Shapes and Context: In-the-Wild Image Synthesis & Manipulation

Authors: Aayush Bansal, Yaser Sheikh, Deva Ramanan, Carnegie Mellon University

Shan Lyu, Minqiang Hu, Weijia Xing
Problem Statement

- Input: semantic label input masks
- Goal: image synthesis through a data-driven approach
Related Work: Parametric Methods

- **Parametric Machine Learning Models:**
  Deep neural networks trained with adversarial losses and perceptual losses

- **Training datasets:** one dataset with a specific data distribution (cityscapes, faces, facades)
Weaknesses:

(1) poor generalization because of dataset bias

(2) one-to-one mapping for outputs
Related Work: Non-Parametric Methods

- Nonparametric Nearest Neighbors Methods:
  
  Find the most similar ones with the nearest distances
Weaknesses of Previous Work:

Global shape fitting for rigid and non-deformable objects
Weaknesses of Previous Work:

Global shape fitting for rigid and non-deformable objects
Strengths for *Shape and Context*

- Not limited to specific training data distributions
- Multiple candidate outputs for one-to-many mappings
- High quality images with parts and pixels fitting
Method
Hierarchical Composition

Get a synthesized image from a semantic and an instance label map in a hierarchical filtering methods.

- Global scene contextual filtering for the training dataset-COCO
- Instance shape matching
- Local part consistency
- Pixel-to-pixel consistency
Hierarchical matching for image synthesis

https://www.youtube.com/watch?v=CdHzrIP7Bo&feature=youtu.be
Generating Various output

Multiple images output by matching different shapes

https://www.youtube.com/watch?v=CdHizrlP7Bo&feature=youtu.be
Generating Various output

Multiple images output by matching different shapes

https://www.youtube.com/watch?v=CdHzrlP7Bo&feature=youtu.be
Generating Various output

Multiple images output by matching different shapes.
Global Scene Context

Different from the classical non-parametric method, this method prune the list of training datasets

Step 1: To check the semantic and instance labels from the training images

- To check the set of labels
Global Scene Context

Step2:

- **Global coverage**: to ensure the label map in training dataset has similar distribution of the input label mask
  - Get normalized histogram of label distribution of both
  - Compute the L2 distance
- **Pixel coverage**: to ensure the selected images have maximum pixel-to-pixel overlap
  - Resize to 100*100
  - It cares about the location while global coverage does not

Consequence: This reduces the search space from hundred thousand images to a few hundreds.
Instance shape matching

Each shape is represented by $x_1, x_2, \ldots, x_n$. $L$ is the number of unique labels.

Use a bounding box for a shape $x_i$ as a rectangular convolutional filter $w_i$ to retrieve similar shapes from the training data.

Logical operator: the part of a shape $x_i$ in the filter $w_i$ is set to 1, the remaining part is set to -1.

Contextual operator: traverse in the unit of pixels to check the consistency of the context.
Contextual operator: traverse in the unit of pixels to check the consistency of the context.

With the context matching, the contextual information will also be retrieve.
Use a bounding box for a shape $x_i$ as a rectangular convolutional filter ($w_i$) to retrieve similar shapes from the training data.

**Logical operator**: the part of a shape ($x_i$) in the filter ($w_i$) is set to 1, the remaining part is set to -1.

**Contextual operator**:

- $N_s$: number of pixels (50*50).
- $I$: indicator function.

\[
S_{\text{shape}}(w_i, w_j) = w_i^l * w_j^l + \sum_{k=1}^{N_s} I(w_{i,k} - w_{j,k})
\]
Ignore the shape for the output if the ratio of their aspect-ratio to that of query component is either less than 0.5 or greater than 2.

The shape based matching will fail if the dataset is limited (no matching shape in the library).
Part Consistency

Why is part consistency necessary?- when global shape is not found, the local part can help

One of the advantage against the the classical non-parametric method:

More friendly for the matching for insufficient shape data and non-rigid (deformable) objects.
How is local part consistency checked?

- From the top-k global shapes, resize it 256*256, 16*16 patches.
- No longer need to look at large window size if weakly aligned global shapes

Resized to 16*16 patches in 256*256
Scoring function for part consistency:

- \( N_p = 256 \times 9 \)
- Each patch: \( w_{i,k}^p \)

\[ S_{part}(w_i^p, w_j^p) = \sum_{k=1}^{N_p} I(w_{i,k}^p - w_{j,k}^p) \]
Pixel-to-pixel Matching

Pixel consistency can fill some minor holes in the shapes even after the shape, and part consistency being applied.
Pixel-to-pixel Matching

Method: Similar to part-consistency algorithms
- Conduct on every pixel
- Every pixel is segmented into a 11*11 surrounding window
- Then look in surrounding 5*5 regions from a low-res 128*128 pixel input map
Experiment
Experiment

Prior parametric approach:

- Pix2Pix
- Pix2Pix-HD

Training details:

- 1 month on a single GPU
Experiment

Training datasets: COCO

- Including 134 different objects and stuff categories.
- Restrict ourselves to COCO training data.

Other scenario: Cityscapes
Result - User Intervention & manipulation

- original label
- synthesized output
- add shape
- modified label
- new RGB component
- manipulated output
Score we use to evaluate

- FID
- Mask-RCNN
FID
Mask-RCNN
## Results

### Figure 1: FID Scores on COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>#examples</th>
<th>Oracle</th>
<th>FID score (256×256)</th>
<th>FID score (64×64)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pix2Pix [33]</td>
<td>1</td>
<td>x</td>
<td>70.43</td>
<td>41.45</td>
</tr>
<tr>
<td>Pix2Pix-HD [55]</td>
<td>1</td>
<td>x</td>
<td>157.13</td>
<td>109.49</td>
</tr>
<tr>
<td>Ours (shapes)</td>
<td>1</td>
<td>x</td>
<td>37.26</td>
<td>23.22</td>
</tr>
<tr>
<td>Ours (shapes+parts)</td>
<td>1</td>
<td>x</td>
<td>32.62</td>
<td>18.02</td>
</tr>
<tr>
<td>Ours (shapes+parts+pixels)</td>
<td>1</td>
<td>x</td>
<td><strong>31.63</strong></td>
<td><strong>16.61</strong></td>
</tr>
</tbody>
</table>

### Figure 2: Mask-RCNN Scores on COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>#examples</th>
<th>Oracle</th>
<th>PC</th>
<th>AC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix2Pix [33]</td>
<td>1</td>
<td>x</td>
<td>17.9</td>
<td>8.9</td>
<td>4.9</td>
</tr>
<tr>
<td>Non-Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>1</td>
<td>x</td>
<td>44.5</td>
<td>31.0</td>
<td>20.9</td>
</tr>
<tr>
<td>Ours</td>
<td>5</td>
<td>✓</td>
<td>58.2</td>
<td>41.2</td>
<td>31.4</td>
</tr>
</tbody>
</table>
Results

Ask Volunteers to evaluate outputs from our methods and comparing methods

- 51.2%  Our method
- 7.8%   Pix2Pix
- 41%    None of method looks real image
## Results on Cityscapes

<table>
<thead>
<tr>
<th>Method</th>
<th>#examples</th>
<th>Oracle</th>
<th>PC</th>
<th>AC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parametric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix2Pix [33]</td>
<td>1</td>
<td>✗</td>
<td>72.5</td>
<td>29.5</td>
<td>24.6</td>
</tr>
<tr>
<td>CRN [10]</td>
<td>1</td>
<td>✗</td>
<td>49.0</td>
<td>22.5</td>
<td>18.2</td>
</tr>
<tr>
<td>Pix2Pix-HD [55]</td>
<td>1</td>
<td>✗</td>
<td>79.0</td>
<td>43.3</td>
<td>37.8</td>
</tr>
<tr>
<td><strong>Semi-Parametric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMS [44]</td>
<td>1</td>
<td>✗</td>
<td>68.6</td>
<td>35.1</td>
<td>28.1</td>
</tr>
<tr>
<td><strong>Non-Parametric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours (top-25)</td>
<td>1</td>
<td>✗</td>
<td>67.1</td>
<td>38.0</td>
<td>30.5</td>
</tr>
<tr>
<td>Ours (top-25)</td>
<td>5</td>
<td>✓</td>
<td>71.3</td>
<td>39.6</td>
<td>32.4</td>
</tr>
</tbody>
</table>

Figure 3: PSP-Net Scores on Cityscapes
Results from Pix2Pix method

(a). constrained vs. in-the-wild data distribution

(b). simple examples with varying foreground/background
Strengths

Non-parametric approach

- Generate large number of outputs by varying shapes and parts.
- User-Controllable (short demo)
- Not limited to specific training data
Weakness

- Small boundary on objects
- Return top k images while k has influence on result
- Unable to reflect object relationship between different objects
- Highly depends on what we have in the database
Future Work and Applications

- Further work to make the output image more realistic
  - More dataset
  - Address smarter ways of combining shapes and parts information
- Explore in-the-wild video synthesis and manipulation.
Questions & Comments