Depth-Aware Mirror Segmentation

Haiyang Mei[1] Bo Dong[2,*] Wen Dong[1] Pieter Peers[3] Xin Yang[1,*] Qiang Zhang[1] Xiaopeng Wei[1,*]
[1] Dalian University of Technology [2] SRI International [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, *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 *et al*. [55] introduced MirrorNet, a convolutional neural network, that leverages contextual contrasted features to detect content discontinuities inside and outside the mirror. Lin *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).

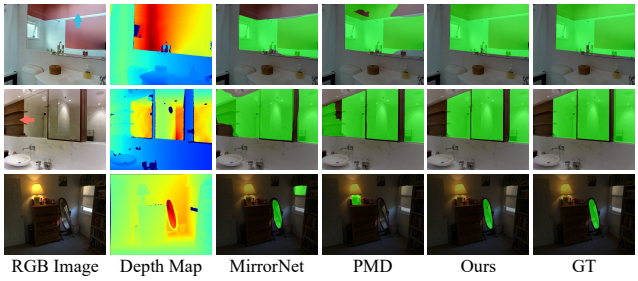


Figure 1. Existing mirror segmentation methods such as MirrorNet [55] and PMD [25] often fail when there is a large variation in contextual contrast/correlation inside the mirror ($1st$ row, blue arrow), large variation outside the mirror in a mirror-like region ($2nd$ row, red arrow), or when the differences are too subtle ($3rd$ row). In contrast, our depth-aware solution is able to accurately segment the mirrors.

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 *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 (*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

* 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 (PDNet). As the name suggests, PDNet 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 recalibrate, 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.

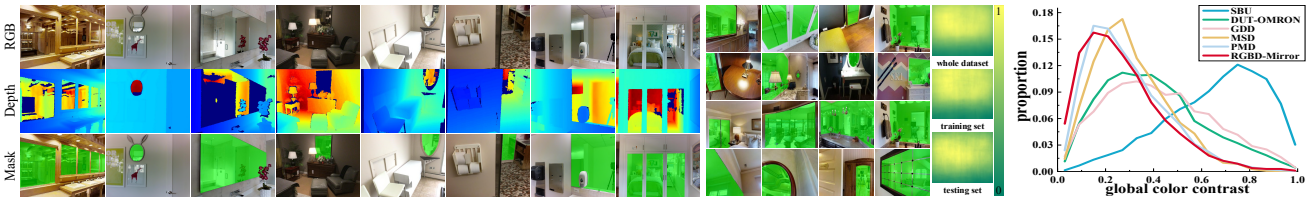
Salient Object Detection (SOD) identifies the most visually distinctive objects/regions in an image of a scene. Current state-of-the-art solutions employ convolutional neural networks (CNNs) to exploit different RGB cues and strategies such as multi-level feature aggregation [22, 17, 60, 67, 37, 33], 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 (*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.

Mirror 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



(a) Mirror image, Depth map, and Mirror-mask triplets (b) Mirror location distribution (c) Color contrast distribution

Figure 2. RGBD-Mirror dataset examples and statistics.

Dataset	Images	Train	Test	Scenes
Matterport3D [2]	1,789	1,153	636	78
SUNRGBD [45]	576	291	285	17
ScanNet [8]	593	484	109	102
2D3DS [1]	91	72	19	5
Total	3,049	2,000	1,049	202

Table 1. Composition of the RGBD-Mirror dataset.

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 χ^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 (PDNet) to exploit discontinuity and correlation in both RGB and depth to efficiently segment mirrors. PDNet (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 3×3 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.* $D = D_l \oplus D_g$). Given a feature F, the local discontinuity feature D_l is defined as the difference between a local region and its surroundings:

$$D_l = \mathcal{R}(\mathcal{N}(f_l(F, \Theta_l) - f_s(F, \Theta_s))), \quad (1)$$

where f_l, with corresponding parameters Θ_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 Θ_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

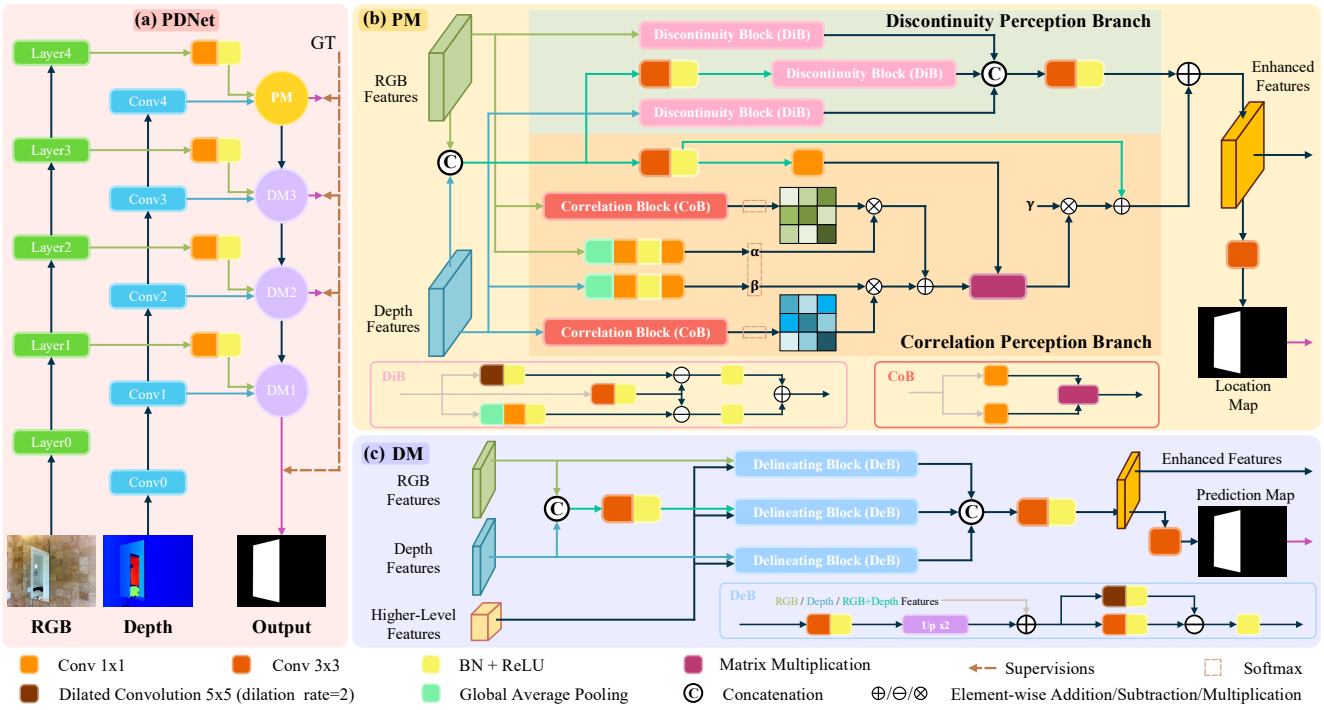


Figure 3. (a) Overview of our positioning and delineating network (PDNet) and its two main building blocks: (b) a positioning module (PM) and (c) a delineating module (DM).

discontinuity feature captures the differences between local regions and their surroundings, under certain viewpoints, the reflected mirror image has little overlap with its surroundings. This case is represented by the global discontinuity feature:

$$D_g = \mathcal{R}(\mathcal{N}(f_l(F, \Theta_l) - f_g(\mathcal{G}(F), \Theta_g))), \quad (2)$$

where $\mathcal{G}$ is a global average pooling operation, and f_g (with corresponding parameters Θ_g) is a 1×1 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:

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

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:

$$Y = g(F)\kappa(F), \quad (4)$$

where $\kappa(F) = \mathit{softmax}(\theta(F)^T\phi(F))$. g, θ, and ϕ 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:

$$Y = g(F^{rd})(\alpha\kappa(F^r) + \beta\kappa(F^d)), \quad (5)$$

$$F^{rd} = \mathcal{R}(\mathcal{N}(\psi_{3\times3}(F^r \copyright F^d))), \quad (6)$$

$$g(F) = \psi_{1\times1}(F), \quad (7)$$

$$\theta(F) = \psi_{1\times1}(F),\ \phi(F) = \psi_{1\times1}(F), \quad (8)$$

$$\alpha = \frac{e^{\mu(F^r)}}{e^{\mu(F^r)} + e^{\mu(F^d)}},\ \beta = 1 - \alpha, \quad (9)$$

$$\mu(F) = \psi_{1\times1}(\mathcal{R}(\mathcal{N}(\psi_{1\times1}(\mathcal{G}(F))))), \quad (10)$$

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 γ: $C^{\text{CPB}} = \gamma Y \oplus F^{rd}$.

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:

$$T = \mathcal{R}(\mathcal{N}(f_l(F \oplus F^{hg}, \Theta_l) - f_s(F \oplus F^{hg}, \Theta_s))), \quad (11)$$

$$F^{hg} = U_2(\mathcal{R}(\mathcal{N}(\psi_{3\times3}(F^h)))), \quad (12)$$

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: $S = \psi_{3\times3}(X)$, where X is the output feature from either the PM or DM:

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

where ℓ_{bce} is a binary cross-entropy (BCE) loss [9], ℓ_{iou} is a map-level IoU loss [32], ℓ_{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:

$$\mathcal{L}_{overall} = \mathcal{L}_{pm} + 2\mathcal{L}_{dm3} + 3\mathcal{L}_{dm2} + 4\mathcal{L}_{dm1}. \quad (14)$$

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 5×10^{-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 416×416 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_β^w) [31], mean absolute error (MAE), and balance error rate (BER) [35]. Note that for IoU and F_β^w 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 10*th* and 11*th* 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-

Methods	Pub. Year	IoU↑	F_β^w↑	MAE↓	BER↓
Statistics	-	19.25	0.190	0.538	37.85
ICNet$^\circ$ [63]	ECCV'18	37.43	0.464	0.122	28.59
PSPNet$^\circ$ [64]	CVPR'17	61.83	0.686	0.056	17.42
DenseASPP$^\circ$ [54]	CVPR'18	63.50	0.700	**0.050**	16.27
BiSeNet$^\circ$ [56]	ECCV'18	62.36	0.694	0.062	15.90
PSANet$^\circ$ [65]	ECCV'18	56.98	0.643	0.057	20.72
DANet$^\circ$ [14]	CVPR'19	63.81	0.708	0.057	16.48
CCNet$^\circ$ [18]	ICCV'19	65.09	0.715	0.055	14.92
DSS$^\triangle$ [17]	TPAMI'19	57.58	0.614	0.087	18.60
PiCANet$^\triangle$ [27]	CVPR'18	64.80	0.682	0.064	14.99
RAS$^\triangle$ [4]	ECCV'18	57.96	0.650	0.080	18.23
R^{3}Net$^{\triangle\dagger}$ [10]	IJCAI'18	53.09	0.584	0.073	21.96
CPD$^\triangle$ [51]	CVPR'19	60.41	0.639	0.080	17.61
PoolNet$^\triangle$ [26]	CVPR'19	62.99	0.677	0.074	15.13
BASNet$^\triangle$ [43]	CVPR'19	64.01	0.689	0.072	15.77
EGNet$^\triangle$ [66]	ICCV'19	60.11	0.657	0.077	16.38
F3Net$^\triangle$ [49]	AAAI'20	65.15	0.707	0.069	14.25
MINet-R$^\triangle$ [37]	CVPR'20	60.25	0.669	0.077	16.63
S2MA$^\triangledown$ [28]	CVPR'20	63.66	0.677	0.071	15.09
SSF$^\triangledown$ [59]	CVPR'20	52.83	0.599	0.097	19.54
A2dele$^\triangledown$ [41]	CVPR'20	53.61	0.614	0.087	19.64
CoNet$^\triangledown$ [19]	ECCV'20	50.96	0.576	0.120	17.23
JL-DCF$^\triangledown$ [15]	CVPR'20	68.21	0.727	0.065	13.52
HDFNet$^\triangledown$ [36]	ECCV'20	47.48	0.549	0.095	24.70
ATSA$^\triangledown$ [58]	ECCV'20	60.03	0.664	0.090	14.79
BBS-Net$^\triangledown$ [13]	ECCV'20	71.22	0.736	0.059	11.77
MirrorNet*† [55]	ICCV'19	68.37	0.723	0.062	**8.66**
PMD*† [25]	CVPR'20	**72.27**	**0.775**	0.054	10.71
PDNet *w/o* D*	**Ours**	**73.57**	**0.783**	**0.053**	**9.26**
PDNet*	**Ours**	**77.77**	**0.825**	**0.042**	**7.77**

Table 2. Quantitative performance of state-of-the-art semantic segmentation methods (marked by the $\circ$ symbol), salient object detection methods ($\triangle$), RGB-D saliency detection methods ($\triangledown$), and RGB mirror segmentation methods ($\star$) retrained on the RGBD-Mirror training set and compared over the RGBD-Mirror testing set. Methods that require an additional CRF [20] post-processing step are marked with the $\dagger$ symbol. We also include a threshold method based on the location *statistics* of the mirror masks in the training set. The first, second, and third best results are highlighted in **red**, **green**, and **blue**, respectively. Our method achieves the best performance over all four evaluation metrics.

biguate mirror from non-mirror regions.

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 3×3 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

	Networks	RGBD-Mirror Testing Set			
		IoU↑	F_β^w↑	MAE↓	BER↓
	MirrorNet*† [55]	68.37	0.723	0.062	8.66
	PMD*† [25]	72.27	0.775	0.054	10.71
A	RFE *w/* VGG-16	75.31	0.805	0.052	9.01
B	RFE *w/* ResNeXt-101	77.25	0.817	0.045	7.80
C	DFE *w/* 64-64-64-64-64	76.00	0.809	0.049	7.95
D	DFE *w/* VGG-16	78.87	0.836	0.044	7.55
E	DFE *w/* ResNeXt-101	**79.31**	**0.837**	**0.041**	**7.31**
F	PDNet *w/o* D	73.57	0.783	0.053	9.26
G	PDNet *w/* RGB+Gray	67.14	0.719	0.064	11.65
H	PDNet *w/* RGB+Black	65.98	0.724	0.059	15.60
I	B	70.08	0.764	0.058	11.68
J	B + DPB	73.81	0.794	0.052	9.91
K	B + CPB	72.48	0.774	0.058	8.74
L	B + PM	75.54	0.807	0.047	8.61
M	B + DM	73.60	0.789	0.052	10.47
N	PDNet *w/o* DW	76.74	0.816	0.047	8.48
O	PDNet	77.77	0.825	0.042	7.77

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.

efficiency into account (Table 4), PDNet's configuration offers a good balance between effectiveness and efficiency.

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 3*rd* column); (ii) employ the regular depth-aware PDNet but at testing replace the depth with the grayscale version of its corresponding RGB images (Table 3 G and Figure 5 4*th* column); and (iii) similar to the previous experiment but replacing the depth map with a pure black image (Table 3 H and Figure 5 5*th* 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 PMD [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 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 DPB, 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_β^w, 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-

Figure 5. Visual ablation comparison of different PDNet variants.

Networks	Input Size	FLOPs (G)	Params (M)	Time (ms)
CCNet [18]	480 × 480	248.579	66.549	11.2
BASNet [43]	256 × 256	127.444	87.060	12.2
F3Net [49]	352 × 352	**16.429**	**25.537**	**10.7**
MINet-R [37]	320 × 320	87.032	162.378	12.7
S2MA [28]	256 × 256	141.064	86.645	13.4
HDFNet [36]	320 × 320	108.680	54.773	27.8
BBS-Net [13]	352 × 352	31.140	49.769	29.2
MirrorNet† [55]	384 × 384	77.656	121.767	32.1(+607.1)
PMD† [25]	384 × 384	101.459	147.661	62.7(+607.1)
DFE w/ VGG-16	416 × 416	188.710	216.403	12.9
DFE w/ ResNeXt-101	416 × 416	104.964	201.769	26.8
PDNet (Ours)	416 × 416	41.059	80.541	12.0

Table 4. Comparison of the computational efficiency of different methods. For each method, we list FLOPs, number of parameters, and inference time. For MirrorNet [55] and PDM [25], we report the CRF [20] post-processing time in Cyan.

portance, we remove the dynamic weighting scheme and retrain the modified model (Table 3 N). Compared to the original PDNet (*i.e.* row O), the modification degrades performance. The two examples shown in Figure 5 have $\beta = 0.49$ and $\beta = 0.66$ assigned by the dynamic weighting respectively. Hence, in the first case, more weight is given to the RGB correlations, whereas the second example puts more weight on the depth correlations. Compared to the equally assigned weight (*i.e.* 'PDNet *w/o* DW'), dynamical weighting improves performance dramatically.

5.4. Computational Cost

Our PDNet is an end-to-end process that does not need any post-processing, unlike MirrorNet [55] and PMD [25] which require post-processing by a computationally costly fully connected conditional random field (CRF) [21]. Table 4 compares the superior computational efficiency of PDNet against MirrorNet [55] and PMD [25] in terms of FLOPs (in G), model parameters (in M), and inference time (in ms). Furthermore, the computational efficiency of PDNet also performs similarly or better to other semantic segmentation methods.

5.5. Limitations

PDNet is not without limitations. Doorways are sometimes incorrectly classified as mirror regions, as illustrated in Figure 6 (top). In this case, the weight assigned to depth

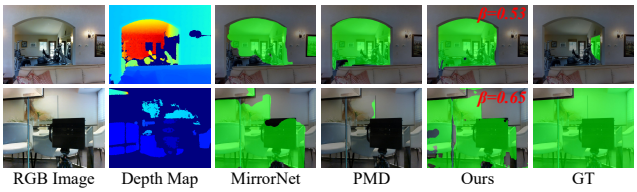


Figure 6. Examples of failure cases such as a doorway (top) and a mirror that covers almost the entire image (bottom).

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

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv:1702.01105*, 2017.

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017.

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018.

[5] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020.

[6] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *CVPR*, 2017.

[7] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *ICIMCS*, 2014.

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias NieBner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

[9] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. In *Annals of Operations Research*, 2005.

[10] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, 2018.

[11] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.

[12] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE TNNLS*, 2020.

[13] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, 2020.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[15] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 2019.

[18] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[19] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, 2020.

[20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[22] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.

[23] Yin Li, Xiaodi Hou, Christof Koch, James Rehg, and Alan Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[24] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *ECCV*, 2016.

[25] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *CVPR*, 2020.

[26] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

[27] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[28] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *CVPR*, 2020.

[29] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[31] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[32] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *ICCV*, 2017.

[33] Haiyang Mei, Yuanyuan Liu, Ziqi Wei, Li Zhu, Yuxin Wang, Dongsheng Zhou, Qiang Zhang, and Xin Yang. Exploring dense context for salient object detection. *IEEE TCSVT*, 2021.

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

[35] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017.

[36] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, 2020.

[37] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, 2020.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

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

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

[41] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, 2020.

[42] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *ICCV*, 2017.

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

[44] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE TIP*, 2017.

[45] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.

[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.

[47] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, 2019.

[48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[49] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020.

[50] Thomas Whelan, Michael Goesele, Steven Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, and Richard Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM TOG*, 2018.

[51] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[52] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.

[53] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[54] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

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

[56] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

[57] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.

[58] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *ECCV*, 2020.

[59] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *CVPR*, 2020.

[60] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[61] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018.

[62] 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.

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

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[65] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.

[66] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, 2019.

[67] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.