

# StyleGAN

cvpr 2019

Notes written by h1astro

提出了真正的1024\*1024 分辨率大规模人脸数据集

基于StyleGAN架构，更容易开展人脸编辑的相关研究

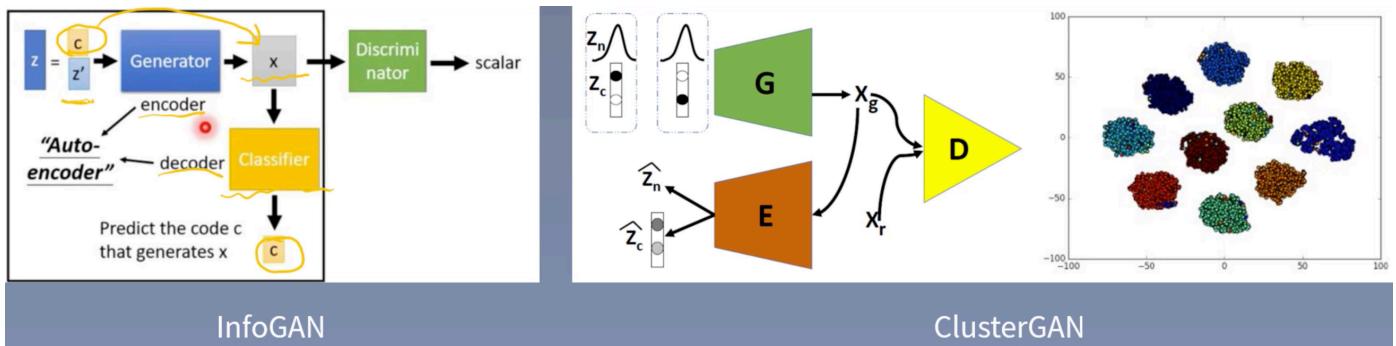
## 核心要点

- 从风格迁移的研究中进行借鉴，提出了GAN的新生成器架构
- 可以自动对图像的高级属性（姿态）和随机变化的图像细节（头发）进行无监督的分离
- 可以直观的、按照特定的尺寸来控制生成效果
- 在SOA的基础上提升了生成质量，并拥有更好的插值性能，还对隐变量进行了更好的解耦
- 提出了两种新方法来对插值质量和隐变量解耦程度进行定量评价
- 提出了一个新的高多样性高分辨率的人脸图像数据集

## 研究背景

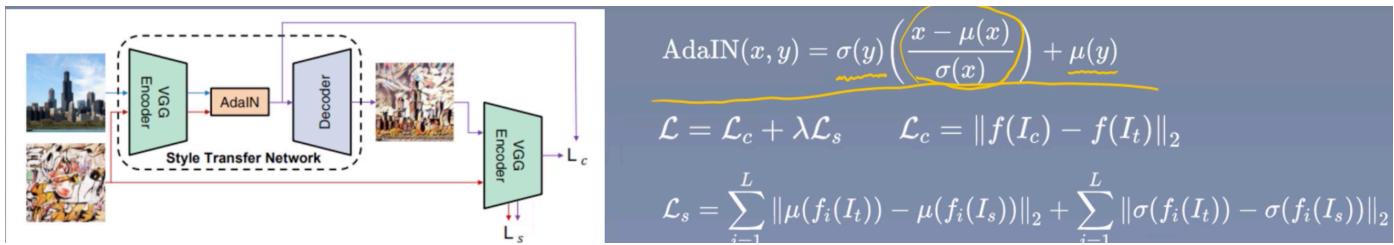
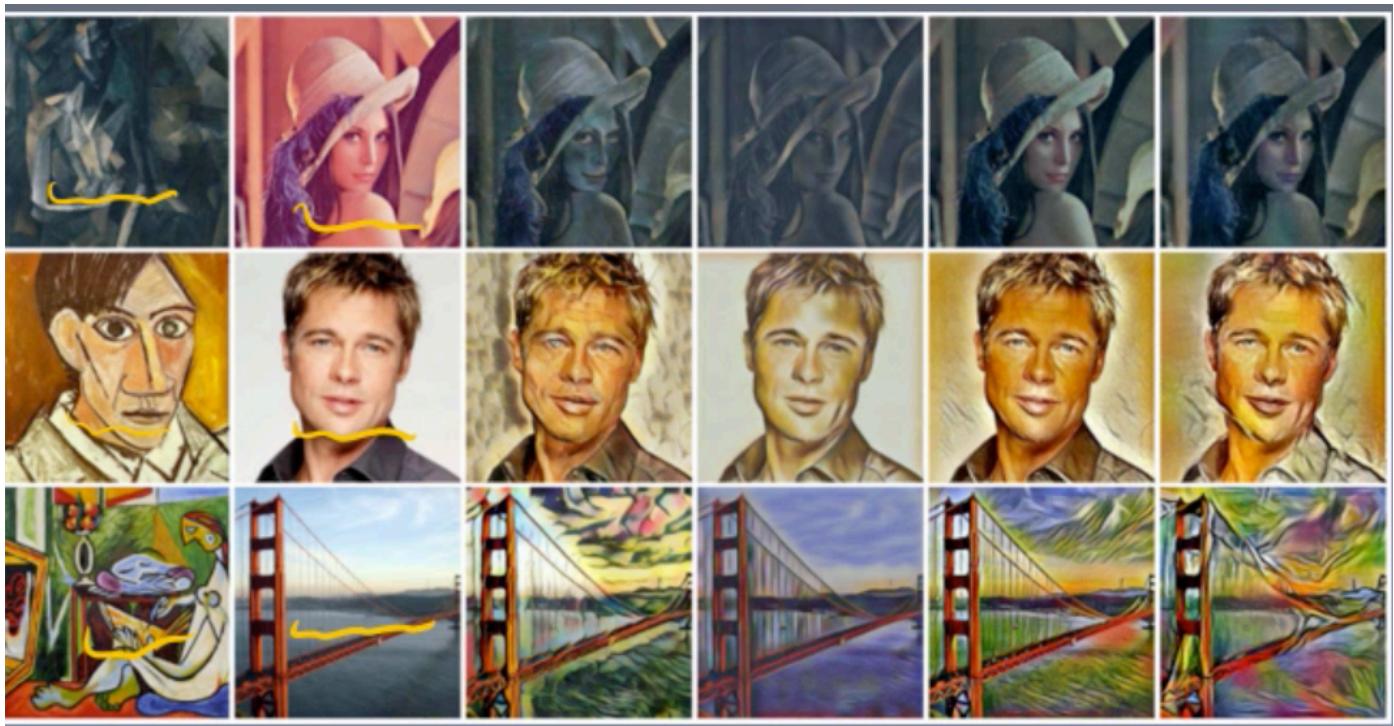
### 隐空间研究

- GAN中的隐变量 $z$ 存在纠缠，使得难以单独控制或改变生成图像的指定特征
- 此前的GAN网络，往往通过添加一个额外的分类器，来对隐变量 $z$ 进行接耦



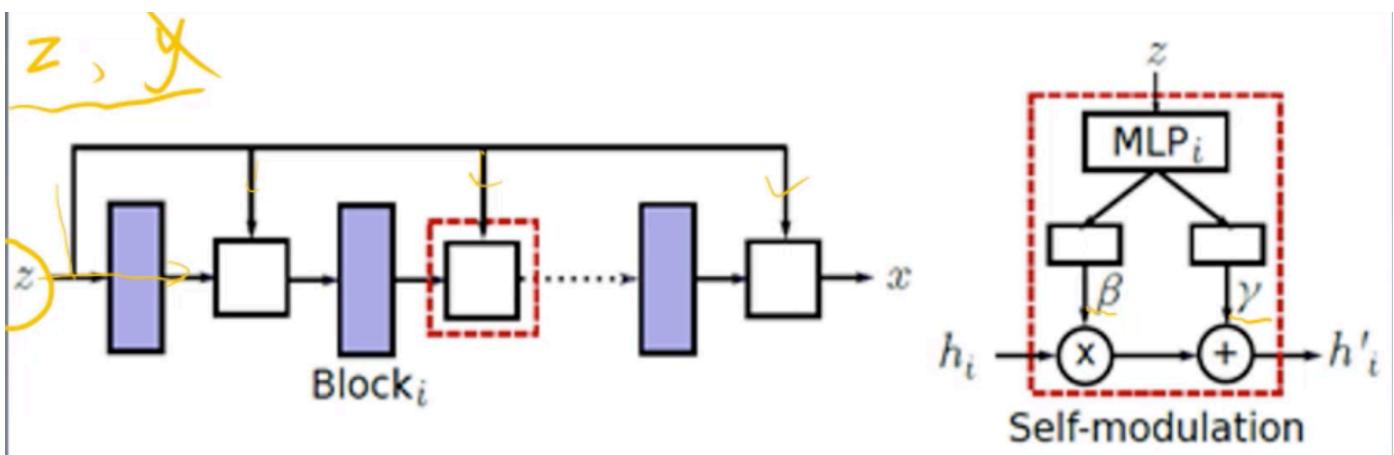
### 图像风格迁移

- 此前的风格迁移方法，一种网络一般只能适应一种风格，并且速度很慢
- 《Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization》，受到Instance normalization启发，发现特征图的均值和方差中带有图像的风格信息
- 基于AdaIN，可以快速实现任意图像风格的转换



### Self-Modulation

- 外部的条件（类别标签）有利于提高GAN的性能，但是这些条件有时并不存在
- 基于无监督的方式，使用噪声 $z$ 代替原本的外部条件
- $\gamma(z)$ 和 $\beta(z)$ 采用两层全连接网络实现
- $h_l' = \gamma_l(z) \odot \frac{h_l - \mu}{\sigma} + \beta_l(z)$

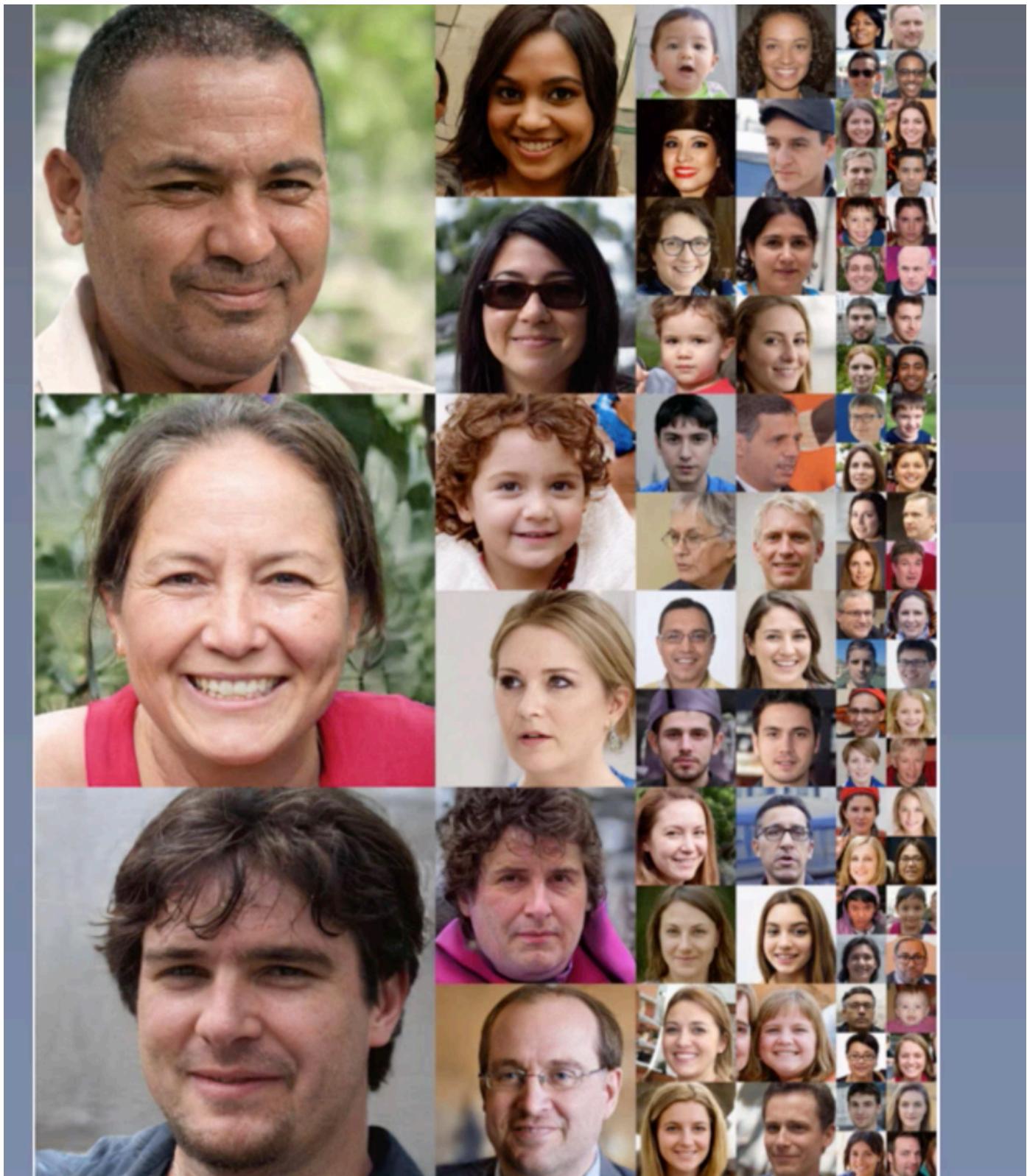


Score		Unconditional		G-Cond		P-cGAN	
		Baseline	Self-Mod	Baseline	Self-Mod	Baseline	Self-Mod
CIFAR10	FID	20.41	<b>18.58</b>	21.08	<b>18.39</b>	16.06	<b>14.19</b>
IMAGENET	FID	81.07	<b>69.53</b>	80.43	<b>68.93</b>	70.28	<b>66.09</b>
CIFAR10	IS	7.89	<b>8.31</b>	8.11	<b>8.34</b>	8.53	<b>8.71</b>
IMAGENET	IS	11.16	<b>12.52</b>	11.16	<b>12.48</b>	13.62	<b>14.14</b>

## 研究成果

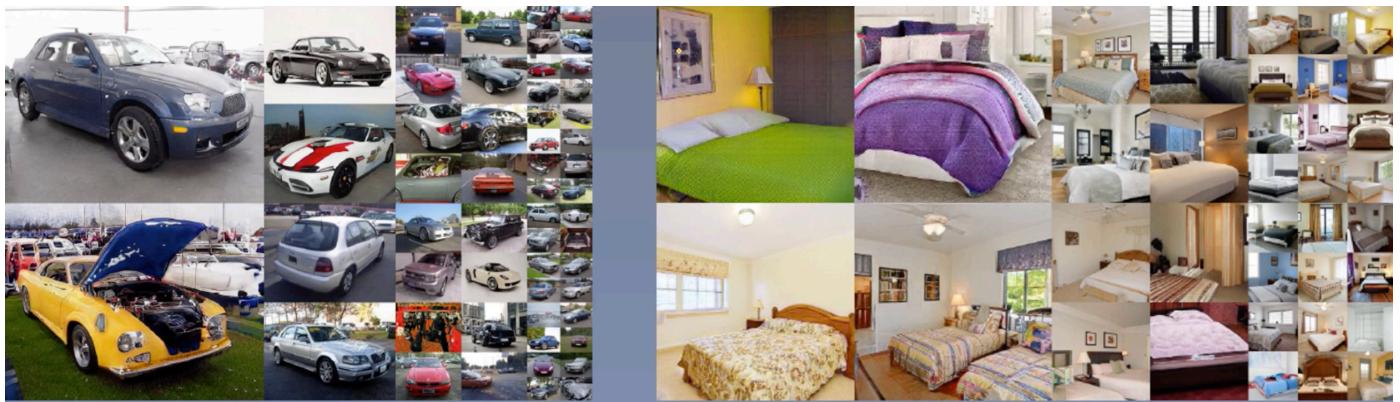
- 进一步提升了人脸生成的质量
- 对隐变量进行解耦，可以控制指定尺度的特征
- 提出了对隐空间进行量化分析对指标
- 提出了真正的1024\*1024 分辨率大规模人脸数据集





## 研究意义

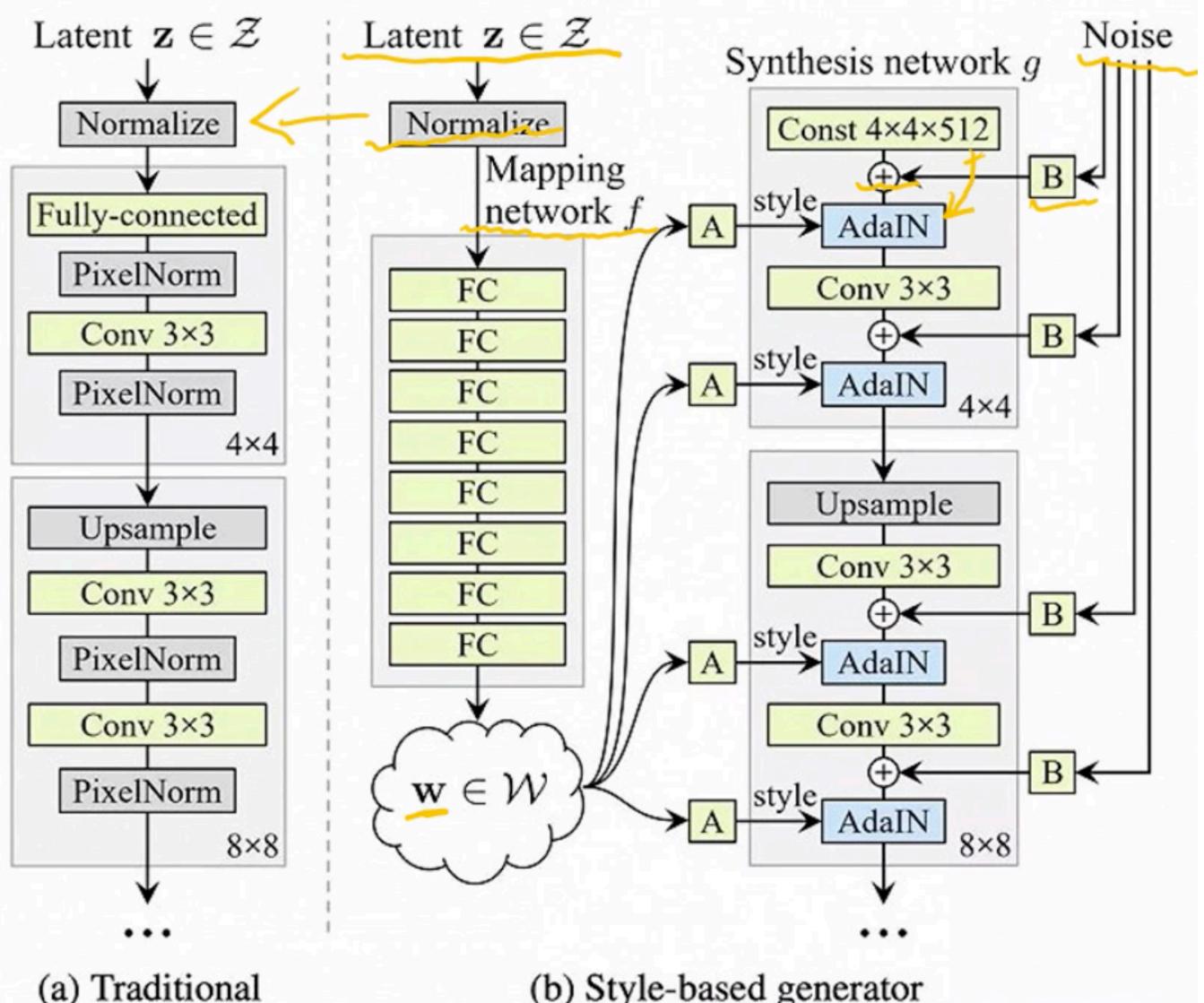
- 基于StyleGAN架构，很容易进一步开展人脸编辑的相关研究
- 启发了对隐空间的进一步探索
- 新的高清人脸数据集，提出了对GAN新的挑战



## 基于样式的生成器

### 总体结构

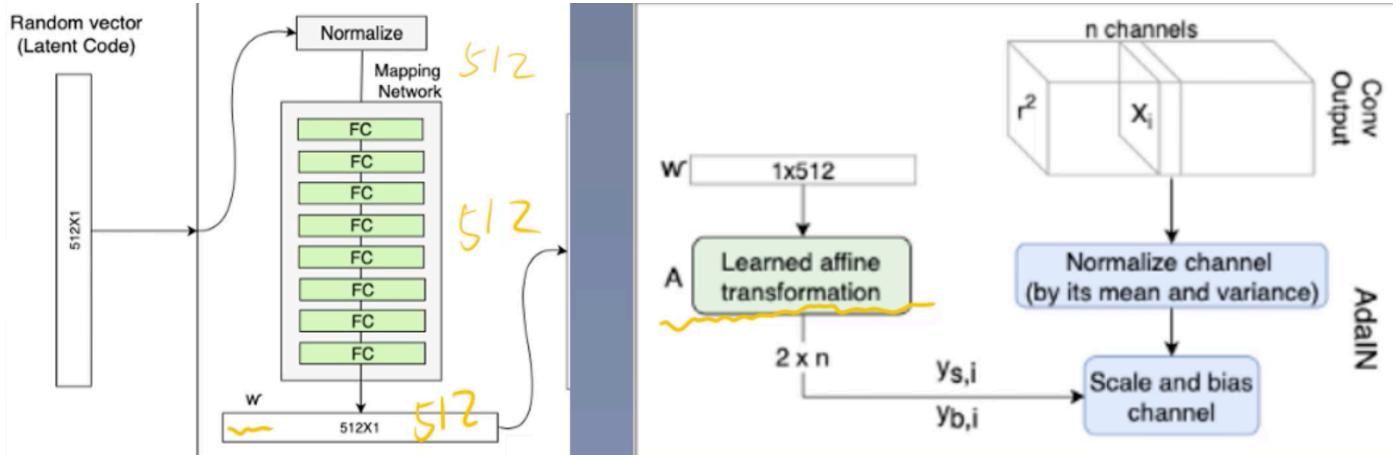
- 把传统生成器第一层输入的随机向量改为常量
- 通过一个全连接映射网络，对隐变量空间采样得到的z进行映射，得到中间隐变量w
- 将AdaIN添加到网络每一层中，使得w可以对每个尺度的整体style进行控制
- 每一层中加入随机高斯噪声，并在B对每个通道的噪声进行缩放，缩放系数可以学习



## Style 控制

- 映射网络是8层的全连接网络，所有层的channel都为512（简单起见）
- 在每个尺度上，通过一个全连接层A学习放射变换，将中间隐变量w转化为样式向量 $y = (y_s, y_b)$
- 对每一层的特征 $x_i$ 进行实例归一化，并用对应层的样式向量 $y_i$ 进行线性变换

$$AdaIN(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}$$



## 训练参数

- 从4\*4到1024\*1024分辨率，每个分辨率两层卷积，所以合成网络有 $9*2=18$ 个卷积层和AdaIN层
- 相同深度和宽度下，传统的参数量为23.1M，而StyleGAN的参数量为26.2M
- 生成器使用的激活函数为 $\alpha=0.2$ 的Leaky ReLU。
- 映射网络使用大学习率时很不稳定，所以令映射网络的学习率 $\lambda' = 0.01 * \lambda$
- 使用和ProGAN相同的判别器、batch size、Adam参数、均衡学习率，生成器推断时同样适用了exponential moving average
- 在CelebA-HQ和FFHQ上训练时，使用了水平翻转的数据增强
- 使用带8个Tesla V100 GPUs的NVIDIA DGX-1来进行训练

## 实验结果

### FFHQ数据集

- 从FLickr上爬去得到，经过了自动对齐和裁剪处理
- 先后使用各种过滤器和AMT众包平台来进行图像筛选，删除了绘画等特殊场景下的人脸
- 7w张1024\*1024分辨率的图像
- 在年龄、种族和图像背景等方面远比CelebA-HQ更多更丰富，并且图像中还包含多种类的配件，例如眼睛、太阳镜、帽子等



- 使用ProGAN作为性能比较的baseline
- 在config B中，上采样由最近邻resize改为bilinear，从8\*8开始生成；在FFHQ上用带R1正则化的原始GAN loss取代WGAN-GP loss，训练时长由12M张图片改为25M；在CelebA-HA的512和1024分辨率上，将学习速率由0.003改为0.002
- 完整配置的StyleGAN即config E，相比于B的ProGAN，FID分数减少20%左右

- 在下面的测试中没有使用截断技巧，如果使用的话，可以只在低分辨率上做截断，这样即使不做额外的正则化也不会带来负效应

Method	CelebA-HQ	FFHQ
A Baseline Progressive GAN [26]	7.79	8.04
B + Tuning (incl. bilinear up/down)	6.11	5.25
C + Add mapping and styles	5.34	4.85
D + Remove traditional input	5.07	4.88
E + Add noise inputs	<b>5.06</b>	4.42
F + Mixing regularization	5.17	<b>4.40</b>

## 生成器的属性分析

### 样式混合

- 由于Instance Normalization会消除输入特征的风格，因此每层的Style向量只能影响当前尺度，到下个AdaIN模块时风格就被重置了
- 训练时，使用两个随机的隐变量z来生成一张图像
- 两个z对应两个中间隐变量w，通过一个随机的crossover point来把两个w合成一下

Mixing regularization	<u>Number of latents during testing</u>			
	1	2	3	4
E 0%	4.42	8.22	12.88	17.41
<u>50%</u>	4.41	6.10	8.71	11.61
<u>F 90%</u>	<b>4.40</b>	<b>5.11</b>	6.88	9.03
<u>100%</u>	4.83	5.17	<b>6.63</b>	<b>8.40</b>



## 随机变化

- 如果像传统生成器一样，在网络的开始部分输入随机向量，这样效率会比较低，并且容易在图像中生成重复的模式
- 由于添加了噪声，使得图像的细节部分发生了一些随机的变化，但是人脸的整体结构等高级特征毫无差异
- 如果网络试图使用噪声来控制整体特征，由于不同像素的噪声没有关联，将很容易产生不一致的内容，从而被判别器惩罚



(a) Generated image    (b) Stochastic variation    (c) Standard deviation

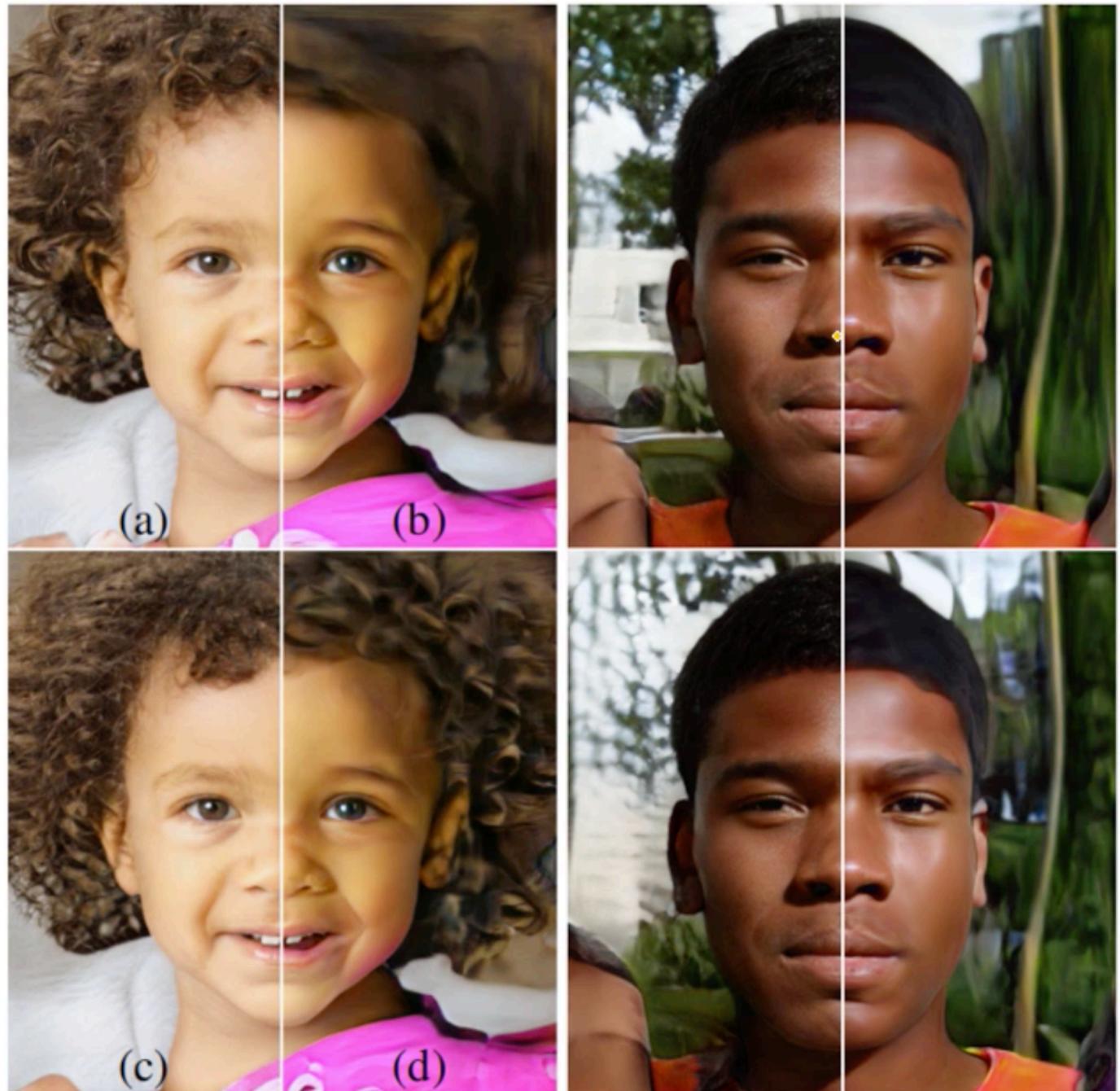
(a) 在每一层添加噪声

(b) 没有噪声

(c) 仅在大分辨率上添加噪声 ( $64^2-1024^2$ )

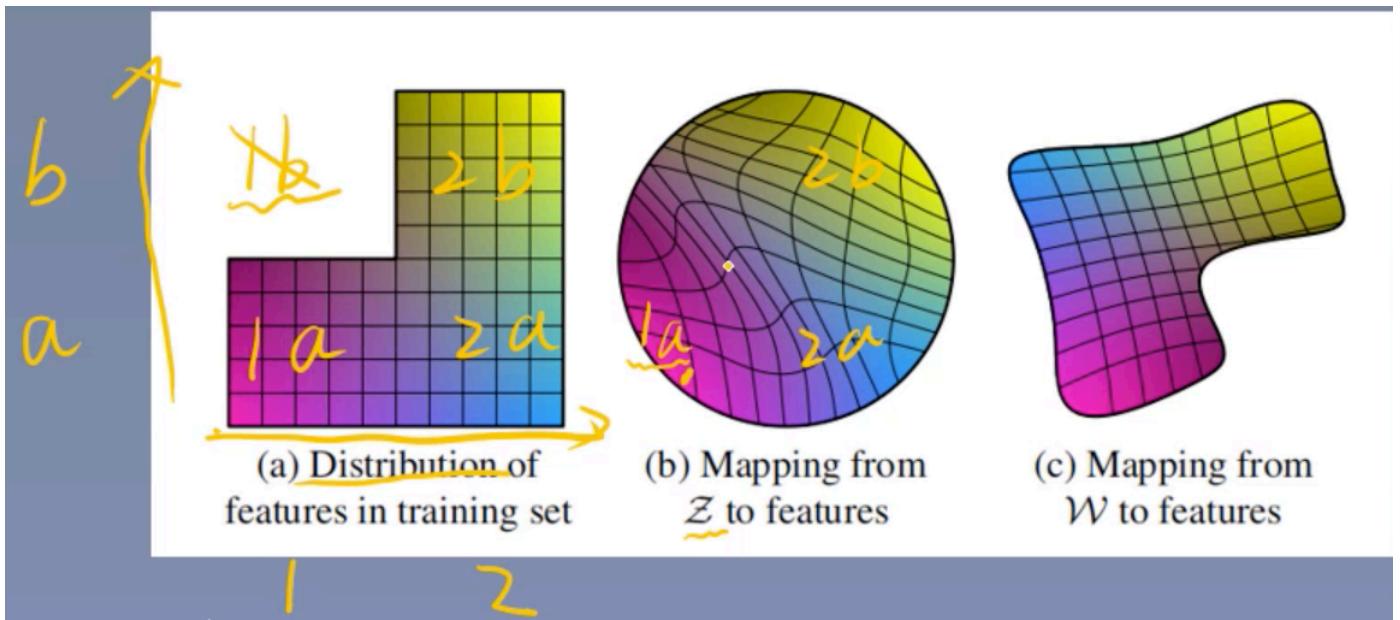
(d) 仅在小分辨率上添加噪声 ( $4^2-32^2$ )

- 没有噪声时会让生成图像过于平滑，更像是绘画
- 在小分辨率上加入噪声会导致头发大幅度的卷曲，并且背景有了更多的内容
- 在大分辨率上加入噪声带来了更多的细节变化



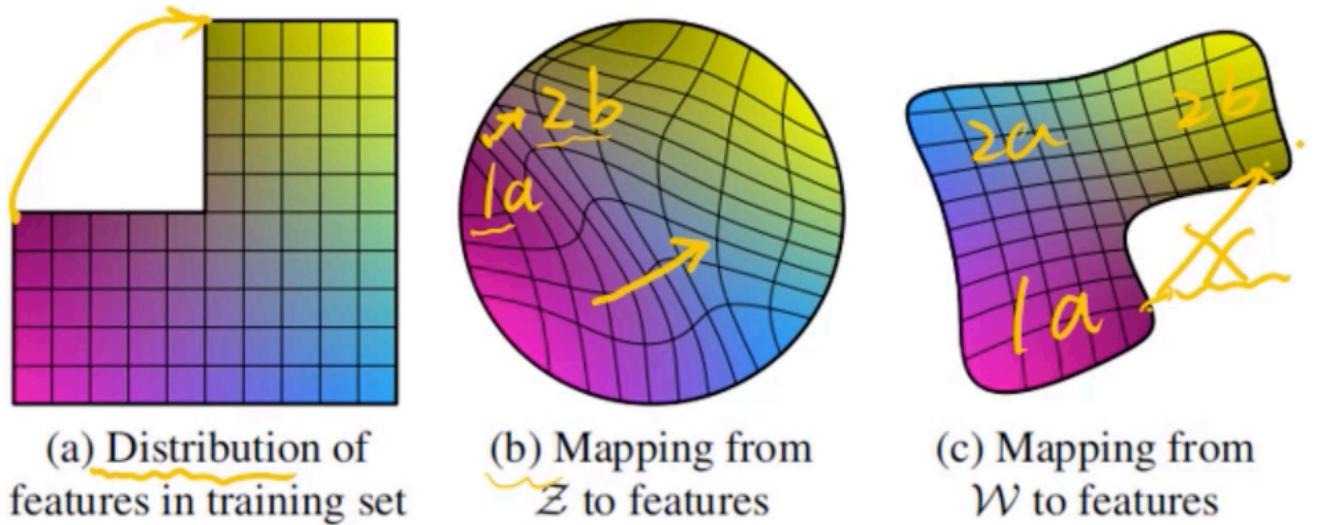
## 隐变量解耦

- 隐变量 $z$ 的分布固定，受训练数据的特征属性分布影响，相对于图像特征的映射空间容易出现卷曲



上图b里有突变 如从1a->2b

- 通过全连接网络生成的中间隐变量w，其分布由 $f(z)$ 决定，是可变的，w到图像特征的映射空间可以随意修改；观察到这个映射空间会更加线性，可能是因为线性的映射对提升生成质量有益
- 此前的用于量化解耦程度的指标，都需要一个额外的编码网络来将图像映射到隐变量。
- 提出了两种没有不需要编码和已知变化因子的量化指标，可以适用于任意生成器和图像数据集



线性变换

### 感知路径长度

- 由于路径长度与图像特征之间的纠缠，对于隐空间的插值可能会在生成图像产生非线形的变化，比如出现两张原始图像中本不存在的特征
- 通过对插值图像的变化程度进行量化，来了解隐空间到图像特征之间的纠缠程度
- 使用VGG16提取的特征的加权差异，来表示一对图像间的感知距离
- 把两个隐变量间的路径进行线形切分，求每一小段傻姑娘两端隐变量所生成图像的感知距离，所有距离之和就是这两个隐变量间的全感知路径长度（PPL）
- 使用10w个样本，分别对z和w计算PPL，由于z已经经过了归一化，所以对z使用球面插值slerp，而对w使用线形插值lerp

$$l_{\mathcal{Z}} = \mathbb{E} \left[ \frac{1}{\epsilon^2} d(G(\text{slerp}(\mathbf{z}_1, \mathbf{z}_2; t)) - G(\text{slerp}(\mathbf{z}_1, \mathbf{z}_2; t + \epsilon))) \right] \quad l_{\mathcal{W}} = \mathbb{E} \left[ \frac{1}{\epsilon^2} d(g(\text{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t)) - g(\text{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t + \epsilon))) \right]$$

- 对于随机噪声的StyleGAN，即config E，其w的PPL明显比config B中z的PPL更小
- 测量全路径的PPL，作者认为会略微偏向于z，因为w路径上的插值向量，可能没有对应的z，也就没有被训练过，从而生成效果很差
- 因此，考虑只测量端点的PPL，即设 $t \in [0, 1]$  此时config D中w的PPL较config B更小
- 加入映射网络后，FID和w上的PPL都更优了，并且映射网络的深度越深越好

Method	Path length		Separability
	full	end	
B Traditional generator $\mathcal{Z}$	412.0	415.3	10.78
D Style-based generator $\mathcal{W}$	446.2	376.6	3.61
E + Add noise inputs $\mathcal{W}$	200.5	160.6	3.54
+ Mixing 50% $\mathcal{W}$	231.5	182.1	3.51
F + Mixing 90% $\mathcal{W}$	234.0	195.9	3.79



Method	FID	Path length		Separability
		full	end	
B Traditional 0 $\mathcal{Z}$	5.25	412.0	415.3	10.78
Traditional 8 $\mathcal{Z}$	4.87	896.2	902.0	170.29
Traditional 8 $\mathcal{W}$	4.87	324.5	212.2	6.52
Style-based 0 $\mathcal{Z}$	5.06	283.5	285.5	9.88
Style-based 1 $\mathcal{W}$	4.60	219.9	209.4	6.81
Style-based 2 $\mathcal{W}$	4.43	217.8	199.9	6.25
F Style-based 8 $\mathcal{W}$	4.40	234.0	195.9	3.79

8层更好

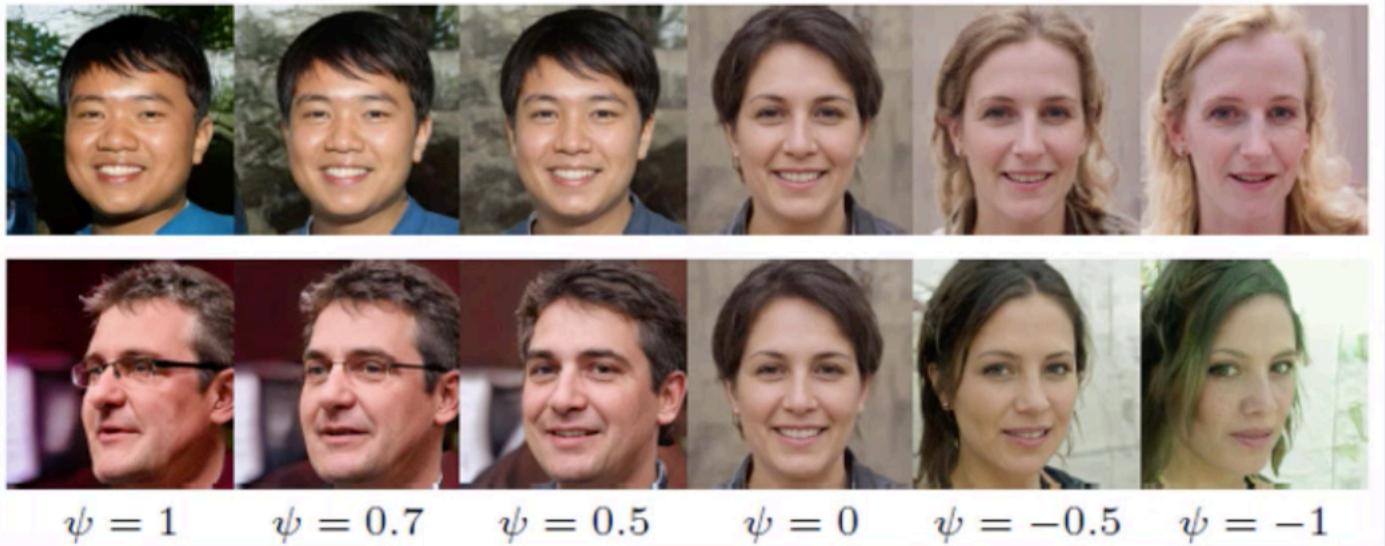
### 线性可分性

- 如果一个隐空间到图像特征之间足够解耦，那么对于一个二分类（男/女）的特征来说，应该能在隐空间中找到一个线性超平面，将隐空间分成对应这个二分类特征的两部分
- 基于CelebA-HQ数据集，使用图像带有**40种属性**进行测试，并训练好美中属性对应分类器
- 使用生成器生成20w个图像，用分类器进行分类，去掉置信度最低的一半，得到10w张隐变量和属性已知的图像
- 对每个属性，拟合一个**线性SVM**来拟合预测这10w张图像的分类，使用 $\exp(\sum_i H(Y_i | X_i))$ 来计算得分，其中X为SVM预测的类别，Y为预训练的分类器预测的类别
- 在所有测试中，w的线性可分性一直优于z，再次说明w与图形特征的纠缠更少

- 增加映射网络的深度，能同时提高图形质量和w的线性可分性，也符合预期
- 在传统生成器前面添加一个映射网络，会导致z的线性可分性发生严重损失，但w的线性可分性和FID都更理想，说明映射网络对于传统生成器同样是work的

## 结论

- 基于Style的生成器，在各方面指标上都超越了传统结构
- 生成图像高级属性和随机效应的分离、以及对隐空间的解耦，对GAN的可解释性和可控性带来了很大帮助
- 感知路径长度很容易作为一个正则化加入到训练中
- 线性可分性的某种变体也许也可以作为一个训练的正则项
- 希望这种直接在训练时塑造中层隐空间的方法，可以在未来产生一些有趣的结果



## 论文总结

### A 关键点

- 结合AdaIN与self-modulation
- 通过中间隐变量w实现映射关系的解耦
- 通过随机噪声进一步提升图像细节和多样性

### B 创新点

- 加入映射网络
- 样式混合
- 对隐空间到图像特征的映射空间线性程度的量化

### C 启发点

- 判别器式模型和生成式模型相辅相成
- 损失函数（如cyclegan）、模型结构（如dcgan）、隐变量分布（如stylegan）、数据集（如stylegan），每一项对于生成式模型都很关键，不可偏废
- 解耦、让映射空间更线性的思想可以广泛应用

纠缠在一起，结果不佳

