Depth-Aware Mirror Segmentation

Haiyang Mei\textsuperscript{1}  Bo Dong\textsuperscript{2,*}  Wen Dong\textsuperscript{1}  Pieter Peers\textsuperscript{3}  Xin Yang\textsuperscript{1,*}  Qiang Zhang\textsuperscript{1}  Xiaopeng Wei\textsuperscript{1,*}

\textsuperscript{1} Dalian University of Technology  \textsuperscript{2} SRI International  \textsuperscript{3} College of William & Mary

https://mhaiyang.github.io/CVPR2021_PDNet/index

Abstract

We present a novel mirror segmentation method that leverages depth estimates from ToF-based cameras as an additional cue to disambiguate challenging cases where the contrast or relation in RGB colors between the mirror reflection and the surrounding scene is subtle. A key observation is that ToF depth estimates do not report the true depth of the mirror surface, but instead return the total length of the reflected light paths, thereby creating obvious depth discontinuities at the mirror boundaries. To exploit depth information in mirror segmentation, we first construct a large-scale RGB-D mirror segmentation dataset, which we subsequently employ to train a novel depth-aware mirror segmentation framework. Our mirror segmentation framework first locates the mirrors based on color and depth discontinuities and correlations. Next, our model further refines the mirror boundaries through contextual contrast taking into account both color and depth information. We extensively validate our depth-aware mirror segmentation method and demonstrate that our model outperforms state-of-the-art RGB and RGB-D based methods for mirror segmentation. Experimental results also show that depth is a powerful cue for mirror segmentation.

1. Introduction

Mirrors are commonly present in human-made scenes, \textit{e.g.}, as personal grooming aids, to create the illusion of enlarged room size, or to enhance safety to enable looking around corners or behind the viewer. Yet, mirrors confuse many vision systems as they are unable to distinguish real from reflected scenes. Hence, the ability to segment mirrors is essential for better scene understanding and to improve practical applications. Automatic mirror segmentation is a challenging task as mirrors do not exhibit relatively fixed patterns or salient features, making it fundamentally different from other objects/saliency based segmentation/detection problems [18, 37, 59, 64]. Early mirror segmentation solutions relied on user interaction [2] or specialized hardware [50]. Recently, Yang \textit{et al.} [55] introduced MirrorNet, a convolutional neural network, that leverages contextual contrasted features to detect content discontinuities inside and outside the mirror. Lin \textit{et al.} [25] further boost performance by looking at relation and edge cues. However, these learning based methods often fail when the mirror or mirror-like regions exhibit large variations or when the contextual contrast and correlations are too subtle (Figure 1).

Just as 3D perception plays an important role in scene understanding in the human visual system [7], so can depth information help in computer vision for mirror segmentation. A key observation is that mirrors yield an apparent depth that is inconsistent with their \textit{true} depth and the depth of the surrounding environment; the observed apparent depth is the depth of the reflected scene. As a result, this creates obvious depth discontinuities at mirror boundaries (\textit{e.g.}, Figure 1, 2nd column), providing a strong cue for delineating mirrors.

To leverage depth information for mirror segmentation and to stimulate further research in depth-aware mirror segmentation, we present the first RGB-D mirror segmentation dataset of 3,049 exemplars. To promote diversity and quality, we curate our RGB-D mirror segmentation dataset from

\textsuperscript{*} Xin Yang (xinyang@dlut.edu.cn) and Xiaopeng Wei are the corresponding authors. Xin Yang and Bo Dong lead this project.
four widely used publicly available datasets, labeled and segmented by professional annotators. In addition, to efficiently leveraging depth information for mirror segmentation, we design a novel positioning and delineating network (PNet). As the name suggests, PNet consists of two key modules: (i) a positioning module (PM) that detects and locates the mirror by exploring global and local discontinuity and correlation cues in both RGB and depth, and (ii) a delineating module (DM) that captures localized discontinuities by performing a local contextual contrast, again in both RGB and depth, for refining the mirror boundaries. We introduce a novel dynamic weighting scheme to fuse the RGB and depth correlations in the PM to address variability in measurement noise and depth ranges.

We perform extensive validation experiments to demonstrate the efficacy of our approach and demonstrate that depth provides a powerful and complimenting cue for mirror segmentation. In summary, our contributions are:

1. the first solution to consider both RGB and depth for mirror segmentation;
2. a new RGB-D mirror segmentation dataset to stimulate research using depth in mirror segmentation;
3. a novel depth-aware mirror segmentation network that leverages both RGB and depth discontinuities and correlations inside and outside the mirror; and
4. a novel dynamic weighting scheme to fuse RGB and depth correlations.

2. Related Work

Semantic Segmentation classifies and assigns a semantic label to each pixel in an image. Recent semantic segmentation methods \[3, 14, 18, 54, 56, 57, 64, 65\] rely on fully convolutional networks (FCNs) \[30\] to model the contextual information. In addition to contextual information, a number of recent methods have leveraged depth information to complement the RGB semantic segmentation by treating depth as an additional input source to recall, explicitly or implicitly, RGB features \[5, 6, 24, 42\] or by regarding the depth data as an additional supervised signal in multi-task learning \[52, 62\]. However, treating mirrors as an additional semantic category fails to produce satisfactory results as the visible mirror content is further semantically classified \[55\]. In this paper, we also leverage depth and contextual information, but employ a novel model specially designed for accurate mirror segmentation.

Salient Object Detection (SOD) identifies the most visually distinctive objects/regions in an image of a scene. Current state-of-the-art solutions employ convolutional neural networks (CNNs) to exploit different RGB cues and strategies such as multi-level feature aggregation \[22, 17, 60, 67, 37, 33\], recurrent and iterative learning strategies \[61, 10, 47, 49\], attention mechanisms \[27, 4, 51\], and edge/boundary cues \[43, 26, 66\]. Despite great progress, these RGB-based SOD methods are less effective in scenarios with cluttered backgrounds, low-intensity environments, or varying lighting conditions. In these situations, depth cues can provide complementary spatially rich information \[39\]. CNN-based RGB-D SOD approaches can be categorized into early fusion methods that regard depth as an additional channel of input \[44\], late fusion methods that process RGB and depth by two separate backbone networks before fusing for final prediction \[12\], and the recently popular middle fusion methods that fuse intermediate depth and RGB features \[40, 59, 58, 36, 13, 28\]. Our approach falls in this last category. SOD methods cannot directly address mirror segmentation due to a lack of salient features. More importantly, depth features can significantly differ inside a mirror region while conversely those of salient object regions are typically the same. Consequently, leveraging RGB-D information for mirror segmentation requires a carefully designed solution to fully take advantage of both RGB and depth cues.

Mirror Segmentation detects and segments mirror regions in an image of a scene. Early work relied on user interaction \[2\] or specialized hardware \[50\]. Recently, Yang et al. \[55\] introduced MirrorNet, a CNN-based solution that leverages contextual contrast cues in an RGB image to segment mirrors. Lin et al. \[25\] further exploit relation and edge cues to improve mirror segmentation. However, for certain view angles, contextual contrast and correlation inside and outside the mirror regions become too subtle, resulting in a significant degradation in accuracy. Instead of further optimizing segmentation using only RGB cues, our approach leverages an additional depth modality to better identify contextual discontinuities and correlations.

3. RGB-D Mirror Segmentation Dataset

Our first contribution is the introduction of a new RGB-D mirror segmentation dataset, named RGBD-Mirror, which contains 3,049 RGB images and corresponding depth maps. Instead of capturing the RGB-D images ourselves, we compose the RGBD-Mirror from selected exemplars from four popular datasets (\textit{i.e.}, Matterport3D \[2\], SUN-RGBD \[45\], ScanNet \[8\], and 2D3DS \[1\]) to ensure a wide diversity and broad coverage; see Table 1 for a summary and Figure 2(a) for representative examples. Each selected image contains at least one mirror region, and the pixel-level accurate reference mirror-masks are created by professional annotators. To the best of our knowledge, RGBD-Mirror is the first RGB-D mirror segmentation dataset.

Mirrors Location Statistics: a wide distribution of mirror locations and sizes in the dataset is necessary to avoid memorization of the mirror-location instead of learning mirror segmentation. Figure 2(b) plots the probability that a pixel is inside a mirror region. As can be seen, the spatial...
distribution is not center-biased, and the distributions are consistent between testing and training subsets as well as with the whole dataset.

**Color Contrast Statistics:** Ideally, color contrast between regions inside and outside the mirror should be small, otherwise salient color features can bias the mirror segmentation. Figure 2(c) shows the color contrast distributions, measured as a $\chi^2$ distance between the RGB histograms inside and outside the mirror regions [23, 11, 55], for our RGBD-Mirror dataset and other selected datasets (i.e., the shadow detection dataset SBU [46], the saliency detection dataset DUT-OMRON [53], the glass detection dataset GDD [34], and the mirror segmentation datasets MSD [55] and PMD [25]). From this, we can see that our RGBD-Mirror has the lowest color contrast of these datasets.

### 4. Methodology

Our approach builds on two key observations of mirrors. First, mirrors introduce a discontinuity in semantics and in depth. The former can be detected in the RGB domain and has been exploited by prior mirror segmentation work [25, 55]. The latter, depth discontinuity, is a result of depth sensors reporting the depth of the reflected scene rather than the physical depth of the mirror surface. Second, mirrors also induce a correlation between inside and outside the mirror regions. Besides semantic correlation that can be efficiently detected in the RGB domain, there is also a depth correlation since the apparent depth of the reflected scene is typically deeper than the true depth of the mirror and its surroundings. We design our Positioning and Delineating Network (PDBNet) to exploit discontinuity and correlation in both RGB and depth to efficiently segment mirrors. PDBNet (illustrated in Figure 3(a)) feeds an RGB-D image through two different multi-level feature extractors to obtain RGB and depth features. Depth features are extracted by 5 cascaded $3 \times 3$ convolutional blocks (with 8-16-32-64-128 channel configuration) followed by max pooling. We choose ResNet-50 [16] for extracting the RGB features. For computational efficiency, the extracted RGB features are passed through an additional channel reduction convolution, before feeding them, together with the depth features, into either a positioning module (b) or a delineating module (c). The positioning module (PM) estimates the mirror’s initial location using both global and local features in both RGB and depth. The delineating module (DM) refines the mirror boundary based on local discontinuity and the features from the previous level. The prediction from the last DM is used as the final mirror segmentation.

#### 4.1. Positioning Module

Given the highest level RGB and depth features, the PM estimates the initial mirror location, as well as corresponding features for guiding the subsequent DM modules, based on global and local discontinuity and correlation cues in both RGB and depth. Training of the PM is supervised by ground truth mirror masks. Our PM module (Figure 3(b)) consists of two subbranches: a Discontinuity Perception Branch (DPB) and a Correlation Perception Branch (CPB).

The Discontinuity Perception Branch extracts and fuses discontinuity features for RGB ($D^r$), depth ($D^d$), and RGB+depth ($D^{rd}$). Each of these features (we will drop the $r$, $d$, and $rd$ superscript for clarity) is extracted by a common discontinuity block, and is the element-wise addition of local and global discontinuity features, $D_l$ and $D_g$, respectively (i.e. $D = D_l + D_g$). Given a feature $F$, the local discontinuity feature $D_l$ is defined as the difference between a local region and its surroundings:

$$D_l = \mathcal{R}(\mathcal{N}(f_l(F, \Theta_l) - f_s(F, \Theta_s)))$$

where $f_l$, with corresponding parameters $\Theta_l$, extracts features from a local area using a convolution with a kernel size of 3 and a dilation rate of 1, followed by a Batch Normalization (BN) and a ReLU activation function. $f_s$, with corresponding parameters $\Theta_s$, extracts features from the surroundings using a convolution with kernel size 5 and a dilation rate of 2, followed by BN and ReLU. While the local

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Train</th>
<th>Test</th>
<th>Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matterport3D</td>
<td>1,789</td>
<td>1,153</td>
<td>636</td>
<td>78</td>
</tr>
<tr>
<td>SUNRGBD [45]</td>
<td>576</td>
<td>291</td>
<td>285</td>
<td>17</td>
</tr>
<tr>
<td>ScanNet [8]</td>
<td>593</td>
<td>484</td>
<td>109</td>
<td>102</td>
</tr>
<tr>
<td>2D3DS [1]</td>
<td>91</td>
<td>72</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,049</strong></td>
<td><strong>2,000</strong></td>
<td><strong>1,049</strong></td>
<td><strong>202</strong></td>
</tr>
</tbody>
</table>

Table 1. Composition of the RGBD-Mirror dataset.
discontinuity feature captures the differences between local regions and their surroundings, under certain viewpoints, the reflected image has little overlap with its surroundings. This case is represented by the global discontinuity feature:

\[ D_g = \mathcal{R}(\mathcal{N}(f_k(F, \Theta_l) - f_g(\mathcal{G}(F), \Theta_g))), \]

(2)

where \( \mathcal{G} \) is a global average pooling operation, and \( f_g \) (with corresponding parameters \( \Theta_g \)) is a \( 1 \times 1 \) convolution followed by BN and ReLU. The discontinuity block is applied to RGB, depth, and RGB+depth, and the resulting features \( D^r, D^d \) and \( D^{rd} \) are fused to produce the final output of the DPB:

\[ D^{DPB} = \mathcal{R}(\mathcal{N}(\psi_{3\times3}(D^r, D^d, D^{rd}))), \]

(3)

where \([\cdot]\) denotes the concatenation operation over the channel dimension, and \( \psi_{t \times t} \) represents a convolution with a kernel size of \( t \).

The Correlation Perception Branch models correlations inside and outside the mirror. The CPB is inspired by the non-local self-attention model [48] augmented with a dynamic weighting to robustly fuse the RGB and depth correlations. The regular non-local self-attention model [48] is defined as:

\[ Y = g(F)\kappa(F), \]

(4)

where \( \kappa(F) = \text{softmax}(\theta(F)^T\phi(F)) \), \( g, \theta, \) and \( \phi \) are learnable linear embedding functions, and \( F \) is the feature extracted from some input domain. In our case, both RGB and depth can provide such features that yield non-local self-attention cues. Simply combining the RGB and depth features, ignores cases where one of the domains does not exhibit meaningful correlations. For example, the RGB features do not provide meaningful information if there is little or no overlap between the reflected image and the mirror’s surroundings, whereas the depth information might still exhibit strong cues. Similarly, the depth information captured by a depth sensor may be noisier than the RGB information, degrading the relative quality of potential depth correlations. To resolve this issue, we introduce a dynamic weighting that adjusts the importance of an input domain during fusion based on its quality:

\[ Y = g(F^{rd})(\alpha\kappa(F^r) + \beta\kappa(F^d)), \]

(5)

\[ F^{rd} = \mathcal{R}(\mathcal{N}(\psi_{3\times3}(F^r \oplus F^d))), \]

(6)

\[ g(F) = \psi_{1 \times 1}(F), \]

(7)

\[ \theta(F) = \psi_{1 \times 1}(F), \phi(F) = \psi_{1 \times 1}(F), \]

(8)

\[ \alpha = \frac{e^{\mu(F^r)}}{e^{\mu(F^r)} + e^{\mu(F^d)}}, \beta = 1 - \alpha, \]

(9)

\[ \mu(F) = \psi_{1 \times 1}(\mathcal{R}(\mathcal{N}(\psi_{1 \times 1}(\mathcal{G}(F)))))), \]

(10)

where \( F^r \) and \( F^d \) are the input RGB and depth features, and
© is the channel-wise concatenation operator. Finally, to enhance fault tolerance, we use a residual connection with a learnable scale parameter $\gamma$: $C^{\text{CPB}} = \gamma Y \oplus F^{rd}$.

4.2. Delineating Module

Given high-level mirror detection features, either from the PM or the previous level’s DM, the DM refines the mirror boundaries (Figure 3(c)). The core of the DM is a delineating block that takes advantage of local discontinuities in both RGB and depth to delineate the mirror boundaries. Since such refinements should only occur in the region around the mirror, we leverage higher-level features from the previous module (either PM or DM) as a guide to narrow down the potential refinement areas. Given a feature $F$ and corresponding higher-level feature $F^h$, we compute a feature $T$ as:

$$T = R(N(f_i(F \oplus F^h, \Theta_i)) - f_s(F \oplus F^h, \Theta_s)), \quad (11)$$

$$F^h = U_2(R(N(\psi_{3 \times 3}(F^h))), \quad (12)$$

where $U_2$ is a bilinear upscaling (by a factor 2). Similar as before, we apply the delineating block to RGB, depth, and RGB+depth and fuse the features similar as in Eq. 3 to obtain the final output feature $T^{DM}$.

4.3. Loss Function

The PM and three DMs (Figure 3(a)) are trained by supervision. Specifically, we compute the loss between the reference $G$ and mirror segmentation map $S$ predicted according to each of the features generated by the four modules as: $S = \psi_{3 \times 3}(X)$, where $X$ is the output feature from either the PM or DM:

$$L = w_b \ell_{bce}(S, G) + w_i \ell_{iou}(S, G) + w_e \ell_{edge}(S, G), \quad (13)$$

where $\ell_{bce}$ is a binary cross-entropy (BCE) loss [9], $\ell_{iou}$ is a map-level IoU loss [32], $\ell_{edge}$ is a patch-level edge preservation loss [67], and $w_b = 1$, $w_i = 1$, and $w_e = 10$ are the corresponding weights for each of the three loss terms.

The BCE loss is the most widely used loss in the foreground-background segmentation tasks, which calculates the loss for each foreground and background pixel equally and independently. In many cases, the number of background pixels outnumbers the foreground pixels, resulting in biased loss. To compensate for such cases, we also include a map-level IoU loss. Additionally, we also use a patch-level edge preservation loss to assign more attention to the foreground boundary.

The final loss function is then defined as:

$$L_{\text{overall}} = L_{\text{pm}} + 2L_{\text{dm3}} + 3L_{\text{dm2}} + 4L_{\text{dm1}}. \quad (14)$$

5. Experiments

5.1. Experimental Setup

We implemented PDNet in PyTorch [38] and use the stochastic gradient descent (SGD) optimizer for training with momentum set to 0.9, weight decay equal to $5 \times 10^{-4}$, batch size of 18, and using the poly strategy [29] (basic learning rate of 0.001 and power equals 0.9). Training takes around 12 hours for 600 epochs on an 8-core i7-9700K 3.6 GHZ CPU, 64 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU; the same configuration is used to execute all experiments in this paper. We use the RGBD-Mirror dataset for training augmented with random horizontal flipping and multi-scale resizing. Input images are resized to $416 \times 416$ before inference, and the resulting mirror segmentation is resized back to the original size of the input image; we use bilinear interpolation for both resizing operations.

5.2. Comparison to Prior Work

To demonstrate the effectiveness of PDNet, we extensively compare our method against 27 related SOTA approaches (Table 2): 7 semantic segmentation methods, 10 salient object detection methods, 8 RGB-D saliency detection methods, and 2 RGB mirror segmentation methods. All methods are retrained and tested on our RGBD-Mirror dataset. We quantitatively validate each method using 4 different metrics: intersection over union (IoU), weighted F-measure ($F_\beta$) [31], mean absolute error (MAE), and balance error rate (BER) [35]. Note that for IoU and $F_\beta$ higher is better. In contrast, for MAE and BER lower is better. From the results in Table 2, we can see that PDNet outperforms all competing SOTA methods by a large margin on all evaluation metrics.

Figure 4 qualitatively compares PDNet with two prior mirror segmentation methods (i.e., MirrorNet [55] and PMD [25]) as well as the best approach from each of the three other categories (i.e., semantic segmentation method CCNet [18], salient object detection method F3Net [49], and RGB-D saliency detection method BBS-Net [13]). The first three rows show segmentation examples of small mirrors. In the first two examples, only PDNet accurately segments the mirror regions behind the lamp. In the third example, all other methods (except BBS-Net [13]) are confused by the painting in the top-left. Thanks to the depth correlation cues, only PDNet succeeds in a pixel-accurate segmentation of the mirror. PDNet can also correctly handle large mirrors (row 4-6) and multiple mirrors (row 7-9) by virtue of taking both global discontinuity and correlation relations inside and outside the mirror regions into account. The examples in the 10th and 11th rows show challenging cases with similar boundaries and similar appearance, respectively. While the regions are similar in RGB, PDNet benefits from the additional depth cue to correctly disam-
To better understand the benefits of including depth cues, we conduct the following three experiments: (i) retrain PDNet without including the depth branch from scratch and test without depth information (Table 3 F and Figure 5 3rd column); (ii) employ the regular depth-aware PDNet but at testing replace the depth with the grayscale version of its corresponding RGB image (Table 3 G and Figure 5 4th column); and (iii) similar to the previous experiment but replacing the depth map with a pure black image (Table 3 H and Figure 5 5th column). Compared to the original PDNet (Table 3 O and the penultimate column in Figure 5), none the variants achieves the same quality. Noteworthy is that the first variant outperforms (also included in Table 2 as ‘PDNet w/o D’) both MirrorNet [55] (except for BER) and PDNet [25].

### Benefits of Depth Cues

To better understand the benefit of including depth cues, we conduct the following three experiments: (i) retrain PDNet without including the depth branch from scratch and test without depth information (Table 3 F and Figure 5 3rd column); (ii) employ the regular depth-aware PDNet but at testing replace the depth with the grayscale version of its corresponding RGB image (Table 3 G and Figure 5 4th column); and (iii) similar to the previous experiment but replacing the depth map with a pure black image (Table 3 H and Figure 5 5th column). Compared to the original PDNet (Table 3 O and the penultimate column in Figure 5), none the variants achieves the same quality. Noteworthy is that the first variant outperforms (also included in Table 2 as ‘PDNet w/o D’) both MirrorNet [55] (except for BER) and PDNet [25].

### Effectiveness of the Positioning and Delineating Modules

We first define and train a base model ‘B’ which is based on PDNet but without both the PM and DM. We replace the PM with a simple fusing scheme (Eq. 6). Similarly for the DM, we fuse RGB and depth features (Eq. 6) and element-wise add higher-level guidance features upsamped according to Eq. 12. The performance of the base model ‘B’ is shown in Table 3 I. Noteworthy is that the base model can achieve comparable results to MirrorNet [55]; again demonstrating the importance of depth information for the mirror segmentation. Starting from the base model, we gradually re-introduce the DBP, CPB, full PM, and full DM (Table 3 J-M). From this we can conclude that: (i) efficiency into account (Table 4), PDNet’s configuration offers a good balance between effectiveness and efficiency.

### Table 2: Quantitative ablation results indicate that each component in PDNet contributes to the overall performance. ‘RFE’ and ‘DFE’ denote the feature extractor for the RGB and depth map respectively. ‘B’ denotes our base network, ‘DPB’ and ‘CPB’ are the discontinuity and correlation perception block respectively, ‘PM’ and ‘DM’ represent the positioning and delineating module respectively, and ‘DW’ is the dynamic weighting.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Networks</th>
<th>Pseudo Labeling</th>
<th>Dilation</th>
<th>BER</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2MA+ [28]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>SSF+ [59]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>DB2le [41]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>CoNet+ [19]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>JL-DCF+ [15]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>HDFNet+ [36]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>ATSA+ [55]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>BBS-Net+ [13]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>MirrorNet†+ [55]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>PMD†+ [25]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>PDNet*</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
</tbody>
</table>

Table 2: Quantitative ablation results indicate that each component in PDNet contributes to the overall performance. ‘RFE’ and ‘DFE’ denote the feature extractor for the RGB and depth map respectively. ‘B’ denotes our base network, ‘DPB’ and ‘CPB’ are the discontinuity and correlation perception block respectively, ‘PM’ and ‘DM’ represent the positioning and delineating module respectively, and ‘DW’ is the dynamic weighting.

### Table 3: Quantitative ablation results indicate that each component in PDNet contributes to the overall performance. ‘RFE’ and ‘DFE’ denote the feature extractor for the RGB and depth map respectively. ‘B’ denotes our base network, ‘DPB’ and ‘CPB’ are the discontinuity and correlation perception block respectively, ‘PM’ and ‘DM’ represent the positioning and delineating module respectively, and ‘DW’ is the dynamic weighting.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Networks</th>
<th>Pseudo Labeling</th>
<th>Dilation</th>
<th>BER</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2MA+ [28]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>SSF+ [59]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>DB2le [41]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>CoNet+ [19]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>JL-DCF+ [15]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>HDFNet+ [36]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>ATSA+ [55]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>BBS-Net+ [13]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>MirrorNet†+ [55]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>PMD†+ [25]</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
<tr>
<td>PDNet*</td>
<td>PDNet w/o D</td>
<td>0.077</td>
<td>0.077</td>
<td>15.09</td>
</tr>
</tbody>
</table>

Table 2: Quantitative ablation results indicate that each component in PDNet contributes to the overall performance. ‘RFE’ and ‘DFE’ denote the feature extractor for the RGB and depth map respectively. ‘B’ denotes our base network, ‘DPB’ and ‘CPB’ are the discontinuity and correlation perception block respectively, ‘PM’ and ‘DM’ represent the positioning and delineating module respectively, and ‘DW’ is the dynamic weighting.

### 5.3. Ablation Study

We conduct an extensive ablation study to validate the effectiveness of each key component in PDNet. Table 3 and Figure 5 summarize our findings.

#### Impact of Different Feature Extractors

PDNet uses relatively simple backbone feature extractors (i.e., ResNet-50 [16] for RGB and cascading $3 \times 3$ convolutional layers for depth). In the first ablation study, we investigate the performance of more advanced feature extractors. From Table 3 (A-E), we observe that: (i) more advanced backbone structures for RGB does not boost performance (i.e., $B$ is lower than $O$); and (ii) stronger depth feature extractors can further boost the performance (i.e., $D$ and $E$ are higher than $O$), indicating that depth information is essential for achieving high quality results. Furthermore, taking computational efficiency into account (Table 4), PDNet’s configuration offers a good balance between effectiveness and efficiency.

#### Effectiveness of the Positioning and Delineating Modules

We first define and train a base model ‘B’ which is based on PDNet but without both the PM and DM. We replace the PM with a simple fusing scheme (Eq. 6). Similarly for the DM, we fuse RGB and depth features (Eq. 6) and element-wise add higher-level guidance features upsampled according to Eq. 12. The performance of the base model ‘B’ is shown in Table 3 I. Noteworthy is that the base model can achieve comparable results to MirrorNet [55]; again demonstrating the importance of depth information for the mirror segmentation. Starting from the base model, we gradually re-introduce the DBP, CPB, full PM, and full DM (Table 3 J-M). From this we can conclude that: (i)
Figure 4. Visual comparison of PDNet against state-of-the-art segmentation methods retrained on the RGBD-Mirror dataset. PDNet outperforms competing methods on scenes with small mirrors (rows 1-3), large mirrors (rows 4-6), and multiple mirrors (rows 7-9), and challenging scenes with similar boundaries and/or appearance (rows 10-12).

DPB and CPB can boost performance, demonstrating the effectiveness of the two blocks in the PM; (ii) adding DPB improves results more than adding CPB, indicating that discontinuity plays a more important role than correlation for locating a mirror; (iii) adding the full PM achieves better results than both ‘B+DPB’ and ‘B+CPB’, indicating that discontinuity and correlation cues complement each other in locating a mirror; (iv) adding the PM on ‘B+DM’ (i.e., ‘PDNet’) further improves performance, indicating the effectiveness of the PM; (v) introducing the DM further enhances mirror segmentation compared to the base model; and (vi) adding the DM on ‘B+PM’ (i.e., ‘PDNet’) gains a 2.23% and 1.80% performance improvement in IoU and $F_w^β$, respectively. This shows that the DM indeed helps to refine the mirror boundaries. This is further corroborated by the qualitative comparison between ‘B+PM’ and ‘PDNet’ in Figure 5 (top).

Effectiveness of Dynamic Weighting. A key contribution of our approach is the dynamic weighting to fuse RGB and depth correlations in the PM. To demonstrate its im-
domain is $\beta = 0.53$. Hence, we think the reason is that the depth cues from the arched doorway confuses our method and mislead it to give more weight to depth cue. Based on the estimated results from MirrorNet [55] and PMD [25], we can see that it is also a hard case for RGB-based approaches. Another failure case is when the mirror covers almost the entire image, as demonstrated in Figure 6 (bottom). In this case, the discontinuities between inside and outside the mirror are hard to quantify. More importantly, without a global view of inside versus outside the mirror, correlations between both become less meaningful. However, such a case would also be difficult for humans.

### 6. Conclusion

We present PDNet, a convolutional neural network based depth-aware mirror segmentation method. Our solution is the first to leverage discontinuities and correlations in both RGB and depth to segment the mirror. PDNet builds on two key components: a positioning module that locates the mirror based on global and local discontinuities and correlations in RGB and depth between the regions inside and outside the mirror, and a delineating module that leverages local contextual contrast in RGB and depth to refine the mirror boundaries. We also introduce an RGB-D mirror segmentation dataset to train PDNet and stimulate further research in this area. We show that our approach outperforms state-of-the-art mirror segmentation methods and demonstrate that depth is a powerful cue for mirror segmentation.

### Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 91748104, Grant 61972067, Grant U1908214, in part by the Innovation Technology Funding of Dalian (Project No. 2020JJ26GX036). Pieter Peers was partially supported by NSF grant IIS-1909028.
References


[18] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Cnet: Criss-cross attention for semantic segmentation. In \textit{ICCV}, 2019. 1, 2, 5, 6, 8


[34] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfei He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don’t hit me! glass detection in real-world scenes. In CVPR, 2020. 3


[37] Youwei Pang, Xiaqoi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In CVPR, 2020. 1, 2, 6, 8


[40] Yongri Piao, Wei Ji, Jingjing Li, Mao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In ICCV, 2019. 2


[43] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In CVPR, 2019. 2, 6, 8


[46] Tomás Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In ECCV, 2016. 3


[55] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In ICCV, 2019. 1, 2, 3, 5, 6, 8


[64] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In CVPR, 2019. 2

[65] Hengshuang Zhao, Xiaojian Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In ECCV, 2018. 6

[66] Hengshuang Zhao, Jianping Shi, Xiaojian Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017. 1, 2, 6

