GAN在图像生成应用综述（论文解读）

 2019-01-23

[GAN](http://www.twistedwg.com/pages/tags.html#GAN)  1780

GAN在图像生成上取得了巨大的成功，这无疑取决于GAN在博弈下不断提高建模能力，最终实现以假乱真的图像生成。GAN自2014年诞生至今也有4个多年头了，大量围绕GAN展开的文章被发表在各大期刊和会议。以改进和分析GAN的数学研究、提高GAN的生成质量研究、GAN在图像生成上的应用（指定图像合成、文本到图像，图像到图像、视频）以及GAN在NLP和其它领域的应用。图像生成是研究最多的，并且该领域的研究已经证明了在图像合成中使用GAN的巨大潜力。本博客围绕[An Introduction to Image Synthesis with Generative Adversarial Nets](https://arxiv.org/abs/1803.04469)一文对GAN在图像生成应用做个综述。

**论文引入**

著名的物理学家Richard Feynman说过：“What I cannot create, I do not understand”（对于我创造不出的事物，我是无法理解它的）。我们现阶段接触到的AI产品，都是在尝试去看懂人类可以看懂的，例如对Imagenet的图像分类、AlphaGo、智能对话机器人等。然而，我们仍然不能断定这些算法具有真正的“智能”，因为知道如何做某事并不一定意味着理解某些东西，而且真正智能的机器人理解其任务是至关重要的。

如果机器可以去create，这也就可以说明机器对它的输入数据已经可以自主的建模，这是否可以说明机器在朝着更加“智慧”迈进了一步。这种create在机器学习的领域下，目前最为可行的方法是生成模型。通过学习的生成模型，机器甚至可以绘制不在训练集中但遵循相同分布的样本。

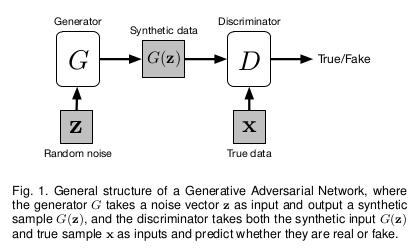
在生成模型中比较有影响力的有VAE[1]、PixelCNN[2]、Glow[3]、GAN[4]。其中在2014年提出的GAN可谓是生成模型中最受欢迎的，即使不能说GAN是一骑绝尘但也可谓是鹤立鸡群。GAN由两个神经网络组成，一个生成器和一个判别器组成，其中生成器试图产生欺骗判别器的真实样本，而判别器试图区分真实样本和生成样本。这种对抗博弈下使得生成器和判别器不断提高性能，在达到纳什平衡后生成器可以实现以假乱真的输出，但是这种纳什平衡只存在于理论中，实际GAN的训练伴随着一些问题的限制。一个是GAN训练不稳定性另一个是模式崩溃，导致问题的理论推导在之前[博客](http://www.twistedwg.com/2018/01/30/GAN-problem.html)有所推证。

GAN存在的问题并没有限制GAN的发展，不断改进GAN的文章层出不穷，在这几年的铺垫下GAN已经发展的蛮成熟的。从这几年关于GAN的高质量文章可以看出，18年以后的文章更多关注的是GAN在各个领域的应用，而之前的文章则是集中在GAN存在问题的改进。GAN在图像生成应用最为突出，当然在计算机视觉中还有许多其他应用，如图像绘画，图像标注，物体检测和语义分割。在自然语言处理中应用GAN的研究也是一种增长趋势，如文本建模，对话生成，问答和机器翻译。然而，在NLP任务中训练GAN更加困难并且需要更多技术，这也使其成为具有挑战性但有趣的研究领域。

[An Introduction to Image Synthesis with Generative Adversarial Nets](https://arxiv.org/abs/1803.04469)一文是概述GAN图像生成中使用的方法，并指出已有方法的优缺点。这篇博客则是对这篇论文进行个人的理解和翻译，并对其中的一些方法结合个人实际应用经验进行分析。

**GAN的基础**

接触过GAN的学者如果对GAN的结构已经很熟悉，这一部分可以自行跳过。我们看一下GAN的基础结构：

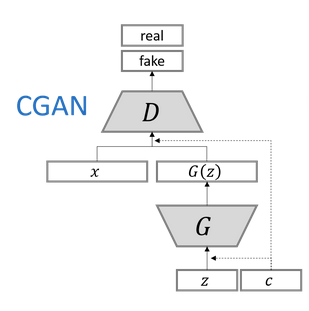


GAN可以将任意的分布作为输入，这里的ZZ就是输入，在实验中我们多取Z∼N(0,1)Z∼N(0,1)，也多取[−1,1][−1,1]的均匀分布作为输入。生成器G的参数为θθ，输入ZZ在生成器下得到G(z;θ)G(z;θ)，输出可以被视为从分布中抽取的样本G(z;θ)∼pgG(z;θ)∼pg。对于训练样本xx的数据分布为pdatapdata，生成模型G的训练目标是使pgpg近似pdatapdata。判别器D便是为了区分生成样本和真实样本的真假，训练发生器和判别器通过最小 - 最大游戏，其中发生器G试图产生逼真的数据以欺骗判别器，而判别器D试图区分真实数据和合成数据。这种博弈可公式化为：

minGmaxDV(D,G)=Ex∼pdata(x)[logD(x)]+Ez∼pz(z)[log(1−D(G(z)))]minGmaxDV(D,G)=Ex∼pdata(x)[logD(x)]+Ez∼pz(z)[log(1−D(G(z)))] 最初的GAN 使用完全连接的层作为其构建块。后来，DCGAN [5]提出使用卷积神经网络实现了更好的性能，从那以后卷积层成为许多GAN模型的核心组件。然而，当判别器训练得比发生器好得多时，D可以有信心地从G中拒绝来自G的样本，因此损失项log(1−D(G(z)))log(1−D(G(z)))饱和并且G无法从中学到任何东西。为了防止这种情况，可以训练G来最大化logD(G(z))logD(G(z))，而不是训练G来最小化log(1−D(G(z)))log(1−D(G(z)))。虽然G的改变后的损失函数给出了与原始梯度不同的梯度，但它仍然提供相同的梯度方向并且不会饱和。

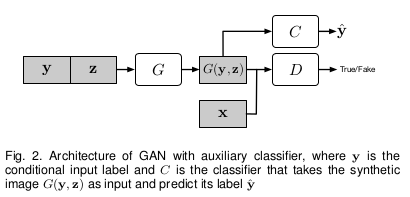
**条件GAN**

在原始GAN中，无法控制要生成的内容，因为输出仅依赖于随机噪声。我们可以将条件输入cc添加到随机噪声ZZ，以便生成的图像由G(c,z)G(c,z)定义。这就是CGAN[6]，通常条件输入矢量cc与噪声矢量zz直接连接即可，并且将得到的矢量原样作为发生器的输入，就像它在原始GAN中一样。条件cc可以是图像的类，对象的属性或嵌入想要生成的图像的文本描述，甚至是图片。



**辅助分类器GAN(ACGAN)**

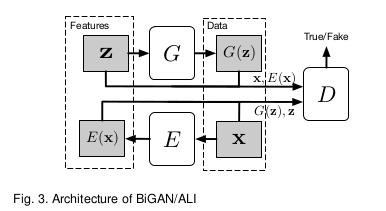
为了提供更多的辅助信息并允许半监督学习，可以向判别器添加额外的辅助分类器，以便在原始任务以及附加任务上优化模型。这种方法的体系结构如下图所示，其中C是辅助分类器。 添加辅助分类器允许我们使用预先训练的模型（例如，在ImageNet上训练的图像分类器），并且在ACGAN [7]中的实验证明这种方法可以帮助生成更清晰的图像以及减轻模式崩溃问题。 使用辅助分类器还可以应用在文本到图像合成和图像到图像的转换。



**GAN与Encoder的结合**

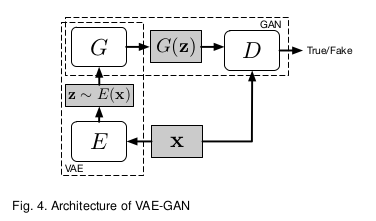
尽管GAN可以将噪声向量z转换为合成数据样本G(z)，但它不允许逆变换。如果将噪声分布视为数据样本的潜在特征空间，则GAN缺乏将数据样本x映射到潜在特征z的能力。为了允许这样的映射，两个并发的工作BiGAN[8]和ALI[9]在原始GAN中添加编码器E，如下图所示。令ΩxΩx为数据空间，ΩzΩz为潜在特征空间，编码器E将x∈Ωxx∈Ωx作为输入，并产生特征向量E(x)∈ΩzE(x)∈Ωz作为输出。修正判别器D以将数据样本和特征向量都作为输入来计算P(Y|x,z)P(Y|x,z)，其中Y=1Y=1表示样本是真实的而Y=0Y=0表示数据由G生成。用数学公式表示为：

minGmaxDV(D,G)=Ex∼pdata(x,E(x))[logD(x)]+Ez∼pz(z)[log(1−D(G(z),z))]minGmaxDV(D,G)=Ex∼pdata(x,E(x))[logD(x)]+Ez∼pz(z)[log(1−D(G(z),z))]



**GAN与VAE的结合**

VAE生成的图像是模糊的，但是VAE生成并没有像GAN的模式崩溃的问题，VAE-GAN[10]的初衷是结合两者的优点形成更加鲁棒的生成模型。模型结构如下：



但是实际训练过程中，VAE和GAN的结合训练过程也是很难把握的。

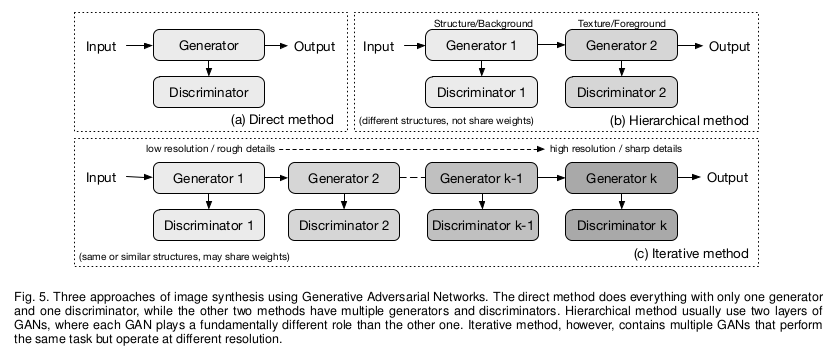
**处理模式崩溃问题**

虽然GAN在图像生成方面非常有效，但它的训练过程非常不稳定，需要很多技巧才能获得良好的结果。GAN不仅在训练中不稳定，还存在模式崩溃问题。判别器不需要考虑生成样品的种类，而只关注于确定每个样品是否真实，这使得生成器只需要生成少数高质量的图像就足以愚弄判别者。例如在MNIST 数据集包含从0到9的数字图像，但在极端情况下，生成器只需要学会完美地生成十个数字中的一个以完全欺骗判别器，然后生成器停止尝试生成其他九位数，缺少其他九位数是类间模式崩溃的一个例子。类内模式崩溃的一个例子是，每个数字有很多写作风格，但是生成器只学习为每个数字生成一个完美的样本，以成功地欺骗鉴别器。

目前已经提出了许多方法来解决模型崩溃问题。一种技术被称为小批量（miniBatch）特征，其思想是使判别器比较真实样本的小批量以及小批量生成的样本。通过这种方式，判别器可以通过测量样本在潜在空间中的距离来学习判断生成的样本是否与其他一些生成的样本过于相似。尽管这种方法运行良好，但性能在很大程度上取决于距离计算中使用的特征。 MRGAN [11]建议添加一个编码器，将数据空间中的样本转换回潜在空间，如BiGAN 它的编码器和生成器的组合充当自动编码器，重建损失被添加到对抗性损失中以充当模式正则化器。同时，还训练判别器以区分重构样本，其用作另一模式正则化器。 WGAN [12]使用Wasserstein距离来测量真实数据分布与学习分布之间的相似性，而不是像原始GAN那样使用Jensen-Shannon散度。虽然它在理论上避免了模式崩溃，但模型收敛的时间比以前的GAN要长。为了缓解这个问题，WGAN-GP [13]建议使用梯度惩罚，而不是WGAN中的权重削减。 WGAN-GP通常可以产生良好的图像并极大地避免模式崩溃，并且很容易将此培训框架应用于其他GAN模型。SAGAN[14]将谱归一化的思想用在判别器，限制判别器的能力。

**GAN在图像生成方法**

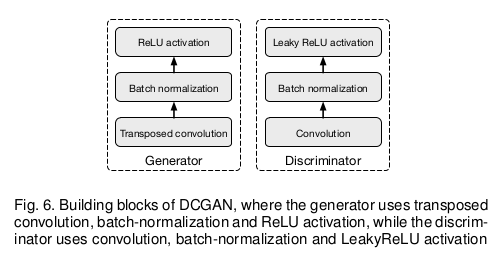
GAN在图像生成中的主要方法为直接方法，迭代方法和分层方法，这三种方法可由下图展示：



区分图像生成方法是看它拥有几个生成器和判别器。

**直接法**

该类别下的所有方法都遵循在其模型中使用一个生成器和一个判别器的原理，并且生成器和判别器的结构是直接的，没有分支。许多最早的GAN模型属于这一类，如GAN [4]、DCGAN [5]、ImprovedGAN [15]，InfoGAN [16]，f-GAN [17]和GANINT-CLS [18]。其中，DCGAN是最经典的之一，其结构被许多后来的模型使用，DCGAN中使用的一般构建块如下图所示，其中生成器使用反卷积，批量归一化和ReLU激活，而判别器使用卷积，batchnormalization和LeakyReLU激活，这也是现在很多GAN模型网络设计所借鉴的。

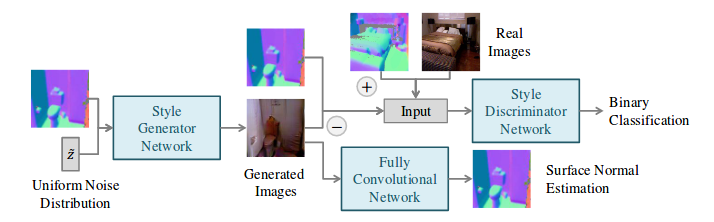


与分层和迭代方法相比，这种方法设计和实现相对更直接，并且通常可以获得良好的结果。

**分层法**

与直接法相反，分层方法下的算法在其模型中使用两个生成器和两个鉴别器，其中不同的生成器具有不同的目的。这些方法背后的想法是将图像分成两部分，如“样式和结构”和“前景和背景”。两个生成器之间的关系可以是并联的或串联的。

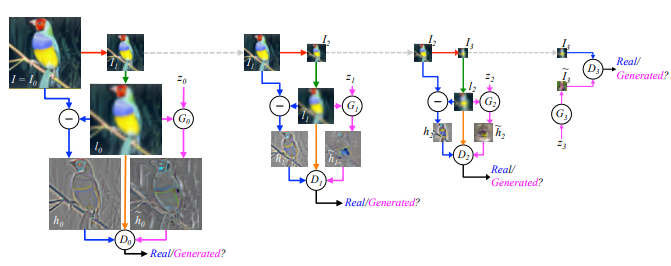
SS-GAN [19]使用两个GAN，一个Structure-GAN用于从随机噪声z^z^作为输入并输出图像，整体结构可由下图展示：



**迭代法**

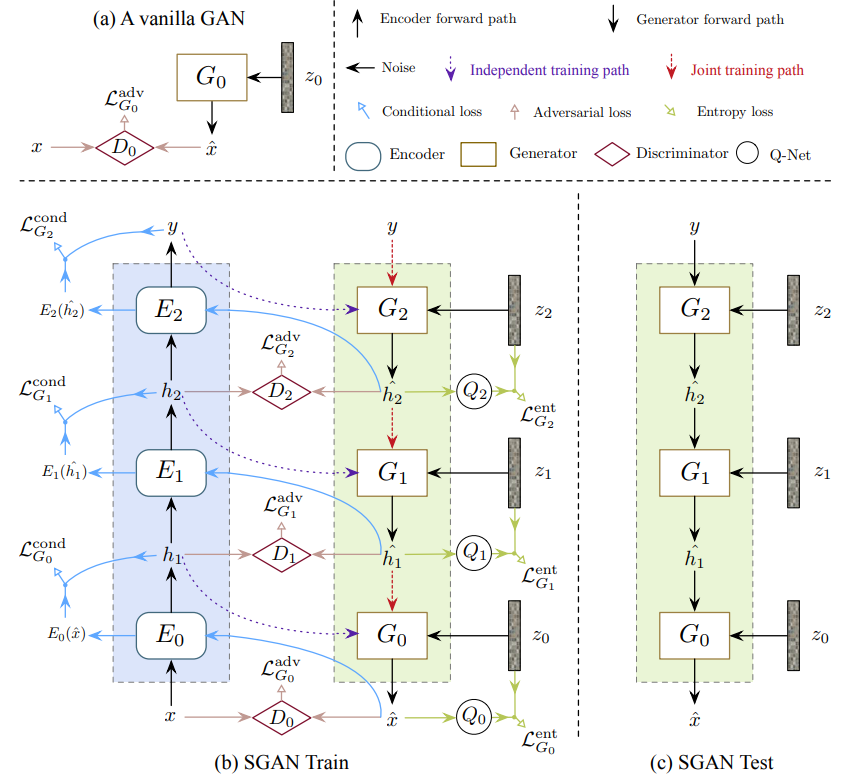
迭代法不同于分层法，首先，不使用两个执行不同角色的不同生成器，此类别中的模型使用具有相似或甚至相同结构的多个生成器，并且它们生成从粗到细的图像，每个生成器重新生成结果的详细信息。当在生成器中使用相同的结构时，迭代方法可以在生成器之间使用权重共享，而分层方法通常不能。

LAPGAN [20]是第一个使用拉普拉斯金字塔使用迭代方法从粗到细生成图像的GAN。LAPGAN中的多个生成器执行相同的任务：从前一个生成器获取图像并将噪声矢量作为输入，然后输出再添加到输入图像时使图像更清晰的细节（残留图像）。这些发生器结构的唯一区别在于输入/输出尺寸的大小，而一个例外是最低级别的生成器仅将噪声向量作为输入并输出图像。LAPGAN优于原始GAN 并且表明迭代方法可以生成比直接方法更清晰的图像。



StackGAN [21]作为一种迭代方法，只有两层生成器。第一个生成器接收输入(z,c)(z,c)，然后输出模糊图像，可以显示粗略的形状和对象的模糊细节，而第二个生成器采用(z,c)(z,c)和前一个生成器生成的图像，然后输出更大的图像，可以得到更加真实的照片细节。

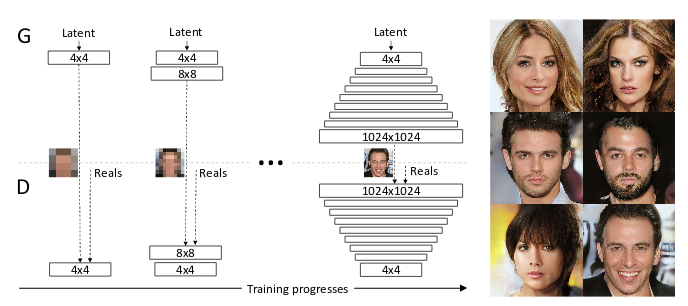
迭代法的另一个例子是SGAN [22]，其堆叠生成器，其将较低级别的特征作为输入并输出较高级别的特征，而底部生成器将噪声矢量作为输入并且顶部生成器输出图像。对不同级别的特征使用单独的生成器的必要性是SGAN关联编码器，判别器和Q网络（用于预测P(zi|hi)P(zi|hi)的后验概率以进行熵最大化，其中hihi每个生成器的第i层的输出特征），以约束和改善这些特征的质量。



**其它方法**

与前面提到的其他方法不同，PPGN[23]使用激活最大化来生成图像，它基于先前使用去噪自动编码器（DAE）学习的采样。为了生成以特定类别标签y为条件的图像，而不是使用前馈方式（例如，如果通过时间展开，可以将循环方法视为前馈），PPGN优化过程为生成器找到输入z这使得输出图像高度激活另一个预训练分类器中的某个神经元（输出层中与其类标签y对应的神经元）。

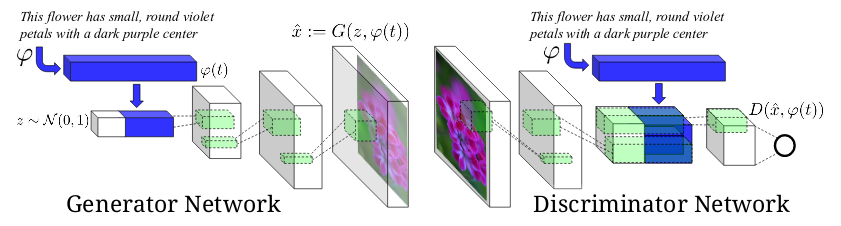
为了生成更好的更高分辨率的图像，ProgressiveGAN [24]建议首先训练4×44×4像素的生成器和判别器，然后逐渐增加额外的层，使输出分辨率加倍至1024×10241024×1024。这种方法允许模型首先学习粗糙结构，然后专注于稍后重新定义细节，而不是必须同时处理不同规模的所有细节。



**GAN在文本到图像的应用**

GAN应用于图像生成时，虽然CGAN [6]这样的标签条件GAN模型可以生成属于特定类的图像，但基于文本描述生成图像仍然是一个巨大的挑战。文本到图像合成是计算机视觉的里程碑，因为如果算法能够从纯粹的文本描述中生成真实逼真的图像，我们可以高度确信算法实际上理解图像中的内容。

GAN-INT-CLS [18]是使用GAN从文本描述生成图像的第一次尝试，这个想法类似于将条件向量与噪声向量连接起来的条件GAN，但是使用文本句子的嵌入而不是类标签或属性的区别。



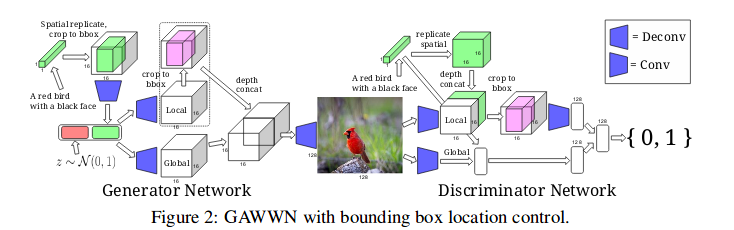
GAN-INT-CLS开创性的区分两种错误来源：不真实的图像与任何文本，以及不匹配的文本的真实图像。为了训练判别器以区分这两种错误，在每个训练步骤中将三种类型的输入馈送到判别器：{真实图像，匹配文本}，{真实图像，不匹配文本}和{伪图像，真实文本}。这种训练技术对于生成高质量图像非常重要，因为它不仅告诉模型如何生成逼真的图像，还告诉文本和图像之间的对应关系。

TAC-GAN [25]是GAN-INT-CLS [18]和ACGAN [7]的组合。

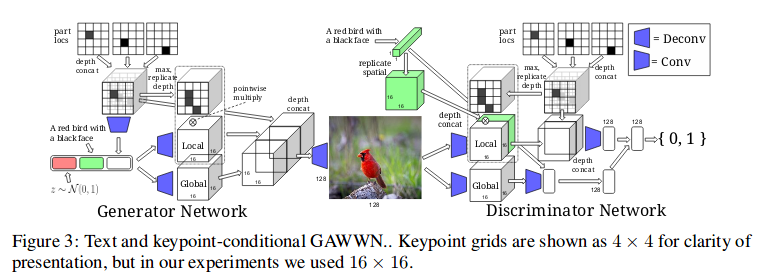
**位置约束的文本到图像**

尽管GAN-INT-CLS [18]和StackGAN [21]可以基于文本描述生成图像，但是它们无法捕获图像中对象的定位约束。为了允许编码空间约束，GAWWN [26]提出了两种可能的解决方案。

GAWWN 提出的第一种方法是通过空间变换网络对空间复制的文本嵌入张量进行学习，从而学习对象的边界框。空间变换器网络的输出是与输入具有相同维度的张量，但是边界外的值都是零。空间变换器的输出张量经过几个卷积层，以将其大小减小回一维向量，这不仅保留了文本信息，而且还通过边界框提供了对象位置的约束。这种方法的一个好处是它是端到端的，不需要额外的输入。



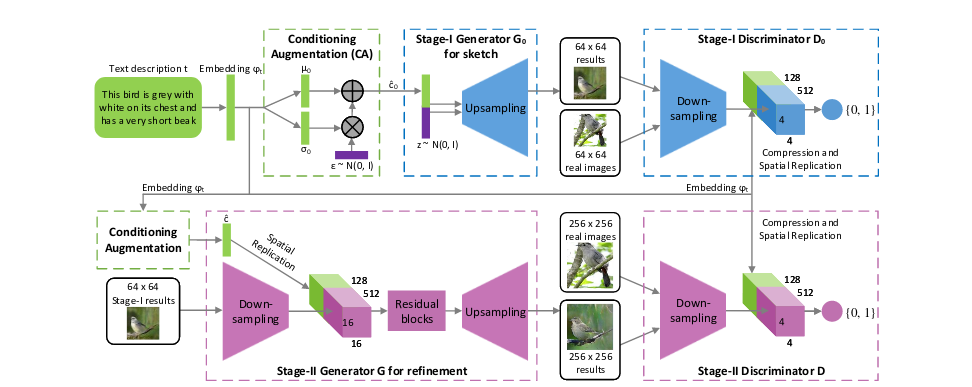
GAWWN 提出的第二种方法是使用用户指定的关键点来约束图像中对象的不同部分（例如头部，腿部，手臂，尾部等）。对于每个关键点，生成一个掩码矩阵，其中关键点位置为1，其他为0，所有矩阵通过深度级联组合形成一个形状[M×M×K][M×M×K]的掩码张量，其中MM是掩码的大小，KK是数字关键点。然后将张量放入二进制矩阵中，其中1指示存在关键点，否则为0，然后在深度方向上复制以成为要被馈送到剩余层中的张量。虽然此方法允许对对象进行更详细的约束，但它需要额外的用户输入来指定关键点。



尽管GAWWN提供了两种可以对生成的图像强制执行位置约束的方法，但它仅适用于具有单个对象的图像，因为所提出的方法都不能处理图像中的多个不同对象。

**堆叠GAN的文本到图像**

StackGAN [21]建议使用两个不同的生成器进行文本到图像的合成，而不是只使用一个生成器。第一个生成器负责生成包含粗糙形状和颜色的对象的低分辨率图像，而第二个生成器获取第一个生成器的输出并生成具有更高分辨率和更清晰细节的图像，每个生成器都与其自己的判别器相关联。



StackGAN ++ [27]建议使用更多对生成器和判别器而不是仅仅两个，为判别器增加无条件图像合成损失，并使用由均值平均损失计算的色彩一致性正则化项和真实和虚假图像之间的差异。

AttnGAN [28]通过在图像和文本特征上使用注意机制进一步扩展了StackGAN ++ [27]的体系结构。在AttnGAM中，每个句子都嵌入到全局句子向量中，并且句子的每个单词也嵌入到单词向量中。全局句子向量用于在第一阶段生成低分辨率图像，然后以下阶段使用前一阶段的输入图像特征和单词向量作为对关注层的输入并计算将使用的词语上下文向量。与图像特征组合并形成生成器的输入，将生成新的图像特征。

**文本到图像模型的局限性**

目前的文本到图像模型在每个图像具有单个对象的数据集上表现良好，例如CelebA中的人脸，CUB中的鸟以及ImageNet中的一些对象。此外，他们可以在LSUN 中为卧室和起居室等场景合成合理的图像，即使场景中的物体缺乏清晰的细节。然而，在一个图像中涉及多个复杂对象的情况下，所有现有模型都工作得很糟糕。

当前模型在复杂图像上不能很好地工作的一个合理的原因是模型只学习图像的整体特征，而不是学习其中每种对象的概念。这解释了为什么卧室和起居室的合成场景缺乏清晰的细节，因为模型不区分床和桌子，所有它看到的是一些形状和颜色的图案应放在合成图像的某处。换句话说，模型并不真正理解图像，只记得在哪里放置一些形状和颜色。

生成性对抗网络无疑提供了一种有前途的文本到图像合成方法，因为它产生的图像比迄今为止的任何其他生成方法都要清晰。为了在文本到图像合成中迈出更进一步的步骤，需要找到新的方法实现算法的事物概念。一种可能的方法是训练可以生成不同种类对象的单独模型，然后训练另一个模型，该模型学习如何基于文本描述将不同对象（对象之间的合理关系）组合成一个图像。然而，这种方法需要针对不同对象的大型训练集，以及包含难以获取的那些不同对象的图像的另一大型数据集。另一个可能的方向可能是利用Hinton等人提出的胶囊理念，]因为胶囊被设计用于捕获物体的概念，但是如何有效地训练这种基于胶囊的网络仍然是一个需要解决的问题。

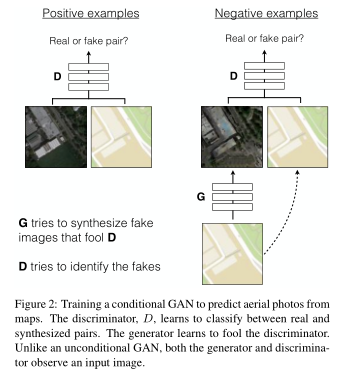
分割线（篇幅有点长，可以选择下次阅读）

**GAN在图像到图像的应用**

图像到图像的转换被定义为将一个场景的可能表示转换成另一个场景的问题，例如图像结构图映射到RGB图像，或者反过来。该问题与风格迁移有关，其采用内容图像和样式图像并输出具有内容图像的内容和样式图像的样式的图像。图像到图像转换可以被视为风格迁移的概括，因为它不仅限于转移图像的风格，还可以操纵对象的属性（如在面部编辑的应用中）。

**有监督下图像到图像转换**

Pix2Pix [29]提出将CGAN的损失与L1正则化损失相结合，使得生成器不仅被训练以欺骗判别器而且还生成尽可能接近真实标注的图像，使用L1而不是L2的原因是L1产生较少的模糊图像。



有条件的GAN损失定义为：

LCGAN(G,D)=Ex,y∼pdata(x,y)[logD(x,y)]+Ex∼pdata,z∼pz(z)[log(1−D(x,G(x,z)))]LCGAN(G,D)=Ex,y∼pdata(x,y)[logD(x,y)]+Ex∼pdata,z∼pz(z)[log(1−D(x,G(x,z)))]

约束自相似性的L1损失定义为：

LL1(G)=Ex∼pdata,z∼pz(z)[∥y−G(x,z)∥1]LL1(G)=Ex∼pdata,z∼pz(z)[‖y−G(x,z)‖1]

总的损失为：

G∗,D∗=argminGmaxDLCGAN+λLL1(G)G∗,D∗=argminGmaxDLCGAN+λLL1(G)

其中λλ是一个超参数来平衡两个损失项，Pix2Pix 的生成器结构基于UNet ，它属于编码器 - 解码器框架，但增加了从编码器到解码器的跳过连接，以便绕过共享诸如对象边缘之类的低级信息的瓶颈。

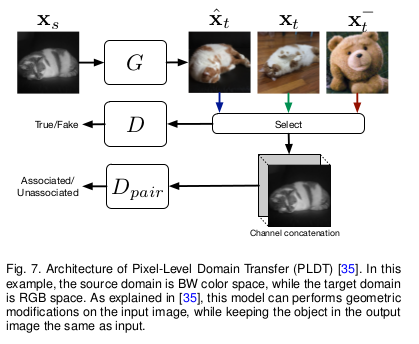
**配对监督下图像到图像转换**

PLDT [30]提出了另一种进行监督图像到图像转换的方法，通过添加另一个判别器DpairDpair来学习判断来自不同域的一对图像是否相互关联。的体系结构如下图所示，给定来自源域的输入图像来学习判断来自不同域的一对图像是否相互关联。PLDT的体系结构如下图所示，给定来自源域的输入图像XsXs，其目标域中的真实图像，其目标域中的真实图像XtXt，目标域中的无关图像，目标域中的无关图像Xt~Xt~，以及生成器将，以及生成器G将XsXs传输到图像传输到图像Xt^Xt^中。中。DpairDpair的损失可表示为：

Lpair=−t⋅log[Dpair(xs,x)]+(t−1)⋅log[1−Dpair(xs,x)]Lpair=−t⋅log[Dpair(xs,x)]+(t−1)⋅log[1−Dpair(xs,x)]

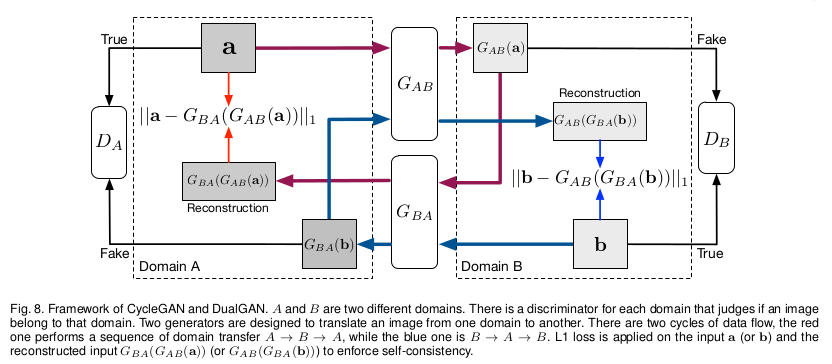
when,x=xtorx=xt^,t=0.when,x=xtorx=xt^,t=0.

when,x=xt~,t=1when,x=xt~,t=1



**无监督图像到图像转换**

两个并发工作CycleGAN [31]和DualGAN [32]采用重构损失，试图在转换周期后保留输入图像。CycleGAN和DualGAN共享相同的框架，如下图所示。可以看到，两个生成器GABGAB和GBAGBA正在进行相反的转换，这可以看作是一种双重学习。此外，DiscoGAN [33]是另一种利用与下图相同的循环框架的模型。



以CycleGAN为例，在CycleGAN中，有两个生成器，GABGAB用于将图像从域A传输到B，GBAGBA用于执行相反的转换。此外，还有两个判别器DADA和DBDB可预测图像是否属于该域。

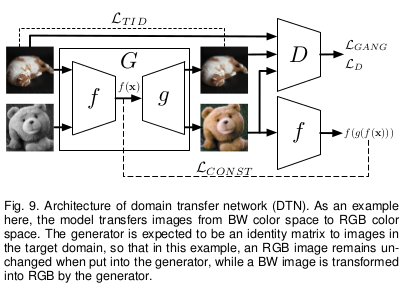
尽管CycleGAN 和DualGAN具有相同的模型结构，但它们对生成器使用不同的实现。CycleGAN使用卷积架构的生成器结构，而DualGAN遵循U-Net结构。

**距离约束下无监督图像到图像转换**

DistanceGAN [34]发现，源域A中两个图像之间的距离的dk=∥xi−xj∥dk=‖xi−xj‖与目标域B中对应图像d,k=∥GAB(xi)−GAB(xj)∥dk,=‖GAB(xi)−GAB(xj)‖的距离高度正相关。高度相关性下∑dkd,k∑dkdk,也应该为高。源域中的成对距离dkdk是固定的，并且最大化∑dkd,k∑dkdk,导致具有大值的dkdk支配损失，这是不希望的。因此作者建议最小化∑|dk−d,k|∑|dk−dk,|。

**特征稳定下无监督图像到图像转换**

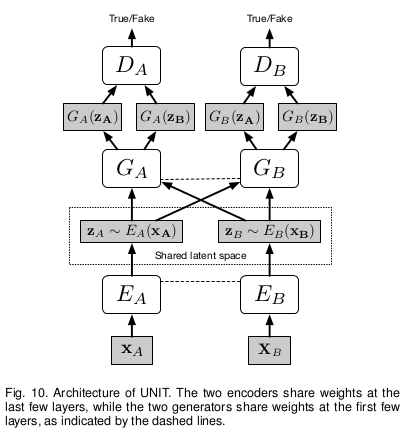
除了最小化原始像素级别的重建误差外，还可以在更高的特征级别进行此操作，这在DTN [35]中进行了探讨。DTN的体系结构如下图所示，其中发生器G由两个神经网络组成，一个卷积网络f和一个反卷积网络g，使得G=f∘gG=f∘g。



这里f充当特征提取器，并且DTN 尝试在将输入图像传输到目标域之后保留输入图像的高级特征。给定输入图像x∈xsx∈xs生成器的输出为G(x)=g(f(x))G(x)=g(f(x))，然后可以使用距离度量d(DTN使用均方误差(MSE))定义特征重建错误。这篇论文之前我们进行过详细解读，可参看[这篇博客](http://www.twistedwg.com/2018/09/05/DTN.html)。

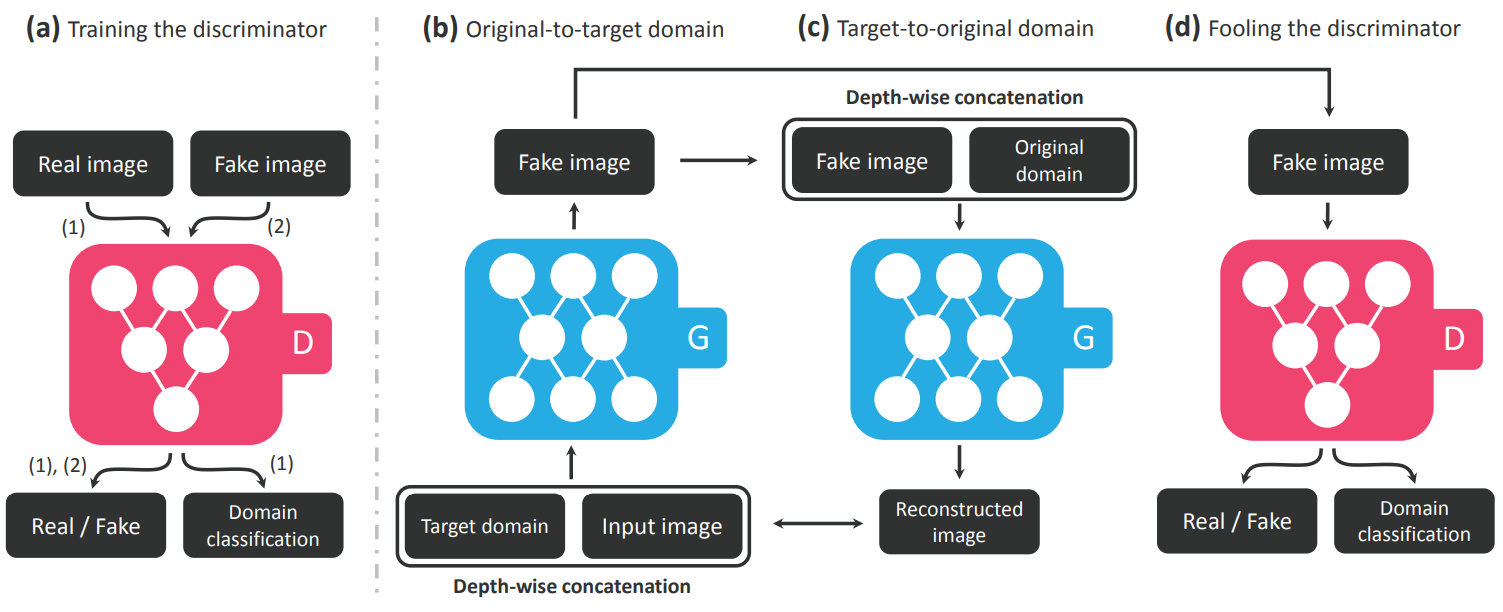
**借助VAE和权重分享下无监督图像到图像转换**

UNIT [36]建议将VAE添加到CoGAN [37]用于无监督的图像到图像转换，如下图所示。此外，UNIT假设两个编码器共享相同的潜在空间，这意味着xA,xBxA,xB是不同域中的相同图像，然后共享潜在空间意味着EA(xA)=EB(xB)EA(xA)=EB(xB)。基于共享潜在空间假设，UNIT强制在编码器的最后几层之间以及发生器的前几层之间进行权重共享。 UNIT的目标函数是GAN和VAE目标的组合，不同之处在于使用两组GAN / VAE并添加超参数λsλs来平衡不同的损耗项。



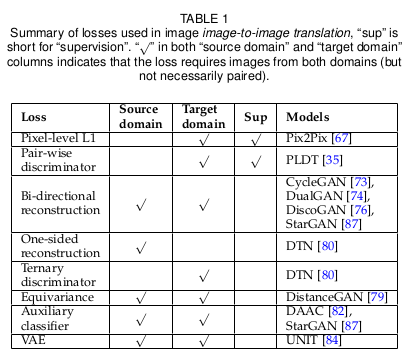
**无监督的多域图像到图像转换**

以前的模型只能在两个域之间转换图像，但如果想在几个域之间转换图像，需要为每对域训练一个单独的生成器，这是昂贵的。为了解决这个问题，StarGAN [38]建议使用一个可以生成所有域图像的生成器。StarGAN不是仅将图像作为条件输入，而是将目标域的标签作为输入，并且生成器用于将输入图像转换为输入标签指示的目标域。与ACGAN类似，StarGAN使用辅助域分类器，将图像分类到其所属的域中。此外，循环一致性损失用于保持输入和输出图像之间的内容相似性。为了允许StarGAN在可能具有不同标签集的多个数据集上进行训练，StarGAN使用额外的单一向量来指示数据集并将所有标签向量连接成一个向量，将每个数据集的未指定标签设置为零。



**图像到图像转换总结**

之前讨论的图像到图像转换方法，它们使用的不同损失总结在下表中：



最简单的损失是像素方式的L1L1重建损失，这需要成对的训练样本。单侧和双向重建损失都可以被视为像素方式L1L1重建损失的无监督版本，因为它们强制执行循环一致性并且不需要成对的训练样本。额外的VAE损失基于源域和目标域的共享潜在空间的假设，并且还意味着双向循环一致性损失。然而，等效损失不会尝试重建图像，而是保留源和目标域之间图像之间的差异。

在所有提到的模型中，Pix2Pix [29]产生最清晰的图像，即使L1L1损失只是原始GAN模型的简单附加组件。将L1L1损失与PLDT中的成对判别器结合起来可以改善模型在涉及图像几何变化的图像到图像转换上的性能。此外，Pix2Pix可能有利于保留源域和目标域中图像之间的相似性信息，如在一些无监督方法如CycleGAN [31]和DistanceGAN [34]中所做的那样。至于无监督方法，虽然它们的结果不如Pix2Pix等监督方法生成效果，但它们是一个很有前途的研究方向，因为它们不需要配对数据，并且在现实世界中收集标记数据是非常昂贵的。

**图像到图像转换的应用**

图像到图像转换已经应用在很多领域，比如在人脸面部编辑、图像超分辨率、视频预测以及医学图像转换，这一部分就不具体展开，因为这方面的工作实在过于庞大。

**GAN生成图像的评价指标**

生成图像的质量很难去量化，并且像RMSE这样的度量并不合适，因为合成图像和真实图像之间没有绝对的一对一对应关系。一个常用的主观指标是使用Amazon Mechanical Turk (AMT)，它雇用人类根据他们认为图像的真实程度对合成和真实图像进行评分。然而，人们通常对好的或坏的看法不同，因此我们还需要客观的指标来评估图像的质量。

Inception score (IS)[15]在将类别放入预先训练的图像分类器时，基于类概率分布中的熵来评估图像。初始得分背后的一个直觉是图像x越好，条件分布p(y|x)p(y|x)的熵就越低，这意味着分类器对图像的高度信任。此外，为了鼓励模型生成各种类型的图像，边际分布p(y)=∫p(y|x=G(z))dzp(y)=∫p(y|x=G(z))dz应具有高熵。结合这两个讨论，初始分数由exp(Ex∼G(z)DKL(p(y|x)∥p(y)))exp(Ex∼G(z)DKL(p(y|x)‖p(y)))计算。Inception score既不对标签的先前分布敏感，也不对适当的距离测量敏感。此外，Inception score受到类内模式崩溃的影响，因为模型只需要为每个类生成一个完美样本以获得完美的Inception score，所以Inception score不能反应生成模型到底有没有模式崩溃。

与初始得分相似，FCN-score[29]采用的理念是，如果生成图像是真实的，那么在真实图像上训练的分类者将能够正确地对合成图像进行分类。然而，图像分类器不需要输入图像非常清晰以给出正确的分类，这意味着基于图像分类的度量可能无法在分辨两个图像之间细节上很小的差异。更糟糕的是，分类器的决定不一定取决于图像的视觉内容，但可能受到人类不可见的噪声的高度影响，FCN-score的度量也是存在问题。

Fréchet Inception Distance (FID) [39]提供了一种不同的方法。首先，生成的图像嵌入到初始网络的所选层的潜在特征空间中。其次，将生成的和真实的图像的嵌入视为来自两个连续多元高斯的样本，以便可以计算它们的均值和协方差。然后，生成的图像的质量可以通过两个高斯之间的Fréchet距离来确定：

FID(x,g)=∥μx−μg∥22+Tr(∑x+∑g−2(∑x∑g)12)FID(x,g)=‖μx−μg‖22+Tr(∑x+∑g−2(∑x∑g)12)

上式(μx，μg)(μx，μg)和(∑x,∑g)(∑x,∑g)分别是来自真实数据分布和生成样本的均值和协方差。FID与人类判断一致，并且FID与生成图像的质量之间存在强烈的负相关。此外，FID对噪声的敏感度低于IS，并且可以检测到类内模式崩溃。

**总结**

此博客在论文[An Introduction to Image Synthesis with Generative Adversarial Nets](https://arxiv.org/abs/1803.04469)基础上回顾了GAN的基础知识，图像生成方法的三种主要方法，即直接方法，分层方法和迭代方法和其它生成方法，如迭代采样。也讨论了图像合成的两种主要形式，即文本到图像合成和图像到图像的转换。希望本文可以帮助读者理清GAN在图像生成方向的指导，当然限于原论文（本博客多数内容为翻译原文），还有很多篇精彩的GAN在图像生成方向的论文没有涉及，读者可以自行阅读。

**参考文献**

[1] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.https://arxiv.org/pdf/1312.6114.pdf

[2] van den Oord, Aaron, et al. “Conditional image generation with pixelcnn decoders.” *Advances in Neural Information Processing Systems*. 2016.https://arxiv.org/abs/1606.05328

[3] Kingma, Durk P., and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions.” *Advances in Neural Information Processing Systems*. 2018.https://arxiv.org/pdf/1807.03039.pdf

[4] Goodfellow, Ian, et al. “Generative adversarial nets.” *Advances in neural information processing systems*. 2014.https://arxiv.org/abs/1406.2661

[5] A. Radford, L. Metz, and S. Chintala, “Unsupervised represetation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.https://arxiv.org/abs/1511.06434

[6] M. Mirza and S. Osindero, “Conditional generative adversarial nets,”arXiv preprint arXiv:1411.1784, 2014.https://arxiv.org/abs/1411.1784

[7] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” arXiv preprint arXiv:1610.09585,2016.https://arxiv.org/abs/1610.09585

[8] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” arXiv preprint arXiv:1605.09782, 2016.https://arxiv.org/abs/1605.09782

[9] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, “Adversarially learned inference,”arXiv preprint arXiv:1606.00704, 2016.https://arxiv.org/abs/1606.00704

[10] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther,“Autoencoding beyond pixels using a learned similarity metric,”arXiv preprint arXiv:1512.09300, 2015.https://arxiv.org/abs/1512.09300

[11] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, “Mode regularized generative adversarial networks,” arXiv preprint arXiv:1612.02136, 2016.https://arxiv.org/abs/1612.02136

[12] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” arXiv preprint arXiv:1701.07875, 2017.https://arxiv.org/abs/1701.07875

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of wasserstein gan,” arXiv preprint arXiv:1704.00028, 2017.https://arxiv.org/abs/1704.00028

[14] Miyato, Takeru, et al. “Spectral normalization for generative adversarial networks.” *arXiv preprint arXiv:1802.05957* (2018).https://arxiv.org/abs/1802.05957

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford,and X. Chen, “Improved techniques for training gans,” in Advances in Neural Information Processing Systems, 2016, pp. 2226–2234.https://arxiv.org/abs/1606.03498

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in Advances In Neural Information Processing Systems, 2016, pp. 2172–2180.https://arxiv.org/abs/1606.03657

[17] S. Nowozin, B. Cseke, and R. Tomioka, “f-gan: Training generative neural samplers using variational divergence minimization,”arXiv preprint arXiv:1606.00709, 2016.https://arxiv.org/abs/1606.00709

[18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee,“Generative adversarial text to image synthesis,” arXiv preprint arXiv:1605.05396, 2016.https://arxiv.org/abs/1605.05396

[19] X. Wang and A. Gupta, “Generative image modeling using style and structure adversarial networks,” arXiv preprint arXiv:1603.05631, 2016.https://arxiv.org/abs/1603.05631

[20] E. L. Denton, S. Chintala, a. szlam, and R. Fergus, “Deep generative image models using a laplacian pyramid of adversarial networks,” in Advances in Neural Information Processing Systems Curran Associates, Inc., 2015, pp. 1486–1494.https://arxiv.org/abs/1506.05751

[21] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” arXiv preprint arXiv:1612.03242, 2016.https://arxiv.org/abs/1612.03242

[22] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, “Stacked generative adversarial networks,” arXiv preprint arXiv:1612.04357, 2016.https://arxiv.org/abs/1612.04357

[23] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune,“Plug & play generative networks: Conditional iterative generation of images in latent space,” arXiv preprint arXiv:1612.00005,2016.https://arxiv.org/abs/1612.00005

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017.https://arxiv.org/abs/1710.10196

[25] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Z. Afzal, and M. Liwicki,“Tac-gan-text conditioned auxiliary classifier generative adversarial network,” arXiv preprint arXiv:1703.06412, 2017.https://arxiv.org/abs/1703.06412

[26] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee,“Learning what and where to draw,” in Advances in Neural Information Processing Systems, 2016, pp. 217–225.https://arxiv.org/abs/1610.02454

[27] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” CoRR, vol. abs/1710.10916,2017. http://arxiv.org/abs/1710.10916

[28] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang,and X. He, “Attngan: Fine-grained text to image generation with attentional generative adversarial networks,” arXiv preprint arXiv:1711.10485, 2017.https://arxiv.org/abs/1711.10485

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, 2016.https://arxiv.org/abs/1611.07004

[30] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, “Pixel-level domain transfer,” in European Conference on Computer Vision. Springer, 2016, pp. 517–532.https://arxiv.org/abs/1603.07442

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,”arXiv preprint arXiv:1703.10593, 2017.https://arxiv.org/abs/1703.10593

[32] Z. Yi, H. Zhang, P. T. Gong et al., “Dualgan: Unsupervised dual learning for image-to-image translation,” arXiv preprint arXiv:1704.02510, 2017.https://arxiv.org/abs/1704.02510

[33] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,”arXiv preprint arXiv:1703.05192, 2017.https://arxiv.org/abs/1703.05192

[34] S. Benaim and L. Wolf, “One-sided unsupervised domain mapping,” arXiv preprint arXiv:1706.00826, 2017.https://arxiv.org/abs/1706.00826

[35] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” arXiv preprint arXiv:1611.02200, 2016.https://arxiv.org/abs/1611.02200

[36] neural information processing systems, 2014, pp. 2366–2374.M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in Advances in Neural Information Processing Systems, 2017, pp. 700–708.https://arxiv.org/abs/1703.00848

[37] M.-Y. Liu and O. Tuzel, “Coupled generative adversarial networks,” in Advances in neural information processing systems, 2016,pp. 469–477.https://arxiv.org/abs/1606.07536

[38] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” arXiv preprint arXiv:1711.09020,2017.https://arxiv.org/abs/1711.09020

[39] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler,G. Klambauer, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a nash equilibrium,” CoRR, vol. abs/1706.08500, 2017. http://arxiv.org/abs/1706.08500

[40] Huang, He, Phillip S. Yu, and Changhu Wang. “An Introduction to Image Synthesis with Generative Adversarial Nets.” *arXiv preprint arXiv:1803.04469* (2018).http://arxiv.org/abs/1803.04469