Image-to-Image Translation with Conditional Adversarial Networks

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros
Berkeley AI Research (BAIR) Laboratory, UC Berkeley
Agenda

- Motivating the problem
- Introduction to GANs
- Method
- Dataset & tasks
- Evaluation Metrics
- Results
- Strength & Weakness
- Future Extension
Introduction: Image-to-Image Translation

Previous Methods


Required loss functions and architectures designed specifically for the task at hand.
Motivation

Goal: A general-purpose solution to image-to-image translation problems.

This paper: Use the same architecture and objective for each image-to-image translation task.
Quick Overview of Generative Adversarial Networks
Generative Adversarial Networks (GANs)

\[ \mathcal{L}_{GAN}(G, D) = \mathbb{E}_y [\log D(y)] + \mathbb{E}_{x, z} [\log (1 - D(G(x, z)))]. \]
Conditional Generative Adversarial Networks (cGANs)

\[ \mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y} [\log D(x, y)] + \mathbb{E}_{x,z} [\log (1 - D(x, G(x, z)))], \]

Objective

\[ \mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))], \quad (1) \]

\[ \mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_{x,z}[\log(1 - D(G(x, z)))]. \quad (2) \]

\[ \mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]. \quad (3) \]

\[ G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G). \quad (4) \]
Generator

Encoder-Decoder Network

U-Net

Skip Connections

Discriminator

PatchGAN

Discriminator

PixelGAN

PatchGAN

ImageGAN
Optimization and Inference

- Alternate between one gradient descent step on D and one gradient descent step on G
- Minibatch SGD and Adam optimization
- At inference time, the generator net is run in exactly the same manner as during training phase

Datasets & Tasks

CityScape

CMP Facades

Map to Aerial
Datasets & Tasks

Edge to Photo  Sketch to Photo  Day to Light
Evaluation Metrics

---

AMT Perceptual Studies

Left or Right?
Evaluation Metrics

---

FCN-score

pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$

mean accuracy: $\left(1/n_{cl}\right) \sum_i n_{ii} / t_i$

mean IU: $\left(1/n_{cl}\right) \sum_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right)$
Experiment Results

Analysis of the Objective Function

<table>
<thead>
<tr>
<th>Loss</th>
<th>Per-pixel acc.</th>
<th>Per-class acc.</th>
<th>Class IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.42</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>GAN</td>
<td>0.22</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>cGAN</td>
<td>0.57</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>L1+GAN</td>
<td>0.64</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>L1+cGAN</td>
<td>0.66</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>Ground truth</td>
<td>0.80</td>
<td>0.26</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Experiment Results

Analysis of generator architecture
Experiment Results

Receptive Field of Discriminator

<table>
<thead>
<tr>
<th>Discriminator receptive field</th>
<th>Per-pixel acc.</th>
<th>Per-class acc.</th>
<th>Class IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1×1</td>
<td>0.39</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>16×16</td>
<td>0.65</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td>70×70</td>
<td>0.66</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>286×286</td>
<td>0.42</td>
<td>0.16</td>
<td>0.11</td>
</tr>
</tbody>
</table>
## Experiment Results

### Perceptual Validation

<table>
<thead>
<tr>
<th>Loss</th>
<th>Photo → Map % Turkers labeled real</th>
<th>Map → Photo % Turkers labeled real</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>2.8% ± 1.0%</td>
<td>0.8% ± 0.3%</td>
</tr>
<tr>
<td>L1+cGAN</td>
<td>6.1% ± 1.3%</td>
<td>18.9% ± 2.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>% Turkers labeled real</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 regression from [58]</td>
<td>16.3% ± 2.4%</td>
</tr>
<tr>
<td>Zhang et al. 2016 [58]</td>
<td>27.8% ± 2.7%</td>
</tr>
<tr>
<td>Ours</td>
<td>22.5% ± 1.6%</td>
</tr>
</tbody>
</table>
Experiment Results

---

Rich Community Driven Results

Edges2Cats  
Background Masking
Strength & Weaknesses

---

**Strength:**

1. An easily generalizable structure to conduct image-to-image translation tasks.
2. A good combination of existing techniques and well-designed loss function to ensure the generation of high-quality synthesized image.

**Weakness:**

1. Required large number of 1-to-1 paired images for training, which are expensive to collect.
2. Does not provide decent comparisons between this network with the other task-specific model to defend its claim for good generalizability.
Future Extension

---

1. Modify the structure such that it can learn the translation mapping with unpaired images in two different domains.

2. Further improvement to generate photorealistic image.