MEGVII 旷视

Real World GAN

Applications in Computer Vision

shuchang.zhou@gmail.com

Oct. 2020

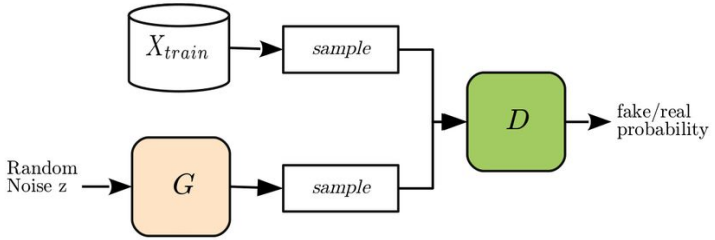
What's GAN?

https://github.com/deepfakes/faceswap

Watch 1.5k Star 32.7k Fork 10.2k

Generative Adversarial Nets - NIPS Proceedings

Generative stochastic **networks** [4] are an example of a **generative** ma
trained with exact backpropagation rather than the numerous ap-.
by I Goodfellow · 2014 · Cited by 23080 · Related articles



GAN is related to

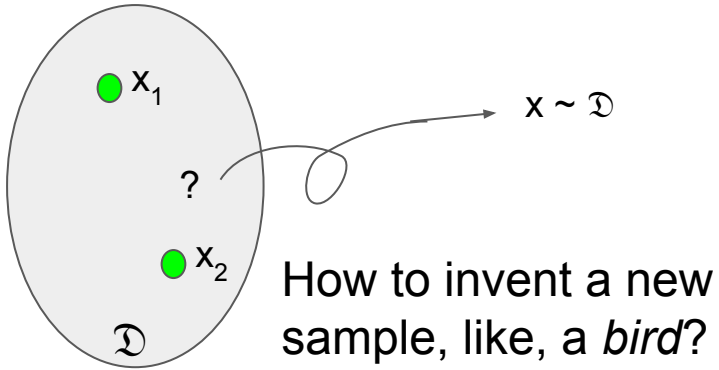
- AutoEncoder
- Metric Learning
- Game Theory
- Adversary Training

Applications of GAN

- Characterizing a Distribution
  - Minimizing distance between distributions
  - Inducing structure in Feature Space: Disentangled Representation
  - Pyramid Structure
- Inventing Plausible Details
  - Context-aware Image Editing and Inpainting
- Exploiting Unpaired Data with Cycle Consistency
  - Translation between Unpaired Data
  - Examples: Text, Graphics primitives, Degraded Images
- Exact Object Transfiguration
  - Object Swapping and Information Preservation

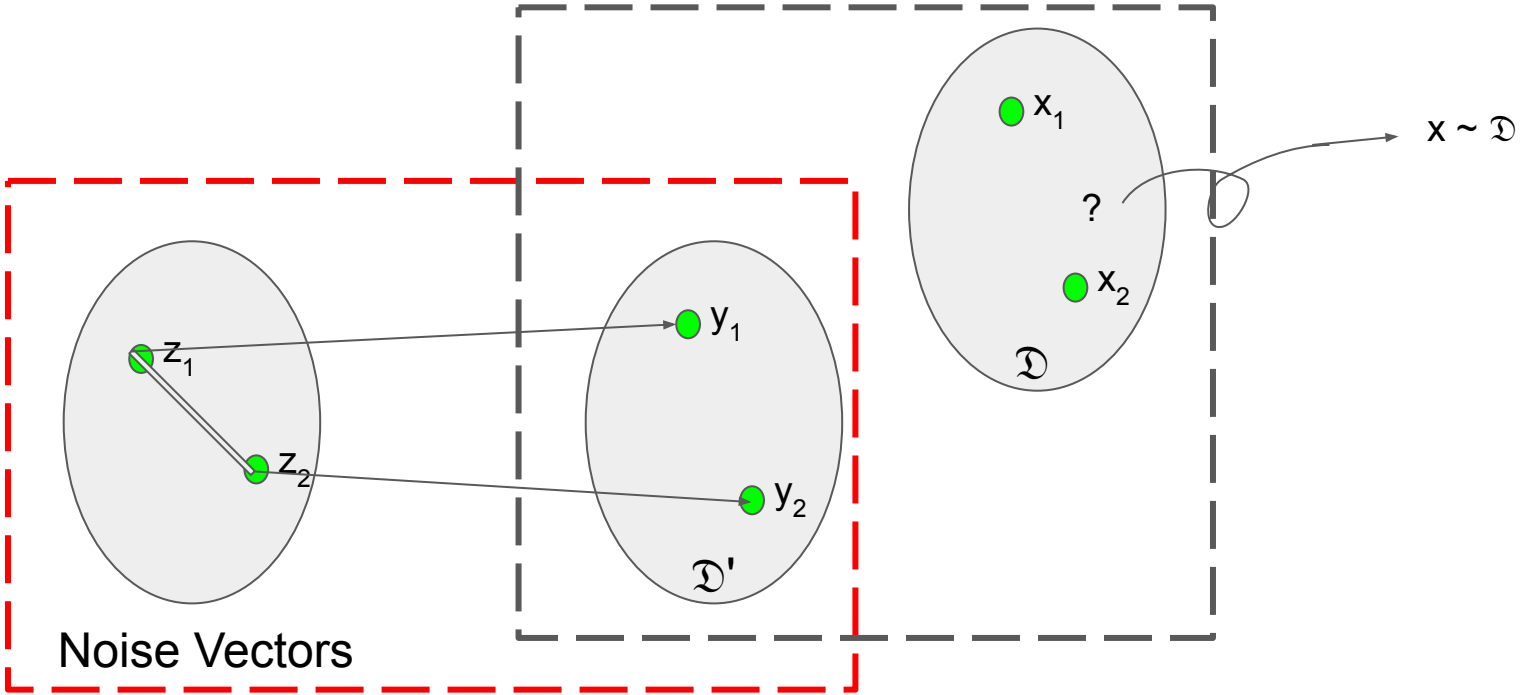
"New" Samples from a Distribution

- Generation task: sampling from a distribution characterized by exemplars
  - Given {x_i | x_i ~ $\mathfrak{D}$}, sample a new x ~ $\mathfrak{D}$ (*unconditional*)
  - Given {x_i | x_i ~ $\mathfrak{D}$}, sample a new x ~ $\mathfrak{D}$, and P(x) = c (*conditional*), P(x) = c are constraints to satisfy
- *How to know $\mathfrak{D}$?*



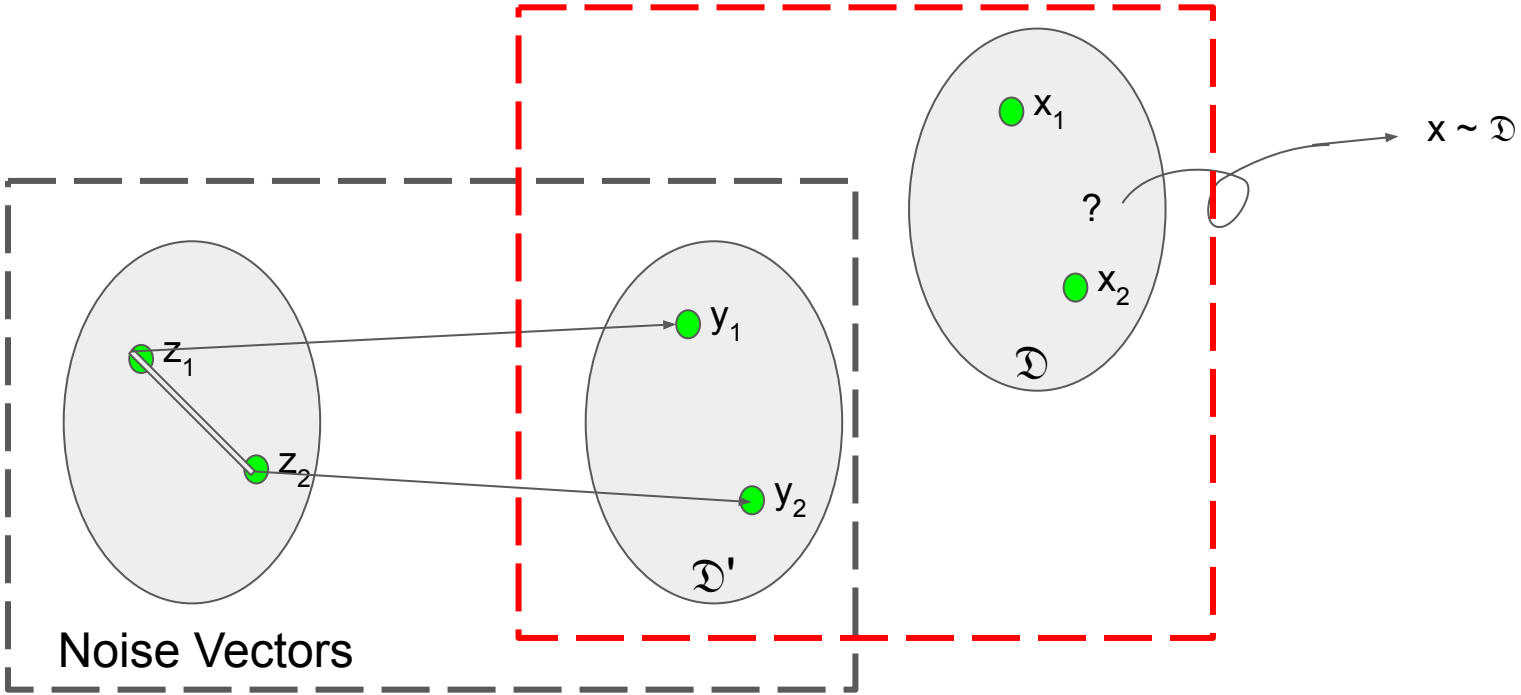
How to invent a new sample, like, a *bird*?

GAN = noise-to-image mapping (G) + adversary discriminator loss (D)



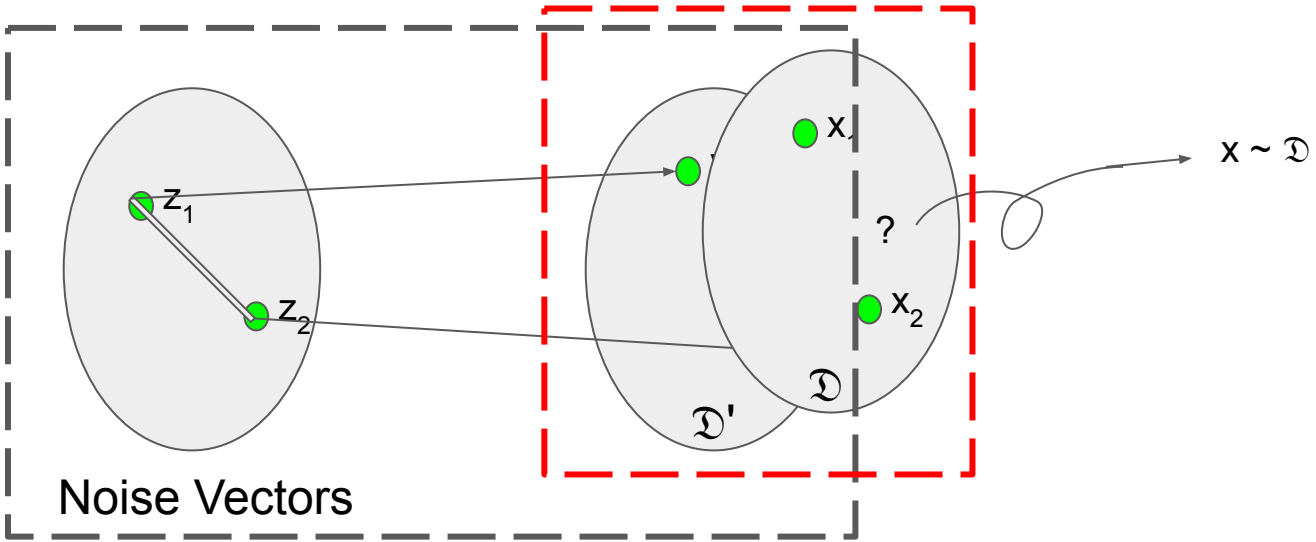
Generator G produces $\mathfrak{D}'$

GAN = noise-to-image mapping (G) + adversary discriminator loss (D)



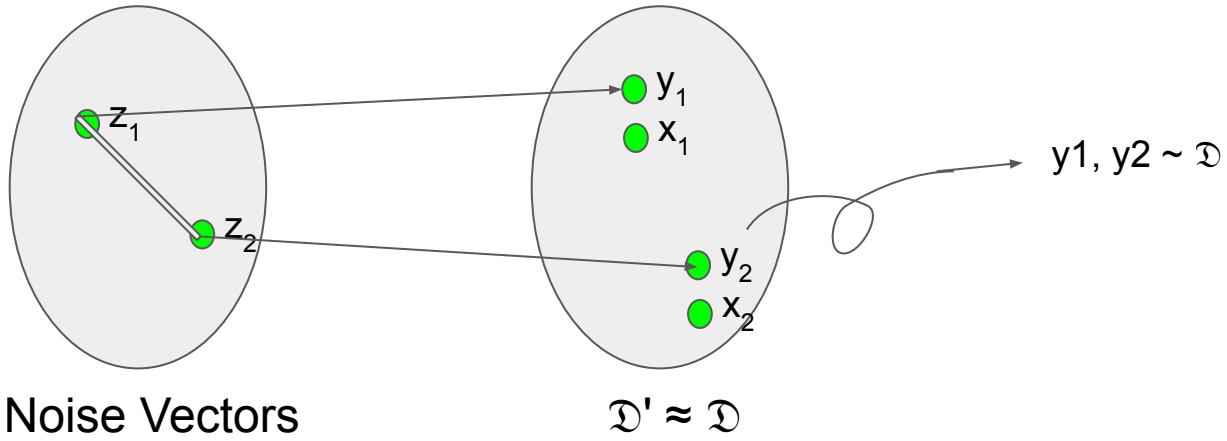
Discriminator D brings 𝔇' close to 𝔇

GAN = noise-to-image mapping (G)
+ adversary discriminator loss (D)



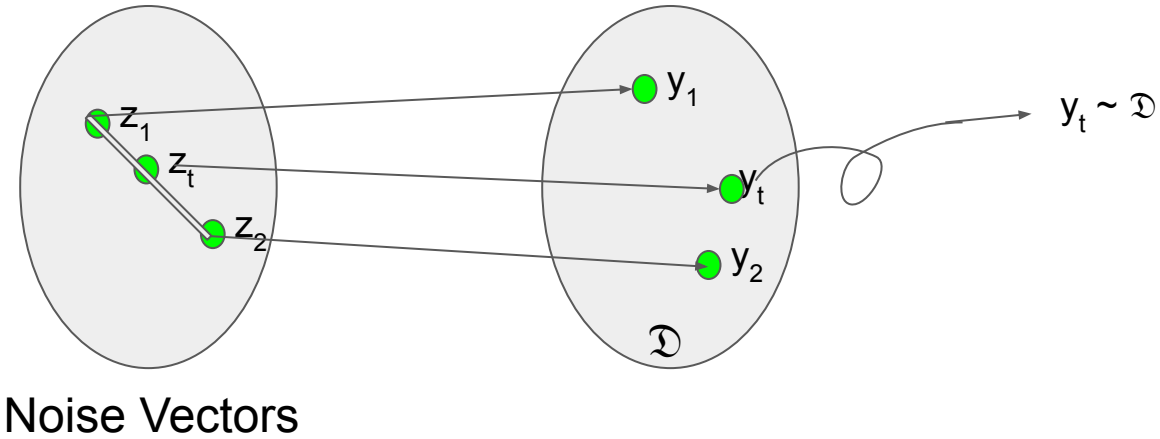
Discriminator D brings $\mathfrak{D}'$ close to $\mathfrak{D}$

When $\mathfrak{D}' \approx \mathfrak{D}$, the noise-to-image map induces structure in Image Space $\mathfrak{D}$

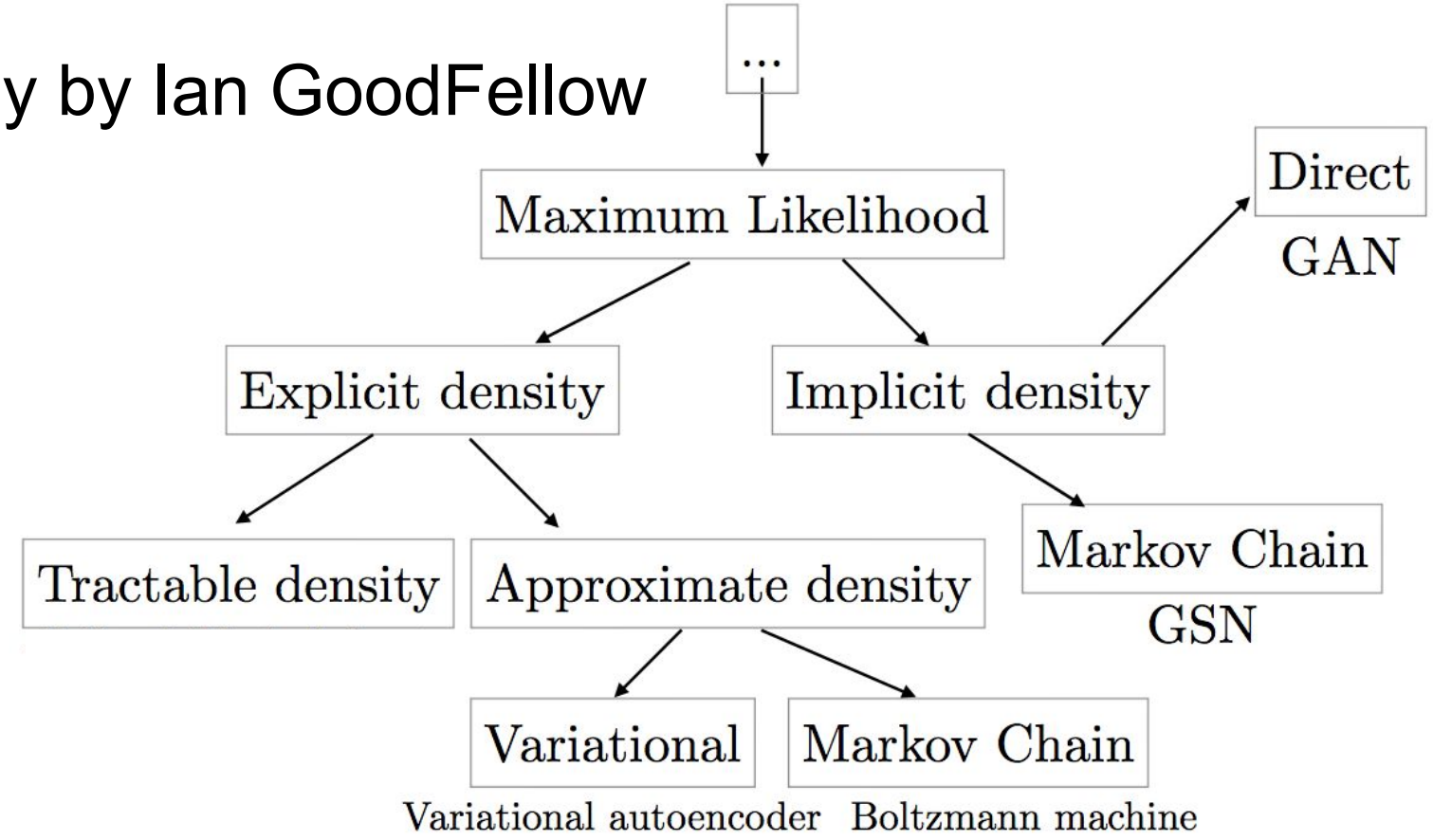


When $\mathfrak{D}' \approx \mathfrak{D}$, the noise-to-image map induces structure in Image Space $\mathfrak{D}$

- Noise as source helps avoiding trivial solution
  - mapping not constant: $\mathfrak{D}'$ contains multiple images
  - still suffer from mode collapse
- Sampling in Feature Space => Sampling in Image Space

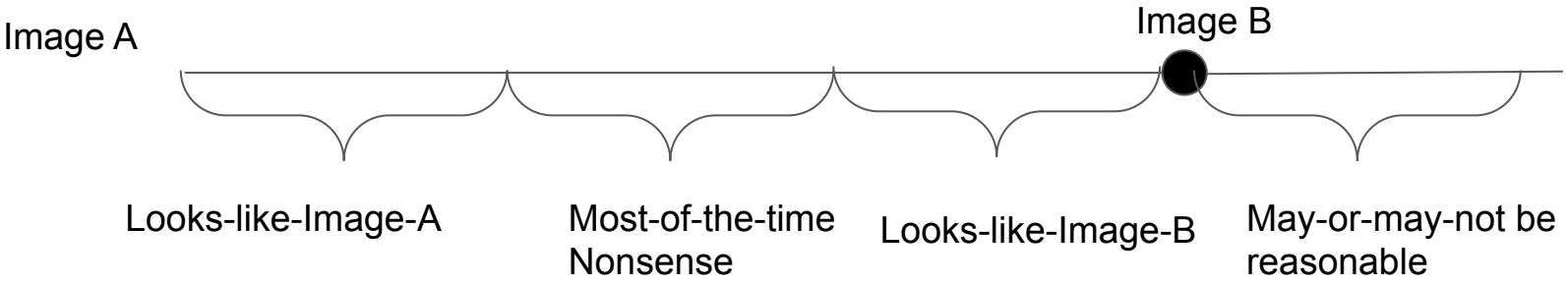


Taxonomy by Ian GoodFellow



Feature Space: Geometry

- Geometry: interpolation is safe, extrapolation is outrageous



- Geometry: disentangled representation, admits vector arithmetic

Feature Space Interpolation: one dimensional

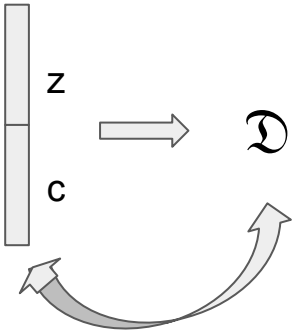
BigGAN '18

Feature Space Interpolation: two dimensional

GeneGAN '17

Feature Space Decomposition: Disentangled Representation by maximizing Mutual Information

InfoGAN ('16)

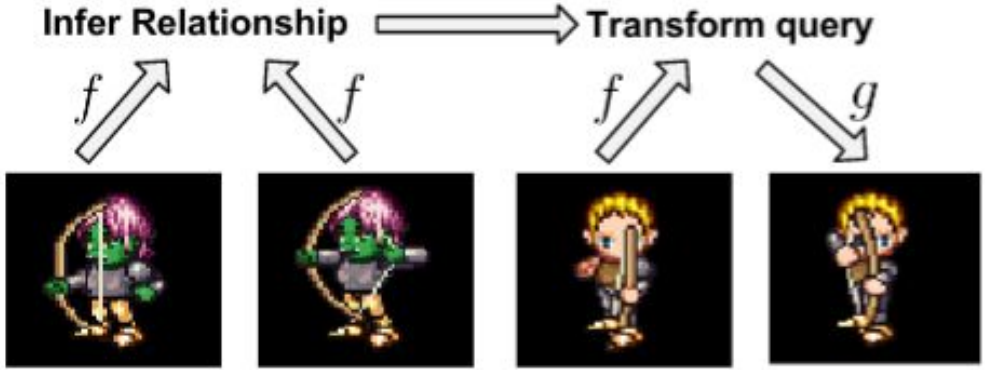


(a) Azimuth (pose) (b) Elevation

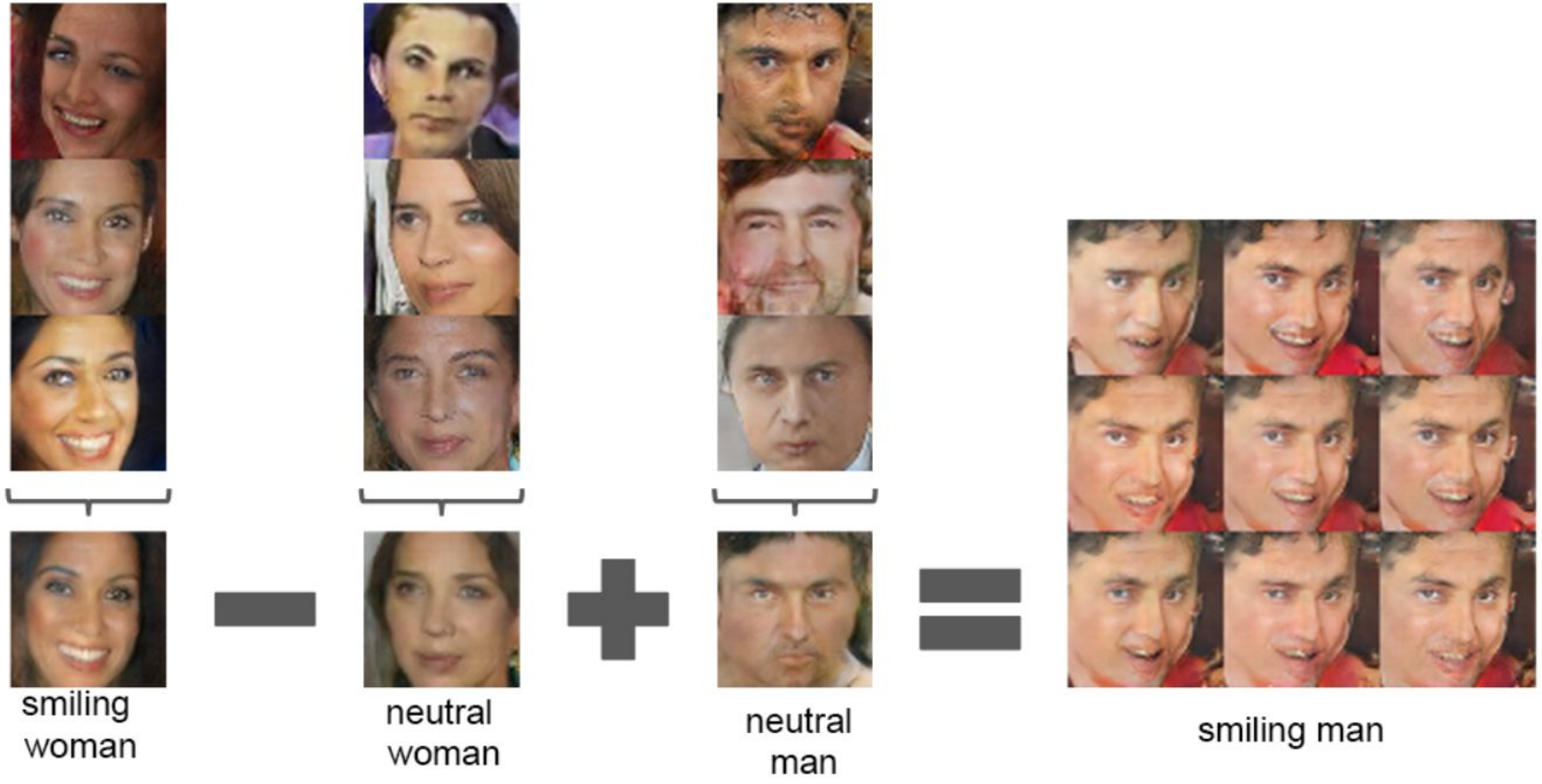
(c) Lighting (d) Wide or Narrow

Disentangled Feature Space: Vector Arithmetic

- Visual Analogy-Making
  - open-bow vector = φ(a person opening bow) - φ(the same person not opening bow)
  - Can transfer the exact "open-bow" to another person (requires paired data, will see later how to remove this requirement)



Deep Visual Analogy-Making '15



DCGAN '16

Feature Space: Disentangling Factors Approach

- Does not require availability of four-tuple training data
- For training, still require availability of paired data
  - Same person with multiple viewpoints (Background: person, Object: viewpoint)
  - Find out which part of feature contains the viewpoints information by statistical independence

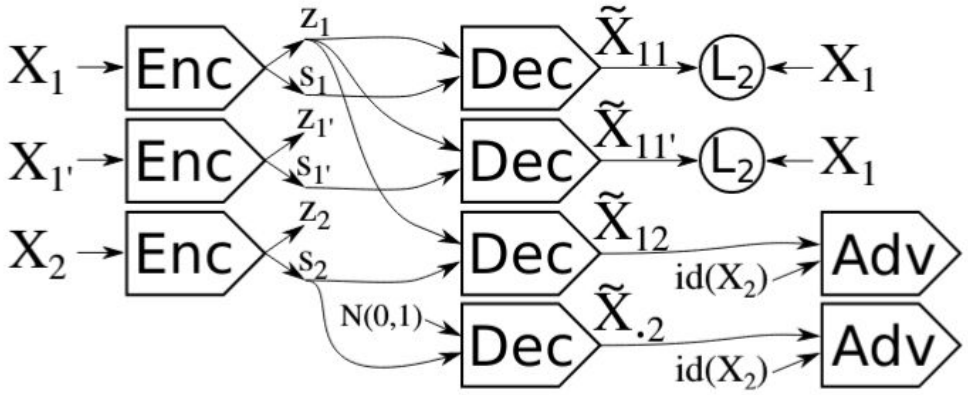


Figure 1: Training architecture. The inputs x_1 and x_1' are two different samples with the same label, whereas x_2 can have any label.

Failed Cases: Entangled Features

- Still hard
- SEAN '20
  - Change of style also changes gender

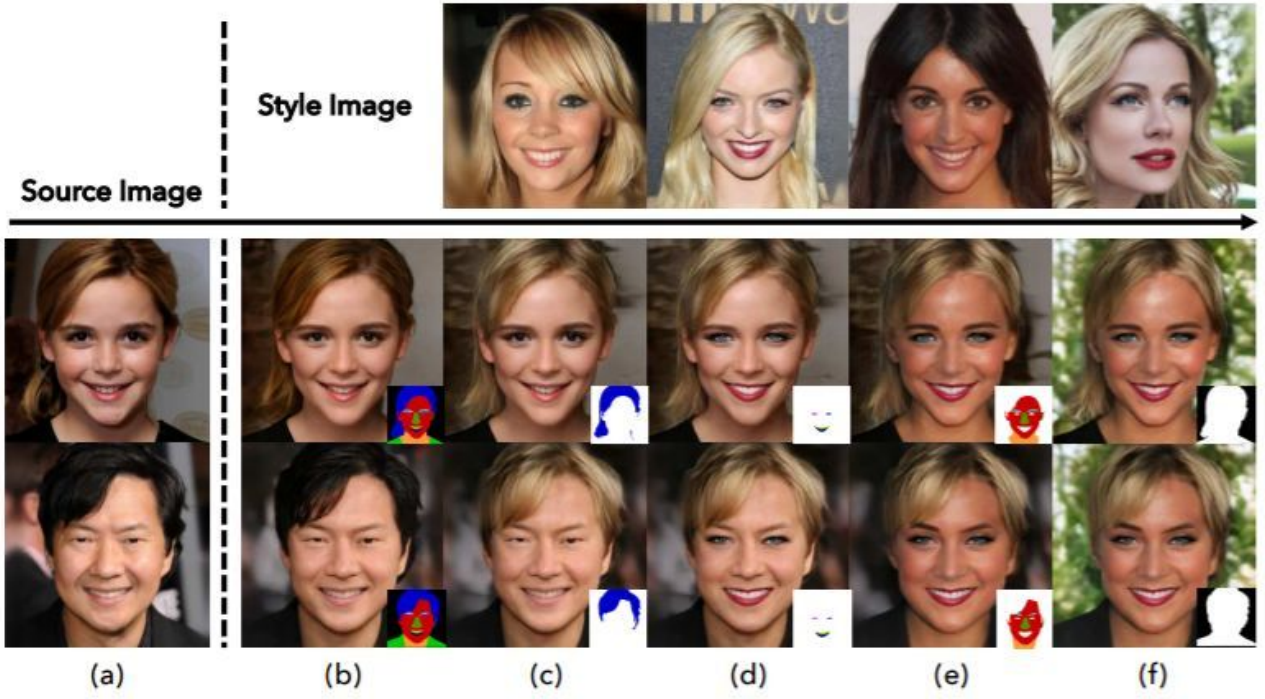


Figure 1: Face image editing controlled via style images and segmentation masks. a) source images. b) reconstruction of the source image; segmentation mask shown as small inset. c - f) four separate edits; we show the image that provides new style information on top and show the part of the segmentation mask that gets edited as small inset. The results of the successive edits are shown in row two and three. The four edits change hair, mouth and eyes, skin tone, and background, respectively.

Image Space: Long-Range Dependence leads to Structure

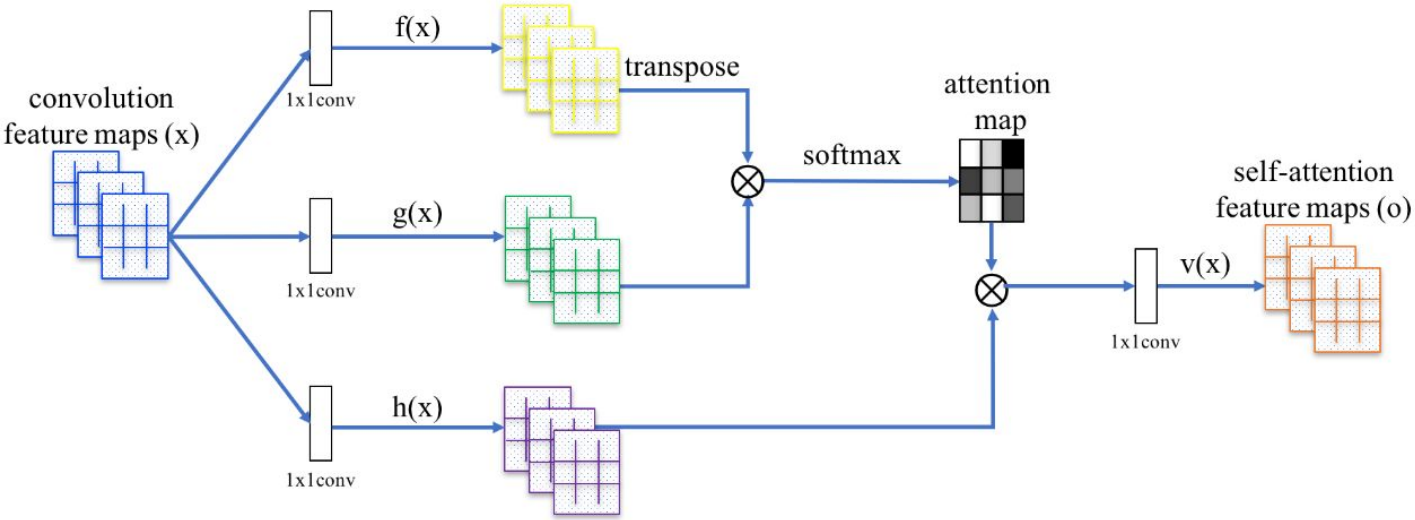
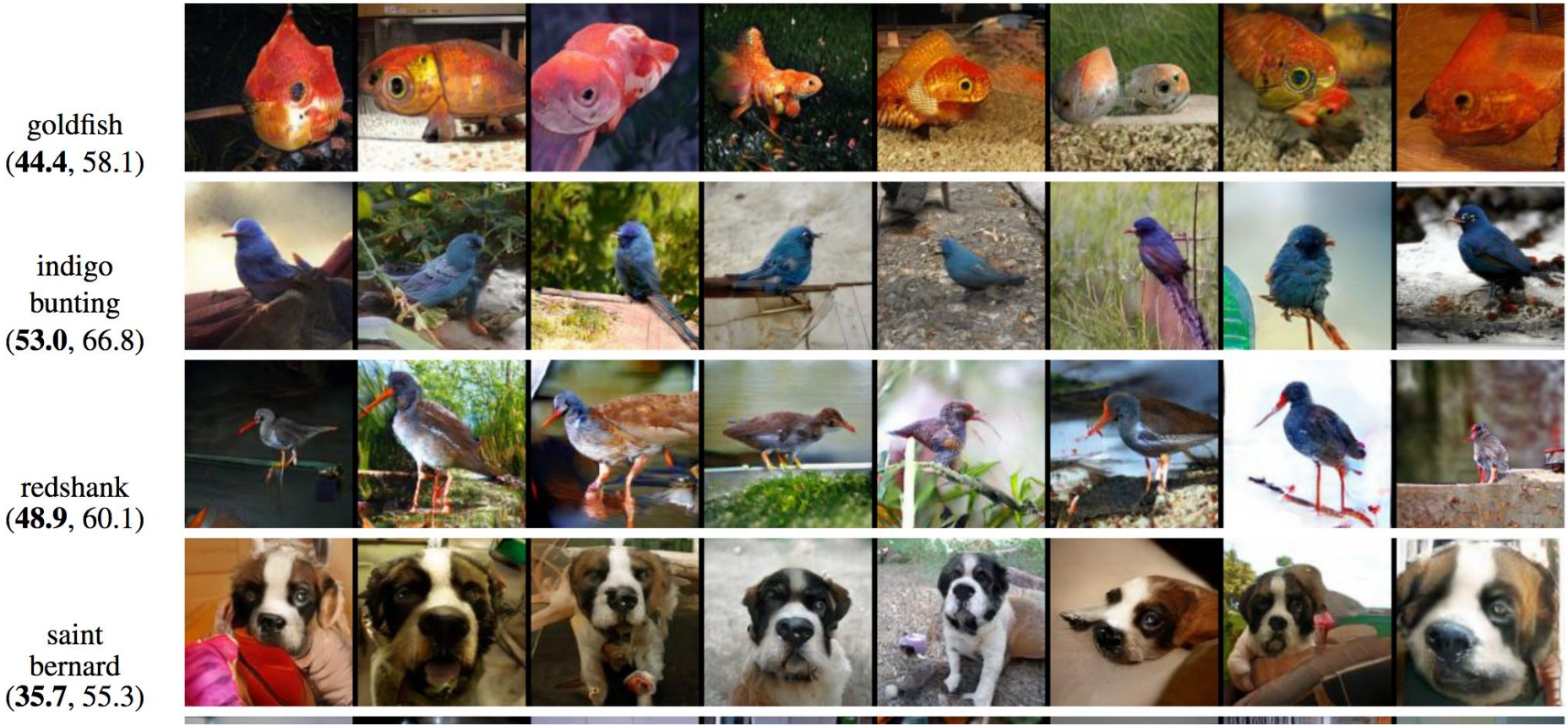


Figure 2. The proposed self-attention module for the SAGAN. The ⊗ denotes matrix multiplication. The softmax operation is performed on each row.

SAGAN '18


goldfish
(**44.4**, 58.1)
indigo
bunting
(**53.0**, 66.8)
redshank
(**48.9**, 60.1)
saint
bernard
(**35.7**, 55.3)

SAGAN '18

Image Space: Pyramid Structure Improves Details

LapGAN '15

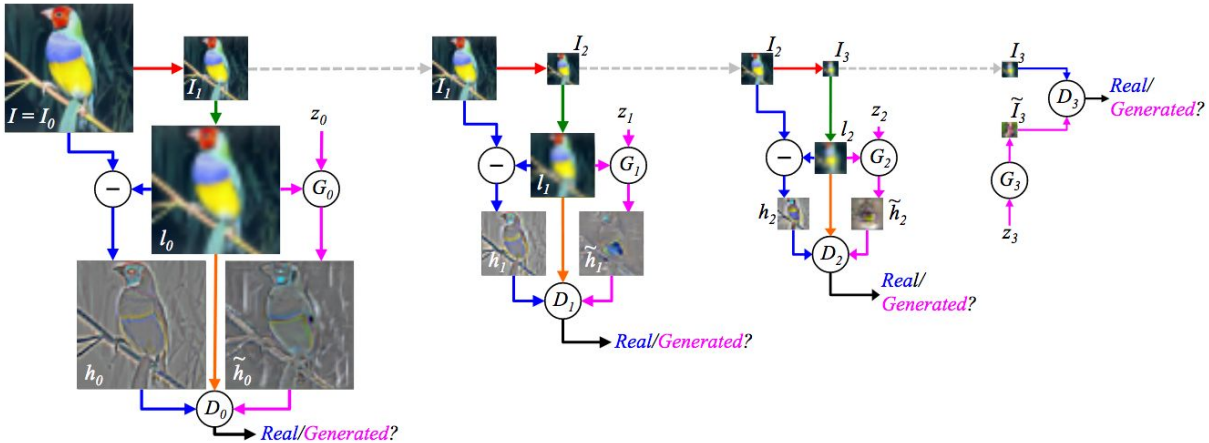


Figure 2: The training procedure for our LAPGAN model. Starting with a 64x64 input image I from our training set (top left): (i) we take $I_0 = I$ and blur and downsample it by a factor of two (red arrow) to produce I_1; (ii) we upsample I_1 by a factor of two (green arrow), giving a low-pass version l_0 of I_0; (iii) with equal probability we use l_0 to create *either* a real *or* a generated example for the discriminative model D_0. In the real case (blue arrows), we compute high-pass $h_0 = I_0 - l_0$ which is input to D_0 that computes the probability of it being real vs generated. In the generated case (magenta arrows), the generative network G_0 receives as input a random noise vector z_0 and l_0. It outputs a generated high-pass image $\tilde{h}_0 = G_0(z_0, l_0)$, which is input to D_0. In both the real/generated cases, D_0 also receives l_0 (orange arrow). Optimizing Eqn. 2, G_0 thus learns to generate realistic high-frequency structure $\tilde{h}_0$ consistent with the low-pass image l_0. The same procedure is repeated at scales 1 and 2, using I_1 and I_2. Note that the models at each level are trained independently. At level 3, I_3 is an 8×8 image, simple enough to be modeled directly with a standard GANs G_3 & D_3.

Image Space: Pyramid Structure Improves Details

ProGAN '17

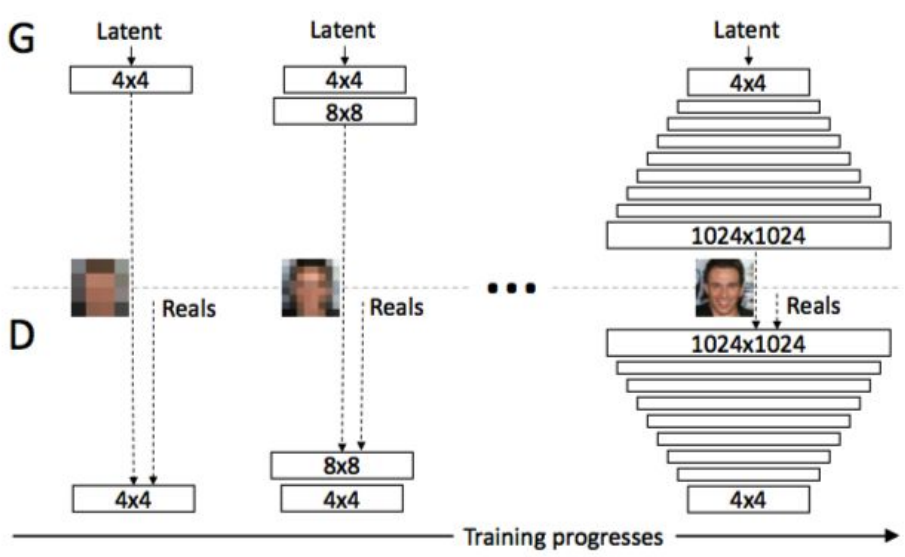
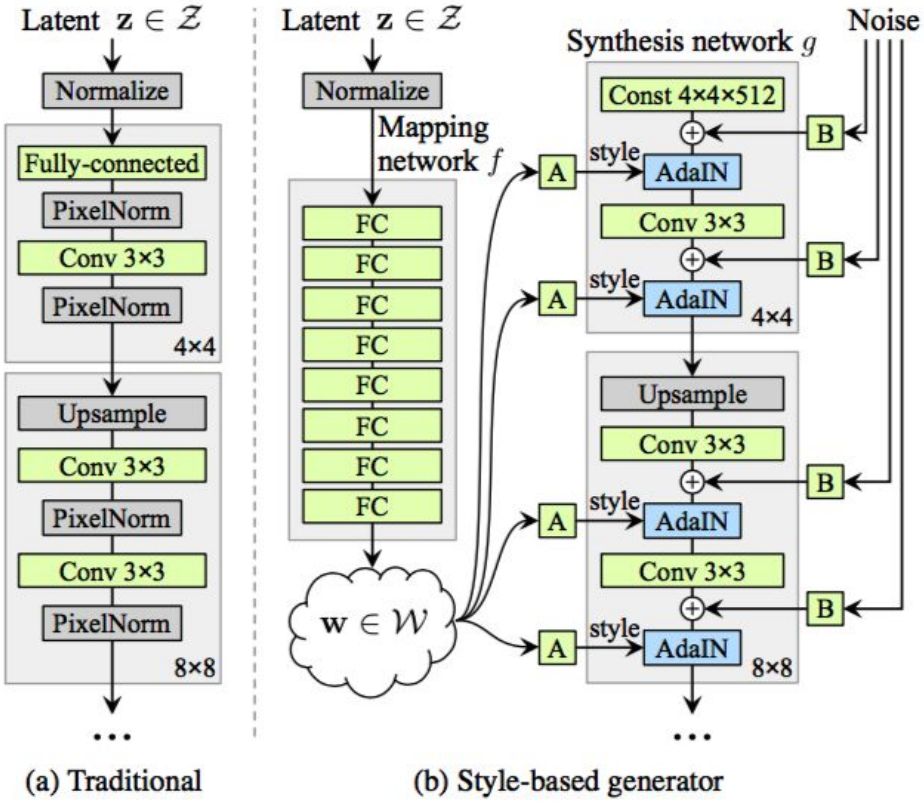


Figure 1: Our training starts with both the generator (G) and discriminator (D) having a low spatial resolution of 4×4 pixels. As the training advances, we incrementally add layers to G and D, thus increasing the spatial resolution of the generated images. All existing layers remain trainable throughout the process. Here $N \times N$ refers to convolutional layers operating on $N \times N$ spatial resolution. This allows stable synthesis in high resolutions and also speeds up training considerably. One the right we show six example images generated using progressive growing at 1024×1024.

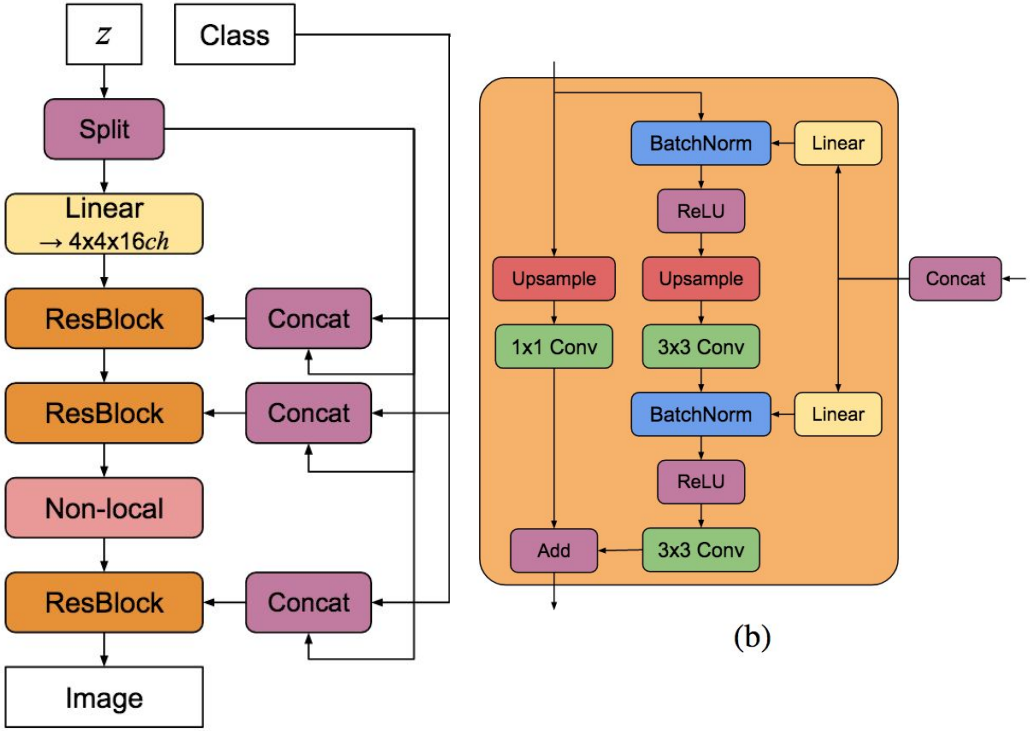
StyleGAN '18

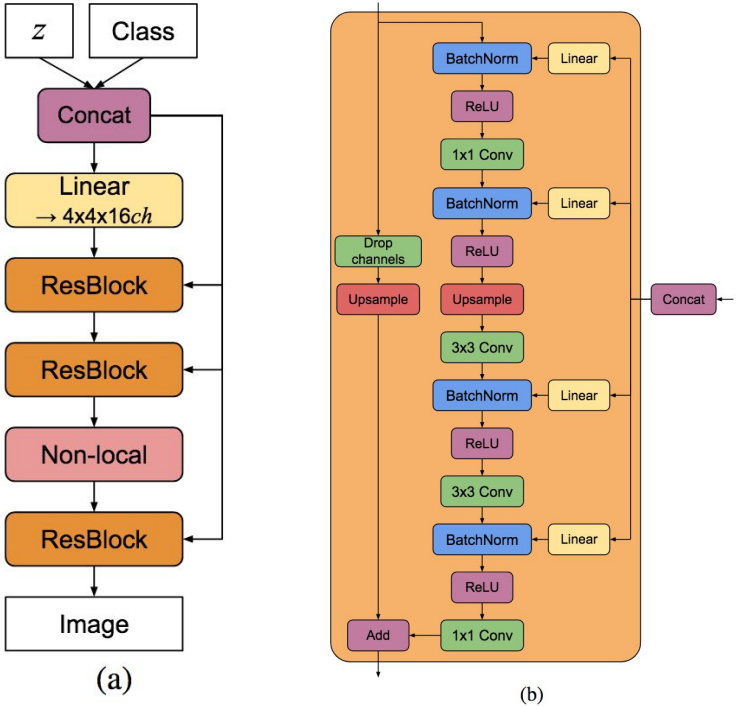

Latent $\mathbf{z} \in \mathcal{Z}$
Normalize
Fully-connected
PixelNorm
Conv 3×3
PixelNorm
4×4
Upsample
Conv 3×3
PixelNorm
Conv 3×3
PixelNorm
8×8
Latent $\mathbf{z} \in \mathcal{Z}$
Normalize
Mapping network f
FC
FC
FC
FC
FC
FC
FC
FC
$\mathbf{w} \in \mathcal{W}$
Synthesis network g
Noise
Const 4×4×512
B
style
AdaIN
A
Conv 3×3
4×4
Upsample
8×8
(a) Traditional
(b) Style-based generator

StyleGAN '18

Text-like scribbles

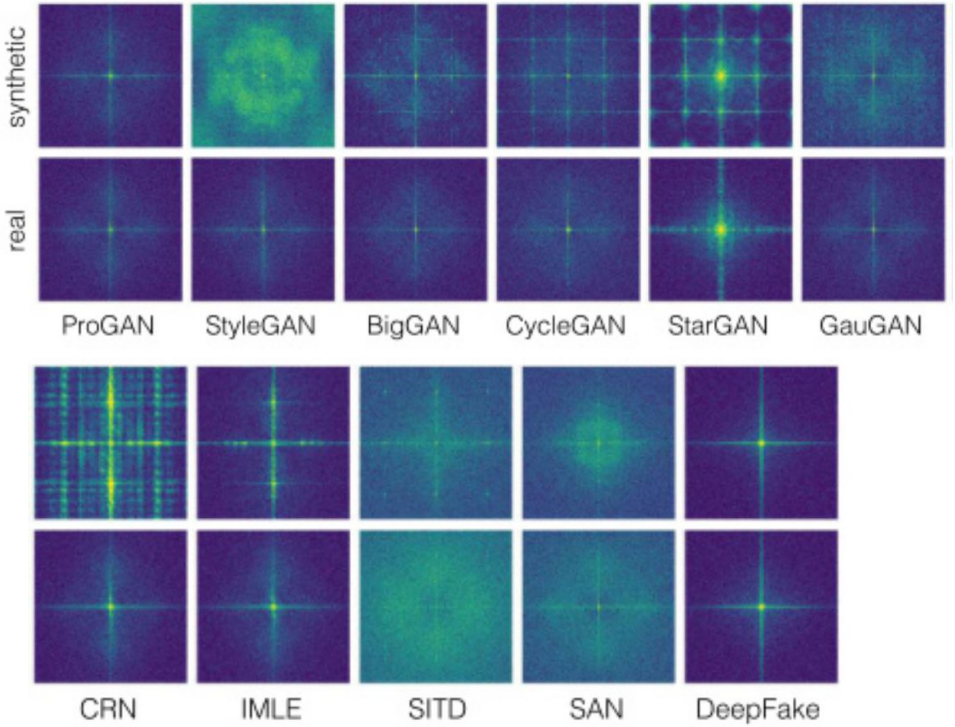
Class Injected to Pyramid Structure





BigGAN '18

Failed cases: Spectral Artifacts despite U-net Structure



CNN-generated images are surprisingly easy to spot… for now '19

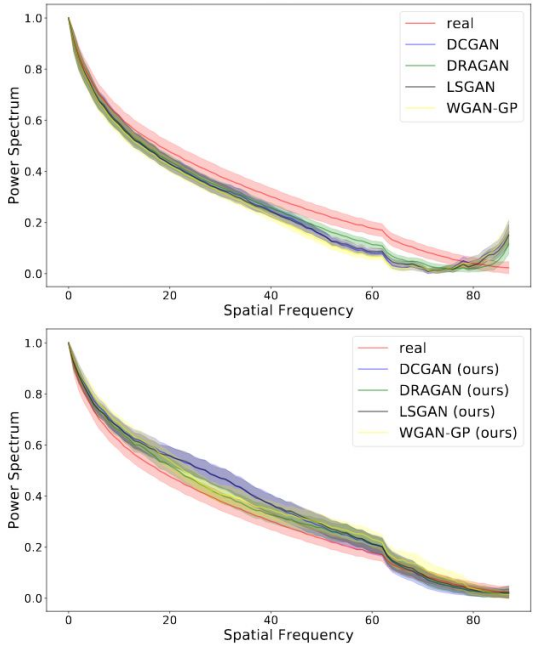


Figure 1: Common up-convolution methods are inducing heavy spectral distortions into generated images. The **top** figure shows the statistics (mean and variance) after azimuthal integration over the power-spectrum (see Section 2.1) of real and GAN generated images. Evaluation on the *CelebA* [34] data set, here all GANs (DCGAN [47], DRAGAN [32], LSGAN [37], WGAN-GP [20]) are using "transposed convolutions" (see Section 2.2) for up-sampling. **Bottom:** Results of the same experiments as above, adding our proposed spectral loss during GAN training.

Watch your up-convolutions '20

Applications of GAN

- Characterizing a Distribution
  - Minimizing distance between distributions
  - Inducing structure in Feature Space: Disentangled Representation
  - Pyramid Structure
- Inventing Plausible Details
  - Context-aware Image Editing and Inpainting
- Exploiting Unpaired Data with Cycle Consistency
  - Translation between Unpaired Data
  - Examples: Text, Graphics primitives, Degraded Images
- Exact Object Transfiguration
  - Object Swapping and Information Preservation

Conditional Image Generation: from labels

- Conditioned by labels
- Note! In practice, the invented details are not that useful for classification

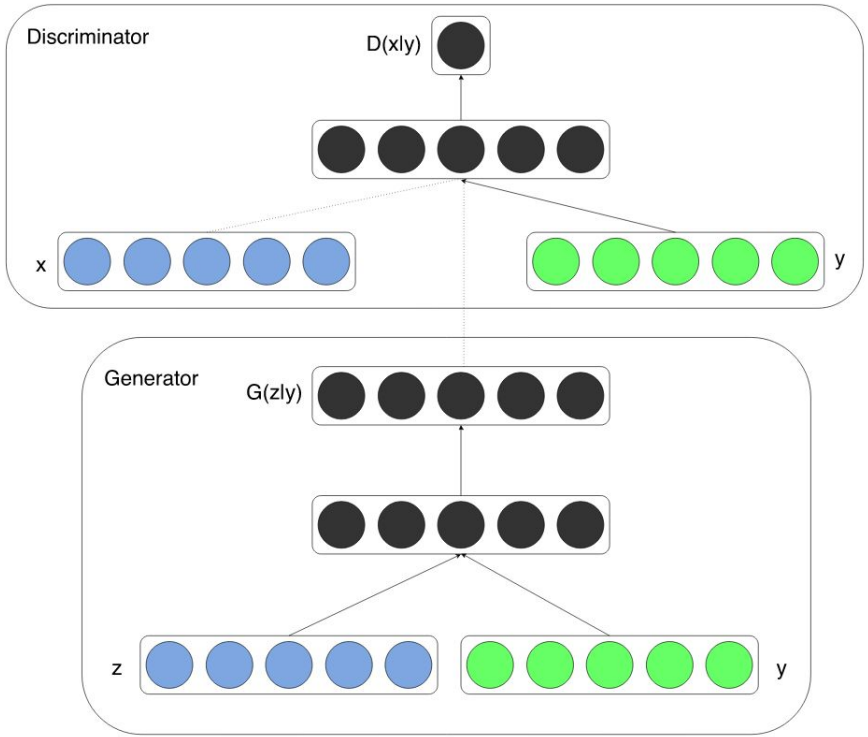


Table 2: CAS for VQ-VAE-2 model reconstructions and BigGAN-deep models at different truncation levels at 256×256 resolution.

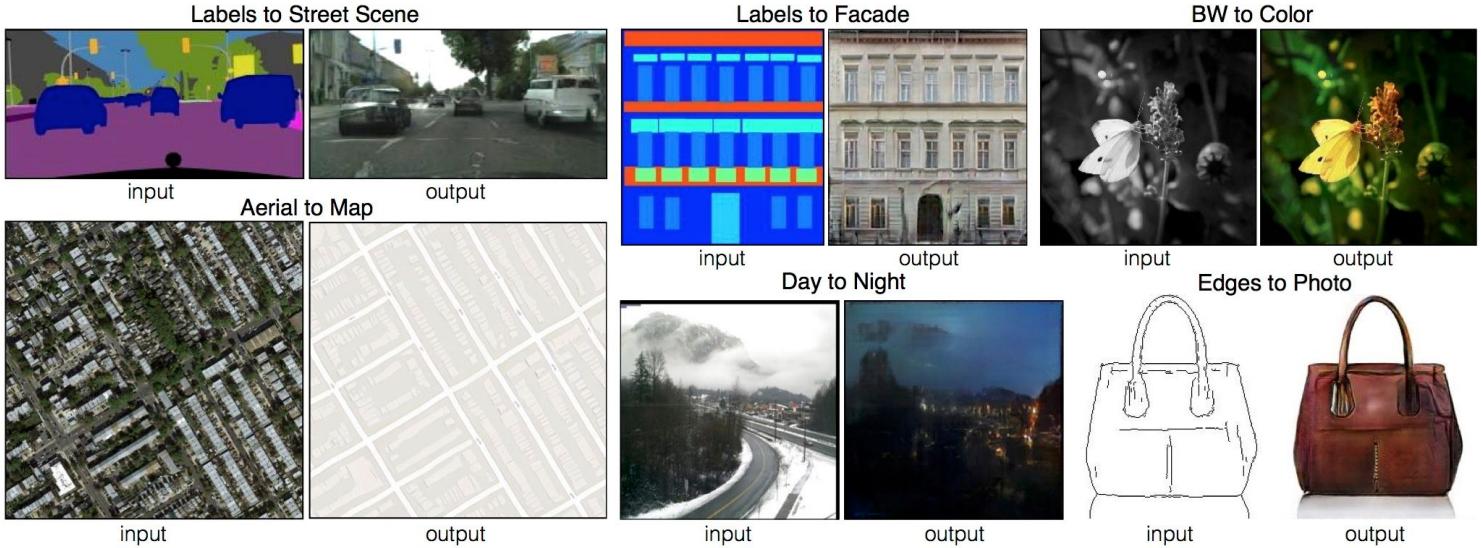
Training Set	Truncation	Top-5 Accuracy	Top-1 Accuracy	IS	FID-50K
BigGAN-deep	0.20	13.24%	5.11%	**339.06 ± 3.14**	20.75
BigGAN-deep	0.42	28.68%	13.30%	324.62 ± 3.29	15.93
BigGAN-deep	0.50	32.88%	15.66%	316.31 ± 3.70	14.37
BigGAN-deep	0.60	45.01%	25.51%	299.51 ± 3.20	12.41
BigGAN-deep	0.80	56.68%	32.88%	258.72 ± 2.86	9.24
BigGAN-deep	1.00	62.97%	39.07%	214.64 ± 2.01	**7.42**
BigGAN-deep	1.50	65.92%	42.65%	109.39 ± 1.56	11.78
BigGAN-deep	2.00	64.37%	40.98%	49.54 ± 0.98	28.67
VQ-VAE-2 reconstructions	-	**89.46%**	**69.90%**	203.89 ± 2.55	8.69
Real	-	91.47%	73.09%	331.83 ± 5.00	2.47

Classification Accuracy Score for Conditional Generative Models

BigGAN-deep '19

Conditional Image Generation from segmentations

- Conditioned by Segmentations (Silhouette)
  - Can also use degraded images as conditions (superresolution)
- In practice: successful!



Pix2pix '17

Generative Image Inpainting with Contextual Attention '18

Figure 1: Example inpainting results of our method on images of natural scene, face and texture. Missing regions are shown in white. In each pair, the left is input image and right is the direct output of our trained generative neural networks without any post-processing.

"AI bias": Invented Details may not be Useful

PULSE '20

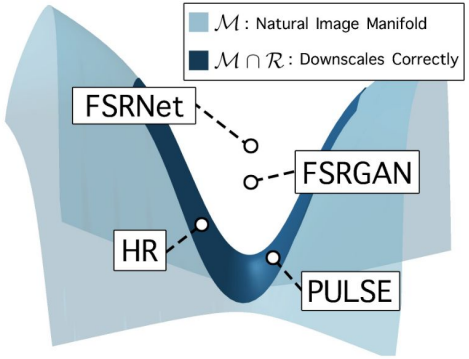
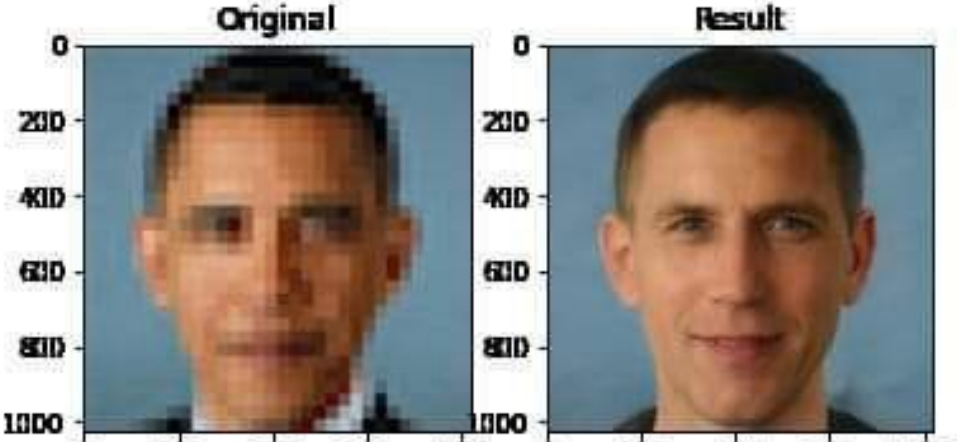


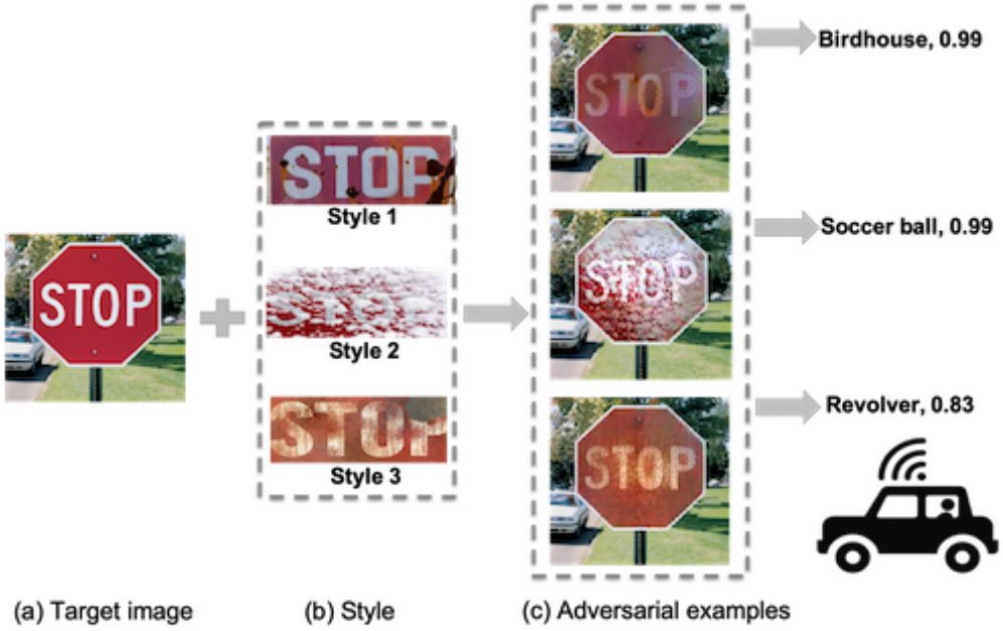
Figure 2. FSRNet tends towards an average of the images that downscale properly. The discriminator loss in FSRGAN pulls it in the direction of the natural image manifold, whereas PULSE always moves along this manifold.



Race							Gender	
Black	East Asian	Indian	Latino/Hispanic	Middle Eastern	Southeast Asian	White	Female	Male
79.2%	87.0%	87.4%	90.2%	87.0%	87.4%	83.4%	91.4%	88.6%

Table 3. Success rates (frequency with which PULSE finds an image in the outputs of the generator that downscales correctly) of PULSE with StyleGAN-FFHQ across various groups, evaluated on FairFace. See "Failure to converge" in Section 6 for full explanation of this analysis and its limitations.

Inventing "Adversary" Details



Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles '20

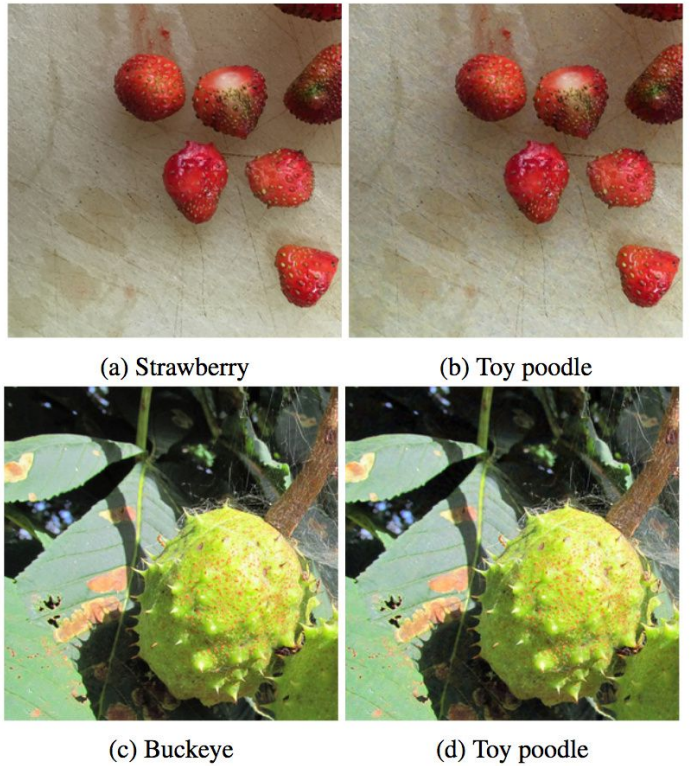


Figure 4: Examples from an ImageNet-compatible set, and the labels denote corresponding classification results Left: original benign images; right: adversarial images generated by AdvGAN against Inception_v3.

Generating Adversarial Examples with Adversarial Networks '18

Graphics Primitive to Image

- DC-IGN '15

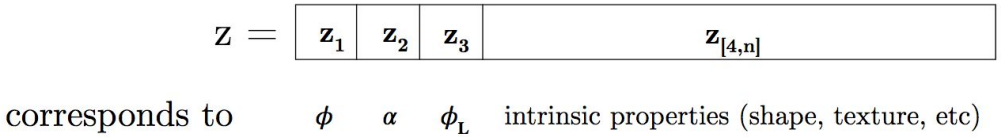
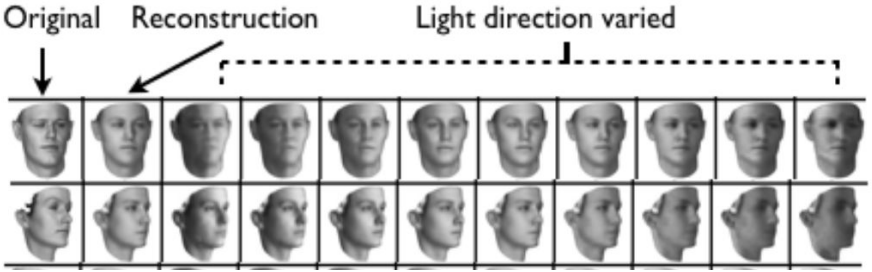
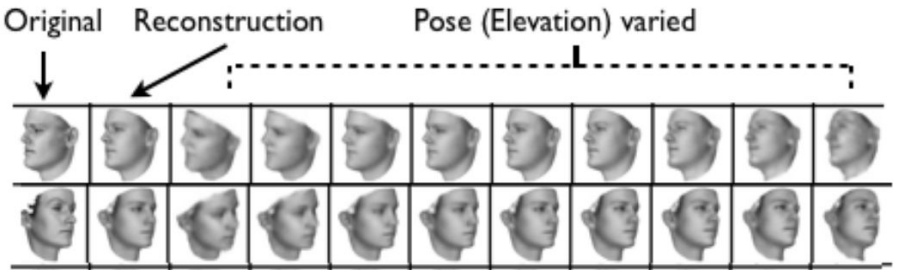


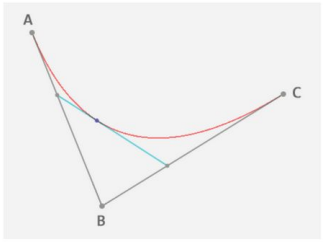
Figure 2: **Structure of the representation vector.** ϕ is the azimuth of the face, α is the elevation of the face with respect to the camera, and ϕ_L is the azimuth of the light source.





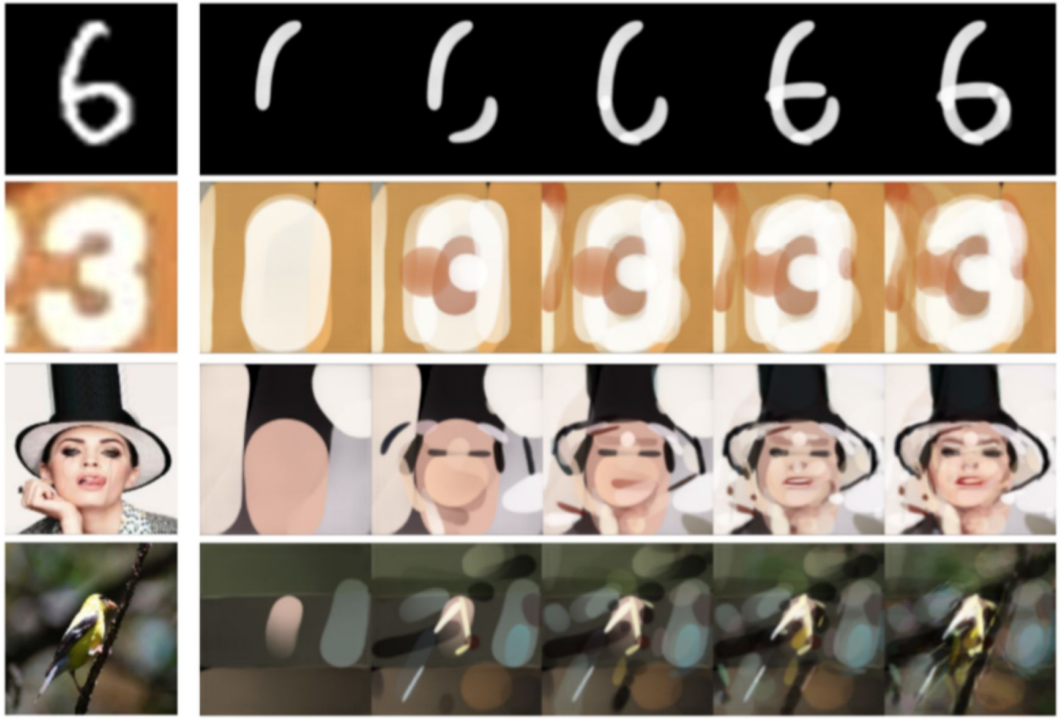
Graphics Primitive to Image

- Learning-to-paint '19



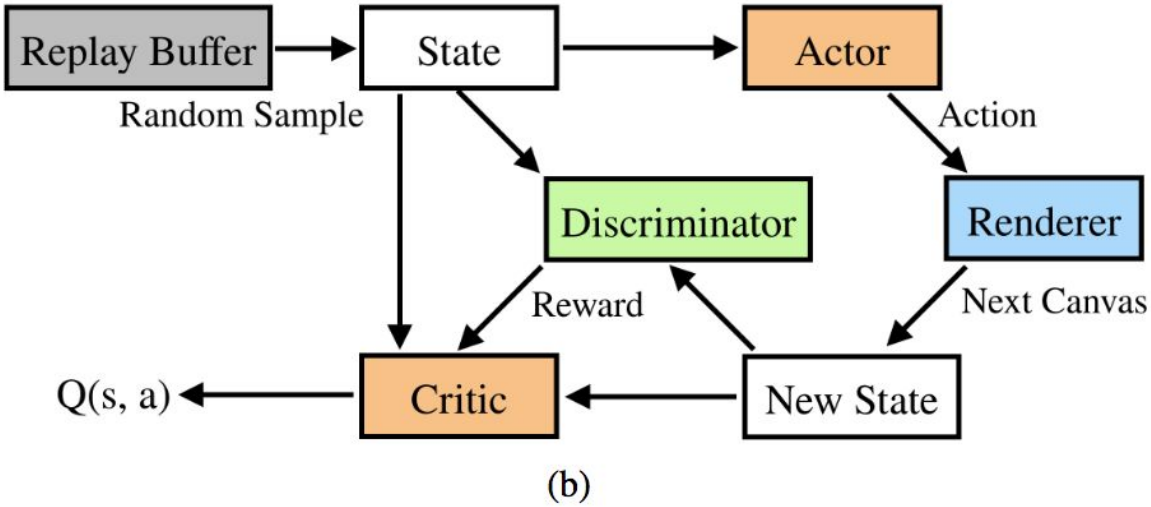
Bézier curve

random strokes

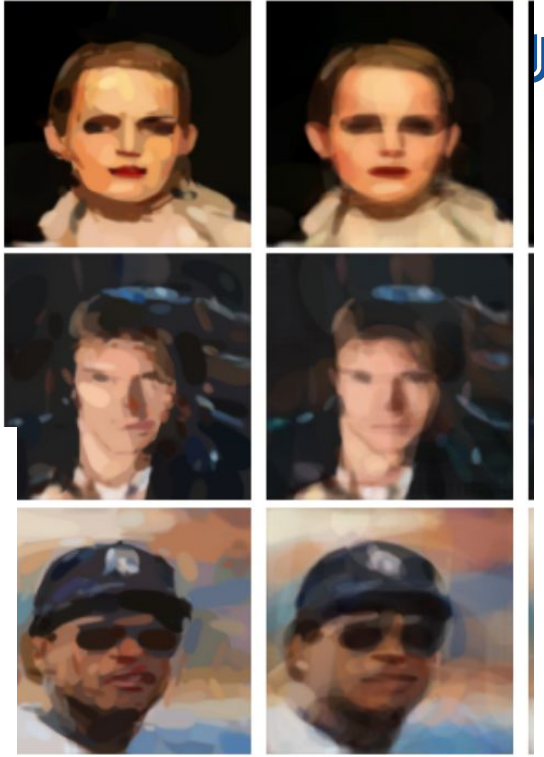


Graphics Primitive to Image

- Learning-to-paint '19



(b)



(c) (d)

GAN loss ℓ_2 loss

Graphics Primitive to Image: IG-GAN '20

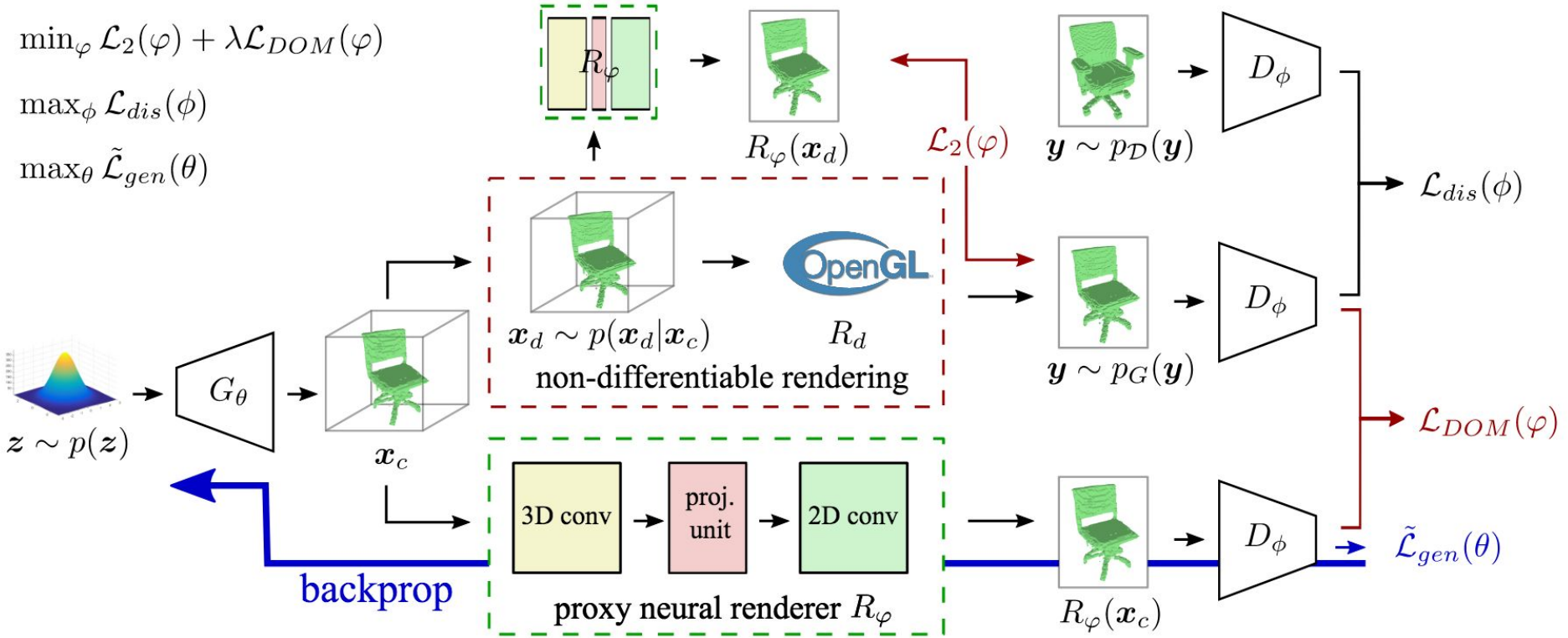


Figure 2. The architecture and training setup for IG-GAN.

Generative Adversarial Imitation Learning '16

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\begin{aligned} &\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta), \\ &\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}] \end{aligned} \tag{18}$$

6: **end for**

Applications of GAN

- Characterizing a Distribution
  - Minimizing distance between distributions
  - Inducing structure in Feature Space: Disentangled Representation
- Inventing Plausible Details
  - Context-aware Image Editing and Inpainting
- Exploiting Unpaired Data with Cycle Consistency
  - Translation between Unpaired Data
  - Examples: Text, Graphics primitives, Degraded Images
- Exact Object Transfiguration
  - Object Swapping and Information Preservation

Learning from (more diverse) Unpaired Data

Extended-Yale-B
Paired
28 human subjects

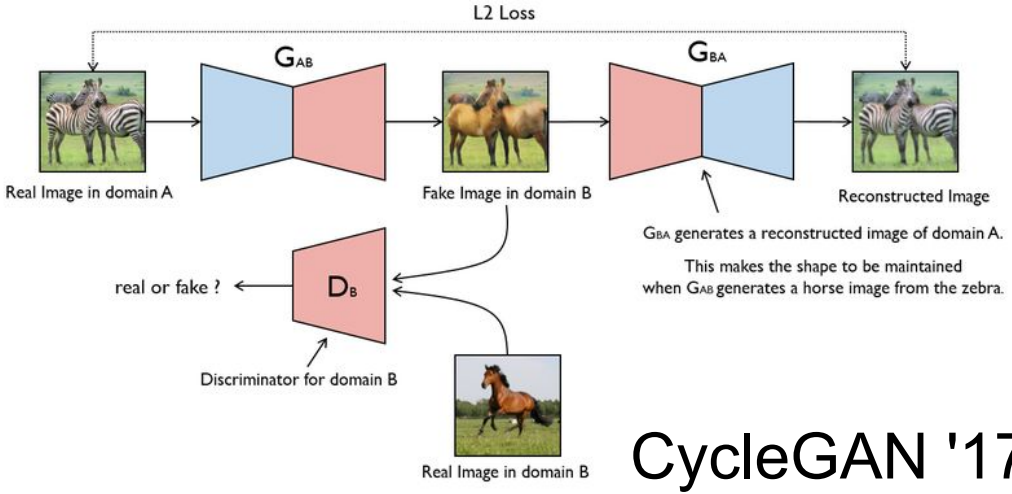
CelebA, unpaired, 10K human subjects

Learning from (more diverse) Unpaired Data

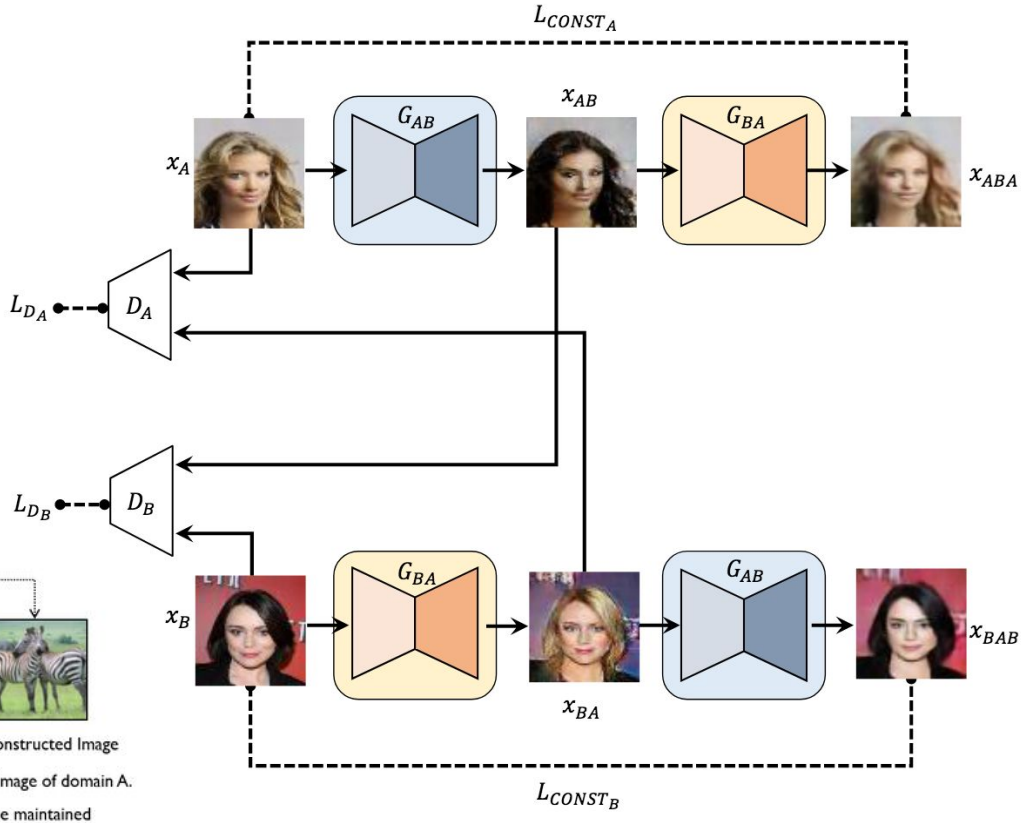
- Unpaired Data
  - Low-resolution, high-resolution
  - Noisy, denoised
  - Artistic Style A, Artistic Style B
  - Image, Text descriptions
  - Language A, Language B
  - ...

Image Translation: Cycle Consistency

- Loop closing: cycle consistency loss improves stability of training



CycleGAN '17

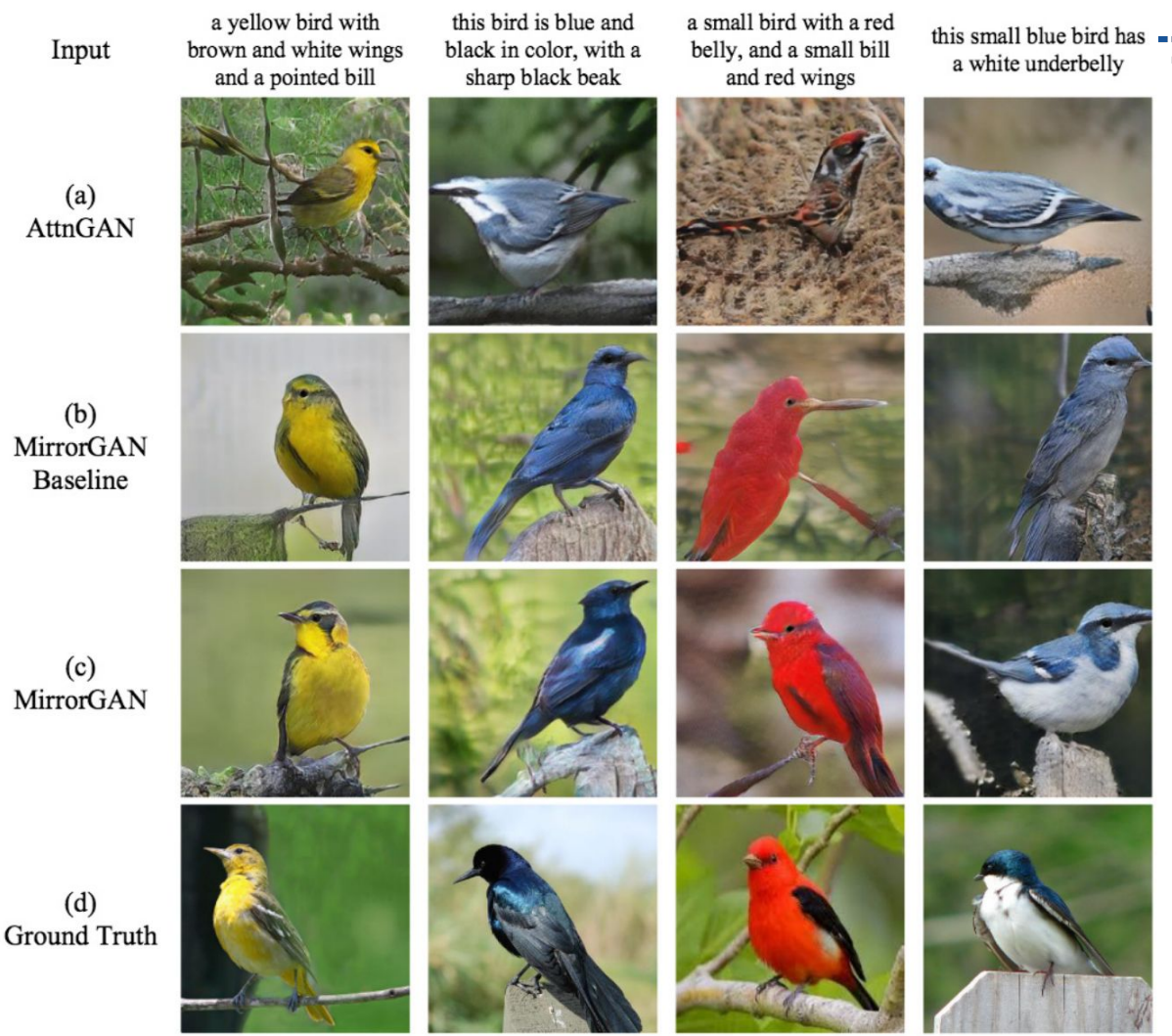


DiscoGAN '17

Text to Image

MirrorGAN '19

- Overcomes scarcity of text-image paired data



Applications of GAN

- Characterizing a Distribution
  - Minimizing distance between distributions
  - Inducing structure in Feature Space: Disentangled Representation
  - Pyramid Structure
- Inventing Plausible Details
  - Context-aware Image Editing and Inpainting
- Exploiting Unpaired Data with Cycle Consistency
  - Translation between Unpaired Data
  - Examples: Text, Graphics primitives, Degraded Images
- Exact Object Transfiguration
  - Object Swapping and Information Preservation

Exact Object Transfiguration

- Exact Object useful for
  - Data Augmentation
  - Image Editing

How to get the exact same eyeglasses to another person?

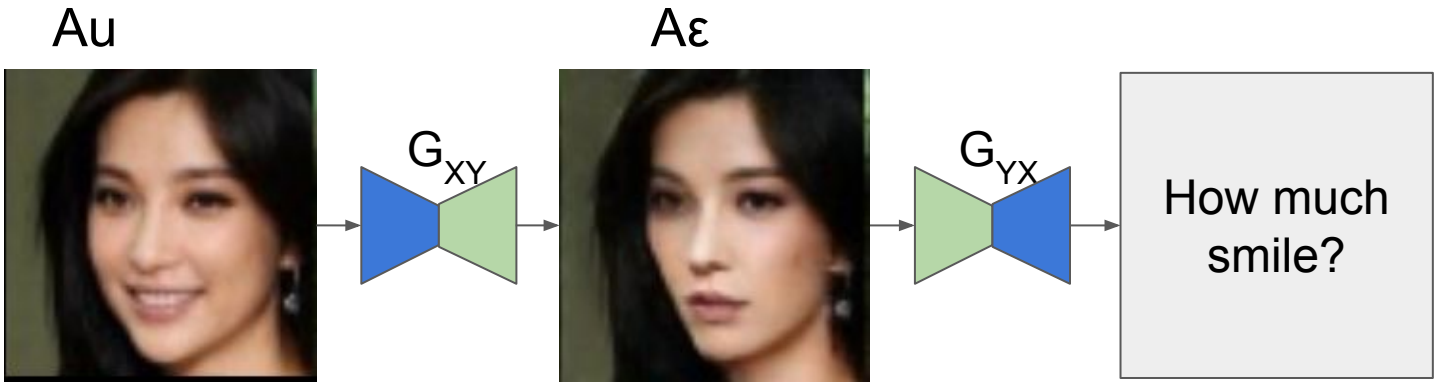
DiscoGAN/CycleGAN approaches

- Exploits identity:
  - Assume X is non-smiling
  - $X + \varphi(\text{glasses}) - \varphi(\text{glasses}) = X$
- Use Generative Adversarial Network to ensure
  - $\varphi^{-1}(X)$ is indeed not wearing glasses
  - $\varphi^{-1}(X + \text{glasses})$ is indeed wearing glasses
- Problem
  - DiscoGAN/CycleGAN Requires source domain and target domain to have the same intrinsic dimensionality for the cyclic reconstruction loss to work
  - For example, for eyeglasses, the *non-wearing-glasses* domain is smaller than the glasses domain in dimensionality. Would not be able to reconstruct the original eyeglass from a "non-wearing-glasses" face,

Image Translation: Preservation of Information

- Not only Glasses, but Smiling etc.

Information is lost as V(Au) > V(Aε)



Transform Vector from Cluster Centers

- Deep Feature Interpolation '16
  - To deal with unpaired data, use differences of cluster centers
  - Use off-the-shelf feature extractor
- Problem: using cluster centers destroys the diversity

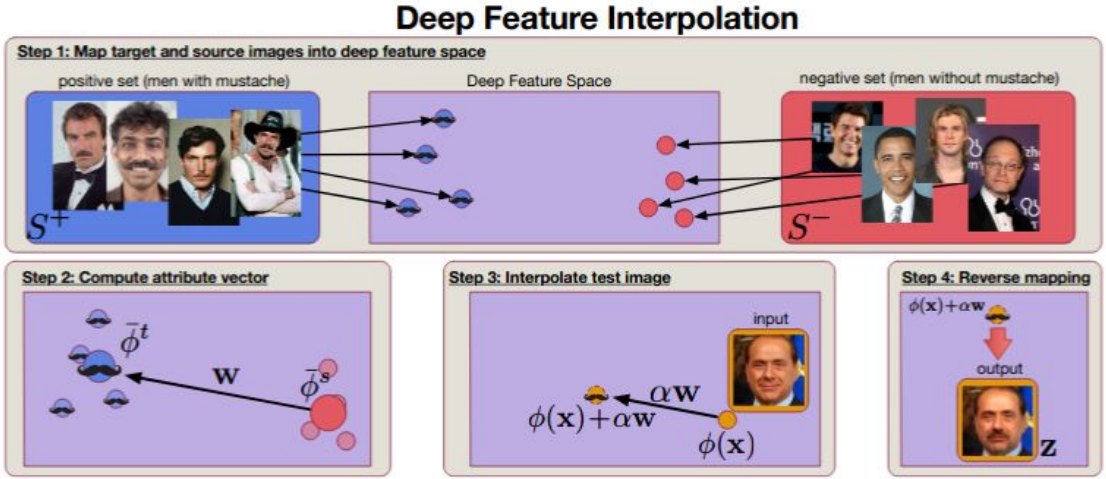
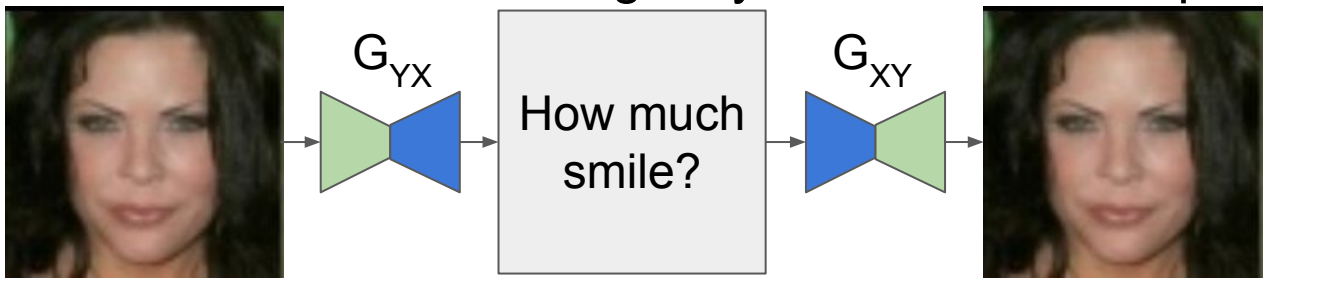


Image Translation: Preservation of Information

Generation out of nothing may suffer mode collapse

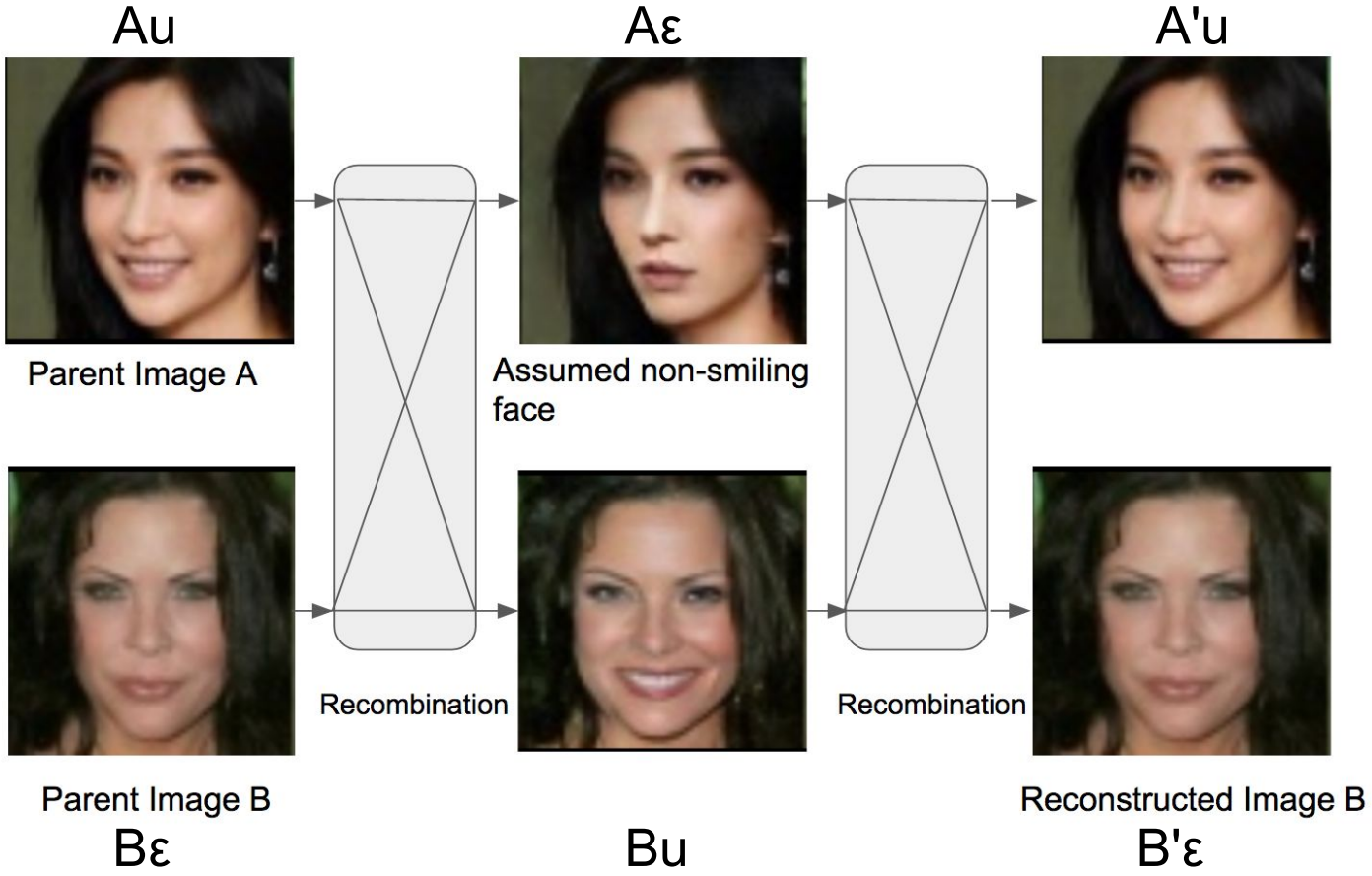


Parent Image B

Reconstructed Image

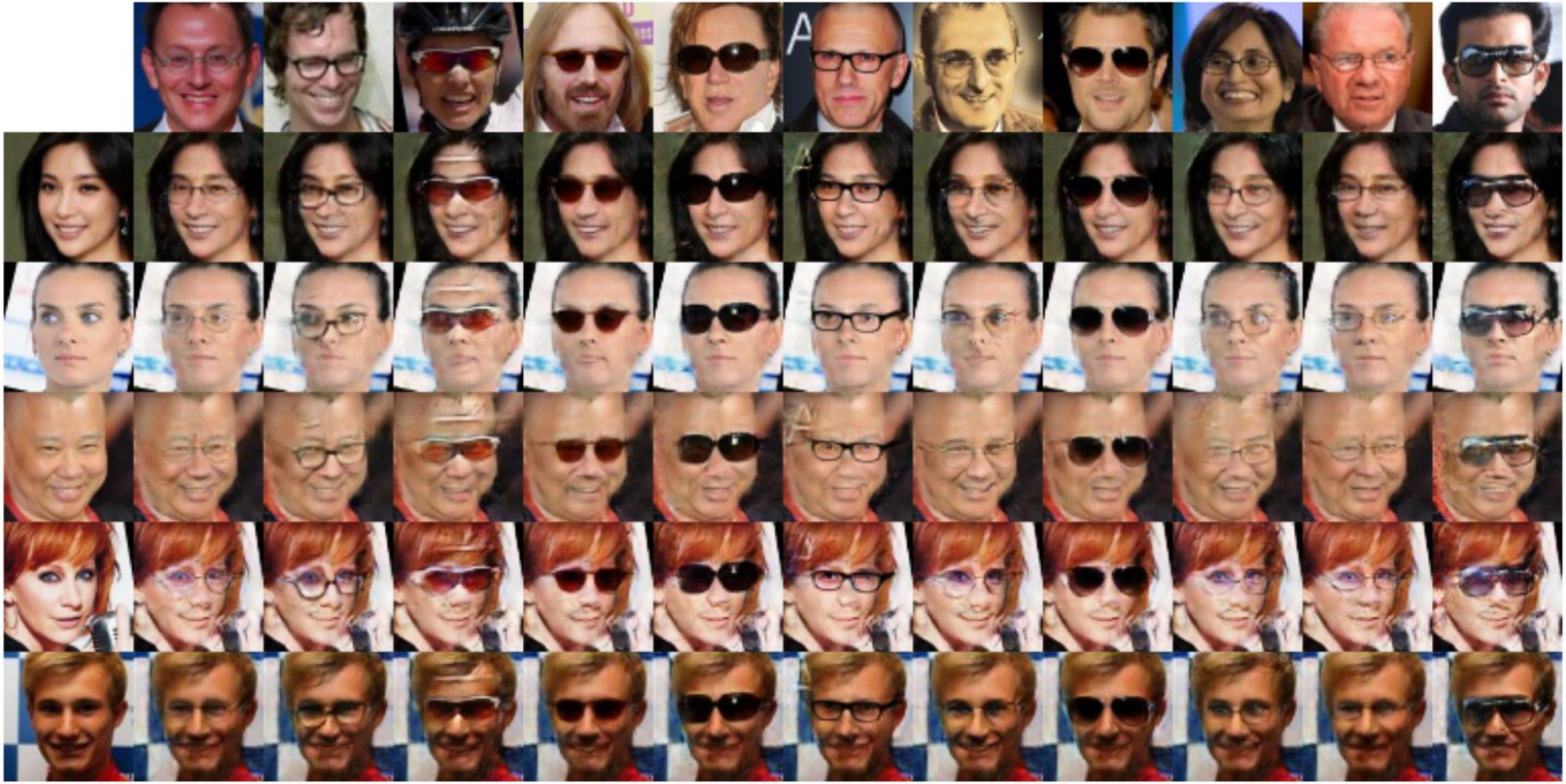
Bε

Original styles of glasses are lost.



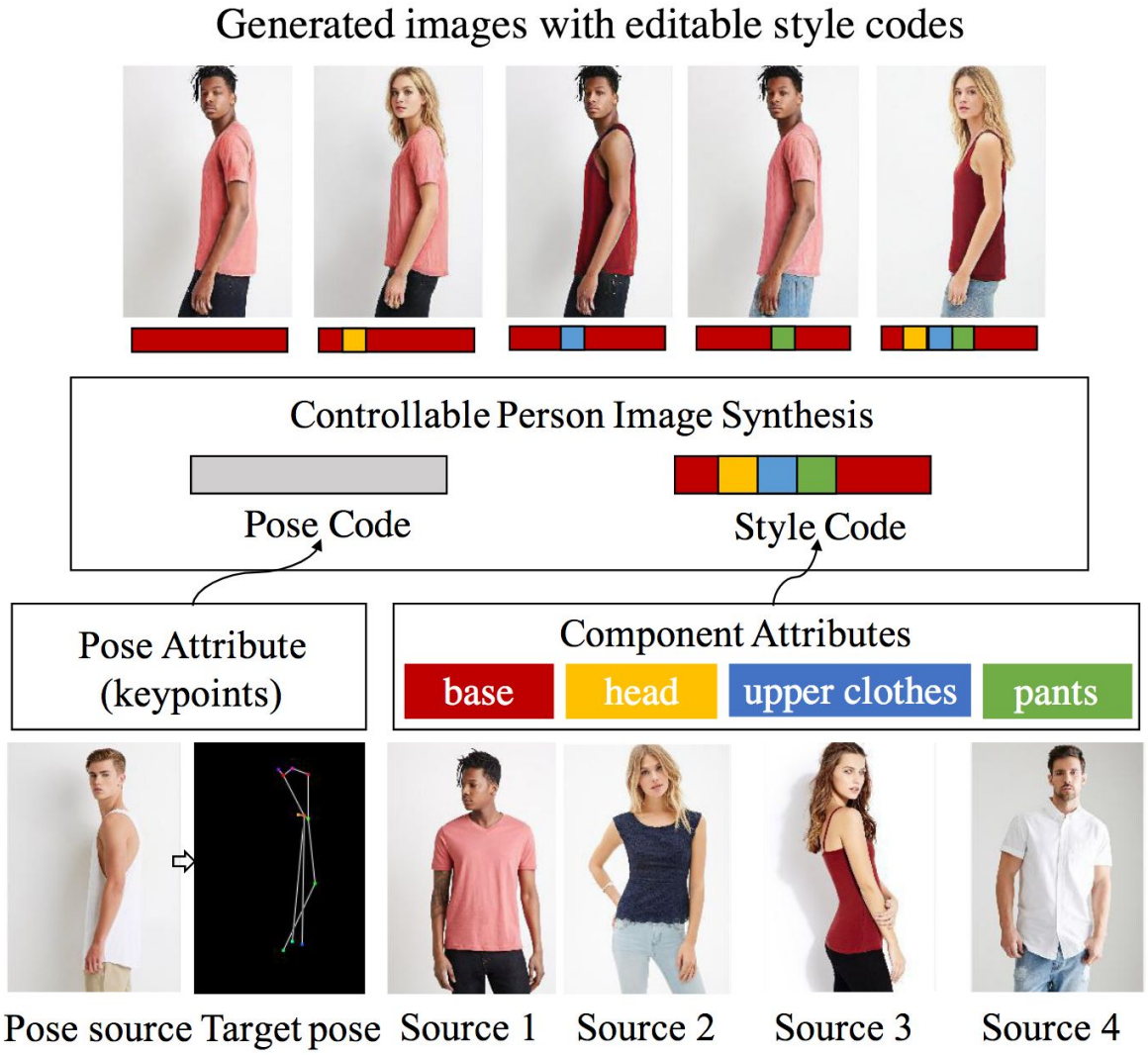
Preservation of Information by Swapping

$V(Au) + V(B\varepsilon) \approx V(A\varepsilon) + V(Bu)$



GeneGAN '17

Attribute-Decomposed GAN '20



What's next?

- Semi-Supervised Learning + GAN

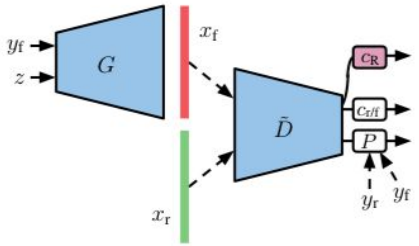
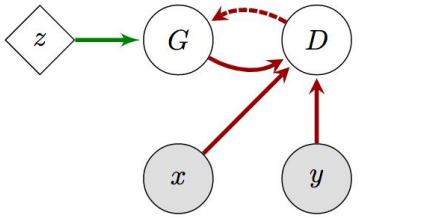


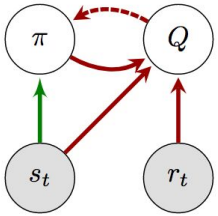
Figure 6. Self-supervision by rotation-prediction during GAN training. Additionally to predicting whether the images at its input are real or generated, the discriminator is trained to predict

https://arxiv.org/pdf/1903.02271.pdf

- GAN + Actor-Critic



(a) Generative Adversarial Networks [4]



(b) Deterministic Policy Gradient [7] / SVG(0) [8] / Neurally-Fitted Q-learning with Continuous Actions [9]

https://arxiv.org/pdf/1610.01945.pdf

- Video, Game, etc.

DVD-GAN '19

References

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).

Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." *Advances in neural information processing systems*. 2016.

https://github.com/deepfakes/

Goodfellow, Ian. "NIPS 2016 tutorial: Generative adversarial networks." *arXiv preprint arXiv:1701.00160* (2016).

Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale gan training for high fidelity natural image synthesis." *arXiv preprint arXiv:1809.11096* (2018).

Zhou, Shuchang, et al. "Genegan: Learning object transfiguration and attribute subspace from unpaired data." *arXiv preprint arXiv:1705.04932* (2017).

References

Denton, Emily L., Soumith Chintala, and Rob Fergus. "Deep generative image models using a laplacian pyramid of adversarial networks." *Advances in neural information processing systems*. 2015.

Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." *arXiv preprint arXiv:1710.10196* (2017).

Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." *arXiv preprint arXiv:1703.05192* (2017).

References

Reed, Scott E., et al. "Deep visual analogy-making." *Advances in neural information processing systems*. 2015.

Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

Zhu, Peihao, et al. "SEAN: Image Synthesis with Semantic Region-Adaptive Normalization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Zhang, Han, et al. "Self-attention generative adversarial networks." *International Conference on Machine Learning*. PMLR, 2019.

Wang, Sheng-Yu, et al. "CNN-generated images are surprisingly easy to spot... for now." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 7. 2020.

Ravuri, Suman, and Oriol Vinyals. "Classification accuracy score for conditional generative models." *Advances in Neural Information Processing Systems*. 2019.

Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

References

Yu, Jiahui, et al. "Generative image inpainting with contextual attention." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Menon, Sachit, et al. "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Duan, Ranjie, et al. "Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Xiao, Chaowei, et al. "Generating adversarial examples with adversarial networks." *arXiv preprint arXiv:1801.02610* (2018).

Kulkarni, Tejas D., et al. "Deep convolutional inverse graphics network." *Advances in neural information processing systems*. 2015.

Huang, Zhewei, Wen Heng, and Shuchang Zhou. "Learning to paint with model-based deep reinforcement learning." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

Lunz, Sebastian, et al. "Inverse Graphics GAN: Learning to Generate 3D Shapes from Unstructured 2D Data." *arXiv preprint arXiv:2002.12674* (2020).

References

Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in neural information processing systems*. 2016.

http://cvc.yale.edu/projects/yalefaces/yalefaces.html

Liu, Ziwei, et al. "Large-scale celebfaces attributes (celeba) dataset." *Retrieved August* 15 (2018): 2018.

Qiao, Tingting, et al. "Mirrorgan: Learning text-to-image generation by redescription." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

https://www.goggles4u.com/hollywood-top-celebrities-wearing-eyeglasses

Upchurch, Paul, et al. "Deep feature interpolation for image content changes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Men, Yifang, et al. "Controllable person image synthesis with attribute-decomposed gan." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Lucic, Mario, et al. "High-fidelity image generation with fewer labels." *arXiv preprint arXiv:1903.02271* (2019).

References

Pfau, David, and Oriol Vinyals. "Connecting generative adversarial networks and actor-critic methods." *arXiv preprint arXiv:1610.01945* (2016).

Clark, Aidan, Jeff Donahue, and Karen Simonyan. "Adversarial video generation on complex datasets." *arXiv* (2019): arXiv-1907.