가천대 회화·조소과 AI 특강

2021-06-24

조형래

SinGAN

ICCV 2019 Review [2] Best Paper SinGAN : Learning a Generative Model from a Single Natural Image

SinGAN: Learning a Generative Model from a Single Natural Image

Tamar Rott Shaham
Technion

Tali Dekel
Google Research

Tomer Michaeli
Technion

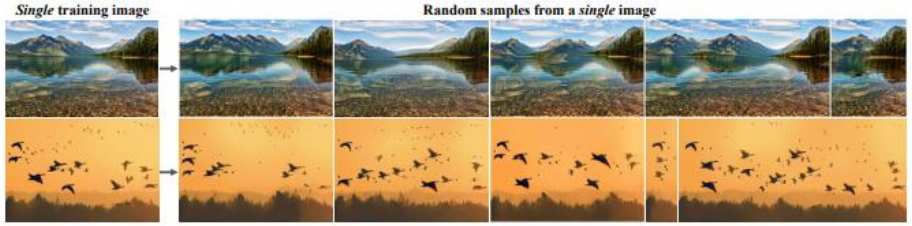


Figure 1: **Image generation learned from a single training image.** We propose *SinGAN*–a new unconditional generative model trained on a *single natural image*. Our model learns the image's patch statistics across multiple scales, using a dedicated multi-scale adversarial training scheme; it can then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

Abstract

We introduce SinGAN, *an unconditional generative model that can be learned from a single natural image. Our model is trained to capture the internal distribution of patches within the image, and is then able to generate high quality, diverse samples that carry the same visual content as the image. SinGAN contains a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of the image. This allows generating new samples of arbitrary size and aspect ratio, that have significant variability, yet maintain both the global structure and the fine textures of the training image. In contrast to previous single image GAN schemes, our approach is not limited to texture images, and is not conditional* (i.e. *it generates samples from noise). User studies confirm that the generated samples are commonly confused to be real images. We illustrate the utility of SinGAN in a wide range of image manipulation tasks.*

(*e.g.* ImageNet [12]), is still considered a major challenge and often requires conditioning the generation on another input signal [6] or training the model for a specific task (*e.g.* super-resolution [30], inpainting [41], retargeting [45]).

Here, we take the use of GANs into a new realm – *unconditional* generation learned from a *single natural image*. Specifically, we show that the internal statistics of patches within a single natural image typically carry enough information for learning a powerful generative model. *SinGAN*, our new single image generative model, allows us to deal with general natural images that contain complex structures and textures, without the need to rely on the existence of a database of images from the same class. This is achieved by a pyramid of fully convolutional light-weight GANs, each is responsible for capturing the distribution of patches at a different scale. Once trained, *SinGAN* can produce diverse high quality image samples (of arbitrary dimensions), which semantically resemble the training image, yet contain new object configurations and structures[1] (Fig. 1).

https://arxiv.org/pdf/1905.01164.pdf

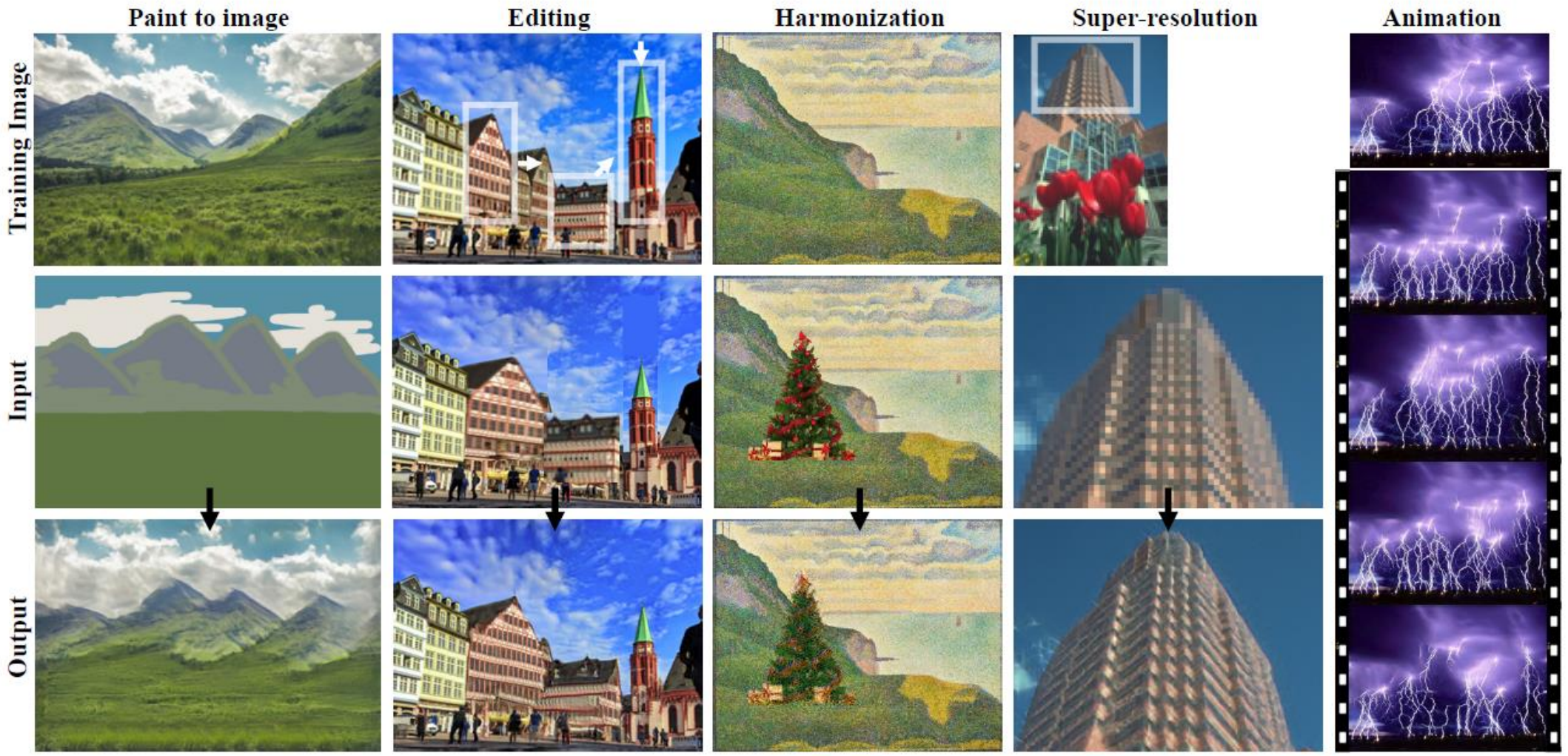
SinGAN 이전의 연구

- conditional natural single image translation
natural image 한장으로부터 생성하는 사전 연구의 경우 conditional GAN 이라서 image에서 image를 맵핑 해야 하고, 랜덤 샘플을 생성하기 어려운 한계가 있다.
<- SinGAN 프레임 워크의 경우 랜덤 노이즈에서 샘플을 생성하는 것이다. 따라서 다양한 이미지 **매니플레이션**(manipulation, 마술같은 교묘한 처리) 작업이 가능하다.

- unconditional texture single image generation
unconditional 한 방법으로 single 이미지를 학습하는 이전 연구에서는 texture 이미지에 국한되고 결과가 좋지 안았다.
SinGAN은 texture 이미지 생성 뿐만 아니라 다양한 자연적인 이미지에 적용에 좋은 성능을 보임

Figure 3: **SinGAN vs. Single Image Texture Generation.** Single image models for texture generation [3, 16] are not designed to deal with natural images. Our model can produce realistic image samples that consist of complex textures and non-reptititve global structures.

Image manipulation 사례



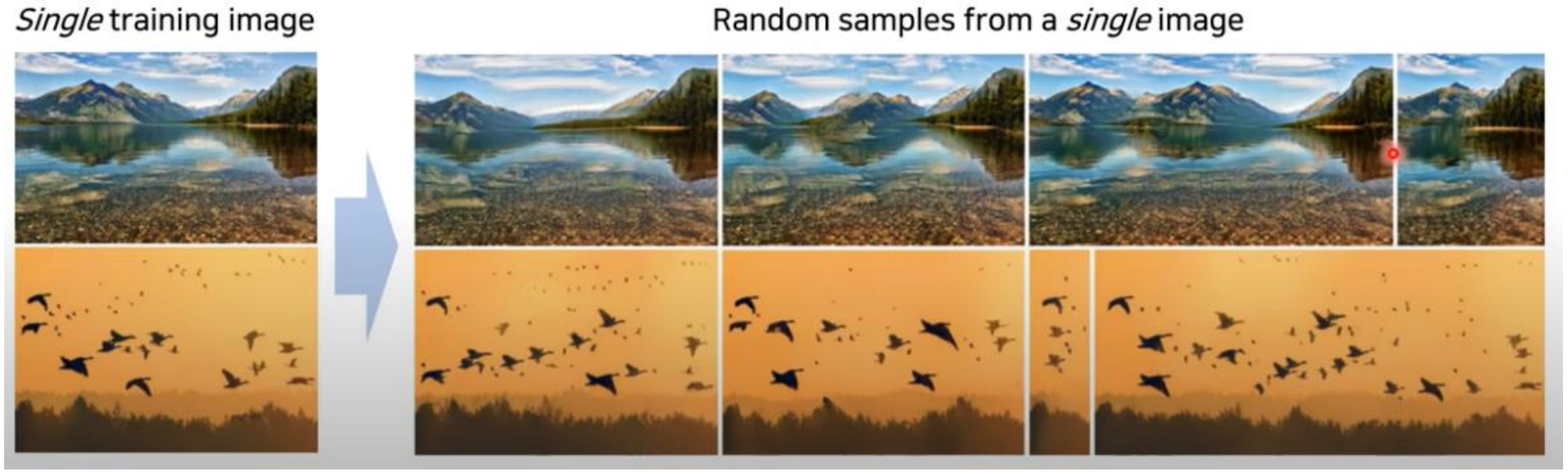
Introduction

적은 수의 image 또는 1장만 가지고 GAN을 학습 시키는 것
noise로부터 image를 생성하는 방식(unconditional)

fully-convolution GAN의 피라미드 구조이다. 이 구조를 통해이미지의 다양한 스케일에서 패치 분포를 학습한다.

장점: 새로운 샘플들을 생성할 수 있고, 학습 이미지의 전체적인 구조나 자세한 texture 들은 유지할 수 있다.

Unconditional하게, 즉 noise로부터 image를 생성하는 방식을 사용



Multi-scale architecture

unconditional generative model을 만드는 것이 목표
그러기 위해선
Global한 특징(Image의 배열과 Object들의 모양 등)과 Fine한 특징(Object의 디테일한 정보, Texture 등)을 모두 배울 수 있어야 한다.
이를 위해서
multi-scale GAN 구조를 사용: coarse-to-fine 하게 image를 생성하는 것

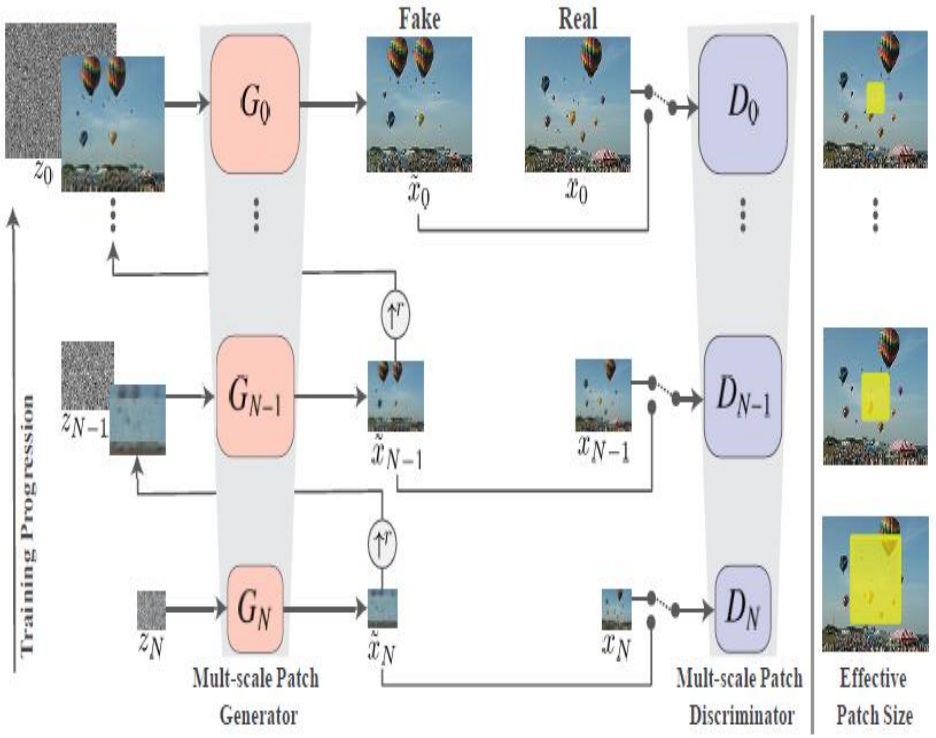
모델은 $\{G_0, \ldots, G_N\}$ 의 생성자의 피라미드 구조로 구성되어 있다.

이미지 또한 $\{x_0, \ldots, x_N\}$ 의 버전으로 되어 있다.

x_n은 이미지 x 의 다운샘플된 버전으로 downsample factor 에 의해 다운샘플 된다.

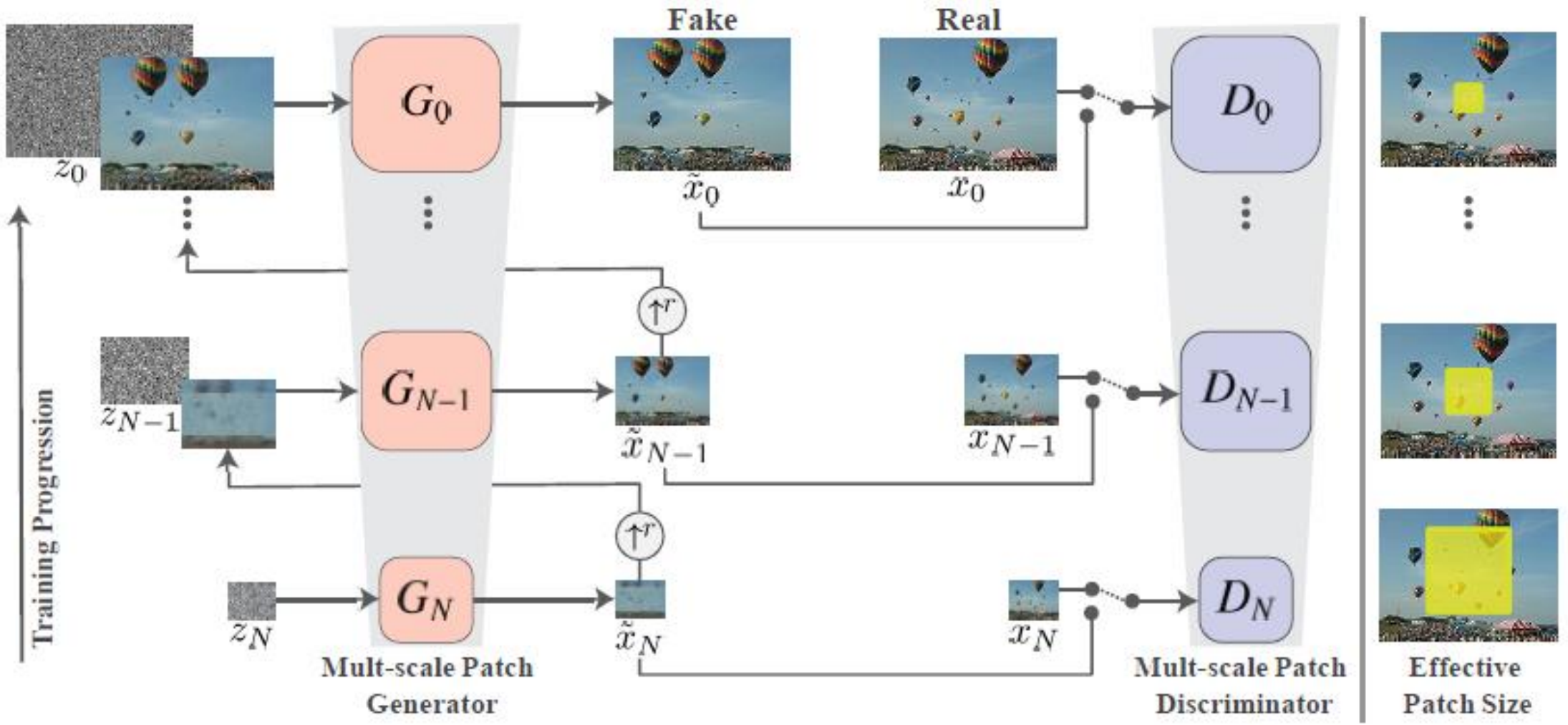
각 생성자 G_n은 이에 상응하는 x_n에 대한 이미지를 학습하고, 판별자 D_n는 생성된 이미지의 패치와 x_n 이미지로부터의 패치를 구분하도록 시도한다.

Multi-scale architecture



- xo 는 training image 이고, 아래로 한 단계씩 내려갈수록 r 배(r>1)씩 downsampling
- 각 단계마다 Generator는 noise와 이전 단계에서 생성된 결과 image를 input으로 하여 image를 생성하고,
- 그 단계의 Discriminator는 downsampling된 GT와 생성된 image를 구분하도록 학습
  - 예외로 제일 첫 단계(맨 밑)에서는 noise만 이용하여 image를 생성한다.
  - 앞 단계에서는 downsampling된 GT를 생성하도록 학습을 하다 보니 coarse한, global한 특징에 집중을 하여 생성을 하게 되고, 위로 갈수록 fine한 영역에 집중하여 생성
  - Generator 구조가 kernel 수만 다르고 연산자들은 같다 보니 동일한 receptive field를 갖게 되고, 생성하는 image의 크기만 다르다 보니 그림의 맨 오른쪽 부분처럼 **Effective Patch Size** 가 달라지면서 coarse-to-fine 하게 학습이 된다.

SinGAN's multi-scale pipeline



coarse-to-fine 학습 (전체에서 세밀한 부분으로)

이미지의 전반적인 구조를 먼저 학습하고 점점 세밀한 구조를 학습한다.

가장 처음은 전체적인 구조 (coarsest scale)에서 시작한다.

coarsest scale에서는 G_N이 z_N가우시안 노이즈를 이미지 샘플 $\tilde{x}_N$으로 맵핑한다.

더 세밀한 스케일의 생성자 G_n $(n < N)$은 이전의 스케일에서 생성되지 않은 디테일이 들어감

각 생성자 G_n은 이전의 더 거친 스케일의 이미지 버전을 업스케일 하려고 시도한다.

Generator 구조: Single Scale Generation

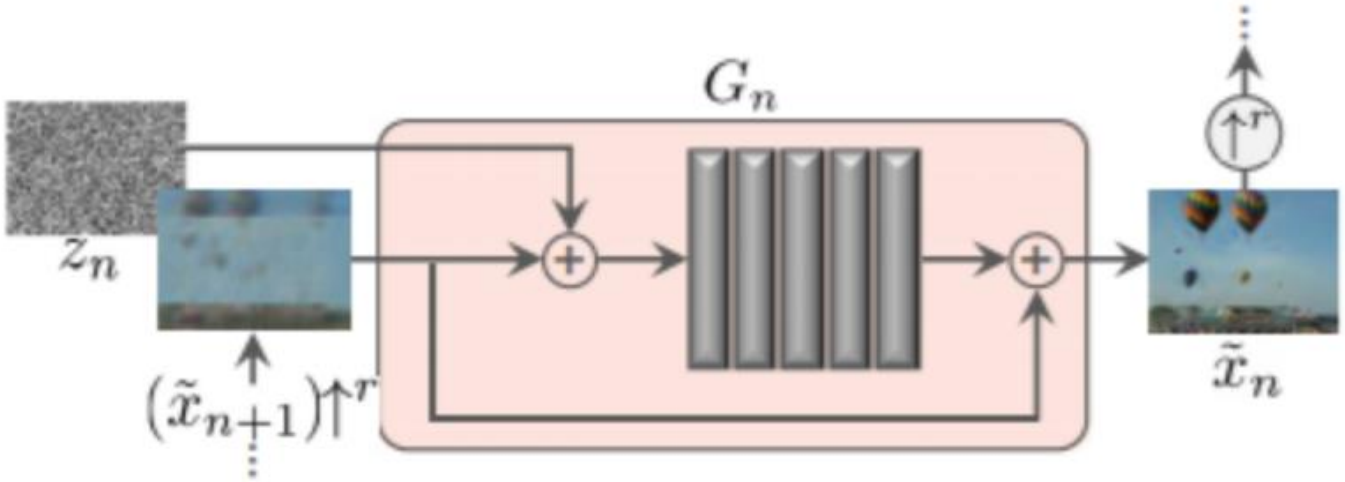


Figure 5: **Single scale generation.** At each scale n, the image from the previous scale, $\tilde{x}_{n+1}$, is upsampled and added to the input noise map, z_n. The result is fed into 5 conv layers, whose output is a residual image that is added back to $(\tilde{x}_{n+1})\uparrow^r$. This is the output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

생성자 오퍼레이션:

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

생성자는 동시에
1.업스케일 하면서
2.생성한다.

- ψ_n는 Conv(3x3)-BatchNorm-LeakyReLU로 구성된 5개의 컨볼루션 블록으로 이루어진 fully convolutional net이다.

- residual 구조를 이용하여 이미지를 생성

Training Loss

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

Adversarial loss와 Reconstruction loss로 구성
Adversarial loss는 WGAN-GP Loss를 사용

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

Reconstruction loss는 생성한 이미지와 원래의 이미지 간의 차이를 줄이는 방향으로 학습하기 위함이다.

- Reconstruction loss는 Generator가 생성한 image와 그 단계의 GT(downsampled) image간의 pixel간의 차이를 줄이는 방향으로 학습하기 위해 squred loss를 사용

- 학습을 위해서 각 단계의 주입되는 noise 셋팅은, 가장 첫 단계인 N 단계에만 고정된 noise를 주입하고, 나머지 단계에서는 noise를 주입하지 않았다. 즉 image의 pixel 차이를 줄이는 것에만 집중

실험평가; 정량적 평가

정량적인 평가]
1) Amazon Mechanical Turk(AMT)를 이용
사람들 투표를 통해 image가 Fake인지 Real인지를 구분
2) 기존에 GAN 연구에서 자주 사용되던 Frechet Inception Distance(FID) 를 Single Image에 맞게 변형한 Single Image FID(SIFID) 지표를 고안하여 비교

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

AMT perceptual study

2가지 실험 환경에서user study를 수행
1) 첫번째 실험 환경
Real image와 SinGAN을 통해 생성한 Fake image를 둘 다 보여주고 어느 쪽이 Fake인지 맞히는 **Paired** 실험
2) 두번째 실험 환경
Real image 혹은 Fake image 한 장만 보여준 뒤 얼마나 헷갈렸는지를 보는 **Unpaired** 실험

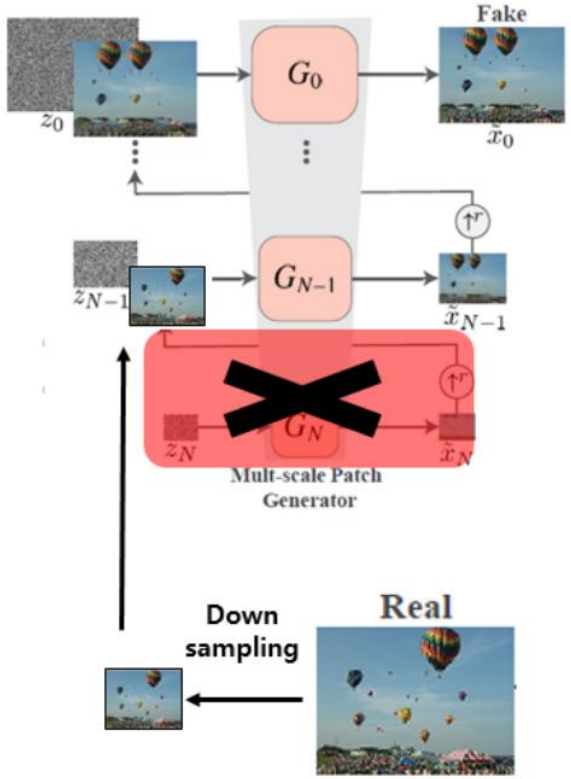
각각 실험당 1초의 시간이 주어지고, 한 명의 Worker당 50장의 image를 보여준다.
coarsest한(N) scale로부터 inferenc를 할 때와, 그 위의 N-1 scale에서 inferenc를 할 때의 결과를 비교하였다.

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	$47.04\% \pm 0.8\%$

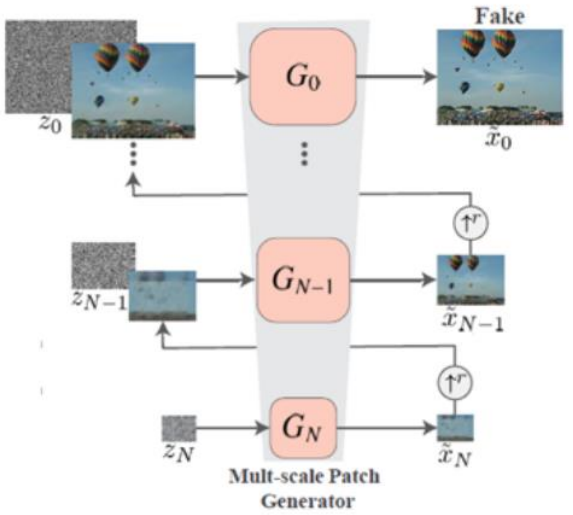
Table 1: **"Real/Fake" AMT test.** We report confusion rates for two generation processes: Starting from the coarsest scale N (producing samples with large diversity), and starting from the second coarsest scale $N-1$ (preserving the global structure of the original image). In each case, we performed 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].

AMT perceptual study

N scale에서 inference를 하는 것은 noise로부터 생성을 하는 것을 의미하고,
N-1 scale에서 inference를 하는 것은 input image를 downsampling한 뒤 N-1 번째 Generator의 input으로 넣어주는 방식을 의미



Result of scale N-1



Result of scale N

Training & Inference image

[Generation from different scales]

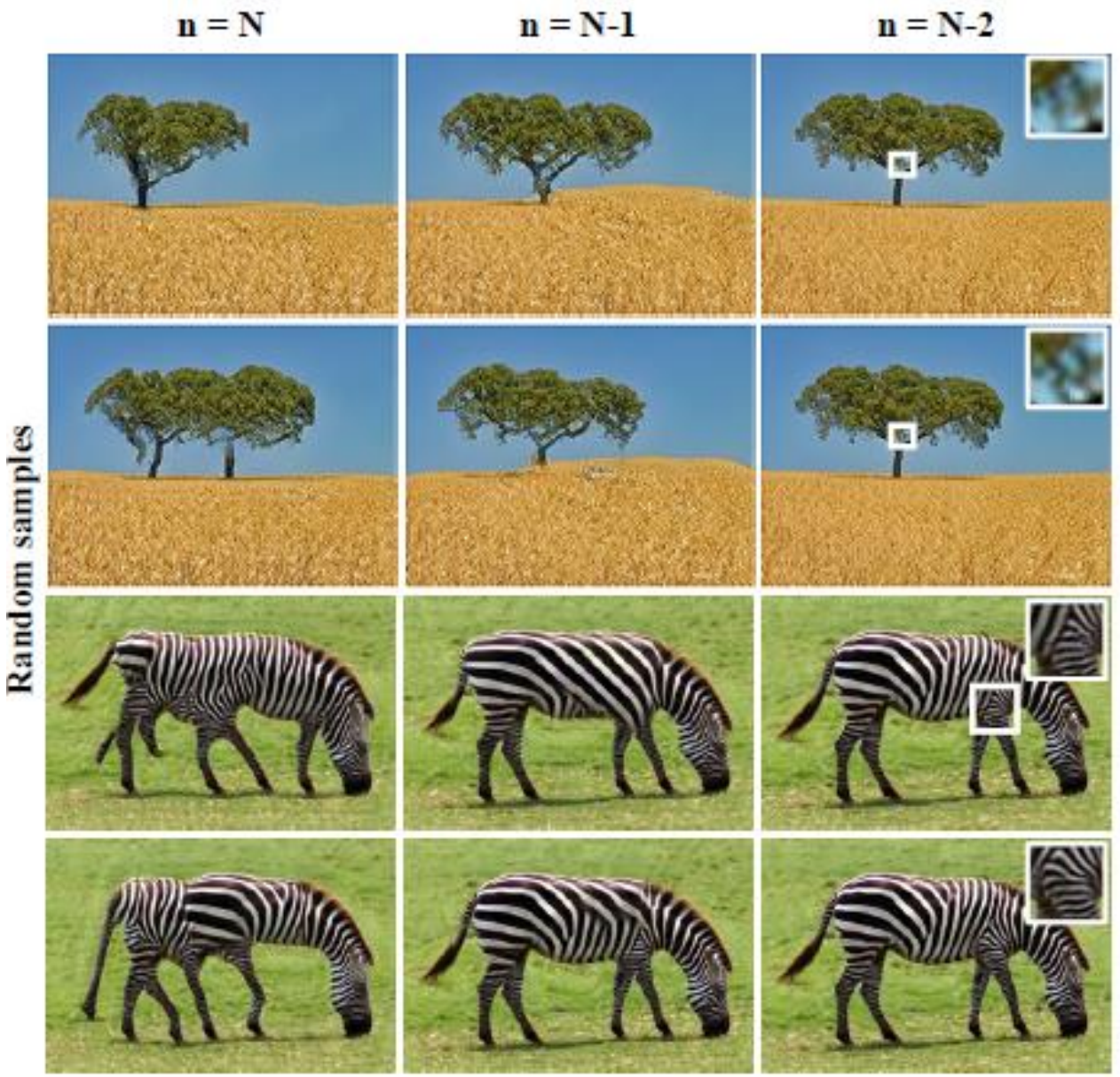
AMT perceptual study

[Generation from different scales 결과 예시]

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].

[AMT perceptual study 결과]



Single Image Super Resolution

Super Resolution 실험을 위해 reconstruction loss에 100의 weight를 주고 low-resolution image 한 장으로 학습을 시킨 뒤, test시에는 upsampling한 image를 제일 마지막 Generator G0에 noise와 함께 입력을 해준다.

High-resolution outpu을 얻기 위해 이러한 과정을 k 번 반복하였고, 결과 비교를 위해 Internal metho인 Deep Image Prior(DIP), Zero-Shot Super Resolution(ZSSR) 방식, 학습 데이터를 많이 필요로 하는 external metho인 SRGAN, EDSR 등과 결과를 비교

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.

Single Image Super Resolution

	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

Table 3: **Super-Resolution evaluation.** Following [5], we report distortion (RMSE) and perceptual quality (NIQE [40], lower is better) on BSD100 [35]. As can be seen, SinGAN's performance is similar to that of SRGAN [30].

distortion quality인 RMSE(낮을수록 좋음)는 높지만 perceptual quality인 NIQE(낮을수록 좋음)은 다른 internal method들에 비해 낮은 것을 확인할 수 있고,
학습 데이터를 필요로 하는 SRGAN과 비슷한 perceptual quality를 보임을 확인할 수 있습니다.

생성된 결과 이미지는 1장의 이미지로 Super Resolution을 나타냄

Paint-to-Image Style Transfer

Paint image로부터 Image로 생성해내는 Style Transfer 실험]
input으로 paint image를 scale N-1 혹은 N-2 에 downsampling하여 넣어주면
학습에 사용한 image의 style을 가진 image로 생성

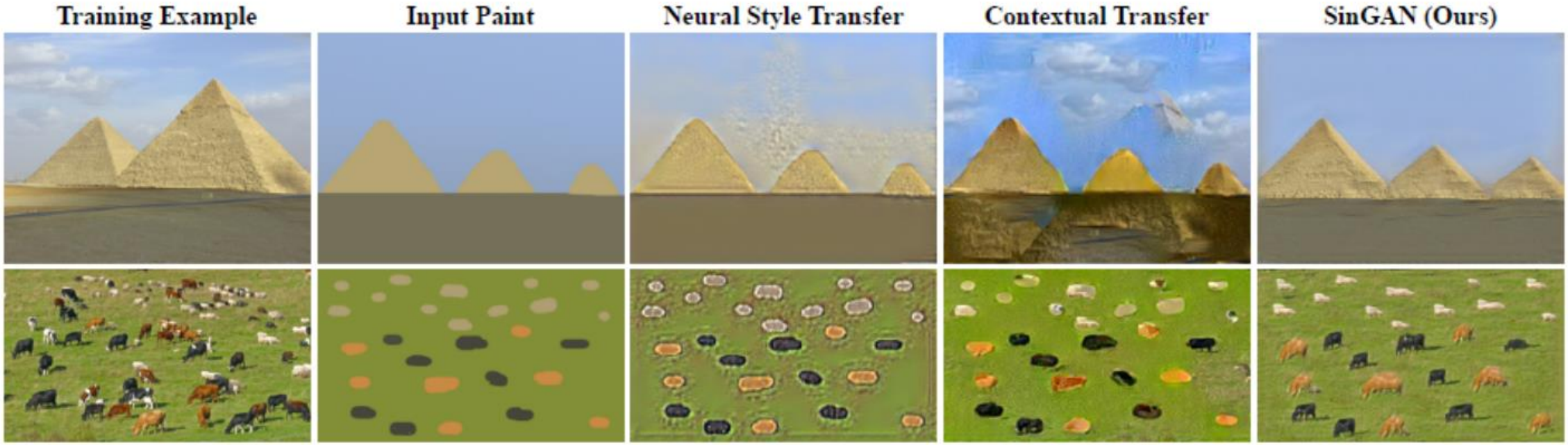


Figure 11: **Paint-to-Image.** We train SinGAN on a target image and inject a downsampled version of the paint into one of the coarse levels at test time. Our generated images preserve the layout and general structure of the clipart while generating realistic texture and fine details that match the training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

Harmonization

Harmonization task에 SinGAN을 적용한 결과]
image에 다른 object를 삽입하였을 때, 주변 배경과 조화를 이루며 object를 변형시켜 보여줌

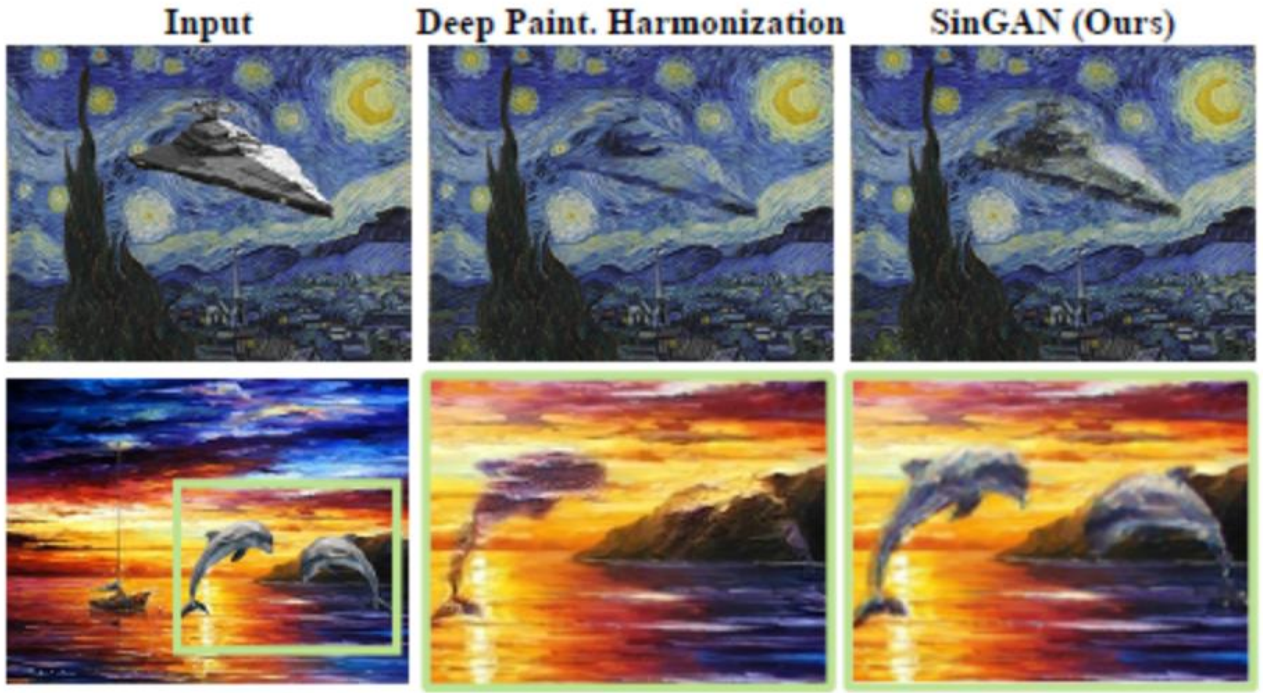


Figure 13: **Harmonization.** Our model is able to preserve the structure of the pasted object, while adjusting its appearance and texture. The dedicated harmonization method [34] overly blends the object with the background.

Editing

Image에서 일부 영역들을
복사+붙여넣기 하여
편집을 했을 때, 편집된
image를 주변에
자연스럽게 어우러지도록
만들어주는 Editing]
포토샵에 들어있는 기능인
Content Aware Move
기능보다 더 자연스러운
결과를 보여주는 것을 확인

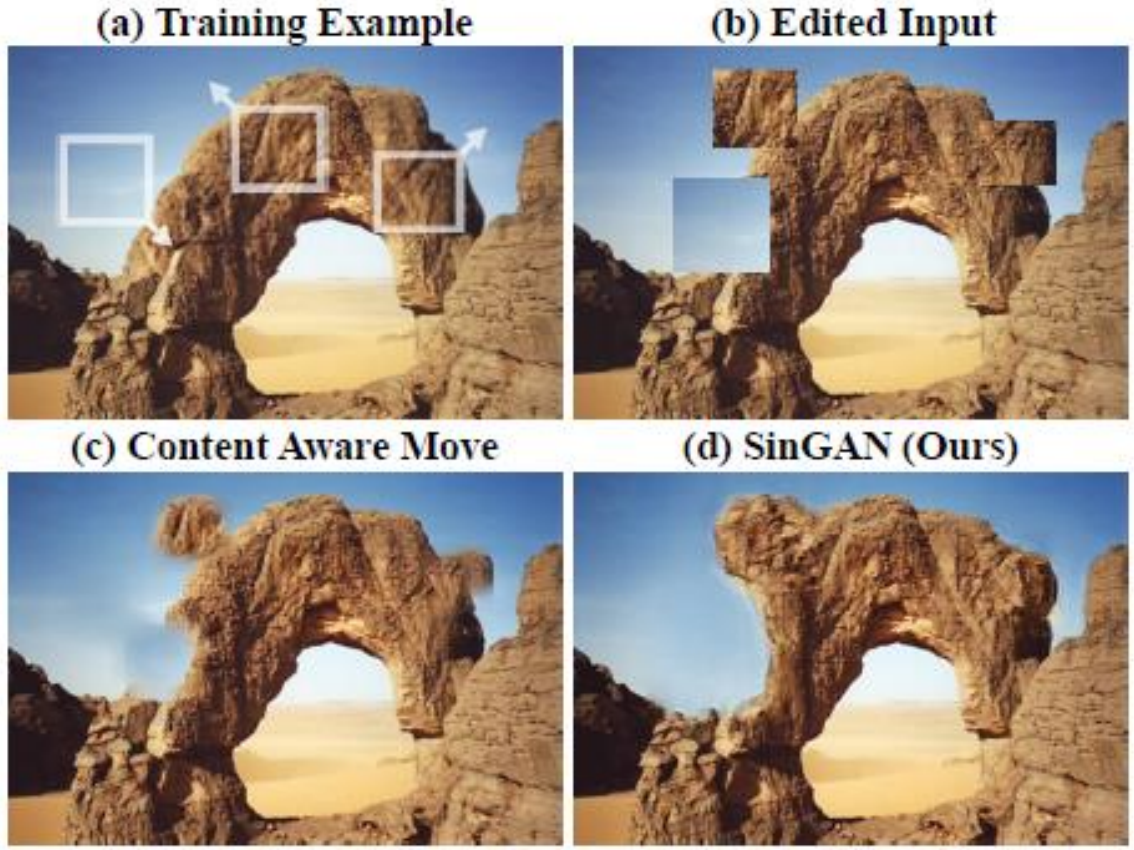


Figure 12: **Editing.** We copy and paste a few patches from the original image (a), and input a downsampled version of the edited image (b) to an intermediate level of our model (pretrained on (a)). In the generated image (d), these local edits are translated into coherent and photo-realistic structures.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

Single Image Animation

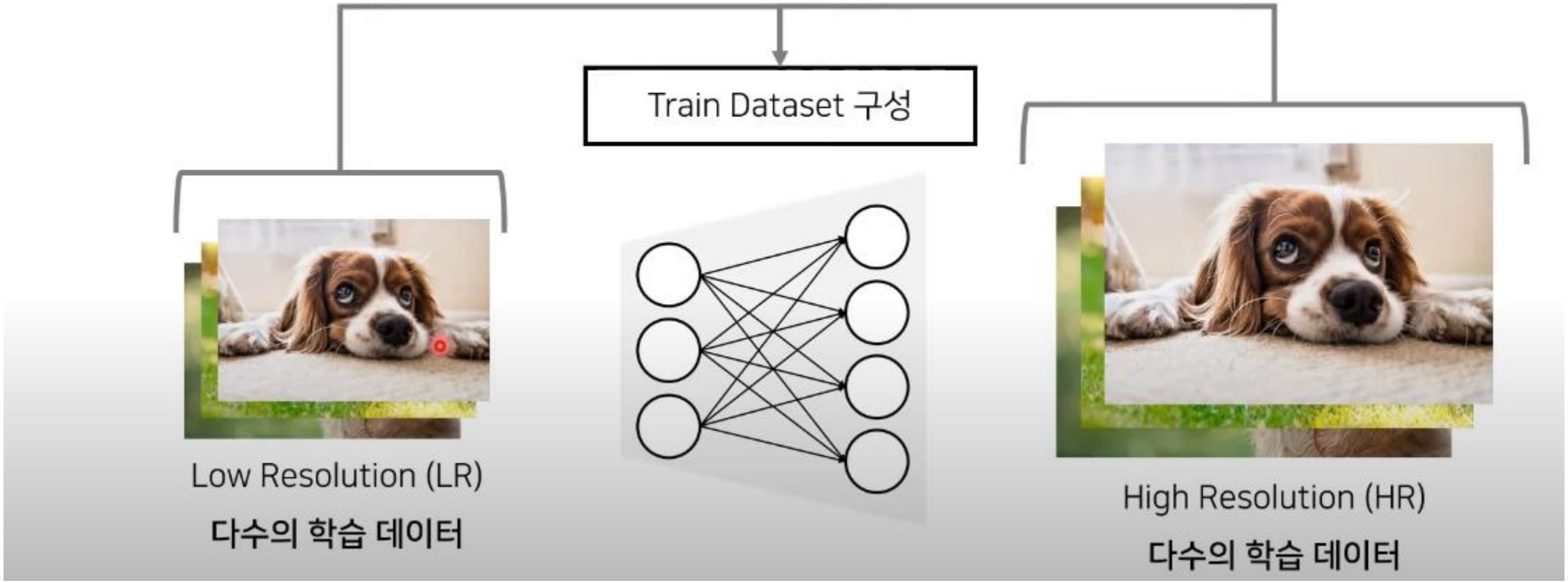
한 장의 Image로부터
짧은 video clip을 생성

https://www.youtube.com/watch?v=xk8bWLZk4DU

Externally Trained Network(Supervised SISR)

- **학습 시기**: 다수의 HR-LR 쌍에 대하여 학습을 진행합니다.

$$\mathbf{I}_{LR}^{\mathbf{k}} = (\mathbf{I}_{HR} * \mathbf{k}) \downarrow_s + \mathbf{n}$$



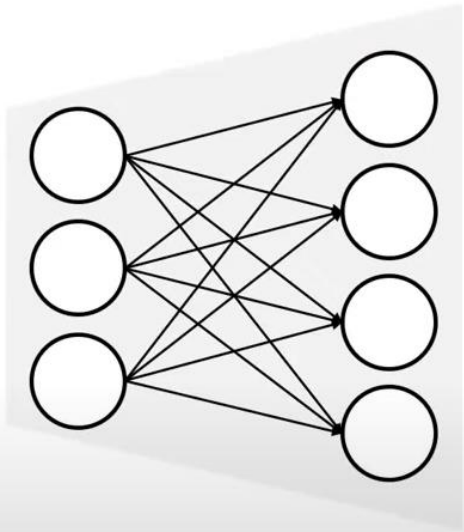
K;커널, s;다운sampling, n(노이즈)

Externally Trained Network(Supervised SISR)

- **테스트 시기**: 다수의 데이터로 학습된 정보를 토대로 현재 테스트 데이터에 대한 고해상도 결과를 예측합니다.

Low Resolution (LR)

테스트 데이터



High Resolution (HR)

예측 결과

SINGAN 실습

https://github.com/artjow/-AI-/blob/main/ART/SinGAN.ipynb