SinGAN: Learning a Generative Model from a Single Natural Image

Best paper of ICCV 2019

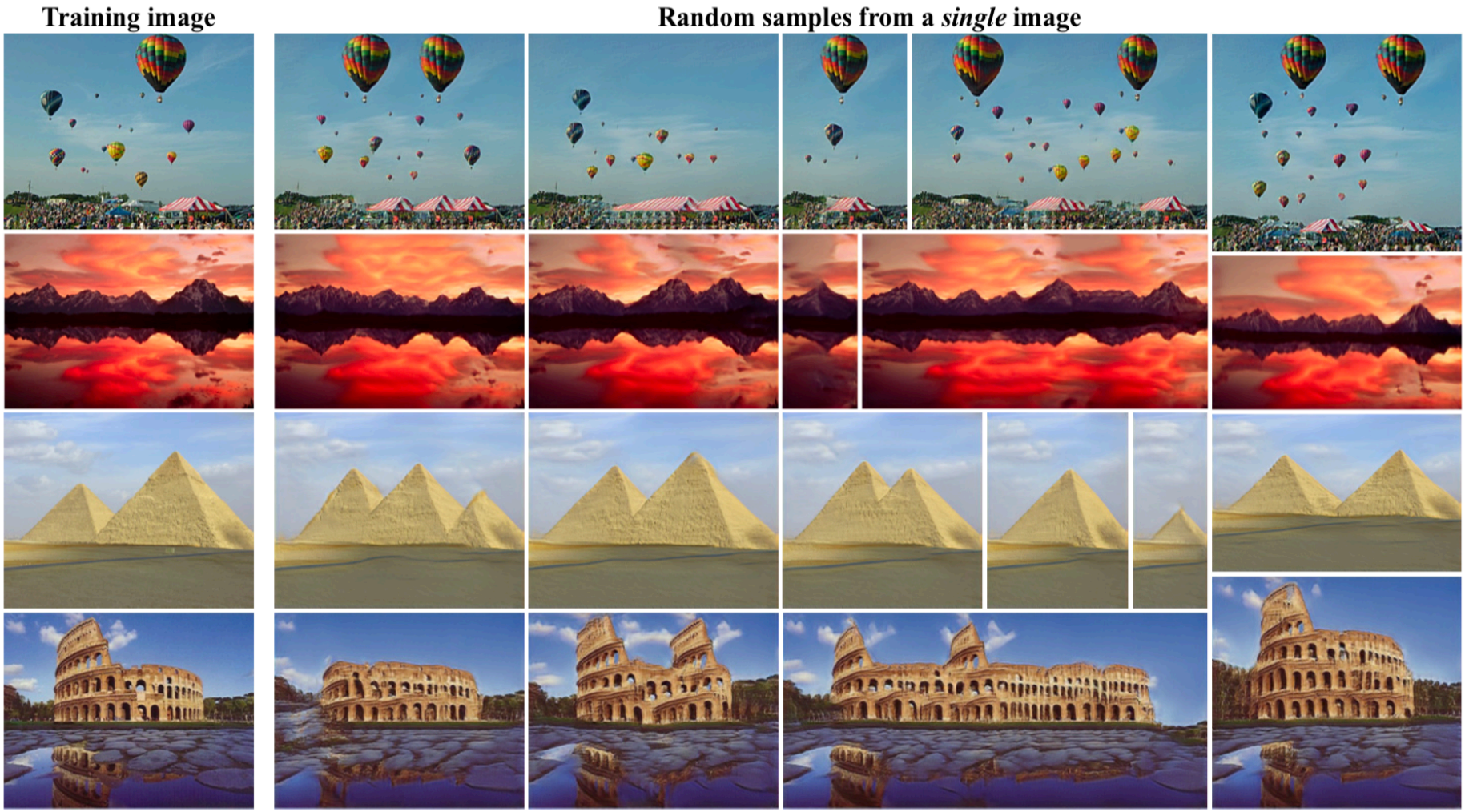
Bochan Kim
kmc2048@gmail.com
2019. 11. 04 Mon

SinGAN: Learning a (Unconditional) Generative Model from a Single Natural Image (For General Image Manipulation)

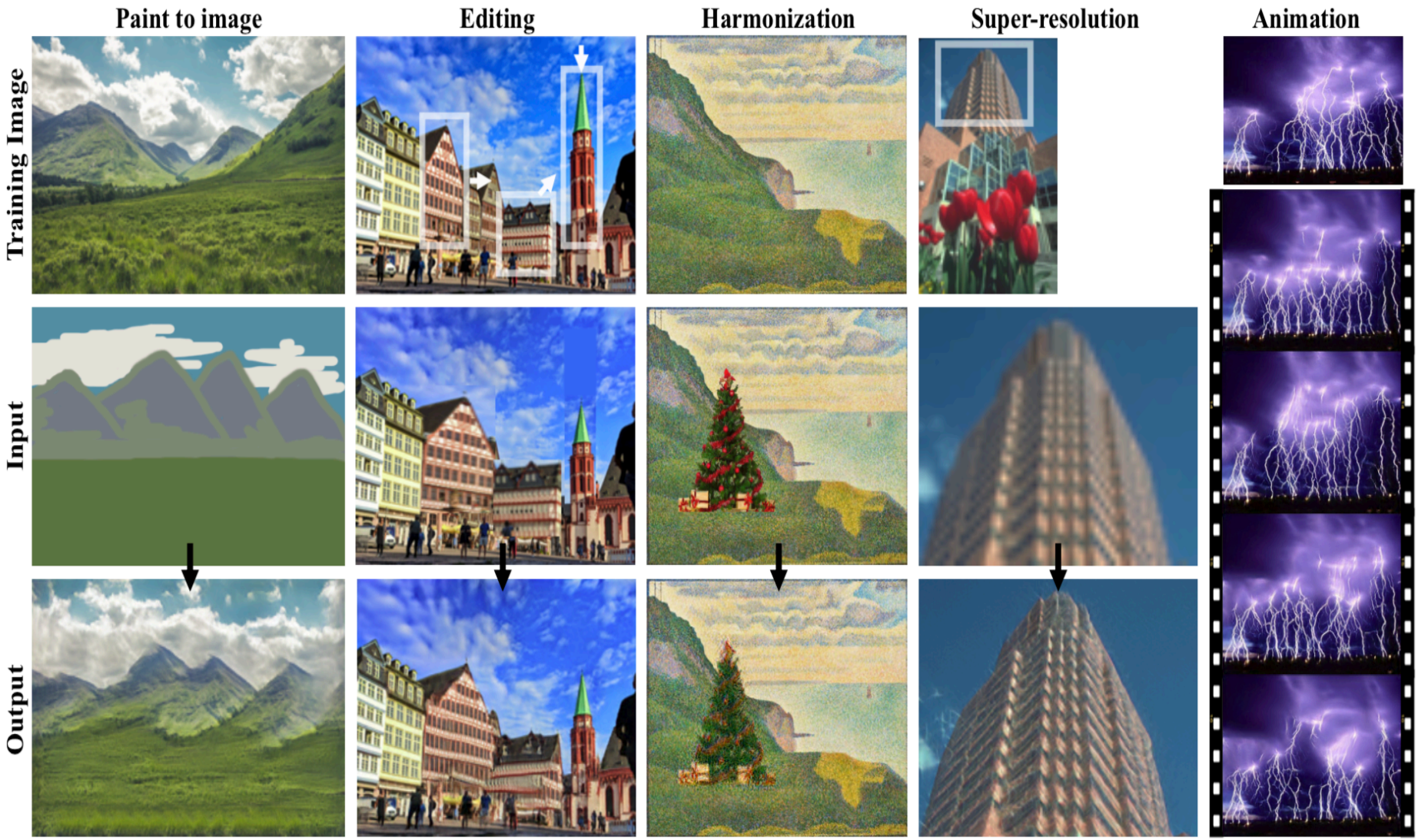
Best paper of ICCV 2019

Bochan Kim
kmc2048@gmail.com
2019. 11. 04 Mon

Results



Results


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

1. Unconditional & from Single Image

2. Model Architecture

3. Training Methods and Details

4. Results and Evaluations

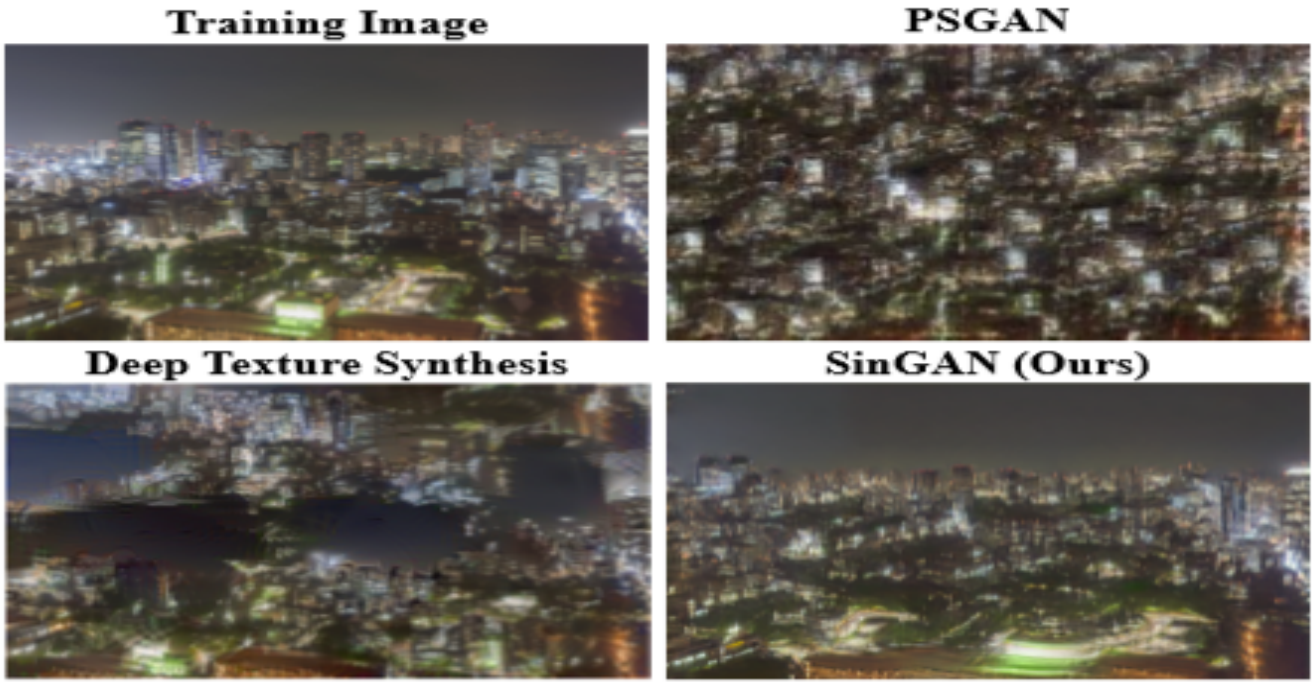
5. Various Applications

6. Conclusions

Unconditional Generative Model

- Generate Sample totally from random noise

- Does not restrained for texture generation
  - generate meaningful samples from non-texture images (real image)
  - generate complex textures or non-repetitive global structures and arrangement



From a Single Real Image

- Trained to capture the internal distribution of patches within the image

- Totally free from existence of dataset
  - regardless of paired or unpaired

- Does not designed for specific task(e.g. SR)
  - can be used for general image manipulation

- Can generate more diverse samples
  - both photo-realism and diversity

1. Unconditional & from Single Image

2. Model Architecture

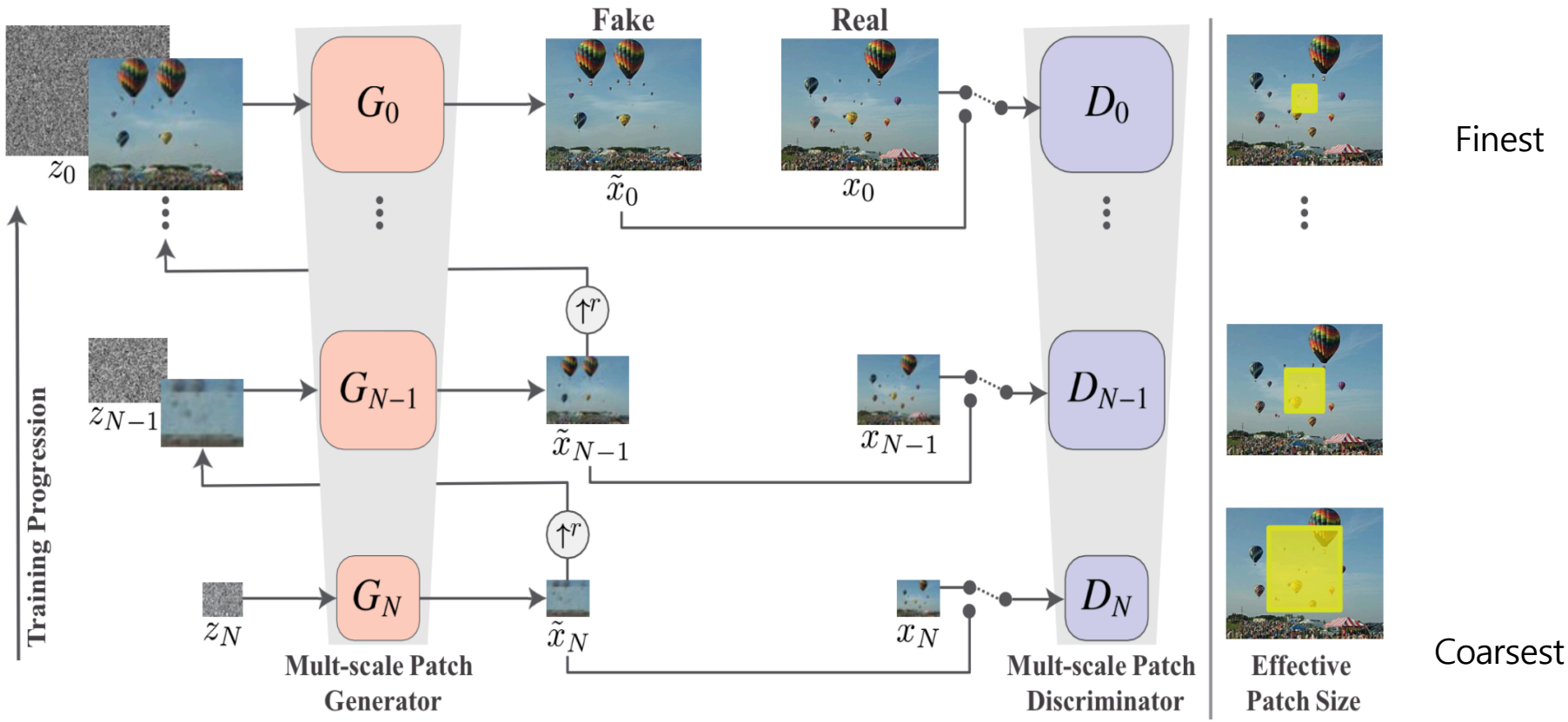
3. Training Methods and Details

4. Results and Evaluations

5. Various Applications

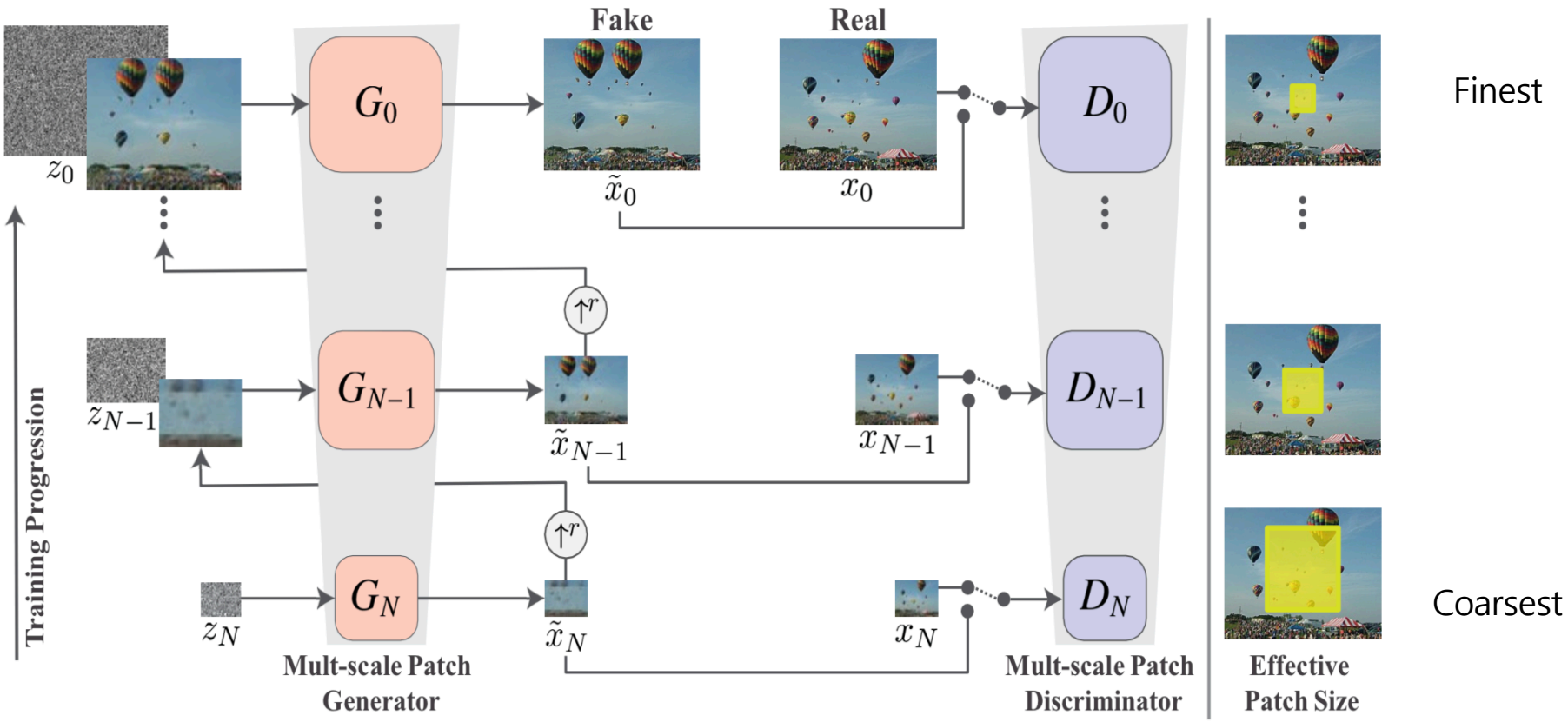
6. Limitations and Conclusions

Model Architecture



- x_n is a downsampled version of x by a factor γ^n ($\gamma > 1$)
- Always set minimal dimension at coarsest scale 25px
- Set the number of scales, N s.t. scaling factor γ is as close as 4/3
- $N = int\left(\frac{\log\left(\frac{imageSize}{25}\right)}{\log(\gamma)}\right)$

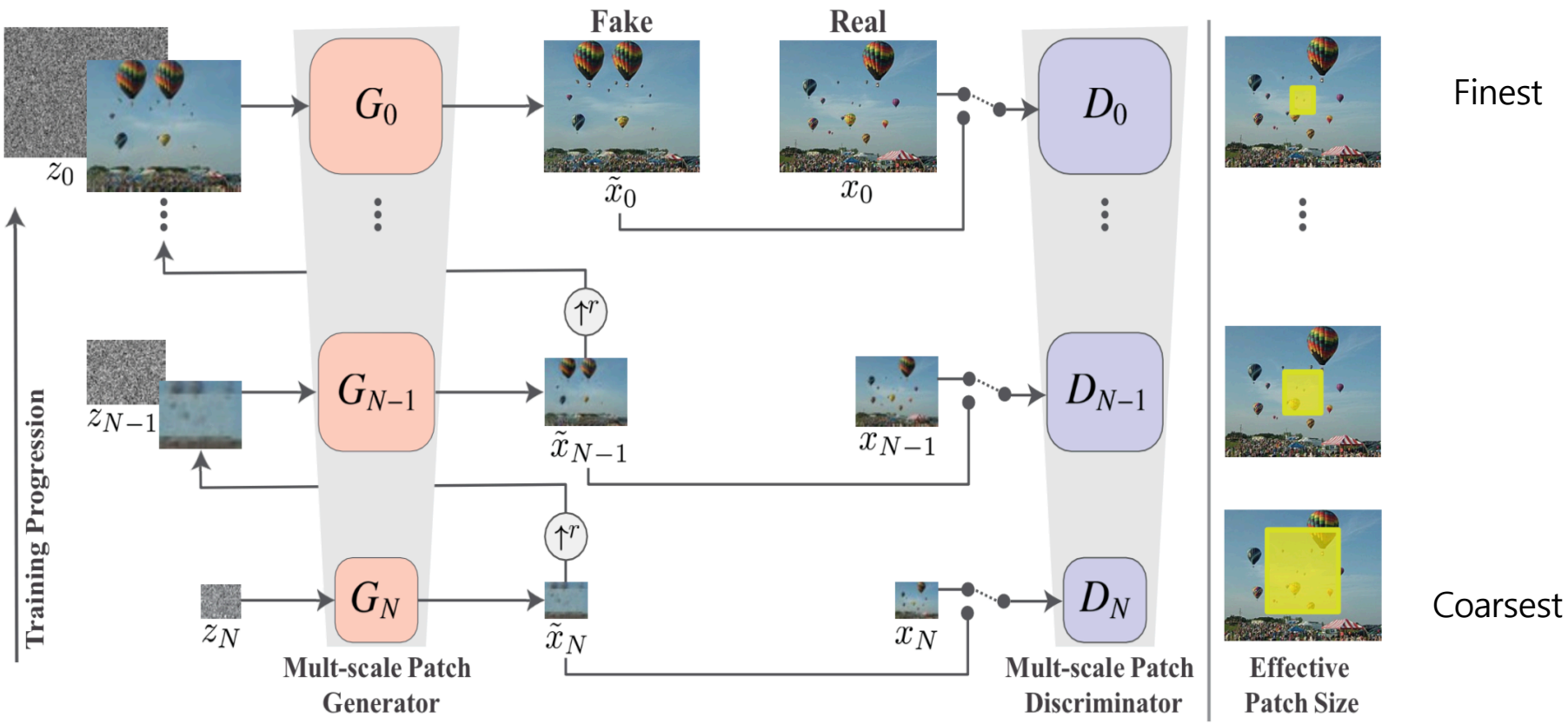
Model Architecture



$$\tilde{x}_N = G_N(z_N). \qquad n = N,$$

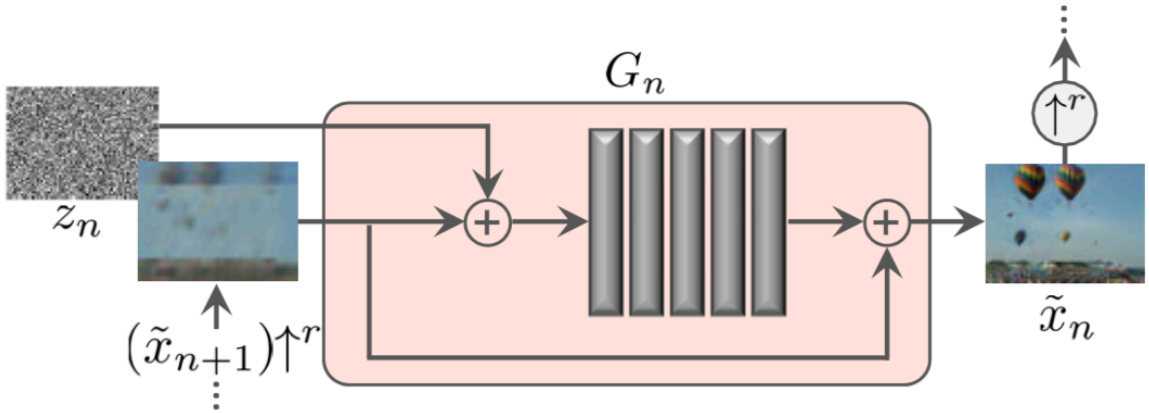
$$\tilde{x}_n = G_n\left(z_n, \left(\tilde{x}_{n+1}\right)\uparrow^r\right), \qquad n < N.$$

Model Architecture



- Train to capture the internal distribution of patches within the image
- Finer scale adds details that are not generated for previous scale
- All the generators have a similar architecture, only filters differs a little.

Model Architecture



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

- ψ_n is a fully conv net with 5 conv-blocks of Conv(3X)-BN-LeakyRelu.
- The role of the conv layer is to generate the missing details.
- Because they are fully convolutional, can generate images of arbitrary size and aspect ratio at test time.
- Light weight : takes only 30 minutes for one 1080Ti for an image of 256 X 256 size

1. Unconditional & from Single Image

2. Model Architecture

3. Training Methods and Details

4. Results and Evaluations

5. Various Applications

6. Limitations and Conclusions

Training Methods

Objective Function for Nth GAN

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

Training Details

- Sequentially train multi-scale architecture, from the coarsest scale to the finest one.
- Once each GAN is trained, it is kept fixed.
- At each scale, the weights of the generator and discriminator are initialized to those from the previous trained scale.
- Train each scale for 2000 iterations. In each iteration, alternate between 3 gradient steps for the generator and 3 gradient steps for the discriminator

Objective Function for nth GAN

Adversarial Loss Term

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} [D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} [D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2 \right]}_{\text{Our gradient penalty}} .$$

- WGAN – GP (Wasserstein with Gradient Penalty)
  - set lambda as 0.1
- Patch GAN

Reconstruction Loss Term

$$\begin{aligned} \text{for } n < N, \quad & \mathcal{L}_{\text{rec}} = \|G_n(0, (\tilde{x}_{n+1}^{\text{rec}}) \uparrow^r) - x_n\|^2, \\ n = N, \quad & \mathcal{L}_{\text{rec}} = \|G_N(z^*) - x_N\|^2. \end{aligned}$$

$$\{z_N^{\text{rec}}, z_{N-1}^{\text{rec}} \ldots, z_0^{\text{rec}}\} = \{z^*, 0, \ldots, 0\}$$

- set alpha as 10

More Optimization Details

- Adam optimizer with a lr 0.0005 (decreased by a factor of 0.1 after 1600 iterations) and $\beta 1 = 0.5$, $\beta 2 = 0.999$.
- The generator's last conv-block activation at each scale is Tanh instead of ReLU.
- The descriminator's last conv-block at each scale includes neither normalization nor activation.
- Leaky Relu with slope 0.2 for negative values.

1. Unconditional & from Single Image

2. Model Architecture

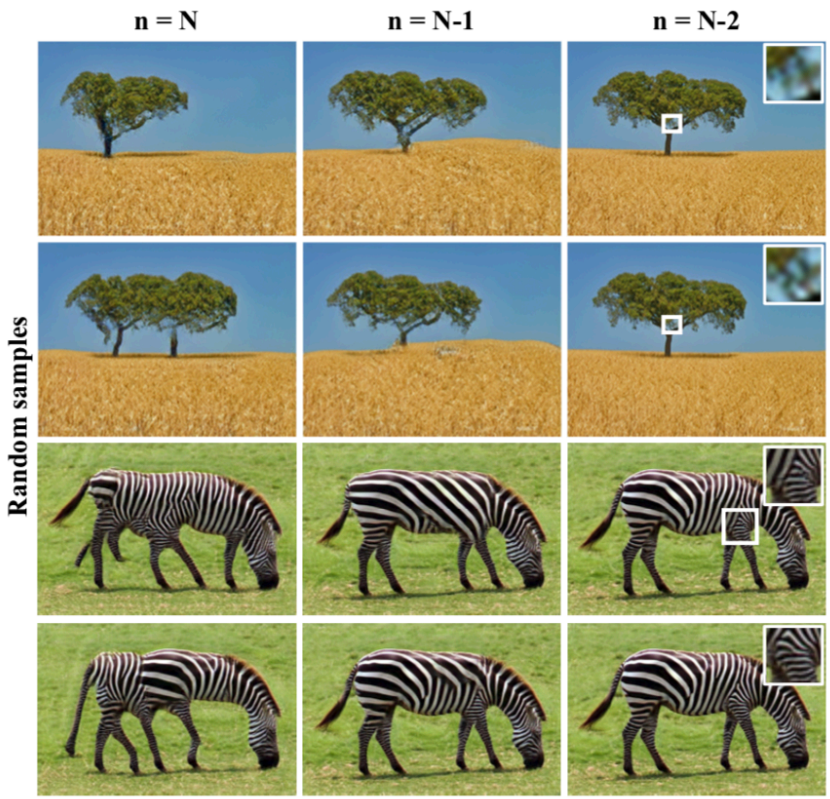
3. Training Methods and Details

4. Results and Evaluations

5. Various Applications

6. Limitations and Conclusions

Effects of Scales at Test Time



Starting from coarsest scale(n = N)

- large variability in the global structure
- in case of large object, can lead to unrealistic samples

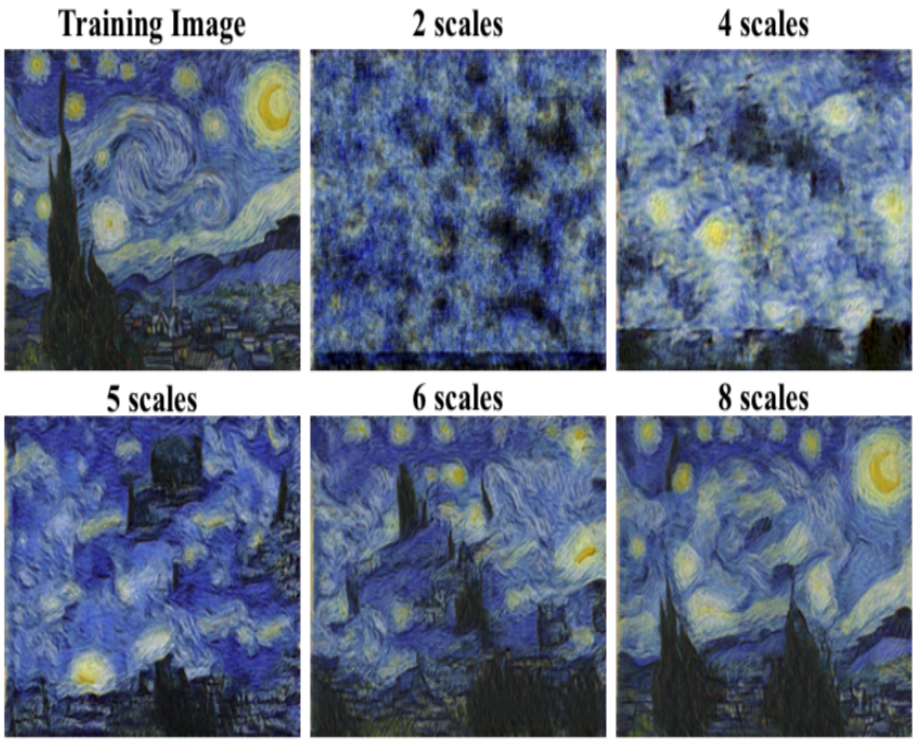
Starting from finer scale(n = N-1, N-2)

- keep the global structure intact
- alter only finer image features

Set noise, $Z = \{z_N^{rec}, \dots, z_{n+1}^{rec}, \mathrm{z_n}, \dots, \mathrm{z_0}\}$

Multi-Scale architecture can control the amount of variability, by choosing the scale from which to start at test time

Effects of Scales at Training Time



With a small number of scales

- can only capture fine textures.

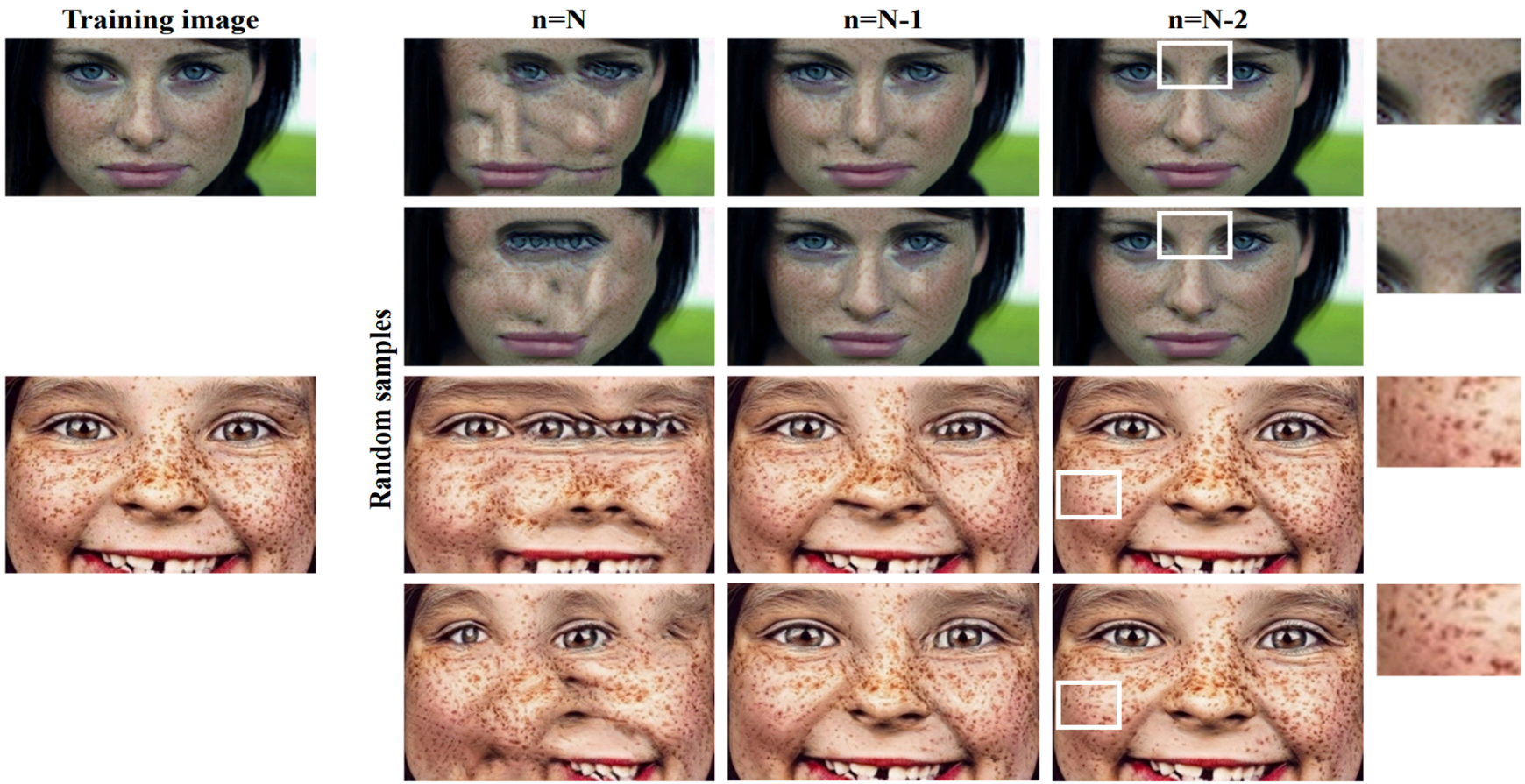
As a number of scales increase

- can capture larger structures and global arrangement of objects

With a small number of scales, the effective receptive field at coarsest level is smaller, allowing to capture only fine features

Failure Case

- SinGAN may produce unrealistic results when the global structure of the image is of high importance
- but, can avoid this by starting the generation process at a finer scale.



Quantitative Evaluation

Amazon Mechanical Turk (AMT)

- real / fake

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	$47.04\% \pm 0.8\%$

Single Image Frechet Inception Distance (SIFID)

- feature distance at the output of the conv layer just before the second pooling layer

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

Quantitative Evaluation

AMT Survey Images

Real Image

Fake Image(low diversity)

1. Unconditional & from Single Image

2. Model Architecture

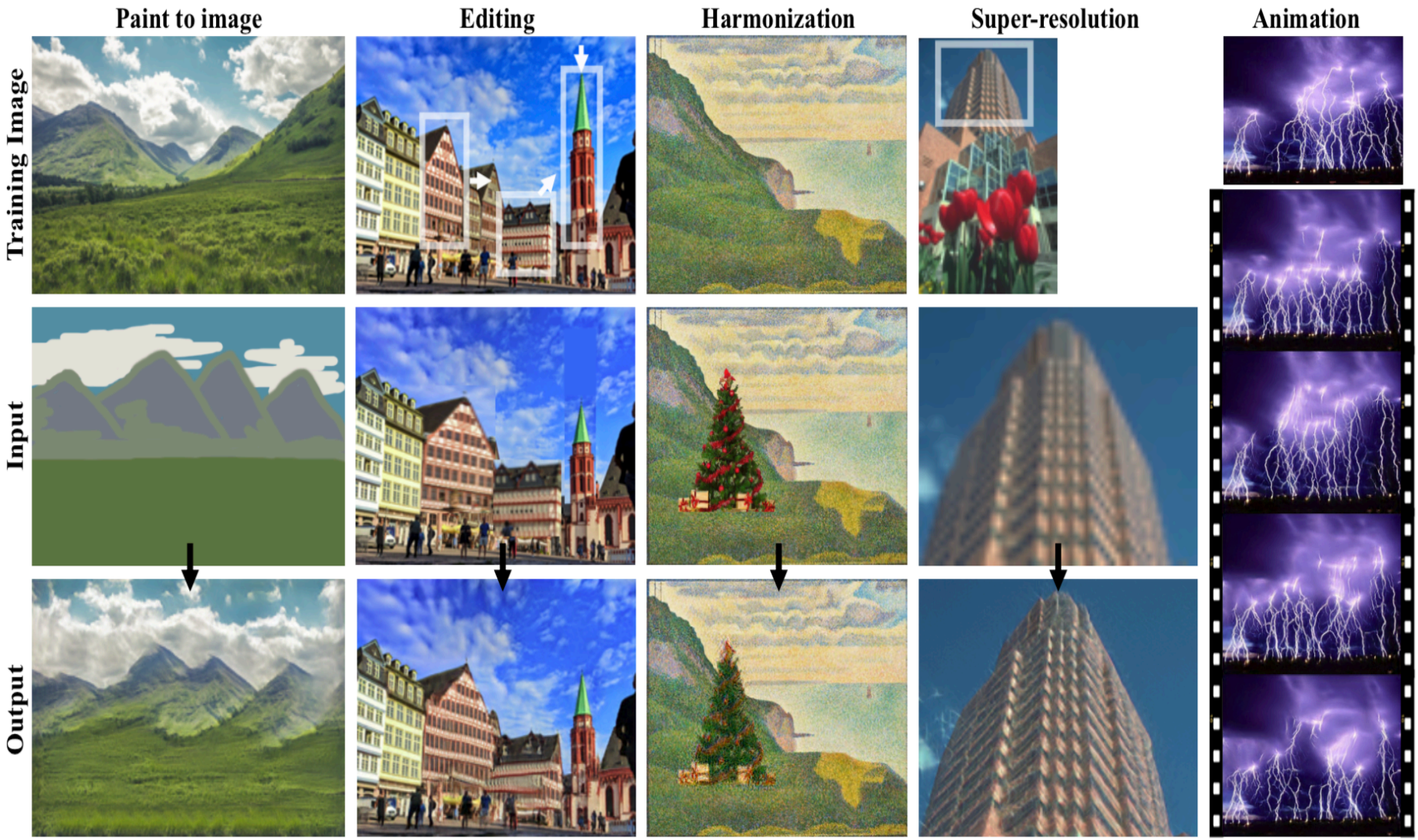
3. Training Methods and Details

4. Results and Evaluations

5. Various Applications

6. Limitations and Conclusions

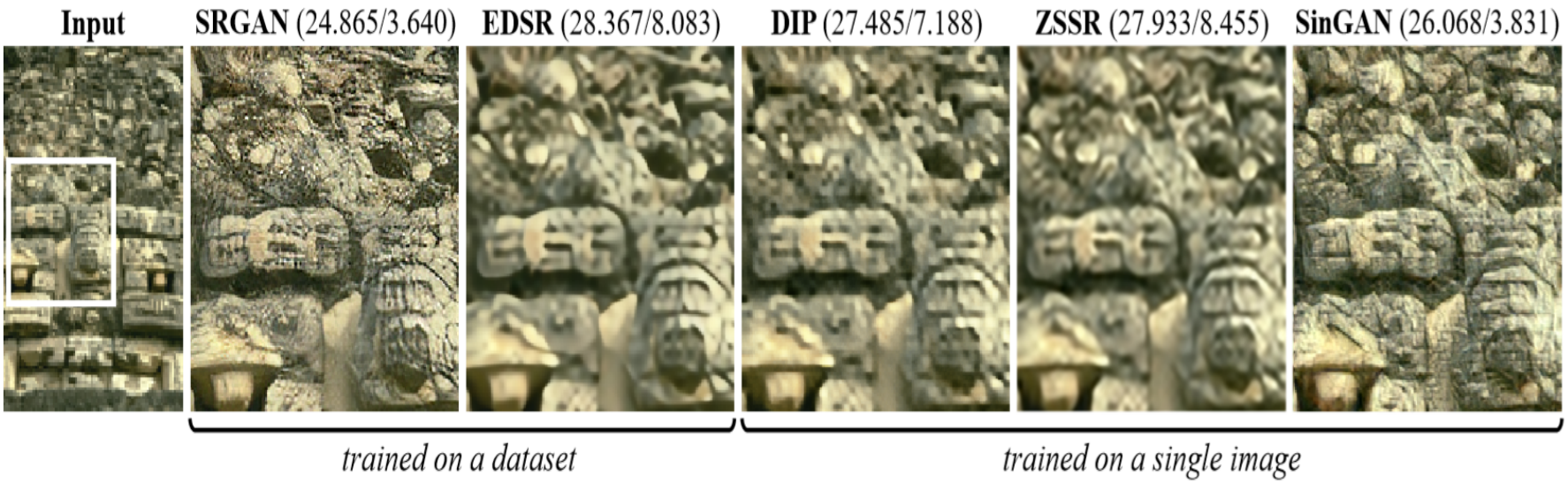
Various Applications



Applications

Super Resolution

- Increase the resolution of an input image by a factor s.
- Set reconstruction loss weight α to 100 and scale factor $\gamma = \sqrt[k]{s}$ for some $k \in \mathbb{N}$
- Train model on low-resolution image(LR)
- At test time, upsample LR images by γ and inject it to the last generator, G_0
- repeat this procedure k times



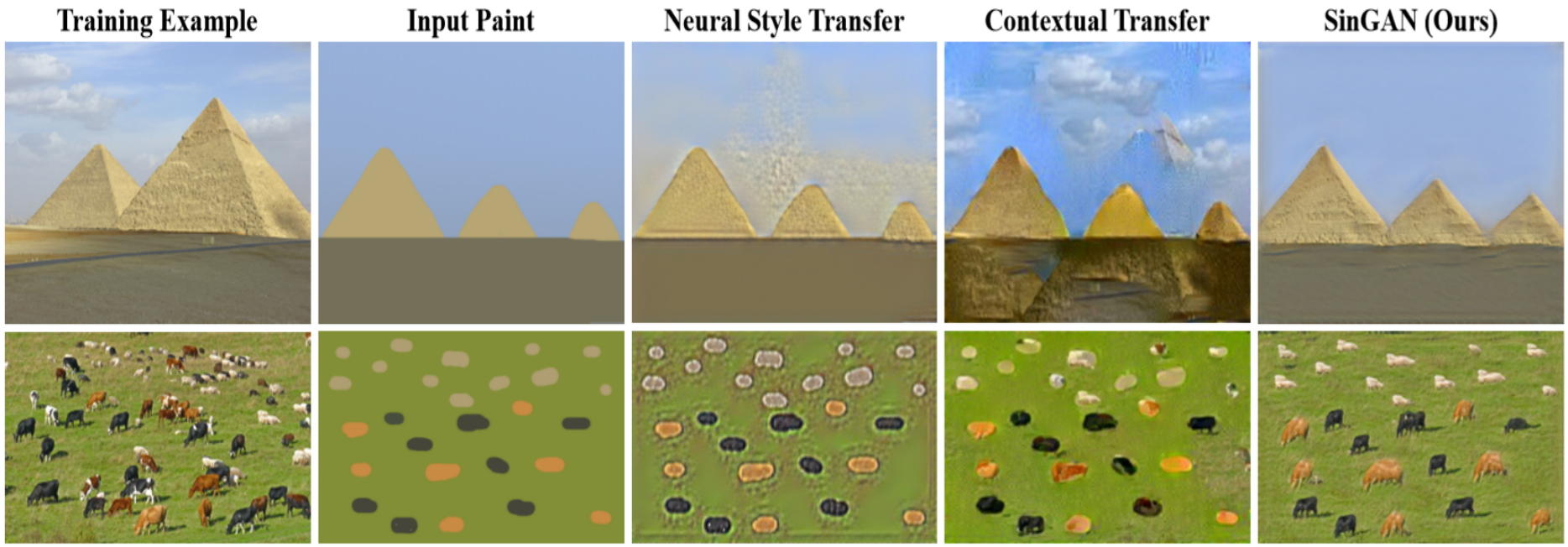
Applications

Super Resolution

Applications

Paint-to-Image

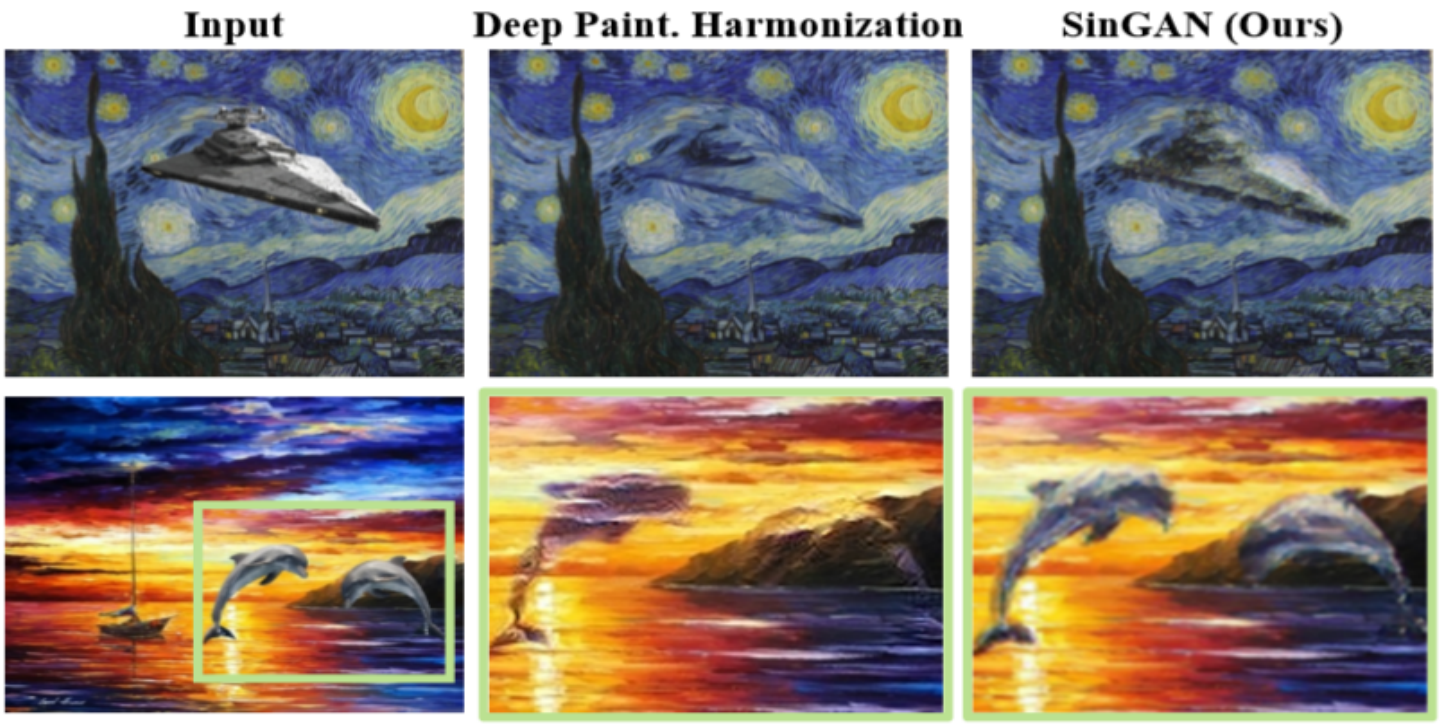
- Transfer a clipart into a photo-realistic image
- train as usual, background image
- downsample the clipart image and feed it into one of the coarse scales



Applications

Harmonization

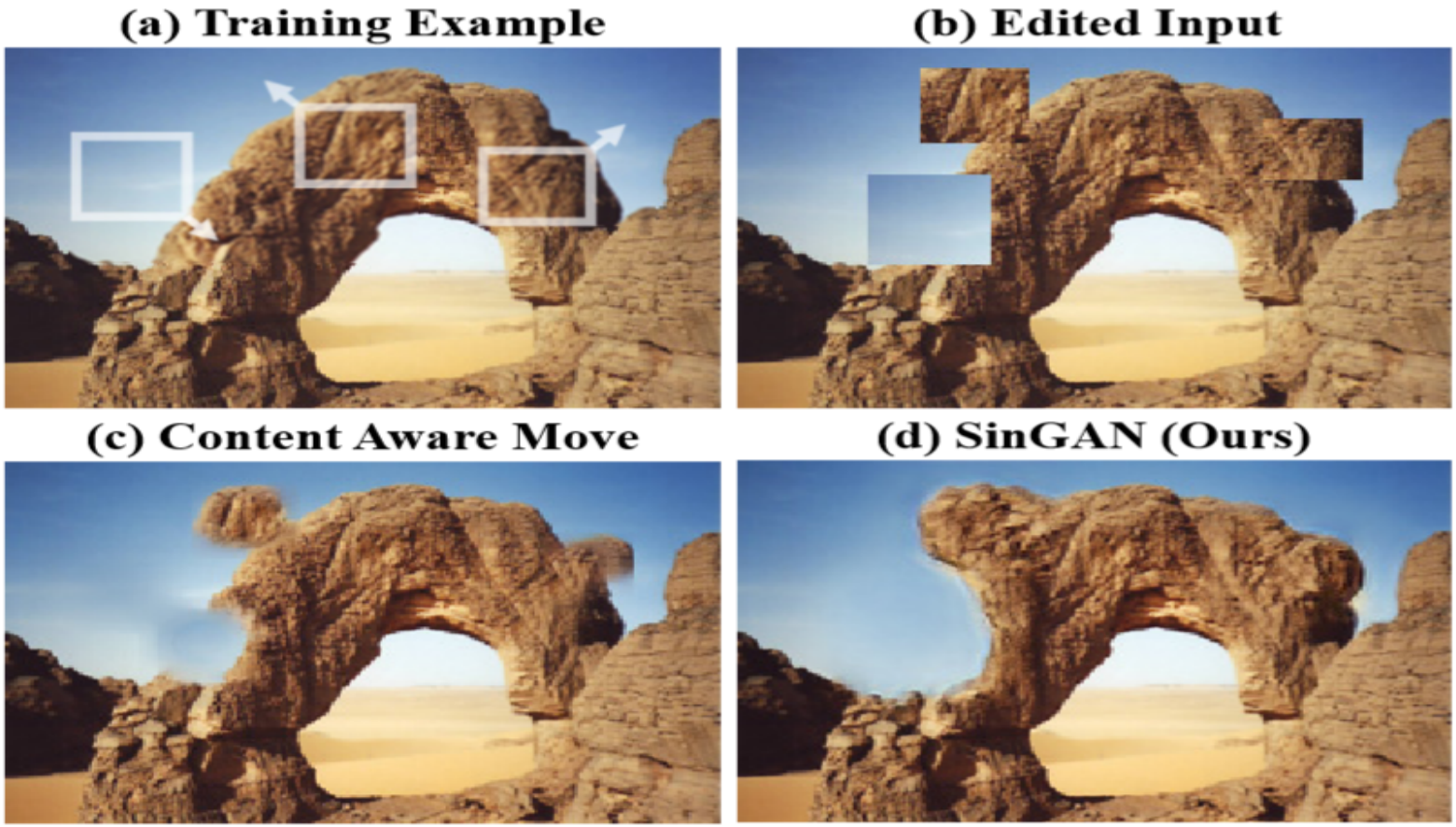
- Realistically blend a pasted object with a background image
- train as usual, background image
- downsample the composite and feed it into one of the coarse scales
- combine the generated image with the original background.



Applications

Editing

- Produce a seamless composite in which image regions have been copied and pasted in other locations



Applications

Single Image Animation

- Create a short video clip with realistic object motion, from a single input image
- Using SinGAN, can travel along the manifold of all appearances of the object
- Realistic effect is achieved by a random walk in z-space

https://www.youtube.com/watch?v=xk8bWLZk4DU&feature=youtu.be

1. Unconditional & from Single Image

2. Model Architecture

3. Training Methods and Details

4. Results and Evaluations

5. Various Applications

6. Conclusions

Conclusion

- unconditional generative model that can be learned from a single natural image
- contains a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of the image.
- generated samples are high quality and SinGAN can be used in a wide range of image manipulation tasks.
- Internal learning is inherently limited in terms of semantic diversity compared to externally trained generation methods.
- For example, if the training image contains a single dog, SinGAN will not generate samples of different dog breeds.

(My) Implementation

Paint to Image

Harmonization

https://github.com/bocharm/sinGAN

Q & A

감사합니다.