IJCAI 2018

Annotation-Free and One-Shot Learning for Instance Segmentation of Homogeneous Object Clusters

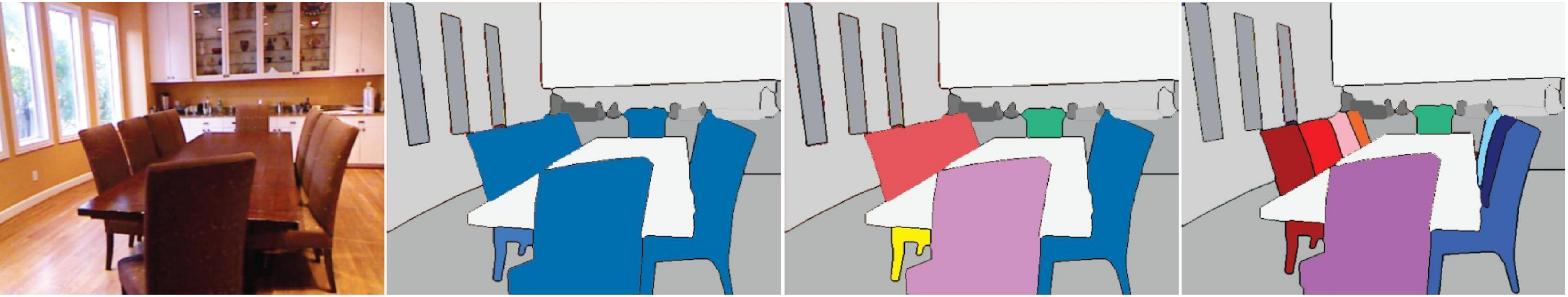
51184506047
邢琛聪

Preliminary - Homogeneous Object Clusters

"Homogeneous object clusters (HOC) are ubiquitous. From microscopic cells to gigantic galaxies, they tend to cluster together. "

Preliminary - Instance Segmentation

实例分割与语义分割的联系和区别



Input Image | Semantic Segmentation | Boundary Segmentation | Semantic Instance Segmentation

Motivation

Directly applying current best performing instance segmentation methods(Mask-RCNN) ,many of which are based on some kind of Deep Convolutional Neural Network (DCNN), meets a bottleneck:

unaffordable annotation cost. All of these segmentation models require a large number of annotated images for training purpose.

Contribution

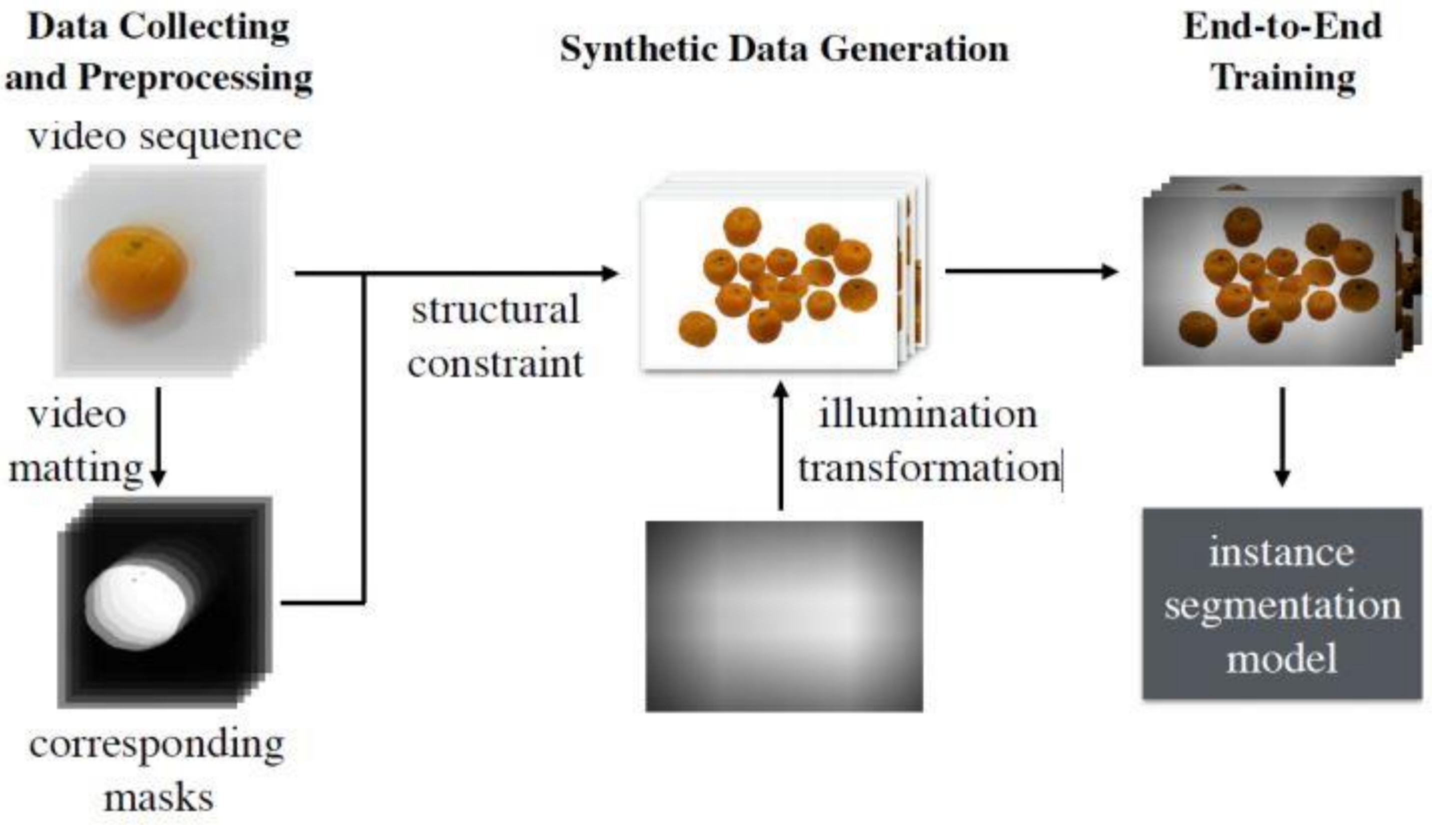
1. propose an efficient framework for instance segmentation of HOCs. Our proposed method significantly reduces the cost of data collection and labeling.

2. propose an efficient method to generate realistic synthetic training data which significantly improves instance segmentation performance.

3. build a dataset consisting of HOC images. The dataset is used to evaluate our proposed method. The dataset and codes will be published with the paper.

System Pipeline

1) System takes single-object video as input, extracts the mask for each frame
2) generates synthetic images of homogeneous objects in cluster
3) use the synthetic images to train an instance segmentation model.



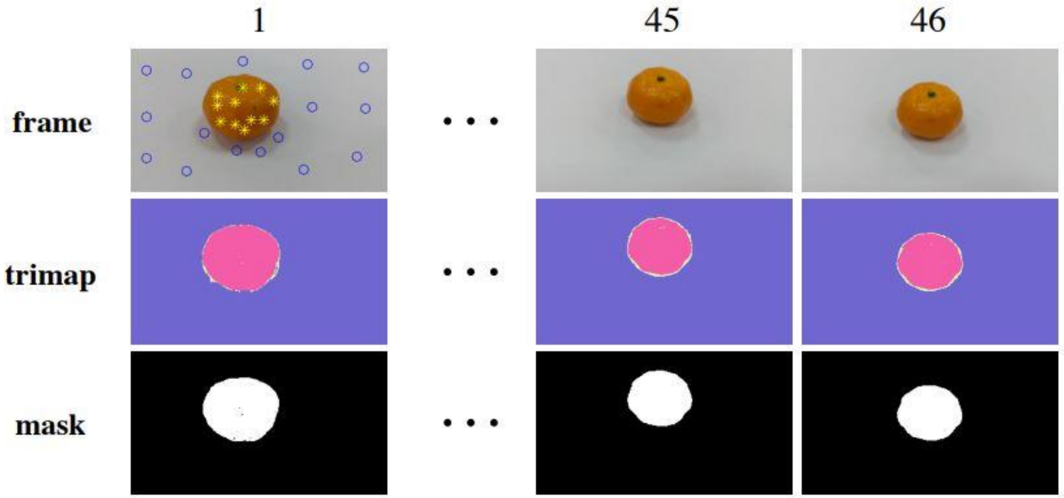
Methods(I) - Data Collection and Preprocessing

Suppose we want to generate a HOC dataset about oranges.

One-Shot Video Collection

put one single orange at the center of a contrastive background, and take a video of the orange at different angles and positions. Such video typically lasts for about 20 seconds

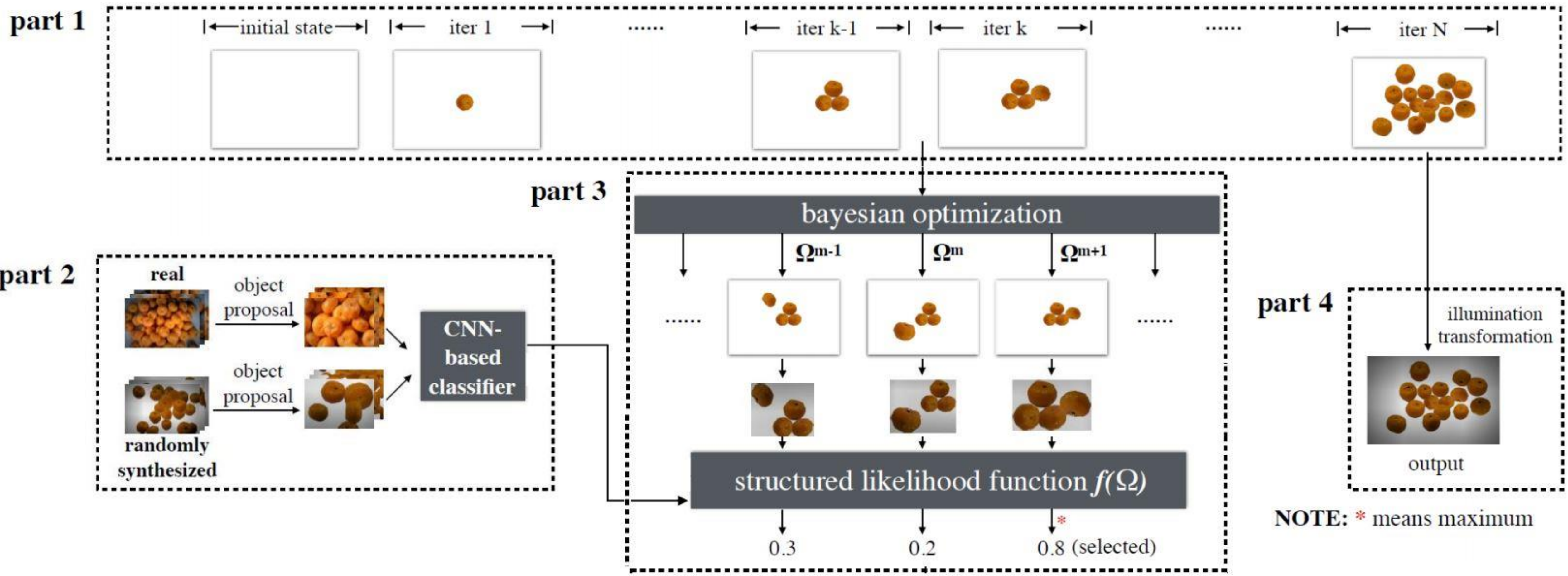
Video Matting



For the first frame, seeds of foreground and background are automatically sampled based on color and location priors. Trimaps are interpolated across the video volume using optical flow. Matting technique uses the flowed trimaps to yield high-quality masks of the moving orange

Methods(II) – Synthetic Data Generation

Overall Algorithm



Methods(II) – Synthetic Data Generation

Structural Constraint(Structured Likelihood Function)

$$I_k = g(I_{k-1}, O_k, \Omega_k)$$

$$\overline{\Omega_k} = \arg\max_{\Omega_k \in D} P(g(I_{k-1}, O_k, \Omega_k))$$

$$= \arg\max_{\Omega_k \in D} f(\Omega_k; I_{k-1}, O_k)$$

Given N objects $\{O_1, O_2, \dots, O_N\}$

I_k denote the image contain k objects

P(·) denote the likelihood of being a real image

$\Omega = \{\theta, \gamma, x, y\}$ θ denote the rotate, γ denote the resize factor, **x, y** denote the coordinates of center O_k in I_{k-1}

Methods(II) – Synthetic Data Generation

Bayesian optimization framework

Since $f(\Omega_k)$ is a black-box function， we cannot optimize f by computing its derivative， Here, we use bayesian optimization to optimize our continuous objective function $f(\Omega_k)$

Bayesian optimization behaves in an iterative manner. At iteration m, given the observed point set $D_{m-1} = \{\left(\Omega_k^1, f\left(\Omega_k^1\right)\right), \ldots (\Omega_k^{m-1}, f\left(\Omega_k^{m-1}\right))\}$ we model a posterior function distribution by the observed points. Then, the **acquisition function** (i.e., a utility function constructed from the model posterior) is maximized to determine where to sample the next point $\left(\Omega_k^{m-1}, f\left(\Omega_k^{m-1}\right)\right)$. $\left(\Omega_k^{m-1}, f\left(\Omega_k^{m-1}\right)\right)$is collected and the process is repeated. Iteration ends when m = M (M is a user-defined parameter)

Methods(III) – Illumination Transformation

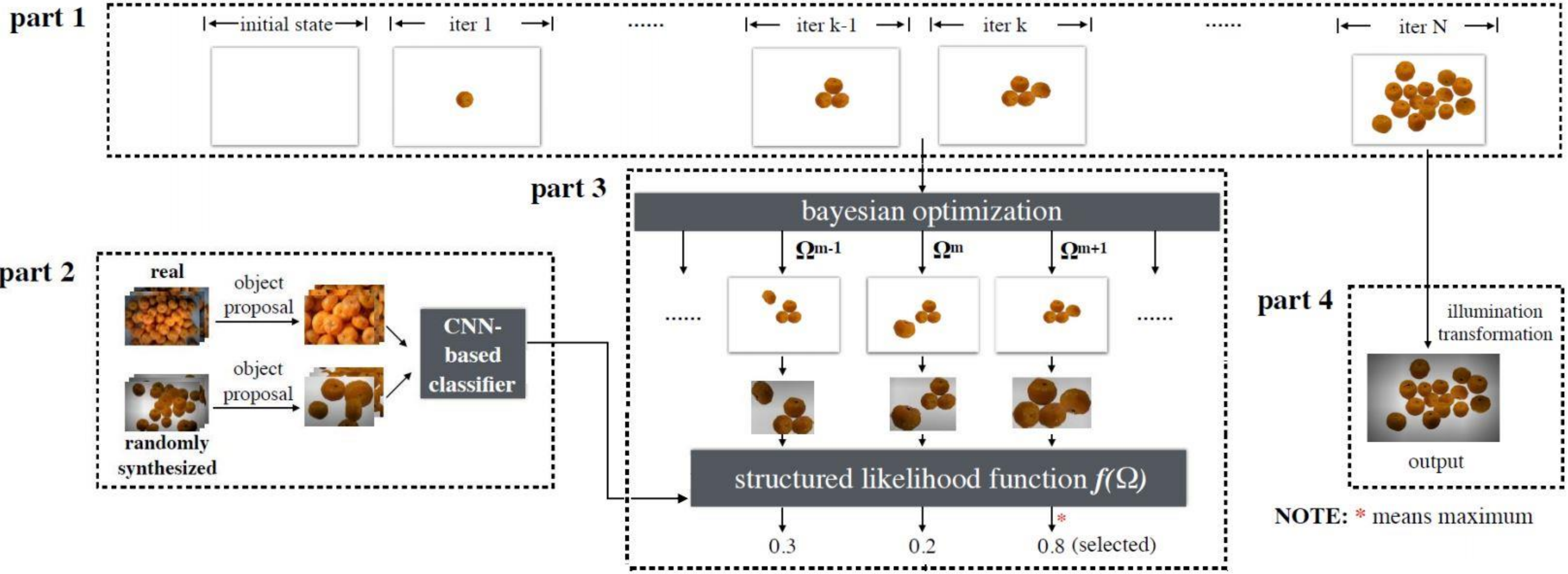
We first convert both the synthetic image and the real image from RGB color space to HSV space, where V represents illumination information. Then, we implement detail removing, by using a large kernel Gaussian smoothing on both images to model general illumination condition

$$V_{syn} = V_{syn} - \text{mean}(V_{syn}) + blur(V_{real})$$
$$V_{real} = V_{real} - \text{mean}(V_{real}) + blur(V_{real})$$

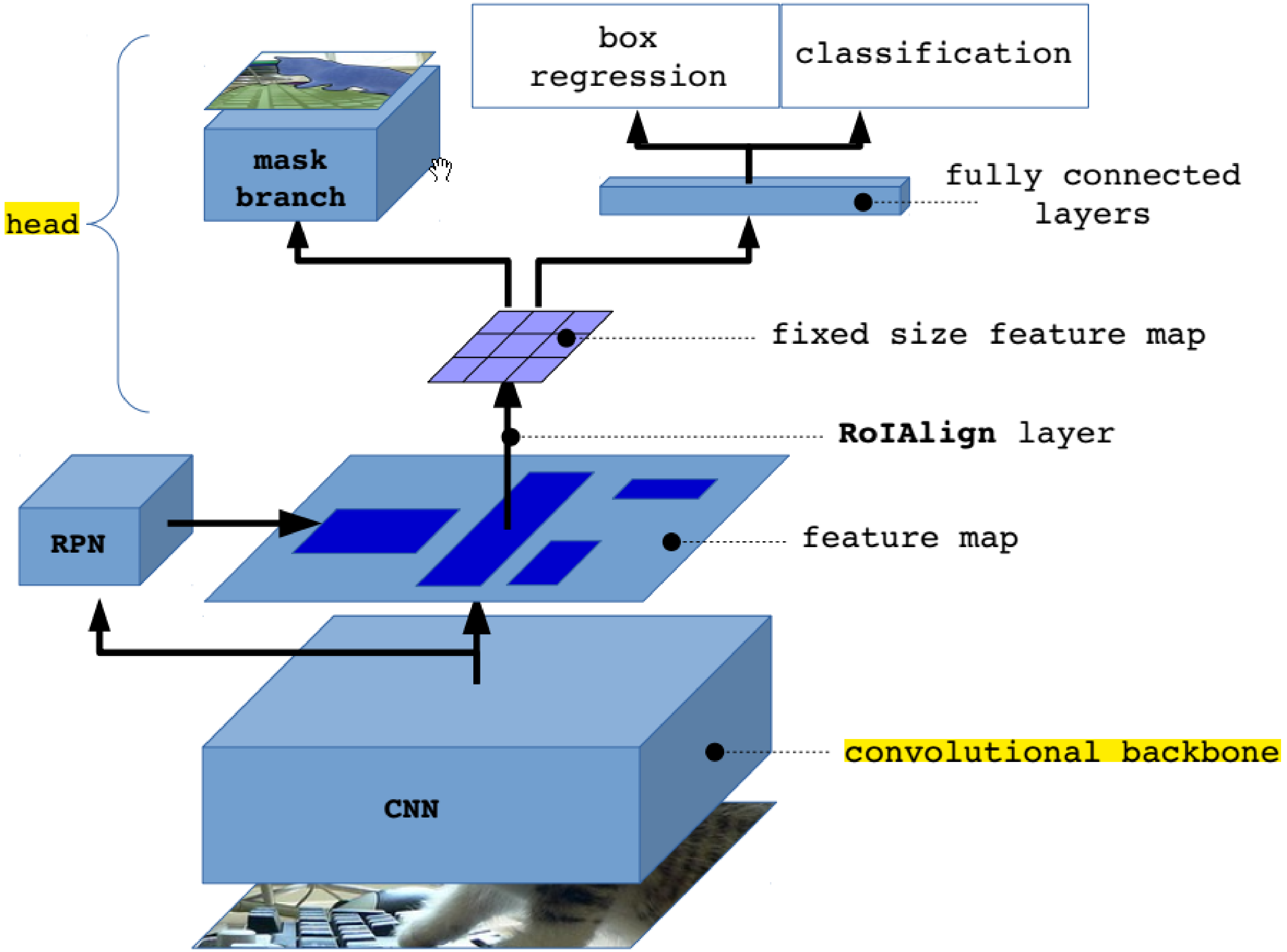
(a) (b) (c) (d)

Methods(II) – Synthetic Data Generation

Why Not Use GAN?



Methods(IV) – End To End Training



This paper adapt Mask RCNN to handle segmentation task.

Using synthetic data to train the model.

Using real images to do the test.

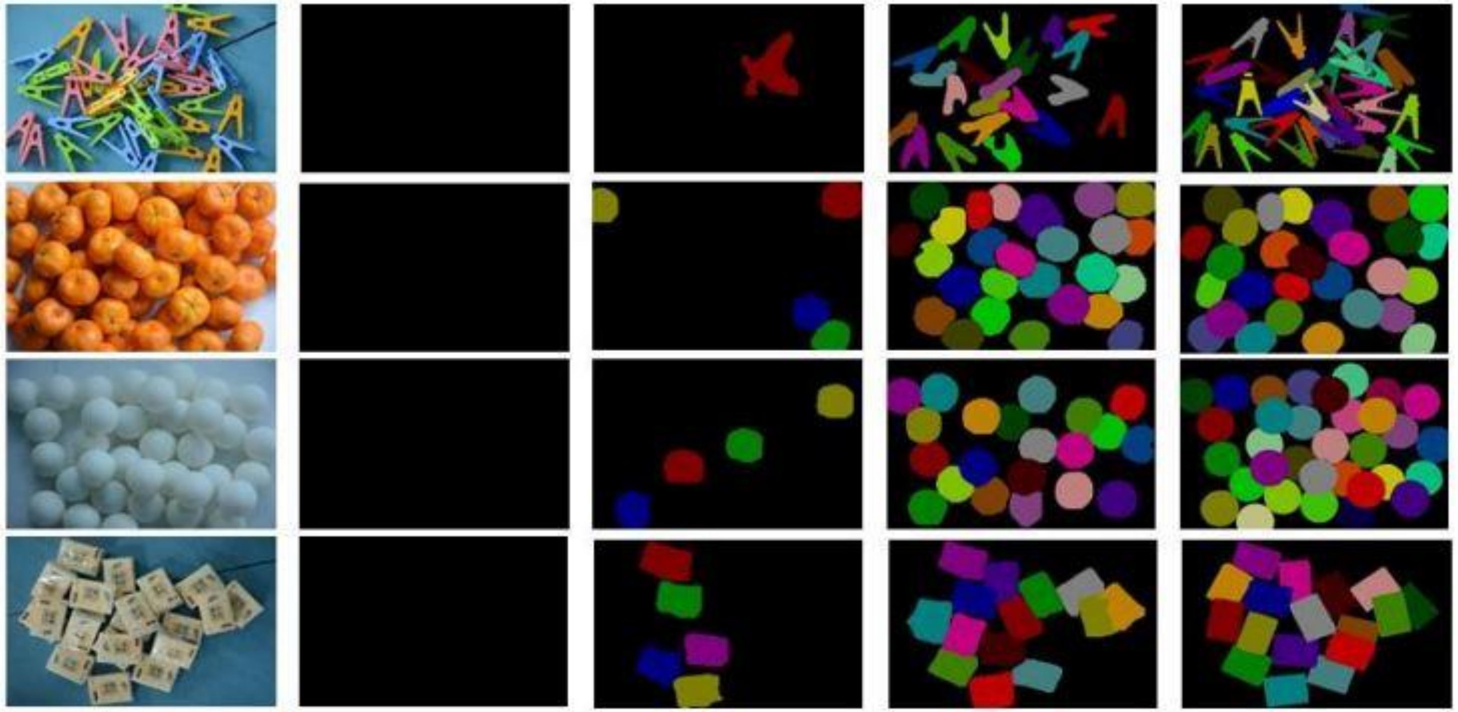


Image Single Random Ours Groundtruth

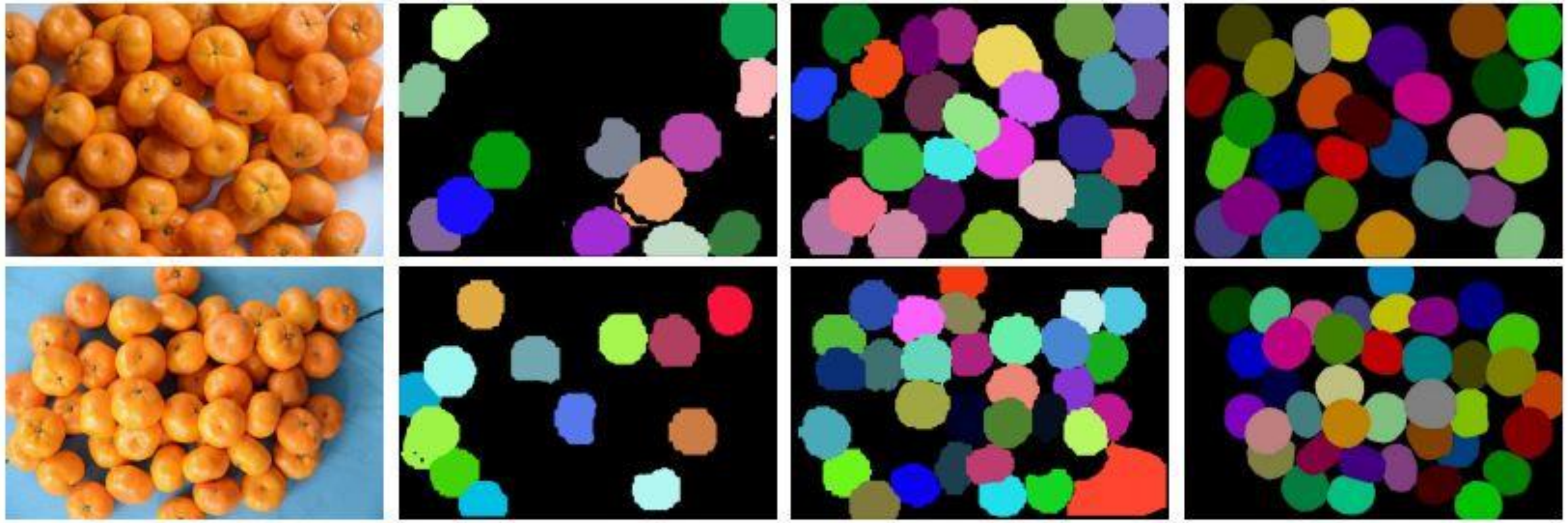


Image COCO pretrained Ours Groundtruth

Experimental Evaluation

dataset has 3, 669 instances in total, each image has 18.3 instances on average

Table 1: Results on mAP^r @0.5 on our dataset. All numbers are percentages %.

	badminton	battery	clothespin	grape	milk	hexagon nut	orange	ping pong	tissue	wing nut	mAP
Single	12.1	23.2	4.4	19.3	8.2	17.5	17.6	14.2	13.1	21.1	15.1
Random	40.6	50.9	38.4	50.8	26.7	52.9	63.8	67.2	83.9	32.9	50.8
Random+illumination	44.6	48.7	34.3	41.6	26.0	46.3	54.9	64.2	68.9	39.4	46.9
Random+structure	34.2	39.3	52.6	**72.7**	31.3	62.8	**90.3**	**81.7**	**90.7**	23.7	57.9
Ours	**53.0**	**69.5**	**67.7**	72.5	**52.6**	**73.6**	90.0	81.2	90.4	**48.4**	**69.9**

Table 2: Results on mAP^r @[0.5:0.95] on our dataset. All numbers are percentages %.

	badminton	battery	clothespin	grape	milk	hexagon nut	orange	ping pong	tissue	wing nut	mAP
Single	8.7	21.2	2.0	16.8	5.0	15.2	16.0	11.8	9.9	18.8	12.5
Random	33.1	44.7	29.3	44.8	20.9	43.4	56.3	59.5	68.4	27.7	42.8
Random+illumination	36.5	43.2	28.4	37.2	20.8	38.3	50.4	56.4	56.8	35.5	40.4
Random+structure	29.8	36.5	36.7	**66.5**	22.0	45.2	83.8	**76.4**	**80.4**	20.7	49.8
Ours	**44.2**	**60.3**	**46.2**	65.4	**41.3**	**54.8**	**84.0**	75.6	**80.4**	**39.3**	**59.2**

Experimental Evaluation