High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

By Clayton Barham
Contents

• Motivation
• Pix2pix Baseline
• Improving Photorealism and Resolution
• Using Instance Maps
• Learning an Instance Level Feature Embedding
• Results
• Interactive Object Editing
Motivation

• Synthesize high resolution, photorealistic images from semantic label maps using conditional GANs
  • Pix2pix – Low resolution
  • Perceptual loss – Lacking detail, repeated patterns

• Allow objects in the image to be added, removed, or edited
pix2pix Baseline

(a) Labels

(b) pix2pix

Blurry Quality
Visual Artifacts
Improving Photorealism and Resolution: Coarse to Fine Generator
Improving Photorealism and Resolution: Multi-Scale Discriminators
Improving Photorealism and Resolution: Improved Adversarial Loss and Perceptual Loss

\[ \mathbb{E}_{(s,x)}[\log D(s, x)] + \mathbb{E}_s[\log(1 - D(s, G(s)))] \]

GAN Loss

\[ \mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(s,x)} \sum_{i=1}^{T} \frac{1}{N_i} \left[ \| D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s)) \|_1 \right] \]

Feature Matching Loss

\[ \lambda \sum_{i=1}^{N} \frac{1}{M_i} \left[ \| F^{(i)}(x) - F^{(i)}(G(s)) \|_1 \right] \]

Perceptual Loss

\[ \min_G \left( \left( \max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G, D_k) \right) + \lambda \sum_{k=1,2,3} \mathcal{L}_{FM}(G, D_k) \right) \]
Using Instance Maps
Learning an Instance Level Feature Embedding
Learning an Instance Level Feature Embedding

• Encoder-Decoder network learns feature values

• Average Pooling layer standardizes feature values for each instance

• Feature vectors concatenated with other input to generator
Learning an Instance Level Feature Embedding
Learning an Instance Level Feature Embedding

- Perform K-Means clustering on per class basis
- Each cluster corresponds to one ‘style’
- Make centroid of each cluster exemplar of that style
- Substitute centroid values to edit object styles
Results: Quantitative Comparison
Results: Quantitative Comparisons

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel acc</td>
<td>78.34</td>
<td>70.55</td>
<td>83.78</td>
<td>84.29</td>
</tr>
<tr>
<td>Mean IoU</td>
<td>0.3948</td>
<td>0.3483</td>
<td>0.6389</td>
<td>0.6857</td>
</tr>
</tbody>
</table>
Results: Perceptual Study

pix2pix  CRN  Coarse to Fine
Results: Human Perceptual Study – Unlimited Time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>93.8%</td>
<td>86.2%</td>
</tr>
<tr>
<td>Ours (w/o VGG)</td>
<td>94.6%</td>
<td>85.2%</td>
</tr>
</tbody>
</table>
Results: Human Perceptual Study – Limited Time
### Results: Generator Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel acc (%)</td>
<td>77.86</td>
<td>78.96</td>
<td>83.78</td>
</tr>
<tr>
<td>Mean IoU</td>
<td>0.3905</td>
<td>0.3994</td>
<td>0.6389</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Our generator</td>
<td>80.0%</td>
<td>76.6%</td>
</tr>
</tbody>
</table>
Results: Discriminator Comparison

<table>
<thead>
<tr>
<th></th>
<th>single D</th>
<th>multi-scale Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel acc (%)</td>
<td>82.87</td>
<td>83.78</td>
</tr>
<tr>
<td>Mean IoU</td>
<td>0.5775</td>
<td>0.6389</td>
</tr>
</tbody>
</table>
Interactive Object Editing

(a) Synthesized result

(b) Application: Change label types

(c) Application: Edit object appearance
Thank you