TECHNISCHE UNIVERSITÄT WIEN

ACIN
AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

Comparison of RGB- and NIR-based Transparent Object Segmentation Methods

DIPLOMARBEIT

Conducted in partial fulfillment of the requirements for the degree of a

Diplom-Ingenieur (Dipl.-Ing.)

supervised by

Ao. Univ.-Prof. Dr. techn. Markus Vincze
Dr. techn. Jean-Baptiste Weibel, MSc.

submitted at the

TU Wien
Faculty of Electrical Engineering and Information Technology
Automation and Control Institute

by

Veronika Rettner
[redacted]

Vienna, September 2022

Vision for Robotics Group
A-1040 Wien, Gußhausstr. 27, Internet: http://www.acin.tuwien.ac.at

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit gemäß dem Code of Conduct, Regeln zur Sicherung guter wissenschaftlicher Praxis, insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Vienna, September 2022

Preamble

I want to thank my supervisors, Jean-Baptiste and Professor Vincze, for their patience and professional support. Especially I'd like to thank them for encouraging me to apply for the ARW22, an experience which I really enjoyed. I also want to thank all the colleagues from the Vision for Robotics research group that contributed to my work. Thank you to Simon and Nikola for 3D modelling of the transparent objects. To Markus Suchi and Bernhard, thank you for instructing me and making the dataset collection easy and enjoyable.

I am also very grateful for the support my friends gave me. Thank you for always having an open ear and for motivating me. Also a shout-out to my pubquiz colleagues for the welcome distraction.

A special thank you goes to my parents, who did everything they could to make my studies possible. And finally, I want to thank Max and my fluffiest friend Flauschhilda for their emotional support and for being there for me every single day.

Vienna, September 2022

Abstract

The ageing population in many western countries leads to new challenges in our society. Here, support by robots can be part of the solution, e.g. by providing assistance for the elderly. However, for this task, robots need to be able to understand their environment better, where machine vision plays a central role. Especially being able to perceive and manipulate transparent objects is essential as they are widely used by humans. To tackle this problem, this thesis compares different methods for mask prediction of transparent objects by evaluating the methods on a new annotated dataset. The dataset consists of RGB-D and infrared images of several scenes with transparent objects. In addition, the camera poses are recorded to enable the annotation of the object poses. The annotation is carried out manually and used to render silhouettes as ground truth images for the comparison. A selection of mask predictions methods for transparent objects is then evaluated on the dataset and the results are compared with the ground truth using pixel-wise metrics, namely F1 score, IoU, precision and recall. The methods for mask prediction selected in this study are an approach using invalid depth and GrabCut [1], an infrared image-based approach adapted from Ruppel et al. [2], and the Convolutional Neural Networks based (CNNs) TOM-Net [3], ClearGrasp [4], TransLab [5] and Trans2Seg [6]. The results of the respective approaches show a varying performance, with deep learning-based methods showing a better performance overall. TransLab, for example, exceeds the other methods with an F1 score of 67.5% and an IoU of 55.8%. The overall performance over the whole dataset is discussed and, furthermore, an in-depth analysis for selected scenes is provided, highlighting similarities as well as challenges for the above approaches. While many approaches successfully predict a rough shape, fine and more complex details like plastic tubes prove to be quite challenging overall.

Kurzzusammenfassung

Die alternde Bevölkerung in vielen westlichen Ländern führt zu neuen Herausforderungen für unsere Gesellschaft. Roboter können hier Teil der Lösung sein, z.B. indem sie zur Unterstützung älterer Menschen eingesetzt werden. Für diese Aufgabe müssen Roboter jedoch ein gutes Verständnis ihrer Umgebung erlangen, wobei maschinelles Sehen eine zentrale Rolle spielt. Insbesondere die Erkennung und Manipulation transparenter Objekte erweist sich hier als schwierig. Zur Untersuchung dieses Problems wird in dieser Arbeit ein Vergleich unterschiedlicher Methoden zur Maskenerkennung transparenter Objekte durchgeführt. Der zu diesem Zweck erstellte Datensatz besteht aus Farb-, Tiefen- und Infrarotbildern transparenter Objekte sowie den zugehörigen Kamerapositionen. Zur Erstellung der Referenzmasken für die Evaluation erfolgt zudem eine manuelle Annotation der Objekte. Eine Auswahl an Methoden zur Maskenerkennung transparenter Objekte wird auf den neu erstellten Datensatz angewandt und die Ergebnisse werden mithilfe der Referenzmasken unter verschiedenen Metriken, nämlich F1-Score, IoU, Precision und Recall, auf Pixelebene ausgewertet. Für diese Studie wurden folgende Methoden ausgewählt: ein Algorithmus, der ungültige Tiefenkamerawerte in Kominbation mit dem GrabCut Algorithmus [1] verwendet, eine Adaption des Algorithmus von Ruppel et al. [2], der auf Infrarotbildern basiert, sowie die neuronalen Netzwerke TOM-Net [3], ClearGrasp [4], TransLab [5] und Trans2Seg [6]. Die jeweiligen Ansätze erzielen sehr unterschiedliche Ergebnisse, wobei die auf maschinelles Lernen basierten Methoden insgesamt am besten abschneiden. TransLab zum Beispiel erzielt den höchsten F1-Score von 67,5% und einen IoU von 55,8%. Die Ergebnisse werden ausführlich diskutiert und darüber hinaus eine eingehende Analyse ausgewählter Szenen erstellt, wobei sowohl die Gemeinsamkeiten als auch die Schwierigkeiten der oben genannten Ansätze aufgezeigt werden. Während die meisten Ansätze erfolgreich grobe Objektmasken erkennen können, erweisen sich feine und komplexere Details wie Kunststoffschläuche insgesamt als schwierig.

Contents

List of Figures

List of Tables

1 Introduction

The perception of transparent objects shapes can be difficult for humans [7], and even more so for machine vision applications. Although great progress in object recognition and reconstruction was achieved in the last decade, solutions for this specific group of objects are still an ongoing challenge. However, transparent objects are indispensable in industry, science and many daily human activities, because they allow easy visibility of their content. Therefore, the ability to grasp such objects is an important task for robots, especially in the context of an ageing population as well as a global pandemic.

1.1 Challenges

The recognition of objects is among the most important topics in the field of robotic vision. To enable robots to grasp unknown objects, precise enough localization and reconstruction of these objects is necessary. Due to cheap RGB-D sensors, object reconstruction from depth data has become much easier recently. Such sensors can already be used for the detection of opaque objects with good success. Due to their working principle, these sensors, however, fail to predict depth of objects with non-Lambertian reflection properties, i.e. objects with transparent or shiny surfaces, like glass and plastic. Here, Figure 1.1 shows an example for images of transparent objects captured with an RGB-D sensor. The sensor either yields zero or invalid depth values for the transparent object, e.g. by wrongly predicting the depth of the surface beneath the objects. Therefore, the raw depth information of such an RGB-D sensor is not suitable for the direct use in robot grasping applications.

While there are sensors that are more suitable for this task, such as light-field cameras, the usage of RGB-D cameras remains attractive due to their ubiquity and lower cost. Until now only few approaches exploit the available depth data to obtain more reasonable shape information. However, various approaches have been proposed in the literature, which predict silhouettes of transparent objects solely from RGB data. These silhouettes combined over multiple views can subsequently be used to reconstruct 3D shapes. 3D reconstruction is expected to be easier and faster via the detour using silhouettes than direct depth prediction.

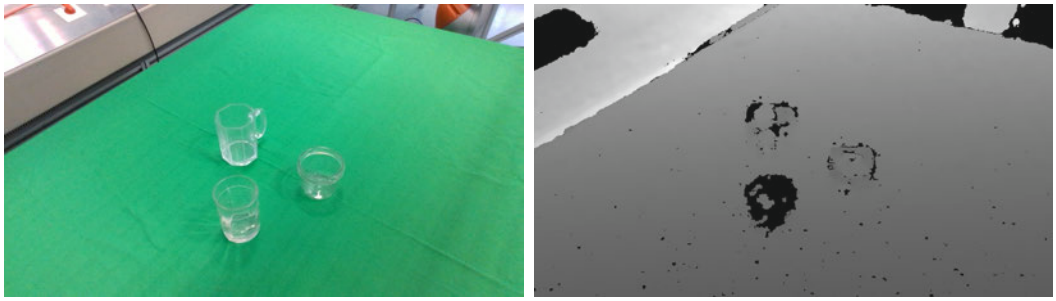


Figure 1.1: RGB image (left) and depth image (right) of transparent objects captured with a RealSense D435 camera. The depth image shows that the depth pixels corresponding to transparent objects are mostly invalid.

1.2 Contribution

This work is focused on providing a comparison between different approaches to find the most suitable segmentation method for several application scenarios. First, common household goods with different geometric complexity are selected as objects, but also a special case of a more complex medical object is investigated. Many approaches in literature are trained on simple, rotationally symmetric objects. Here, the performance on more complex objects is highly interesting, as they are often very common in real applications. In our dataset, these objects are arranged in various scenes, which are captured with a commercial RGB-D sensor. The data is then annotated and used as ground truth for the comparison of different segmentation approaches. Here, classical as well as deep neural network based methods are employed (i.e. [1] [6]). It is investigated which approach for silhouette acquisition has the best performance on our dataset. For evaluation, common performance parameters like recall, precision and F1 score are used in this study.

Our aim is to provide a more concise picture on which approach to choose for various case scenarios with transparent objects. In this work, a vast difference in performance of the different approaches is observed. Best overall performance is obtained by deep learning-based methods. All methods used in the comparison struggle more or less with details like transparent tubes or handles, but also with simpler geometries, especially if the objects' material is thin plastic. Figure 1.2 displays some examples of the evaluation results.

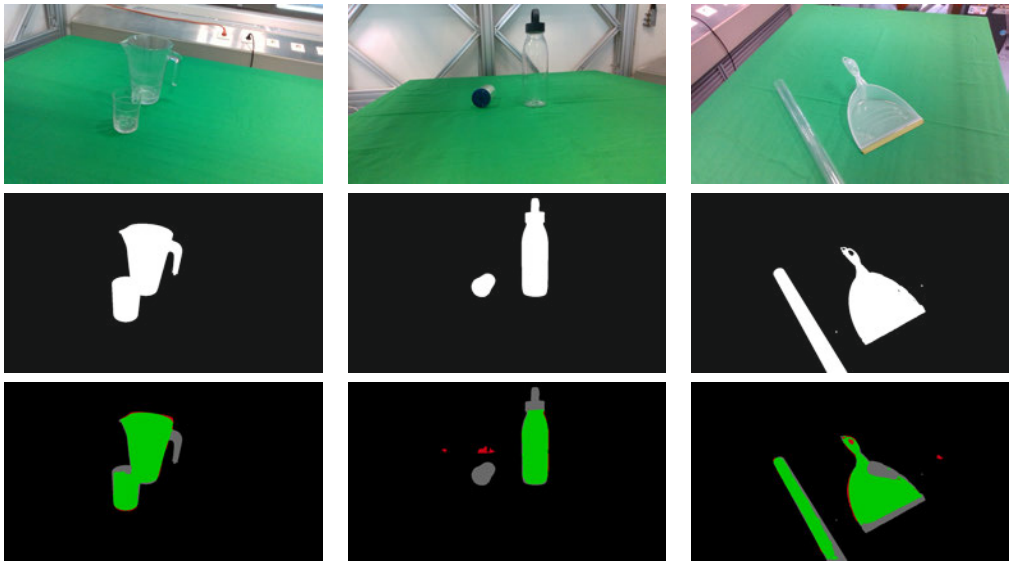


Figure 1.2: Preview of evaluation results. Each column shows a different scene with the respective RGB image, groundtruth mask and an example of a mask prediction. In the last row, for visualisation, correctly predicted pixels are marked green, while incorrect pixels are marked either red (background, i.e. false-positive) or grey (object, i.e. false-negative).

1.3 Outline

In Chapter 2, the task definition is provided as well as background information on the metrics used for the comparison. Furthermore, a detailed literature review on work dealing with detection and segmentation of transparent objects is given. Chapter 3 deals with the experimental aspects of dataset acquisition and annotation. Furthermore, details on invalid depth-based and infrared-based mask prediction are given. The results and their discussion is found in Chapter 4. Here, the main focus lies on the evaluation of the transparent object segmentation of different scenes. These findings are subsequently summarised in Chapter 5 and finally an outlook on possible future work is given.

2 Background

This chapter provides background information on mask prediction of transparent objects. First, a task definition is given and several metrics for the evaluation of image segmentation are presented. In the main section, existing literature on image segmentation is discussed with a special focus on work dealing with the detection and segmentation of transparent objects.

2.1 Task Definition

The segmentation of transparent objects is still an active research area and a key challenge. Although there are already numerous approaches introduced in the literature, a comparison is difficult since they are all trained and evaluated on completely different datasets. In this work, existing methods and algorithms for transparent object segmentation are to be applied on the same real-world dataset, measuring performance using recall, precision, F1 score and intersection over unit (IoU), see Section 2.2. The new dataset introduced in this work is described in Section 3.1. The results are then compared in order to evaluate which approach works best on a dataset with household objects and medical devices. In addition, it is evaluated if the predicted silhouettes are suitable for 3D reconstruction via Shape-from-Silhouette (SfS) techniques to allow tasks like robot grasping.

2.2 Metrics

The prediction of binary segmentation masks is a pixel-wise classification problem. Therefore, similar metrics as for object classification can be applied. To each pixel of a predicted mask one of four labels can be assigned by comparison of the predicted pixel value with the ground truth pixel value. These labels and their descriptions are described in Table 2.1.

A visualisation of these pixel labels for mask prediction is shown in Figure 2.1. Here, green pixels correspond to TP pixels, black pixels to TN pixels, red pixels to FP pixels and grey pixels to FN pixels.

Label	Description
TP - true positive	predicted pixel and ground truth pixel both true
FP - false positive	predicted pixel true and ground truth pixel false
TN - true negative	predicted pixel and ground truth pixel both false
FN - false negative	predicted pixel false and ground truth pixel true

Table 2.1: Pixel labels and their descriptions used for the evaluation of mask predictions.

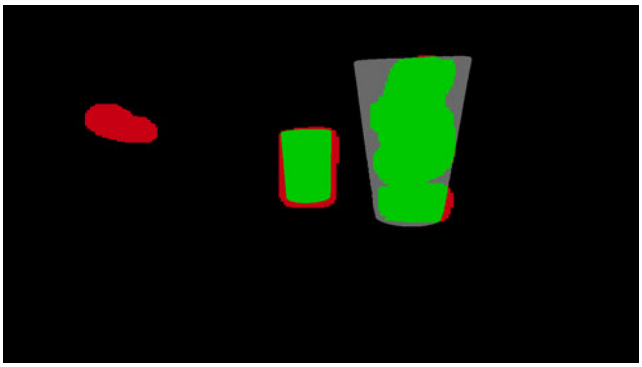


Figure 2.1: Classification of pixels obtained by mask prediction in comparison to ground truth mask pixels. The colours indicate TP (green), TN (black), FP (red) and FN (grey) pixels.

A simple measure for the percentage of correctly classified pixels is represented by the pixel accuracy (PA), which is calculated by

$$PA = \frac{TP + TN}{TP + FP + TN + FN}. \tag{2.1}$$

However, this metric is not very meaningful, if the number of pixels for each class is unbalanced. In a binary image this would be the case, if the biggest part of an image is occupied by the background. Therefore, other types of metrics have to be considered to overcome this problem.

Further widely used metrics for classification evaluation are precision and recall, which can also be used to evaluate pixel-wise labeling. The precision tells how many of the predicted true pixels are actually true positives (Equation 2.2). It is, therefore, an indicator for the quality of the true pixel prediction.

$$Precision = \frac{TP}{TP + FP}. \tag{2.2}$$

In contrast, the recall is a measure for the sensitivity of the recognition of positives: It shows the proportion of correctly classified true pixels to all true

ground truth pixels:

$$Recall = \frac{TP}{TP + FN}. \tag{2.3}$$

These two metrics can be combined into the F1 score, which is the harmonic mean of precision and recall (see Equation 2.4). Thus, the F1 score is more suitable for unbalanced classes than pixel accuracy.

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

Another relevant metric is the Intersection over Union (IoU) or Jaccard Index, which is a measure for the overlapping area of the ground truth mask positives and the predicted mask positives (Equation 2.5). Although usually bounding boxes are used for this metric, it also can be calculated at pixel level.

$$IoU = \frac{TP}{TP + FP + FN}. \tag{2.5}$$

2.3 Related Work

The detection and segmentation of objects is highly relevant in the field of computer vision and therefore subject to a wide range of studies in literature. In this section, previous work related to the task investigated in this thesis is presented. First, a short overview of state-of-the-art image segmentation approaches is given and then the field of transparent object recognition is discussed in more detail. In particular, literature dealing with transparent object detection, segmentation, 3D reconstruction and pose estimation is summarized. Finally, various existing datasets containing transparent objects are presented.

2.3.1 Image Segmentation

The easiest methods to achieve segmentation include e.g. thresholding and morphological operations [8]. Other early segmentation techniques are region splitting and merging, watershed, and clustering. More advanced approaches formulate the segmentation task as an energy minimization problem, e.g. with Markov Random Fields (MRFs) or Conditional Random Fields (CRFs), that can be solved with graph cuts [9] or loopy belief propagation (LBP) [10]. A widely used interactive approach based on iterative graph cuts was presented with the GrabCut segmentation algorithm [1].

With the rapid development of neural networks, most of recent literature focuses on solving the problem of object detection and segmentation with deep

learning networks. Here, the prediction of a segmentation task is treated as a pixel-wise labelling problem. One of the first deep learning networks for image segmentation was a fully convolutional network (FCN) introduced by Long et. al [11] that allows end-to-end learning of arbitrary-sized images. However, conventional FCNs are too slow for real-time inference, have low resolution and do not take advantage of global context [12].

Most deep-learning image segmentation methods are based on some kind of encoder-decoder architecture. An early approach uses a convolutional network like the VGG16 [13] as the encoder while the decoder is a mirrored deconvolutional network for the prediction of the segmentation map [14]. A well-known model developed for medical image segmentation is U-Net [15], in which feature maps from the encoder are copied to the decoder network to preserve pattern information. A more recent work with encoder-decoder architecture is the DeepLabV3+ network [16], which uses dilated ("atrous") convolutions to balance the resolution of features. As encoder the DeepLabV3 model [17] is adapted, which includes parallel atrous convolutions at different dilation rates, called Atrous Spatial Pyramid Pooling (ASPP).

Another way to address the issue of a small receptive field of CNNs is to include attention mechanisms, e.g. by adding an attention module [18]. Recent work showed that the transformer architecture can achieve state-of-the-art image classification results and therefore is promising for image segmentation [19]. This idea was adopted in [20], introducing a transformer-based encoder combined with an FCN decoder. Furthermore, in [21], a pure transformer-based encoder-decoder model for semantic segmentation is presented.

2.3.2 Detection of Transparent Objects

The detection of transparent objects is a special case in the wide field of object detection and reconstruction. The non-Lambertian properties of transparent objects cause difficulties for general object detection algorithms. Many applications use RGB-D sensors, which are common due to their low cost and high flexibility. However, their infrared-based depth calculation frequently fails in the case of transparent and highly reflective objects. Here, different possible solutions for the detection of transparent objects are identified: McHenry et al. [22] were among the first to exploit the properties of glass for the detection of edges and regions of transparent objects, using cues like the distortion of the background texture or highlights to train a Support Vector Machine (SVM) classifier. Alt et al. [23] compared depth maps from multiple views to detect points with depth inconsistencies that indicate the presence of a transparent surface. Hagg et al. [24] combined multiple sensor modalities to enable the recognition of diffuse, reflective, or transparent objects. In particular, they use

the distortion of the infrared pattern of an active RGB-D camera as a cue for transparent materials. More recent approaches rather focus on the application of deep learning networks, showing that transparent object detection is feasible without consideration of the special properties of transparent materials [25].

2.3.3 Segmentation of Transparent Objects

Due to the variety of methods used for the segmentation of transparent objects, the methods are classified based on the input data.

RGB image-based segmentation

Based on [22], Torrez-Gomez et al. proposed a graph-based segmentation algorithm, incorporating edge detection, super-pixel classification and consensus voting [26].

More recently, just like for general image segmentation, deep-learning networks especially for transparent objects were employed. For example, a deep-learning framework for learning transparent object matting from a single image, called TOM-Net, was reported [3]. The output of TOM-Net is an environment matte consisting of an object mask, an attenuation mask and a refractive flow mask. These results can be used to composite the object onto new backgrounds, as seen in Figure 2.2. In contrast to other image matting methods for transparent objects, this approach does not depend on specific backgrounds or patterns as in [27] or [28]. In TOM-Net, the transparent object mask is predicted in the first stage of the framework by CoarseNet, an adapted mirror-link CNN [29] trained from scratch. For training and testing a large-scale synthetic dataset of transparent objects rendered in front of scene images and synthetic patterns was created. Additionally, a real-world dataset containing 14 objects was collected for evaluation. However, this dataset does not include the ground-truth matte, but pictures of the backgrounds without the transparent objects. Although TOM-Net is trained just on synthetic data, it shows good results on both the synthetic dataset as well as the real-world dataset collected by the authors.

Other recent approaches often use CNNs, which were originally trained for general object segmentation. For example, the state-of-the-art instance segmentation network Mask R-CNN [30] was retrained on a new dataset consisting of transparent objects, highlighting the transferability for learning transparent features [31]. Stets et al. [32] showed that a CNN with VGG16-Net backbone trained on a large-scale synthetic dataset of scenes with transparent objects is able to generalise to real-world scenes .

Xie et al. [5] developed a boundary-aware segmentation algorithm called TransLab and trained it on a large-scale dataset of real scenarios with transpar-

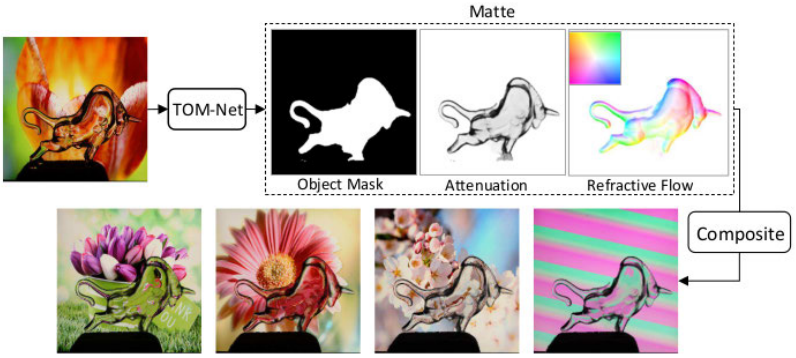


Figure 2.2: TOM-Net results, consisting of an object mask, an attenuation mask and a refractive flow mask, for realistic composition of transparent object. Reproduced from [3].

ent objects. They also re-trained several state-of-the-art networks for semantic segmentation on their own training set and showed that TransLab has overall a better performance. In a follow-up work, they extended their dataset to 11 categories and presented a hybrid CNN-Transformer segmentation pipeline [6], concluding that it outperforms pure CNN-based methods.

Very recently, a cascade network architecture was proposed by [33], introducing modules with residual learning and point-based graph convolution to enhance boundary prediction. They reported that their approach even outperforms TransLab, when trained on the same dataset.

Invalid depth-based segmentation

Some methods exploit the "holes" (missing depth) caused by transparent objects in depth images captured with common RGB-D cameras. For example, boundary label predictions from appearance and depth features were integrated into a Markov Random Field (MRF) model for glass object segmentation in [34]. Zero-depth values and noise region search in multiple RGB images were used to extract silhouettes by Ji et al. using joint GrabCut segmentation [35]. Guo-Hua et al. also used invalid depth information as a cue for segmenting transparent objects in an image in combination with the GrabCut segmentation algorithm [36]. The deviation of feature points due to the background distortion caused by refractive materials was subsequently used to validate the transparent object candidates.

Ruppel et al. [2] proposed a reconstruction pipeline that utilizes only raw infrared images. In a first step, the scattering of the infrared projection pattern on transparent surfaces is used to generate a transparency candidate map. A

standard blob detection algorithm is then employed to detect transparent object candidates. Additional constraints are considered to filter the candidates and the transparency is validated by comparing the brightness difference between pixels inside the object with pixels around the object.

Approaches including other image information

In light-field images, the features of objects with Lambertian properties are distributed almost linearly with respect to the viewpoints, whereas features of reflective objects are inconsistent between viewpoints [37]. These light-field properties were considered in a graph-cut optimization algorithm called TransCut [38], [39]. In the polarization plane, the strength of transparent object texture is increased drastically. Furthermore, Kalra et al. use polarization images as input in their proposed Polarized Mask R-CNN framework [40]. In addition, their method is also capable to distinguish real transparent objects from printouts.

2.3.4 Depth Estimation of Transparent Objects

Several works focus on the challenging depth prediction of transparent objects. Some approaches tackle this by improving the camera-internal depth calculation, e.g. Saygili et al. [41] propose a fully-connected CRF-based model to improve cross-modal stereo between RGB and IR images of the Kinect camera.

More recently, neural networks were also employed in depth estimation for transparent objects. With ClearGrasp [4], a deep learning approach for transparent objects, which provides a 3D shape prediction from a single RGB-D image, is available. Here, the authors predict pixel-wise masks, surface normals and occlusion boundaries using three deep convolutional networks. Therefore they modified Deeplabv3+ models [16] with a dilated residual network as backbone. The predicted masks are used to remove all pixels of the depth image corresponding to the transparent object. These outputs of the neural networks as well as the original depth image are then fed into a global optimization algorithm proposed by Zhang and Funkhouser [42], which completes the missing depth information. The schematic representation of the pipeline from the original contribution is shown in Figure 2.3. The authors also created a synthetic dataset for training and testing, as well as an additional real-world test dataset of transparent objects with the corresponding ground truth, see Section 2.3.7. The real-world test benchmark shows the ability of this approach to generalise to real-world transparent objects as well as unknown objects. A major advantage of this method is that no knowledge of the camera position or 3D model objects is necessary. In addition, the authors incorporated ClearGrasp

in a robotic picking system leading to an improvement of the grasping success rate for transparent objects. Difficulties identified in the study are varying lighting conditions, cluttered environments and sharp caustics (i.e. light effects caused by curved surfaces) and shadows.

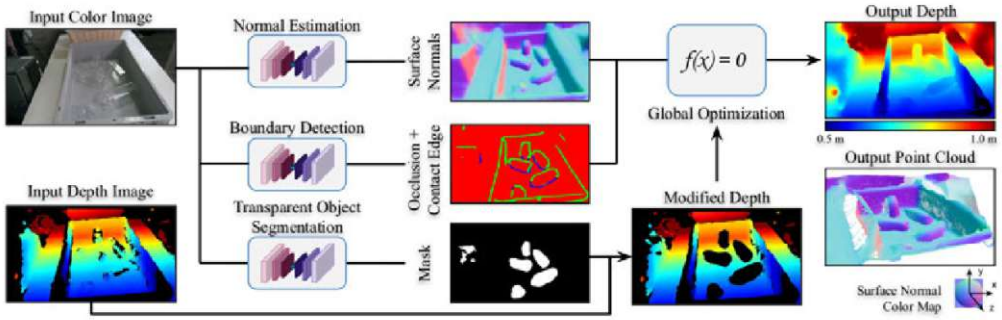


Figure 2.3: Schematic representation of the ClearGrasp pipeline including three neural networks for normal estimation, boundary detection and transparent object segmentation which are subsequently used in a global optimization algorithm to predict depth. Reproduced from[4].

Another method for depth completion was introduced with TranspareNet [43]. In contrast to ClearGrasp, in their work the distorted depth map caused by transparent objects is also taken into account, using a 3D CNN encoder-decoder for a rough point cloud prediction, which is then fed into a depth completion module for refinement.

Very recently, quite different approaches based on implicit neural representations of 3D objects were proposed: For example Zhu et al. [44] introduced a network for learning implicit depth functions defined on ray-voxel pairs and additionally an iterative model for depth prediction refinement.

2.3.5 Multi-View Reconstruction of Transparent Objects

Some approaches have the goal of reconstructing transparent objects as accurately as possible, for instance for virtual reality or graphic rendering. These methods, however, often require special setups containing moving light beams [27], multiple cameras and a monitor that has to be moved manually [45] or placement on a turntable [28]. Since these methods are designed for very specific environments, they are not applicable for independent robotic tasks. However, recently Li et al. [46] provided a physically-based deep NN trained on

synthetic images of glass figures, where no constrained environment is necessary to reconstruct detailed objects.

In contrast, there are also several multi-view reconstruction approaches that were designed explicitly having robot grasping applications in mind. Some of them assume prior knowledge of the object model [47], [48] or are limited to rotationally symmetrical shapes [36], [49], [50]. One way to retrieve 3D models of arbitrary-shaped objects is to estimate their silhouettes (see Section 2.3.3) and to use them in a Shape from Silhouette (SfS) technique, e.g. voxel carving [26], [35], [51]. In Ruppel et al. [2], the focus lies not so much on accurate silhouettes, but rather on a formulation of an optimization problem based on the resulting 2D transparency maps. The problem is then solved iteratively, resulting in a 3D density distribution that can be used to obtain a volumetric object model. The quantitative evaluation shows an average reconstruction error in the order of the voxel grid resolution.. Furthermore, a duration of about one minute for a full reconstruction on a standard desktop computer is reported. A big advantage of this approach is the capability to reconstruct objects with cavities, which are challenging for many other methods.

Quite a different method using a Neural Radiance Field (NeRF) model to recover the geometry of transparent objects from multi-view RGB images was presented by Ichnowski et al. [52]. So far, a drawback of this method is the long training time for NeRF models.

2.3.6 Pose Estimation of Transparent Objects

Earlier approaches used edge fitting [53] or template matching [48] to estimate the pose of known objects. In another work [36], contours from RGB and IR images were extracted and used for stereo matching to determine the pose. With the rise of CNNs, KeyPose was introduced [54]. Here, a dilated CNN is used to estimate poses with the help of keypoints from stereo input. Furthermore, a two-stage approach specifically for 6DoF pose estimation of transparent objects from a single RGB-D image was proposed by Xu et al. [55]. Here, an extended pointcloud representation is used in order to gain higher efficiency. The extended pointcloud is defined by a UV map, surface normals and the estimated 3D plane on which the object is placed.

2.3.7 Datasets of Transparent Objects

Since neural networks need specific training data and in addition test data for validation, there are already several datasets available which contain transparent objects. An overview of datasets featuring transparent objects which are

published online is given in Table 2.2. ClearGrasp is listed twice as both a synthetic and a real-world dataset are reported.

Dataset	Type	# Images	# Objects	Clutter	Mask	Depth	Pose
Stets et al. [32]	synthetic	80,000	600		x	x	
ClearGrasp [4]	synthetic	50,000	6	x	x	x	x
ClearGrasp [4]	real-world	286	10	x	x	x	x
KeyPose [54]	real-world	48,000	15		x	x	x
Trans10K [5]	real-world	10,428	10k+	x	x		
TODD [43]	real-world	15,000	6	x	x	x	
Our dataset	real-world	640	15	x	x	x	x

Table 2.2: Overview of datasets published online at the time of this work, which feature transparent objects.

Large-scale datasets of transparent objects predominantly consist of synthetic images. For instance, one of the largest datasets of transparent objects was introduced by Stets et al. [32]: It contains 80,000 synthetic images of 600 shapes rendered from different viewpoints. Besides groundtruth masks there are also depth maps and normal maps available.
Another large-scale synthetic dataset was created for the depth estimation pipeline ClearGrasp [4]: For the usage as training data, 50,000 images of transparent objects were rendered as well as the corresponding masks, depth maps, normals and occlusion boundaries. It contains only six objects, but in contrast to the dataset of Stets et. al also provides cluttered scenes. In addition as a benchmark, a real-world dataset of 286 RGB-D images and corresponding groundtruth geometries was created. Figure 2.4 shows excerpts of synthetic RGB images with transparent objects from the datasets of Stets et al. and ClearGrasp.

Figure 2.4: Example images of datasets with synthetic transparent objects, taken from [32] (left) and [4] (right).

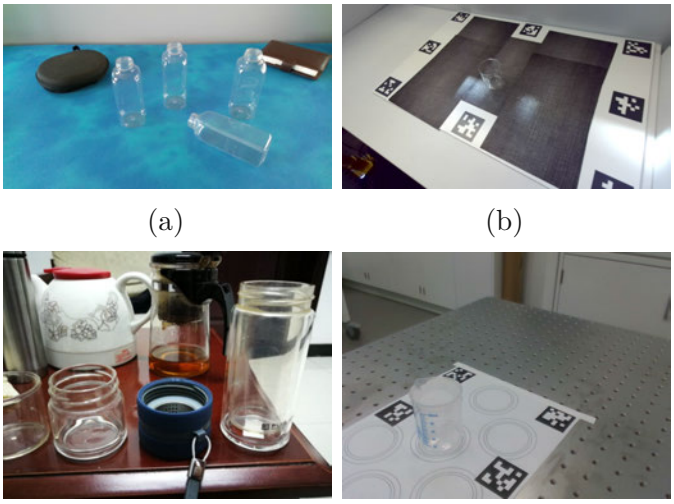
A big advantage of these synthetic datasets is that they allow an easy scale-up. Furthermore, no difficult annotation is needed and other properties such as

normals can be simulated for groundtruth. On the downside, simulation of transparent objects can be highly challenging and real-world influences might be underestimated.

Therefore, the need for real-world datasets arises, either for training or for validation purposes. For example, the KeyPose dataset [54] consists of 48,000 RGB-D images showing 15 different real-world objects and also features annotated 3D keypoints and masks. However, the dataset does not include cluttered scenes or complex object shapes.

Xie et al. [5] present a large-scale real-world dataset for transparent object segmentation called Trans10K consisting of 10,428 images of varying complexity, including occluded objects and large flat surfaces such as windows. In the follow-up work [6], object labels in 11 categories were added, but in both datasets just segmentation masks are available, but no depth information is included as groundtruth.

The datasets mentioned above only include empty transparent vessels. However, the Toronto Transparent Object Depth (TODD) Dataset also features objects filled with liquid, containing more than 15,000 RGB-D images including groundtruth mask, depth and poses [43]. On the downside, only 6 different objects are used. In Figure 2.5, example images from real-world transparent object datasets are displayed.



(a) (b)

(c) (d)

Figure 2.5: Example images of real-world datasets with transparent objects, taken from (a) ClearGrasp [4], (b) KeyPose [54], (c) Trans10K [6] and (d) TODD [56].

Another relevant dataset especially designed for robot grasping is the Dex-

NeRF dataset [52]. It includes several scenes with single or cluttered transparent objects. Up to now, only camera poses, but no groundtruth are provided.

3 Mask Prediction for Transparent Objects

In this thesis, the performance of different approaches for transparent object segmentation is evaluated. A new dataset is captured and annotated to allow a comparison of different methods.

3.1 Introduction of a New Dataset

A dataset with RGB-D images of transparent objects is created, consisting of 640 images in 10 scenes. In addition to RGB and depth images, it also contains infrared images and camera poses. Examples of information included in this dataset are shown in Figure 3.1. Since each scene is captured from 64 views, the dataset can also be utilized to evaluate multi-view reconstruction methods.

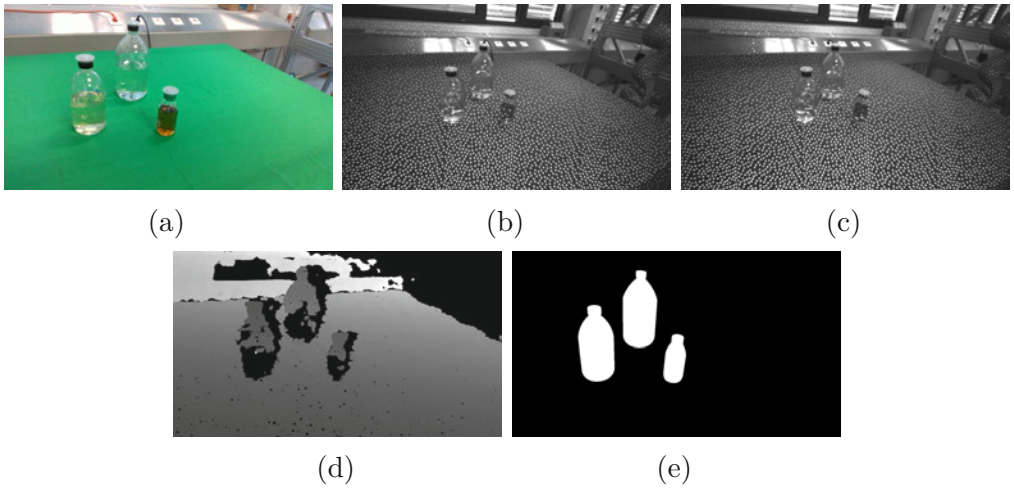


Figure 3.1: Examples of different pieces of information included in the dataset presented in this work: RGB image (a), left (b) and right (c) stereo IR images, depth data (d) and groundtruth mask (e).

Our dataset contains two different types of transparent objects: on the one hand, common transparent household items and on the other hand, transparent

containers and bottles used for various medical applications. The selection covers objects of different complexity, ranging from simple symmetric objects and flat objects to objects with more and finer details as well as vessels filled with contents. Figure 3.2 gives an overview over some scenes of our dataset.

Figure 3.2: Selected scenes from our dataset showing the wide range of objects of different complexity.

3.1.1 Acquisition

The setup for capturing the data is shown in Figure 3.3: It consists of a 7 DoF robotic manipulator with an RGB-D camera attached to its end effector. A RealSense D435 camera is selected since it allows both the collection of RGB-D images and infrared images. The eye-in-hand setup is calibrated using fiducial markers.

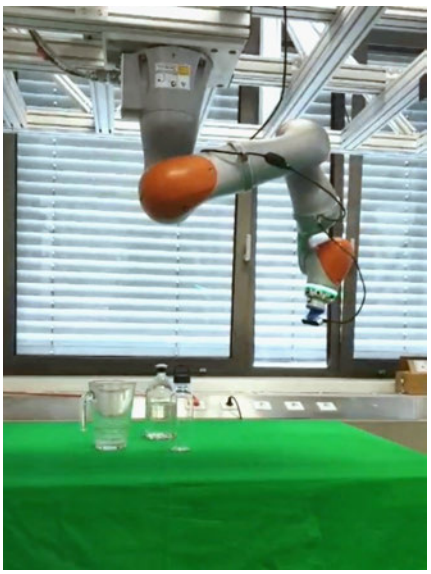


Figure 3.3: Setup consisting of a robot manipulator and an RGB-D camera.

For each scene, one or more objects are placed on a table and the robot arm moves the camera along a predefined trajectory, capturing 64 frames with

known camera poses. The camera is moved circularly around the scene and the angle between the camera and the table plane is changed every 16 frames. A sketch of camera poses corresponding to images taken during one scene is depicted in Figure 3.4.

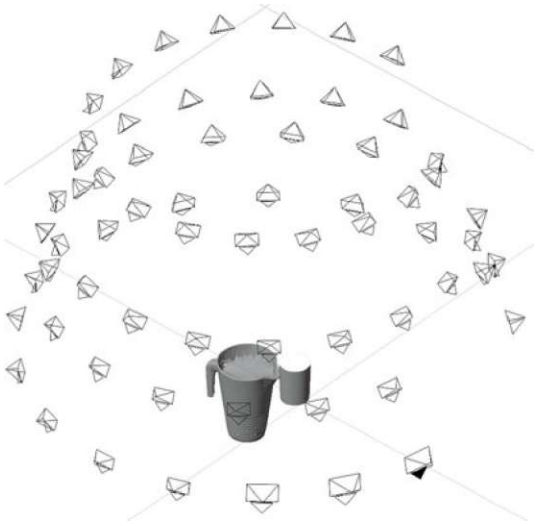


Figure 3.4: Sketch of the camera poses used to capture the data of one scene.

The distance between the camera and the objects remains in the range of 60-100cm throughout capturing the whole dataset. The angle of the camera relative to the table, however, is changed every 16 frames resulting in 4 different angles. The light source is placed head-on above the setup, therefore the camera angle also changes in regard to the light source, respectively.

3.1.2 Annotation

A tool for 3D annotation of objects introduced by [57] is used in order to render ground truth masks suitable for evaluation of the predicted silhouettes. First of all, 3D models of the transparent objects are created. Some of them are modelled using CAD software, others are obtained by spray-painting the objects and then applying object reconstruction methods for opaque objects.

The annotation of the object poses is executed carried out using Blender [58]. With knowledge of intrinsic and extrinsic camera parameters, an object pose can be obtained by alignment of the perspective projection of the 3D model with the RGB images. Figure 3.5 shows the outline of a 3D model projected on the respective RGB frame. The derived object pose can then be used to render silhouettes from any view, hence the masks belonging to the frames captured for the dataset.

(a) (b)

Figure 3.5: Outlines of the 3D models (a) are projected onto RGB images, allowing the alignment with the object contours (b).

3.2 Mask Estimation from Invalid Depth Data

Invalid or zero-value pixels in a depth image can be a cue for the presence of transparent objects. It is investigated if the invalid depth values can be used to extract silhouettes of the objects. In addition, the available RGB information is considered as well to attempt an improvement of the segmentation. Therefore, trimaps are generated from zero depth masks to allow the use of the foreground segmentation algorithm GrabCut [1] in an unsupervised manner.

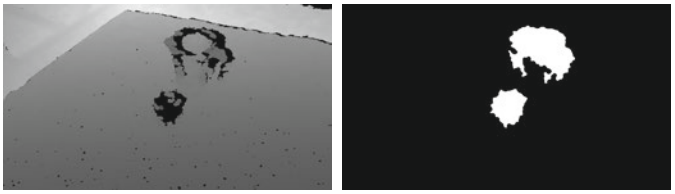
Pre-processing of Depth Images

Several processing steps are applied on the raw depth images captured with the RealSense D435 to filter the zero depth values corresponding to transparent objects. As shown in Figure 3.6 (a), depth shadows and reflective objects in the background also produce invalid depth values. In some cases, not all pixels from transparent objects are zero depth values, resulting in an incomplete silhouette either with holes or disconnected parts. To address this problem, the depth image is inverted and morphological closing with a circular kernel is applied.

Since the previous steps still leave some zero depth values from speckle noise, contour detection is used to find objects bigger than a certain size. Afterwards, again closing with a circular kernel is applied to improve the object masks. Figure 3.6 (b) shows the binary masks resulting after the morphological operations.

Unsupervised Segmentation using GrabCut

The binary masks created from zero depth values indicate that additional information must be considered to obtain silhouettes. Therefore, the zero depth



(a) (b)

Figure 3.6: Depth image of a scene with transparent objects (a) and resulting mask after processing the invalid depth values (b).

masks are used to create rough trimaps as input for image matting. First, the mask is dilated and true pixels are labelled as possible foreground. Then the original zero depth mask is eroded and the remaining pixels are labelled as definite foreground. The trimap achieved with this method is shown in Figure 3.7 (a).

Finally, the implementation of the GrabCut segmentation algorithm [1] provided by the OpenCV library [59] is applied on the RGB images corresponding to the depth images. The trimaps created in the previous step are used as initialization for GrabCut. Figure 3.7 shows the resulting segmentation (b) and binary mask (c).

(a) (b) (c)

Figure 3.7: Trimap created from the mask in Fig. 3.6 obtained from invalid depth information (a), GrabCut segmentation (b) and the new binary mask (c).

3.3 Mask Estimation from Infrared Data

Since not all transparent pixels cause zero or invalid values in the depth maps produced by the RealSense D435, another possibility to detect transparent objects is to directly use the IR images. The IR pattern is hardly visible on the surface of transparent objects, resulting in failure of the intern depth prediction of the camera.

In this work, the algorithm proposed by Ruppel et al. [2] is used to segment transparent objects in IR images. This method exploits the lack of IR speckles on transparent surfaces to obtain a transparency candidate map. Since a Structured Light camera is used in their work, the algorithm was adapted to work with the pattern of the RealSense D435 camera, which predicts depth by Active IR Stereo Vision. A transparency map is created by applying a high-pass filter to the normalised IR image. The image is then blurred with a median filter and dilated multiple times with a circular kernel. Finally, a threshold is applied and the resulting mask is filtered with a blob detection algorithm by size. Figure 3.8 shows the IR image of a scene with transparent objects (a), the transparency candidate map obtained by this method (b) and the corresponding mask resulting after thresholding (c).

(a) (b) (c)

Figure 3.8: Example of an IR image with transparent objects (a) and the transparency candidate map (b) and the corresponding binary mask (c) obtained with the method proposed by Ruppel et al. [2].

3.4 CNN-based Mask Prediction

Different deep learning pipelines for mask prediction of transparent objects are tested on the proposed dataset, discussed in Section 3.1. Only methods that can be used with RGB-D cameras and that have pre-trained models available online are selected. Table 3.1 gives an overview of these approaches and the image size used for inference.

Although TransLab [5] and Trans2Seg [6] also predict the object class, only the resulting binary masks are used for evaluation in this study. Since the authors of Trans2Seg state that the scale of the transformer does not improve without large amounts of pre-trained data, small scale transformer is used for inference.

Approach	Type	Inference Size
TransLab [5]	CNN-based	512 × 512
Trans2Seg [6]	CNN + Transformer	513 × 513
ClearGrasp [4]	CNN-based	256 × 256
TOM-Net [3]	CNN-based	448 × 448

Table 3.1: Overview of the deep learning pipelines for mask prediction of transparent objects selected for comparison.

4 Results and Discussion

The results of different transparent object segmentation techniques applied on the dataset introduced in Section 3.1 are displayed and discussed in this chapter. First, an overall comparison of the results is provided and special scenes are discussed in more detail. The comparison covers a method using invalid depth data in combination with RGB information and GrabCut [1] proposed in Section 3.2, an IR-based approach [2] as well as various CNN-based pipelines, namely TOM-Net [3], ClearGrasp [4], TransLab [5] and Trans2Seg [6]. As a baseline, zero depth masks are also included in the comparison.

4.1 Transparent Object Segmentation

Table 4.1 and Figure 4.1 show the recall, precision, F1 score and IoU achieved on our dataset. The metrics are calculated for each image of the dataset and then averaged over all images.

Method	Recall [%]	Precision [%]	F1 [%]	IoU [%]
Invalid Depth (baseline)	43.29	27.62	30.77	19.92
Depth+GrabCut [1]	58.84	39.25	43.06	30.92
IR-based [2]	37.03	41.56	33.13	25.05
TOM-Net [3]	3.57	3.45	2.96	1.84
ClearGrasp [4]	**75.86**	49.99	56.24	42.72
TransLab [5]	73.50	**71.67**	**67.54**	**55.85**
Trans2Seg [6]	54.02	65.62	52.86	41.86

Table 4.1: Overall evaluation results of selected transparent object segmentation techniques on our dataset. The highest score for each metric is highlighted in bold.

It is immediately noticeable that TransLab shows the highest precision, F1 score and IoU, and even the recall value is rather comparable to ClearGrasp. ClearGrasp has the highest recall of around 76%, but has significantly lower scores than TransLab for all other metrics. The F1 score and IoU of Trans2Seg are comparable to the results of ClearGrasp. The recall of Trans2Seg is much

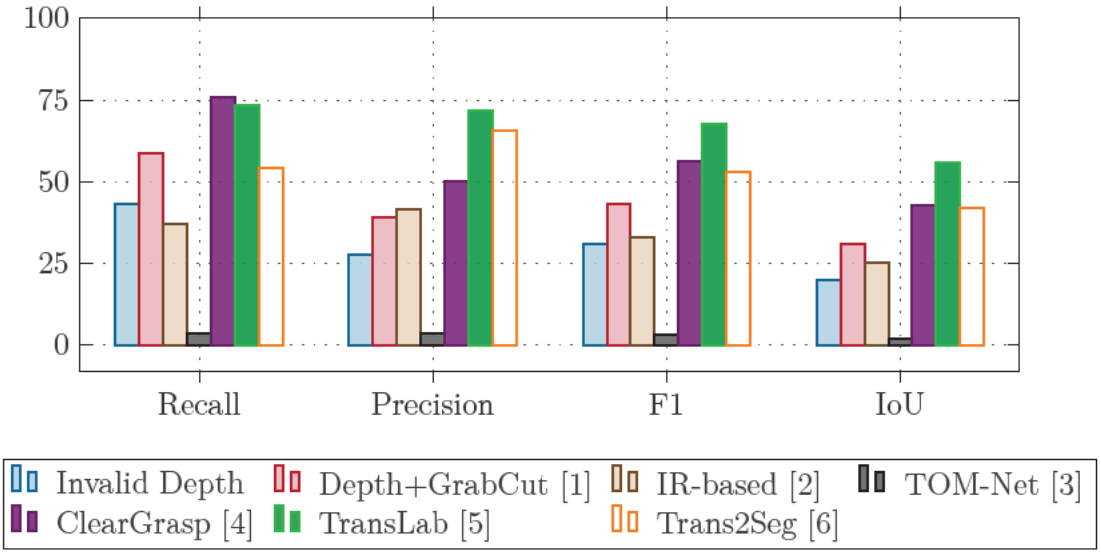


Figure 4.1: Visualisation of the evaluation results from Table 4.1.

lower, but the precision exceeds ClearGrasp by around 15% in absolute value. The GrabCut-based approach using invalid depth data clearly shows an improvement in comparison to invalid depth masks. It reaches a recall of above 58%, but has otherwise lower scores than the CNN pipelines discussed above. The IR-based approach performs worse on average than invalid depth-guided GrabCut, even yielding a recall lower than the baseline masks from invalid depth. In contrast, TOM-Net performs much worse than the invalid depth masks and every other algorithm investigated in this study. TOM-Net achieves average scores of just a few percent for each metric on the dataset.

Effect of camera pose

Since the dataset is captured at predefined camera poses (see Section 3.1), the results can also be evaluated in dependency of the camera pose relative to the setup. Figure 4.2 shows the angular dependence of each metric, highlighting that the segmentation techniques perform very differently over a wide range of camera angles (Here, an angle of 0° corresponds to a perpendicular position, while 90° corresponds to a parallel position of the image plane in relation to the table.): TransLab and Trans2Seg show the best averaged results at an angle of 14° between image and table plane, but slightly declining results at higher angles for all metrics. ClearGrasp and the segmentation methods based on invalid depth data, however, show the best results within an angle range of

30-44°. In contrast, the IR-based approach fails at angles close to perpendicular position to the table, but increases strongly with higher angles, with the best results at an angle of 55°. The TOM-Net pipeline shows a slight tendency to improvement of the results at a smaller angle, but results remain low overall.

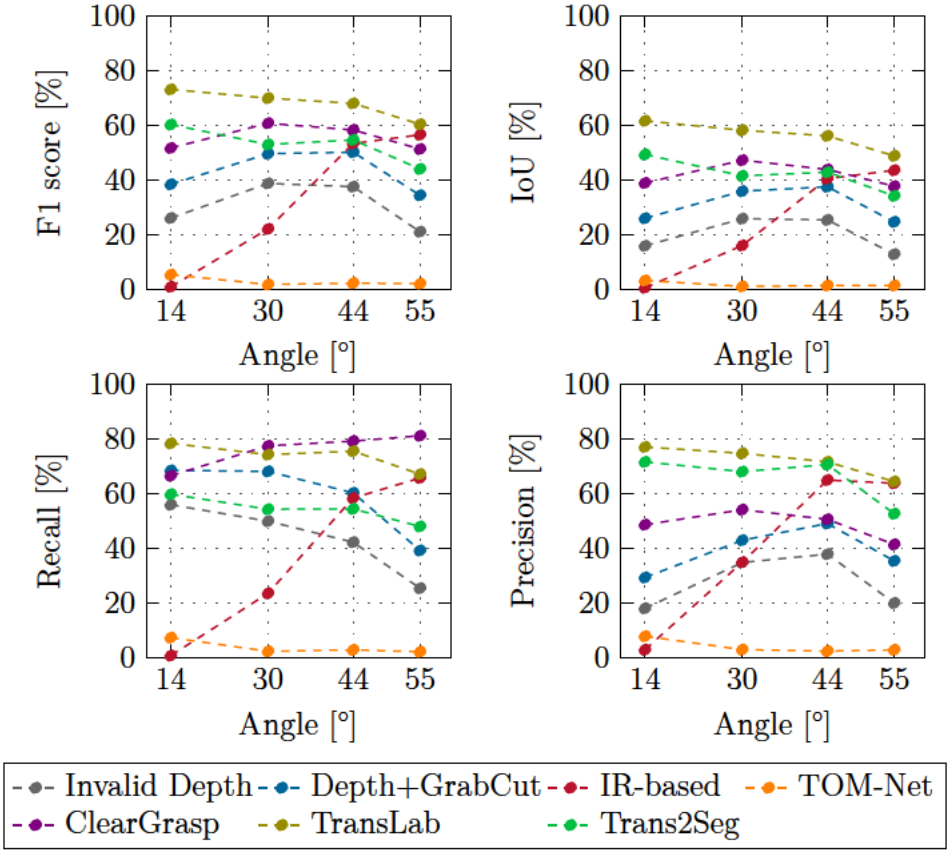


Figure 4.2: Effect of different camera poses on the metrics of the selected segmentation methods. Please note that an angle of 0° corresponds to image plane and table in perpendicular position.

4.1.1 Example 1: Glass objects with thick walls

The first scene discussed in detail contains a pitcher and a drinking glass made out of thick glass walls. Example images of the scene are shown in Figure 4.3. Table 4.2 displays the metrics achieved by methods investigated. TransLab and Trans2Seg show the best results with F1 scores above 90%. The zero depth based method and ClearGrasp also performed well, however, the baseline itself already scored an F1 above 50%. Again, TOM-Net achieved very low metrics, but still performed distinctly better than on the whole dataset.

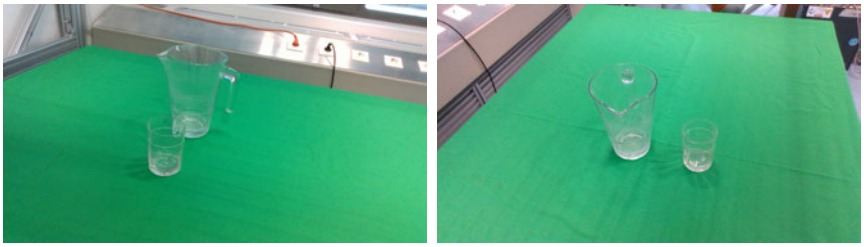
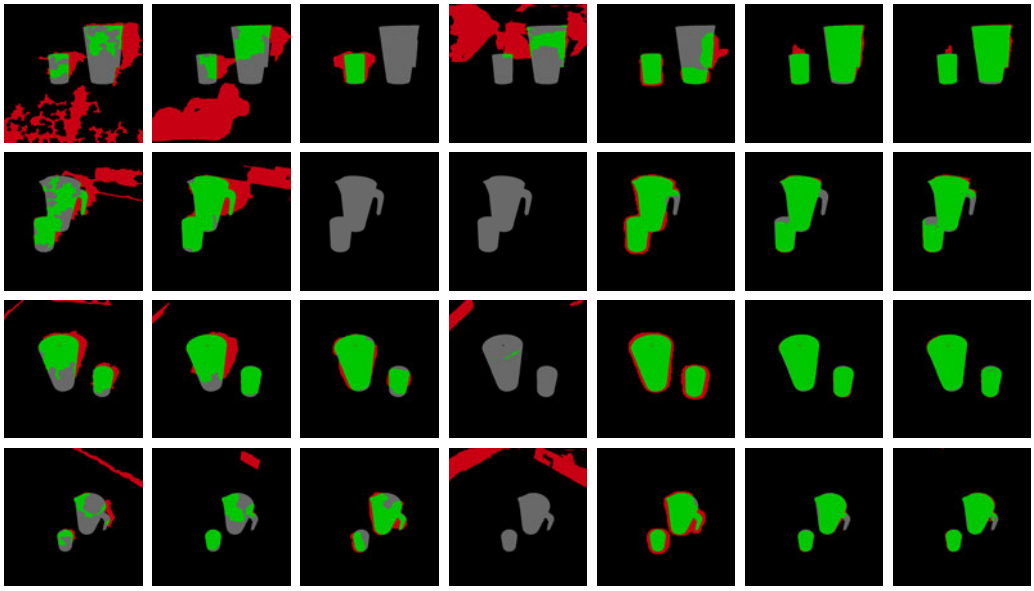


Figure 4.3: Example images of a scene containing two glass objects with thick walls.

Method	Recall [%]	Precision [%]	F1 [%]	IoU [%]
Invalid Depth	61.64	45.79	51.22	36.22
Depth+GrabCut [1]	79.59	60.62	67.5	53.14
IR-based [2]	46.41	51.08	46.79	39.75
TOM-Net [3]	9.64	12.54	9.74	6.74
ClearGrasp [4]	93.04	66.09	76.09	62.61
TransLab [5]	92.78	**93.92**	**93.13**	**87.39**
Trans2Seg [6]	**93.32**	90.98	91.41	85.21

Table 4.2: Evaluation results of transparent object segmentation for a scene containing two glass objects with thick walls.

In Figure 4.4 TransLab and Trans2Seg achieve a decent segmentation of the objects. However, the handle of the pitcher is not segmented by any of the pipelines. Also, both pipelines struggle with the metallic structure in the background of the first frame - a part of the material's shiny surface is falsely considered as part of the drinking glass. ClearGrasp is also able to recognise the objects in most frames, but fails to predict parts of the pitcher in some cases. For example, in the first frame in Figure 4.4, the pitcher stands in front of the edge of the table. Such an inhomogeneous background can



(a) (b) (c) (d) (e) (f) (g)

Figure 4.4: Segmentation results of selected methods on a scene with thick-walled glass objects: (a) invalid depth mask, (b) depth + GrabCut [1], (c) IR-based [2], (d) TOM-Net [3], (e) ClearGrasp [4], (f) TransLab [5] and (g) Trans2Seg [6].

cause problems for the mask prediction. In comparison, TOM-Net fails to detect the object in all example frames and falsely labels background pixels as transparent. The IR-based method shows behaviour similar to TOM-Net in the first two frames, but manages to predict rough silhouettes of the objects in the frames taken from a higher angle. The evaluation of the invalid depth masks reveals that only parts of the transparent objects cause zero depth values, so no accurate silhouettes are obtained. The GrabCut algorithm is able to restore the silhouette of the drinking glass in some frames, but struggles with correctly detecting the pitcher.

4.1.2 Example 2: Plastic objects with thin walls

Besides glass, plastic is the other major material category for transparent objects. Therefore, two plastic objects with thin walls are investigated in the scene pictured in Figure 4.5, to be more precise, a bottle standing upright and a cylindrical container lying flat on the table. In Table 4.3 the averaged metrics over all frames of the scene are listed. In this case, TransLab clearly outperforms the other methods in terms of precision, F1 score and IoU. However, values are much lower compared to Example 1 discussed above, highlighting the challenging nature of the materials used in this scene. ClearGrasp shows the highest recall at the cost of lower precision. Surprisingly, the follow-up work of TransLab, Trans2Seg, has a much lower recall than TransLab., and also performs significantly worse in terms in terms of F1 score and IoU. Since TOM-Net wrongly estimates the structure behind the actual transparent objects as transparent pixels, the metrics shown in Table 4.3 do not give the full picture.

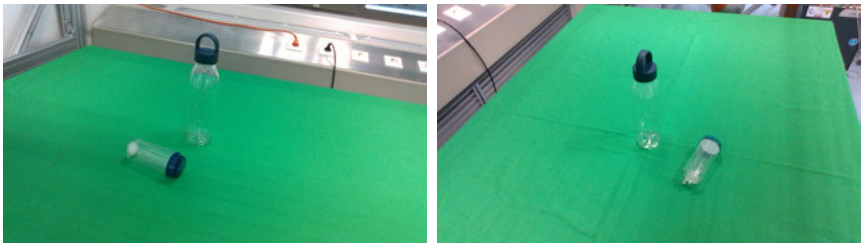
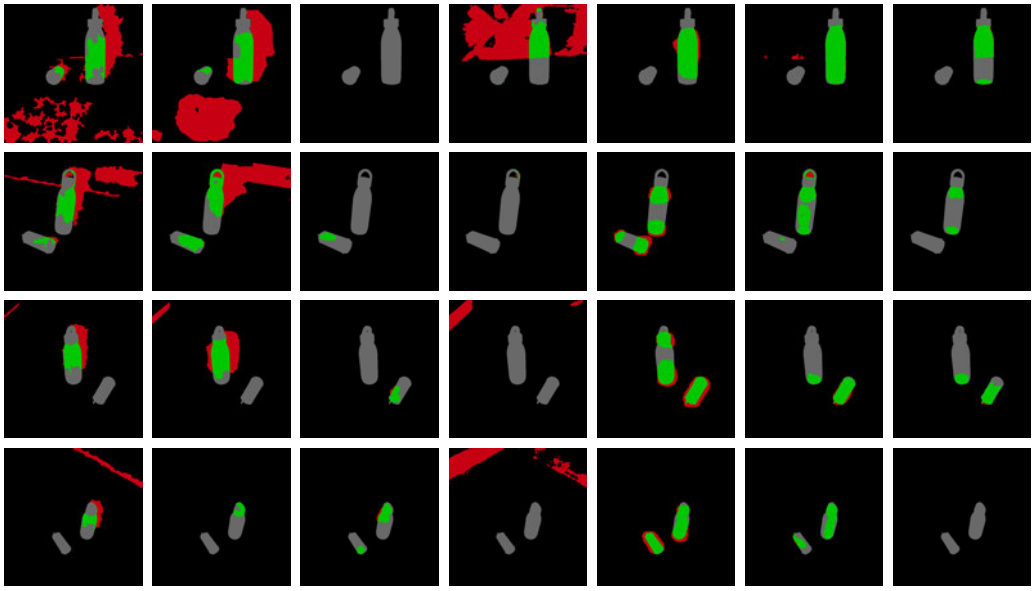


Figure 4.5: Example images of a scene containing two thin-walled plastic objects.

Method	Recall [%]	Precision [%]	F1 [%]	IoU [%]
Invalid Depth	39.84	24.30	28.09	16.71
Depth+GrabCut [1]	51.89	32.36	36.85	23.70
IR-based [2]	16.15	48.42	22.05	14.05
TOM-Net [3]	4.66	4.53	3.32	1.83
ClearGrasp [4]	**72.59**	57.73	61.46	45.76
TransLab [5]	55.81	**88.77**	**66.25**	**51.44**
Trans2Seg [6]	26.01	69.11	33.86	22.53

Table 4.3: Evaluation results of transparent object segmentation for a scene containing two plastic objects with thin walls.

The mask prediction results in Figure 4.6 clearly show that this scene is more challenging than the previous one. The upright standing bottle is recognised



(a) (b) (c) (d) (e) (f) (g)

Figure 4.6: Segmentation results of selected methods on a scene with thin-walled plastic objects: (a) invalid depth mask, (b) depth + GrabCut [1], (c) IR-based [2], (d) TOM-Net [3], (e) ClearGrasp [4], (f) TransLab [5] and (g) Trans2Seg [6].

in some cases by TransLab, Trans2Seg and ClearGrasp, but more often just parts of the bottle are predicted correctly, mostly the bottom parts touching the table. The invalid depth mask on the other hand covers the middle part of the bottle. The container placed with its side on the table is not detected in most frames by any of the methods investigated. However, in the third frame, TransLab, Trans2Seg and ClearGrasp are able to predict the mask of the lying container. A reason could be that in the corresponding RGB image prominent light speckles are visible on the surface of the container. Light speckles are a strong cue for transparent surfaces, what presumably has been learned by the CNNs during training stage, revealing the very importance of lighting for transparent objects detection. The IR image-based mask prediction approach recognises some parts of the transparent objects in the frames captured at wider angles (44-55°), but is not able to recover a decent silhouette. TOM-Net [3] could not predict any transparent pixels of the objects. Instead, the metallic structure in the background of the scene was falsely labelled as transparent.

4.1.3 Example 3: Objects lying flat on the table

Another scene investigated more thoroughly contains objects like a pipe and a dust pan, shown in Figure 4.7. Training data for transparent object segmentation mostly contains cylindrical vessels like cups or bottles that stand upright and where a big part of the object is not in contact with the ground. The objects in this scene however are lie flat and, in addition, the dust pan is made of milky semi-transparent plastic. Therefore, this scene might drastically differ from training data for the learning-based methods investigated in this study. Accordingly, the results in Table 4.4 show that of all methods only TransLab yields a recall, an F1 score and IoU above 50%. For precision, the IR based methods and Trans2Seg yield values above 50%, which is still much lower than the approx. 91% obtained by TransLab. The visualisation of the results in Figure 4.8 paints the same picture. TransLab achieves the overall best segmentation results, followed by Trans2Seg and ClearGrasp.

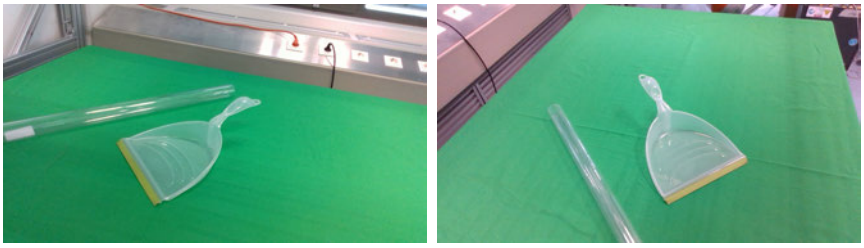


Figure 4.7: Example images of a scene containing two objects lying flat.

Method	Recall [%]	Precision [%]	F1 [%]	IoU [%]
Invalid Depth	16.99	22.70	16.32	9.45
Depth+GrabCut [1]	24.26	36.06	25.23	16.46
IR-based [2]	20.16	55.39	27.71	17.93
TOM-Net [3]	7.40	7.63	6.17	3.50
ClearGrasp [4]	36.75	44.78	37.78	25.59
TransLab [5]	**69.23**	**90.89**	**76.56**	**64.40**
Trans2Seg [6]	39.37	71.67	46.36	35.09

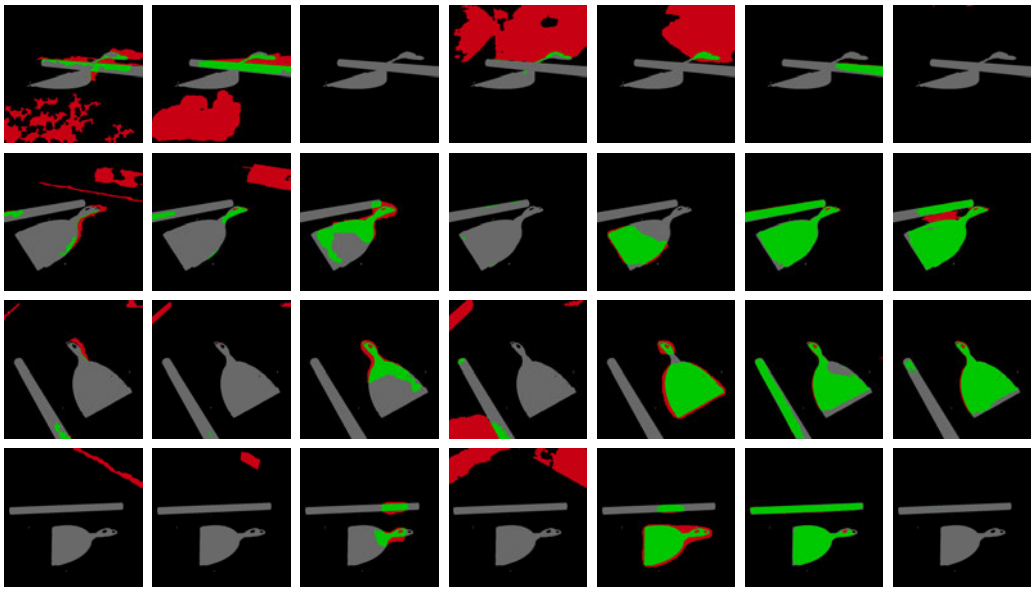
Table 4.4: Evaluation results of transparent object segmentation for a scene containing two flat lying objects.

The first example frame in Figure 4.8 presents a case, where none of the tested methods is able to predict the object silhouettes. Only TransLab could detect a part of the pipe in the image. For Trans2Seg and ClearGrasp, the

pipe was quite challenging. Both networks just predicted small parts of the pipe or nothing at all, resulting in a recall under 50% for the whole scene. A reason could be its unusual dimensions that may not be covered in the training datasets.

Regarding the dust pan, Trans2Seg and ClearGrasp show promising results, but the predicted silhouettes are not always complete. Especially the handle of the dust pan causes problems for ClearGrasp.

The depth-based methods and TOM-Net struggle to detect the transparent objects in this scene and incorrectly label lots of background pixels as transparent instead. The IR-based approach however is able to correctly predict some parts of the transparent objects, but does not provide decent silhouettes.



(a) (b) (c) (d) (e) (f) (g)

Figure 4.8: Segmentation results of selected methods on a scene with two flat lying objects: (a) invalid depth mask, (b) depth + GrabCut [1], (c) IR-based [2], (d) TOM-Net [3], (e) ClearGrasp [4], (f) TransLab [5] and (g) Trans2Seg [6].

4.1.4 Example 4: Sterility kit

Furthermore, a scene with a rather complex medical object is discussed, i.e. the sterility kit shown in Figure 4.9. Difficulty arises here from the proximity of the two containers to each other and finer details including the plastic tubes. Also, transparent packaging is placed near the sterility kit. As seen in Example 2 in Section 4.1.2, plastic objects especially if thin-walled prove to be quite challenging for the approaches selected in this study.

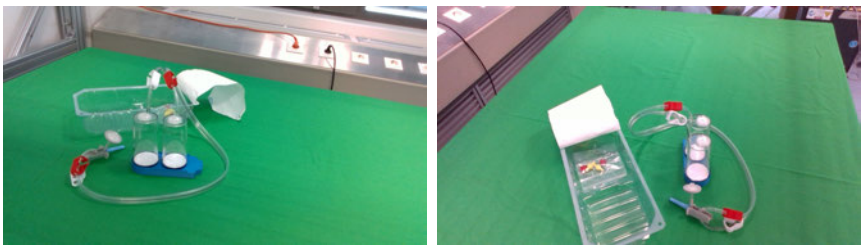


Figure 4.9: Example images of a scene containing a medical kit and packaging.

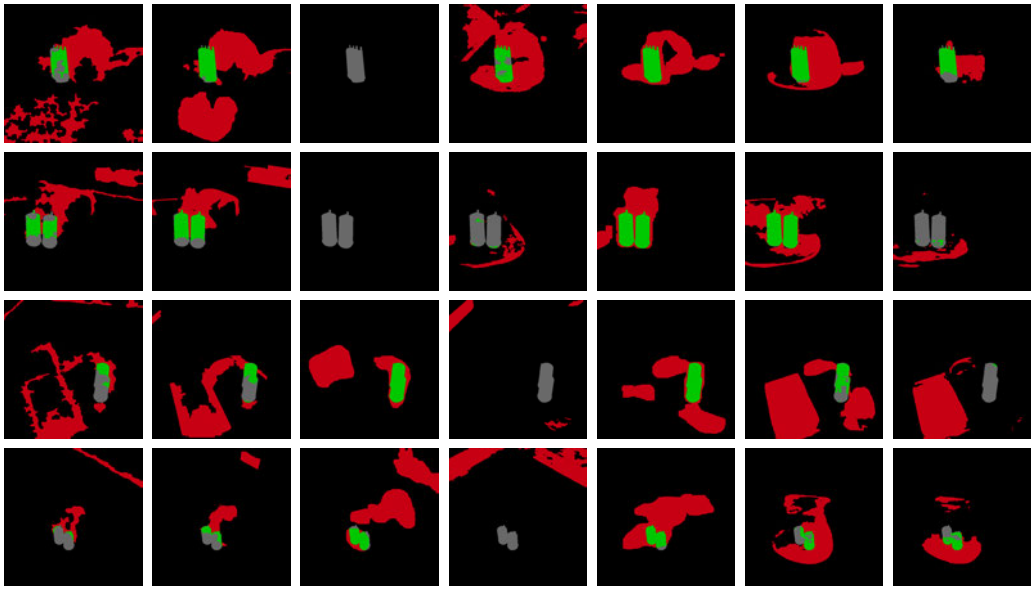
Please note that the ground truth available in this case only covers the containers. Therefore, the other transparent components are treated incorrectly in the evaluation. For example, some approaches recognised parts of the tube or the packaging, which is not considered in the results in Table 4.5. Therefore, the recall of TransLab, Trans2Seg and ClearGrasp is relatively high, but many "false positive" pixel predictions cause a lower precision than it is actually the case.

Method	Recall [%]	Precision [%]	F1 [%]	IoU [%]
Invalid Depth	47.46	8.39	13.78	7.52
Depth+GrabCut [1]	65.13	12.79	20.62	11.70
IR-based [2]	36.55	6.95	11.25	6.92
TOM-Net [3]	14.14	4.69	5.74	3.18
ClearGrasp [4]	**87.91**	**20.02**	**31.39**	**19.52**
TransLab [5]	71.47	17.59	27.45	16.70
Trans2Seg [6]	54.08	15.45	22.58	13.87

Table 4.5: Evaluation results of transparent object segmentation for a scene with a medical object.

The recognition of the containers proves to be rather difficult: The two containers are not separated from each other or from the tubes and other object parts in the segmentation masks. In Figure 4.10, TransLab quantitatively

achieves the best mask prediction: It partly recognises the containers, tubes and the packaging. ClearGrasp effectively locates the containers but struggles with the tubes and the packaging. Trans2Seg fails to detect the containers in some example frames, but is able to segment parts of the tubes and the packaging. TOM-Net detects parts of the objects in the close-up frames, but fails when the image is taken from further away. The IR-based approach is able to segment the transparent containers in some of the frames, but mislabels the white paper of the packaging as transparent.



(a) (b) (c) (d) (e) (f) (g)

Figure 4.10: Segmentation results of selected methods on a scene with a medical object: (a) invalid depth mask, (b) depth + GrabCut [1], (c) IR-based [2], (d) TOM-Net [3], (e) ClearGrasp [4], (f) TransLab [5] and (g) Trans2Seg [6].

4.2 3D Reconstruction

The predicted silhouettes of transparent objects can be used for 3D reconstruction via SfS techniques. For qualitative evaluation, the mask predictions from 16 views of the scene in Section 4.1.1 were used as input for the for mesh reconstruction. An open source implementation of the voxel carving algorithm in C++ is used, supporting Truncated Signed Distance Function (TSDF) fusion and marching cubes [60]. Figure 4.11 shows the meshes created from the groundtruth silhouettes (a) and from the masks predicted by GrabCut (b), ClearGrasp (c) and TransLab (d) by this technique.

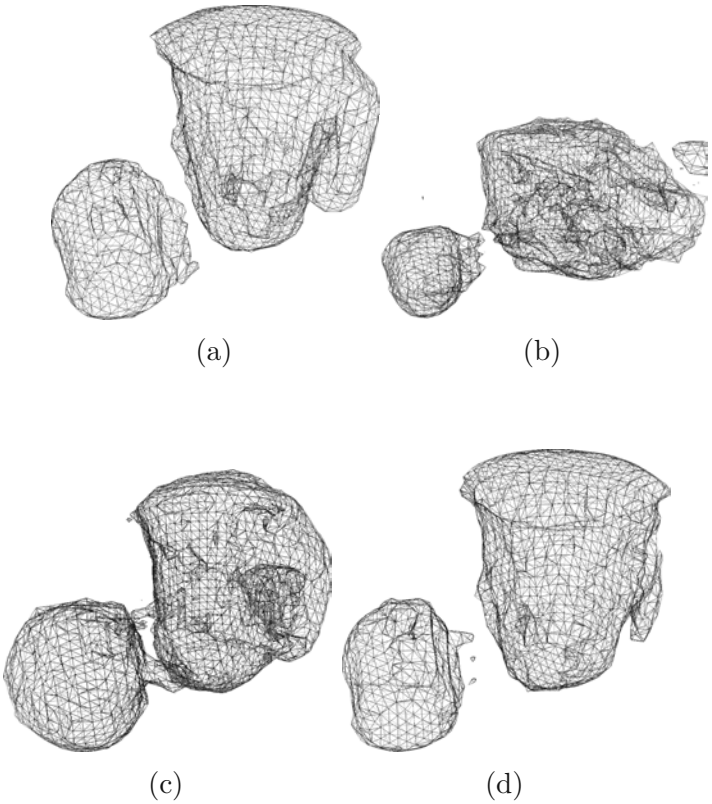


Figure 4.11: Comparison of meshes obtained by voxel carving [60] from (a) groundtruth silhouettes and from the masks predicted by (b) GrabCut, (c) ClearGrasp and (d) TransLab.

The mesh obtained from carving groundtruth silhouettes (see Figure 4.11 (a)) shows that the reconstruction with this method is not perfect, but good enough to get a recognisable shape. The GrabCut silhouettes produce a rough shape for the glass, but fail to reconstruct the water jug. Nevertheless, the object are separated from each other. In contrast, this is not the case with the

mesh generated from ClearGrasp silhouettes. Here, the two object were not separated completely. The drinking glass appears ball-like, but the shape of the water jug is better recognisable than in the previous case. The quality of the mesh carved from TransLab silhouettes is comparable to the 3D model created from groundtruth data. The shape of the glass and the water jug are recognisable and well separated, only the handle as a finer detail could not be reconstructed completely. Therefore, we could successfully demonstrate that 3D reconstruction via Sfs techniques is feasible for transparent objects. The quality of silhouettes is vital for the overall 3D reconstruction, but as demonstrated silhouettes from e.g. TransLab are already of sufficient quality to obtain recognisable 3D shapes of simple transparent objects.

4.3 Discussion

Trans2Seg [6] is the follow-up work of TransLab [5]. Although it was trained on an extended dataset, it does not perform better than TransLab. A reason for the inferior results could be that the "tiny" configuration of the transformer was used for inference. The output masks obtained by ClearGrasp [4] are already dilated, since the pipeline subsequently uses them as input for depth completion. The idea behind this was to ensure that all transparent depth pixels are removed before depth completion. Therefore, the high recall found in our study can partially be attributed to the dilation of the masks.

TOM-Net [3] fails completely on our proposed dataset. It was trained just on synthetic data, where transparent objects were rendered in front of patterned backgrounds. The other networks used in this comparison all were trained on real-world images and performed much better. This is a strong cue that it is not sufficient to use synthetic training data and that more diverse background is also important.

The IR-based method based on Ruppel et al. [2] works very differently from the others tested in this study. It uses a traditional segmentation approach based on morphological operations applied on infrared images. An exact segmentation is not the main goal of the pipeline, but the results suggest that there is potential in using infrared images for the task of transparent object detection and segmentation. However, infrared images captured with the camera close to the scene have the problem of too strong reflections. Therefore, the angular dependency is most pronounced for the IR-based method in this study, since the table plane is closer at low camera angles. Comparing the IR images of scene 1 and scene 2 (see Figure 4.12), it can be observed that for thick glass materials (left) the IR laser speckles are absorbed (low intensity) while the intensity remains nearly the same as on the table for plastic objects (right). An approach using the dilution of IR speckles therefore works much better with glass objects and has to be adapted for different transparent materials.

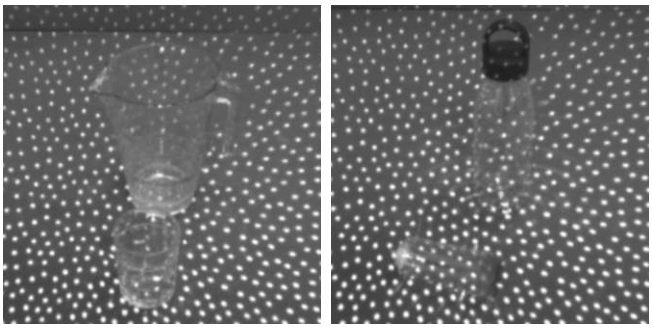


Figure 4.12: Comparison of IR images of different scenes.

Overall, it also must be mentioned that deviations of the camera poses and the semi-manual annotation of the ground truth impose a certain error. Especially the annotation of the scenes containing the sterility kit is not complete, since it only takes into account the containers and not the cables and other non-rigid (semi-)transparent details.

5 Conclusion

This work provides a comparison between different state-of-the-art approaches for mask prediction of transparent objects in order to point out which cases are difficult for current methods. To achieve this, a new real-world dataset was introduced featuring different scenes showing transparent objects. Mainly vessels like bottles or cups were used, but also objects like a dust pan or a pipe were included as well as a more complex medical object. The dataset provides RGB-D images, infrared images and camera poses, as well as groundtruth masks that were annotated manually. This dataset was used to compare the performance of a method combining invalid depth and GrabCut, an IR-based approach by Ruppel et al. [2], TOM-Net [3], ClearGrasp [4], TransLab [5] and Trans2Seg [6]. For evaluation, metrics were calculated and averaged over all frames of the dataset. Here, the performance varied greatly between the compared approaches. The best overall performance was achieved by TransLab in regards of precision, F1 score and IoU, yielding values up to 71.67%, 67.54% and 55.85%, respectively. The average highest recall was observed for ClearGrasp with 75.86%, but TransLab also shows a quite comparable recall of 73.50%. This clearly suggests that TransLab is the most effective approach in this comparison. This work also indicates that TOM-Net is the least suitable for our dataset with metrics well below 4%. However, it has to be considered that only pre-trained methods are used in this study and TransLab is trained on the most extensive and varied dataset of the compared pipelines.

Overall, the mask prediction worked well for simple and thick-walled objects, but struggled with thin plastic objects and transparent objects in close contact with the ground. Complex objects like the medical kit were very challenging for all approaches due to the fine details like tubes and the cluttered appearance.

Although invalid depth values are a strong cue for transparency, they are not robust enough for detection of transparent objects depending on the material. For instance, thin transparent plastic objects appear nearly invisible in depth images. Trimaps generated from this invalid depth are also not accurate enough for use with image matting approaches. Furthermore, the predicted silhouettes were incorporated in a voxel carving algorithm and it could be shown that it is possible to get a rough 3D shape estimation.

5.1 Outlook

The presented work shows that the segmentation of transparent objects is an ongoing challenge. Transparent objects containing liquids, very thin plastic objects, semi-transparent objects and fine structures like handles or cables are cases that still need to be tackled. Especially, an approach that takes into account all these cases and still is able to handle opaque objects is needed. Many approaches just consider one of these problems, like for instance the approach by Song et al. [61], which focuses on depth estimation of translucent boxes.

While CNNs show the best results in this work, the evaluation indicates that there could be an improvement by training with more diverse data. Although there are approaches that directly predict the depth of transparent objects, improving the segmentation of transparent objects is still promising since depth prediction of transparent objects, like e.g. ClearGrasp [4], benefits from an accurate mask prediction, as is shown by Xie et al. [5].

Bibliography

[1] C. Rother, V. Kolmogorov, and A. Blake, „"GrabCut": interactive foreground extraction using iterated graph cuts," *ACM SIGGRAPH 2004 Papers*, 2004.

[2] P. Ruppel, M. Görner, N. Hendrich, and J. Zhang, „Detection and Reconstruction of Transparent Objects with Infrared Projection-based RGB-D Cameras," in *International Conference on Cognitive Systems and Information Processing (ICCSIP)*, 2020.

[3] G. Chen, K. Han, and K.-Y. K. Wong, „TOM-Net: Learning Transparent Object Matting from a Single Image," in *CVPR*, 2018.

[4] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, „Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3634 3642.

[5] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, „Segmenting Transparent Objects in the Wild," *arXiv preprint arXiv:2003.13948*, 2020.

[6] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, „Segmenting Transparent Objects in the Wild with Transformer," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, 2021.

[7] N. Schlüter and F. Faul, „Visual shape perception in the case of transparent objects," *Journal of Vision*, vol. 19, no. 4, p. 24, 2019.

[8] R. Szeliski, *Computer Vision: Algorithms and Applications, 2nd ed.* Springer London, 2022, ISBN: 978-1-84882-934-3. [Online]. Available: `http://szeliski.org/Book/`.

[9] Y. Boykov, O. Veksler, and R. Zabih, „Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222 1239, 2001.

[10] J. S. Yedidia, W. Freeman, and Y. Weiss, „Generalized Belief Propagation," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13, MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf.

[11] J. Long, E. Shelhamer, and T. Darrell, „Fully Convolutional Networks for Semantic Segmentation," Nov. 2014. arXiv: 1411.4038 [cs.CV].

[12] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, „Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1 1, 2021.

[13] K. Simonyan and A. Zisserman, „Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014. arXiv: 1409.1556 [cs.CV].

[14] H. Noh, S. Hong, and B. Han, „Learning Deconvolution Network for Semantic Segmentation," May 2015. arXiv: 1505.04366 [cs.CV].

[15] O. Ronneberger, P. Fischer, and T. Brox, „U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015. arXiv: 1505.04597 [cs.CV].

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, „Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," Feb. 2018. arXiv: 1802.02611 [cs.CV].

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking Atrous Convolution for Semantic Image Segmentation*, 2017.

[18] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, and T. S. Huang, „CCNet: Criss-Cross Attention for Semantic Segmentation," Nov. 2018. arXiv: 1811.11721 [cs.CV].

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2020.

[20] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, *Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers*, 2020.

[21] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, „Segmenter: Transformer for Semantic Segmentation," May 2021. arXiv: 2105.05633 [cs.CV].

[22] K. McHenry, J. Ponce, and D. A. Forsyth, „Finding glass,“ *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 973 979 vol. 2, 2005.

[23] N. Alt, P. Rives, and E. Steinbach, „Reconstruction of transparent objects in unstructured scenes with a depth camera,“ in *2013 IEEE International Conference on Image Processing*, IEEE, 2013.

[24] A. Hagg, F. Hegger, and P. G. Plöger, „On Recognizing Transparent Objects in Domestic Environments Using Fusion of Multiple Sensor Modalities,“ in *RoboCup 2016: Robot World Cup XX*, Springer International Publishing, 2017, pp. 3 15.

[25] M. P. Khaing and M. Masayuki, „Transparent Object Detection Using Convolutional Neural Network,“ in, Springer Singapore, 2018, pp. 86 93.

[26] A. Torres-Gomez and W. Mayol-Cuevas, „Recognition and reconstruction of transparent objects for augmented reality,“ in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, IEEE, 2014.

[27] S.-K. Yeung, T.-P. Wu, C.-K. Tang, T. F. Chan, and S. Osher, „Adequate reconstruction of transparent objects on a shoestring budget,“ in *CVPR 2011*, IEEE, 2011.

[28] B. Wu, Y. Zhou, Y. Qian, M. Cong, and H. Huang, „Full 3D reconstruction of transparent objects,“ *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1 11, 2018.

[29] J. Shi, Y. Dong, H. Su, and S. X. Yu, „Learning Non-Lambertian Object Intrinsics Across ShapeNet Categories,“ in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, „Mask R-CNN,“ Mar. 2017. arXiv: `1703.06870 [cs.CV]`.

[31] A. H. Madessa, J. Dong, X. Dong, Y. Gao, H. Yu, and I. Mugunga, „Leveraging an Instance Segmentation Method for Detection of Transparent Materials,“ IEEE, 2019.

[32] J. Stets, Z. Li, J. R. Frisvad, and M. Chandraker, „Single-Shot Analysis of Refractive Shape Using Convolutional Neural Networks,“ in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2019.

[33] H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, V. Prinet, and L. Weng, „Enhanced Boundary Learning for Glass-like Object Segmentation,“ Mar. 2021. arXiv: `2103.15734 [cs.CV]`.

[34] T. Wang, X. He, and N. Barnes, „Glass object segmentation by label transfer on joint depth and appearance manifolds," in *2013 IEEE International Conference on Image Processing*, IEEE, 2013.

[35] Y. Ji, Q. Xia, and Z. Zhang, „Fusing Depth and Silhouette for Scanning Transparent Object with RGB-D Sensor," vol. 2017, pp. 1 11, 2017.

[36] C. Guo-Hua, W. Jun-Yi, and Z. Ai-Jun, „Transparent object detection and location based on RGB-D camera," *Journal of Physics: Conference Series*, vol. 1183, p. 012 011, 2019.

[37] K. Maeno, H. Nagahara, A. Shimada, and R.-I. Taniguchi, „Light Field Distortion Feature for Transparent Object Recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2013.

[38] Y. Xu, H. Nagahara, A. Shimada, and R.-I. Taniguchi, „TransCut: Transparent Object Segmentation from a Light-Field Image," IEEE, 2015.

[39] , „TransCut2: Transparent Object Segmentation From a Light-Field Image," *IEEE Transactions on Computational Imaging*, vol. 5, no. 3, pp. 465 477, 2019.

[40] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, „Deep Polarization Cues for Transparent Object Segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8599 8608.

[41] G. Saygili, L. V. D. Maaten, and E. A. Hendriks, „Hybrid Kinect Depth Map Refinement for Transparent Objects," in *2014 22nd International Conference on Pattern Recognition*, IEEE, 2014.

[42] Y. Zhang and T. Funkhouser, „Deep Depth Completion of a Single RGB-D Image," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2018.

[43] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, „Seeing Glass: Joint Point Cloud and Depth Completion for Transparent Objects," Sep. 2021. arXiv: `2110.00087 [cs.CV]`.

[44] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, „RGB-D Local Implicit Function for Depth Completion of Transparent Objects," IEEE, 2021.

[45] Y. Qian, M. Gong, and Y.-H. Yang, „3D Reconstruction of Transparent Objects with Position-Normal Consistency," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016.

[46] Z. Li, Y.-Y. Yeh, and M. Chandraker, „Through the Looking Glass: Neural 3D Reconstruction of Transparent Shapes," IEEE, 2020.

[47] U. Klank, D. Carton, and M. Beetz, „Transparent object detection and reconstruction on a mobile platform," in *2011 IEEE International Conference on Robotics and Automation*, IEEE, 2011.

[48] I. Lysenkov and V. Rabaud, „Pose estimation of rigid transparent objects in transparent clutter," in *2013 IEEE International Conference on Robotics and Automation*, IEEE, 2013.

[49] C. J. Phillips, M. Lecce, and K. Daniilidis, „Seeing Glassware: from Edge Detection to Pose Estimation and Shape Recovery," in *Robotics: Science and Systems*, 2016.

[50] A. Xompero, R. Sanchez-Matilla, A. Modas, P. Frossard, and A. Cavallaro, „Multi-View Shape Estimation of Transparent Containers," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020.

[51] S. Albrecht and S. R. Marsland, „Seeing the Unseen: Simple Reconstruction of Transparent Objects from Point Cloud Data," 2013.

[52] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, „Dex-NeRF: Using a Neural Radiance Field to Grasp Transparent Objects," *ArXiv*, vol. abs/2110.14217, 2021.

[53] I. Lysenkov, V. Eruhimov, and G. Bradski, „Recognition and Pose Estimation of Rigid Transparent Objects with a Kinect Sensor," in *Robotics: Science and Systems VIII*, Robotics: Science and Systems Foundation, 2012.

[54] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, „KeyPose: Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2020.

[55] C. Xu, J. Chen, M. Yao, J. Zhou, L. Zhang, and Y. Liu, „6DoF Pose Estimation of Transparent Object from a Single RGB-D Image," vol. 20, no. 23, p. 6790, 2020.

[56] H. Xu, *TODD dataset*, 2021.

[57] M. Suchi, B. Neuberger, T. Patten, and M. Vincze, „3D-SADT: Simple Annotation & Dataset Toolkit for Robotic Vision," Submitted to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2022.

[58] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: `http://www.blender.org`.

[59] Itseez, *Open Source Computer Vision Library*, `https://github.com/itseez/opencv`, 2015.

[60] *Vacancy: A Voxel Carving implementation in C++*, online, 2021. [Online]. Available: `https://github.com/unclearness/vacancy.git`.

[61] S. Song and H. Shim, „Depth Reconstruction of Translucent Objects from a Single Time-of-Flight Camera Using Deep Residual Networks," in *Computer Vision ACCV 2018*, Springer International Publishing, 2019, pp. 641 657.