Depth Inconsistency-based Spatial-channel Attention Gate for Mirror Segmentation

Ritsuki Kurohiji
kurohiji.ritsuki@g.wakayama-u.jp
Hirotaka Hachiya†
hhachiya@wakayama-u.ac.jp
Wakayama University
Wakayama, Japan

Abstract

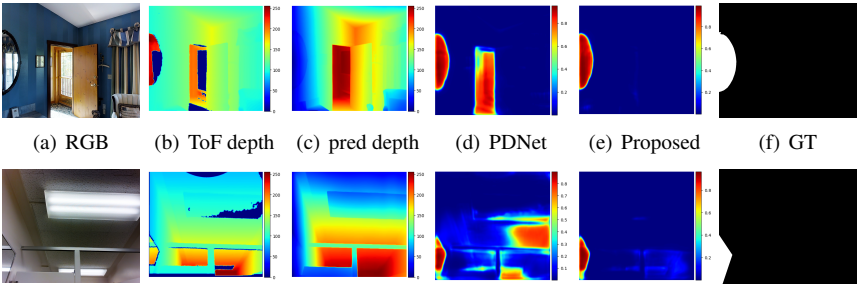
We propose a novel depth-aware mirror and specular object segmentation framework that leverages the contextual inconsistency between two types of depth information: time-of-flight (ToF) measurements and depth predicted by a deep neural network. ToF depth values are often significantly overestimated near mirrors due to indirect light paths, while predicted depth maps tend to be overly smoothed in those regions. We observe that this contextual discrepancy serves as a strong cue for identifying mirror regions. However, direct comparison between these depths in either the image or feature space results in high false positive rates, primarily due to their inherently different value ranges and responses to scene content. To address this issue, we introduce a depth inconsistency-based spatial-channel Attention Gate (discAG), which adaptively highlights informative regions in ToF feature maps based on spatial- and channel-wise extracted depth context inconsistencies. By integrating discAG into a segmentation architecture, our method achieves superior performance compared to state-of-the-art approaches. Experimental results on the RGB-D Mirror Segmentation dataset and our newly constructed Specular Object dataset demonstrate the effectiveness of our approach, particularly in challenging scenarios involving small or frameless mirrors.

1 Introduction

Mirrors and specular objects play essential roles in daily life, retail, and industry. Examples include hand and stand mirrors used for shaving and makeup, wall mirrors for expanding perceived room space, curved mirrors for enhancing safety in traffic and retail settings, and mirror-polished molds for producing smooth-surfaced plastic parts in automated manufacturing. However, these objects introduce significant challenges for machine vision systems, as reflections create ambiguity between real and mirrored scenes, often leading to perception errors. Therefore, developing accurate segmentation methods is essential to enhance scene understanding, improve robotic perception, and ensure reliable object recognition.

Deep model-based mirror segmentation methods have been proposed to extract mirror-specific multi-level features [2, 3, 4, 7, 13]. These include contextual contrast (CC) [13], contextual similarity [7], region-wise similarity [2], symmetricity [4], and combinations of static specular flow and highlights [3]. To further improve segmentation accuracy, recent

†Denotes corresponding authors.



(a) RGB (b) ToF depth (c) pred depth (d) PDNet (e) Proposed (f) GT

(g) RGB (h) ToF depth (i) pred depth (j) PDNet (k) Proposed (l) GT

Figure 1: PDNet [9], which leverages the co-occurrence of RGB and ToF depth discontinuities, fails when similar co-occurrences appear in non-mirror scenes—for example, the open door (first row) and the transom above the restroom door (second row). In contrast, our proposed method, which exploits the contextual inconsistency between ToF and predicted depths, can suppress false positive detections while accurately detecting mirrors.

studies have incorporated depth information measured by time-of-flight (ToF) sensors [9]. For example, PDNet [9] detects mirrors based on the co-occurrence of discontinuities in both RGB and ToF depth feature maps.

As illustrated in Fig. 1, PDNet performs well when mirrors exhibit strong CC, such as frames in RGB images and extreme or missing values in ToF depth, i.e., 0 indicates out-of-range measurements. However, similar co-occurrences can also arise in non-mirror scenes, such as open doors or transoms, where frames and depth discontinuities also occur. In such cases, PDNet may mistakenly classify these regions as mirrors, resulting in false positives as shown in Fig. 1.

To address this limitation, we examine the relationship between depth measured by ToF sensors and that predicted by deep models. As shown in the predicted depth of Fig.1, data-driven models like Depth Anything V2 [12] tend to estimate mirror surfaces as if they were ordinary physical surfaces because the unique patterns within mirrors are underrepresented in standard datasets. In contrast, for typical non-mirror scenes (e.g., open doors), predicted depths are relatively accurate. Therefore, the depth discrepancy between ToF and predicted depths tends to be larger in mirror regions than in non-mirror regions.

Based on this observation, we propose a new mirror and specular object segmentation framework that leverages the inconsistency between ToF and predicted depths. A key challenge is that directly comparing these depths (in image or feature space) often results in false positives due to differences in depth scale and response characteristics. To overcome this, we introduce a novel attention module—depth inconsistency-based spatial-channel Attention Gate (discAG)—which adaptively highlights informative regions in ToF feature maps using spatial- and channel-wise extracted inconsistency contexts.

We evaluate our method on the RGB-D Mirror Segmentation dataset [9] and a newly constructed Specular Object dataset, consisting of multiple real-world specular objects. The results show that our method effectively suppresses false detections made by PDNet and achieves superior segmentation performance.

2 Related work

This section formulates the problem and reviews existing related works.

2.1 Problem Formulation

Let $I_{\text{rgb}} \in \mathbb{R}^{3\times H\times W}$, $I_{\text{dep}} \in \mathbb{R}^{1\times H\times W}$, and $Y \in \{0,1\}^{H\times W}$ be an RGB image, the depth map measured by a ToF sensor, and the mirror map where $Y_{hw} = 1$ if the specular surface exists and $Y_{hw} = 0$ otherwise at pixel (h,w), respectively. Here, H and W are the height and width of the image.

Let $t \in \mathcal{T}$ be the label of specular objects, e.g., mirror and metal plate, and $\mathcal{D}^{\text{tr}}$ and $\mathcal{D}^{\text{te}}$ denote training and test data consisting of sets of corresponding Y, I_{rgb}, I_{dep}, and t as follows:

$$\mathcal{D}^{\text{tr}} = \{(Y^n, I^n_{\text{rgb}}, I^n_{\text{dep}}, t^n)\}_{n=1}^{N^{\text{tr}}}, \quad \mathcal{D}^{\text{te}} = \{(Y^n, I^n_{\text{rgb}}, I^n_{\text{dep}}, t^n)\}_{n=1}^{N^{\text{te}}}, \tag{1}$$

where N^{tr} and N^{te} are the numbers of training and test data.

Let $\mathcal{X}_{\text{rgb}}$ and $\mathcal{X}_{\text{dep}}$ be multi-level feature pyramids extracted from the RGB I_{rgb} and depth map I_{dep}, respectively, using L-layer CNNs as follows:

$$\mathcal{X}_{\text{rgb}} = \{X^l_{\text{rgb}}\}_{l=1}^{L} = \text{CNN}(I_{\text{rgb}}), \quad \mathcal{X}_{\text{dep}} = \{X^l_{\text{dep}}\}_{l=1}^{L} = \text{CNN}_{\boldsymbol{\theta}}(I_{\text{dep}}), \tag{2}$$

where X^l_{rgb} and $X^l_{\text{dep}} \in \mathbb{R}^{C^l\times H^l\times W^l}$ represent the l-th layer feature map, with the number C^l of channels, height H^l, and width W^l.

Then, let us consider a segmentation model $f_\Theta(\cdot)$ to generate a mirror map Y given an RGB I_{rgb} and depth map I_{dep} as follows:

$$\widehat{Y} = f_\Theta(I_{\text{rgb}}, I_{\text{dep}}), \tag{3}$$

where Θ is a set of parameters of the model to be tuned using training data $\mathcal{D}^{\text{tr}}$.

2.2 RGB-based mirror segmentation

Deep model-based methods based on RGB I_{rgb} have been proposed by combining a multi-level feature pyramid with the relational properties between a mirror and its surroundings [2, 3, 4, 7, 13]. There tends to exist a boundary between a mirror image and its surroundings due to reflections and mirror frames. MirrorNet [13] proposed a contextual contrast (CC) feature using the difference between feature maps obtained using small and large receptive fields. In addition, there tends to exist a common content between a mirror image and its surroundings. PMDNet [7] proposed contextual similarity (CS), the similarity pattern in 8 cardinal directions.

The concept of CS has been extended in recent works, by incorporating semantic segmentation for region-wise similarity as in SANet [2] and a symmetry-aware attention mechanism for capturing chirality as in SATNet [4]. In addition, SSSNet [3] integrates low-level features such as specular highlights (SH) [10] and static specular flow (SSF) [1] in multi-level features with CS and CC modules [3] to detect specular objects with complex shapes and ambiguous boundaries.

However, these RGB-based methods inherently struggle to detect mirrors in low-light environments and in frameless mirrors, where the boundary and similarity between the mirror and its surroundings become indistinct.

2.3 RGB-D based mirror segmentation

To overcome the limitations of RGB-based methods, PDNet [9] has proposed an RGB-D-based mirror segmentation method using not only RGB I_{rgb} but also depth I_{dep} captured by a ToF sensor. The ToF depth for specular surfaces is often incorrectly measured as being much farther than the actual physical surface, since infrared light reflects off the surface and returns to the sensor via an indirect path. This results in depth discontinuities between the mirror and its surroundings.

PDNet consists of two main components: the Positioning Module (PM) and the Delineating Module (DM). In the PM, local and global CC modules—which respectively employ $\text{Conv}_{3\times3}$ versus $\text{DilatedConv}_{5\times5}$ convolutions for local context, and $\text{Conv}_{3\times3}$ versus global average pooling (GAP) for global context—are applied to the feature maps X^l_{rgb}, X^l_{dep}, and their concatenation $\text{Concat}(X^l_{\text{rgb}}; X^l_{\text{dep}})$. Following this, spatial self-attention is independently applied to X^l_{rgb} and X^l_{dep}, which are then fused through adaptive weighting. This process effectively suppresses noise in the depth features and dynamically emphasizes mirror regions by capturing the co-occurrence of RGB and depth discontinuities.

The DM further refines the mirror boundaries by applying localized CC features to the outputs from the PM as well as from other high-level layers, enhancing spatial precision around mirror edges.

However, as discussed in Sec. 1 and shown in Fig. 1, PDNet may incorrectly detect non-mirror regions as mirrors. This is because the similar co-occurrence of RGB and depth discontinuities as mirror boundaries can also occur in ordinary scenes.

3 Proposed method

To overcome the limitation of the existing RGB-D based mirror segmentation method, PDNet, we propose leveraging the contextual inconsistency between the ToF depth I_{dep} and the predicted depth $\widehat{I_{\text{dep}}}$. As discussed in Sec. 1, data-driven deep models, i.e., Depth Anything V2 [12], predict depth as if mirror surfaces were ordinary physical surfaces, and thus significant discrepancies are more likely to arise between I_{dep} and $\widehat{I_{\text{dep}}}$ in mirror regions, compared to non-mirror regions.

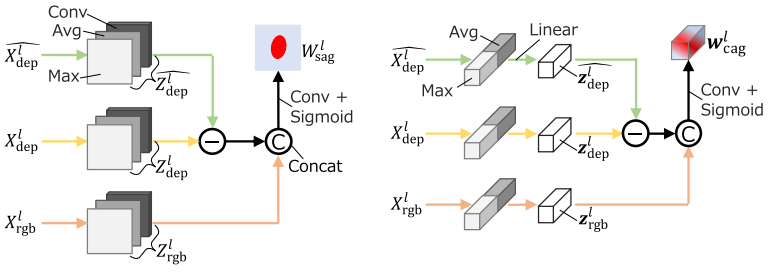
Let $\widehat{I_{\text{dep}}} \in \mathbb{R}^{1\times H\times W}$ and $\widehat{\mathcal{X}_{\text{dep}}}$ be a depth map predicted by a deep model $g(\cdot)$, e.g., Depth Anything V2, given the RGB I_{rgb} and multi-level feature pyramids extracted from the predicted depth map $\widehat{I_{\text{dep}}}$ using a L-layer CNN, respectively, as follows:

$$\widehat{I_{\text{dep}}} = g(I_{\text{rgb}}) \in \mathbb{R}^{1\times H\times W}, \quad \widehat{\mathcal{X}_{\text{dep}}} = \{\widehat{X^l_{\text{dep}}}\}_{l=1}^{L} = \text{CNN}_{\boldsymbol{\phi}}(\widehat{I_{\text{dep}}}). \tag{4}$$

3.1 Depth inconsistency-based spatial and channel attention gate

Ideally, depth inconsistency can be obtained by computing the difference between the ToF and the predicted depths in either image space, i.e., I_{dep} and $\widehat{I_{\text{dep}}}$ or feature space, i.e., X^l_{dep} and $\widehat{X^l_{\text{dep}}}$. However, this may lead to false positives for several reasons.

- Predicted depth is often overly smooth especially around object boundaries.
- ToF depth may contain missing values due to out-of-range or infrared-absorbing materials.



(a) disAG (b) dicAG

Figure 2: Diagram of contextual depth inconsistency-based spatial/channel attention gates.

- Since the predicted depth represents relative values, ToF depth is min-max normalized to match its scale, i.e., $[0,255]$; however, the resulting depth values may still be inconsistent across modalities.

As shown in the ToF depth images of Fig. 1, such issues can lead to spurious depth contradictions. For example, missing ToF values around a fluorescent light (second row) or mismatches in near-wall regions.

Inspired by scAG (spatial-channel attention gate) [8], to address this issue, we propose a depth inconsistency-based spatial-channel attention gate (discAG), which adaptively highlights important regions in ToF feature maps X^l_{dep} based on the inconsistency between two depth contexts extracted along with each of the spatial and channel axes. More specifically, we statistically and adaptively aggregate each of the feature maps X^l_{dep} and $\widehat{X_{\text{dep}}}$, along with the spatial or channel using max, mean, and convolution operations to extract corresponding and important contexts as follows:

$$\begin{aligned} Z^l &= \Big[\text{Max}(X^l, d=0); \text{Avg}(X^l, d=0); \text{Conv}(X^l)\Big] \in \mathbb{R}^{(2+C^{\text{Conv}})\times H^l \times W^l}, \\ \boldsymbol{z}^l &= W\Big[\text{Max}(X^l, d=[1,2]); \text{Avg}(X^l, d=[1,2])\Big] \in \mathbb{R}^{C^l \times 1} \end{aligned} \tag{5}$$

where ';' indicates the row breaking, C^{Conv} is the number of output channels of $\text{Conv}(\cdot)$, Z^l and $\boldsymbol{z}^l$ are spatial and channel contexts, respectively. $\text{Max}(\cdot,d)$ and $\text{Avg}(\cdot,d)$ are maximum and mean operations along with the axis d. $W \in \mathbb{R}^{C^l \times (C^l \times 2)}$ are weights for linear projection.

Then, we take the difference between two depth contexts and fuse it with RGB features to obtain spatial and channel attention maps $W^l_{\text{sag}} \in \mathbb{R}^{1\times H^l \times W^l}$ and $\boldsymbol{w}^l_{\text{cag}} \in \mathbb{R}^{C^l \times 1 \times 1}$ as depicted in Fig. 2 and as follows:

$$W^l_{\text{sag}} = \text{Conv}\left(\left[Z^l_{\text{dep}} - \widehat{Z^l_{\text{dep}}}; Z^l_{\text{rgb}}\right]\right), \quad \boldsymbol{w}^l_{\text{cag}} = \text{Conv}\left(\left[\boldsymbol{z}^l_{\text{dep}} - \widehat{\boldsymbol{z}^l_{\text{dep}}}; \boldsymbol{z}^l_{\text{rgb}}\right]\right). \tag{6}$$

Finally, we apply these attention maps to the depth map X^l_{dep} to emphasize important spatial and channel contexts as depicted in Fig. 3 and as follows:

$$\widehat{X^l_{\text{dep}}} = X^l_{\text{dep}} \otimes \text{Sigmoid}(W^l_{\text{sag}}) \otimes \text{Sigmoid}(\boldsymbol{w}^l_{\text{cag}}), \tag{7}$$

where $\text{Sigmoid}(\cdot)$ is a sigmoid function.

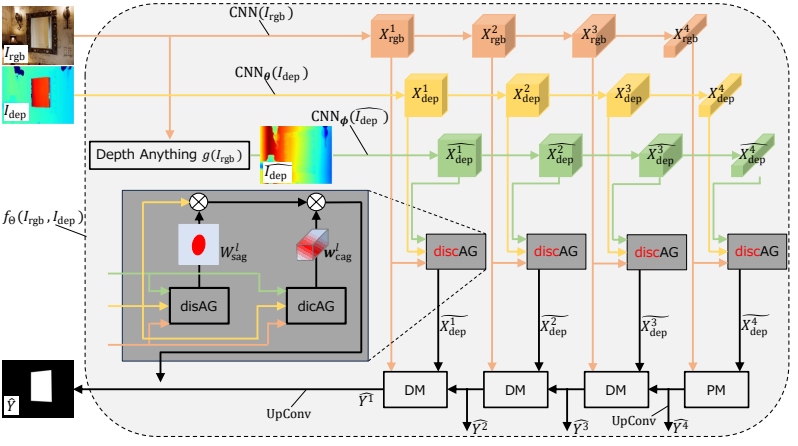


Figure 3: Network structure of the proposed method, consisting of a pre-trained CNN(I_{rgb}), newly trained CNN$_{\boldsymbol{\theta}}$(I_{dep}) and CNN$_{\boldsymbol{\phi}}$($\widehat{I_{\mathrm{dep}}}$) (in Eqs. 2 and 4), Depth Anything V2 [17], proposed discAG modules (Sec. 3.1), and PM and DM in PDNet (in Sec. 2.3).

3.2 Network architecture

The entire network structure of the proposed model $f_{\Theta}(I_{\mathrm{rgb}}, I_{\mathrm{dep}})$ is depicted in Fig. 3 where we combine our proposed discAG modules with PDNet's PM and DM (see Sec 2.3) to extract discontinuities in RGB and ToF depth. More specifically, depth map $\widehat{I_{\mathrm{dep}}}$ is predicted using a pre-trained model $g(I_{\mathrm{rgb}})$, i.e., Depth Anything V2 [17] (in Eq. 4). Multi-level appearance feature pyramids $\mathcal{X}_{\mathrm{rgb}}$, $\mathcal{X}_{\mathrm{dep}}$, and $\widehat{\mathcal{X}_{\mathrm{dep}}}$ are extracted from RGB I_{rgb}, ToF depth I_{dep}, and predicted depth $\widehat{I_{\mathrm{dep}}}$ through pre-trained CNN$(\cdot)$, e.g., ResNeXt-50 [11], and newly trained CNNs CNN$_{\boldsymbol{\theta}}(\cdot)$ and CNN$_{\boldsymbol{\phi}}(\cdot)$ (in Eqs. 2 and 4).

Given feature maps $\{X^l, X^l_{\mathrm{dep}}, \widehat{X^l_{\mathrm{dep}}}\}$ at each layer, spatial context W^l_{sag} and channel context $\boldsymbol{w}^l_{\mathrm{cag}}$ are extracted through disAG and dicAG modules (in Eq. 6 and Fig. 2). Then, each ToF depth feature map X^l_{dep} is weighted by contexts W^l_{sag} and $\boldsymbol{w}^l_{\mathrm{cag}}$ in discAG module (in Eq. 7).

Through the PM and DM modules [9], a predicted mirror map $\widehat{Y}^l \in \mathbb{R}^{1\times H^l \times W^l}$ is produced at each layer. The final output is obtained by upsampling the first-layer prediction: $\widehat{Y} = \mathrm{Upconv}(\widehat{Y^1})$, where the resolution is expanded to $1 \times H \times W$, matching the input image size I_{rgb} and I_{dep}.

4 Experimental evaluation

In this section, we show the effectiveness of the proposed methods through experiments on mirror and specular object segmentation tasks.

4.1 Data and metrics

We used two datasets as follows:

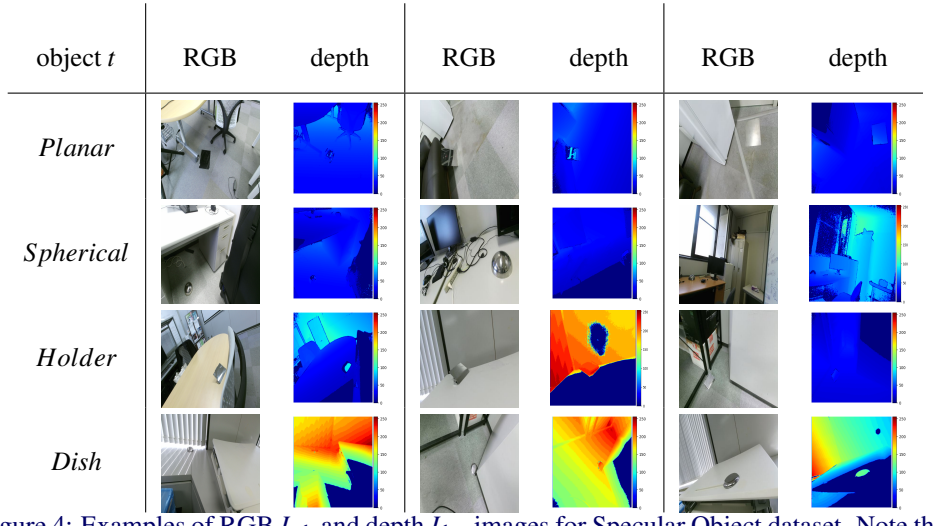


Figure 4: Examples of RGB I_{rgb} and depth I_{dep} images for Specular Object dataset. Note that the depth values are min-max normalized per image and multiplied by 255.

- **RGB-D Mirror Segmentation dataset** [9]

- **Specular Object dataset**: A newly created dataset collected in office and home environments using Kinect V2 [6], consisting of four types of specular objects, as illustrated in Fig. 4, and defined as $\mathcal{T} \equiv \{Planar, Spherical, Holder, Dish\}$. For each object type, we created 40 data samples: $\{(I^n_{\text{rgb}}, I^n_{\text{dep}}, \widehat{I^n_{\text{dep}}}, t^n)\}_{n=1}^{40}$, resulting in a total of 160 samples. We conducted 4-fold cross-validation, where one object type was used for testing and the remaining three for training. This yields $N^{\text{tr}} = 120$ training samples and $N^{\text{te}} = 40$ test samples per fold. In contrast to the RGB-D Mirror Segmentation dataset, the mirror regions in this dataset are relatively small, such as small dishes and bathroom accessories. Although the resolutions of the RGB and depth images differ—1920×1080 for I_{rgb} and 512×424 for I_{dep}—they are resized to a common resolution of $H \times W = 416 \times 416$ for training and evaluation.

4.2 Comparison methods

We compare the performance of the proposed method with the state-of-the-art method PDNet [9], as well as a baseline method (i.e., the proposed method without the discAG module). The baseline is designed to evaluate the effectiveness of the proposed discAG by directly computing the difference between the features X^l_{dep} and $\widehat{X^l_{\text{dep}}}$, which are extracted from ToF and predicted depths, respectively, using a shared $\text{CNN}_{\boldsymbol{\theta}}$—Eq. 7 in the discAG is replaced as follows:

$$\widetilde{X^l_{\text{dep}}} = \text{CNN}_{\boldsymbol{\theta}}(I_{\text{dep}})^l - \text{CNN}_{\boldsymbol{\theta}}(\widehat{I_{\text{dep}}})^l. \tag{8}$$

4.3 Results

Table 1 presents the average performance of mirror and specular object segmentation. These tables show that the proposed method consistently outperforms the SOTA RGB-D-based

Table 1: Performance comparison using β, IoU, and MAE on RGB-D mirror segmentation and specular object datasets. Values represent the average of the top five trials out of ten, with values in parentheses indicating their standard deviation. The best result for each metric is highlighted in **bold**.

dataset	area ratio/object	model	$F_{\beta=0.5}\uparrow$	$F_{\beta=1.0}\uparrow$	IoU↑	MAE ($\times 10^{-2}$)↓
RGB-D Mirror Segmentation	all	PDNet [9]	86.80 (0.53)	85.41 (0.44)	68.85 (1.13)	8.16 (0.47)
	$N^{\text{te}} = 1049$	baseline	85.74 (1.39)	84.33 (1.31)	66.05 (3.03)	9.48 (0.92)
		proposed	**88.13 (0.33)**	**86.68 (0.33)**	**70.94 (0.88)**	**7.89 (0.27)**
	ratio $> 50\%$	PDNet [9]	93.14 (0.29)	93.58 (0.21)	65.88 (3.17)	29.41 (2.27)
	$N^{\text{te}} = 101$	baseline	90.65 (0.92)	91.23 (0.74)	58.66 (6.60)	34.30 (4.69)
		proposed	**93.47 (0.26)**	**94.00 (0.20)**	**68.01 (3.46)**	**28.13 (2.46)**
	$10\% <$ ratio $\leq 50\%$	PDNet [9]	91.96 (0.33)	91.00 (0.31)	77.62 (0.79)	8.47 (0.48)
	$N^{\text{te}} = 401$	baseline	89.97 (1.33)	88.97 (1.38)	72.79 (3.11)	10.14 (1.04)
		proposed	**92.53 (0.33)**	**91.54 (0.48)**	**78.41 (1.25)**	**8.31 (0.33)**
	ratio $\leq 10\%$	PDNet [9]	81.84 (0.87)	79.80 (0.72)	62.97 (1.39)	4.00 (0.35)
	$N^{\text{te}} = 547$	baseline	81.73 (1.62)	79.66 (1.47)	62.47 (2.47)	4.41 (0.34)
		proposed	**83.91 (0.59)**	**81.77 (0.43)**	**66.01 (0.77)**	**3.86 (0.21)**
Specular Object	*Planar*	PDNet [9]	40.87 (3.60)	39.11 (3.34)	9.39 (1.37)	4.93 (0.25)
		baseline	22.35 (8.61)	24.05 (7.15)	2.81 (1.16)	7.55 (3.44)
		proposed	**41.92 (1.70)**	**39.11 (1.34)**	**12.66 (1.69)**	**4.90 (0.11)**
	Spherical	PDNet [9]	84.73 (1.21)	79.51 (1.23)	56.97 (0.36)	**2.07 (0.19)**
		baseline	44.67 (14.69)	40.06 (12.65)	17.58 (5.18)	3.53 (0.92)
		proposed	**86.07 (0.73)**	**81.73 (1.55)**	**60.66 (2.23)**	2.12 (0.13)
	Holder	PDNet [9]	67.22 (2.38)	65.68 (2.06)	44.15 (1.24)	1.84 (0.34)
		baseline	43.97 (8.10)	42.87 (7.92)	24.94 (2.23)	2.21 (0.57)
		proposed	**69.32 (3.35)**	**67.58 (2.58)**	**48.40 (0.98)**	**1.57 (0.17)**
	Dish	PDNet [9]	79.48 (3.08)	77.35 (2.59)	57.35 (3.10)	1.73 (0.57)
		baseline	43.35 (17.36)	41.25 (16.68)	22.58 (12.64)	4.14 (2.19)
		proposed	**83.36 (1.09)**	**81.26 (1.50)**	**63.07 (2.46)**	**1.60 (0.43)**

mirror segmentation method, PDNet [9][1], as well as the baseline method, across almost all metrics and datasets. Additionally, we observe that the performance gain of our proposed method over PDNet and the baseline becomes increasingly pronounced as the mirror-to-image area ratio decreases, i.e., the IoU scores of PDNet and our method are 65.88 and 68.01 respectively for ratio $> 50\%$, and 62.97 and 66.01 for ratio $\leq 10\%$. This trend supports the analysis in Secs. 1 and 2.3, and Fig. 1, which highlighted that the co-occurrence of RGB and depth discontinuities in non-mirror regions causes frequent false positives—particularly in small-mirror cases.

On the Specular Object dataset, although performance varies depending on the shape and complexity of the target object, the proposed method shows notably larger improvements for compact and geometrically challenging objects such as the paper holder and small dish. In particular, the baseline method, which directly compares ToF and predicted feature maps (Eq. 8), suffers from significantly degraded accuracy. This degradation is largely due to the field-of-view mismatch between the RGB I_{rgb} and the ToF depth map I_{dep} (see Fig. 4), which leads to spatial inconsistencies between the two depth modalities, I_{dep} and $\widehat{I_{\text{dep}}}$. Especially, the performance for *Planar* is significantly lower than others because the relatively large mirror surface reflects different objects in different regions, causing the ToF behavior to vary

[1] Although we used the publicly available implementation [8] and re-implemented parts of the training procedure based on the descriptions in the paper [9], the performance scores in our table for PDNet on the RGB-D Mirror Segmentation dataset are lower than those reported in [9]. This discrepancy arises because we report the average score over the best five trials out of ten to ensure a fairer comparison.

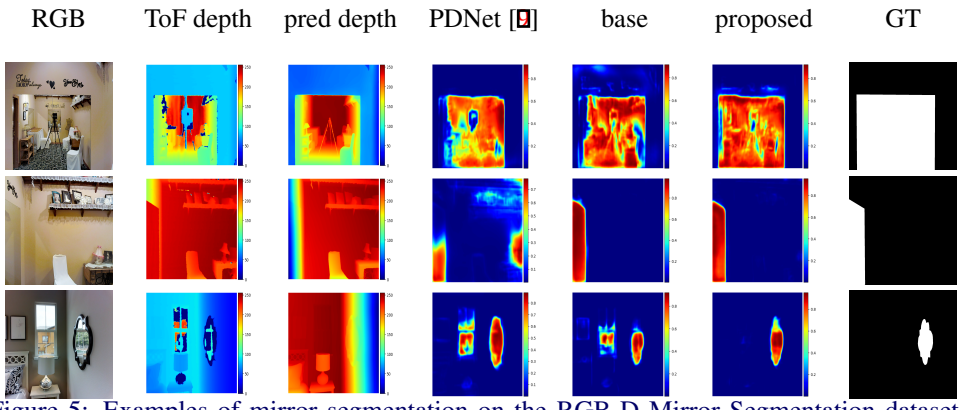


Figure 5: Examples of mirror segmentation on the RGB-D Mirror Segmentation dataset, showing a large (ratio $>$ 50%), medium-sized (10% $<$ ratio $\leq$ 50%), and small mirror (ratio $\leq$ 10%) from top to bottom.

across the entire surface and resulting in only partial detection. In contrast, the proposed method mitigates this issue by aggregating features in a spatial- and channel-wise manner before performing the comparison, thereby reducing the adverse effects of such geometric misalignments.

Fig. 5 shows representative predicted mirror maps $\widehat{Y}$ from the RGB-D Mirror Segmentation dataset for large (ratio $>$ 50%), medium (10% $<$ ratio $\leq$ 50%), and small (ratio $\leq$ 10%) mirrors—Fig. 1 further illustrates the performance on small mirrors. These figures show that in ratio $>$ 50%, there is little difference between PDNet [9] and the proposed method because mirrors occupy a large portion of the image, making the co-occurrence of depth and RGB discontinuities in non-mirror objects less likely. As the proportion of mirror area decreases, the influence of such co-occurrences in non-mirror objects becomes more prominent, leading to more false detections by PDNet. In contrast, the proposed method successfully suppresses these false detections by leveraging the contextual inconsistency between the two depth sources.

Fig. 6 presents $\widehat{Y}$ on the Specular Object dataset. As shown, PDNet and the baseline misclassify non-mirror regions as specular objects due to misleading co-occurrences of discontinuities in RGB and ToF depth, especially across diverse object types. In contrast, the proposed method substantially reduces false positives by relying on depth inconsistency cues.

4.4 Ablation study

To verify the effectiveness of each component of the proposed discAG module, we conducted an ablation study using the RGB-D Mirror Segmentation dataset as shown in Table 2. The results indicate that incorporating proposed modules (disAG and dicAG) and predicted depth $\widehat{I_{\text{dep}}}$ are beneficial for mirror segmentation. In particular, removing disAG (w/o dicAG) leads to a larger performance drop than removing dicAG, suggesting that spatial attention plays a more critical role. Further removing $\widehat{I_{\text{dep}}}$ (w/o $\widehat{I_{\text{dep}}}$ + w/o disAG) shows the largest performance drop (see $F_{\beta=1.0}$ and IoU). Overall, the ablation study validates the complementary effectiveness of each component in the discAG module.

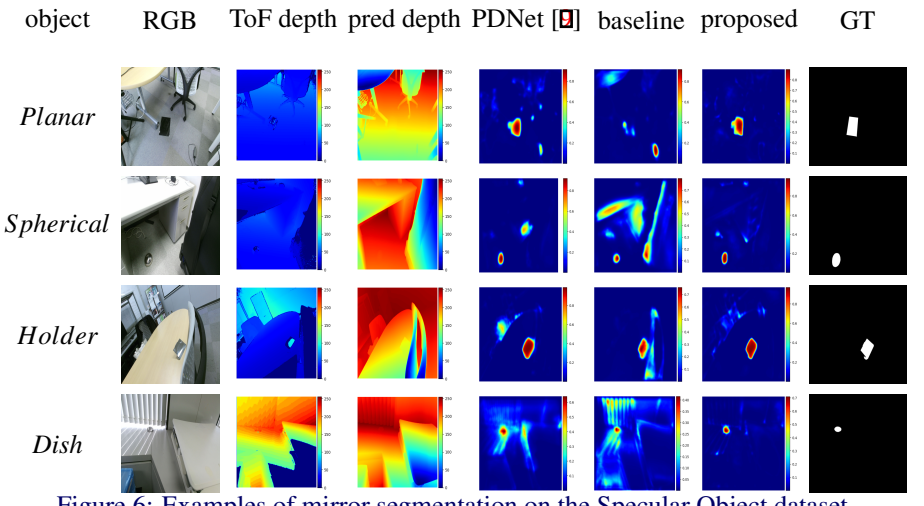


Figure 6: Examples of mirror segmentation on the Specular Object dataset.

Table 2: Ablation study on the RGB-D Mirror Segmentation dataset (all) evaluating the contributions of disAG, dicAG, and depth prediction $\widehat{I_{\text{dep}}}$. The condition is same as Table 1.

model	$F_{\beta=0.5}\uparrow$	$F_{\beta=1.0}\uparrow$	IoU↑	MAE ($\times 10^{-2}$)↓
proposed	**88.13 (0.33)**	**86.68 (0.33)**	**70.94 (0.88)**	**7.89 (0.27)**
w/o disAG	87.15 (0.55)	85.83 (0.52)	68.77 (1.03)	8.37 (0.61)
w/o dicAG	87.73 (0.85)	86.32 (0.80)	69.40 (1.31)	8.34 (0.42)
w/o $\widehat{I_{\text{dep}}}$	87.38 (0.66)	85.90 (0.63)	68.40 (1.08)	8.99 (0.70)
+ w/o disAG	85.94 (0.54)	84.55 (0.57)	66.43 (0.62)	9.24 (0.33)
+ w/o dicAG	85.86 (0.21)	84.60 (0.19)	67.07 (0.70)	9.16 (0.23)

5 Conclusion

In this study, we proposed a novel attention module, discAG, designed to capture depth inconsistencies between time-of-flight (ToF) depth and predicted depth, and integrated it into a deep learning framework for RGB-D mirror and specular object segmentation. We also constructed a new RGB-D Specular Object dataset composed of multiple specular objects to complement the existing RGB-D Mirror Segmentation dataset. Through extensive experiments on both datasets, we demonstrated that the proposed method effectively reduces false positives caused by co-occurring discontinuities in RGB and ToF depth images, which often mislead existing methods.

One current limitation of our method lies in its computational cost, primarily due to the use of Depth Anything V2 as the depth prediction backbone as shown in Table 2 (supplementary material). While this model ensures high-quality depth estimation, it increases overall processing time. As future work, we plan to explore more lightweight depth estimation approaches to improve the efficiency of our framework without compromising segmentation accuracy.

References

[1] Andrey DelPozo and Silvio Savarese. Detecting specular surfaces on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383215.

[2] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 5941–5950, 2022.

[3] Hirotaka Hachiya and Yuto Yosimura. Specular surface detection with deep static specular flow and highlight. *Machine Vision Applications*, 35(126), 2024.

[4] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson W.H. Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 935–943, 2023.

[5] Trinh Le Ba Khanh, Duy-Phuong Dao, Ngoc-Huynh Ho, Hyung-Jeong Yang, Eu-Tteum Baek, Gueesang Lee, Soo-Hyung Kim, and Seok Bong Yoo. Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. *Applied Sciences*, 10(17), 2020.

[6] KinectV2. Kinect for windows sdk 2.0. https://www.microsoft.com/en-us/download/details.aspx?id=44561(Accessed: 2025-05-6).

[7] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3694–3702, 2020.

[8] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Cvpr2021_pdnet. https://github.com/Mhaiyang/CVPR2021_PDNet(Accessed: 2025-05-5).

[9] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[10] Alexandre Morgand and Mohamed Tamaazousti. Generic and real-time detection of specular reflections in images. In *Proceedings of International Conference on Computer Vision Theory and Applications*, volume 1, pages 274–282, 2014.

[11] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.

[13] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8809–8818, 2019.