Deep Polarization Cues for Transparent Object Segmentation

Agastya Kalra[1], Vage Taamazyan[1], Supreeth Krishna Rao[1], Kartik Venkataraman[1], Ramesh Raskar*[1,2], and Achuta Kadambi*[1,3]

[1]Akasha Imaging
[2]MIT Media Lab
[3]University of California, Los Angeles (UCLA)

Abstract

Segmentation of transparent objects is a hard, open problem in computer vision. Transparent objects lack texture of their own, adopting instead the texture of scene background. This paper reframes the problem of transparent object segmentation into the realm of light polarization, i.e., the rotation of light waves. We use a polarization camera to capture multi-modal imagery and couple this with a unique deep learning backbone for processing polarization input data. Our method achieves instance segmentation on cluttered, transparent objects in various scene and background conditions, demonstrating an improvement over traditional image-based approaches. As an application we use this for robotic bin picking of transparent objects.

1. Introduction

Transparent objects occur in manufacturing, life sciences, and automotive industries. In contrast to conventional objects, transparent objects lack texture of their own. As a result, it is hard to segment transparent objects captured with standard imaging—segmentation algorithms do not have any texture to latch on to. In this paper, we bring transparent object segmentation to the realm of polarization imaging. As shown in Figure 1, the polarization imagery of transparent objects visualizes their very unique texture. There is a geometry-dependent signature on edges and a very unique pattern arises in the angle of linear polarization. The object's *intrinsic texture* is more visible in the polarization than in just in the intensity. Unfortunately, the peculiar texture of polarization requires re-examination for deep learning in the context of polarization imagery.

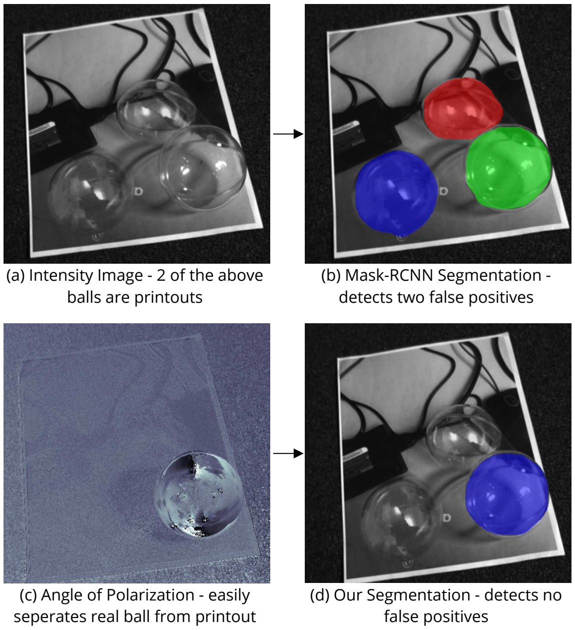
(a) Intensity Image - 2 of the above balls are printouts
(b) Mask-RCNN Segmentation - detects two false positives
(c) Angle of Polarization - easily seperates real ball from printout
(d) Our Segmentation - detects no false positives

Figure 1: **Polarized CNNs leverage unique texture from polarized imagery for robust segmentation.** Standard instance segmentation is unable to differentiate between a print-out spoof and a real ball. Our segmentation is able to robustly segment the real ball using the unique polarization texture. Our paper studies this textural motif and designs a customized deep learning architecture.

In this paper, we introduce a new deep learning framework for polarization-based segmentation of transparent objects. We refer to our framework as a Polarized Convolutional Neural Network (Polarized CNN). Our framework consists of a novel backbone that is suitable for processing the peculiar texture of polarization and can be coupled with architectures like Mask R-CNN (e.g. Polarized Mask R-CNN) to produce accurate and robust solution for instance segmentation of transparent objects. We summarize our

*Contributions from these authors occurred wholly during their consulting periods at Akasha Imaging.

Method	Clutter Accuracy	Environment Assumption	Print-Out spoofs	Sensors	Physical Limitations
Intensity Mask R-CNN [19]	Medium	Seen during Training	Not robust	1	Optical
Light Field [45]	Not designed for instance segmentation	None	Robust	1	Range limits
RGBD [44, 12]	Not designed for instance segmentation	Indoors	Robust	2+	Range limits
Polarized CNN **Proposed**	Medium-High	None	Robust	1	Optical

Table 1: **Polarization input has a large application scope.** It is monocular - therefore only optically limited, works in cluttered conditions, robust to novel backgrounds and lighting conditions, and robust to print-out spoofs. Other modalities suffer in at least one of these categories.

contributions as:

- A robust method for transparent object instance segmentation that outperforms previous methods.
- Polarized CNNs: an attention-fusion based framework to process polarization data. We compare against several simpler Polarization + CNN baselines in a detailed ablation study.
- The first single-frame monocular instance segmentation approach that is significantly more robust to print-out spoofs.
- Application scenario: using Polarized CNNs in the context of robotic bin picking of cluttered transparent balls.

2. Scientific Background and Related Work

Difficulty of Transparency The interaction between light and transparent objects is rich and complex, but the material of an object determines its transparency in visible light. For many household objects, the majority of visible light passes straight through and a small portion ($\sim$ 4-8% dependent on the refractive index) is reflected. This is because visible wavelengths are insufficient in energy to excite atoms in the transparent object. This leads the texture behind the object to dominate the image. This unique property leads to the following difficulties when attempting instance segmentation.

1. **Clutter:** Clear edges are hard to see in densely cluttered scenes with transparent objects. In extreme cases, the edges are not visible at all (see Figure 3 (b) for an example), creating ambiguities in the exact shape of the objects [45, 12].
2. **Novel Environments:** Low reflectivity in the visible spectrum causes these objects to appear different, *out-of-distribution*, in novel environments leading to poor generalization.
3. **Print-Out Spoofs:** Any algorithm using a single RGB image is very susceptible to print-out spoofs [46, 24] due to the perspective ambiguity. While other non-monocular algorithms for semantic segmentation of transparent objects exist [45, 44], they are range limited and unable to handle instance segmentation (see Table 1).

Physics of Polarization Image Formation A light ray hitting the camera has three measured components: the intensity of light (intensity image/I), the percentage of light that is linearly polarized (degree of linear polarization/DOLP/ρ) and the direction of that linear polarization (angle of linear polarization/AOLP/ϕ). These properties encode information about the surface curvature and material of the object being imaged [4]. Measuring I, ρ, and ϕ at each pixel requires 3+ images of a scene taken behind a polarizing filter at different angles, ϕ_{pol}. Then we solve for ϕ, ρ, I using the following equation:

$$I_{\phi_{pol}} = I(1 + \rho\cos(2(\phi - \phi_{pol}))). \tag{1}$$

In our case we use a FLIR Blackfly S multi-polar camera that gives us I_0,I_{45},I_{90},I_{135} in a single capture.

Shape from Polarization (SfP) theory [4] gives us the following relationship between the refractive index (n), azimuth angle (θ_a) and zenith angle (θ_z) of the surface normal of an object and the ϕ and ρ components of the light ray coming from that object. When diffuse reflection is dominant:

$$\rho = \frac{(n - \frac{1}{n})^2 \sin^2(\theta_z)}{2 + 2n^2 - ((n + \frac{1}{n})^2 \sin^2\theta_z + 4\cos\theta_z\sqrt{n^2 - sin^2\theta_z}}, \tag{2}$$

$$\phi = \theta_a. \tag{3}$$

And when the specular reflection is dominant:

$$\rho = \frac{2\sin^2\theta_z\cos\theta_z\sqrt{n^2 - \sin^2\theta_z}}{n^2 - \sin^2\theta_z - n^2\sin^2\theta_z + 2\sin^4\theta_z}, \tag{4}$$

$$\phi = \theta_a - \frac{\pi}{2}. \tag{5}$$

Note that in both cases ρ increases exponentially as θ_z increases and if the refractive index is the same, specular reflection is much more polarized than diffuse reflection.

Deep Instance Segmentation: There are many approaches for deep instance segmentation: semantic segmentation based [5, 20, 3], proposal-based [14], and even RNN based [36, 34]. Our framework can be applied to any of them. In this work we focus on the state-of-the-art Mask R-CNN [14] architecture. Mask R-CNN works by taking an input image x, which is an HxWx3 tensor of image intensity values, and running it through a backbone network:

$C = B(x)$. $B(x)$ is responsible for extracting useful features from the input image and can be any standard CNN architecture e.g. ResNet-101 [15]. The backbone network outputs a set of tensors, $C = \{C_1, C_2, C_3, C_4, C_5\}$, where each tensor C_i represents a different resolution feature map. These feature maps are then combined in a feature pyramid network (FPN) [22], processed with a region proposal network (RPN) [35], and finally passed through an output subnetwork [35, 14] to produce classes, bounding boxes, and pixel-wise segmentations. These are merged with non-maximum suppression for instance segmentation. More details are available in [14].

Deep Learning for Multi-Modal Input Deep learning has been used to combine many modalities including depth and RGB [11, 43, 24], polarimetric and RGB [47] and more [46, 43, 41]. All of these architectures have unique backbones for each input signal. The depth at which they are fused classifies them as one of the following: early-fusion[43], mid-fusion [33] and late-fusion [47]. In this work we apply mid-fusion, as described in [33], since it is easily extendable to Mask R-CNN. Mid-fusion is defined as follows: Assume there are two input images x_a and x_b. First, each image is fed into a unique backbone $B_a(x_a)$ and $B_b(x_b)$, then the output of each backbone is fused at each scale i with a fusion function f.

$$C_i = f(C_{ia}, C_{ib}). \tag{6}$$

There are four main fusion methods:

- $C_{ia} + C_{ib}$ addition/averaging of the tensors [11].
- $G([C_{ia}, C_{ib}])$ concatenation of the tensors along the depth axis followed by a 1x1 Convolution G to reduce dimensionality [32, 33].
- $w_{ia}C_{ia} + w_{ib}C_{ib}$ mixture of experts (MoE) [28, 43] where a sub-network predicts two scalar weights w_{ia} and w_{ib} such that $w_{ia} + w_{ib} = 1$ allowing the network to dynamically weight each input.
- $G([SE(C_{ia}), SE(C_{ib})]$ and squeeze-excitation (SE) fusion [46] where each input tensor has it's channels re-weighted using an SE block [16] and then follows the concatenation procedure above.

None of these fusion methods allow for spatially dynamic weighting.

Transparent Object Instance Segmentation Previous work [19, 21] uses deep learning trained on existing RGB image datasets [7, 23, 13] for detection of transparent objects. These can easily be extended to instance segmentation by replacing the SSD [19] or R-CNN [21] with Mask R-CNN [14]. We call this Intensity Mask R-CNN and use this as our baseline.

There are several other approaches to do detection and segmentation with more complex imaging setups that would be more robust to print-out spoofs. However these approaches are not extendable to instance segmentation. Multiple existing approaches use an RGB + depth sensor (e.g. Kinect) to do transparent object segmentation [44, 12, 39], pose estimation [25, 26] and even 3D reconstruction [38, 2]. However, in a cluttered environment a depth sensor provides no information on instance boundaries for transparent objects. Other approaches also include using light field information for segmentation [45] and camera motion utilization for shape reconstruction of transparent objects [6].

Polarization in Computer Vision Shape from polarization can be used for 3D reconstruction of objects such as shiny metals [30], diffuse dielectrics [4], and transparent/translucent objects [29, 37, 8]. Polarization is also used for problems in 3D imaging [17, 10, 18, 49], reflection separation [31, 40], face scanning [27], underwater de-scattering [42], and semantic segmentation [48].

Polarization for semantic segmentation of roads [47] is the only other work using polarization with deep learning. They do not treat it as multi-modal fusion, rather they concatenate $[I_0, I_{45}, I_{90}, I_{135}]$ and feed it to a deep network. While this works, the model struggles to learn the physical priors, which leads to poor performance. Our framework leverages the physics of polarization described in equation 1 to create three unique input images I, ρ, and ϕ. These images a fused in a multi-modal fashion in our unique backbone.

3. Our Method

In what follows, we derive the polarization image formation model, motivate why this image contains better texture for transparent object segmentation, and then present our Polarized CNN framework for adding those cues in deep learning models.

3.1. Polarization Image Formation (Transparency)

Light rays coming from a transparent objects have two components: a reflected portion, consisting of I_r, ρ_r, ϕ_r and the refracted portion I_t, ρ_t, ϕ_t. The intensity of a single pixel in the resulting image can be written as:

$$I = I_r + I_t. \tag{7}$$

When we add a polarizing filter in front of the camera we get:

$$I_{\phi_{pol}} = I_r(1+\rho_r \cos(2(\phi_r - \phi_{pol}))) + I_t(1+\rho_t \cos(2(\phi_t - \phi_{pol}))). \tag{8}$$

To understand the impact this has on our measured ρ and ϕ from 1, we solve for ρ and ϕ in terms of I_r,ρ_r,ϕ_r,I_t,ρ_t,ϕ_t:

$$\rho = \frac{\sqrt{(I_r\rho_r)^2 + (I_t\rho_t)^2 + 2I_t\rho_t I_r\rho_r \cos(2(\phi_r - \phi_t))}}{I_r + I_t}, \tag{9}$$

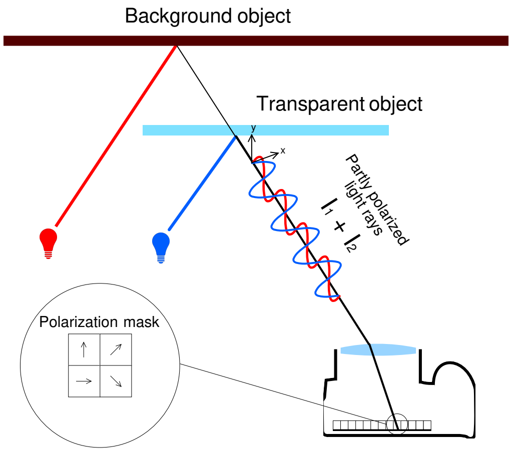

Figure 2: **Polarization image formation model for transparent objects.** A light ray hitting the multi-polar sensor contains polarization information from both a transparent object and the background object. The small fraction of reflected light from a transparent object is heavily polarized, and thus has a large impact on the polarization measurement.

$$\phi = \arctan\left(\frac{I_r\rho_r \sin(2(\phi_r - \phi_t))}{I_t\rho_t + I_r\rho_r \cos(2(\phi_r - \phi_t))}\right) + \phi_r. \quad (10)$$

Equations 7, 9, and 10 give us the image formation model for I, ρ, and ϕ in the case of transparency. We use these equations to show the superiority of the ρ and ϕ images for transparent object segmentation when compared to I. We verify this through an ablation analysis in Table 3 rows 1-3.

Here we motivate why ρ and ϕ can show texture when objects appear textureless in I. An object's texture appears invisible in I because it is strictly dependent on the ratio of $\frac{I_r}{I_t}$ (see equation 7). Unlike opaque objects where $I_t = 0$, transparent objects transmit most light and only reflect a small portion. This is why we bring this problem into the realm of polarization, where the strength of a transparent objects texture is instead dependent on $\phi_r - \phi_t$ and the ratio of $\frac{I_r\rho_r}{I_t\rho_t}$ (see equations 9, 10). We can safely assume that $\phi_r \neq \phi_t$ and $\theta_{zr} \neq \theta_{zt}$ for the majority of pixels, i.e. the geometry of the background and transparent object are different. And we know that ρ_r follows the specular reflection curve [29], meaning it is highly polarized, and at Brewster's angle (approx. 60°), it is 1.0 (see equation 4). Therefore we can be certain that at the appropriate zenith angles, $\rho_r \geq \rho_t$, and if the background is diffuse or has a low zenith angle, $\rho_r \gg \rho_t$. We can see this effect in Figure 1 where the sphere's texture dominates when $\theta_z \approx 60°$. This leads us to believe that in many cases:

$$\frac{I_r}{I_t} \leq \frac{I_r\rho_r}{I_t\rho_t}. \quad (11)$$

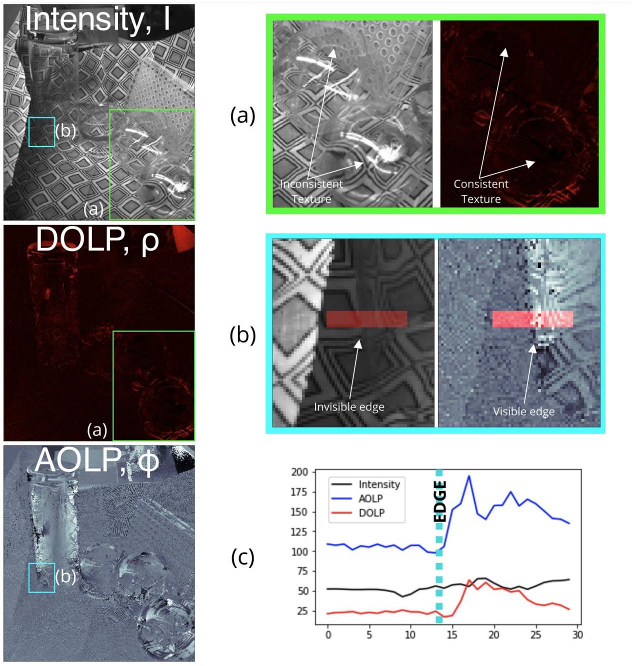

Figure 3: **The polarization texture of the transparent objects improves input quality.**(a) The texture of the 2 balls is inconsistent in the intensity image due to the change in background, highlighting problem (2). In the DOLP this doesn't happen. (b) The edge is practically invisible in the intensity image, but is much brighter in the AOLP. (c) A cross-section of the edge shows the edge has much higher contrast in the AOLP and DOLP when compared to the intensity.

Thus even if the texture of the transparent object appears invisible in I, it may be more visible in ϕ and ρ, motivating this paper. We discuss specific cases further in the supplement and discuss the implications on the key problems below.

Clutter In clutter, a key difficulty lies in seeing the edges of a texture-less transparent object, see Figure 3 (b) and (c) for an example. Since the texture appears more visible in ϕ and ρ, some edges are better visible.

Novel Environments Other than increasing the strength of the transparent object texture, the ρ image also reduces the impact of diffuse backgrounds like textured clothes. This allows the transparent object to appear similar even though the environment has changed. We show an example in Figure 3 (a) and verify the effectiveness in Table 3.

Print-Out Spoofs Paper is flat, leading to a mostly uniform ϕ and ρ. Transparent objects have some amount of surface variation, which will appear very non-uniform in ϕ and ρ as shown in Figure 1.

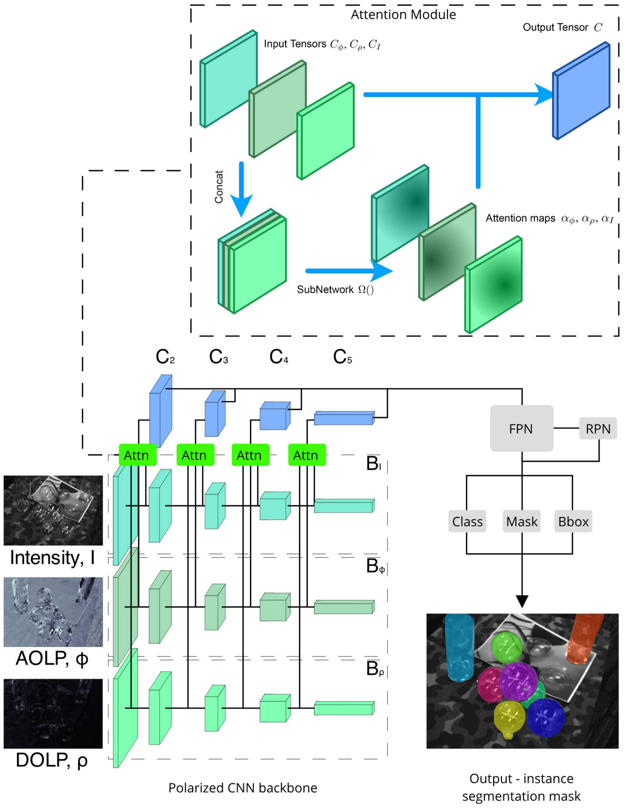

Figure 4: **Our Polarized CNN framework applied to Mask-RCNN.** We use three separate backbones and merge them with attention-fusion to produce high quality instance segmentation with Mask R-CNN. See Section 3.2.

3.2. Polarized CNN Framework:

Polarized CNNs, as shown applied to Mask R-CNN in Figure 4, is a framework for effectively leveraging the extra information contained in polarized images using deep learning. Applying this framework requires three changes to a CNN architecture: (1) *Input Image:* Applying the physical equations of polarization to create the right input images. (2) *Attention-fusion Polar Backbone:* Treating the problem as a multi-modal fusion problem. (3) *Geometric Data Augmentations:* Correctly augmenting the data to reflect the physics of polarization.

Input Image We propose feeding in three input images: the AOLP (ϕ), the DOLP (ρ), and the intensity image (I) from equation 1 as the optimal input for transparent object. These images are computed from I_0,I_{45},I_{90}, and I_{135}, normalized to be in the range [0-255] and turned into three-channel gray scale images to allow for easy transfer learning from MSCoCo [23] pre-trained weights.

Multi-Modal Fusion Each input image is fed through a unique backbone: $B_I(I)$, $B_\rho(\rho)$, $B_\phi(\phi)$. We propose a novel spatially-aware attention-fusion mechanism to perform multi-modal fusion. The output feature maps from each backbone B_I, B_ρ, B_ϕ at each scale i, $C_{i,I}$, $C_{i,\rho}$, $C_{i,\phi}$ are concatenated and processed through a set of convolutional layers Ω_i. Ω_i outputs a 3-channel image with the same height and width as the input. This is followed by a softmax giving us pixel-wise attention weights α:

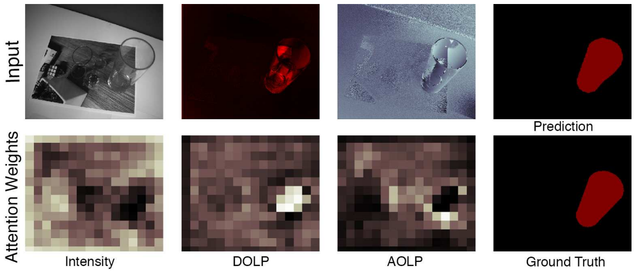

Figure 5: **Our attention module allows for interpretable multi-modal fusion.** The learned attention weights are brightest on the AOLP and DOLP to avoid the ambiguous print-out spoof in the intensity image. More examples available in the supplement.

$$[\alpha_{i,\phi}, \alpha_{i,\rho}, \alpha_{i,I}] = softmax(\Omega_i([C_{i,\phi}, C_{i,\rho}, C_{i,I}])). \tag{12}$$

These attention weights are used to perform a weighted average per channel:

$$C_i = \alpha_{i,\phi} C_{i,\phi} + \alpha_{i,\rho} C_{i,\rho} + \alpha_{i,I} C_{i,I}. \tag{13}$$

The attention module allows the model to weight the different inputs depending how relevant they are to a given portion of the scene. Results are available in Table 2 and discussion in Section 4. Figure 4 describes this model and architecture in detail. Attention maps are visualized in Figure 5 and in the supplement.

Geometric Data Augmentations In small training datasets, affine transformations are an important data augmentation to achieve good generalization performance. Naively applying this to the ϕ image doesn't work. The AOLP is an angle from 0-360 that reflects the direction of the electromagnetic wave with respect to the camera coordinate frame. If a rotation operator is applied to the image, then this is the equivalent of rotating the camera around it's Z-axis. This rotation will change the orientation of the xy plane of the camera, and thus will change the relative direction of the wave. To account for this change, the pixel values of the AOLP must be rotated accordingly in the opposite direction. We apply this same principal to other affine transformations. This is key to achieving good performance as we show later in Section 4.

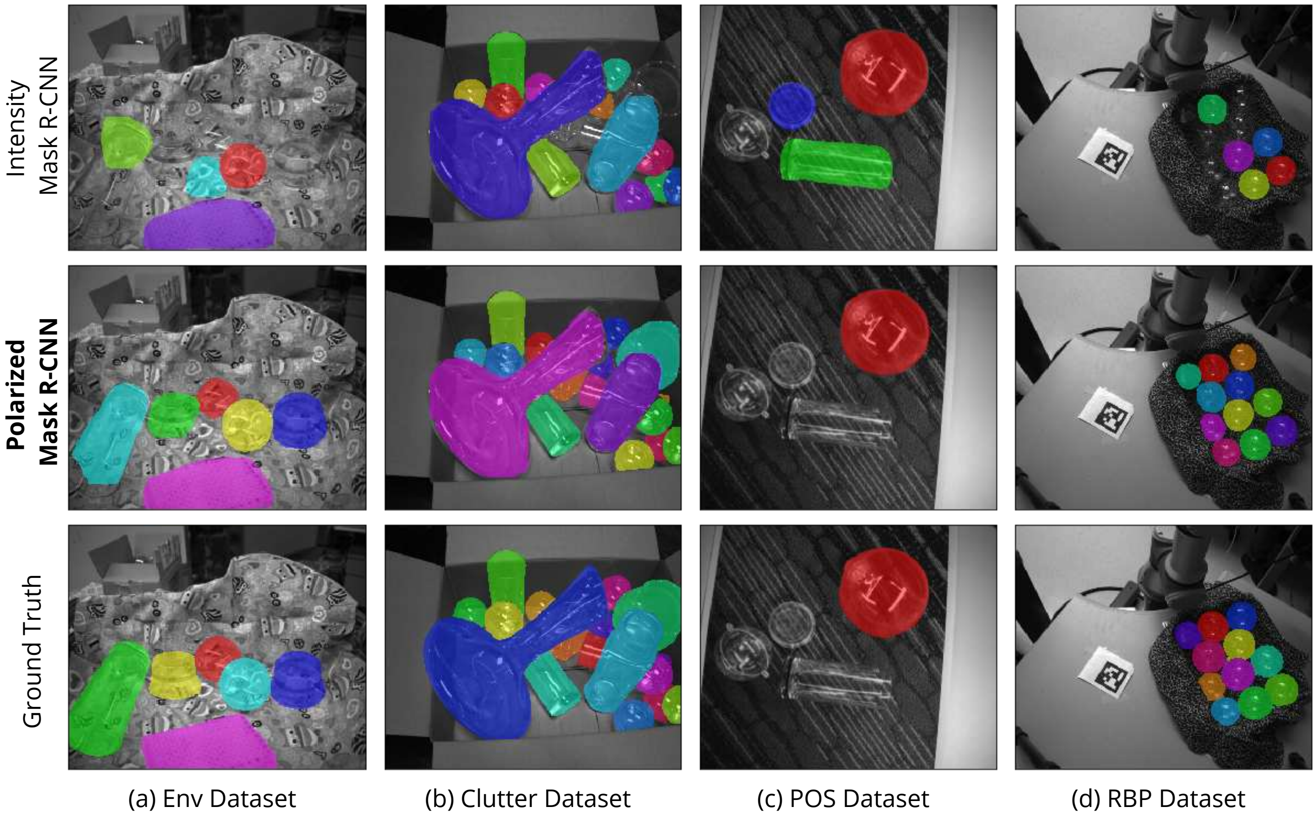

Figure 6: **Qualitative comparisons showing improvement from Intensity to Polarized Mask R-CNNs.** (a) Polarization helps accurately segment clutter where it is ambiguous in the intensity image. (b) The Intensity Mask R-CNN completely fails to adapt to the novel environment while the polar model succeeds. (c) The Intensity Mask R-CNN is fooled by the printed paper however the Polarized Mask R-CNN is robust. (d) An example image from our robotic bin picking application. Polarization adapts better to this novel environment with poor lighting.

4. Experiments

We perform a thorough set of experiments to determine the effectiveness of the proposed Polarized CNN and each individual component.

4.1. Experimental Details

Implementation To conduct the experiments, we use a keras [9] implementation of Mask R-CNN [1]. All experiments were run on an AWS p3.2xlarge instance with a single P100 GPU.

Training Data Our transparent object training set contains 1000 images with over 20000 instances of transparent objects in 15 unique environments from a 6 possible classes: plastic cups, plastic trays, glasses, ornaments, and other.

Evaluation Data We construct 4 test sets to properly evaluate problems (1)-(3). An example image from each test set is available in Figure 6.

- *Clutter* This test set contains 200 images of cluttered transparent objects in environments similar to the training set with no print-outs - used to test problem (1).
- *Novel Environments (Env)* This test set contains 50 images taken of 6 objects per image with environments not available in the training set. The backgrounds contain harsh lighting, textured cloths, shiny metals, and more - testing problem (2).
- *Print-Out Spoofs (POS)* This test set contains 50 images, each containing a 1-6 printed objects and 1-2 real objects.
- *Robotic Bin Picking (RBP)* This set contains 300 images taken from a live demo of our robotic arm picking up ornaments. This set is used to test the instance segmentation performance in a real world application.

For each dataset, we use two metrics to measure accuracy: mean average precision in range of IoUs 0.5-0.7

($mAP_{.5:.7}$), and mean average precision in the range of IoUs 0.75-0.9 ($mAP_{.75:.9}$). These two metrics measure coarse segmentation and fine-grained segmentation respectively. To further test generalization, we test all models on object detection as well using the Faster R-CNN component of Mask R-CNN.

Capture Setup All images are taken with a Flir Blackfly S Monochrome Polar Camera. To allow all models to be trained on the exact same set of images, the intensity baseline is done with monochrome images. This is fair because the transparent objects in our dataset are colorless, RGB data does not add value for transparent object segmentation. We verify this in the supplement.

4.2. Polarized vs. Intensity Mask R-CNN

We test the Intensity Mask R-CNN [19] and our Polarized Mask R-CNN on the four test sets mentioned above. Qualitative examples from each dataset are visible in Figure 6 and quantitative results in Table 2. Our average improvement in coarse segmentation is 14.3% mAP, and in fine-grained segmentation is 17.2% mAP. The performance improvement in problem (1) is more visible when doing fine-grained segmentation where the gap in performance goes from 1.1% mAP to 4.5% mAP. This supports our thesis that polarization data provides useful edge information allowing the model to more accurately segment objects. For generalization to new environments we see much larger gains for both fine-grained and coarse segmentation supporting our thesis that the the intrinsic texture of a transparent object appears more visible in the polarized images. Our architecture shows a similarly large improvement in robustness against print-out spoofs, achieving almost 90% mAP. This demonstrates a monocular solution that is robust to perspective projection issues such as print-out spoofs. All of these results help explain the dramatic improvement in performance shown for an uncontrolled and cluttered environment like Robotic Bin Picking (RBP). The results in Table 2 highlight the benefits of Polarized CNNs for robust instance segmentation of transparent objects.

4.3. Polarization + CNN Comparisons

We create many different Polarization + CNN baselines to compare against our Polarized CNNs framework.

Input Images Our first set of baselines uses the following inputs independently, ρ, ϕ, I [19], and $I_0 - I_{135}$ [47]. We train a Mask R-CNN on each input type, and test all on four test sets in Table 3. Each input is good for a different problem. Both ϕ and ρ are much better than I at avoiding print-out spoofs. ρ is the most useful signal on the *RBP* and *Env* datasets. I is the best at handling cluttered environments previously seen. It achieves the slightly better performance on the *Clutter* test set, but is significantly worse in all 3 other test sets. [46]'s method for processing polar input performs is worse than ρ in novel backgrounds and worse than ϕ at avoiding print-out spoofs. Hence the first four rows of Table 3 show that while using all 4 channels independently, as in [47] is good, there is more to be gained by adopting physical priors in the deep learning model by using ρ and ϕ.

Multi-Modal Fusion After verifying that I, ρ, and ϕ are the appropriate input channels, we evaluate the different fusion method baselines. We compare to four major standard approaches [11, 33, 28, 46] described in Section 3.2. Overall, attention-based fusion of polarization data leads to improved robustness for transparent object instance segmentation across all tests. We visualize the ablation analysis in the supplement.

We also demonstrate that multi-modal fusion is necessary by comparing against concatenating all three images into a single 3-channel image and using a single backbone. The model is unable to take advantage of all three channels and learns a very sub-optimal policy which is on average 8 mAP worse than our attention-fusion mechanism.

Geometric Data Augmentations We verify the necessity for geometrically accurate data augmentations in an ablation study with results reported in Table 4. It shows that using normal augmentations actually hurts performance in some cases, whereas geometric augmentations improve performance across all four test sets.

4.4. Application: Pick and Place

Bin picking of transparent and translucent (non-Lambertian) objects is an incredibly hard and open problem. To show the difference high quality, robust segmentation makes, we compare Intensity Mask R-CNN with our Polar Mask R-CNN as part of a proof of concept end-to-end system to bin pick different sized cluttered transparent ornaments.

A bin picking solution contains three components, a segmentation component to isolate each object, a depth estimation component, and a pose estimation component. To understand the effect of segmentation, we use a simple depth estimation and pose where we have the robot arm move to the center of the segmentation and stop when it hits a surface. This only works because the objects are perfect spheres. A slightly inaccurate segmentation can cause an incorrect estimate and a false pick. This application allows us to compare both Polarized Mask R-CNN and Intensity Mask R-CNN. We test our system in 5 tough environments outside the training set. For each environment we stack 15 balls, then measure the number of correct/incorrect (missed) picks the robot arm makes to pick up all 15 balls or makes

Evaluation Criteria		**Mean Score**		**Clutter**		**Env**		**POS**		**RBP**	
Model	Task	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$
Intensity Mask R-CNN [19]	Instance Seg.	0.656	0.454	0.878	0.689	0.281	0.146	0.685	0.616	0.779	0.364
Polarized Mask R-CNN (Ours)	Instance Seg.	**0.793**	**0.635**	**0.889**	**0.733**	**0.511**	**0.351**	**0.893**	**0.841**	**0.877**	**0.614**
Intensity Mask R-CNN [19]	Detection	0.662	0.434	0.885	0.694	0.277	0.13	0.681	0.546	0.803	0.364
Polarized Mask R-CNN (Ours)	Detection	**0.796**	**0.601**	**0.893**	**0.723**	**0.516**	**0.299**	**0.893**	**0.758**	**0.883**	**0.624**

Table 2: **Polarized Mask R-CNN outperforms Intensity Mask R-CNN for both detection and instance segmentation.**

Model Info		**Mean Score**		**Clutter**		**Env**		**POS**		**RBP**	
Input Type	Backbone	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$	$mAP_{.5:.7}$	$mAP_{.75:.9}$
I_{un} [19]	ResNet-101	0.656	0.454	0.878	0.689	0.281	0.146	0.685	0.616	0.779	0.364
ϕ	ResNet-101	0.702	0.531	0.84	0.605	0.28	0.145	0.872	0.807	0.816	0.573
ρ	ResNet-101	0.738	0.561	0.867	0.653	0.447	0.256	0.8	0.716	0.838	0.609
I_0,I_{45},I_{90},I_{135} [47]	Concat + ResNet-101	0.743	0.545	**0.89**	0.711	0.386	0.221	0.868	0.803	0.829	0.444
I_{un}, ϕ, ρ	Concat + ResNet-101	0.711	0.538	0.864	0.656	0.278	0.134	0.833	0.765	0.87	0.596
I_{un}, ϕ, ρ	Mid-Fusion + Mean [11] *	0.787	0.624	**0.892**	0.734	0.493	0.337	0.886	**0.842**	**0.879**	0.582
I_{un}, ϕ, ρ	Mid-Fusion + Concat [32, 33] *	0.768	0.606	**0.892**	0.727	0.469	0.297	0.843	0.786	0.869	**0.615**
I_{un}, ϕ, ρ	Mid-Fusion + MoE [28, 43] *	0.777	0.616	**0.889**	**0.738**	0.468	0.287	0.871	0.825	**0.878**	**0.615**
I_{un}, ϕ, ρ	Mid-Fusion + SE Merge [46] *	0.764	0.603	**0.894**	**0.740**	0.448	0.298	0.844	0.794	0.870	0.578
I_{un}, ϕ, ρ	Mid-Fusion + Attention (Ours)	**0.793**	**0.635**	**0.889**	0.733	**0.511**	**0.351**	**0.893**	**0.841**	**0.877**	**0.615**

Table 3: **Instance Segmentation Ablation Analysis Input and Backbone Ablation Results.** We bold any result with 0.005 mAP of the best result. * Adapted to our task.

Input Type	Mean	Clutter	Env	POS	RBP
AOLP	0.476	0.550	0.136	0.707	0.514
AOLP + Augs	0.486	0.591	0.080	0.746	0.528
AOLP + Geometric Augs	**0.531**	**0.605**	**0.145**	**0.807**	**0.573**

Table 4: **Geometric data augmentations are vital for improved performance.**

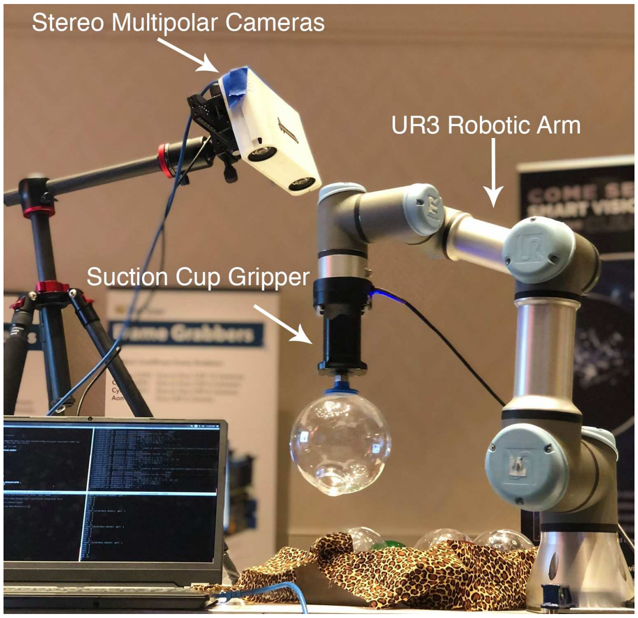

Figure 7: **Transparent object bin picking setup:** A custom polar stereo camera pair calibrated with a UR3 robotic arm with a suction cup gripper, picking a bin of stacked transparent objects.

Model Type	Picked	False Picks	Remaining	Total
Intensity Mask R-CNN [19]	60	56	30	90
Polarized Mask R-CNN (Ours)	**90**	**18**	**0**	**90**

Table 5: **Polarized CNNs allow the robot to empty the bin with minimal false picks.**

15 incorrect picks, whichever comes first. The results are available in Table 4. Intensity based model is unable to empty the bin consistently because the robotic arm consistently misses certain picks due to poor segmentation quality. The polar model on the other hand, picks all 90 balls successfully, with approximately 1 incorrect pick for every 6 correct picks. These results validate the effect that a difference of 20 mAP can make.

5. Conclusion

Transparent objects have more prominent textures in the polarization domain. This unique texture is best exploited with our Polarized CNN framework, which we demonstrate on instance segmentation with Mask R-CNN. We support this with our experiments and demonstrate it's importance with an application to robotic bin picking. We also demonstrated a passive monocular system that is robust to printout spoof attacks. We hope to spur future work in computer vision that exploits polarization with data driven problems and explore novel camera configurations for applications in broad areas such as robotics, autonomous driving, and face authentication.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://

github.com/matterport/Mask_RCNN, 2017. 6

[2] Nicolas Alt, Patrick Rives, and Eckehard Steinbach. Reconstruction of transparent objects in unstructured scenes with a depth camera. In *2013 IEEE International Conference on Image Processing*, pages 4131–4135. IEEE, 2013. 3

[3] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 441–450, 2017. 2

[4] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006. 2, 3

[5] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5221–5229, 2017. 2

[6] Moshe Ben-Ezra and Shree K Nayar. What does motion reveal about transparency? In *Proceedings of the IEEE International Conference on Computer Vision*, page 1025. IEEE, 2003. 3

[7] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 3

[8] Tongbo Chen, Hendrik PA Lensch, Christian Fuchs, and Hans-Peter Seidel. Polarization and phase-shifting for 3d scanning of translucent objects. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 3

[9] François Chollet et al. Keras. https://keras.io, 2015. 6

[10] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2017. 3

[11] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Fabian Duffhauss, Claudius Glaeser, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *arXiv preprint arXiv:1902.07830*, 2019. 3, 7, 8

[12] Chen Guo-Hua, Wang Jun-Yi, and Zhang Ai-Jun. Transparent object detection and location based on rgb-d camera. In *Journal of Physics: Conference Series*, volume 1183, page 012011. IOP Publishing, 2019. 2, 3

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[17] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3370–3378, 2015. 3

[18] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1-3):34–51, 2017. 3

[19] May Phyo Khaing and Mukunoki Masayuki. Transparent object detection using convolutional neural network. In *International Conference on Big Data Analysis and Deep Learning Applications*, pages 86–93. Springer, 2018. 2, 3, 7, 8

[20] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2017. 2

[21] Po-Jen Lai and Chiou-Shann Fuh. Transparent object detection using regions with convolutional neural network. In *IPPR Conference on Computer Vision, Graphics, and Image Processing*, pages 1–8, 2015. 3

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 5

[24] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 2, 3

[25] Ilya Lysenkov, Victor Eruhimov, and Gary Bradski. Recognition and pose estimation of rigid transparent objects with a kinect sensor. *Robotics*, 273:273–280, 2013. 3

[26] Ilya Lysenkov and Vincent Rabaud. Pose estimation of rigid transparent objects in transparent clutter. In *2013 IEEE International Conference on Robotics and Automation*, pages 162–169. IEEE, 2013. 3

[27] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007. 3

[28] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156. IEEE, 2016. 3, 7, 8

[29] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Transparent surface modeling from a pair of polarization images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):73–82, 2004. 3, 4

[30] Olivier Morel, Fabrice Meriaudeau, Christophe Stolz, and Patrick Gorria. Polarization imaging applied to 3d reconstruction of specular metallic surfaces. In *Machine Vision Applications in Industrial Inspection XIII*, volume 5679, pages 178–186. International Society for Optics and Photonics, 2005. 3

[31] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. Separation of reflection components using color and polarization. *International Journal of Computer Vision*, 21(3):163–186, 1997. 3

[32] Andreas Pfeuffer and Klaus Dietmayer. Optimal sensor data fusion architecture for object detection in adverse weather conditions. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2018. 3, 8

[33] Hazem Rashed, Ahmad El Sallab, Senthil Yogamani, and Mohamed ElHelw. Motion and depth augmented semantic segmentation for autonomous navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 7, 8

[34] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6656–6664, 2017. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[36] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In *European conference on computer vision*, pages 312–329. Springer, 2016. 2

[37] Megumi Saito, Yoichi Sato, Katsushi Ikeuchi, and Hiroshi Kashiwagi. Measurement of surface orientations of transparent objects using polarization in highlight. *Systems and Computers in Japan*, 32(5):64–71, 2001. 3

[38] Shreeyak S Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Cleargrasp: 3d shape estimation of transparent objects for manipulation. *arXiv preprint arXiv:1910.02550*, 2019. 3

[39] Viktor Seib, Andreas Barthen, Philipp Marohn, and Dietrich Paulus. Friend or foe: exploiting sensor failures for transparent object localization and classification. In *2016 International Conference on Robotics and Machine Vision*, volume 10253, page 102530I. International Society for Optics and Photonics, 2017. 3

[40] Vage Taamazyan, Achuta Kadambi, and Ramesh Raskar. Shape from mixed polarization. *arXiv preprint arXiv:1605.02066*, 2016. 3

[41] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. Dynamic subtitles: A multimodal video accessibility enhancement dedicated to deaf and hearing impaired users. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[42] Tali Treibitz and Yoav Y Schechner. Active polarization descattering. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):385–399, 2008. 3

[43] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651. IEEE, 2017. 3, 8

[44] Tao Wang, Xuming He, and Nick Barnes. Glass object localization by joint inference of boundary and depth. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3783–3786. IEEE, 2012. 2, 3

[45] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Transcut: Transparent object segmentation from a light-field image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3442–3450, 2015. 2, 3

[46] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. 2, 3, 7, 8

[47] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. 2019. 3, 7, 8

[48] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. Exploration of Deep Learning-based Multimodal Fusion for Semantic Road Scene Segmentation. In *VISAPP 2019 14th International Conference on Computer Vision Theory and Applications*, Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, Czech Republic, Feb. 2019. 3

[49] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019. 3