从去噪扩散隐式模型隐空间中生成新桥型的尝试

张洪俊

（万世先行数智交通科技有限公司，南京 210016）

摘要：使用去噪扩散隐式模型进行桥型创新。将图像加噪、去噪的过程，类比为尸体腐烂、神探还原受害人遇害时情景的过程，来帮助初学者理解。通过通俗易懂的代数方法，推导加噪、去噪的函数公式，便于初学者掌握模型的数学原理。采用对称结构的三跨梁式桥、拱式桥、斜拉桥、悬索桥图像数据集，基于 Python 编程语言、TensorFlow 及 Keras 深度学习平台框架，构建和训练去噪扩散隐式模型。从隐空间采样，能够生成不对称结构的新桥型。去噪扩散隐式模型能够在人类原创桥型的基础上，将不同结构部件进行有机搭配，创造生成新桥型。

关键词：生成式人工智能；桥型创新；扩散模型；隐空间；深度学习

中图分类号：U448.2；TP181　　**文献标志码：**A

An attempt to generate new bridge types from latent space of denoising diffusion Implicit model

Hongjun Zhang

Wanshi Antecedence Digital Intelligence Traffic Technology Co., Ltd, Nanjing, 210016, China

Abstract: Use denoising diffusion implicit model for bridge-type innovation. The process of adding noise and denoising to an image can be likened to the process of a corpse rotting and a detective restoring the scene of a victim being killed, to help beginners understand. Through an easy-to-understand algebraic method, derive the function formulas for adding noise and denoising, making it easier for beginners to master the mathematical principles of the model. Using symmetric structured image dataset of three-span beam bridge, arch bridge, cable-stayed bridge and suspension bridge , based on Python programming language, TensorFlow and Keras deep learning platform framework , denoising diffusion implicit model is constructed and trained. From the latent space sampling, new bridge types with asymmetric structures can be generated. Denoising diffusion implicit model can organically combine different structural components on the basis of human original bridge types, and create new bridge types.

Keywords: generative artificial intelligence; bridge-type innovation; diffusion model; latent space; deep learning

0　引言

仅作者尝试的六种初级的生成式人工智能算法（变分自编码器、生成对抗网络、像素卷积神经网络、标准化流、基于能量的模型、去噪扩散隐式模型），就已经基本满足桥型创意设计的要求了。考虑到其它更高级的算法存在，因此生成式人工智能技术达到了桥梁方案设计师虚拟助手的能力标准。该项技术可以轻易横向扩展到景观、建筑设计等领域。(注：这里仅讨论图像生成。实际上生成式人工智能技术还可以文本、语音、视频等生成，这在工程领域也有很大应用价值，如工程标书、总说明书的编制和审查。)

该项技术在桥梁设计领域落地的思路有：①人工智能企业、高校、创业团队等，将人工智能技术拓展到桥梁领域来完善生态布局；②大型设计单位自建、收购研发团队，或者与外部共同开发、研发任务外委等方式。一般来说，回避基础算法，在现有预训练模型上，使用桥梁专业数据集进行微调，是成本低、快速出成果的技术路线，产品性能也能够满足行业要求。受宏观经济环境、营销、人力和硬件成本等因素制约，短期内盈利较为困难。目前该项技术仅仅能够提供造型创意，生成完整的方案设计还很困难，更不能生成初步设计、施工图设计，因此还有相当大的局限性和技术提升空间。

最近的大型扩散模型，如 OpenAI 的 DALL-E 2 和 Google 的 Imagen，表现出令人难以置信的文本到图像生成能力。扩散模型作为一种强大的生成式建模技术[1-3]，在图像生成质量上已经超过了经典的生成对抗网络（GAN），同时易于训练和扩展，已经成为开发者（尤其是文生图应用）之首选。扩散是指将图像一步步地转换为噪声的过程。使用经过训练的扩散模型，可以模拟扩散的反向操作，从噪声中逐步去噪来生成图像。

本文建立去噪扩散隐式模型（Denoising Diffusion Implicit Model，DDIM）[2,4]，采用之前同样的桥型图片数据集[5-9]，进一步尝试桥型的几何形态组合创新（本文数据集与源代码开源地址https://github.com/QQ583304953/Bridge-DDIM）。

1 去噪扩散隐式模型简介

1.1 概述

扩散模型的突破性论文（去噪扩散概率模型）发表于2020年夏天[1]，之后演化出去噪扩散隐式模型[2]、Latent Diffusion Models[3]等。目前比较火的AI绘画模型如DALL-E 2和Stable Diffusion等，核心算法都是基于Latent Diffusion Models。

给定图像数据集，我们逐步添加一点噪音。每一步，图像都会变得越来越不清晰，噪音成分越来越多，原图像信息越来越少。然后，我们学习一个机器学习模型，可以撤销每个这样的步骤，于是我们就有了一个可以从纯随机噪音生成图像的模型。

去噪扩散隐式模型可以直观理解为：生成模型如同一个神探，看到受害人最终腐烂状态的尸体，只要知道死亡时间，就能逐步分析还原出受害人遇害时的情景，无论尸体经历过多少种类的侵蚀。这个神探是如何拥有这项本领的？答：他之前在大学学习时，在实验室做了许多尸体腐烂的试验，观察各种侵蚀手段、时长对尸体状态的影响，不断修正自己的认知，最终修炼成功。

1.2 图像添加噪音的操作

假如有一幅图像X_0，我们希望通过一系列步骤来逐渐对图像进行各种侵蚀，使得最终的侵蚀结果X_T看起来与标准高斯噪声差不多。这个步骤相当于一个数学函数q作用，它可以给一幅图像X_{t-1}添加少量高斯噪音，来生成一幅新的图像X_t。如果我们一直使用该函数q，我们将生成一系列渐进式噪音图像$(X_0, X_1, \cdots, X_{t-1}, X_t, \cdots, X_{T-1}, X_T)$，如下图所示。

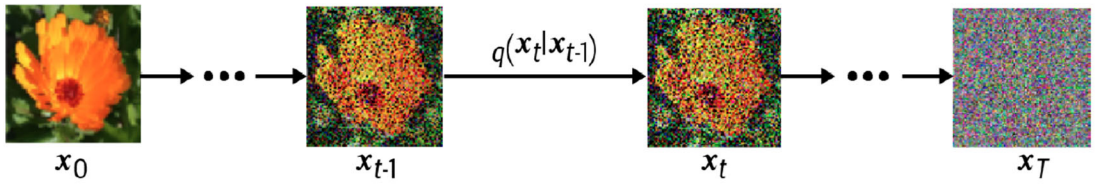


图1 图像添加噪音示意图

Fig.1 Schematic diagram of adding noise to image

为了便于计算，先将数据集进行数据标准化预处理，即图像X_0服从标准正态分布。（注：此步不是加噪，而是图像格式无损转换。）

1.为了让每幅噪音图像都服从标准正态分布，需要如下构造函数q：

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}\epsilon_{t-1} \tag{1}$$

式中：ϵ_{t-1}为噪音，服从标准正态分布；$\sqrt{1-\beta_t}$为图像X_{t-1}信息的比例系数signal_rates；$\sqrt{\beta_t}$为噪音信息的比例系数noise_rates。

X_{t-1}与ϵ_{t-1}相互独立、均服从标准正态分布，$(1-\beta_t)$为$\sqrt{1-\beta_t}X_{t-1}$的方差，β_t为$\sqrt{\beta_t}\epsilon_{t-1}$的方差，$(1-\beta_t)+\beta_t=1$，故$X_t$也服从标准正态分布。

公式（1）中，ϵ_{t-1}是随机抽取的，因此每一步的结果都有无数种可能性。

2. 从一幅图像X_0到X_t，仅需1次函数q的运算，而无需t次函数q的运算。推导如下：

令$\alpha_t = 1-\beta_t$，则公式（1）变为：

$$X_t = \sqrt{\alpha_t}X_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{2}$$

$$= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}X_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-2}\right) + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{3}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}X_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1} \tag{4}$$

ϵ_{t-2}与ϵ_{t-1}相互独立、均服从标准正态分布，$\alpha_t(1-\alpha_{t-1})+(1-\alpha_t) = 1-\alpha_t\alpha_{t-1}$，故对于加噪操作而言，以下等式符合要求：

$$\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1} = \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon' \tag{5}$$

故：

$$X_t = \sqrt{\alpha_t\alpha_{t-1}}X_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon' \tag{6}$$

令$\bar{\alpha} = \alpha_t\alpha_{t-1}\ldots\alpha_1$，依据归纳法可得：

$$X_t = \sqrt{\bar{\alpha}}X_0 + \sqrt{1-\bar{\alpha}}\epsilon \tag{7}$$

式中：ϵ为噪音，服从标准正态分布；$\sqrt{\bar{\alpha}}$为图像X_0信息的比例系数 signal_rates；$\sqrt{1-\bar{\alpha}}$为噪音信息的比例系数 noise_rates。

3. 采用公式（7），将数据集转换为噪音图像，示例见下图（采用有偏置项的余弦扩散计划）：

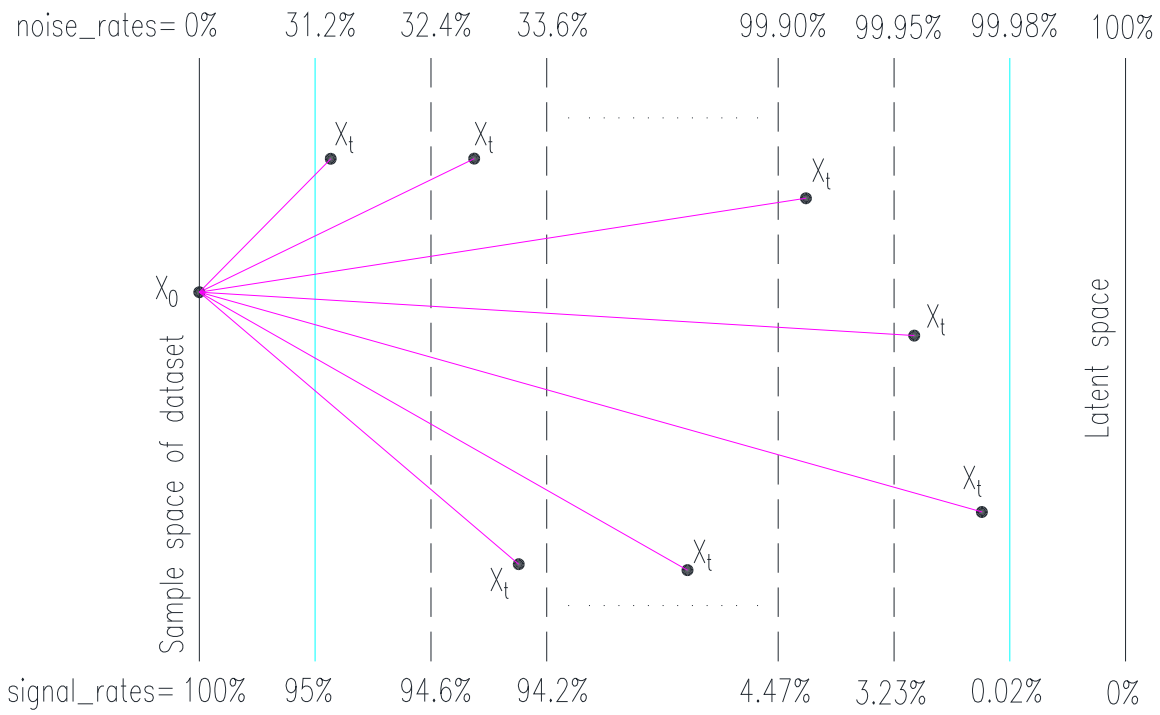


图 2 数据集转换为噪音图像的示意图

Fig. 2 Schematic diagram of converting dataset into noisy images

上图设定图像X_0信息的比例系数 signal_rates 范围为 95%~0.02%，对应噪音信息的比例系数 noise_rates 范围为 31.2%~99.98%。图中示意的隐空间是人为假设的，对应 100%的噪音，是X_0经过无限多次函数q运算的理论结果，实际上按照上图的扩散计划是不可能出现的。

4、图 2 中每条洋红色线段，代表采用公式（7）的一次加噪操作。一般情况下图中是不存在共线的情景。但是也可以人为构造出共线的情景，方法如下：

$$X_{20} = \sqrt{\bar{\alpha}_{20}}X_0 + \sqrt{1-\bar{\alpha}_{20}}\epsilon_{20} \tag{8}$$

$$X_{30} = \sqrt{\bar{\alpha}_{30}}X_0 + \sqrt{1-\bar{\alpha}_{30}}\epsilon_{30} \tag{9}$$

ϵ_{20}与ϵ'相互独立、均服从标准正态分布，$(1-\bar{\alpha}_{30}-\sigma^2)+\sigma^2 = 1-\bar{\alpha}_{30}$，故对于加噪操作而言，以下等式符合要求：

$$\sqrt{1-\bar{\alpha}_{30}-\sigma^2}\epsilon_{20} + \sigma\epsilon' = \sqrt{1-\bar{\alpha}_{30}}\epsilon_{30} \tag{10}$$

联立公式（9）、（10）：

$$X_{30} = \sqrt{\bar{\alpha}_{30}}X_0 + \sqrt{1-\bar{\alpha}_{30}-\sigma^2}\epsilon_{20} + \sigma\epsilon' \tag{11}$$

其中σ可以取任意值。当σ=0 时：

$$X_{30} = \sqrt{\bar{\alpha}_{30}}X_0 + \sqrt{1-\bar{\alpha}_{30}}\epsilon_{20} \tag{12}$$

观察公式（8）、（12），两次不同的加噪操作，噪音取值相同，这意味着X_{20}是X_{30}的中途值，故X_{20}、X_{30}两次加噪操作共线。

公式（11）中当$\sigma \neq 0$ 时，代表X_{20}、X_{30}两次加噪操作不共线，也即噪音取值不相同。

1.3 训练神经网络预测噪音

公式（7）中，一共 4 个变量，加噪时，已知$\bar{\alpha}$、X_0、ϵ三个值，通过代数运算求解X_t。

如果仅知道$\bar{\alpha}$、X_t两个值，通过代数运算是无法求解ϵ 、X_0的。

但是图像特征具有平移不变性、空间层次结构等特性，仅利用$\bar{\alpha}$、X_t两个变量，经过训练，神经网络可以预测ϵ 、X_0。例如一具尸体左侧被某种未知化学物质腐蚀，那么神探可以利用人体对称性将其还原。

1. U-Net 神经网络预测噪音

U-Net 包含两半：一半是降采样，输入图像在空间上压缩，在通道上扩展；另一半是上采样，在空间上扩展，而通道数则减少。这与变分自编码器类似。但是，与变分自编码器不同的是，在网络的降采样和上采样部分中，对于同样空间形状的层，U-Net 是存在跳跃连接的，允许信息在网络的一部分上形成短路并直接流向后面的层。

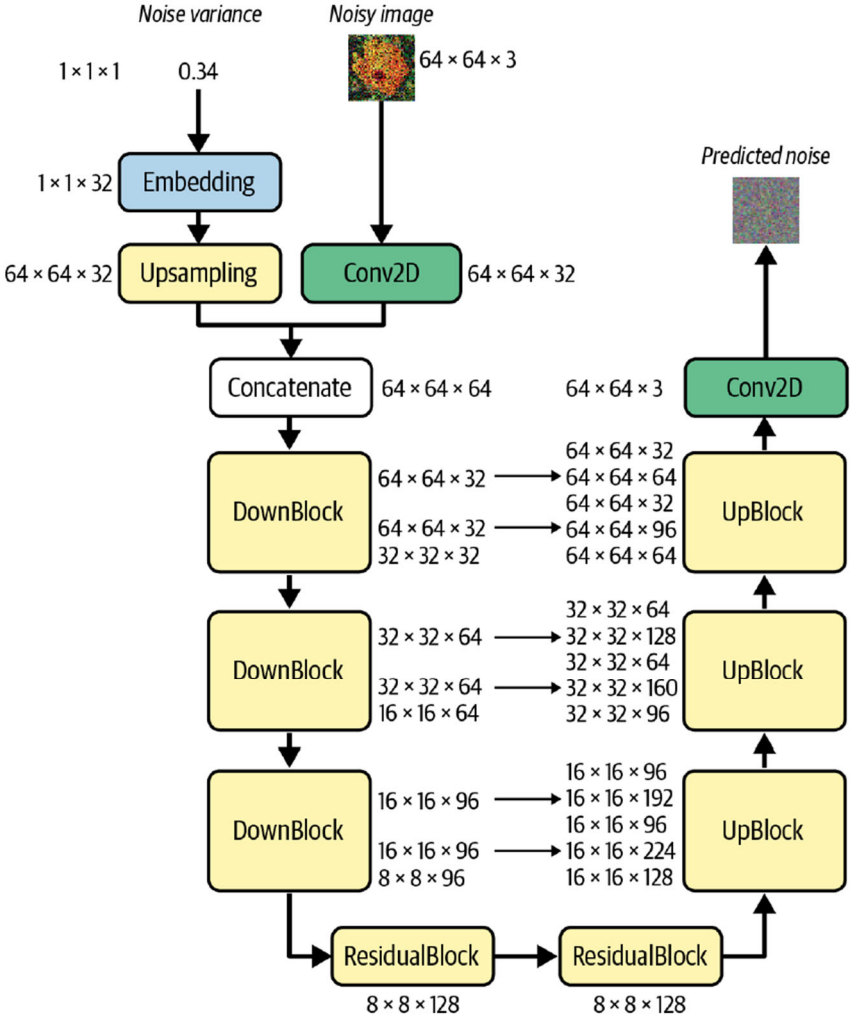


图 3 U-net 架构图

Fig.3 U-net architecture diagram

公式（7）中，$1-\bar{\alpha}$是$\sqrt{1-\bar{\alpha}}\epsilon$的方差，也即 noise_rates 的平方。

U-Net 神经网络的功能是：输入方差、噪音图像，计算得到噪音ϵ的预测值。再利用公式（7）即可得到X_0的预测值。

将真实噪音ϵ与预测噪音ϵ的偏差绝对值作为损失函数，训练神经网络。

1.4 隐空间采样去除噪音生成图像的操作

从隐空间随机噪声X_t出发，使用模型来逐步去除噪声，直到我们得到一幅有意义的图像X_0。

模型是一步到位得到噪音ϵ、X_0的预测值。可以想象一步到位生成一副有意义的图像是困难的，模型性能达不到这种能力，毕竟隐空间采样点是 100%纯噪音，之前训练时最大噪音也就 99.98%。

因此需要逐步去除噪声，先生成X_{t-1}，然后通过循环迭代逐步生成最终图形。

人为构造出共线的情景，从而通过模型预测的ϵ、X_0，推算出X_{t-1}。过程如下：

$$X_{t-1}=\sqrt{\bar{\alpha}_{t-1}}X_0+\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_{t-1} \tag{13}$$

ϵ_t与ϵ相互独立、均服从标准正态分布，$(1-\bar{\alpha}_{t-1}-\sigma_t^2)+\sigma_t^2=1-\bar{\alpha}_{t-1}$，故对于加噪操作而言，以下等式符合要求：

$$\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_t+\sigma_t\epsilon=\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_{t-1} \tag{14}$$

联立公式（13）、（14）：

$$X_{t-1}=\sqrt{\bar{\alpha}_{t-1}}X_0+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_t+\sigma_t\epsilon \tag{15}$$

其中σ_t可以取任意值。当σ_t=0 时：

$$X_{t-1}=\sqrt{\bar{\alpha}_{t-1}}X_0+\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t \tag{16}$$

因此，公式（16）实现了利用模型预测的ϵ_t、X_0，推算出X_{t-1}。接着采用同样方法，生成X_{t-2}，直到X_0。步数 t 取值越大，生成图像逼近理论中的X_0精度越高。

分 20 步循环迭代生成图像的过程图示如下：

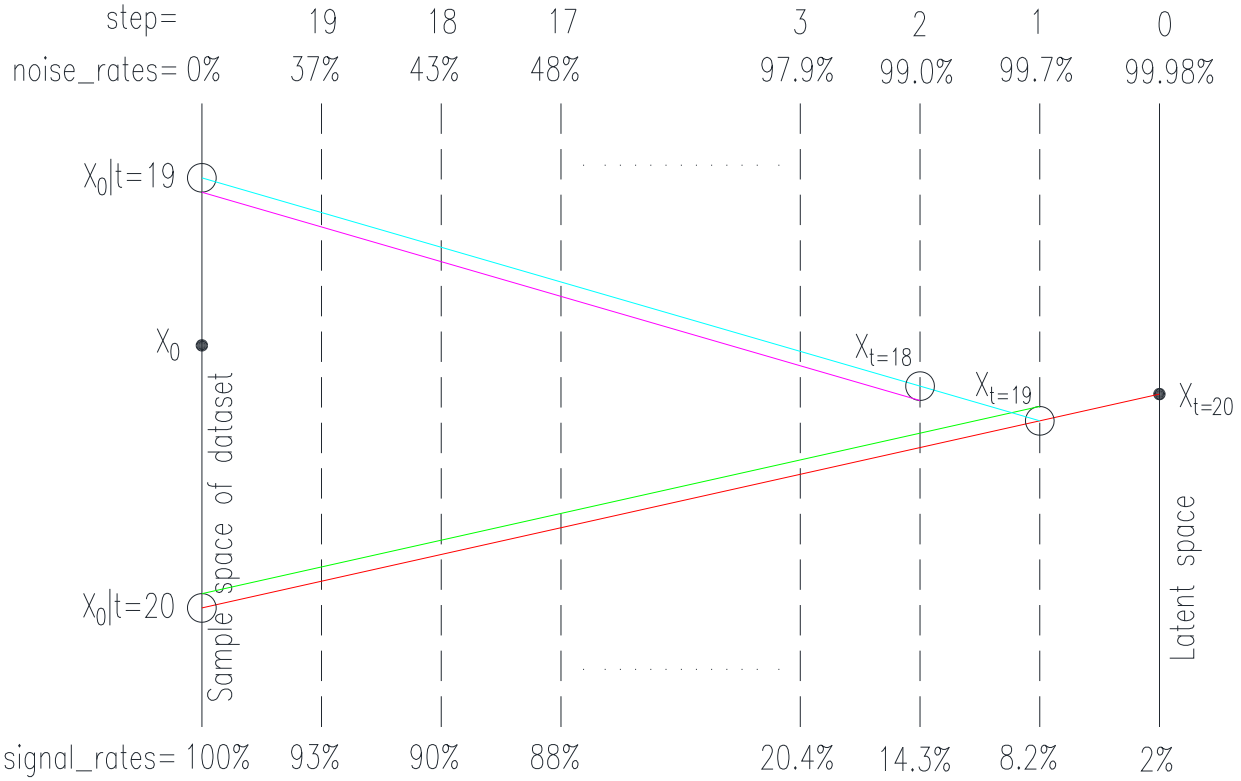


图 4 隐空间采样去除噪音生成图像的示意图

Fig. 4 Schematic diagram of generating images by removing noise from latent space

公式（15）中人为设定σ_t=0，这牺牲了每个采样点生成图像的多样性。而这种特性也正是我们希望的，因为我们希望隐空间采样点与像素空间是确定的映射关系。

2 从去噪扩散隐式模型隐空间中生成新桥型的尝试

2.1 数据集

采用作者之前论文[5-9]的数据集，即每种桥型两种子类（分别为等截面梁式桥、V 形墩刚构梁式桥、上承式拱式桥、下承式拱式桥、竖琴式斜拉桥、扇式斜拉桥、竖吊杆悬索桥、斜吊杆悬索桥），且均为三跨（梁式桥为 80+140+80m，其它桥型均为 67+166+67m），结构对称。

为了能够快速训练模型，故将图片尺寸由 512x128 减少为 192x48 像素（缺点是清晰度降低了许多）。

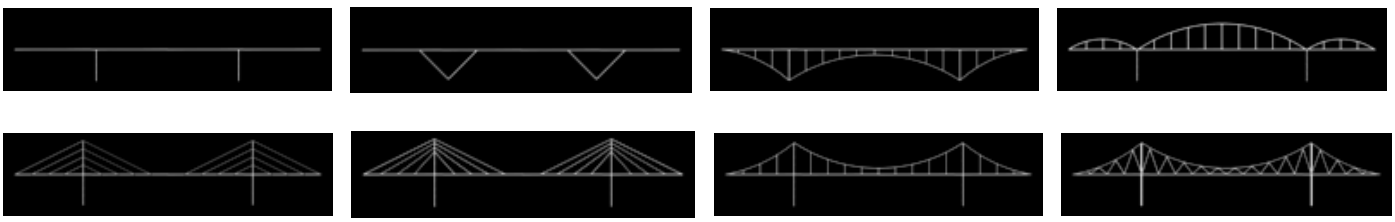


图 5 各桥型立面灰度图

Fig. 5 Grayscale image of each bridge facade

每个子类桥型 1200 张各不相同的图片，整个数据集共 9600 张图片。

2.2 模型构建和训练

基于 Python 编程语言、TensorFlow 及 Keras 深度学习平台框架，构建和训练去噪扩散隐式模型[4]。

1. U-Net 神经网络中的四个组成部分

（1）噪声方差的正弦嵌入(the sinusoidal embedding of the noise variance)

将一个标量数值（噪声方差）转换到一个不同的高维向量，该向量可以为网络的下游提供一个更复杂的表

示。

（2）残差模组（the ResidualBlock）

一个残差模组包含从输入连向输出的跳跃连接。残差模块帮助我们构建可以学习更复杂模式的更深网络，而无需遭受梯度消失、退化问题。

（3）下采样模组（the DownBlock）

一个下采样模组通过 block_depth(这里取 2)个残差模组来增加通道数，同时也最终接入一个 AveragePooling2D 层来将图像尺寸减半。每个残差模组都被加入到一个列表里，以便通过 U-Net 跳跃连接直连的对应上采样层能够利用起来。

（4）上采样模组（the UpBlock）

一个上采样模组首先应用 UpSampling2D 层来将尺寸通过双线性插值翻倍。每个连续的上采样模组通过 block_depth（这里取 2）个残差模组来减小通道数，同时也将通过 U-Net 跳跃连接直连的下采样模组的输出连接起来。

2. 训练

训练过程的损失变化曲线见下图：

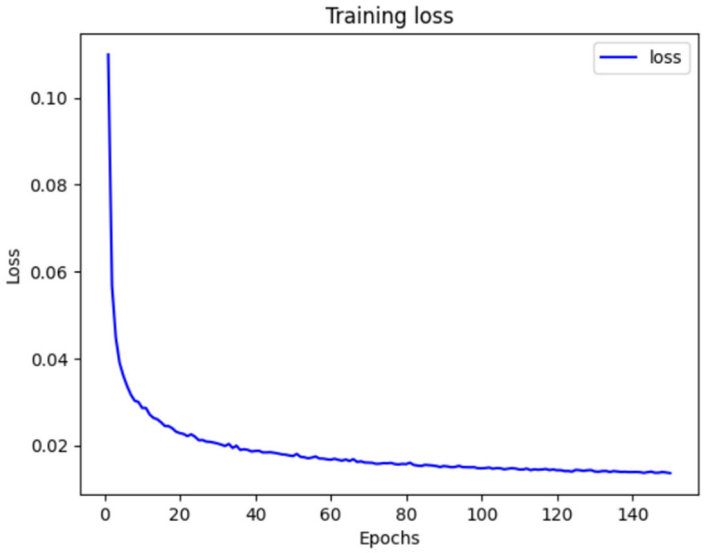
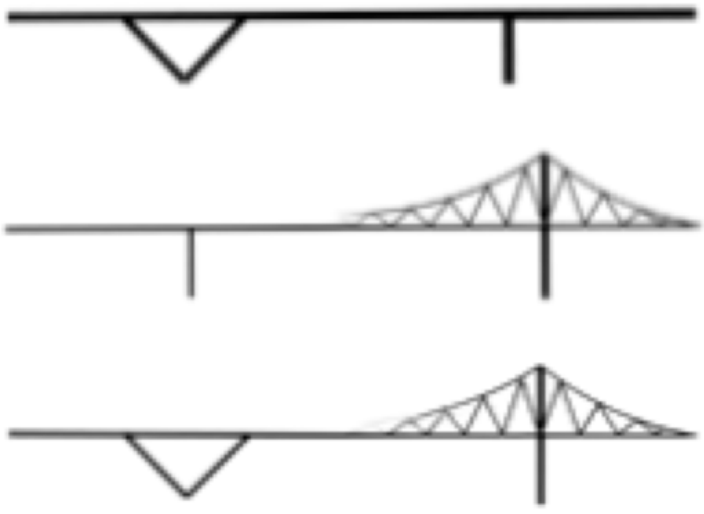


图 6 训练损失曲线

Fig.6 Training loss curve

2.3 隐空间采样探索新桥型

在隐空间中随机采样，然后人工基于工程结构思维筛选，得到了与训练集完全不同的 5 种技术可行的新桥型:



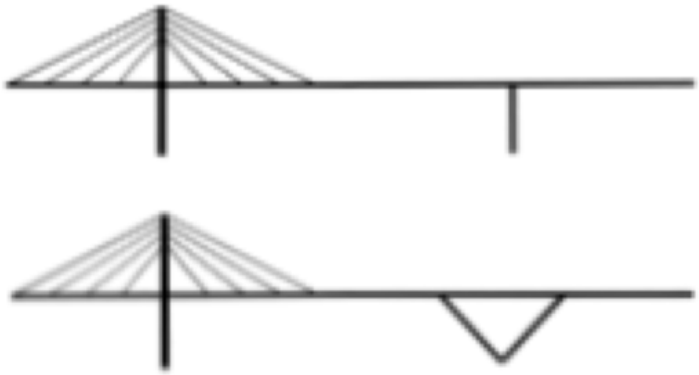


图 7 5 种技术可行的新桥型

Fig.7 Five new bridge types with feasible technology

这里新桥型是指数据集中未曾出现过，而神经网络根据算法创作出的，这代表了模型具有创新能力。

3 结语

去噪扩散隐式模型的性能与生成对抗网络、标准化流类似，比变分自编码器、基于能量的模型更有创造力，能够在人类原创桥型的基础上，将不同结构部件进行有机搭配，创造生成新桥型，一定程度上具有类似人类的原创能力，它能够打开想象空间，给予人类启发。

参考文献

[1]Jonathan Ho, Ajay Jain, Pieter Abbeel. Denoising Diffusion Probabilistic Models[J]. arXiv preprint,2020,arXiv:2006.11239.

[2]Jiaming Song, Chenlin Meng,Stefano Ermon.DENOISING DIFFUSION IMPLICIT MODELS[J]. arXiv preprint,2020,arXiv: 2010.02502.

[3] Robin Rombach, Andreas Blattmann,Dominik Lorenz, et al.High-Resolution Image Synthesis with Latent Diffusion Models[J]. arXiv preprint,2021,arXiv: 2112.10752.

[4] David Foster. Generative Deep Learning[M].2nd Edition. America: O’Reilly,2023.

[5]张洪俊. 从变分自编码器隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2023-11-06]. http://www.paper.edu.cn/releasepaper/content/202311-5.

[6]张洪俊. 从生成对抗网络隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2023-12-25]. http://www.paper.edu.cn/releasepaper/content/202312-73.

[7]张洪俊. 从像素卷积神经网络隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2024-01-08]. http://www.paper.edu.cn/releasepaper/content/202401-1.

[8]张洪俊. 从标准化流隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2024-01-25]. http://www.paper.edu.cn/releasepaper/content/202401-64.

[9]Hongjun Zhang. An attempt to generate new bridge types from latent space of energy-based model[J]. arXiv preprint,2024,arXiv: 2401.17657.