SinGAN: Learning a Generative Model from a Single Natural Image Single

ICCV 2019：best paper

Prior

- 图像中的一些特征信息
- 在某些图像去雾应用中，由于问题的ill-posed&under-constrain也就导致图像的解比较多，于是利用图像的先验我们可以缩小解空间，减小图像不确定性，增强结果，简化参数之类的。

Patch

- 是站在一个更宏观的角度去理解像素容器
- 如果你说你的图像是100*100(px)，那么我们把他分成10*10去划分，对每一个10*10的图像块就相当于论文中的image patch
- 在深度学习中patch可以和window一样去理解，就是patch一般专注于多个具有共同属性或是分布的像素群。
- 对于某些图像处理算法比如去模糊，去噪任务，在图像块上的操作会比整个图像的操作更容易。

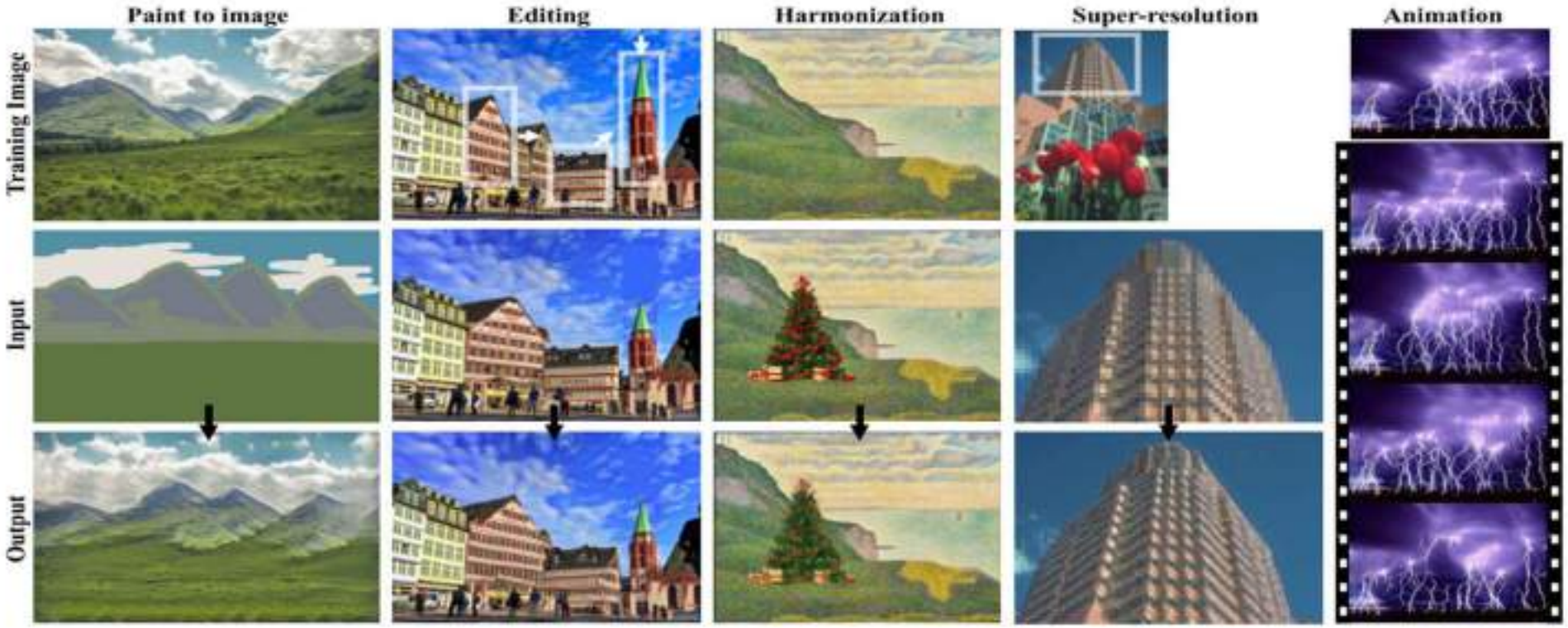
Motivation

- Capturing the distribution of highly diverse datasets with multiple object classes is hard.
- Solving many image manipulation tasks,requires conditioning the generation on another input signal or training the model for a specific task.
- Modeling the internal distribution of patches within a single natural image has been long recognized as a powerful prior in many computer vision tasks.

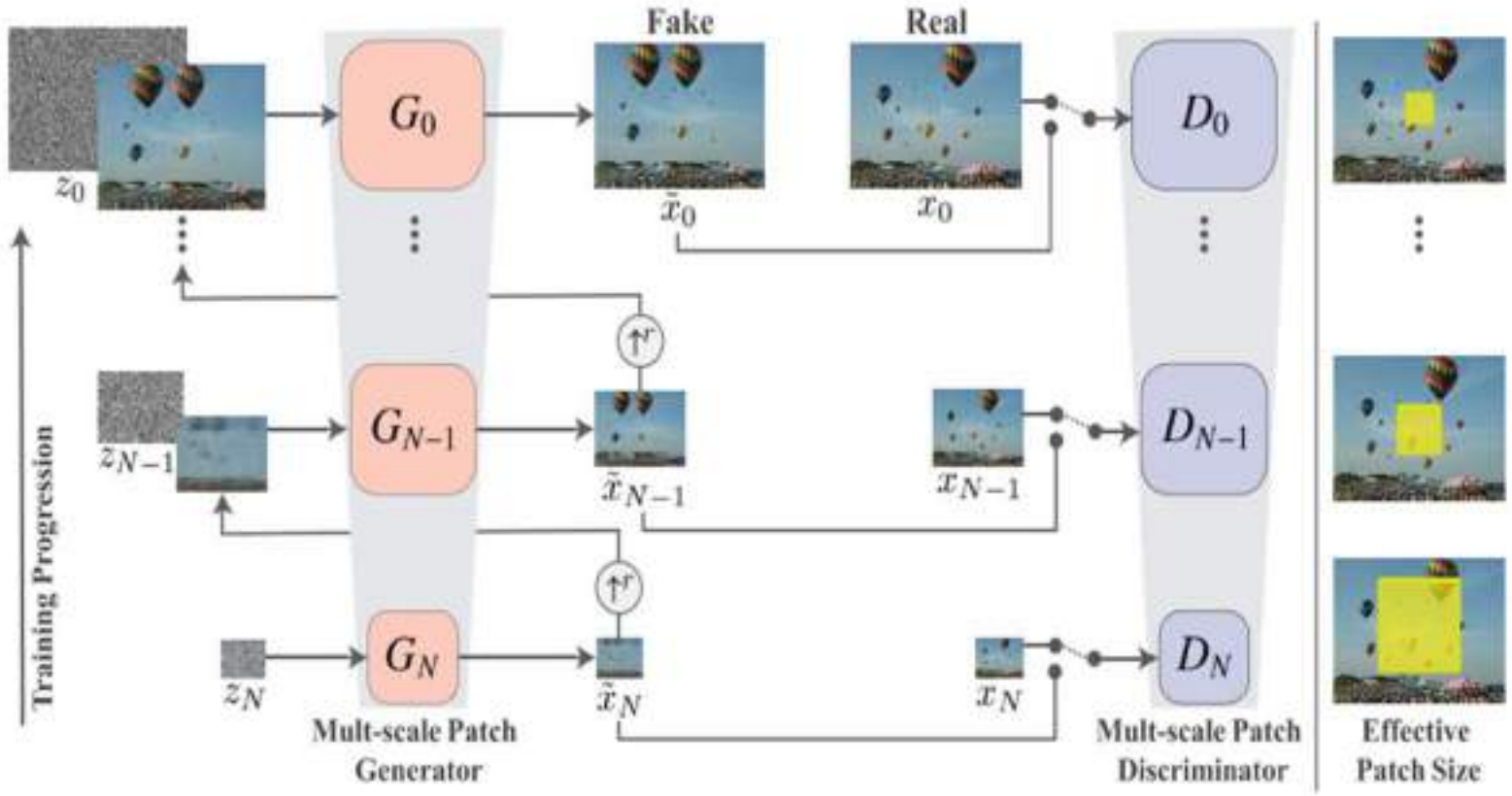
Innovation

- Single image.
- A simple unified learning framework.
- Without any additional information or further training beyond the original training image.

Result


Paint to image
Editing
Harmonization
Super-resolution
Animation
Training Image
Input
Output

Multi-scale architecture



Multi-scale architecture

- lowest level $\tilde{x}_N = G_N(z_N)$.
- other level:

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

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

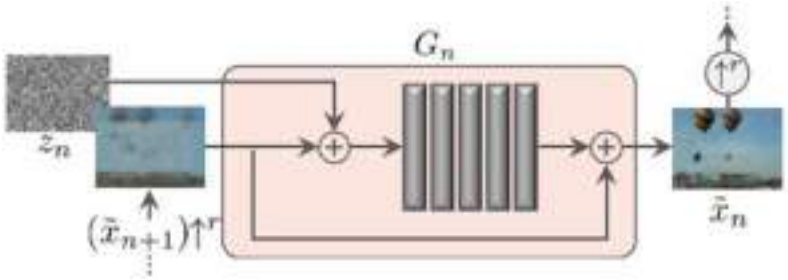


Figure 5: **Single scale generation.** At each scale n, the image from the previous scale, $\tilde{x}_{n+1}$, is upsampled and added to the input noise map, z_n. The result is fed into 5 conv layers, whose output is a residual image that is added back to $(\tilde{x}_{n+1})\uparrow^r$. This is the output $\tilde{x}_n$ of G_n.

- we can generate images of arbitrary size and aspect ratio at test time by changing the dimensions of the noise maps.
- 全卷积网络(FCN)，具备任意的尺寸和纵横比

Training

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

Adversarial loss:calculate the distribution distance between $\tilde{x}_N$ and x_N

- WGAN-GP loss increase training stability.

(whole image losss instead of patch loss)

- The architecture of D_n is the same as the net ψ_n within G_n

Reconstruction loss: insure produce

- D_n have the same architecture as G_n.

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

$$\text{and for } n = N, \text{we use } \mathcal{L}_{\text{rec}} = \|G_N(z^*) - x_N\|^2.$$

Results

- AMT perceptual study

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	$47.04\% \pm 0.8\%$

Table 1: **"Real/Fake" AMT test.** We report confusion rates for two generation processes: Starting from the coarsest scale N (producing samples with large diversity), and starting from the second coarsest scale $N-1$ (preserving the global structure of the original image). In each case, we performed both a paired study (real-vs.-fake image pairs are shown), and an unpaired one (either fake or real image is shown). The variance was estimated by bootstrap [14].

- A new single-image version of the FID

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

Table 2: **Single Image FID (SIFID).** We adapt the FID metric to a single image and report the average score for 50 images, for full generation (first row), and starting from the second coarsest scale (second row). Correlation with AMT results shows SIFID highly agrees with human ranking.

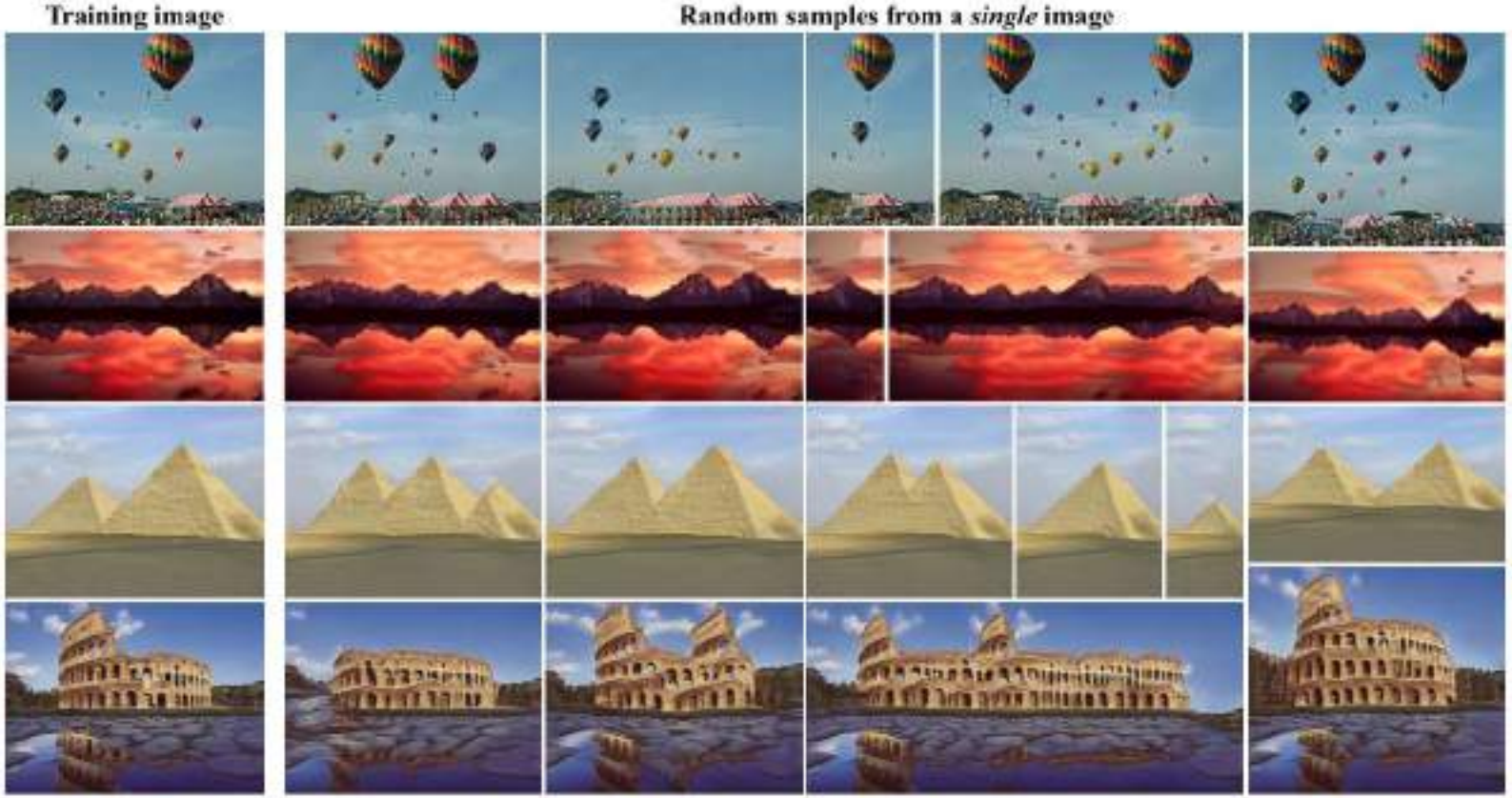


Figure 6: **Random image samples.** After training SinGAN on a single image, our model can generate realistic random image samples that depict new structures and object configurations, yet preserve the patch distribution of the training image. Because our model is fully convolutional, the generated images may have arbitrary sizes and aspect ratios. Note that our goal is not image retargeting – our image samples are random and optimized to maintain the patch statistics, rather than preserving salient objects. See SM for more results and qualitative comparison to image retargeting methods.

Effect of Scales During Training

- 小数字尺度：感受野比较小。
- 大数字尺度：捕捉到全局信息。

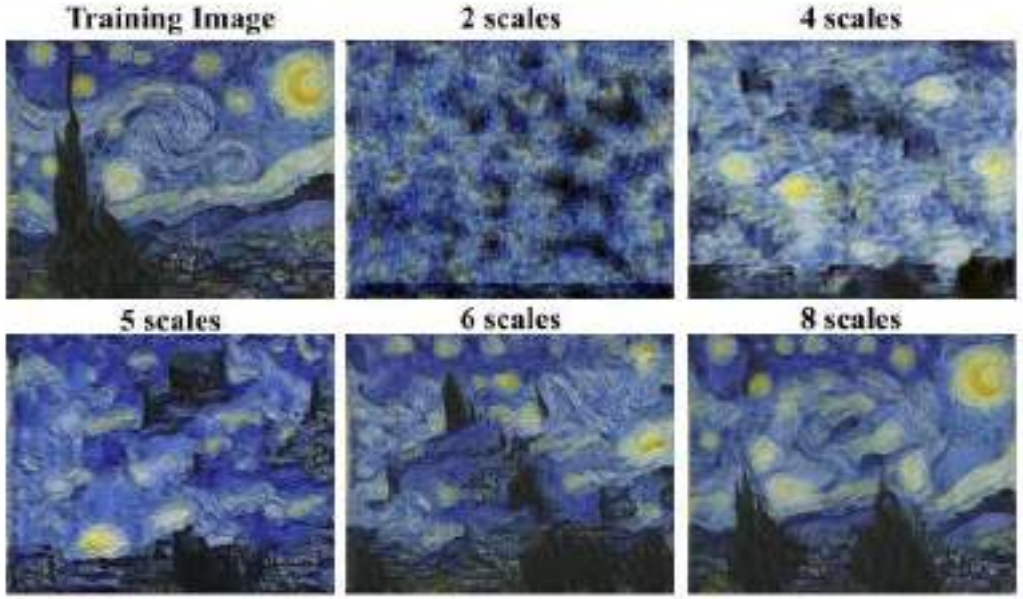


Figure 9: **The effect of training with a different number of scales.** The number of scales in SinGAN's architecture strongly influences the results. A model with a small number of scales only captures textures. As the number of scales increases, SinGAN manages to capture larger structures as well as the global arrangement of objects in the scene.

Effect of Scales at a Test Time

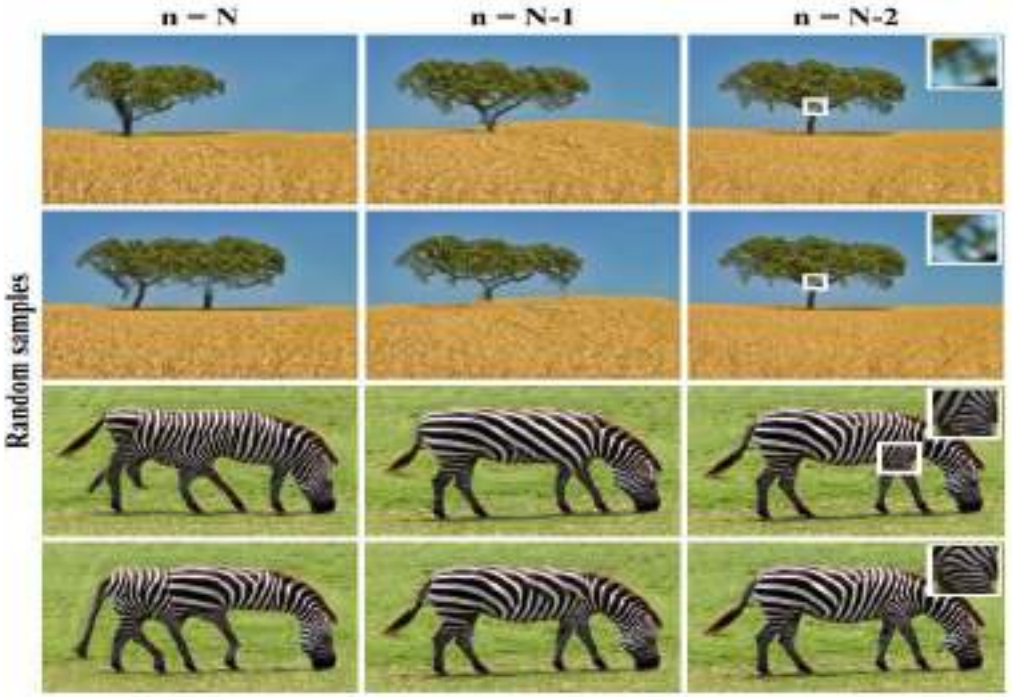


Figure 8: **Generation from different scales (at inference).** We show the effect of starting our hierarchical generation from a given level n. For our full generation scheme ($n = N$), the input at the coarsest level is random noise. For generation from a finer scale n, we plug in the downsampled original image, x_n, as input to that scale. This allows us to control the scale of the generated structures, *e.g.*, we can preserve the shape and pose of the Zebra and only change its stripe texture by starting the generation from $n = N - 1$.

- n=N：从粗糙尺度下进行生成样本，我们得到的结果是不自然的，
- n=N-1：从精细一点的尺度去生成样本那么这个整体的结构不会出现不自然情况，只会出现斑纹级别尺度的不同。

Application-Super-Resolution

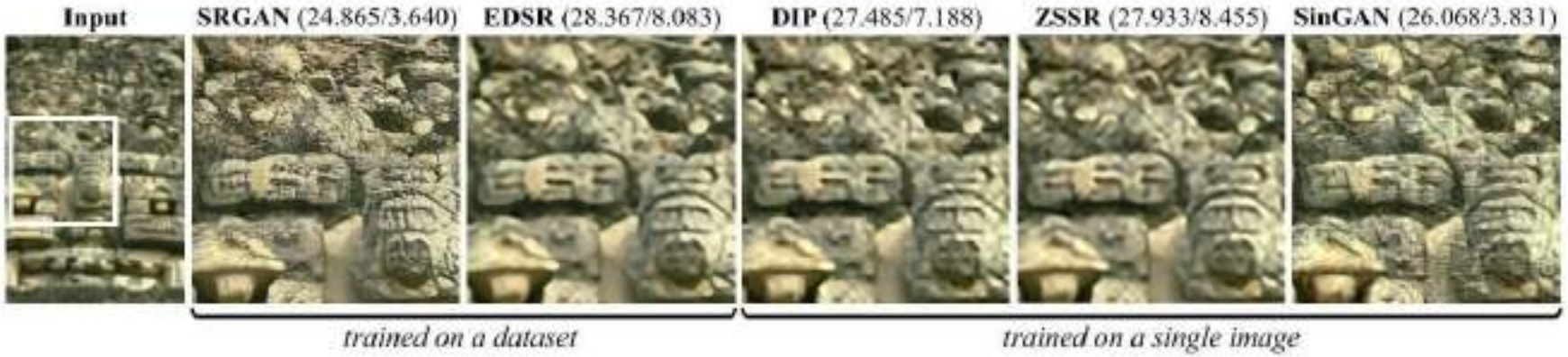


Figure 10: **Super-Resolution.** When SinGAN is trained on a low resolution image, we are able to super resolve. This is done by iteratively upsampling the image and feeding it to SinGAN's finest scale generator. As can be seen, SinGAN's visual quality is better than the SOTA internal methods ZSSR [46] and DIP [51]. It is also better than EDSR [32] and comparable to SRGAN [30], external methods trained on large collections. Corresponding PSNR and NIQE [40] are shown in parentheses.

- at test time， we upsample the LR image by a factor of r and inject it (together with noise) to the last generator, G0. $r = \sqrt[k]{s}$
- repeat this k times to obtain the final high-res output($k \in N$)

Paint-to-Image

Figure 11: **Paint-to-Image.** We train SinGAN on a target image and inject a downsampled version of the paint into one of the coarse levels at test time. Our generated images preserve the layout and general structure of the clipart while generating realistic texture and fine details that match the training image. Well-known style transfer methods [17, 38] fail in this task.

- 下采样目标图
- 喂给粗糙层

Figure 7: **High resolution image generation.** A random sample produced by our model, trained on the 243×1024 image (upper right corner); new global structures as well as fine details are realistically generated. See 4Mpix examples in SM.

Editing

- 下采样
- 添加进粗糙层

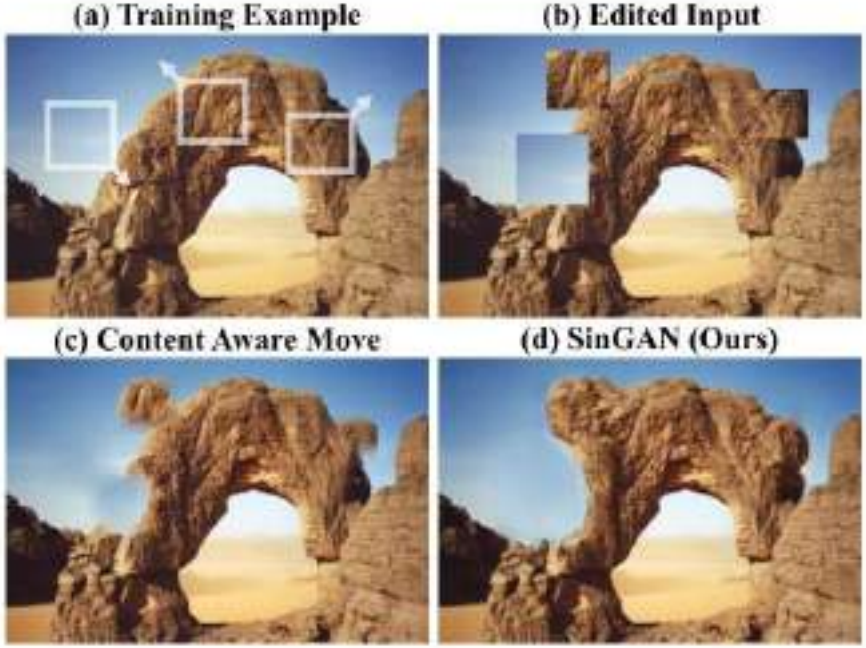


Figure 12: **Editing.** We copy and paste a few patches from the original image (a), and input a downsampled version of the edited image (b) to an intermediate level of our model (pretrained on (a)). In the generated image (d), these local edits are translated into coherent and photo-realistic structures. (c) comparison to Photoshop content aware move.

Single Image Animation

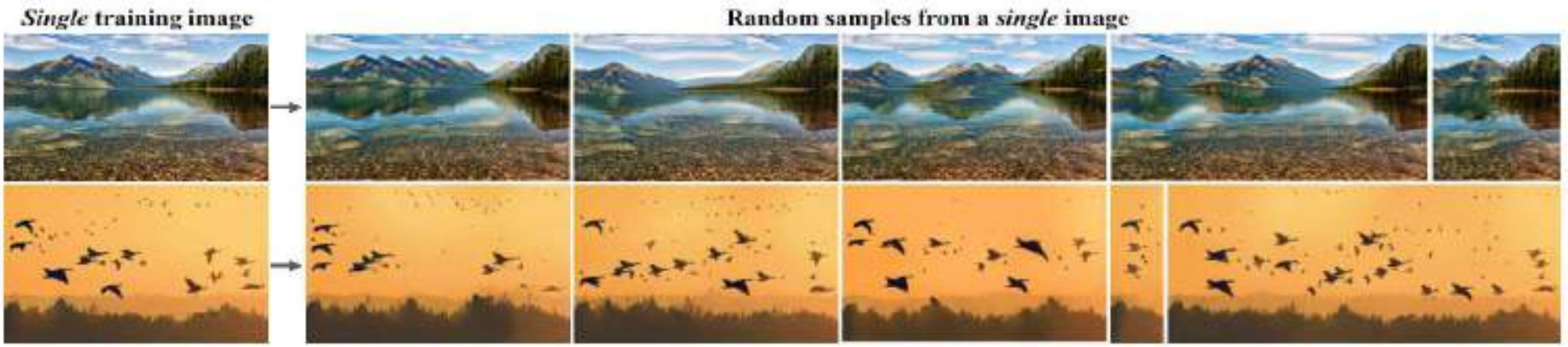


Figure 1: **Image generation learned from a single training image.** We propose *SinGAN*–a new unconditional generative model trained on a *single natural image*. Our model learns the image's patch statistics across multiple scales, using a dedicated multi-scale adversarial training scheme; it can then be used to generate new realistic image samples that preserve the original patch distribution while creating new object configurations and structures.

we can travel along the manifold of all appearances of the object in the image, thus synthesizing motion from a single image. We found that for many types of images, a realistic effect is achieved by a random walk in z-space.

Conclusion

- sinGAN 广泛应用于各种图像任务
- Internal learning：单一种类生成