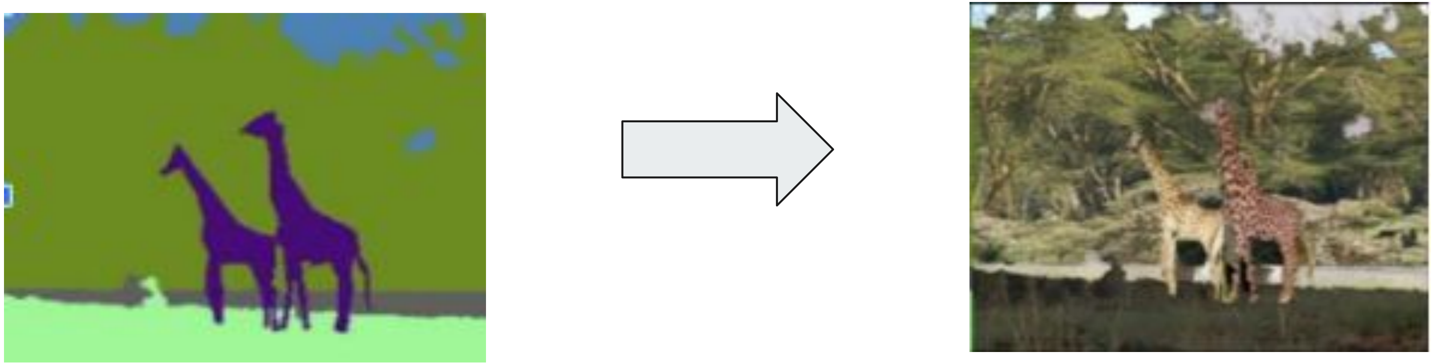
Shapes and Context: In-the-Wild Image Synthesis & Manipulation

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

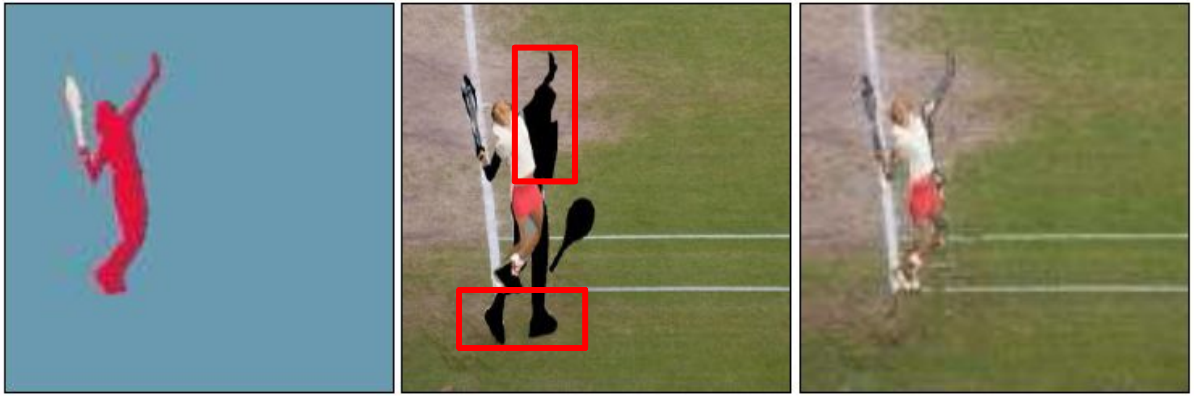
Related Work: Non-Parametric Methods

- 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



input ours (global shapes) ours (full)

Weaknesses of Previous Work:

Global shape fitting for rigid and non-deformable objects

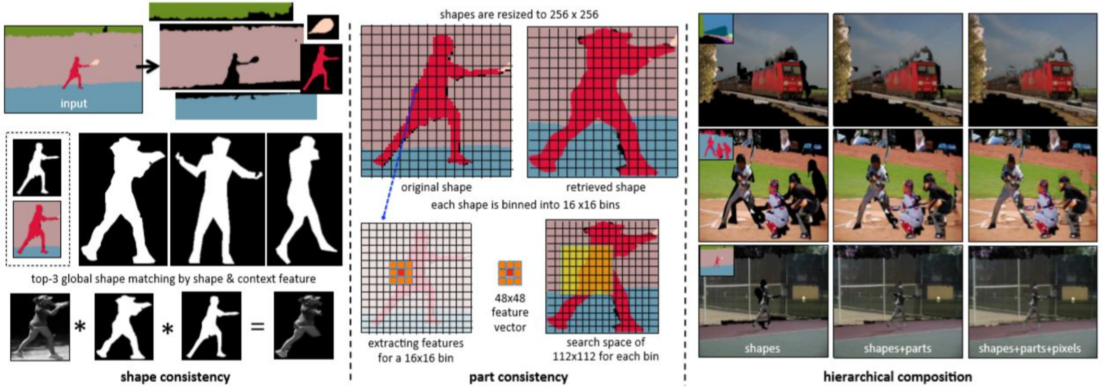
input | ours (global shapes) | ours (full)

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

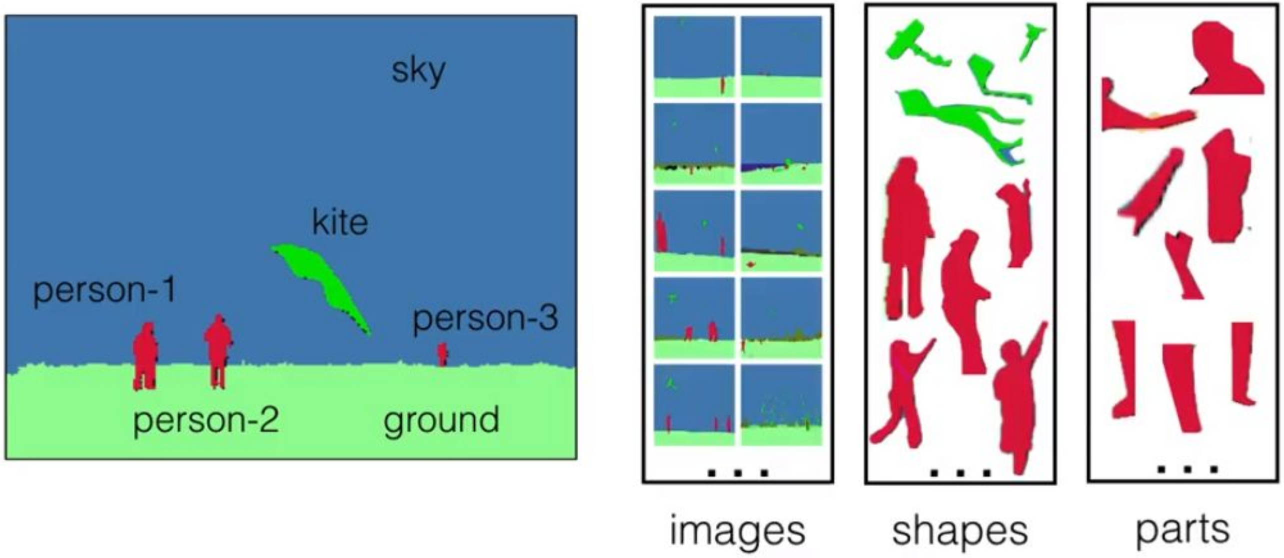
Method

Hierarchical Composition



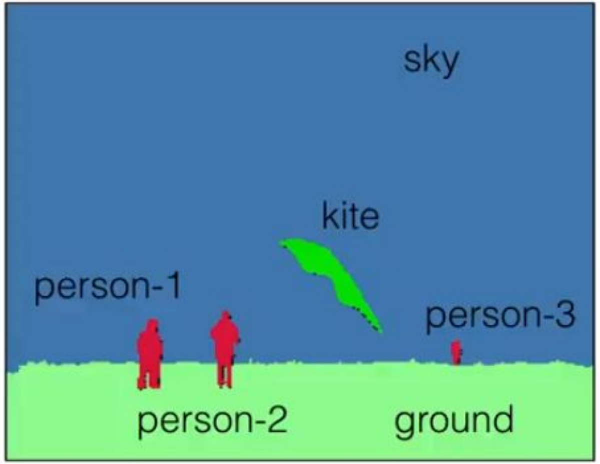
Get a synthesized image from a semantic and an instance label map in a hierarchical filtering methods.

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



Hierarchical matching for image synthesis

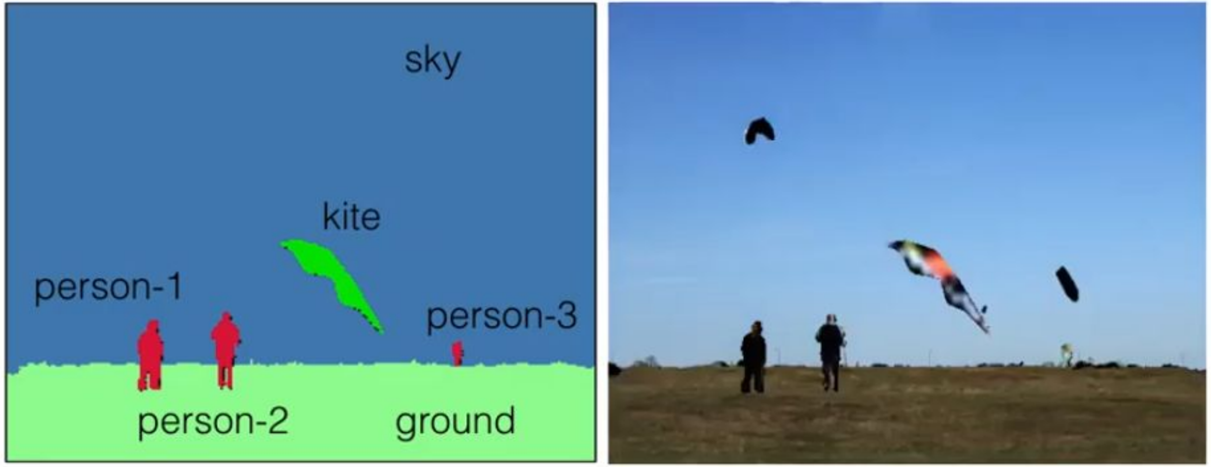
https://www.youtube.com/watch?v=CdHIzrlP7Bo&feature=youtu.be



Generating Various output

Multiple images output by matching different shapes

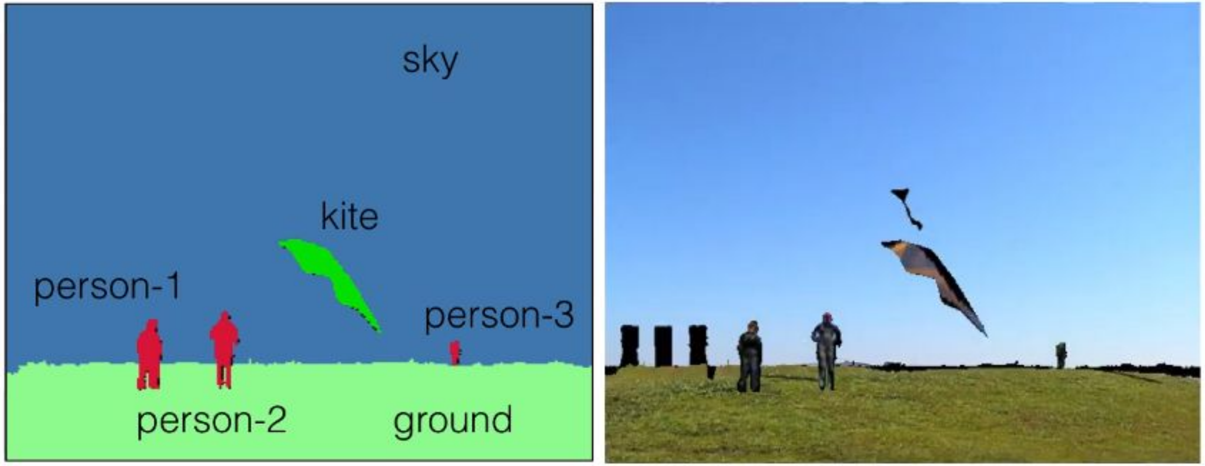
https://www.youtube.com/watch?v=CdHIzrlP7Bo&feature=youtu.be



Generating Various output

Multiple images output by matching different shapes

https://www.youtube.com/watch?v=CdHIzrlP7Bo&feature=youtu.be



Generating Various output

Multiple images output by matching different shapes https://www.youtube.com/watch?v=CdHIzrlP7Bo&feature=youtu.be

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**

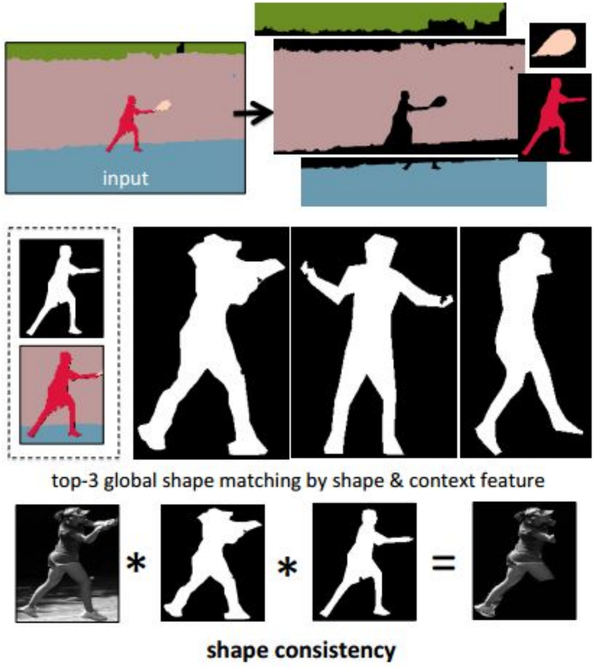
Global Scene Context

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
    - Computer 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 x1, x2, … , xn. L is the number of unique labels.

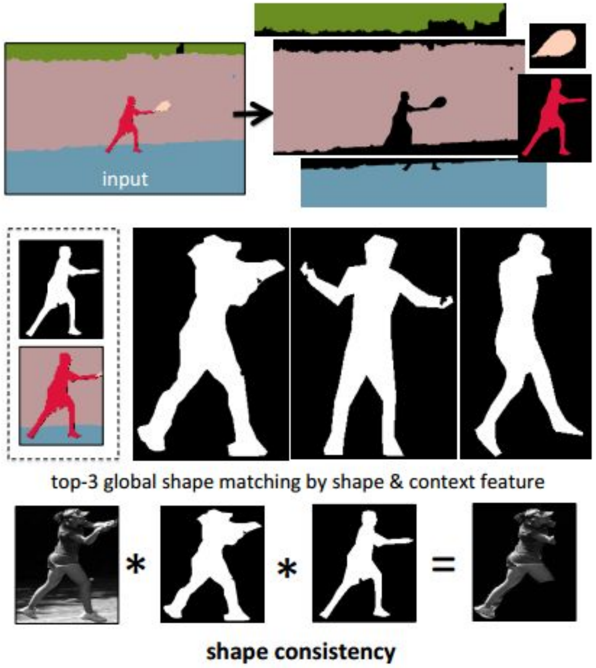
Use a bounding box for a shape xi as a rectangular convolutional filter(wi) to to retrieve similar shapes from the training data.

Logical operator:the part of a shape (xi) in the filter (wi) 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.

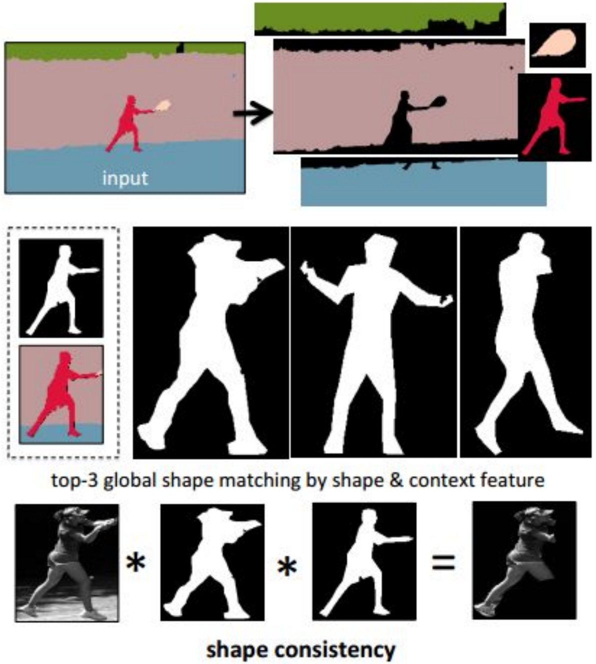


Use a bounding box for a shape xi as a rectangular convolutional filter(wi) to to retrieve similar shapes from the training data.
Logical operator:the part of a shape (xi) in the filter (wi) is set to 1, the remaining part is set to -1.

Contextual operator:

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

$$S_{shape}(w_i, w_j) = w_i^l * w_j^l + \sum_{k=1}^{N_s} I(w_{i,k}^c - w_{j,k}^c),$$



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

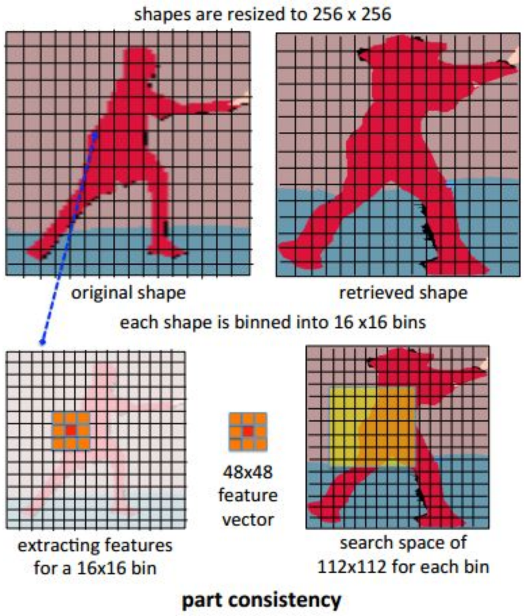
Why is part consistency necessary?- when global shape is not found, the local part can help

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

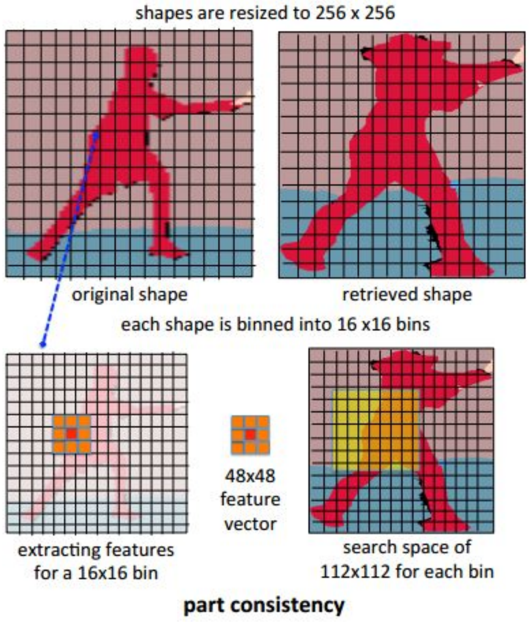
More friendly for the matching for insufficient shape data and non-rigid (deformable) objects.

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



$$S_{part}(w_i^p, w_j^p) = \sum_{k=1}^{N_p} I(w_{i,k}^p - w_{j,k}^p)$$

Scoring function for part consistency:

- Np: 256*9
- Each patch: $w_{i,k}^p$

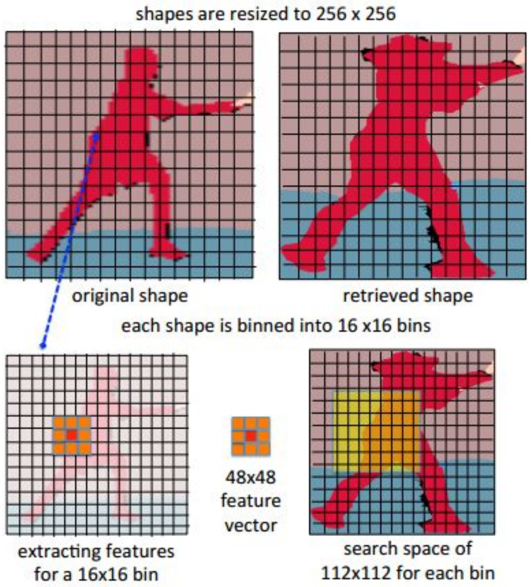
Pixel-to-pixel Matching

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



part consistency

Experiment

Experiment

Prior parametric approach:

- Pix2Pix
- Pix2Pix-HD

Training details:

- 1 month on a single GPU

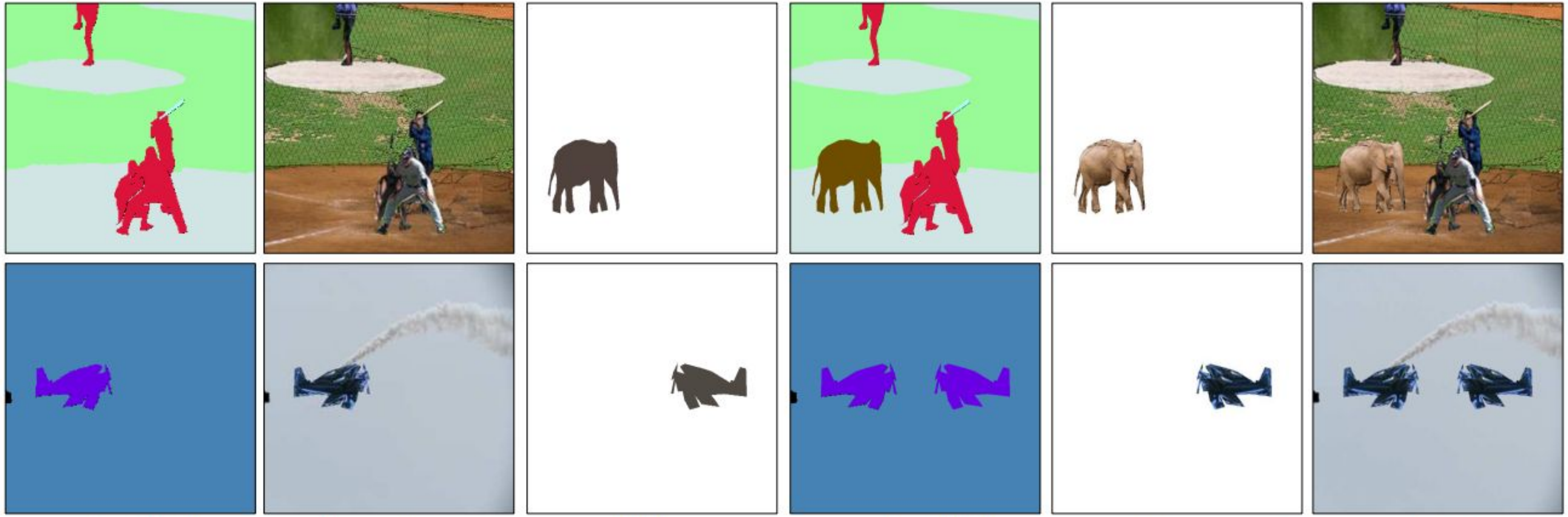
Experiment

Training datasets: COCO

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

Other scenario: Cityscapes

Result-User Intervention & manipulation

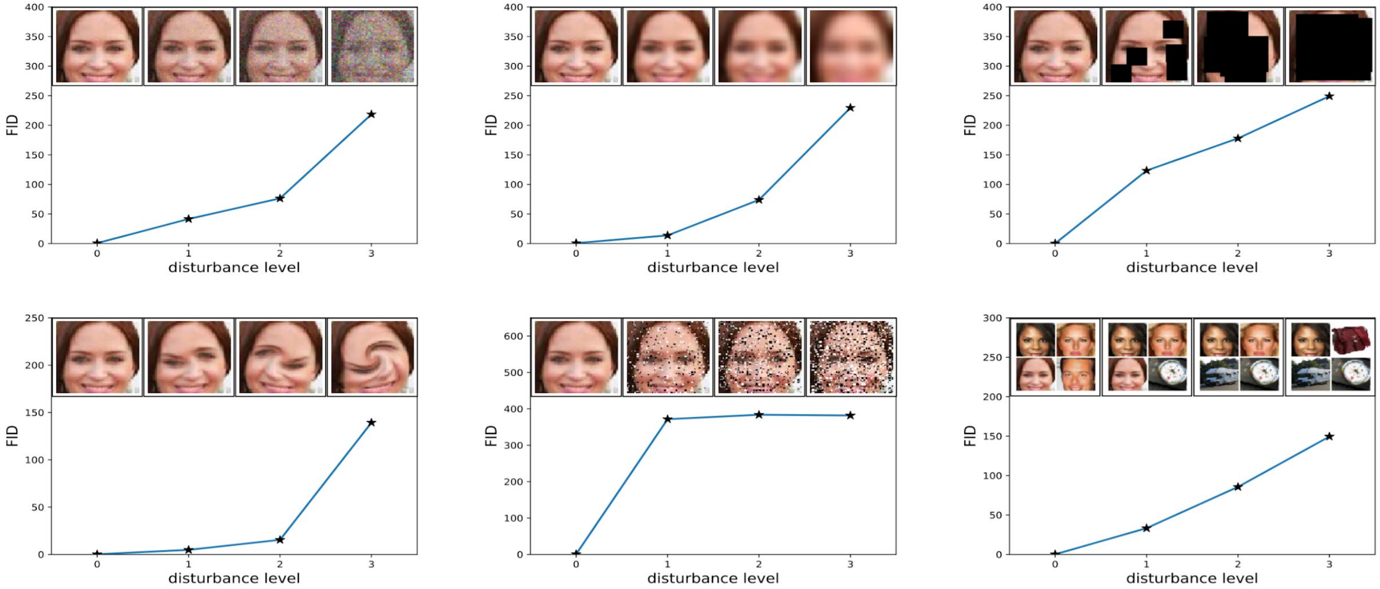


original label | synthesized output | add shape | modified label | new RGB component | manipulated output

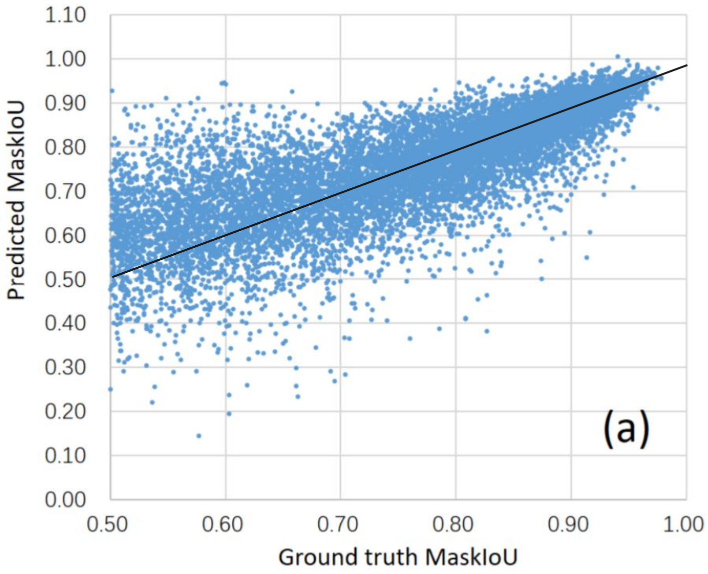
Score we use to evaluate

- FID
- Mask-RCNN

FID


FID
disturbance level
0
1
2
3
0
50
100
150
200
250
300
350
400
600
500

Mask-RCNN



1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Predicted MaskIoU
0.50
0.60
0.70
0.80
0.90
1.00
Ground truth MaskIoU
(a)

Results

Method	#examples	Oracle	FID score (256×256)	FID score (64×64)
Pix2Pix [33]	1	✗	70.43	41.45
Pix2Pix-HD [55]	1	✗	157.13	109.49
Ours (shapes)	1	✗	37.26	23.22
Ours (shapes+parts)	1	✗	32.62	18.02
Ours (shapes+parts+pixels)	1	✗	**31.63**	**16.61**

Figure 1: FID Scores on COCO

Method	#examples	Oracle	PC	AC	IoU
Parametric					
Pix2Pix [33]	1	✗	17.9	8.9	4.9
Non-Parametric					
Ours	1	✗	44.5	31.0	20.9
Ours	5	✓	58.2	41.2	31.4

Figure 2: Mask-RCNN Scores on COCO

Results

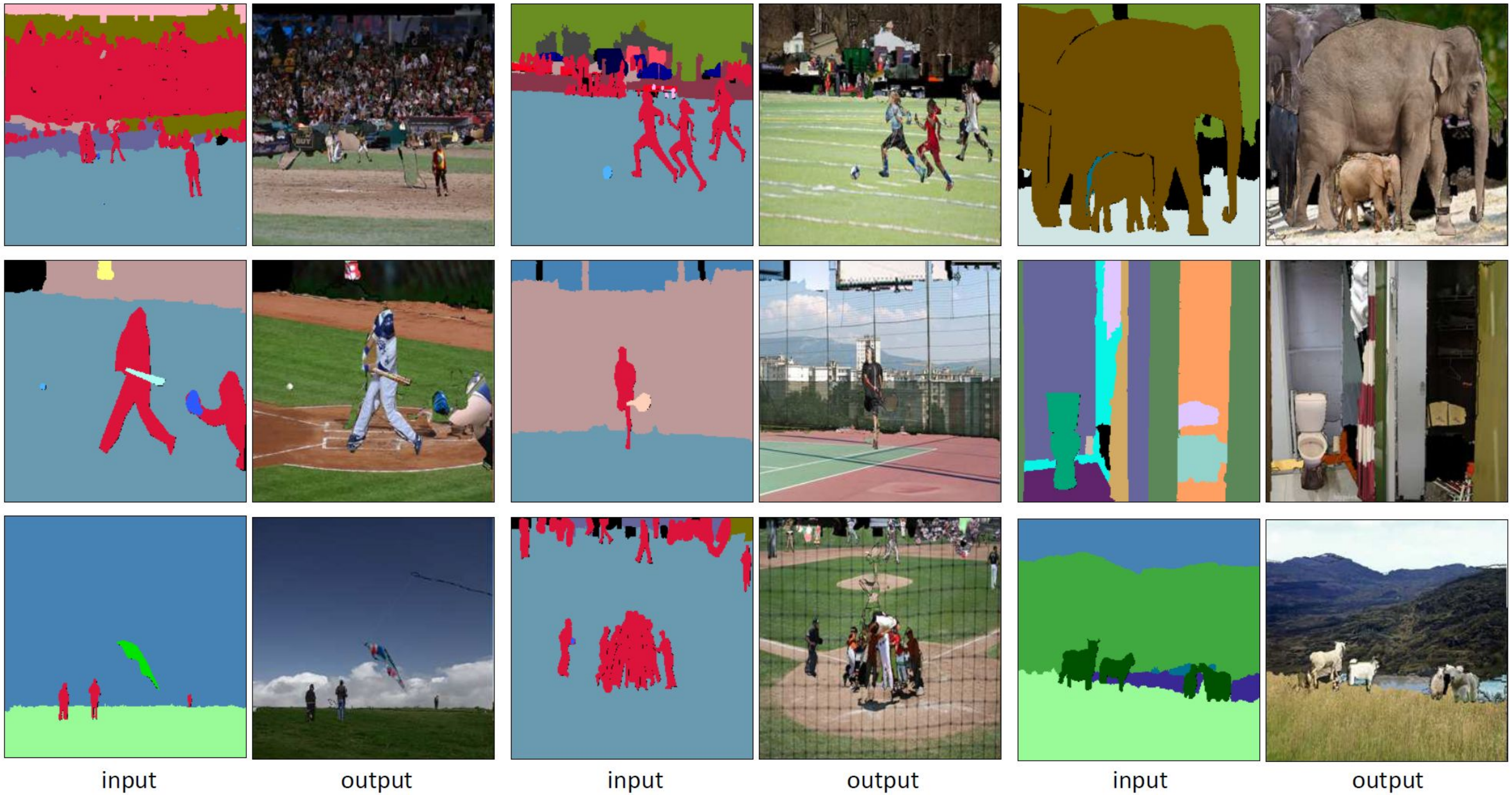
Ask Volunteers to evaluate outputs from our methods and comparing methods

- 51.2% Our method
- 7.8% Pix2Pix
- 41% None of method looks real image

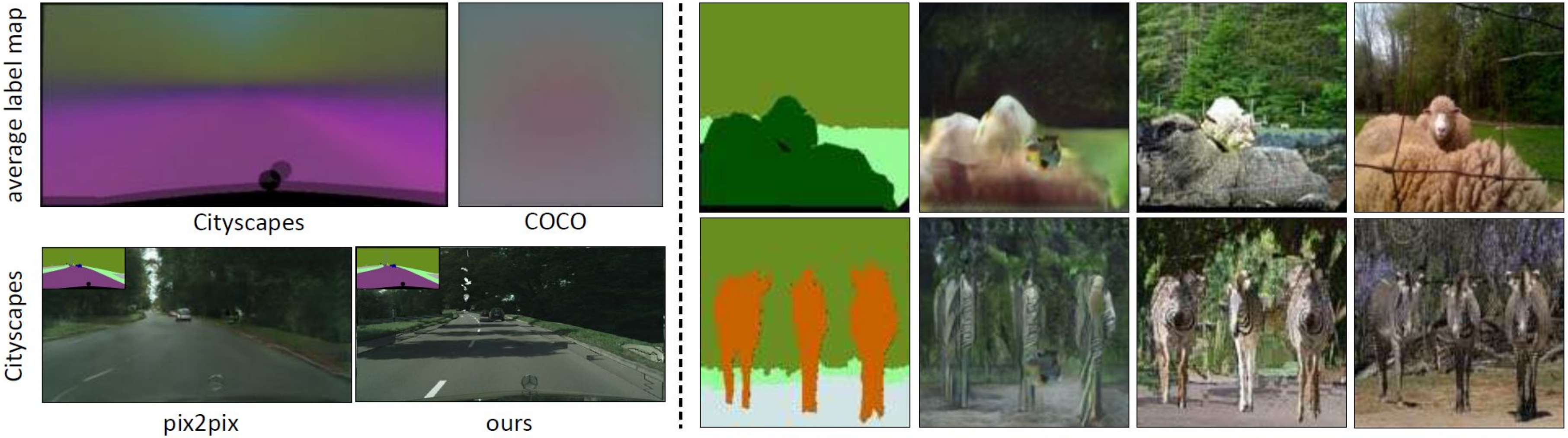
Results on Cityscapes

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

Figure 3: PSP-Net Scores on Cityscapes


input
output
input
output
input
output

Results from Pix2Pix method



an example of label map similar to average label map

(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)
- http://www.cs.cmu.edu/~aayushb/OpenShapes/
- 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