*The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018*

# Instance Segmentation of Visible and Occluded Regions for Finding and Picking Target from a Pile of Objects

Kentaro Wada  Shingo Kitagawa  Kei Okada  Masayuki Inaba

University of Tokyo, JSK Laboratory

We present a robotic system of picking target from a pile of objects that is capable to find and pick the target object by removing obstacles away in the appropriate order. The key idea to achieve this is segmenting instances regarding both visible and occluded masks, which we call 'instance occlusion segmentation' to find which objects are occluding the target object. To achieve this, we extend existing instance segmentation model with a novel 'relook' architecture, in which the model explicitly learns the inter-instance relationship. With extension to existing image synthesis, we also make the system to be capable to handle novel objects without human annotations, in consideration of the future applications like warehouse picking. The experimental results show the effectiveness of the relook architecture compared with the conventional model and image synthesis compared with the human annotations for instance occlusion segmentation. We also demonstrate the capability of our picking system for picking a target in a cluttered environment.

## 1. Introduction

With recent progress in deep learning especially convolutional neural network, the robotics community has been improved the ability of the robot to find and pick various target objects in clutter including some novel objects [Jonschkowski 16, Zeng 18]. However, these works restrict the environment with less occluded target objects, or tackle the problem by picking objects in an indiscriminate order and recognizing its label afterward. Our goal in this work is to develop a framework that enables robot to pick various target objects in the appropriate order in an environment with heavy occlusion (e.g. a pile of objects in a bin).

Picking target objects from a pile (ex. Fig.1) is difficult especially with the variety of shapes and flexibility of objects. Estimating the environmental state of stacked objects is impractical and usually impossible because of unavailability of the mesh model and infinite possible patterns of stacking even with a perfect detection of visible part of objects. In order to plan the picking order correctly in this situation, it is necessary to understand the scene with the occlusion relationship among objects. And this motivates us to introduce *instance occlusion segmentation*, in which the occluded region of each object is segmented as well as the visible one.

In this paper, we propose a vision system that detects object instances and recognizes their occlusion status simultaneously. For the application to various objects, we have designed the system to be able to handle novel objects without gathering any task-specific training data for them. To achieve this, our system consists of two components: 1) image synthesis of a stack of objects only with the instance images of these objects, in which various pattern of stacking and occlusion status is generated with ground truth of visible and occluded region mask for each object instance; 2) instance segmentation model with a novel 'relook' architecture designed for occlusion segmentation, in which we extend the recent works [He 17, Do 17] to recognize and use the density of objects to segment multi-class: visible and occluded, for each instance.

Contact: Kentaro Wada. Graduation School of Information Science and Technology, The University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. +81-3-5841-7416, wada@jsk.imi.i.u-tokyo.ac.jp.

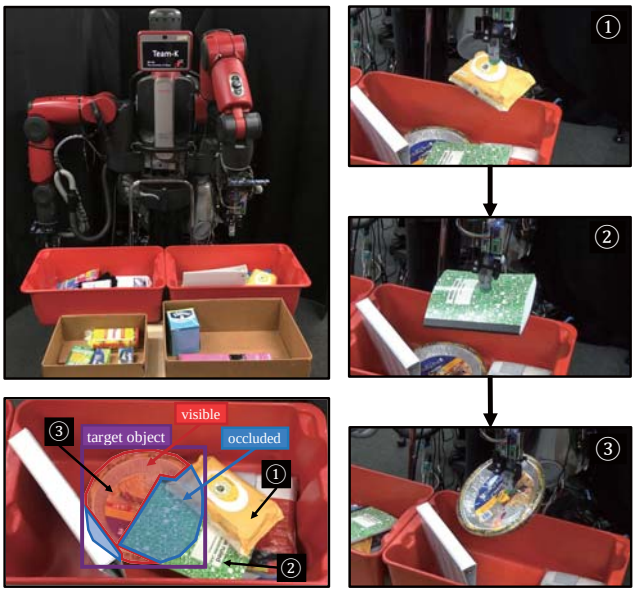



Fig. 1: **Our system.** It recognizes occlusion status of objects via instance occlusion segmentation, and plans an appropriate picking order to pick the target object from the stacked objects.

We also propose a metric for instance occlusion segmentation; it is the extension of instance segmentation which only segments the visible mask. In the experiments, we provide evaluation results of instance occlusion segmentation using various objects that are used in the Amazon Robotics Challenge (ARC), and demonstrates the ability of our system to find and pick the target from a pile of objects.

## 2. Related works

### Instance segmentation

Instance segmentation is aimed at predicting object region mask and its label at the same time. Since instance segmentation is a compound task of bounding box detection and pixel-wise semantic segmentation, previous works propose models that solve the 2 tasks *sequentially* or *concurrently*. In the *sequential* approach, past works [Pinheiro 16, Dai 16] propose models which learn to propose mask segmentation first and classifies them afterward. On the other

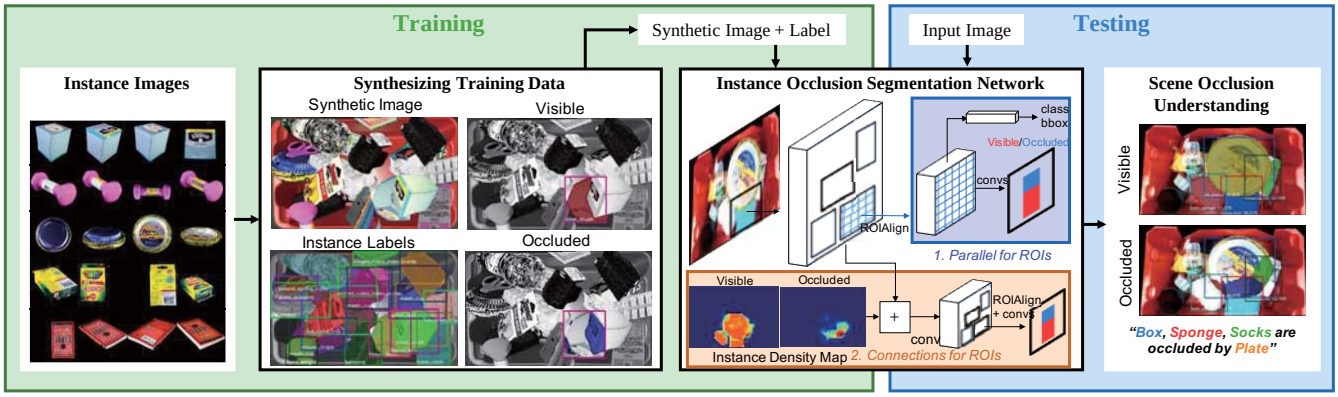



Fig. 2: **System overview.**
At the train time, the system receives instance images as its input and trains instance occlusion segmentation model.
At the test time, the system receives real image as the input and outputs scene occlusion status for the image.

hand, recently *concurrent* prediction of segmentation proposals and object detection is proposed [He 17, Do 17]. In these models, object classes, boxes, and masks are simultaneously predicted, and it turns out that these concurrent approaches are effective for not only speed but accuracy.

In this paper, we extend the cutting-edge work [He 17, Do 17], for instance occlusion segmentation by seeing it as the multi-class (visible and occluded) extension of the conventional instance segmentation (visible only). Compared to the visible region segmentation, the relationship between nearby objects is more important in occlusion segmentation. This motivates us to extend the previous models to have a connection between predicted object instances, while previous models concurrently predict visible region masks for each box proposals and there is no connection among them.

### Image synthesis for object detection

Recent successes in machine vision tasks are mainly based on deep learning, but since deep learning requires a large amount of data, it naturally motivates researchers to acquire training data from synthesizing rather than using human annotations. A naive approach of synthesizing training data is using 3D mesh models to render on existing 2D image. Past works use mesh models to learn viewpoint estimation [Su 15], and object detection [Hinterstoisser 17] for objects on 2D image. They have been working on how can it possible to make synthetic image near to the real images. On the other hand, recently it has been found that synthesizing only with 2D instance images of objects is also effective to train object detection models [Georgakis 17, Dwibedi 17]. The base idea for this is if we could generate infinite synthetic images at random and train learning model by it, the model would generalize to the real images. Our approach is closer to the latter and we extend the past works to generate ground truth of object masks (visible and occluded) not only the bounding box. Since it is impractical to gather realistic mesh models for various objects, 2D synthesizing is more practical than 3D one. We also show that the number of instance images can be small to achieve detection performance using human annotations in the experiments, while the past works [Georgakis 17, Dwibedi 17] use a huge number of instance images for each object.

## 3. Methods

### 3.1 System overview

Our proposed system (Fig.2) has two components:

- Instance occlusion segmentation neural networks trained using the generated images;
- 2D image synthesis of cluttered scene generated from object instance images to handle various objects.

At the training time, the system receives instance images of objects as the input to train instance occlusion segmentation model from the generated synthetic image. At the testing time, a real image is the input and occlusion status of the scene is predicted and returned. Since our proposed system only requires instance images of objects of interest, we can easily gather them from web or with a standard camera; this is important for the vision system to handle various objects without human labeling which is especially hard for instance occlusion segmentation, and for the applications to warehouse picking in e-commerce services for example.

### 3.2 Instance occlusion segmentation networks

#### Mask R-CNN

Mask R-CNN is composed of three branches of convolutional layers after the feature extraction and its ROI (region of interest) transformation (ROIAlign): classification, bounding box regression and mask segmentation. It is fast because the prediction of each instance information is done in *parallel* and *independently* using the feature extracted by ROIAlign for each proposed ROI. However, since the prediction about each instance after ROI feature extraction is independent of other ROIs/instances, this motivates us to seek another architecture which can learn the relationship or *dependency* among ROIs/instances. Mask R-CNN is proposed as a model for instance segmentation of the visible region, in which the relationship among instances may not be so important, but for instance occlusion segmentation, the relationship is more necessary because there is overlap among occluded masks of objects while there is no overlap among visible masks. This leads us to introduce the inter-instance connection in following.

#### Relook architecture: inter-instance connection

In order to learn relationship and dependency between instances, it is necessary to have connections among the representations of each instance. To do this, we convert instance masks predicted in the first stage to a density map and use it to predict instance masks in the second stage (middle of Fig.2). The two instance masks of first and second stage are added in pixel-wise (fused) as the final result and segmentation loss is computed for the fused result. The second stage can be interpreted as a "*relook*" architecture for learning and predicting inter-instance connection, so we call this

model as Mask R-CNN (relook).

The first stage is Mask R-CNN (softmax) (§2.), and the final layer predicts 3 masks: visible, occluded and the other, so the density map is also generated for each of the three. For learning inter-instance connection, we concatenate the density map with the feature extracted by the feature extractor; we use ResNet50-C4 and ResNet101-C4 (fourth stage output of ResNetX [Nair 10]) as in Mask R-CNN [He 17], and apply a convolutional layer. After that, ROIAlign, 'res5' and a deconvolutional layer is applied using shared parameter with the first stage. On top of that, the final convolutional layer is applied to predict 3 masks for each instance. We use ReLU [Nair 10] as the activation function for the hidden layers, and the kernel size of convolutional layer is 3 for hidden layer and 1 for the output layer.

### 3.3 Image synthesis for learning instance occlusion segmentation

**Instance image gathering**

The background of the instance image (left in the Fig.2 is static, black cloth, however, we found that it is difficult to extract foreground mask for each instance especially of object who has various colors (e.g. Robot Book) or transparent (e.g. Wine Glass) characteristics. To overcome this difficulty, the previous work [Dwibedi 17] trained a convolutional neural network (ConvNet) model [Long 15] using the mask acquired from thresholding of the depth image as the ground truth. And we applied the same approach using the mask acquired from pixel value thresholding using instance images of 112 objects.

**Image synthesis of stacked objects**

We generate the ground truth labels and masks in addition to the synthetic image of stacked objects. The left figures in Fig.2 show an example of that and visualizes how the ground truth looks like. While stacking the instance images onto the background image, we apply color and geometric data augmentation for each instance image at random. The foreground mask of each instance image is acquired by ConvNet and since the filled region of already stacked instances are known while synthesizing, we can get masks of both visible and occluded region as shown in the left of Fig.2.

### 3.4 A metric for evaluation of instance occlusion segmentation

Recently, a metric, Panoptic Quality ($PQ$), to evaluate the accuracy of both detection and segmentation is proposed in [Kirillov 18] to evaluate instance segmentation, and we decided to extend this metric to multi-class segmentation. $PQ = DQ \cdot SQ$ is represented as the multiplication of Detection Quality ($DQ = |TP|/(|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|)$) and Segmentation Quality ($SQ = \frac{\sum_{(p,g)\in TP} IoU(p,q)}{|TP|}$): where $TP$, $FP$ and $FN$ represent the set of true positive, false positive and false negative instances, and $p$, $g$ represent predicted and ground truth instances. $PQ$ is computed for each object class and the $PQ$s for all classes are averaged as "*mean of PQ*" ($mPQ$). For computation of $DQ$, we use the visible mask of predicted and ground truth to find the matched instances with IoU threshold of 0.5 between these predicted masks. Since the $SQ$ was proposed as the metric of instance segmentation, which is a single mask segmentation for each object, we extended the metric to be able to evaluate multi masks segmentation for each instance as $SQ_{multi} = \frac{\sum_{(p,g)\in TP} mIoU(p,q)}{|TP|}$, where $mIoU(p,q) = \sum_{m\in M} IoU_m(p,q)/|M|$. and $M$ represents the set of possible instance masks, of which we have three in the instance occlusion segmentation: background, visible and occluded. $IoU$ is the intersect over union of predicted and ground truth masks for a single mask-class, and the $IoU$s for each mask-class is averaged as $mIoU$. where $p$ and $g$ is the predicted and ground truth masks for a single mask-class.

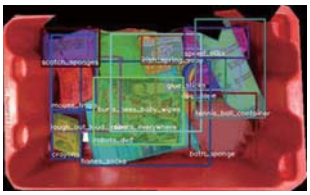
## 4. Experiments

**Objects for evaluation**

We evaluate our system with the 40 objects used in the Amazon Robotics Challenge 2017 (ARC2017). whose 4-6 instance images were distributed at the competition. We believe these objects has a large diversity and appropriate to demonstrate the picking from a pile of objects because they were selected for the worldwide warehouse picking competition, Amazon Robotics Challenge.

**Human-annotated dataset for evaluation**

For evaluation of both the model and image synthesis, we created a dataset of instance occlusion segmentation shown in Fig.3. Since occlusion mask is too difficult even for human to annotate, we created a set of sequential camera frames in which a pile of objects are cleared from top. With annotating the visible mask of objects in all frames of a video captured from a fixed camera, the visible masks are backtraced to acquire the occluded masks. We created 21 videos (splitted to train:test = 14:7) in which the 40 objects appear in 7 times (3-5 times in the train split).



(a) Visible Masks that includes Baby Wipes.

(b) Occluded mask of Baby Wipes.

Fig. 3: **Annotations of instance occlusion segmentation.** Occluded mask of objects is visualized separately. Note that its color (Baby Wipes) corresponds to the mask of visualization of visible masks.

**Model evaluation: softmax vs. relook**

We evaluated the proposed relook architecture with a human-annotated dataset for both training and testing, with the comparison to the softmax extension of Mask R-CNN [Do 17]. Table 1 shows the quantitative results, in which "Softmax" denotes Mask R-CNN Softmax, "Softmax_x2" denotes Softmax whose mask loss is scaled by 2, and "Relook" denotes Mask R-CNN with the relook architecture. Since learning result of RPN has lots of noise caused by randomness, we show results averaged in 10 - 15 times experiments for each model. For fair comparison, we use ResNet50 feature extractor using learning rate 0.0375 with 3 GPUs in which the learning rate is scaled by the number of GPUs in following experiments without note: $0.00375 = 3 \times 0.00125$.

The results of mPQ in Table 1 show that the proposed relook architecture surpasses the existing models, and effective to instance occlusion segmentation. We also show results of mAP (mean averaged precision) that has been used as the metric of instance segmentation of visible mask [Lin 14]. The results of mAP show that our model excels other models also in instance segmentation of the visible masks.

Table 1: **Softmax vs. Relook.**

| Model | **mPQ** | mSQ | mDQ | mAP |
|---|---|---|---|---|
| Softmax [Do 17] | 13.4 | 24.7 | 40.7 | 46.1 |
| Softmax_x2 [Do 17] | 13.8 | 25.2 | 41.9 | 47.1 |
| **Relook** (Ours) | **14.4** | **26.0** | **42.8** | **48.6** |

**Dataset evaluation: annotated vs. synthetic**

We evaluated our image synthesis framework with the proposed model using both synthetic and human-annotated training data to compare the performance training with the framework. The training result is evaluated with the performance using the test split of human-annotated dataset. Table 2 shows the averaged results of 3 experiments using the image synthesis referring results in §4.. It shows that our image synthesis using 4 - 6 instance images (14.2) is as effective as the result using a small human annotated dataset (14.4), in which each object appears 3 - 5 times, for learning instance occlusion segmentation.

Table 2: **Human-annotated vs. Synthetic.**

| Model | Dataset | mPQ |
|---|---|---|
| Softmax [Do 17] | Annotated | 13.4(±0.3) |
| | Synthetic | 13.5(±0.5) |
| Softmax_x2 [Do 17] | Annotated | 13.8(±0.4) |
| | Synthetic | 13.6(±0.6) |
| Relook (ours) | Annotated | 14.4(±0.5) |
| | Synthetic | **14.2**(±1.2) |

### 4.1 Application to Warehouse Picking

As an application of our system, we demonstrate the picking task of a target object from a pile of objects as shown in Fig.4. The recognition result shows the input image, predicted visible and occluded masks, and the targe object decided based on the occlusion understanding. In this experiment, we set the threshold of occlusion ratio to 0.3 in order to judge each instance is occluded; the occlusion ratio is the ratio of occluded pixels comparing with the whole pixels of instance. We set the threshold of inter-instance occlusion ratio to 0.1 to judge the instance is occluded by the other. And for the model of this demonstration, we use ReNet101 feature extractor as the backbone of our proposed model in order to have a better recognition accuracy to achieve the task. The successful clearing of obstacle objects and picking of targets in the demonstration shows that our proposed system's effectiveness and applicability in the real-world picking task.

## 5. Conclusions

We presented a vision system that only requires a few instance images of objects to learn instance occlusion segmentation; it is a new vision task proposed in this work. The system consists of 1) image synthesis with having ground truth of occluded region mask of each instance, and 2) instance segmentation networks that learn inter-instance relationship which is an important information for especially occlusion segmentation. We evaluated the proposed image synthesis and segmentation model via the ablation studies, and presented the effectiveness of the proposed system in the real picking task from a pile of objects.

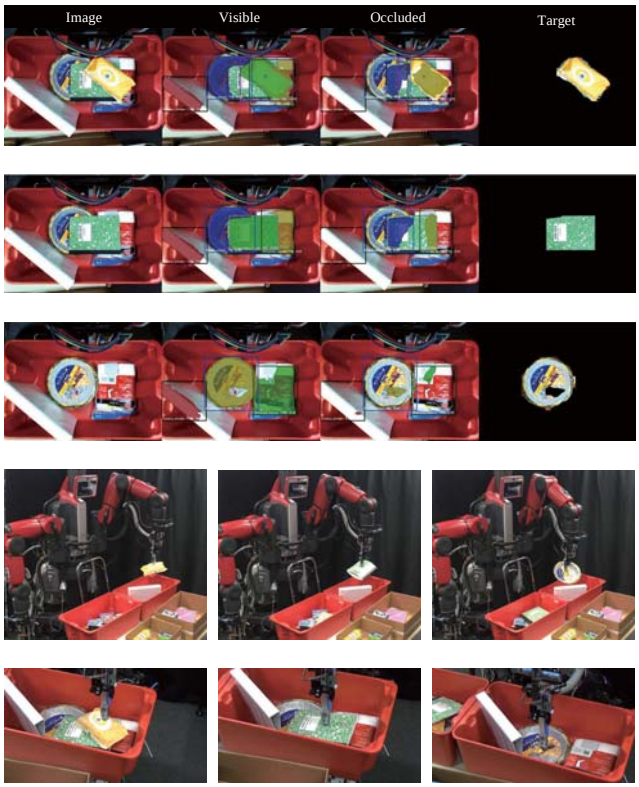



Fig. 4: **Picking a target from a pile of objects.**

## References

[Dai 16] Dai, J., He, K., Li, Y., Ren, S., and Sun, J.: Instance-sensitive Fully Convolutional Networks, *ECCV*, pp. 1–15 (2016)

[Do 17] Do, T.-T., Nguyen, A., Reid, I., Caldwell, D. G., and Tsagarakis, N. G.: AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection (2017)

[Dwibedi 17] Dwibedi, D., Misra, I., and Hebert, M.: Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection (2017)

[Georgakis 17] Georgakis, G., Mousavian, A., Berg, A. C., and Kosecka, J.: Synthesizing Training Data for Object Detection in Indoor Scenes (2017)

[He 17] He, K., Gkioxari, G., Dollár, P., and Girshick, R.: 1 1 Mask R-CNN, *ICCV* (2017)

[Hinterstoisser 17] Hinterstoisser, S., Lepetit, V., Wohlhart, P., and Konolige, K.: On Pre-Trained Image Features and Synthetic Images for Deep Learning (2017)

[Jonschkowski 16] Jonschkowski, R., Eppner, C., Höfer, S., Martín-Martín, R., and Brock, O.: Probabilistic multi-class segmentation for the Amazon picking challenge, *IEEE International Conference on Intelligent Robots and Systems*, No. i, pp. 1–7 (2016)

[Kirillov 18] Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P.: Panoptic Segmentation, pp. 1–9 (2018)

[Lin 14] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dolí, P.: Microsoft COCO: Common Objects in Context (2014)

[Long 15] Long, J., Shelhamer, E., and Darrell, T.: Fully convolutional networks for semantic segmentation, *CVPR*, Vol. 07-12-June, pp. 3431–3440 (2015)

[Nair 10] Nair, V. and Hinton, G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines, *Proceedings of the 27th International Conference on Machine Learning*, No. 3, pp. 807–814 (2010)

[Pinheiro 16] Pinheiro, P. O., Lin, T. Y., Collobert, R., and Dollár, P.: Learning to refine object segments, *Lecture Notes in Computer Science*, Vol. 9905 LNCS, pp. 75–91 (2016)

[Su 15] Su, H., Qi, C. R., Li, Y., and Guibas, L. J.: Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views, *Proceedings of the IEEE International Conference on Computer Vision*, Vol. ICCV, No. Sec 2, pp. 2686–2694 (2015)

[Zeng 18] Zeng, A., Song, S., Yu, K.-T., Donlon, E., Hogan, F., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E., Fazeli, N., Alet, F., Chavan Dafle, N., Holladay, R., Morona, I., Nair, P. Q., Green, D., Taylor, I., Liu, W., Funkhouser, T., and Rodriguez, A.: Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching, *ICRA* (2018)