Text-to-Image Generation

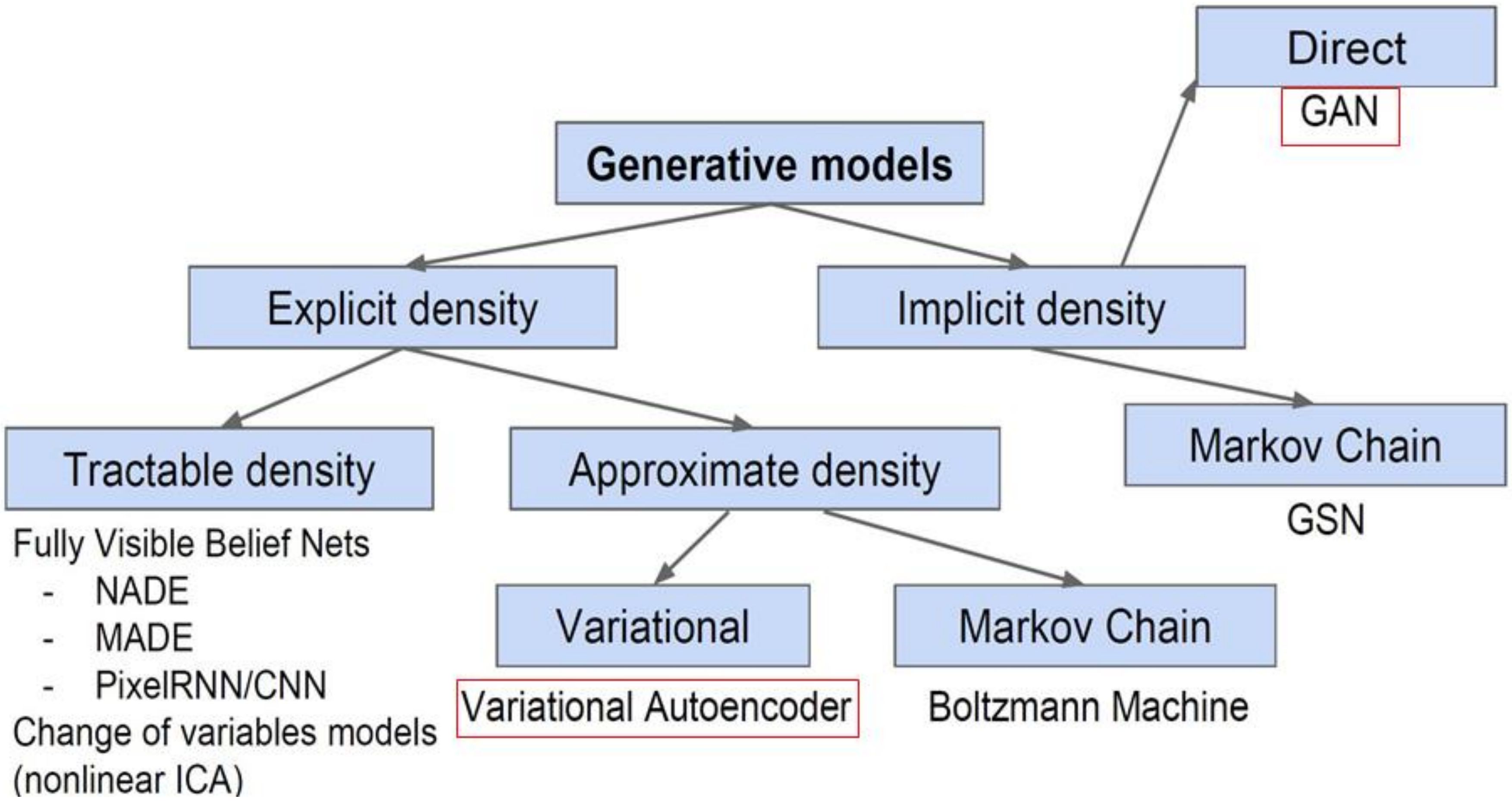
Yu Cheng

Microsoft

Text-to-Image Synthesis

- Text-to-Image Synthesis
  - StackGAN, AttnGAN, TAGAN, ObjGAN
- Text-to-Video Synthesis
  - GAN-based methods, VAE-based methods, StoryGAN
- Dialogue-based Image Synthesis
  - ChatPainter, CoDraw, SeqAttnGAN

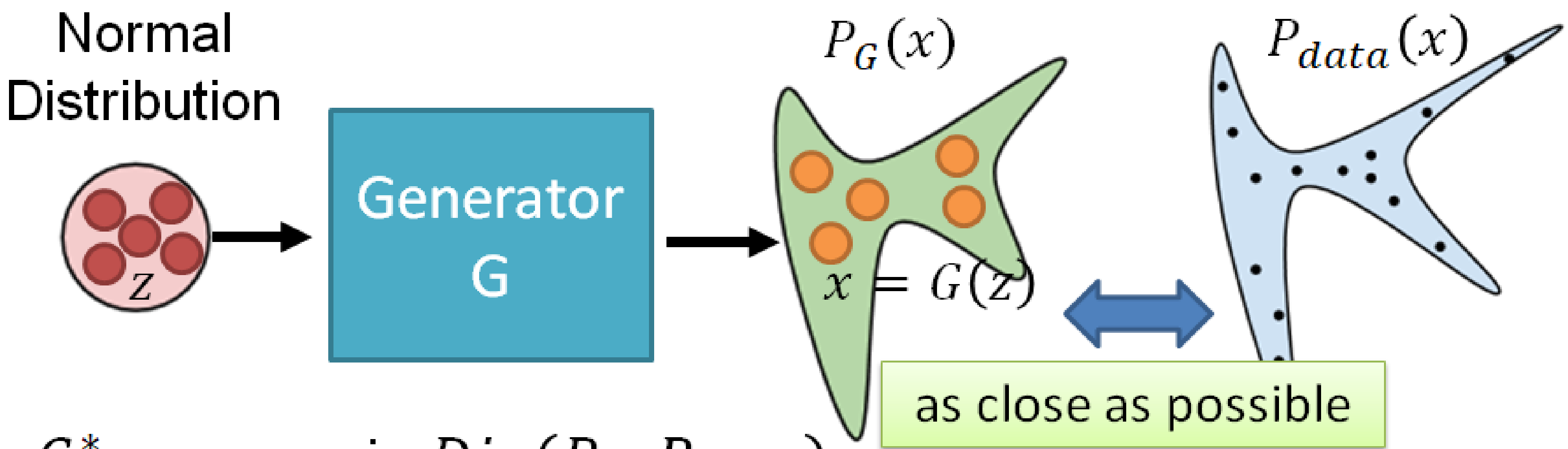
Generative Models



*Slides from Ian Goodfellow's tutorial

Generative Adversarial Networks (GAN)

- A generator G is a network. The network defines a probability distribution P_G



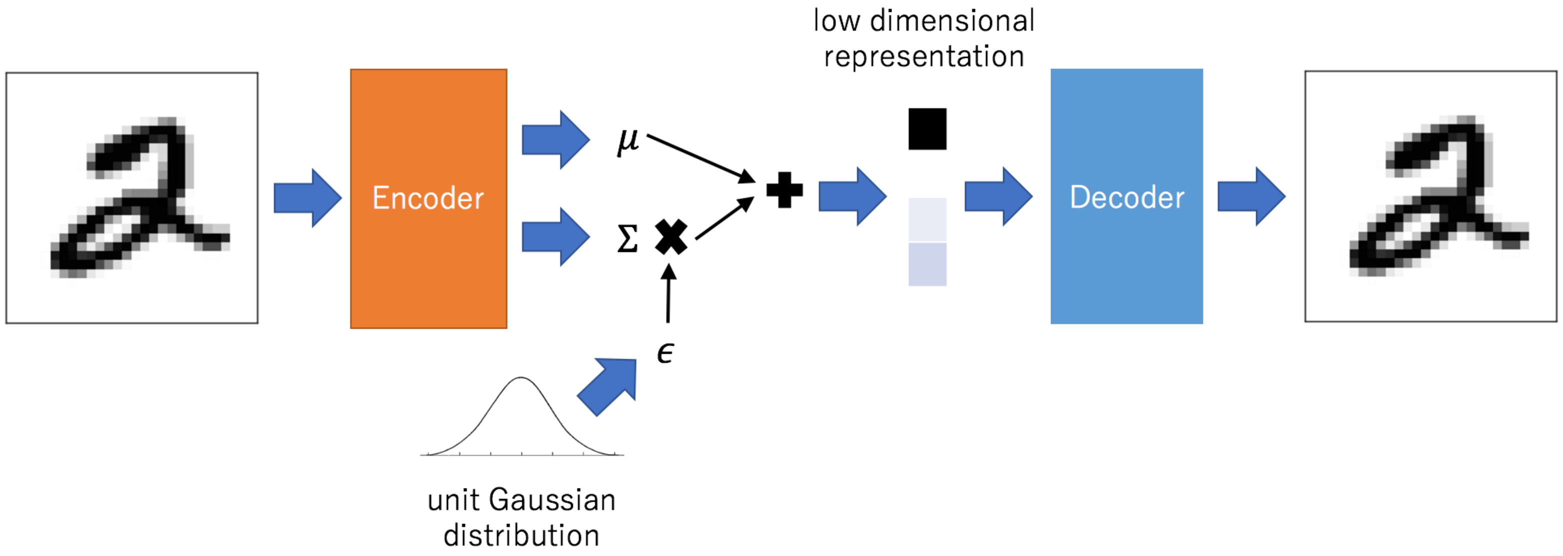
$$G^* = arg \min_G \underline{Div(P_G, P_{data})}$$

Divergence between distributions P_G and P_{data}

Goodfellow et al., 2014. Generative Adversarial Networks

Variational Autoencoder (VAE)

- VAE is an autoencoder whose encodings distribution is regularised during the training in order to ensure that its latent space has good properties allowing us to generate new data



Kingma and Welling, 2014. Auto-Encoding Variational Bayes

Two Paradigms for Generative Modeling

GAN

StyleGAN

[Karras, et al., 2019]

VAE

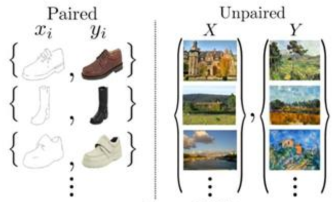
VQ-VAE-2

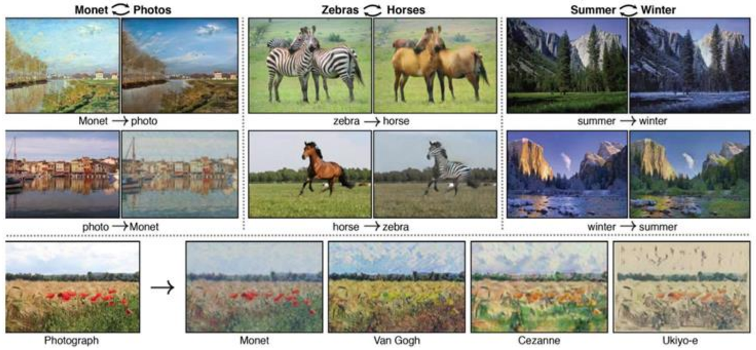
[Razavi, et al., 2019]

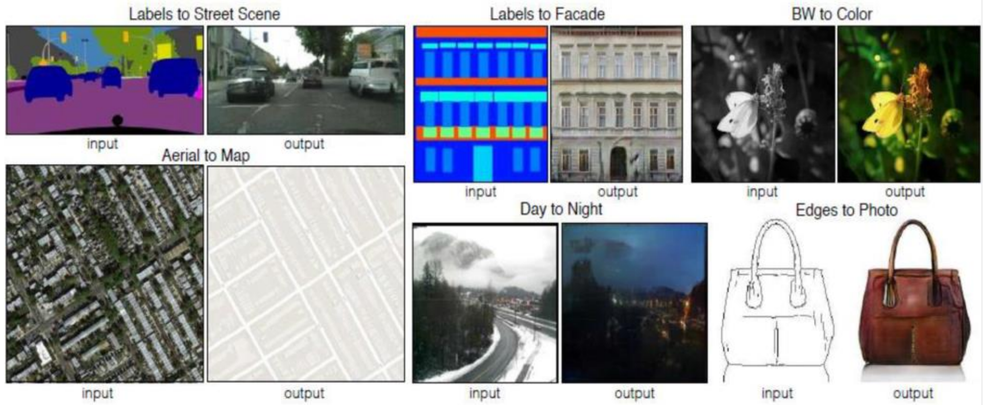
Conditional Image Synthesis

Cycle GAN

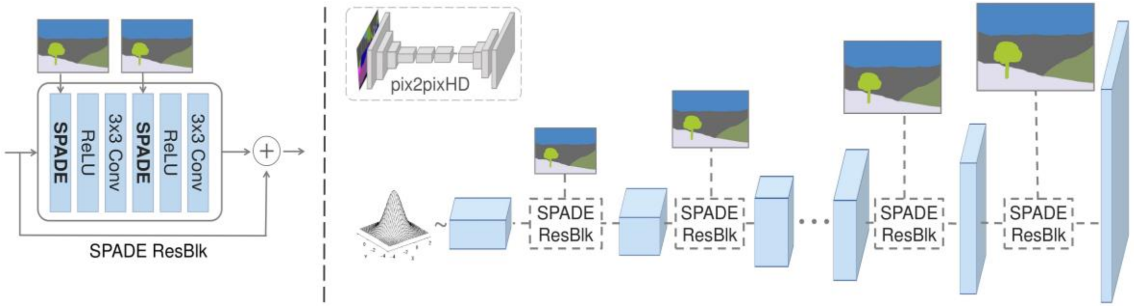
https://arxiv.org/abs/1703.10593





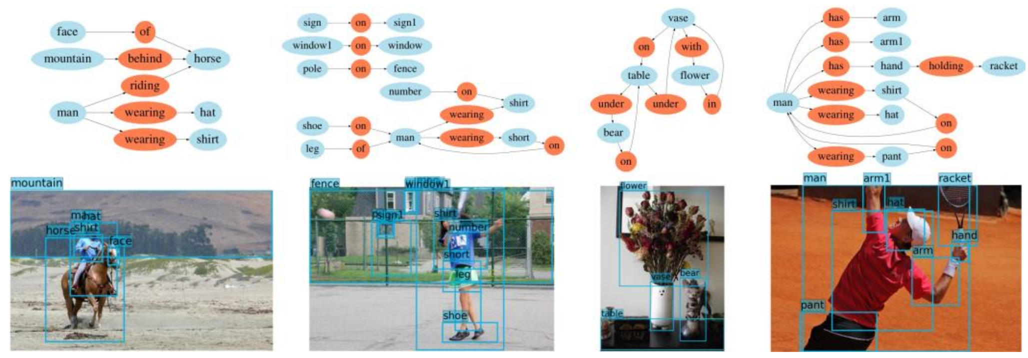


Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", arXiv preprint, 2016

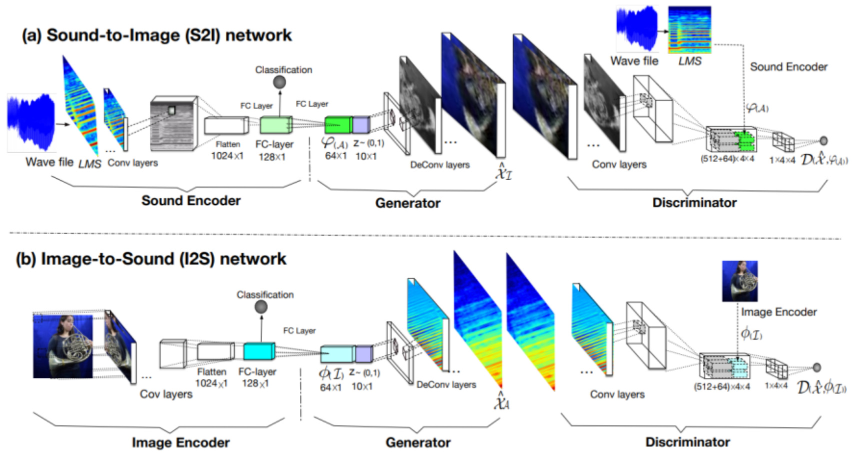


SPADE [Park et al., 2019]

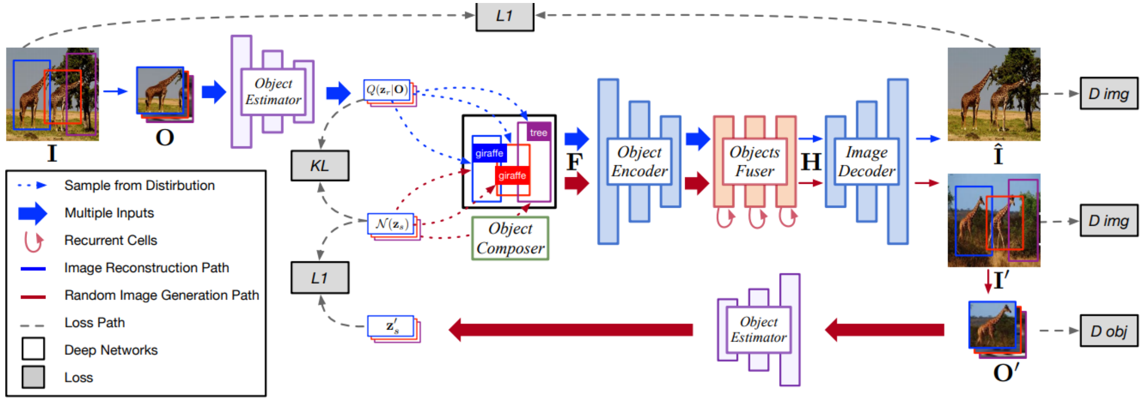
Conditional Image Synthesis



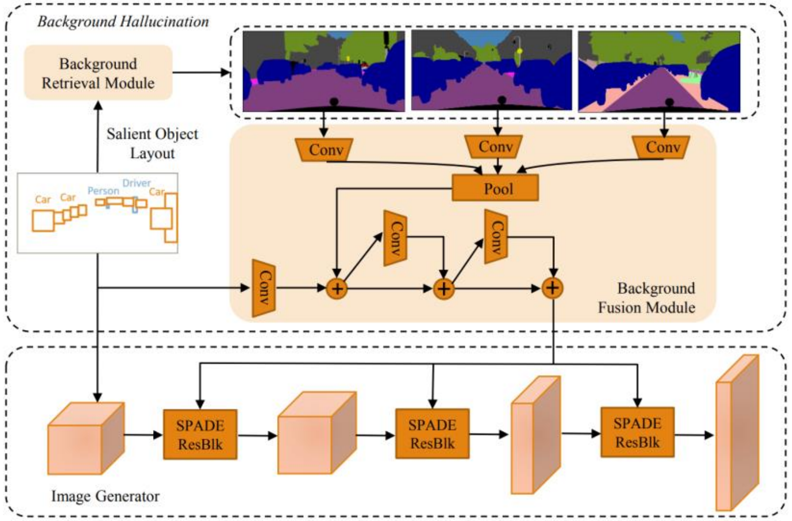
SceneGraph2img [Johnson et al., 2018]



Audio2img [Chen et al., 2019]

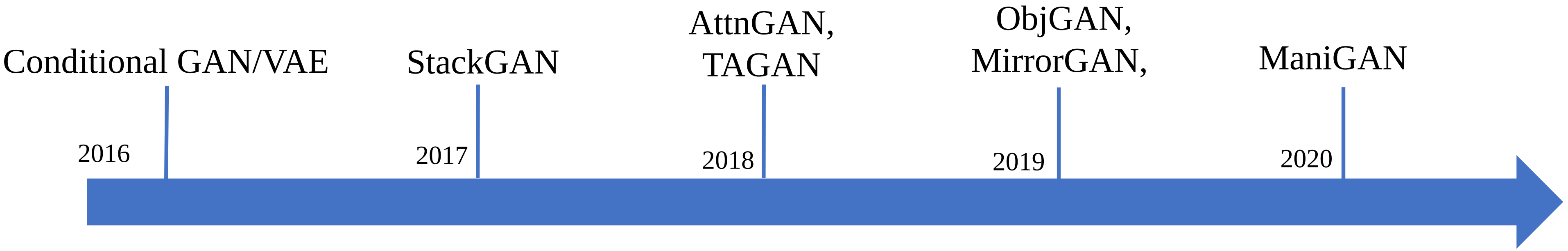


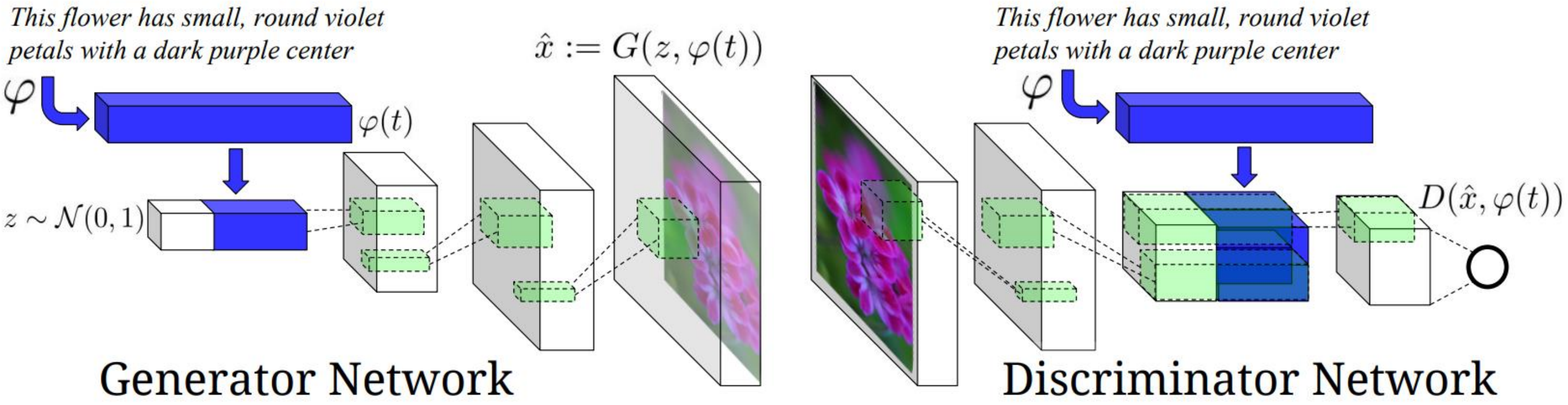
Layout2img [Zhao et al., 2019]



BachGAN [Li et al., 2020]

Text-to-Image Synthesis





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

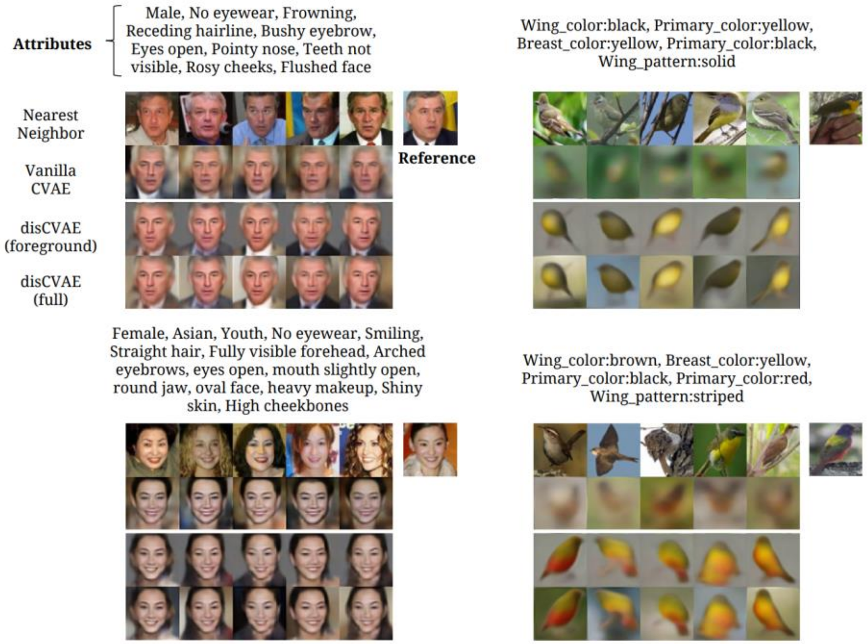
Text-to-Image Synthesis

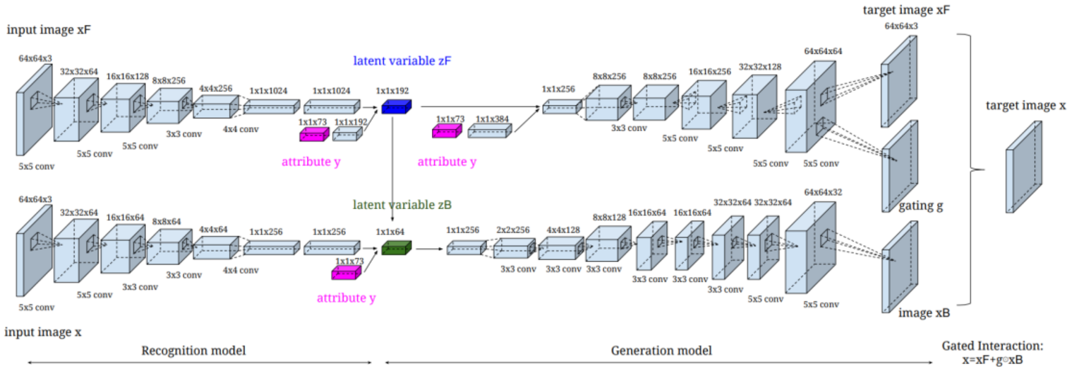
"red flower with black center"

Caption	Image
this flower has white petals and a yellow stamen	
the center is yellow surrounded by wavy dark purple petals	
this flower has lots of small round pink petals	

Text-to-Image Synthesis

- Text(attribute) to image generation with Conditional VAE



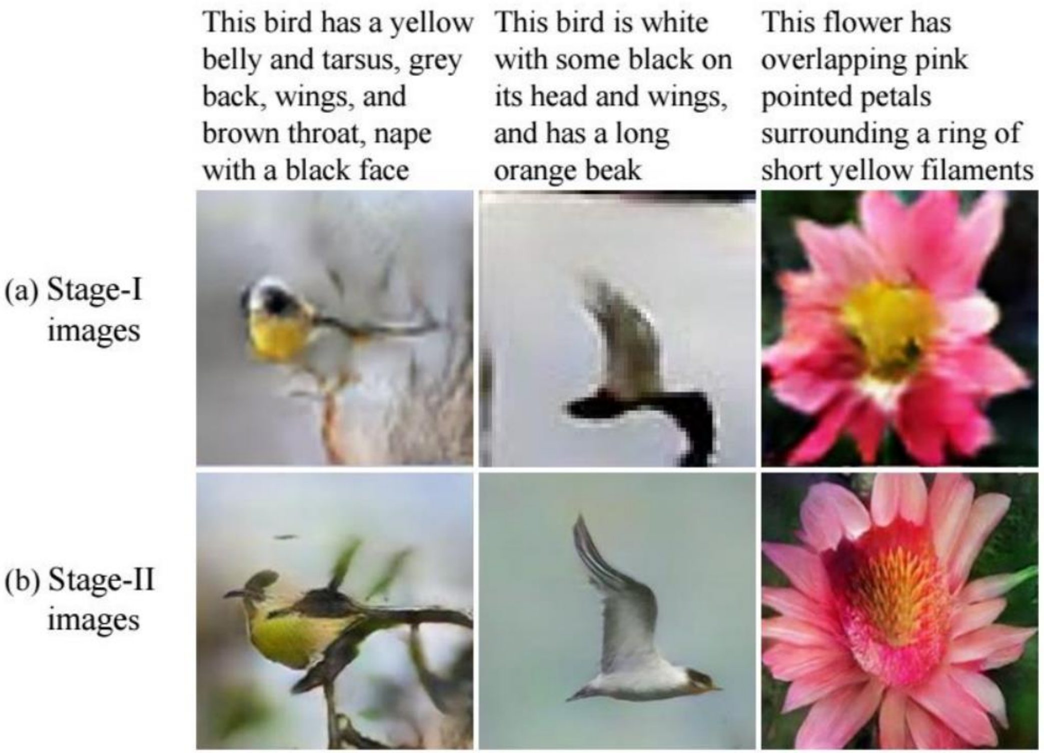


Yan et al, 2016. Attribute2Image: Conditional Image Generation from Visual Attributes

StackGAN

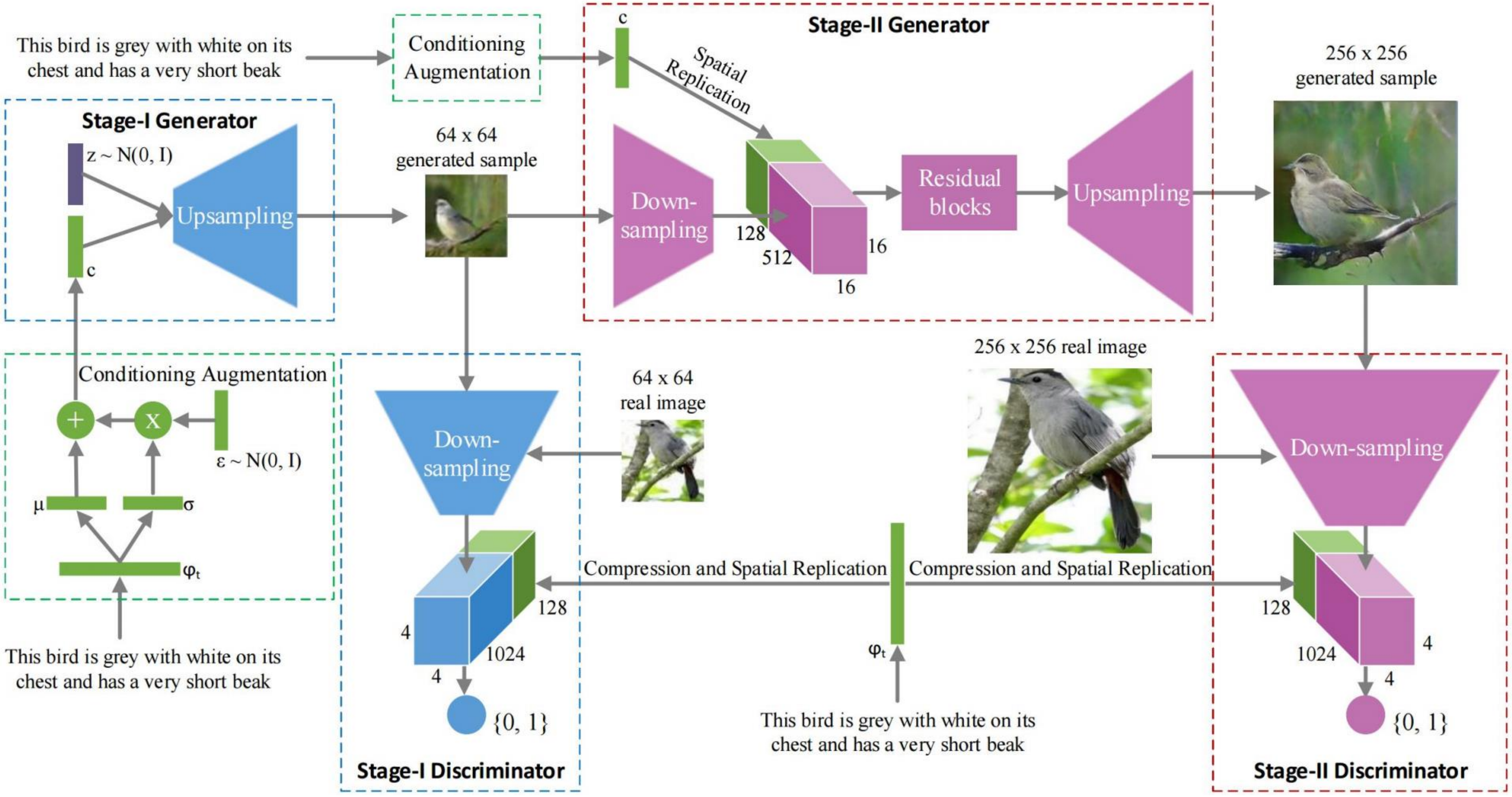
- Stage 1.
  - Generates 64x64 images
  - Structural information
  - Low detail
- Stage 2.
  - Requires Stage 1. output
  - Upsamples to 256x256
  - Higher detail, photorealistic

Both stages take in the same conditioned textual input



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

StackGAN

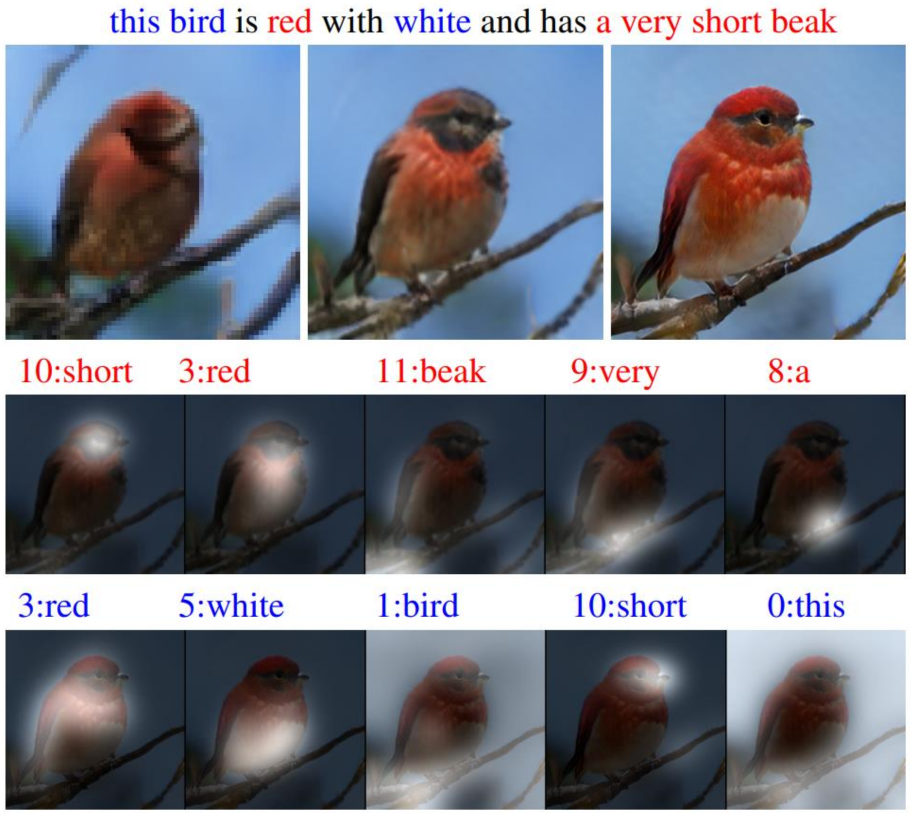

This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
Stage-II Generator
c
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~ N(0, I)
Upsampling
c
64 x 64 generated sample
Down-sampling
128
512
16
16
Residual blocks
Upsampling
Conditioning Augmentation
ε ~ N(0, I)
μ
σ
φt
This bird is grey with white on its chest and has a very short beak
Down-sampling
64 x 64 real image
256 x 256 real image
Down-sampling
Compression and Spatial Replication
Compression and Spatial Replication
128
4
1024
4
{0, 1}
Stage-I Discriminator
φt
This bird is grey with white on its chest and has a very short beak
128
1024
4
4
{0, 1}
Stage-II Discriminator

StackGAN


Text description
This bird is blue with white and has a very short beak
This bird has wings that are brown and has a yellow belly
A white bird with a black crown and yellow beak
This bird is white, black, and brown in color, with a brown beak
The bird has small beak, with reddish brown crown and gray belly
This is a small, black bird with a white breast and white on the wingbars.
This bird is white black and yellow in color, with a short black beak
Stage-I images
Stage-II images

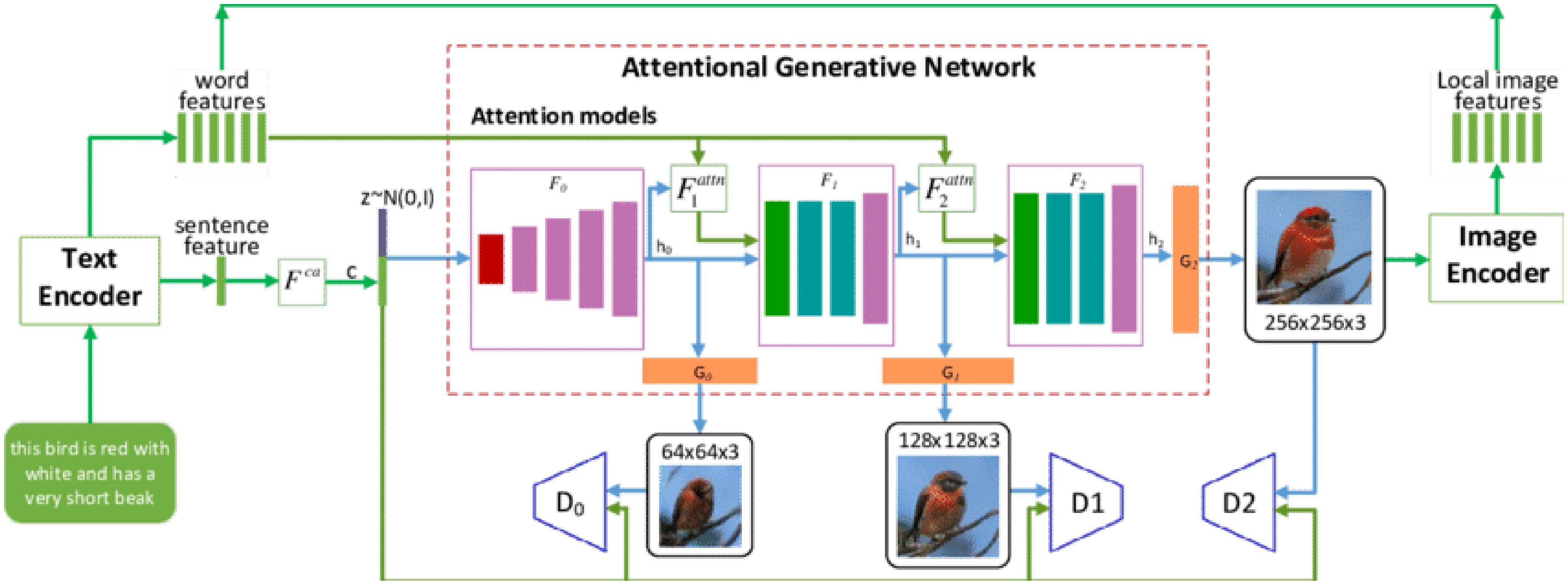
AttnGAN

- Paying attentions to the relevant words in the natural language description
- Capture both both the global sentence level information and the fine-grained word level information



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

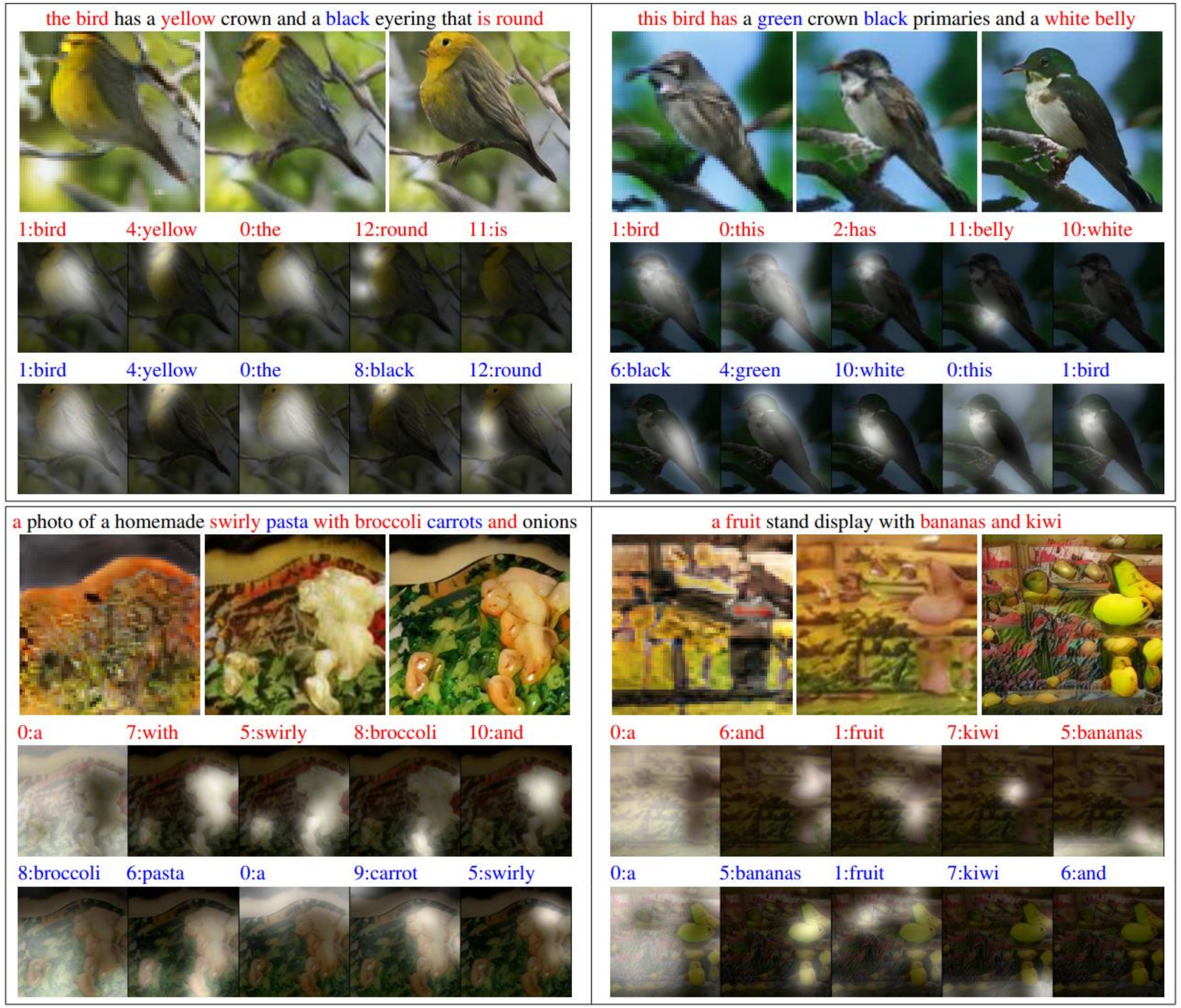
AttnGAN



word features
Attentional Generative Network
Attention models
Local image features
sentence feature
Text Encoder
z~N(0,I)
c
F^{ca}
F_0
F_1^{attn}
F_1
F_2^{attn}
F_2
h_0
h_1
h_2
G_0
G_1
G_2
256x256x3
Image Encoder
64x64x3
128x128x3
D_0
D1
D2
this bird is red with white and has a very short beak

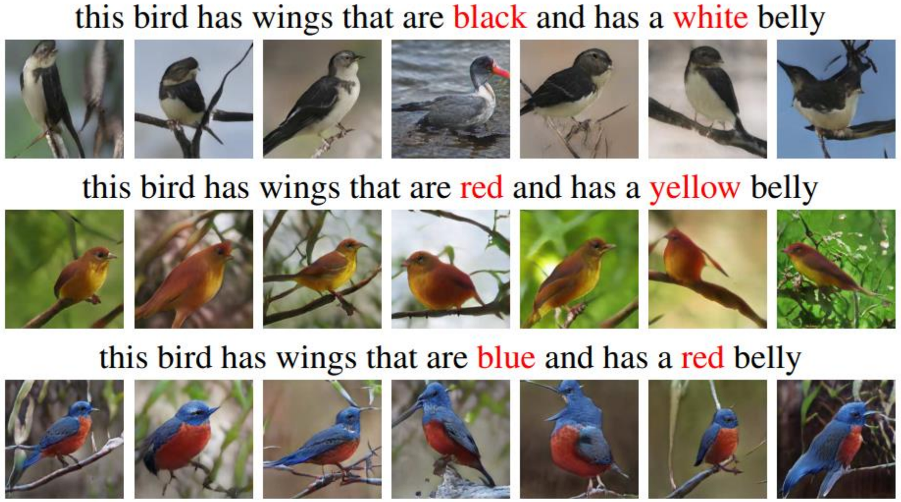
AttnGAN

- AttnGAN can generation more object detailed information



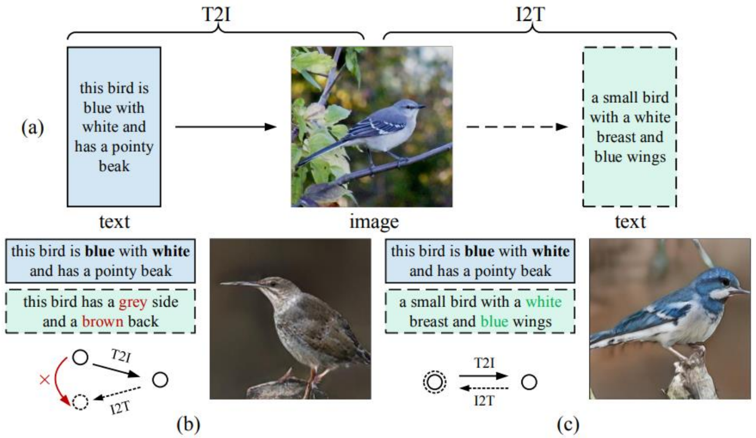
AttnGAN

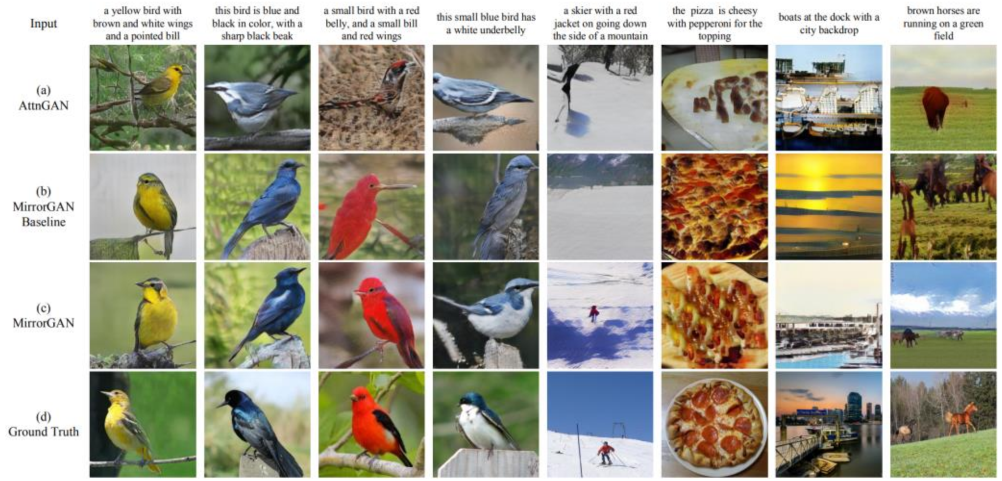
Dataset	GAN-INT-CLS [20]	GAWWN [18]	StackGAN [36]	StackGAN-v2 [37]	PPGN [16]	Our AttnGAN
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.84 \pm .06$	/	$\mathbf{4.36 \pm .03}$
COCO	$7.88 \pm .07$	/	$8.45 \pm .03$	/	$9.58 \pm .21$	$\mathbf{25.89 \pm .47}$



MirrorGAN

- Using a semantic-preserving text-to-image-to-text framework





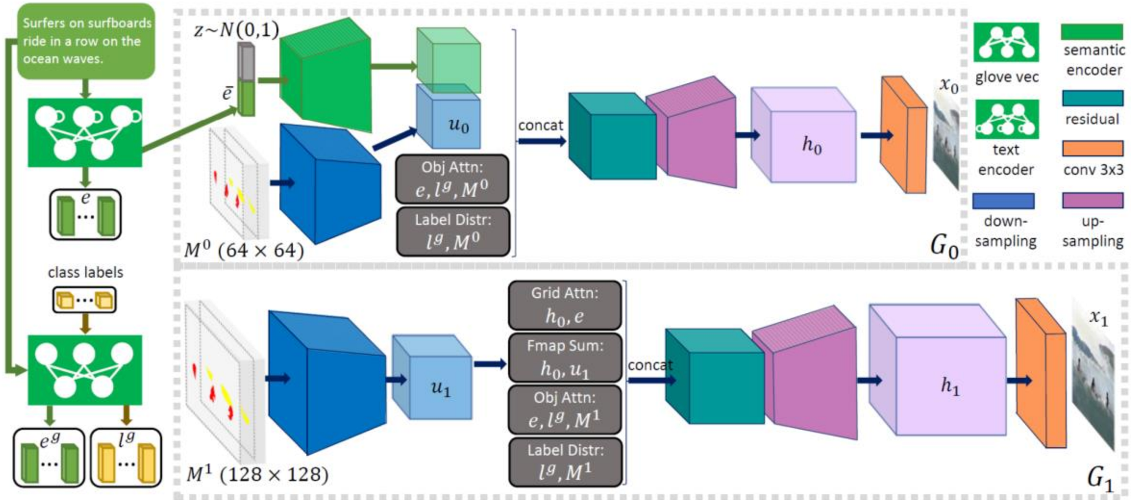
Qiao et al., 2019. MirrorGAN: Learning Text-to-image Generation by Redescription

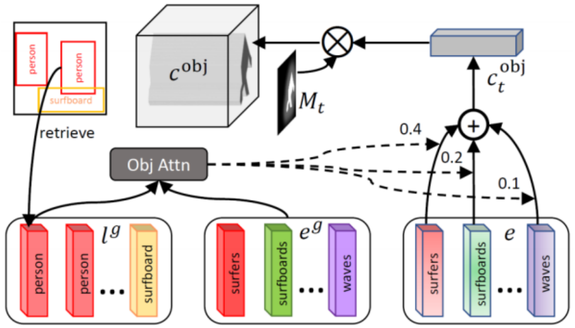
Text-to-Image Synthesis

- Current approaches follows StackGAN, AttenGAN
  - Generation quality is very good on CUB, flowers datasets
  - But not that good on complicated one, such as COCO
- What Evaluations?
  - IS, FID and human evaluation
- Technique challenges
  - How to handle large vocabulary
  - How to generate multiple objects and model their relations

ObjGAN

- Object-centered text-to-image synthesis for complex scenes





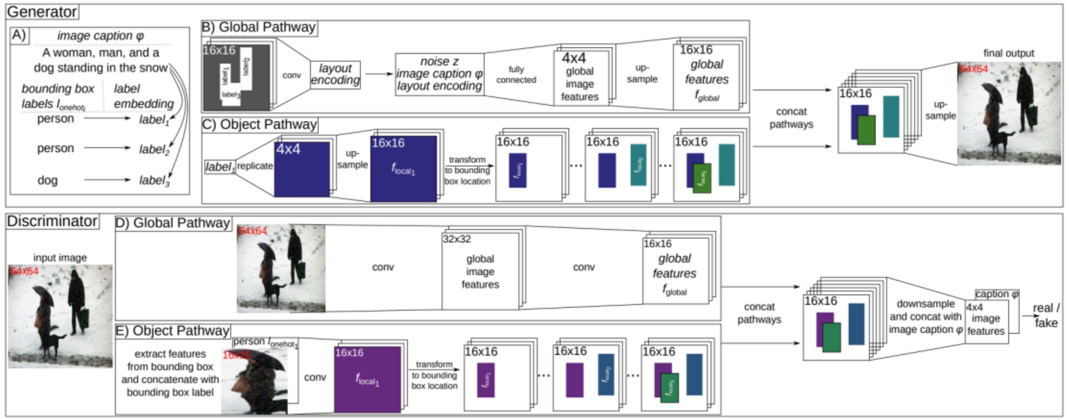
Li et al., 2019. Object-driven Text-to-Image Synthesis via Adversarial Training

ObjGAN

Methods	Inception ↑	FID ↓	R-prcn (%) ↑
Obj-GAN^0	**27.37 ± 0.22**	**25.85**	86.20 ± 2.98
Obj-GAN^1	27.96 ± 0.39*	24.19*	88.36 ± 2.82
Obj-GAN^2	29.89 ± 0.22**	20.75**	89.59 ± 2.67
P-AttnGAN w/ Lyt^0	18.84 ± 0.29	59.02	65.71 ± 3.74
P-AttnGAN w/ Lyt^1	19.32 ± 0.29	54.96	68.40 ± 3.79
P-AttnGAN w/ Lyt^2	20.81 ± 0.16	48.47	70.94 ± 3.70
P-AttnGAN	26.31 ± 0.43	41.51	86.71 ± 2.97
Obj-GAN w/ SN^0	26.97 ± 0.31	29.07	**86.84 ± 2.82**
Obj-GAN w/ SN^1	27.41 ± 0.17	27.26	88.70 ± 2.65*
Obj-GAN w/ SN^2	28.75 ± 0.32	23.37	89.97 ± 2.56**
Reed *et al.* [23]†	7.88 ± 0.07	n/a	n/a
StackGAN [32]†	8.45 ± 0.03	n/a	n/a
AttnGAN [29]	23.79 ± 0.32	28.76	82.98 ± 3.15
vmGAN [35]†	9.94 ± 0.12	n/a	n/a
Sg2Im [12]†	6.7 ± 0.1	n/a	n/a
Infer [9]^0†	11.46 ± 0.09	n/a	n/a
Infer [9]^1†	11.94 ± 0.09	n/a	n/a
Infer [9]^2†	12.40 ± 0.08	n/a	n/a
Obj-GAN-SOTA^0	30.29 ± 0.33	25.64	91.05 ± 2.34
Obj-GAN-SOTA^1	30.91 ± 0.29	24.28	92.54 ± 2.16
Obj-GAN-SOTA^2	32.79 ± 0.21	21.21	93.39 ± 2.08

Object Pathways

- Using a separate net to model the objects/relations



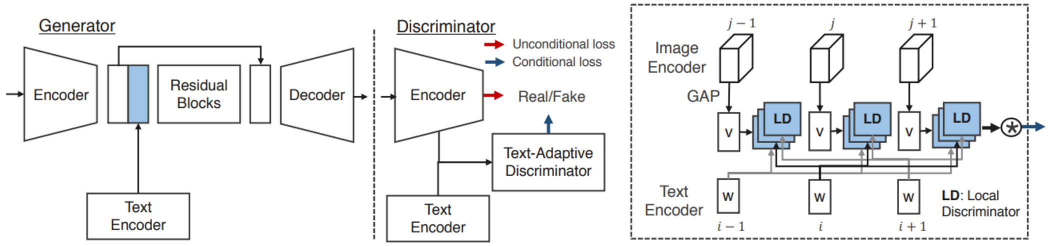
Hinz et al., 2019. Generating Multiple Objects at Spatially Distinct Locations

Text-Adaptive GAN (TAGAN)

- Task: manipulating images using natural language description

This particular bird with a **red head and breast** and features **grey wings**.

This small bird has a **blue crown** and **white belly**.



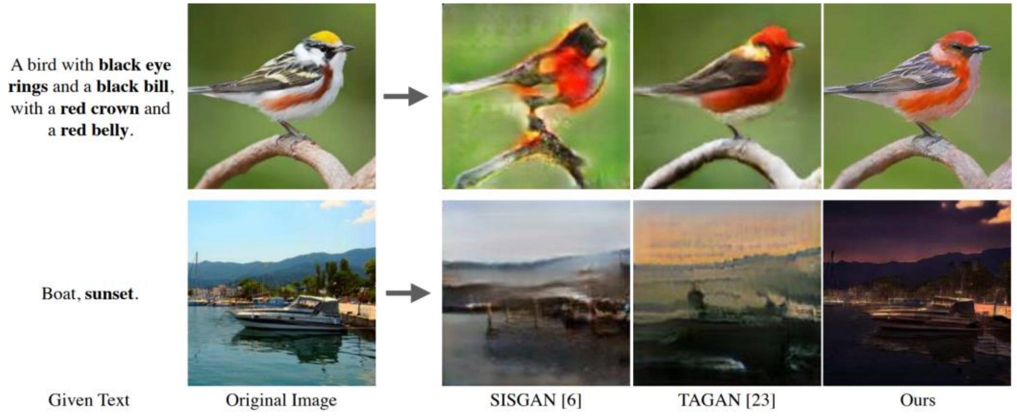
(a) GAN structure

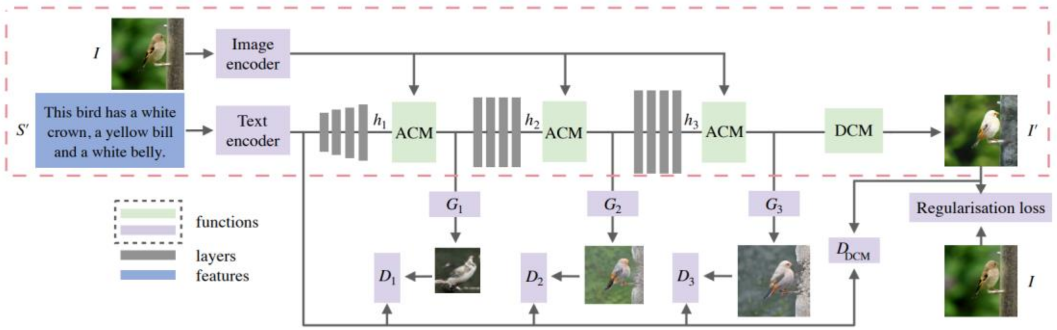
(b) Text-adaptive discriminator

Nam et al., 2018. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language

ManiGAN

- Consists of text-image affine combination module (ACM) and detail correction module (DCM)

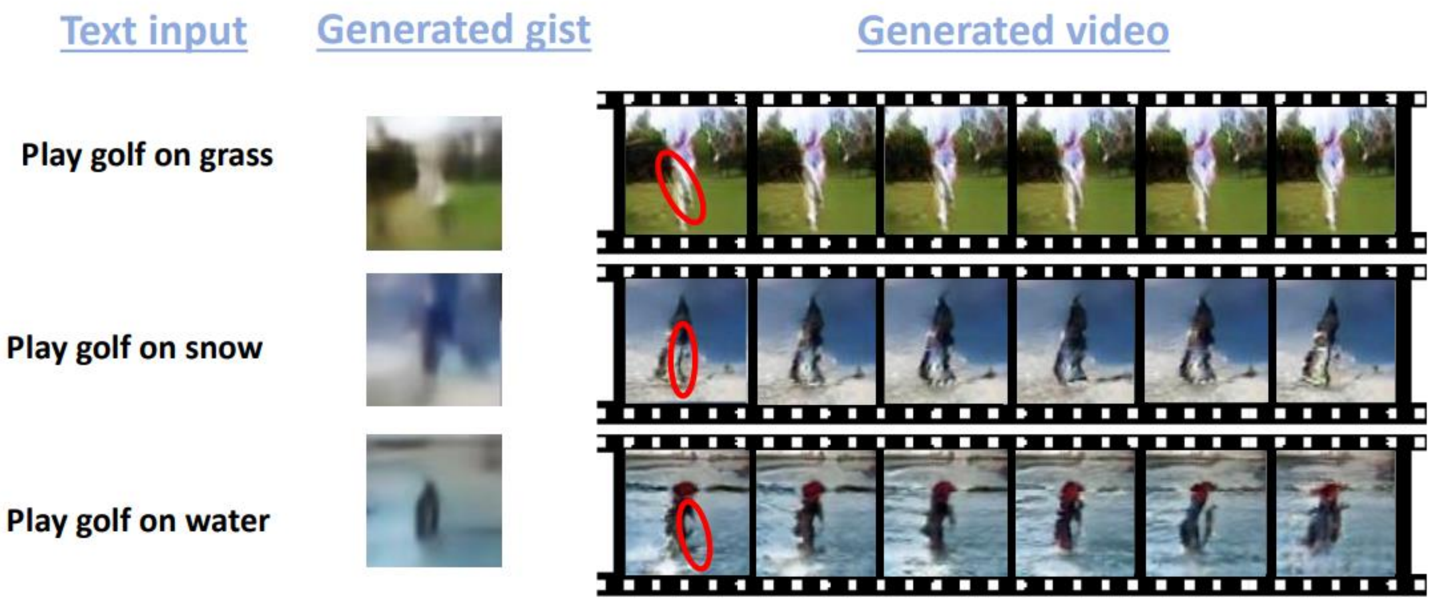




Li et al., 2020. ManiGAN: Text-Guided Image Manipulation

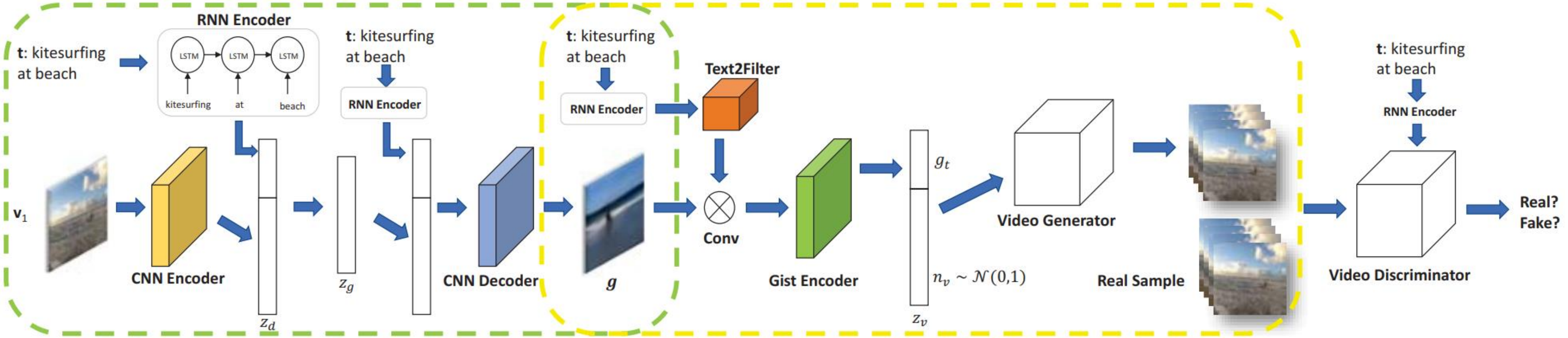
Text-to-Video Synthesis

- Task: generating a sequence of image given text description



T2V

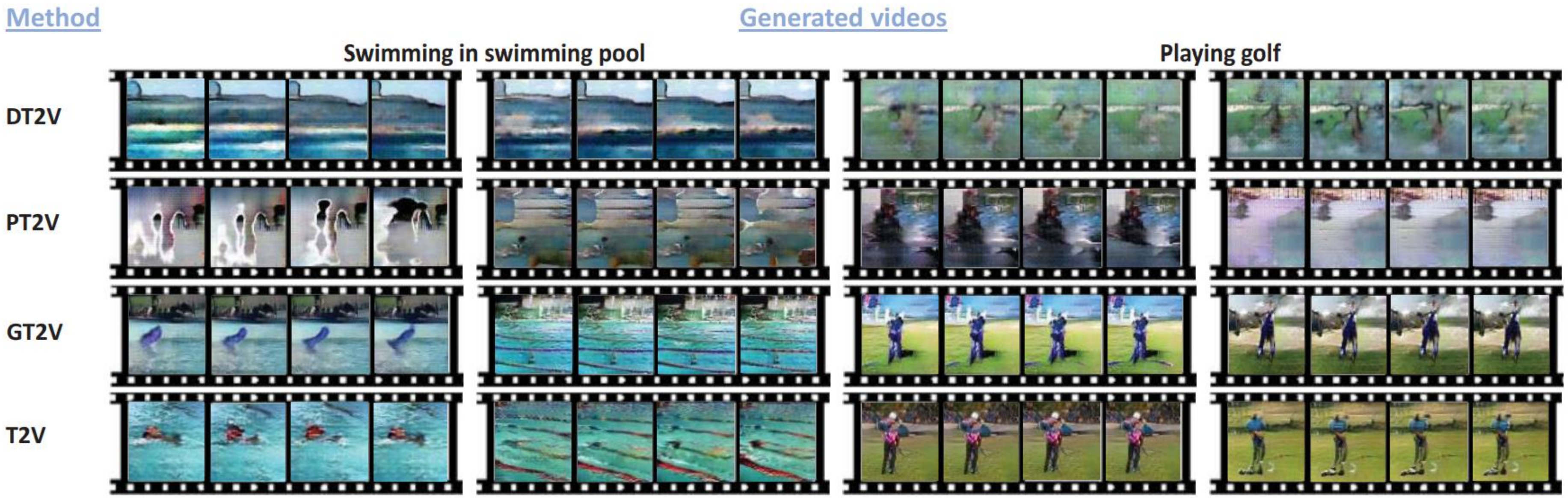
T2V: a VAE framework combining the text and gist information



Li et al., 2018. Video Generation from Text

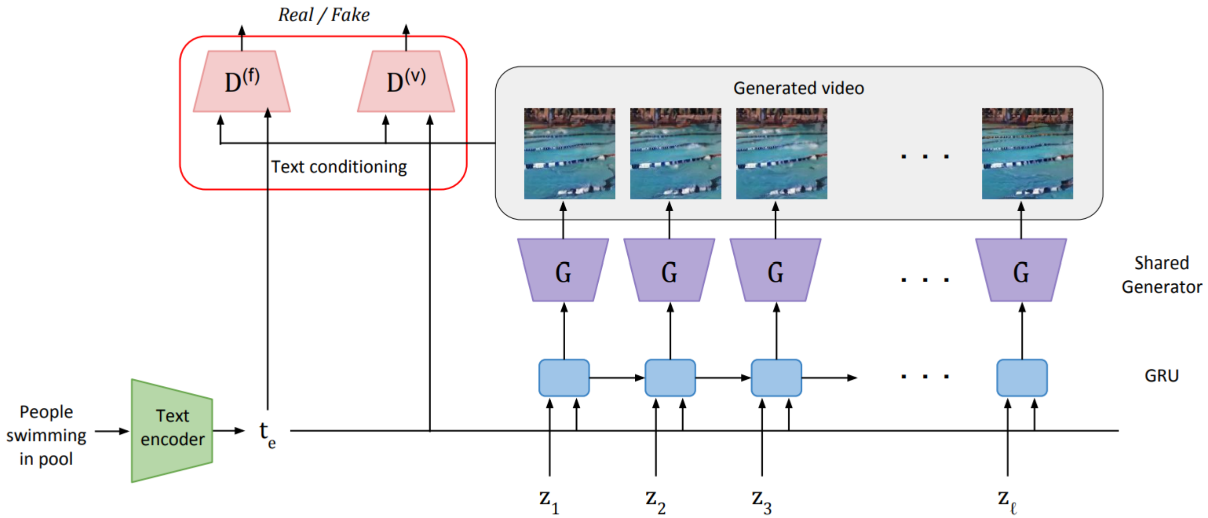
T2V

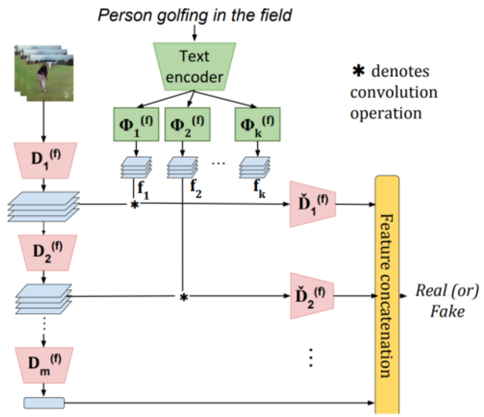
	In-set	DT2V	PT2V	GT2V	T2V
Accuracy	0.781	0.101	0.134	0.192	0.426



TFGAN

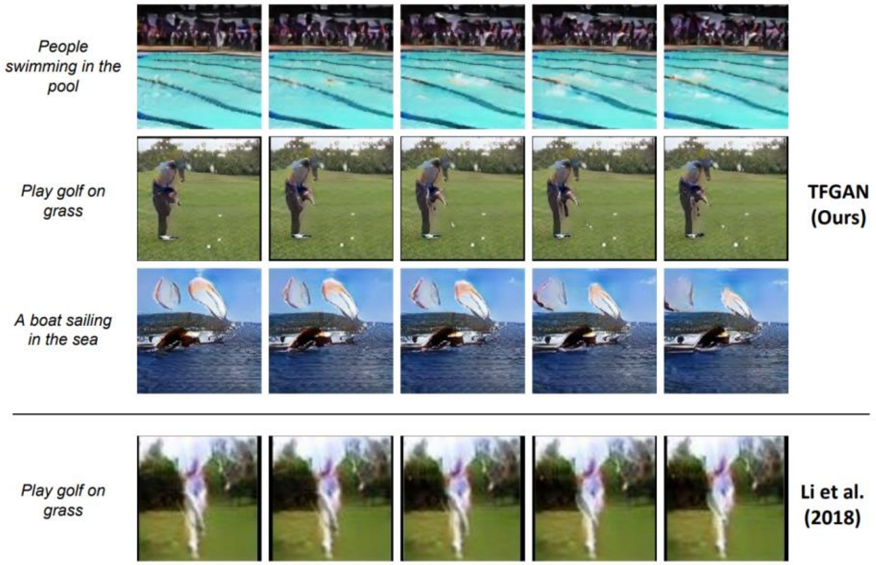
- GAN with multi-scale text-conditioning scheme based on convolutional filter generation

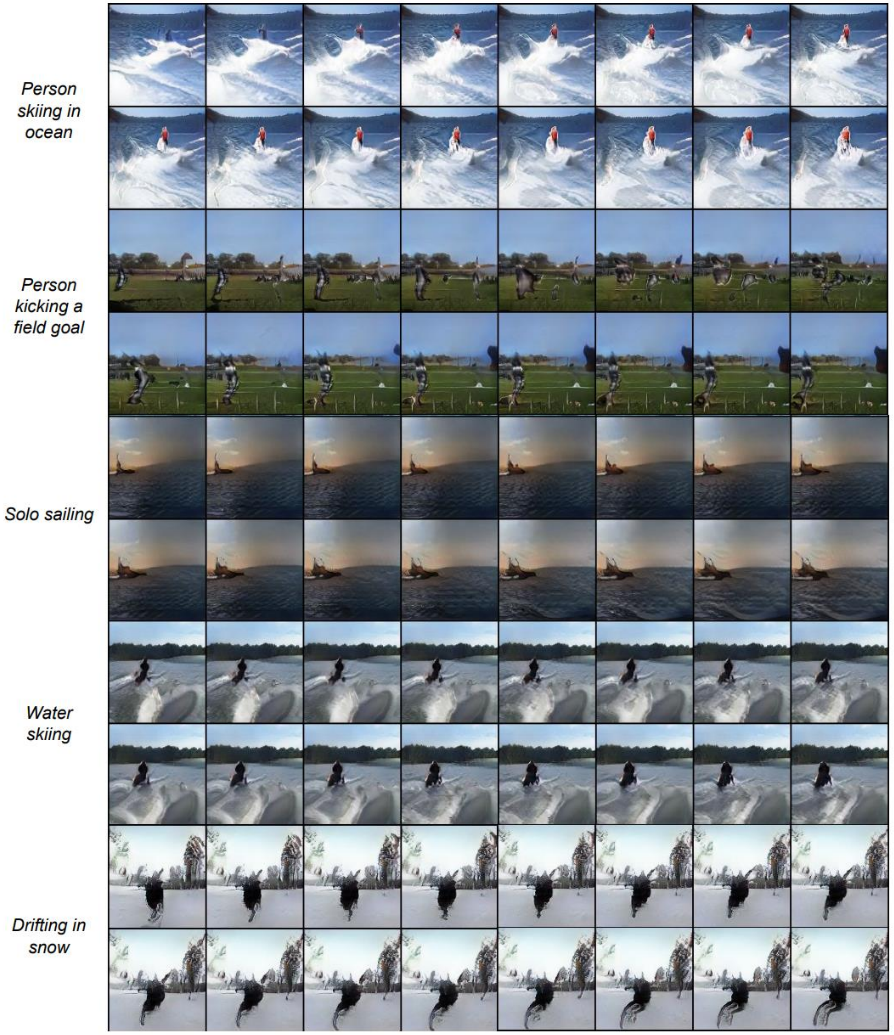




Balaji et al,. 2018. TFGAN: Improving Conditioning for Text-to-Video Synthesis

TFGAN


People swimming in the pool
Play golf on grass
A boat sailing in the sea
TFGAN (Ours)
Play golf on grass
Li et al. (2018)


Person skiing in ocean
Person kicking a field goal
Solo sailing
Water skiing
Drifting in snow

StoryGAN

- Short story (sequence of sentences) → Sequence of images

Image Generation

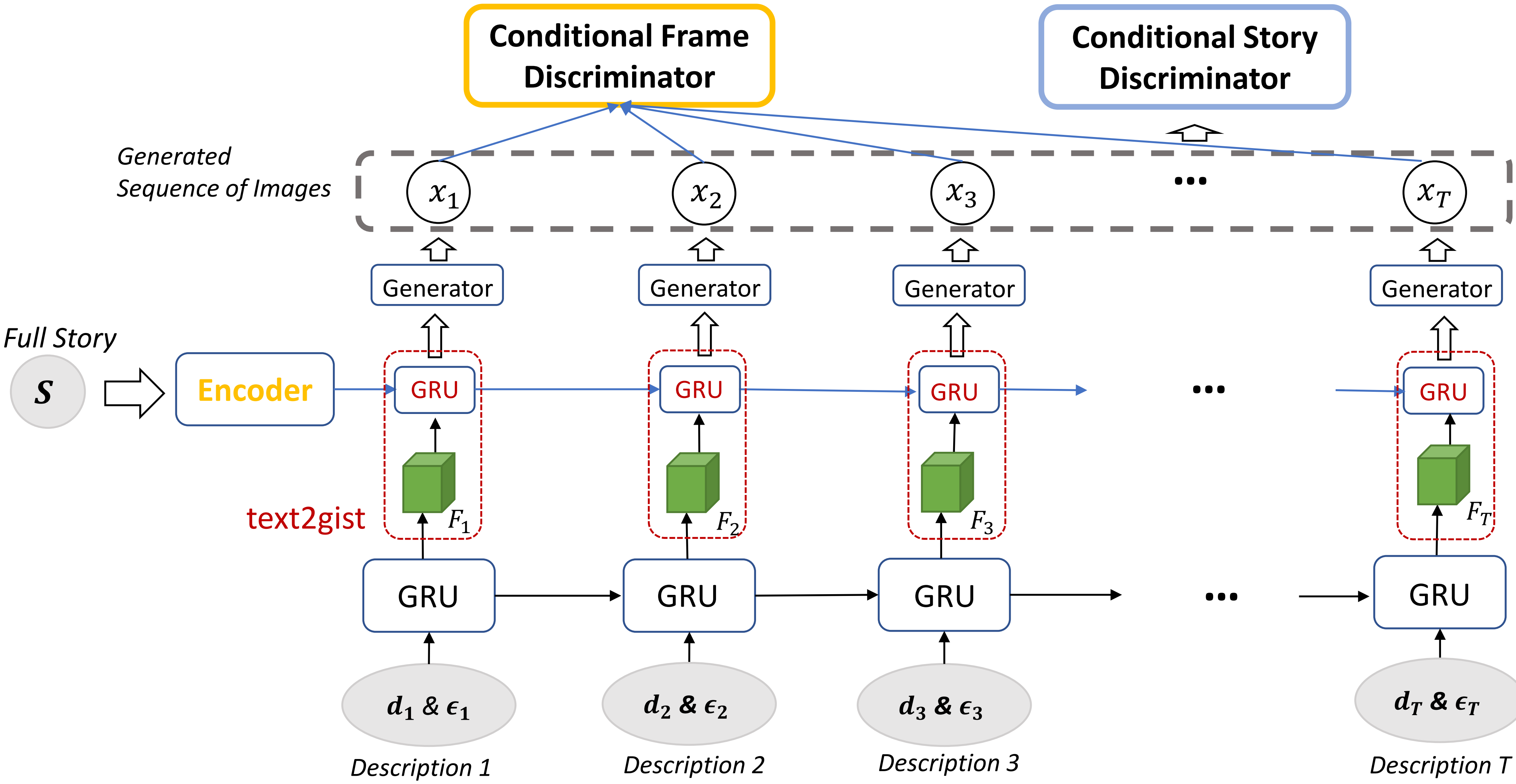
"A small yellow bird with a black crown and beak."

Story Visualization

"Pororo and Crong fishing together. Crong is looking at the bucket. Pororo has a fish on his fishing rod."

Li et al., 2018. StoryGAN: A Sequential Conditional GAN for Story Visualization

StoryGAN



Conditional Frame Discriminator
Conditional Story Discriminator
Generated Sequence of Images
x_1
x_2
x_3
…
x_T
Generator
Generator
Generator
Generator
Full Story
S
Encoder
GRU
GRU
GRU
…
GRU
text2gist
F_1
F_2
F_3
F_T
GRU
GRU
GRU
…
GRU
d_1 & ϵ_1
d_2 & ϵ_2
d_3 & ϵ_3
d_T & ϵ_T
Description 1
Description 2
Description 3
Description T

CLEVR Dataset: Result I

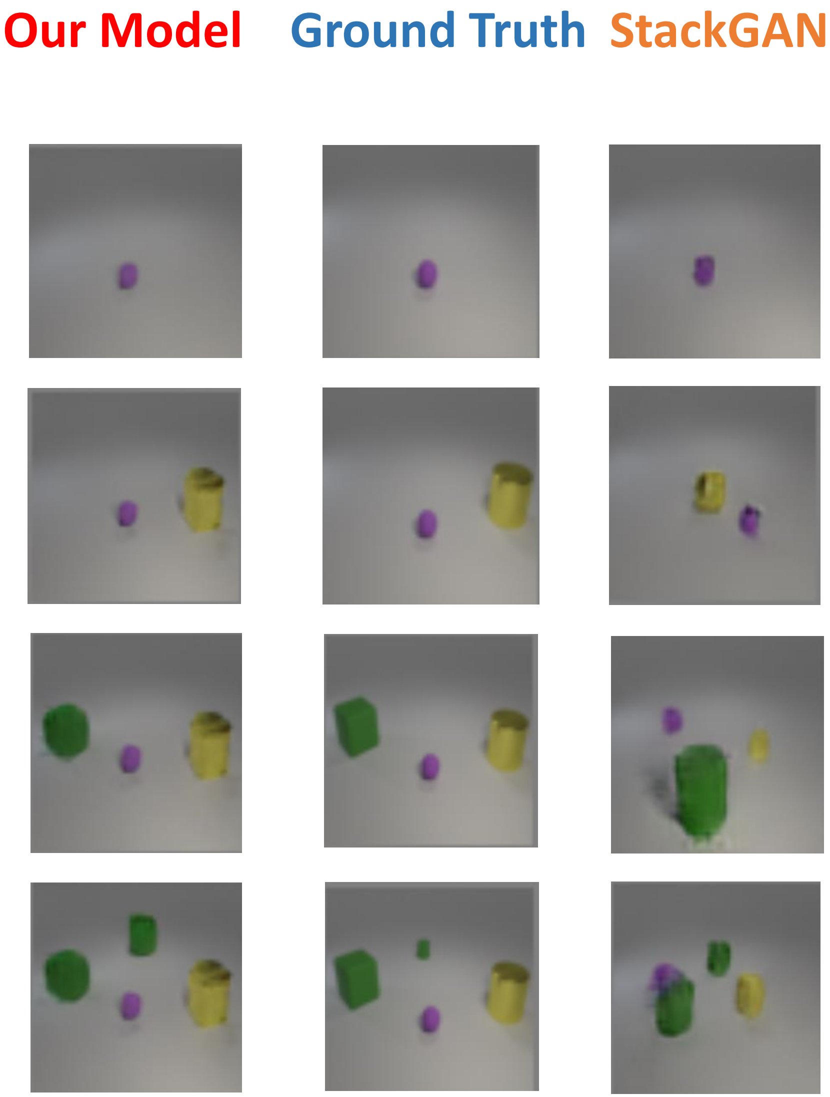
- Given attributes of objects, generate the image

“Small purple rubber sphere, position is 1.4, -0.7.”

“Large yellow metallic cylinder, position is 2.1, 2.6.”

“Large green rubber cube, position is -2.0, -1.2.”

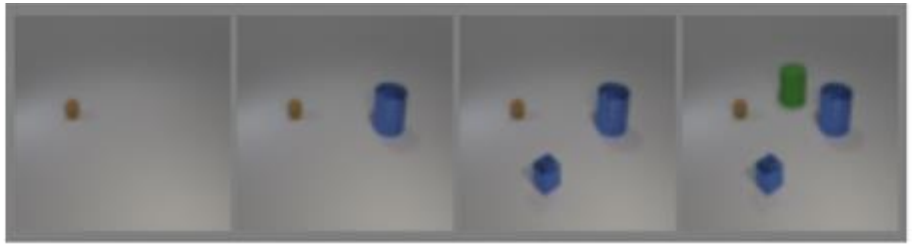
“Small green rubber cylinder, position is -2.5, 1.6.”



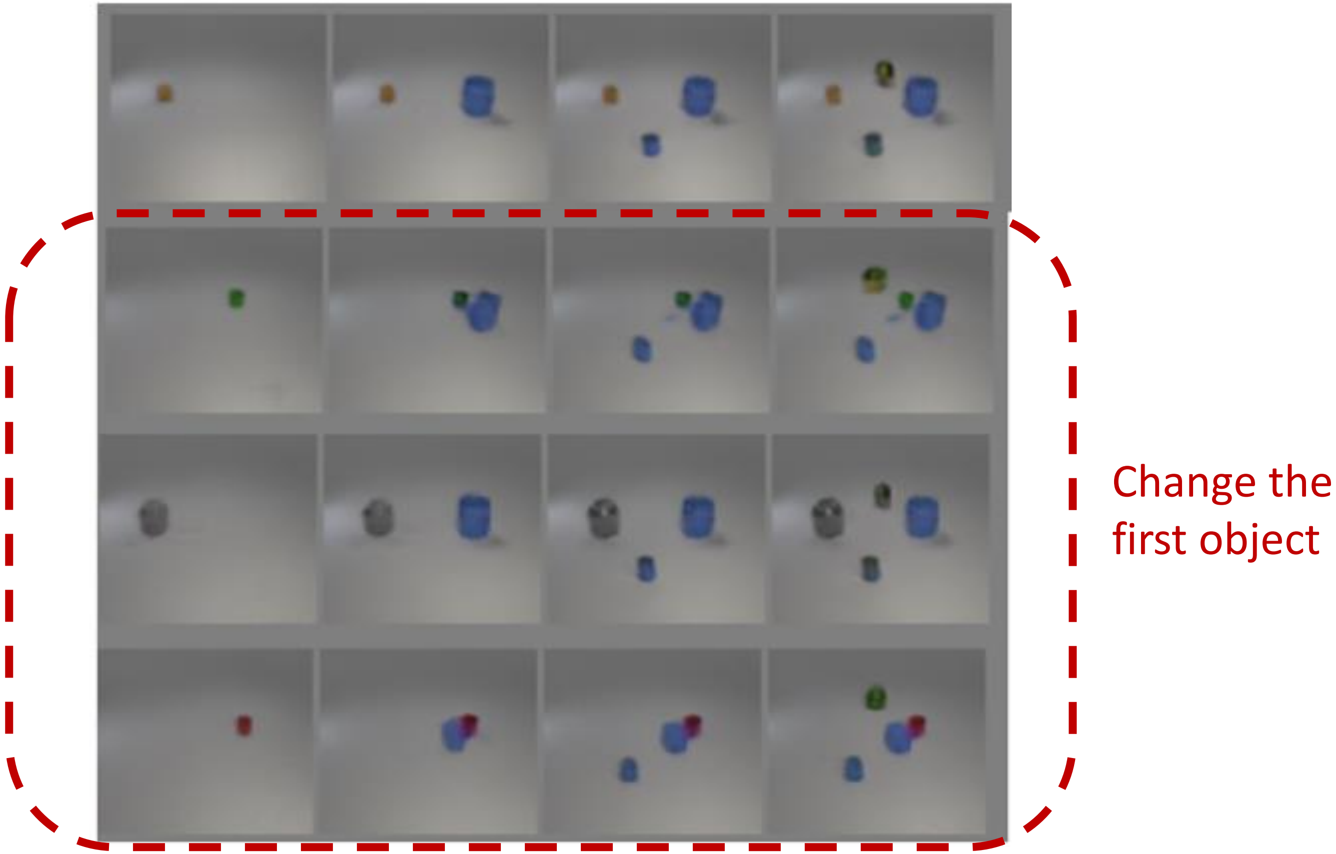
CLEVR Dataset: Result II

- Validate consistency (ongoing)

Real Images



Generated Images



Pororo Dataset: Result I

- Given text descriptions of a short story, generate a sequence of images

Pororo arrives at the top. Pororo is surprised. Pororo opens a red car. Pororo is ready to get down. Pororo takes off from the top.

The forest is covered with snow. Loopy is seated beside a house. Loopy is reading a book. A princess is looking at a mirror on the wall. Loopy gets surprised.

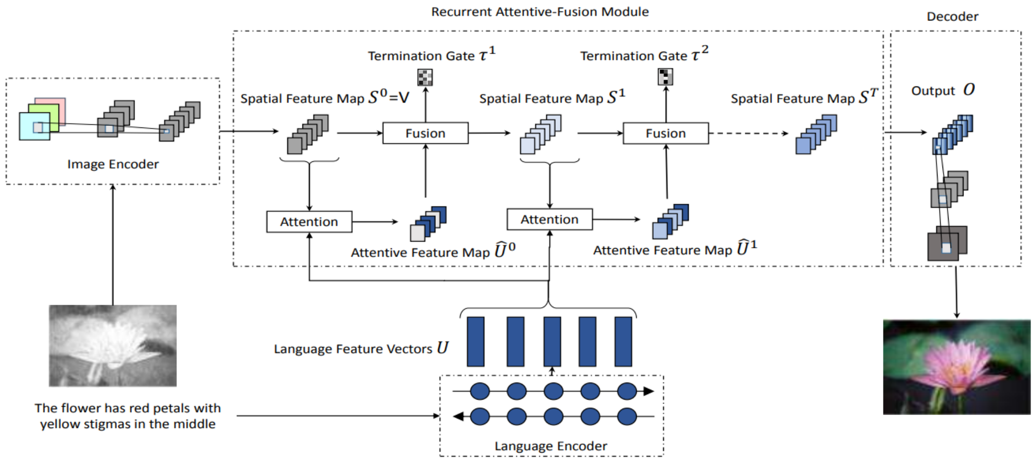
Pororo Dataset: Result II

- Given text descriptions of a short story, generate a sequence of images

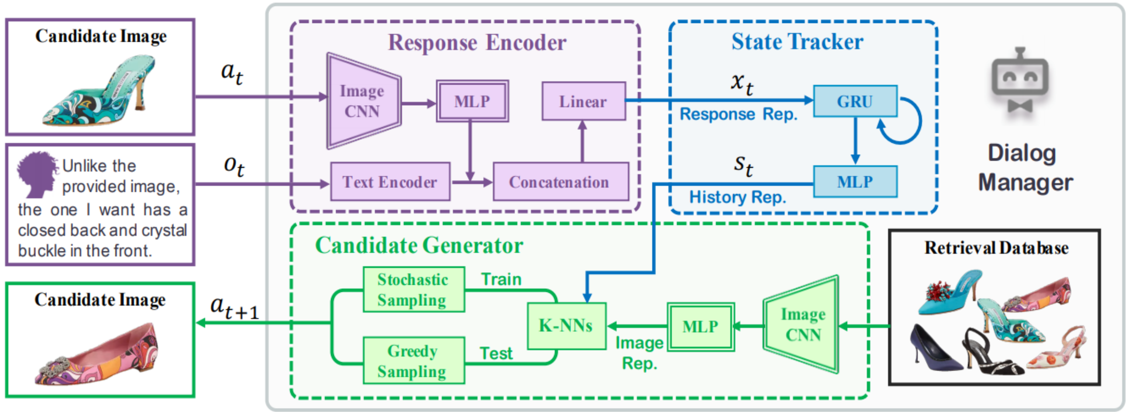
The woods are covered with snow. The sky is blue and clear. Pororo went to Loppy's house. Pororo saw crong. They are in front of a door. Crong looked at his friends. Loppy smiled at Crong.

Loopy is in a wooden house looking at Pororo. Loopy wants Pororo to come in. They are in a wooden house. Loopy is coming closer to Pororo. Loopy finds Crong. Pororo is sitting on a green couch. Pororo is asking why Loopy has come to his house. Loppy is stretching his arms and saying let's go to play ground.

Dialogue-based Image Synthesis



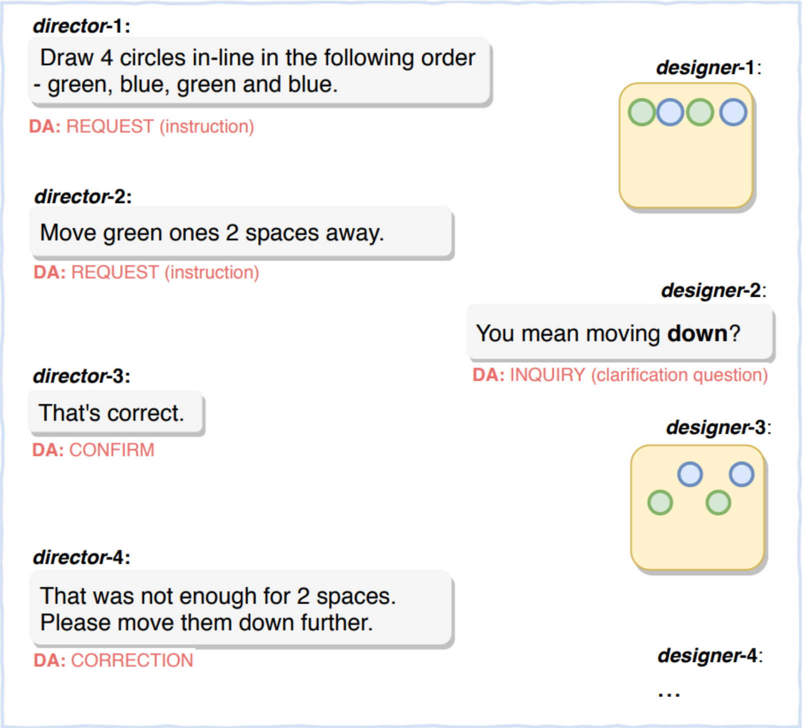
Text-based image editing
[Chen et al., 2018]

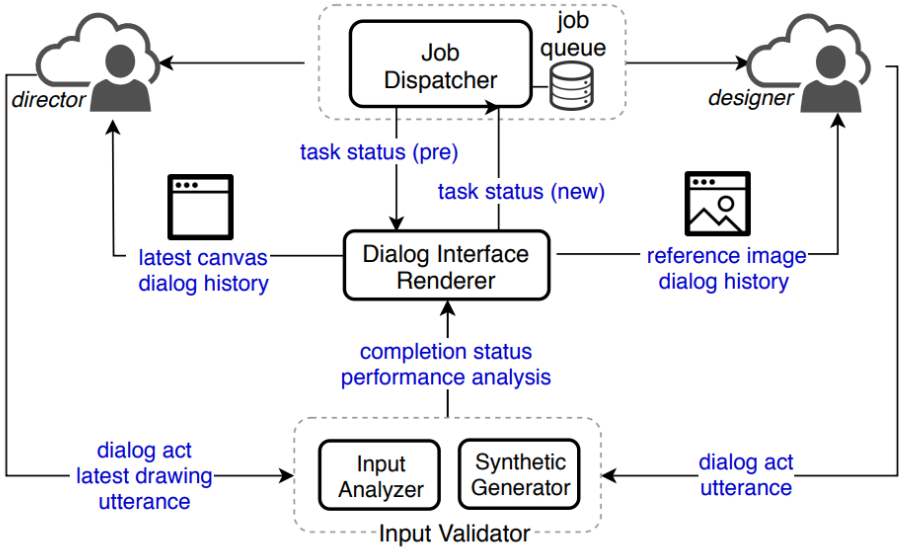


Dialogue-based image retrieval
[Guo et al., 2018]

Chat-crowd

- A Dialog-based Platform for Visual Layout Composition

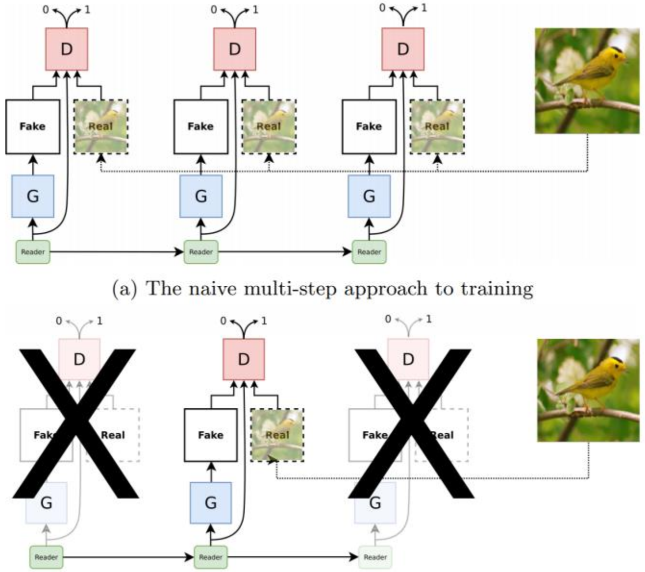




Bollina et al., 2018. Chat-crowd: A Dialog-based Platform for Visual Layout Composition

Neural Painter

- Randomly sample a sequence each time and only backprop through the GAN for that step in the sequence



Benmalek et al., 2018. The Neural Painter: Multi-Turn Image Generation

ChatPainter

- A new dataset of image generation based on multi-turn dialogues

(a) A flock of birds flying in a blue sky.

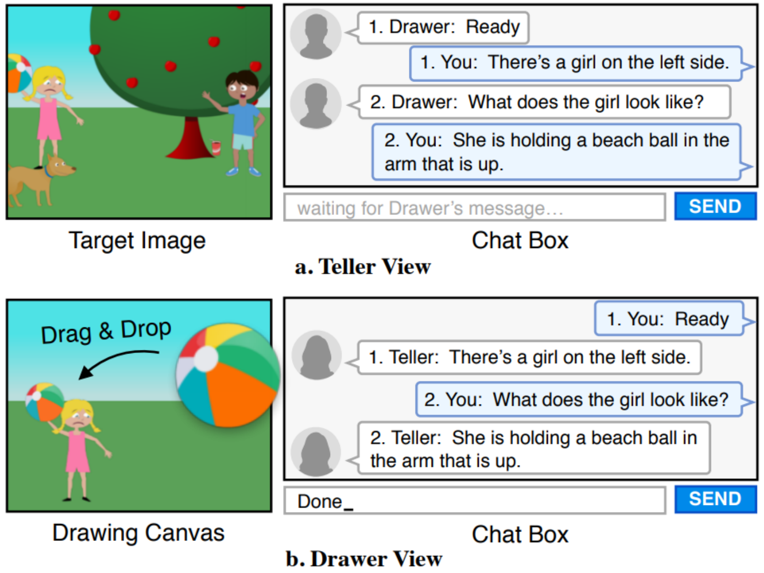
(b) A flock of birds flying in an overcast sky

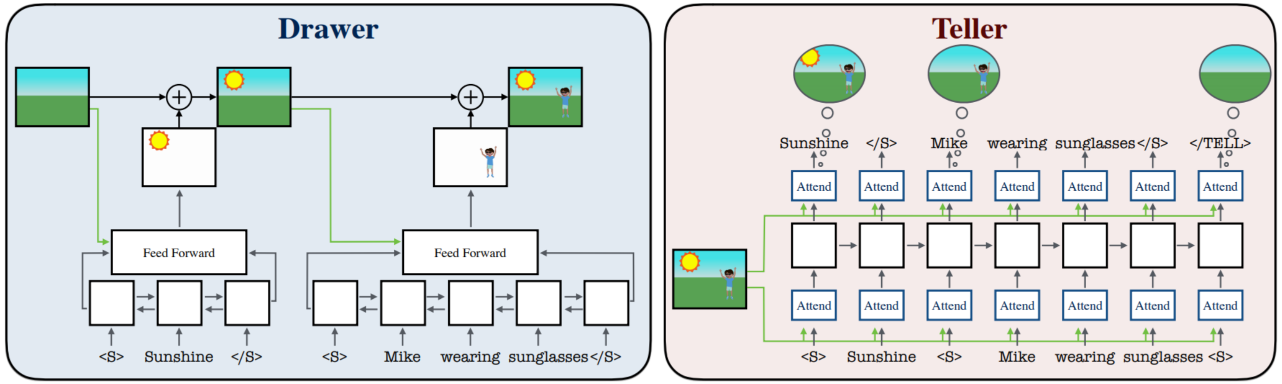
Input		**Dataset image**	**Generated image**
Caption: adult woman with yellow surfboard standing in water.			
Q: is the woman standing on the board?	A: no she is beside it.		
Q: how much of her is in the water?	A: up to her midsection.		
Q: what color is the board?	A: yellow.		
Q: is she wearing sunglasses?	A: no.		
Q: what about a wetsuit?	A: no she has on a bikini top.		
Q: what color is the top?	A: orange and white.		
Q: can you see any other surfers?	A: no.		
Q: is it sunny?	A: the sky isn't visible but it appears to be a nice day.		
Q: can you see any palm trees?	A: no.		
Q: what about mountains?	A: no.		

Sharma, et al., 2018. ChatPainter: Improving Text to Image Generation using Dialogue

CoDraw

- A goal-driven collaborative task involves two players: a Teller and a Drawer

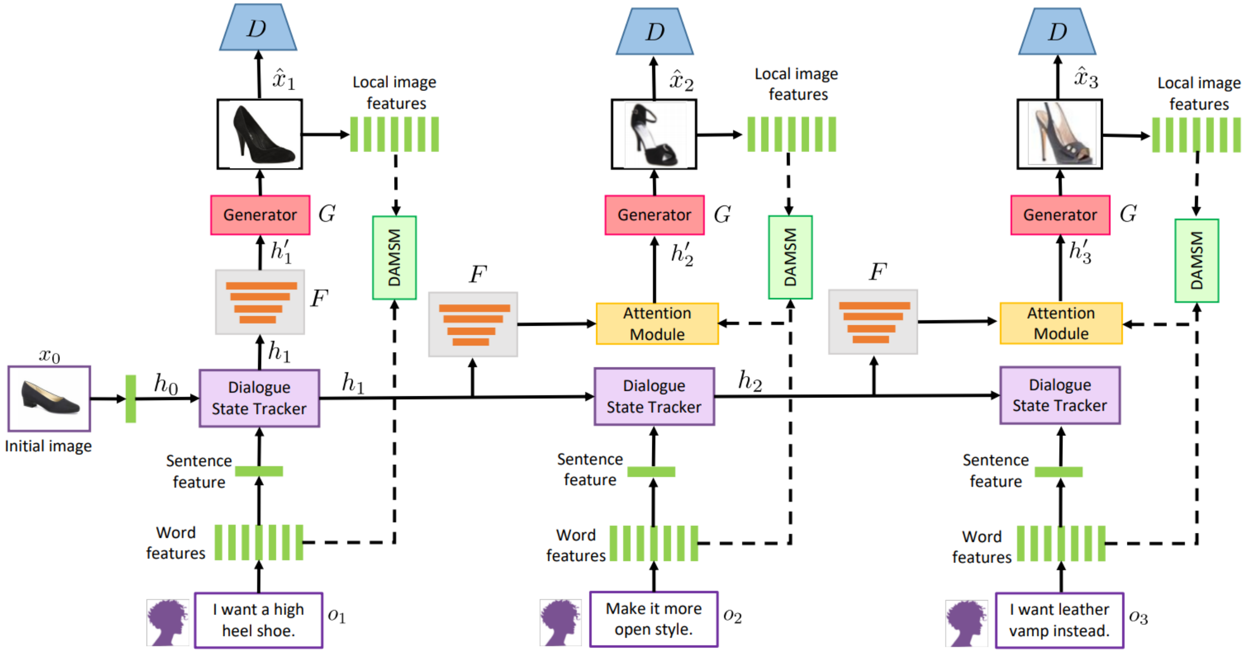




Kim et al., 2019. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication

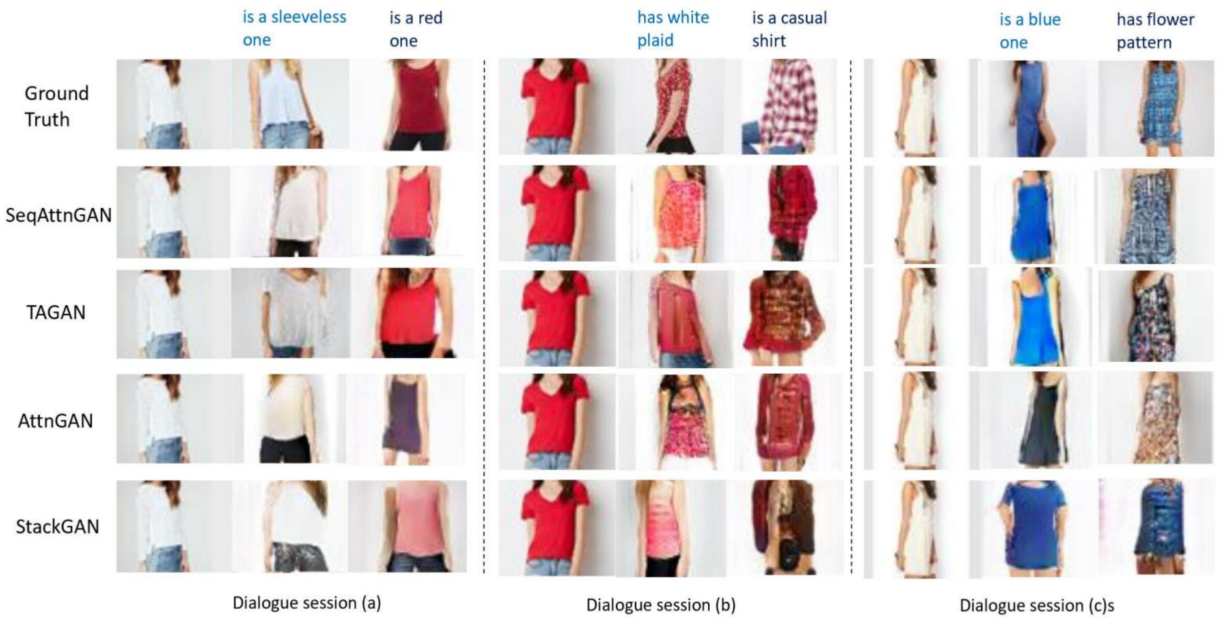
SeqAttnGAN

- Two new datasets: Zap-Seq and DeepFashion-Seq
- A method is extended from AttnGAN using sequential attention



Cheng et al., 2019. Sequential Attention GAN for Interactive Image Editing via Dialogue

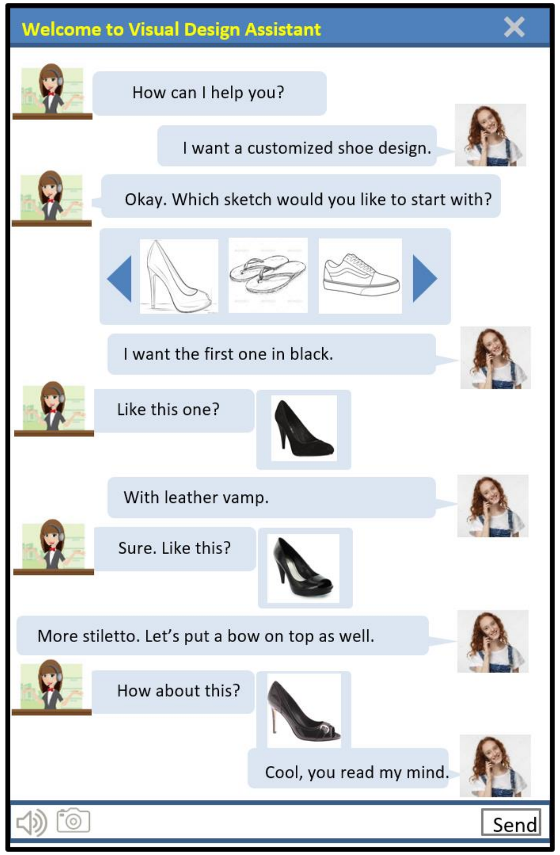
SeqAttnGAN


is a sleeveless one
is a red one
has white plaid
is a casual shirt
is a blue one
has flower pattern
Ground Truth
SeqAttnGAN
TAGAN
AttnGAN
StackGAN
Dialogue session (a)
Dialogue session (b)
Dialogue session (c)s


has a thick wedge heel
has a buckle and kitten heel
is a full boo
is a gray and leather one
are high-heeled shoes
have a button on the top
Ground Truth
SeqAttnGAN
TAGAN
AttnGAN
StackGAN
Dialogue session (a)
Dialogue session (b)
Dialogue session (c)

Text (Dialogue)-to-Video Synthesis

- There are several trials in recent years
  - Problem definition, datasets efforts
  - Some preliminary results are shown
- Technique challenges and solutions
  - Good (high quality) benchmarks
  - New evaluations
  - Generation consistency, disentangled learning, compositional generation



Thank you!

Q & A