Conditional Image Synthesis by Generative Adversarial Modeling

Zhifei Zhang
Conditional Image Synthesis by Generative Adversarial Modeling
Conditional Image Synthesis by Generative Adversarial Modeling

Instable and hard to train

Limited on empirical studies
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Generative Adversarial Network (GAN)

Noise ~ N(0,1) → Generative Model
Generative Adversarial Network (GAN)

Advantage of GAN:
Achieving more perceptually/visually realistic images.

https://github.com/hindupuravinash/the-gan-zoo
Generative Adversarial Network (GAN)

Advantage of GAN --- An example of super-resolution
Generative Adversarial Network (GAN)

Traditional loss for image synthesis ---
pixel-wise distance

GAN loss --- perceptual distance

Perceptual distance is more sensitive to edges and texture
Generative Adversarial Network (GAN)

[D: Discriminator (Detective)]

[X: Original Data]

[G: Generator (Counterfeit)]

[z: Input for generator]

[G. Ramachandra, 2017]
The objective function:

$$\min_G \max_D \mathbb{E}_{x \sim q(x)}[\log(D(x))] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Instability of GAN

- Mode missing
- Gradient vanishing

Perfect Discriminator causes gradient vanishing

Target

Step 0  Step 5k  Step 10k  Step 15k  Step 20k  Step 25k

[L. Metz et al., 2017]
Instability of GAN

The objective function

$$\min_G \max_D \mathbb{E}_{x \sim q(x)} [\log(D(x))] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]$$

Since,

$$\mathbb{E}_{x \sim q(x)} [p(x)] = \int_x q(x)p(x) \, dx$$

$$\min_G \max_D \int_x q(x) \log(D(x)) \, dx + \int_z p(z) \log(1 - D(G(z))) \, dz$$
Instability of GAN

Alternative update of $G$ and $D$

**Fix $G$,**

$$\min_G \max_D \int_x q(x) \log(D(x)) \, dx + \int_z p(z) \log(1 - D(G(z))) \, dz$$

$$= \max_D \int_x q(x) \log(D(x)) + p_g(x) \log(1 - D(x)) \, dx$$

**Fix $D^*$,**

$$= \min_G \int_x p_g(x) \log \left(1 - \frac{q(x)}{q(x) + p_g(x)} \right) \, dx$$

$$= \min_G \int_x p_g(x) \log \left(\frac{p_g(x)}{q(x) + p_g(x)} \right) \, dx$$

$$= \min_G D_{KL}(p_g \| q + p_g)$$

$$= \min_G D_{KL}(p_g \| \frac{q + p_g}{2}) - 2 \log 2$$

How does it come? (click to see)

Both mode missing and gradient vanishing are caused by KL-divergence
Instability of GAN

KL-divergence is an unsymmetrical measurement

\[ \nabla_x KL\left(\frac{p_g + q}{2}\right) \]

- \( p_g(x) > q(x) \): Unrealistic samples
- \( q(x) > p_g(x) \): Mode missing

Reasons of mode missing:
- Punish more on unrealistic samples
- Punish less on mode missing

Reason of gradient vanishing:
- KL-divergence is constant (i.e., zero) if two distribution are not overlapped. Then, the gradient will be zero.

\[
\begin{align*}
\min_G D_{KL}(p_g \parallel \frac{q + p_g}{2}) & - 2 \log 2 \\
= \min_G \int p_g(x) \log \left( \frac{p_g(x)}{q(x) + p_g(x)} \right) dx
\end{align*}
\]
**Conditional GAN**

Conditional GAN can relax the instabilities

In GAN, $D(x | p_g) = D(x | q)$, which is indirect matching.

Ideally, $p_g(x) = q(x)$, which is direct matching, avoiding mode missing.

An extra loss based on Euclidian distance relaxes gradient vanishing.
Conditional GAN

GAN: $\mathbb{E}_{x \sim p_x} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$

Conditional GAN: $\mathbb{E}_{x \sim p_x} [\log (D(x)(1 - D(\mathcal{H}(x)))) + \lambda \mathcal{L}(x, \mathcal{H}(x))]$

where $\lambda$ balances the effect of reconstruction error, and $\mathcal{H}(x) = G(E(x))$. 
Conditional GAN --- Summary

- Theatrical analysis of the instability in the GAN
- How conditional GAN stabilizes the training
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Image Synthesis by Conditional GAN

Becomes a common framework for image synthesis
Image Synthesis by Conditional GAN

Image condition [Isola et al., CVPR 2017]
Image Synthesis by Conditional GAN

Label condition [Reed et al., CVPR 2016]

- 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.
Face Aging by Conditional GAN

Regression/Rejuvenation

Given face

Progression/Aging

10

20

35 years old

40

60

Image synthesis conditioned on age and identity
Face Aging by Conditional GAN

[Zhang et al., CVPR 2017]

Conditions

Age: 10

Synthesis
Previous Works

Kemelmacher, et al., CVPR2014

Shu, et al., ICCV2015

Wang, et al., CVPR2016
Main Idea --- Manifold Traversing
### Previous Works

- Group-wise learning
- Unidirectional transition
- Label required in testing

### Ours

- Joint learning
- Bidirectional transition
- No label in testing
Conditional Adversarial Autoencoder

$x$ $\rightarrow$ $E$ $\rightarrow$ Identity ($z$) $\rightarrow$ $G$ $\rightarrow$ $D$

Real/fake conditioned on age

Age label

9/20/2018
Conditional Adversarial Autoencoder (CAAE)

Encoder $E$

Input face

64x64x64
32x32x128
16x16x256
8x8x512

Conv_1 Conv_2 Conv_3 Conv_4 Reshape

FC_1 FC_2 FC_3 FC_4 Reshape

Deconv_1 Deconv_2 Deconv_3 Deconv_4

128x128x3

Discriminator on face image -- $D_{tmg}$

Input/output face

Conv_1 Conv_2 Conv_3 Conv_4 Reshape

128x128x3

Discriminator on z -- $D_z$

$z$ or $p(z)$

Prior distribution of $z$ (uniform)

1x1xn

Resize to 64x64x10

64x64x(n+16)

$z$ or $p(z)$

Generator $G$

128x128x64
64x64x128
32x32x256
16x16x512
8x8x1024

Conv_1 Conv_2 Conv_3 Conv_4 Reshape

FC_1

Deconv_1 Deconv_2 Deconv_3 Deconv_4

128x128x3

L2 loss

1x1x1024

Label

Resize to 64x64x10

64x64x(n+16)
Conditional Adversarial Autoencoder (CAAE)

Objective function

\[
\min_{E,G} \max_{D_z,D_{img}} \lambda \mathcal{L}(x, G(E(x), l)) + \gamma TV(G(E(x), l))
\]

\[
+ \mathbb{E}_{z^* \sim p(z)} [\log D_z(z^*)]
\]

\[
+ \mathbb{E}_{x \sim p_{data}(x)} [\log (1 - D_z(E(x)))]
\]

\[
+ \mathbb{E}_{x,l \sim p_{data}(x,l)} [\log D_{img}(x, l)]
\]

\[
+ \mathbb{E}_{x,l \sim p_{data}(x,l)} [\log (1 - D_{img}(G(E(x), l)))]
\]

where TV(\cdot) denotes the total variation which is effective in removing the ghosting artifacts. The coefficients \(\lambda\) and \(\gamma\) balance the smoothness and high resolution.
Conditional Adversarial Autoencoder (CAAE)

Effect of the Discriminator on $z$

With $D_z$

Without $D_z$
Conditional Adversarial Autoencoder (CAAE)

Effect of the Discriminator on image resolution

With $D_{\text{img}}$

Without $D_{\text{img}}$

GAN boosts the image resolution.
Experimental Evaluation

<table>
<thead>
<tr>
<th>Input</th>
<th>Others</th>
<th>Ours</th>
<th>0~5</th>
<th>6~10</th>
<th>11~15</th>
<th>16~20</th>
<th>21~30</th>
<th>31~40</th>
<th>41~50</th>
<th>51~60</th>
<th>61~70</th>
<th>71~80</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>31~40</td>
<td>31~40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>69~80</td>
<td>71~80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>16~20</td>
<td>16~20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>60~80</td>
<td>61~70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continuously bidirectional aging

Project page: [https://zzutk.github.io/Face-Aging-CAAE](https://zzutk.github.io/Face-Aging-CAAE)
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Drawback of Conditional GAN

Need to carefully balance GAN loss and reconstruction loss

Assume the weight of AE is 1, and the weight of GAN varies from 0 to 0.1

Too low $\rightarrow$ blurry

Too high $\rightarrow$ noisy
AE and GAN are tied in the parts of decoder/generator. Therefore, the reconstruction loss and adversarial loss interact/compete with each other, potentially causing unstable results.

The path of backpropagation is decoupled that avoids the interaction between the two losses, and thus relaxes the effort on balancing them.
Decoupled Learning --- AE//GAN

Input $I$ → Enc $\rightarrow$ Dec $\rightarrow$ Output $\hat{I}$

$L_1$ $\rightarrow$ Residual $I_G$ $\rightarrow$ Reconstruction $I_{ED}$

Real/Fake

Diagrams and images showing the process of decoupled learning with AE//GAN.
Experimental Evaluation

<table>
<thead>
<tr>
<th></th>
<th>0.001</th>
<th>0.01</th>
<th>0.1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE//GAN</td>
<td><img src="image1.jpg" alt="Images" /></td>
<td><img src="image2.jpg" alt="Images" /></td>
<td><img src="image3.jpg" alt="Images" /></td>
<td><img src="image4.jpg" alt="Images" /></td>
</tr>
<tr>
<td>AE+GAN</td>
<td><img src="image5.jpg" alt="Images" /></td>
<td><img src="image6.jpg" alt="Images" /></td>
<td><img src="image7.jpg" alt="Images" /></td>
<td><img src="image8.jpg" alt="Images" /></td>
</tr>
</tbody>
</table>
Evaluation Metric of GANs

• It is challenge to evaluate image quality/GAN performance
• There is no convincing metric for evaluating image quality

Inception score

- Generated Samples → Pre-trained Classifier → Clearly Classified?

Downsides of inception score:
• Heavily depend on the pre-trained classifier
• The results may vary with different datasets
• Absolut score misaligns actual performance
New Evaluation Metric of GANs

Idea: relative comparison of models

Which is better?

Project page: https://github.com/ZZUTK/Decoupled-Learning-Conditional-GAN
Conditional Image Synthesis --- Summary

- Conditional GAN outperforms previous works in image synthesis
- Decoupled learning could further stabilize the training of Conditional GAN
- The relative comparison is proposed for image quality evaluation.
Contents

1. Introduction to generative adversarial network (GAN)
2. Instability of GAN and Stabilization by Conditional GAN
3. Image Synthesis by Conditional GAN --- Face Aging
4. Further Stabilize Conditional GAN --- Decoupled Learning
5. Reference-Conditioned Super-Resolution
Reference-Conditioned Super-Resolution

**Conditions**

- LR
- Reference

**Synthesis**

- with Reference
- w/o Reference
Our Approach

1 Pre-SR

2 Feature Swapping

3 Ref-SR

VGG

Similar patches
Reference-Conditioned Super-Resolution

Bicubic (4x)  SRCNN  SRGAN  SRNTT (ours)  Reference

Project page: http://web.eecs.utk.edu/~zzhang61/project_page/SRNTT/SRNTT.html
What If A Bad Reference

Negative effect from the reference is introduced to the output
Our Approach

1. Pre-SR

2. Feature Swapping

3. Ref-SR

SR

VGG

Similar patches

Ref-SR

Weight map

9/20/2018
Effect of the Weight Map

The weight reduces the negative effect from the reference.
Reference-Conditioned Super-Resolution

Texture loss

\[ L_t = \frac{1}{4V^2} \| Gr(\phi(I^{SR}) \otimes M^s) - Gr(M^t) \|_F \]
Reference-Conditioned Super-Resolution

Texture loss:

$$\mathcal{L}_t = \frac{1}{4V^2} \left\| Gr(\phi(I^{SR}) \otimes M^s) - Gr(M^t) \right\|_F$$

$$s_{i,j} = \left\langle \frac{p_i^{LR}}{||p_i^{LR}||}, \frac{p_j^{LRef}}{||p_j^{LRef}||} \right\rangle$$
Reference-Conditioned Super-Resolution

Texture loss:

$$\mathcal{L}_t = \frac{1}{4V^2} \left\| Gr(\phi(I^{SR}) \otimes M^s) - Gr(M^t) \right\|_F$$
Reference-Conditioned Super-Resolution

Texture loss: \[ \mathcal{L}_t = \frac{1}{4V^2} \| Gr(\phi(I^{SR}) \otimes M^s) - Gr(M^t) \|_F \]
Data Collection --- CUFED5

There is no benchmark dataset for evaluating the reference-based SR methods

| HR | XH | H | M | L | XL |

Project page: [http://web.eecs.utk.edu/~zzhang61/project_page/SRNTT/SRNTT.html](http://web.eecs.utk.edu/~zzhang61/project_page/SRNTT/SRNTT.html)
Experimental Results
Experimental Results

- **Project page:** [http://web.eecs.utk.edu/~zzhang61/project_page/SRNTT/SRNTT.html](http://web.eecs.utk.edu/~zzhang61/project_page/SRNTT/SRNTT.html)

<table>
<thead>
<tr>
<th></th>
<th>XH</th>
<th>H</th>
<th>M</th>
<th>L</th>
<th>XL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>24.57</td>
<td>24.58</td>
<td>24.56</td>
<td>24.63</td>
<td>24.69</td>
</tr>
</tbody>
</table>
Summary of Contributions

1. Theoretically analyze GAN, i.e., drawbacks and improvements. GAN \rightarrow \text{conditional GAN} \rightarrow \text{decoupled learning}

2. Demonstrate the advantages of conditional GAN in image synthesis through the face aging task.

3. Extend conditional image synthesis to a traditional area, i.e., super-resolution, significantly boosting the visual quality.

4. Exploit relative comparison of GAN-based models, providing an alternative for image evaluation.

5. Extensive experimental evaluation is conducted to support the proposed designs.
Thank you