CS 2770: Conditional Generative Models

PhD. Nils Murrugarra-Llerena
nem177@pitt.edu

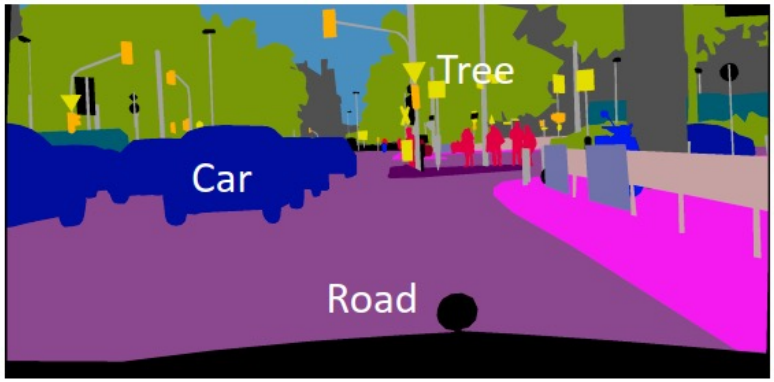
Conditional Generative Models: Problem Statement

Input Output

Goal: synthesize a photograph given an input image

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Generative Models: Problem Statement



Input

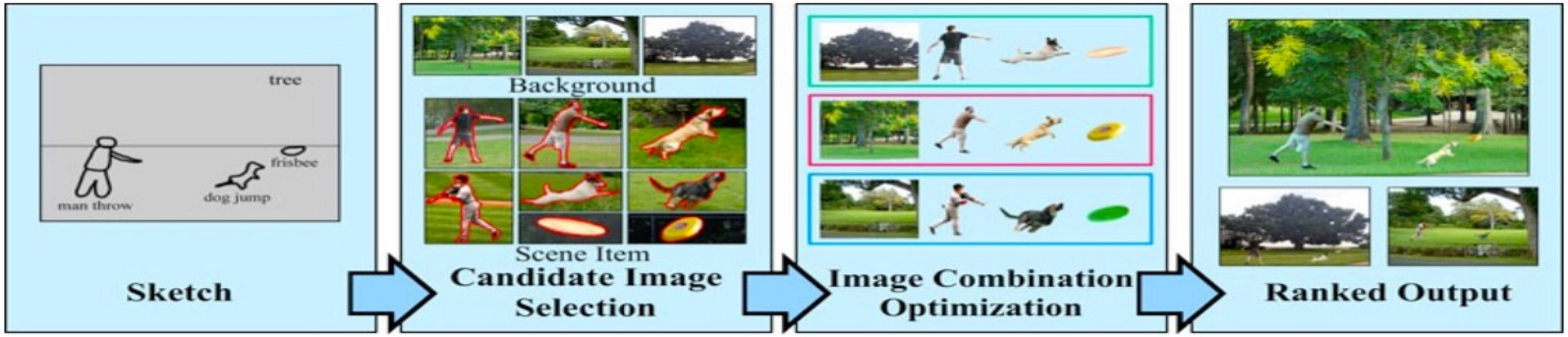
Output

Goal: synthesize a photograph given an input image

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Generative Models: Early Work

Semantic Photo Synthesis [Johnson et al., Eurographics 2006]



Sketch2Photo [Tao et al., SIGGRAPH Asia 2009]

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Generative Models

Two ways to do it:

1. Model p(x,y), then do inference to get conditional p(y|x)

$$\arg\max_y p(x, y) = \arg\max_y p(y|x)$$

2. Directly model p(y|x)

Conditional Generative Models: Structured Prediction

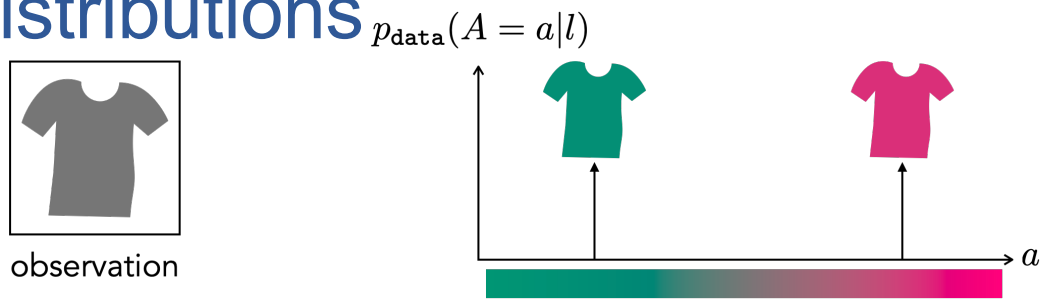
X is high-dimensional

Model *joint* distribution of high-dimensional data $P(\mathbf{X}|\mathbf{Y}=\mathbf{y})$

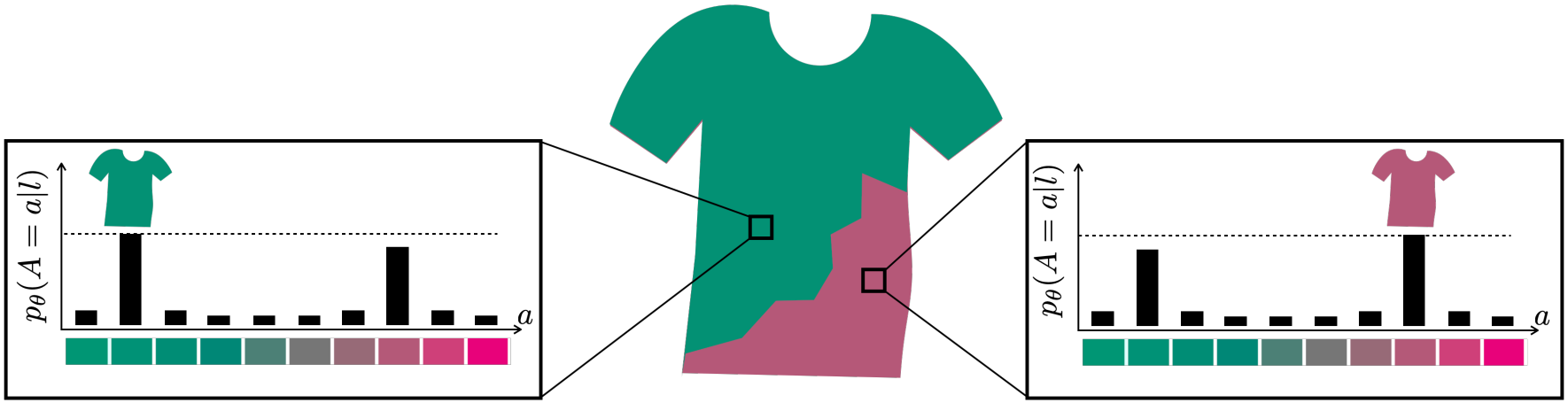
In vision this is usually what we are interested in

Unstructured: $\prod_i p(X_i|\mathbf{Y}=\mathbf{y})$

The failure of Point Prediction: Multimodal Distributions



The failure of Point Prediction: Joint Structure

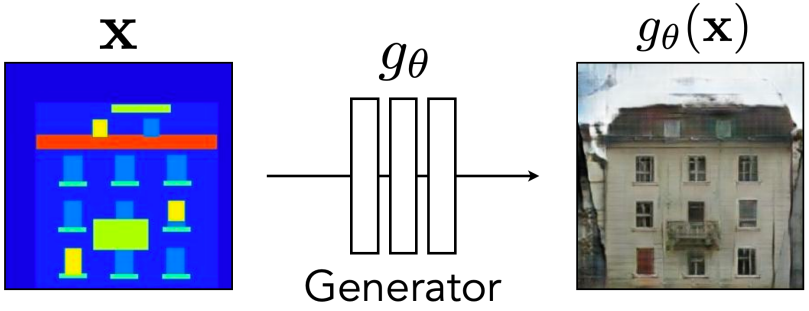


Generative models have two important properties for structured prediction:

1. They can model a multimodal distribution
2. They can model joint dependences between multidimensional predictions

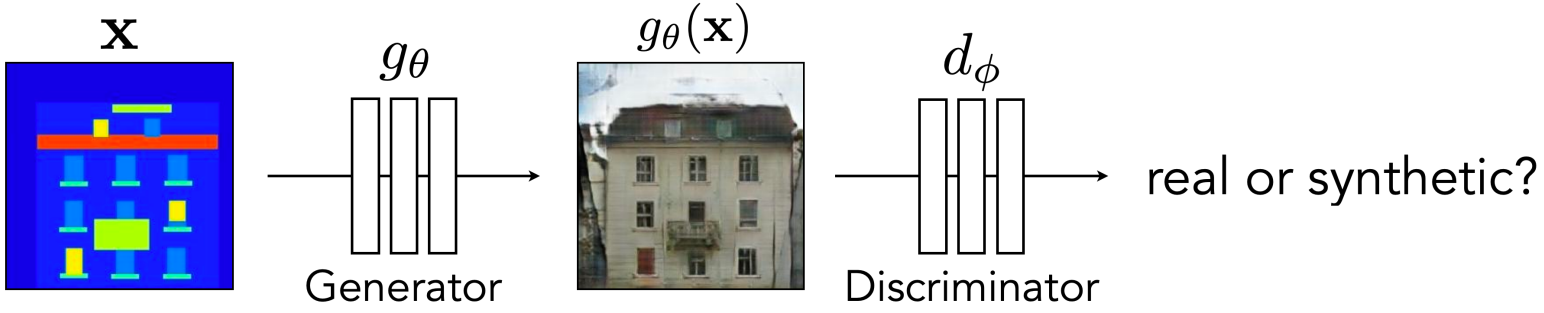
Conditional GANs

Conditional GANs



For example: pix2pix [Isola et al. 2017]

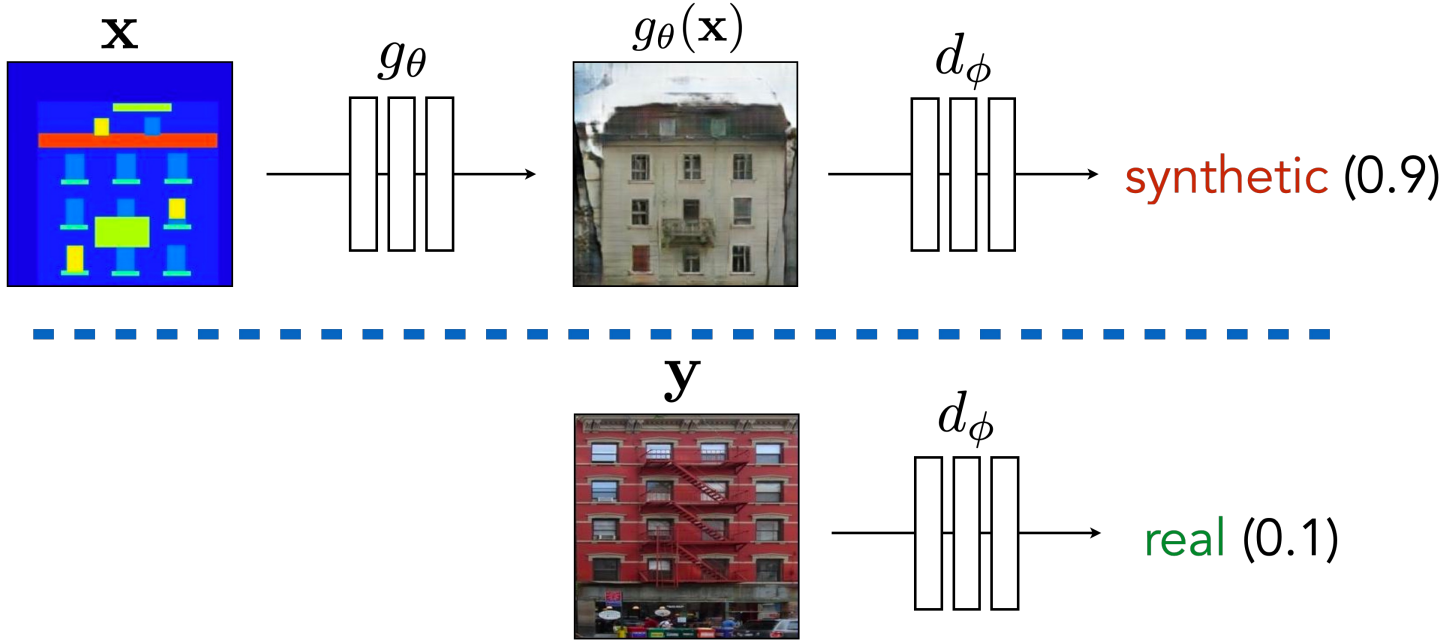
Conditional GANs



g tries to synthesize fake images that fool d

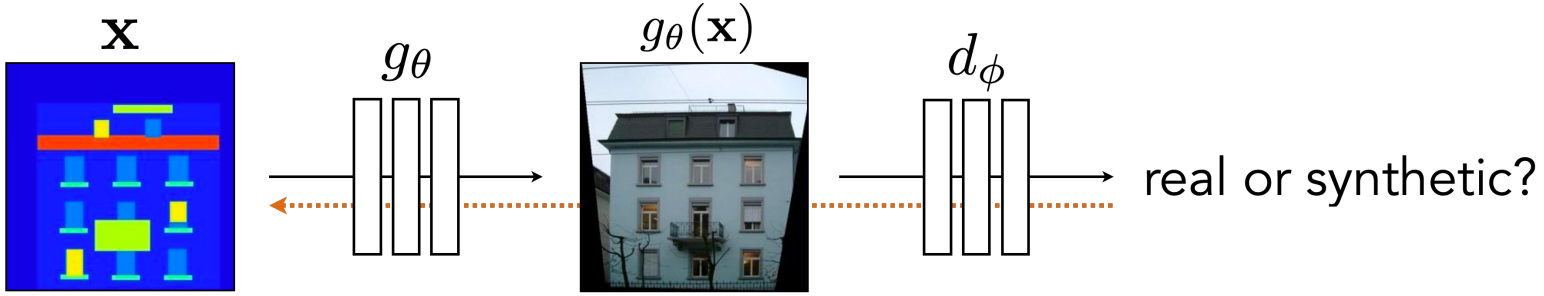
d tries to identify the fakes

Conditional GANs



$$d_\phi^* = \arg\max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y})) \right]$$

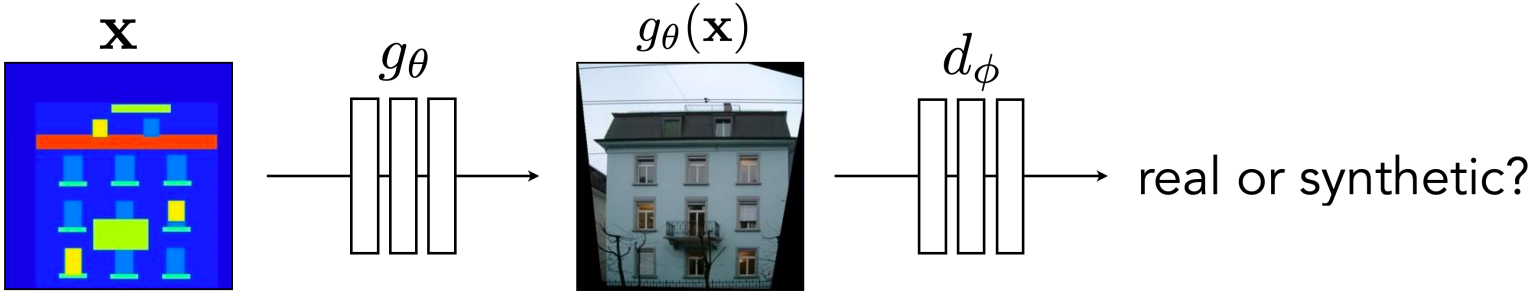
Conditional GANs



g tries to synthesize fake images that *fool* d:

$$g_\theta^* = \boxed{\arg\min_\theta} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))]$$

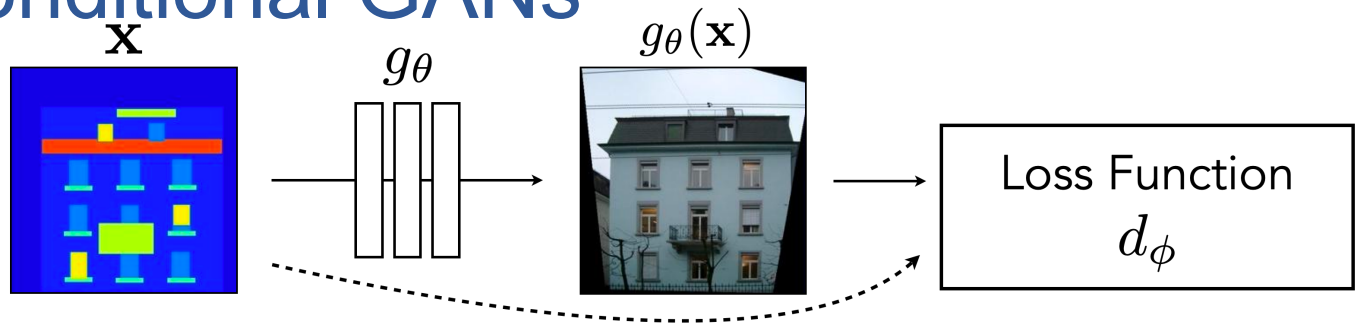
Conditional GANs



g tries to synthesize fake images that *fool* the *best* d:

$$\arg \min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{y}))]$$

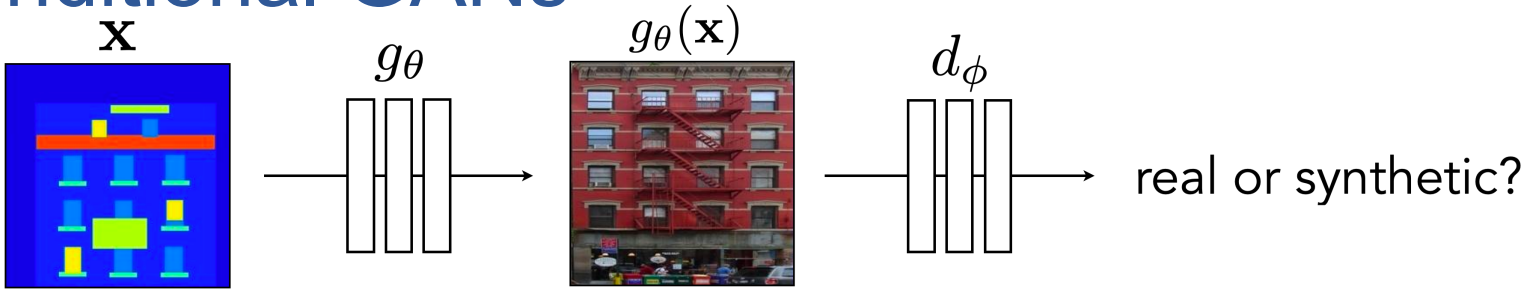
Conditional GANs



g's perspective: d is a loss function.

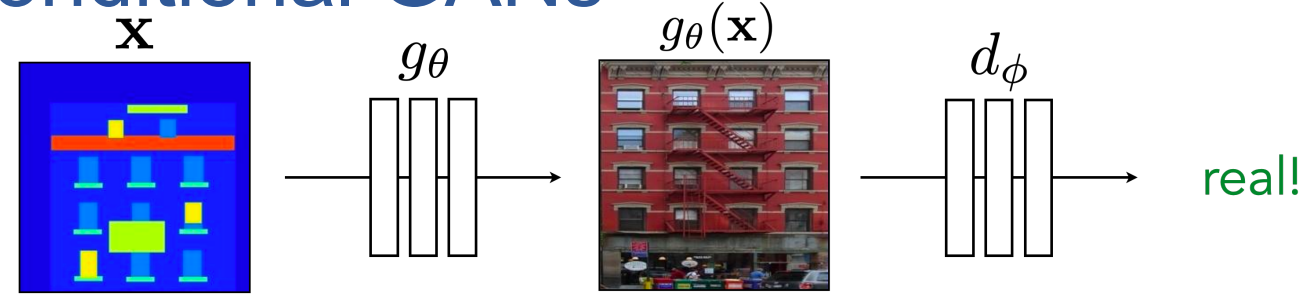
Rather than being hand-designed, it is *learned* and *highly structured*.

Conditional GANs



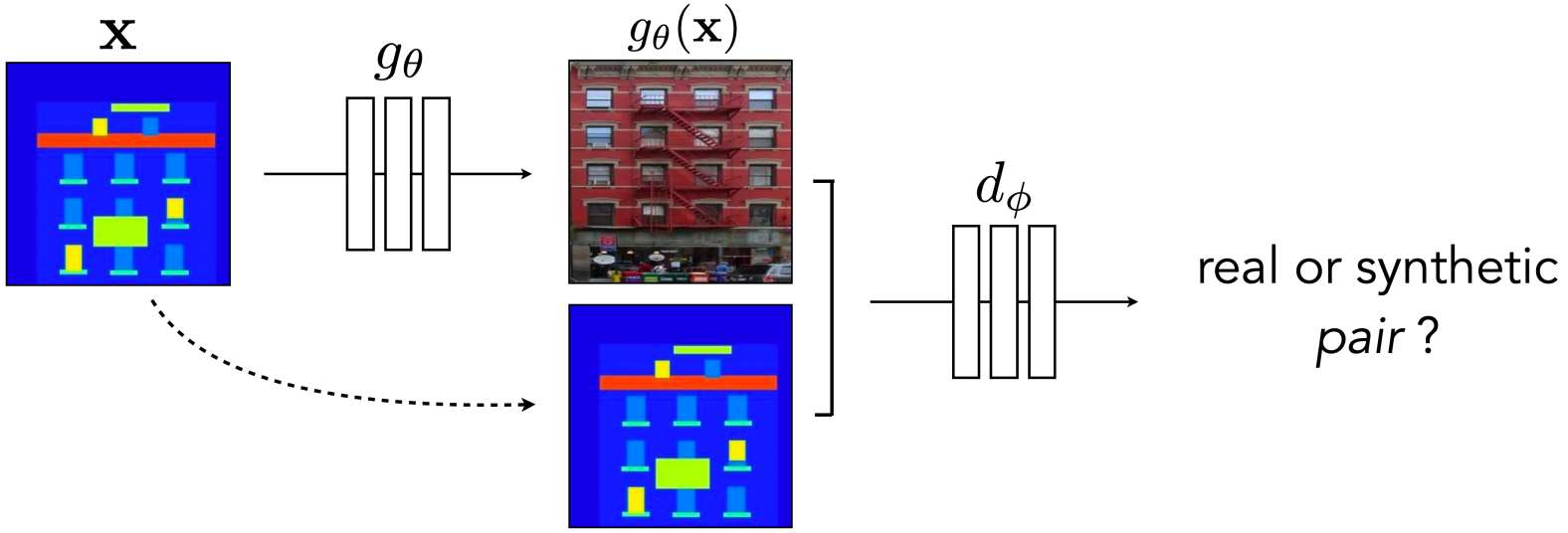
$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))+\log(1-d_\phi(\mathbf{y}))]$$

Conditional GANs



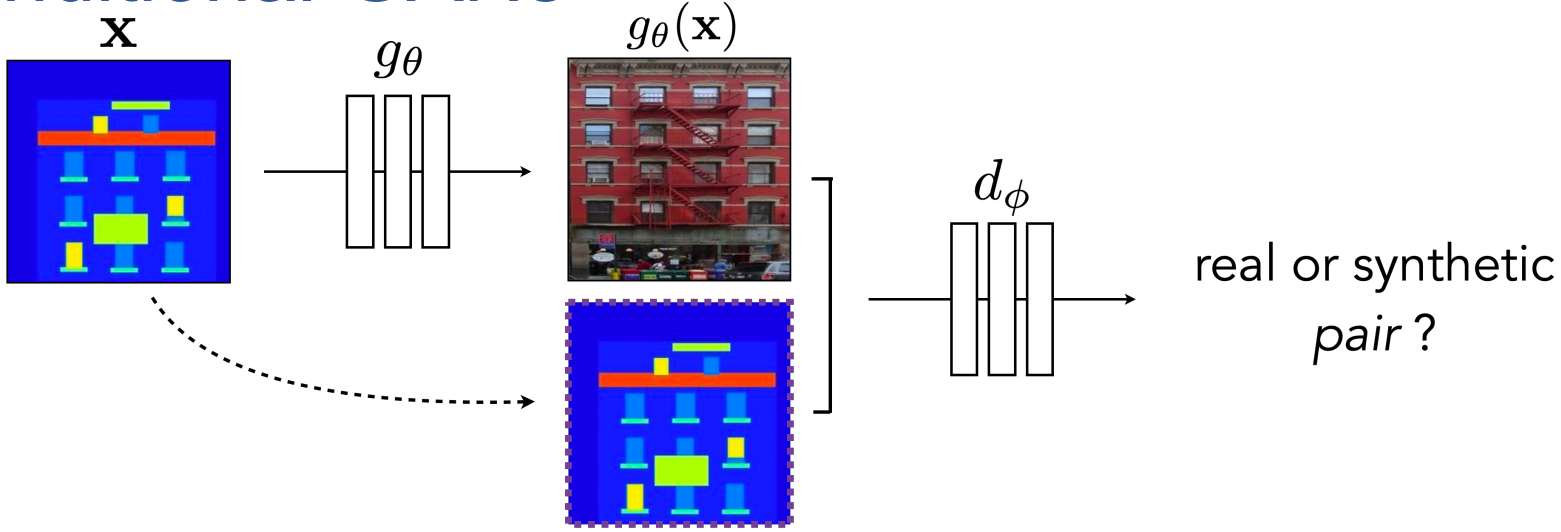
$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))+\log(1-d_\phi(\mathbf{y}))]$$

Conditional GANs



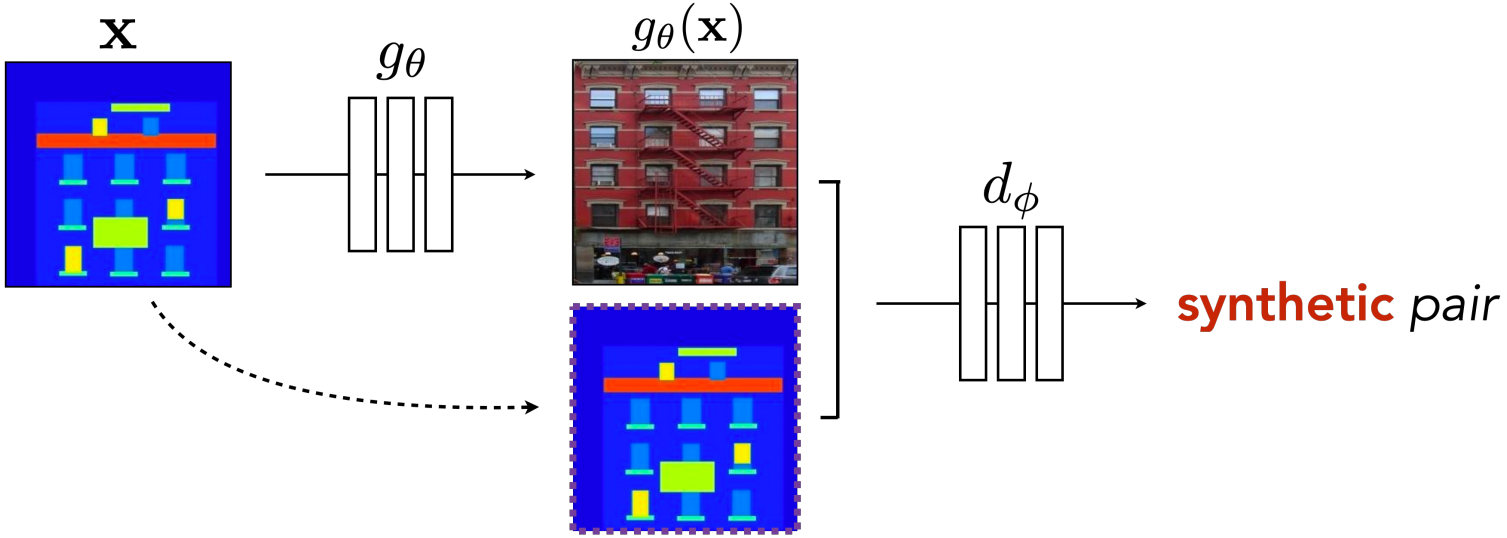
$$\arg\min_{\theta}\max_{\phi}\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(g_\theta(\mathbf{x}))+\log(1-d_\phi(\mathbf{y}))]$$

Conditional GANs



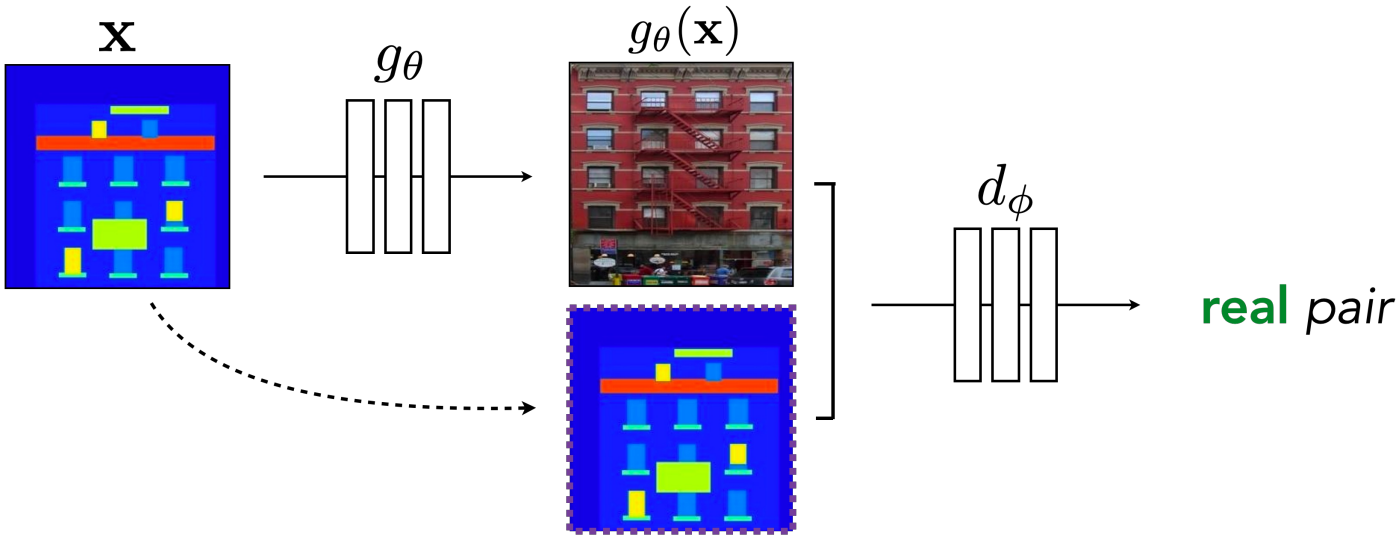
$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x}, \mathbf{y}))]$$

Conditional GANs



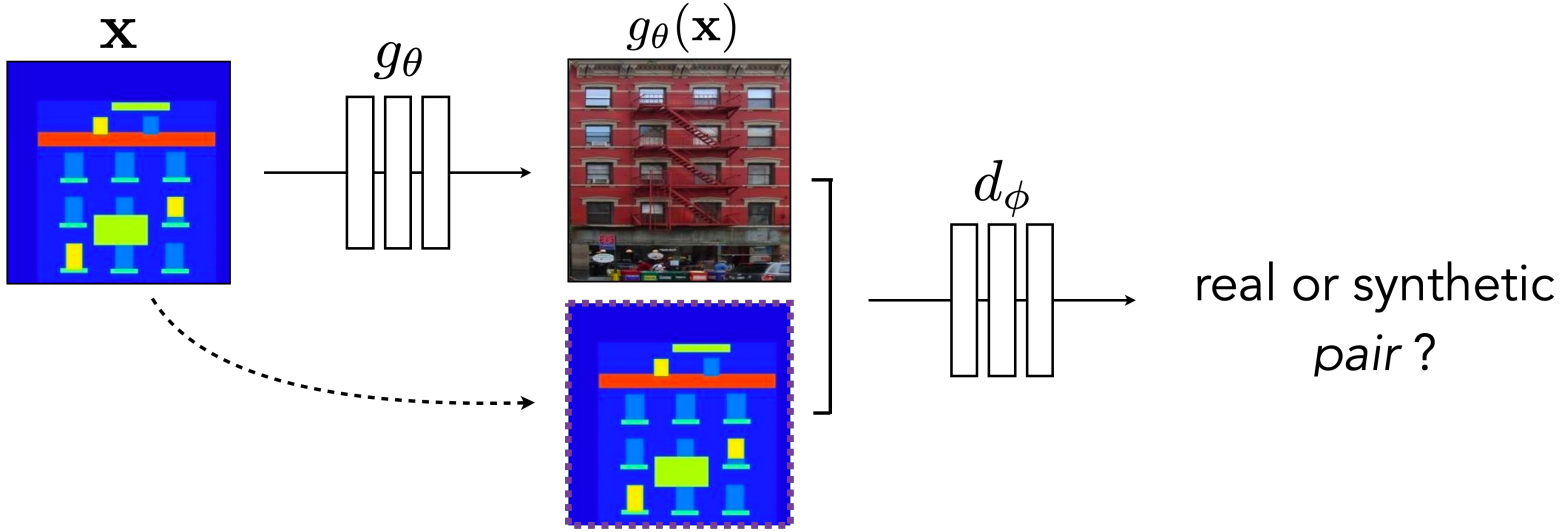
$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x}, \mathbf{y}))]$$

Conditional GANs



$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x}, \mathbf{y}))]$$

Conditional GANs

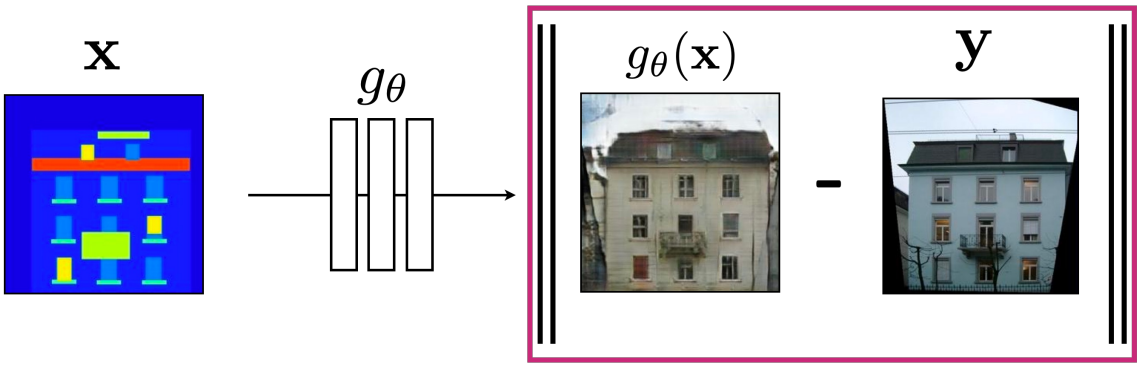


$$\arg\min_\theta \max_\phi \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log d_\phi(\mathbf{x}, g_\theta(\mathbf{x})) + \log(1 - d_\phi(\mathbf{x}, \mathbf{y}))]$$

Conditional GANs

Training Details: Loss function

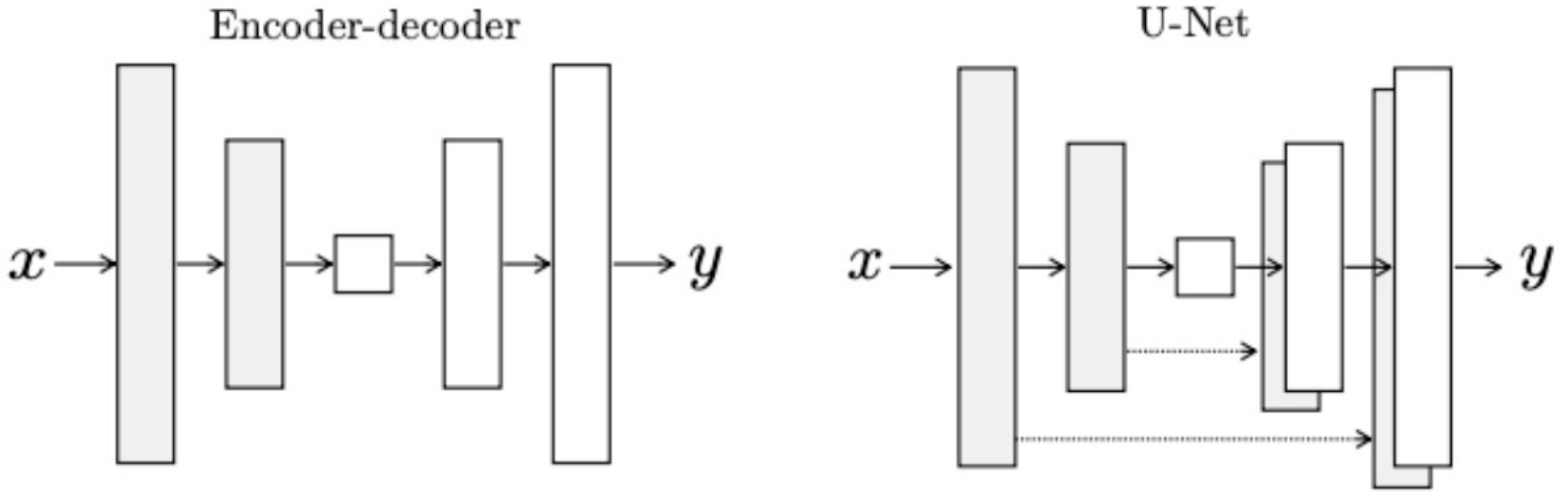
$$g_\theta^* = \arg\min_\theta \max_\phi \mathcal{L}_{\texttt{cGAN}}(\theta, \phi) + \lambda \mathcal{L}_{\texttt{L1}}(\theta)$$



Stable training + fast convergence

[c.f. Pathak et al. CVPR 2016]

Conditional GANs: pix2pix Generator



U-Net [Ronneberger et al.]: popular CNN backbone for biomedical image segmentation
U-Net: preserve high-frequency information (e.g., edge) of the input image.
Encoder-decoder: lose high-frequency details due to the information bottleneck

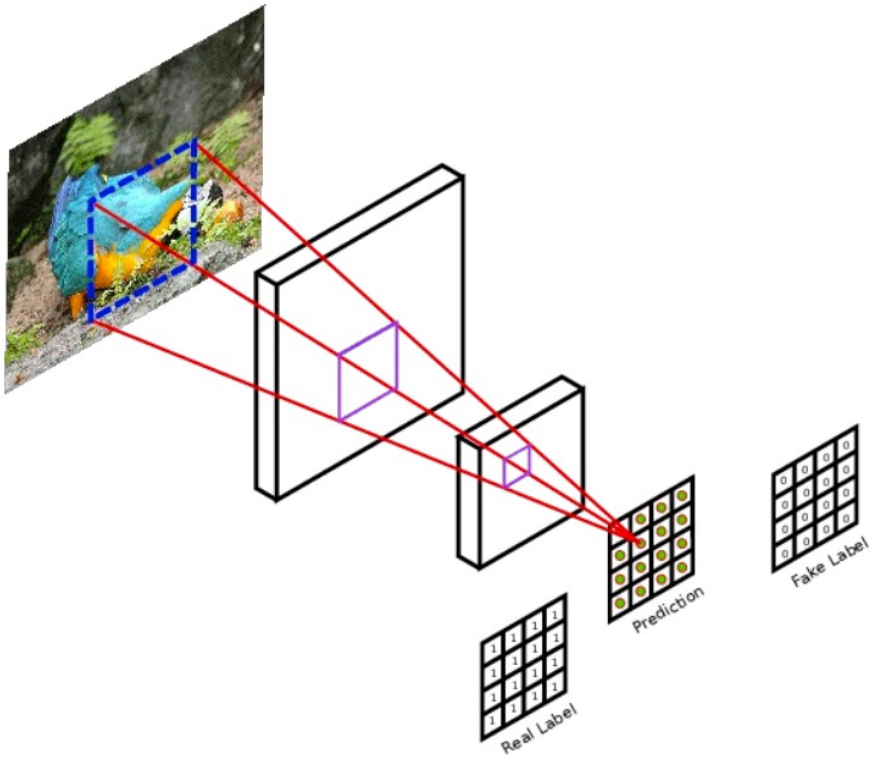
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional GANs: pix2pix Discriminator

- Rather than penalizing if output image looks fake, penalize if each overlapping patches looks fake
- Focus on local visual cues (color, textures).
- Global structure: the input image has already encoded global structure. L1 loss can help as well.

Advantages:

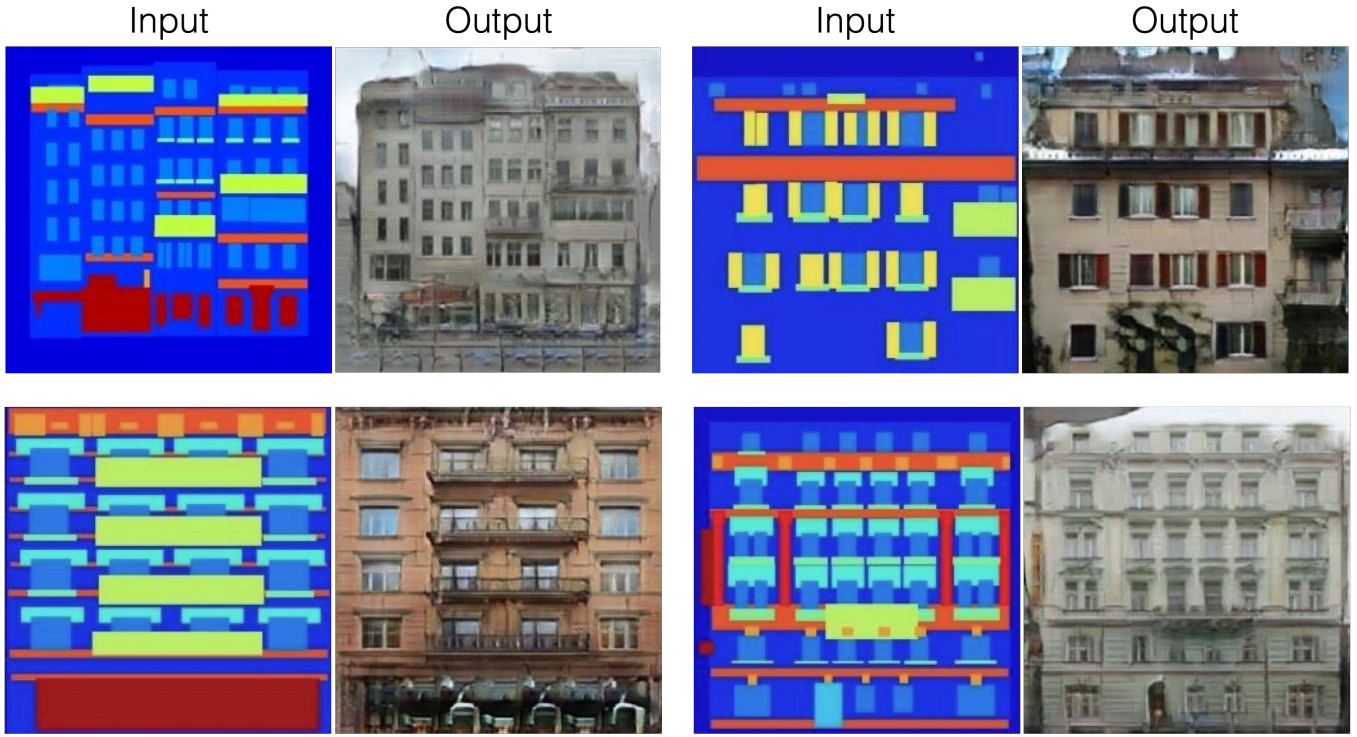
- Faster, fewer parameters
- More supervised observations
- Applies to arbitrarily large images



Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional GANs

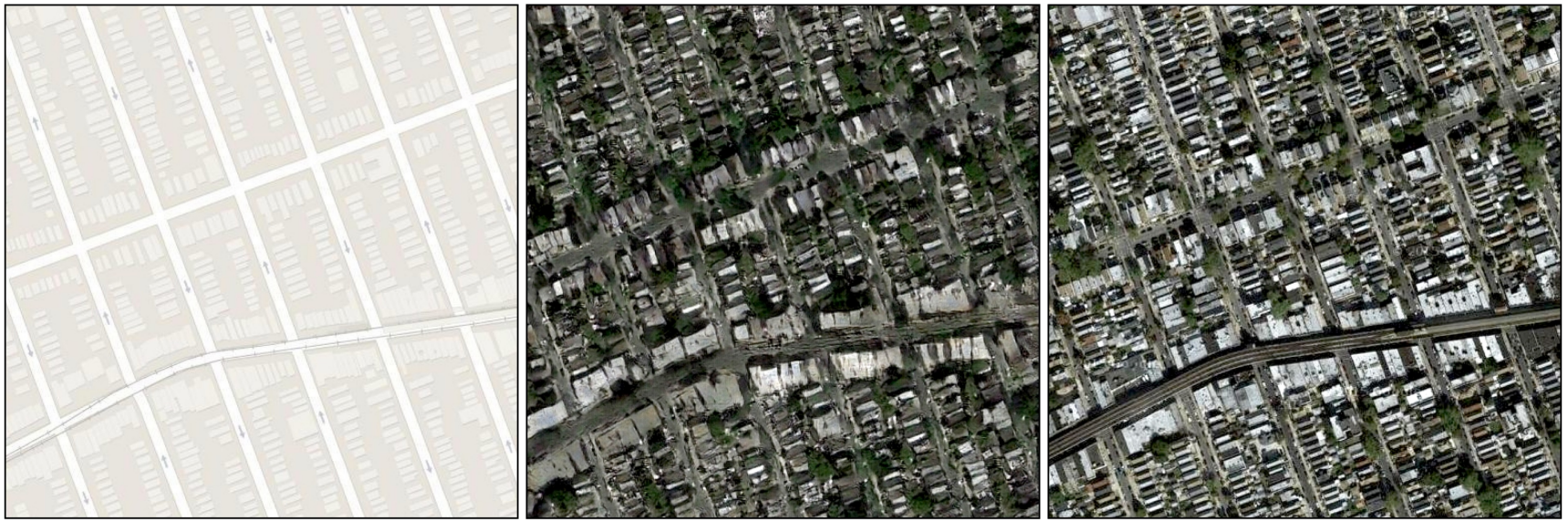
Labels → Facades



Data from [Tylecek, 2013]

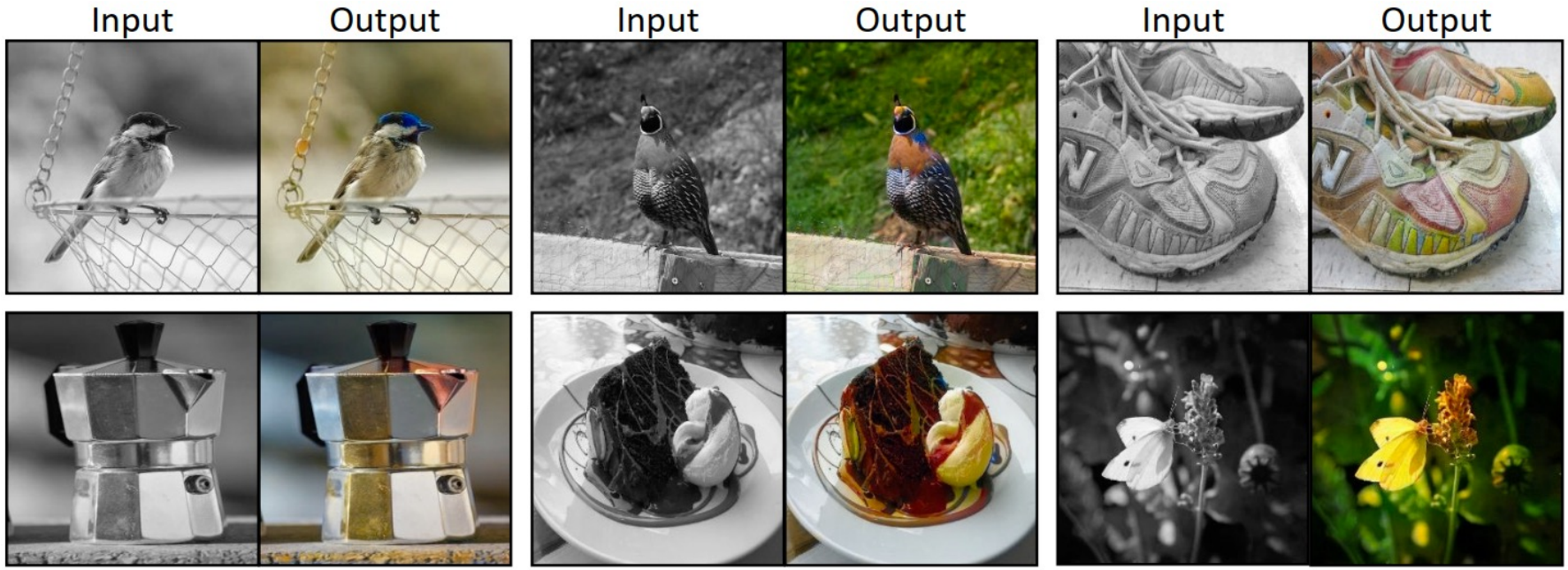
Conditional GANs

Input | Output | Groundtruth



Data from [maps.google.com]

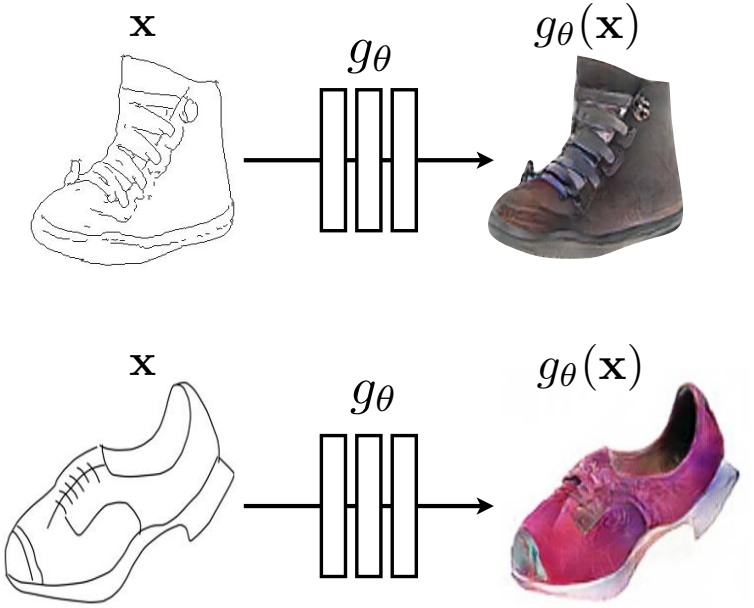
Conditional GANs: Automatic Colorization



Data from [Russakovsky et al. 2015]

Conditional GANs

[HED, Xie & Tu, 2015]

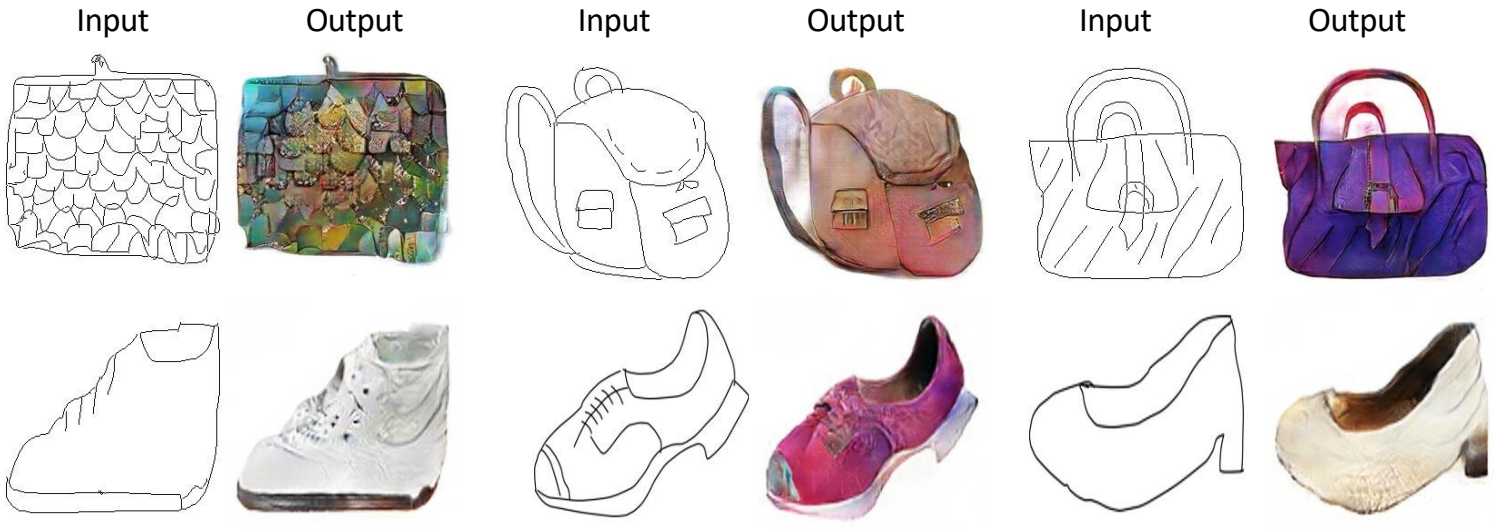


Conditional GANs: Edges to Images

Edges from [Xie & Tu, 2015]

Pix2pix / CycleGAN

Conditional GANs: Sketches to Images



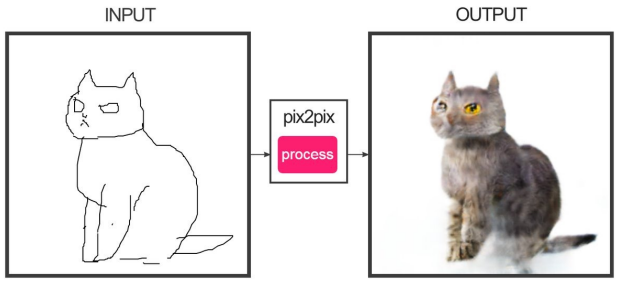
Trained on Edges → Images

Data from [Eitz, Hays, Alexa, 2012]

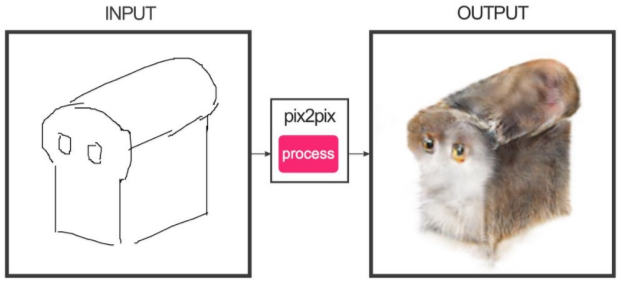
Pix2pix / CycleGAN

Conditional GANs: Edges to Images

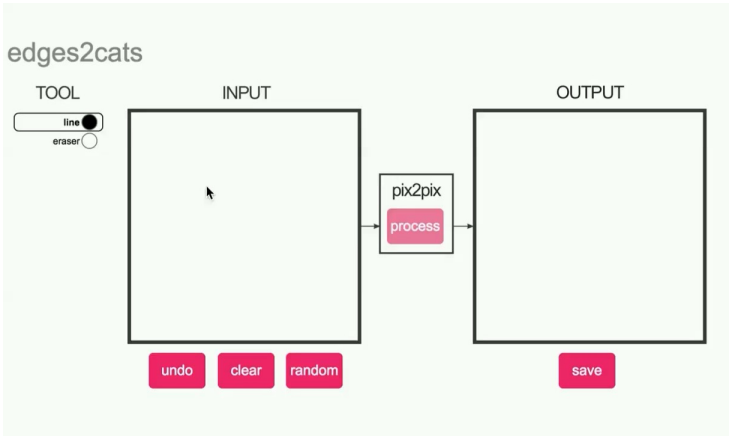
#edges2cats [Christopher Hesse]



@gods_tail



Ivy Tasi @ivymyt



@matthematician

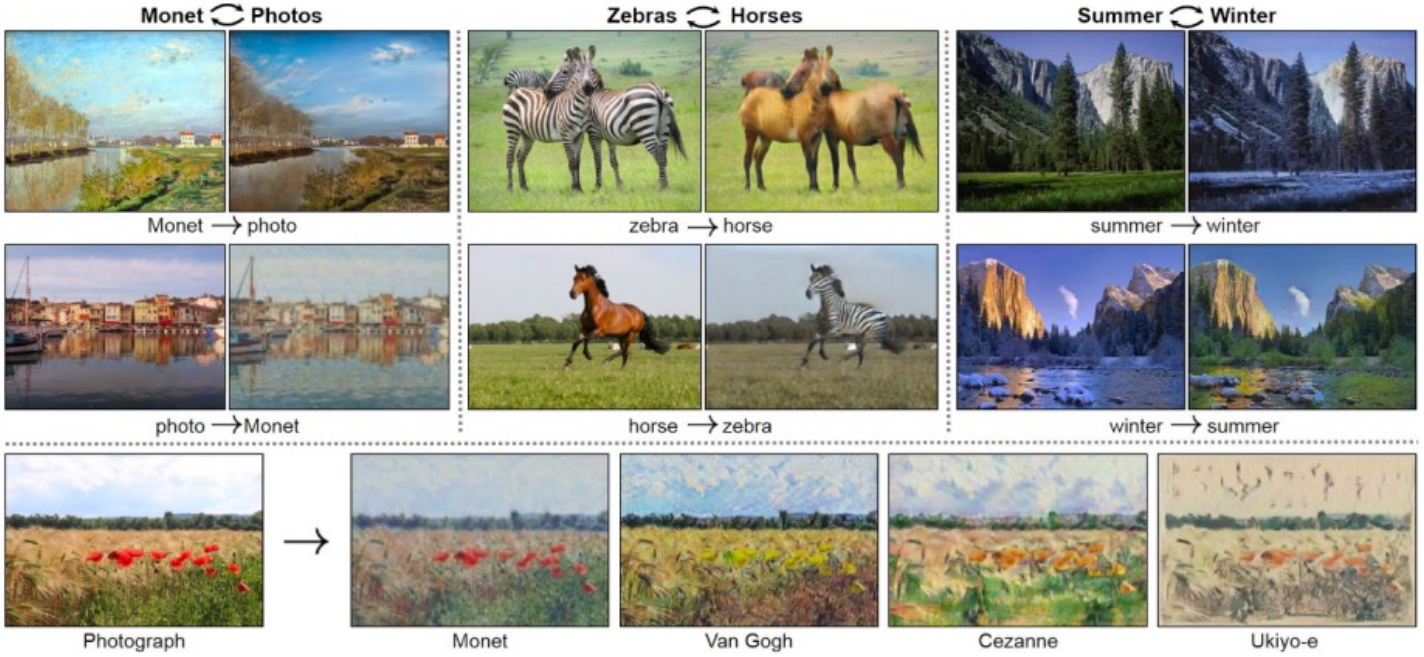
Vitaly Vidmirov @vvid

https://affinelayer.com/pixsrv/

Pix2pix / CycleGAN

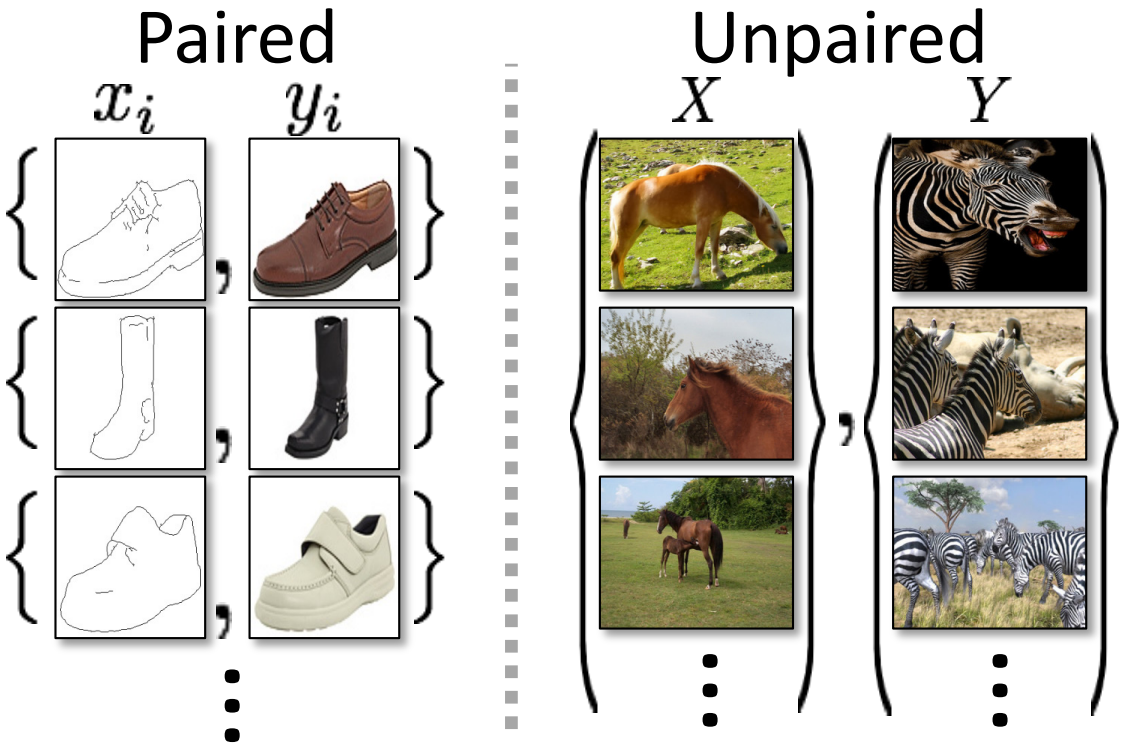
Conditional GANs: Unpaired Translations

Style transfer problem: change the style of an image while preserving the content.



Data: Two unrelated collections of images, one for each style

Conditional GANs: Unpaired Translations - CycleGAN



Conditional GANs: CycleGAN

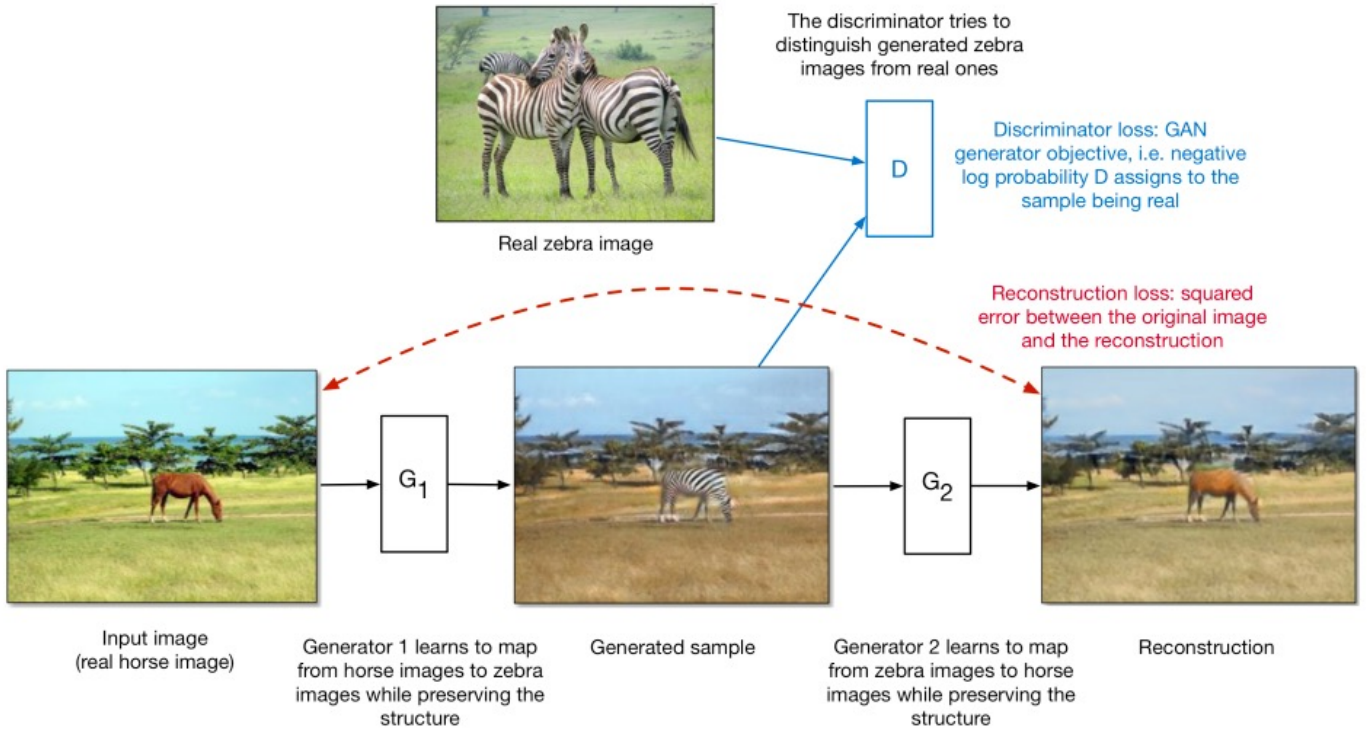
If we had paired data (same content in both styles), this would be a supervised learning problem. But this is hard to find!

The CycleGAN architecture learns to do it from unpaired data.

- Train two different generator nets to go from style 1 to style 2, and vice versa.
- Make sure the generated samples of style 2 are indistinguishable from real images by a discriminator net.
- Make sure the generators are cycle-consistent:

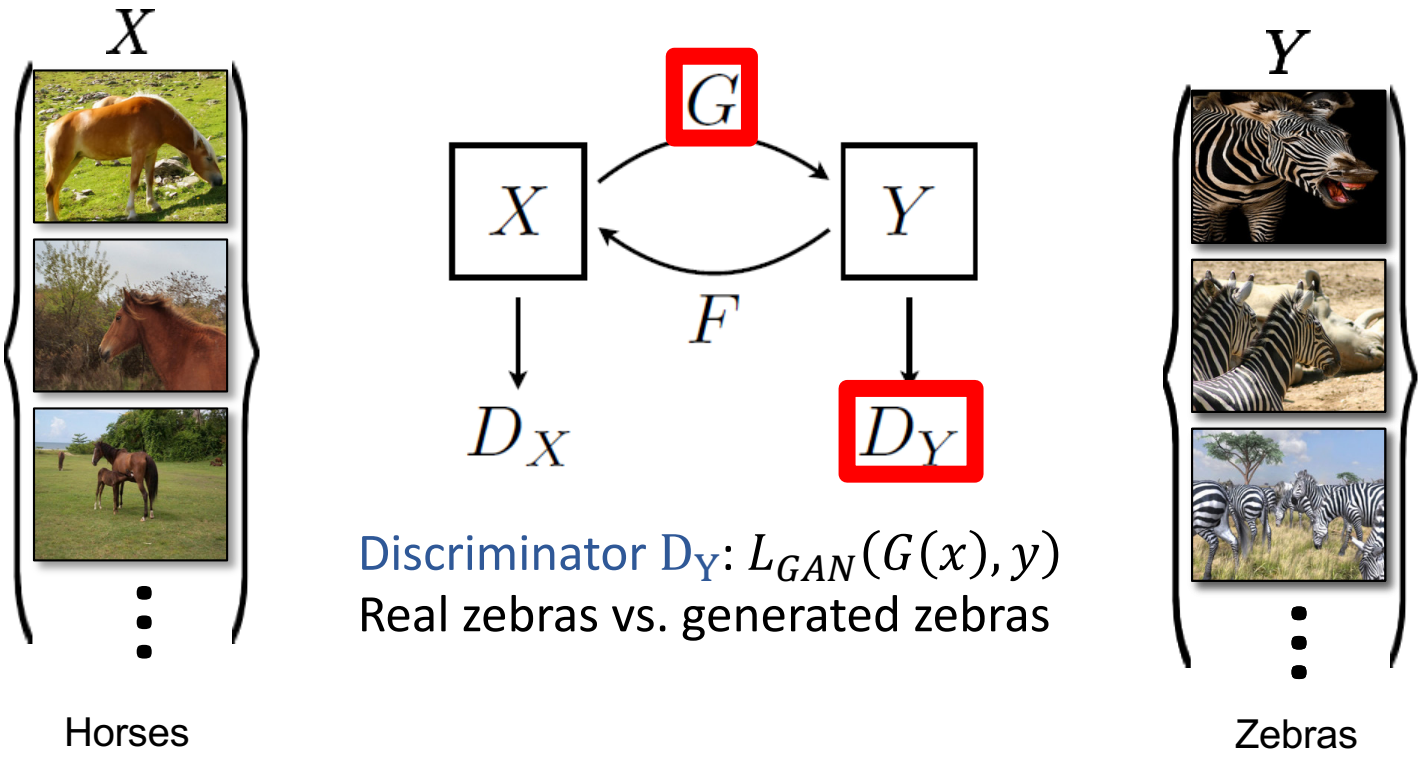
  Mapping from style 1 to style 2 and back again should give you almost the original image.

Conditional GANs: CycleGAN



Total loss = discriminator loss + reconstruction loss

Cycle GAN: Cycle Consistency

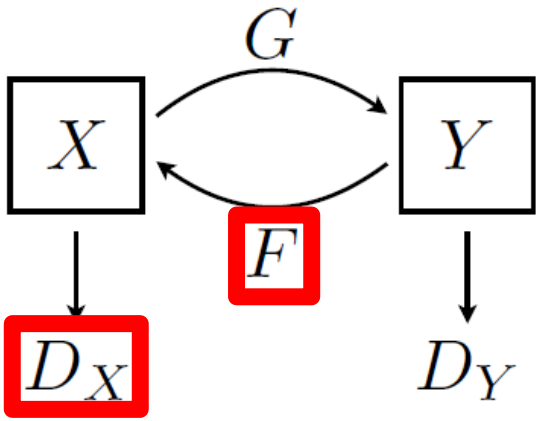


Discriminator D_Y: $L_{GAN}(G(x), y)$
Real zebras vs. generated zebras

Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Cycle Consistency

Horses



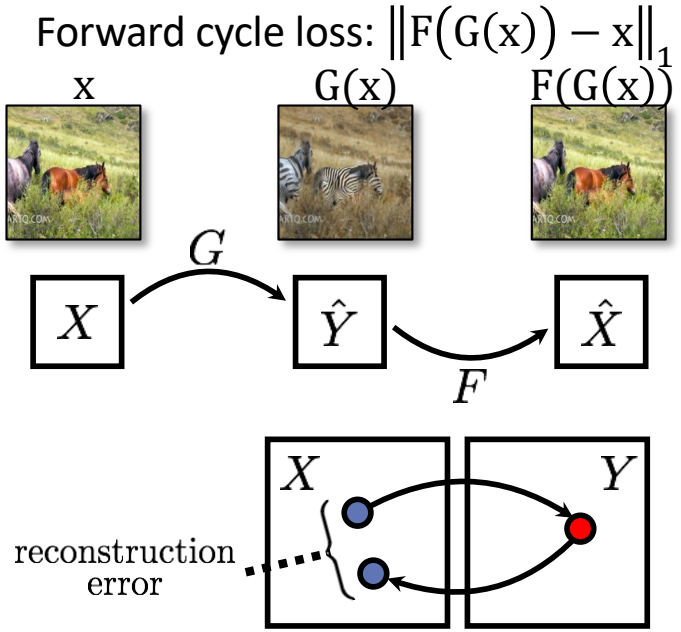
Discriminator D_Y: $L_{GAN}(G(x), y)$
Real zebras vs. generated zebras

Discriminator D_X: $L_{GAN}(F(y), x)$
Real horses vs. generated horses

Zebras

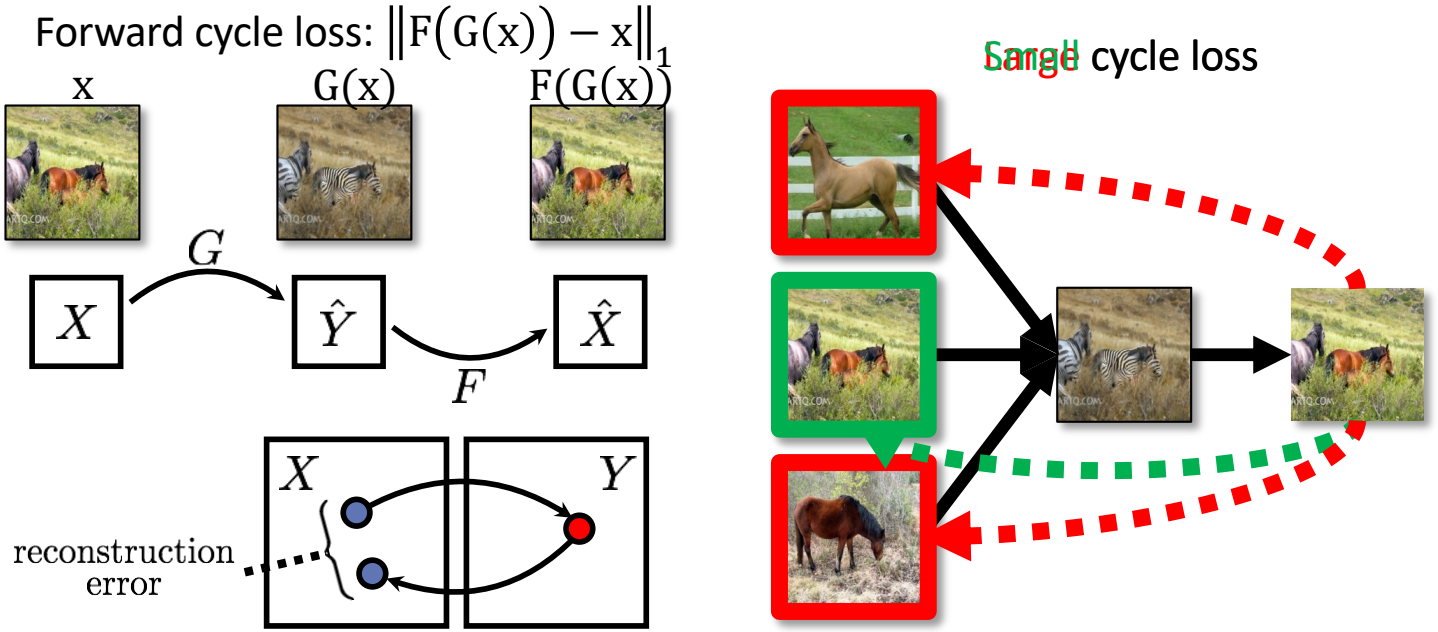
Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Cycle Consistency



Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Cycle Consistency



Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Training

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) =& \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] \\ &+ \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))],\end{aligned}$$

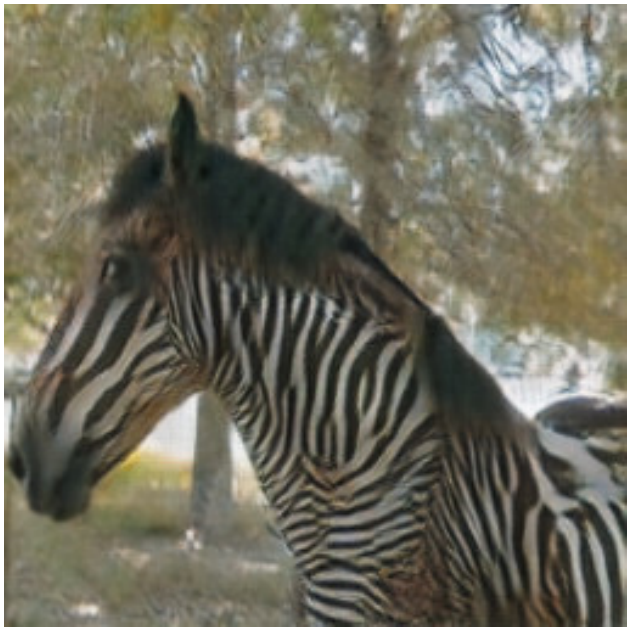
$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) =& \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ &+ \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].\end{aligned}$$

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) =& \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ &+ \lambda \mathcal{L}_{\text{cyc}}(G, F),\end{aligned}$$

$$G^*, F^* = \arg\min_{G,F} \max_{D_x, D_Y} \mathcal{L}(G, F, D_X, D_Y).$$

Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

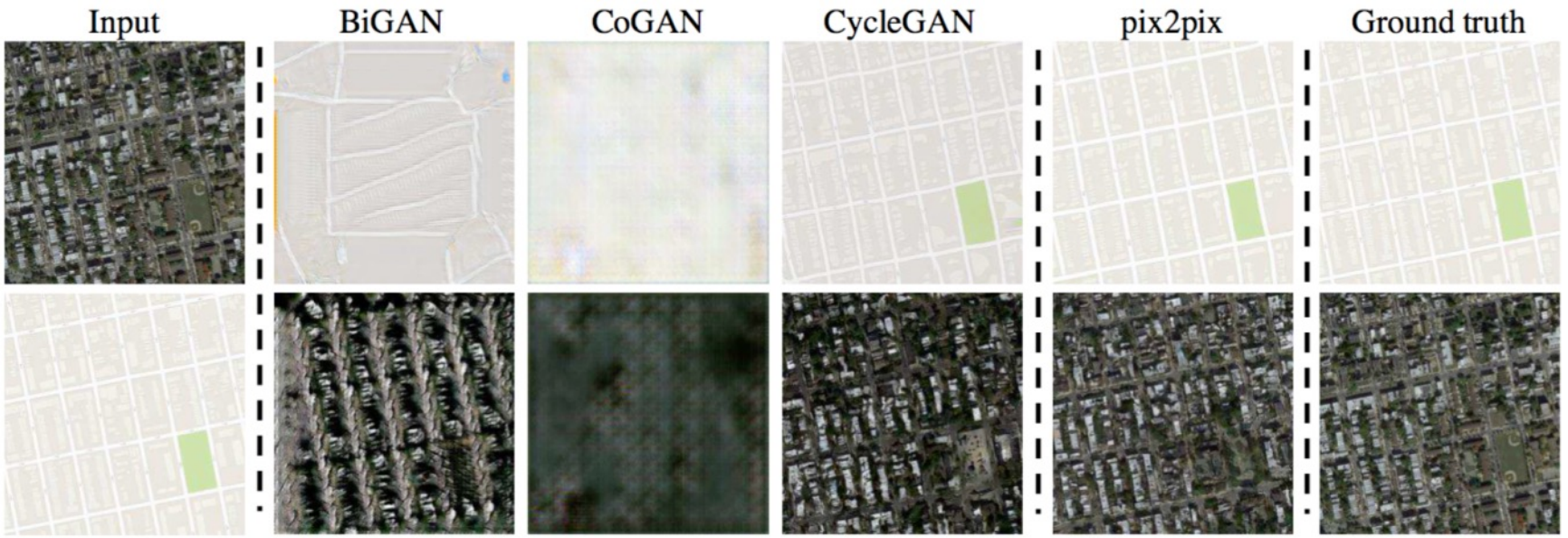
Cycle GAN: Examples



Pix2pix / CycleGAN

Cycle GAN: Application

Style transfer between aerial photos and maps:



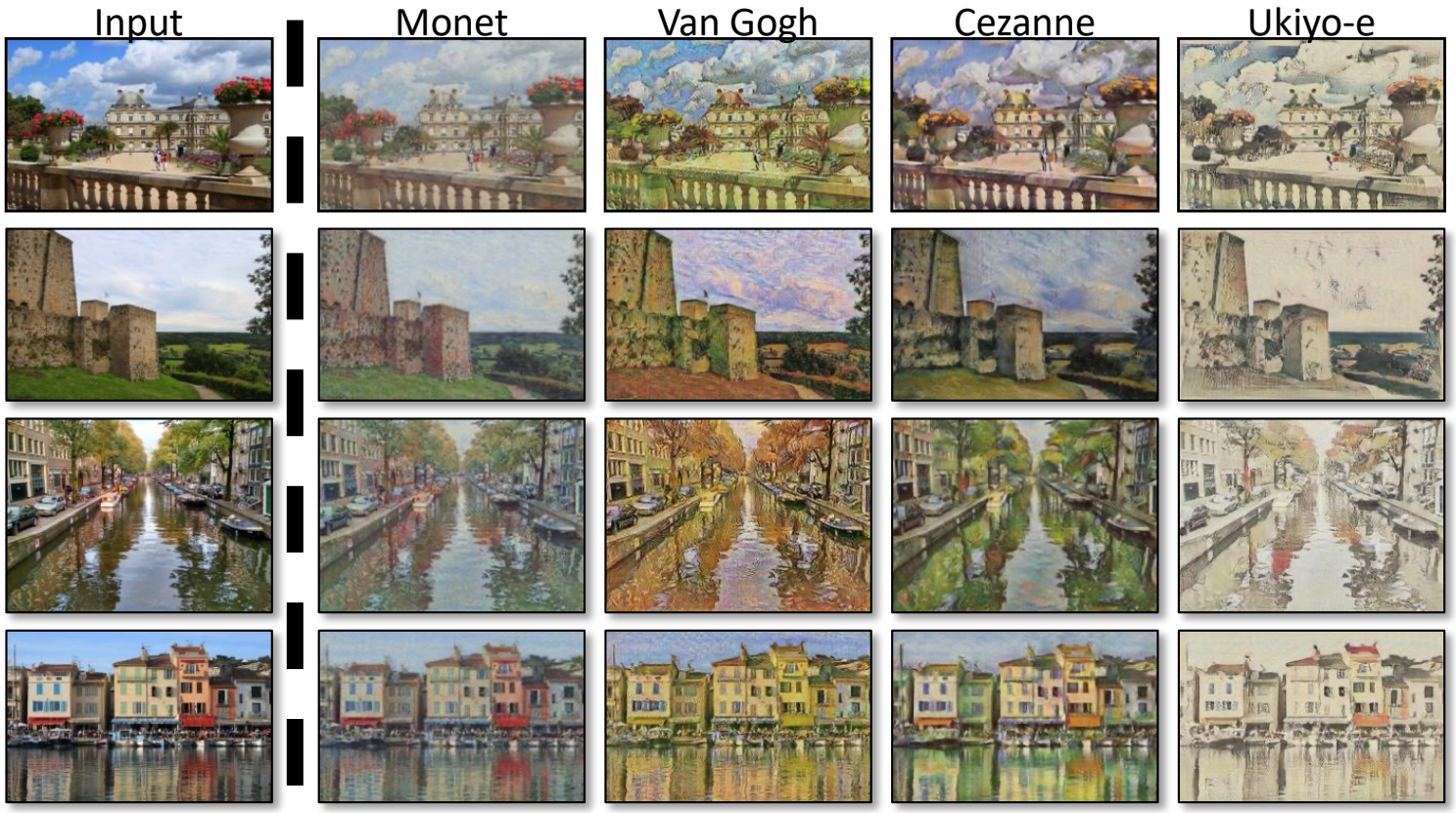
Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Application

Style transfer between road scenes and semantic segmentations (labels of every pixel in an image by object category):

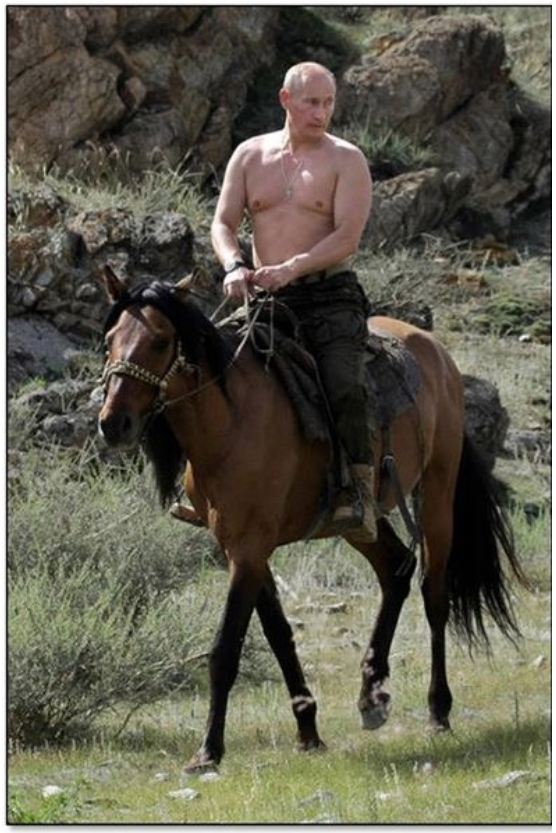
Zhu et al., "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks", ICCV 2017

Cycle GAN: Application



Pix2pix / CycleGAN

Cycle GAN: Failure Example



Pix2pix / CycleGAN

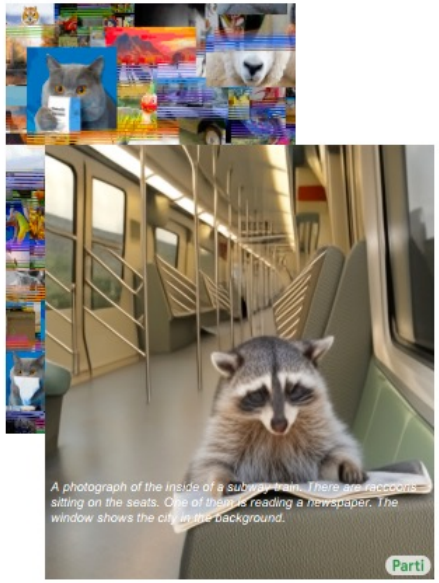
Cycle GAN: Failure Example

Pix2pix / CycleGAN

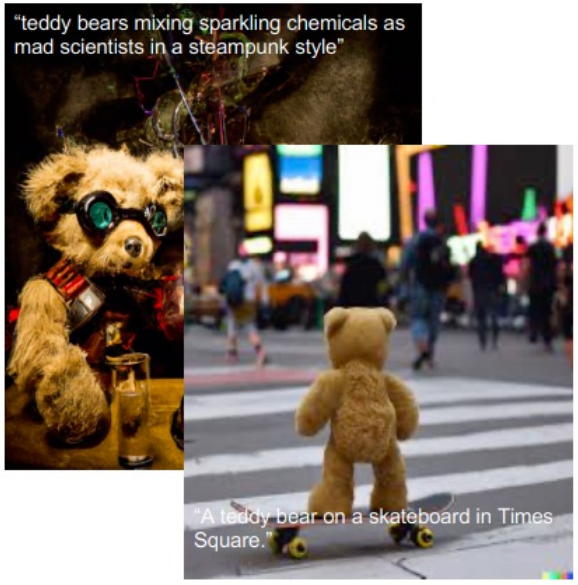
Conditional GANs

What other modalities?

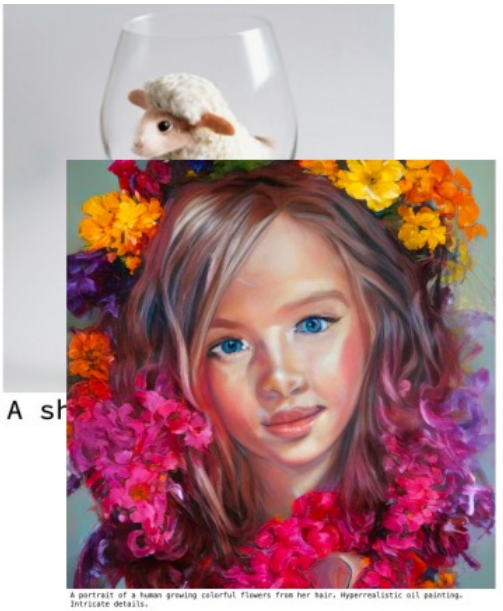
Conditional Textual GANs: Motivation



Autoregressive models
(Image GPT, Parti)



Diffusion models
(DALL-E 2, Imagen)



GANs, Masked GIT
(GigaGAN, MUSE)

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Start

First the farmer gives hay to the goat. Then the farmer gets milk from the cow.

Step 1: Image Selection.
Step 2: Layout Optimization (Minimum overlap, Centrality, Closeness)

A Text-to-Picture Synthesis System for Augmenting Communication
Xiaojin Zhu, Andrew Goldberg, Mohamed Eldawy, Charles Dyer, and Bradley Strock. AAAI 2007

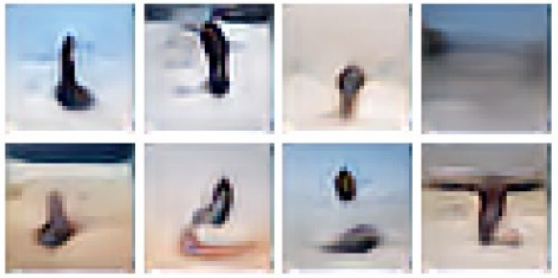
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Start DL

A stop sign is flying in blue skies.

A herd of elephants flying in the blue skies.

A toilet seat sits open in the grass field.

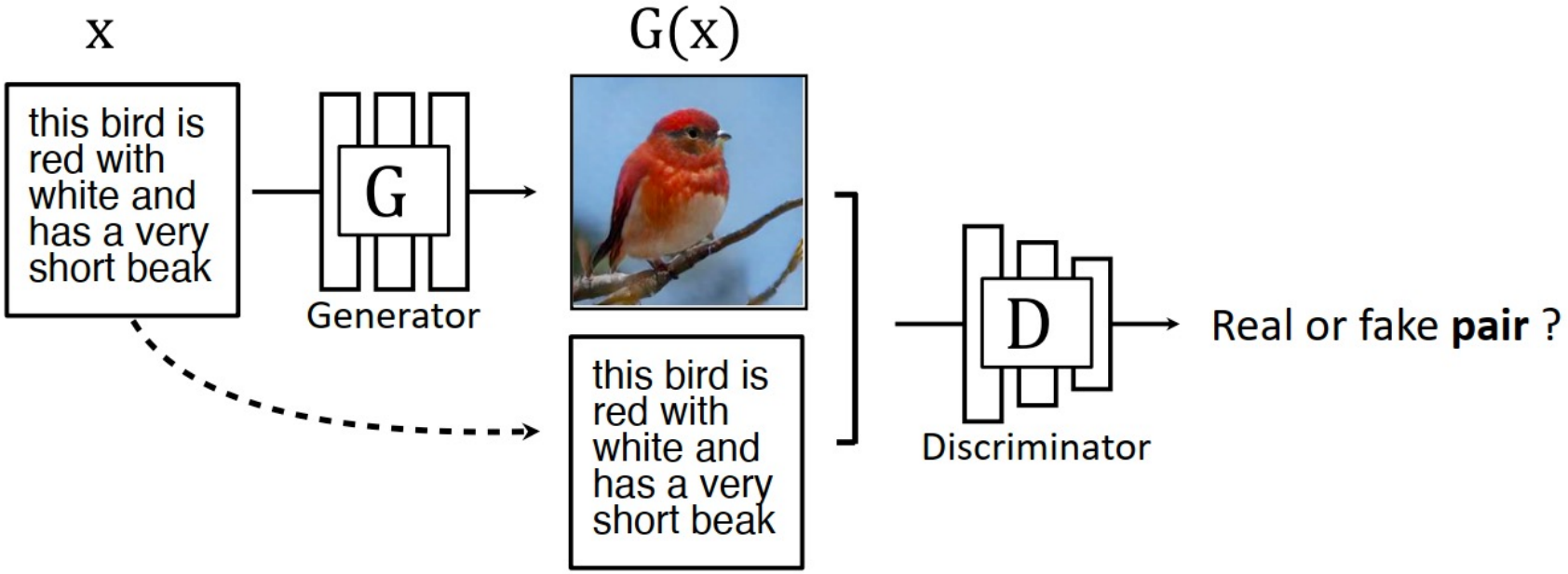


A person skiing on sand clad vast desert.

Generating Images from Captions with Attention.
Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. ICLR 2016

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GAN

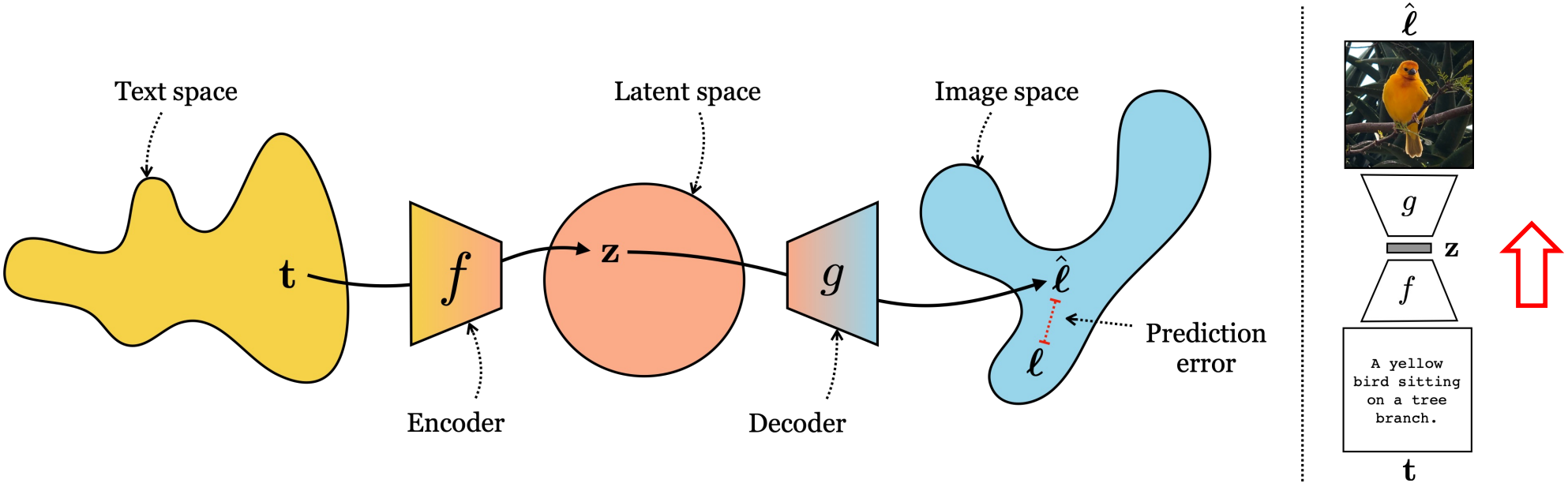


Input: **Text** → Output: **Photo**

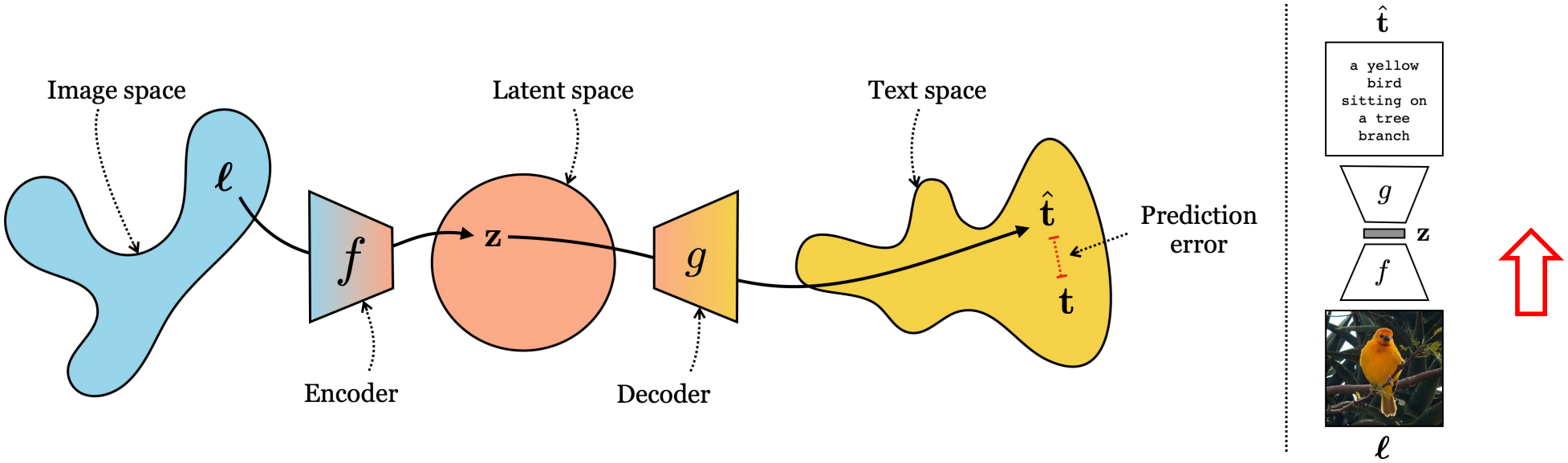
Text-to-Image Synthesis

StackGAN, StackGAN++ [Zhang et al., 2016 and 2017], AttnGAN [Xu et al., 2018]

Conditional Textual GANs: Representation



Conditional Textual GANs: Representation

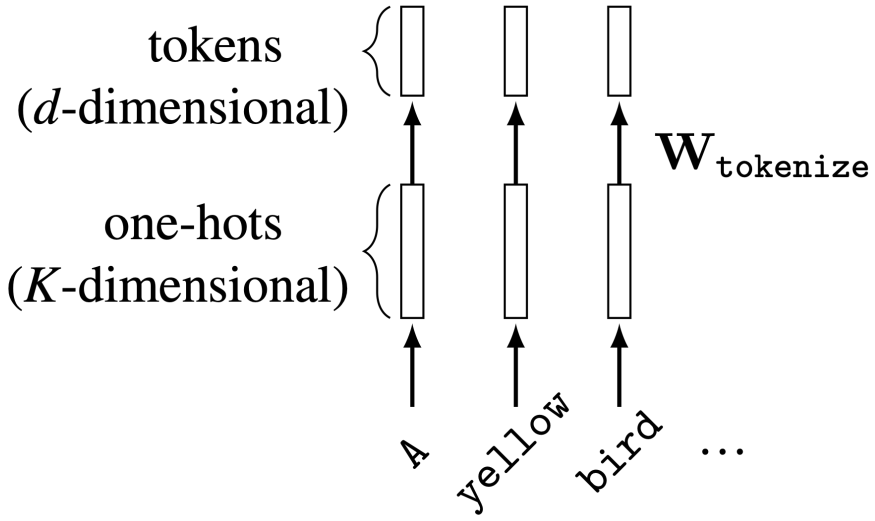


Conditional Textual GANs: Representation

How to model Text?

Conditional Textual GANs: Representation

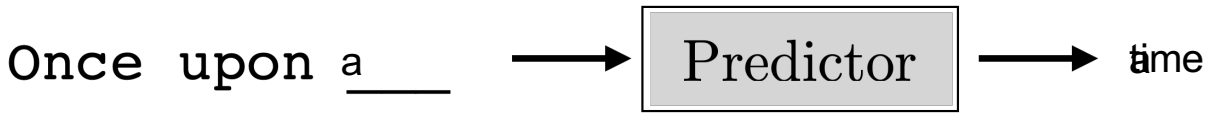
How to represent text as tokens?

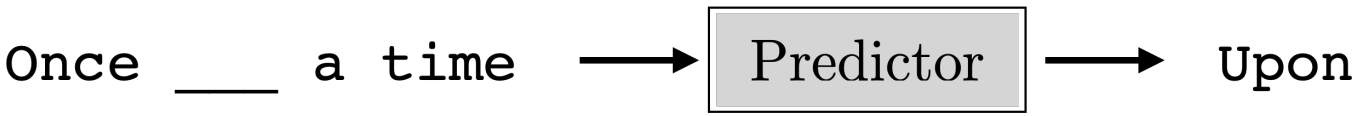


Note: sometimes the word "token" is used to refer to a unit of the discrete vocabulary we will model (the one-hots here). We use a more general definition, where a token can be discrete or continuous — each layer of a transformer consists of a set of tokens.

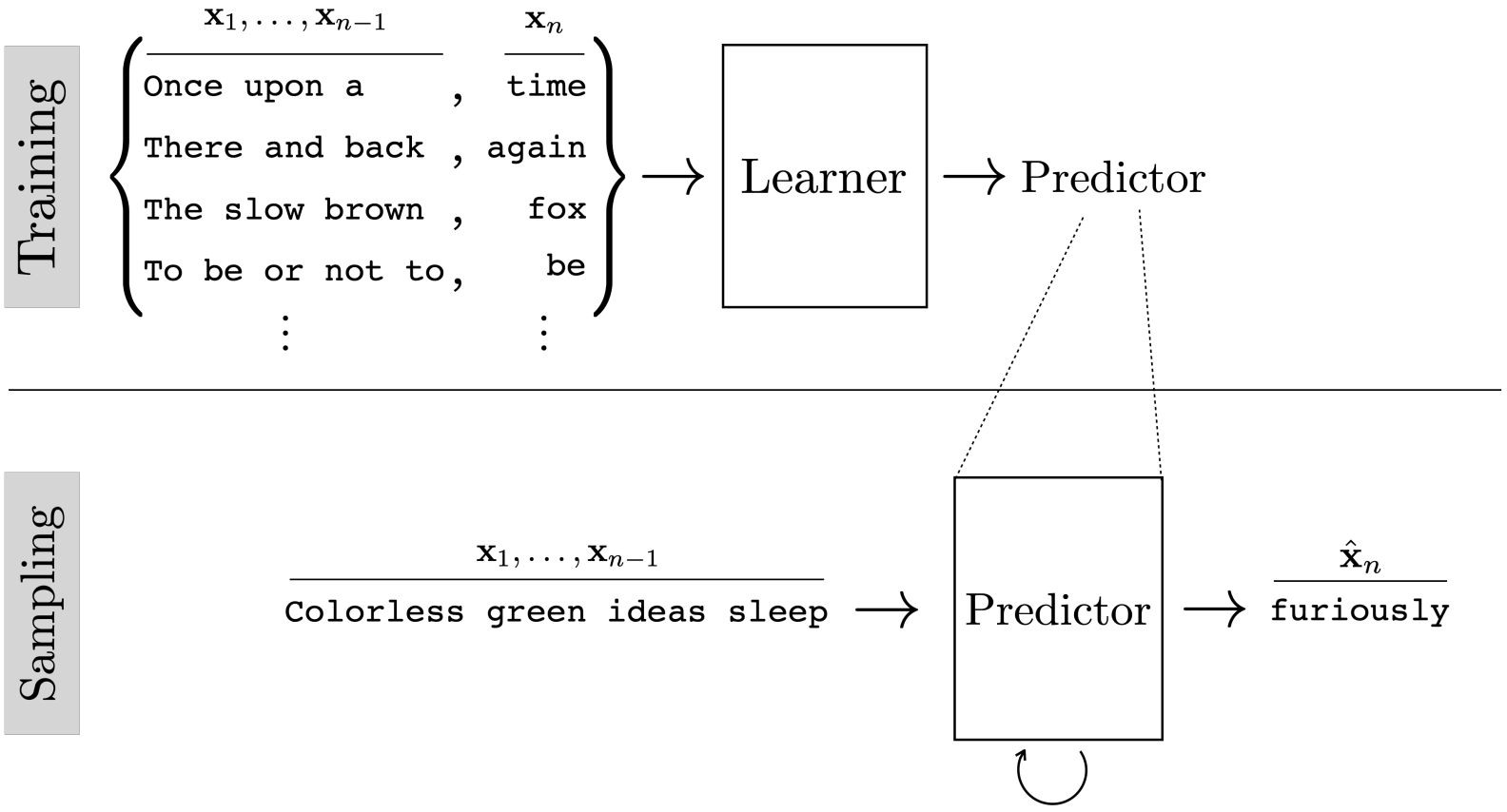
Conditional Textual GANs: Learning

Language Models — Autoregressive



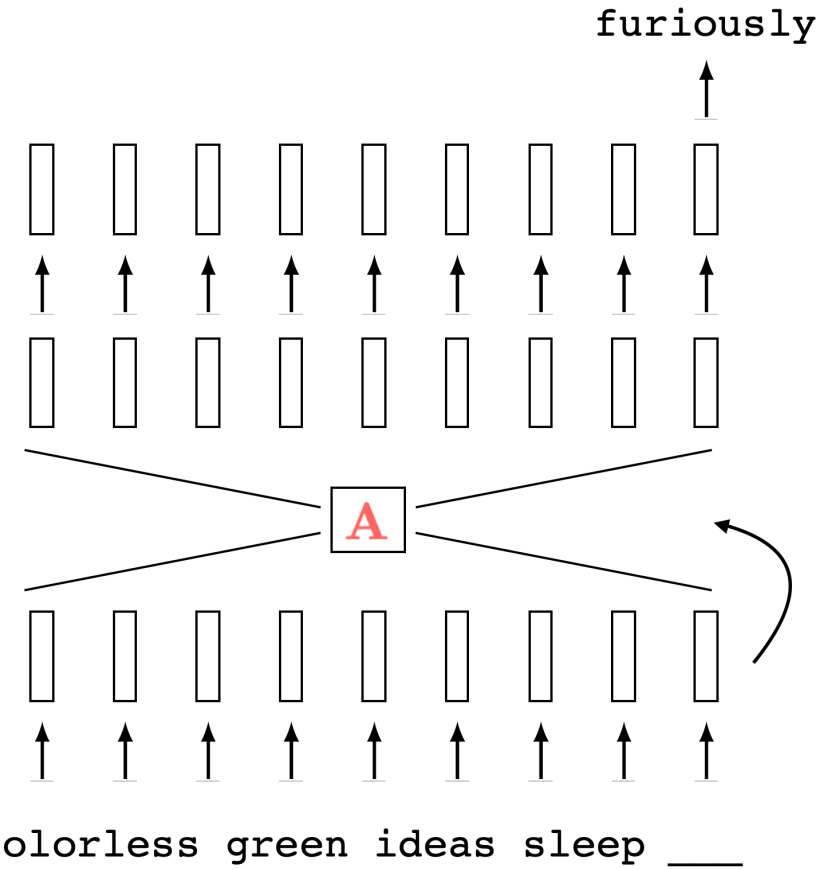


Conditional Textual GANs: Learning



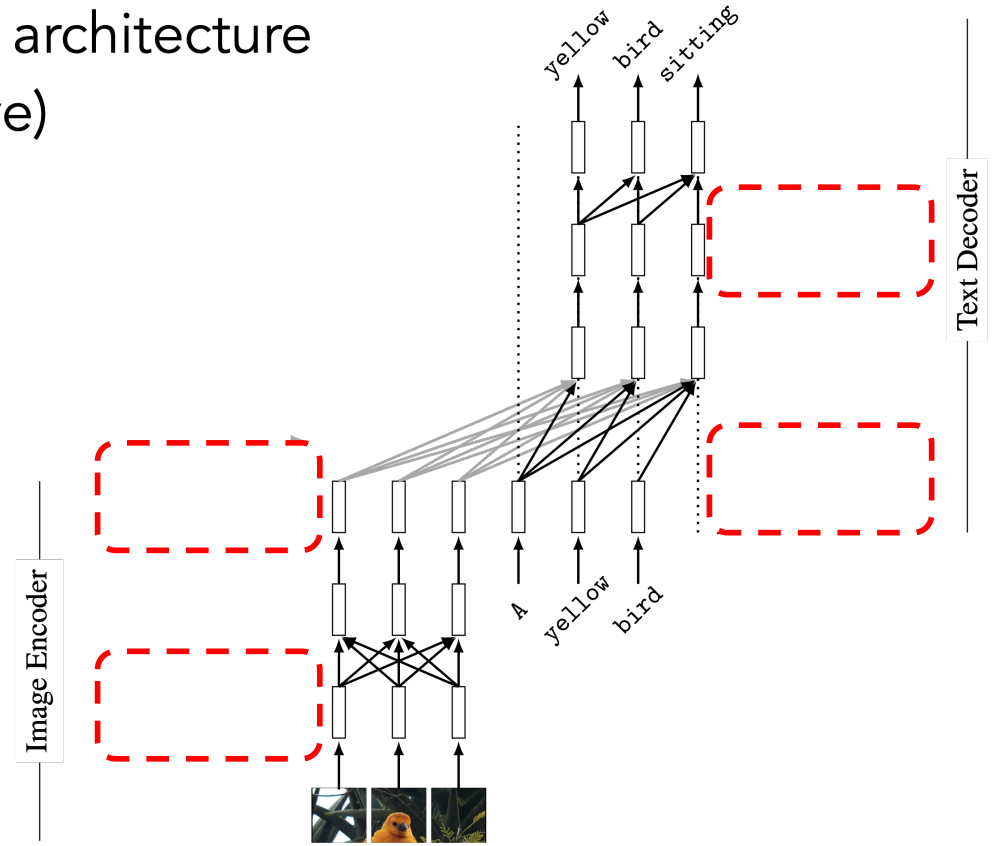
Conditional Textual GANs: Learning

Language Models — Autoregressive



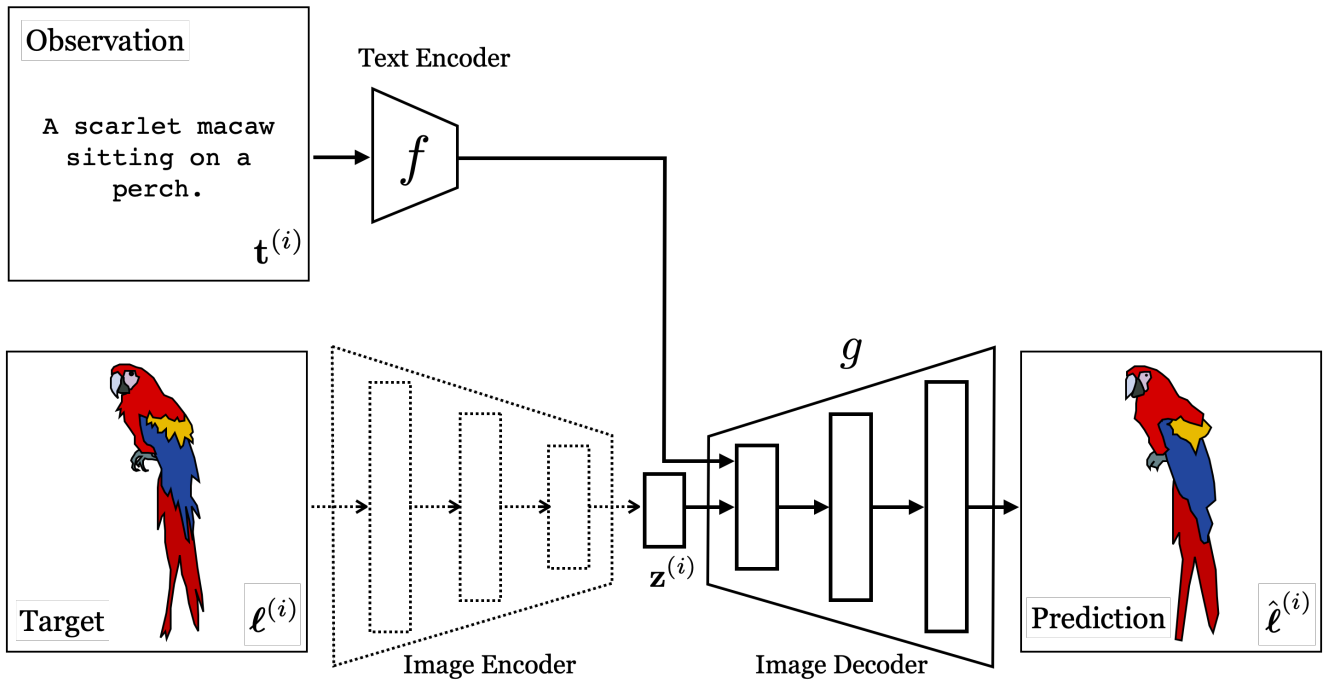
Conditional Textual GANs: Learning

Image-to-text architecture
(autoregressive)



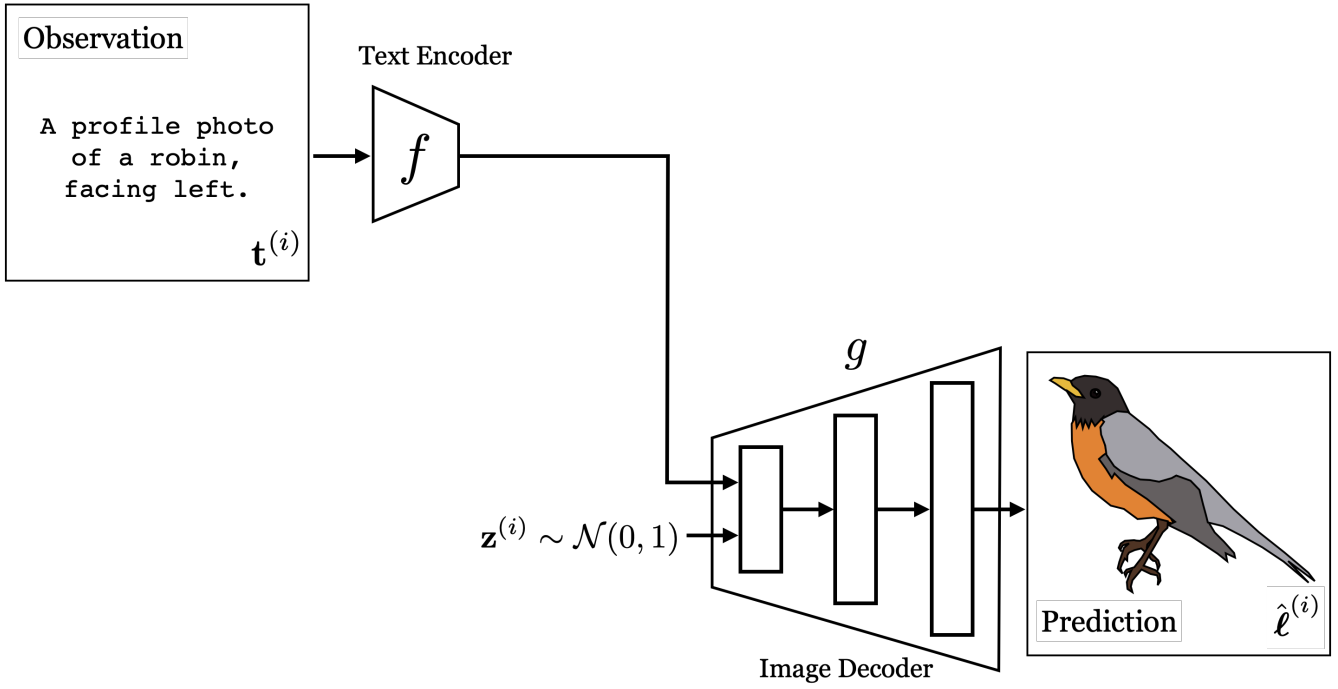
Conditional Textual GANs: Learning

Text-to-image architecture (cVAE) — training

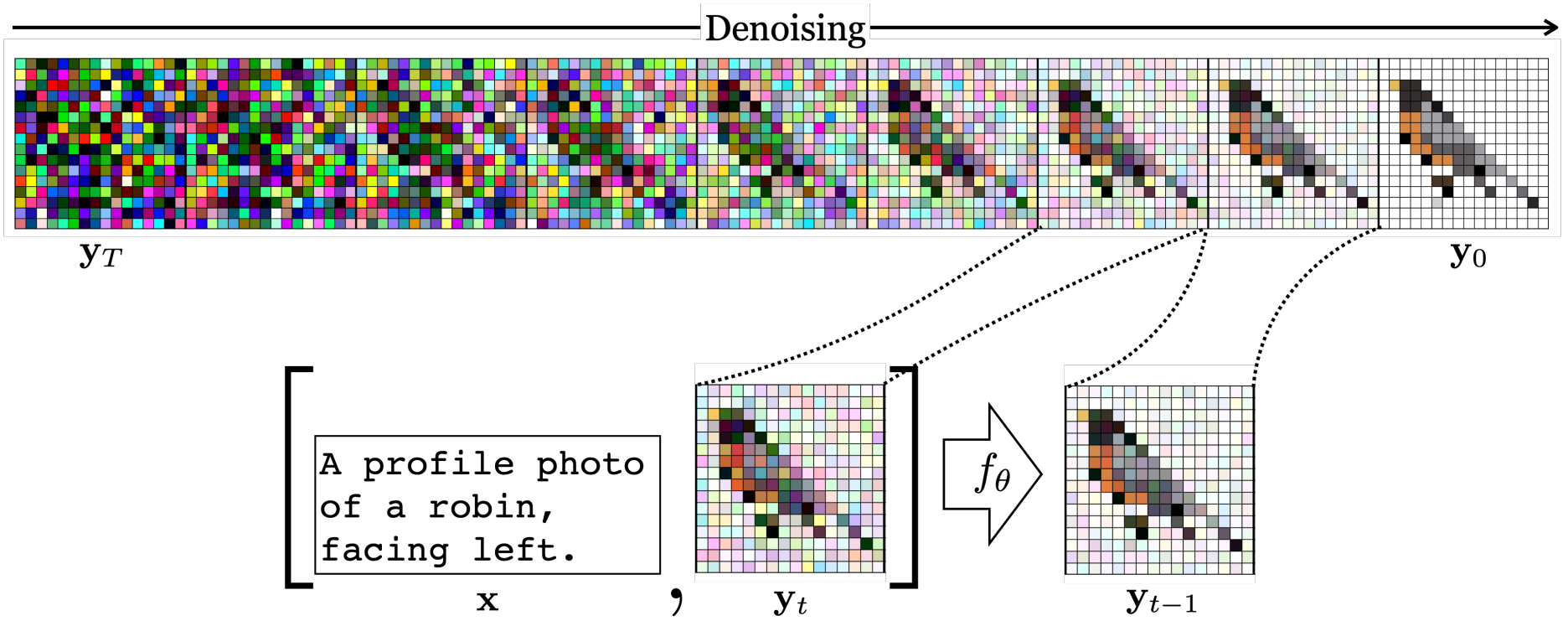


Conditional Textual GANs: Learning

Text-to-image architecture (cVAE) — inference



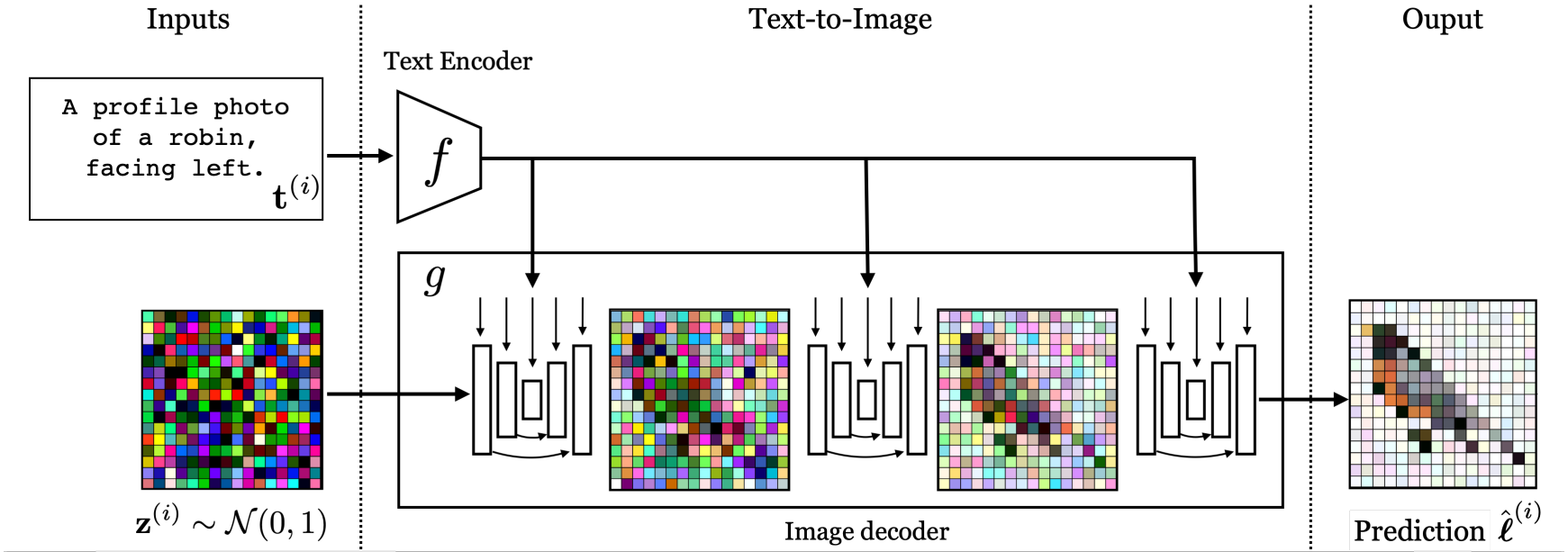
Conditional Textual GANs: Learning - Diffusion



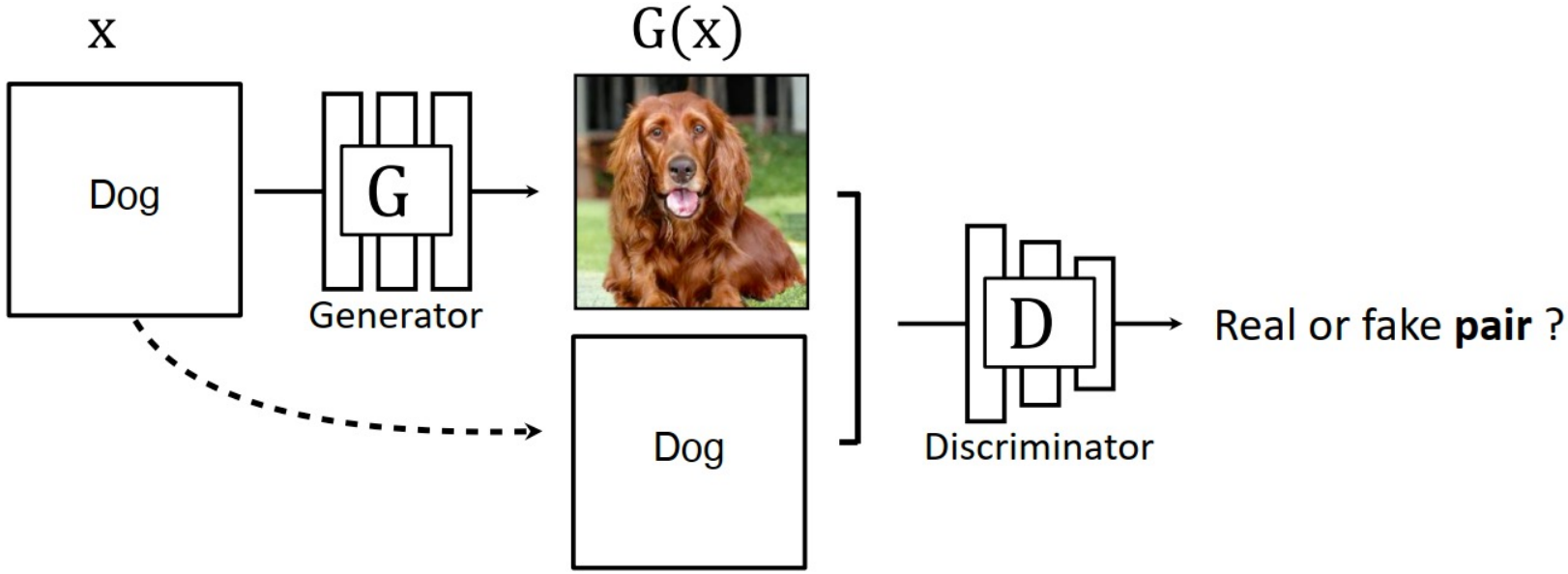
For example: DALL-E 2 [Ramesh et al. 2022], Stable Diffusion [Rombach*, Blattman* et al. 2022]

Conditional Textual GANs: Learning

Text-to-image architecture (diffusion)



Conditional Textual GANs: Class / Category

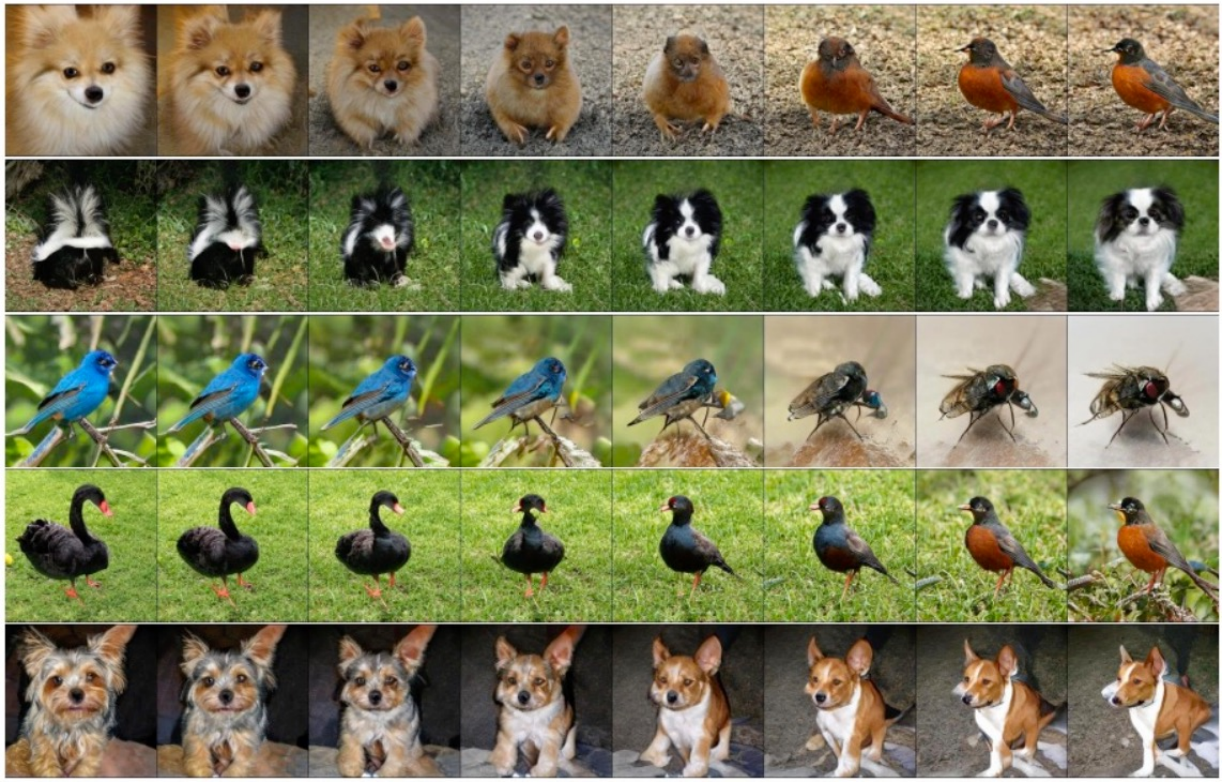


Input: **Class** → Output: **Photo**

Class-conditional GANs

cGANs [Mirza and Osindero. 2014], SAGAN [Zhang et al., 2018], BigGAN [Brock et al., 2019] StyleGAN-XL [Sauer et al., 2022]

Conditional Textual GANs: Class Conditioned – BigGAN - Feature Space Interpolation



cGANs [Mirza and Osindero. 2014], SAGAN [Zhang et al., 2018], BigGAN [Brock et al., 2019] StyleGAN-XL [Sauer et al., 2022]

Conditional Textual GANs: Details

this small bird has a pink breast and crown, and black primaries and secondaries.

this magnificent fellow is almost all black with a red crest, and white cheek patch.

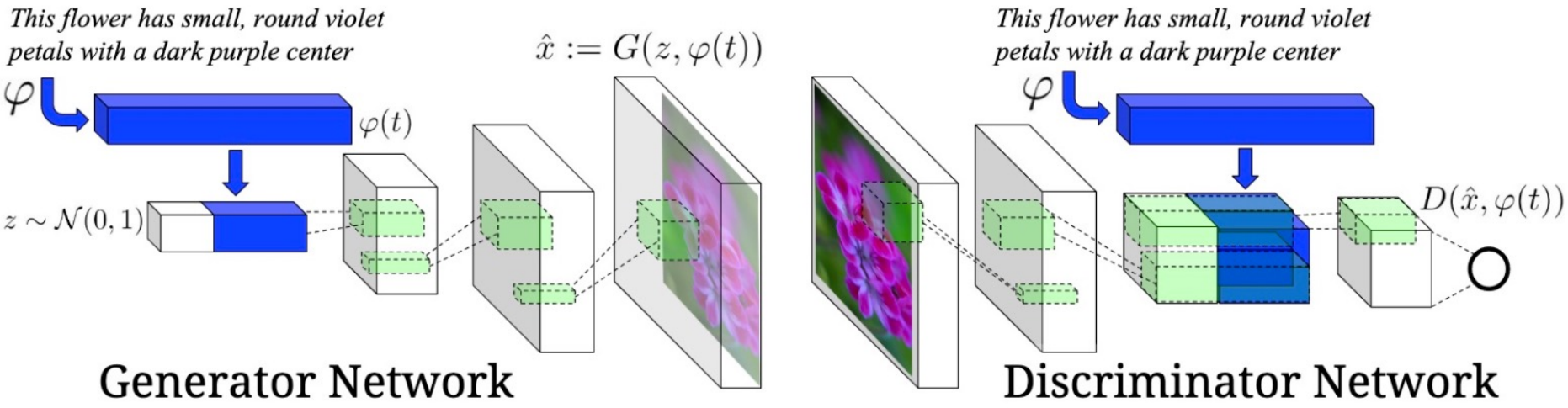
the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

Generative Adversarial Text to Image Synthesis
Scott Reed et al., ICML 2016

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Details



Conditional GAN + CNN + concatenation

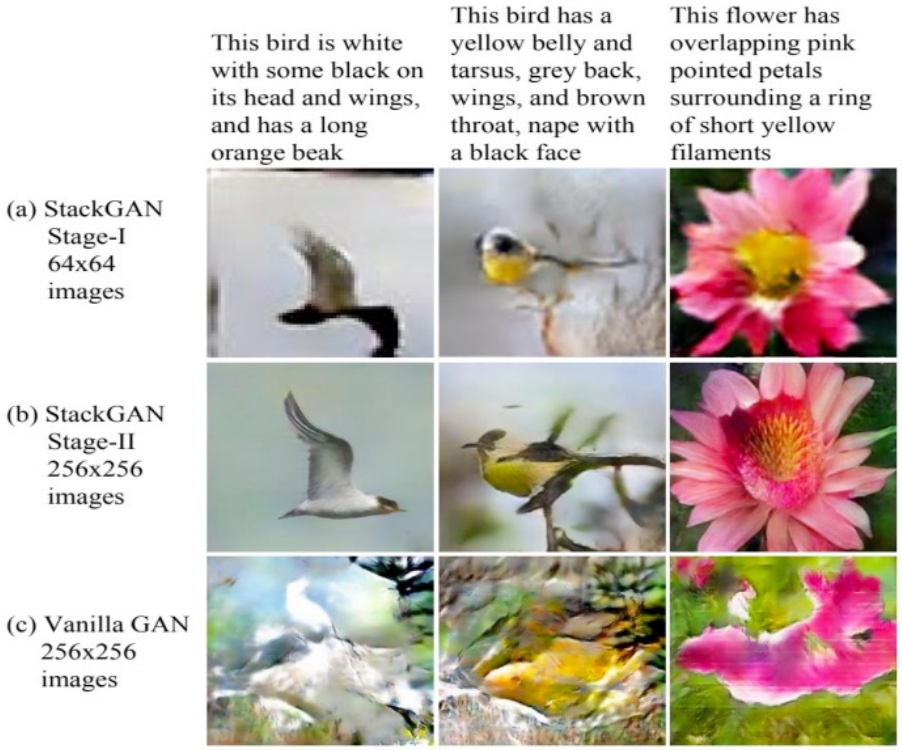
Generative Adversarial Text to Image Synthesis
Scott Reed et al., ICML 2016

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs

How to improve resolution?

Conditional Textual GANs: Two-stage Model

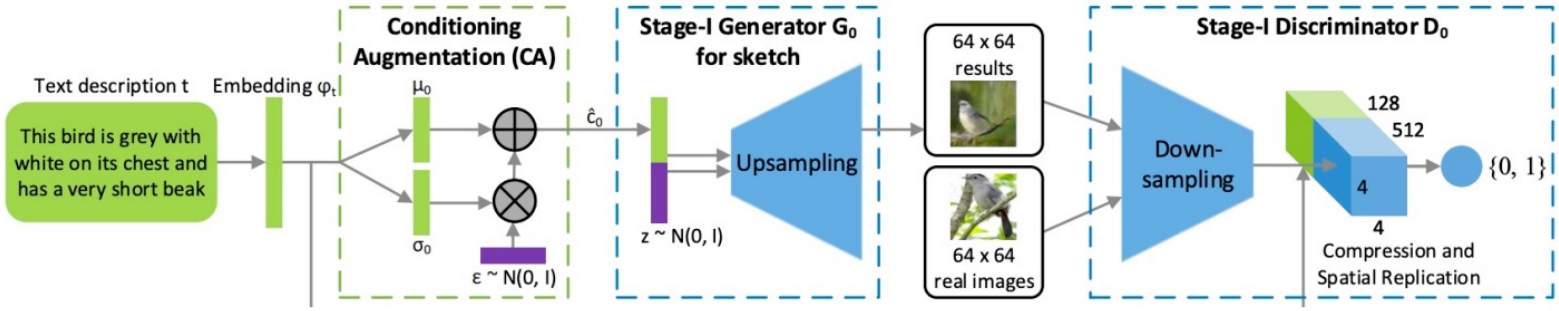


Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
Han Zhang et al., ICCV 2017

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Two-stage Model



Two-stage Conditional GAN + CNN + concatenation

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
Han Zhang et al., ICCV 2017

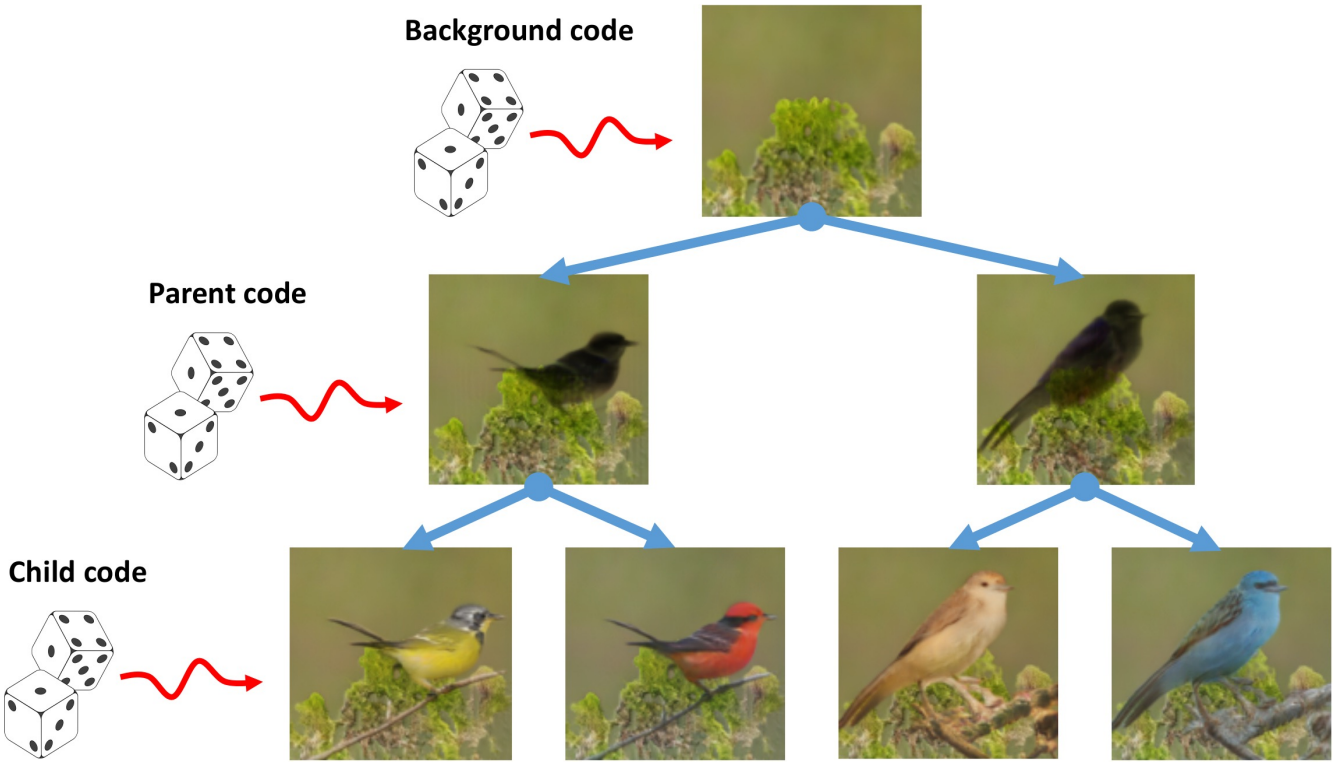
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Two-stage Model

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
Han Zhang et al., ICCV 2017

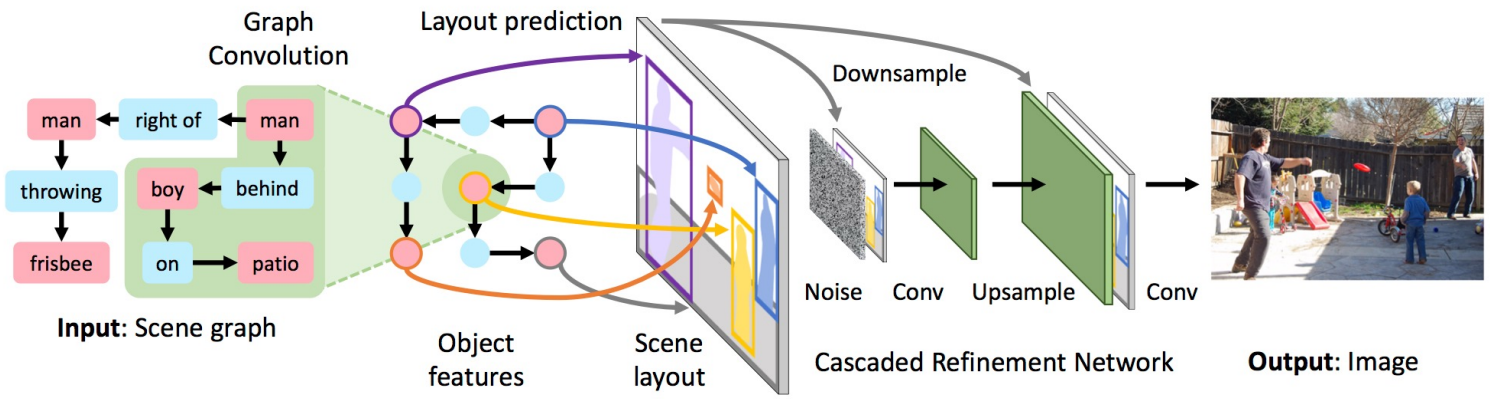
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional GANs: Multi-stage Model



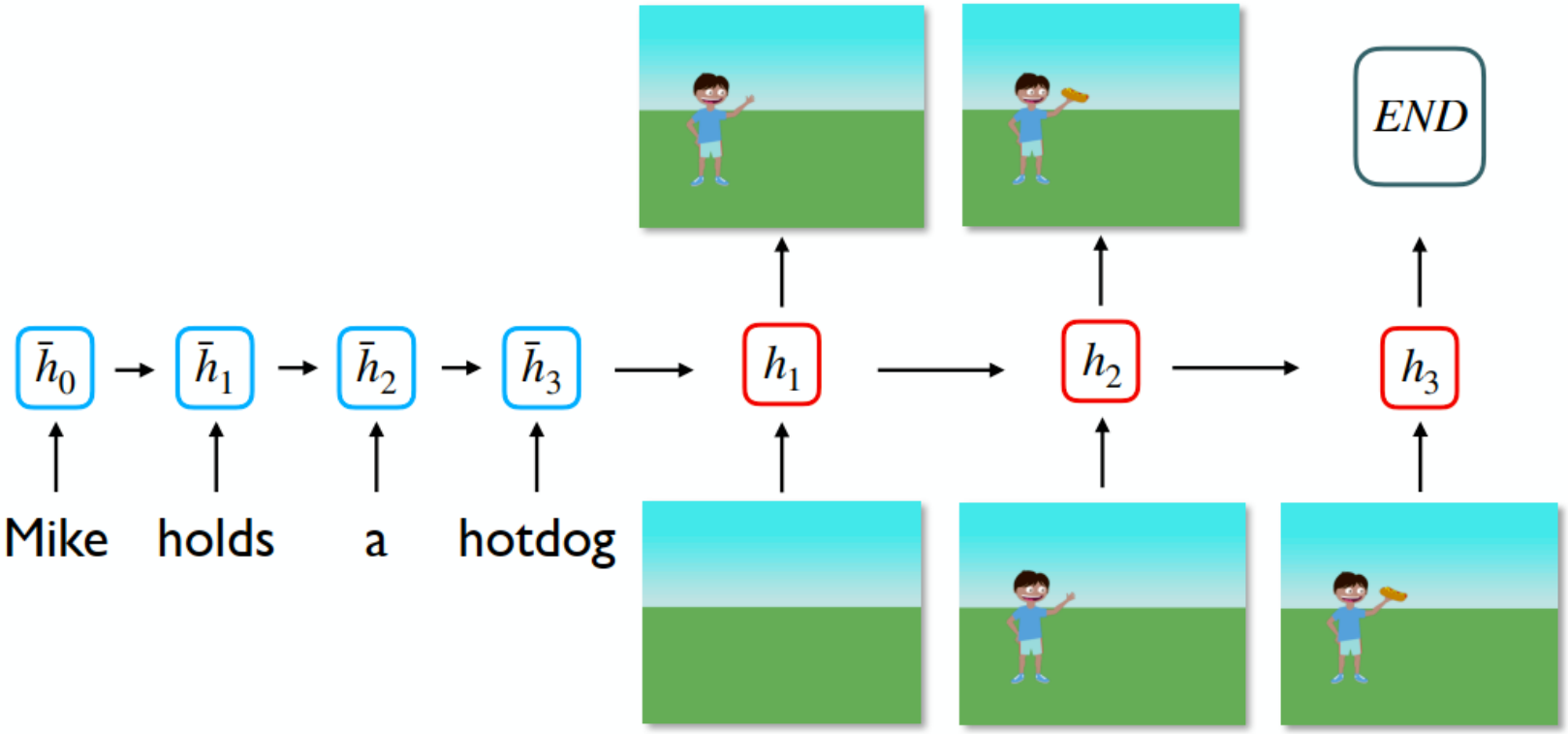
Singh et al., "FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery", CVPR 2019

Conditional GANs: Multi-stage Model



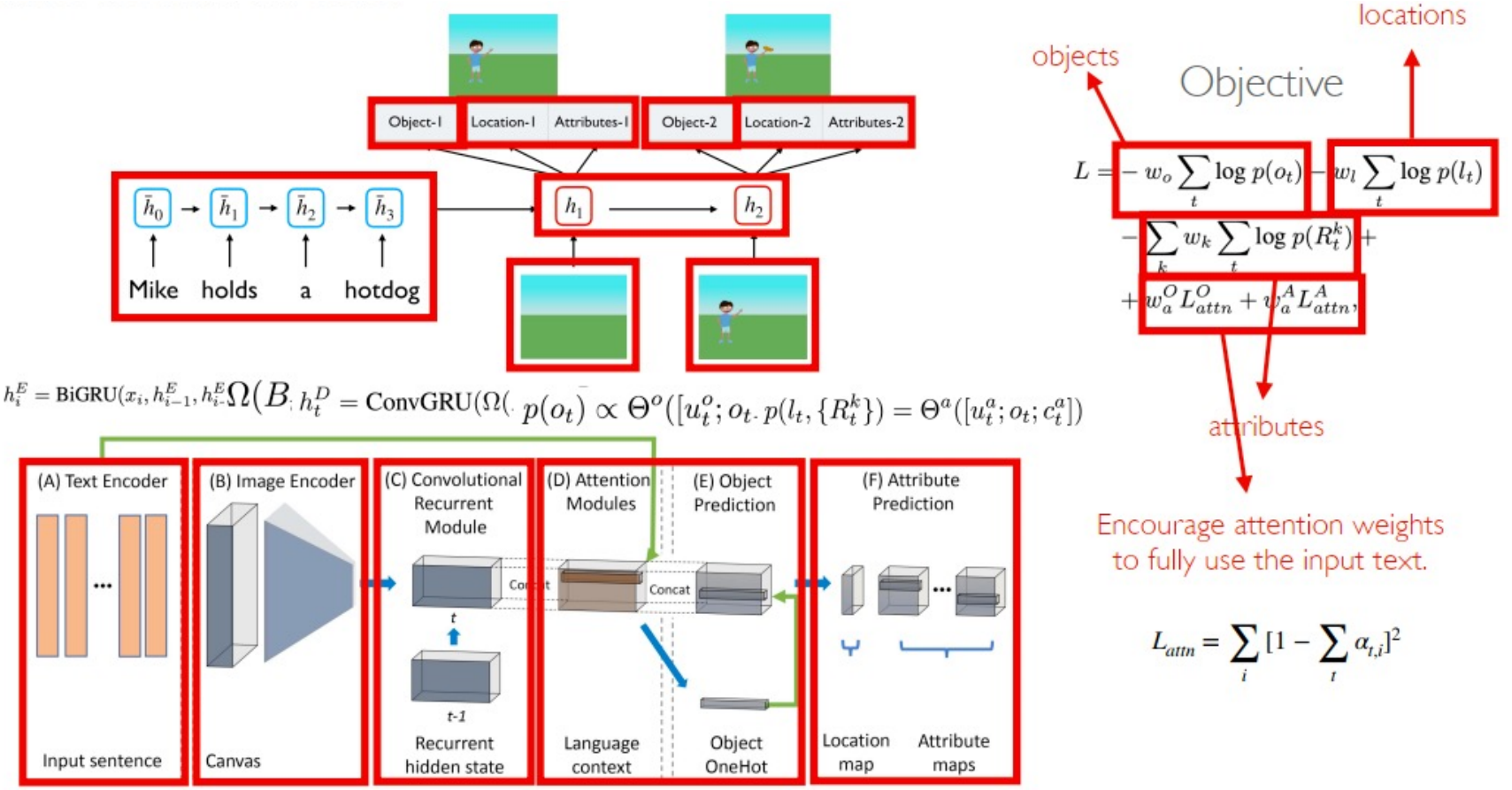
Johnson et al., "Image Generation from Scene Graphs", CVPR 2018

Text to Scene as Machine Translation!



Text2Scene: Generating Compositional Scenes from Textual Descriptions Fuwen Tan, Song Feng, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. CVPR 2019

Text to Scene as Machine Translation!



$$h_i^E = \text{BiGRU}(x_i, h_{i-1}^E, h_{i-1}^E) \quad \Omega(B \quad h_t^D = \text{ConvGRU}(\Omega(\quad p(o_t) \propto \Theta^o([u_t^o; o_t \quad p(l_t, \{R_t^k\}) = \Theta^a([u_t^a; o_t; c_t^a])$$

$$L = -w_o \sum_t \log p(o_t) - w_l \sum_t \log p(l_t)$$
$$- \sum_k w_k \sum_t \log p(R_t^k) +$$
$$+ w_a^O L_{attn}^O + w_a^A L_{attn}^A,$$

$$L_{attn} = \sum_i [1 - \sum_t \alpha_{t,i}]^2$$

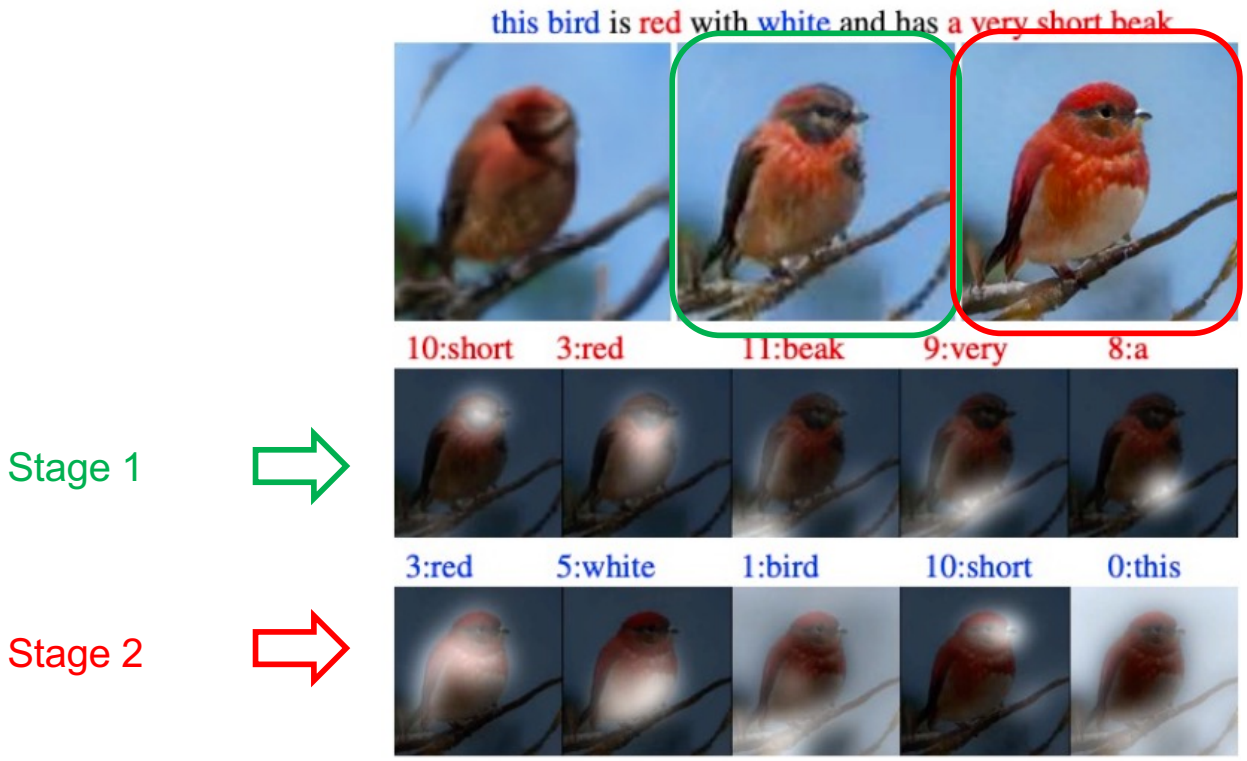
Text2Scene: Generating Compositional Scenes from Textual Descriptions Fuwen Tan, Song Feng, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. CVPR 2019

Text to Scene as Machine Translation!

https://vislang.ai/text2scene

Text2Scene: Generating Compositional Scenes from Textual Descriptions Fuwen Tan, Song Feng, Vicente Ordonez. Intl. Conference on Computer Vision and Pattern Recognition. CVPR 2019

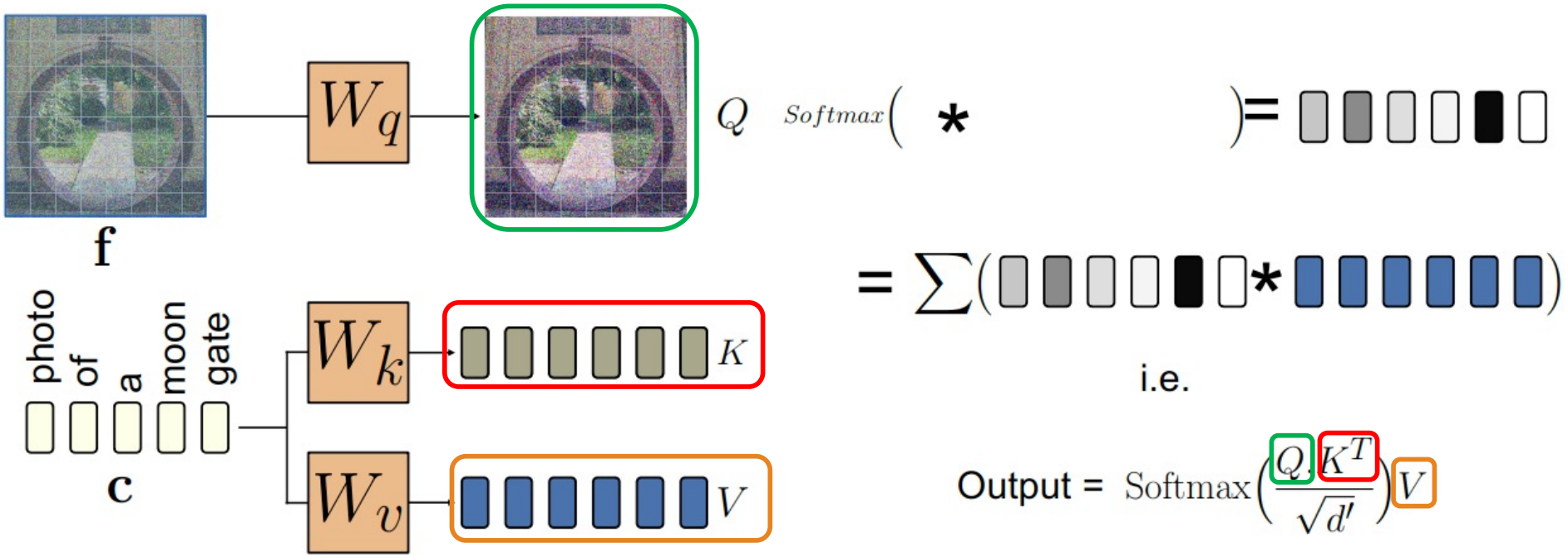
Conditional Textual GANs: + Attention



AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
Tao Xu et al., CVPR 2018

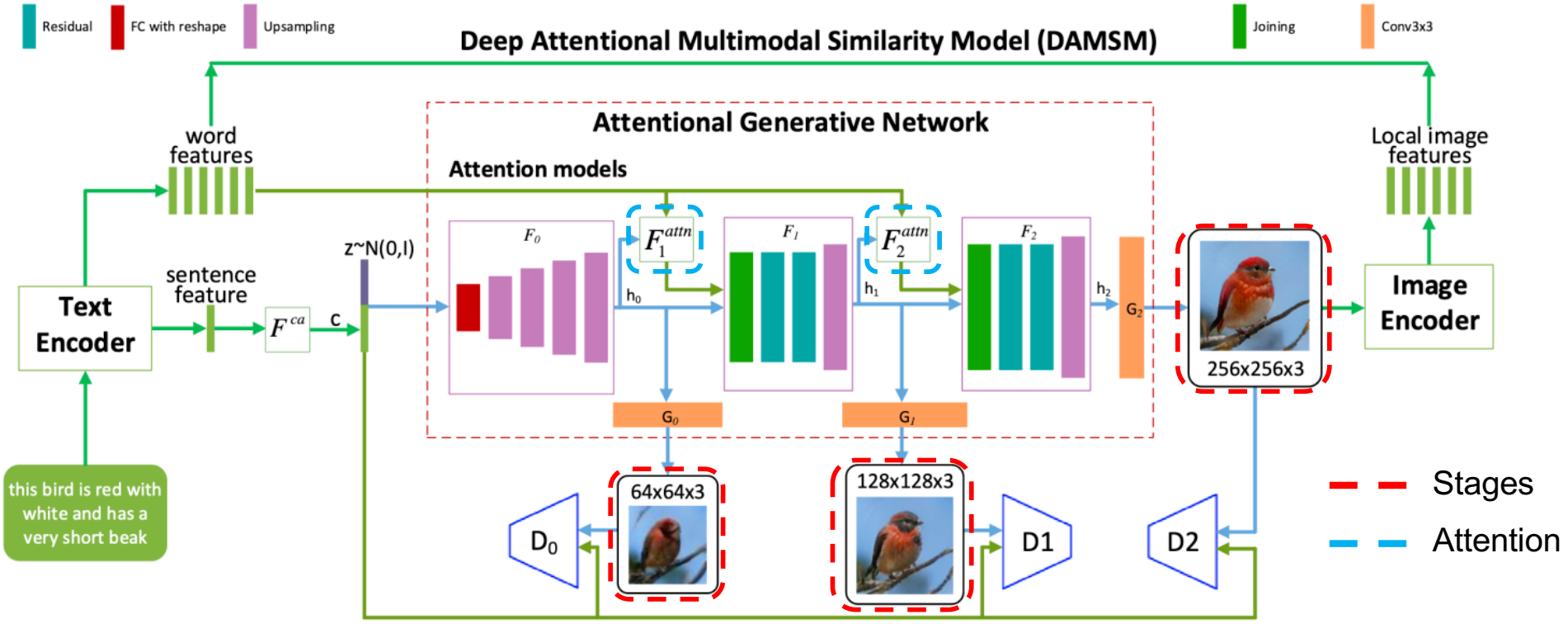
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Cross-attention



Slides from [Kumari et al., CVPR 2023]

Conditional Textual GANs: Cross-attention



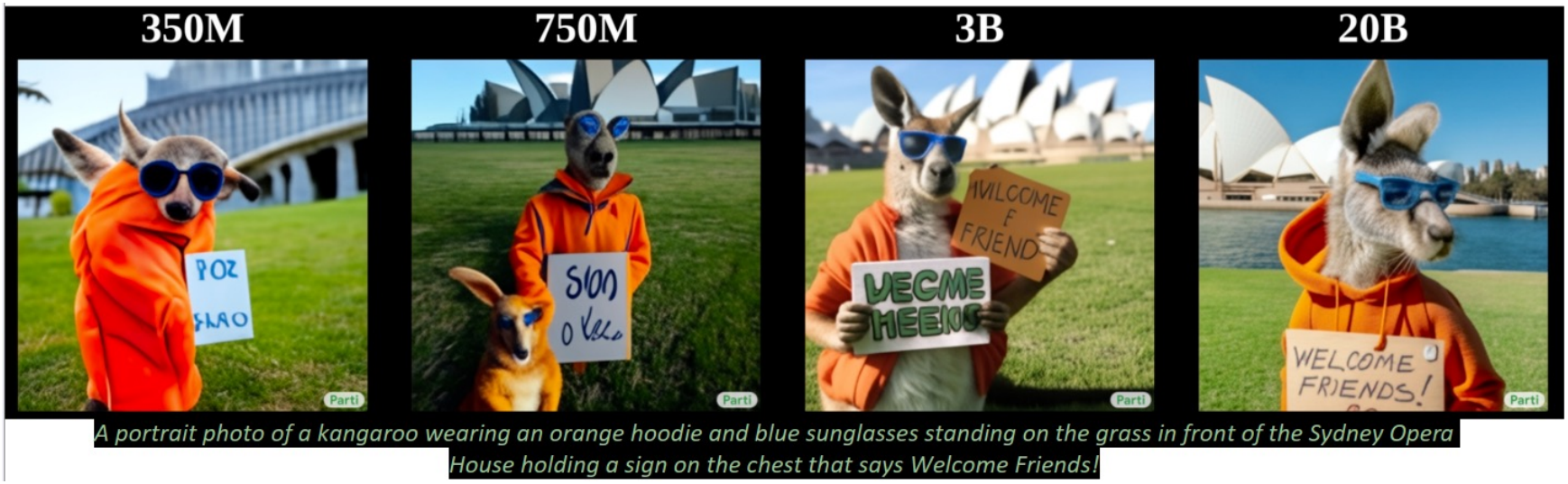
AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
Tao Xu et al., CVPR 2018

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Scaling VQGAN

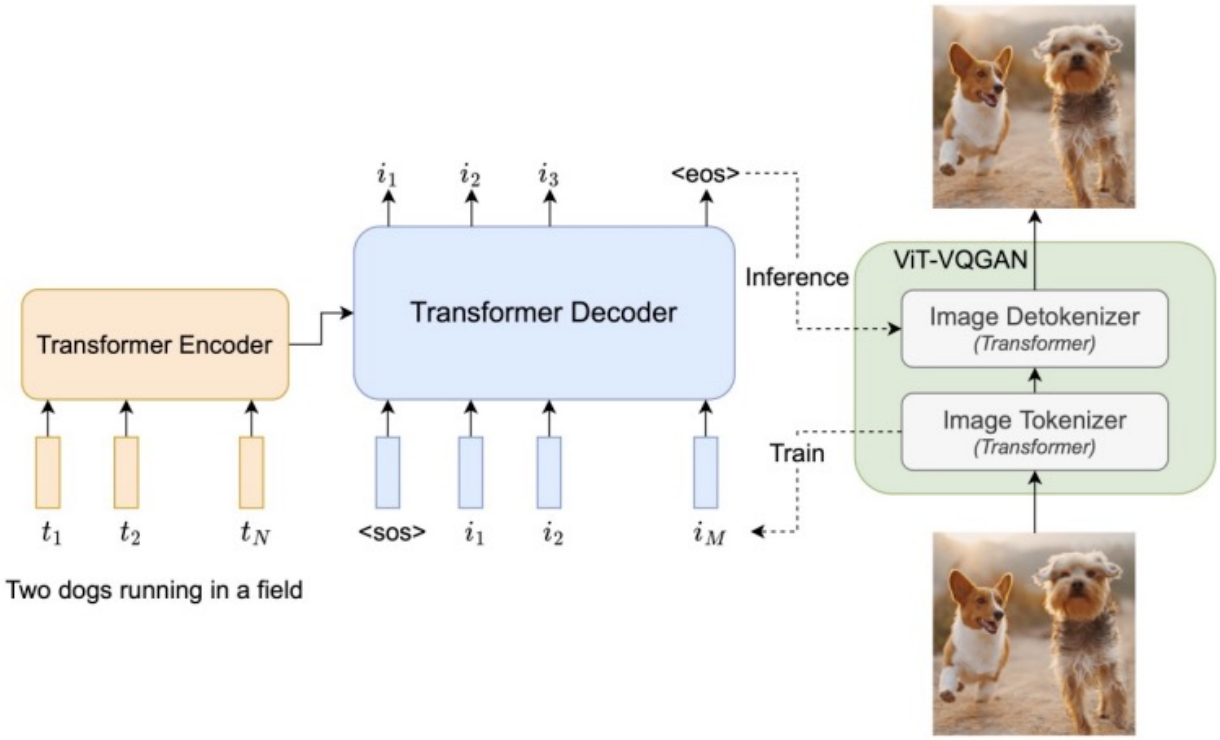
See released “Parti” paper by Google (text-to-image model)

- https://parti.research.google/



Slide credit: Robin Rombach

Conditional Textual GANs: Scaling VQGAN



Transformer-based Encoder/Decoder + Transformer-based Autoregressive models

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Diffusion Models!

great results for image synthesis

Denoising Diffusion Probabilistic Models

Jonathan Ho, Ajay Jain, et al

https://arxiv.org/abs/2006.11239

Diffusion Models beat GANs on Image Synthesis

Prafulla Dhariwal, Alex Nichol

https://arxiv.org/abs/2105.05233

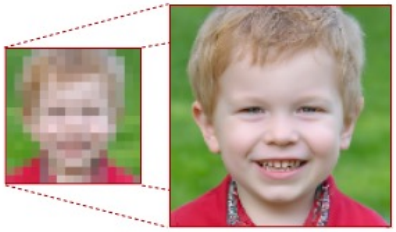


Image Super-Resolution via Iterative Refinement

Chitwan Saharia, et al

https://arxiv.org/abs/2104.07636

… but very expensive :(

Slide credit: Robin Rombach

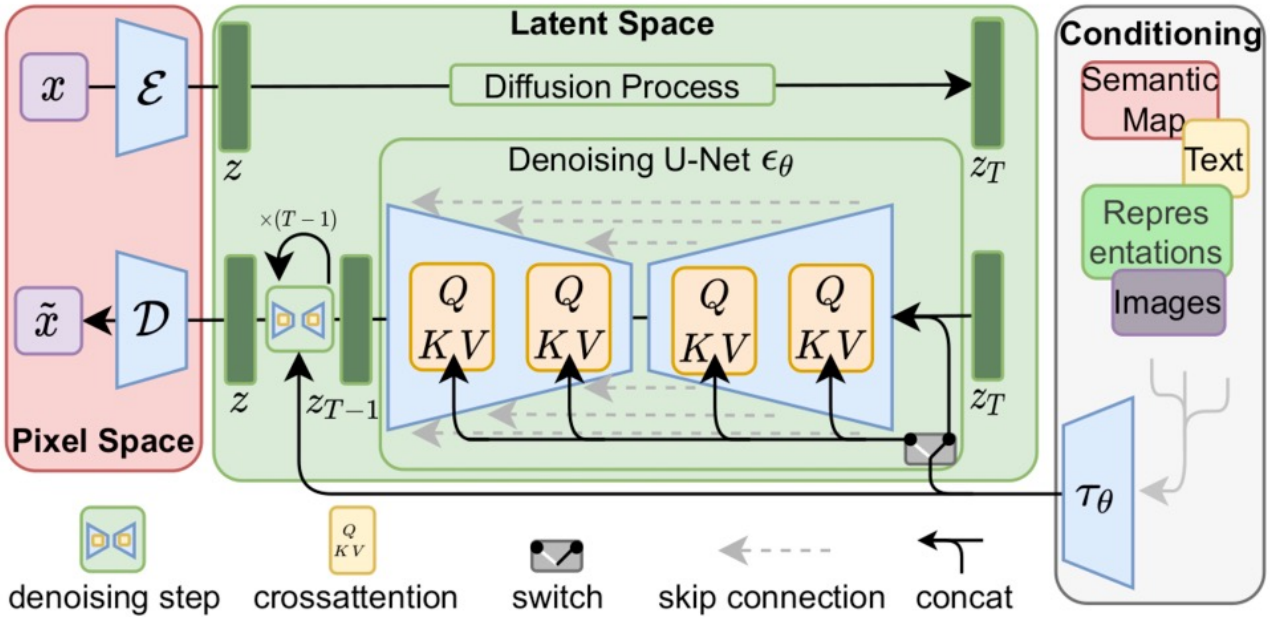
Conditional Textual GANs: Latent Diffusion Modeling (LDM)

Autoencoder with KL or VQ regularization.

VQ-reg.: $\mathcal{L}_{\text{total}} = \mathcal{L}_{rec} + \mathcal{L}_{VQ} + \lambda \mathcal{L}_{GAN}$

KL-reg.: $\mathcal{L}_{\text{total}} = \mathcal{L}_{rec} + \beta \mathcal{L}_{KL} + \lambda \mathcal{L}_{GAN}$

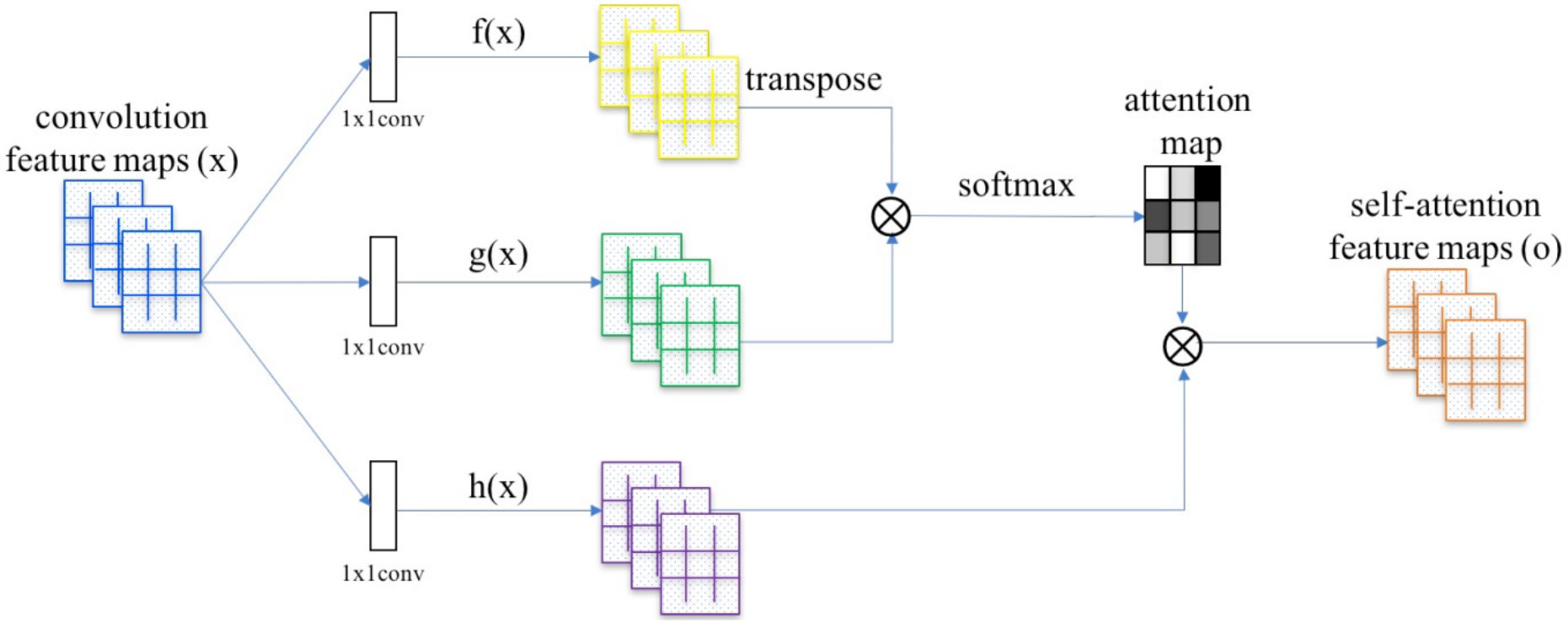
where $\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{rec}]}{\nabla_{G_L}[\mathcal{L}_{GAN}]+\delta}$



Slide credit: Robin Rombach

Conditional Textual GANs: Latent Diffusion Modeling (LDM)

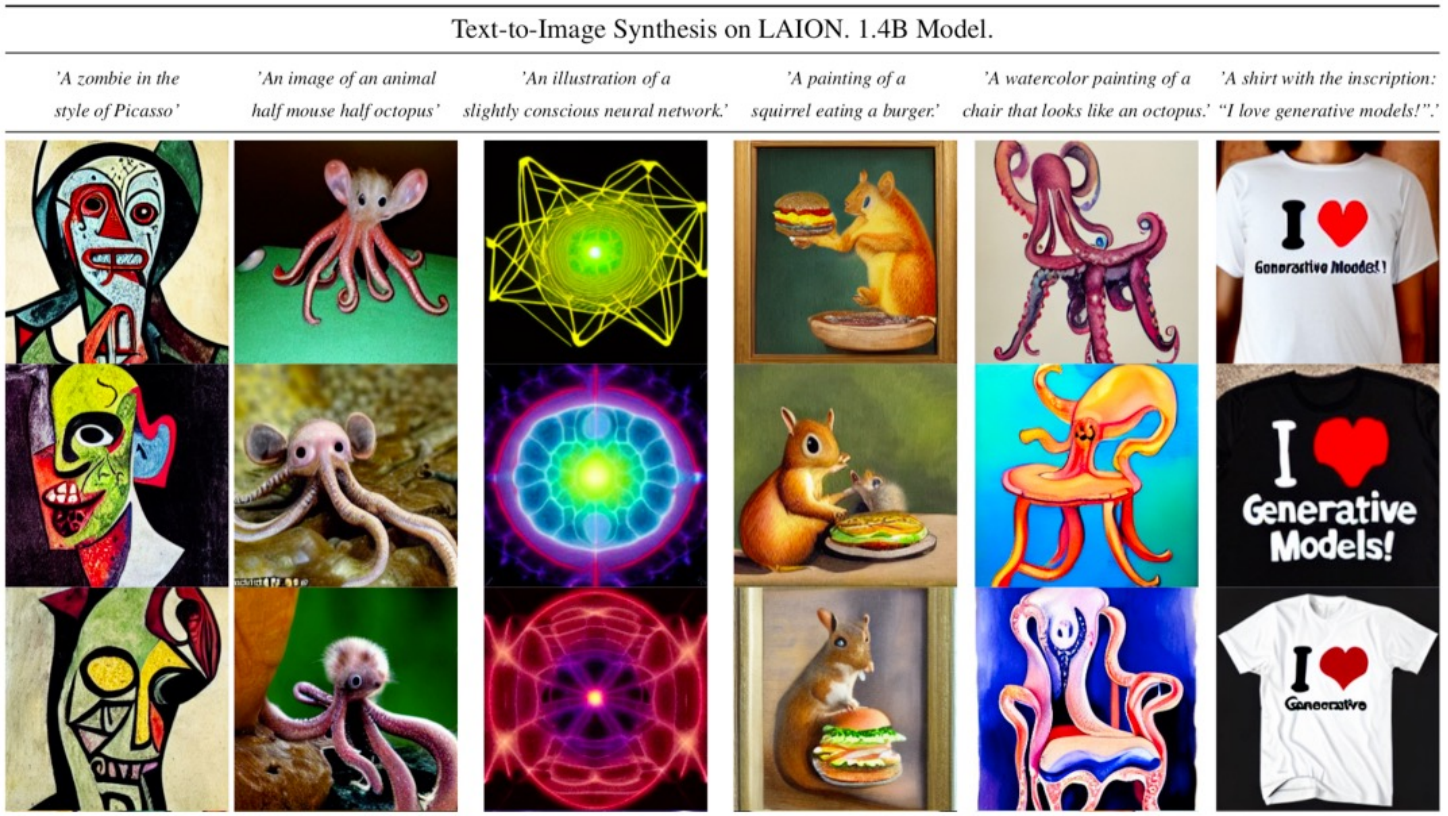
(Spatial) Self-attention Layer



Han Zhang et al, “Self-Attention Generative Adversarial Networks”, ICML 2018

Conditional Textual GANs: LDM results

- 32x32 cont. space
- 600M Transformer
- 800M UNet
- 400M Image/Text Pairs

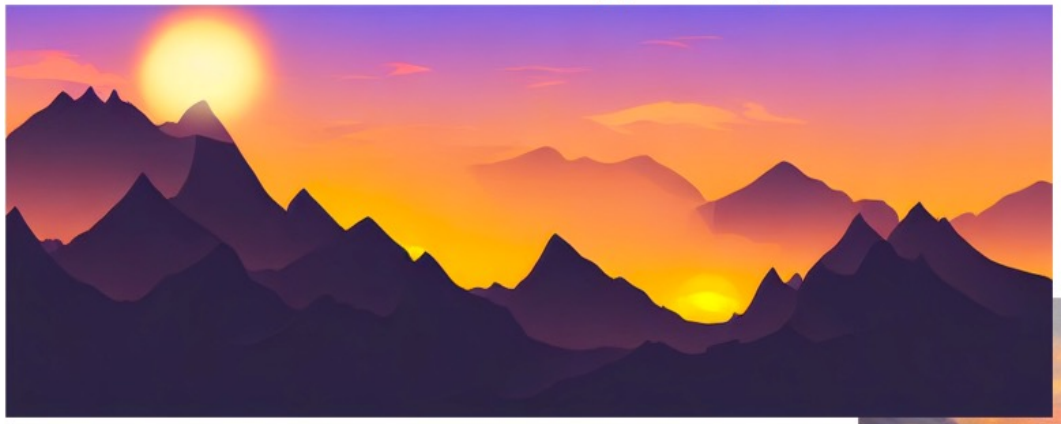


Slide credit: Robin Rombach

Conditional Textual GANs: LDM results

convolutional sampling (train on 256^2, generate on $>256^2$)

"A sunset over a mountain range, vector image"

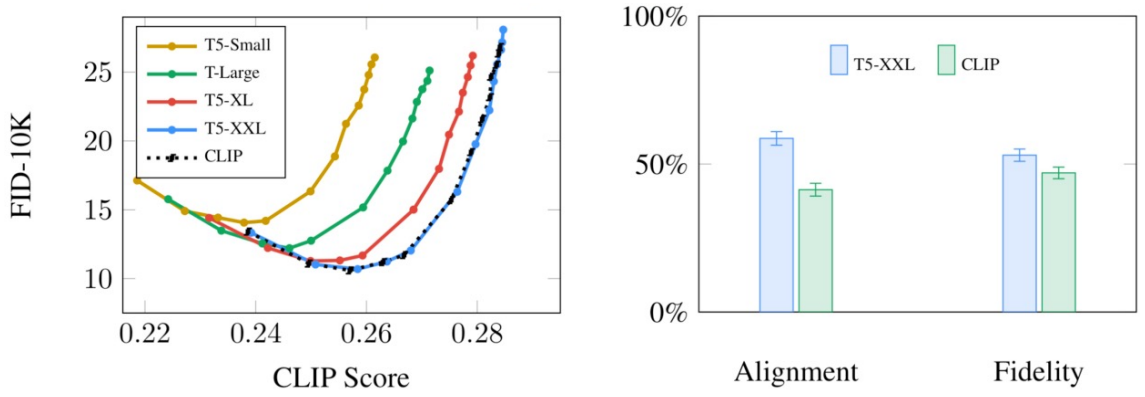


"A sunset over a mountain range, oil on canvas"

Slide credit: Robin Rombach

Conditional Textual GANs: Stable Diffusion

- Goal: achieve a small model that people can actually run locally on "small" GPUs (~10GB VRAM)
- Progressive training: pretrain on 256x256, then continue on 512x512
- Fix text encoder (as in Imagen)
- → choose CLIP (ViT-L/14) since performance/size tradeoff seems significant



(a) Pareto curves comparing various text encoders. (b) Comparing T5-XXL and CLIP on DrawBench.

Figure from Imagen, https://arxiv.org/abs/2205.11487

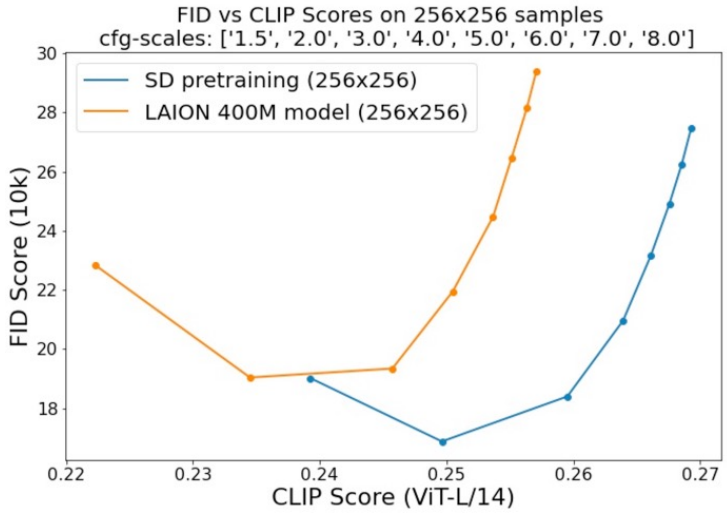
Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Stable Diffusion

Stage 1: Pretraining @256x256

- 237k steps at resolution 256x256 on LAION 2B(en)
- batch-size = 2048
- ~ 64 A100 GPUs

* FID score is a metric used to evaluate the quality of images generated by generative models, with lower scores indicating a better match between generated and real images.



10k random COCO val captions / 50 decoding steps

FID Score: https://en.wikipedia.org/wiki/Fr%C3%A9chet_inception_distance

Figure from Imagen, https://arxiv.org/abs/2205.11487

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Stable Diffusion

Stage 2: Training @512x512. batch-size=2048, #gpus=256

[Part 1 (v1.1)]

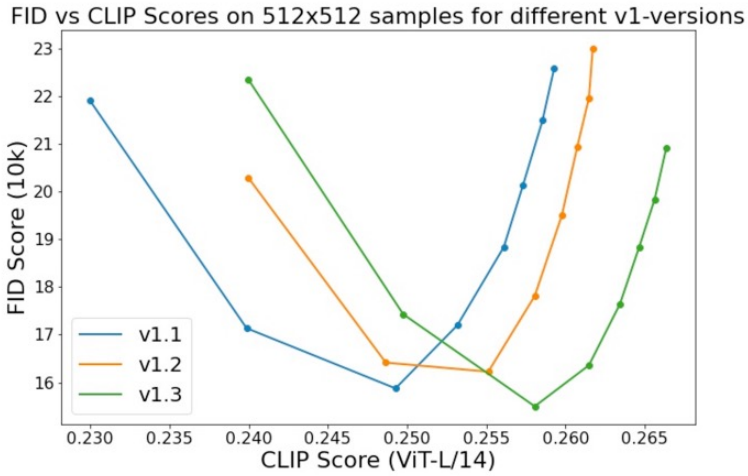
- 194k steps at resolution 512x512 on laion-high-resolution (170M examples from LAION-5B with resolution >= 1024x1024).

[Part 2 (v1.2)]

- 515k steps at resolution 512x512 on "laion-improved-aesthetics" (a subset of laion2B-en, filtered to images with an original size >= 512x512, estimated aesthetics score > 5.0, and an estimated watermark probability < 0.5

[Part 3/4 (v1.3/v1.4)]

- 195k/225k steps at resolution 512x512 on "laion-improved aesthetics" and 10% dropping of the text-conditioning



10k random COCO val captions / 50 decoding steps

FID Score: https://en.wikipedia.org/wiki/Fr%C3%A9chet_inception_distance

Figure from Imagen, https://arxiv.org/abs/2205.11487

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Stable Diffusion

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

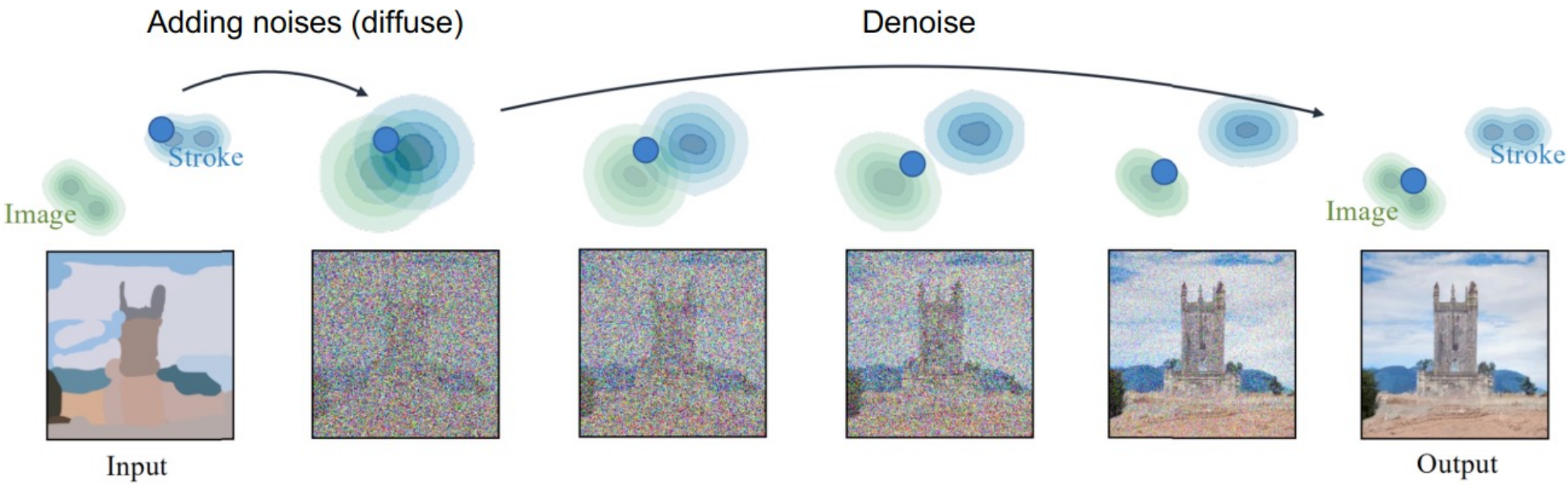
Conditional Textual GANs: Stable Diffusion

Stable Diffusion Demo: link

Slide credit: Robin Rombach

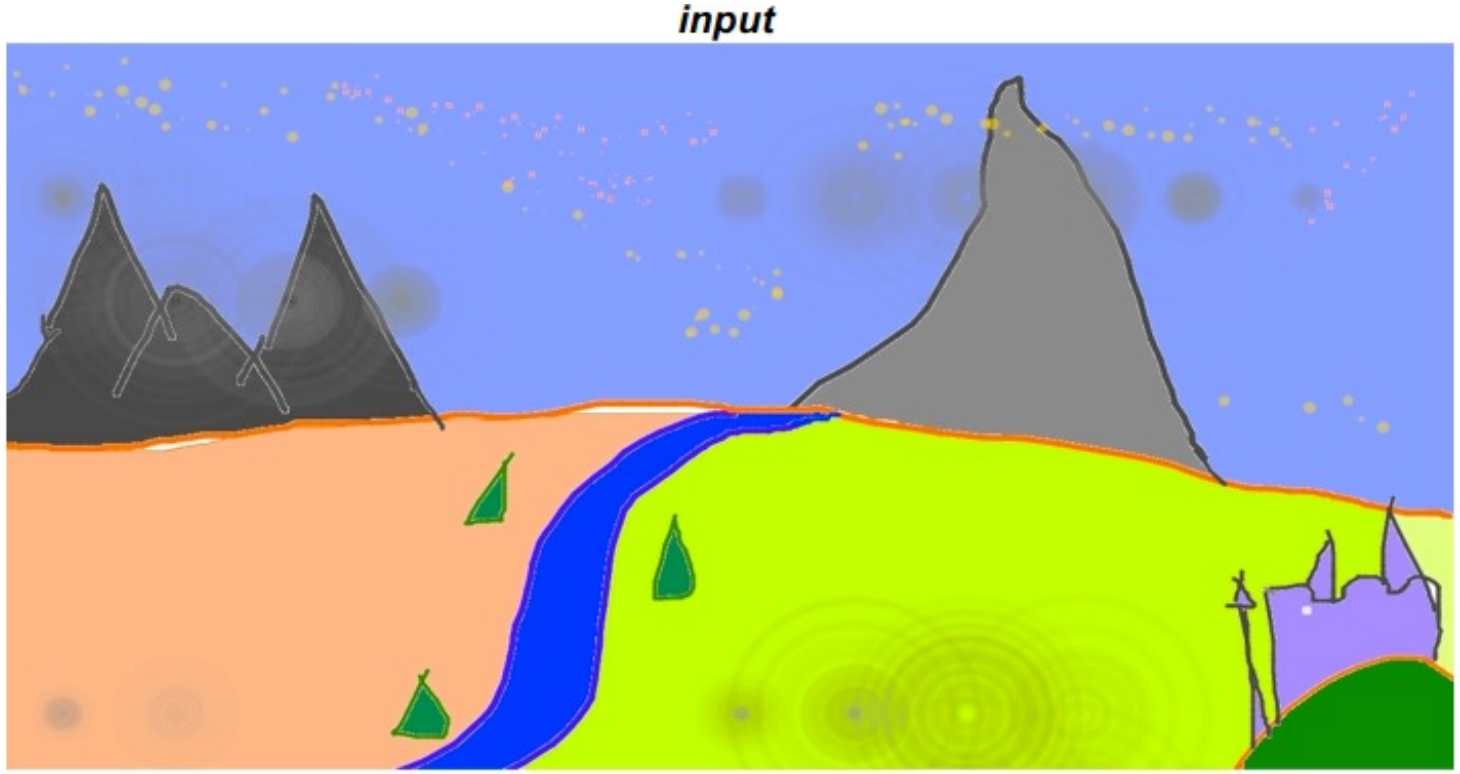
Conditional Textual GANs: Stroke-guided Image-to-image

SDEdit (https://arxiv.org/abs/2108.01073) recipe: diffuse → denoise



Slide credit: Robin Rombach

Conditional Textual GANs: Text-guided Image-to-image



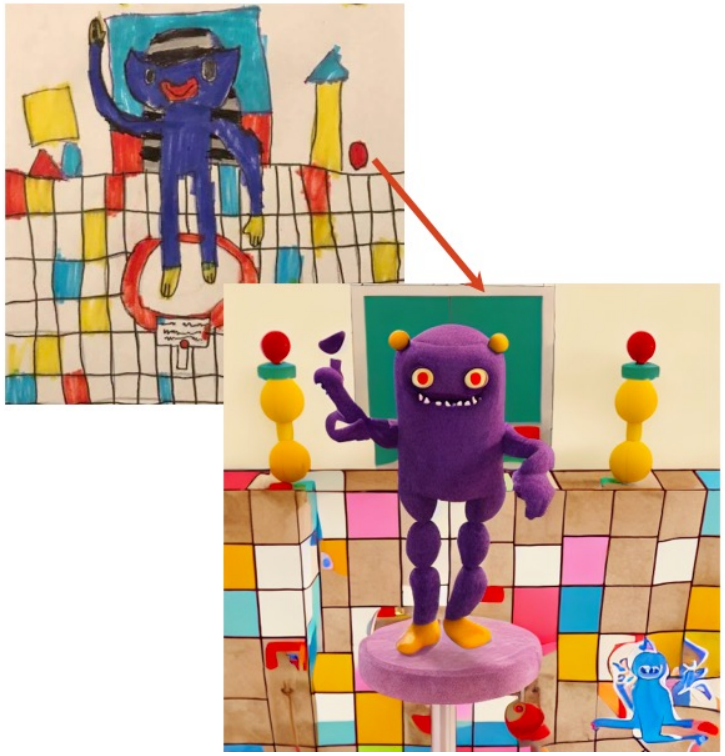
Slide credit: Robin Rombach

Conditional Textual GANs: Text-guided Image-to-image

Slide credit: Robin Rombach

Conditional Textual GANs: Text-guided Image-to-image

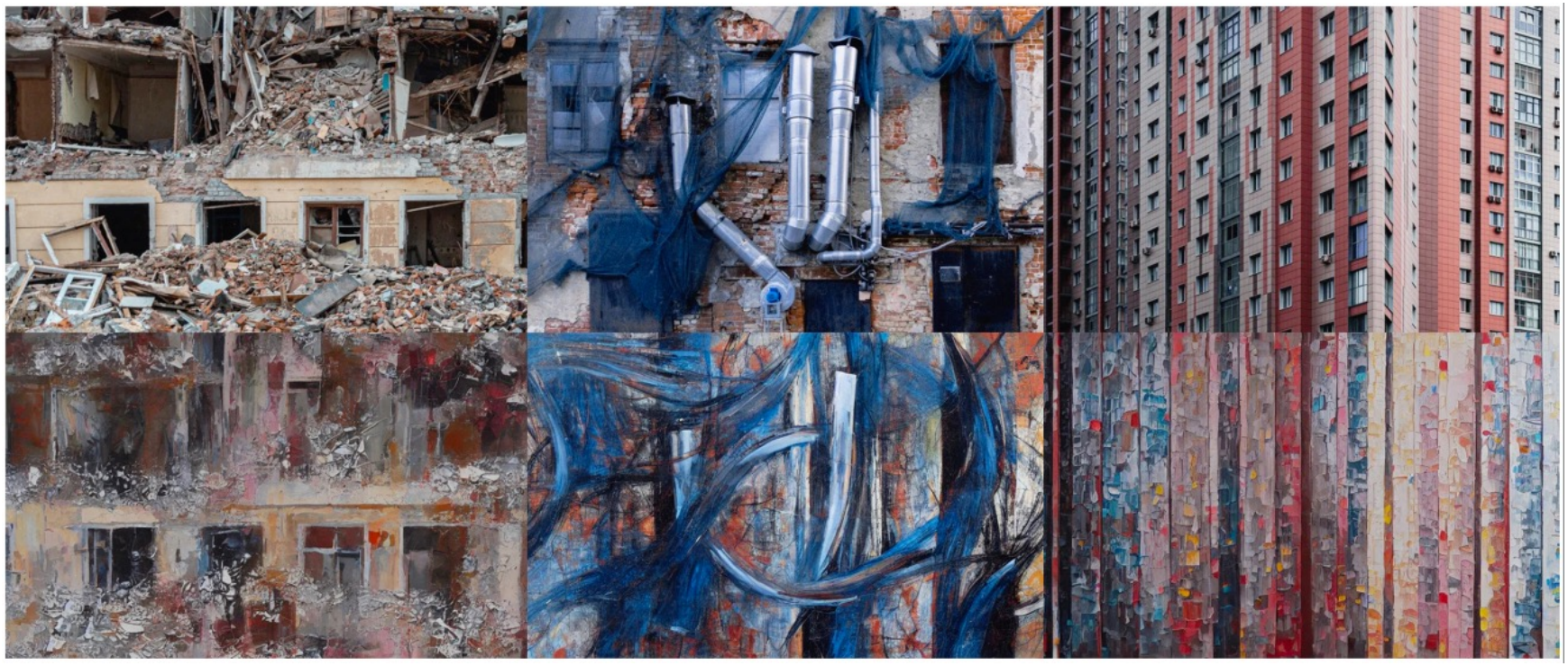
“Upgrade” your child’s artwork
original post: https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/



Slide credit: Robin Rombach

Conditional Textual GANs: Text-guided Image-to-image

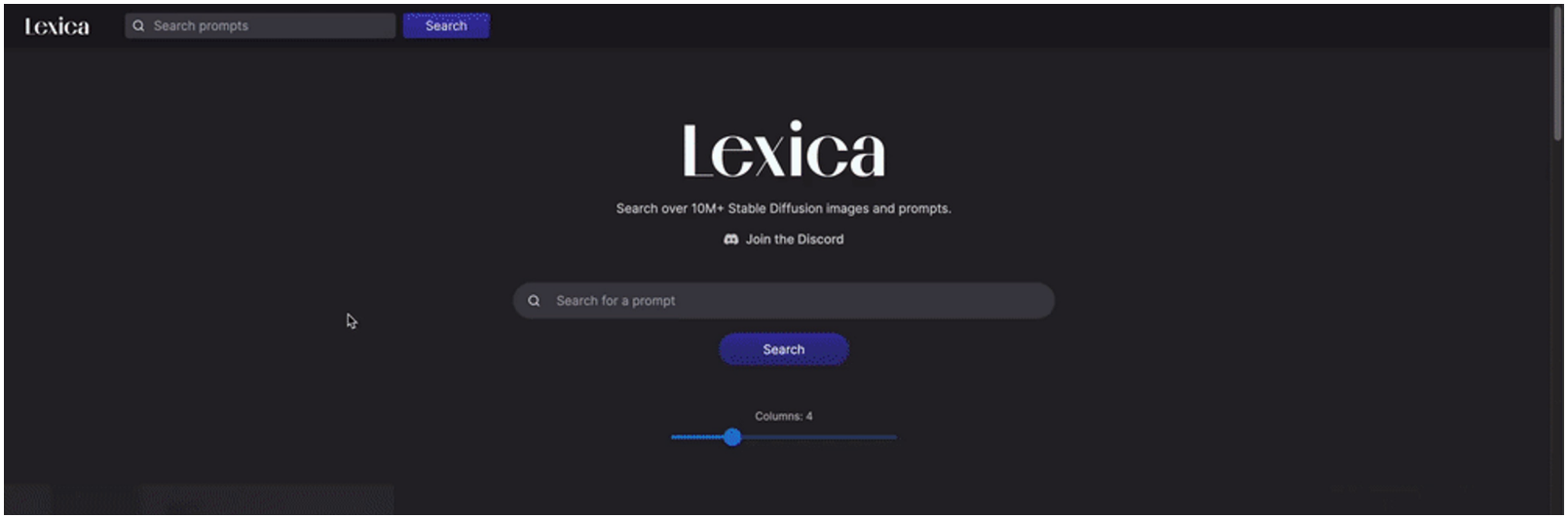
original post by u/Pereulkov:
https://www.reddit.com/r/StableDiffusion/comments/xhhyad/i_made_abstract_art_from_my_photos/



Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs: Prompting

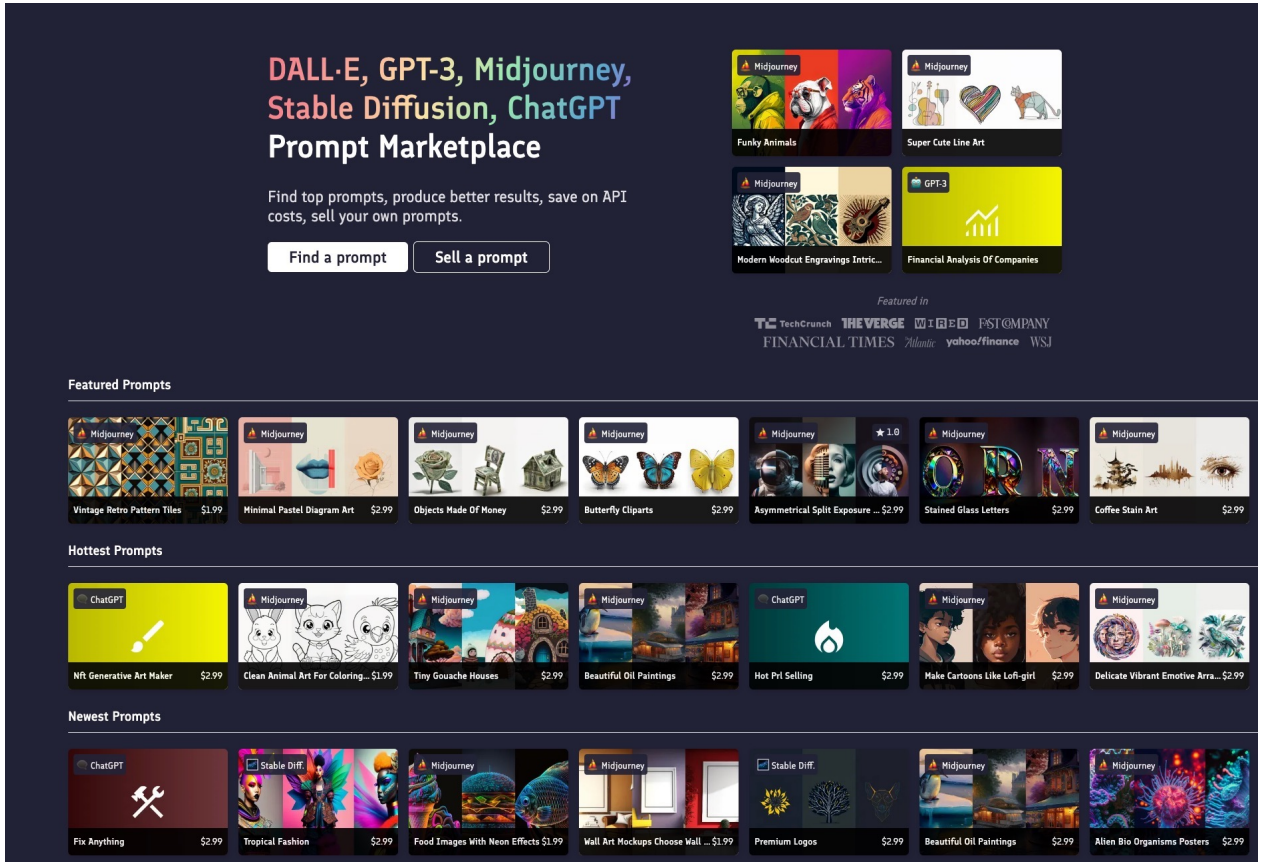
Prompt Search Engine (lexica.art)



Slide credit: Robin Rombach

Conditional Textual GANs: Prompting

Prompt Marketplace
(promptbase.com)



Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional Textual GANs

What if you have 1,000+ GPUs/TPUs

DALL-E (v1)

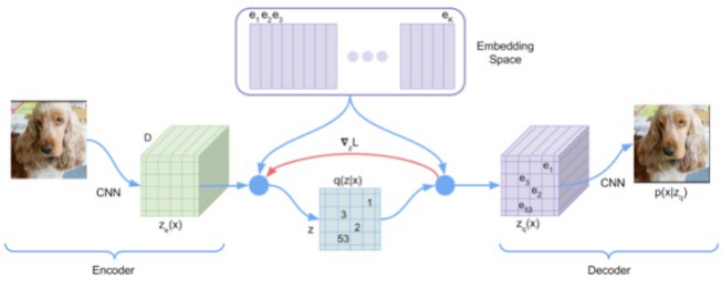
Zero-Shot Text-to-Image Generation

OpenAI Feb 2021

Aditya Ramesh [1] Mikhail Pavlov [1] Gabriel Goh [1] Scott Gray [1]
Chelsea Voss [1] Alec Radford [1] Mark Chen [1] Ilya Sutskever [1]

Step 1:

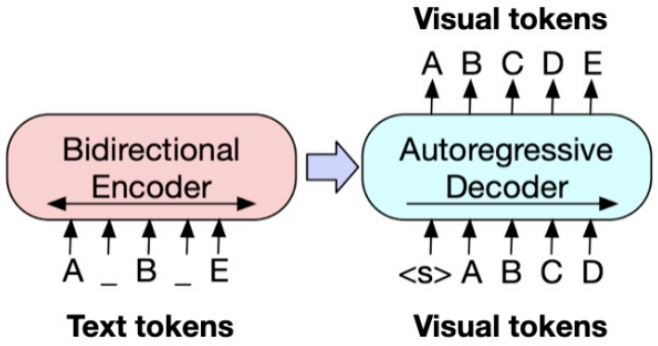
Learn Discrete Dictionary of Visual Tokens



VQVAE — Oord, Vinyals, Kavukcuoglu, 2017
VQGAN — Esser, Rombach, Ommer, 2021
dVAE - DALL-E — Ramesh et al 2021

Step 2:

Build a scene as a composition of discrete visual tokens



BART, GPT-3, etc

DALL-E (v1): Example

an armchair in the shape of an avocado. . . .

Conditional Textual GANs: DALL·E 2, Imagen

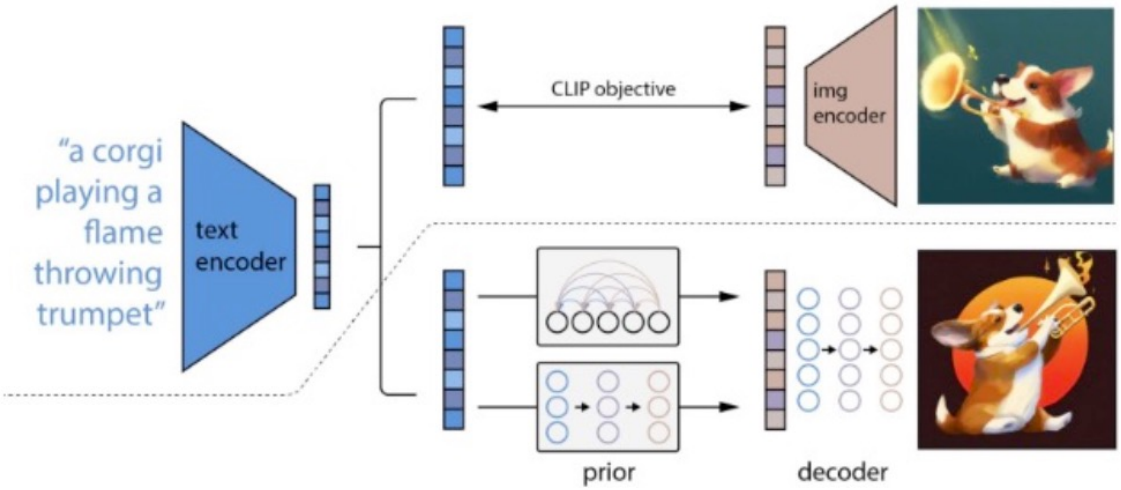
- Pixel-based Diffusion (No encoder-decoder)
- Pre-trained text encoder (CLIP, t5)
- Diffusion model + classifier-free guidance
- Cascaded models: 64->128->512

https://cdn.openai.com/papers/dall-e-2.pdf
https://arxiv.org/abs/2205.11487

DALL.E 2 | OpenAI

Conditioning on CLIP embeddings

- Helps capture multimodal representations
- The bi-partite latent enables several text-controlled image manipulation tasks



Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen - https://arxiv.org/abs/2204.06125

DALL.E 2 | OpenAI

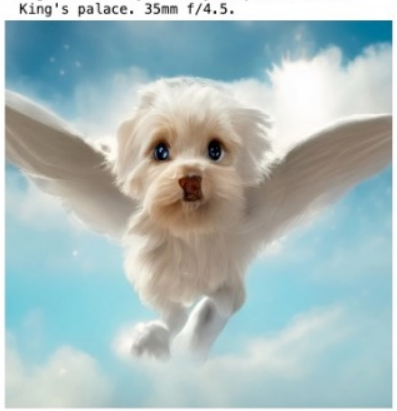
- 1k x 1k text-conditioned image generation
- Uses a **prior** to produce CLIP embeddings conditioned on the text-caption
- Uses a **decoder** to produce images conditioned on the CLIP embeddings

Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen - https://arxiv.org/abs/2204.06125

Conditional Textual GANs: GigaGAN – Scaling up GAN

A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.

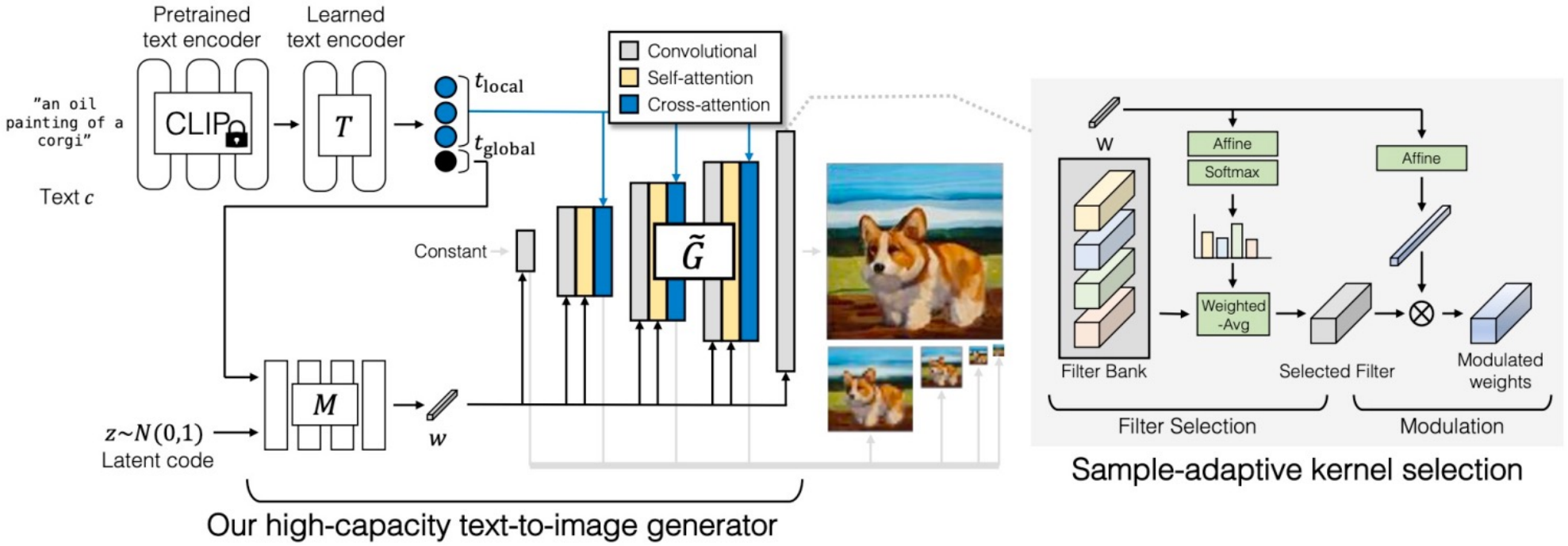
A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.



a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

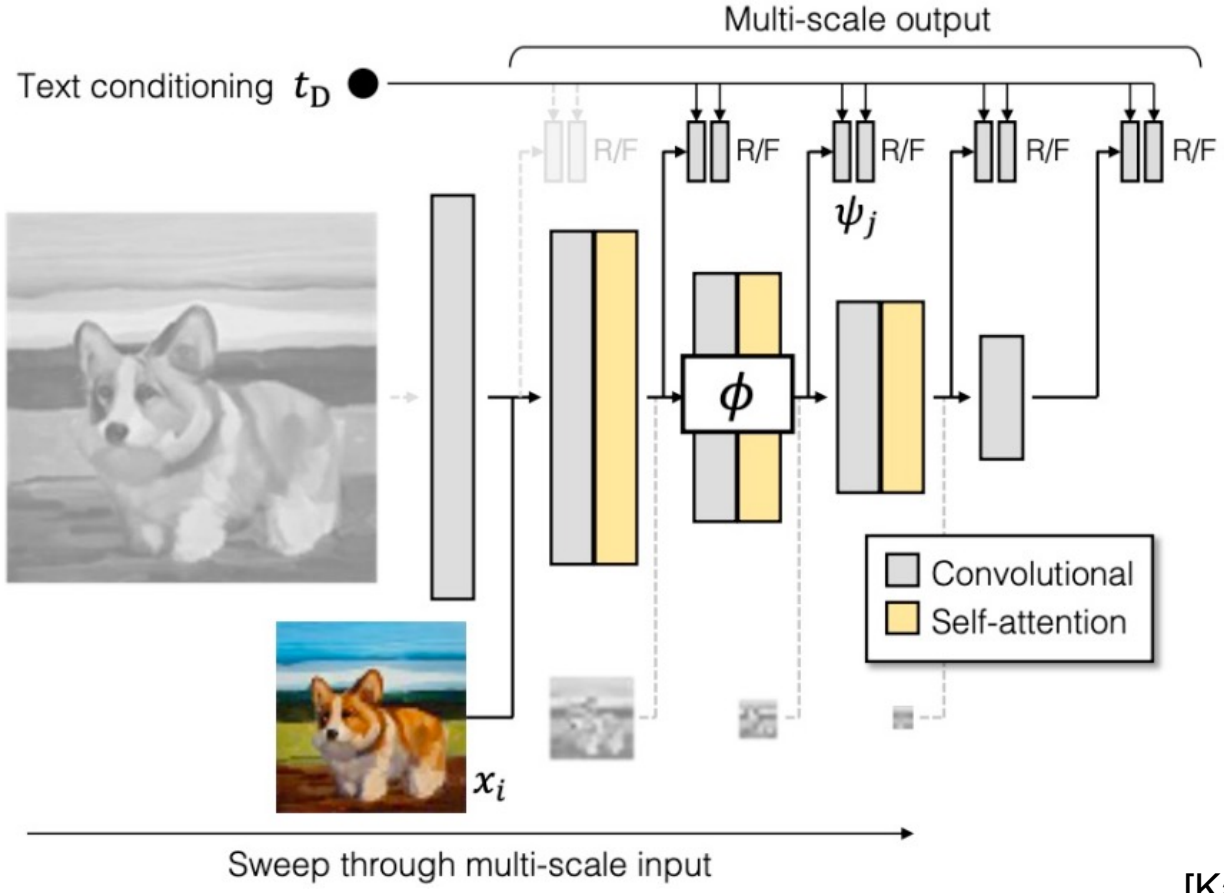
[Kang et al., CVPR 2023]

Conditional Textual GANs: GigaGAN Generator



[Kang et al., CVPR 2023]

Conditional Textual GANs: GigaGAN Discriminator

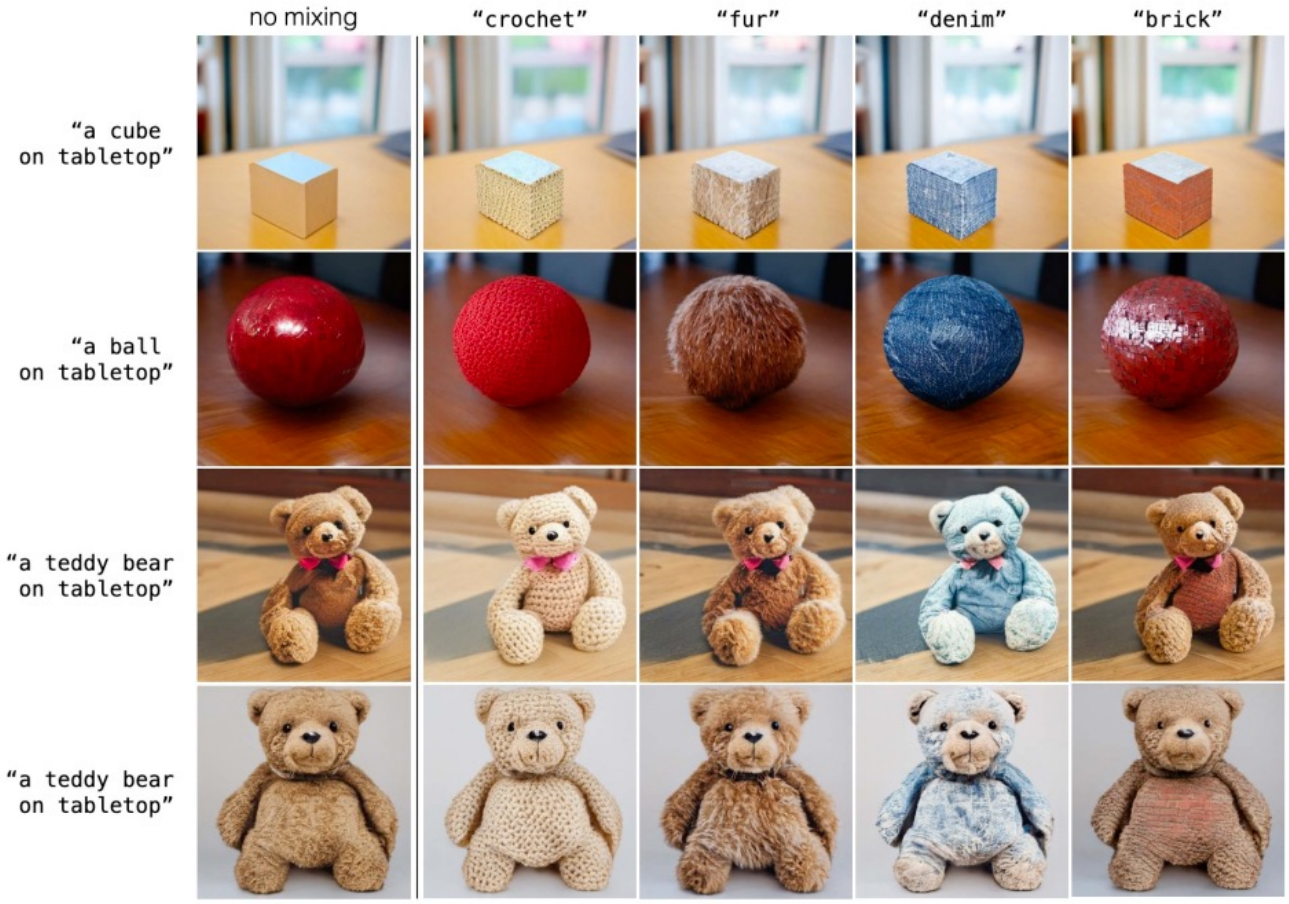


[Kang et al., CVPR 2023]

Conditional Textual GANs: GigaGAN Style Mixing

[Kang et al., CVPR 2023]

Conditional Textual GANs: GigaGAN Prompt Mixing



[Kang et al., CVPR 2023]

Conditional Textual GANs: GigaGAN Results

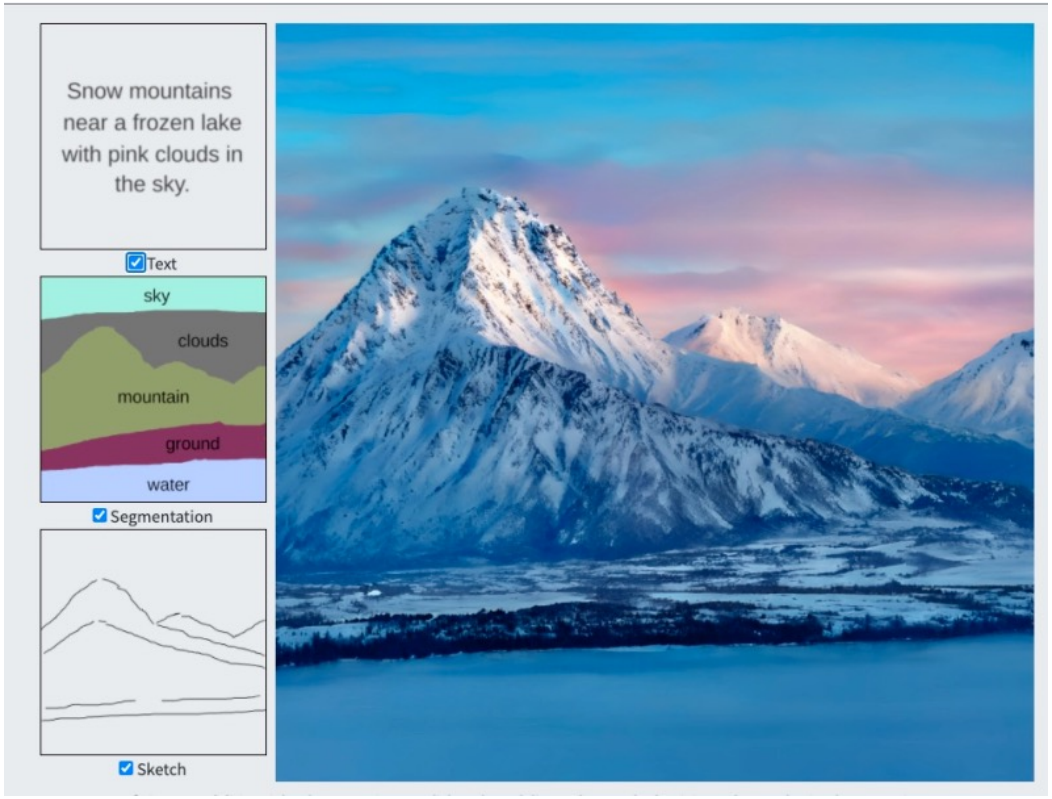
[Kang et al., CVPR 2023]

Slide credit: Jun-Yan Zhu - Learning-Based Image Synthesis

Conditional GANs: Applications [Example-Guided Translation]

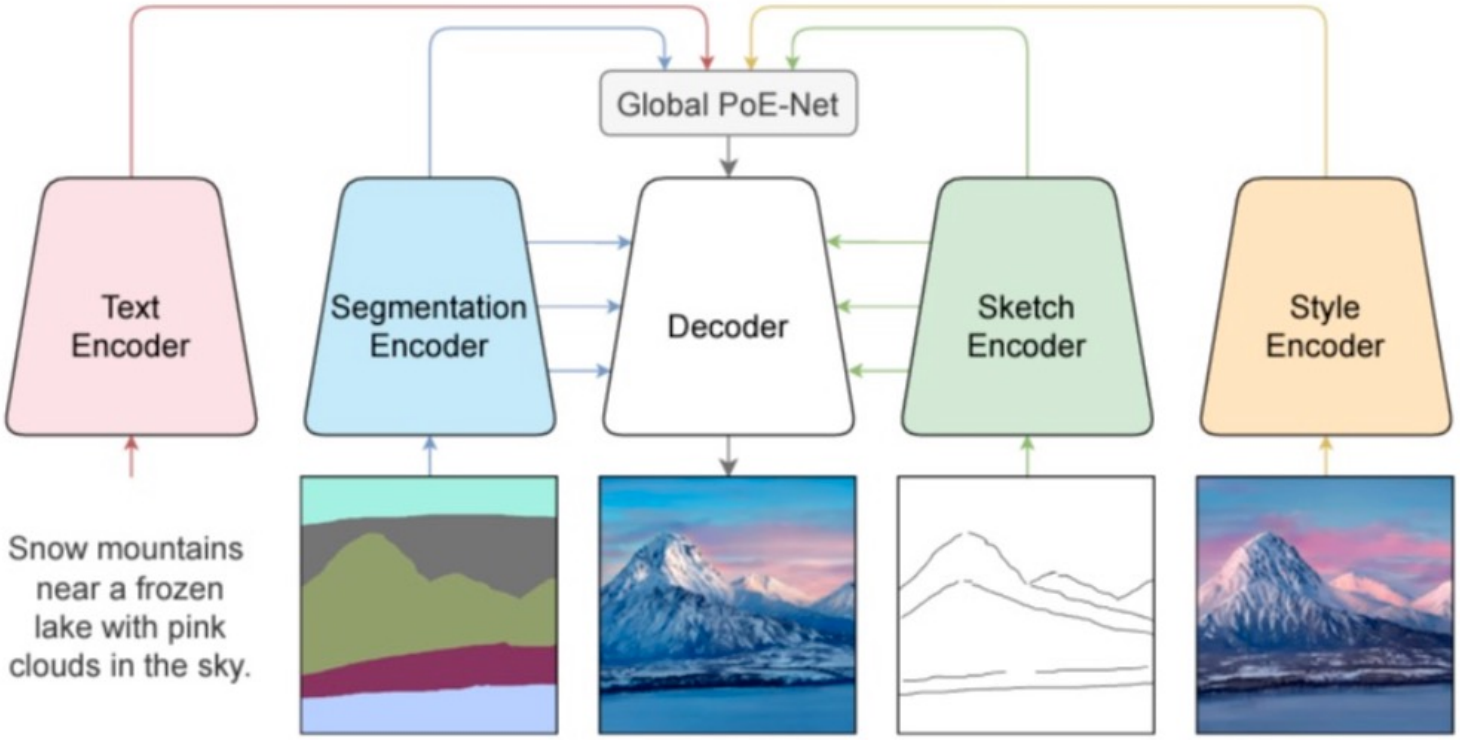
Multimodal Unsupervised Image-to-Image Translation - Xun Huang et al. ECCV 2018

Conditional GANs: Applications



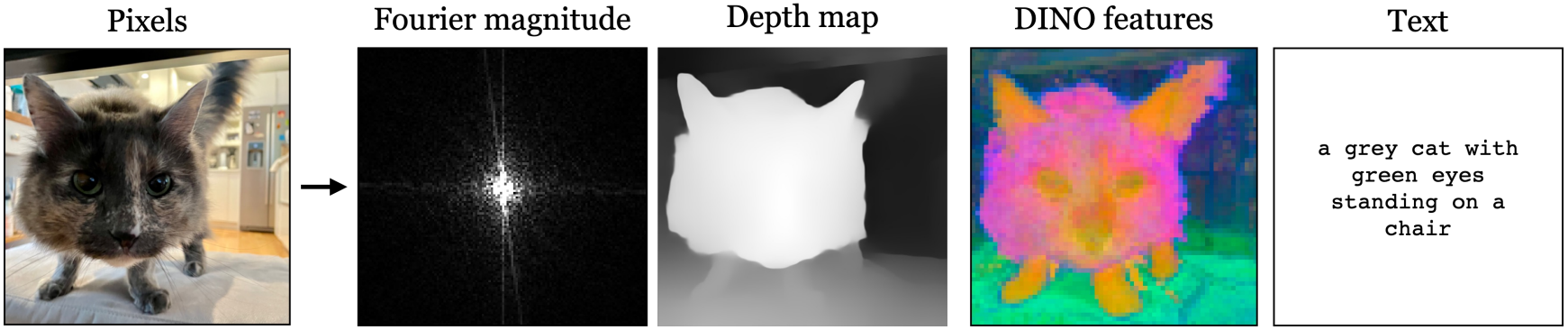
Multimodal Conditional Image Synthesis with Product-of-Experts GANs [Huang et al., 2021]

Conditional GANs: Applications



Multimodal Conditional Image Synthesis with Product-of-Experts GANs [Huang et al., 2021]

Conditional Textual GANs: Visual Representations



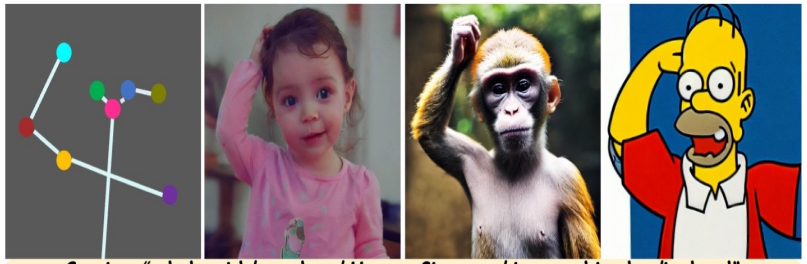
Different kinds of visual representations

GLIGEN: Open-Set Grounded Text-to-Image Generation

Caption: "A woman sitting in a restaurant with a pizza in front of her "
Grounded text: table, pizza, person, wall, car, paper, chair, window, bottle, cup

Caption: "Elon Musk and Emma Watson on a movie poster"
Grounded text: Elon Musk, Emma Watson; Grounded style image: blue inset

Caption: "A dog / bird / helmet / backpack is on the grass"
Grounded image: red inset



Caption: "a baby girl / monkey / Hormer Simpson / is scratching her/its head"
Grounded keypoints: plotted dots on the left image

https://huggingface.co/spaces/gligen/demo, https://gligen.github.io/

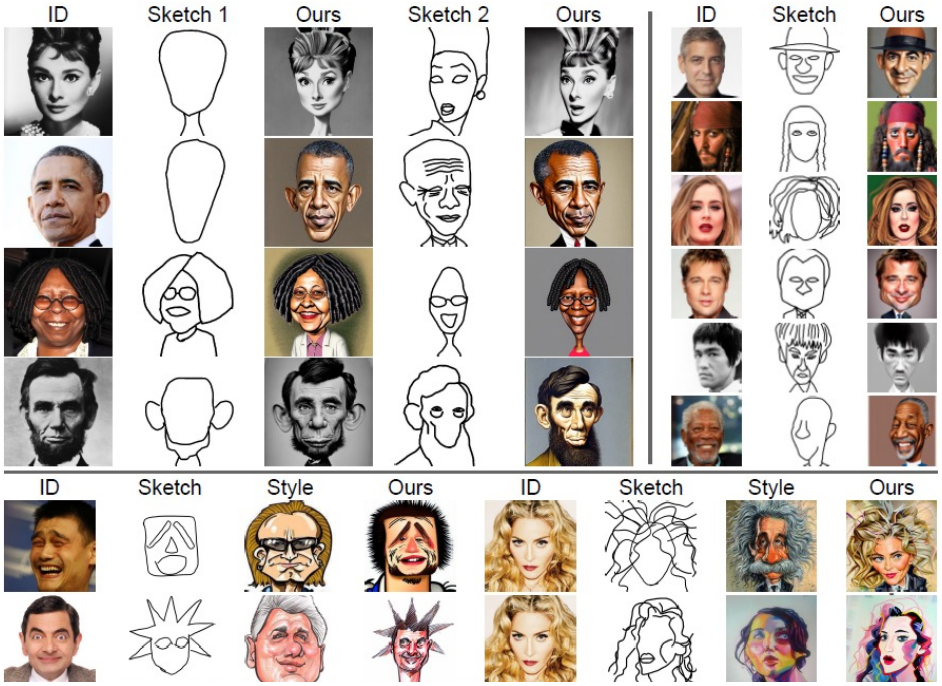
GLIGEN: Open-Set Grounded Text-to-Image Generation

Yong Jae Lee, https://huggingface.co/spaces/gligen/demo, https://gligen.github.io/

DemoCaricature: Democratising Caricature Generation with a Rough Sketch

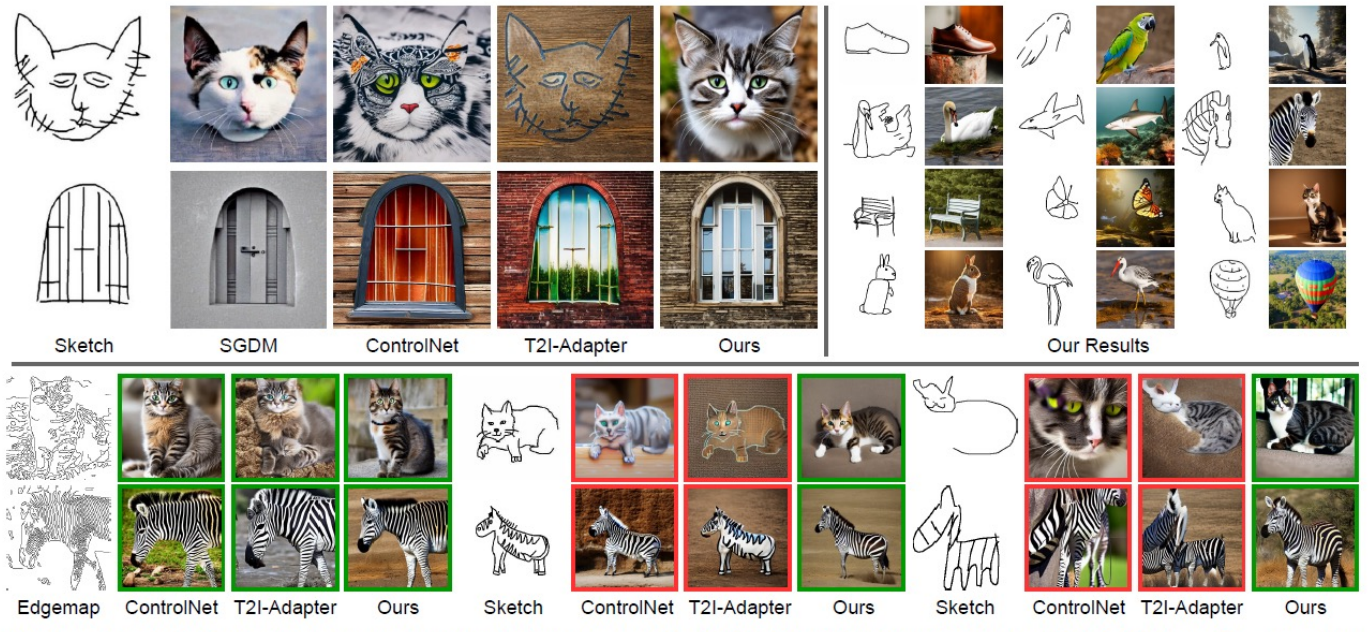
Dar-Yen Chen Subhadeep Koley Aneeshan Sain Pinaki Nath Chowdhury
Tao Xiang Ayan Kumar Bhunia Yi-Zhe Song
SketchX, CVSSP, University of Surrey, United Kingdom.
{s.koley, a.sain, p.chowdhury, t.xiang, a.bhunia, y.song}@surrey.ac.uk

https://democaricature.github.io



It's All About Your Sketch: Democratising Sketch Control in Diffusion Models

Subhadeep Koley[1,2] Ayan Kumar Bhunia[1] Deeptanshu Sekhri[1] Aneeshan Sain[1,2]
Pinaki Nath Chowdhury[1,2] Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{s.koley, a.bhunia, d.sekhri, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk



RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models

Ozgur Kara[1*] Bariscan Kurtkaya[2*†] Hidir Yesiltepe[4] James M. Rehg[1,3] Pinar Yanardag[4]

[1]Georgia Tech [2]KUIS AI Center [3]UIUC [4]Virginia Tech

okara7@gatech.edu, bkurtkaya23@ku.edu.tr, hidir@vt.edu, jrehg@uiuc.edu, pinary@vt.edu

Project Webpage: https://rave-video.github.io