生成对抗网络中一种基于样式的生成器结构

Tero Karras
NVIDIA
tkarras@nvidia.com

Samuli Laine
NVIDIA
slaine@nvidia.com

Timo Aila
NVIDIA
taila@nvidia.com

摘 要

我们提出了一种用于生成对抗网络的新生成器体系结构，借鉴了样式转移的相关文献。 新的体系结构导致自动学习地、无监督地分离高级属性（例如，在人脸上训练时的姿势和身份）和生成的图像中的随机变化（例如，雀斑，头发），并且能够直观地、按特定尺度地控制合成。 新的生成器在传统的生成质量指标方面提高了最新技术水平值，可以显著提高插值特性，并且可以更好地解决变异的潜在因素。 为了量化插值质量和分解，我们提出了两种适用于任何生成器架构的新的自动化方法。最后，我们介绍了一个新的，高度多样化和高质量的人脸数据集。

1. 介 绍

由生成方法产生的图像的分辨率和质量 - 尤其是生成对抗性网络（GAN）[21] - 近来已经迅速改善[29,43,5]。然而，生成器继续作为黑匣子运行，尽管最近的努力[3]，仍然缺乏对图像合成过程的各个方面的理解，例如随机特征的组织。潜在空间的性质也很难理解，并且通常展示的潜在空间插值[12,50,35]没有提供将不同生成器相互比较的定量方法。

在样式转移文献[26]的推动下，我们以一种揭示控制图像合成过程的新方法的方式重新设计了生成器结构。我们的生成器从学习常量输入开始，并基于潜码调整每个卷积层图像的“样式”，因此直接控制不同尺度的图像特征的强度。结合直接注入网络的噪声，这种架构变化导致高级属性的自动，无监督解缠（例如，姿势，身份）来自所生成图像中的随机变化（例如，雀斑，头发），并且能够进行直观的尺度特定混合和插值操作。我们不以任何方式修改鉴别器或损失函数，因此我们的工作与正在进行的关于 GAN 损失函数，正则化和超参数的讨论是互为独立的[23,43,5,38,42,34] 。

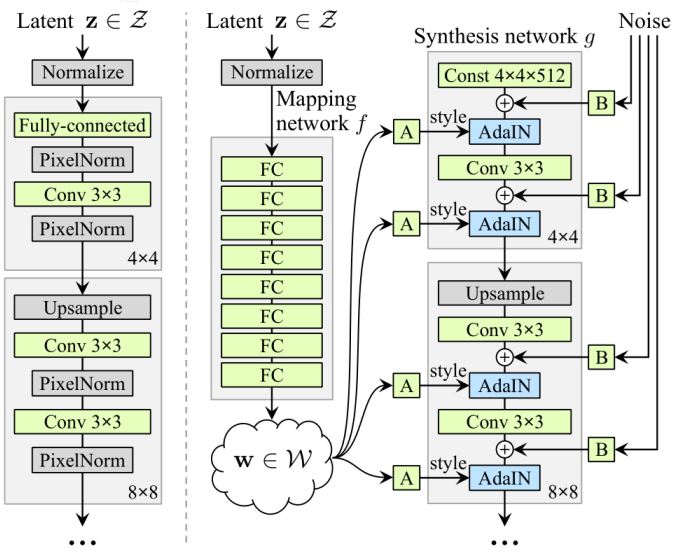
我们的生成器将输入的潜在编码嵌入到中间的潜在空间中，这对于如何在网络中表示噪音因子具有深远的影响。输入潜在空间必须遵循训练数据的概率密度，并且我们认为这会导致某种程度的不可避免的纠缠。我们的中间潜在空间不受限制，因此可以解开。由于以前估算潜在空间解缠的方法并不直接适用于我们的情况，我们提出了两个新的自动度量 - 感知路径长度和线性可分性 - 用于量化生成器的这些方面。使用这些指标，我们表明，与传统的生成器架构相比，我们的生成器允许更多线性，更少纠缠的不同变化因素的表示。

最后，我们提出了一个新的人脸数据集（Flickr-Faces-HQ，FFHQ），它提供了更高的质量，涵盖了比现有高分辨率数据集更广泛的变化（附录 A）。我们已将此数据集与我们的源代码和预先训练的网络一起公开提供，见底部链接 1，随附的视频可在同一链接下找到。

2. 基于样式的生成器

传统上，潜在编码通过输入层，即前馈网络的第一层（图 1a）提供给生成器。 我们通过省略输入层并从学习的常量开始而偏离此设计（图 1b，右）。 给定输入潜在空间 Z 中的潜在代码 z，非线性映射网络 $f : \mathcal{Z} \rightarrow \mathcal{W}$ 首先产生 w ∈ W（图 1b，左）。 为简单起见，我们设置了两者的维度

[1] https://github.com/NVlabs/stylegan



(a) 传统生成器 (b) 基于 Style 的生成器

图 1.虽然传统的生成器[29]仅通过输入层提供潜在编码，但我们首先将输入映射到中间潜在空间 W，然后在每一个卷积层通过自适应实例规范化（AdaIN）控制生成器。 在评估非线性之前，在每次卷积之后添加高斯噪声。 这里“A”代表学习的仿射变换，“B”将学习的每通道缩放因子应用于噪声输入。 映射网络 f 由 8 个层组成，合成网络 g 由 18 层组成 - 每个分辨率两层（分辨率从$4^2 - 1024^2$）。使用单独的 1×1 卷积将最后一层的输出转换为 RGB，类似于 Karras 等人[29]。 我们的生成器共有 2620 万可训练的参数，而传统生成器的参数量为 2310 万。

空间为 512，映射 f 使用 8 层 MLP 实现，我们将在 4.1 节中进行分析。 然后，学习的仿射变换将 w 专门化为样式 $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$，其控制在合成网络 g 的每个卷积层之后的自适应实例归一化（AdaIN）[26,16,20,15]运算。 AdaIN 操作定义为

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}, \qquad (1)$$

其中每个特征映射 xi 分别标准化，然后使用样式 y 中相应的标量组件进行缩放和偏置。 因此，y 的维度是该层上的特征映射的数量的两倍。

比较我们的样式转换方法，我们从矢量 w 而不是示例图像计算空间不变的样式 y。 我们选择为 y 使用“style”这个词，因为类似的网络架构已经用于前馈式传输[26]，无监督的图像到图像转换[27]和域混合[22]。 与更一般的特征变换相比[36,55]，由于其效率和紧凑的表现，AdaIN 特别适合我们的目的。

Method	CelebA-HQ	FFHQ
A Baseline Progressive GAN [29]	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	**5.06**	4.42
F + Mixing regularization	5.17	**4.40**

表 1.各种生成器设计的 Frechet'起始距离（FID）（越低越好）。在本文中，我们使用从训练集中随机抽取的 50,000 张图像来计算 FID，并报告在训练过程中遇到的最低距离。

最后，我们通过引入显式噪声输入为我们的生成器提供直接的方法来生成随机细节。 这些是由不相关的高斯噪声组成的单通道图像，并且我们将专用噪声图像馈送到合成网络的每一层。

使用学习的每个特征缩放因子将噪声图像广播到所有特征图，然后将其添加到相应卷积的输出中，如图 1b 所示。添加噪声输入的含义在第 3.2 和 3.3 节中讨论。

2.1. 生成图像的质量

在研究我们的生成器的特性之前，我们通过实验证明重新设计不能保证图像质量，但事实上，它可以大大改善它。表 1 给出了 CELEBA-HQ [29]和我们新的 FFHQ 数据集（附录 A）中各种生成器架构的 Frechet'起始距离（FID）[24]。其他数据集的结果在附录 E 中给出。我们的基线配置（A）是 Karras 等人的 Progressive GAN 设置 [29]，我们默认继承了网络和所有超参数除非另有说明。我们首先通过使用双线性上/下采样操作[62]，更长的训练和调整的超参数来切换到改进的基线（B）。附录 C 中包含了对训练设置和超参数的详细描述。然后我们通过添加映射网络和 AdaIN 操作（C）进一步改进了这个新的基线，并且惊讶地观察到网络不再从将潜在编码馈送到第一个卷积层以获益。因此，我们通过移除传统的输入层并从学习的 4 × 4 × 512 恒定张量（D）开始图像合成来简化架构。我们发现，即使只通过控制 AdaIN 操作的样式接收输入，合成网络也能够产生有意义的结果。

最后，我们介绍了进一步改善结果的噪声输入（E），以及对相邻样式进行去相关的新型混合正则化（F），并对生成的图像进行更细粒度的控制（第 3.1 节）。

我们使用两种不同的损失函数来评估我们的方法：对于 CELEBA-HQ，我们依赖于 WGAN-GP [23]，

图 2.由基于样式的生成器（config F）和 FFHQ 数据集生成的未经验证的图像集。 在这里，我们使用截断技巧[40,5,32]的变化，对于分辨率$4^2 - 32^2$，使用ψ=0.7。请参阅随附的视频以获得更多结果。

而 FFHQ 使用 WGAN-GP 进行配置 A 和非饱和损耗[21]，其中 R1 正则化[42,49,13]用于配置 B-F。我们发现这些选择可以产生最佳效果。我们的贡献不需要修改损失函数。

我们观察到基于样式的生成器（E）相比传统生成器（B）显著提高了 FID，几乎达到了 20%，证实了并行工作中的大规模 ImageNet 测量[6,5]。图 2 显示了使用我们的生成器从 FFHQ 数据集生成的一组未经验证的新图像。正如 FID 所证实的那样，平均质量很高，甚至眼镜和帽子等配件也能成功合成。对于这个图，我们使用所谓的截断技巧避免了从 W 的极端区域采样[40,5,32] - 附录 B 详述了如何在 W 而不是 Z 中执行该技巧。注意我们的生成器允许应用仅选择性地截断为低分辨率，以便不影响高分辨率细节。

本文中的所有 FID 都是在没有截断技巧的情况下计算出来的，我们在图 2 和视频中仅将其用于说明目的。所有图像均以1024^2分辨率生成。

2.2. 现有技术

关于 GAN 架构的大部分工作都集中在通过例如使用多个鉴别器[17,15]，多分辨率判别[58,53]或自注意力机制[61]来改进鉴别器。 生成器侧的工作主要集中在输入潜在空间的精确分布[5]或通过高斯混合模型[4]，聚类[46]或鼓励凸性[50]对输入潜在空间进行整形。

最近的条件生成器通过单独的嵌入网络将类标识符提供给生成器[44]中的大量层，而潜在的仍然通过输入层提供。一些作者已经考虑将潜在编码的一部分馈送到多个生成器层[9,5]。 在平行工作中，陈等人[6]使用 AdaIN“自我调制”生成器，与我们的工作类似，但不考虑中间潜在空间或噪声输入。

3. 基于样式的生成器的属性

我们的生成器架构可以通过对样式的特定尺度修改来控制图像合成。我们可以将映射网络和仿射变换视为从学习的分布中为每种样式绘制样本的方式，并且将合成网络视为基于样式集合生成新颖图像的方式。每种样式的效果都在网络中定位，即：可以预期修改样式的特定子集仅影响图像的某些方面。

为了了解这种定位的原因，让我们考虑 AdaIN 操作（方程 1）如何首先将每个通道归一化为零均值和单位方差，然后才应用基于样式的尺度和偏差。根据样式的要求，新的每通道统计数据修改了后续卷积操作的特征的相对重要性，但由于规范化，它们不依赖于原始统计数据。因此，每个样式在被下一个 AdaIN 操作覆盖之前仅控制一个循环。

3.1. 样式混合

为了进一步鼓励样式进行本地化，我们采用混合正则化，其中使用两个随机潜码而不是训练期间的一个潜在编码生成给定百分比的图像。 当生成这样的图像时，我们简单地从一个潜在编码切换到另一个 - 我们称之为样式混合的操作 - 在合成网络中随机选择的点。 具体来说，我们运行两个潜码 z1， z2 通过映射网络，并具有相应的 w1，w2 控制样式，以便 w1 在交叉点之前应用，w2 在交叉点之后应用。 这种正则化技术可以防止网络假设相邻的样式是相关的。

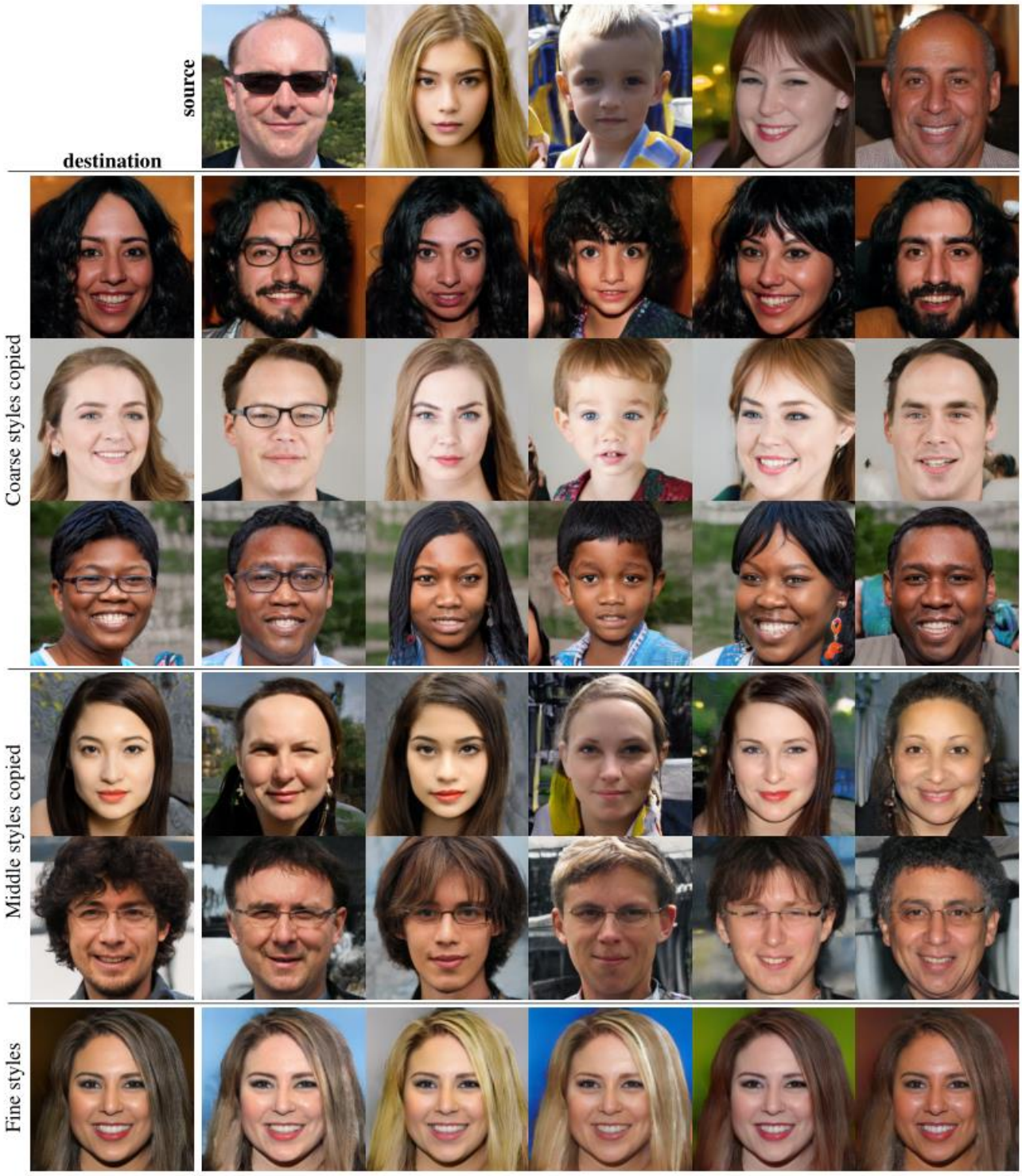


图 3.通过使一个潜在编码（源）生成的样式覆盖另一个（目标）的样式的子集，可视化生成器中样式的效果。 覆盖与粗糙空间分辨率相对应的图层样式（$4^2 - 8^2$），从源头复制高级属性，如姿势，通用发型，面部形状和眼镜，同时所有颜色（眼睛，头发，灯光）和保留目标的更精细的面部特征。 如果我们改为复制中间层（$16^2 - 32^2$）的样式，我们会继承较小比例的面部特征，发型，从源头打开/关闭的眼睛，同时保留来自目标的姿势，一般面部形状和眼镜。 最后，复制与精细分辨率（$64^2 - 1024^2$）相对应的样式主要带来来自源的颜色方案和微结构。

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

表 2.通过对不同百分比的训练样例进行混合正则化训练的网络的 FFHQ 中的 FID。 在这里，我们通过随机化 1...4 潜码和它们之间的交叉点来对受过训练的网络进行压力测试。 混合正则化可显著提高对这些不良操作的耐受性。 标签 E 和 F 参见表 1 中的配置。

(a) 生成图片 (b) 随机变化 (c) 标准偏差

图 4.随机变化的例子。 （a）两个生成的图像。 （b）放大输入噪声的不同实现。 虽然整体外观几乎相同，但个别毛发的放置方式却截然不同。 （c）100 个不同实现中每个像素的标准偏差，突出显示图像的哪些部分受到噪声的影响。 主要区域是头发，轮廓和背景的一部分，但眼睛反射也有有趣的随机变化。 身份和姿势等全局方面不受随机变化的影响。

表 2 显示了在训练期间启用混合正则化如何显著改善定位，通过在测试时混合多个潜码的情况下改进的 FID 来表示。图 3 给出了通过混合不同尺度的两个潜码合成的图像的例子。我们可以看到每个样式子集控制图像的有意义的高级属性。

3.2. 随机变化

人类肖像中有许多方面可以被视为随机的，例如毛发，茬，雀斑或皮肤毛孔的确切位置。只要它们遵循正确的分布，任何这些都可以在不影响我们对图像的感知的情况下进行随机化。

让我们考虑一下传统生成器如何实现随机变化。鉴于网络的唯一输入是通过输入层，网络需要发明一种方法来生成空间变化的伪随机数

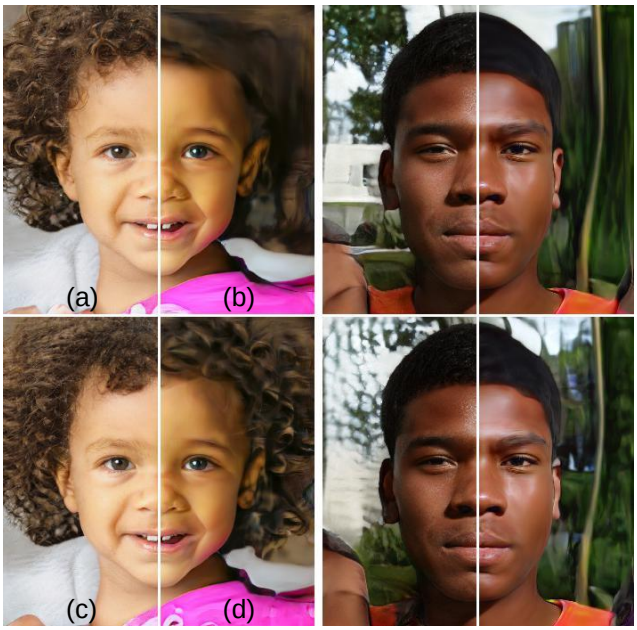
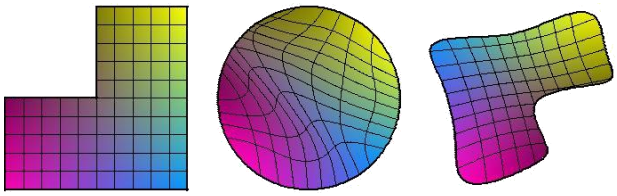


图 5.在我们的生成器的不同层的噪声输入的影响。（a）噪音适用于所有层。（b）没有噪音。（c）仅有精细层的噪音（$64^2 - 1024^2$）。（d）仅粗糙层的噪音（$4^2 - 32^2$）。 我们可以看到，人为地忽略噪音会导致无意义的“绘画”外观。 粗糙的噪音会导致头发大规模卷曲和出现更大的背景特征，而细微的噪音则会带来更精细的头发卷曲，更细致的背景细节和皮肤毛孔。

从需要时的早期激活开始。这消耗了网络容量并且隐藏了生成信号的周期性是困难的 - 并且并不总是成功的，正如生成的图像中常见的重复模式所证明的那样。我们的架构通过在每次卷积后添加每像素噪声来回避这些问题。

图 4 显示了使用具有不同噪声实现的生成器产生的相同非图像的随机实现。我们可以看到，噪声仅影响随机方面，使整体构成和身份等高级属性完整无缺。图 5 进一步说明了将随机变化应用于不同的子集设置的效果。由于这些效果最好在动画中看到，请参阅随附的视频，以了解如何更改一层的噪声输入导致匹配比例的随机变化。

我们发现有趣的是，噪声的影响似乎在网络中紧密地定位。我们假设在生成器的任何一点，都有压力尽快引入新的内容，而我们的网络创建随机变化的最简单方法是依靠提供的噪声。每一层都有一组新的噪声，因此没有动力从早期的激活中产生随机效应，从而产生局部效应。



(a) 训练集中的特征分布 (b) 从 Z 到特征的映射 (c) 从 W 到特征的映射

图 6.具有两个变异因素的示例性示例（图像特征，例如，男性气质和头发长度）。（a）一个示例训练集，其中一些组合（例如，长发男性）几乎不存在。（b）这会强制从 Z 到图像特征的映射变为弯曲，以便禁止组合在 Z 中消失，以防止对无效组合进行采样。（c）然而，从 Z 到 W 的学习映射能够“消除”大部分翘曲。

3.3. 从随机性中分离全局效应

前面的部分以及随附的视频表明，虽然样式的变化具有全局效果（改变姿势，身份等），但噪声仅影响无关紧要的随机变化（不同梳理的头发，胡须等）。这种观察符合样式转换文献，其中已经确定空间不变统计量（格式矩阵，通道平均值，变量等）可靠地编码图像的样式[19,37]。空间变化的特征编码特定的实例。

在我们基于样式的生成器中，样式会影响整个图像，因为完整的要素图将使用相同的值进行缩放和双向化。因此，可以同时控制姿势，灯光或背景样式等全局效果。同时，噪声被独立地添加到每个像素，因此理想地适合于控制随机变化。如果网络试图控制例如使用噪声的姿势，那将导致空间不一致的决定，然后由鉴别器惩罚。因此，网络学会在没有明确指导的情况下，适当地使用全局和局部渠道。

4. 解缠研究

解缠有各种定义[52,48,2,7,18]，但共同的目标是由线性子空间构成的潜在空间，每个空间控制一个变异因子。 然而，Z 中每个因子组合的采样概率需要与训练数据中的对应密度相匹配。 如图 6 所示，这排除了与典型数据集和输入潜在分布完全解开的因素

我们的生成器架构的一个主要好处是中间潜在空间 W 不必根据任何固定分布支持采样;它的采样密度是由学习的分段连续映射 f（z）引起的。该映射可以适于“解除”W，以使变化因子变得更加线性。我们认为生成器有这样的压力，因为基于解开的表示而不是基于纠缠的表示来生成逼真的图像应该更容易。因此，我们期望训练在无人监督的环境中，即当事先不知道变异因素时产生较少陷入的 W [10,33,47,8,25,30,7]。

不幸的是，最近提出的用于量化解缠的指标[25,30,7,18]需要编码器网络将输入图像映射到潜码。由于我们的基线 GAN 缺少这样的编码器，因此这些方法不适合我们的目的。虽然可以为此目的添加额外的网络[8,11,14]，但我们希望避免将工作投入到不属于实际解决方案的组件中。为此，我们描述了量化非解缠结的两种新方法，这两种方法都不需要编码器或已知的变化因子，因此可以计算任何图像数据集和生成器。

4.1. 感知路径长度

正如 Laine [35]所指出的，潜在空间矢量的插值可能会在图像中产生令人惊讶的非线性变化。例如，任一端点中不存在的特征可能出现在线性插值路径的中间。这表明潜在的空间被纠缠，并且变化的因素没有被恰当地分开。为了量化这种效果，我们可以测量在潜在空间中执行插值时图像的剧烈变化。直观地说，较小弯曲的潜在空间应该比高度弯曲的潜在空间更容易过渡。

作为我们度量的基础，我们使用基于感知的成对图像距离[63]，其被计算为两个 VGG16 [56]嵌入之间的加权差异，其中权重是适合的，使得度量与人类的视觉相似性判断一致。如果我们将潜在空间插值路径细分为线性段，我们可以将该分段路径的总感知长度定义为每个段上的感知差异的总和，如图像距离度量所报告的。在无限精细细分下，这个总和的自然定义将是这个和的极限，但在实践中我们使用一个小的细分小量 $\epsilon = 10^{-4}$来近似它。潜在空间 Z 中的平均感知路径长度，超过所有可能的终点，因此

$$l_{\mathcal{Z}} = \mathbb{E}\left[\frac{1}{\epsilon^2} d\big(G(\mathrm{slerp}(\mathbf{z}_1, \mathbf{z}_2;\, t)),\; G(\mathrm{slerp}(\mathbf{z}_1, \mathbf{z}_2;\, t+\epsilon))\big)\right], \tag{2}$$

其中 $\mathbf{z}_1, \mathbf{z}_2 \sim P(\mathbf{z}), t \sim U(0,1)$，G 是生成器（即

[2] 设计用于解缠研究的少数人工数据集（例如，[41,18]）将预定变异因子的所有组合制成具有均匀频率，从而隐藏了该问题。

	Method	Path length full	Path length end	Separability
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

表 3. FFHQ 中各种生成器架构的感知路径长度和可分性得分（越低越好）。 我们在 Z 中为传统网络执行测量，在 W 中执行基于样式的测量。 使网络抵抗样式混合似乎有点扭曲中间潜在空间 W。我们假设混合使得 W 更难以有效地编码跨越多个尺度的变化因子。

	Method	FID	Path length full	Path length end	Separability
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**

表 4. FFHQ 中映射网络的影响。 方法名称中的数字表示映射网络的深度。 我们看到 FID，可分离性和路径长度都可以从映射网络中受益，这适用于基于样式和传统的生成器体系结构。 此外，更深的映射网络通常比浅层映射网络表现更好。

$g \circ f$ 用于基于样式的网络），d（.,.）评估的是结果图像之间的每个视距。 这里 slerp 表示球面插值运算[54]，它是我们归一化输入潜在空间中最合适的插值方式[59]。 为了专注于面部特征而不是背景，我们在评估成对图像度量之前裁剪生成的图像以仅包含面部。 由于度量 d 恰好是自然二次[63]，我们除以 ϵ^2 而不是 ϵ 为了抵消对细分粒度的不必要依赖。 我们通过获取 100,000 个样本来计算期望值。

计算 W 中的平均感知路径长度以类似的方式执行：

$$l_{\mathcal{W}} = \mathbb{E}\left[\frac{1}{\epsilon^2} d\big(g(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t)), g(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t+\epsilon))\big)\right], \tag{3}$$

唯一的区别是插值发生在 W 空间中。因为 W 中的向量没有以任何方式标准化，我们使用线性插值（lerp）。

表 3 显示，对于具有噪声输入的基于样式的生成器，该全路径长度明显更短，这表明 W 在感知上比 Z 更线性。然而，这种测量实际上略微偏向于输入潜在空间 Z。如果 W 确实是 Z 的解缠和“平坦”映射，它可能包含不在输入流形上的区域 - 因此由生成器较差重建 - 甚至在从输入流形映射的点之间，而输入潜在空间 Z 根据定义没有这样的区域。因此，如果我们将度量限制在路径端点，即 $t \in \{0, 1\}$，则结果应该是可期望的，我们应该获得更小的 $l_{\mathcal{W}}$ 而 $l_{\mathcal{Z}}$ 不受影响。这确实是我们在表 3 中观察到的。

表 4 显示了映射网络如何影响路径长度。我们看到传统和基于样式的生成器都受益于映射网络，并且附加深度通常会改善感知路径长度以及 FID。有趣的是，虽然 $l_{\mathcal{W}}$ 有所改善

在传统的生成器里，$l_{\mathcal{Z}}$ 却变得更加糟糕，说明我们的主张输入潜在空间确实可以任意纠缠在 GAN 中。

4.2. 线性可分性

如果潜在空间被充分解开，则应该可以找到始终与各个变化因素相对应的方向向量。我们提出了另一个度量，通过测量潜在空间点通过线性超平面分成两个不同的集合的程度来量化这种效应，以便每个集合对应于图像的特定二进制属性。

为了标记所生成的图像，我们为许多二进制属性训练辅助分类网络，例如，以区分男性和女性面部。在我们的测试中，分类器具有与我们使用的判别器相同的架构（即，与[29]中相同），并且使用 CelebA-HQ 数据集进行训练，该数据集保留原始 CelebA 数据集中可用的 40 个属性。为了测量一个属性的可分离性，我们用 z ~ P(z) 生成 200,000 个图像，并使用辅助分类网络对它们进行分类。然后，我们根据分类器置信度对样本进行排序，并移除最不自信的一半，产生 100,000 个标记的潜在空间向量。

对于每个属性，我们拟合线性 SVM 来预测基于潜在空间点的标签 - 传统的 z 和基于样式的 w - 并且通过该平面对点进行分类。为了测量超平面能够将点分成正确的组的程度，我们计算条件对数 H（Y | X），其中 X 是由 SVM 预测的类，Y 是由预训练的分类器确定的类。因此，条件熵告诉我们确定样本的真实类需要多少附加信息，假设我们知道它所在的超平面的哪一侧。直觉是，如果相关的变异因子或其组合具有不一致的拉伸空间方向，则通过超平面分离样本点将更加困难，从而产生高条件

图 7. FFHQ 数据集在年龄，种族，观点，光照和图像背景方面提供了很多种类。

熵。低值表示易于分离，因此对于特定因素或属性对应的一组因子，更一致的潜在空间方向。

我们将最终的可分性得分计算为 $\exp(\sum_i \mathrm{H}(Y_i|X_i))$，其中我列举了 40 个属性。与初始得分[51]类似，取幂使得从对数到线性域的值更容易比较。

表 3 和表 4 显示 W 一直比 Z 更好，这表明纠缠的表示更少。此外，增加映射网络的深度改善了 W 中的图像质量和可分离性，这符合合成网络当然有利于解缠的输入表示的假设。有趣的是，在传统生成器前面添加一个映射网络会导致 Z 中可分离性的严重损失，但是会证明中间潜在空间 W 中的情况，并且 FID 也会改善。这表明，当我们引入一个不必跟随训练数据分布的中间潜在空间时，即使是传统的生成器架构也能表现得更好。

5.结论

基于我们的结果和 Chen 等人的并行工作[6]，越来越清楚的是，传统的 GAN 生成器架构在各方面都不如基于样式的设计。在已建立的质量指标方面也是如此，我们进一步认为，我们对高级属性和随机效应分离的研究以及中间潜在空间的线性将在提高 GAN 理论的理解和可控性方面取得丰硕成果。

我们注意到，我们的平均路径长度度量可以很容易地用作训练期间的正则化器，也许线性可分性度量的一些变体也可以作为一个变量。总的来说，我们希望在训练期间直接塑造中间潜码空间的方法将为未来的工作提供有趣的途径。

6.致谢

我们感谢 Jaakko Lehtinen，David Luebke 和 TuomasKynka¨anniemi 的深入讨论和有益的评论；Janne Hellsten，Tero Kuosmanen 和 PekkaJanis，负责计算基础架构和代码发布方面的帮助。

A. FFHQ 数据集

我们收集了一个新的人脸数据集，Flickr-Faces-HQ（FFHQ），由1024^2分辨率的 70,000 个高质量图像组成（图 7）。该数据集在年龄，种族和图像背景方面包含了比 CelebA-HQ [29]更多的变化，并且还有更多的配件覆盖范围，如眼镜，太阳镜，帽子等。图像是从 Flickr 爬行的（从而加入了该网站的所有偏见）并自动对齐和裁剪。只收集了许可证下的图像。各种自动过滤器用于修剪套装，最后 Mechanical Turk 允许我们删除偶然的雕像，绘画或照片。我们已经公布了数据集：

https://github.com/NVlabs/ffhq-dataset

B. W 中的截断技巧

如果我们考虑训练数据的分布，很明显低密度区域的代表性很差，因此生成器很难学习。这是所有生成建模技术中一个重要的开放性问题。然而，众所周知，从截断的[40,5]或以其他方式缩小的[32]采样空间中绘制潜在向量倾向于改善平均图像质量，尽管丢失了一些量的变化。

我们可以遵循类似的策略。首先，我们计算 W 的质心为 $\bar{\mathbf{w}} = \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f(\mathbf{z})]$。在 FFHQ 的情况下，这一点代表一种平均面（图 8，ψ = 0）。然后我们可以将给定 w 与中心的偏差缩放为 $\mathbf{w}' = \bar{\mathbf{w}} + \psi(\mathbf{w} - \bar{\mathbf{w}})$，其中ψ < 1。

而 Brock 等人[5]观察到，即使使用正交正则化，只有一部分网络能够适应这种截断，即使不改变损失函数，W 空间中的截断似乎也能可靠地工作。

C.超参数和训练细节

我们建立在 Karras 等人的正式 TensorFlow [1]实施渐进式 GAN 的基础上[29]，我们继承了大部分的训练细节（见页底 3）这个原始设置对应于表 1 中的配置 A。特别是，我们使用相同的鉴别器结构，分辨率相关的小批量大小，Adam [31]超参数，

[3] https://github.com/tkarras/progressive growing of gans

图 8.截断技巧作为样式比例函数ψ的影响。 当我们褪色ψ → 0，所有面孔都汇聚到 FFHQ 的“均值”面。 对于所有受过训练的网络来说，这个面部是相似的，并且对它的插入似乎永远不会导致伪影。 通过对样式应用负缩放，我们得到相应的相反或“反面”。 有趣的是，在对立面之间翻转的各种高级属性，包括视点，眼镜，年龄，着色，头发长度和性别。

和生成器的指数移动平均值。我们为 CelebA-HQ 和 FFHQ 启用了镜像扩充，但是为 LSUN 禁用了它。使用 8 个 Tesla V100 GPU 的 NVIDIA DGX-1 训练时间约为一周。

对于我们改进的基线（表 1 中的 B），我们进行了几项修改以提高整体结果质量。我们用双线性采样替换两个网络中的最近邻/上/下采样，我们通过在每个上采样层之后和每个下采样层之前用可分离的二阶二次滤波器对激活进行低通滤波来实现[62]。我们以与 Karras 等人相同的方式实施渐进式增长[29]，但我们从8^2的图像而不是4^2的图像开始。对于 FFHQ 数据集，我们使用γ = 10 从 R1 正则化[42]切换到 WGAN-GP 到非饱和损失[21]。我们发现了 R1 与 WGAN-GP 相比，FID 分数持续下降的时间要长得多，因此我们将训练时间从 12M 增加到 25M。我们使用与 Karras 等人相同的学习率 [29]对于 FFHQ，但我们发现将学习率设置为 0.002 而不是 0.003（对于512^2和1024^2）可以使 CelebA-HQ 获得更好的稳定性。

对于我们基于样式的生成器（表 1 中的 F），我们使用 Leaky ReLU [39]和α = 0.2 以及所有层的均衡学习率[29]。我们在卷积层中使用与 Karras 等人相同的特征映射计数[29]。我们的映射网络由 8 个完全连接层组成，所有输入和输出激活的维度 - 包括 z 和 w - 是 512。我们发现在高学习率下增加映射网络的深度往往会使训练不稳定。因此，我们将映射网络的学习速率降低了两个数量级，即$\lambda' = 0.01 \cdot \lambda$。我们使用 N（0，1）初始化卷积，完全连接和仿射变换层的所有权重。偏差和噪声缩放因子初始化为零，除了与我们初始化为 1 的 y 相关的偏差。

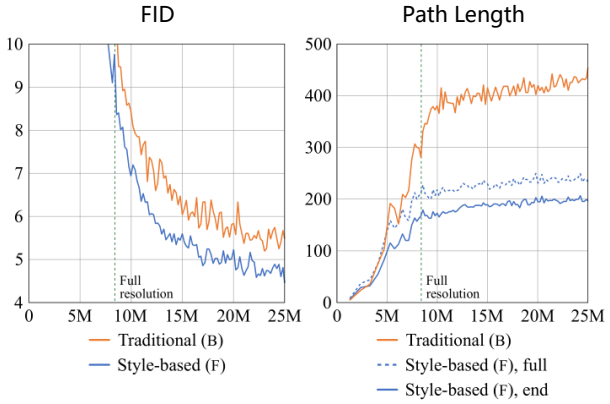


图 9.使用 FFHQ 数据集在我们的配置 B 和 F 中训练过程中的 FID 和感知路径长度指标。横轴表示鉴别器所看到的训练图像的数量。 8.4M 张图像的虚线垂直线标志着训练进展到完全1024^2分辨率时的点。 在右边，我们只显示传统生成器路径长度测量的一条曲线，因为 Z 中的全路径和端点采样之间没有明显的差异（4.1 节）。

我们的可分性度量（第 4.2 节）使用的分类器与我们的鉴别器具有相同的体系结构，即禁用了小批量标准偏差[29]。我们使用10^{-3}的学习率，8 的小批量大小，Adam 优化器，以及 150,000 张图像的训练长度。 分类器独立于生成器进行训练，并且相同的 40 个分类器（每个 CelebA 属性一个）用于测量所有生成器的可分性度量。 我们将发布预训练的分类器网络，以便我们的测量可以再现。

我们的网络中不使用批量归一化[28]，谱正则化[43]，注意力机制[61]，Dropout[57]或像素特征向量归一化[29]。

D. 训练融合

图 9 显示了在使用 FFHQ 数据集训练配置 B 和 F 期间 FID 和感知路径长度度量如何演变。 R1 正则化在两种配置中都有效，随着训练的进行，FID 继续缓慢下降，激励我们选择将训练时间从 12M 图像增加到 25M 图像。 即使训练达到了1024^2分辨率，缓慢上升的路径长度也表明 FID 的改进是以更纠缠的表示为代价的。 考虑到未来的工作，这是一个有趣的问题，这是否是不可避免的，或者是否有可能在不影响 FID 收敛的情况下鼓励更短的路径长度。

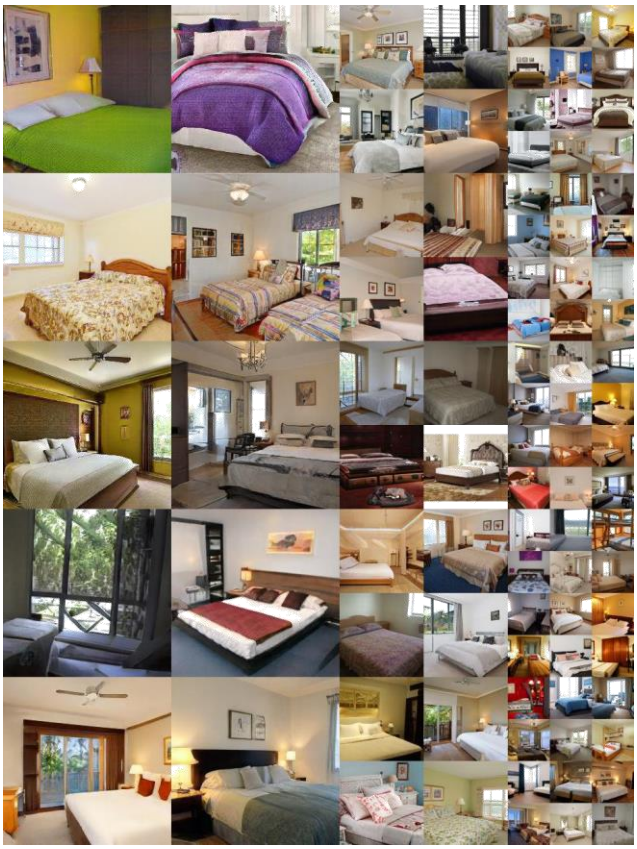


图 10.由基于样式的生成器（config F）生成的未经计算的图像集，其中 LSUN BEDROOM 数据集位于256^2。

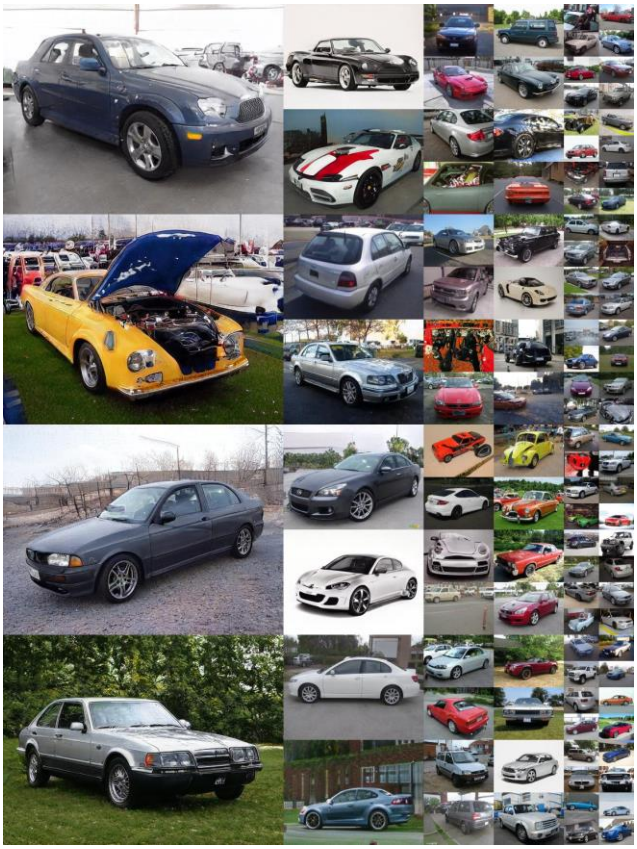


图 11.由基于样式的生成器（config F）生成的未经验证的图像集，其中 LSUN CAR 数据集为 512 × 384。

E. 其他数据集

图 10,11 和 12 分别显示了 LSUN [60] BEDROOM，CARS 和 CATS 的一组未经验证的结果。在这些图像中，我们使用附录 B 中的截断技巧，对于分辨率$4^2 - 32^2$，使用ψ = 0.7。附带的视频提供了样式混合和随机变化测试的结果。从中可以看出，在 BEDROOM 的情况下，粗糙的样式基本上控制了相机的视点，中间样式选择特定的家具，而精细的样式处理颜色和材料的较小细节。在 CARS 中，效果大致相似。随机变化主要影响卧室中的织物，CARS 中的背景和前照灯，以及毛皮，背景，以及有趣的是，爪子在 CATS 中的定位。有点令人惊讶的是，汽车的车轮似乎永远不会根据随机输入进行旋转。

这些数据集使用与 FFHQ 相同的设置进行训练，持续时间为卧室和 CATS 的 70M 图像，以及 CARS 的 46M 图像。我们怀疑 BEDROOM 的结果开始接近训练数据的极限，因为在许多图像中，最令人反感的问题是从低质量训练数据继承的严重压缩伪像。 CARS 有高得多的质量训练数据也允许更高的空间分辨率（512 × 384 而不是256^2），并且由于姿势，缩放级别和背景的高内在变化，CATS 仍然是一个困难的数据集。

参考文献

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: a system for large-scale machine learning. In Proc. 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16, pages 265– 283, 2016. 8

[2] A. Achille and S. Soatto. On the emergence of invariance and disentangling in deep representations. CoRR, abs/1706.01350, 2017. 6

[3] Anonymous. Visualizing and understanding generative adversarial networks. Submitted to ICLR 2019, https://openreview.net/forum?id=Hyg X2C5FX, 2018. 1

[4] M. Ben-Yosef and D. Weinshall. Gaussian mixture genera-tive adversarial networks for diverse datasets, and the unsu-pervised clustering of images. CoRR, abs/1808.10356, 2018. 3

图 12.由基于样式的生成器（config F）生成的未经计算的图像集，其中 LSUN CAT 数据集位于 2562。

[5] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. CoRR, abs/1809.11096, 2018. 1, 3, 8

[6] T. Chen, M. Lucic, N. Houlsby, and S. Gelly. On self modulation for generative adversarial networks. CoRR, abs/1810.01365, 2018. 3, 8

[7] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolat-ing sources of disentanglement in variational autoencoders. CoRR, abs/1802.04942, 2018. 6

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: interpretable representation learn-ing by information maximizing generative adversarial nets. CoRR, abs/1606.03657, 2016. 6

[9] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a Laplacian pyramid of ad-versarial networks. CoRR, abs/1506.05751, 2015. 3

[10] G. Desjardins, A. Courville, and Y. Bengio. Disentan-gling factors of variation via generative entangling. CoRR, abs/1210.5474, 2012. 6

[11] J. Donahue, P. Krahenb¨uhl,¨ and T. Darrell. Adversarial fea-ture learning. CoRR, abs/1605.09782, 2016. 6

[12] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. CoRR, abs/1411.5928, 2014. 1

[13] H. Drucker and Y. L. Cun. Improving generalization perfor-mance using double backpropagation. IEEE Transactions on Neural Networks, 3(6):991–997, 1992. 3

[14] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned in-ference. In Proc. ICLR, 2017. 6

[15] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio. Feature-wise transforma-tions. Distill, 2018. https://distill.pub/2018/feature-wise-transformations. 2

[16] V. Dumoulin, J. Shlens, and M. Kudlur. A learned represen-tation for artistic style. CoRR, abs/1610.07629, 2016. 2

[17] I. P. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. CoRR, abs/1611.01673, 2016. 3

[18] C. Eastwood and C. K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In Proc. ICLR, 2018. 6

[19] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In Proc. CVPR, 2016. 6

[20] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. CoRR, abs/1705.06830, 2017. 2

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Gen-erative Adversarial Networks. In NIPS, 2014. 1, 3, 9

[22] W.-S. Z. Guang-Yuan Hao, Hong-Xing Yu. MIXGAN: learn-ing concepts from different domains for mixture generation. CoRR, abs/1807.01659, 2018. 2

[23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. CoRR, abs/1704.00028, 2017. 1, 2

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In NIPS, pages 6626– 6637. 2017. 2

[25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In Proc. ICLR, 2017. 6

[26] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. CoRR, abs/1703.06868, 2017. 1, 2

[27] X. Huang, M. Liu, S. J. Belongie, and J. Kautz. Mul-timodal unsupervised image-to-image translation. CoRR, abs/1804.04732, 2018. 2

[28] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, abs/1502.03167, 2015. 9

[29] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and varia-tion. CoRR, abs/1710.10196, 2017. 1, 2, 7, 8, 9

[30] H. Kim and A. Mnih. Disentangling by factorising. In Proc. ICML, 2018. 6

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 8

[32] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. CoRR, abs/1807.03039, 2018. 3, 8

[33] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In ICLR, 2014. 6

[34] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly. The gan landscape: Losses, architectures, regularization, and normalization. CoRR, abs/1807.04720, 2018. 1

[35] S. Laine. Feature-based metrics for exploring the latent space of generative models. ICLR workshop poster, 2018. 1, 6

[36] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In Proc. NIPS, 2017. 2

[37] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. CoRR, abs/1701.01036, 2017. 6

[38] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bous-quet. Are GANs created equal? a large-scale study. CoRR, abs/1711.10337, 2017. 1

[39] A. L. Maas, A. Y. Hannun, and A. Ng. Rectifier nonlin-earities improve neural network acoustic models. In Proc. International Conference on Machine Learning (ICML), vol-ume 30, 2013. 9

[40] M. Marchesi. Megapixel size image creation using genera-tive adversarial networks. CoRR, abs/1706.00082, 2017. 3, 8

[41] L. Matthey, I. Higgins, D. Hassabis, and A. Lerch-ner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 6

[42] L. Mescheder, A. Geiger, and S. Nowozin. Which train-ing methods for GANs do actually converge? CoRR, abs/1801.04406, 2018. 1, 3, 9

[43] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. CoRR, abs/1802.05957, 2018. 1, 9

[44] T. Miyato and M. Koyama. cGANs with projection discrim-inator. CoRR, abs/1802.05637, 2018. 3

[45] G. Mordido, H. Yang, and C. Meinel. Dropout-gan: Learn-ing from a dynamic ensemble of discriminators. CoRR, abs/1807.11346, 2018. 3

[46] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan. Cluster-GAN : Latent space clustering in generative adversarial net-works. CoRR, abs/1809.03627, 2018. 3

[47] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep genera-tive models. In Proc. ICML, 2014. 6

[48] K. Ridgeway. A survey of inductive biases for factorial representation-learning. CoRR, abs/1612.05299, 2016. 6

[49] A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. CoRR, abs/1711.09404, 2017. 3

[50] T. Sainburg, M. Thielk, B. Theilman, B. Migliori, and T. Gentner. Generative adversarial interpolative autoencod-ing: adversarial training on latent space interpolations en-courage convex latent distributions. CoRR, abs/1807.06650, 2018. 1, 3

[51] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In NIPS, 2016. 8

[52] J. Schmidhuber. Learning factorial codes by predictability minimization. Neural Computation, 4(6):863–879, 1992. 6

[53] R. Sharma, S. Barratt, S. Ermon, and V. Pande. Improved training with curriculum gans. CoRR, abs/1807.09295, 2018. 3

[54] K. Shoemake. Animating rotation with quaternion curves. In Proc. SIGGRAPH '85, 1985. 7

[55] A. Siarohin, E. Sangineto, and N. Sebe. Whitening and col-oring transform for GANs. CoRR, abs/1806.00420, 2018. 2

[56] K. Simonyan and A. Zisserman. Very deep convolu-tional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 6

[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neu-ral networks from overfitting. Journal of Machine Learning Research, 15:1929–1958, 2014. 9

[58] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. CoRR, abs/1711.11585, 2017. 3

[59] T. White. Sampling generative networks: Notes on a few effective techniques. CoRR, abs/1609.04468, 2016. 7

[60] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Con-struction of a large-scale image dataset using deep learning with humans in the loop. CoRR, abs/1506.03365, 2015. 10

[61] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. CoRR, abs/1805.08318, 2018. 3, 9

[62] R. Zhang. Making convolutional networks shift-invariant again, 2019. 2, 9

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a percep-tual metric. In Proc. CVPR, 2018. 6, 7