Supplement: 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)

1. Overview

In the supplement, we provide additional information and qualitative comparisons that we were unable to fit in the main paper. First, in section 2 we explore the main result of the paper further with both some qualitative and quantitative analysis. In section 3 we add the specific cases to equations 8-11 of the main paper. In section 4 we compare RGB imaging vs Grayscale imaging for transparent object segmentation. In section 5 visualize and interpret attention maps from our attention fusion backbone. In section 6 we extend section 3.1 of the paper with more examples. Finally in section 7 we include more experiments for an ablation study.

2. Intensity vs Polarized CNNs - More Results

Table 1 shows a per class breakdown for the clutter dataset. This dataset contains the most examples of each class. We notice that the largest gains occur in objects that are the most transparent. For example plastic trays are very translucent - almost opaque, and therefore quite visible in intensity imaging. There we see no significant increase in performance. For Plastic Cups, Glasses, and Ornaments, all of which are transparent, we see an improvement in performance, especially for fine-grained classification. This supports our thesis that polarization improves segmentation for transparent objects.

More qualitative examples, similar to Figure 5. of the main paper are available in Figures 1 - 4

3. More Analysis on Polarization Image Model

The appearance of a transparent object is dependant $I_r\rho_r$ and $I_t\rho_t$ and $\phi_r - \phi_t$ as defined by equations 7, 8, 9, 10, 11 of the main paper. Here we analyse different cases of the above variables. At each pixel only one of the following three cases is possible:

1. $I_t\rho_t \ll I_r\rho_r$. In this case ϕ and ρ_s mostly depend on reflected component, effectively making the transparent object opaque in the DOLP/AOLP channels. The light polarization is similar to one reflected from opaque object, and encodes the shape of the object, making edges contrasting in both AOLP and DOLP channels as well as making the polarization of internal part of the object not depending on background. In fact, this case addresses all the challenges (1)-(3) of the transparent object segmentation. Example of such case is shown on Figure 1 of main paper.

2. $I_t\rho_t \sim I_r\rho_r$. Here everything depends on the value of $\Delta\phi \ (= |\phi_r - \phi_t|)$ with three possibilities: (1) $\Delta\phi \sim \pi/2$: then, according to equation 8 of the paper, $I\rho = |I_r\rho_r - I_t\rho_t|$ and equation 9 gives us that ϕ is either equal to ϕ_R or ϕ_T - this doesn't guarantee a consistent look for the transparent object, but still helps contrasting edges in DOLP (and AOLP if $I_r\rho_r > I_t\rho_t$); (2) $\Delta\phi \sim 0$: then $\phi = \phi_R$ which still corresponds to object's surface normal, making the AOLP measurement consistent with the object's shape, and therefore, with the measurements in neighboring pixels; (3) $0 < \Delta\phi < \pi/2$: whereby both ϕ and ρ_s are different compared to the refracted component, making the transparent object semi-opaque.

3. $I_t\rho_t \gg I_r\rho_r$. Reflected component just slightly changes ϕ and ρ_s. However, as mentioned, refraction also changes the polarization state of light which makes ϕ and ρ_s different from the case where the transparent object did not exist. If this happens closer to the edge of the transparent object - it provides sufficient contrast thereby helping with clutter and novel environments.

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

Model Info		All Classes		Ornaments		Plastic Cups		Glasses		Plastic Trays		Other Transparent	
Model	Task	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$	$\text{mAP}_{.5:.7}$	$\text{mAP}_{.75:.9}$
Intensity Mask-RCNN [1]	Inst. Seg.	0.878	0.689	0.915	0.733	0.787	0.656	**0.743**	0.49	**0.932**	**0.789**	0.733	0.541
Polarized Mask R-CNN (Ours)	Inst. Seg.	**0.889**	**0.733**	**0.928**	**0.777**	**0.813**	**0.692**	**0.745**	**0.547**	**0.934**	**0.787**	**0.756**	**0.579**

Table 1: Per class results breakdown for the clutter dataset between Intensity Mask R-CNN and Polarized Mask R-CNN. Polarized Mask R-CNN seems improves performance more for more transparent object

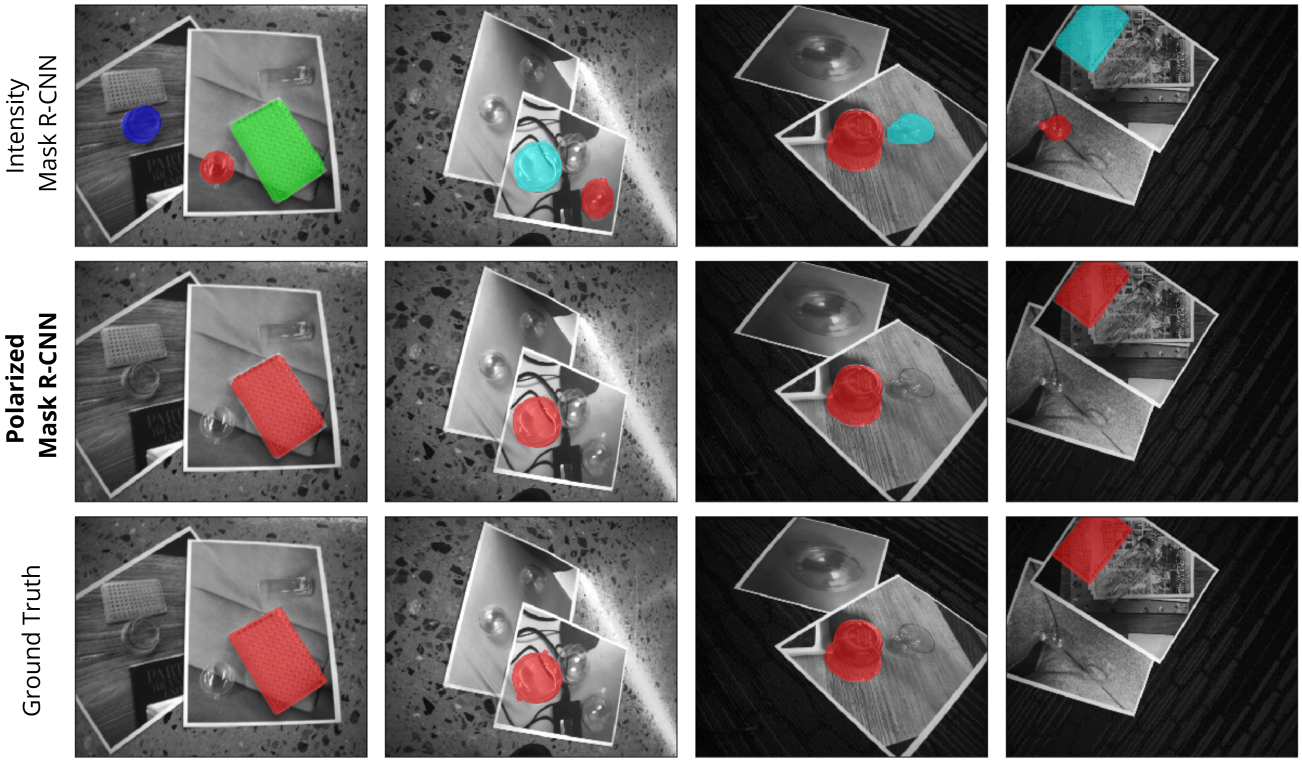


Figure 1: Qualitative comparisons between intensity and polarized CNNs for the case of transparent object segmentation in the presence of printout spoofs (POS dataset).

Examples of each of the above cases are visualize in supplement Figure 8.

4. RGB vs Gray

As mentioned in lines 577-581 of the paper, the comparison of the proposed method to the gray-scale intensity image based CNN is reasonable and using RGB images instead as a baseline won't add any value. To support this argument we train the model on a new dataset of RGB images and corresponding gray images and compare the results. The dataset contains 1125 training images and 285 validation images of ornaments. This dataset is very similar to our clutter dataset, except taken with an RGB camera instead of a polar camera. The training was done the exact same way as the baseline from the paper. The results of the comparison are presented in the Table 2. Here we see that the grayscale intensity based model performs slightly better since all the objects are colorless, and therefore removing color information reduces overfitting slightly. The numbers available in Table 2 are very similar to the ornament class for Intensity Mask R-CNN in Table 1.

Input Type	$\text{mAP}_{0.50:0.70}$	$\text{mAP}_{0.75:0.95}$
RGB Mask R-CNN	0.918	0.687
Intensity Mask R-CNN	**0.924**	**0.697**

Table 2: Comparison of the mAP score for RGB and Intensity based Mask R-CNN models. The small difference in favor of Intensity Mask R-CNN can be explained by the reduction in overfitting to color data when segmenting colorless objects.

5. Attention Maps

As mentioned in the main paper, we visualize the attention maps from our attention fusion and try to understand

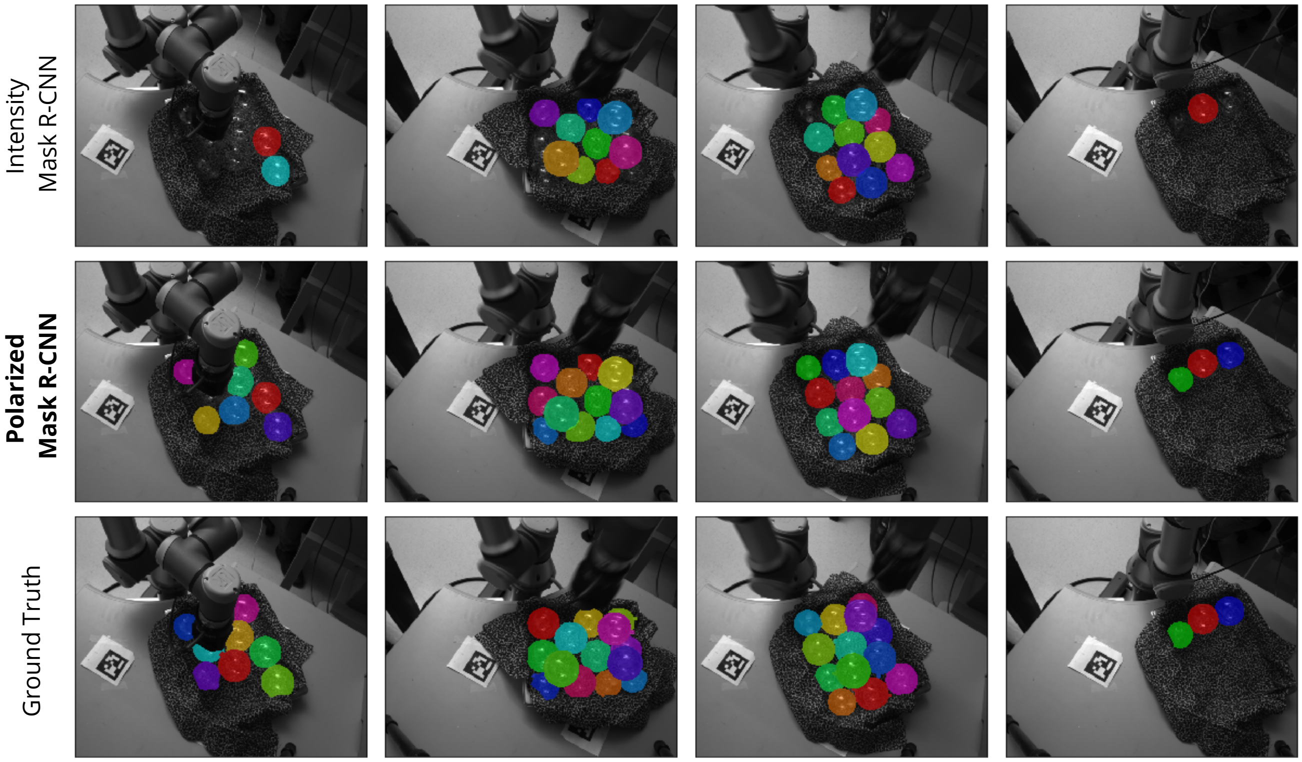


Figure 2: Qualitative comparisons between intensity and polarized CNNs for the case of transparent object segmentation for robotic bin picking (RBP dataset).

the effect different modalities have on the output. We give three examples of these attention maps in Figures 5 - 7. Overall we see that the model relies heavily on the DOLP and AOLP to determine which objects are not print-outs, identifying the objects in a novel background, and finding the edges for fine-grained segmentation - explaining the gains we see in the segmentation results.

6. Polarization Examples

The paper provides qualitative motivation for using the light polarization as a way to enhance transparent object segmentation. In lines 298-364 we describe possible cases for the $I_t\rho_t$ and $I_r\rho_r$ relationship. Here we support the qualitative discussion with the examples from the collected dataset (Figure 8).

7. More Ablation

Here we provide some more qualitative comparison examples to support Table 3 Figures 9 and 10. We also show results for detection ablation analysis and parameter/runtime for each model in Tables 3 and 4.

We show more qualitative results of the Polarized Mask R-CNN compared to Intensity Mask R-CNN on Figure 1 - 4.

References

[1] 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

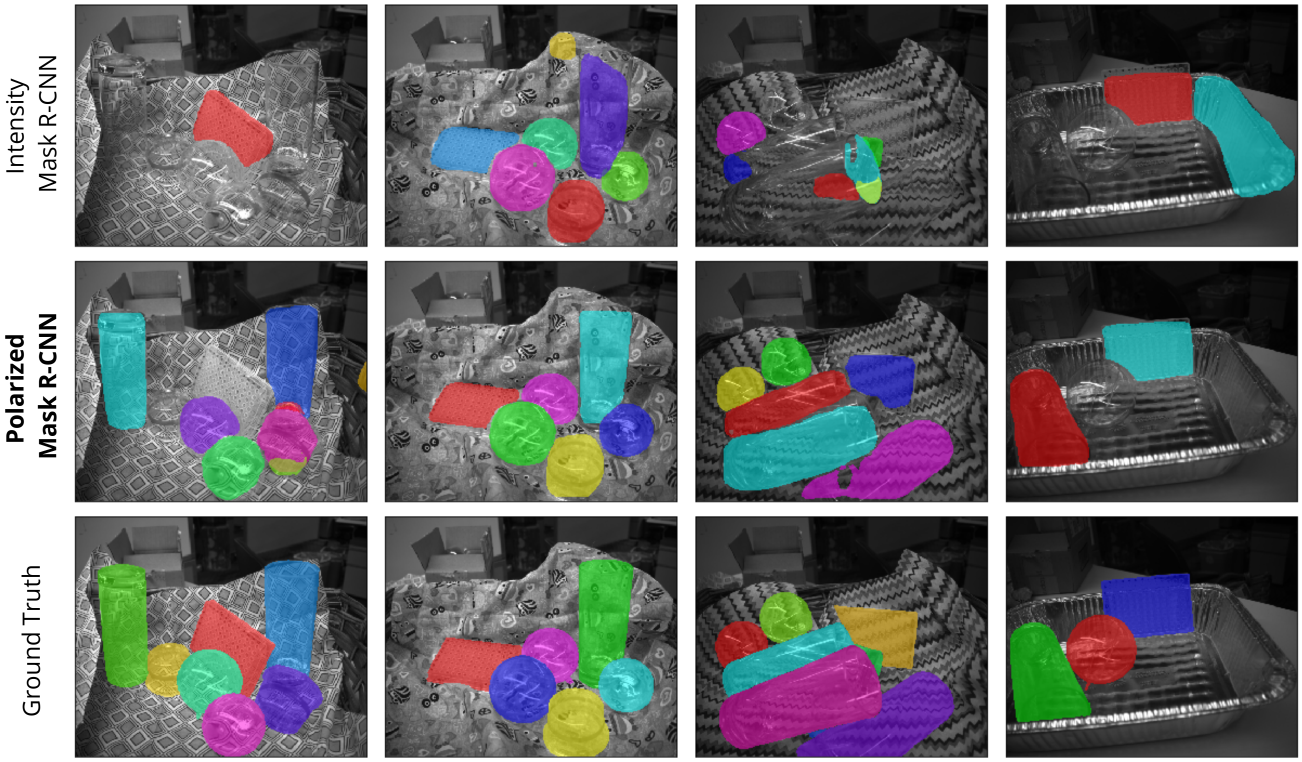


Figure 3: Qualitative comparisons between intensity and polarized CNNs for the case of transparent object segmentation in a new environment (Env dataset).

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	ResNet-101	0.662	0.434	0.885	0.694	0.277	0.13	0.681	0.546	0.803	0.364
ϕ	ResNet-101	0.696	0.507	0.847	0.6	0.283	0.157	0.833	0.694	0.822	0.577
ρ	ResNet-101	0.740	0.551	0.871	0.667	0.446	0.21	0.8	0.698	0.842	0.629
I_0,I_{45},I_{90},I_{135}	Concat + ResNet-101	0.752	0.560	**0.895**	0.722	0.397	0.217	0.868	**0.793**	0.848	0.508
I, ϕ, ρ	Concat + ResNet-101	0.718	0.544	0.872	0.663	0.287	0.114	0.833	0.742	**0.878**	**0.655**
I, ϕ, ρ	Mid-Fusion + Mean	**0.794**	**0.599**	**0.894**	0.723	**0.512**	0.291	0.886	0.772	**0.883**	0.608
I, ϕ, ρ	Mid-Fusion + Concat	0.771	0.586	**0.894**	0.715	0.471	0.261	0.843	0.746	0.874	0.623
I, ϕ, ρ	Mid-Fusion + MoE	0.780	0.578	**0.89**	0.711	0.476	0.249	0.871	0.718	**0.883**	0.634
I, ϕ, ρ	Mid-Fusion + SE Merge	0.768	0.588	**0.896**	**0.741**	0.453	0.213	0.844	0.764	**0.879**	0.632
I, ϕ, ρ	Mid-Fusion + Attention	**0.796**	**0.601**	**0.893**	0.723	**0.516**	**0.299**	**0.893**	0.758	**0.883**	0.624

Table 3: Detection ablation analysis. Here attention is on par with other methods rather then slightly better. This is because one benefit of spatially-aware fusion is the ability to select the strongest features along the edge, but here there is no segmentation of the edge and thus this value is decreased.

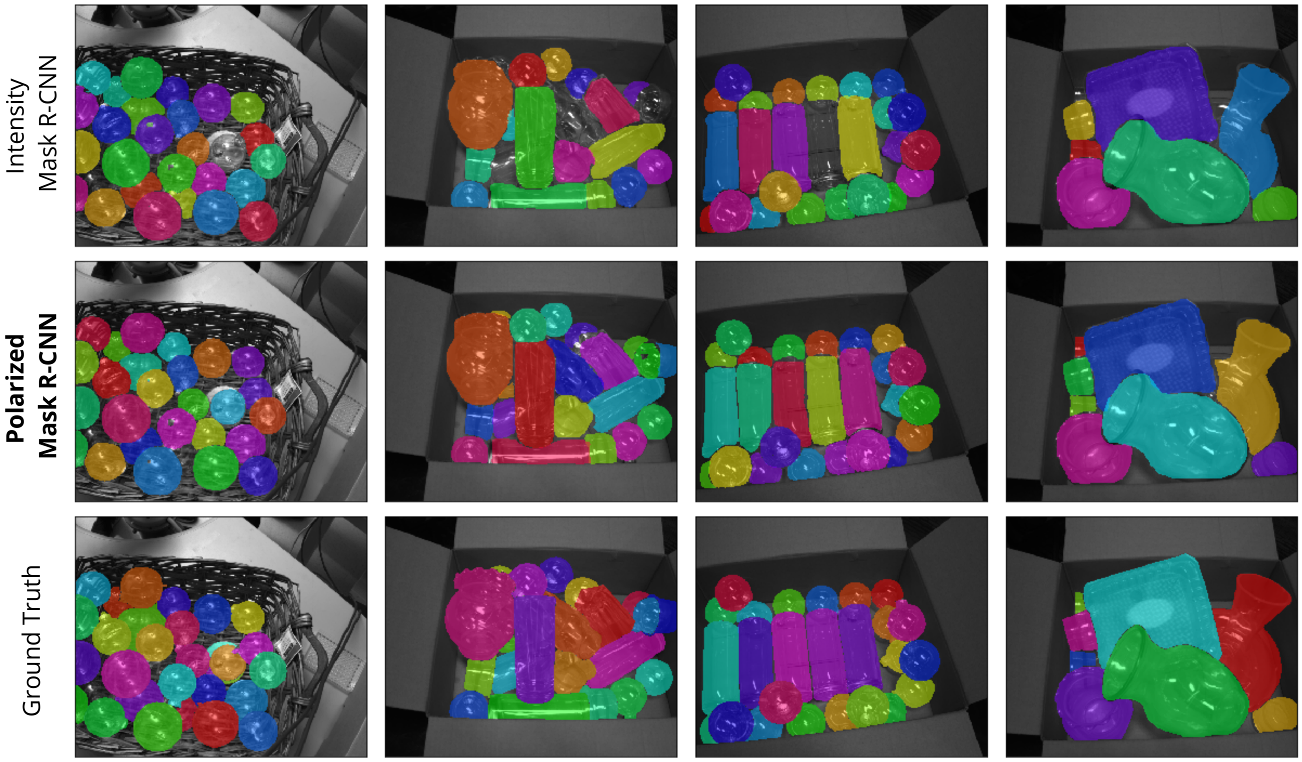


Figure 4: Qualitative comparisons between intensity and polarized CNNs for the case of transparent object segmentation in the presence of clutter (Clutter Dataset).

Model Info			
Input Type	Backbone	Runtime (seconds)	Param Count
I	ResNet-101	0.109	63,760,316
ϕ	ResNet-101	0.106	63,760,316
ρ	ResNet-101	0.107	63,760,316
I_0,I_{45},I_{90},I_{135}	Concat + ResNet-101	0.113	63,760,316
I, ϕ, ρ	Concat + ResNet-101	0.110	63,760,316
I, ϕ, ρ	Mid-Fusion + Mean	0.152	149,076,668
I, ϕ, ρ	Mid-Fusion + Concat	0.160	165,792,188
I, ϕ, ρ	Mid-Fusion + MoE	0.155	149,831,624
I, ϕ, ρ	Mid-Fusion + SE Merge	0.167	167,893,388
I, ϕ, ρ	Mid-Fusion + Attention	0.155	149,814,984

Table 4: Runtime results and parameter count for ablation analysis. Everything is measured on a single P100 GPU. Here adding the mid-fusion backbone adds almost 2.5x the paramters, but only 1.5x the runtime, and it is still only 150ms. The backbones requiring concatenation add an extra 10-12 million parameters to account for the increased size of the tensor.

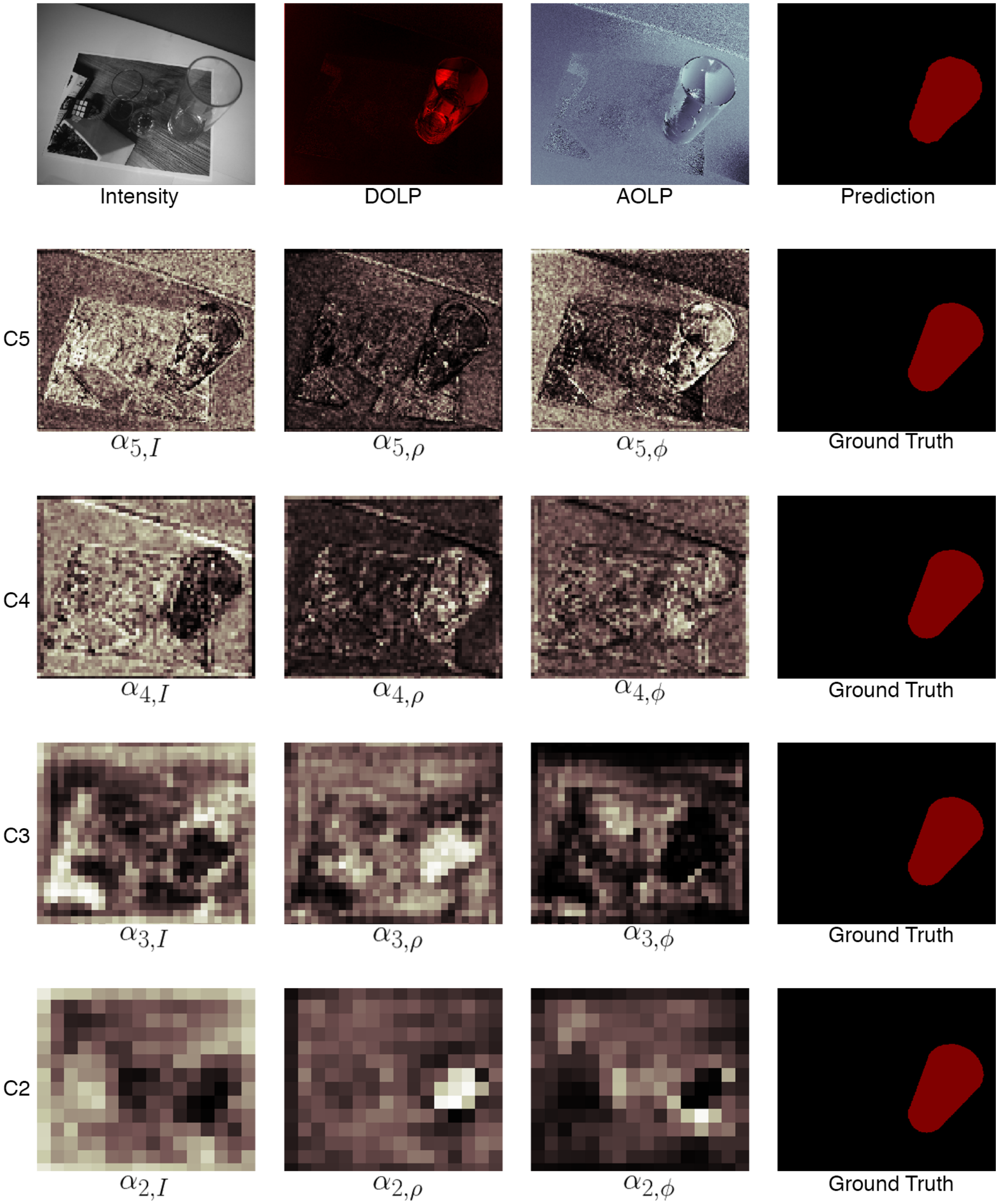


Figure 5: Attention maps visualized for a print-out spoofs. Here we see that the AOLP/DOLP are being used heavily to determine where the object is real or not. In α_5 we see that the majority of weight on the real object comes from the AOLP. Then in α_4,α_3,α_2 the DOLP is being heavily used to highlight the main object again. This allows the model to effectively be robust to print-out attacks, as it is using those two channels to decide if an object is real or fake.

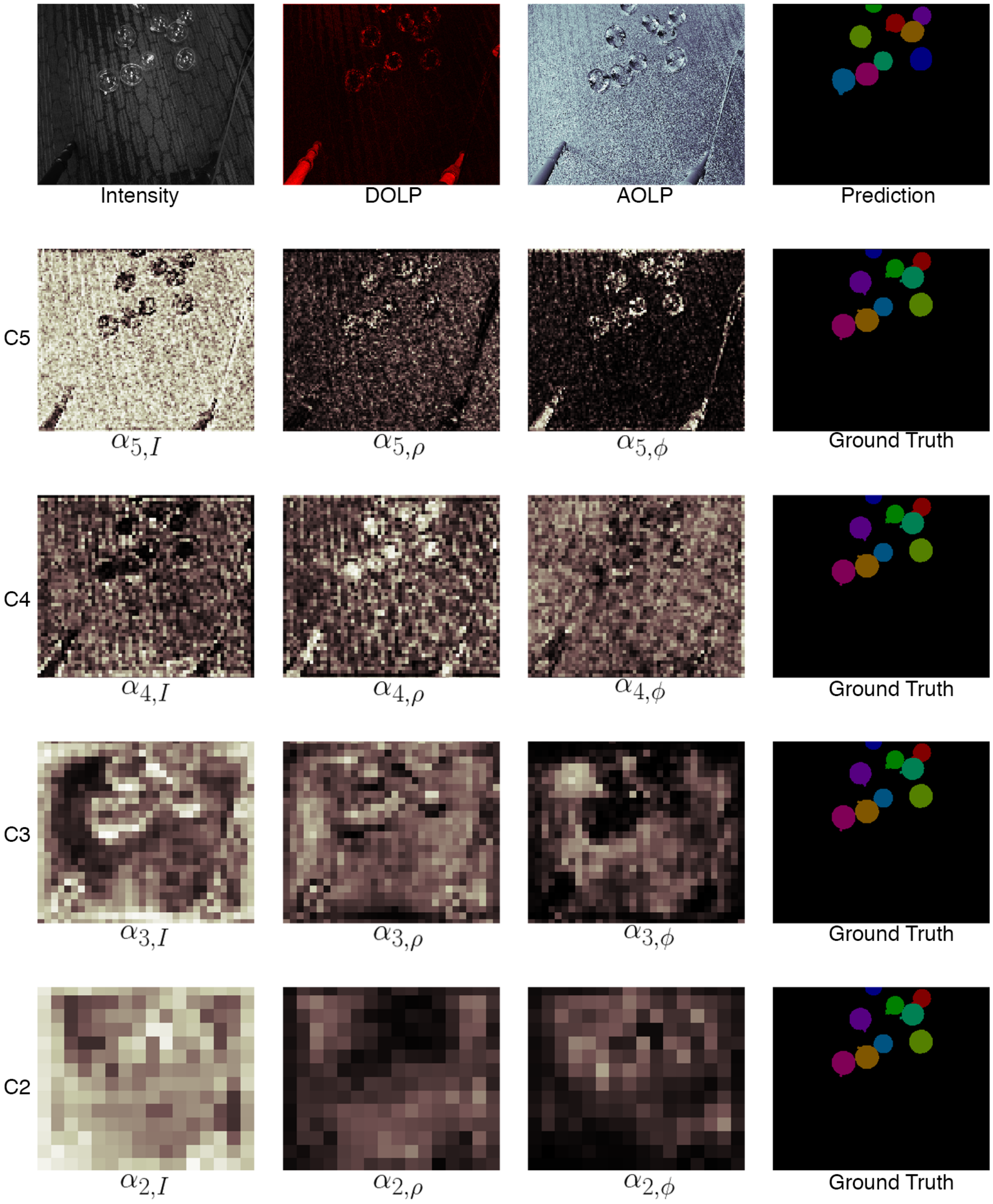


Figure 6: Attention visualization on 7 ornaments. Here we see that the AOLP is being used in α_5 to give features near the edges of the ornament at the highest resolution. This helps explain the improvement in performance we see in the fine-grained segmentation using the attention mechanism.

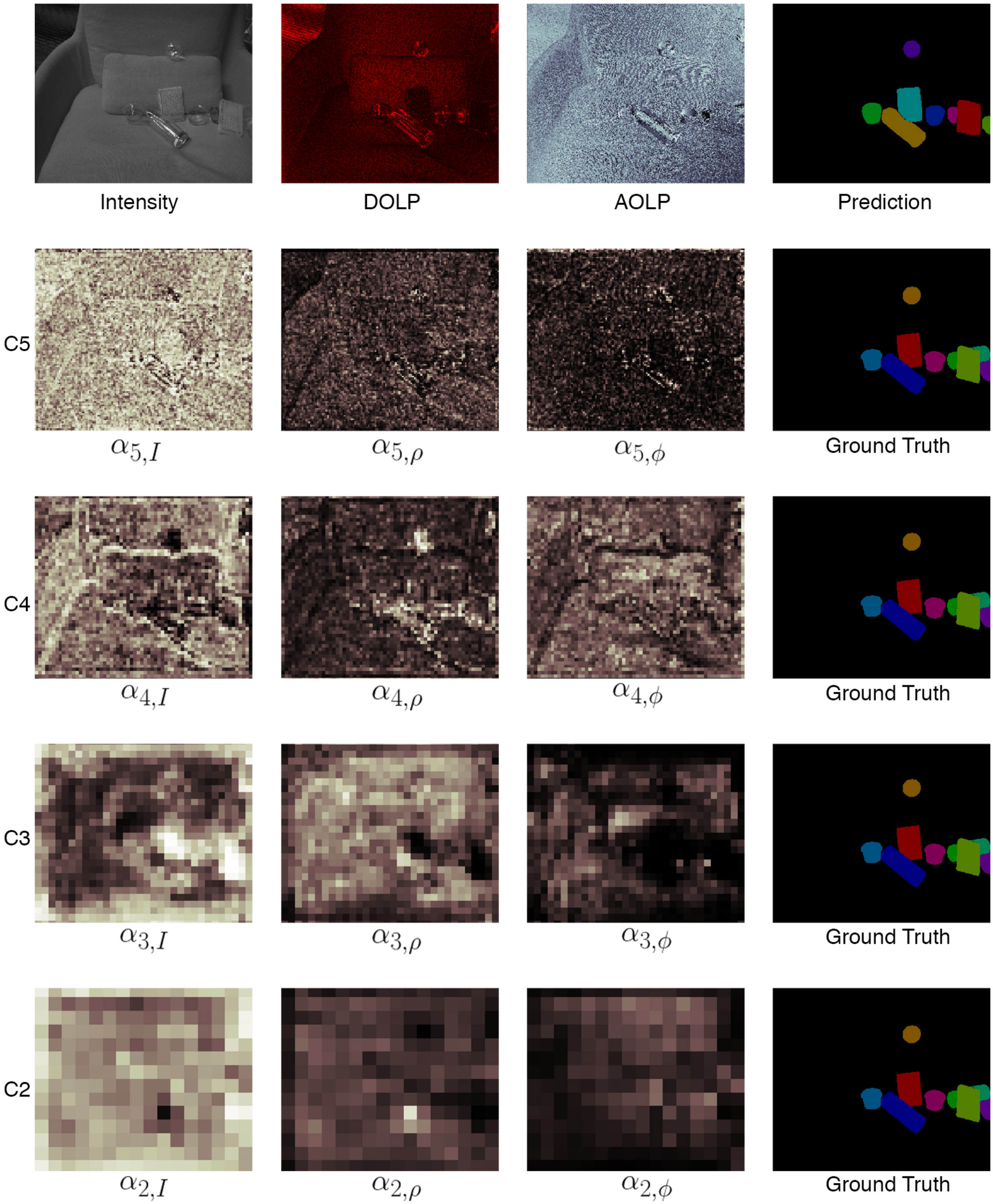


Figure 7: Attention visualization on couch scene. $\alpha_{4,\rho}$ shows very clearly that the DOLP is being used to identify objects in different backgrounds, which makes sense considering the great scores it achieved in the novel background dataset.

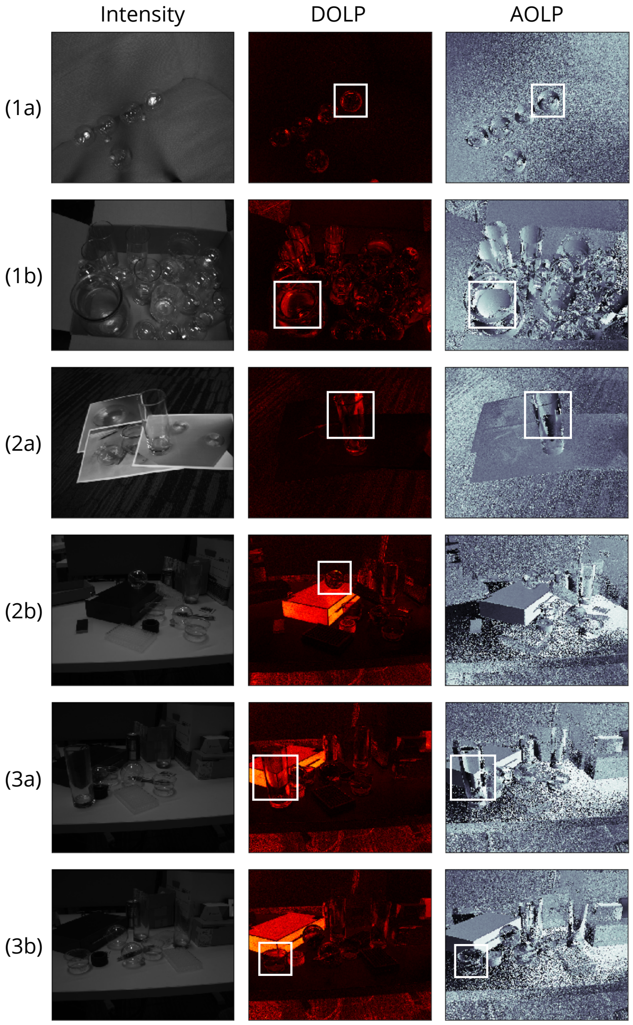
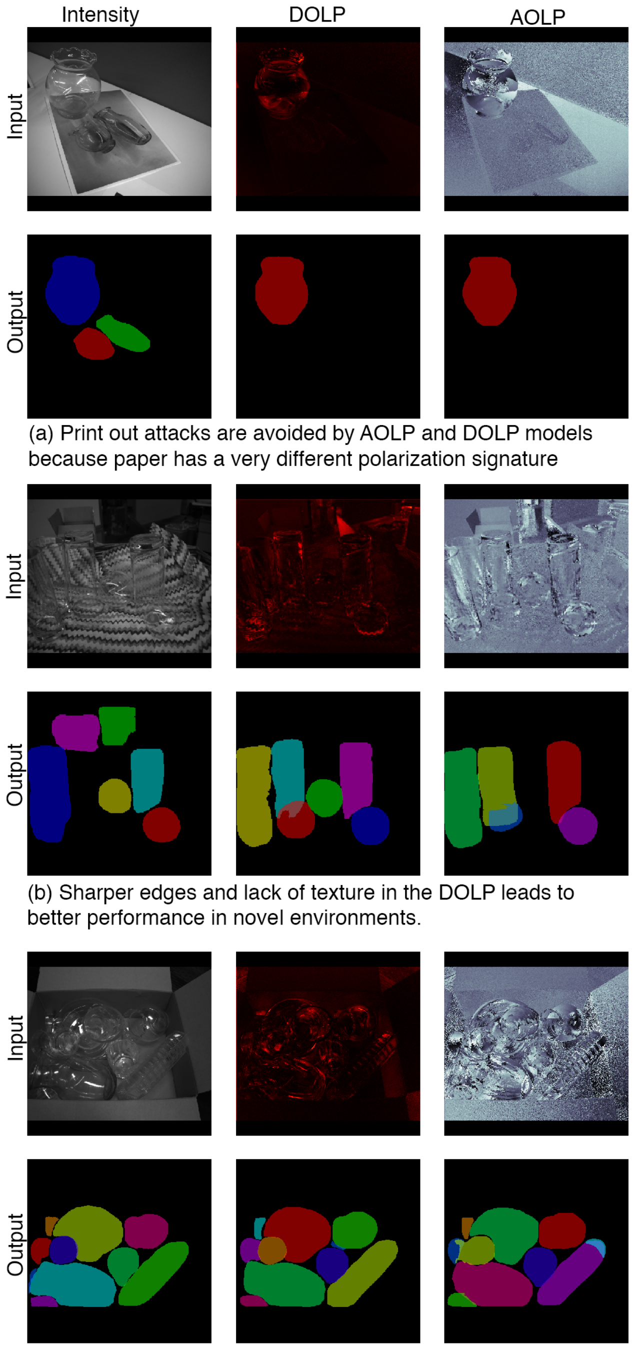


Figure 8: (1) - $I_t\rho_t \ll I_r\rho_r$. (1a) - background light's polarized component is close to 0, hence, transparent balls are contrasting both in AOLP and DOLP, polarizing the light as if they would be opaque. (1b) - again, transparent object is polarizing the light much stronger than the background, so the AOLP signal smoothly follows the shape of the transparent object. (2) - $I_t\rho_t \sim I_r\rho_r$. With comparable polarized components the resulting AOLP can become low, contrasting the object on the polarized background, as in (2b) or making the internal part of the object inconsistent as in (2a). (3) - $I_t\rho_t \gg I_r\rho_r$. Both (3a) and (3b) show how the highly polarized object in the background makes transparent object look transparent in the AOLP and DOLP channels too. However, even in this case transparent object is still contrasting, because the polarization properties of the background light change after refraction through the transparent object.



(a) Print out attacks are avoided by AOLP and DOLP models because paper has a very different polarization signature

(b) Sharper edges and lack of texture in the DOLP leads to better performance in novel environments.

(c) In a cluttered scene, all models struggle, but intensity images do well as they contain the most information.

Figure 9: Here we give three examples, each showing the different polarization channels and it's effect on the output.

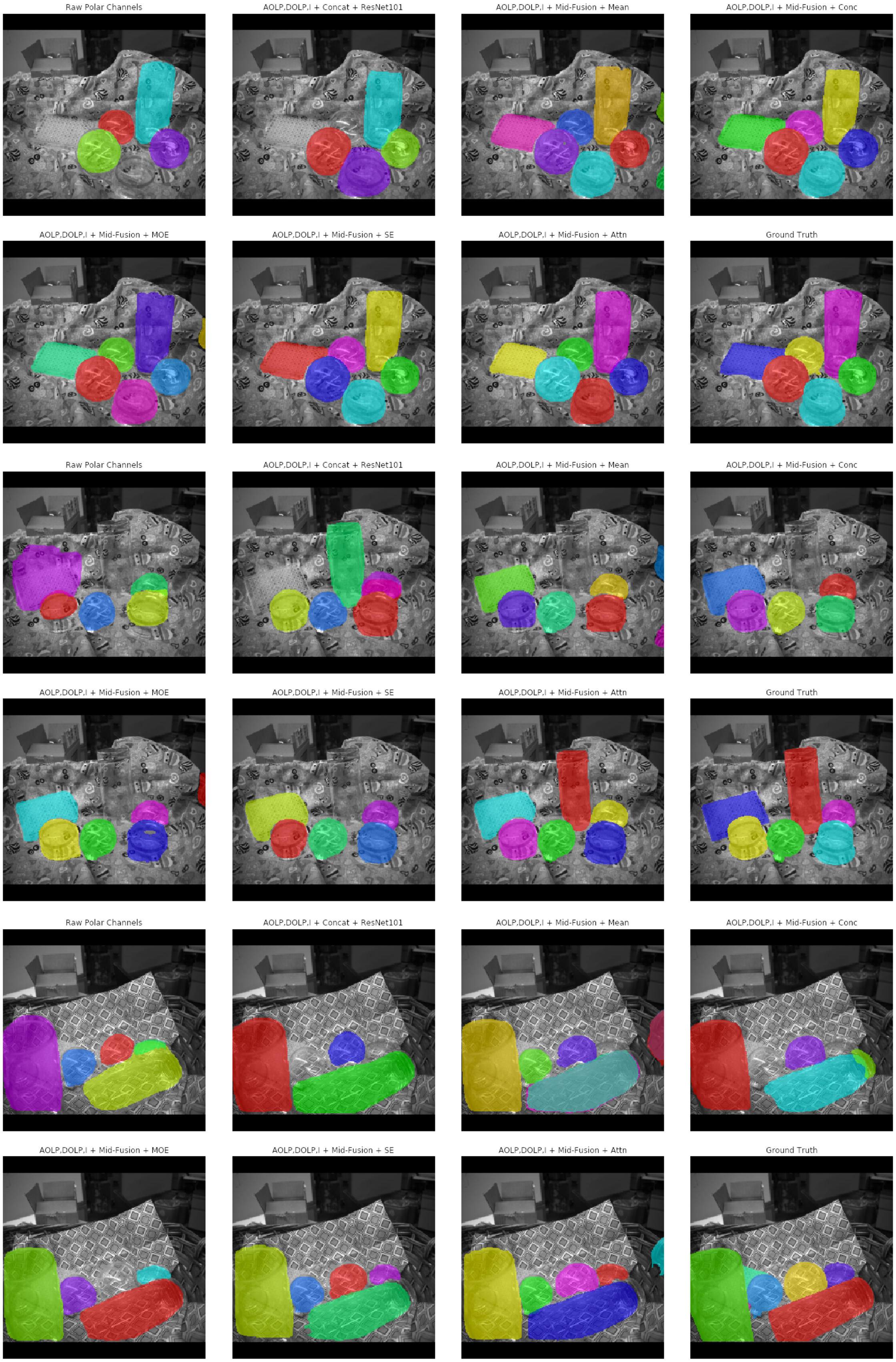

Raw Polar Channels
AOLP,DOLP,I + Concat + ResNet101
AOLP,DOLP,I + Mid-Fusion + Mean
AOLP,DOLP,I + Mid-Fusion + Conc
AOLP,DOLP,I + Mid-Fusion + MOE
AOLP,DOLP,I + Mid-Fusion + SE
AOLP,DOLP,I + Mid-Fusion + Attn
Ground Truth
Raw Polar Channels
AOLP,DOLP,I + Concat + ResNet101
AOLP,DOLP,I + Mid-Fusion + Mean
AOLP,DOLP,I + Mid-Fusion + Conc
AOLP,DOLP,I + Mid-Fusion + MOE
AOLP,DOLP,I + Mid-Fusion + SE
AOLP,DOLP,I + Mid-Fusion + Attn
Ground Truth
Raw Polar Channels
AOLP,DOLP,I + Concat + ResNet101
AOLP,DOLP,I + Mid-Fusion + Mean
AOLP,DOLP,I + Mid-Fusion + Conc
AOLP,DOLP,I + Mid-Fusion + MOE
AOLP,DOLP,I + Mid-Fusion + SE
AOLP,DOLP,I + Mid-Fusion + Attn
Ground Truth

Figure 10: Here we give three examples, each showing one of the 7 models from the second half of the ablation and its performance on a novel environment.