中图法分类号:TP18;TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)03-0745-10

论文引用格式:Zhan R Y, Fan Y, Zhou L N, Xie Y B, Chen J X, Yang H Y, Huang D and Wang Y H. 2026. Dynamically distribution-aware quantization for diffusion models. Journal of Image and Graphics, 31(3):0745-0754(占瑞乙, 樊轶, 周丽娜, 谢宇宝, 陈佳鑫, 杨鸿宇, 黄迪, 王蕴红. 2026. 分布范围动态感知的扩散模型量化. 中国图象图形学报, 31(3):0745-0754)[DOI:10.11834/jig.250319]

分布范围动态感知的扩散模型量化

占瑞乙[1,2],樊轶[3],周丽娜[3],谢宇宝[3],陈佳鑫[1,2*],杨鸿宇[1,4],黄迪[2],王蕴红[1,2]

1. 北京航空航天大学虚拟现实技术与系统全国重点实验室,北京　100191; 2. 北京航空航天大学计算机学院,北京　100191; 3. 中国电子科学研究院,北京　100041; 4. 北京航空航天大学人工智能学院,北京　100191

摘　要: **目的**　扩散模型在图像生成、艺术创作等领域具有广泛的应用前景。然而,现有扩散模型存在存储开销大、推理时延长等问题。已有工作通过模型量化减少存储及推理时间消耗,但在量化过程中面临诸多挑战。一方面,噪声估计网络结构复杂,不同模块对量化误差的敏感性存在显著差异;另一方面,其多采样时间步特性使量化模型在推理时存在误差累积问题。现有量化方法未能考虑模块间量化差异和多时间步推理特性,因而量化模型性能较全精度模型仍有较大差距。为此,设计了一种分布范围动态感知的训练后量化方法。**方法**　在校准过程中,首先基于模块的量化误差对量化参数进行选择;再基于输入激活值的分布范围对不同网络模块中的量化参数进行微调;最后,依次计算全精度扩散模型与量化模型在每个采样时间步下估计噪声之间的均方误差,抑制多采样时间步推理过程中的累积量化误差。**结果**　在CIFAR-10(Canadian Institute for Advanced Research)、LSUN-Bedroom(large-scale scene understanding)与LSUN-Church这3个公开数据集上采用DDIM(denoising diffusion implicit model)与LDM(latent diffusion model)两种扩散模型,使用不同量化位宽(W8A8,W4A8,W6A6),并与主流扩散模型量化方法进行对比。实验结果表明,所提方法将DDIM量化至W6A6时,在CIFAR-10数据集上取得了(IS:9.40,FID:4.61)的性能表现,与全精度DDIM性能(IS:9.12,FID:4.12)相近;相较于对比量化方法,所提方法将IS(inception score)值提高了0.34,FID(Frechet inception distance score)值降低了1.96。在LSUN-Church数据集上,相较于已有工作,所提方法量化LDM至W8A8时,将FID值降低了0.34。此外,所提方法与现有量化方法结合均能进一步取得一致性提升。**结论**　本文所提出的分布范围动态感知的训练后量化方法同时考虑了扩散模型的多时间步推理特性与模块间量化差异,显著提升了量化扩散模型在低激活比特位宽上的性能,且该方法可作为即插即用模块与已有量化方法结合取得更优的性能。

关键词: 图像生成;扩散模型(DM);模型压缩;推理加速;模型量化;训练后量化

Dynamically distribution-aware quantization for diffusion models

Zhan Ruiyi[1,2], Fan Yi[3], Zhou Lina[3], Xie Yubao[3], Chen Jiaxin[1,2*], Yang Hongyu[1,4], Huang Di[2], Wang Yunhong[1,2]

1. State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing　100191, China; 2. School of Computer Science and Engineering, Beihang University, Beijing　100191, China; 3. China Academy of Electronics and Information Technology, Beijing　100041, China; 4. School of Artificial Intelligence, Beihang University, Beijing　100191, China

收稿日期:2025-07-16;**修回日期**:2025-09-12;**预印本日期**:2025-09-19

*通信作者:陈佳鑫 jiaxinchen@buaa.edu.cn

基金项目:国家自然科学基金项目(62202034,82441024);北京市自然科学基金项目(4242044,L259044);中央高校基本科研业务费专项资金资助(JKF-2025073186377)

Supported by: National Natural Science Foundation of China(62202034, 82441024); Beijing Municipal Natural Science Foundation (4242044, L259044); Fundamental Research Funds for the Central Universities(JKF-2025073186377)

Abstract: **Objective** In recent years, the diffusion model has become a promising alternative to conventional generative models including generative adversarial network and variational autoencoder due to the high quality and diversity of its generated images, as well as its stable training process. It has a wide range of applications such as super-resolution, graph generation, and image restoration. Generally, the generation process of diffusion models involves gradually adding Gaussian noise to image data and then iteratively removing the noise step by step through a noise estimation network. As this process typically takes hundreds or even thousands of steps to find sampling trajectories for denoising, the diffusion model usually requires tremendous storage overhead and inference time cost. The high computational complexity of diffusion models is mainly attributed to the following two issues. First, generating a single image requires hundreds or even thousands of denoising steps, which involve repeatedly executing the estimation network. Second, the estimation network alone involves significant computational cost due to the high network complexity. Although many approaches have been proposed to reduce the number of estimation steps, balancing the number of steps with the quality of generated images remains an unsolved problem. In this paper, we address the second issue, i. e. , accelerating the U-Net-based noise estimation network. Quantization has been a promising solution by converting the floating-point weights and activations to low-bit-width integers. Typically, quantizing a full-precision model to 8-bit can theoretically accelerate the inference process by about 2. 2 times, while further reducing to 4-bit achieves an additional 59% improvement. However, directly applying the quantization methods designed for general-purpose to diffusion models often yields poor performance, as the diffusion models using the same network to denoise inputs with different distributions at various timesteps, which is neglected by most existing approaches. Some works incorporate multi-timestep calibration into the quantization process or focus on the temporal characteristics within the estimation network to mitigate the impact of multi-timestep distribution. Despite these advancements, a significant performance drop persists when models are quantized to bit-widths lower than 8-bit using post-training techniques. To overcome the above drawbacks, we analyze quantization errors within the estimation network across different timesteps. Our analysis reveals that the reconstruction granularities employed in quantization are often inappropriate for diffusion models, leading to pronounced discrepancies among quantized modules. Moreover, we identify substantial quantization errors in specific modules characterized by a wide range of activation distributions across timesteps, leading to diminished performance when quantizing to lower bit-widths. **Method** In this study, we propose a novel method called temporal distribution-aware quantization for diffusion models to deal with the uneven distribution of internal quantization errors within the network and the accumulation of external quantization errors over multiple sampling timesteps. We evaluate the significance of different quantized modules by analyzing relative quantization errors and select quantization parameters for fine-tuning on the basis of the significance score. Meanwhile, the fine-tuning strength is dynamically set according to the range of inputs. Furthermore, to reduce the accumulation of quantization errors across multiple sampling timesteps in diffusion models, we calculate the mean squared error between output of the full-precision model at each sampling timestep with that of quantized models. **Result** Our method outperforms the state-of-the-art compared approaches, achieving an improvement of 0. 34 (9. 40 vs. 9. 06) in inception score (IS) and 1. 96 (4. 61 vs. 6. 57) in Frechet inception distance score (FID) on CIFAR-10, respectively. The quantized LDM by our method outperforms TFMQ-DM by an FID reduction of 0. 34 on the LSUN-Church benchmark. In the meantime, the computational cost for our method remains consistent with that of baseline methods. The results demonstrate improved performance across various datasets and distinct models. Additionally, our method can be incorporated into existing quantization methods as a plug-and-play module. **Conclusion** In this study, we propose a novel temporal distribution-aware quantization method for diffusion models, which reduces the accumulating quantization errors arising from the substantial distribution variations across distinct layers and blocks at different timesteps. Our method is a plug-and-play method that is applicable to existing quantization approaches. Extensive experiments on various benchmarks clearly demonstrate the effectiveness of our method.

Key words: image generation; diffusion model (DM); model compression; inference acceleration; model quantization; post-training quantization

0 引 言

近年来，图像生成技术已广泛应用于商业与日常社会生活。对于低分辨率、受损和缺失的图像，可以使用图像生成技术对其进行编辑、补全和修复，从而提升图像质量。此外，图像生成还可用于生成人脸图像、病灶图像等，对计算机视觉等下游任务小规模数据集进行扩充，通过增加训练数据的多样性提高深度学习模型的泛化性能。在图像生成模型中，扩散模型（diffusion models）不仅在生成图像的质量及多样性（Ho 等，2020；Song 等，2021）上优于生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2020）、变分自编码器（variational autoencoder，VAE）（Kingma 和 Welling，2013），其训练时表现也更加稳定。这使得扩散模型成为广受关注的生成模型之一，并在超分辨率（Rombach 等，2022）、图生成（Niu 等，2020）、图像翻译（Sasaki 等，2021）、视频生成（郑天鹏 等，2025）以及视频隐写（李文琪等，2025）等领域得到广泛应用。

在使用经典的基于概率的扩散模型（denoising diffusion probabilistic models，DDPM）生成图像时，模型会在真实数据中逐渐加入高斯噪声，然后通过噪声估计网络逐步预测并去除高斯噪声以得到生成的数据（Ho 等，2020）。在上述马尔可夫过程中，噪声估计网络需要进行上千轮噪声估计，较 GAN 与 VAE 等传统生成模型而言，扩散模型的存储、计算与时间开销较为显著。例如，基于 DPM-Solver（high-order solver for diffusion probabilistic models）（Lu 等，2022）使用 Stable Diffusion（Rombach 等，2022）生成一幅分辨率为 512 × 512 像素的图像需要在 RTX 3090 上运行 1 s，并占用 16 GB 内存与 10 GB 显存（He 等，2023）。因此，为了将扩散模型应用至更多计算资源受限的场景中，需要加速扩散模型的推理过程。然而，以往的工作集中于寻找更优的采样路径、减少噪声估计轮次，忽视了噪声估计网络自身的计算与存储开销。

对于噪声估计过程的优化，目前的工作主要使用量化、蒸馏和剪枝等方法简化扩散模型中基于 U-Net（Ronneberger 等，2015）的噪声估计网络。其中，模型量化是主流方法（state-of-the-art，SOTA）之一，其通过将深度学习模型中的权重与激活从浮点数变为低位宽表示的整数以减小计算量及存储空间。然而，由于通用模型量化方法未考虑扩散模型特点，将其直接应用于扩散模型量化，得到的量化模型性能表现较差，与全精度模型差距较大。因此，部分工作针对扩散模型设计了特定的量化方法。

在对扩散模型进行量化时，考虑到其推理开销主要源于噪声估计网络，现有工作通常选择对该网络进行量化。同时，由于模型训练所需开销较大，相关研究中又以训练后量化方法为主，此类方法所需资源少、可移植性强且量化速度快。

具体地，PTQ4DM（post-training quantization on diffusion models）（Shang 等，2023）、Q-Diffusion（quantizing diffusion models）（Li 等，2023）通过对校准数据集分布进行分析，指出均匀采样来自不同时间步的图像可以在校准过程中减小量化误差。同时，Q-Diffusion 进一步提出了针对基于 U-Net 结构噪声估计网络的优化策略（Li 等，2023），指出 U-Net 网络中的残差结构会因量化程度不同而放大量化误差，并据此设计了通道分离的量化方案，随后在 CIFAR-10（Canadian Institute for Advanced Research）（Krizhevsky 和 Hinton，2009）、LSUN（large-scale scene understanding）（Yu 等，2015）数据集上进行了 W8A8 与 W4A8 量化，取得了与全精度模型近似的生成效果。PTQD（accurate post-training quantization for diffusion models）（He 等，2023）考虑到量化扩散模型时产生的量化误差中含有高斯噪声，将其与扩散模型中的高斯噪声融合，通过调整高斯噪声的方差减小量化误差。此外，PTQD 还基于信噪比衡量量化效果，确定不同时间步下的最佳量化位宽，并采用混合精度量化策略。TDQ（temporal dynamic quantization for diffusion models）（So 等，2023）使用一个 3 层感知机将采样时间编码信息映射到微调参数用于修正量化参数；EfficientDM（efficient quantization-aware fine-tuning of low-bit diffusion models）（He 等，2024）采用 QLoRA（efficient finetuning of quantized llms）（Dettmers 等，2023）的设计在量化校准过程中对量化参数进行微调。TFMQ-DM（temporal feature maintenance quantization for diffusion models）（Huang 等，2024）构建了针对采样时间步的特定量化模块用于矫正量化时间信息编码及嵌入模块时产生的误差；APQ-DM（towards accurate post-training quantization for diffusion models）（Wang 等，2024）通过对不同采样时间步

进行分组，设置不同采样时间步间的共享参数，为量化参数寻找合适的映射范围以优化模型性能。

然而，现有工作在量化扩散模型时忽视了量化噪声估计网络时不同网络模块间的复杂度差异，且未有效缓解模型推理时多采样时间步引发的累积量化误差。

针对上述问题，本文提出一种分布范围动态感知的扩散模型量化方法，对扩散模型进行量化压缩时分别关注扩散模型网络内部量化误差与外部随时间步的累积量化误。主要贡献为：1）减小了扩散模型量化后噪声估计网络不同模块内的量化误差；2）减小了扩散模型量化后随采样时间累积的量化误差；3）在CIFAR-10、LSUN-Bedroom及LSUN-Church多种分辨率的数据集上对DDIM（denoising diffusion implicit model）与LDM（latent diffusion model）两种不同参数规模的扩散模型进行了模型量化实验，并与同类量化方法对比证明了提出方法的有效性；4）将本文方法作为即插即用模块应用于同类量化方法进行扩展性实验，验证了本文方法的高可扩展性。

1 相关概念

本节对扩散模型量化过程涉及的相关概念进行介绍，主要分为扩散模型与模型量化两部分。

1.1 扩散模型

基于概率的扩散模型通过使用马尔可夫过程生成图像，在该过程中分T个时间步依次对原始数据$\boldsymbol{x}_0 \sim \boldsymbol{q}(\boldsymbol{x})$加入高斯噪声，其中$\boldsymbol{q}(\boldsymbol{x})$表示原始数据符合的数据分布，加入噪声后依次得到模糊程度不同的图像$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T$。

$$\boldsymbol{q}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = N(\boldsymbol{x}_t; \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{1}$$

$$\boldsymbol{q}(\boldsymbol{x}_{1:T}|x_0) = \prod_{t=1}^{T} \boldsymbol{q}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{2}$$

式中，$\beta_1, \beta_2, \cdots, \beta_T \in (0,1)$用于控制每个时间步添加的高斯噪声的强度，$N$表示正态分布。最终得到的加噪图像$\boldsymbol{x}_T$即可近似为高斯噪声。因此，生成图像时即可随机采样一个高斯噪声$\boldsymbol{x}_T \sim N(0, \boldsymbol{I})$作为起点，逐步执行逆马尔可夫过程去除噪声以生成图像。然而条件概率$\boldsymbol{p}_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$是未知的，所以扩散模型通过学习噪声估计网络拟合该概率分布，具体为

$$\boldsymbol{p}_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = N(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_{\theta,t}(\boldsymbol{x}_t), \tilde{\beta}_t \boldsymbol{I}) \tag{3}$$

在经过重参数化后，均值$\tilde{\boldsymbol{\mu}}_{\theta,t}(\boldsymbol{x}_t)$与方差$\tilde{\beta}_t$为

$$\tilde{\boldsymbol{\mu}}_{\theta,t}(\boldsymbol{x}_t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\delta}}_{\theta,t}\right) \tag{4}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5}$$

式中，$\alpha_t = 1-\beta_t$，$\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$，$\hat{\boldsymbol{\delta}}_{\theta,t}$表示在第$t$个时间步使用参数$\theta$生成的噪声图像。最终，$\boldsymbol{x}_{t-1}$可以表示为

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\delta}}_{\theta,t}\right) + \sigma_t \boldsymbol{z} \tag{6}$$

式中，$\boldsymbol{z} \sim N(0, \boldsymbol{I})$，$\sigma_t$为与采样方式相关的常数。

1.2 模型量化

在深度神经网络中，参数以32位全精度浮点数或16位半精度浮点数进行存储与计算。对此，模型量化过程通过压缩网络中权重与激活值的数据位宽以减小模型大小。同时，由于量化方法通常会将数据压缩并映射至8比特、4比特甚至更低比特数的整数，全精度模型中的浮点运算也被转换为相应的整数运算，因此量化模型的推理速度也会得到提升。

在使用线性映射对全精度浮点数模型进行量化时，一般会对该模型的权重和激活进行量化，将其转换为低比特位宽的整数表示$\bar{\boldsymbol{W}}$，可以表示为

$$\bar{\boldsymbol{W}} = Clip\left(R\left(\frac{\boldsymbol{W}}{S}\right) + Z, C_{\min}, C_{\max}\right) \tag{7}$$

式中，$\boldsymbol{W}$代表模型参数，S代表缩放因子（scaling factor），Z代表偏移量（zero point），$C_{\min}$与$C_{\max}$表示映射范围的上下界，其范围也称为量化范围，$R(\cdot)$与$Clip(\cdot)$分别表示取整和映射函数。

2 问题描述与方法设计

本节主要介绍本文提出的分布范围动态感知的训练后量化框架。该方法分为分布范围感知的量化参数微调与累积误差动态抑制两个模块，其分别针对扩散模型中噪声估计网络内部量化误差差异性与外部量化误差随采样时间步累积问题进行了分析与改进。

2.1 分布范围感知的量化参数微调

在量化DDIM中噪声估计网络的单个采样时间步内，不同量化粒度下的模块、网络中不同位置模块的激活值范围及量化误差存在较大差异。分析其原因，一方面，基于BRECQ（block reconstruction quanti-

zation)(Li 等,2021)的训练后量化方法为网络中不同量化粒度的模块使用全局统一的量化超参数配置,而训练后量化过程中使用的校准数据样本较少,一般只有200～5 000幅图像样本,其数量远小于预训练模型在训练时所使用的数据集样本数,这意味着在同一个量化网络的不同模块中可能同时存在过拟合问题与欠拟合问题。

图1展示了量化参数校准过程中不同模块内激活量化器的量化重建情况,表明不同模块的量化损失收敛时所需的重建轮次并不相同。这意味着在统一的量化重建超参数设置下,网络中将既有处于过拟合状态的激活量化器也有处于欠拟合状态的激活量化器。为了进一步说明该问题,本文基于BRECQ采用不同的量化重建超参数,对网络中部分处于欠拟合状态的激活量化器进行观察,发现其量化损失存在至多50%的下限,图2展示了使用本文方法量化 DDIM 至 W4A6 时模块 29(对应模块为 up. 1. upsample. conv)的激活值均方误差。对此,本文通过设定误差阈值以感知量化重建误差,缓解了过拟合与欠拟合问题对量化重建的影响。对于噪声估计网络内部的量化误差,本文向网络中的激活量化器添加部分微调参数,在BRECQ量化重建的过程中利用校准数据实现逐模块的量化参数微调,并且在微调结束后将其融入量化参数,避免增加额外的量化参数。

对于网络中不同模块量化重建收敛程度不同的问题,基于下式,假设激活值分布范围 $\boldsymbol{X}_{\text{range}}$ 及相对量化误差 Q_{err} 与模块欠恢复程度 η 成正相关,即激活值分布范围越广,在量化时产生的截断误差越大,此时截断误差占量化误差比重较舍入误差更大,从而使得此类分布范围与模块中的全精度激活值被量化至低比特时产生的误差分布相似。具体为

$$\eta \propto \boldsymbol{X}_{\text{range}}, Q_{\text{err}} \tag{8}$$

此外,综合考虑重建时部分模块的激活值分布范围较小但因模块功能差异也会产生较大的相对量化误差,本文也将相对量化误差纳入考虑。因而,本文选择使用量化重建时的激活值分布范围与量化相对误差对欠恢复程度 η 进行评估,并向其中 $\boldsymbol{X}_{\text{range}}$ 排列前 β%或相对量化误差大于 γ 的模块插入微调参数对量化参数进行微调,加快其量化参数的收敛速度,此处取 $\beta = 10$,$\gamma = 20$。

一般而言,在进行量化模拟时会采用如式(7)对模拟量化误差进行模拟,其通过量化与反量化的过程向深度学习模型中引入量化时产生的截断误差与舍入误差,其可以进一步表示为量化与反量化两个计算式,具体表述为

$$\boldsymbol{X}_{\text{quant}} = Clip\left(R\left(\frac{\boldsymbol{X}_{fp}}{S}\right) + Z, C_{\min}, C_{\max}\right) \tag{9}$$

$$\boldsymbol{X}_{\text{dequant}} = (\boldsymbol{X}_{\text{quant}} - Z) \cdot S \tag{10}$$

首先,使用缩放因子 S 对全精度数据 $\boldsymbol{X}_{fp}$ 进行缩放,再通过取整函数 $R(\cdot)$ 进行舍入;在进一步加入偏移量 Z 后,使用映射函数 $Clip(\cdot)$ 将数据映射至低比特数据范围。所以,通过应用式(9)即可得到量化器运行时的理论量化结果 $\boldsymbol{X}_{\text{quant}}$。随后,通过量化参数 S 与 Z 即可将运行时模拟的量化数据 $\boldsymbol{X}_{\text{quant}}$ 还原为全精度数据 $\boldsymbol{X}_{\text{dequant}}$,用以代替全精度模型在此处的值 $\boldsymbol{X}_{fp}$ 继续传递至后续模块。而在插入微调参数 F_S 与 F_Z 对量化参数分别进行微调后,微调运算记做 $F(\cdot)$,量化计算式可以表示为

$$F(\boldsymbol{X}_{\text{quant}}) = Clip(F(\boldsymbol{X}_{\text{int}}), C_{\min}, C_{\max}) \tag{11}$$

$$F(\boldsymbol{X}_{\text{dequant}}) = (F(\boldsymbol{X}_{\text{quant}}) - Z - R(F_Z)) \cdot S \cdot F_S \tag{12}$$

如图2所示,在W4A8与W4A6的量化位宽下,对DDIM进行分布范围感知的量化参数微调所得到的误差曲线W4A6-微调表现优于W4A6误差曲线,即在W4A6量化位宽设置下进行微调时的量化误差更小,其结果更接近W4A8的结果,说明该方法在将激活值量化至更低比特时仍能保持较低的量化损失与较好的重建质量,这进一步证明了该方法在减小单一时间步内噪声估计网络中不同模块量化误差时的有效性。

2.2 累积误差动态抑制

式(3)表明,噪声估计网络中每个采样时间步的输入依赖于上一个时间步的输出结果,这意味着随着采样时间步的迭代,量化后的噪声估计网络产生的量化误差也会进行累积,从而不断偏离全精度网络的输出结果。以量化至W4A8的DDIM使用100个采样时间步进行推理为例,其量化误差逐渐增加且在40～50个采样时间步后达到饱和。因此,对于多采样时间步下的累积量化损失,本文基于时间步设计了逐采样时间步的后处理方法。在对量化网络进行后处理时,使用全精度网络作为参考,在每个时间步进行噪声估计时,将量化模型的输出结果与全精度网络的输出结果计算均方误差(mean squared

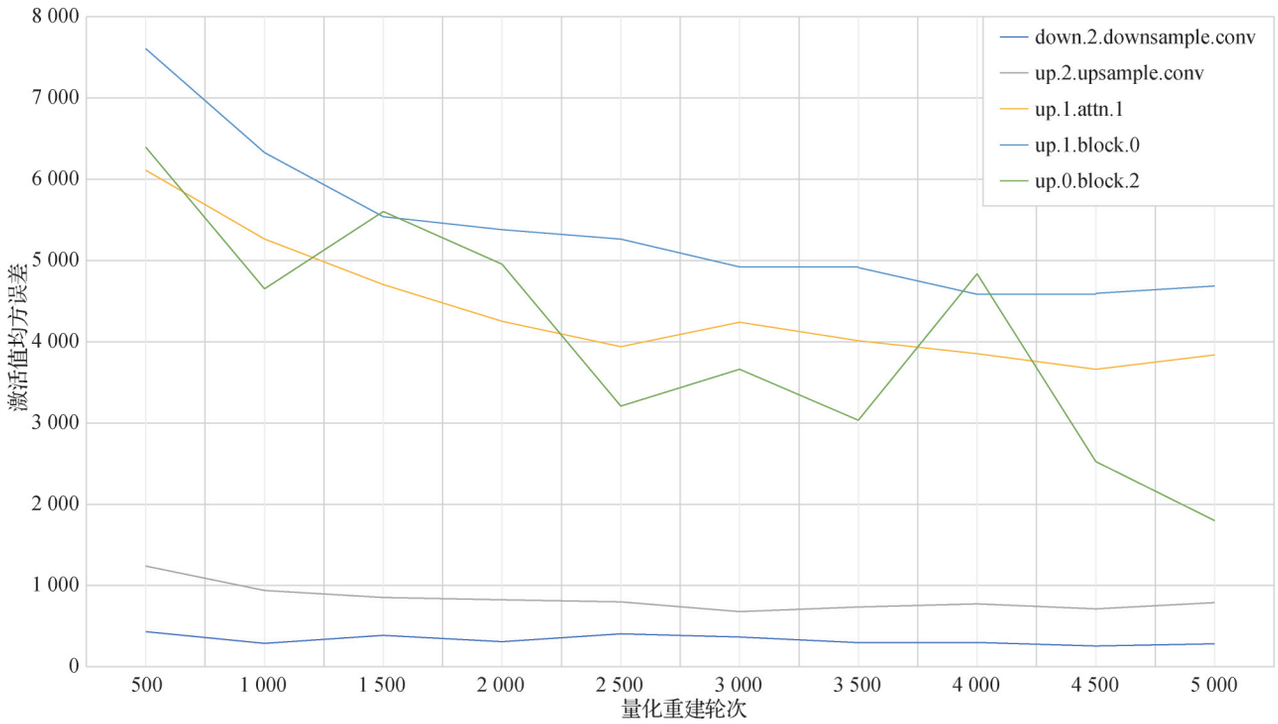

图1　量化不同模块时收敛所需的重建轮次对比

Fig. 1　Comparison of the reconstruction epochs required for convergence when quantizing different modules

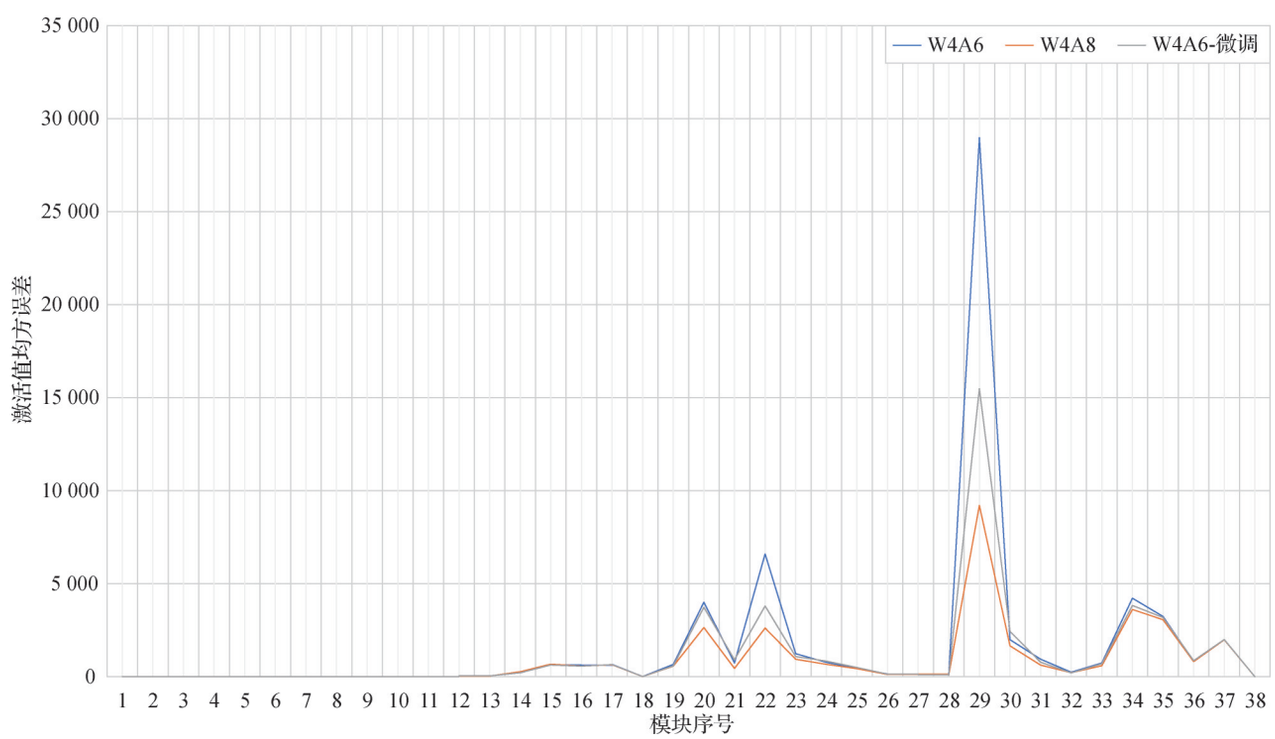

图2　不同量化位宽下激活值量化误差对比

Fig. 2　Comparison of the MSE of quantized activation under different quantization bit-widths settings

error，MSE)损失，并进行反向传播优化量化参数，以达到逐时间步矫正量化模型输出的目的，减少随采样时间累积的量化误差，具体为

$$L_t = \| \boldsymbol{n}_{\mathrm{fp},t} - \boldsymbol{n}_{\mathrm{quant},t} \|^2 \tag{13}$$

式中，L_t表示时间步t时对应的损失，$\boldsymbol{n}_{\mathrm{fp},t}$与$\boldsymbol{n}_{\mathrm{quant},t}$分别表示量化前后的噪声估计网络在时间步$t$时预测

的噪声图像。

3 实验结果与分析

本节主要对本文提出的面向扩散模型的训练后量化方法的性能进行评估，先后对实验设置与实验结果进行介绍。

3.1 实现细节

本文对基于CIFAR-10、LSUN数据集训练得到的预训练扩散模型进行量化。为了评估扩散模型的图像生成效果，本文使用PyTorch中的torch_fidelity库计算IS（inception score）（Salimans等，2016）与FID（Frechet inception distance score）（Heusel等，2017）作为其评价指标。

本文主要基于Q-Diffusion框架进行实验，DDIM在单卡NVIDIA GeForce RTX 3080 Ti上运行，LDM在单卡NVIDIA GeForce RTX 3090上运行。选择在DDIM上进行所有技术方案的消融、迁移等实验，并在LDM上以多种量化位宽进行扩展实验，验证本文提出的量化方法。

在对噪声估计网络进行量化时，对权重部分使用AdaRound量化器（Nagel等，2020）进行量化，对激活部分使用均匀量化器（Krishnamoorthi，2018）进行量化。校准量化参数时，使用5 120幅均匀采样自20个时间步（每个采样时间步各256幅图像）的图像作为校准数据，校准时批处理大小设置为32。在进行量化重建时，本文基于BRECQ中提出的训练后量化框架实现，对网络的残差模块、注意力模块使用块粒度重建，对其余部分使用层粒度进行重建。其中，权重重建轮次设置为40 000，激活重建轮次设置为5 000，初始学习率设置为0.000 4，使用Adam优化器作为优化器。在生成图像时，通过DDIM采样方法使用100个采样时间步按批处理大小32生成50 000幅图像。

3.2 实验结果

为评估本文方法的模型量化效果，本节基于Q-Diffusion应用本文提出的量化方法，并将其模型量化结果与现有的其他在相同数据集、相同量化位宽上实现的面向扩散模型的量化方法进行对比。

表1展示了量化DDIM在CIFAR-10数据集、量化LDM在LSUN-Church与LSUN-Bedroom数据集上的实验结果。由结果可知，在将LDM量化至W4A8时，本文量化方法应用至Q-Diffusion后的性能表现较TFMQ-DM有一定差距，但其FID分数均优于原始方法Q-Diffusion。对此，本文在3.4节中对TFMQ等量化方法进行进一步扩展性实验验证。此外，在W8A8量化位宽下，本文方法的性能表现均达到最优。

表1 本文方法与主流训练后量化方法性能对比

Table 1 Comparison of the performance between method proposed with SOTA quantization methods

方法	位宽	CIFAR-10		LSUN-Church	LSUN-Bedroom
		IS ↑	FID ↓	FID ↓	FID ↓
全精度	W32A32	9.12	4.22	4.12	2.98
PTQ4DM	W8A8	9.31	14.18	4.80	4.75
Q-Diff	W8A8	9.48	3.75	4.41	4.51
TDQ	W8A8	8.85	5.99	8.17	–
PTQD	W8A8	–	–	4.89	3.75
APQ-DM	W8A8	9.07	4.24	4.02	3.88
TFMQ-DM	W8A8	9.07	4.24	4.02	**3.14**
本文	W8A8	**9.67**	**3.38**	**3.68**	3.73
PTQ4DM	W4A8	–	–	4.97	20.72
Q-Diff	W4A8	9.12	4.93	4.66	6.40
PTQD	W4A8	–	–	5.10	5.94
TFMQ-DM	W4A8	9.13	4.78	**4.14**	**3.68**
本文	W4A8	**9.49**	**4.08**	4.17	3.71

注：加粗字体表示对应位宽及数据集的最优结果。“–”表示该方法未在相应数据集进行测试。“↑”表示值越高越好，“↓”表示值越低越好。

由于同类量化方法在进行量化时未考虑激活值分布范围差异性为模型量化带来的负面影响，使用其他方法将激活量化至8比特以下均会产生较大性能损失，而本文提出的方法基于激活值分布范围进行优化，将其纳入量化参数的微调方案中，对此，本文在W6A6量化位宽下对DDIM进行了进一步量化实验，其结果如表2所示，表明本文方法仍可获得较好的模型性能表现。

3.3 消融实验

本文在Q-Diffusion的基础上实现了分布范围感知的量化参数动态微调方法与累积误差抑制方法，表3为在DDIM上进行不同量化位宽消融实验结果，

表2　量化DDIM至W6A6时的实验结果

Table 2　Results for quantized DDIM on W6A6

方法	位宽	CIFAR-10	
		IS ↑	FID ↓
全精度	W32A32	9.12	4.22
Q-Diff	W6A6	8.76	9.19
APQ	W6A6	9.06	6.57
TFMQ-DM	W6A6	8.84	9.59
本文	W6A6	**9.40**	**4.61**

注：加粗字体表示各列最优结果。“↑”表示值越高越好，“↓”表示值越低越好。

表3　量化方法消融实验

Table 3　Ablation for proposed quantization methods

方法	位宽	CIFAR-10	
		IS ↑	FID ↓
全精度	W32A32	9.12	4.22
基准方法	W8A8	9.38	3.75
+ DA	W8A8	9.48	3.85
+ PT	W8A8	9.46	3.72
+ DA + PT	W8A8	**9.67**	**3.38**
基准方法	W4A8	9.12	4.93
+ DA	W4A8	9.47	4.11
+ PT	W4A8	9.18	**3.96**
+ DA + PT	W4A8	**9.49**	4.08
基准方法	W6A6	8.76	9.19
+ DA	W6A6	8.96	6.75
+ PT	W6A6	9.31	8.10
+ DA + PT	W6A6	**9.40**	**4.61**

注：加粗字体表示对应位宽及数据集的最优结果。“↑”表示值越高越好，“↓”表示值越低越好。

表3中基准方法指代Q-Diffusion方法。实验结果表明，在W8A8、W6A6与W4A8三种量化位宽下分布范围感知的量化参数动态微调方法与累积误差抑制方法效果均优于基准方法。表中，+DA（distribution aware）表示应用了分布范围感知的量化参数动态微调方法，+PT（parameters finetuning）表示应用了累积误差抑制方法。

如表3所示，在W8A8量化位宽下，分别应用DA与PT方法后，其IS指标均有提升，然而其FID指标较基准方法无显著差异，但在两方法同时使用时，IS与FID指标均出现了明显提升，其原因在于量化器原有量化参数表征能力有限，其性能出现瓶颈，而在进行多维度优化时可以帮助原量化器跳出局部最优，找到更优的量化器参数。在W4A8与W6A6量化位宽下，分别使用DA与PT方法在IS与FID指标上均可获得较基准方法更优的结果，其同时使用后的模型性能也较基准方法更优。然而，在W4A8量化位宽下仅使用PT方法时获得的FID优于同时使用DA与PT的结果，其原因在于使用PT方法对参数进行微调时，未对量化器中的量化参数与DA中插入的微调参数进行进一步区分，而这两者的数值规模及对量化误差的敏感程度不同，从而导致两种方法混合使用后性能有所下降。

3.4　扩展性验证

表4中的方法均基于相关工作源码复现，+本文方法表示在其量化方法的基础上进一步应用本文方法。然而由于不同方法的适用范围及可供参考的实验设置不同，本组实验使用了不同的量化位宽进行量化。本组实验结果表明将本文方法应用于不同量化方法均可获得性能提升，从可行性上验证了该框架可以作为插件进一步应用于不同的扩散模型量化方法，意味着本文方法可以在已有量化方法的基础上进一步减小量化误差以提高量化模型性能表现。

表4　本文方法的扩展性验证

Table 4　Demonstration of the expansibility of proposed method

方法	位宽	CIFAR-10	
		IS ↑	FID ↓
全精度模型	W32A32	9.12	4.22
TDQ	W8A8	9.58	3.77
+ 本文	W8A8	**9.58**	**3.47**
Q-Diff	W4A8	9.12	4.93
+ 本文	W4A8	**9.43**	**4.02**
EfficientDM	W4A8	9.30	4.67
+ 本文	W4A8	**9.32**	**4.18**
TFMQ-DM	W4A8	9.03	7.68
+ 本文	W4A8	**9.09**	**6.79**

注：加粗字体表示对应位宽及数据集的最优结果。“↑”表示值越高越好，“↓”表示值越低越好。

4 结 论

基于扩散模型的图像生成方法应用广泛，但模型存在复杂度高、存储开销大以及推理速度慢等问题。易于快速实现与迁移的训练后量化方法在量化扩散模型时表现出了较大潜力。对此，本文研究了面向扩散模型的量化方法，并对目前已有的训练后量化方法进行了改进，结合扩散模型中噪声估计网络的特点，提出了分布范围动态感知的训练后量化方法，减小了量化模型在推理时产生的累积误差。实验结果表明，本文方法在相同量化位宽下基本显著优于目前最优的训练后量化压缩方法，且可作为即插即用模块与其他量化方法结合，进一步提升低比特扩散模型的性能表现，具有较高的可扩展性。

具体而言，针对扩散模型中噪声估计网络模块间激活值范围及量化误差存在较大差异的问题，本文分析其原因为不同模块在量化重建时的收敛程度不同，继而提出分布范围感知的量化参数微调方法，基于模块间量化误差及输入范围分布评估模块的欠恢复程度，定位量化误差来源，并在校准过程中基于多源信息对量化因子进行动态微调。

其次，针对扩散模型中噪声估计网络产生的量化误差会随采样时间步迭代而累积的问题，本文在量化后校准阶段引入全精度模型作为引导，分别在每个时间步下对量化参数进行微调，从而在推理阶段动态抑制了随时间步累积的量化误差。

为验证本文提出量化方法的有效性，实验通过对 DDIM 与 LDM 两种研究及应用广泛的扩散模型进行量化及性能评估，在 CIFAR 及 LSUN 数据集上验证了 W8A8、W6A6 与 W4A8 三种量化位宽设置下的结果，表明本文方法在性能上基本达到同类最优。此外，本文在主流扩散模型量化框架 TDQ、Q-Diffusion、EfficientDM 及 TFMQ-DM 的基础上进一步集成了提出方法，实验结果表明其具有良好的可移植性。

然而，本文工作仍然存在不足。一方面，对于量化重建时的粒度划分策略，本文工作沿用了 BRECQ、Q-Diffusion 等方法的设置，但网络内分布不匀的量化误差表明其粒度划分并不均匀，需要对网络内部的量化粒度进行更细致的分析；另一方面，不同采样时间步不同模块的输入范围分布对最终生成图像的贡献度并不相同，其贡献度即权重应该被纳入微调策略，用以缓解训练后量化过程中校准数据较少而可能带来的过拟合问题。

参考文献（References）

Dettmers T, Pagnoni A, Holtzman A and Zettlemoyer L. 2023. QLORA: efficient finetuning of quantized LLMs//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #441

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

He Y F, Liu J, Wu W J, Zhou H and Zhuang B H. 2024. EfficientDM: efficient quantization-aware fine-tuning of low-bit diffusion models// Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

He Y F, Liu L P, Liu J, Wu W J, Zhou H and Zhuang B H. 2023. PTQD: accurate post-training quantization for diffusion models// Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #580

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: ACM: 6629-6640

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #574

Huang Y S, Gong R H, Liu J, Chen T L and Liu X L. 2024. TFMQ-DM: temporal feature maintenance quantization for diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7362-7371 [DOI: 10.1109/CVPR52733.2024.00703]

Kingma D P and Welling M. 2013. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Krishnamoorthi R. 2018. Quantizing deep convolutional networks for efficient inference: a whitepaper [EB/OL]. [2025-12-11]. https://arxiv.org/pdf/1806.08342.pdf

Krizhevsky A and Hinton G. 2009. Learning multiple layers of features from tiny images [EB/OL]. [2025-12-11]. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

Li W Q, Zhou K Y, Hu X X, Zhang X P, Li S and Qian Z X. 2025. Generative video steganography via secret mapping in diffusion models. Journal of Image and Graphics, 30(11): 3465-3478 (李文琪, 周科扬, 胡宵宵, 张新鹏, 李晟, 钱振兴. 2025. 扩散模型中秘

密信息映射的生成式视频隐写. 中国图象图形学报, 30(11): 3465-3478)[DOI: 10.11834/jig.250079]

Li X Y, Liu Y J, Lian L, Yang H R, Dong Z and Kang D. 2023. Q-diffusion: quantizing diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 17489-17499 [DOI: 10.1109/ICCV51070.2023.01608]

Li Y H, Gong R H, Tan X, Yang Y, Hu P, Zhang Q, et al. 2021. BRECQ: pushing the limit of post-training quantization by block reconstruction//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Lu C, Zhou Y H, Bao F, Chen J F, Li C X and Zhu J. 2022. DPM-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #418

Nagel M, Amjad R A, Van B M, Louizos C and Blankevoort T. 2020. Up or down? Adaptive rounding for post-training quantization//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: PMLR: #667

Niu C H, Song Y, Song J M, Zhao S J, Grover A and Ermon S. 2020. Permutation invariant graph generation via score-based generative modeling//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy: PMLR: 4474-4484

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X. 2016. Improved techniques for training GANs//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: ACM: 2234-2242

Sasaki H, Willcocks C G and Breckon T P. 2021. UNIT-DDPM: unpaired image translation with denoising diffusion probabilistic models [EB/OL]. [2025-12-11]. https://arxiv.org/pdf/2104.05358.pdf

Shang Y Z, Yuan Z H, Xie B, Wu B Z and Yan Y. 2023. Post-training quantization on diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1972-1981 [DOI: 10.1109/CVPR52729.2023.00196]

So J, Lee J, Ahn D, Kim H and Park E. 2023. Temporal dynamic quantization for diffusion models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: IEEE: #2113

Song J M, Meng C L and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Wang C Y, Wang Z W, Xu X W, Tang Y S, Zhou J and Lu J W. 2024. Towards accurate post-training quantization for diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 16026-16035 [DOI: 10.1109/CVPR52733.2024.01517]

Yu F, Seff A, Zhang Y D, Song S R, Funkhouser T and Xiao J X. 2015. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop [EB/OL]. [2025-12-11]. https://arxiv.org/pdf/1506.03365.pdf

Zheng T P, Chen Y X, Wen X Z, Li Y C and Wang Z Y. 2025. Diffusion model-generated video dataset and detection benchmarks. Journal of Image and Graphics, 30(4): 1059-1071 (郑天鹏, 陈雁翔, 温心哲, 李严成, 王志远. 2025. 扩散模型生成视频数据集及其检测基准研究. 中国图象图形学报, 30(4): 1059-1071) [DOI: 10.11834/jig.240259]

作者简介

占瑞乙,男,硕士研究生,主要研究方向为计算机视觉与模式识别。E-mail:zry@buaa.edu.cn

陈佳鑫,通信作者,男,副教授,主要研究方向为计算机视觉和机器学习。E-mail:jiaxinchen@buaa.edu.cn

樊轶,男,高级工程师,主要研究方向为计算机应用。E-mail:fanyi@cetc.com.cn

周丽娜,女,高级工程师,主要研究方向为计算机应用。E-mail:biranshiwo@163.com

谢宇宝,男,高级工程师,主要研究方向为计算机应用。E-mail:592471@qq.com

杨鸿宇,女,副教授,主要研究方向为计算机视觉和模式识别。E-mail:hongyuyang@buaa.edu.cn

黄迪,男,教授,主要研究方向为计算机视觉、模式识别和具身智能。E-mail:dhuang@buaa.edu.cn

王蕴红,女,教授,主要研究方向为生物特征识别、计算机视觉和模式识别。E-mail:yhwang@buaa.edu.cn