Abstract

Transparent and semi-transparent materials pose significant challenges for existing scene understanding and segmentation algorithms due to their lack of RGB texture which impedes the extraction of meaningful features. In this work, we exploit that the light-matter interactions on glass materials provide unique intensity-polarization cues for each observed wavelength of light. We present a novel learning-based glass segmentation network that leverages both trichromatic (RGB) intensities as well as trichromatic linear polarization cues from a single photograph captured without making any assumption on the polarization state of the illumination. Our novel network architecture dynamically fuses and weights both the trichromatic color and polarization cues using a novel global-guidance and multi-scale self-attention module, and leverages global cross-domain contextual information to achieve robust segmentation. We train and extensively validate our segmentation method on a new large-scale RGB-Polarization dataset (RGBP-Glass), and demonstrate that our method outperforms state-of-the-art segmentation approaches by a significant margin.

1. Introduction

Autonomous robots, aerial drones, and self-driving vehicles rely on an array of sophisticated sensors and algorithms that enable them to sense and understand their environment. However, objects with transparent or semi-transparent materials remain an open challenge for existing scene understanding methods. In contrast to opaque materials, transparent materials typically lack texture, and their complex dynamic appearance depends over various local and global properties, ranging from light-matter interactions (i.e., reflection, refraction, and transmission), object shape, and background, resulting in out-of-distribution observations that are difficult to model.

The majority of existing segmentation methods for transparent materials leverage either contextual information [27, 41] or rely on boundary detection [11, 40]. Both strategies operate in the RGB domain where the interactions between light waves and transparent materials only produce weak cues. A few works have investigated leveraging richer representations of light-matter interactions for transparent material recognition, such as light fields [23, 34, 43] and polarization [17, 19, 20, 37, 39]. However, these method also rely on strong assumption on the target size and reflectivity, or assume restricted capture conditions.

In this work, based on that glass materials often provide a distinctive spectral-polarimetric response, we lever-
age both trichromatic intensity and trichromatic linear polarization cues from images captured in-the-wild to infer rich contextual information for robust transparent material segmentation. Linear polarization cues, described by the degree of linear polarization (DoLP) and the angle of polarization (AoLP), can provide strong cues \cite{17} for transparent object segmentation (Figure 1) and can be thought of as intrinsic object textures for transparent materials. However, depending on the view and lighting conditions, these cues might not be equally informative over all three wavelengths, or even confound valid RGB intensity cues.

To address these challenges, we design a Polarization Glass Segmentation Network, which we dub “PGSNet”, that utilizes an Early Dynamic Attention (EDA) module to dynamically estimate three global scaling weights for each channel of the trichromatic DoLP and AoLP. The weighted DoLP and AoLP, together with the RGB image features, are fed into a Conformer \cite{31} backbone network to extract robust global and local features. The multi-modal local features are then fused by a Dynamic Multimodal Feature Integration (DMFI) module guided by the global features, and subsequently used by a Global Context Guided Decoder (GCGD).

To train PGSNet, we introduce a large-scale RGB-Polarization dataset, dubbed RGBP-Glass, which contains 4,511 manually annotated RGB intensity images and the corresponding trichromatic (i.e., RGB) AoLP and DoLP images. To ensure diversity, we capture the images in the RGBP-Glass dataset from different real-world scenes that have significant variations in location, type, shape, color contrast, and light conditions.

We demonstrate the effectiveness of our approach and show the importance of multi-chromatic polarization cues for glass segmentation. Our extensive experiments show that our method significantly outperforms competing methods. We make the following contributions:

- the first learning based method to exploit multi-chromatic polarization cues for glass segmentation on photographs taken in-the-wild;
- a novel attention-based glass segmentation network that dynamically fuses RGB and multi-chromatic polarization cues; and
- a new and unique large-scale RGB-P glass segmentation dataset.

2. Background and Related Work

**Polarization.** Light is composed of transverse waves of electric and magnetic fields, and its polarization state describes the orientation of the transverse electric field. Within a non-zero finite time of observation, this orientation can be randomly distributed (unpolarized), biased toward a single direction (linearly polarized), or in between the two extremes (partially linearly polarized). We focus our discussion on linear polarization supported by emerging polarization-array CMOS sensors, and omit polarization states such as circular and elliptical polarization. Typically, these ‘polarization’ cameras record four linear polarization states of light: $I_0$, $I_{45}$, $I_{90}$, and $I_{135}$, where $I_x$ describes the image captured by a linear polarizer at the angle $x$.

The polarization state of light can be described using a Stokes vector $S = [S_0, S_1, S_2, S_3]$, where $S_0$ stands for the total light intensity, $S_1$ and $S_2$ describe the ratio of the $0^\circ/45^\circ$ linear polarization over its perpendicular counterpart, and $S_3$ is the circular polarization power. The Stokes elements $S_0, S_1, S_2$ can be computed from the measurements $I_0, I_{45}, I_{90},$ and $I_{135}$ as:

$$
S_0 = I_0 + I_{90} = I_{45} + I_{135},
S_1 = I_0 - I_{90},
S_2 = I_{45} - I_{135}.
$$

The degree of linear polarization (DoLP) and angle of linear polarization (AoLP) are then formally defined as:

$$
\text{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \quad \text{AoLP} = \frac{1}{2} \arctan \left( \frac{S_2}{S_1} \right).
$$

The type and composition of materials are known to be highly correlated to the DoLP and AoLP observations \cite{4} as illustrated for transparent glass materials in Figure 2. However, this correlation is often challenging to analytically characterize for real-world scenes due to the many factors that contribute to the observations, and a key challenge that we address through the various components that comprise PGSNet (section 4).

We are not the first to consider polarization cues. The use of polarization cues has a rich history in computer vision for a wide range of tasks such as estimating shape and/or surface normals (e.g., \cite{1–3, 6, 16, 33}), reflectance component separation (e.g., \cite{19, 20, 37}), and semantic segmentation (e.g., \cite{17, 39}).

**Transparent Object Segmentation.** The majority of glass object segmentation techniques work on regular RGB images \cite{11, 27, 40, 41, 46}. While these methods have been able to achieve impressive results, RGB images only provide weak glass segmentation cues and the efficacy of these methods is reduced for cluttered scenes and print-out spoofs \cite{17}. To improve robustness, richer records of light-matter interactions have been considered for transparent and semi-transparent object segmentation, such as distortions due to transparency in light-fields \cite{23, 34, 43} and depth information \cite{10, 32}. Despite the richer input sources, these methods still rely on additional assumptions such as weak specular reflections \cite{23, 34, 43}, limited object shapes \cite{10}, or isolated objects \cite{32}, thereby limiting their generality.

Closest related to our work is the glass segmentation network of Kalra et al. \cite{17} that takes as input both intensity image as well as polarization cues (i.e., AoLP and
Figure 2. RGBP-Glass Examples. For each exemplar we show two rows, with in the first column the RGB intensity (top) and reference glass segmentation (bottom), and in the last three columns the polarization measurements for the red, green, and blue channels, respectively (top: AoLP, bottom: DoLP). The top exemplar exhibits clear glass cues in both RGB and polarization. The middle exemplar features weak intensity cues, but a strong polarization cues in the red channel. The bottom exemplar does not show strong cues in either RGB or polarization.

DoLP). However, Kalra et al. focus on robotic bin picking and train their network on a proprietary training set of 1,600 monochromatic images of small transparent objects, ignoring potential wavelength dependent cues embedded in the AoLP and the DoLP. The lack of a large-scale dataset containing in-the-wild transparent objects such as glass walls and windows precludes the exploitation of polarization cues for more general application scenarios. While we also exploit polarization cues, our glass segmentation network (PGSNet) differs in two critical aspects from the method of Kalra et al. First, we use trichromatic polarization cues and introduce a publicly-available large-scale RGB-P dataset of in-the-wild transparent objects. Second, whereas Kalra et al. only leverage local contextual attention, our method is guided by both global and local contextual attention.

3. RGB-P Glass Segmentation Dataset

We collected a large-scale polarization glass segmentation dataset, named RGBP-Glass, using a trichromatic polarizer-array camera (LUCID PHX050S) that records four different linear-polarization directions (0°, 45°, 90°, and 135°) for each color channel (i.e., R, G, and B) at a 612 × 512 resolution per polarization direction. RGBP-Glass contains 4,511 RGB intensity and corresponding pixel-aligned trichromatic AoLP and DoLP images with manually annotated pixel-level accurate reference glass-masks and associated bounding-boxes. Each image in RGBP-Glass contains at least one in-the-wild glass object. To ensure diversity of scenes, we capture the dataset from different locations, view angles, lighting conditions, types of glass, and shapes of glass. The polarization filter mask of the camera reduces the light efficiency of the sensor, and we compensate for this by using a f/1.6 aperture and manually adjust the exposure time. Table 1 compares RGBP-Glass to other similar datasets, and Figure 2 provides representative examples. To avoid overfitting to glass location, object size or number of glass instances, we ensure RGBP-Glass covers a wide distribution of glass locations (Figure 3(a)), ratio of glass area (Figure 3(b)), and number of glass instances per image (Figure 3(b)). To the best of our knowledge, RGBP-Glass is the most extensive publicly-available RGB-P-based dataset for glass-like object segmentation tasks.

4. Spectral-Polarimetric Glass Segmentation

The three selected examples in Figure 2 show that polarization measurements can provide strong additional cues for glass segmentation. However, naively including these measurements in existing glass segmentation networks does not necessarily yield the expected improvement in performance. In typical cases, both RGB and polarization observations provide meaningful cues for glass segmentation (e.g., Figure 2(a)). However, under certain light conditions
and/or view angles, the polarization cues may be weak or even non-existent, providing no meaningful cues for segmentation (e.g., Figure 2(c)). Similarly, under adverse conditions (e.g., fog), RGB intensities might not provide meaningful cues either. Furthermore, even within a modality, the cues provided by the different color channels might not be equally important (e.g., Figure 3(b)), or even provide contradictory cues. Effectively and dynamically fusing between and within the multimodal cues is essential for robust multimodal glass segmentation.

We introduce a novel Polarization Glass Segmentation Network (PGSNet) that aims to dynamically fuse multimodal intensity and polarization measurements for robust segmentation by leveraging both local and global contextual information. PGSNet follows an encoder-decoder architecture, summarized in Figure 4(a). During encoding, an early dynamic attention module (EDA; subsection 4.1) estimates global scaling weights for balancing the different color channels within each of the trichromatic AoLP and DoLP. Next, the weighted trichromatic AoLP and DoLP measurements are passed into three separate Conformer [31] branches for feature extraction. The goal of the Conformer stage is to balance differences between glass and non-glass objects within each of the different sources. For example, if there is no or little polarization observed on glass-like objects, then PGSNet should leverage any potential global and local contextual information between glass and non-glass objects in the polarization cues. In the final encoding step, we employ a novel

Dynamic Multimodal Feature Integration (DMFI) module (subsection 4.2) to dynamically fuse together the extracted local features from the three input sources (i.e., RGB, AoLP, and DoLP) guided by the global features.

During decoding, we rely on the global contextual cues to avoid over-segmentation. To avoid diluting global context features with subsequent steps in the decoding pipeline, we introduce a novel Global Context Guided Decoder (GCGD; subsection 4.3) that employs an Attention Enhancement (AE) module to dynamically provide global guidance based on the multimodal global features from the three Conformer branches.

### 4.1. Early Dynamic Attention (EDA)

The purpose of the EDA module is to estimate global weight factors to balance the color channels in the AoLP and DoLP measurements. We employ a ResNet-18 [13] (with shared weights between color channels) followed by a fully connected layer and a SoftMax operator to estimate appropriate weights for each of the color channels. Formally, the EDA module can be denoted as:

\[
    \begin{align*}
        w_r, w_g, w_b &= \sigma(G(p_r), G(p_g), G(p_b)), \\
        P &= [w_r p_r, w_g p_g, w_b p_b],
    \end{align*}
\]

where \( p_{\{r,g,b\}} \) are the red, green, or blue polarization measurements (AoLP or DoLP) with weights \( w_{\{r,g,b\}} \) respectively; \( [\cdot, \cdot, \cdot] \) indicates the concatenation operation over the channel dimension; \( \sigma \) is the SoftMax function; \( \langle \cdot, \cdot, \cdot \rangle \) denotes a vector; and \( G \) is the weight estimation network.
4.2. Dynamic Multimodal Feature Integration (DMFI)

The importance of the cues gathered from the different modalities (i.e., RGB intensity, AoLP, and DoLP), is scene-dependent (cf. Figure 2). A naive combination of these cues can dilute the impact of strong cues with weak signals, or even amplify adverse effects of confounding cues. A Dynamic Multimodal Feature Integration (DMFI) addresses the robust fusing of features from the three input domains by leveraging global and local information. The DMFI module, illustrated in Figure 4(b), consists of two blocks: a Dynamic Fusion (DF) block and a Multi-Scale Dependency Perception (MSDP) block.

**Dynamic Fusion (DF).** The DF block first generates three spatial attention maps on the three sequences of token embeddings provided by three Conformers [31] for each of the three input modalities (see the supplemental material for details on Conformers). The extracted convolution features are subsequently weighted by the attention maps and fused (summer) together:

\[
M^1, M^2, M^3 = \sigma(\Omega(T^1), \Omega(T^2), \Omega(T^3)),
\]

\[
F_{DF} = M^1 \otimes C^1 + M^2 \otimes C^2 + M^3 \otimes C^3, \tag{4}
\]

where \(M\) are the attention maps generated from \(I, \phi, \) and \(\rho,\) the RGB intensity, AoLP, and DoLP input respectively, and \(\Omega\) is a function that first reduces the dimensions of every token embedding to one via a fully connected layer, and then subsequently reshapes the resulting embedding to a 2D map. \(C\) and \(T\) are the convolution features and token embeddings generated by the \(\text{conv}\) and the \(\text{trans}\) branch in the Conformer [31], respectively, where the superscript denotes the index of Conformer’s internal block, and \(\otimes\) is the element-wise multiplication.

**Multi-Scale Dependency Perception (MSDP).** To reduce the impact of shape variations and locations of the glass objects, the MSDP block enhances the global dependencies for locating glass objects in the dynamically fused feature \(F_{DF}\) using a specially designed multi-scale self-attention mechanism. By varying the perceptive scales, the MSDP block can effectively detect correlations between regions at different scales. Formally:

\[
F_V = \psi^{br}_\nu(F_{DF}),
\]

\[
F^n_{DP} = \Upsilon^n(F_V) = F_V + \alpha \ast \mathcal{U}(\mathcal{A}^n(F_V)),
\]

\[
F_{MSDP} = [F_{DF}, F^0_{DP}, F^1_{DP}, F^2_{DP}, F^{11}_{DP}], \tag{5}
\]

where \(\psi^{br}_\nu\) is a \(k \times k\) convolution layer followed by a Batch Normalization (BN) and ReLU activation function. \(\mathcal{A}^n\) is an adaptive average pooling with target size \(n \times n,\) \(\mathcal{U}\) is a bilinear upsampling, and \(\alpha\) is a learnable parameter. \(\mathcal{N}(x)\) is the self-attention operation defined as \(\mathcal{V}(x)(\sigma(\mathcal{K}(x)^T \mathcal{Q}(x))\); \(\mathcal{Q}, \mathcal{K}, \) and \(\mathcal{V}\) are three learnable linear embedding functions, implemented as three fully connected layers. Our MSDP block is similar to existing attention schemes (e.g., PPM [47], ASPP [5] non-local attention [35]). We refer to the supplementary material for additional experiments validating that MSDP outperforms prior schemes.

The final output of the DMFI block applies an additional 3 \(\times\) 3 convolution to the output features of the MSDP block:

\[
F_{DMFI} = \psi^{br}_\nu(F_{MSDP}).
\]

### 4.3. Global Context Guided Decoder (GCGD)

Global contextual cues are essential to avoid over-segmentation during the decoding phase. Typically, these global contextual cues are injected in the decoder via the high-level features. However, as the decoding process proceeds to lower-level features, the influence of the global contextual features dilutes. To retain the global contextual information during the decoding process, we introduce a novel Global Context Guided Decoder (GCGD) that consists of a Global Context Generation (GCG) module (Figure 4(c)) that forms global guidance cues across the three input domains, and an Attention Enhancement (AE) module (Figure 4(d)) that leverages these global guidance cues to enhance the low-level features.

**Global Context Generation (GCG).** Key to the GCG is the observation that the token embeddings \(T^1, T^2, \) and \(T^3\) from the Conformers [31] are inherently global-aware characteristics. We leverage these token embeddings by first computing a set of cross-correlation features:

\[
F_{xy} = \mathcal{N}(T^3_x, T^4_y),
\]

\[
en = \mathcal{R}(T^3_x, T^4_y),
\]

\[
= T^4_x + \zeta(\mathcal{Q}(T^3_y)\mathcal{K}(T^3_x)^T/\sqrt{d})\mathcal{V}(T^3_x), \tag{6}
\]

where \(xy \in \{I_\phi, I_\rho, I_\phi I_\rho, I_\phi \rho I, \rho_\phi\},\) \(\zeta\) is the sigmoid function, and \(d\) denotes the length of a token embedding. These cross-correlation features are then combined via a linear projection \(\Gamma,\) implemented by a fully connected layer:

\[
T = \Gamma([F_{I_\phi}, F_{I_\rho}, F_{I_\phi I_\rho}, F_{I_\phi \rho I}, F_{\rho_\phi}]). \tag{7}
\]

**Attention Enhancement (AE).** The AE utilizes the combined features from the GCG module to enhance the input features by computing and combining a spatial enhancement map \(E\) and channel features \(e.\) In the GCGD, we deploy four AE blocks, and the decoder features go through the 4th AE block first. Mathematically, the \(j\)-th AE block is defined as:

\[
e_j = \mathcal{R}(F^j) \ast \mathcal{R}(T_g)
\]

\[
E_j = \mathcal{P}_c(F^{j'}) \ast \mathcal{P}_r(t_s, T_g),
\]

\[
F^{j''} = F^{j'} \ast E_j + F^{j'} ,
\]

\[
F^j_{AE} = \psi^{br}_\nu(F^{j''}) \quad j \in [1, 4], \tag{8}
\]
where $F^4 = F_{DMF1}$ and $F^i = F_{BD} = \psi_{ibr}(C'_j + U(\psi_{ibr}(F^{i+1}_{AE})))$, $i \in [1, 3]$; $R(x)$ is the channel feature generator defined as $\zeta(\psi_i(\psi_{ibr}(A^i(x))))$; $\mathcal{P}_C(x)$ is a spatial map generator based on convolution features, defined as $\zeta(\psi_T(x))$; and $\mathcal{P}_T(x, y)$ is also a spatial map generator but based on token embeddings, defined as $\zeta(\Omega(y + Y(x, y)))$. $T_{\phi}$ and $t_{\omega}$ are $n$ glass and segmentation tokens in $T$.

4.4. Loss Function

We supervise both the encoder and decoder during training. For the encoder, we follow the training process for Conformers [31], and apply two loss functions, $\mathcal{L}_m^C$ and $\mathcal{L}_m^T$, for the conv and the trans-branches:

$$\mathcal{L}_E = \sum_{m \in \{I, \phi, \rho\}} \mathcal{L}_{m}^C + \mathcal{L}_{m}^T,$$

where $\mathcal{L}_m^C$ and $\mathcal{L}_m^T$ are both the sum of a binary cross-entropy (BCE) loss $l_{bce}$ and a IoU loss $l_{iou}$ [25].

For the decoder, we apply supervision on the features generated by the deepest three AE modules and the features generated by the GCG module:

$$\mathcal{L}_D = \sum_{i=2}^{4} \mathcal{L}_{AE}^i + \mathcal{L}_{GCG},$$

where the losses on the AE modules and the GCG module are computed again as: $l_{bce} + l_{iou}$. Finally, we combine the losses for both the encoder $\mathcal{L}_E$ and decoder $\mathcal{L}_D$ with the BCE and IoU loss on the final output mask. To promote clear mask boundaries, we also add an edge loss $l_{edge}$ [48] (weighted by $\omega = 10$ empirically determined):

$$\mathcal{L} = \mathcal{L}_E + \mathcal{L}_D + l_{bce} + l_{iou} + \omega l_{edge},$$

5. Assessment

We implemented PGSNet in PyTorch [30] and train our network for 180 epochs with a batch size of 16 using stochastic gradient descent with a momentum of 0.9 and a weight decay of $5 \times 10^{-4}$. We employ the poly strategy [22] and set the initial learning rate and power to 0.001 and 0.9, respectively. We initialize PGSNet randomly, except EDA which is initialized with ResNet-18 [13] and the Conformer-B model [31] which is initialized with a model pre-trained on ImageNet. All input images are resized to 416 × 416 for both training and testing, and the final output is bilinearly resized back to the original input resolution.

We use four metrics for validation and ablation: intersection over union (IoU), weighted F-measure ($F^w_{\beta}$) [24], mean absolute error (MAE), and balance error rate (BER) [28]. For IoU and $F^w_{\beta}$, higher is better, while for MAE and BER, lower is better. We refer to the supplementary materials for a formal definition of each metric.

5.1. Qualitative and Quantitative Evaluation

We extensively compare the effectiveness of our method to 22 state-of-the-art methods across different related tasks such as instance/semantic, salient/camouflaged objects, shadow/mirror segmentation, and glass region/instance segmentation (Table 2). For a fair comparison, all methods are re-trained and tested on the RGB-P Glass segmentation dataset. Of the compared methods, EAFNet [39] and P Mask R-CNN [17] are the only two that also leverage polarization cues. GDNet [27], TransLab [40], Trans2Seg [41], GSD* [21], EAFNet* [39], P Mask R-CNN* [17], and GSD* [21] are in-the-wild glass segmentation methods, but only rely on RGB intensity input. From Table 2 we can see that the proposed method offers the best performance for all four metrics, outperforming the other competing methods by a significant margin. The two polarization-based approaches, P Mask R-CNN [17] and EAFNet [39], do not perform well. P Mask R-CNN [17] extends Mask R-CNN [12] with a cross-domain attention scheme. Mask R-CNN work well on small objects, as is the case for Kalra et al.’s intended task of robotic bin picking, but its performance suffers when segmenting larger objects, even when including polarization cues. Furthermore, P Mask R-CNN

![Table 2](image)

Table 2. Quantitative comparison against state-of-the-art: instance/semantic segmentation methods (marked by the ◦ symbol), salient object detection methods (△), camouflaged object segmentation methods (▽), medical image segmentation method (†), shadow detection method (⋆), mirror segmentation method (⊗), RGB glass segmentation methods (•), RGB+P semantic segmentation method (⊙), monochromatic intensity, and polarization-based glass segmentation methods (∇). All methods are retrained and tested on the RGBP-Glass dataset (except the last row which demonstrates that PGSNet generalizes to other datasets). Methods that require an additional CRF [18] post-processing step are marked with the † symbol. The first, second, and third best results are highlighted in red, green, and blue, respectively.
only uses monochromatic cues for both intensity and polarization, which is less effective than using trichromatic cues. While EAFNet [39] also explored multichromatic DoLP and AoLP, Xiang et al. concluded that EAF-A (i.e., RGB+AoLP) performs best for semantic segmentation with EAFNet, and in our comparisons we follow this approach. However, as our ablation study will show (subsection 5.2), the DoLP is more informative than AoLP for glass segmentation. The lower accuracy of EAFNet is partially because it is designed to solve a more general problem (semantic vs. glass segmentation) and partially because it places a higher emphasis on performance than PGSNet. We refer to the supplemental material for a performance comparison. Finally, we also trained and tested PGSNet on the smaller ZJU-RGB-P dataset (last row of Table 2), demonstrating that PGSNet generalizes well to other datasets with similar performance gains. Figure 5 further qualitatively demonstrates the benefits of our method:

1. The reflections in the glass in the bathroom scene share the same texture as the wall. Only our method is able to accurately segment the glass. The monochromatic polarization information leveraged by P Mask R-CNN as well as the employed fusion scheme are not powerful enough to successfully segment the glass.

2. Glass in metal door-frame: all methods except PGSNet and Trans2Seg confuse the metal material for glass. Trans2Seg’s glass segmentation is less accurate than our method’s result which leverages both the strong polarization cues as well as global contextual information to achieve the best performance.

3. In the 3rd and 4th example, even though the glass is invisible in the RGB intensity image, we still observe strong AoLP and DoLP cues. Despite also leveraging polarization cues, P Mask R-CNN fails on the 4th example. In contrast, our method succeeds thanks to our dynamic context-aware attention-based fusion.

5.2. Ablation Study

Next, we investigate (a) the impact of spectral polarization cues and (b) influence of each component in PGSNet. For each experiment we fully retrain each model.

Impact of Spectral Polarization Cues. We conduct a series of ablation experiments to demonstrate the effects of spectral polarization cues on glass segmentation Table 3: (A) PGSNet baseline; (B) with RGB intensity cues only; (C) with AoLP, but without DoLP; (D) with DoLP, but without AoLP; (E) monochromatic intensity plus monochromatic polarization cues; and (F) RGB intensity cues with monochromatic polarization cues. Comparing B (RGB only) with C, D, or F, we can see that adding any form of polarization cues to the RGB intensity cues improves the segmentation accuracy. Furthermore, we observe that DoLP cues (D) have a greater impact than AoLP cues (C). In contrast to the findings by Kalra et al. [17], the differences between E and F indicate that spectral RGB intensity information has a major impact. Finally, the differences between our baseline (A) and (F) further demonstrates that spectral polarization cues are more informative than monochromatic polarization cues. Figure 6 visually supports the above quantitative observations.

Influence of Early Dynamic Attention (EDA). The EDA module balances the different spectral components in both the DoLP and AoLP. Comparing Table 3 A (with EDA) versus G (without EDA) shows significant performance gain when including EDA, validating the dynamically balancing the contributions of each wavelength.

Influence of PGSNet Components. We demonstrate the influence and importance of each of the components that comprise PGSNet by gradually removing different components. First, we ablate the decoder by removing the GCG from the GCGD (Table 3 H) which results in a reduction in performance compared to the baseline (Table 3 A). Next,

<table>
<thead>
<tr>
<th>RGB Image</th>
<th>Trichromatic Polarization Cues</th>
<th>GDNet</th>
<th>TransLab</th>
<th>Trans2Seg</th>
<th>GSD</th>
<th>P Mask R-CNN</th>
<th>PGSNet (Ours)</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Qualitative comparison of PGSNet against state-of-the-art glass segmentation methods retrained on the RGBP-Glass dataset.
Table 3. Quantitative ablation comparisons showing that: a) spectral and polarization cues promote more robust glass segmentation, and b) all component of PGSNet contributes to the overall performance. We denote the backbone network (EDA + Conformer) with ‘B’, where ‘EDA’ is the Early Dynamic Attention module. ‘BI’ denotes a basic integration unit (i.e., element-wise addition), used for ablating the Dynamic Multimodal Feature Integration (DMFI) module, and ‘BD’ denotes a Basic Decoder used to ablate the Global Context Generation (‘GCG’) module.

5.3. Limitations

When polarization only provides weak or no cues, the effectiveness of our method decreases; Figure 7 demonstrates such a case. However, even without polarization cues, our method (Table 3 B) still performs well compared to prior glass segmentation methods. Even with RGB only input, our method still outperforms existing glass segmentation methods that leverage polarization cues. In addition, PGSNet expects at least one glass object in the photograph, and it fails when no such object is present. Note that this can be resolved by training on RGBP-Glass augmented with images without glass objects from ZJU-RGB-P [39].

6. Conclusion

In this paper we presented a robust glass segmentation network, PGSNet, to dynamically fuse trichromatic intensity and polarization cues recorded in-the-wild. The proposed network includes several novel modules. On the encoder side, a DMFI module integrates multimodal trichromatic measurements by leveraging multi-scale pixel-wise dependencies to dynamically enhance local contextual cues. On the decoder side, a novel GCGD leverages cross-modal global contextual information to provide robust segmentation. To promote polarization as a valuable cue for vision tasks, we also introduce a large-scale RGBP-Glass dataset that we also use to train PGSNet. Our validation and ablations demonstrate the value of trichromatic polarization cues as well as the effectiveness and robustness of PGSNet.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 61972067/U21A20491/U1908214, National Key Research and Development Program of China (2021ZD0112400), and the Innovation Technology Funding of Dalian (2020JJ26GX036). Pieter Peers was supported by NSF grant IIS-1909028. Felix Heide was supported by an NSF CAREER Award (2047359), a Sony Young Faculty Award, and a Project X Innovation Award.
References

[9] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenhui Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In CVPR, 2020. 1, 2
[16] Haiyang Mei, Gepeng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Dengping Fan. Camouflaged object segmentation with distraction mining. In CVPR, 2021. 6
[17] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiong Yan, Xiaoqiang Wei, and Rynson W.H. Lau. Don’t hit me! glass detection in real-world scenes. In CVPR, 2020. 1, 2, 3, 6
[22] Dorian Tsai, Donald G. Dansereau, Thierry Peynot, and Peter Corke. Distinguishing refracted features using light field cameras with application to structure from motion. IEEE Robotics and Automation Letters, 2019. 1, 2
[38] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019. 6