中山大學
SUN YAT-SEN UNIVERSITY

基于深度生成模型的多模态视觉合成

NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion

中山大学智能工程学院

方桂安，唐迅，凌海涛，张书戬，潘嘉雯

指导老师：刘梦源

目录

CONTENTS

Part.01

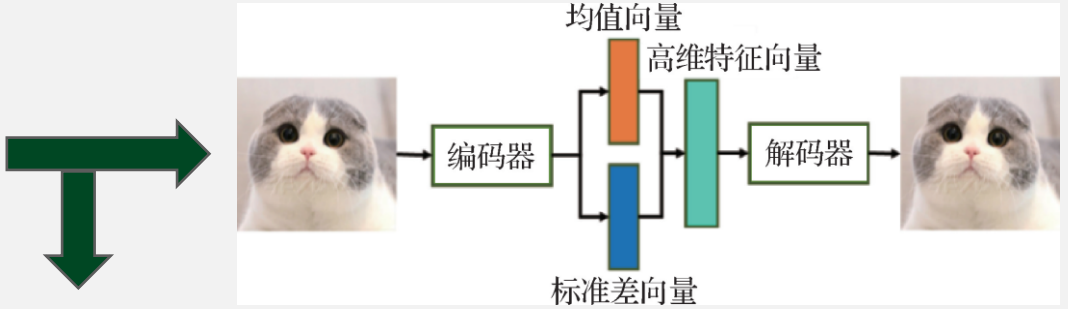
问题背景解读

1.1 研究背景

研究背景解读

First

随着图像和视频成为新的信息载体并在许多实际应用中使用，视觉合成成为了越来越热门的研究课题，其目的是为各种视觉场景建立能够生成新的或操作现有视觉数据(即图像和视频)的模型。



1.1 研究背景

研究背景解读

Second

与生成对抗网络模型相比，自回归模型由于其明确的密度建模和稳定的训练优势，在视觉合成任务中扮演着重要的角色。

早期自回归模型：

PixelCNN，PixelRNN、Image Transformer、iGPT和Video Transformer

存在的缺陷：

都是以“逐个像素”的方式进行视觉合成的

↓ 在高维视觉数据上的计算代价很高

只能应用于低分辨率的图像或视频，并且很难放大

1.1 研究背景

研究背景解读

Third

与此前以“逐个像素”的方式进行视觉合成的自回归模型相比，VQ-VAE模型的提出让视觉生成任务变得更加高效。虽然取得了巨大的成功，但这样的解决方案仍然有局限性：它们将图像和视频分开处理

SOTA：

女娲算法

(Visual Synthesis Pre-training for Neural visUal World creAtion)

1.2 研究意义

A. 视觉合成可以为计算机视觉任务提供大量的数据，并保证其标注质量和多样性

B. 视觉合成可以生成难以在现实世界捕获的数据如交通冲突区的视觉内容

C. 视觉合成还可以减少对手工搜集和标注数据的依赖

1.3 研究现状

国内外现状分析

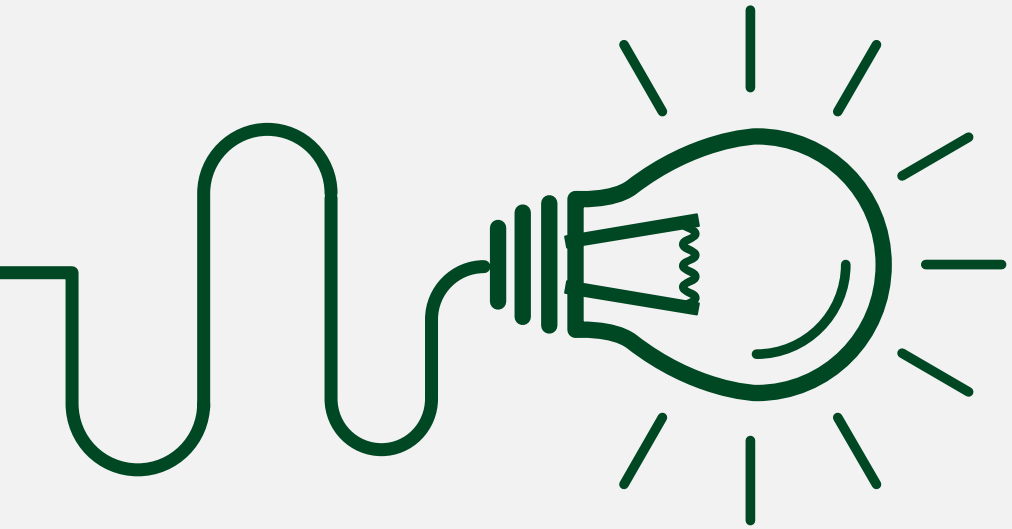
主要研究热点

1. 利用视觉自回归模型如DALL-E、CogView等，生成高分辨率的图像和视频。
2. 利用稀疏自注意力机制处理图像生成的问题。

研究应用领域

1. 在视觉设计、图像/ 视频制作、艺术创作和电商广告等众多领域有广泛应用。
2. 在医疗图像分析领域中有着至关重要的作用，可以用于医疗图像的生成、 分割 、重构、检测等。
3. 用于人工智能领域中数据集的扩充，解决数据集类别不均衡等问题。

1.4 主要创新点



1 提出了一个能统一表示文字、图像、视频和素描的标记规范。$X \in \mathbb{R}^{h \times w \times s \times d}$

2 定义了一个统一的3D Nearby Self-Attention(3DNA)模块，能够支持自注意力和交叉注意力。$Y = 3DNA(X, C; W)$

3 基于3DNA引入了3D Encoder-Decoder，实现了文本、图像和视频生成的大一统。

1.5 解读的具体过程

随着Web的普及，图像和视频成为新的信息载体并在许多实际应用中使用

⇩

视觉合成成为了越来越热门的研究课题

⇩

生成对抗网络

⇨

传统自回归模型

⇧

VQ-VAE：一种离散视觉标记化方法

⇧

NÜWA：一种统一的多模态预训练模型

旨在同时支持图像和视频的视觉合成任务

Part.02

解决方案与联系

深度生成模型的目标函数是数据分布与模型分布之间的距离，可以用**极大似然法**进行求解。从处理极大似然函数的方法的角度，可将传统深度生成模型分成如下三种，

深度生成模型概述

第一类方法是通过变分或抽样的方法求似然函数的近似分布

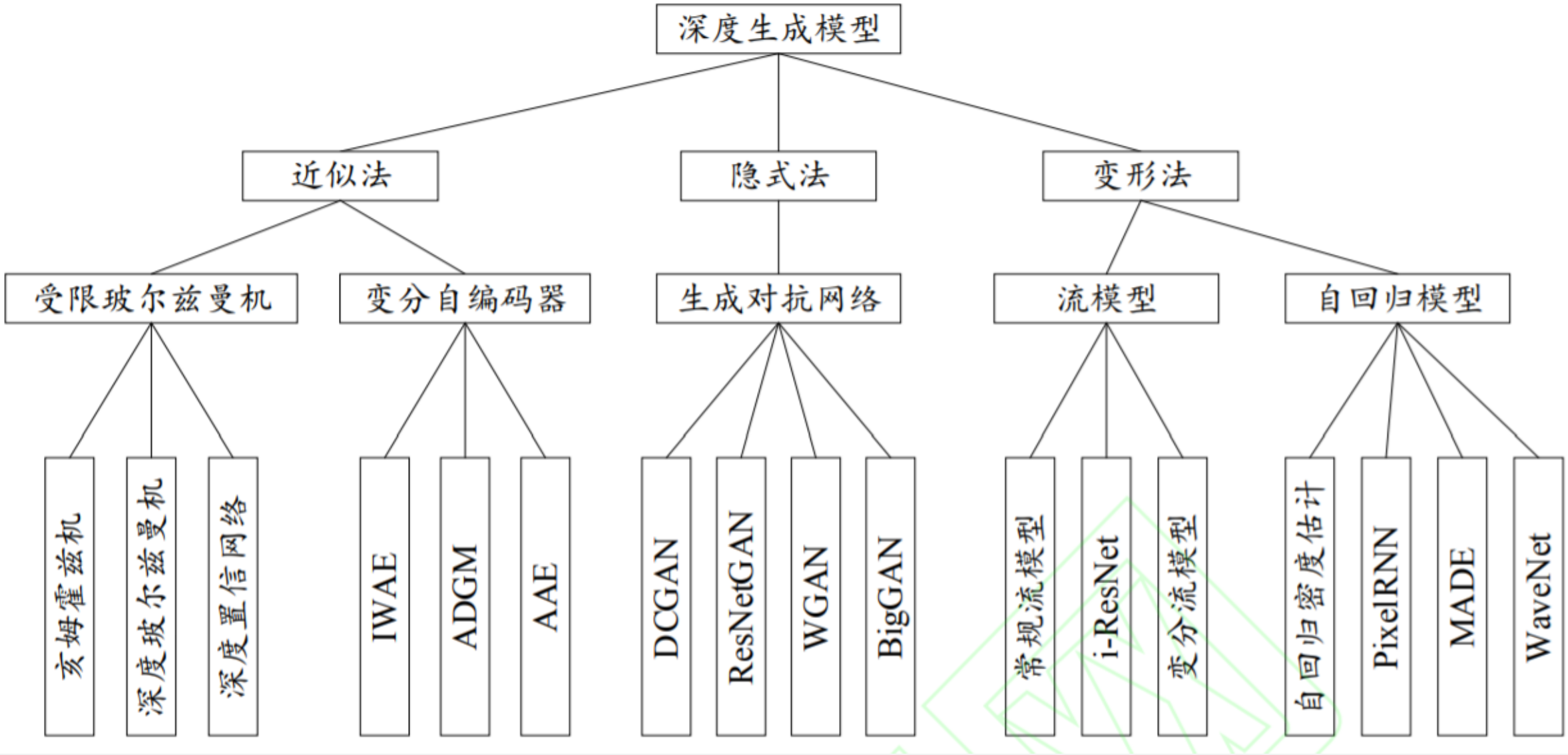
包括受限玻尔兹曼机，变分自编码器。**变分自编码器**用似然函数的变分下界作为目标函数。

第二类方法是避开求极大似然过程的隐式方法

代表模型是生成对抗网络。**生成对抗网络**利用神经网络的学习能力来拟合两个分布之间的距离，巧妙地避开了求解似然函数的难题。

第三种方法是对似然函数进行适当变形

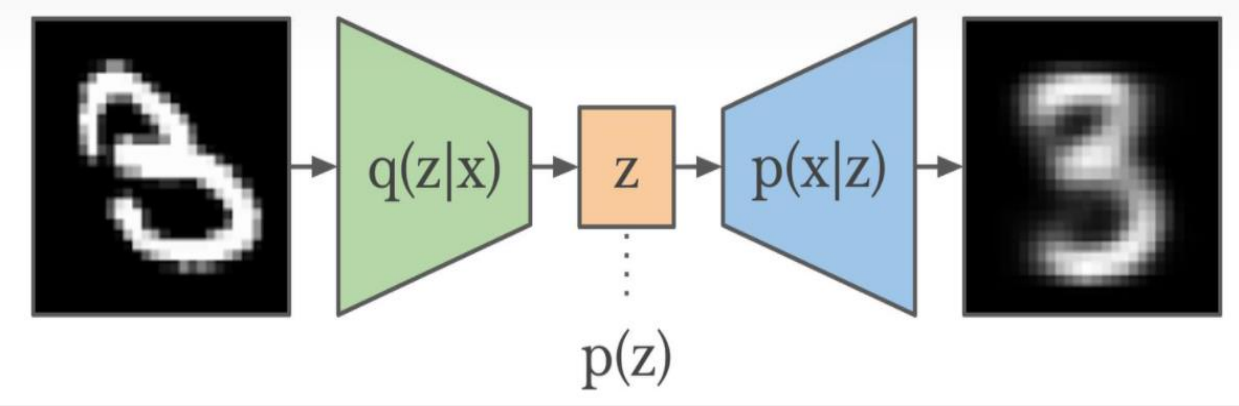
包括自回归模型等。**自回归模型**将目标函数分解为条件概率乘积的形式

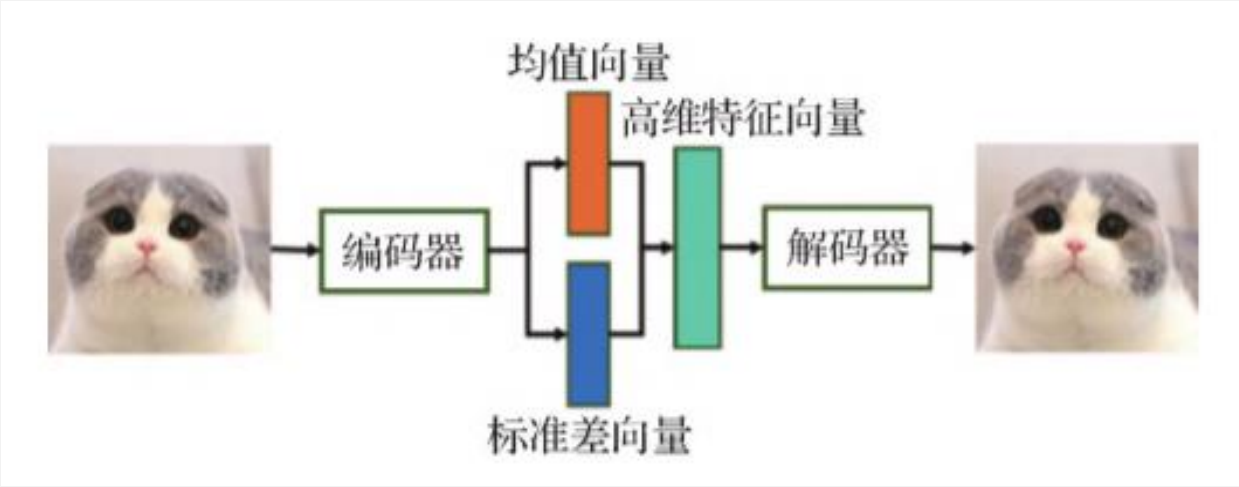

深度生成模型
近似法
隐式法
变形法
受限玻尔兹曼机
变分自编码器
生成对抗网络
流模型
自回归模型
亥姆霍兹机
深度玻尔兹曼机
深度置信网络
IWAE
ADGM
AAE
DCGAN
ResNetGAN
WGAN
BigGAN
常规流模型
i-ResNet
变分流模型
自回归密度估计
PixelRNN
MADE
WaveNet

2.1 变分自编码器（VAE）

2.1.1 VAE简介

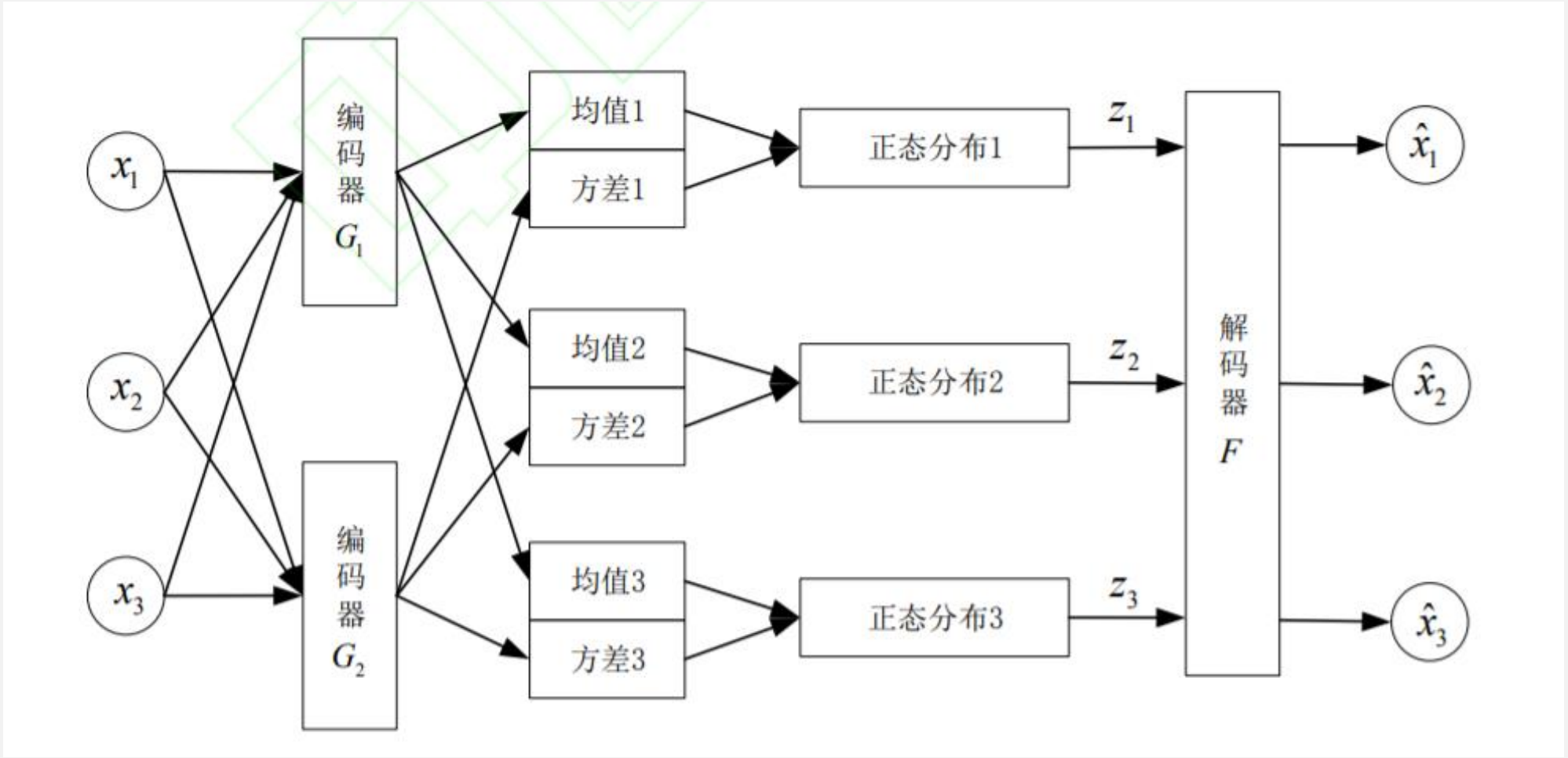
变分自编码器是 Kingma 和 Welling(2014)基于编码器(encoder)和解码器(decoder)结构提出的一种经典深度视觉生成模型。





2.1 变分自编码器（VAE）

2.1.2 VAE结构



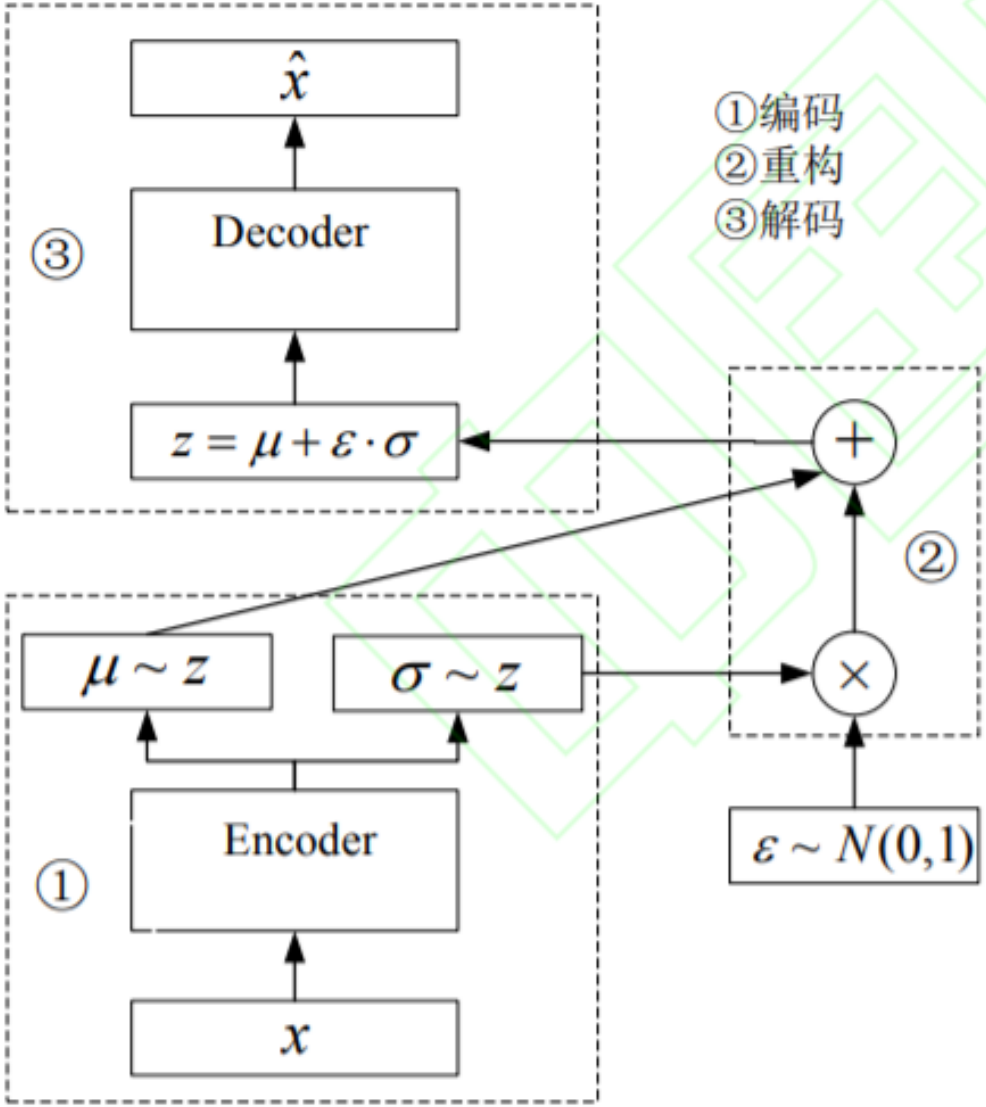
2.1 变分自编码器（VAE）

2.1.3 VAE训练过程

1. 第一个阶段是编码过程，样本通过两个神经网络分别获得正态分布的均值和方差。

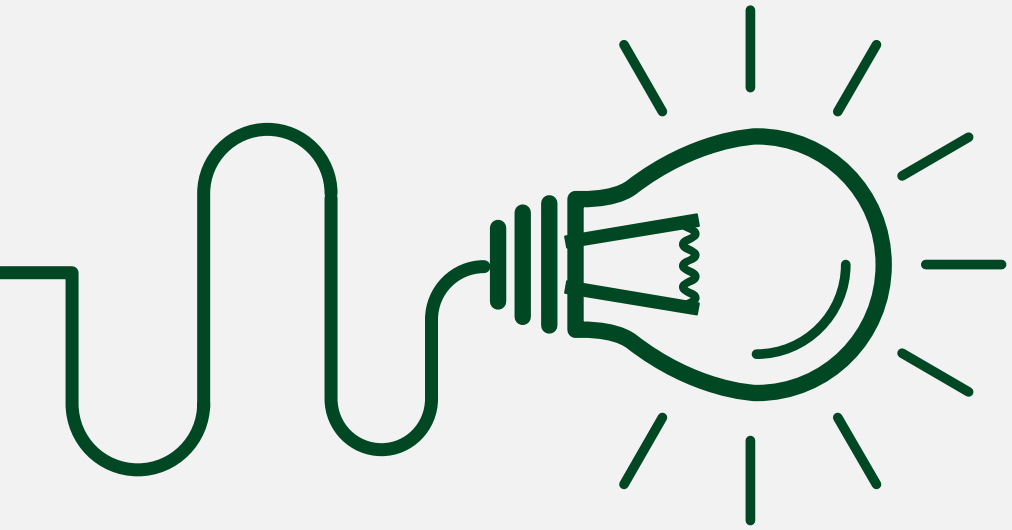
2. 第二个阶段是重参数化，以便从后验分布中抽样并能够用反向传播训练模型参数。

3. 第三个阶段是解码过程，将重参数化后的变量通过生成模型生成新样本。



2.1 变分自编码器（VAE）

2.1.4 VAE的优劣



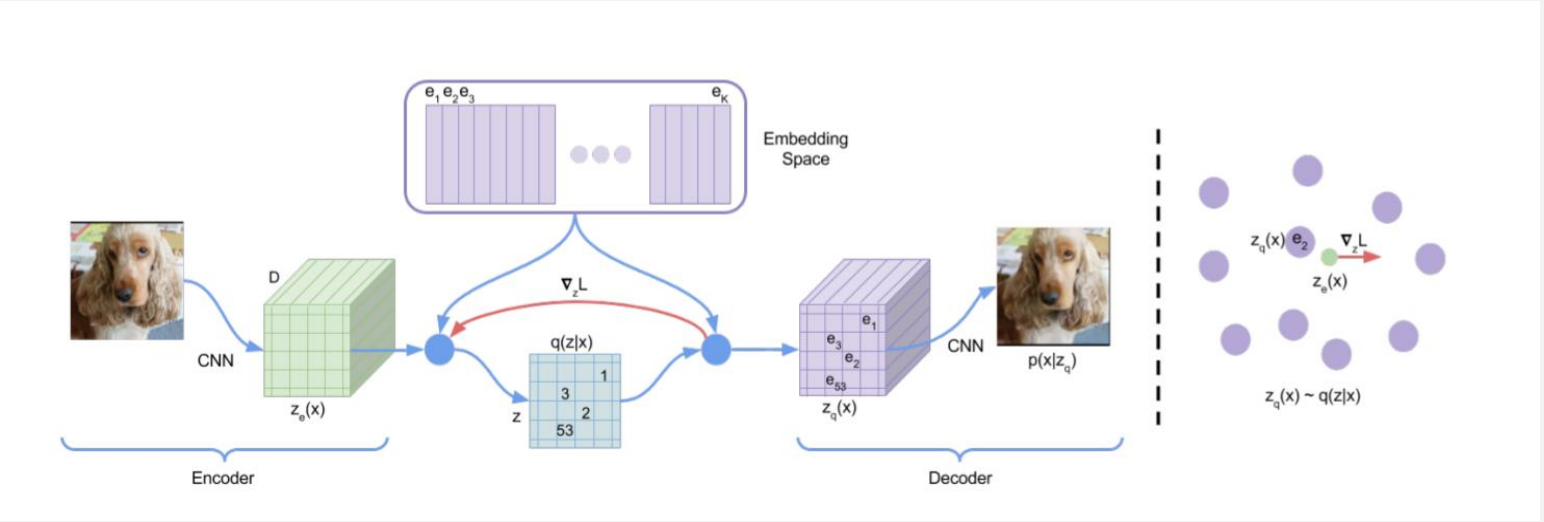
1 VAE广泛应用于数据降维和数据生成(Zhu等,2020a;Zhu等,2020b) 等方面,具有训练快、稳定等优势。

2 VAE 由于自身结构的固有缺点使模型生成的图片样本带有大量的噪声，大部分 VAE 结构很难生成高清的图片样本。

2.1 变分自编码器（VAE）

2.1.5 向量量化变分自编码器(VQ-VAE)

VQ-VAE是首个使用离散化隐藏变量的 VAE 模型。



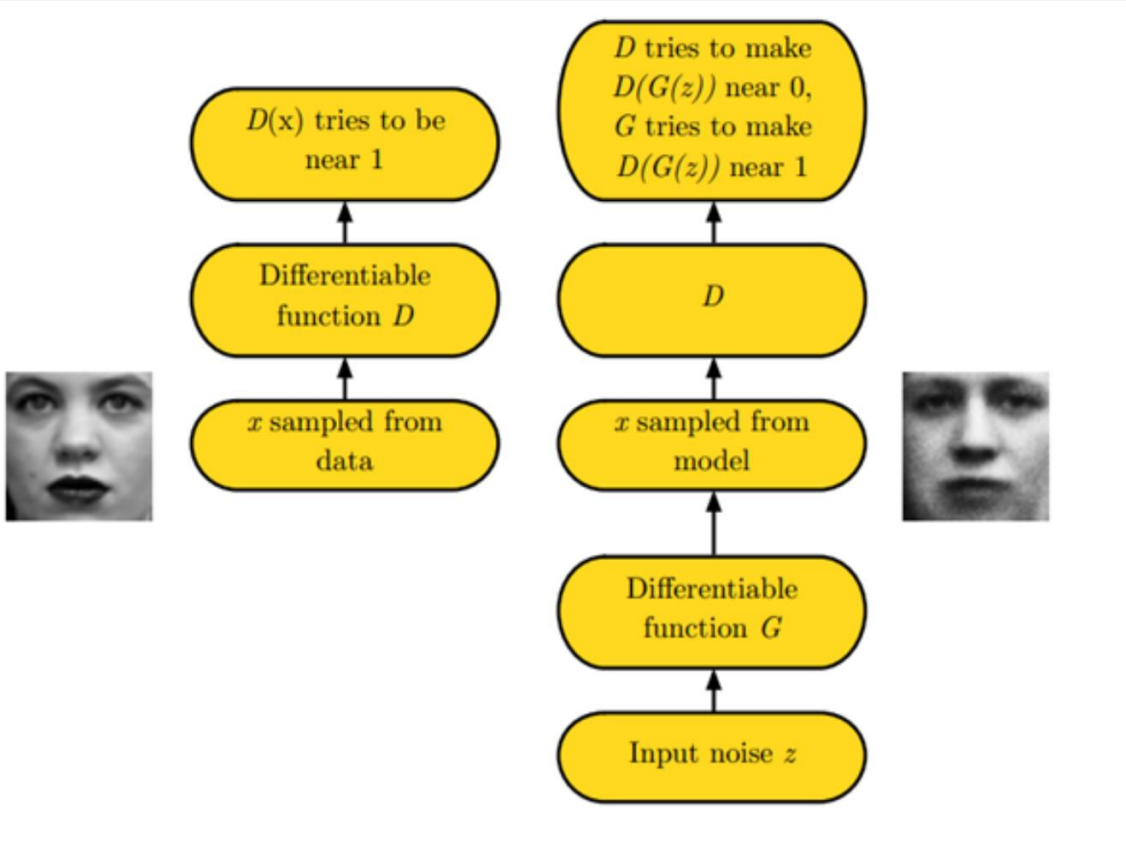
VQ-VAE 受到向量量化，方法的启发而提出了新的训练方法：后验概率分布和先验概率分布有明确分类，从这些分类明确的概率分布中提取样本，利用嵌入表示进行索引，得到的嵌入表示输入到解码器中。这种训练方法和有效的离散表达形式共同限制了解码器的学习过程，避免后验崩溃（posterior collapse）现象。

2.2 生成对抗网络（GAN）

2.2.1 GAN简介

生成对抗网络（Generative Adversarial Nets, GANs）在图像生成领域占有绝对优势。GAN 本质上是将难以求解的似然函数转化成神经网络，让模型自己训练出合适的参数拟合似然函数。

GAN 及其变体的本质是解决一个分布匹配问题(distribution matching problem)。 模型对已有数据进行学习,获得匹配已有数据分布(该分布通常很难直接描述)的能力,进而生成符合目标分布的图像或视频。



2.2 生成对抗网络（GAN）

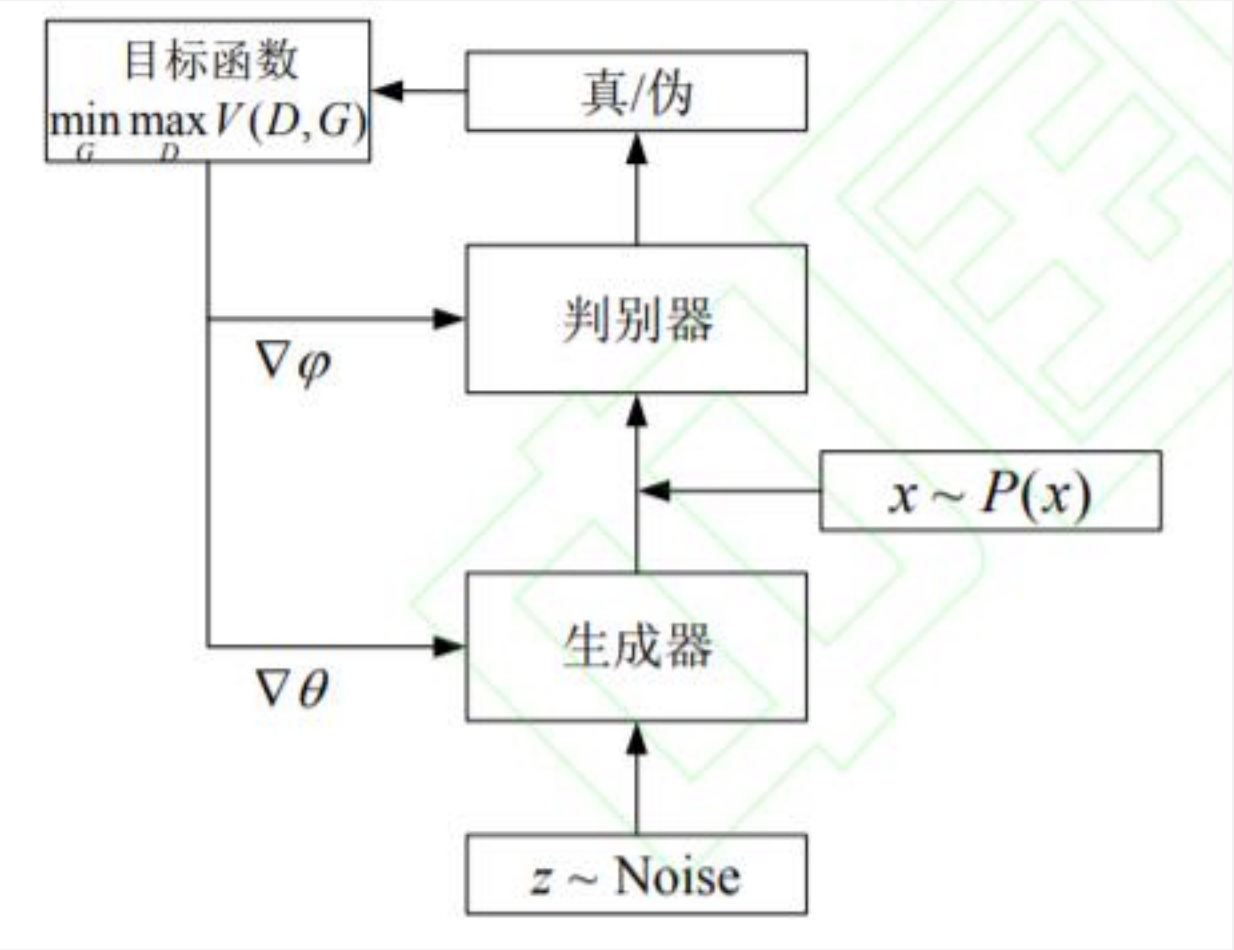
2.2.2 GAN模型结构

博弈方：一个生成器+一个判别器，

生成器的目标：是生成逼真的伪样本让判别器无法判别出真伪；
判别器的目标：是正确区分数据是真实样本还是来自生成器的伪样本。

在博弈的过程中，两个竞争者需要不断优化自身的生成能力和判别能力，而博弈的结果是找到两者之间的**纳什均衡。**

当判别器的识别能力达到一定程度却无法正确判断数据来源时，就获得了一个学习到真实数据分布的生成器。



2.2 生成对抗网络（GAN）

2.2.3 GAN训练算法

GAN 的训练机制由**生成器优化**和**判别器优化**两部分构成，

优化判别器：固定生成器后 $G(z;\theta)$ ，优化判别器 $D(x;\varphi)$ ，由于判别器是二分类模型，目标函数选用交叉熵函数：

$$\max_D V(D) = \mathrm{E}_{x\sim P_r}[\log D(x)] + \mathrm{E}_{x\sim P_g}[\log(1-D(x))]$$

优化生成器：固定训练好的判别器参数，考虑优化生成器模型参数。生成器希望学习到真实样本分布，因此优化目的是生成的样本可以让判别器误判为 1，即最大化 $\mathrm{E}_{x\sim\rho_s}[\log(D(x))]$ ，所有生成器的目标函数为：

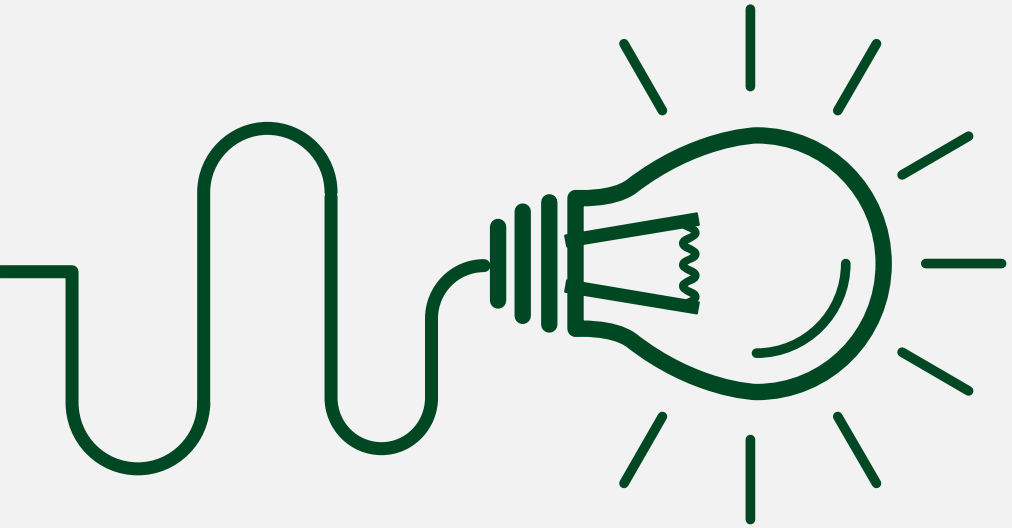
$$\min_G V(G)\mathrm{E}_{x\sim P_g}[\log(1-D(x))]$$

后又提出改进函数：

$$\min_G V(G)\mathrm{E}_{x\sim p_g}[-\log D(x)]$$

2.2 生成对抗网络（GAN）

2.2.4 初代GAN算法的缺陷



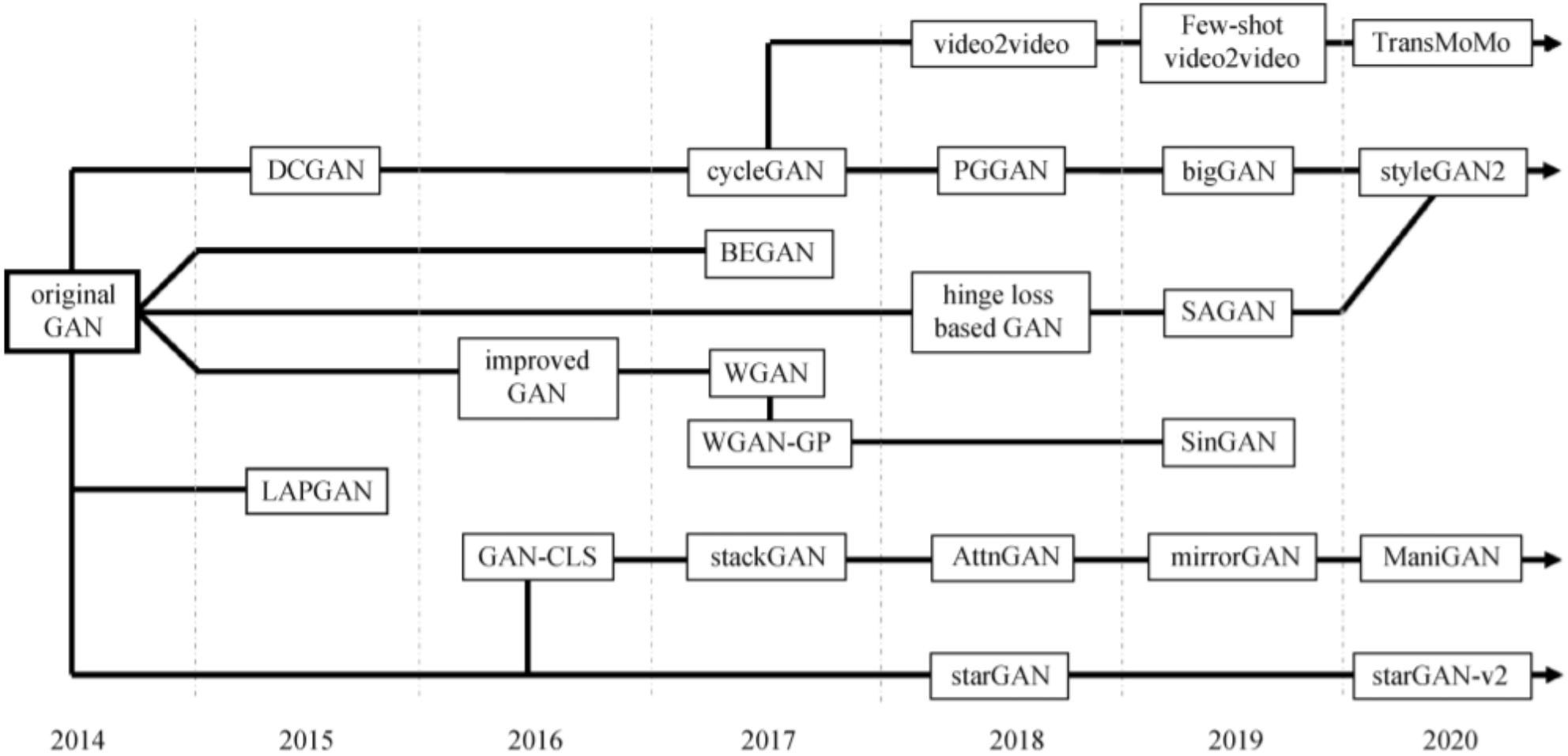
1 模型难以训练，经常出现梯度消失导致模型无法继续训练；生成器形式过于自由，训练时梯度波动极大造成训练不稳定；需要小心地平衡生成器和判别器的训练程度，使用更新一次判别器后更新k 次生成器的交替训练法并不能很好的缓解训练问题；

2 出现模式崩溃（model collapse），具体表现为生成样本单一，无法生成其它类别的样本；

3 目标函数的形式导致模型在训练过程没有任何可以指示训练进度的指标。

2.2 生成对抗网络（GAN）

2.2.5 GAN发展历程



2.3 自回归模型（AR）

2.3.1 AR简介

自回归是统计学中处理时间序列的方法，用同一变量之前各个时刻的观测值预测该变量当前时刻的观测值。用条件概率表示可见层数据相邻元素的关系，以条件概率乘积表示联合概率分布的模型都可以称为自回归网络。

AR序列建模过程

1. 步骤 1 对序列作白噪声检验，若经检验判定序列为白噪声，建模结束；否则转步骤 2.
2. 步骤 2 对序列作平稳性检验，若经检验判定为非平稳，则进行序列的平稳化处理，转步骤 1；否则转步骤 3.
3. 步骤 3 对模型进行识别，估计其参数，转步骤 4.
4. 步骤 4 检验模型的适用性，若检验通过，则得到拟合模型并可对序列做预测；否则转步骤 3.

2.3 自回归模型（AR）

2.3.2 AR的基本形式

三种基本形式

线性自回归网络

线性自回归网络是自回归网络中最简单的形式，没有隐藏单元、参数和特征共享。

神经自回归网络

神经自回归网络的提出是为了用条件概率分解似然函数，避免如 DBN 等传统概率图模型中高维数据引发的维数灾难。

NADE

2.3 自回归模型（AR）

2.3.3 NADE

观测数据的有序排列起源于完全可见贝叶斯网络（Fully Visible Bayes Nets, FVBN），该算法最早定义了将高维数据的概率通过链式法则分解为条件概率乘积的方法。神经自回归分布估计模型（Neural Autoregressive Distribution Estimation，NADE)根据这种方法进行建模：

$$P(x)=\prod_{d=1}^{D}P(x_{o_d}\mid x_{o_d})$$

其中 $x_{o_{<d}}$ 表示观测 D 维观测数据中位于 x_{o_d} 左侧的所有维数，表明该定义中第 i 个维数的数值只与其之前的维数有关，与之后的维数无关。

2.3 自回归模型（AR）

2.3.4 PixelRNN

像素循环神经网络（Pixel Recurrent NeuralNetwork, PixelRNN）将图片的像素作为循环神经网络的输入，本质上是自回归神经网络在图片处理上的应用，该模型利用深度自回归网络预测图片的像素值，并提出三种不同结构的深度生成模型。

PixelCNN

该模型直接利用卷积神经网络（Convolutional Neural Network, CNN）处理像素，然后用特殊结构的掩码避免生成样本时出现缺少像素的问题。

Row LSTM

这种模型结构能捕捉到更多邻近像素的信息，该模型对 LSTM 的输出进行行卷积，且三个门也由卷积产生。

Diagonal BiLSTM

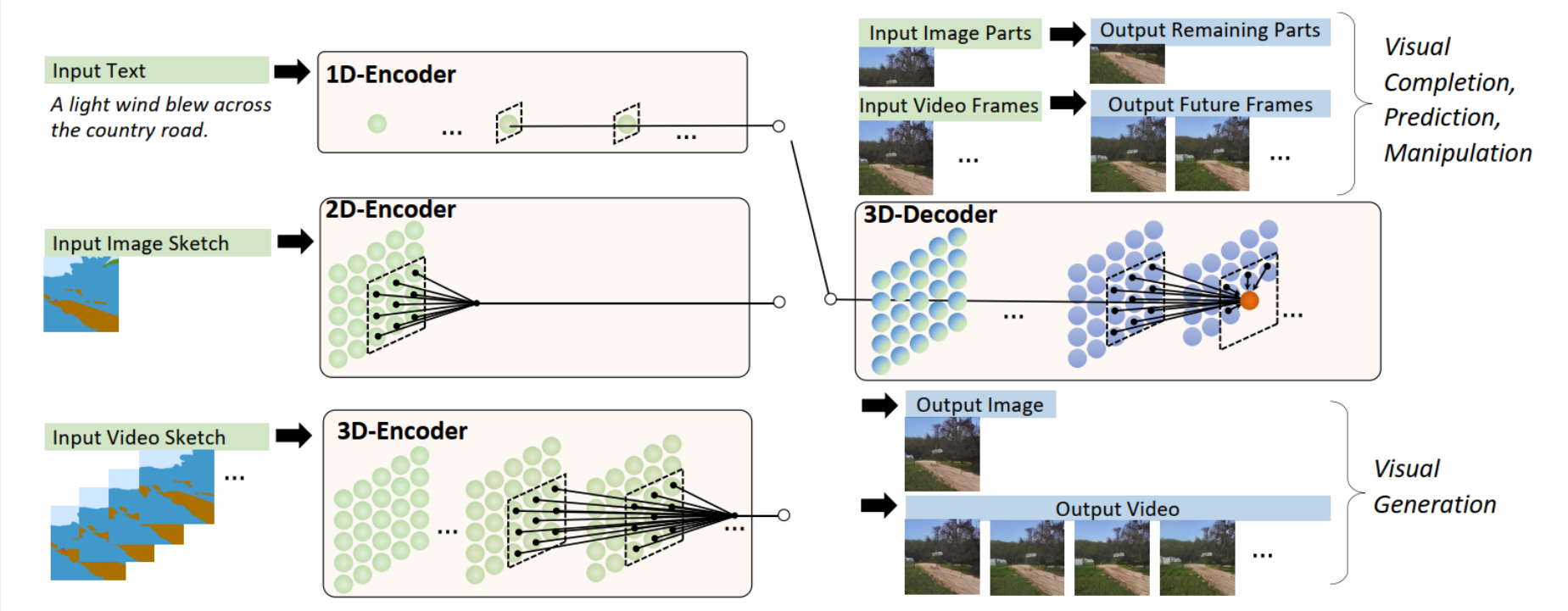
该模型通过重新构造像素位置的方法使 LSTM 的输入不存在遗漏像素，即双向长短时记忆网络 BiLSTM。

2.4 NÜWA算法

2.4.1 NÜWA概述

NÜWA包含一个支持不同条件的自适应编码器和一个受益于图像和视频数据的预训练解码器。

对于图像完成、视频预测、图像处理和视频处理任务，输入的部分图像或视频都能直接输入到解码器中。



2.4.2 关键技术

- **3D Data Representation**

用 $X \in \mathbb{R}^{h \times w \times s \times d}$ 统一表示文本、图像、视频和素描数据。 其中h和w分别表示height和weight，s表示时间轴上token的数量，d表示每个token的维度。

Texts 仿照transformer编码，用小写字节对编码(BPE)的方法将文本用 $\mathbb{R}^{1 \times 1 \times s \times d}$ 表示，其中占位符1是为了表示它不具空间维度。

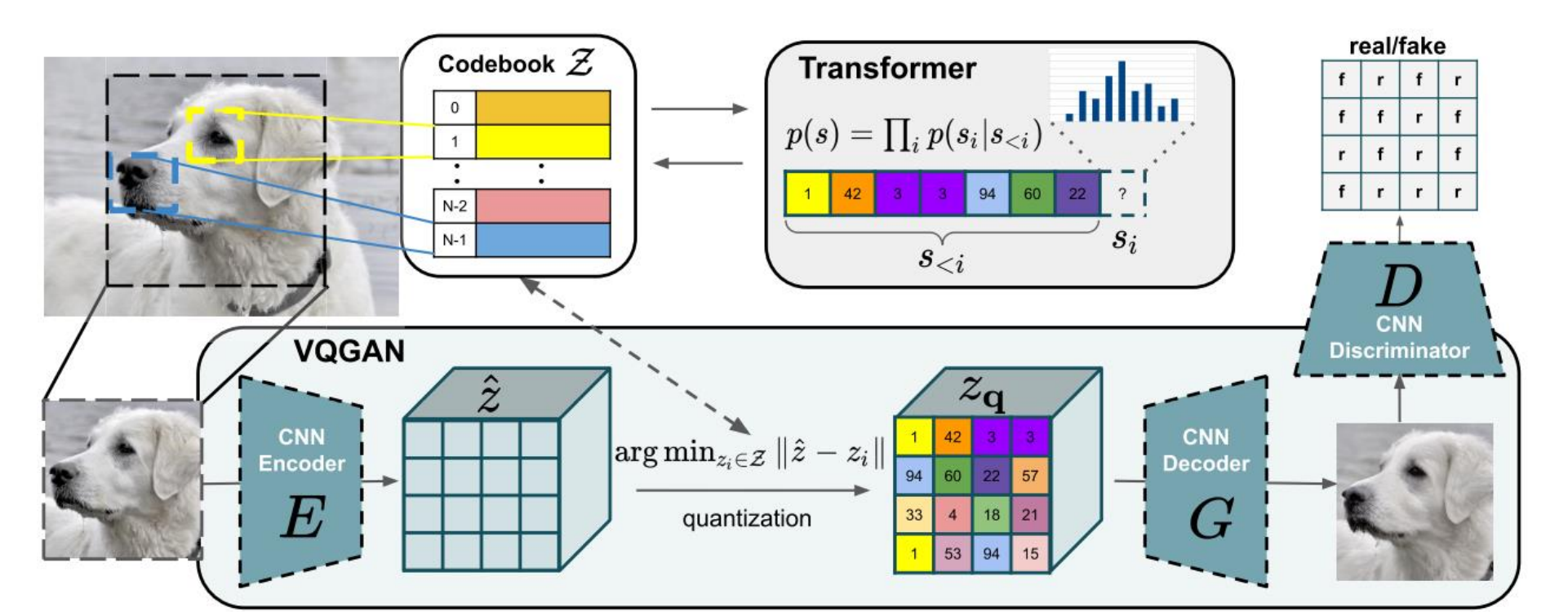
Images 通过输入图片，训练VQ-GAN，将最终得到的 $B[z] \in \mathbb{R}^{h \times w \times 1 \times d}$ 来表示图片，其中占位符1是为了表示它在时间轴上不具token。

Videos 视频可以看成是许多图片在时间轴上变化的数据，本文使用2D VQ-GAN编码视频的每一帧，将得到的结果 $\mathbb{R}^{h \times w \times s \times d}$ 用来表示视频，其中s用来表示视频的帧数。

Sketches 素描可以看成是具有特殊通道的图片。用上述方法，可用 $\mathbb{R}^{h \times w \times 1 \times d}$ 表示素描图片，用 $\mathbb{R}^{h \times w \times s \times d}$ 表示素描视频。

2.4.2 关键技术

- 3D Data Representation中的VQ-GAN



2.4.2 关键技术

- **3D Nearby Self-Attention**

基于3D Data Representation，定义3D Nearby Self-Attention为 $Y = 3DNA(X, C; W)$ 。

其中X和C均为上述基于3D Data Representation得到的数据，W为可学习的权重。

当X=C时，Y表示目标X的自注意力；当X≠C时，Y表示目标X以C为条件的的交叉注意力。

2.4.2 关键技术

- **3D Encoder-Decoder**

基于3D Data Representation 和 3D Nearby Self-Attention，为了在 $C \in \mathbb{R}^{h' \times w' \times s' \times d^{in}}$ 的条件下生成目标 $Y \in \mathbb{R}^{h \times w \times s \times d^{out}}$ ，C和Y的位置编码可由下面的公式进行迭代更新。

$$Y_{ijk} := Y_{ijk} + P_i^h + P_j^w + P_k^s$$

$$C_{ijk} := C_{ijk} + P_i^{h'} + P_j^{w'} + P_k^{s'}$$

2.4.2 关键技术

- **3D Encoder-Decoder**

随后，将条件C输入到具有 l 层的3DNA编码器以模拟自注意力交互，第 l 层计算公式如下：

$$C^{(l)} = 3DNA(C^{(l-1)}, C^{(l-1)})$$

类似地，解码器由一叠 l 层的3DNA构成，同时计算生成结果的自注意力和生成结果与条件的交叉注意力，第 l 层计算公式如下：

$$Y_{ijk}^{(l)} = 3DNA(Y_{<i,<j,<k}^{(l-1)}, Y_{<i,<j,<k}^{(l-1)}) + 3DNA(Y_{<i,<j,<k}^{(l-1)}, C^{(L)})$$

2.4.3 实验步骤

01

设置3D大小表示：

文本，三维表示的大小:1×1×77×1280
图像，三维表示的大小:21×21×1×1280
视频，三维表示的大小:21×21×10×1280

02

图像与视频的VQ-GAN模型

设置:

网格特征的大小441×256，
码本的大小为12288

03

稀疏范围设置：

文本设置为$(e^w, e^h, e^s) = (1, 1, \infty)$
图像和图像草图设置为 $(e^w, e^h, e^s) = (3, 3, 1)$
视频和视频草图设置为 $(e^w, e^h, e^s) = (3, 3, 3)$

04

训练参数设置：

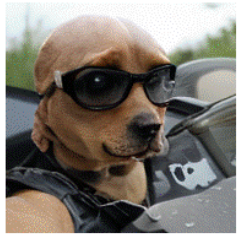
在64个A100 GPU上进行为期两周的预训练图层设置为24，使用Adam优化器，学习率为1e-3，批量大小为128，预热总步骤为50M的5%

2.5 结果展示与对比

NÜWA是一个统一的多模态预训练模型，可以生成新的或处理现有的视觉数据（即图像和视频），用于以下8个视觉合成任务。

Text-To-Image (T2I)

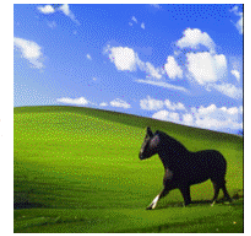
A sad looking puppy staring at the camera.



Sketch-To-Image (S2I)

Image Completion (I2I)

Image Manipulation (TI2I)

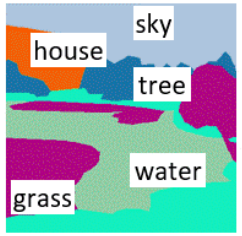


a horse is running on the grassland

Text-To-Video (T2V)

A person is preparing some art.

Sketch-To-Video (S2V)



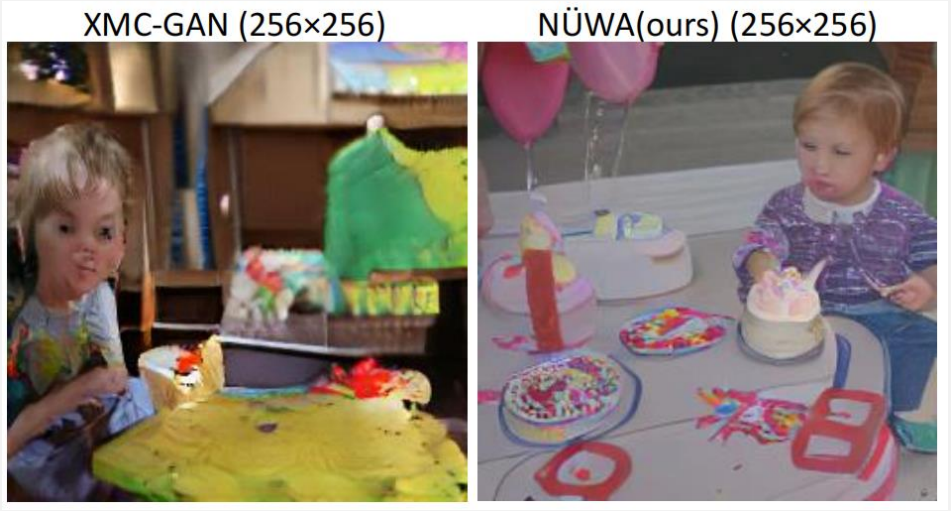
Video Prediction (V2V)

Video Manipulation (TV2V)

The car is reversing

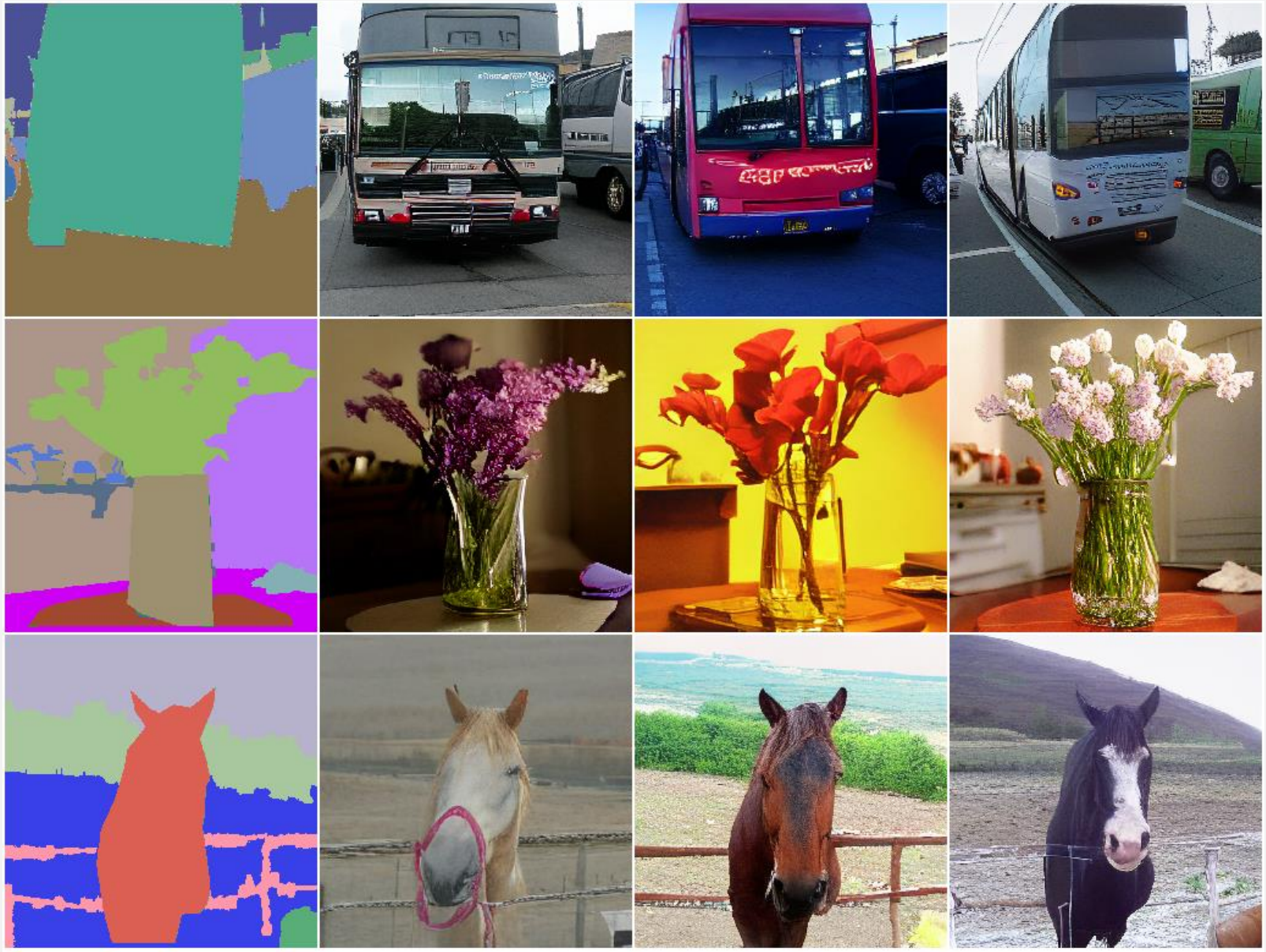
2.5 结果展示与对比

文本 - 图像 (T2I)：
通过比较NÜWA在 MSCOCO 数据集上的性能可以发现，NÜWA明显优于 CogView，其中 FID-0 为 12.9，CLIPSIM 为 0.3429 。尽管 XMC-GAN 的 FID-0 为 9.3，优于NÜWA，但NÜWA能生成更逼真的图像，如下图所示。NÜWA生成的男孩脸更清晰，并且男孩旁边的气球也很逼真。



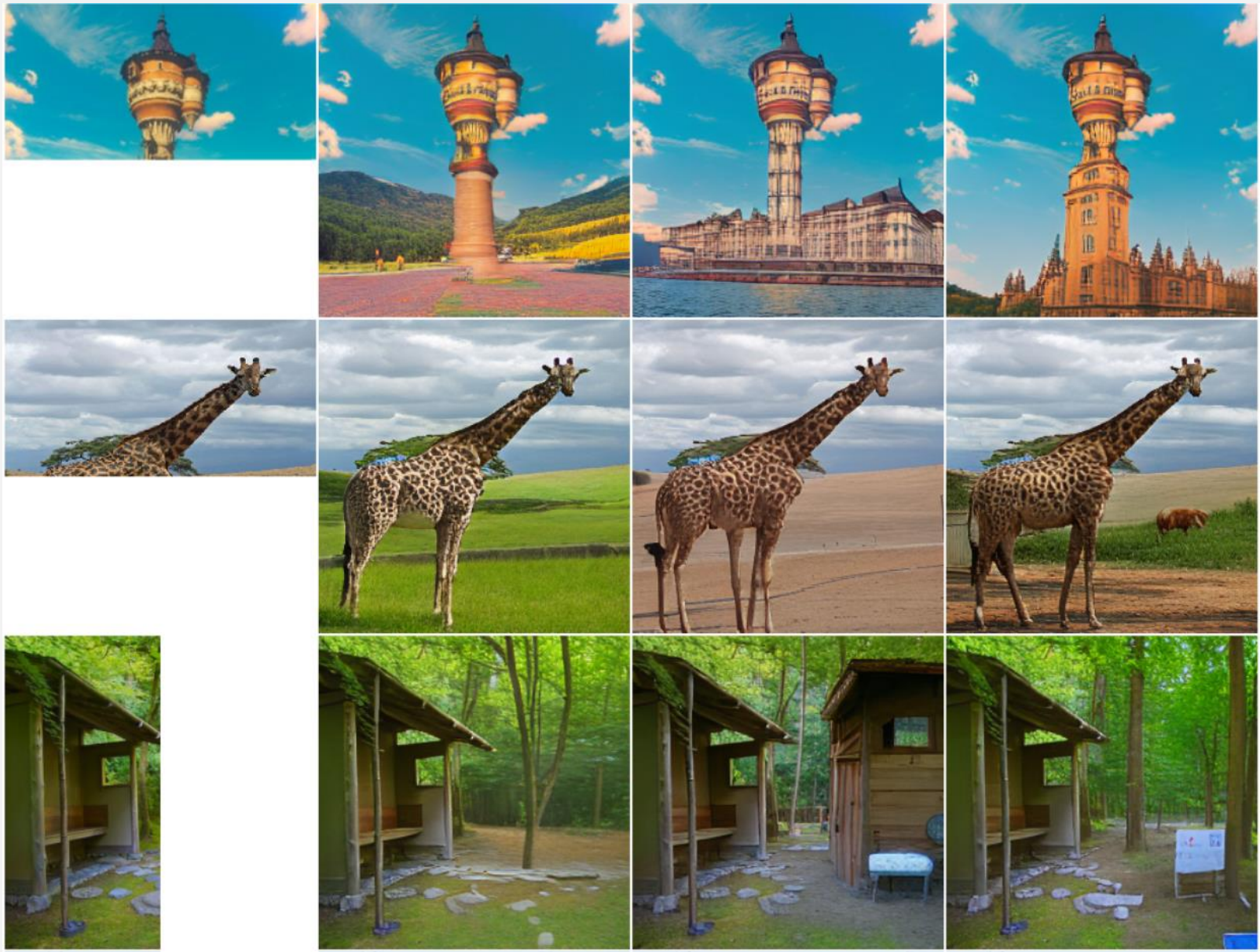
2.5 结果展示与对比

草图 - 图像 (S2I):
NÜWA算法在草图向图像的转换在MSCOCO stuff 上进行。如下图 所示，NÜWA算法与 Taming-Transformers 和 SPADE 相比，NÜWA算法生成了种类繁多的逼真汽车， 甚至巴士车窗的反射也清晰可见。



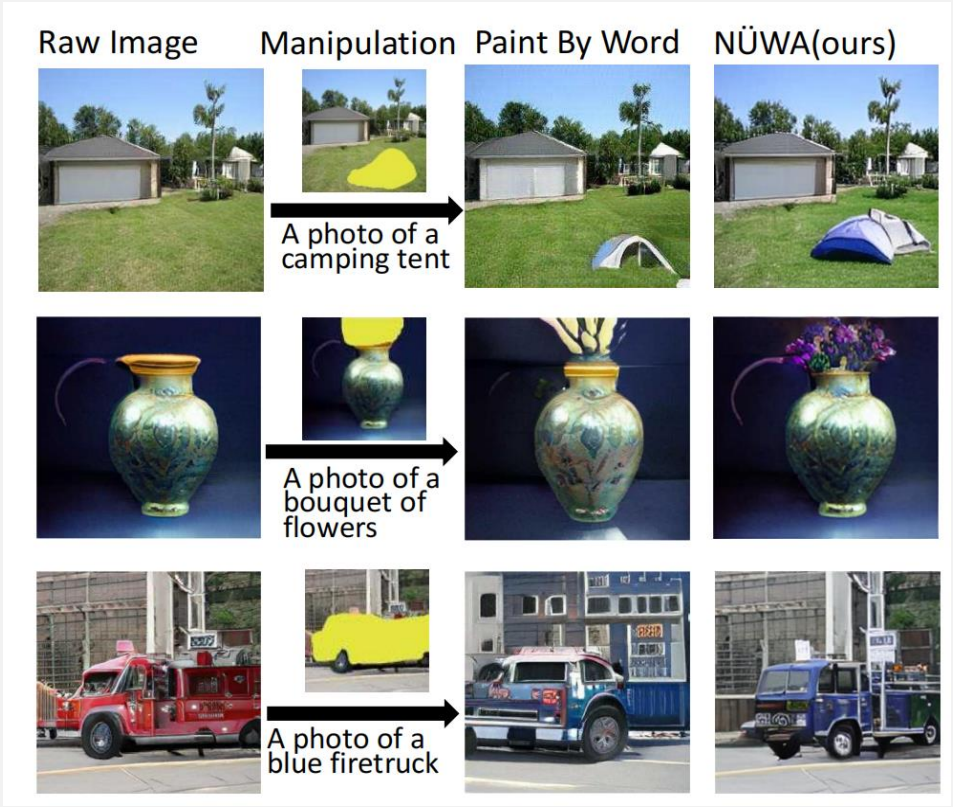
2.5 结果展示与对比

图像补全 (I2I) 零样本评估：
给定塔楼的上部，与 Taming Transformers 模型进行比较，NÜWA算法可以生成对塔楼下半部分更丰富的想象，包括生成周围建筑物、湖泊、花草、树木、山脉等。



2.5 结果展示与对比

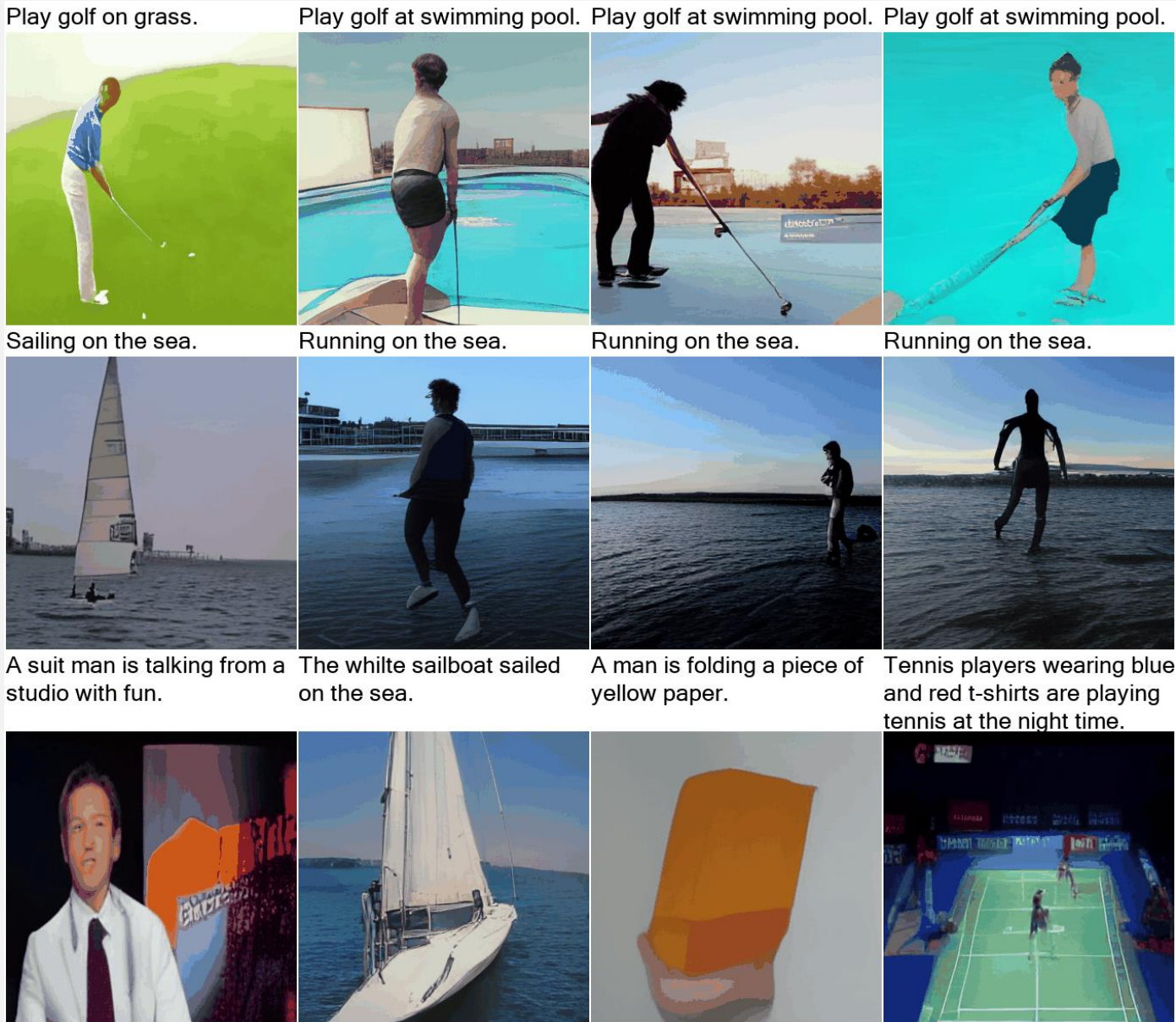
文本 - 指导图像处理 (TI2I) 零样本评估：NÜWA算法显示了其强大的处理能力，可以生成高质量的文本一致性结果，而不会改变图像的其他部分。



2.5 结果展示与对比

文本 - 视频 (T2V)：
在 Kinetics 数据集上评估NÜWA算法的结果如图所示，NÜWA算法在所有指标上实现了最好的性能。
另外在右图还展示了NÜWA算法强大的零样本生成能力，可以生成没见过的图像，例如：在游泳池里打高尔夫球，在海里奔跑等。

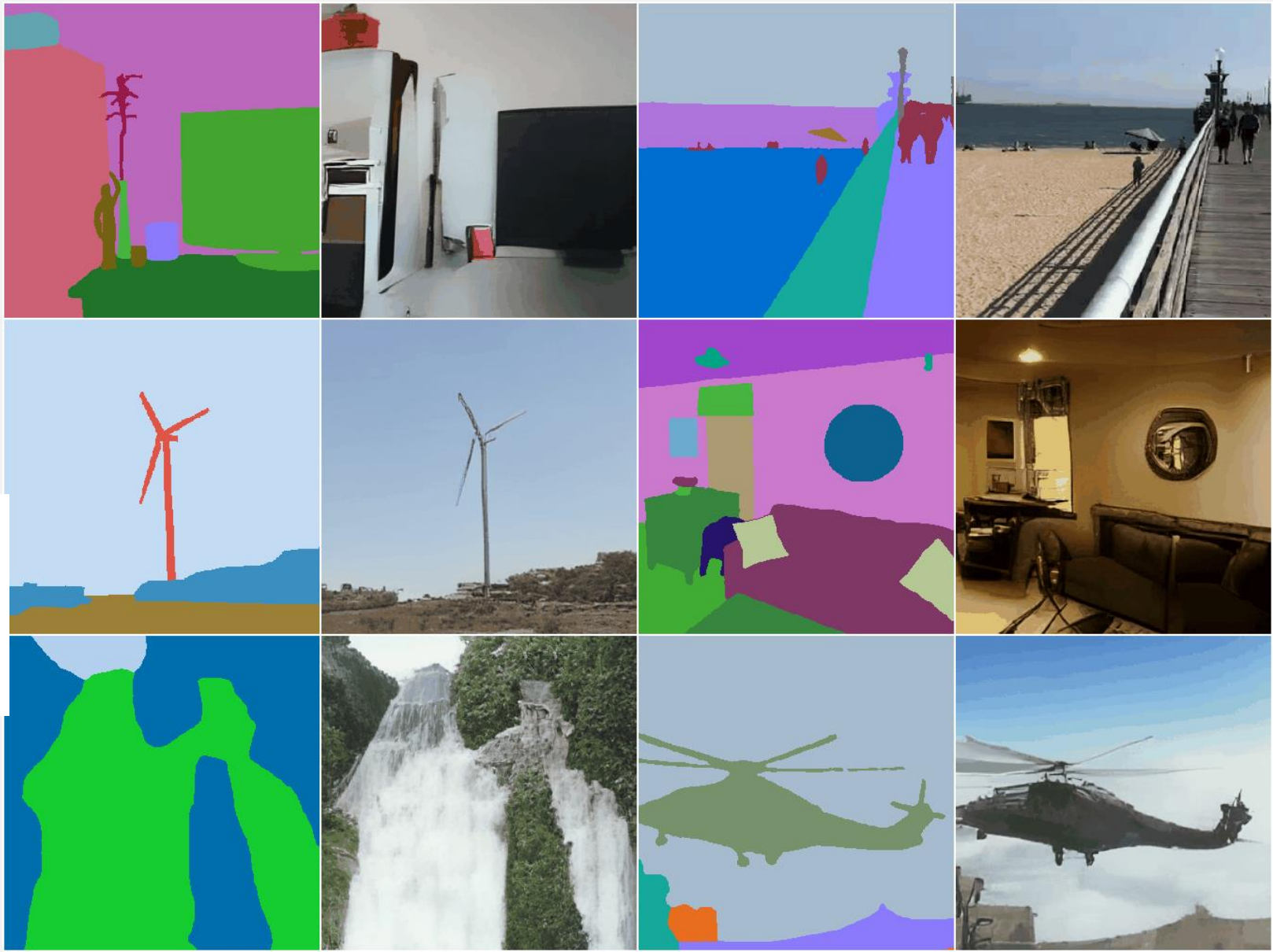
Model	Acc↑	FID-img↓	FID-vid↓	CLIPSIM↑
T2V (64×64) [21]	42.6	82.13	14.65	0.2853
SC (128×128) [2]	74.7	33.51	7.34	0.2915
TFGAN (128×128) [2]	76.2	31.76	7.19	0.2961
NÜWA (128×128)	**77.9**	**28.46**	**7.05**	**0.3012**



2.5 结果展示与对比

草图 - 视频(S2V):
在 VSPW 数据集上评估NÜWA算法的结果如图所示，NÜWA算法在所有指标上实现了最好的性能。
另外在右图还展示了NÜWA算法强大生成能力，可以生成逼真的视频片段。

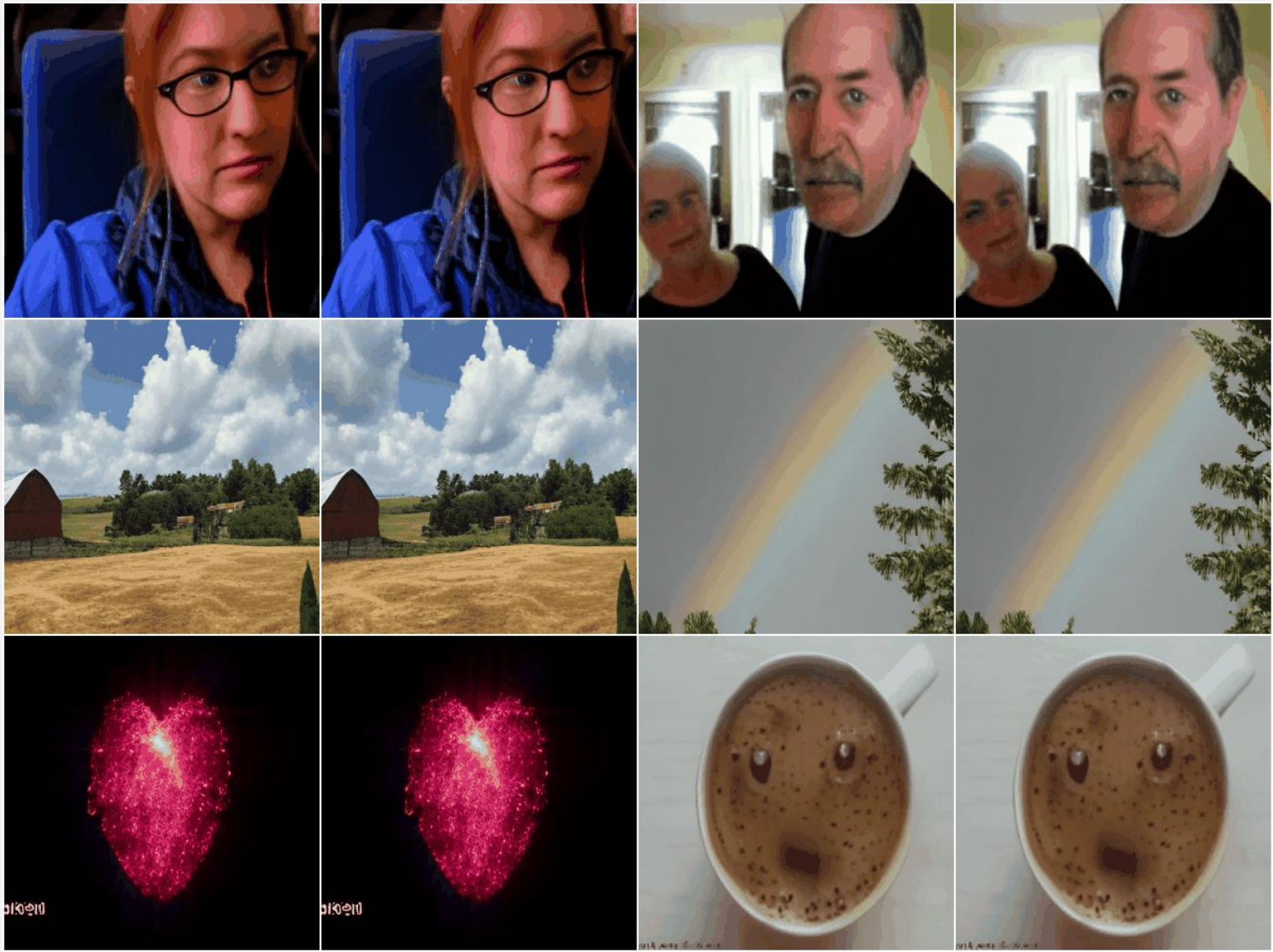
Model	Encoder	Decoder	FID-vid↓	Detected PA↑
NÜWA-FF	Full	Full	35.21	0.5220
NÜWA-NF	Nearby	Full	33.63	0.5357
NÜWA-FN	Full	Nearby	32.06	0.5438
NÜWA-AA	Axis	Axis	29.18	0.5957
NÜWA	Nearby	Nearby	**27.79**	**0.6085**



2.5 结果展示与对比

视频预测 (V2V)：
该研究在 BAIR Robot Pushing 数据集上对NÜWA和其他模型进行了比较，结果如下表所示。为了进行公平比较，所有模型都使用 64×64 分辨率。**虽然只给出了一帧作为条件（Cond.），**但NÜWA仍然将 SOTA FVD 得分从 94±2 降到 86.9 。

Model	Cond.	FVD↓
MoCoGAN [37]	4	503
SVG-FP [8]	2	315
CNDA [12]	2	297
SV2P [1]	2	263
SRVP [13]	2	181
VideoFlow [18]	3	131
LVT [31]	1	126±3
SAVP [20]	2	116
DVD-GAN-FP [7]	1	110
Video Transformer (S) [44]	1	106±3
TriVD-GAN-FP [23]	1	103
CCVS [25]	1	99±2
Video Transformer (L) [44]	1	94±2
NÜWA	1	**86.9**



2.5 结果展示与对比

文本指导视频处理(TV2V):
第一张显示了原始视频帧，潜水员在潜水；第二张为潜水员正在向水面游；第三张显示可以让潜水员游到海底，另外NÜWA还可以实现生成让潜水员飞向天空的图片，从第四张图中可以看出，潜水员像火箭一样飞向天空。

Part.03

小组新见解

3.1 发展趋势

发展趋势概述

NLP和CV的趋势是统一的

-相似的主干（转化器）
-相似的表示格式
（文本标记和离散的视觉标记）
-相似的预训练任务
（自动回归解码、
指代自动编码、
对比性学习）

多模态AI成为创新的前沿阵地

-建立通用的视觉-语言表征
（预训练）
-从文本中生成图像/视频
-为多模态AI任务设计新的基准和衡量标准
-设计新的模型可视化和解释机制

AI将应用于视觉内容创作

-多模态搜索引擎/问题搜索/对话系统
-视觉广告/ppt/新闻创作
-AI辅助的图像/视频编辑
-AR/VR/Metaverse等

3.2 实践难点

样本多样性	如何使深度生成模型生成的图像、文本和语音等样本具有多样性是一个值得研究的问题。度量多样性最基本的标准是熵，把训练样本看作多个概率分布的噪声混合后的随机变量，提取不同噪声的特征表示，得到不同层次的特征表示，在训练目标函数里显式地引入不同的归纳偏置。
泛化能力	机器学习理论认为好的模型要具有更好的泛化能力。重新思考深度学习的泛化能力，从模型复杂性、偏差-方差权衡等观点，理论上讨论各种深度生成模型的学习机制，丰富模型的理论基础
效率	代表着最先进的一批生成模型如 BigGAN 、 Glow 和VQ-VAE等已经可以生成足够清晰的图片样本，但这样的大型模型背后是远超常规的计算量，是所有大型生成模型的弊端

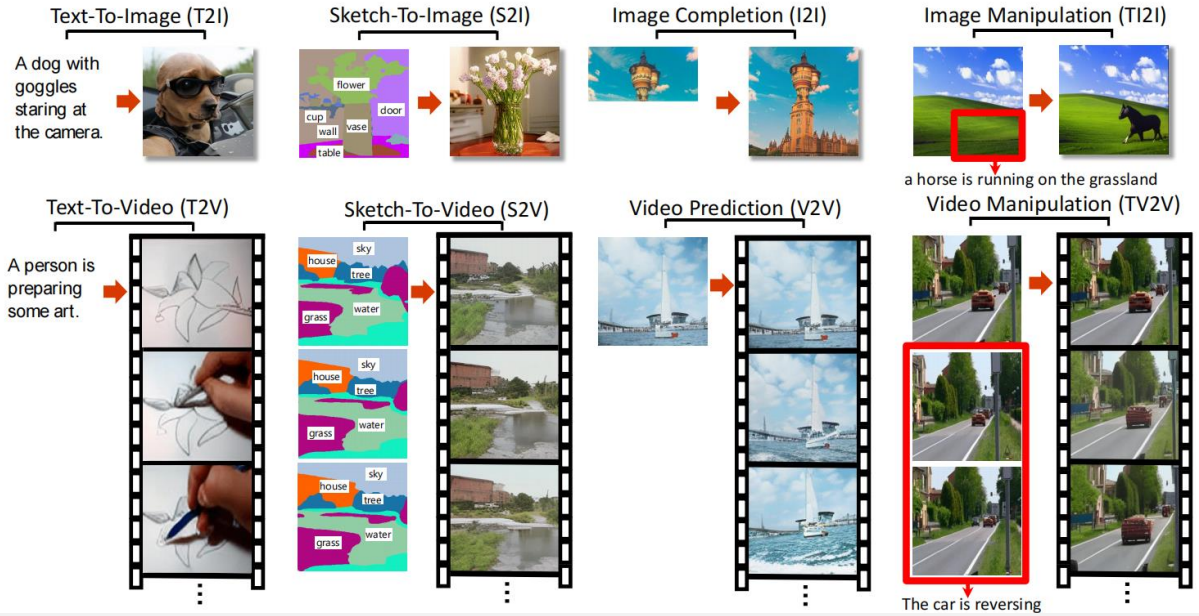
3.3 解决方法


样本多样性
进行数据增强算法的研究
泛化能力
优化模型的同时，提高数据集的规模与难度
效率
通过一些条件独立性来简化模型

Part.04

论文总结

4.1 NÜWA算法总结



NÜWA是一个统一多模态预训练模型，该模型可以为各种视觉合成任务生成新的或操作现有的视觉数据（即图像和视频）。

为了在不同场景下同时覆盖语言、图像和视频，NÜWA设计了一个3D transformer编解码框架，该框架不仅可以将视频作为3D数据处理，还可以分别将文本和图像作为一维和二维数据处理。

NÜWA提出了一种3D邻近注意力（3DDNA）机制来考虑视觉数据的性质，从而达到降低计算复杂度的效果。

与其他同类算法相比，NÜWA在8项下游任务中均获得了非常优秀的性能。

4.2 NÜWA获得的成就

01 对图像与视频具有强大的处理能力，可以生成高质量的与文本一致性结果，而不会改变图像或视频的其他部分。

02 提出的3D邻近注意(3DNA)机制不仅降低了计算复杂度，而且提高了生成结果的视觉质量。

03 在多个领域都取得了SOTA的结果，例如文本生成图片、文本生成视频、视频预测等领域。

4.3 亮点与不足

亮点之处

- 提出了一个大一统的、涵盖了文本、图像、视频的3D编码解码器；
- 提出了一个结合了空间和时间特征的近邻稀疏注意力机制；
- 进行了全面的实验，包括视觉合成最主要的8个任务。

不足之处

- 虽然该模型能实现热门的8个任务，但是不能实现逆转化，例如从images到texts和videos到texts的转化。
- 作为预训练模型，其鲁棒性还有待提高。
- 生成的图像或视频仍无法与人类的视觉认知达成高度一致。

4.4 面临的挑战

01 评估指标

由于训练过程复杂、结构不易理解和使用、训练速度慢等问题，在大规模数据上学习模型很困难，在不同的应用领域应该有相应的有效评估指标和实用的评估系统。

02 不确定性

实际过程限于求解的难度不得不进行近似和简化，使模型偏离原来的目标。训练好的模型难以在理论上分析透彻，只能借助实验结果反向判断调整方法，对生成模型的训练造成很大困扰

03 应用领域扩展

深度生产模型的应用范围相对较小，如何将其他深度生成模型的思想以及成果运用在常见场景中、如何加速与这些领域的融合，是未来进一步发展深度生成模型的关键方向

Part.05

参考文献

文献目录

- [1] WU C, LIANG J, JI L, et al. Nüwa: Visual synthesis pre-training for neural visual world creation[Z]. 2021.
- [2] 谭明奎,许守恺,张书海,陈奇. 2021. 深度对抗视觉生成综述. 中国图象图形学报,26(12):2751-2766
- [3] Bond-Taylor S， Leach A， Long Y , et al. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models[J]. 2021.
- [4] Oord A , Vinyals O , Kavukcuoglu K . Neural Discrete Representation Learning[J]. 2017.
- [5] 胡铭菲，刘建伟，左信．深度生成模型综述．自动化学报

中山大學
SUN YAT-SEN UNIVERSITY

感谢阅读

中山大学智能工程学院

方桂安，唐迅，凌海涛，张书戬，潘嘉雯

指导老师：刘梦源老师