arXiv:1905.01164v2 [cs.CV] 4 Sep 2019

SinGAN：从单个自然图像中学习生成模型

Tamar Rott Shaham
Technion

Tali Dekel
Google Research

Tomer Michaeli
Technion

图 1：**从单个训练图像中学到的图像生成。** 我们提出 SinGAN —— 一种在单个自然图像上训练的新的无条件生成模型。我们的模型使用专用的多尺度对抗训练方法来跨多个尺度学习图像的斑块（patch）统计信息；然后可以将其用于生成新的逼真的图像样本，该样本在创建新的对象配置和结构的同时保留原始斑块分布。

摘 要

我们介绍 SinGAN，这是一种无条件的生成模型，可以从单个自然图像中学习。我们的模型经过训练，可以捕获图像内斑块的内部分布，然后能够生成高质量、多样的样本，并承载与图像相同的视觉内容。SinGAN 包含一个完全卷积的 GAN 金字塔，每个 GAN 负责学习不同图像比例的斑块分布。这样就可以生成具有任意大小和纵横比的新样本，这些样本具有显著的可变性，同时又可以保持训练图像的整体结构和精细纹理。与之前的单图像 GAN 方法相比，我们的方法不仅限于纹理图像，而且不是有条件的（即它从噪声中生成样本）。用户研究证实，生成的样本通常被误认为真实图像。我们将说明 SinGAN 在各种图像处理任务中的实用性。

1. 介绍

生成对抗网络（GAN）[19]在建模可视数据的高维分布方面取得了巨大飞跃。特别是，在针对特定类别的数据集（例如，面孔[33]，卧室[47]）进行训练时，无条件 GAN 在生成逼真的，高质量样本方面已取得了显著成功。但是，捕获具有多个对象类的高度多样化的数据集的分布（例如 ImageNet [12]），仍然被认为是一项重大挑战，并且通常需要根据另一个输入信号[6]调整生成或为特定任务训练模型（例如超分辨率[30]，修复[41]，重新定位[45]）。

在这里，我们将 GAN 的使用带入了一个新领域-从单个自然图像中学到的无条件生成。具体而言，我们表明，单个自然图像中斑块的内部统计信息通常具有足够的信息，可以学习强大的生成模型。新的单图像生成模型 SinGAN 使我们能够处理包含复杂结构和纹理的普通自然图像，而不必依赖于存在来自同一类别的图像数据库。这是通过完全卷积的轻型 GAN 金字塔实现的，每个 GAN 负责捕获不同规模的斑块分布。经过训练后，SinGAN 可以生成各种质量的图像样本（任意尺寸），其语义在外观上类似于训练图像，但包含新的对象配置和结构（见底部 1，图 1）。

长期以来，在许多计算机视觉任务中，对单个自然图像中斑块内部分布进行建模已被认为是一项艰巨的任务[64]。典型示例包括去噪[65]，去模糊[39]，超分辨率[18]，去雾[2、15]和图像编辑[37、21、9、11、50]。在这种情况下，与之最相关的工作是[48]，其中定义并优化了双向斑块相似性度量，以确保操纵后的图像斑块与原图像的斑块相同。

1 代码已开源在：https://github.com/tamarott/SinGAN

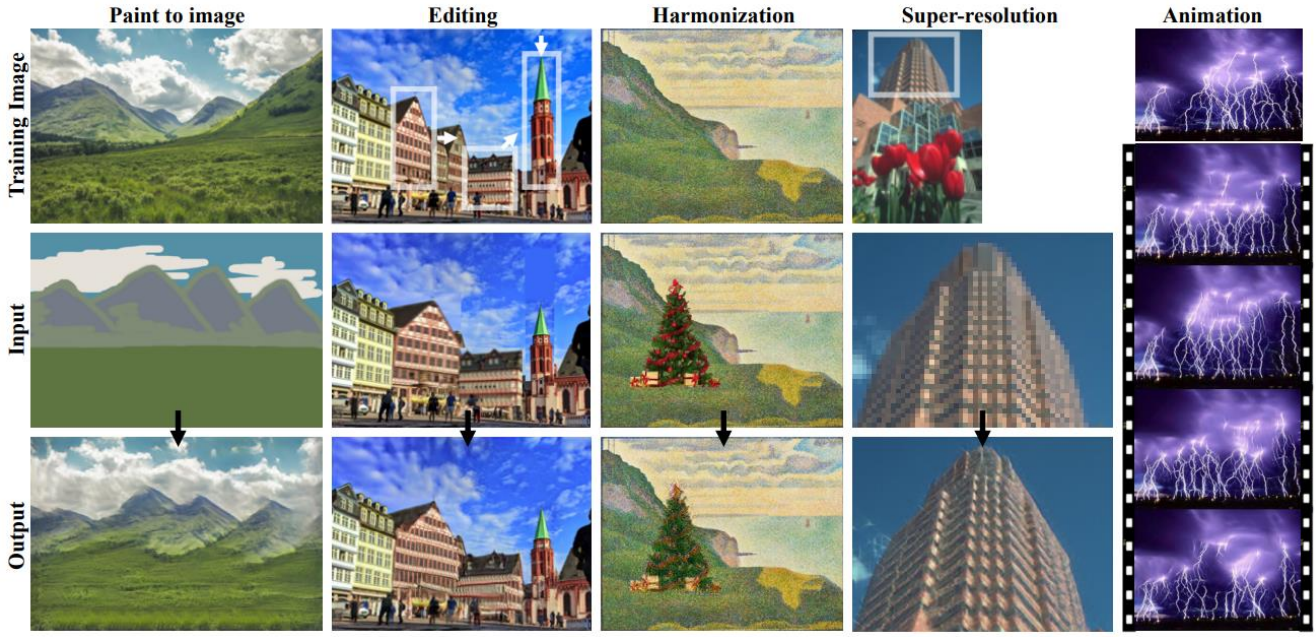

图 2：**图像处理。** SinGAN 可用于各种图像处理任务，包括：将涂料（剪贴画）转换为逼真的照片，重新排列和编辑图像中的对象，将新对象融合入图像，图像超分辨率以及从单个图像创建动画输入。在所有这些情况下，我们的模型仅观察训练图像（第一行），并且对于所有应用程序都以相同的方式进行训练，而无需进行架构更改或进一步调整（请参见第 4 节）。

受这些作品的激励，在这里我们展示了如何在简单的统一学习框架内使用 SinGAN 来解决各种图像处理任务，包括从单个图像进行图像绘制，编辑，协调，超分辨率和动画处理。在所有这些情况下，我们的模型都会产生高质量的结果，并保留训练图像的内部斑块统计信息（请参见图 2 和我们的项目网页）。使用相同的生成网络即可完成所有任务，而无需任何其他信息或超出原始训练图像的进一步训练。

1.1 相关工作

单一图像深度模型 最近有几项工作提出将深度模型“过度拟合”为单个训练示例[51、60、46、7、1]。 但是，这些方法是为特定任务设计的（例如，超分辨率[46]，纹理扩展[60]）。Shocher 等人 [44，45]是第一个为单个自然图像引入基于 GAN 的内部模型，并在重定位的情况下对其进行了说明。然而，它们的生成取决于输入图像（即将图像映射到图像），并且不用于绘制随机样本。相反，我们的框架纯粹是生成性的（即将噪声映射到图像），因此适合许多不同的图像处理任务。仅在纹理生成的背景下研究了无条件的单图像 GAN [3，27，31]。当在非纹理图像上训练时，这些模型不会产生有意义的样本（图 3）。另一方面，我们的方法不限于纹理，并且可以处理一般的自然图像（例如，图 1）。

图 3：**SinGAN 与单图像纹理生成。**用于生成纹理的单图像模型[3，16]并非设计用于处理自然图像。我们的模型可以得出由复杂纹理和非隐身全局结构组成的逼真的图像样本。

图像处理的生成模型 在许多不同的图像处理任务中，最近基于 GAN 的方法已经证明了对抗学习的力量[61、10、62、8、53、56、42、53]。 示例包括交互式图像编辑[61、10]，草图到图像生成 [8、43]和其他图像到图像翻译任务[62、52、54]。 但是，所有这些方法都是在特定类别的数据集上进行训练的，在这里，通常也要根据另一个输入信号来限制生成。我们对捕获同一类图像之间的共同特征并不感兴趣，而是

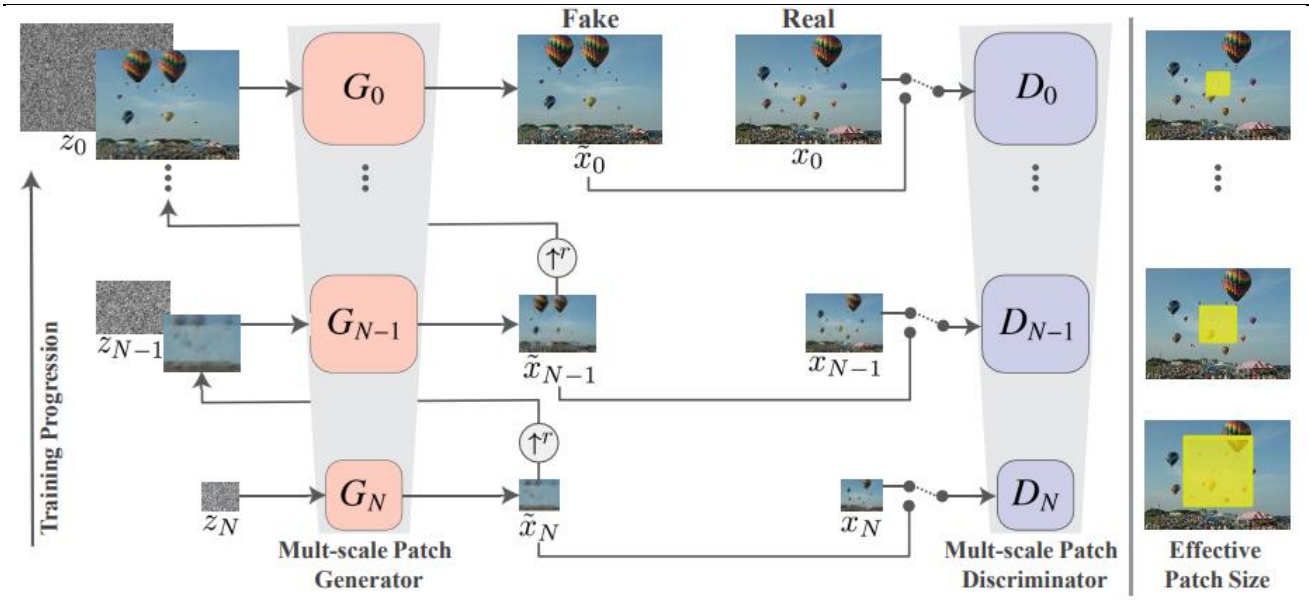

图 4：**SinGAN 的多尺度通道。**我们的模型由 GAN 金字塔组成，其中训练和推理均以从粗到精的方式进行。在每个尺度上，G_n学会生成图像样本，其中判别器D_n无法将所有重叠的块与下采样的训练图像x_n中的块区分开。随着我们上采样金字塔，有效斑块的大小会减小（在原始图像上以黄色标记为说明）。G_n的输入是一个随机噪声图像z_n，并且从前一个尺度$\tilde{x}_n$生成的图像被上采样到当前分辨率（除了最粗糙的级别，它纯粹是生成的）。在 n 级的生成过程涉及所有生成器$\{G_N \dots G_n\}$和所有噪声映射$\{z_N, \dots, z_n\}$ 都达对应级别。在第 2 节上查看更多详细信息。

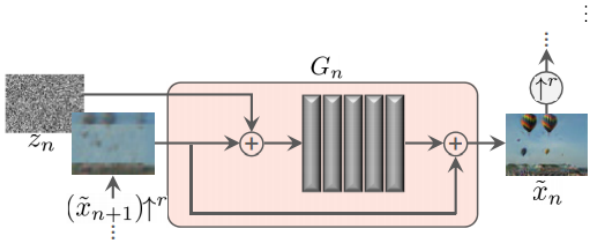

图 5：**单尺度生成。** 在每个尺度 n 处，对前一个尺度$\tilde{x}_{n+1}$的图像进行上采样并添加到输入噪声图z_n中。 结果被馈送到 5 个卷积层，其输出是残差图像，该残差图像被加回到$(\tilde{x}_{n+1})\uparrow^r$。这是$G_n$的输出$\tilde{x}_n$。

考虑不同的训练数据源–所有重叠的块都在单个自然图像的多个尺度上。我们表明，可以从此数据中学习强大的生成模型，并将其用于许多图像处理任务中。

2. 方法

我们的目标是学习一个无条件的生成模型，该模型捕获单个训练图像 x 的内部统计数据。此任务在概念上与常规 GAN 设置相似，不同之处在于，此处的训练样本是单个图像的斑块，而不是数据库中的整个图像样本。

我们选择超纹理生成，并处理更一般的自然图像。这需要以许多不同的比例来捕获复杂图像结构的统计信息。例如，我们要捕获全局属性，

例如图片中大型物体的排列和形状（例如顶部的天空，底部的地面）以及精细的细节和纹理信息。为了实现这一点，我们的生成框架如图 4 所示，由斑块 GAN（马尔可夫判别器）[31，26]的层次结构组成，其中每个斑块 GAN 负责捕获斑块分布的不同规模。GAN 的感受野小且容量有限，从而阻止了它们记忆单个图像。虽然在常规 GAN 设置中已经探索了类似的多尺度体系结构（例如[28、52、29、52、13、24]），但我们还是第一个为从单个图像进行内部学习而探索的体系结构。

2. 1. 多尺度架构

我们的模型由金字塔生成器$\{G_0, \dots, G_N\}$组成。针对 x：$\{x_0, \dots, x_N\}$的图像金字塔训练，其中x_n是 x 的降采样后的r^n倍，其中 r>1。每个生成器G_n负责生成真实的图像样本，即关于相应图像x_n中的斑块分布。这是通过对抗训练实现的，在对抗训练中，G_n学会欺骗对应的判别器D_n，该判别器试图将生成的样本中的斑块与x_n中的斑块区分开。

图像样本的生成从最粗糙的级别开始，并依次通过所有生成器直到最精细的级别，并且在每个级别都注入了噪声。所有的生成器和判别器都具有相同的感受野，因此随着我们的生成过程的进行，感受野的大小逐渐减小。在最粗糙的尺度上，生成是纯粹生成的，即G_n将空间中的白高斯噪声z_N映射到图像样本$\tilde{x}_N$，

$$\tilde{x}_N = G_N(z_N). \quad (1)$$

在此级别上，有效的感受野通常是图像高度的约 1/2，因此G_n会生成图像的总体布局以及对象的全局结构。每个生成器G_n都以较小的比例（n <N）添加了在之前尺度上未生成的详细信息。 因此，除了空间噪声z_n之外，每个生成器G_n还考虑较粗的尺度，即，

$$\tilde{x}_n = G_n\left(z_n, (\tilde{x}_{n+1})\uparrow^r\right), \qquad n < N. \quad (2)$$

如图 5 所示，所有生成器都具有类似的体系结构。具体而言，在将噪声z_n馈送到一系列卷积层之前，将其添加到图像$(\tilde{x}_{n+1})\uparrow^r$。GAN 不会忽略噪声，就像在涉及随机的条件方法中经常发生的那样[62，36，63]。卷积层的作用是生成$(\tilde{x}_{n+1})\uparrow^r$（残差学习[22，57]）。 即$G_n$执行操作

$$\tilde{x}_n = (\tilde{x}_{n+1})\uparrow^r + \psi_n\left(z_n + (\tilde{x}_{n+1})\uparrow^r\right), \quad (3)$$

其中ψ_n是具有 5 个卷积块的完全卷积网络，形式为 Conv(3×3)-BatchNorm-LeakyReLU [25]。我们以每个块最粗的尺度从 32 个卷积核开始，并将此数字每 4 个尺度增加 2 倍。由于生成器是完全卷积的，因此我们可以在测试时生成任意大小和纵横比的图像（通过更改噪声图的尺寸）。

2. 2. 训练

从最粗到最细的顺序，我们依次训练我们的多尺度体系结构。训练完每个 GAN 后，它就会保持固定。我们对第 n 个 GAN 的训练损失包括一个对抗损失和一个重建损失，

$$\min_{G_n} \max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha\mathcal{L}_{\text{rec}}(G_n). \quad (4)$$

对抗损失$\mathcal{L}_{\text{adv}}$对x_n中斑块分布与生成样本$\tilde{x}_n$中斑块分布之间的距离进行惩罚。重建损失$\mathcal{L}_{\text{rec}}$确保存在可以产生x_n的特定噪声图集，x_n是图像处理的重要特征（第 4 节）。接下来我们详细描述$\mathcal{L}_{\text{adv}}$和$\mathcal{L}_{\text{rec}}$。有关优化的详细信息，请参见补充材料（SM）。

对抗损失 每个生成器G_n都与一个马尔可夫判别器D_n耦合，后者将其输入的每个重叠斑块分类为真实或假冒[31，26]。我们使用 WGAN-GP 损失[20]，我们发现该损失可提高训练稳定性，其中最终判别得分是斑块判别图上的平均值。与用于纹理的单图像 GAN 相对（例如[31，27，3]），此处我们定义了整个图像的损失，而不是随机作物（批大小为 1）的损失。这允许网络学习边界条件（请参阅 SM），这是我们设置中的重要功能。D_n的架构与G_n中的网络 n 相同，因此其斑块大小（网络的感受野）为 11×11。

重建损失 我们要确保存在一组特定的输入噪声图，这些图会生成原始图像 x。我们专门选择$\{z_N^{\text{rec}}, z_{N-1}^{\text{rec}} \ldots, z_0^{\text{rec}}\} = \{z^*, 0, \ldots, 0\}$，其中$z^*$是一些固定的噪声图（绘制一次并在训练期间保持固定）。当使用这些噪声图时，以$\tilde{x}_n^{\text{rec}}$表示在第 n 个尺度上生成的图像。然后对于 n <N，

$$\mathcal{L}_{\text{rec}} = \|G_n(0, (\tilde{x}_{n+1}^{\text{rec}})\uparrow^r) - x_n\|^2, \quad (5)$$

并且对于 n = N，我们使用$\mathcal{L}_{\text{rec}} = \|G_N(z^*) - x_N\|^2$。

重建图像$\tilde{x}_n$在训练过程中还有另一个作用，即确定每个尺度中噪声z_n的标准偏差σ_n。 具体而言，我们将σ_n与$(\tilde{x}_{n+1}^{\text{rec}})\uparrow^r$和$x_n$之间的均方根误差（RMSE）成比例，这指示了需要在该比例下添加的细节数量。

3. 结果

我们对跨越大范围场景的各种图像进行了定性和定量测试，包括城市和自然风光以及艺术和纹理图像。我们使用的图像来自伯克利细分数据库（BSD）[35]，场景[59]和网络。我们始终将最粗略的最小尺寸设置为 25px，并选择比例数 N 其约束比例因子 r 尽可能接近 4/3。对于所有结果（除非另有说明），我们将训练图像的大小调整为最大尺寸 250px。

我们生成的随机图像样本的定性示例如图 1，图 6 所示，SM 中还包含更多示例。对于每个示例，我们显示了许多随机样本，这些样本的纵横比与原始图像相同，并且每个轴上的维数有减小和扩展。可以看出，在所有这些情况下，生成的样本都描述了对象的新的现实结构和配置，同时保留了训练图像的视觉内容。我们的模型成功地预先存储了对象的全局结构，例如山脉（图 1），气球或金字塔（图 6），以及精细的纹理信息。由于网络的感受野有限（小于整个图像），因此它可以生成训练图像中不存在的色块的新组合。此外，我们观察到，在许多情况下，反射和阴影是可以实际合成的，可以在图 6 和图 1（以及图 8 的第一个示例）中看到。请注意，SinGAN 的体系结构与分辨率无关，因此可以在高分辨率图像上使用，如图 7 所示（请参见 SM 中的 4M pix 结果）。同样，在这里，从天空、云层和山脉的整体布局到雪的精细纹理，都可以很好地生成各种尺度的结构。

测试时尺度的影响 我们的多尺度体系结构允许通过选择在测试时开始生成的尺度来控制样本之间的可变性。首先从尺度 n 开始，我们修正了噪声图

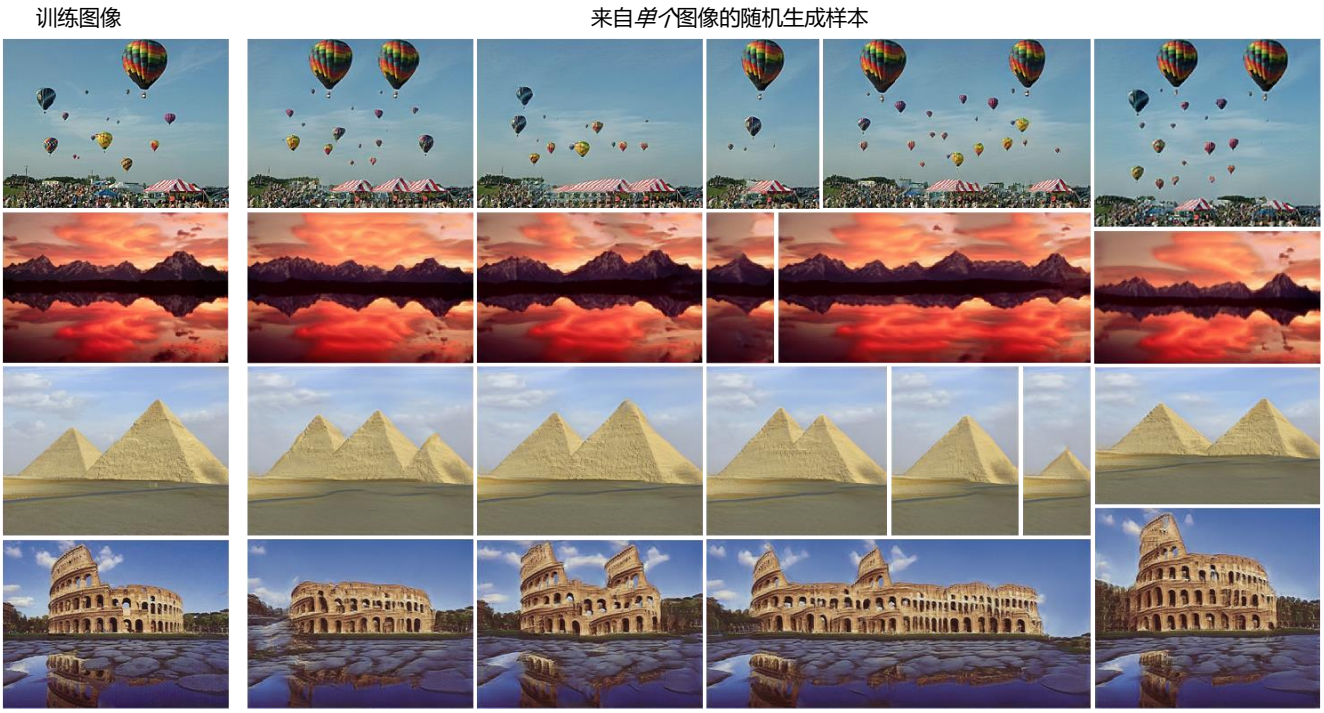

图 6：**随机图像样本。** 在单幅图像上训练 SinGAN 之后，我们的模型可以生成描述新结构和对象配置的逼真的随机图像样本，同时保留训练图像的斑块分布。因为我们的模型是完全卷积的，所以生成的图像可能具有任意大小和纵横比。请注意，我们的目标不是图像重新定向-我们的图像样本是随机的，并且经过优化以维护斑块统计信息，而不是保留显著的对象。有关更多结果以及与图像重新定向方法的定性比较，请参见 SM。

图 7：**高分辨率图像生成。**由我们的模型产生的随机样本，在 243×1024 图像（右上角）上训练；现实地产生了新的全局结构以及细节。 请参阅 SM 中的 4Mpix 示例。

到这个尺度变成$\{z_N^{\text{rec}},\ldots,z_{n+1}^{\text{rec}}\}$，并且仅对$\{z_n,\ldots,z_0\}$使用随机抽奖，效果如图 8 所示。可以看出，以最粗略的尺度（n = N）开始生成会导致全局结构的较大变化。在某些带有大型显著物体的情况下，例如斑马图像，这可能会导致不切实际的采样。但是，从更精细的比例开始生成，可以使整体结构保持完整，同时仅更改更精细的图像功能（例如斑马的条纹）。 有关更多示例，请参见 SM。

尺度在训练过程中的影响 图 9 显示了尺度较小时训练的效果。 在较小尺度下，最粗糙水平的有效感受野是更小的，只允许捕获精细的纹理。随着尺度的增加，出现了更大的支撑结构，并且更好地保留了全局对象的排列。

3.1. 定量评估

为了量化生成的图像的真实性以及它们捕获训练图像的内部统计数据的程度，我们使用两个指标：（i）Amazon Mechanical Turk（AMT）“真实/伪造”用户研究，以及（ii）一个新的图像的 Frechet´初始距离[23]。

AMT 感知研究 我们遵循[26，58]的协议，并在 2 种环境下进行了感知实验。

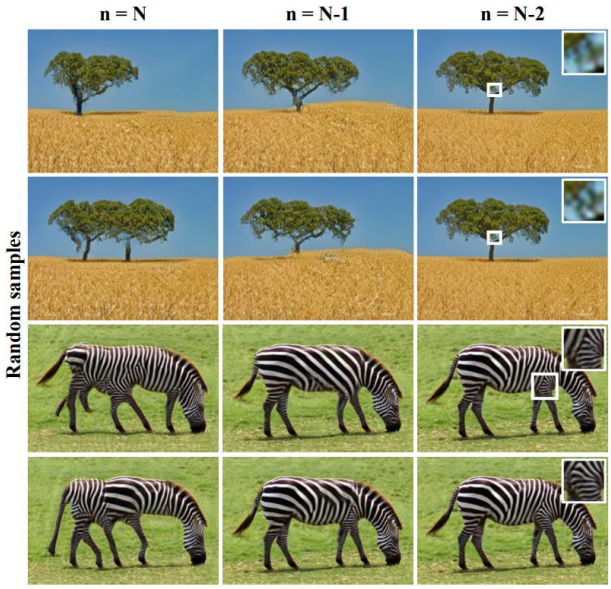

图 8：**从不同尺度生成（按推断）。**我们展示了从给定级别 n 开始分层生成的效果。对于我们的全尺度生成方法（n = N），最粗略的输入是随机噪声。为了从更精细的尺度 n 生成图像，我们插入了降采样后的原始图像，x_n，作为该比例的输入。这使我们可以控制生成的结构的尺度，例如，我们可以保留斑马的形状和姿势，并仅通过从 n = N - 1 开始生成来更改其条纹纹理。

（i）配对（真实与虚假）：向工人展示了 50 个实验的顺序，其中每个实验都将一个伪造的图像（由 SinGAN 生成）与其真实的训练图像一起呈现 1 秒钟。工人被要求挑选假图。

（ii）不成对（真实或伪造）：向工人展示一张图像 1 秒钟，并询问其是否为伪造。总共以随机顺序向每个工人展示了 50 张真实图像和 50 张假图像的不相交的集合。

我们针对两种类型的生成过程重复了这两个协议：从最粗糙的（Nth）尺度和 N-1 尺度开始生成（如图 8 所示）。这样，我们在两个不同的可变性水平中评估了结果的真实性。为了量化生成图像的多样性，对于每个训练示例，我们计算了 100 张生成图像上每个像素的强度值的标准偏差（std），将其平均化到所有像素上，并应用于训练图像以进行正则化。

真实图像是从“场景”数据库[59]中随机选择的，这些数据库来自以下子类别：山脉，丘陵，沙漠，天空。在这 4 个测试中的每个测试中，我们有 50 个不同的参与者。在所有测试中，前 10 个试验都是包含反馈的教程。结果记录在表 1 中。

不出所料，在没有配对参考的不成对情况下，混淆率始终较高。另外，很明显混乱率随着

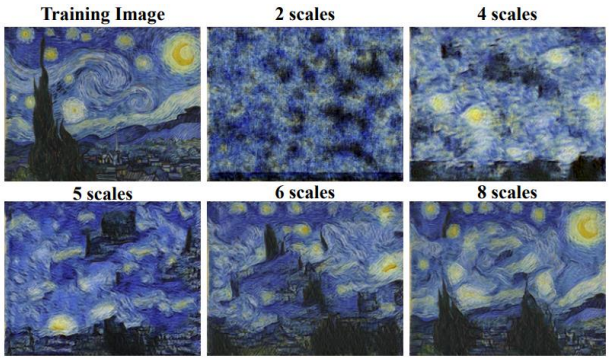

图 9：**使用不同数量的量表进行训练的效果。**SinGAN 体系结构中的尺度数量极大地影响了结果。尺度较小的模型只能捕获纹理。随着比例尺数量的增加，SinGAN 设法捕获更大的结构以及场景中对象的全局布置。

1st Scale	Diversity	Survey	Confusion
N	0.5	paired	$21.45\% \pm 1.5\%$
		unpaired	$42.9\% \pm 0.9\%$
$N-1$	0.35	paired	$30.45\% \pm 1.5\%$
		unpaired	$47.04\% \pm 0.8\%$

表 1：**“真实/伪造”AMT 测试。**我们报告了两个生成过程的混淆率：从最粗的尺度 N（产生具有较大多样性的样本）开始，以及从第二个最粗尺度 N-1（保留原始图像的全局结构）开始。在每种情况下，我们都进行了配对研究（显示了成对真假图像）和未配对研究（显示了或假或真的一张图像），通过 bootstrap 估计方差[14]。

生成的图像的多样性而降低。但是，即使更改了大型结构，我们生成的图像也很难与真实图像区分开（50%的分数意味着真实与伪造之间的完美混淆）。完整的测试图像集包含在 SM 中。

单个图像的 Frechet 初始距离 我们接下来将量化 SinGAN 捕获 x 的内部统计数据的能力。GAN 评估的常用指标是 Frechet 初始距离（FID）[23]，该距离可测量生成图像的深层特征的分布与真实图像的分布之间的偏差。但是，在我们的环境中，我们只有一个真实的图像，并且对其内部斑块统计非常感兴趣。因此，我们提出了 Single Image FID（SIFID）指标。而不是使用 Inception Network [49] 中最后一个汇聚层之后的激活向量（每个图像一个向量），我们使用卷积层输出的深层特征的内部分布，即第二个汇聚层之前（特征图中每个位置一个向量）。我们的 SIFID 是真实图像和生成的样本中这些特征的统计量之间的 FID。

如表 2 所示，平均 SIFID 在

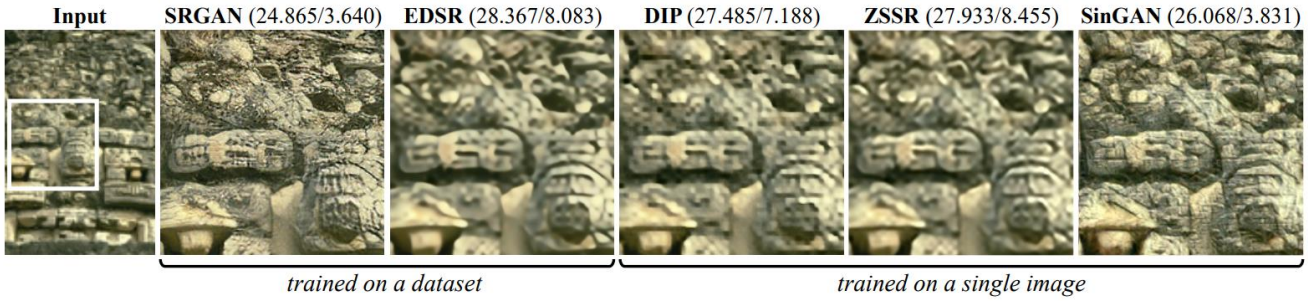

图 10：**超分辨率。** 当在低分辨率图像上训练 SinGAN 时，我们能够超分辨率。这是通过反复对图像进行升采样并将其馈送到 SinGAN 最好的尺度生成器来完成的。可以看出，SinGAN 的视觉质量优于 SOTA 内部方法 ZSSR [46]和 DIP [51]。它也比 EDSR [32]更好，并且与 SRGAN [30]相当，后者是针对大型收集数据集训练的外部方法（基于像素）。 括号中显示了相应的 PSNR 和 NIQE [40]。

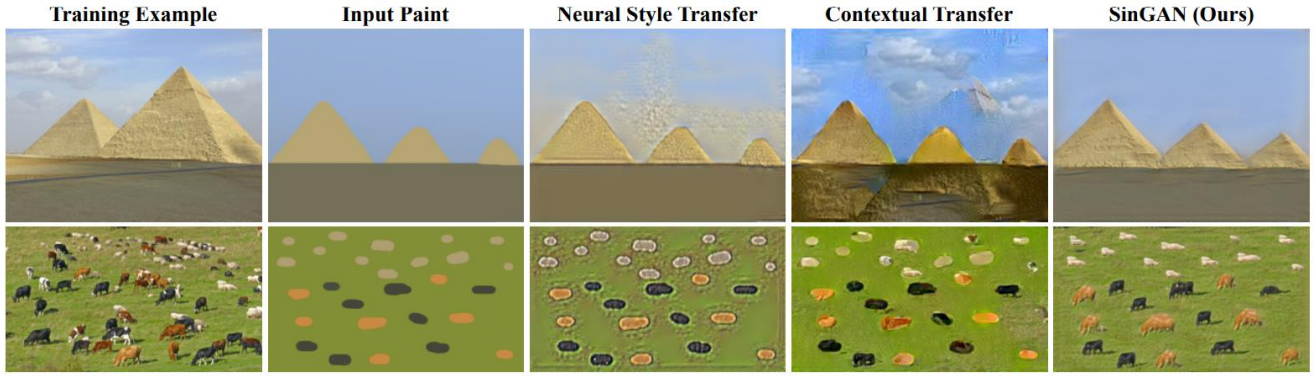

图 11：**画图。** 我们在目标图像上训练 SinGAN，并在测试时将下采样的涂料注入其中一个粗糙级别。我们生成的图像保留剪贴画的布局和一般结构，同时生成逼真的纹理和与训练图像匹配的精细细节。众所周知的样式转换方法[17，38]无法完成此任务。

1st Scale	SIFID	Survey	SIFID/AMT Correlation
N	0.09	paired	-0.55
		unpaired	-0.22
$N-1$	0.05	paired	-0.56
		unpaired	-0.34

表 2：**单映像 FID（SIFID）。**我们将 FID 算术调整为一张图像，并报告 50 幅图像的平均得分，包括完整图像（第一行）和第二个最粗糙的比例（第二行）。与 AMT 结果的相关性表明 SIFID 与人类排名高度吻合。

尺度为 N-1 的生成上比从尺度为 N 的生成上更低，这与用户研究结果一致。我们还报告了 SIFID 分数与伪造图像的混淆率之间的相关性。请注意，两者之间存在显著（反）相关性，这意味着较小的 SIFID 通常是较大混淆率的良好指标。由于 SIFID 是成对的度量（它在成对x_n，$\tilde{x}_n$上操作），因此对于成对的测试相关性更强。

4. 应用

我们探索将 SinGAN 用于许多图像处理任务。为此，我们在训练后使用我们的模型，而无需进行架构更改或进一步调整，并针对所有应用遵循相同的方法。这个想法是利用 SinGAN 只能产生与训练图像具有相同斑块分布的图像这一事实。因此，可以通过将图像（可能是下采样版本）以 n <N 的某个比例注入到生成金字塔中，并通过生成器进行馈送以使其斑块分布与训练图像的斑块分布相匹配来进行操纵。不同的尺度采样会导致不同的效果。我们考虑以下应用（有关更多结果和尺度采样效应，请参见 SM）。

超分辨率 *将输入图像的分辨率提高 s 倍。*我们在低分辨率（LR）图像上训练模型，重建损失权重为α = 100，并且金字塔比例因子对于某些$k \in \mathbb{N}$为$r = \sqrt[k]{s}$。由于小尺度倾向于在自然尺度上重复出现场景[18]，在测试时，我们将 LR 图像上采样 r 倍，并将其（连同噪声）注入最后一个生成器G_0。我们重复此 k 次以获得最终的高分辨率输出。示例结果如图 10 所示。可以看出，我们重建的视觉质量超过了最新的内部方法（基于语义）[51，46]以及旨在实现 PSNR 最大化的外部方法（基于像素）[32]。

	External methods		Internal methods		
	SRGAN	EDSR	DIP	ZSSR	SinGAN
RMSE	16.34	12.29	13.82	13.08	16.22
NIQE	3.41	6.50	6.35	7.13	3.71

表 3：**超分辨率评估。** 继[5]之后，我们报告了 BSD100 [35]的失真（RMSE）和感知质量（NIQE [40]，越低越好）。可以看出，SinGAN 的性能类似于 SRGAN [30]。

有趣的是，尽管仅暴露于单个图像，但它可与外部训练的 SRGAN 方法[30]相比。继[4]之后，我们在失真（RMSE）和感知质量（NIQE [40]）方面比较了 BSD100 数据集[35]的表 3 中的这 5 种方法，这是两个基本矛盾的要求[5]。可以看出，SinGAN 的感知质量非常出色。它的 NIQE 分数仅次于 SRGAN，其 RMSE 则稍好。

绘画成图 *将剪贴画转换为逼真的图像。*这是通过对剪贴画图像进行下采样并将其馈送到粗略尺度之一（例如 N-1 或 N-2）来完成的，从图 5 和 6 中可以看出。如图 2 和 11 所示，保留了绘画的整体结构，同时真实地生成了与原始图像相匹配的纹理和高频信息。就视觉质量而言，我们的方法优于样式转换方法[38，17]（图 11）。

协调 *现实地将粘贴的对象与背景图像融合在一起。*我们在背景图像上训练 SinGAN，并在测试时注入直接粘贴的合成文件的下采样版本。在这里，我们将生成的图像与原始背景结合在一起。从图 2 和图 13 可以看出，我们的模型对粘贴对象的纹理进行了裁剪以使其与背景相匹配，并且通常比[34]更好地保留其结构。尺度 2,3,4 通常会在保持对象的结构与传递背景纹理之间取得良好的平衡。

编辑 *产生无缝合成，其中图像区域已被复制并粘贴到其他位置。*同样，在这里，我们将组合的下采样版本注入到一个粗尺度中。然后，我们将 SinGAN 在编辑区域的输出与原始图像结合在一起。如图 2 和图 12 所示，SinGAN 可以重新生成精细的纹理并无缝缝合粘贴的零件，比 Photoshop 的 Content-Aware-Move 产生更好的结果。

单图像动画 *从单个输入图像创建具有逼真的对象运动的简短视频剪辑。*自然图像经常包含重复，这些重复会在同一动态物体上及时显示出不同的“快照” [55]（例如，一群鸟的图像揭示了一只鸟的所有机翼姿势）。使用 SinGAN，我们可以沿着对象在图像中所有外观的整体运动，从而从单个图像中合成运动。我们发现，对于许多类型的图像，可以通过在 z 空间中随机行走来实现逼真的效果，从所有生成比例的第一

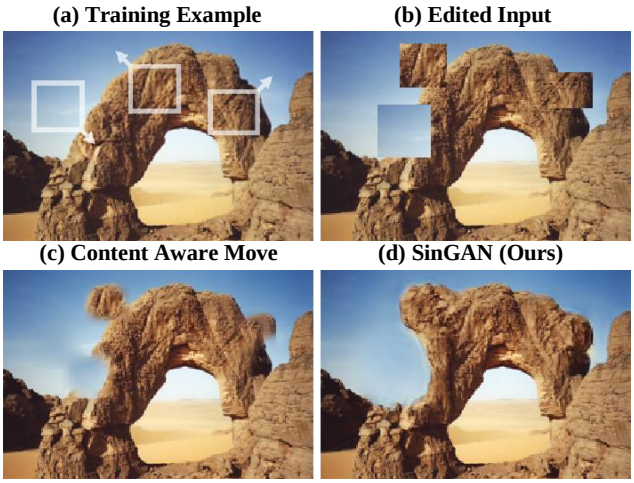

图 12：**编辑。** 我们从原始图像（a）复制并粘贴一些斑块，然后将编辑图像的下采样版本（b）输入到模型的中间级别（在（a）上进行了预训练）。在生成的图像（d）中，这些局部编辑被转换为连贯且逼真的结构。（c）与 Photoshop 内容感知动作比较。

图 13：**协调。** 我们的模型能够保留粘贴对象的结构，同时调整其外观和纹理。专用的协调方法[34]将对象与背景过度融合。

帧的 z^{rec} 开始。结果可访问在：https://youtu.be/xk8bWLZk4DU。

5. 结论

我们介绍了 SinGAN，这是一种新的无条件生成方法，可以从单个自然图像中学习。我们证明了它具有超越纹理的能力，并能生成自然逼真的图像的各种逼真的样本。与外部训练的生成方法相比，内部学习在语义多样性方面固有地受到限制。例如，如果训练图像只包含一条狗，则我们的模型将不会生成不同犬种的样本。不过，正如我们的实验所证明的，SinGAN 可以为各种图像处理任务提供非常强大的工具。

致谢 感谢 Idan Kligvasser 的宝贵见解。这项研究得到了以色列科学基金会（852/17 赠款）和奥伦多夫基金会的支持。

参考文献

[1] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. Surprising effectiveness of few-image unsupervised feature learning. arXiv preprint arXiv:1904.13132, 2019. 2

[2] Yuval Bahat and Michal Irani. Blind dehazing using internal patch recurrence. In 2016 IEEE International Conference on Computational Photography (ICCP), pages 1–9. IEEE, 2016. 1

[3] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learn-ing texture manifolds with the periodic spatial GAN. arXiv preprint arXiv:1705.06566, 2017. 2, 4

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on per-ceptual image super-resolution. In European Conference on Computer Vision Workshops, pages 334–355. Springer, 2018. 8

[5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition, pages 6228–6237, 2018. 8

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018. 1

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. arXiv preprint arXiv:1808.07371, 2018. 2

[8] Wengling Chen and James Hays. Sketchygan: towards di-verse and realistic sketch to image synthesis. In Proceed-ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9416–9425, 2018. 2

[9] Taeg Sang Cho, Moshe Butman, Shai Avidan, and William T Freeman. The patch transform and its applications to image editing. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. 1

[10] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3511– 3520, 2018. 2

[11] Tali Dekel, Tomer Michaeli, Michal Irani, and William T Freeman. Revealing and modifying non-local variations in a single image. ACM Transactions on Graphics (TOG), 34(6):227, 2015. 1

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009. 1

[13] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adver-sarial networks. In Advances in neural information process-ing systems, pages 1486–1494, 2015. 3

[14] Bradley Efron. Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics, pages 569–593. Springer, 1992. 6

[15] Gilad Freedman and Raanan Fattal. Image and video upscal-ing from local self-examples. ACM Transactions on Graph-ics (TOG), 30(2):12, 2011. 1

[16] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Tex-ture synthesis using convolutional neural networks. In Ad-vances in neural information processing systems, pages 262– 270, 2015. 2

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Im-age style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2414–2423, 2016. 7, 8

[18] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In 2009 IEEE 12th Interna-tional Conference on Computer Vision (ICCV), pages 349– 356. IEEE, 2009. 1, 7

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014. 1

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In Advances in Neural Information Pro-cessing Systems, pages 5767–5777, 2017. 4

[21] Kaiming He and Jian Sun. Statistics of patch offsets for im-age completion. In European Conference on Computer Vi-sion, pages 16–29. Springer, 2012. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceed-ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 4

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equi-librium. In Advances in Neural Information Processing Sys-tems, pages 6626–6637, 2017. 5, 6

[24] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5077–5086, 2017. 3

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co-variate shift. arXiv preprint arXiv:1502.03167, 2015. 4

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adver-sarial networks. arXiv preprint, 2017. 3, 4, 5

[27] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. Tex-ture synthesis with spatial generative adversarial networks. Workshop on Adversarial Training, NIPS, 2016. 2, 4

[28] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017. 3

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948, 2018. 3

[30] Christian Ledig, Lucas Theis, Ferenc Huszar,´ Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative ad-versarial network. In Proceedings of the IEEE conference on

computer vision and pattern recognition, pages 4681–4690, 2017. 1, 7, 8

[31] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In European Conference on Computer Vision, pages 702–716. Springer, 2016. 2, 3, 4

[32] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for sin-gle image super-resolution. In Proceedings of the IEEE Con-ference on Computer Vision and Pattern Recognition Work-shops, pages 136–144, 2017. 7

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, pages 3730–3738, 2015. 1

[34] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep painterly harmonization. arXiv preprint arXiv:1804.03189, 2018. 8

[35] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In null, page 416. IEEE, 2001. 4, 8

[36] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, 2015. 4

[37] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. In 2018 IEEE Win-ter Conference on Applications of Computer Vision (WACV), pages 1368–1376. IEEE, 2018. 1

[38] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In Proceedings of the European Conference on Com-puter Vision (ECCV), pages 768–783, 2018. 7, 8

[39] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In European Conference on Com-puter Vision, pages 783–798. Springer, 2014. 1

[40] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Mak-ing a completely blind image quality analyzer. IEEE Signal Processing Letters, 20(3):209–212, 2013. 7, 8

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE con-ference on computer vision and pattern recognition, pages 2536–2544, 2016. 1

[42] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Alvarez. Invertible conditional GANs for image edit-ing. arXiv preprint arXiv:1611.06355, 2016. 2

[43] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In Proceedings of the IEEE Con-ference on Computer Vision and Pattern Recognition, pages 5400–5409, 2017. 2

[44] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and remapping the “DNA” of a natural im-age. arXiv preprint arXiv: arXiv:1812.00231, 2018. 2

[45] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. InGAN: Capturing and Remapping the “DNA” of a Natu-ral Image. International Conference on Computer Vision (ICCV), 2019. 1, 2

[46] Assaf Shocher, Nadav Cohen, and Michal Irani. Zero-Shot Super-Resolution using Deep Internal Learning. In Proceed-ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3118–3126, 2018. 2, 7

[47] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, pages 746–760. Springer, 2012. 1

[48] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similar-ity. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8. IEEE, 2008. 1

[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015. 6

[50] Tal Tlusty, Tomer Michaeli, Tali Dekel, and Lihi Zelnik-Manor. Modifying non-local variations across multiple views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6276–6285, 2018. 1

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 7

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image syn-thesis and semantic manipulation with conditional GANs. arXiv preprint arXiv:1711.11585, 2017. 2, 3

[53] Xiaolong Wang and Abhinav Gupta. Generative image mod-eling using style and structure adversarial networks. 2016. 2

[54] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Tex-turegan: Controlling deep image synthesis with texture patches. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 2

[55] Xuemiao Xu, Liang Wan, Xiaopei Liu, Tien-Tsin Wong, Liansheng Wang, and Chi-Sing Leung. Animating animal motion from still. ACM Transactions on Graphics (TOG), 27(5):117, 2008. 8

[56] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with con-textual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5505– 5514, 2018. 2

[57] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing, 26(7):3142–3155, 2017. 4

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In European conference on computer vision, pages 649–666. Springer, 2016. 5

[59] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Tor-ralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Advances in neural information processing systems, pages 487–495, 2014. 4, 6

[60] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. arXiv preprint arXiv:1805.04487, 2018. 2

[61] Jun-Yan Zhu, Philipp Krahenb¨uhl,¨ Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In European Conference on Computer Vision (ECCV), pages 597–613. Springer, 2016. 2

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE International Con-ference on Computer Vision, 2017. 2, 4

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Dar-rell, Alexei A Efros, Oliver Wang, and Eli Shechtman. To-ward multimodal image-to-image translation. In Advances in Neural Information Processing Systems, pages 465–476, 2017. 4

[64] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In CVPR 2011, pages 977–984. IEEE, 2011. 1

[65] Maria Zontak, Inbar Mosseri, and Michal Irani. Separat-ing signal from noise using patch recurrence across scales. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1195–1202, 2013. 1