VIDEO-TO-VIDEO SYNTHESIS

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro
GENERATIVE ADVERSARIAL NETWORKS
Unconditional GANs

\[ z \sim Z \rightarrow \text{Generator} \rightarrow \text{Discriminator} \rightarrow \text{False} \]

\[ \sim \rightarrow \text{Discriminator} \rightarrow \text{True} \]

Image credit: Celebrity dataset, Jensen Huang, Founder and CEO of NVIDIA, Ian Goodfellow, Father of GANs.
After training for a while using NVIDIA DGX1 machines

Fun sampling time begin

$z_1, z_2, z_3, \ldots \to \text{Generator}$

Image credit: NVIDIA StyleGAN
CONDITIONAL GANS
Allow user more control on the sampling process

Modeling
\( p_{X|Y} \)

( training )
Generated result
Given info (e.g. image, text)

Sampling
\( z \sim Z, y \sim Y \)

( testing )
output style
Given info (e.g. image, text)
SKETCH-CONDITIONAL GANS

Image credit: NVIDIA pix2pixHD
IMAGE-CONDITIONAL GANS

Image credit: NVIDIA MUNIT
MASK-CONDITIONAL GANS

Semantic Image Synthesis
Semantic Image Synthesis

- MASK

- CONDITIONAL

GANS
LIVE DEMO


- It is running live in GTC

PROBLEM WITH PREVIOUS METHODS

<table>
<thead>
<tr>
<th>input</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td></td>
</tr>
<tr>
<td>grass</td>
<td></td>
</tr>
</tbody>
</table>
PROBLEM WITH PREVIOUS METHODS

Batch Norm (Ioffe et al. 2015)

\[ y = \frac{x - \mu}{\sigma} \cdot \gamma + \beta \]

removes label information

same output!
PROBLEM WITH PREVIOUS METHODS

input  result

sky

grass
PROBLEM WITH PREVIOUS METHODS

• Do not feed the label map directly to network
• Use the label map to generate normalization layers instead
**SPADE** (Spatially Adaptive DENormalization)

\[
y = \frac{x - \mu}{\sigma} \cdot \gamma + \beta
\]

**Diagram:**
- Network input \( x \) (label free)
- Parameter-free Batch Norm
- Element-wise operation
- Network output \( y \)

- Convolution operations:
  - \( \text{conv} \)
  - \( \beta \)
  - \( \gamma \)
SPADE
SPatially Adaptive DE-normalization
SPADE RESIDUAL BLOCKS

SPADE ResBlk
SPADE GENERATOR
PROBLEM WITH PREVIOUS METHODS

<table>
<thead>
<tr>
<th>input</th>
<th>w/o SPADE</th>
<th>w/ SPADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td></td>
<td></td>
</tr>
<tr>
<td>grass</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IMAGE RESULTS
VIDEO-TO-VIDEO SYNTHESIS
IMAGE-TO-IMAGE SYNTHESIS
VIDEO-TO-VIDEO SYNTHESIS
VIDEO-TO-VIDEO SYNTHESIS
VIDEO-TO-VIDEO SYNTHESIS
VIDEO-TO-VIDEO SYNTHESIS
MOTIVATION

• AI-based rendering

Traditional graphics

Geometry, texture, lighting

Machine learning graphics

Data
MOTIVATION

• AI-based rendering

• High-level semantic manipulation

Largely explored

little explored (this work)

High-level representation

Segmentation
Keypoint Detection
etc

Original image

Image/video synthesis

New image

Edit here!
PREVIOUS WORK

Image translation

Video style transfer

Unconditional synthesis

Video prediction

pix2pixHD [2018], CRN [2017], pix2pix [2017]

COVST [2017], ArtST [2016]

MoCoGAN [2018], TGAN [2017], VGAN [2016]

MCNet [2017], PredNet [2017]
PREVIOUS WORK: FRAME-BY-FRAME RESULT
OUR METHOD

• Sequential generator
• Multi-scale temporal discriminator
• Spatio-temporal progressive training procedure
OUR METHOD

Sequential Generator
OUR METHOD

Sequential Generator

Multi-scale Discriminators

Image Discriminator

Video Discriminator
OUR METHOD

Spatio-temporally Progressive Training

Spatially progressive

Temporally progressive

Alternating training
RESULTS

• Semantic → Street view scenes
• Edges → Human faces
• Poses → Human bodies
RESULTS

• Semantic → Street view scenes
• Edges → Human faces
• Poses → Human bodies
STREET VIEW: CITYSCAPES

Semantic map

pix2pixHD

COVST (video style transfer)

Ours
STREET VIEW: BOSTON
STREET VIEW: NYC
RESULTS

• Semantic → Street view scenes
• Edges → Human faces
• Poses → Human bodies
FACE SWAPPING (FACE $\rightarrow$ EDGE $\rightarrow$ FACE)
FACE SWAPPING (SLIMMER FACE)

input  (slimmed) edges  (slimmed) output
FACE SWAPPING (SLIMMER FACE)

input  (slimmed) edges  (slimmed) output
MULTI-MODAL EDGE → FACE

Style 1

Style 2

Style 3
RESULTS

• Semantic $\rightarrow$ Street view scenes
• Edges $\rightarrow$ Human faces
• Poses $\rightarrow$ Human bodies
MOTION TRANSFER (BODY → POSE → BODY)

input  poses  output
MOTION TRANSFER (BODY $\rightarrow$ POSE $\rightarrow$ BODY)
MOTION TRANSFER (BODY → POSE → BODY)

input poses output
MOTION TRANSFER (BODY → POSE → BODY)

input  poses  output
MOTION TRANSFER
EXTENSION: FRAME PREDICTION

• Goal: predict future frames given past frames
• Our method: decompose prediction into two steps
  • 1. predict the semantic map for next frame
  • 2. synthesize the frame based on the semantic map
EXTENSION: FRAME PREDICTION

Ground truth

PredNet

MCNet

Ours
INTERACTIVE GRAPHICS
PATH TO INTERACTIVE GRAPHICS

• Real-time inference
• Combining with existing graphics pipeline
• Domain gap between real input and synthetic input
PATH TO INTERACTIVE GRAPHICS

- Real-time inference
- Combining with existing graphics pipeline
- Domain gap between real input and synthetic input
PATH TO INTERACTIVE GRAPHICS

• Real-time inference
  • FP16 + TensorRT $\rightarrow$ ~5 times speed up
  • 36ms (27.8 fps) for 1080p inference
  • Overall: 15-25 fps
PATH TO INTERACTIVE GRAPHICS

• Real-time inference

• Combining with existing graphics pipeline
  • CARLA: open-source simulator for autonomous driving research
  • Make game engine render semantic maps
  • Pass the maps to the network and display the inference result
PATH TO INTERACTIVE GRAPHICS

• Real-time inference

• Combining with existing graphics pipeline

• Domain gap between **real** input and **synthetic** input
  • Network trained on real data but tested on synthetic data
  • Things that differ: Object shapes/edges, density of objects, camera viewpoints, etc
  • On-going work
ORIGINAL CARLA IMAGE
RENDERED SEMANTIC MAPS
RECORDED DEMO RESULTS
RECORDED DEMO RESULTS
CONCLUSION
CONCLUSION

• What can we achieve?
• What can it be used for?
CONCLUSION

• What can we achieve?
  • Synthesize high-res realistic images
CONCLUSION

• What can we achieve?
  • Synthesize high-res realistic images
  • Produce temporally-smooth videos
CONCLUSION

• What can we achieve?
  • Synthesize high-res realistic images
  • Produce temporally-smooth videos
  • Reinvent interactive graphics
CONCLUSION

• What can we achieve?
• What can it be used for?
  • AI-based rendering
  • High-level semantic manipulation