# Leveraging an Instance Segmentation Method for Detection of Transparent Materials

Amanuel Hirpa Madessa
*Department of Computer Science and*
Technology
*Ocean University of China*
Qingdao, China
amanuel.hirpa2003@stu.ouc.edu.cn

Junyu Dong
*Department of Computer Science and*
Technology
*Ocean University of China*
Qingdao, China
dongjunyu@ouc.edu.cn

Xinghui Dong
*Center for Image Sciences*
*University of Manchester*
Manchester, United Kingdom
xinghui.dong@manchester.ac.uk

Ying Gao
*Department of Computer Science and Technology*
*Ocean University of China*
Qingdao, China
gaoying@stu.ouc.edu.cn

Hui Yu
*School of Creative Technologies*
*Univeristy of Portsmouth*
Portsmouth, United Kingdom
hui.yu@port.ac.uk

Israel Mugunga
*Department of Computer Science and*
Technology
*Ocean University of China*
Qingdao, China
mugungaisrael@stu.ouc.edu.cn

***Abstract*—Automatic detection of transparent materials (e.g., glass, plastic, etc.) is essential in many computer vision tasks. For example, a robot could use such a system to navigate around transmissive materials or operate tasks with these materials without causing damage. Nevertheless, it is challenging task as such materials exhibit less texture or background scenes dominate visual perception. Existing methods used either hand-engineered or leaned features to detect and segment transparent objects. We argue that pixel-wise detection and segmentation of transmissive materials improve detection performance and provide the fine-grained information compared to detecting bounding boxes of objects (i.e., localisation task). In this paper, we leverage a robust and state-of-the-art instance segmentation method namely, Mask R-CNN, in order to detect transparent materials. To be specific, we train the model on a new dataset with an evaluation based on publicly available dataset. Experimental results show that the adopted method significantly enhances the performance of transparent material detection. In particular, the resulting binary masks provides the pixel-level information for an improved understanding and analysis of transparency.**

***Index Terms*—Transparent Material, Instance Segmentation, Material Detection, Mask R-CNN**

## I. Introduction

Transparent materials, such as glass and plastic, can be found nearly everywhere in our surrounding environment. Due to the substance that they are made of and the nature of transmissiveness, these materials are sensitive to damage or mishandling. Thus, extra caution is required when handling and navigating around these materials immediately after they are first perceived. Psychological studies show that the human visual system perceives transparency (i.e., one surface is seen through another) when the Michelson contrast (i.e., the difference between the highest and lowest luminance) occurs [1]. Though we exhibit varying visual capability, our visual system can fail to recognise transparent surfaces as they have limited recognisable features or they transmit the background scene. The same challenges make it difficult for a computer system to recognise transparent materials.

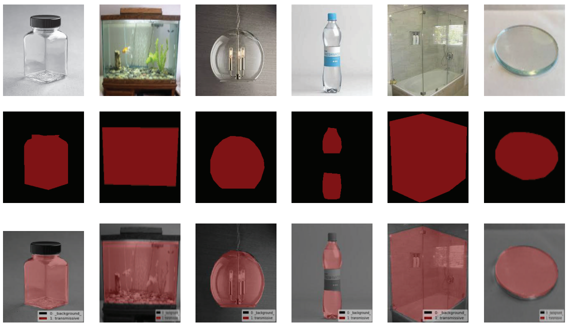

Fig. 1: Examples for training images: (top row) RGB images, (middle row) ground-truth masks, and (bottom row) segmentation class visualization.

With the growth of intelligent machines, such as self-driving vehicles and cleaning robots, automatic detection of transparent materials becomes important for the continued development of these technologies. For example, in the chemistry and biology laboratories, with the successful detection or understanding of transparency, an intelligent machine or robot can manipulate these materials and move around freely without causing any damage. Moreover, in the field of computer vision, recognition of transparent materials is significant for the understanding and analysis of multimedia data.

In the past years, studies have been conducted for recognition of transparent materials [2]–[7]. Although some progress has been made, recognizing transparent materials is still open problem. Most existing methods use hand-engineered features or other traditional methods as in [2], [3], [8]–[11], and few implemented learning methods to detect transparent objects [5]–[7]. The traditional methods impose several constraints and require prior information that make an inference of transparency computationally expensive and challenging. On the other hand, the learning-based methods do not detect objects regions at the pixel-level and cannot separate overlapping transparent materials though they may succeed in detecting transparent regions. In this paper, we therefore aim to address these problems by adapting an instance segmentation based on a deep leaning method, namely, Mask R-CNN [12].

The major contributions of this paper are summarized as follows. First, we present a framework for analysis and understanding of transparent material detection at the pixel-level using an instance segmentation method (i.e., Mask - RCNN). Our results provide insights into solving the feature descriptor problem with transparent object detection. Second, we introduce a new annotated dataset for use in the detection or segmentation of transparent objects.

## II. Related Work

Existing methods for detection of transparent materials can be classified into two categories based on the algorithm that they employ: traditional methods [2], [3], [8]–[11], [13]–[15] and leaning methods [5]–[7](e.g., deep learning based methods). In this section, we review the recent and most popular works related to transparent material detection.

Transparent materials share dominant features, such as highlight, blurring, overlay-consistency and texture distortion [4]. Most of these materials refract light from the background, which causes distortion. Although different materials result in different distortions amounts, they present similar characteristics of transparency. Compared to opaque materials, transparent materials are known for a few deterministic features. In this context, Kompella and Sturm [2] defined transparency as an inverse measure of the total number of discrete characteristics of an object. Relying on this definition, they proposed the collective-reward (CR) approach for the detection and localisation of semi-transparent objects. The principle follows that semi-transparent objects and surrounding pixels share similar features that result from the refraction and reflection of light through them. The algorithm classifies a semi-transparent region by aggregating support fitness functions and a feature reward function. Let $Cr_i^f$ denote the collective reward for every feature-cue $f$= {transparent feature cues}, and for each point $p_i$ of the background to the region $R$ (an assumed region randomly selected), the reward is

$$Cr_i^f = \frac{1}{\mu_1'}(\mu_1 Cr_{i,1}^f + \ \mu_2 Cr_{i,2}^f + \ldots + \ \mu_n Cr_{i,n}^f \qquad (1)$$

where $\{ \ \mu_1, \mu_{21}, \ \ldots, \mu_n\}$ are the results from calculating the fitness values of the connections. This method involves hand-engineered parameters and prior assumptions. In general, the algorithm fails when the region $R$ is large and when a comparison is performed far from the transparent region. In addition, some false positive results dominate whenever the image contains shadows of transparent objects.

Moving beyond the methods that focus on the deterministic features for detecting transparent objects, Wang et al. [3] instead proposed the use of depth information and multi-mode sensors. These features jointly predict glass edges and regions by building a Markov Random Field (MRF) model. Depth information is also used in the work of Luo et al. [10] and Hagg et al. [11]. The former used depth information with transparent cues, such as colour similarity and intensity consistency between the transparent region and surrounding pixels, while the latter used reflectance.

Although the above methods utilize the features observed in a transparent object, Maeno et al. [9] proposed a detection scheme, namely, the light field distortion (LFD) feature that relies on the distortion of the background scene. They claimed that a transparent object's form largely depends on the background scene instead of the object, which offers less information about its presence.

While traditional methods show reasonable progress toward detecting transparent materials, they significantly depend on prior knowledge and constraints. Also, they are computationally expensive and challenging to use in real time scenarios. These challenges motivate the use of learning-based methods, which are more robust, computationally inexpensive, and provide nonlinear solution. However, the implementation based on deep learning for detection of transparent materials is rare. One example is the work that Fuh and Lai [5] adapted in which a region convolutional neural network(R-CNN) method [16] was used to detect a transparent object. To improve the region proposal algorithm, they used highlights and colour similarity cues to remove identified regions which do not contain highlights.

More recently, Khaing and Masayuki [6] reported successful transparent object detection by leveraging a convolutional neural network (CNN) method, i.e., the Single Shot MultiBox Detector (SSD). This simple and effective method eliminates the need for proposals and feature sampling stages by computing everything in a single forward pass network [17]. The approach follows the assumption that if an object is presented in an image, there must be a window and label to which it is well aligned. In the process of training, SSD minimizes a joint regression and classification loss as

$$L\left(x,c,l,g\right) = \ \frac{1}{N}(L_{conf}\left(x,c\right) + \ \alpha L_{loc(x,l,g)}) \qquad (2)$$

where $N$ is the total number of default boxes, $l$ is the predicted box, and $g$ is the ground truth box parameters.; $L_{loc(x,l,g)}$ in the above formula is the localization loss denoted as

$$L_{loc(x,l,g)} = \sum_{i\in Pos}^{N} \ \sum_{m\in\{cx\ ,cy,w,h\}} x_{ij}^k smooth_{L1}(l_i^m - \hat{g}_j^m)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$

$$\hat{g}_j^{cw} \ = \log\frac{g_j^w}{d_i^w} \qquad \hat{g}_j^{ch} \ = \log\frac{g_j^h}{d_i^h}$$

where $c$ is the class confidence. Although SSD performs efficiently in the detection task, it does not handle well small transparent materials due to its shallowness. In this paper, we adapt the state-of-the-art object detection and instance segmentation method called Mask R-CNN [12] in order to better deal with a range of small to large transparent materials and identify them at the pixel level. Moreover, Mask R-CNN offers the advantage of identifying instances of the same materials and overlapping materials in different orientations within the image.

## III. Detection of Transparent Materials

The objectives of this study are to detect and segment transparent materials and examine the latency and transferability of Mask R-CNN for learning transparent features. Furthermore, we contribute an annotated dataset which is useful for the understanding, analysis, and detection of transparent objects. The framework for detection and segmentation of transparent materials is shown in Fig. 2.

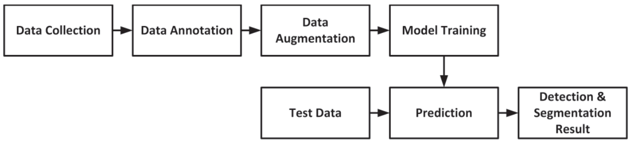

Fig. 2: The framework for detection and segmentation of transparent materials

The first stage of the proposed framework is collecting training data (e.g., images of transparent materials), and then we manually annotate the transparent regions of each image using an open source software tool. The next step applies an augmentation technique to increase the training size and variation of the data. Finally, we pass the data through the Mask R-CNN module for model training. During prediction, test data from [2] is feed into the inference module, which detects and segments the transparent materials from the images. The obtained visual and quantitative result is graphically displayed for interpretation.

### A. The Dataset

Mask R-CNN requires a considerable amount of labeled data in order to train without overfitting. However, obtaining images containing transparent materials with the mask information is tedious. In this study, we collected 1050 images with transparent materials (glass and plastic) from the Internet and annotated them manually (see Fig. 1). Then, we used the Mask R-CNN with the pre-trained weights trained on ImageNet for training.

We chose glass and plastic images because of high-volume availability. While gathering the images, we assumed a transparent material is one that clearly shows all or a part of its background scene. The size of the collected images varies from 127x127 to 259x259 pixels. During the annotation, we labeled all pixels belonging to the transparent material in each image using an open source annotator software. Since our collection for the training data is small, the model could easily overfit. Thus, in addition to using pre-trained weights, we incorporated an augmentation technique to increase the data variation. Upon publication and recommendation, we will make our dataset publicly available on https://github.com/AmanuelHirpa/TMD2.

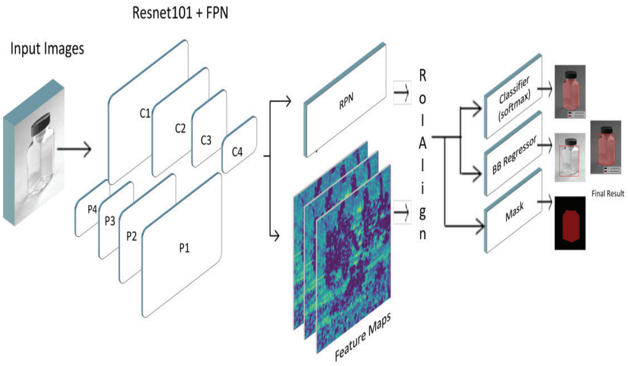

Fig. 3: A simplified graphical representation of the adapted Mask R-CNN architecture [12].

### B. Experimental Setup

We employed the open-source package of Mask R-CNN for training and prediction. The experiment was conducted on a single GPU (Tesla K40, 2880 cores, 12GB RAM) until it converge. For the backbone network, we used ResNet-101(with ResNet-101 and ImageNet pre-trained weights the model registered better results) with a minibatch size of two images. We retained the default parameters except for the RPN anchor scale size, image dimension and mean pixel values. We set the length of the square anchor side to 16, 32, 64, 128 and 256 pixels because we observed there are many small transparent materials in our dataset. Since our images are collected from the Internet whose quality may be low, we set the maximum dimension to 512 pixels and the mean pixel values to (43.53, 39.56, 48.22).

### C. Mask R-CNN Method

The Mask R-CNN method [12] is an object instance segmentation technique [18] that extends Faster R-CNN [19] by adding a module to predict the mask of instances from a detected bounding box [12]. As can be seen from Fig. 3, the Mask R-CNN architecture includes five main modules [12]: (1) a backbone network (ResNet [20]) serving as a feature extractor, (2) a Feature Pyramid Network (FPN) [21] enabling the accessibility of lower and higher level features from every level used to handle objects at multiple scales, (3) a Region Proposal Network (RPN) to generate the region of interest (ROI), (4) an ROI classifier and Bounding Box Regressor to predict classes of each ROI and the refine ROI assisted by ROIAlign and (5) a Mask Network to predict the mask at the pixel level.

Mask R-CNN implements a multi-task loss function that combines classification, bounding box, and segmentation mask losses [12] denoted as:

$$L = L_{cls} + L_{box} + L_{mask} \tag{3}$$

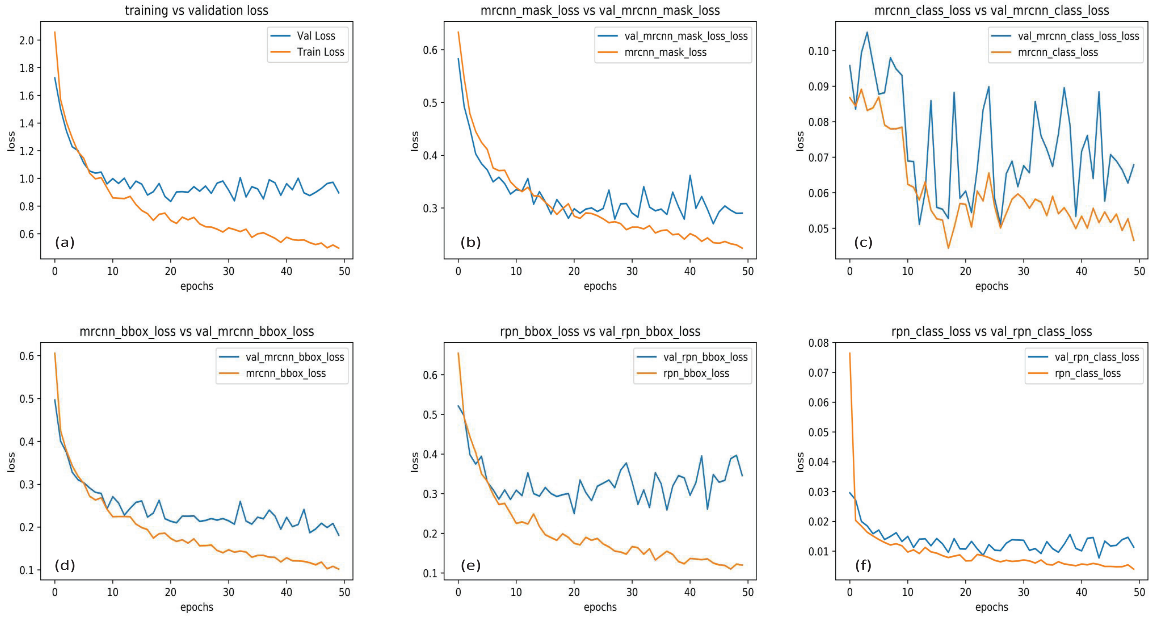

Fig. 4: Graph visualization of losses for Mask R-CNN optimization. (a) Training and validation L1-Loss; (b) training and validation of Mask R-CNN mask loss; (c) training and validation of Mask R-CNN class loss; (d) training and validation of Mask R-CNN bounding box loss; (d) training and validation of RPN bounding box loss; (e) training and validation of RPN class loss.

where, $L_{cls}$ and $L_{box}$ are the classification and bounding box losses used in [19] respectively. The $L_{mask}$ loss is a mask loss function defined as the average binary cross-entropy loss.

## IV. Evaluation and Results

In this study, we aim to investigate the effectiveness of Mask R-CNN model for detecting transparent materials. Table I shows the quantitative comparison between the results of our experiment and those derived using two different transparent object detection models [2], [6].

TABLE I: Quantitative comparison of transparent object detection on test images taken from [2].

| Metrics | Mask-RCNN | SSD [6] | CR [2] |
|---|---|---|---|
| AP@50 | **0.73** | 0.48 | - |
| AP@75 | **0.53** | 0.31 | - |
| AP@95 | **0.09** | 0.001 | - |
| Preciison | **0.82** | 0.78 | 0.75 |
| Recall | **0.77** | 0.43 | 0.66 |

Using the parameters and experimental setup described in Section III our adapted method generated superior performance to the work of Khaing and Masayuki [6] by a margin of 25%, 22%, and 8.9% on the average precision (averaged over the IoU thresholds) at thresholds 0.5, 0.75, and 0.95, respectively. Compared with the traditional method employed by Kompella and Sturm [2], our adapted method outperforms by a margin of 7% in precision and 11% in recall. This result demonstrates that Mask R-CNN better addressed the challenge of identifying transparent materials.

During optimization, we observed that the validation loss of the pixel-to-pixel segmentation mask was as smooth as bounding box loss even though bounding boxes are not an exact fit of the objects. This can be observed from apparent from the optimized loss curves depicted in Fig. 4. The finding demonstrates the effectiveness of the pixel-to-pixel segmentation branch of Mask R-CNN for locating transparent materials. We attribute the effective segmentation to the fact that the fully convolutional network is able to preserve the spatial dimension during mask prediction as well as the accurate alignment between the extracted features and their RoI using the RoIAlign function [12]. It is more evident on the qualitative comparison as shown in Fig. 5.

As can be seen in Fig. 5, the predicted results on the test set show that the adapted method detected small materials better than the existing methods. The segmentation ability of the adopted method is an additional function useful for the manipulation or interpretation of a selected block of pixels for further tasks. We also observed that Mask R-CNN is better at avoiding the non-transparent regions that have similar features to the transparent regions, such as shape and colour. As can be seen from the results above, the existing methods failed to differentiate between the shapes and colours in non-

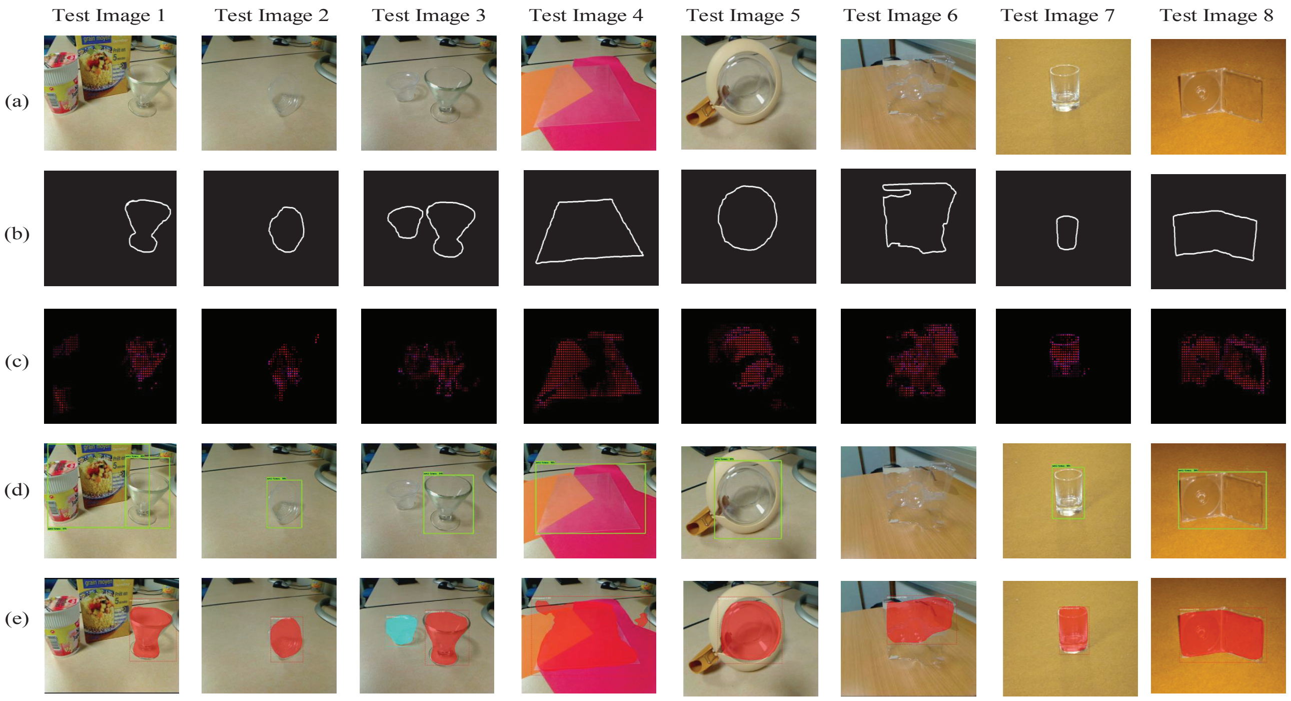

Fig. 5: Qualitative results of the transmissive object detection task on a test set from [2]. Top row (a) test images; second row (b) ground truth masks; third row (c) result from [2], fourth row (d) results of [6], bottom row (e) our result.

transparent regions that are similar to a transparent region. Also, some clear (i.e., a high degree of textureless) transparent materials were detected by the adapted method while Khaing and Masayuki [6] failed to detect the same features. All the experimental results show that Mask R-CNN is able to detect and segment transparent material effectively. We believe that the depth of the network and the pixel-wise learning capabilities of the Mask R-CNN are the key factors that enabled it to detect small and clear transparent materials.

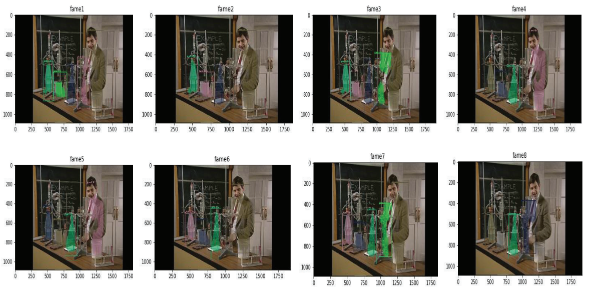

Fig. 6: Test result in indoor challenging lab video. For a better view, please zoom in the electronic version.

To visually inspect the quality of our trained model on videos, we conducted a test on a video obtained from the Internet. We further challenged the detection algorithm by selecting a video filmed indoor with several overlapping transparent materials as well as a moving person. As can be seen from Fig. 6, despite the challenging nature of the task and the test data, the model detected transparent materials successfully. However, in some frames, the person's face was also detected as a transparent material due to the pre-trained weight was trained on ImageNet, which contains faces. This issue might be avoided in future work if a large, ad-hoc dataset is available for detection of transparent materials to train the model from scratch.

To understand how well and what the model had learned, we inspected the weight and bias distributions along with the backbone network feature map (see Fig. 7). We observed that weights and bias were properly distributed. The feature map extracted from the backbone layer shows that some features, such as reflections and shininess, lead to false negative results. Although these features were mostly used as cues for transparent detection in previous studies [2], [5], they inclined to generate incorrect results when non-transparent and shiny objects presented in the image. This finding primarily affects the classification task between the transmittance properties, such as transparent and translucent, because these features are common to both properties. In such cases, avoiding these features may be an option to build a better recognition algorithm.

We also observed a few misalignments of segmentations produced by Mask R-CNN especially when the transparent material was entirely clear. We believe that the failures were due to insufficient training images. To be exact, the training

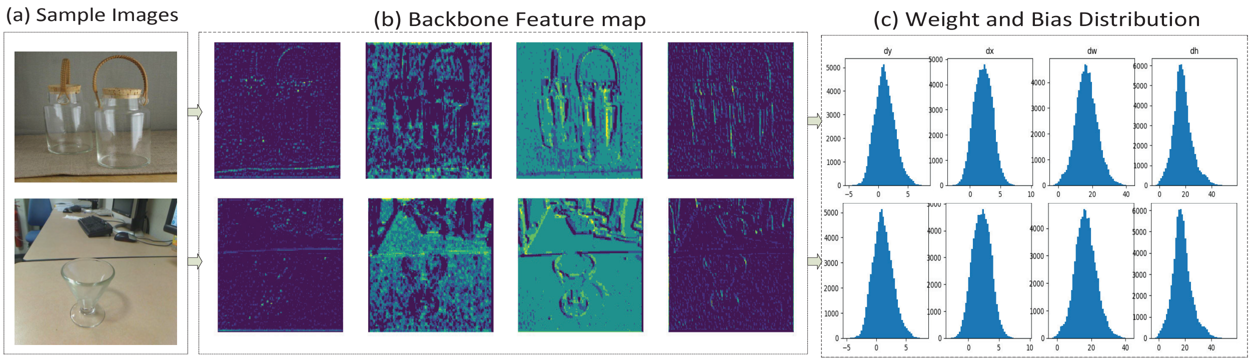

Fig. 7: Visualizing result; (a) sample Image; (b) backbone network feature map (resnet101); (c) shows how well the weight and bias distributed.

data is not enough for detecting very clear transparent materials when the non-transparent regions' visual features dominate the transparent regions in the image.

## V. Conclusion and Future Work

In this paper, we applied the instance segmentation method, namely, Mask R-CNN, to localization of the precise area of an image containing transparent behaviours or features from materials, such as glass and plastic. To this end, we performed extensive experiments on both the new dataset that we collected and a public test set. The comparison between the adapted method and two existing approaches suggested that our approach performed better than the two counterparts. By analyzing the results, we conclude that promising detection of transparent materials can be achieved using pixel-wise instance segmentation method. Despite Mask R-CNN was initially built for detection and segmentation of opaque objects, it can also be used to more difficult tasks, such as detecting transparent materials and locating pixels belong to these features with the minimal adjustment.

Further work is required in order to improve the performance of the detection of transparent materials by integrating the removal of negative artefacts through an end-to-end deep learning instance segmentation method. Also, a dataset with an adequate number of high resolution images should be considered for obtaining better results.

## References

[1] M. Singh and B. L. Anderson. Toward a perceptual theory of transparency. Psychological Review, 109(3):492–519, 2002. 106, 116.

[2] V. R. Kompella and P. Sturm, "Collective-reward based approach for detection of semi-transparent objects in single images," Computer Vision and Image Understanding, vol. 116, no. 4, pp. 484-499, 2012.

[3] T.Wang, X.He, and N.Barnes, "Glass Object Localization by Joint Inference of Boundary and Depth," in 21st International Conference on Pattern Recognition (ICPR 2012), (Japan), p.4, November 2012.

[4] K.McHenry, J.Ponce, and D.Forsyth, "Finding glass," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), IEEE, 2005.

[5] P.J. Lai, C.S.Fuh, "Transparent object detection using regions with convolutional neural network," In: IPPR Conference on Computer Vision, Graphics, and Image Processing, pp. 1–8 (2015).

[6] M.P.,Khaing, M.,Masayuki, "Transparent Object Detection Using Convolutional Neural Network," In: Zin T., Lin JW. (eds) Big Data Analysis and Deep Learning Applications.ICBDL 2018. Advances in Intelligent Systems and Computing, vol 744, 2018.

[7] V. Seib, A. Barthen, P. Marohn, and D. Paulus, "Friend or Foe: Exploiting Sensor Failures for Transparent Object Localization and Classification," in International Conference on Robotics and Machine Vision, Volume 10253, 2017.

[8] I. Lysenkov, V. Eruhimov, and G. Bradski, "Recognition and Pose Estimation of Rigid Transparent Objects with a Kinect Sensor," Proc. Robot. Sci. Syst., p. 8, 2012

[9] Y. Xu, K. Maeno, H. Nagahara, A. Shimada, and R. I. Aniguchi, "Light field distortion feature for transparent object classification," Comput. Vis. Image Underst., vol. 139, pp. 122–135, 2015

[10] R. C. Luo, P. J. Lai, and V. W. Sen Ee, "Transparent object recognition and retrieval for robotic bio-laboratory automation applications," IEEE Int. Conf. Intell. Robot. Syst., vol. 2015–Decem, pp. 5046–5051, 2015

[11] A. Hagg, F. Hegger, and P. G. Pl, "On Recognizing Transparent Objects in Domestic Environments Using Fusion of Multiple Sensor Modalities," in Robot World Cup, pp. 3–15, 2016.

[12] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask R-CNN. In International Conference on Computer Vision (ICCV), 2017.

[13] R. W. Fleming, F. Jäkel, L. T. Maloney, R. W. Fleming, and J. Gießen, "Visual perception of thick transparent materials," vol. 3855, 2011.

[14] T. Wang, X. He, and N. Barnes. Glass object segmentation by label transfer on joint depth and appearance manifolds. In ICIP, 2013.

[15] Y. Xu, H. Nagahara, A. Shimada, R. Taniguchi, "Transcut: Transparent object segmentation from a light-field image", Proc. IEEE Int. Conf. Comput. Vis, pp. 3442-3450, 2015.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. In ECCV, 2016.

[18] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path aggregation network for instance segmentation" in CVPR, 2018

[19] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", Proc. 28th Int. Conf. Neural Inf. Process. Syst., pp. 91-99, 2015

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778, 2016.

[21] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In Computer Vision and Pattern Recognition (CVPR), 2017.