Computer Science and Application 计算机科学与应用, 2023, 13(12), 2243-2252
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312225

Hans汉斯

基于稳定扩散模型的AR显示装置像质提升与智能交互方法

张馨月[1*]，李凌霄[2#]，高　蕾[2]，周晓强[2]，赵　芫[2]

[1]重庆理工大学计算机科学与工程学院，重庆
[2]重庆理工大学理学院，重庆

收稿日期：2023年11月15日；录用日期：2023年12月15日；发布日期：2023年12月25日

摘　要

本文针对现有AR显示装置在图像质量提升领域的固有缺陷，以扩散生成模型为基础，改进设计了一种基于稳定扩散模型的AR显示装置像质提升与智能交互方法。该像质提升方法将通用Diffusion扩散模型和Encoder-Dedecoder的结构相结合，将场景图像输入经过训练的像质提升模型中，通过编码器将输入的场景图像转换为隐变量特征，然后通过反向扩散模块中训练好的深度神经网络层按时间节点的逆向顺序对隐变量特征进行反向迭代计算逐层生成降噪隐变量特征，直至得到最终的降噪隐变量特征，最后通过解码器将最终的降噪隐变量特征转换为降噪处理后的像质提升图像。通过相关试验证明本文方法相较于通用Diffusion扩散模型迭代速度更快，生成性能指标更好。在此基础上，本文还进一步设计了相应的智能交互方法，能够通过稳定扩散模型实现场景图像的降噪增强以及场景图像与所需交互信息之间的智能融合，且还能够将稳定扩散模型进行本地终端部署以避免不必要的数据远程传输消耗。

关键词

AR显示，像质提升，智能交互

Image Quality Improvement and Intelligent Interaction Method of AR Display Device Based on Stable Diffusion Model

Xinyue Zhang[1*], Lingxiao Li[2#], Lei Gao[2], Xiaoqiang Zhou[2], Yuan Zhao[2]

[1]School of Computer Science and Engineering, Chongqing University of Technology, Chongqing
[2]School of Science, Chongqing University of Technology, Chongqing

[*]第一作者。
[#]通讯作者。

文章引用: 张馨月, 李凌霄, 高蕾, 周晓强, 赵芫. 基于稳定扩散模型的 AR 显示装置像质提升与智能交互方法[J]. 计算机科学与应用, 2023, 13(12): 2243-2252. DOI: 10.12677/csa.2023.1312225

Received: Nov. 15th, 2023; accepted: Dec. 15th, 2023; published: Dec. 25th, 2023

Abstract

The invention specifically relates to a method for image quality improvement and intelligent interaction of an AR display device based on a stable diffusion model. Image quality improvement methods include: The scene image is input into the trained image quality improvement model, and the input scene image is converted into hidden variable features through the encoder. Then, the trained deep neural network layer in the reverse diffusion module performs reverse iterative calculation on the hidden variable features according to the reverse order of time nodes to generate the noise reduction hidden variable features layer by layer until the final noise reduction hidden variable features are obtained. Finally, the final hidden variable features of noise reduction are converted into the image quality improvement image after noise reduction by decoder. The invention further discloses a corresponding intelligent interaction method. The invention can realize the noise reduction and enhancement of the scene image and the intelligent fusion between the scene image and the required interactive information through the stable diffusion model, and can also deploy the stable diffusion model to the local terminal to avoid unnecessary consumption of remote data transmission.

Keywords

AR Display, Pixel Enhancement, Intelligent Interaction

1. 引言

近年来随着人工智能领域[1]和 AR (Augmented Reality，增强现实)技术[2]的不断发展，各类 AR 眼镜产品[3] (显示装置)逐渐开始民用化和商业化。AR 技术能够基于现实物理环境来构建虚拟景象，从而带给使用者全新体验，因此应用 AR 技术的近眼显示装置日益受到关注

AR 技术涉及多类技术领域，发展符合人眼视觉特性的近眼显示技术成为增强现实的技术制高点。近眼显示技术以沉浸感提升与眩晕控制为主要发展趋势，而眩晕控制是其技术难点。从人眼视觉特性看，业界公认的眩晕感主要源自三方面。一是显示画质[4]，二是视觉与其他感官通道的冲突[5]，三是辐辏调节冲突[6]。

显示画质是 AR 技术的研究重点。如公开号为 CN109068125A 的中国专利就公开了《一种 AR 系统》，包括穿戴式显示器[7]及云端服务器[8]；穿戴式显示器与云端服务器相互通信；穿戴式显示器用于采集当前图像信息并将采集到的当前图像信息发送至云端服务器；云端服务器用于将当前图像信息与云端服务器的信息库中的预设信息进行匹配获取当前图像信息对应的目标匹配信息，并发送目标匹配信息至穿戴式显示器；穿戴式显示器还用于显示目标匹配信息。

上述现有方案需要将 AR 显示装置采集的图像信息发送至云端服务器处理，并由云端服务器将处理后的图像信息回传至 AR 显示装置进行显示，以此实现 AR 显示装置的像质提升和智能交互。该方案需建立 AR 显示装置和云端的网络通信，涉及数据采集、数据传输、数据处理和数据回传等过程。然而，

数据传输和数据回传会耗费大量时间，这会导致 AR 显示装置像质提升和智能交互的效率偏低，延时较大，在网络环境不好时甚至无法正常实施，从而大大影响这类方案的环境适应性和稳定性。因此，如何设计一种能够高效、稳定地提高 AR 显示装置像质和实时交互性的新方法是目前亟需解决的技术问题。

针对上述现有技术的不足，本文所要解决的技术问题是：如何提供一种基于稳定扩散模型[9]的 AR 显示装置像质提升与智能交互方法，使得该方法既能够实现对场景图像的降噪增强[10]，也能将场景图像与所需交互信息进行智能融合，并且还能够将算法模型快速进行本地终端部署以避免不必要的数据远程传输，从而提高 AR 显示装置在像质提升和智能交互过程中的效率和稳定性，保证用户丝滑、流畅的使用体验。

2. 设计思路

本文设计的 AR 显示装置的逻辑框图如下图 1 所示，分别包含显示交互单元，视频采集单元，模式选择单元，综合信息处理单元，每个单元的特征如下：

显示交互单元，用于显示 AR 图像；

视频采集单元[11]，用于采集用户视场内的场景图像；

模式选择单元，用于供用户选择显示交互单元待执行的模式，包括像质提升模式和智能交互模式；

综合信息处理单元，用于根据用户选择的显示交互单元待执行的模式。

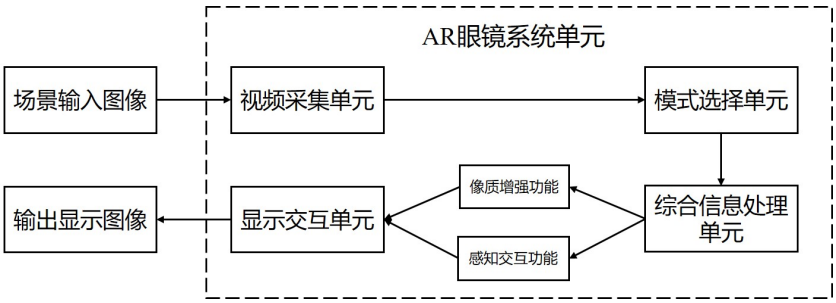

Figure 1. Logic block diagram of AR display device system
图 1. AR 显示装置系统的逻辑框图

本文设计的 AR 显示装置的工作流程图如下图 2 所示，所述 AR 显示装置的像质提升与智能交互方法，具体包括以下 4 个步骤：

1) 通过所述视频采集单元采集人眼视场内的场景图像；

2) 通过所述模式选择单元对 AR 眼镜的功能模式进行选择或切换，所述功能模式具体包括像质提升模式和智能交互模式。若选择所述像质提升模式，则进入 3)，若选择所述智能交互模式，则进入 4)；

3) 所述像质提升模式通过综合信息处理单元集成的一种基于稳定扩散模型的深度神经网络生成算法，利用该算法对视频采集单元获取的图像进行降噪增强处理，同时将处理后的图像输出到显示识别单元与所述 AR 眼镜进行实时交互；

4) 所述智能交互模式同样通过综合信息处理单元内置的深度神经网络生成算法，将人工选择的体验场景作为文本附加条件输入到稳定扩散模型中从而生成智能化交互图像，并与视频采集单元获取的现实图像场景进行融合，最后将融合后的图像输出到显示识别单元中实现与所述 AR 眼镜进行智能交互。

基于稳定扩散模型的 AR 显示装置系统，其鲜明特征在于：综合信息处理单元，而综合信息处理单元包括像质提升和智能交互两个模块。其中，像质提升模块，用于在显示交互单元待执行的模式为像质提升模式时，通过本文的像质提升模型对场景图像进行图像去噪并生成对应的像质提升图像，进而将像

质提升图像作为显示交互单元的输出图像进行显示；而对于智能交互模块，用于在显示交互单元待执行的模式为智能交互模式时，通过本文的智能交互模型将当前场景图像与用户选择的交互体验模式对应的文本描述信息融合并生成对应的重构交互图像，进而将重构交互图像作为显示交互单元的输出图像进行显示。

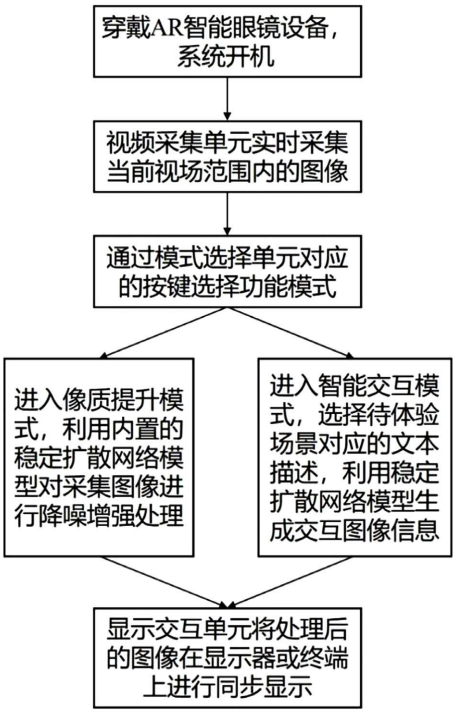

Figure 2. Workflow diagram of AR display device system
图 2. AR 显示装置系统的工作流程图

基于稳定扩散模型的 AR 显示装置像质提升与智能交互方法与现有技术相比，具有显著优势。训练智能交互模型时，通过前向扩散的方式在纯净的样本图像上按时间节点的先后顺序依次添加标准高斯噪声用作各层深度神经网络层的训练输入，同时将时间节点的编码特征向量以及交互体验模式的文本描述信息作为深度神经网络层的训练输入，并将各层深度神经网络层输出的预测噪声(含有文本描述信息)和添加的标准高斯噪声作为损失函数进行参数优化，训练后的每层深度神经网络层能够实现对应层级的特征和文本描述信息融合，使得经过训练的智能交互模型能够通过训练好的各层深度神经网络层以反向扩散的方式，按时间节点的逆向顺序对输入的隐变量特征进行反向迭代计算，逐层融合交互体验模式的文本描述信息并生成重构隐变量特征，这种逐层信息融合的方式仅需在仿照上述像质提升模型对应的逐层降噪过程基础上，每次多加入一个文本描述信息作为额外输入就能够实现交互体验模式下文本描述信息和场景图像的智能融合，模型复用性强、可扩展性能好，因此能够有效提高 AR 显示装置智能交互的体验效果。

3. 稳定扩散模型的建立

本文面向增强现实技术领域，具体设计的基于稳定扩散模型的 AR 显示装置像质提升与智能交互方法。其逻辑框图和工作流程图分别如下图 3、图 4 所示。

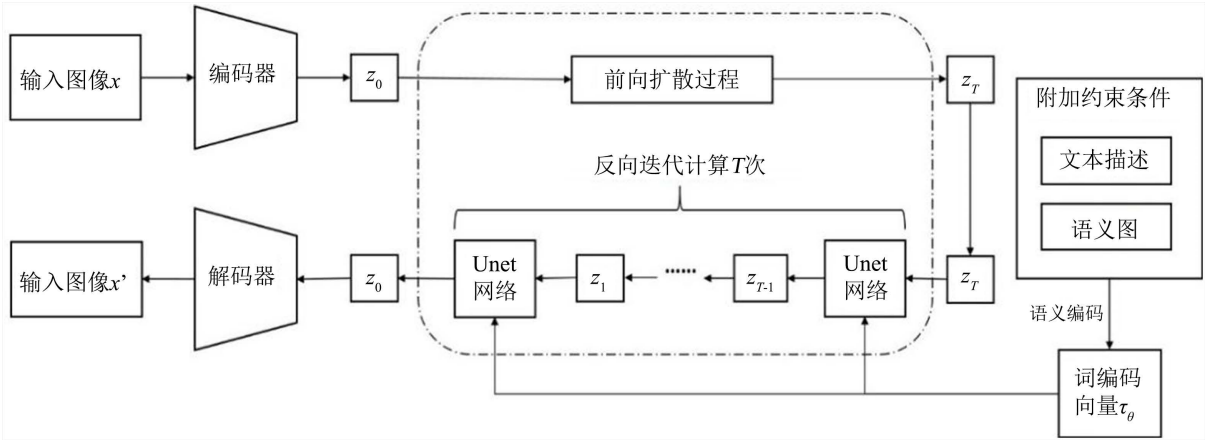

Figure 3. Logic diagram of image quality improvement methods and intelligent interaction methods for AR display devices
图 3. AR 显示装置像质提升方法和智能交互方法的逻辑框图

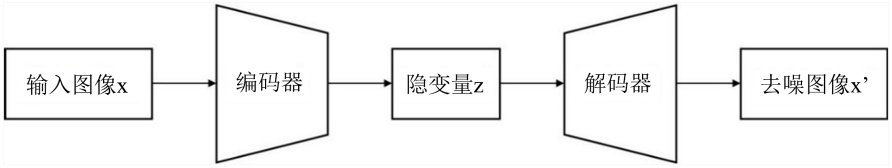

Figure 4. Flow chart of image quality improvement and intelligent interaction methods for AR display devices
图 4. AR 显示装置像质提升和智能交互方法的流程图

本文基于稳定扩散模型的 AR 显示装置智能交互方法的主要控制器件和模块包括：编码器、前向扩散模块、反向扩散模块和解码器，具体操作流程如下：

(1)：获取场景图像和用户选择的交互体验模式，并将该交互体验模式转换为对应的文本描述信息；

(2)：将场景图像和文本描述信息同时输入经过训练的智能交互模型中，输出融合了文本描述信息的重构交互图像。

其中在模型训练时：如图 5 所示，首先将训练样本图像和对应交互体验模式的文本描述信息作为智能交互模型的输入，利用编码器将样本图像转换为对应的隐变量特征；之后在前向扩散模块中采样 T 个离散的时间节点 $t^k \sim U\left(\{1,\cdots,T\}\right)$，在每个时间节点处随机生成标准高斯噪声 $\varepsilon_t^k \sim N\left(0,I\right)$，然后根据时间节点的先后顺序依次将各节点处对应的标准高斯噪声逐层添加到隐变量特征上，最终生成 T 个带噪隐变量特征；反向扩散模块包含 T 个与前向扩散模块各时间节点一一对应的深度神经网络层，各深度神经网络层用于将反向扩散过程中各节点处的带噪隐变量特征、该时间节点对应的编码特征向量以及该交互体验模式对应的文本描述信息作为输入，并输出对应的预测噪声，进而通过最小化约束神经网络输出的预测噪声和在该时间节点上前向扩散过程添加的标准高斯噪声作为损失函数实现网络训练；最后重复训练 T 个深度神经网络层直至网络收敛；在具体实施中，对时间节点进行嵌入编码转换为长度固定的时间编码特征向量 t_e^k (此处可调用 pytorch 深度学习库中的 torch.nn.Embedding 方法通过一个简单的正余弦编码实现)，而深度神经网络可选用现有常见的 Denoising-Unet 网络来进行搭建。

在模型训练后：首先将场景图像作为智能交互模型的输入；其次通过编码器将输入的场景图像转换为对应的隐变量特征；然后通过反向扩散模块中训练好的 T 个深度神经网络层按时间节点的逆向顺序对隐变量特征进行反向迭代计算，逐层生成融合了对应文本描述信息的重构隐变量特征，直至得到最终的重构隐变量特征；最后通过解码器将最终的重构隐变量特征转换为融合了交互体验模式文本信息的重构

交互图像并输出。

(3)：将重构交互图像作为对应场景图像的 AR 处理图像。

本文在训练和实际应用的过程中，先将原始图像转换为隐变量特征，然后对图像的隐变量特征进行前向扩散和反向扩散，由于隐变量特征的信息量大小比原始图像小很多，这使得本文能够显著降低图像前向扩散和反向扩散的计算量和处理难度，进而能够实现将智能交互模型(即稳定扩散模型)直接部署在 AR 显示装置所对应的本地边缘计算终端上，避免了数据通过网络传输造成的延迟和损耗，从而能够进一步提高 AR 显示装置智能交互的数据同步效率，保证了用户更快更好的使用体验。

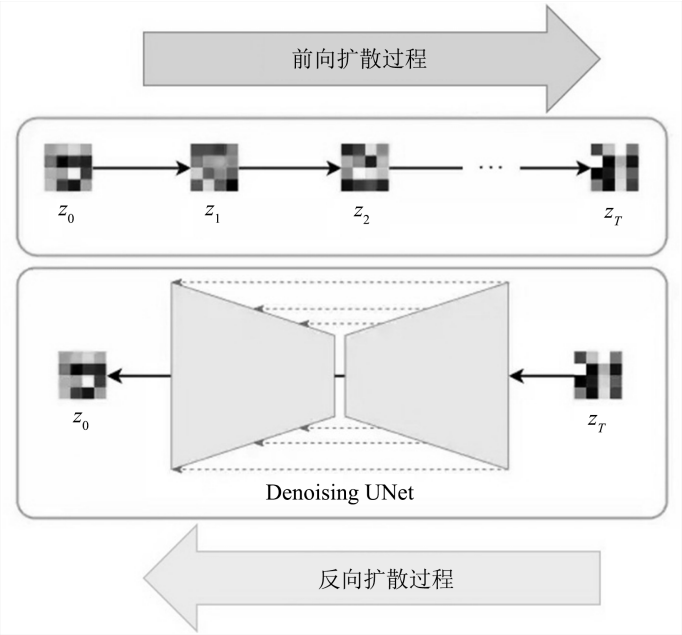

Figure 5. Schematic diagram of the process of forward and backward diffusion of latent variable features

图 5. 隐变量特征前向扩散和反向扩散的过程示意图

基于稳定扩散模型的 AR 显示装置像质提升方法，在具体实施中首先通过如下公式将训练样本图像转换为对应的隐变量特征：

$$z_0^k = Encode\left(x_0^k\right) \tag{1}$$

式中：z_0^k 表示第 k 个训练样本图像的隐变量特征；$Encode$ 表示编码操作；x_0^k 表示第 k 个训练样本图像。

基于稳定扩散模型的 AR 显示装置像质提升方法，通过如下公式生成带噪隐变量特征：

$$z_t^k = \overline{\alpha_t^k} z_0^k + \overline{\beta_t^k} \varepsilon_t^k \tag{2}$$

其中：

$$\left(\overline{\alpha_t^k}\right)^2 + \left(\overline{\beta_t^k}\right)^2 = 1; \tag{3}$$

$$\overline{\alpha_t^k} = \alpha_t^k \times \alpha_{t-1}^k \times \cdots \times \alpha_1^k, \overline{\beta_t^k} = \beta_t^k \times \beta_{t-1}^k \times \cdots \times \beta_1^k; \ \left(\alpha_t^k\right)^2 + \left(\beta_t^k\right)^2 = 1; \tag{4}$$

式中：z_t^k 表示第 k 个训练样本图像在时间节点 t 处的带噪隐变量特征；ε_t^k 表示第 k 个训练样本图像在时间节点 t 处对应的标准高斯噪声；z_0^k 表示第 k 个训练样本图像的隐变量特征；$\overline{\alpha_t^k}$ 和 $\overline{\beta_t^k}$ 表示第 k 个训练

样本图像在时间节点 t 处的噪声强度关联参数。

基于稳定扩散模型的 AR 显示装置像质提升方法，通过如下公式计算像质提升模型中各个深度神经网络层的损失函数：

$$L'(\phi)=\frac{\left(\beta_t^k\right)^2}{\left(\alpha_t^k\right)^2}\left\|\varepsilon_t^k-\varepsilon_\phi^k\left(z_t^k,t_e^k\right)\right\|^2; \tag{5}$$

式中：$L'(\phi)$ 表示像质提升模型中深度神经网络层的损失函数；$\varepsilon_\phi^k\left(z_t^k,t_e^k\right)$ 表示像质提升模型中与时间节点 t 对应的深度神经网络层输出的预测噪声；z_t^k 表示第 k 个训练样本图像在时间节点 t 处的带噪隐变量特征；t_e^k 表示第 k 个训练样本图像在时间节点 t 处的时间编码特征向量；ε_t^k 表示第 k 个训练样本图像在前向扩散过程中对应时间节点 t 处的标准高斯噪声；ϕ 表示深度神经网络层待优化的参数；α_t^k 和 β_t^k 表示第 k 个训练样本图像在时间节点 t 处的噪声强度关联参数 $\overline{\alpha_t^k}$ 和 $\overline{\beta_t^k}$ 中的中间系数。

基于稳定扩散模型的 AR 显示装置像质提升方法，通过如下公式生成降噪隐变量特征：

$$z_{t-1}=\frac{1}{\alpha_t}\left(z_t-\beta_t\varepsilon_\phi\left(z_t,t_e\right)\right)+\beta_t\varepsilon; \tag{6}$$

式中：z_{t-1} 表示时间节点 $t-1$ 处的降噪隐变量特征；$\varepsilon_\phi\left(z_t,t_e\right)$ 表示与时间节点 t 对应的深度神经网络层输出的预测噪声；z_t 表示时间节点 t 处的带噪隐变量特征；α_t 和 β_t 表示时间节点 t 处的噪声强度关联参数 $\overline{\alpha_t}$ 和 $\overline{\beta_t}$ 中的中间系数；ε 表示满足标准高斯分布的随机噪声；t_e 表示时间节点 t 处的时间编码特征向量。

通过如下公式将最终的降噪隐变量特征转换为像质提升图像：

$$x'=Decode\left(z_0\right); \tag{7}$$

式中：x' 表示像质提升图像；z_0 表示最终的降噪隐变量特征；$Decode$ 表示解码操作。

基于稳定扩散模型的 AR 显示装置智能交互方法，通过如下公式表示智能交互模型中深度神经网络层的损失函数：

$$L''(\phi)=\frac{\left(\beta_t^k\right)^2}{\left(\alpha_t^k\right)^2}\left\|\varepsilon_t^k-\varepsilon_\phi^k\left(z_t^k,t_e^k,\tau_\theta(y)\right)\right\|^2; \tag{8}$$

式中：$L''(\phi)$ 表示智能交互模型中深度神经网络层的损失函数；$\varepsilon_\phi^k\left(z_t^k,t_e^k,\tau_\theta(y)\right)$ 表示智能交互模型中与时间节点 t 对应的深度神经网络层输出的预测噪声；z_t^k 表示第 k 个训练样本图像在时间节点 t 处的带噪隐变量特征；t_e^k 表示第 k 个训练样本图像在时间节点 t 处的时间编码特征向量；$\tau_\theta(y)$ 表示文本描述信息的词编码向量(可以通过调用 python 机器学习库中 tokenizer 库进行选择，从而将文本描述信息 y 转换为词编码向量)；ε_t^k 表示第 k 个训练样本图像在时间节点 t 处的标准高斯噪声；ϕ 表示深度神经网络层待优化的参数，用于判断深度神经网络层是否收敛。

4. 试验与分析

为了测试上述设计方法的效果，本文对基于稳定扩散模型的 AR 显示装置进行了测试，并在测试过程中对稳定扩散模型的中间系数 β_t^k 进行了调整。在之前的公式(2)~(4)中，一般是将 β_t^k 设置为一个数值从小到大的等差序列，如 $\beta_t^k\in[1e-4,1e-2]$，然后通过上述公式计算前向扩散过程中的 z_t^k。这种方法虽然简单，但容易造成在前向扩散时靠后的时间节点上添加的噪声过多，导致在反向生成采样的时候这部分图像的贡献度较低，使得模型性能受限。因此本文采用一种改进的 β_t^k 生成策略，可以使得前向扩散过程中各时刻添加的标准高斯噪声强度更加均匀合理，其公式为：

$$\beta_t^k = 1 - \frac{s_t}{s_{t-1}} \tag{9}$$

$$s_t = \cos\left(\frac{t/T + step}{1 + step} \cdot \frac{\pi}{2}\right)^2 \tag{10}$$

上式中 step 为设置步长，其值一般为 0.008。利用上式进行计算时限定 $0 \le \beta_t^k \le 1$。利用本文提出的改进的 β_t^k 生成方法，如下图 6 所示，可以看出该方法添加的高斯噪声(第二行)比原始扩散模型添加的高斯噪声(第一行)在图像序列中的分布更加均匀，哪怕是序列中靠后时间间隔的带噪图像依然保留了原始图像中的部分特征，因此有利于提高网络模型的训练速度。另一方面，从下图 7 所绘制的折线分布图中也可以看到，相比于原始扩散模型的扩散策略(蓝色折线所示)，我们设计改进的扩散策略(黄色虚线所示)在同样的采样次数下的生成性能指标 FID (Frechet Inception Distance score，表示计算真实图像与生成图像在特征层上的距离，值越小代表生成质量越好)更低，由此证明该方法性能更好，效率更高。

Figure 6. Intelligent generation strategy diagram for intermediate coefficients
图 6. 中间系数智能生成策略图

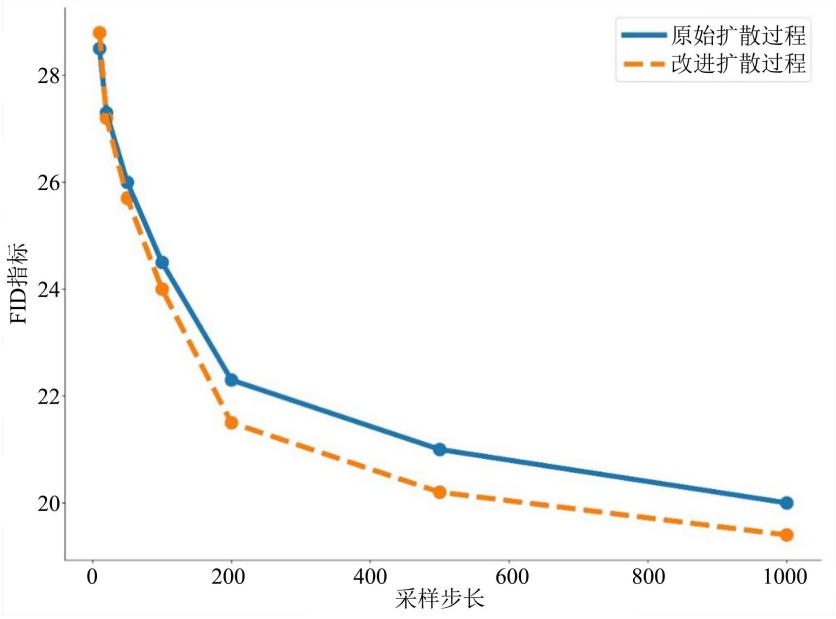

Figure 7. Comparison diagram of generation strategies
图 7. 生成策略的对比示意图

通过上述公式中的损失函数作为智能交互模型中各个深度神经网络层的损失函数，使得训练后的每层深度神经网络层能够实现对应层级的特征和文本描述信息融合，进而使得经过训练的智能交互模型能够通过训练好的各层深度神经网络层以反向扩散的方式，按时间节点的逆向顺序对输入的隐变量特征进

行反向迭代计算，逐层融合交互体验模式的文本描述信息并生成重构隐变量特征，这种逐层融合信息的方式能够更好的实现文本描述信息和场景图像的融合，进而能够有效实现场景图像的智能交互体验。最终利用该模型训练完成后可以快速实现对 AR 显示装置获取的图像进行快速去雾、夜视增强等各项像质优化功能，典型效果示例如下图 8 所示：

Figure 8. Example of image quality improvement effect of AR display device
图 8. AR 显示装置的像质提升效果示例

5. 总结

本文通过将稳定扩散模型与 AR 显示技术相结合，设计实现了一种基于稳定扩散模型的 AR 显示装置像质提升与智能交互方法，各项试验结果显示本文方法的迭代速度更快，生成性能指标更好，从而表明该方法既能够实现对场景图像的降噪增强，也能将场景图像与所需交互信息进行智能融合，并且还能够将算法模型快速进行本地终端部署以避免不必要的数据远程传输，从而提高 AR 显示装置在像质提升和智能交互过程中的效率和稳定性，保证用户丝滑、流畅的使用体验。

基金项目

重庆理工大学本科生科研立项项目(KLB22039)；重庆市基础研究与前沿探索专项(重庆市自然科学基金)一般项目(CSTB2022NSCQ-BHX0693)；重庆理工大学科研启动基金资助项目(2020ZDZ002)。

参考文献

[1] 毛毅. 人工智能研究热点及其发展方向[J]. 技术与市场, 2008(3): 4.

[2] 魏三强. Unity3D 与原生代码交互技术在 AR 开发中的应用[J]. 重庆理工大学学报(自然科学版), 2017(11): 166-171.

[3] 梁美玉. 阵列波导透视式 AR 眼镜光学系统设计[J]. 长春工程学院学报(自然科学版), 2019, 20(1): 121-123.

[4] Gross, H. (2008) Handbook of Optical Systems. Wiley-VCH, Weinheim. https://doi.org/10.1002/9783527699247

[5] 尤哈尼·帕拉斯玛. 肌肤之目——视觉与感官[M]. 北京: 中国建筑工业出版社, 2008.

[6] 蔡建奇. 3D 显示技术发展和偏光式眼镜测试方法研究[J]. 中国眼镜科技杂志, 2012: 7. http://xueshu.baidu.com/usercenter/paper/show?paperid=ad02cfa8aa3bcd5298267a2c2d5d089f&site=xueshu_se

[7] Zeng, K., Shi, X., Tang, C., Liu, T. and Peng, H. (2023) Design, Fabrication and Assembly Considerations for Electronic Systems Made of Fibre Devices. *Nature Reviews Materials*, **8**, 552-561. https://doi.org/10.1038/s41578-023-00573-x

[8] 张泽立. 云服务器 4G 传输模块的比较研究[J]. 计算机械工程学报, 2015, 38(2): 370-402.

[9] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. (2015) Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *International Conference on Machine Learning*, Edinburgh, 26 June-1 July 2012, 2256-2265.

[10] Purwins, H., Li, B., Virtanen, T., Schluter, J., Chang, S.-Y., and Sainath, T. (2019) Deep Learning for Audio Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*, **13**, 206-219. https://doi.org/10.1109/JSTSP.2019.2908700

[11] 高俊岭, 陈志飞, 章佩佩. 基于 FPGA 的实时视频图像采集处理系统设计[J]. 电子技术应用, 2018, 44(2): 10-12, 19.