测与识别两项核心任务,评估各方法在仓库盘点场景下的性能表现。7 种对比方法在 WareSegNet 和 SpatioTrack-360 数据集上的评估指标如表 1 所示。

由表 1 可知,FCN 在检测任务中性能较差,其 F1 值仅为 0.31 和 0.33。这主要由于 FCN 在多次降采样过程中丢失了大量细节信息,对大范围上下文的理解能力不足,难以充分提取目标的判别性特征。PSPNet 检测任务相较于 FCN 在误检方面有所改善,且 mAP 在两个数据集上分别为 0.62 和 0.69,但其 F1 值仍偏低,分别为 0.28 和 0.32。在识别任务中,PSPNet 精确度为 0,表明该方法未能正确识别任何目标物品类别。尽管 PSPNet 通过金字塔池化模块扩展了感受野并捕获更多高维特征,但在训练过程中对不确定样本的错误归类导致过拟合问题。因此,尽管其平均精确度高于 FCN,但精确度和 F1 值反而低于 FCN。

UPerNet 作为一种性能优异的语义分割方法,在多任务学习与细节信息保留方面表现突出,其编码器-解码器结构与特征金字塔网络有助于保留更多空间细节,从而提升分割精度。与 UPerNet 相比,DANet 在物品检测精确度上提升约 20%,其检测 mAP 分别达到 0.86 和 0.88,F1 值分别为 0.54 和 0.55,表明该方法在精确度与召回率之间取得了较好的平衡。

DMNet 在识别与检测精确度方面较 DANet 提升有限,但其 F1 值与平均精度均有显著提高。该方法能够动态调整网络结构以适应不同尺度目标,在检测任务中 F1 值分别达到 0.61 和 0.60。相比之下,SAM v2 在检测 mAP 上进一步提升约 8%,识别 mAP 提升约 4%;检测 F1 值提升约 30%,识别 F1 值提升约 15%。SAM v2 借助时空记忆-注意力机制,在实现多尺度特征融合的同时保持跨帧一致性,从而在动态仓库场景中实现精确度、召回率性能平衡。与 SAM v2 相比,DC-Swin v2 在检测 mAP 上略低约 2%,检测 F1 值下降约 2%,识别 F1 值下降约 3%。

表 1 定量实验结果

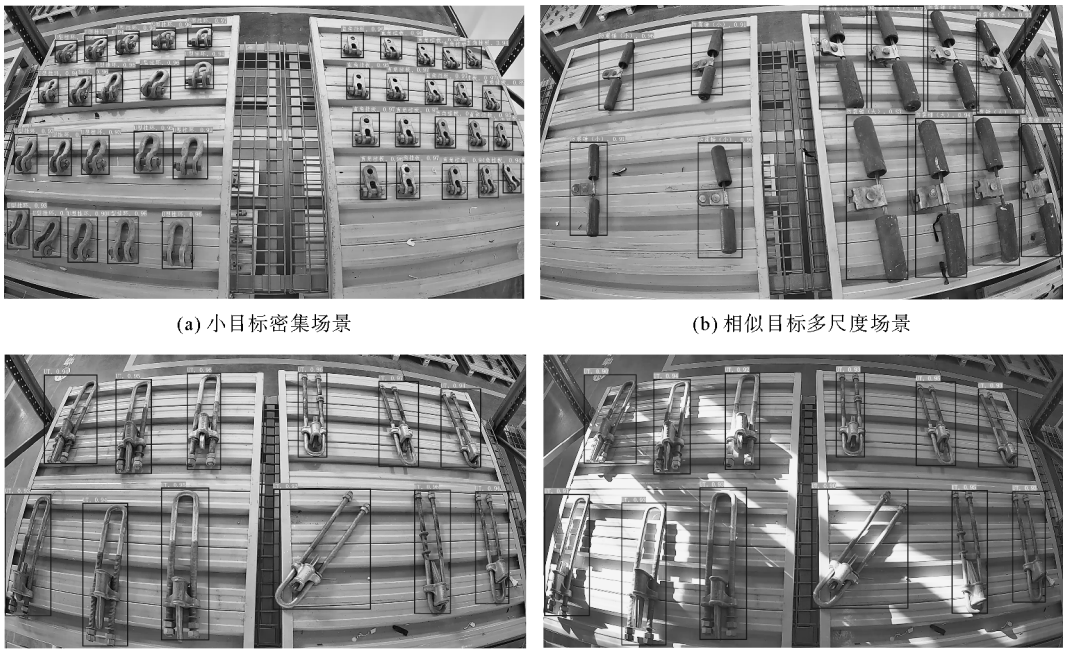
Table 1 Quantitative experiment results

方法	任务	WareSegNet				SpatioTrack-360			
		m_{AP}	R	F_1	P	m_{AP}	R	F_1	P
FCN	检测	0.53	1.00	0.31	0.19	0.55	1.00	0.33	0.20
	识别	0.65	1.00	0.08	0.04	0.59	1.00	0.12	0.07
PSPNet	检测	0.62	1.00	0.28	0.17	0.69	1.00	0.32	0.19
	识别	0.75	0	—	0	0.77	1.00	—	0.02
UPerNet	检测	0.71	1.00	0.32	0.19	0.74	1.00	0.32	0.19
	识别	0.83	1.00	0.03	0.01	0.85	1.00	0.02	0.02
DANet	检测	0.86	1.00	0.55	0.38	0.88	1.00	0.55	0.39
	识别	0.88	0	—	0	0.87	0	—	0
DMNet	检测	0.87	1.00	0.61	0.44	0.88	1.00	0.60	0.44
	识别	0.89	1.00	0.09	0.05	0.91	1.00	0.04	0.03
SAM v2	检测	0.95	0.98	0.92	0.87	0.94	0.97	0.93	0.89
	识别	0.93	1.00	0.77	0.63	0.94	1.00	0.83	0.71
DC-Swin v2	检测	0.93	0.97	0.90	0.84	0.92	0.95	0.89	0.83
	识别	0.91	1.00	0.67	0.50	0.92	1.00	0.80	0.57
MDSVP-WIM	检测	0.97	0.96	0.95	0.93	0.96	0.95	0.93	0.89
	识别	0.97	1.00	0.75	0.60	0.95	1.00	0.73	0.57

MDSVP-WIM 在检测与识别任务中均显著优于上述 7 种对比方法,展现出较好的综合性能。在检测任务中,MDSVP-WIM 在两个数据集上的 mAP 分别达到 0.97 和 0.96,远高于其他方法;在识别任务中,其 mAP 与 F1 值优势明显。需要指出的是,SAM v2 在识别精确度方面略高于本研究模型,这主要源于其零样本迁移能力与基于大规模预训练所带来的泛化性能,使其在相似品类物资的区分上表现更佳。而 MDSVP-WIM 在检测精度上的优势则源于其多维空间去噪模块对仓储环境干扰的有效抑制,以及跨帧跟踪模块对动态目标的稳定定位能力。尽管在识别精确度上略低于 SAM v2,但 MDSVP-WIM 在密集遮挡等复杂仓储场景中表现出更强的检测鲁棒性,更符合实际盘点任务对低误检率与高定位精度的要求。总之,MDSVP-WIM 在保持高召回率的同时,显著提升了检测与识别的精确度与平衡性,其低误报率和高鲁棒性使其成为当前任务中的最优解决方案。

3.3.3 定性分析

本研究进一步通过四类典型复杂场景定性分析验证 MDSVP-WIM 的有效性,包括小目标密集场景、相似目标多尺度识别场景、无光照干扰场景和强光照干扰场景。图 4 中方框标注了模型的目标识别结果,显示在各类复杂场景下其识别精确度均超过 90%。在图 4(a)所示的小目标密集场景中,MDSVP-WIM 基于 BSM 和轻量级编码-解码架构实现了精准识别。图 4(b)展示了在相似目标多尺度识别场景下的实验结果,模型展现出卓越的特征提取与识别能力。为验证模型在外部干扰条件下的鲁棒性,本研究特别设计了强光照变化的实验场景。图 4(c)和图 4(d)分别展示了无光照干扰和强光照干扰条件下的识别结果。实验表明,MDSVP-WIM 在常规条件下具有极高的识别精度。当受到光照等外部干扰时,模型通过跨帧位置跟踪、语义分割及多数投票机制保证识别精度。综上,该模型适用于复杂仓储环境,具有较高的实际应用价值。



(a) 小目标密集场景　(b) 相似目标多尺度场景

(c) 无光照干扰场景　(d) 强光照干扰场景

图 4　定性实验结果

Fig. 4　Results of qualitative experiments

3.3.4　消融实验

为分析各关键模块的贡献,本研究在 SpatioTrack-360 数据集上对 MDSVP-WIM 进行了消融实验,结果如表 2 所示。移除多维空间去噪模块的 MDSVP-WIM_A,在检测与识别任务中各指标均出现下降,表明该模块对提升本研究方法检测与识别性能具有正向作用,能够为后续处理提供更纯净的输入特征。在 MDSVP-WIM_B 中移除多维空间视觉感知模块后,检测与识别指标下降更为显著,说明该模块在提升本研究方法、平均精度方面至关重要,能够为本研究方法提供重要的像素级分类信息,缺少该模块会显著影响本研究方法在复杂场景中的表现。在 MDSVP-WIM_C 中移除跨帧跟踪-联合投票模块后,检测与识别性能出现小幅下降,但整体仍维持在较高水平,表明该模块有助于提升方法在动态场景下的稳定性。

表 2　消融实验结果

Table 2　Results of the ablation experiment

方法	任务	m_{AP}	R	F_1	P
MDSVP-WIM_A	检测	0.90	0.93	0.70	0.56
	识别	0.92	0.60	0.11	0.07
MDSVP-WIM_B	检测	0.89	0.91	0.65	0.51
	识别	0.92	1.00	0.11	0.07
MDSVP-WIM_C	检测	0.91	0.95	0.68	0.53
	识别	0.92	1.00	0.15	0.09
MDSVP-WIM	检测	0.96	0.95	0.92	0.89
	识别	0.95	1.00	0.71	0.57

4　结论

本研究针对仓储盘点任务中存在的小目标易漏检、密集目标易误检以及动态场景适应性差等问题,提出了 MDSVP-WIM。其中,多维空间去噪模块通过引入高斯混合模型实现动态更新背景,融合连续帧间像素变化序列与空间维度信息,有效去除环境干扰,为精准盘点奠定基础;多维视觉感知模块提升多尺度特征提取能力;跨帧跟踪-联合投票模块提升了动态目标的定位稳定性。实验结果表明该方法在检测精度和鲁棒性方面具备显著优势,尤其在复杂场景下表现优异。未来研究可进一步推进方法的轻量化设计,并拓展其在更多工业场景中的应用。

参考文献:

[1] ZHOU B L,ZHAO H,PUIG X,et al. Scene parsing through ADE20K dataset[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:633-641.

[2] ZACCARIA M,MONICA R,ALEOTT J. A comparison of deep learning models for pallet detection in industrial warehouses[C]//Proceedings of the 16th IEEE International Conference on Intelligent Computer Communication and Processing. Piscataway:IEEE Press,2020:417-422.

[3] 刘冰,郭李丽,刘如飞,等.全景影像快速加载与精确量测算法研究[J].山东科技大学学报(自然科学版),2024,43(4):57-66.
LIU Bing,GUO Lili,LIU Rufei,et al. Research on fast loading and accurate measurement algorithm for panoramic images[J]. Journal of Shandong University of Science and Technology(Natural Science),2024,43(4):57-66.

[4] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2014:580-587.

[5] LINDEBERG T. Scale invariant feature transform[J]. International Journal of Computer Vision,2012,91(3):1-42.

[6] 王紫芮,蒋德钧.基于超低延迟 SSD 的页交换机制关键技术[J].计算机研究与发展,2024,61(3):557-570.
WANG Zirui,JIANG Dejun. Key techniques of swapping mechanism based on ultra-low latency SSD[J]. Journal of Computer Research and Development,2024,61(3):557-570.

[7] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 17th IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2005,1:886-893.

[8] REN S Q,HE K M,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems,2015,28:91-99.

[9] JIANG P Y,ERGU D,LIU F Y,et al. A review of YOLO algorithm developments[J]. Procedia Computer Science,2022,199:1066-1073.

[10] GOODFELLOW I,BENGIO Y,COURVILLE A. Deep Learning[M]. Cambridge:MIT Press,2016.

[11] SHELHAMER E,LONG J,DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2015:3431-3440.

[12] ZHAO H S,SHI J P,QI X J,et al. Pyramid scene parsing network[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:2881-2890.

[13] XIAO T,LIU Y C,ZHOU B L,et al. Unified perceptual parsing for scene understanding[J/OL]//Lecture Notes in Computer Science,2018. DOI:10.1007/978-01228-1_26.

[14] FU J,LIU J,TIAN H J,et al. Dual attention network for scene segmentation[C]//Proceedings of the 31th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:3146-3154.

[15] LI W S,ZHANG X Y,PENG Y D,et al. DMNet:A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images[J]. IEEE Sensors Journal,2020,20(20):12190-12202.

[16] RAVI N,GABEUR V,HU Y,et al. SAM 2:Segment anything in images and videos[C]//Proceedings of the Thirteenth International Conference on Learning Representation. Open Directory,2025:571-615.

[17] LIU Z,HU H,LIN Y T,et al. Swin transformer V2:Scaling up capacity and resolution[C]//Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2025:12009-12019.

(责任编辑:傅 游)