**基于GAN的文本引导图像生成器**

生成对抗网络(GANs)是在2014年就已经出现的图像生成模型，并在后续逐渐发展和丰富，成为图像生成的早期范式。GAN将图像的生成任务看作生成器模型与判别器模型之间的对抗与竞争，通过生成器网络尽可能生成真实的图片，而判别器则尽可能区分生成的图片与真实图片。通过生成器与判别器之间的竞争，使得生成器最终得到能够欺骗判别器的图片。

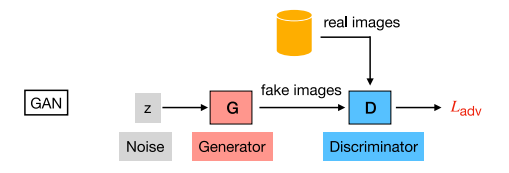
在本节中，我们从GAN的基础内容与变体出发，逐步过渡到基于GAN的T2I合成模型。我们将在第一节讲述GAN生成器的基础生成方法，以及如何使用条件控制GAN的生成效果；接着，在第二节，将会重点讲述以文本作为条件的GAN模型设计发展过程；而在第三节，将对基于GAN的生成范式的优缺点进行总结。

值得一提的是，基于GAN的T2I模型早在2016年就已经提出，直到现在已经经过了长时间的发展和改进，提出了各式各样数量丰富的模型，因此，我们并没有尝试对所有的GAN工作进行总结，而是选择具有代表性的部分工作，简单介绍其发展过程和态势。

**Generative adversarial networks**

传统的GAN方法在2014年提出，包括两个神经网络：生成器(generator)网络G(z)，利用从噪声先验分布中采样的随机噪声作为输入，合成图片；以及判别器(discriminator)网络D(x)，对真实图片与合成图片进行判别。

GAN将图像生成的训练过程看作两个神经网络之间的博弈和竞争。其中，判别器尝试区分真实图像与生成器生成的图像，生成器则尝试捕获真实数据的分布，从而产生能够愚弄判别器的，接近于真实数据分布的生成图像。



如[1]中所属，整个训练过程可以看作”two-player minimax game”：

其中，生成器G(z)尝试最小化上述公式，意味着生成器需要最大化判别器对生成图片的判别分数，也就是愚弄判别器的判别结果，使其赋予生成图片的判别结果接近真实的评价；判别器则尝试最大化上述公式，希望对于真实图片赋予真实的评价，对于生成图片赋予“不真实”的评价。

在GANs发展过程中，等人提出StyleGAN，将风格编码解耦出来，提高生成质量的同时，使得GAN的特征空间具备解耦的性质，允许对风格编码进行编辑，从而实现图像编辑，成为无条件GAN生成器的主流模型。StyleGAN模型结构如下：

**Conditional GAN**

GAN的生成范式使得无条件图像生成成为可能，但是如果能够在生成过程中引入控制条件，提高图像生成的控制性，将会使得图像生成任务的应用场景更加广泛，并且能够提高图像生成的应用价值。因此，在现有的生成范式中，如何引入条件的控制能力，成为后续研究的重要领域。

条件GAN最早在2014年由Mirza等人提出。通过在生成器和鉴别器上额外引入条件变量y，控制模型生成特定的数字。在他们的工作中，通过将随机采样的与条件表征y一起输入到多层感知网络(MLP)中，学习两者的联合隐藏特征应用于生成过程；而在判别器中，将条件与图像输入到判别器进行判别。如Mirza and Osindero(2014)所述，式(1)变为式(2)：

条件GAN的基础范式提出后，后续的工作继续在条件形式（标签、图片、文本等）、生成器判别器的条件表示形式、判别器判别目标与打分方式等方面进一步发展改进，延伸出了各式各样的工作。而我们所重点研究的，就是以文本作为条件控制生成的子任务。

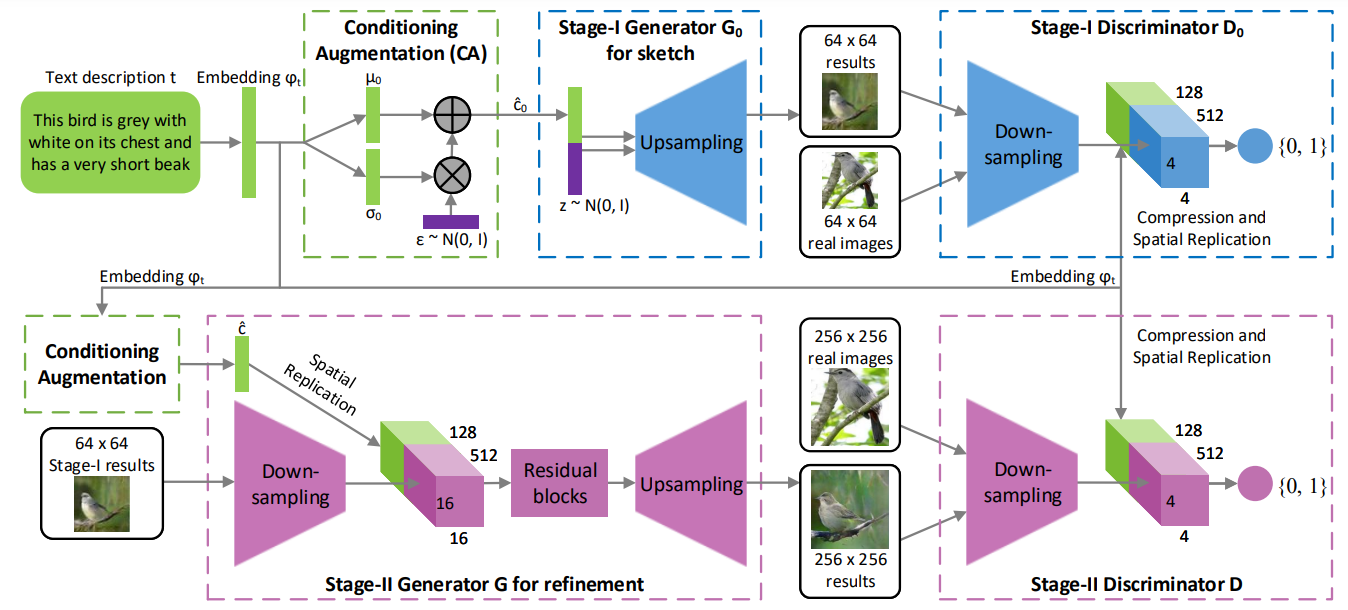
**基于GAN的T2I模型发展**

利用文本作为条件控制GAN的生成，可以看作cGAN的一种具体形式。与最早的以标签作为条件的GAN相比，文本条件蕴含更加丰富的信息内容，因此如何获取文本中蕴含的信息，以及如何控制文本信息对于图像生成的监督能力，使得生成的图像与文本更加匹配；此外，也有的工作尝试生成更高分辨率的图像。

在前一章中回顾了GAN和cGAN的基础范式之后，我们将进一步研究基于GAN的T2I模型发展历程。在Reed等人在2016年提出的第一个T2I方法后，基于GAN的T2I模型便沿着对应的范式，在模型结构设计、特征融合、条件监督等方面不断发展。

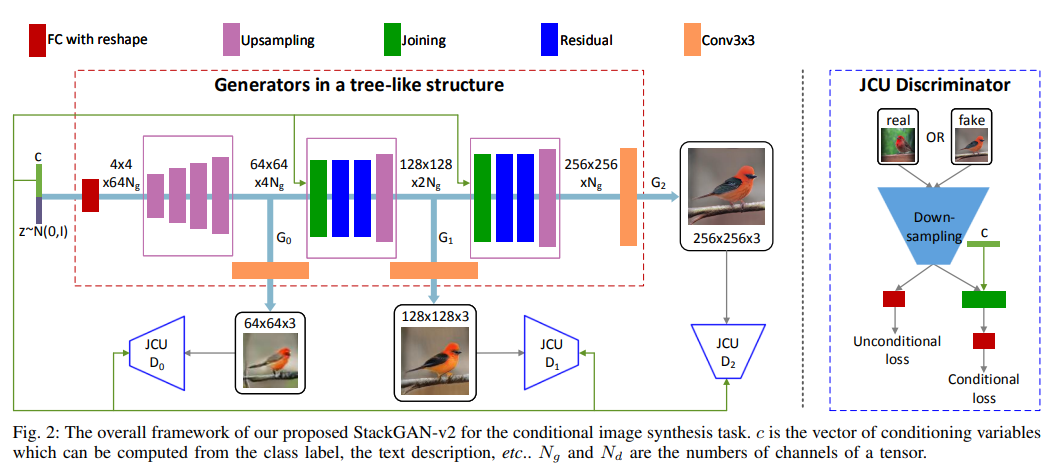
基于GAN的第一个T2I模型是cGAN的简单扩展，利用预训练文本编码器提取文本的全局句子特征(sentence embedding)，替换cGAN中的标签作为GAN的生成条件，判别器则判别真实的文本对与合成的文本对；为了是的生成器与判别器更加关注于文本与图像之间的匹配，在GAN-INT-CLS模型中，判别器需要对真实图像与匹配文本、生成图像与匹配文本、真实图像与不匹配文本间进行判别。

在初步解决文本-图像匹配问题后，基于GAN的T2I模型向着更大分辨率的生成发展。为了生成较高分辨率(256 x 256)的图像，StackGAN提出了堆叠式生成器结构，并被后续的工作不断沿用和改进。



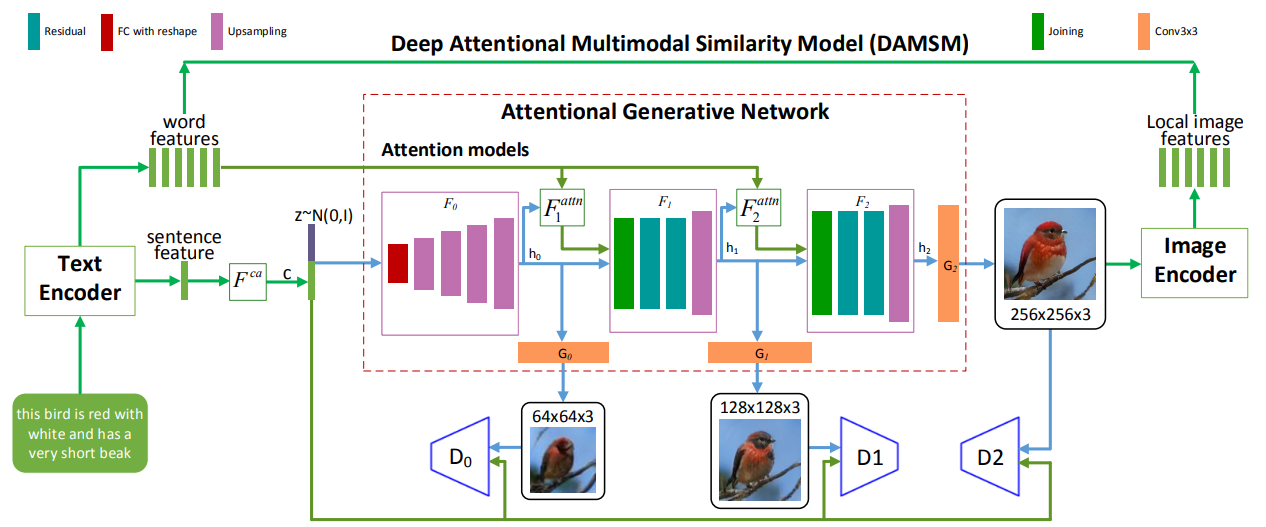
如上图所示，StackGAN使用堆叠式的生成器架构，其生成过程分为两个阶段。在第一阶段，利用文本获取sentence embedding后，使用文本向量对随机采样的正态分布噪声进行条件增强，然后输入到生成器中生成图像，并与真实图像进行判别。在第二阶段，初始粗糙图片和文本输入到二阶段生成器中，在特征层面继续进行增强融合，实现后续生成判别。而每个阶段的判别器均尝试区分匹配和非匹配的图像-文本对。

随后，StackGAN++ (Zhang等人，2017)通过端到端框架进一步改进了架构：



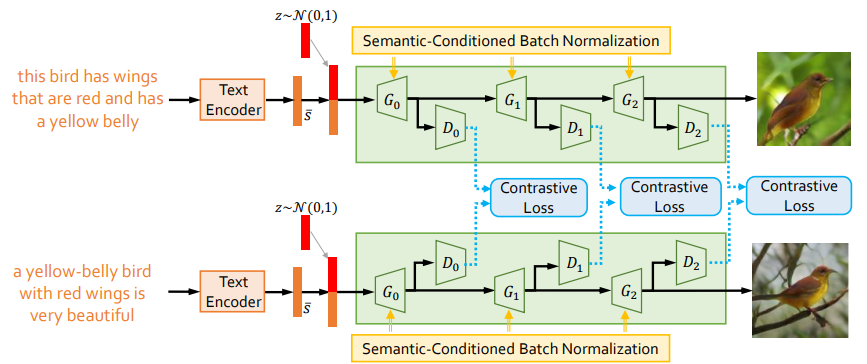
如上图，StackGAN++使用三级级联生成结构，每一级结构负责特定分辨率图像的生成和判别。这种级联范式被后续的许多工作继承和发展。

沿着StackGAN的框架和思路，AttnGAN在网络中引入了注意力机制，使得文本条件能够更加充分的融入到生成过程中。如下图，AttnGAN采用了与StackGAN++同样的级联式生成结构：



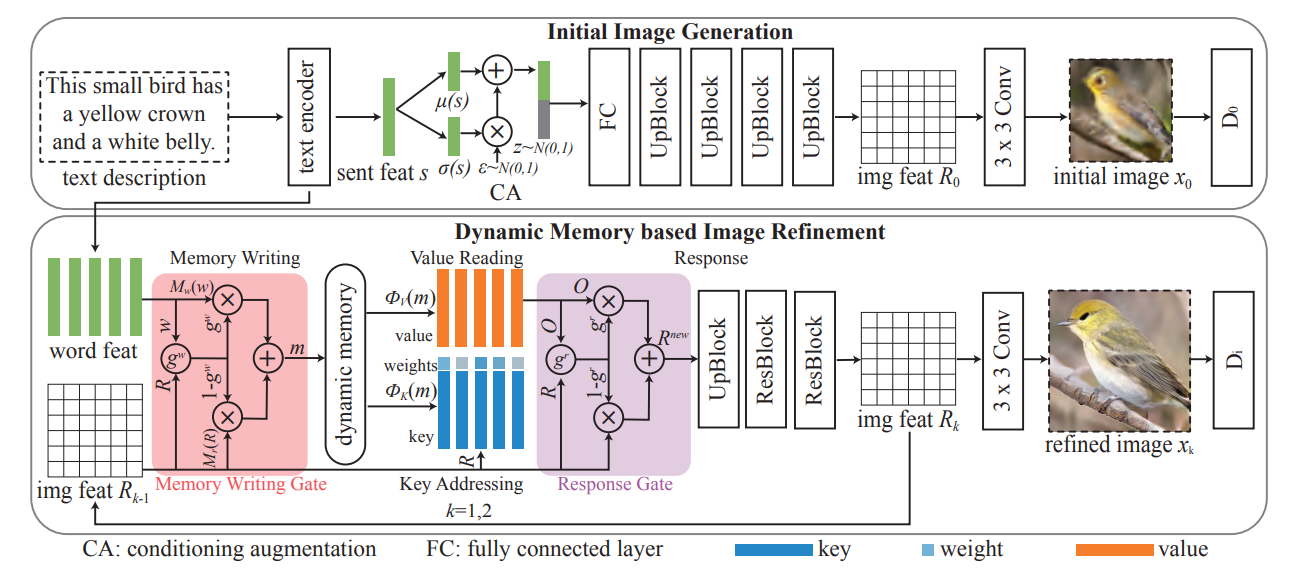
如上图，在标准的级联结构下，AttnGAN更加强调文本中细节内容的生成。AttnGAN认为，传统的T2I生成GAN网络，直接使用sentence feature作为文本信息控制生成，这种sentence feature仅仅捕获文本的全局信息，对于特定的单词和细节，没有很好的注意。因此，AttenGAN额外提取单词的word feature作为辅助信息，利用注意力框架将辅助信息融合到每一级的生成模块中，在全局句子向量的基础上，根据相关单词合成细粒度的细节，从而实现更好的图像-文本对齐；此外，他们还将图像划分为各个子区域，关注子区域与单词之间的关联，在训练阶段使用深度注意多模态相似性模型损失(DAMSM)来衡量句子、单词与图像、区域之间的相似性。后续工作沿用AttnGAN所提出的word feature、注意力融合框架、DAMSM等内容，进一步做出改进。

后续工作的关注重点进一步放在提升文本与生成图像的匹配性上，为此，许多工作尝试引入孪生网络、记忆网络、各式对比损失等机制，提升文本信息的融合和监督能力。SD-GAN (Yin等人，2019)尝试在生成框架中引入孪生网络，使用两个分支实现生成：



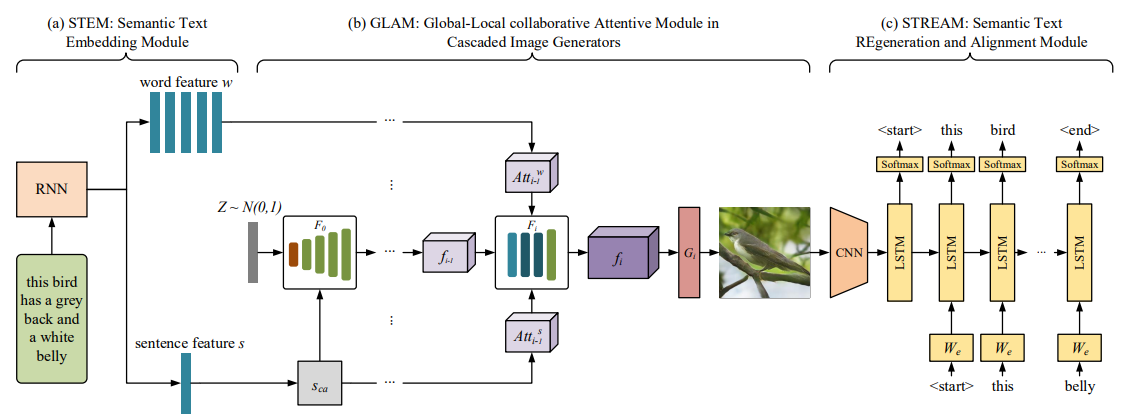
如图所示，SD-GAN处理两条变体文本，使用共享参数的双分支模型，分别生成两张图象，在双分支的级联生成过程中，每一层级都是用对比损失，如果两条文本语义相同，则使得模型最大化生成的两张图片之间的相似度，如果表示的语义不同，则最小化两张图片的相似度。通过对比损失的监督，使得生成器能够捕获文本的公共域，学习到具体的语义特征。

而DM-GAN为首的工作，则尝试利用记忆网络来提升文本特征信息的引导能力：



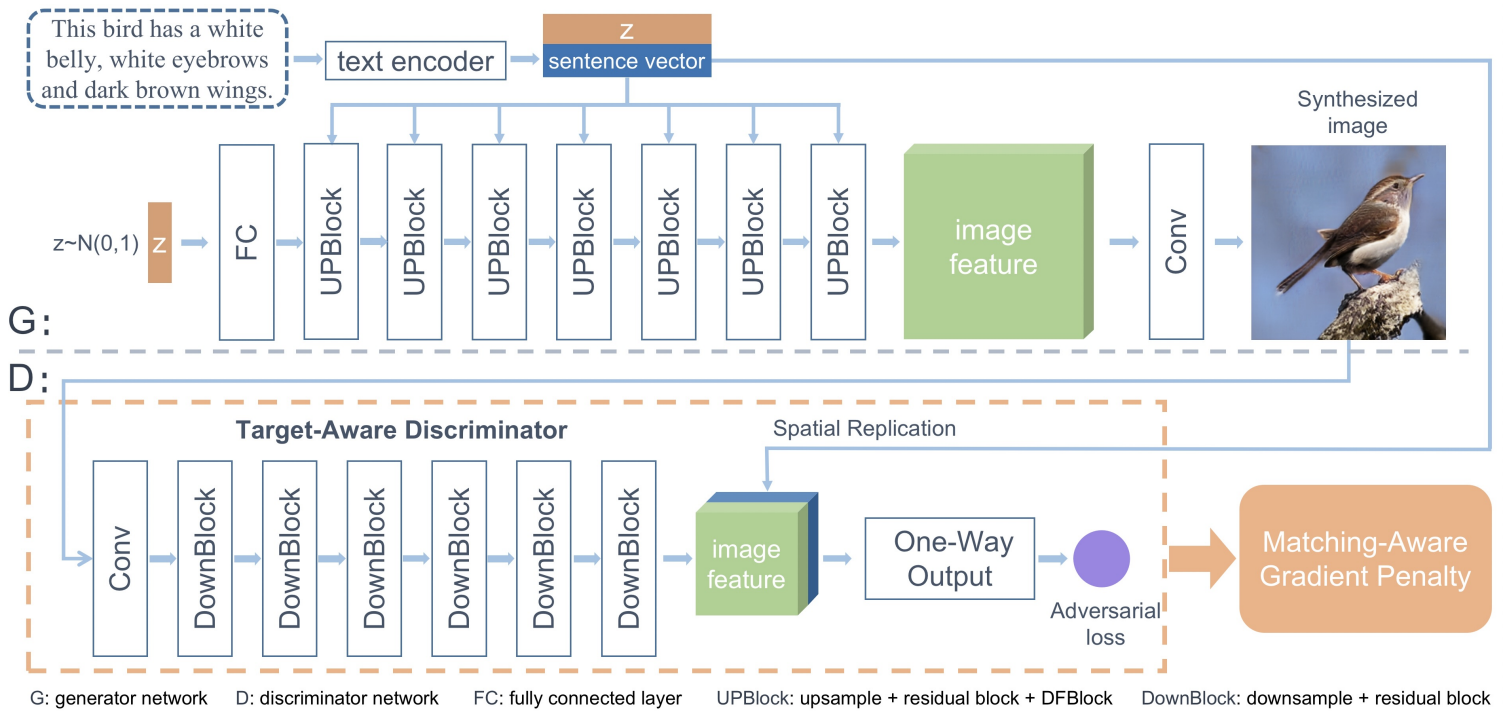
DM-GAN依然遵循着多阶段的生成框架，只是在生成精细图片的后续阶段，引入记忆网络，实现细节特征的进一步优化。记忆写入门(Written Gate)以初始的图像与单词特征作为输入，通过计算每个单词的重要性，结合单词与图像特征写入记忆网络中存储。在后续的生成时，计算记忆网络中键值与图像特征间的相似度，对记忆模块中存储的细节特征进行加权求和并返回到后续生成模块中作为输入，从而更新图像特征，使得图像生成的信息与细节更加丰富。

此外，MirrorGAN除了利用判别器保证图像与文本之间的匹配性之外，额外引入对齐流模块，对生成的图像，额外使用image caption相关模型为生成的图像预测文本，使得预测的文本与原始的文本对齐：



如上图所示，通过语义对齐模块的显式文本预测与对齐，使得模型的图文匹配性进一步提高。

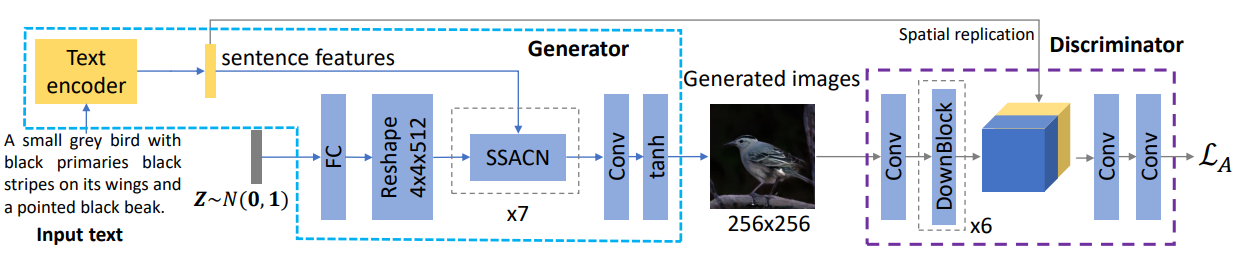
从StackGAN开始逐渐流行的级联式生成结构，虽然有助于生成高分辨率图像，但是DF-GAN认为，这种级联的训练方式，使得模型不同层级间的生成器产生耦合，导致在级联式生成过程中，最终的生成结果仅仅是之前层级特征、细节的简单组合。为此，DF-GAN将级联的生成结构进行简化，在单层级结构中，堆叠大量的特征融合块，仅仅使用单个生成器和判别器，实现高分辨率图像的生成。



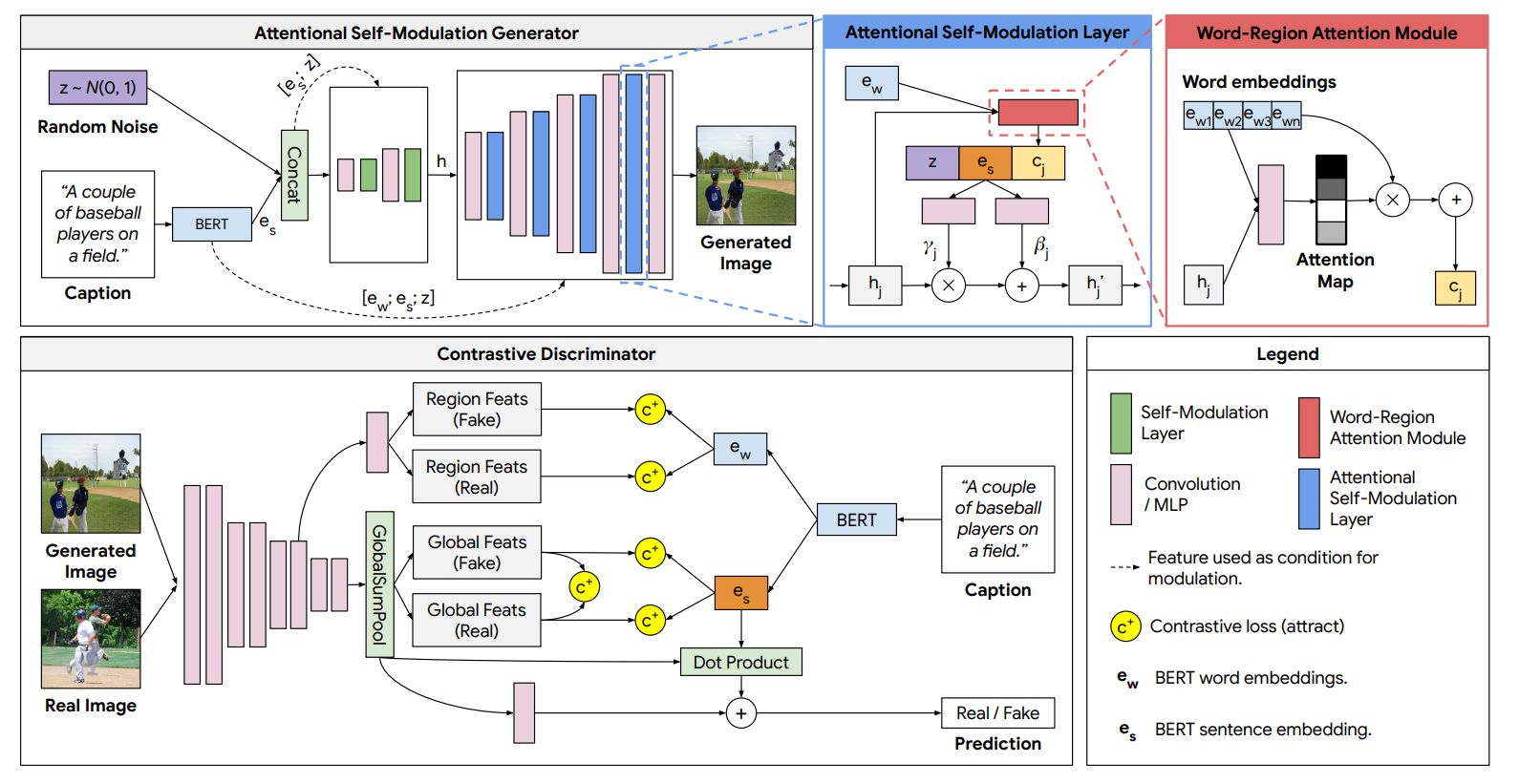
如上图所示，DF-GAN使用单级生成结构，并在生成器中堆叠大量的UPBlock，实现图像特征的逐级扩大，从而最终得到高分辨率的图像；此外，在单个UPBlock中，设计相应的特征融合结构，实现文本特征的深层融合。

从DF-GAN之后，人们逐渐意识到，在生成高分辨率图像时，不需要多级生成器、判别器架构，只需要在模型层面上利用更高的特征融合处理层进行处理，通过加深网络就可以实现图像生成，因此后续也有许多工作从这一角度入手，沿着单阶段图像生成范式进行图像生成，能够缓解多个生成判别模型难以训练、推断时间较长带来的问题。

SSA-GAN就是在单级图像生成架构上进行的尝试。如下图所示，SSA-GAN沿着与DF-GAN相似的架构进行设计，只是在生成器中堆叠了自主设计的多个SSACN模块。SSACN模块为了更好的实现特征融合，引入了conditional Batch Normalization机制，利用文本条件构造BN过程中所需要的γ与β参数，实现文本特征的融合；此外，在SSACN模块中，利用神经网络构建的Mask预测模块，预测目前层级图像仍然需要优化的细节部分，从而更好的控制需要进一步优化、提升的图像区域。

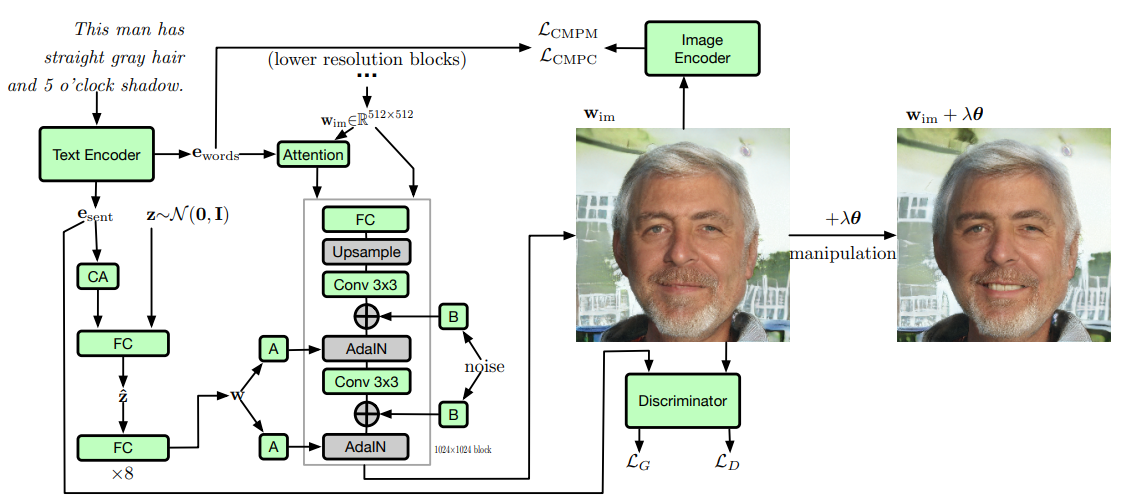


此外，利用对比学习损失提升图像与文本匹配性，也是基于GAN的T2I模型经常使用的技术。XMC-GAN就在单个模型中使用三种对比损失，强调图像与文本、单词与区域、区域与区域之间的关联。通过这种对比损失的引入，显著增强了模型对于图像与文本间关联的理解与控制能力，也使得图像生成的质量显著提高。



上述工作均以StackGAN++架构作为起点，使用自主设计的网络模块架构实现T2I模型的训练和生成，此外，也有部分工作沿着StyleGAN这一经典的无条件生成器架构，尝试将文本引入到StyleGAN当中，控制模型的生成。

TextStyleGAN就是尝试将文本引入StyleGAN的经典工作。他们使用与StackGAN相似的条件增强机制，将文本与随机采样的噪声融合，进一步得到StyleGAN所需要的风格编码；此外，利用提取的单词特征与注意力机制，将文本单词信息融入到每一层风格调制模块当中，并在损失函数中额外控制生成图像与文本特征间相似度，控制模型的图文匹配能力。



鉴于StyleGAN更加强大的图像生成与风格解耦能力，现如今的许多工作往往沿着StyeGAN模型的架构进行设计。Lafite系列模型就是以StyleGAN2作为生成器架构，在其中引入条件信息。只是Lafite模型为了减少训练过程中对于训练数据文本对的需求，采用了无文本生成训练的设定，因此利用图像的CLIP特征，加上噪声扰动构建伪文本，从而完成无监督或半监督的生成器训练；而Lafite2则进一步优化伪文本构建过程，利用CLIP本身的词汇库与自己设定的一系列模板，构建候选文本语句，与输入图像计算CLIP相似度，筛选得到可能的候选文本，并在CLIP空间中对文本进行进一步迭代优化。

基于StyleGAN架构的T2I模型，

T2I模型评估指标

T2I模型数据集