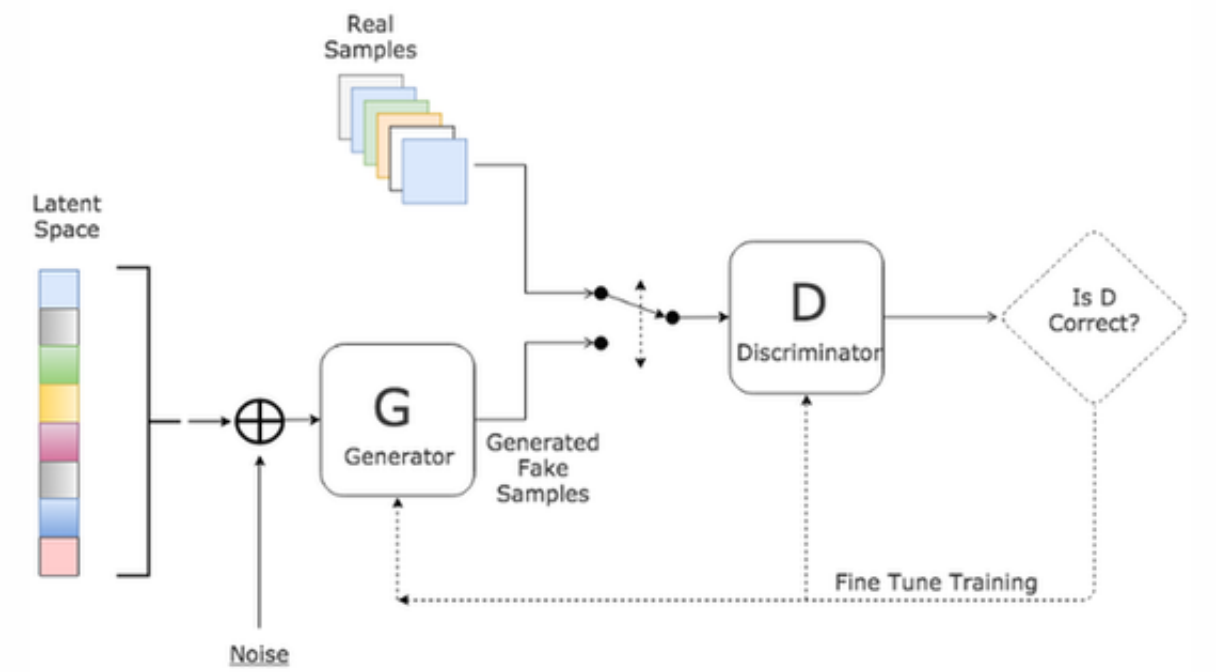
SS-GAN reading notes (CVPR2019)

Background Knowledges

Ian Goodfellow's GAN

> GAN在结构上受博弈论中的二人零和博弈（即二人的利益之和为零，一方的所得正是另一方的所失）的启发，设定参与游戏双方分别为一个生成器（Generator）和一个判别器（Discriminator）。生成器的目的是尽量去学习和捕捉真实数据样本的潜在分布，并生成新的数据样本；判别器是一个二分类器，目的是尽量正确判别输入数据是来自真实数据还是来自生成器。为了取得游戏胜利，这两个游戏参与者需要不断优化，各自提高自己的生成能力和判别能力，这个学习优化过程就是一个极小极大博弈(Minimax game)问题，目的是寻找二者之间的一个纳什均衡，使生成器估测到数据样本的分布
> 生成器G类似造假币的，一个劲地学习如何骗过D；而判别器D，类似稽查警察，一个劲地学习如何分辨出G的造假技巧



理论推导大概包含以下几个步骤：

1. （形式化）用数学语言形式化这个GAN模型的价值函数（想法也许来自二进制交叉熵（Binary Cross Entropy））

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[log(D(x))] + E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

2. （最优解）证明上式的最优解 G 和 D 是存在且唯一的，分三步

   2.1. 基于Radon-Nikodym定理中的等式来对V(D, G)作等价变换（这样来证，生成器G不需要满足可逆条件）

$$E_{z\sim p_z(z)}[log(1-D(G(z)))] = E_{x\sim p_G(x)}[log(1-D(x))]$$

   2.2. 给定生成器G，最大化V(D, G)而得出最优判别器D（实践中并不是可计算的），$D^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_G(x)}$

   2.3. 将得到的D(x)代入V(D, G)，化简成由KL散度（或JS散度）表达的形式，显然就完成了证明

$$\min_G V(D^*,G) = -2log2 + KL(p_{data}(x)||\frac{p_{data}(x)+p_G(x)}{2}) + KL(p_G(x)||\frac{p_{data}(x)+p_G(x)}{2})$$

$$= -2log2 + 2\cdot JSD(p_{data}(x)||p_G(x))$$

3. （收敛）证明给定足够的训练数据和正确的环境，训练过程将收敛到最优解（$P_G = P_{data}$）

有木有感觉！GAN就是典型的先有了实验, 然后为了写paper强行套公式！！！

更多推导细节可以参考机器之心这篇文章，机器之心GitHub项目：GAN完整理论推导与实现，Perfect！

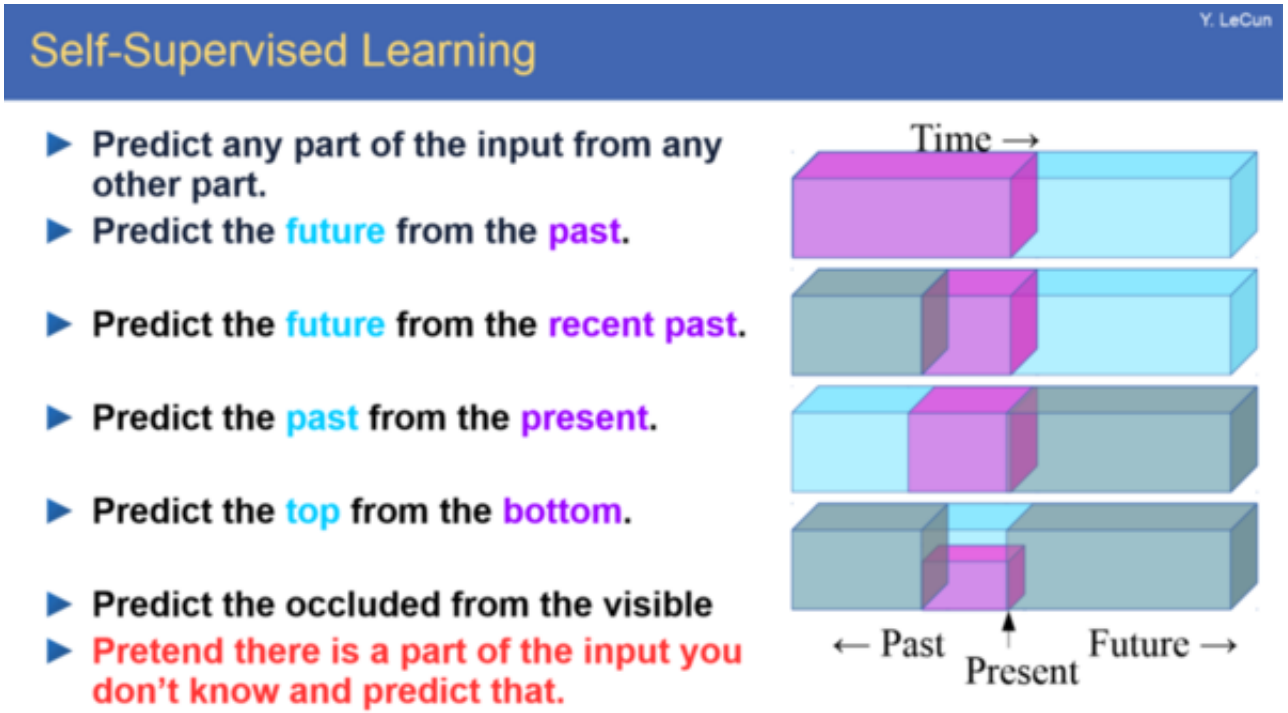
Self-Supervised Learning

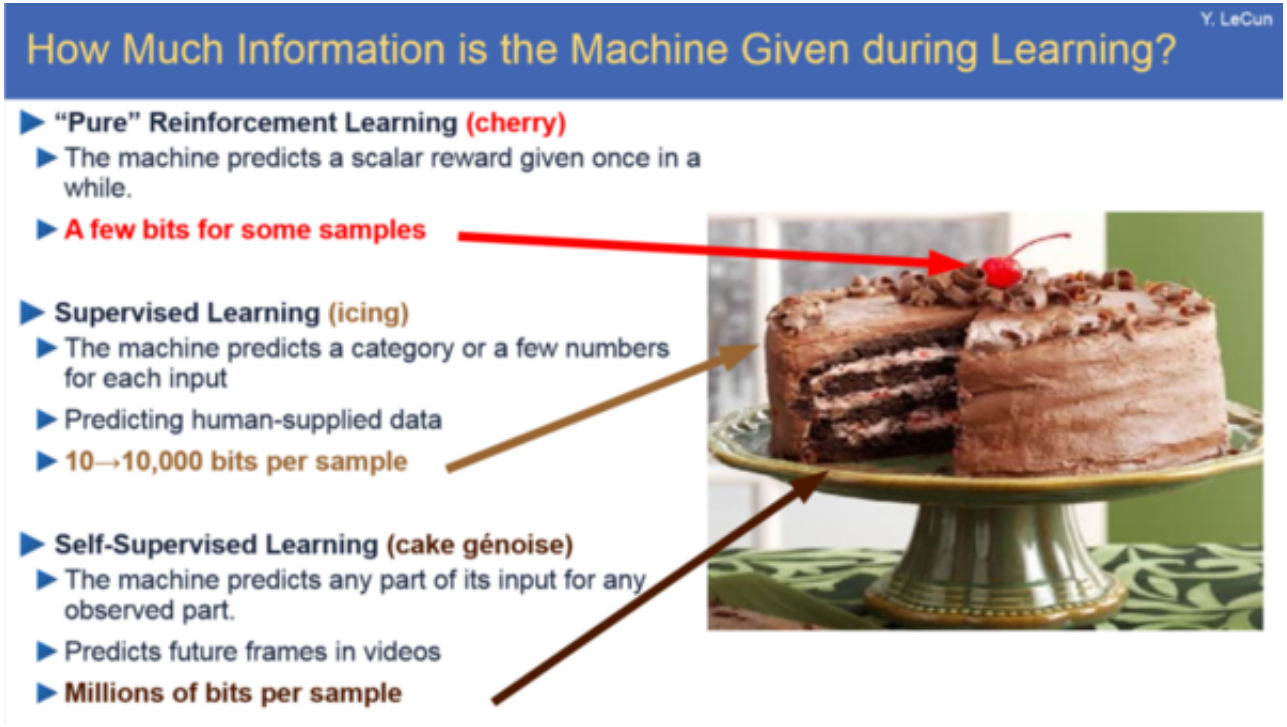
Self-supervised learning不能说完全就是自动生成标签，更好的理解是自动学习更多的目标对象的知识（例如表征，推测），并将其作为一个pretext task（辅助任务）。就好像人的学习一样，通过看到某一种情景，能够很快自动了解这个情景的不同角度的特征，而不是像幼稚园小朋

友一样需要老师的解释（类似监督学习中的贴标签），而且最好这种特征是可以记忆的，所以也许可以借鉴迁移学习的思想。但是根据目前的进展来看，不少使用self-supervised learning而成功的例子都有种碰运气的感觉

Yann LeCun's beautiful explaination

更多Yann LeCun在IJCAI 2018上演讲的内容具体见How Could Machines Learn Like Animals & Humans?





Self-supervised representation learning

> Many ideas have been proposed for self-supervised representation learning on images. A common workflow is to train a model on one or multiple pretext tasks with unlabelled images and then use one intermediate feature layer of this model to feed a multinomial logistic regression classifier on ImageNet classification. The final classification accuracy quantifies
> $how\ good\ the\ learned\ representation\ is$.
>
> Recently, some researchers proposed to train supervised learning on labelled data and self-supervised pretext tasks on unlabelled data simultaneously with $shared\ weights$, like in Zhai et al, 2019 and Sun et al, 2019.

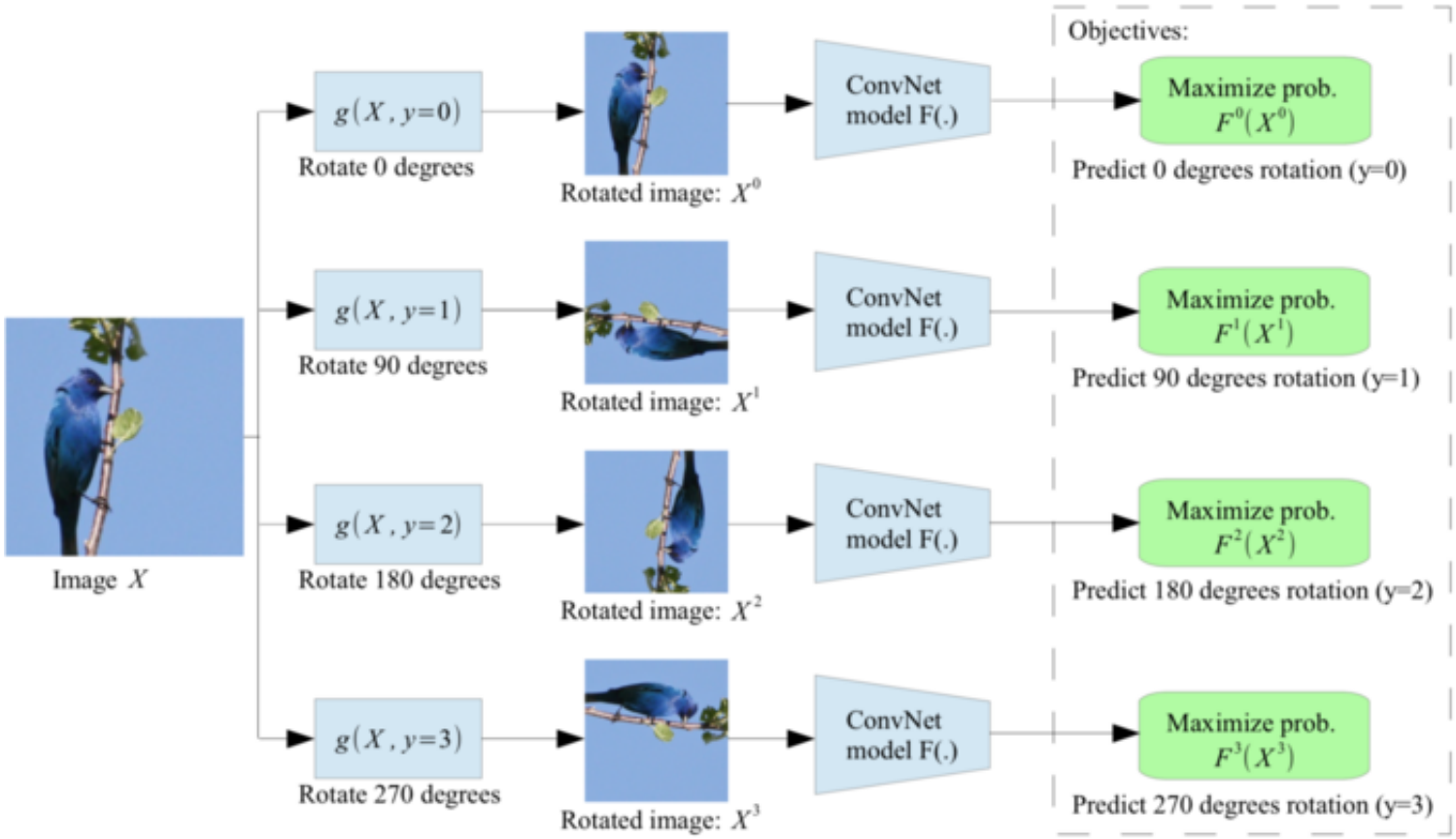
这里简单介绍基于图像的自监督表征学习（当然还有基于视频的、基于控制的），具体参考Self-Supervised Representation Learning [zh-cn]

*Unsupervised Representation Learning by Predicting Image Rotations*中的模型被训练来在语义内容保持不变的情况下修改输入图像，认为图像上的轻微失真不会改变其原始语义或几何形式，因此预计学习到的特征也不会失真

每个输入图像首先随机旋转90°的倍数，分别对应于$[0^\circ, 90^\circ, 180^\circ, 270^\circ]$。模型经过训练可以预测应用了哪种旋转角度，从而得出4分类问题。为了识别旋转了不同角度的同一图像，模型必须学会识别高级对象部分，如头部、鼻子和眼睛，以及这些部分的相对位置，而不是局部模式。这个pretext task驱动模型以这种方式学习对象的语义概念

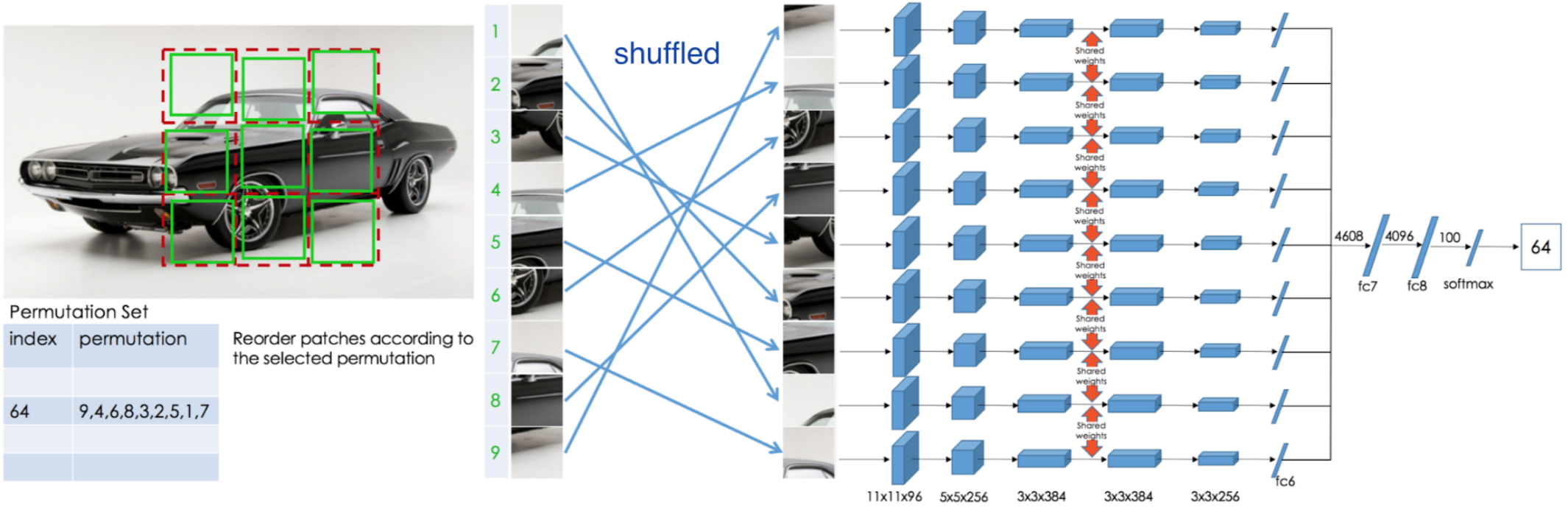
那么为何加了判断图像旋转角度的分类器后，网络可以更好的提取feature?

因为论文中作者假设，如果网络不能很好地理解图像中包含的信息，网络就不可能很好地判断出图像旋转的角度。在GAN中，对于discriminator来说，通过约束discriminator让它能识别出旋转角度，有助于让discriminator更好、更全面地理解图像中包含的信息，从而间接引导discriminator学出来一个通用、稳定的feature提取方法



*Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*中的模型被训练来将9个打乱的patches放回原来的位置，模型学习到了这些patches之间的关系，也即学习到了一些语义表征

下图的卷积网络使用权值共享来独立处理每个patch，并输出一个概率向量，概率值最大的索引在排序集（permutation set）中找到对应的排序。为了控制拼图游戏的难度（排序种类不能太多，论文中是100种），论文按照预先定义的排序集对patch进行洗牌，然后使用置换不变图卷积网络（GCN）替换卷积网络能够加快训练速度，具体参考*PIC: Permutation Invariant Critic for Multi-Agent Deep Reinforcement Learning*



Discriminator Forgetting

这里列出一些常见的GAN训练时的问题以及相应的一些解决方法，更多内容具体参考这份攻略帮你「稳住」反复无常的 GAN

- instability（不稳定性，由WGAN-GP彻底解决，即不再需要小心平衡生成器和判别器的训练程度）
- divergence（发散，也即收敛性的问题，传统GAN不能基于loss的值来判别GAN的收敛性，好在WGAN-GP给出了像交叉熵、准确率这样的

数值来指示训练的进程）
- cyclic behavior（循环行为，可以理解为训练过程有很多反复，GAN的记忆机制差，这篇论文对其有所改进）
- mode collapse（模式崩溃，即生成样本的多样性差，由WGAN-GP基本解决）

作者认为出现这些问题的原因之一是生成器和判别器在一种非稳定的环境下学习，即生成器G的参数改变时，分布P_G随之发生改变，这导致判别器处于一种非静态的学习环境，也就引出了这篇论文的Key Issue：Discriminator的健忘性

以往使用CGAN能够基本解决这个问题，但是CGAN也是有缺点的，它需要大量标记数据，所以这篇论文提出了自监督学习，不必标记数据但能接近CGAN的效果

论文中的这段话举了一个栗子来说明Discriminator的健忘性

> In the context of GANs, learning varying levels of detail, structure, and texture, can be considered different tasks. For example, if the generator first learns the global structure, the discriminator will naturally try to build a representation which allows it to efficiently penalize the generator based only on the differences in global structure, or the lack of local structure. As such, one source of instability in training is that the discriminator is not incentivised to maintain a useful data representation as long as the current representation is useful to discriminate between the classes.

论文中还用了几个实验进一步证实了Discriminator的健忘性

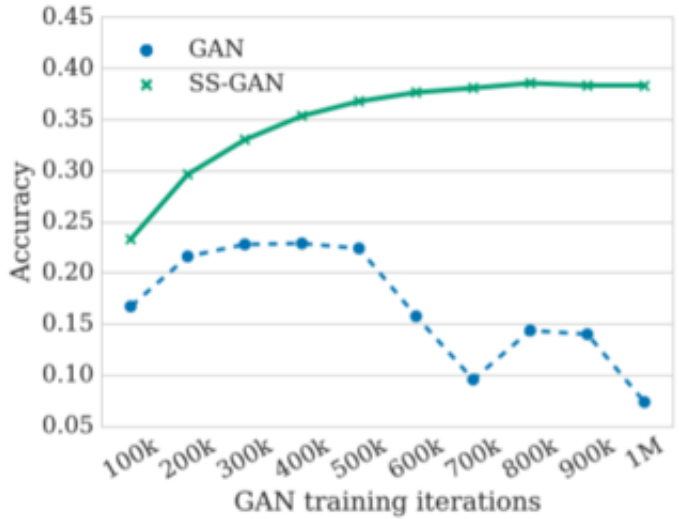


Figure 2: Performance of a linear classification model, trained on IMAGENET on representations extracted from the final layer of the discriminator. Uncond-GAN denotes an unconditional GAN. SS-GAN denotes the same model when self-supervision is added. For the Uncond-GAN, the representation gathers information about the class of the image and the accuracy increases. However, after 500k iterations, the representations lose information about the classes and performance decreases. SS-GAN alleviates this problem. More details are presented in Section 4.

Figure 2表示每经过100k次迭代，从discriminator的最后一层提取出表征，喂给一个在IMAGENET数据集上训练好了的线性分类器，比较Uncond-GAN和SS-GAN对应的分类器分类的准确率。从实验结果来看，从500k次迭代开始，Uncond-GAN的discriminator的表征开始往复地丢失和获得类别信息，即似乎开始“健忘”。这说明了Uncond-GAN的discriminator学到的表征很不稳定，相对应的，SS-GAN的discriminator学到的表征就显得很稳定

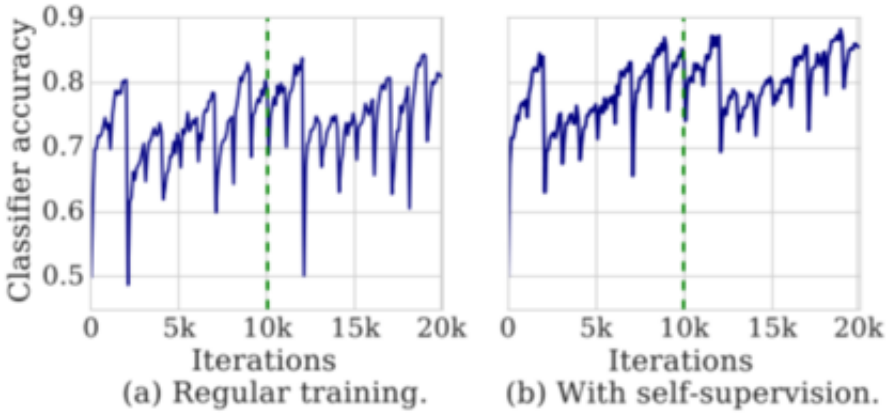


Figure 3: Image classification accuracy when the underlying class distribution shifts every 1k iterations. The vertical dashed line indicates the end of an entire cycle through the tasks, and return to the original classification task at $t = 0$. *Left*: vanilla classifier. *Right*: classifier with an additional self-supervised loss. This example demonstrates that a classifier may fail to learn generalizable representations in a non-stationary environment, but self-supervision helps mitigate this problem.

Figure 3的实验方法与Figure 2大致相同，区别在于这次的数据集是CIFAR-10，而且是依次使用CIFAR-10中的10个类来训练，每个类训练1k次迭代，然后将从discriminator提取的特征喂给线性分类器获得准确率。从左图可以看出，每次训练图像的类型发生变化时，分类器的性能明显下降，10k次迭代后，看上去像是从头学习，之前学习到的方法似乎都被“遗忘”了，而相对应的，右图SS-GAN对应的准确率整体不断上升，显得有一定的记忆力

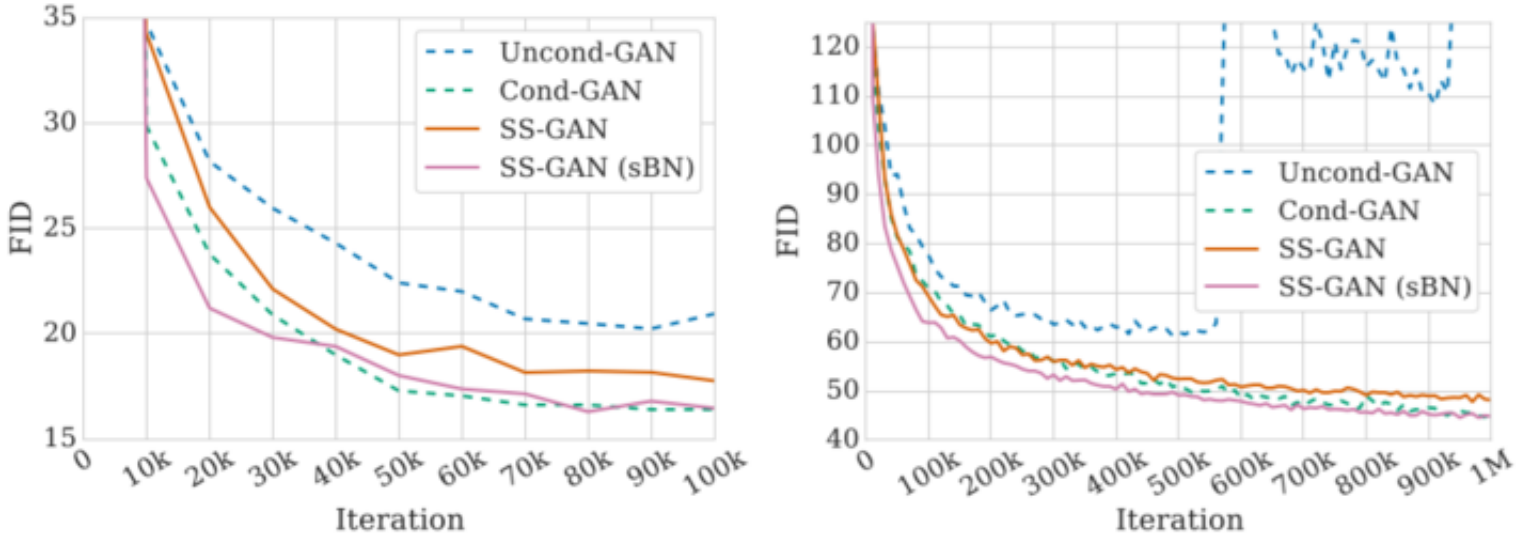
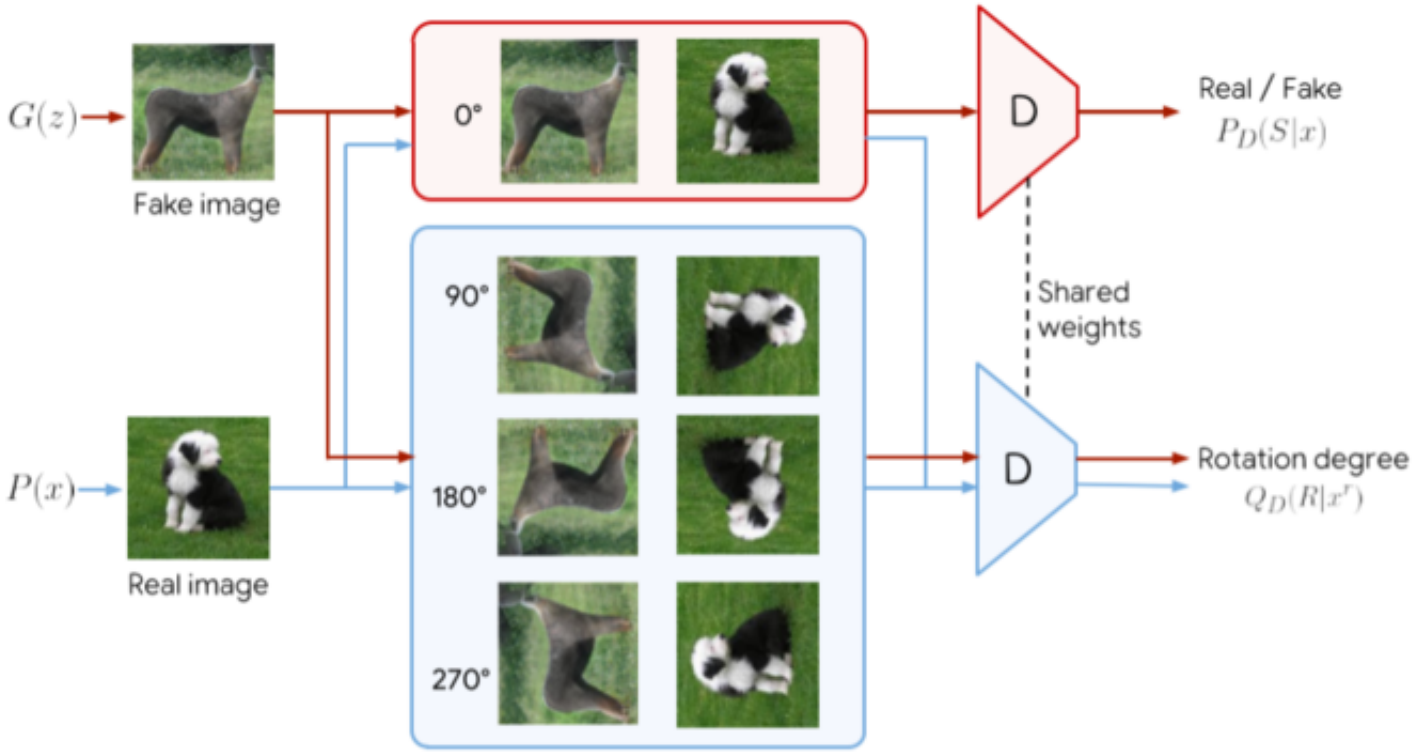


Figure 4: FID learning curves on CIFAR10 and IMAGENET. The curves show the mean performance across three random seeds. The unconditional GAN (Uncond-GAN) attains significantly poorer performance than the conditional GAN (Cond-GAN). The unconditional GAN is unstable on IMAGENET and the runs often diverge after 500k training iterations. The addition of self-supervision (SS-GAN) stabilizes Uncond-GAN and boosts performance. Finally, when we add the additional self-modulated Batch Norm (sBN) [7] to SS-GAN, which mimics generator conditioning in the unconditional setting, this unconditional model attains the same mean performance as the conditional GAN.

Figure 4表示Uncond-GAN在训练过程中FID值可能会发生剧烈的抖动（如右边IMAGENET上500k迭代之后），这说明了训练过程的不稳定性，也间接说明Discriminator的健忘性。然后也可以看出，SS-GAN(sBN)的FID值是很逼近Cond-GAN的，所以说在生成图像的质量和多样性的衡量上，SS-GAN也体现出了它的优越性

Workflow



SS-GAN与传统GAN在搭建网络方面的区别简单而明显，大致有以下这么几点：

1. 生成器和判别器都使用包含5个残差模块的ResNet网络结构

ResNet网络中最主要的两个概念是residual block（残差模块）和skip connection（跳跃连接）

2. 判别器由Real / Fake的判断和Rotation degree的判断两部分组成，且这两部分权值共享

具体实现：

- Discriminator按之前的方法，输出real / fake的判别结果
- 取Discriminator倒数第二层的输出，作为feature，加上一个linear层，预测出旋转的类型

3. 生成器和判别器之间加入对图像的旋转操作

判别器中Real / Fake的判断部分的输入是正常的图像（即没有旋转过的），GPU上训练时batch_size=64

判别器中Rotation degree的判断部分的输入是所有旋转过了的图像（而且是带标签的），旋转角度是满足$R = \{0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\}$，GPU上训练时num_rotated_examples=16，batch_size=num_rotated_examples*4=64

4. 对于Real / Fake的判断使用的是hinge loss和gradient penalty

$$L_G^{GAN} = -E_{z \sim P_z}[D(G(z))]$$

$$L_D^{GAN} = \underbrace{E_{x \sim P_{data}}[max(0, 1 - P_D(x))] + E_{z \sim P_z}[max(0, 1 + P_D(G(z)))]}_{hinge\ loss} + \underbrace{\lambda E[(|\nabla P_D(\alpha x - (1 - \alpha G(z)))| - 1)^2]}_{gradient\ penalty}$$

$$L_G^{SS-GAN} = L_G^{GAN} + \alpha E_{x \sim P_G} E_{r \sim R}[logQ_D(x^r)]$$

$$L_D^{SS-GAN} = L_D^{GAN} + \underbrace{\beta E_{x \sim P_{data}} E_{r \sim R}[logQ_D(x^r)]}_{rotation\ augmented\ loss}$$

这里罗列一下各类GAN loss的数学表达式

除最后一行的Hinge Loss使用梯度下降法更新L_D和L_G，其它都使用梯度上升法更新L_D和L_G

GAN	Arxiv	$L_D^{GAN} = E\big[\log\big(D(x)\big)\big] + E\big[\log\big(1 - D(G(z))\big)\big]$ $L_G^{GAN} = E\big[\log\big(D(G(z))\big)\big]$

LSGAN	Arxiv	$L_D^{LSGAN} = E[(D(x)-1)^2] + E[D(G(z))^2]$ $L_G^{LSGAN} = E[(D(G(z))-1)^2]$
WGAN	Arxiv	$L_D^{WGAN} = E[D(x)] - E[D(G(z))]$ $L_G^{WGAN} = E[D(G(z))]$ $W_D \leftarrow clip\_by\_value(W_D, -0.01, 0.01)$
WGAN-GP	Arxiv	$L_D^{WGAN\_GP} = L_D^{WGAN} + \lambda E[(\lvert\nabla D(\alpha x - (1-\alpha G(z)))\rvert - 1)^2]$ $L_G^{WGAN\_GP} = L_G^{WGAN}$
DRAGAN	Arxiv	$L_D^{DRAGAN} = L_D^{GAN} + \lambda E[\left(\lvert\nabla D(\alpha x - (1-\alpha x_p))\rvert - 1\right)^2]$ $L_G^{DRAGAN} = L_G^{GAN}$
CGAN	Arxiv	$L_D^{CGAN} = E[\log(D(x,c))] + E[\log(1 - D(G(z),c))]$ $L_G^{CGAN} = E[\log(D(G(z),c))]$

infoGAN	Arxiv	$L_{D,Q}^{infoGAN} = L_D^{GAN} - \lambda L_I(c,c')$ $L_G^{infoGAN} = L_G^{GAN} - \lambda L_I(c,c')$
ACGAN	Arxiv	$L_{D,Q}^{ACGAN} = L_D^{GAN} + E[P(class = c\lvert x)] + E[P(class = c\lvert G(z))]$ $L_G^{ACGAN} = L_G^{GAN} + E[P(class = c\lvert G(z))]$
EBGAN	Arxiv	$L_D^{EBGAN} = D_{AE}(x) + \max(0, m - D_{AE}(G(z)))$ $L_G^{EBGAN} = D_{AE}(G(z)) + \lambda \cdot PT$
BEGAN	Arxiv	$L_D^{BEGAN} = D_{AE}(x) - k_t D_{AE}(G(z))$ $L_G^{BEGAN} = D_{AE}(G(z))$ $k_{t+1} = k_t + \lambda(\gamma D_{AE}(x) - D_{AE}(G(z)))$

Hinge Loss	Arxiv	$L_D = E[max(0, 1 - D(x))] + E[max(0, 1 + D(G(z)))]$ $L_G = -E[D(G(z))]$

5. 判别器对于Real / Fake的判断是对立的，但是对于Rotation degree的判断是协作的

$$L_G^{SS-GAN} = L_G^{GAN} + \alpha E_{x \sim P_G} E_{r \sim R}[logQ_D(x^r)]$$

$$L_D^{SS-GAN} = L_D^{GAN} + \underbrace{\beta E_{x \sim P_{data}} E_{r \sim R}[logQ_D(x^r)]}_{rotation\ augmented\ loss}$$

对于判别器对于Real / Fake的判断是对立的这点好理解（不然怎么还能叫GA(Adversarial)N），至于对于Rotation degree的判断是协作的，看上面的L_G^{SS-GAN}和L_D^{SS-GAN}，生成器基于生成图像的rotation augmented loss来更新参数（这鼓励了生成器生成易于检测旋转的图像），而判别器基于真实图像的rotation augmented loss来更新参数（为了使得判别器不那么容易检测生成器生成图像的旋转，所以没有使用生成图像的rotation augmented loss），以上这些都体现出是为了更好地训练判别器（生成器在Rotation degree任务这方面更像是对判别器的一个辅助），使其获得更好更稳定的表征，所以看上去像是一个协作的过程

Experiments

Frechet Inception Distance

$$FID = || \mu_x - \mu_g ||_2^2 + Tr(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}})$$

μ_x（μ_g）：真实图片（生成图片）的所有feature向量的均值

$\sum_x$（$\sum_g$）：真实图片（生成图片）的所有feature向量的协方差矩阵

基本原理：

计算Inception Score（IS）时只考虑了生成样本，没有考虑真实样本，即IS无法反映真实样本和生成样本之间的距离。而Frechet Inception Distance（FID）度量了真实样本和生成样本的feature向量（提取预训练好的Inception Net-V3的全连接层之前（池化层之后）的2048维向量）的距离，其中用到了均值和协方差，将真实样本和的生成样本的所有feature向量分别归纳为多维高斯分布（$X_r \sim N(\mu_x, \sum_x), X_g \sim N(\mu_g, \sum_g)$），即FID是在衡量两个多维高斯分布的距离

三个优点：

1. 生成器的训练集和Inception Net-V3的训练集可以不同
2. 计算FID时同时用到了生成数据和真实数据，比起IS来更灵活。可以理解成，IS判断真实性与否，是把生成数据和ImageNet数据做比较，而FID是把生成数据和真实数据做比较，因此更reasonable
3. 以优化FID为目标，不会产生对抗样本，即生成图片不会失真。因为优化的是lantent space feature，而不是最终的输出图片

在不同数据集上比较不同GAN模型

实验在四个数据集IMAGENET，CIFAR-10，LSUN-BEDROOM，CELEBA-HQ上运行并获得最佳的FID值

硬件配置是一个P100 GPU，其中IMAGENET训练了1M次，其它三个各训练了100k次

Uncond-GAN使用的是*Spectral Normalization for Generative Adversarial Networks*，在学会分类ImageNet的1000个类这个任务上面明显优于AC-GAN，后者被认为是目前最常用的条件GAN，既能生成图像又能进行分类

Cond-GAN使用的*cGANs with Projection Discriminator*，该模型提出不久，性能优于经典的AC-GAN

SS-GAN是在Uncond-GAN的基础上加入了基于图像旋转的自监督

SS-GAN(sBN)是在网络中加入了batch normalization，为了和Cond-GAN相对应（Cond-GAN实现中有batch normalization）

从实验结果来看，一般来说SS-GAN(sBN)优于SS-GAN优于Uncond-GAN，然后SS-GAN(sBN)比较逼近Cond-GAN

DATASET	METHOD	FID
CIFAR10	Uncond-GAN	19.73
	Cond-GAN	**15.60**
	SS-GAN	17.11
	SS-GAN (sBN)	15.65
IMAGENET	Uncond-GAN	56.67
	Cond-GAN	**42.07**
	SS-GAN	47.56
	SS-GAN (sBN)	43.87
LSUN-BEDROOM	Uncond-GAN	16.02
	SS-GAN	13.66
	SS-GAN (sBN)	**13.30**
CELEBA-HQ	Uncond-GAN	**23.77**
	SS-GAN	26.11
	SS-GAN (sBN)	24.36

Table 1: Best FID attained across three random seeds. In this setting the proposed approach recovers most of the benefits of conditioning.

SS-GAN对超参的鲁棒性

在不同超参的设置下，SS-GAN对比Uncond-GAN，不仅FID的平均值小，标准差也小，说明了对超参的鲁棒性强

涉及到的超参来自于：

1. gradient penalty（超参λ）和spectral normalization这两种Lipschitz限制手法
2. Adam optimizer的超参β_1和β_2
3. 每训练一次生成器，就训练DITERS次判别器，这源于GAN训练的一个技巧：往往判别器训练的次数多于生成器最终效果好

TYPE	λ	β_1	β_2	D ITERS	CIFAR10 UNCOND-GAN	CIFAR10 SS-GAN	IMAGENET UNCOND-GAN	IMAGENET SS-GAN
GRADIENT PENALTY	1	0.0	0.900	1	121.05 ± 31.44	**25.8 ± 0.71**	183.36 ± 77.21	**80.67 ± 0.43**
				2	28.11 ± 0.66	**26.98 ± 0.54**	85.13 ± 2.88	**83.08 ± 0.38**
		0.5	0.999	1	78.54 ± 6.23	**25.89 ± 0.33**	104.73 ± 2.71	**91.63 ± 2.78**
	10	0.0	0.900	1	188.52 ± 64.54	**28.48 ± 0.68**	227.04 ± 31.45	**85.38 ± 2.7**
				2	29.11 ± 0.85	**27.74 ± 0.73**	227.74 ± 16.82	**80.82 ± 0.64**
		0.5	0.999	1	117.67 ± 17.46	**25.22 ± 0.38**	242.71 ± 13.62	**144.35 ± 91.4**
SPECTRAL NORM	0	0.0	0.900	1	87.86 ± 3.44	**19.65 ± 0.9**	129.96 ± 6.6	**86.09 ± 7.66**
				2	20.24 ± 0.62	**17.88 ± 0.64**	80.05 ± 1.33	**70.64 ± 0.31**
		0.5	0.999	1	86.87 ± 8.03	**18.23 ± 0.56**	201.94 ± 27.28	**99.97 ± 2.75**

Table 2: FID for unconditional GANs under different hyperparameter settings. Mean and standard deviations are computed across three random seeds. Adding the self-supervision loss reduces the sensitivity of GAN training to hyperparameters.

在IMAGENET上获得最佳FID值

为了在IMAGENET上获得SS-GAN最佳的FID值，扩大SS-GAN模型的容量使与BigGAN相匹配，后者目前被认为是最强的GAN，相对来说规模比较大，参数比较多，当然稳定性也就比较好

硬件配置是128核的Google TPU v3 Pod，训练了500k次，使用的batch_size=2048

SS-GAN获得了FID=23.6 ± 0.1的好成绩，而且可以看出鲁棒性很强（标准差才0.1），然后通过随机选取种子进行训练，最终得到最优的FID=23.4，这是在IMAGENT上，目前已知的最好的无监督GAN的结果

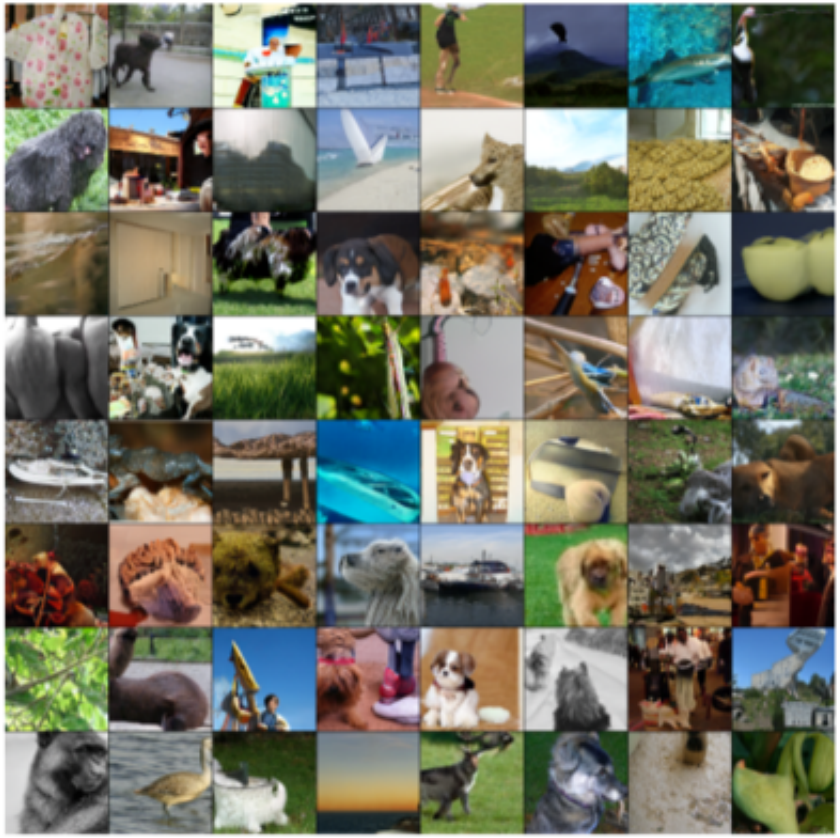


Figure 5: A random sample of unconditionally generated images from the self-supervised model. To our knowledge, this is the best results attained training unconditionally on IMAGENET.

评价Discriminator表征学习的结果

为了证明自监督学习的确提高了Discriminator学到的表征的质量，论文作者采用了常用的表征学习评价方法来评价。具体来说就是训练一个逻辑斯蒂回归分类器，对从判别器网络的每个ResNet Block后提取出来的feature进行多分类，计算top-1分类准确度

从Table 6和Table 7的实验结果可以看出，不论是CIFAR-10数据集，还是IMAGENET数据集，SS-GAN(sBN)的表现有目共睹，分类准确率不仅

按ResNet Block顺序是不断递增的，而且是各模型中最优的

Method	Uncond-GAN	Cond-GAN	Rot-only	SS-GAN (sBN)
Block0	0.719 ± 0.002	0.719 ± 0.003	0.710 ± 0.002	$\mathbf{0.721 \pm 0.002}$
Block1	0.762 ± 0.001	0.759 ± 0.003	0.749 ± 0.003	$\mathbf{0.774 \pm 0.003}$
Block2	0.778 ± 0.001	0.776 ± 0.005	0.762 ± 0.003	$\mathbf{0.796 \pm 0.005}$
Block3	0.776 ± 0.005	0.780 ± 0.006	0.752 ± 0.006	$\mathbf{0.799 \pm 0.003}$
Best	0.778 ± 0.001	0.780 ± 0.006	0.762 ± 0.003	$\mathbf{0.799 \pm 0.003}$

Table 6: Top-1 accuracy on CIFAR10 with standard variations.

Method	Uncond-GAN	Cond-GAN	Rot-only	SS-GAN (sBN)
Block0	0.074 ± 0.074	0.156 ± 0.002	0.147 ± 0.001	$\mathbf{0.158 \pm 0.001}$
Block1	0.063 ± 0.103	0.187 ± 0.010	0.134 ± 0.003	$\mathbf{0.222 \pm 0.001}$
Block2	0.073 ± 0.124	0.217 ± 0.007	0.158 ± 0.003	$\mathbf{0.250 \pm 0.001}$
Block3	0.083 ± 0.142	0.272 ± 0.014	0.202 ± 0.005	$\mathbf{0.327 \pm 0.001}$
Block4	0.077 ± 0.132	0.253 ± 0.040	0.196 ± 0.001	$\mathbf{0.358 \pm 0.005}$
Block5	0.074 ± 0.126	0.337 ± 0.010	0.195 ± 0.029	$\mathbf{0.383 \pm 0.007}$
Best	0.083 ± 0.142	0.337 ± 0.010	0.202 ± 0.005	$\mathbf{0.383 \pm 0.007}$

Table 7: Top-1 accuracy on IMAGENET with standard variations.