StackGAN：具有堆叠生成对抗网络的文本到照片真实感图像合成

韩章1,陶旭2,李洪生3,，

张少亭、王晓刚、黄晓蕾2、美他赛1 2 4

利哈伊大学罗格斯大学中国香港大学百度研究

{han.zhang，dnm}@cs。罗格斯大学。教育部，{tax313，xih206}@lehigh。埃杜

{hsli，xgwang}@ee。中大。埃杜。香港，zhangshaoting@baidu.com

# 摘要

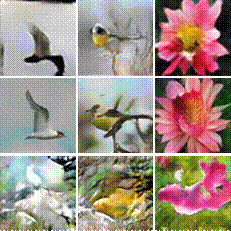
*从文本描述中合成高质量的图像是计算机视觉中一个具有挑战性的问题，具有许多实际应用。现有的文本到图像方法生成的样本可以大致反映给定描述的含义，但它们无法包含必要的细节和生动的对象部分。在本文中，我们提出了堆叠生成对抗网络（StackGAN）来生成256*\**256张以文字描述为条件的照片逼真图像。我们通过草图细化过程将困难问题分解为更易于管理的子问题。Stage-I GAN根据给定的文本描述绘制对象的基本形状和颜色，生成Stage-I低分辨率图像。第二阶段GAN将第一阶段的结果和文本描述作为输入，并生成具有照片真实细节的高分辨率图像。它能够纠正第一阶段结果中的缺陷，并通过细化过程添加引人注目的细节。为了提高合成图像的多样性并稳定条件GAN的训练，我们引入了一种新的条件增强技术，该技术鼓励潜在条件流形中的平滑性。在基准数据集上进行的大量实验和与现有技术的比较表明，该方法在生成基于文本描述的照片真实感图像方面取得了显著的改进。*

# 1.  介绍

|  |
| --- |
|  |
|  |  |

从文本生成照片真实感图像是一个重要的问题，具有巨大的应用，包括照片编辑、计算机辅助设计等。最近，生成对抗网络（GAN）[8,5,23]在合成真实世界图像方面显示了良好的结果。以给定的文本描述为条件，有条件- GANs[26，24]能够生成与文本含义高度相关的图像。

这种鸟有一个白色的黄色的腹部和重叠的粉红色与一些黑色的跗骨，灰色的背部，尖尖的花瓣它的头部和翅膀，翅膀，周围有一个棕色的环，有一个长长的喉咙，颈背上有黄色的短喙和黑色的脸丝

（a） stackGAN

第一阶段64x64

图像

（b） 斯塔克根

第二阶段256x256图像

（c） 香草甘蓝256x256图像

图1。比较所提出的StackGAN和普通单级GAN生成256×256图像。（a） 给出文本描述后，StackGAN的第一阶段勾勒出物体的大致形状和基本颜色，生成低分辨率图像。（b） StackGAN的第二阶段将第一阶段的结果和文本描述作为输入，并生成具有照片真实细节的高分辨率图像。（c） 结果由一个普通的256×256 GAN产生，它简单地为最先进的GAN-INT-CLS添加了更多的上采样层[26]。它无法生成任何256×256分辨率的合理图像。

然而，训练GAN从文本描述生成高分辨率照片真实感图像是非常困难的。只需在最先进的GAN模型中添加更多上采样层，即可生成高分辨率（例如256⇥256）图像通常会导致训练不稳定并产生无意义的输出（见图1（c））。GANs生成高分辨率图像的主要困难在于，自然图像分布和隐含模型分布的支持在高维像素空间中可能不会重叠[31，1]。随着图像分辨率的增加，此问题更加严重。里德等人仅成功地生成了可信的64⇥64图像以文字描述为条件[26]，文字描述通常缺乏细节和生动的物体部分，例如鸟类的喙和眼睛。此外，他们无法合成更高的分辨率（例如128）⇥128）图像，而不提供对象的附加注释[24]。

与人类画家的绘画方式类似，我们将文本到照片真实感图像合成问题分解为两个更容易处理的子问题，即堆叠生成对抗网络（StackGAN）。低分辨率图像首先由第一阶段GAN生成（见图1（a））。在第一阶段GAN的顶部，我们堆叠第二阶段GAN以生成真实的高分辨率（例如256⇥256）基于第一阶段结果和文本描述的图像（见图1（b））。通过再次对第一阶段的结果和文本进行调节，第二阶段GAN学习捕获第一阶段GAN忽略的文本信息，并为对象绘制更多细节。由大致对齐的低分辨率图像生成的模型分布的支持与图像分布的支持有更好的相交概率。这就是II级GAN能够生成更好的高分辨率图像的根本原因。

此外，对于文本-图像生成任务，训练文本-图像对的数量有限常常导致文本条件流形中的稀疏性，这种稀疏性使得训练GAN变得困难。因此，我们提出了一种新的条件增强技术，以鼓励在潜在的条件流形平滑。它允许调节流形中的小随机扰动，并增加合成图像的多样性。

该方法的贡献有三：（1）我们提出了一种新的堆叠生成对抗网络，用于从文本描述合成照片真实感图像。它将生成高分辨率图像的难题分解为更易于管理的子问题，并显著提高了技术水平。StackGAN首次生成256幅图像⇥256分辨率，具有来自文本描述的照片逼真细节。（2） 提出了一种新的条件增强技术来稳定条件GAN训练，并提高了生成样本的多样性。（3） 大量的定性和定量实验证明了整体模型设计的有效性以及各个组件的影响，这为设计未来的条件GAN模型提供了有用的信息。我们的代码可在https://github.com/hanzhanggit/StackGAN.

# 2.  相关工作

生成性图像建模是计算机视觉中的一个基本问题。随着深度学习技术的出现，这方面已经取得了显著的进展。变分自动编码器（VAE）[13，28]用概率图形模型描述问题，其目标是最大化数据可能性的下限。利用神经网络对像素空间的条件分布建模的自回归模型（例如，PixelRNN）[33]也生成了吸引人的合成图像。最近，生成性对抗网络（GAN）[8]在生成更清晰图像方面表现出了良好的性能。但是训练的不稳定性使得GAN模型很难产生高分辨率（例如256⇥256）图像。已经提出了几种技术[23,29,18,1,3]来稳定培训过程并产生令人信服的结果。基于能量的GAN[38]也被提出用于更稳定的训练行为。

在这些生成模型的基础上，还研究了条件图像生成。大多数方法使用简单的条件变量，如属性或类标签[37,34,4,22]。还有一项工作是以图像为条件生成图像，包括照片编辑[2,39]、域转移[32,12]和超分辨率[31,15]。然而，超分辨率方法[31,15]只能为低分辨率图像添加有限的细节，并且不能像我们建议的StackGAN那样纠正较大的缺陷。最近，已经开发了几种从非结构化文本生成图像的方法。Mansimov等人[17]通过学习估计文本和生成画布之间的对齐，建立了AlignDRAW模型。Reed等人[27]使用条件像素CNN，使用文本描述和对象位置约束生成图像。Nguyen等人[20]使用近似的Langevin采样方法生成以文本为条件的图像。然而，他们的抽样方法需要一个低效的迭代优化过程。利用条件GAN，Reed等人[26]成功地生成了似是而非的基于文本描述的64\*64幅鸟类和花卉图像。他们的后续工作[24]能够产生通过在对象零件位置上使用附加注释获得128\*128个图像。

除了使用单个GAN生成图像外，还有工作[36,5,10]利用一系列GAN生成图像。Wang等人[36]将室内场景生成过程分解为结构生成和风格生成，并使用提出的*s*2-GAN。相比之下，StackGAN的第二阶段旨在完成对象细节，并基于文本描述纠正第一阶段结果的缺陷。Denton等人[5]在拉普拉斯金字塔框架内构建了一系列GANs。在金字塔的每一层，根据前一阶段的图像生成剩余图像，然后将其添加回输入图像以生成下一阶段的输入。与我们的工作同时，Huang等人[10]还表明，他们可以通过堆叠多个GANs来重建预先训练的判别模型的多级表示，从而生成更好的图像。然而，他们只成功地生成了32\*32个图像，而我们的方法使用更简单的体系结构生成256\*256幅图像，具有逼真的细节和64倍以上的像素。

# 3.  堆叠生成对抗网络

为了生成具有照片真实感细节的高分辨率图像，我们提出了一种简单而有效的堆叠生成对抗网络。它将文本到图像的生成过程分解为两个阶段（见图2）。

-    第一阶段：它根据给定的文本描述绘制对象的基本形状和基本颜色，并从随机噪声向量中绘制背景布局，生成低分辨率图像。

-    第二阶段GAN：它修正了第一阶段低分辨率图像中的缺陷，并通过再次读取文本描述来完成对象的细节，生成高分辨率照片真实感图像。

## 3.1.  准备

生成性对抗网络（GAN）[8]由两个模型组成，这两个模型交替训练以相互竞争。生成器经过优化，通过生成鉴别器难以区分真实图像的图像来再现真实数据分布。同时，对该算法进行了优化，以区分真实图像和合成图像。总的来说，训练过程类似于两人最小-最大博弈，目标函数为minmaxV（D，G）=E∼[logD（x）]+*G数据DDG十、P数据*

*G D*(1)

E∼[日志（1）− D(G(z))],，*ZPZ*

其中x是来自真实数据分布的真实图像z，是从真实数据分布（例如，*pz*均匀分布或高斯分布）采样的噪声向量

条件GAN[7,19]是GAN的一个扩展，其中生成器和鉴别器都接收额外的条件变量，产生*G*（z，c）和*D*（x，c）。此公式允许G生成以变量c为条件的图像。

## 3.2.  条件反射增强

如图2所示，文本描述首先由编码器编码，产生文本嵌入。在以前的工作[26,24]中，文本嵌入被非线性转换，以生成调节潜在变量作为生成器的输入。然而，文本嵌入的潜在空间通常是高维的（>100维）。由于数据量有限，它通常会导致潜在数据流形的不连续性，这对于学习生成器是不可取的。为了缓解这个问题，我们引入了一种条件增强技术来产生额外的条件变量。与[26,24]中的固定条件文本变量不同，我们从独立的高斯分布①N(μ(t), ②∑(μ(t)),中随机抽取潜在变量，其中平均值①和对角协方差矩阵②是文本嵌入的函数。在给定少量图像文本对的情况下，所提出的条件增强产生更多的训练对，从而鼓励对沿条件流形的小扰动的鲁棒性。为了进一步加强调节歧管上的平滑度并避免过度拟合[6,14]，我们在训练期间将以下正则化项添加到生成器的目标中，

*DKL*（N（µ（&apos;t），∑（&apos;t））| | N（0，I）），(2)

这是标准高斯分布和条件高斯分布之间的Kullback-Leibler散度（KL散度）。条件增强中引入的随机性有利于建立文本到图像的翻译模型，因为同一句话通常对应于具有各种姿势和外观的对象。

## 3.3.  第一阶段GAN

我们没有直接生成以文本描述为条件的高分辨率图像，而是简化了任务，首先使用Stage-I GAN生成低分辨率图像，该GAN只关注对象的粗略形状和正确颜色。

设为给定描述的文本嵌入，该描述由本文中预先训练的编码器[25]生成。用于文本嵌入的高斯条件变量ˆ0从N（µ0（&apos;t），∑0（&apos;t））中采样，以捕获具有变化的文本的含义。在ˆ0和随机变量的条件下，第一阶段GAN通过交替地最大化LDin等式（3）和最小化LGin等式（4）来训练鉴别器和发生器，





其中真实图像和文本描述来自真实数据分布。是从给定分布（本文中为高斯分布）随机采样的噪声向量。是平衡等式（4）中两项的正则化参数。我们所有的实验都设置为=1。使用[13]中介绍的重新参数化技巧，（&apos;t）和∑0（&apos;t）与网络的其余部分一起学习。

|  |
| --- |
| 图2。提出的StackGAN的体系结构。Stage-I生成器通过从给定文本绘制对象的粗略形状和基本颜色，并从随机噪声向量绘制背景，从而绘制低分辨率图像。以第一阶段的结果为条件，第二阶段生成器纠正缺陷，并将引人注目的细节添加到第一阶段的结果中，生成更逼真的高分辨率图像。 |

模型架构。对于生成器*G*0，为了获得文本调节变量*C* 0，首先将文本嵌入馈送到完全连接的层中，以生成高斯分布N（µ0（&apos;t），∑0（&apos;t））的和（是∑0对角线上的值）。然后从高斯分布中取样。我们的维度条件向量ˆ0由ˆ0=µ0+σ0#计算✏ （式中#为元素乘法，⇠ N（0，I））。然后，将ˆ0与维度噪声向量串联以生成一个噪声向量⇥ H0图像由一系列上采样块构成。

对于鉴别器*D*0，首先使用完全连接的层将文本嵌入压缩到维度，然后在空间上复制以形成一个文本张量。同时，图像通过一系列下采样块进行馈送，直到其Nd\*Nd\*Md空间维度。然后，图像滤波器映射沿通道维度与文本张量连接。由此产生的张量进一步馈送到1\*1个卷积层，共同学习图像和文本的特征。最后，使用一个具有一个节点的完全连接层来生成决策得分。

## 3.4.  第二阶段

由第一阶段GAN生成的低分辨率图像通常缺少生动的对象部分，并且可能包含形状扭曲。文本中的一些细节在第一阶段也可能被省略，这对于生成照片真实感图像至关重要。我们的第二阶段GAN基于第一阶段GAN结果生成高分辨率图像。它以低分辨率图像和再次嵌入文本为条件，以纠正第一阶段结果中的缺陷。第二阶段GAN完成以前忽略的文本信息，以生成更逼真的细节。

在低分辨率结果*s*0=G0（z，cˆ0）和高斯潜变量的条件下，通过交替地最大化等式（5）中的LD和最小化等式（6）中的LG来训练阶段II GAN中的鉴别器和生成器，





与原始GAN公式不同的是，在假设随机性已被保留的情况下，在此阶段不使用随机噪声。本阶段中使用的高斯条件变量和第一阶段中使用的共享相同的预训练文本编码器，生成相同的文本嵌入。然而，第一阶段和第二阶段条件强化有不同的完全连接层，用于生成不同的平均值和标准差。通过这种方式，第二阶段GAN学习捕获由第一阶段GAN省略的文本嵌入中的有用信息。

模型架构。我们将Stage II生成器设计为具有剩余块的编解码网络[9]。与前一阶段类似，文本嵌入用于生成维度文本调节向量，该向量在空间上被复制以形成文本张量。同时，由第一阶段GAN生成的第一阶段结果被馈送到几个下采样块（即编码器）中，直到其空间大小为Mg\*Mg毫克。图像特征和文字特征沿通道标注连接。将编码后的图像特征与文本特征相结合，形成若干残差块，用于学习图像和文本特征的多模态表示。最后，使用一系列上采样层（即解码器）来生成高分辨率图像。这种生成器能够帮助纠正输入图像中的缺陷，同时添加更多细节以生成逼真的高分辨率图像。

对于鉴别器，其结构类似于第一阶段鉴别器，仅具有额外的下采样块，因为该阶段的图像尺寸较大。为了明确地强制GAN学习图像和条件文本之间更好的对齐，而不是使用普通鉴别器，我们在两个阶段都采用了Reed等人[26]提出的匹配感知鉴别器。在训练过程中，鉴别器将真实图像及其对应的文本描述作为正样本对，而负样本对由两组组成。第一种是包含不匹配文本嵌入的真实图像，而第二种是包含相应文本嵌入的合成图像。

## 3.5.  实施细节

上采样块包括最近邻上采样，然后是3⇥3步1卷积。除最后一次卷积外，每次卷积后都会应用批标准化[11]和ReLU激活。剩余区块3\*3个步1卷积组成、批量标准化和ReLU。其中使用了两个剩余块128\*128的StackGAN型号，使用4个256\*256个型号。下采样块由4\*4个组成2步卷积、批量标准化和LeakyReLU，但第一步没有批量标准化。

默认情况下，=128、=100、=16、=4、=128、=H0=64和=H=256。对于训练，我们首先通过固定第二阶段GAN，对第一阶段GAN进行600个时期的迭代训练。然后，我们通过固定第一阶段GAN，对第二阶段GAN再进行600个时代的迭代训练和测试。使用批次大小为64且初始学习率为0.0002的ADAM解算器对所有网络进行训练。每100个时代，学习率就会衰减到其先前值的1/2。

# 4.  实验

为了验证我们的方法，我们进行了广泛的定量和定性评估。比较了两种最先进的文本图像合成方法：GAN-INT-CLS[26]和GAWWN[24]。使用作者发布的代码生成两种比较方法的结果。此外，我们还设计了几个基线模型，以研究我们提出的StackGAN的总体设计和重要组件。对于第一个基线，我们直接训练Stage-I GAN以生成64⇥64和256⇥256幅图像，以研究所提出的堆叠结构和条件增强是否有益。然后我们修改StackGAN以生成128⇥128和256⇥256幅图像，以研究我们的方法得到的较大图像是否会导致较高的图像质量。我们还调查了在StackGAN的两个阶段输入文本是否有用。

## 4.1.  数据集和评估指标

CUB[35]包含200种鸟类和11788张图片。由于该数据集中80%的鸟类对象图像大小比率小于0.5[35]，作为预处理步骤，我们裁剪所有图像以确保鸟类边界框的对象图像大小比率大于-0.75。Oxford-102[21]包含来自102个不同类别的8189幅花卉图像。为了展示我们方法的泛化能力，一个更具挑战性的数据集MS COCO[16]也被用于评估。与CUB和Oxford102不同，MS COCO数据集包含具有多个对象和各种背景的图像。它有一个包含80k图像的训练集和一个包含40k图像的验证集。COCO中的每个图像有5个描述，而[25]为CUB和Oxford102数据集中的每个图像提供了10个描述。按照[26]中的实验设置，我们直接使用COCO提供的训练集和验证集，同时我们将CUB和Oxford-102划分为类不相交的训练集和测试集。

评价指标。很难评估生成模型（如GAN）的性能。我们选择最近提出的数值评估方法“初始分数”[29]进行定量评估，

*I*=exp(Ex*DKL*（p（y）|十、)||p(y)),，(7)

其中x表示一个生成的样本，是初始模型预测的标签[30]。这一指标背后的直觉是，好的模型应该生成多样化但有意义的图像。因此，边际分布（y）和条件分布（y | x）之间的KL差异应该很大。在我们的实验中，我们直接使用预先训练好的COCO数据集初始模型。对于细粒度数据集CUB和Oxford-102，我们为它们中的每一个微调初始模型。如[29]所述，我们对每个模型的大量样本（即30k随机选择的样本）评估该指标。

尽管初始分数与人类对样本视觉质量的感知有很好的相关性[29]，但它不能反映生成的图像是否符合给定的文本描述。因此，我们也进行人的评估。我们为每一类CUB和Oxford-102测试集随机选择50个文本描述。对于COCO数据集，从其验证集中随机选择4k文本描述。对于每个句子，每个模型生成5个图像。对于相同的文本描述，要求10名用户（不包括任何作者）用不同的方法对结果进行排序。通过计算人类用户的平均等级来评估所有比较方法。

## 4.2.  定量和定性结果

我们将我们的结果与CUB、Oxford-102和COCO上的最新文本图像方法[24,26]进行比较

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 文本描述 | 这只鸟是红色和棕色的，喙短而粗 | 这只鸟又矮又短，长着一条短尾巴  它身上是黄色的 | 一种有中等橙色喙、白色身体、灰色翅膀和蹼足的鸟 | 这只黑色的小鸟有一张短而略微弯曲的喙和长腿 | 一种眼睛下面有白色的棕色渐变色的小鸟 | 一种黄色的小鸟，有黑色的冠和黑色尖嘴 | 这种小鸟有白色的胸脯，浅灰色的头，黑色的翅膀和尾巴 |   64x64  GAN-INT-CLS  128x128  加文  256x256  斯塔克根  图3。StackGAN、GAWWN[24]和GAN-INT-CLS[26]的示例结果以CUB测试集的文本描述为条件。   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 文本描述 | 这种花有许多小紫色花瓣，呈圆顶状 | 这朵花有粉色、白色和黄色，花瓣上有条纹 | 这种花的花瓣是深粉红色的，有白色的边缘和粉红色的雄蕊 | 这朵花是白色和黄色的，花瓣是波浪状和光滑的 | 一张非常干净的客厅的照片 | 一群滑雪的人  站在雪地里 | 鸡蛋、水果、糖果、坚果和肉放在白色盘子上 | 在一天中的一个交通灯杆上的街道标志 |   64x64  GAN-INT-CLS  256x256  斯塔克根  图4。StackGAN和GAN-INT-CLS[26]的示例结果以牛津-102测试集（最左边）的文本描述为条件 |

四列）和COCO验证集（最右边的四列）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 米制的 | 数据集 | GAN-INT-CLS | 加文 | 我们的斯塔克根 |
| 起始分数 | 幼兽 | 2.88 ± .04 | 3.62 ± .07 | 3.70 ± .04 |
| 牛津 | 2.66 ± .03 | / | 3.20 ± .01 |
| 可可 | 7.88 ± .07 | / | 8.45 ± .03 |
| 人类  等级 | 幼兽 | 2.81 ± .03 | 1.99 ± .04 | 1.37 ± .02 |
| 牛津 | 1.87 ± .03 | / | 1.13 ± .03 |
| 可可 | 1.89 ± .04 | / | 1.11 ± .03 |

表1。我们的StackGAN、GAWWN[24]和GAN-INT-CLS[26]在CUB、Oxford102和MS-COCO数据集上的初始得分和平均人类等级。

数据集。表1中报告了我们提出的StackGAN和比较方法的初始分数和平均人类等级。图3和图4比较了代表性示例。

我们的StackGAN在所有三个数据集上都获得了最佳的初始分数和平均人类等级。与GAN-INT-CLS[26]相比，StackGAN在CUB数据集上的初始得分提高了28.47%（从2.88提高到3.70），在Oxford-102上提高了20.30%（从2.66提高到3.20）。StackGAN较好的平均人类等级也表明，我们提出的方法能够生成基于文本描述的更真实的样本。

如图3所示，由GAN-INT-CLS[26]生成的64\*64个样本只能反映鸟类的一般形状和颜色。在大多数情况下，他们的结果缺乏生动的部分（如喙和腿）和令人信服的细节，这使得他们既不现实，也没有足够高的分辨率。通过对local使用附加调节变量

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | 这只鸟是  文本为蓝色，带有白色描述，具有非常清晰的颜色  短喙 | 这种鸟有棕色的翅膀和黄色的腹部 | 一种有黑色王冠和黄色喙的白色鸟 | 这只鸟有白色、黑色和棕色三种颜色，有棕色的喙 | 这种鸟有小喙，有红棕色的冠和灰色的腹部 | 这是一只黑色的小鸟，白色的胸脯，白色的翼杆。 | 这只鸟是白色、黑色和黄色的，有一个黑色的短喙 |   第一阶段图像  第二阶段图像  图5。我们的StackGAN从CUB测试集中看不见的文本生成的样本。每列列出文字说明、生成的图像 |

根据StackGAN的第一阶段和第二阶段的文本。

从测试集中训练集文本的五个最近邻生成的图像



图6。对于生成的图像（第1列），使用Stage II鉴别器提取视觉特征，检索最近的训练图像（第2-6列）。为最近邻检索计算特征之间的2个距离。*DL*

由于时间限制，GAWWN[24]在CUB数据集上获得了更好的初始分数，仍然略低于我们的分数。它比GAN-INT-CLS生成更高分辨率的图像，细节更多，如图3所示。然而，正如作者所提到的，当GAWWN仅以文本描述为条件时，它无法生成任何合理的图像[24]。相比之下，我们的StackGAN可以生成256⇥256张照片真实感图像，仅来自文本描述。

图5展示了StackGAN生成的第一阶段和第二阶段图像的一些示例。如图5的第一行所示，在大多数情况下，Stage-I GAN能够绘制给定文本描述的对象的粗略形状和颜色。然而，第一阶段的图像通常是模糊的，有各种缺陷和缺少细节，特别是对于前景对象。如第二行所示，StageII GAN生成4\* 更高分辨率的图像具有更令人信服的细节，以更好地反映相应的文本描述。对于第一阶段GAN已经生成合理形状和颜色的情况，第二阶段GAN完成细节。例如，在图5的第1列中，第一阶段的结果令人满意，第二阶段的GAN着重于绘制文本中描述的短喙和白色以及尾巴和腿的细节。在所有其他示例中，将不同程度的细节添加到阶段II图像中。在许多其他情况下，第二阶段GAN能够通过再次处理文本描述来纠正第一阶段结果的缺陷。例如，虽然第5列中的第一阶段图像具有蓝色冠而不是文本中描述的红棕色冠，但缺陷通过第二阶段GAN校正。在某些极端情况下（如图5第7列），即使第一阶段GAN未能绘制合理的形状，第二阶段GAN也能够生成合理的对象。我们还观察到StackGAN能够从第一阶段图像转移背景，并在第二阶段对其进行微调，使其更逼真，分辨率更高。

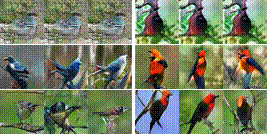
重要的是，StackGAN并不是通过简单地记忆训练样本，而是通过捕捉复杂的语言-图像关系来获得好的结果。我们通过StackGAN的Stage II鉴别器从生成的图像和所有训练图像中提取视觉特征。对于每个生成的图像，可以从训练集中检索其最近的邻居。通过目视检查检索到的图像（见图6），我们可以得出结论，生成的图像与训练样本具有一些相似的特征，但本质上是不同的。*D*

## 4.3.  成分分析

在本小节中，我们使用基线模型分析CUB数据集上StackGAN的不同组件。表2中报告了这些基线的初始分数。

StackGAN的设计。如表2的前四行所示，如果直接使用Stage-I GAN生成图像，初始分数会显著降低。图7中的结果可以很好地说明这种性能下降。如图7第一行所示，第一阶段GAN无法生成任何合理的结果256\*256个样本，不含

一种头部为黑色的小鸟，这种鸟全身呈红色，翅膀为黑色，翅膀为灰色，喙尖

256x256

无CA的I期GAN

256x256

Ⅰ期GAN伴CA

256x256

StackGAN和CA，

文本两次

图7。条件增强（CA）有助于稳定条件GAN的训练，并提高生成样本的多样性。（第1行）在没有CA的情况下，第一阶段GAN无法生成合理的256×256个样本。虽然每列使用不同的噪波向量z，但生成的样本对于每个输入文本描述都是相同的。（第2-3行）使用CA但固定噪声向量z，方法仍然能够生成具有不同姿势和视点的鸟类。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 加利福尼亚州 | 文本两次 | 起始分数 |
| 64×64第一级GAN | 不 | / | 2.66 ± .03  ± |
| 256Ⅰ级GAN  128×128斯塔克根 |  |  | 3.20 ± .03  ± |
| 256×256斯塔克根 |  |  | 3.31 ± .03  ± |

表2。初始分数由StackGAN的不同基线模型生成的30000个样本计算得出。

使用条件强化（CA）。尽管具有CA的I期GAN能够产生更多的多样性256⇥256个样本，这些样本不像StackGAN生成的样本那样真实。论证了提出叠层结构的必要性。此外，通过将输出分辨率从256⇥256降低至128⇥128，初始得分从3.70降至3.35。请注意，所有图像都缩放到299⇥ 299，然后计算初始分数。因此，如果我们的StackGAN只增加图像大小而不添加更多信息，那么对于不同分辨率的样本，初始分数将保持不变。因此，初始得分下降了128⇥128 StackGAN证明了我们的256⇥256 StackGAN确实在较大的图像中添加了更多细节。对于256⇥256如果文本仅输入到第一阶段（表示为“两次无文本”），则初始分数从3.70降至3.45。这表明在第二阶段再次处理文本描述有助于细化第一阶段的结果。从项研究的结果中可以得出同样的结论128⇥128个StackGAN型号。

条件反射增强。我们还研究了建议的条件反射增强（CA）的有效性。将其从StackGAN 中移除256⇥256（表2中表示为“无CA”），初始得分从3.70降至3.31. 图7还显示了256⇥256第一阶段GAN（和StackGAN）使用CA可以生成具有不同姿势的鸟类

这只鸟完全是红色的→ 这只鸟完全是黄色的



这只鸟全身呈红色，翅膀黑色，喙尖→ 这种蓝色的小鸟有一个短而尖的喙，翅膀呈棕色



图8。（从左到右）通过插入两个句子嵌入生成的图像。可以观察到从第一句话的意思到第二句话的意思的逐渐变化。每行的噪声向量固定为零。

以及来自同一文本的观点。相反，在不使用CA的情况下，由256⇥256个StageI-GAN由于GANs不稳定的训练动力学而崩溃为无意义的图像。因此，所提出的条件增强有助于稳定条件GAN训练，并改善生成样本的多样性，因为它能够鼓励对沿潜在流形的小扰动的鲁棒性。

句子嵌入插值。为了进一步证明StackGAN学习了平滑的潜在数据流形，我们使用它从线性插值的句子嵌入生成图像，如图8所示。我们固定了噪声向量，因此生成的图像仅从给定的文本描述推断。第一行中的图像是由我们编造的简单句子生成的。这些句子只包含简单的颜色描述。结果表明，插值嵌入生成的图像能够准确地反映颜色变化，生成合理的鸟形。第二行展示了从更复杂的句子中生成的样本，其中包含了关于鸟类外观的更多细节。生成的图像将其原色从红色更改为蓝色，并将机翼颜色从黑色更改为棕色。

# 5.  结论

在本文中，我们提出了叠加生成对抗网络（StackGAN）与条件增强合成照片真实感图像。该方法将文本到图像的合成分解为一种新的草图细化过程。第一阶段根据给定文本描述中的基本颜色和形状约束绘制对象。第二阶段GAN纠正了第一阶段结果中的缺陷，并添加了更多细节，产生了具有更好图像质量的更高分辨率图像。大量的定量和定性结果证明了该方法的有效性。与现有的文本到图像生成模型相比，我们的方法生成的图像分辨率更高（例如256⇥256分辨率）和更多逼真的细节。

# 工具书类

[1] M.Arjovsky和L.Bottou。培养生成性对抗网络的原则性方法。在ICLR，2017年。2.

[2] 布罗克、林、里奇和韦斯顿。具有内省对抗网络的神经照片编辑。在ICLR，2017年。2.

[3] T.Che、Y.Li、A.P.Jacob、Y.Bengio和W.Li。模式正则化生成对抗网络。在ICLR，2017年。2.

[4] 陈X、段Y、霍索夫R、舒尔曼J、苏茨基和阿比尔。Infogan：通过信息最大化生成性对抗网络进行解释性表征学习。在NIPS，2016年。2.

[5] E.L.Denton、S.Chintala、A.Szlam和R.Fergus。使用拉普拉斯金字塔对抗网络的深层生成图像模型。在NIPS，2015年。1, 2

[6] 多尔斯基。变分自动编码器教程。arXiv:1606.059081016。3.

[7] 高蒂尔。用于卷积人脸生成的条件生成对抗网络。技术报告，2015年。3.

[8] I.J.古德费罗、J.普吉·阿巴迪、M.米尔扎、B.徐、D.沃德·法利、S.奥扎尔、A.C.库尔维尔和Y.本吉奥。生成性对抗网。在NIPS，2014年。1, 2, 3

[9] 何国强、张学友、任志强和孙志强。深度残差学习用于图像识别。在CVPR，2016年。4.

[10] 黄X、李Y、O.Pourseed、J.Hopcroft和S.Belongie。堆叠的生成性对抗网络。在CVPR，2017年。2, 3

[11] S.Ioffe和C.Szegedy。批量标准化：通过减少内部协变量转移来加速深度网络训练。在ICML，2015年。5.

[12] P.Isola，J.-Y.Zhu，T.Zhou和A.A.Efros。使用条件对抗网络进行图像到图像的翻译。在CVPR，2017年。2.

[13] D.P.Kingma和M.Welling。自动编码变分贝叶斯。在ICLR，2014年。2, 3

[14] A.B.L.Larsen、S.K.Sønderby、H.Larochelle和O.Winther。使用学习的相似性度量自动编码像素以外的内容。在ICML，2016年。3.

[15] C.莱迪格、L.泰斯、F.胡萨、J.卡巴莱罗、A.艾特肯、A.特贾尼、J.托茨、Z.王和W.史。使用生成对抗网络的照片逼真单图像超分辨率。在CVPR，2017年。2.

[16] 林泰耀、梅尔、贝隆吉、海斯、佩罗纳、拉马南、多尔和齐特尼克。Microsoft coco：上下文中的常见对象。在ECCV中，2014年。5.

[17] E.曼西莫夫、E.帕里索托、L.J.巴和R.萨拉胡季诺夫。从标题中产生注意力集中的图像。在ICLR，2016年。2.

[18] L.梅茨、B.普尔、D.普福和J.苏尔·迪克斯坦。展开生成性对抗网络。在ICLR，2017年。2.

[19] M.米尔扎和S.奥辛德罗。条件生成对抗网。arXiv:1411.17842014。3.

[20] A.Nguyen、J.Yosinski、Y.Bengio、A.Dosovitskiy和J.Clune。即插即用生成网络：在潜在空间中有条件地迭代生成图像。在CVPR，2017年。2.

[21]M.-E.Nilsback和A.Zisserman。对大量类别进行自动花卉分类。ICCVGIP，2008年。5.

[22]A.Odena、C.Olah和J.Shlens。带辅助分类器的条件图像合成。在ICML，2017年。2.

[23]A.Radford、L.Metz和S.Chintala。深度卷积生成对抗网络的无监督表征学习。在ICLR，2016年。1, 2

[24]S.Reed、Z.Akata、S.Mohan、S.Tenka、B.Schiele和H.Lee。学习画什么和在哪里画。在NIPS，2016年。1, 2, 3, 5, 6, 7

[25]S.里德、Z.赤田、B.希勒和H.李。学习细粒度视觉描述的深层表示。在CVPR，2016年。3, 5

[26]S.Reed、Z.Akata、X.Yan、L.Logeswaran、B.Schiele和H.Lee。生成对抗性文本图像合成。在ICML，2016年。1, 2, 3, 5, 6

[27]S.里德、A.范登奥尔德、N.卡尔希布伦纳、V.巴普斯特、M.博特维尼克和N.德弗雷塔斯。生成具有可控结构的可解释图像。技术报告，2016年。2.

[28]D.J.Rezende、S.Mohamed和D.Wierstra。深层生成模型中的随机反向传播和近似推理。在ICML，2014年。2.

[29]T.Salimans、I.J.Goodfello、W.Zaremba、V.Cheung、A.Radford和X.Chen。改进了训练gans的技术。在NIPS，2016年。2, 5

[30]C.Szegedy、V.Vanhoucke、S.Ioffe、J.Shlens和Z.Wojna。重新思考计算机视觉的初始架构。在CVPR，2016年。5.

[31]C.K.斯德比、J.卡巴莱罗、L.泰斯、W.史和F.胡萨。用于图像超分辨率的摊销贴图推断。在ICLR，2017年。2.

[32]Y.泰格曼、A.波利亚克和L.沃尔夫。无监督的跨域图像生成。在ICLR，2017年。2.

[33]A.van den Oord、N.Kalchbrenner和K.Kavukcuoglu。像素递归神经网络。在ICML，2016年。2.

[34]A.van den Oord、N.Kalchbrenner、O.Vinyals、L.Espeholt、A.Graves和K.Kavukcuoglu。使用像素CNN解码器生成条件图像。在NIPS，2016年。2.

[35]C.Wah、S.Branson、P.Welinder、P.Perona和S.Belongie。加州理工大学UCSD Birds-200-2011数据集。技术报告CNS-TR-2011-001，加利福尼亚理工学院，2011年。5.

[36]X.Wang和A.Gupta。使用风格和结构对抗网络生成图像建模。在ECCV中，2016年。2.

[37]严X.杨J.孙K.孙和李H.李。Attribute2image：从视觉属性生成条件图像。在ECCV中，2016年。2.

[38]J.Zhao、M.Mathieu和Y.LeCun。基于对抗性发电网络。在ICLR，2017年。2.

[39]J.Zhu、P.Krahenb–uhl、E.Shechtman和A.A.Efros。生成-自然图像流形上的操作性视觉操作。在ECCV中，2016年。2.