Evaluation of Generative Models: *Sample Quality*

Hao Dong

Peking University

北京大學
PEKING UNIVERSITY

Evaluation of Generative Models: *Sample Quality*

Hao Dong

Peking University

北京大學
PEKING UNIVERSITY

- Sample Quality
- Density Estimation & Latent Representation
- Practice

Evaluation of Generative Models: Sample Quality

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

北京大學
PEKING UNIVERSITY

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

Known Ground Truth

- You know the ground truth of the generated images
  - Data is paired
  - Directly compare the generated images with ground truth images
  - Metrics: MSE, PSNR, SSIM
  - Example: SRGAN

- SRGAN: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network
  - Given a low-resolution input image
  - To obtain its high-resolution counterpart
  - Generative Adversarial Networks can make it!
  - Generate the corresponding high-resolution counterpart

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

SRGAN - Architecture

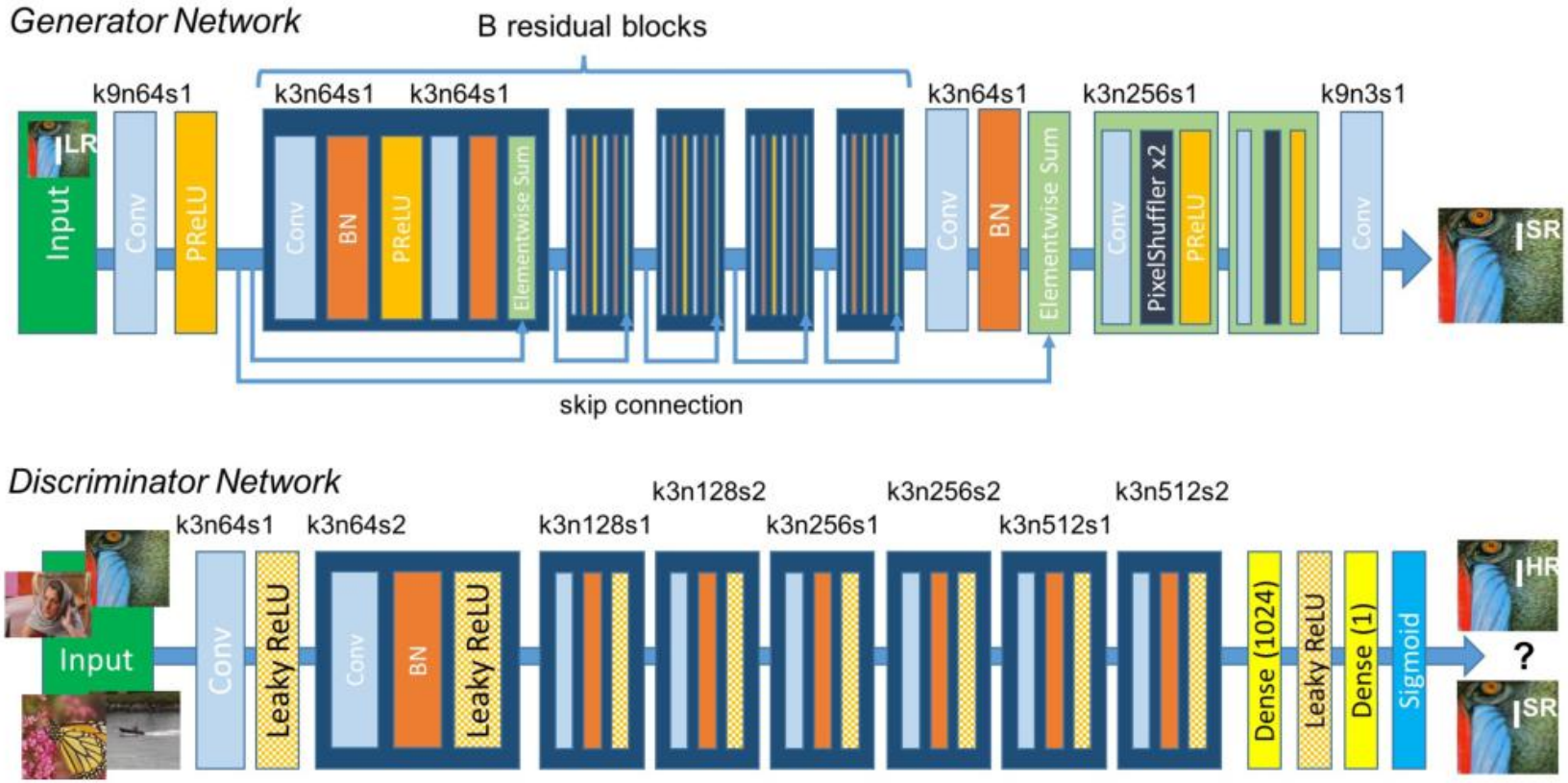


Figure 4: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

SRGAN – Loss Functions

- Generative Adversarial Loss

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))] \quad (2)$$

  - To train a generative model G with the goal of fooling the discriminator D
  - The discriminator D is trained to distinguish super-resolved images from real images.
  - G can learn to create solutions that are highly similar to real images and thus difficult to classify by D
  - Encourages perceptually superior solutions residing in the subspace, the manifold, of natural images.
  - In contrast to SR solutions obtained by minimizing pixel-wise error metrics, such as the MSE.

SRGAN – Loss Functions

- Perceptual Loss

$$l^{SR} = \underbrace{\underbrace{l_{\mathrm{X}}^{SR}}_{\text{content loss}} + \underbrace{10^{-3} l_{Gen}^{SR}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}} \tag{3}$$

  - Content Loss

    1) MSE Loss

$$l_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{4}$$

    2) VGG Loss
    (VGG is a CNN net)

$$l_{VGG/i.j}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \tag{5}$$

  - Adversarial Loss

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{6}$$

SRGAN – Results

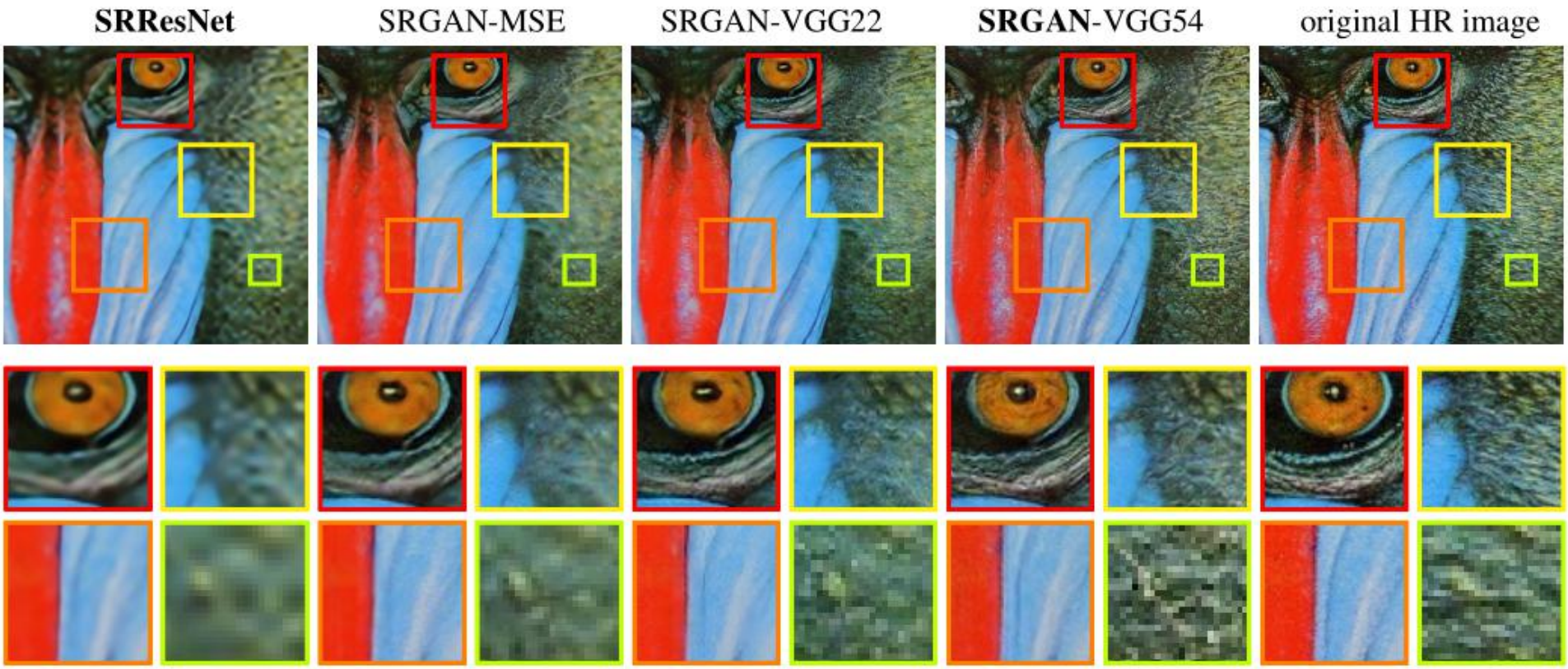


Figure 6: **SRResNet** (left: a,b), SRGAN-MSE (middle left: c,d), SRGAN-VGG2.2 (middle: e,f) and **SRGAN**-VGG54 (middle right: g,h) reconstruction results and corresponding reference HR image (right: i,j). [4× upscaling]

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

MSE: Mean Squared Error

- Reconstruction Evaluation
  - If you know the ground truth … e.g., image super resolution
  - Mean Squared Error (MSE): range $[0, \infty)$ the smaller the better
  - $MSE = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2 = \frac{1}{m}\|x - y\|_2^2$
  - where x_i, y_i are the per pixel value of the pair of images
  - Mean Squared Error can evaluate the similarity of the pair of images
  - The smaller the number is, the more similar the two images are, thus the better the reconstruction is
  - However, the pixel-wise error measurements have limitations (discuss)
  - Not perceptually good, even when the number is small

PSNR: Peak Signal-to-Noise Ratio

- Reconstruction Evaluation
  - If you know the ground truth … e.g., image super resolution

  - Peak Signal to Noise Ratio (PSNR): the max possible power of data/ the power of noise
  - $PSNR(x,y) = 10\log_{10}\left(\frac{R^2}{MSE(x,y)}\right)$
  - For uint8 data, the max possible power is 255
  - For float data, the max possible power is 1
  - The larger the number is, the better the quality of the image is

  - Implementation:

```
# im1 和 im2 都为灰度图像，uint8 类型

# method 1
diff = im1 - im2
mse = np.mean(np.square(diff))
psnr = 10 * np.log10(255 * 255 / mse)

# method 2
psnr = skimage.measure.compare_psnr(im1, im2, 255)
```

SSIM: Structural Similarity

- Reconstruction Evaluation
  - If you know the ground truth … e.g., image super resolution
  - Structure Similarity Index Measure (SSIM): range [0, 1], the higher the better
  - $SSIM(x,y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$

mean variance covariance

$c_1 = (k_1 L), c_2 = (k_2 L)$ are two variables to stabilise the division with the weak denominator. $k_1 = 0.01$, $k_2 = 0.03$ by default. L the dynamic range of the pixel value, e.g., [0, 1] for 1

  - The product of the relative luminance, contrast and structure
  - The larger the number is, the better the quality of the image is
  - Implementation:

compare_ssim

skimage.measure.compare_ssim(*X*, *Y*, *win_size=None*, *gradient=False*, *data_range=None*, *multichannel=False*, *gaussian_weights=False*, *full=False*, ***kwargs*) [source]

Compute the mean structural similarity index between two images.

SRGAN – Evaluations

Table 2: Comparison of NN, bicubic, SRCNN [9], SelfExSR [31], DRCN [34], ESPCN [48], **SRResNet**, **SRGAN**-VGG54 and the original HR on benchmark data. Highest measures (PSNR [dB], SSIM, MOS) in bold. [4× upscaling]

Set5	nearest	bicubic	SRCNN	SelfExSR	DRCN	ESPCN	**SRResNet**	**SRGAN**	HR
PSNR	26.26	28.43	30.07	30.33	31.52	30.76	**32.05**	29.40	∞
SSIM	0.7552	0.8211	0.8627	0.872	0.8938	0.8784	**0.9019**	0.8472	1
MOS	1.28	1.97	2.57	2.65	3.26	2.89	3.37	**3.58**	4.32
Set14									
PSNR	24.64	25.99	27.18	27.45	28.02	27.66	**28.49**	26.02	∞
SSIM	0.7100	0.7486	0.7861	0.7972	0.8074	0.8004	**0.8184**	0.7397	1
MOS	1.20	1.80	2.26	2.34	2.84	2.52	2.98	**3.72**	4.32
BSD100									
PSNR	25.02	25.94	26.68	26.83	27.21	27.02	**27.58**	25.16	∞
SSIM	0.6606	0.6935	0.7291	0.7387	0.7493	0.7442	**0.7620**	0.6688	1
MOS	1.11	1.47	1.87	1.89	2.12	2.01	2.29	**3.56**	4.46

- Several Evaluation Metrics are used
  - PSNR: Peak Signal-to-Noise Ratio
  - SSIM: Structural Similarity
  - MOS: Mean Opinion Score (Human Evaluation)

To summary

- Reconstruction Evaluation
  - If you know the ground truth … e.g., image super resolution
  - Although there are a variety of metrics to evaluate the quality of the generated images, these metrics also have some drawbacks.
  - In SRGAN, although SRResNet performs best in terms of PSNR/SSIM, the perceptual quality of its results is not the best.
  - In terms of MOS (Human Evaluation), SRGAN performs best.
  - The paper has further shown that standard quantitative measures such as PSNR and SSIM fail to capture and accurately assess image quality with respect to the human visual system.
  - Human Evaluation is necessary in the evaluation of generative models

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

Unknown Ground Truth

- You don’t know the ground truth of the generated images
  - Data is unpaired
  - Infeasible to directly compare the generated images with ground truth images
  - Unpaired data is very common
  - Metrics: IS, FID, KID, LPIPS
  - Example: DCGAN, CycleGAN, DRIT, StarGAN

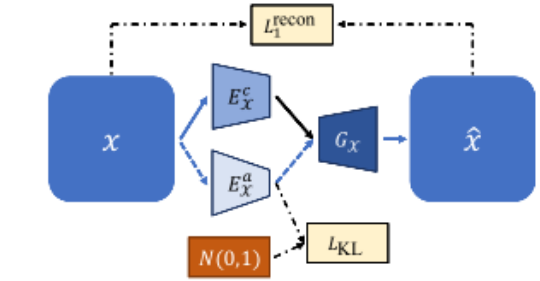
- DRIT: Diverse Image-to-Image Translation via Disentangled Representations
  - Cross-Domain Translation

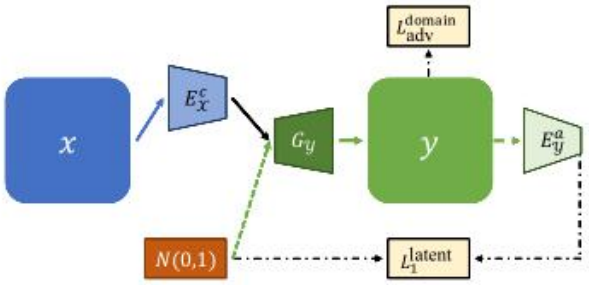
- StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation
  - Multi-Domain Translation within a single model

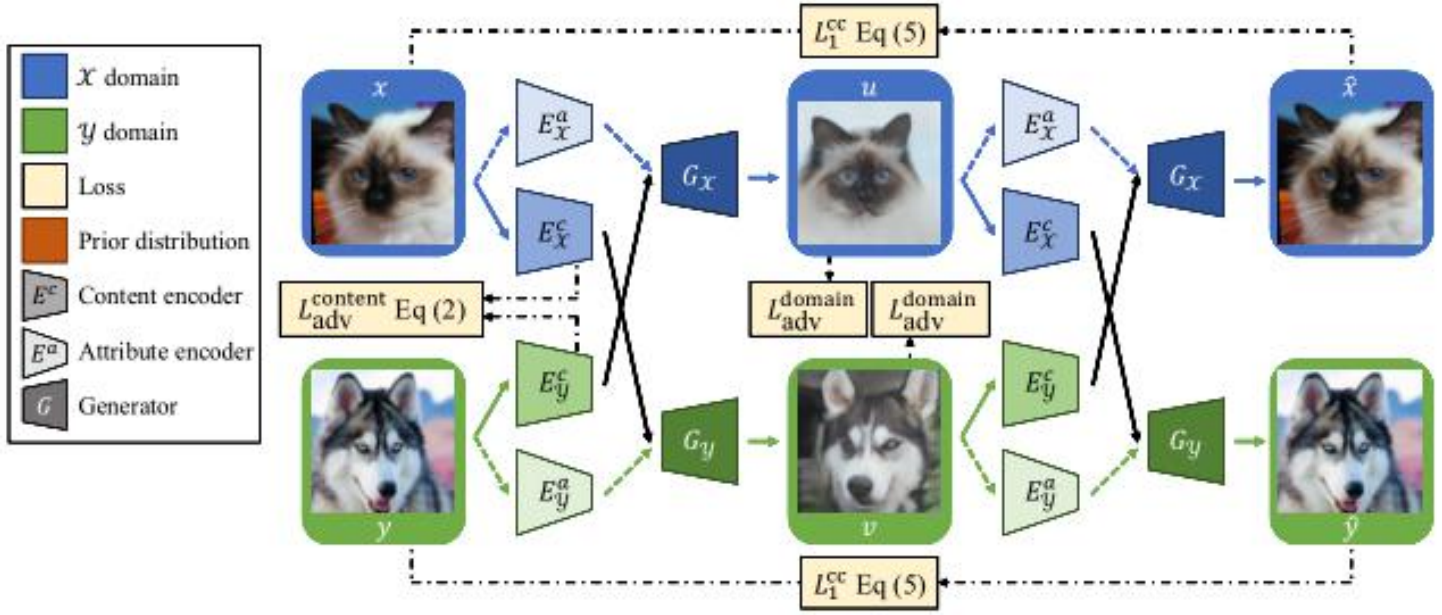
- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

DRIT - Architecture

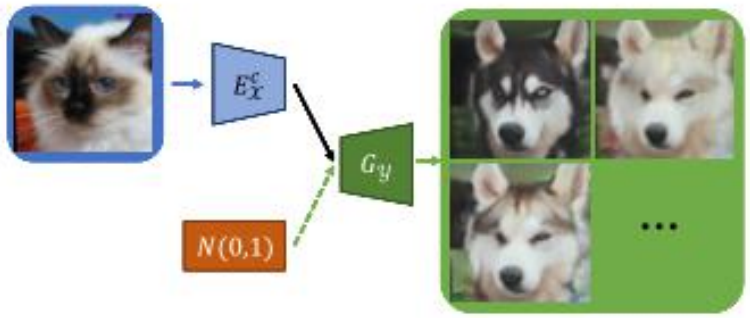
- Cross-Domain
- Diversity
- Appearance + Content



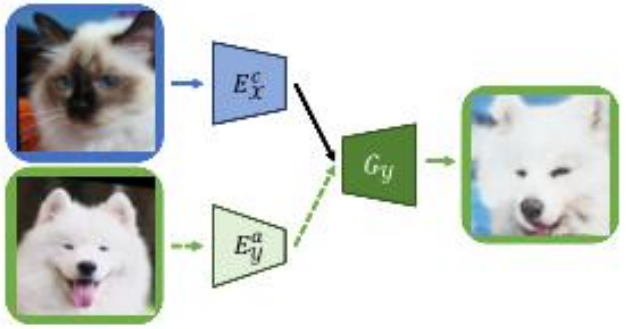




(a) Training with unpaired images

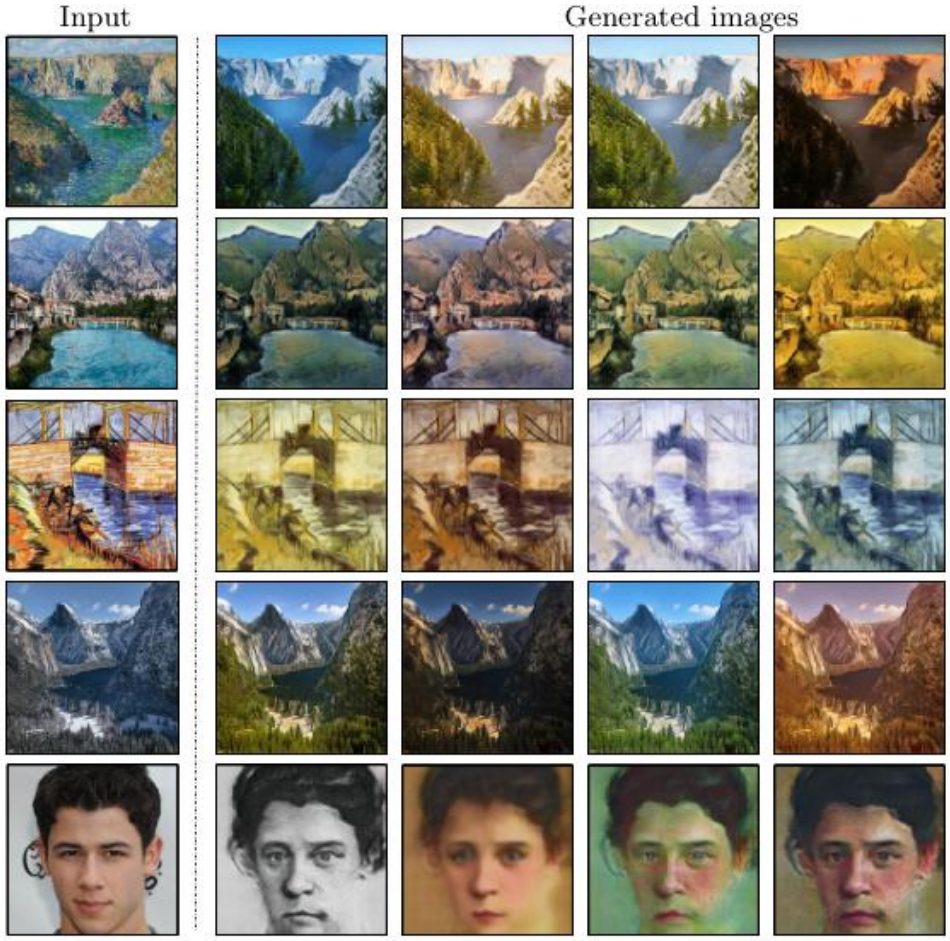


(b) Testing with random attributes



(c) Testing with a given attribute

DRIT - Results



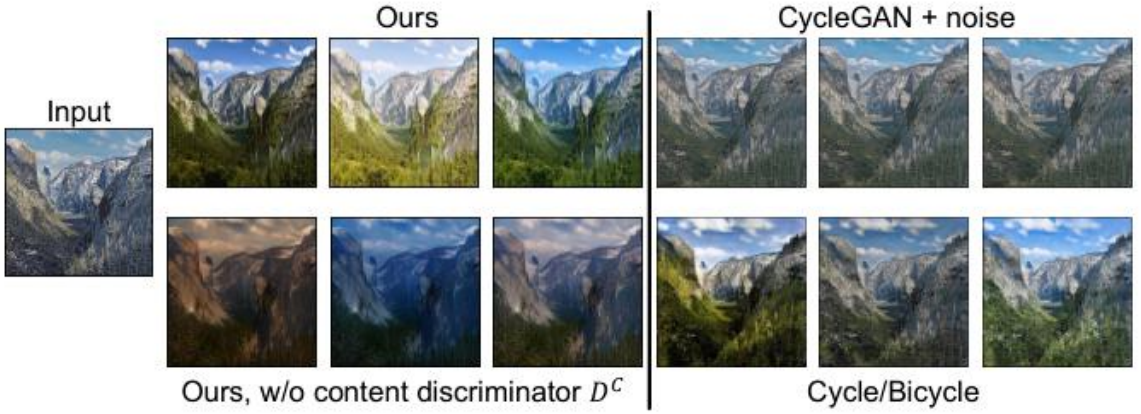


Fig. 6: **Diversity comparison.** On the winter → summer translation task, our model produces more diverse and realistic samples over baselines.

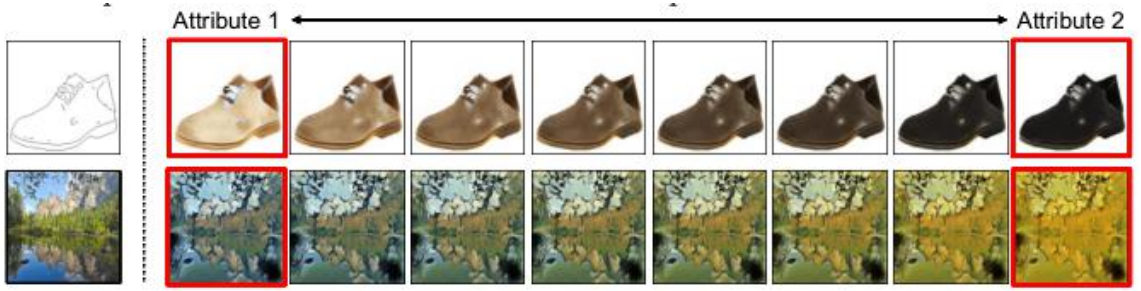
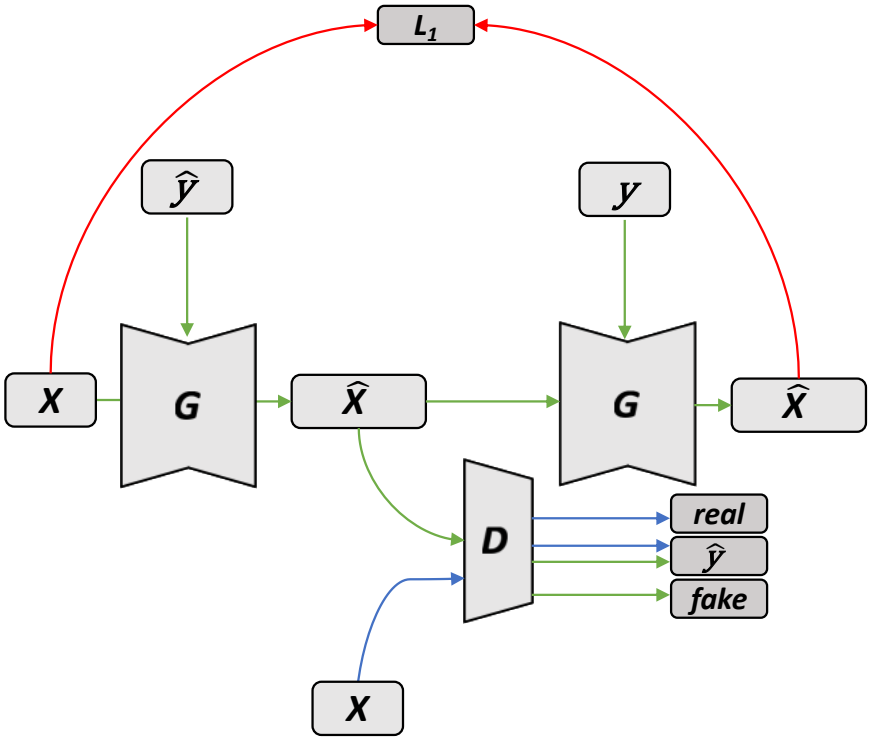
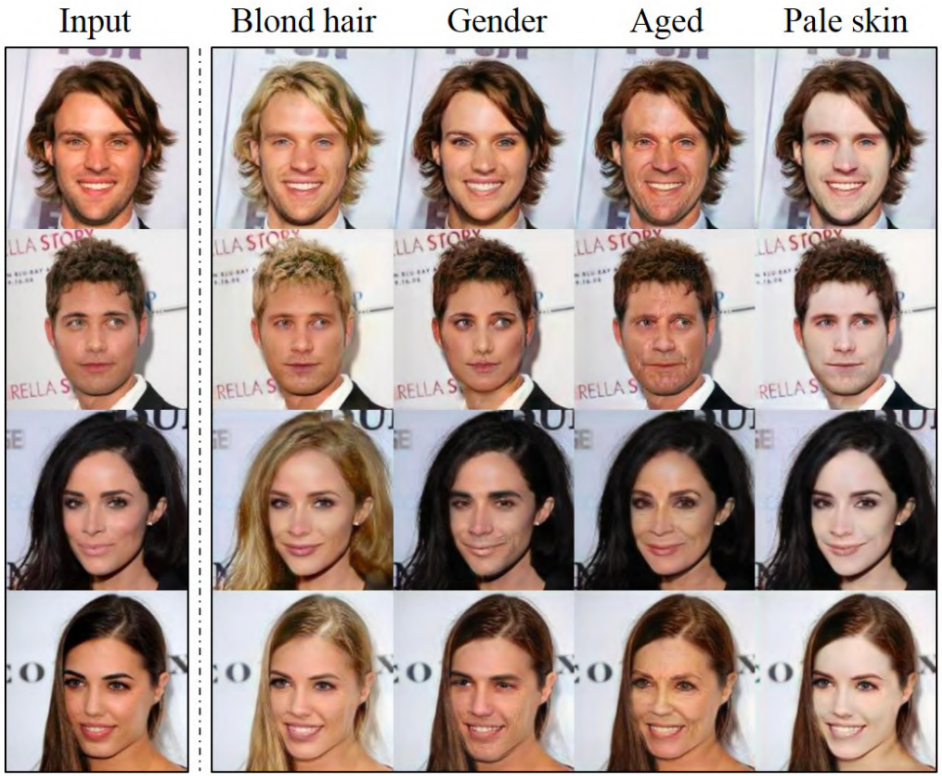


Fig. 7: **Linear interpolation between two attribute vectors.** Translation results with linear-interpolated attribute vectors between two attributes (highlighted in red).

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

StarGAN – Architecture





StarGAN – Results (Multi-Domain Translation in a single model)

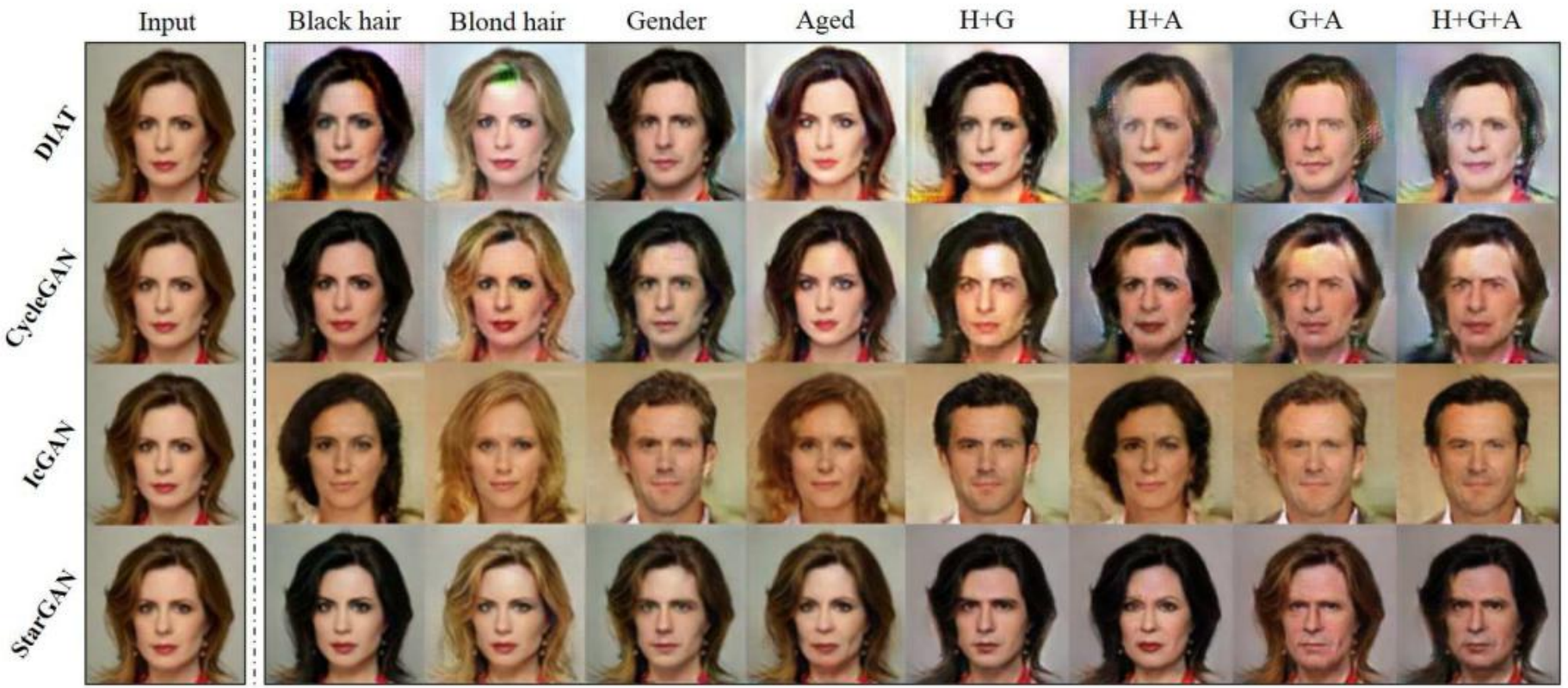


Figure 4. Facial attribute transfer results on the CelebA dataset. The first column shows the input image, next four columns show the single attribute transfer results, and rightmost columns show the multi-attribute transfer results. H: Hair color, G: Gender, A: Aged.

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

StarGAN – Evaluation

- Classification Accuracy
- Human Evaluation

Method	Classification error	# of parameters
DIAT	4.10	52.6M × 7
CycleGAN	5.99	52.6M × 14
IcGAN	8.07	67.8M × 1
StarGAN	**2.12**	**53.2M × 1**
Real images	0.45	-

Table 3. Classification errors [%] and the number of parameters on the RaFD dataset.

Method	Hair color	Gender	Aged
DIAT	9.3%	31.4%	6.9%
CycleGAN	20.0%	16.6%	13.3%
IcGAN	4.5%	12.9%	9.2%
StarGAN	**66.2%**	**39.1%**	**70.6%**

Table 1. AMT perceptual evaluation for ranking different models on a single attribute transfer task. Each column sums to 100%.

Method	H+G	H+A	G+A	H+G+A
DIAT	20.4%	15.6%	18.7%	15.6%
CycleGAN	14.0%	12.0%	11.2%	11.9%
IcGAN	18.2%	10.9%	20.3%	20.3%
StarGAN	**47.4%**	**61.5%**	**49.8%**	**52.2%**

a

Table 2. AMT perceptual evaluation for ranking different models on a multi-attribute transfer task. H: Hair color; G: Gender; A: Aged.

Inception Score

- **Assumption 1:** We are evaluating sample quality for generative models trained on labelled datasets
- **Assumption 2:** We have a good probabilistic classifier $c(y|\boldsymbol{x})$ for predicting the label y for any point $\boldsymbol{x}$
- We want samples from a good generative model to satisfy two criteria: **sharpness** and **diversity**
- **Sharpness (S):**

Low sharpness

High sharpness

$$S = \exp\left(E_{\mathbf{x}\sim p}\left[\int c(y|\mathbf{x})\log c(y|\mathbf{x})\mathrm{d}y\right]\right)$$

Inception Score

- High sharpness implies classifier is confident in making predictions for generated images
- That is, classifier's predictive distribution $c(y|\boldsymbol{x})$ has low entropy
- **Diversity (D):**

Low diversity

High diversity

$$D = \exp\left(-E_{\mathbf{x}\sim p}\left[\int c(y|\mathbf{x})\log c(y)\mathrm{d}y\right]\right)$$

where $c(y) = E_{x\sim p}\left[c(y|\boldsymbol{x})\right]$ is the classifier's marginal predictive distribution

- High diversity implies $c(y)$ has high entropy

Inception Score

- **Inception Scores (IS)** combine the two criteria of sharpness and diversity into a simple metric

$$IS = D \times S$$

- Correlates well with human judgement in practice

Table 2. Quantitative evaluation on animal image translation. This dataset contains 3 domains. We perform bidirectional translation for each domain pair, resulting in 6 translation tasks. We use CIS and IS to measure the performance on each task. To obtain a high CIS/IS score, a model needs to generate samples that are both high-quality and diverse. While IS measures diversity of all output images, CIS measures diversity of outputs conditioned on a single input image.

	CycleGAN		CycleGAN* with noise		UNIT		MUNIT	
	CIS	IS	CIS	IS	CIS	IS	CIS	IS
house cats → big cats	0.078	0.795	0.034	0.701	0.096	0.666	0.911	0.923
big cats → house cats	0.109	0.887	0.124	0.848	0.164	0.817	0.956	0.954
house cats → dogs	0.044	0.895	0.070	0.901	0.045	0.827	1.231	1.255
dogs → house cats	0.121	0.921	0.137	0.978	0.193	0.982	1.035	1.034
big cats → dogs	0.058	0.762	0.019	0.589	0.094	0.910	1.205	1.233
dogs → big cats	0.047	0.620	0.022	0.558	0.096	0.754	0.897	0.901
Average	0.076	0.813	0.068	0.762	0.115	0.826	1.039	1.050

Huang, X., Liu, M.y., Belongie, S., Kautz, J.: Munit: Multimodal unsupervised image-to-image translation. In: ECCV (2018)

Fréchet Inception Distance

- **Fréchet Inception Distance (FID)**
- measures similarities in the feature representations (e.g., those learned by a pretrained classifier) for datapoints sampled from p_θ and the test dataset p_{data}.

- Different from IS, FID takes into account both samples from p_θ and the desired data distribution p_{data}.

Table 1: FID of Different Methods with respect to five attributes. The + (−) represents the generated images by adding (removing) the attribute.

FID	bangs		smiling		mustache		eyeglasses		male	
	+	−	+	−	+	−	+	−	+	−
UNIT	135.41	137.94	120.25	125.04	119.32	131.33	111.49	139.43	152.16	154.59
CycleGAN	**27.81**	33.22	**23.23**	**22.74**	43.58	55.49	**36.87**	**48.82**	60.25	**46.25**
StarGAN	59.68	71.07	51.36	78.87	99.03	176.18	70.40	142.35	70.14	206.21
DNA-GAN	79.27	76.89	77.04	72.35	126.33	127.66	75.02	75.96	121.04	118.67
ELEGANT	30.71	**31.12**	25.71	24.88	**37.51**	**49.13**	47.35	60.71	**59.37**	56.80

Fréchet Inception Distance

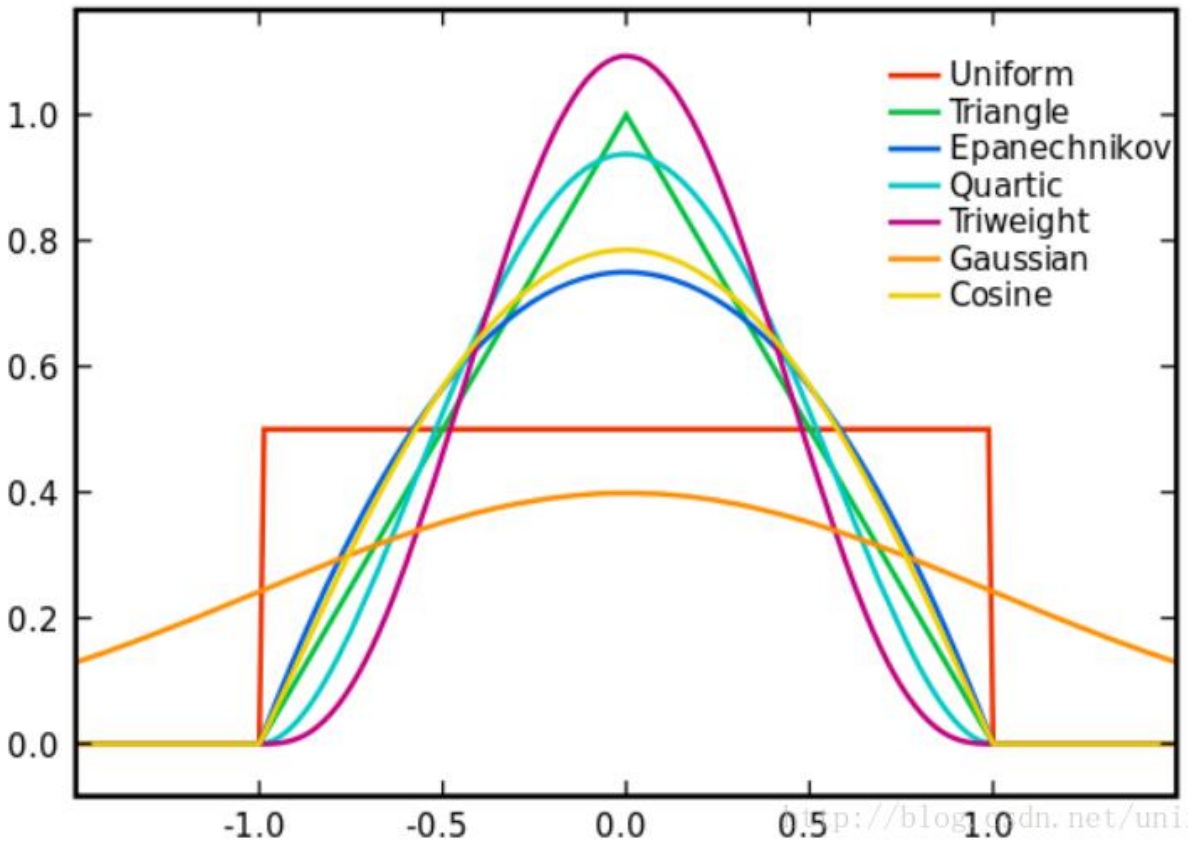
- Computing **Fréchet Inception Distance (FID)**
  - Let G denote the generated samples and T denote the test dataset
  - Compute feature representations F_G and F_T for G and T respectively (e.g., prefinal layer of Inception Net)
  - Fit a multivariate Gaussian to each of F_G and F_T. Let (μ_G, Σ_G) and (μ_T, Σ_T) denote the mean and covariances of the two Gaussians
  - FID is defined as

$$\text{FID} = \|\mu_T - \mu_G\|^2 + \text{Tr}(\Sigma_T + \Sigma_G - 2(\Sigma_T\Sigma_G)^{1/2})$$

- Lower **FID** implies better sample quality
  - [1] Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2018)
  - [2] Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with GAN for transferring multiple face attributes. In: ECCV (2018)

Kernel Inception Distance: Kernel Functions

- The integral of $K(x)$ should be 1
  - There are many kinds of kernel functions
  - Uniform
  - Triangular
  - Biweight
  - Triweight
  - Epanechnikov
  - Normal



- Gaussian Kernel
  - Convenient to use

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

Kernel Inception Distance

- **Maximum Mean Discrepancy (MMD)** is a two-sample test statistic that compares samples from two distributions p and q by computing differences in their moments (mean, variances etc.)
- Key idea: Use a suitable kernel e.g., Gaussian to measure similarity between points

$x^2 + y^2 - 2xy = 0$ when $x = y$

$$MMD(p, q) = E_{\mathbf{x},\mathbf{x}' \sim p}[K(\mathbf{x}, \mathbf{x}')] + E_{\mathbf{x},\mathbf{x}' \sim q}[K(\mathbf{x}, \mathbf{x}')] - 2E_{\mathbf{x} \sim p, \mathbf{x}' \sim q}[K(\mathbf{x}, \mathbf{x}')]$$

- Intuitively, **MMD** is comparing the "similarity" between samples within p and q individually to the samples from the mixture of p and q

Kernel Inception Distance

- **Kernel Inception Distance (KID):** compute the MMD in the feature space of a classifier (e.g., Inception Network)

- **FID vs. KID**
- FID is biased (can only be positive), KID is unbiased
- FID can be evaluated in $O(n)$ time, KID evaluation requires $O(n^2)$ time

- Lower **KID** implies better sample quality
  - [1] Nizan, O., Tal, A.: Breaking the cycle – colleagues are all you need. In: arXiv:1911.10538 (2019)
  - [2] Kim, J., Kim, M., Kang, H., Lee, K.: U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: ICLR (2020)

DRIT - Evaluations

- Classification Accuracy
- Diversity - LPIPS
- Human Evaluation

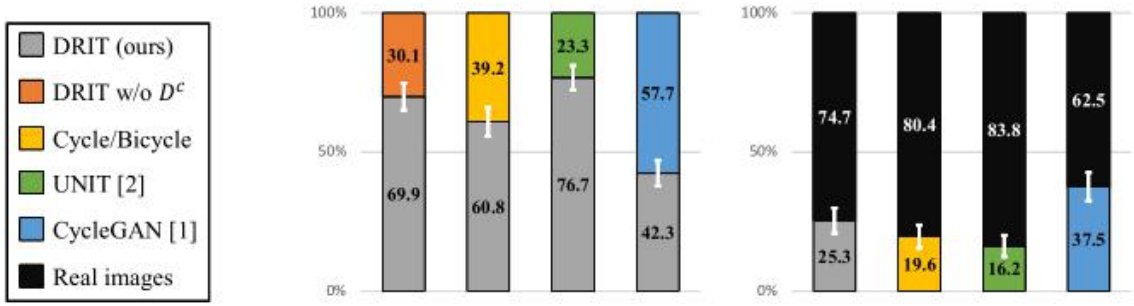


Table 2: **Diversity.** We use the LPIPS metric [47] to measure the diversity of generated images on the Yosemite dataset.

Method	Diversity
real images	.448 ± .012
DRIT	**.424** ± .010
DRIT w/o D^c	.410 ± .016
UNIT [27]	.406 ± .022
CycleGAN [48]	.413 ± .008
Cycle/Bicycle	.399 ± .009

(a) MNIST-M

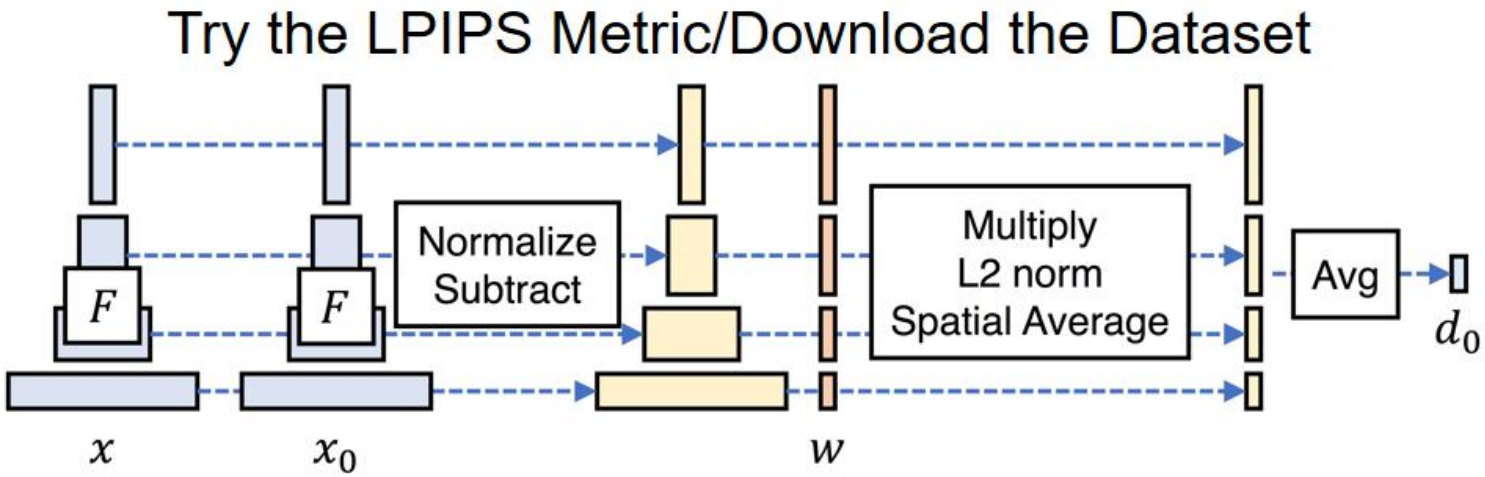
Model	Classification Accuracy (%)
Source-only	56.6
CycleGAN [48]	74.5
Ours, ×1	86.93
Ours, ×3	90.21
Ours, ×5	**91.54**
DANN [13]	77.4
DSN [4]	83.2
PixelDA [3]	**95.9**
Target-only	96.5

(b) Cropped LineMod

Model	Classification Accuracy (%)	Mean Angle Error (°)
Source-only	42.9 (47.33)	73.7 (89.2)
CycleGAN [48]	68.18	47.45
Ours, ×1	95.91	42.06
Ours, ×3	97.04	37.35
Ours, ×5	**98.12**	**34.4**
DANN [13]	99.9	56.58
DSN [4]	**100**	53.27
PixelDA [3]	99.98	**23.5**
Target-only	100	12.3 (6.47)

Evaluation of Diversity

- **Learned Perceptual Image Patch Similarity (LPIPS)**
  - Deep features outperform all previous metrics by large margins.
  - Perceptual similarity is an emergent property shared across deep visual representations.



- Higher **LPIPS** implies better sample quality
  - [1] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep networks as a perceptual metric. In: CVPR (2018)

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

Human Evaluation (User Study)

- As mentioned in SRGAN, machine evaluations explained above has limitations.
- To accurately evaluate the perceptual quality, human evaluation is a better choice
- Ranking: Invite human to give rankings for a group of images
- Contrast: Invite human to choose the better one out of a pair of two images
- Amazon Mechanical Turk

Ranking

- Method
  - 1. For each testing image, a group of synthesised images from different methods (as well as the original image) are presented to the subjects without telling them which images were from which method.
  - 2. The subjects are required to rank the group of synthesised Images based on the given criteria.
  - 3. For each testing image, different methods are ranked starting from 1 for the best, and the ranking is allowed to be tied.
  - 4. Average all human rankings to calculate the scores.

- Lower score indicates higher perceptual quality

Contrast

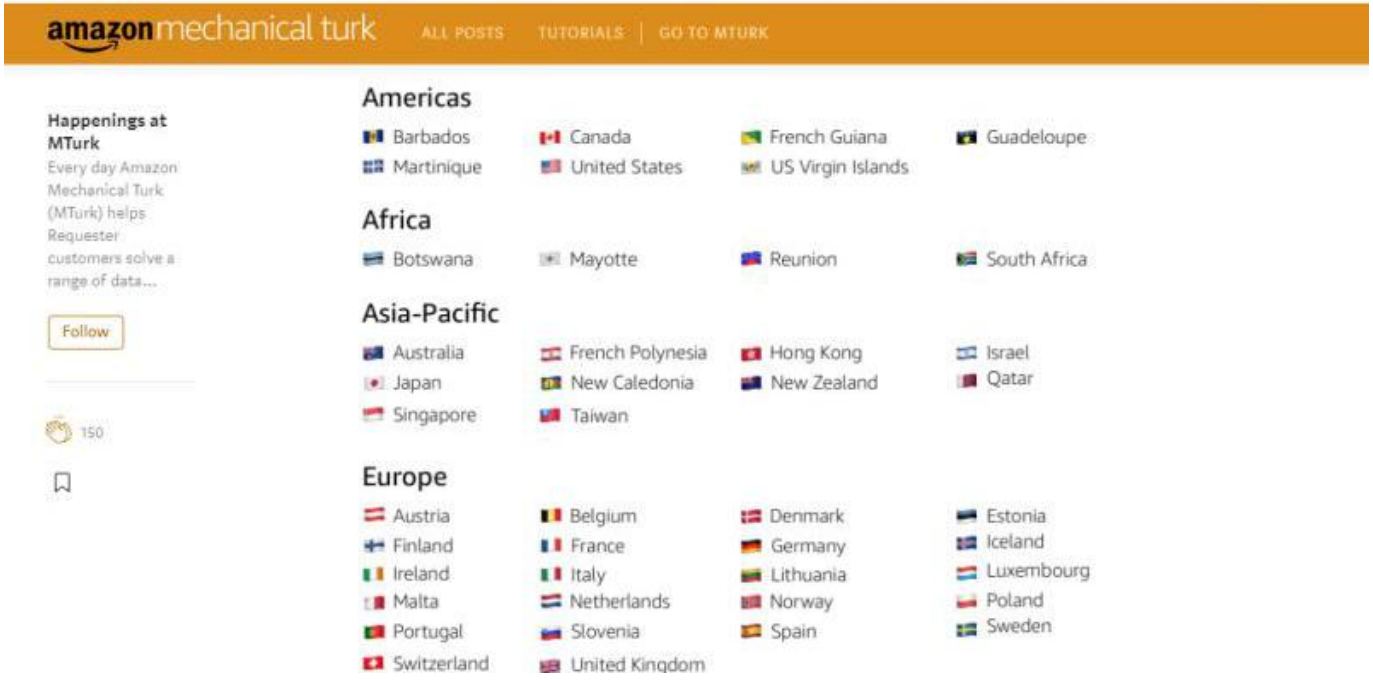
- Method
  - 1. Present the users with image pairs side by side on the screen.
  - 2. For each pair, one image is generated by your own model while the other image is generated by a randomly picked baseline model.
  - 3. To make users not know from which model the images are generated, the image positions are random in the pair, i.e., the image generated by your model on the left or right is of equal possibility.
  - 4. Calculate the ratio of users who favor the results of your model to users who favor a certain competing method.
  - 5. The ratio greater than one indicates your model is better.

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

Amazon Mechanical Turk (AMT)

- Amazon Mechanical Turk is a system that allows humans to complete microtasks on Amazon's platform for money.

- Warning: It can't be used in mainland of China, even through VPN!



Amazon Mechanical Turk (AMT)

- The results of user study through AMT is admitted internationally.
- You' d better ask a foreign friend to help you to send the questionnaires

- A few tips for fair and high-quality comparison:
  - Give unlimited time to the workers to make the selection
  - Each group/pair of images is compared by 5 different workers.
  - Only approve workers with a life-time task approval rate greater than 98% to participate in the evaluation.

Create Questionnaires

- Your questionnaire is released to workers from all over the world, so write it in English.
- Workers access your questionnaire through your link on the AMT.
- You can write a website on a foreign server (e.g. Google Cloud Platform).
- A website is convenient to collect the answers of the workers.
- The website can be written in javascript.
- The last question of the questionnaire should let the worker to fill in his Worker ID provided by AMT to verify his completeness.

Summary: Sampling Quality

- Known Ground Truth
  - SRGAN
  - -MSE PSNR SSIM
- Unknown Ground Truth
  - DRIT
  - StarGAN
  - - Classification
  - - IS FID KID
  - - LPIPS
- Human Evaluation
  - Ranking v.s. Contrast
  - Tools: AMT

北京大學
PEKING UNIVERSITY

Thanks