RGB-D Video Mirror Detection

Mingchen Xu[1] **Peter Herbert**[1] **Yu-Kun Lai**[1] **Ze Ji**[2] **Jing Wu**[1]
[1]School of Computer Science and Informatics, Cardiff University
[2]School of Engineering, Cardiff University
{xum35, herbertp1, laiy4, jiz1, wuj11}@cardiff.ac.uk

Abstract

Mirror detection aims to identify mirror areas in a scene, with recent methods either integrating depth information (RGB-D) or making use of temporal information (video). However, utilizing both data is still under-explored due to the lack of a high-quality dataset and an effective method for the RGB-D Video Mirror Detection (DVMD) problem. To the best of our knowledge, this is the first work to address the DVMD problem. To exploit depth and temporal information in mirror segmentation, we first construct a large-scale RGB-D Video Mirror Detection Dataset (DVMD-D), which contains 17977 RGB-D images from 273 diverse videos. We further develop a novel model, named DVMDNet, which can first locate the mirrors based on triple consistencies: local consistency, cross-modality consistency and global consistency, and then refine the mirror boundaries through content discontinuity, taking the temporal information within videos into account. We conduct a comparative study on the DVMD dataset, evaluating 12 state-of-the-art models (including single-image mirror detection, single-image glass detection, RGB-D mirror detection, video shadow detection, video glass detection, and video mirror detection methods). Code is available from https://github.com/UpChen/2025_DVMDNet.

1. Introduction

Mirror Detection aims to distinguish reflective areas, also known as mirror regions, from a scene. Reflection from mirrors can help with some computer vision tasks, such as locating objects [27], 3D human pose reconstruction [5], and 3D scene reconstruction [25]. However, the presence of mirrors can also affect the performance of existing computer vision tasks. For example, object detection methods may identify objects in the mirror as real objects. Thus, it is necessary for current computer vision systems to have the ability to detect mirrors accurately.

Some mirror detection methods, especially RGB image-based methods [7–10, 12, 22, 29, 32] have been developed in recent years. However, these methods may fail when there is a large variation inside the mirror (first column of Fig. 1), or the reflection of the mirror and the surroundings are too similar. Objects that look like mirrors, e.g., paintings, windows, etc., also pose challenges to these methods. As observed in [16], there are obvious depth discontinuities between the reflections inside the mirror and the surrounding outside the mirror. Consequently, depth information is a strong cue for mirror detection, and has been used in mirror from RGB-D image [16, 18, 35, 36].

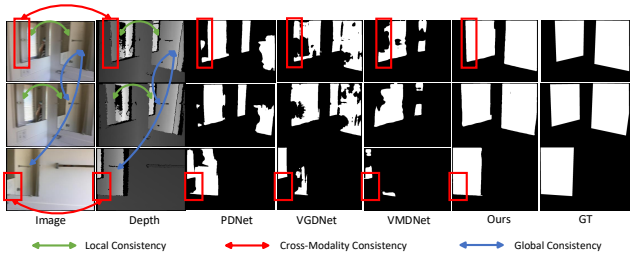


Figure 1. Although existing depth-based mirror detection methods such as PDNet [16] have shown their effectiveness by using spatial consistency in static scenes, and existing video-based mirror detection methods such as VMDNet [11] have proved their effectiveness by using spatial and temporal consistency in semantic sequence, they may fail when there are obvious discontinuities inside the mirror (1st and 2nd rows) or when the number of mirrors is changed (2nd and 3rd rows). The lack of exploiting spatial and temporal consistency in depth sequence (green and blue arrows) and consistency information across the RGB and depth (red arrows) causes the current mirror detection methods to produce unsatisfactory results when applied to the DVMD task. On the contrary, our method can perform well by utilizing the proposed triple consistency module to exploit the spatial, temporal, and cross-modality consistencies in both RGB and depth.

In addition to RGB image-based methods, there has been considerable progress in video mirror detection (VMD) by taking into account dynamic scenes. Compared with static images, dynamic scenes are more challenging because of various motion modes, occlusion, blurring, and object de-

formation. By encoding temporal consistency within dynamic scenes, VMD methods [11,24,30] have demonstrated effectiveness in mitigating these problems.

Detecting mirrors from RGB-D images and videos has received research interest individually. However, there remains a gap in research with combining depth information and temporal information, both of which are important for accurately distinguishing mirrors from scenes. First, the correlation between inside and outside the mirror regions in RGB and depth is exploited in [16] which we call local consistency in this paper, as shown by the green arrow in Fig. 1. Second, we observe that there is consistent information between RGB and depth which we call cross-modality consistency. The red arrow of Fig. 1 shows that content discontinuities are obvious in both RGB and depth, thereby creating consistency across modalities. By considering cross-modality consistency, we can align the RGB and depth features more effectively. Third, we observe that the global consistency between inside and outside the mirror regions not only exists in semantic sequence but also exists in depth sequence. The former can be detected in the RGB domain and has been exploited by prior mirror segmentation works [8, 11, 12]. The latter, global consistency in depth sequence is still not explored. The blue arrow of Fig. 1 shows that there is global consistency in multiple frames' depth discontinuities at mirror boundaries. Based on the above three consistencies, we propose the Triple Consistency (TC) Module to first determine the initial mirror area by considering local consistency between the inside and outside of the mirror and the consistent information between RGB and depth, and then detect the mirror by taking into account the global consistency in both RGB and depth sequences.

In addition, content discontinuity caused by mirrors not only exists in static scenes but also exists in dynamic scenes. Besides discontinuity in static scenes that can be efficiently detected [8, 12, 16, 32], the temporal discontinuity in RGB and depth sequences has still not been explored. The short-term discontinuity can provide spatial prior information on the mirror boundary, and the long-term discontinuity can provide context information on the mirror boundary. Based on this, we propose a Temporal Discontinuity (TD) module that captures local and global contextual contrasted features in RGB and depth sequences, for refining the mirror boundaries.

To further support leveraging both depth and temporal information for mirror detection, we further propose a large-scale dataset for DVMD, which contains 17977 RGB-D images from 273 diverse videos. We have conducted extensive experiments to evaluate our model against state-of-the-art methods. The results demonstrate that our proposed model outperforms existing methods on the proposed large-scale DVMD-D.

Our main contributions can be summarized as:

- We construct a large-scale RGB-D video mirror detection dataset, called DVMD-D. The new dataset contains 17977 RGB-D images from 273 videos with pixel-wise annotated masks.
- We propose a novel network, called DVMDNet, which leverages local, cross-modality, and global consistency via a triple consistency (TC) module to initialize the mirror region and local and global contextual contrasted features in RGB and depth sequences to refine the mirror boundaries.
- Extensive experiments show that our method outperforms existing state-of-the-art methods for mirror detection on our proposed DVM-D dataset.

2. Related Work

2.1. Mirror Segmentation from RGB-D Images

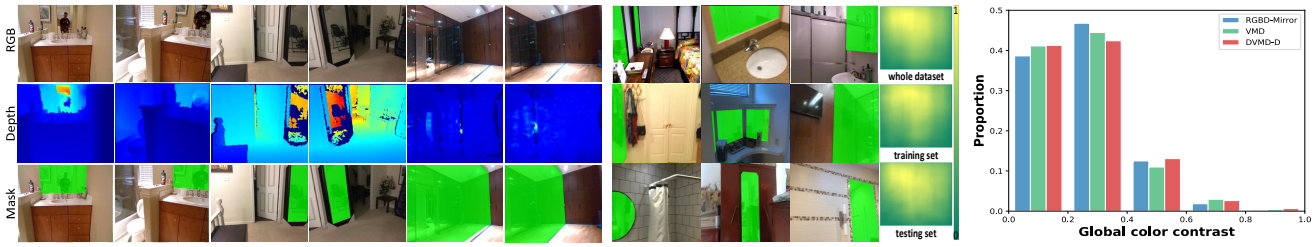
In recent years, numerous studies [16, 18, 35, 36] have been proposed to detect mirrors from RGB-D images such as geometric relationship [18], color and depth discontinuities and correlations [16], global contextual relationship [35] and morphology knowledge [36]. While RGB-D single-image mirror detection models have achieved reliable results, their performance on video data remains sub-optimal due to insufficient exploitation of temporal information.

2.2. Mirror Segmentation from Videos

Video Mirror Segmentation has recently started to gain attention [11, 24, 30]. Lin *et al.* [11] propose the first video detection network, named VMDNet. It focuses on extracting spatial and temporal correspondences between mirror and non-mirror to detect the mirror. Alex *et al.* [24] model motion inconsistency between the mirror and its surroundings for mirror detection. Xu *et al.* [30] model the temporal variation in similarity and contrast to detect the mirror. These methods improve the performance of video mirror segmentation but still perform poorly in some challenging scenarios.

3. RGB-D Video Mirror Detection Dataset (DVMD-D)

Our first contribution is the RGB-**D** **V**ideo **M**irror **D**etetion **D**ataset (DVMD-D), which contains 17,977 RGB-D images from 273 videos and corresponding pixel-level annotations. Instead of collecting all the data ourselves, we construct the DVMD-D by extending the recently proposed ViMirr dataset [31] to ensure a wide variety and extensive range. The composition and split of our dataset are shown in Tab. 1. Our dataset consists of 223 videos from ViMirr and 50 videos self-captured over different scenes. Fig. 2 (a)



(a) Examples of the proposed DVMD-D. (b) Mirror Location Distribution (c) Color Contrast Distribution

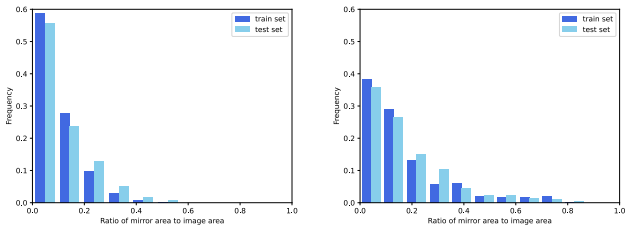
Figure 2. DVMD-D examples and statistics.

shows some examples of DVMD-D. In line with the common practices [16] [11] used for constructing datasets for RGB-D-based and video-based mirror problems, we make sure that each frame in the DVMD-D includes at least one mirror region. To the best of our knowledge, DVMD-D is the first RGB-D Video Mirror Dataset dedicated to mirror segmentation.

Dataset	Images	Videos	Train		Test	
			Images	Videos	Images	Videos
ViMirr [31]	13967	223	7484	114	6483	109
Self-captured	4010	50	2160	27	1850	23
Total	17977	273	9644	141	8333	132

Table 1. Composition of our benchmark for RGB-D video mirror detection. The fourth and fifth columns show the dataset split.

Dataset Construction: The ViMirr dataset [31] contains 19,255 frames from 276 videos. The data comes from three sources: NYUv2 [20], ScanNet [4], and videos captured by themselves. The former two have depth maps as well. So, we first include all data in ViMirr that come from NYUv2 and ScanNet, which is called Part ViMirr below. However, we find that the mirror area distribution of the Part ViMirr is biased. From Fig. 3 (a), we can see that Part ViMirr only contains mirrors covering the range of 0 to 0.6 area ratios. In addition, we also notice that almost 60% of the dataset is less than 0.1 mirror area ratio. To address the above problem, we first manually reduced the number of frames with a mirror area of less than 0.1 only in the beginning or end of videos in Part ViMirr, and then used the Intel RealSense Camera to collect the RGB-D videos with mirrors in daily-life scenes. From Fig. 3 (b), we can see that our dataset includes mirrors that span a broad spectrum of area ratios. The pixel-level mirror masks of our data are created by professional annotators. All collected videos have a frame rate of 30 fps.



(a) Part ViMirr (b) DVMD

Figure 3. Mirror area ratio of the Part ViMirr and DVMD datasets.

Dataset Analysis: Fig. 2 (b) (c) provide the statistical results of our DVMD-D from two dimensions. (1) *Mirror location distribution*: To analyze the spatial distribution of mirrors in DVMD-D, we compute probability maps to illustrate the likelihood of each pixel belonging to a mirror, as shown in Fig. 2 (b). Despite the fact that our DVMD-D includes mirrors in various locations, they predominantly cluster in the upper part of the image. This clustering is reasonable, given that mirrors are typically positioned around human eye level. In addition, the mirror location distribution for the training and test splits aligns closely with the distribution across the entire dataset. (2) *Color contrast distribution*: We analyze the global color contrasts between the contents inside and outside of the mirror by computing the $\mathcal{X}^2$ distance between their RGB histograms. In addition, we compare the distribution between RGBD-Mirror [16] and VMD [11], as shown in Fig. 2 (c). In general, DVMD-D includes more frames with extremely low color contrast (< 0.2) compared to the existing mirror datasets RGBD-Mirror and VMD, which makes the mirror segmentation task more challenging.

4. Method

Our approach builds on two core ideas. First, we observe that consistency between the real object and its corresponding reflection not only exists in the RGB sequence, but also exists in the depth sequence, and there is consistent information across semantics and depth. Based on this idea, we propose a Triple Consistency Module to estimate the mirror location initially. Second, we observe that content discontinuity caused by mirrors occurs not only in static scenes but also in dynamic scenes. Temporal discontinuity is divided into short-term temporal discontinuity and long-term temporal discontinuity. The former obtained from the previous

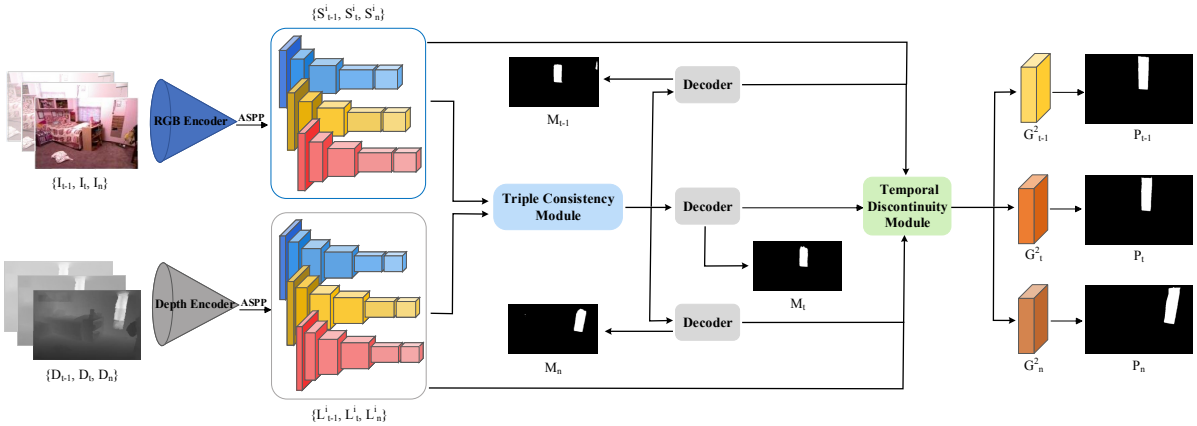


Figure 4. The overview of our proposed model. We first feed three RGB-D images from the same video to the backbone feature extractor to extract multi-scale features, then the TC module extracts triple consistency features to locate the mirror. Each decoder subsequently processes the triple consistency features and produces intermediate mirror maps as output. Second, the TD module extracts temporal contextual contrasted features to refine the mirror boundary.

frame can provide spatial prior for the mirror boundary, and the latter obtained from the random frame can provide contextual information for the mirror boundary. Based on this idea, we propose the Temporal Discontinuity Module to refine the mirror boundary further.

Fig. 4 shows the overall architecture of the proposed DVMDNet. DVMDNet takes three RGB-D image pairs as input and passes them through two different multi-level feature extractors to derive features separately from RGB and depth information. The first two RGB-D images are from adjacent video frames, the third RGB-D images are randomly selected from other frames. We employ SegFormer [28] as our backbone feature extractor to extract multi-scale RGB features. Following [11], we also add an Atrous Spatial Pyramid Pooling (ASPP) after the SegFormer to obtain enhanced high-level features. To reduce computation, we employ a sequence of five cascaded 3×3 convolutional blocks, where each block includes an increasing number of channels (64-128-256-512-512), followed by max pooling to extract depth features. Its channel configuration is the same as that of the RGB backbone feature extractor. Again, we add an ASPP module behind the simple network. After obtaining RGB and depth backbone features, we feed them into a triple consistency module (Fig. 5) and a temporal discontinuity module (Fig. 6). The triple consistency (TC) module estimates the mirror's location using the local, cross-modality, and global consistency features in both RGB and depth sequences. The temporal discontinuity (TD) module refines the mirror boundary based on the temporal discontinuity features from both the short-term (previous frame) and long-term (random frame) information.

4.1. Triple Consistency (TC) Module

Fig. 5 shows the structure of the proposed TC module. The TC module is designed to model local and global consistency and cross-modality consistency in both RGB and depth sequences to first locate the mirror. Then the output features of the TC module will guide the subsequent TD module to refine the mirror boundaries. The gray part in Fig. 5 shows the overall network architecture, in which there are three main blocks: Local Consistency (LC) block, Cross-modality Consistency (CMC) block, and Global Consistency (GC) block. LC and CMC blocks aim to locate the mirror region in one single frame. By considering the temporal consistency in both RGB and depth sequences extracted by the GC block, the TC module can confirm the mirror location in RGB-D dynamic scenes. Specially, given the RGB $\{R_{t-1}, R_t, R_n\}$ and depth features $\{D_{t-1}, D_t, D_n\}$, we first extract the modality-wise local consistency features $\{R^{lc}, D^{lc}\}$ with LC blocks to initially estimate the mirror location (light blue part of Fig. 5), and then confirm the mirror location with CMC blocks by considering consistent information between the semantic and depth (deep blue part of Fig. 5). After locating the mirror region in a single frame, GC blocks extract global consistency $\{F^g_{t-1}, F^g_t, F^g_n\}$ to detect the mirror region in dynamic scenes (green part of Fig. 5). The fusion block is used to combine short-term and long-term global consistency features for F^g_t to obtain the final global consistency feature $\hat{F^g_t}$ for the current frame. Following the [11], our TC module only exploits the low-level features at the 2nd scale and high-level features at the last scale (5th scale for RGB features and 6th scale for depth features).

Local Consistency (LC) Block: Our LC block aims to

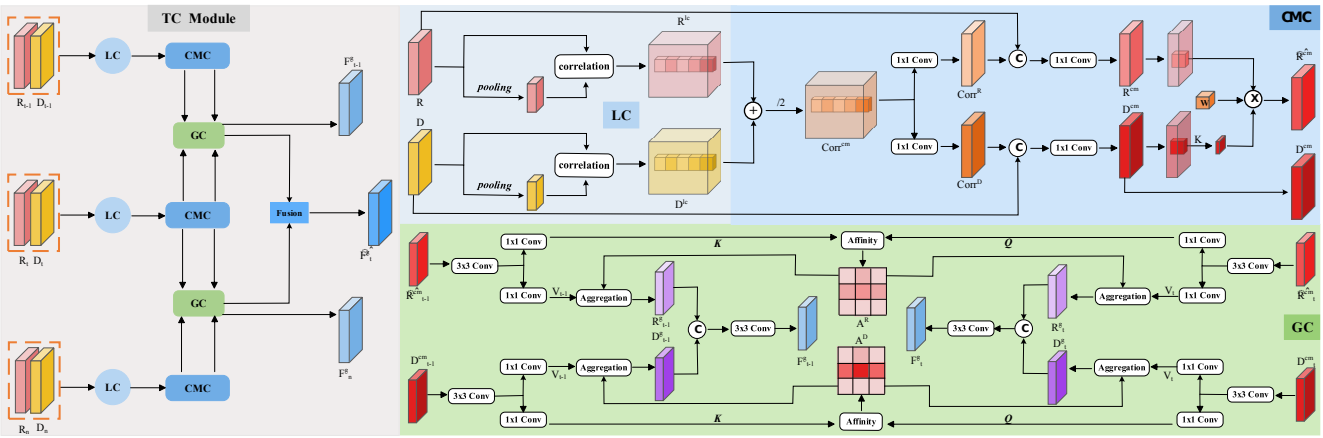


Figure 5. The schematic illustration of our Triple Consistency (TC) module and its three main building blocks: the local consistency block (light blue part), the cross-modality consistency block (green part), and the global consistency block (dark blue part).

extract the local consistency inside and outside of the mirror in one single frame in both RGB and depth. It follows the design of Long-Range Context Information Gathering (LCG) module in [23] which can efficiently and effectively extract the long-range relations in a pixel-patch way. We use LCG to extract the consistency between the real object and its corresponding reflection in the mirror. As the light blue part of Fig. 5 demonstrated, we extract modality-wise local consistency for each pixel of RGB and depth respectively. For computational efficiency, a pooling operation is used to obtain low-resolution template patches. Then the correlation between each pixel and each template patch is calculated and finally, modality-wise local consistency is obtained.

Cross-Modality Consistency (CMC) Block: Our CMC block aims to align RGB and depth features by considering cross-modality consistency. An intuitive way to fuse RGB and depth features is concatenation. However, simply concatenating the RGB and depth features and ignoring some useful consistent information between them is insufficient to handle complex scenes. Toward better performance, we introduce the CMC block that aligns RGB and depth by extracting useful intrinsic consistency features across modalities. It is inspired by the Cross-Modality Context Information Fusion (CCIF) module in [23] which only enhances the RGB features with cross-modality consistency. We extend it by enhancing both RGB and depth features with cross-modality consistency. The reasons we also enhance the depth features are 1) the color and texture information from the RGB image is complimentary for depth features; and 2) the enhanced depth features will help the subsequent GC block to extract more accurate global consistency features. Given the modality-wise local consistency features R^{lc}, D^{lc}, the cross-modality consistency features $Corr^{cm}$ is obtained by a simple average. Then, we reduce the channel of $Corr^{cm}$ to keep the same channel number with corresponding RGB and depth features by the 1×1 convolution layer. With the $Corr^{R}$ and $Corr^{D}$, we enhance the RGB features with layout information and the depth features with the color and texture information. The enhancing process is formally defined as:

$$Corr^{cm} = (R^{lc} + D^{lc})/2 \tag{1}$$

$$R^{cm} = Conv_{1\times1}(Cat(R^{lc}, Corr^{cm})) \tag{2}$$

$$D^{cm} = Conv_{1\times1}(Cat(D^{lc}, Corr^{cm})) \tag{3}$$

where R^{cm} and D^{cm} are the layout-aware RGB features and semantic-aware depth features.

Following the Refinement with Local Depth Structure Information (RLDSI) module in [23], to make up for the loss caused by the relatively low-resolution patch feature obtained by downsampling operation in the LC block, we further refine the cross-consistency aware RGB feature with the depth layout information in a content-adaptive way. Compared with the original RLDSI module which directly uses depth backbone features D to refine the layout-aware RGB features R^{cm}, we choose semantic-aware depth features D^{cm} to refine it. The reason is that depth features can vary significantly within a mirror region, where these in salient object regions are usually the same. Therefore, semantic-aware depth features which are complemented by the mirror consistency based on the color and texture information in RGB can provide more accurate scene layout information. With the D^{cm}, PAC [21] operation is used to further weight the layout-aware RGB features R^{cm} with depth structure information. The refinement process is formally defined as:

$$\hat{R}^{cm}(x_i) = \sum_{j\in\{(-1,-1)(-1,0),\ldots,(1,1)\}} Ker(i,i+j)W(j)R^{cm}(x_{i+j}) \tag{4}$$

$$Ker(i,i+j) = \exp(-\frac{1}{2}\sum_{c=1}^{c}(D^{cm}(x_i) - D^{cm}(x_{i+j}))^2) \tag{5}$$

where W is the convolution filter weights and Ker adaptively calculates the coefficients of feature within each local convolution window according to the Gaussian function. C is the channel number. In the end, we obtain refined layout-aware RGB feature $\hat{R}^{cm}$ and semantic-aware depth feature D^{cm} which are the inputs of the GC block.

Global Consistency (GC) Block: Our GC block extracts and fuses temporal consistency features for RGB sequence ($\hat{R}^{cm}_{t-1}$ and $\hat{R}^{cm}_{t}$) and depth sequence (D^{cm}_{t-1} and D^{cm}_{t}). Each of these features is extracted by a cross-attention module proposed in [11] which can effectively extracting long-range temporal consistency between the contents inside and outside of the mirror across different frames. After obtaining single-modality temporal consistency features R^g_{t-1}, R^g_t, D^g_{t-1} and D^g_t, we concatenate them to obtain cross-modality temporal consistency features F^g_{t-1} and F^g_t which consider temporal consistency based on the color and texture feature from the RGB sequence and the layout feature from the depth sequence. In the green part of Fig. 5, we omitted the global consistency extracting process for RGB ($\hat{R}^{cm}_t$ and $\hat{R}^{cm}_n$) and depth (D^{cm}_t and D^{cm}_n) for clarity.

The fusion block aims to fuse the short-term global consistency features F^g_t and long-term global consistency features $\hat{F}^g_t$. In our experiment, the performance was worsened by simply plus or concatenating them because their distribution ranges of the correlation values are inconsistent. Therefore, we adopt the same design as the SLF module proposed in [31] which can effectively fuse short-term and long-term features by encoding both the appearance of the mirror from the short-term features and the position of the mirror from the long-term features. The decoders following the TC module comprise a 3×3 convolution layer and a 1×1 convolution layer, which generate intermediate mirror maps.

4.2. Temporal Discontinuity (TD) Module

After the TC module locates the mirror, our TD module can refine the mirror boundaries. The core of the TD module takes advantage of both short-term and long-term discontinuities in both RGB and depth sequences to refine the mirror boundary. Unlike the contextual contrasted feature extraction (CCFE) in [32] and the delineating module in [16], which only focus on local contextual contrasted information in the single modality or the single image, our

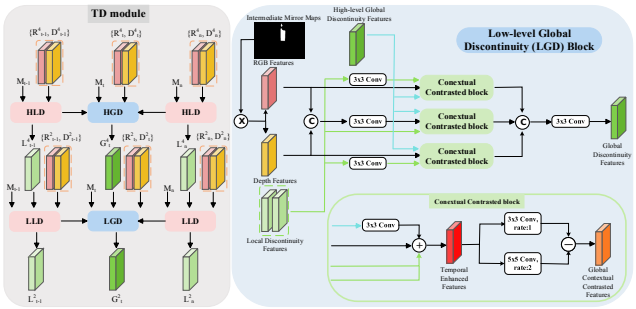


Figure 6. The schematic illustration of Temporal Discontinuity module. The gray part denotes the whole architecture of the Temporal Discontinuity module. The blue part denotes the low-level Global Discontinuity Block.

module exploits temporal discontinuity within RGB-D dynamic scenes. In the TD module, the local discontinuity features we extracted from the previous frame contain the location information of the boundary, and the local discontinuity features extracted from the next frame contain the appearance information of the boundary. After fusing the local discontinuity features of the previous and random frames with the local discontinuity features of the current frame, we obtain global discontinuity features that encode both location information and appearance information.

The gray part of Fig. 6 shows the architecture of the TD module. Our TD module consists of two kinds of blocks: the local discontinuity (LD) block and the global discontinuity (GD) block. LLD and HLD are low-level and high-level local discontinuities. LGD and HGD are low-level and high-level global discontinuity. They are used to deal with high-level (4^{th} scale) and low-level (2^{nd} scale) features. The difference between (LLD, LGD) and (HLD, HGD) is that low-level blocks accept one extra input(high-level discontinuity features) to narrow down the mirror region. The difference between the LD and GD is that GD needs to accept two extra inputs (short-term and long-term discontinuity features from the LD). Given the intermediate mirror maps M, RGB features R, depth features D, high-level global discontinuity features G^4_t and local discontinuity features L^2_{t-1}, L^2_n, we first multiply RGB R and depth D features by their corresponding mirror maps M, which are normalization by a sigmoid function to get the potential mirror area. Then, we extract the global contextual contrasted feature G for RGB, depth, and RGB+depth. The contextual contrasted block is formally defined as:

$$G = f_l(F \oplus L^2_{t-1} \oplus L^2_n \oplus \hat{G}^4_t, \Theta_l) - f_c(F \oplus L^2_{t-1} \oplus L^2_n \oplus \hat{G}^4_t, \Theta_c) \tag{6}$$

$$\hat{G}^4_t = U_4(\mathcal{R}(\mathcal{N}(Conv_{3\times3}(G^4_t)))) \tag{7}$$

where f_l with corresponding parameters Θ_l denotes a convolution operation with a kernel size of 3 and a dilation rate

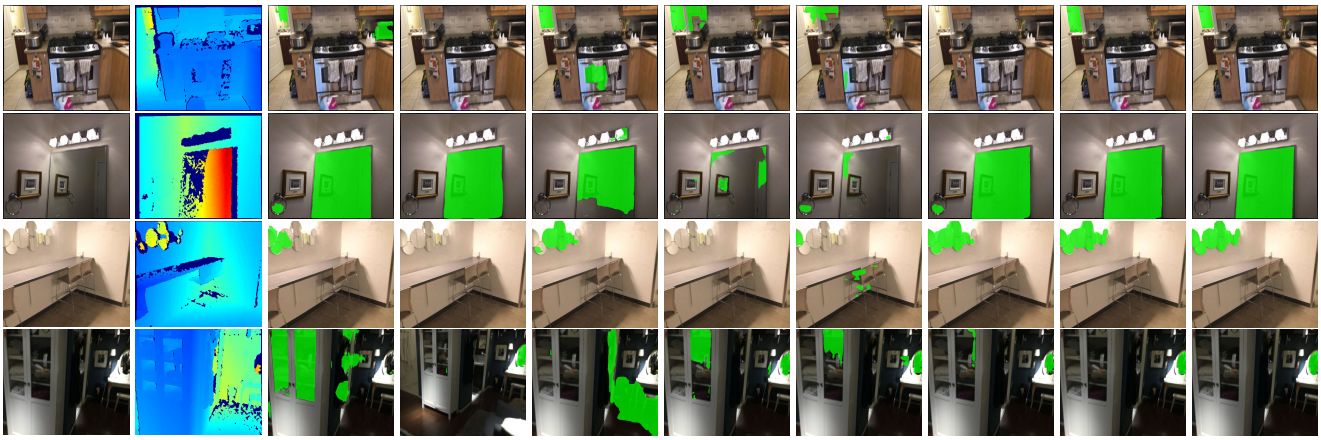


Figure 7. Visual comparison of DVMDNet with state-of-the-art segmentation methods retrained on the DVMD-D demonstrates that DVMDNet outperforms the competing methods on scenes with small mirrors (row 1), large mirrors (row 2), multiple mirrors (row 3), and challenging scenes featuring similar boundaries and appearances (row 4). You can find more visual results in Section 1.1 of the supplementary material.

of 1. f_c with corresponding parameters Θ_c denotes a convolution operation with a kernel size of 5 and a dilation rate of 2. U_2 represents a bilinear upscaling (by a factor of 4). After obtaining glocal contextual contrasted features for RGB, depth, and RGB+depth, we concatenate them together to get the final global discontinuity features. In the end, we forward the output features of the TD module to individual decoders to obtain the final output predictions P_{t-1},P_t,P_n. In the blue part of Fig. 6, we only show the details of the LGD block for clarity.

4.3. Loss Function

Following the suggestion of [2], we adopt the binary cross-entropy (BCE) and the Lovász-hinge loss to both supervise the training of our network. The final loss function is:

$$\mathcal{L} = \sum_{i}^{i\in\{t-1,t,n\}} \mathcal{L}_h(M_i, G_i) + \mathcal{L}_b(M_i, G_i) + \mathcal{L}_h(P_i, G_i) + \mathcal{L}_b(P_i, G_i) \tag{8}$$

where $\mathcal{L}_{hinge}$ and $\mathcal{L}_{bce}$ represent the Lovász-hinge loss and the binary cross-entropy (BCE) loss, respectively. M_i denotes the intermediated predicted map, P_i denotes the final predicted map, and G_i represents the ground truth mirror map.

5. Experiments

5.1. Experimental Setting and Evaluation Metrics

Our proposed segmentation architecture is developed using the PyTorch [19] deep learning framework. The feature extraction encoder parameters are initialized with the weights from the MiT-B3 model, pre-trained for image segmentation on the ADE20K dataset [33, 34] and publicly available on HuggingFace [26]. The other parameters, including those for attention modules and the MLP decoder, are randomly initialized using the "Xavier" method [6]. For training, we employ the AdamW optimizer [15] with an initial learning rate of 1×10^{-5} with a weight decay of 5×10^{-4}. All experiments and ablation studies are conducted for 15 epochs on an NVIDIA A100 GPU with 40GB RAM and a batch size of 5. We adopt four widely recognized metrics to quantitatively assess the performance of the tested methods: intersection over union (IoU), pixel accuracy, F-measure(F_β) [1], and mean absolute error (MAE).

5.2. Comparison to SOTA Techniques

Due to the lack of existing methods for RGB-D video mirror detection, we compare our approach with 12 state-of-the-art methods from related fields. These include TVSD [3] and SCOTCH [14] for video shadow detection, GDNet [17] for glass detection, VGNet [13] for video glass detection, MirrorNet [32], PMD [12], SANet [7], HetNet [8] and SATNet [9] for mirror detection, VMDNet [11] for video mirror detection, and PDNet [16] for RGB-D mirror detection. We trained and tested all baseline methods on our video mirror detection dataset VMD-D using their released codes on the same platform. Tab. 2 presents the quantitative results, demonstrating that our method significantly outperforms all others across all four metrics.

Fig. 7 provides a visual comparison of the results produced by our method with those obtained from two prior mirror segmentation methods (*i.e.*, PDNet [16] and VMDNet [11]) as well as the best approach from each of the three other categories (*i.e.*, glass detection method GDNet [17],

Methods	Pub. Year	IoU↑	F_β ↑	Accuracy↑	MAE↓
TVSD [3]	CVPR'2021	0.4088	0.6748	0.8829	0.1169
SCOTCH [14]	CVPR'2023	0.6211	0.7847	0.9064	0.0836
GDNet [17]	CVPR'2020	0.5602	0.7338	0.9036	0.0936
VGDNet [13]	AAAI'2024	0.5163	0.7197	0.9059	0.0939
MirrorNet [32]	ICCV'2019	0.5442	0.7509	0.9061	0.0937
PMD [12]	CVPR'2020	0.5753	0.7795	0.9201	0.0797
SANet [7]	CVPR'2022	0.5340	0.7133	0.9012	0.0986
HetNet [8]	AAAI'2023	0.4828	0.7420	0.8997	0.1001
SATNet [9]	AAAI'2023	0.6401	0.7985	0.9013	0.0887
CSFwinformer [29]	TIP'2024	0.6638	0.8148	0.9118	0.0780
VMDNet [11]	CVPR'2023	0.5673	0.7873	0.9060	0.0939
PDNet [16]	CVPR'2021	0.5851	0.7925	0.9144	0.0855
DVMDNet w/o D	Ours	0.7024	0.8388	0.9452	0.0547
DVMDNet	Ours	**0.7423**	**0.8581**	**0.9490**	**0.0509**

Table 2. Quantitative comparison between the proposed DVMDNet and 12 state-of-the-art methods from relevant fields. The best results are shown in bold.

single-image mirror detection CSFwinformer [29], video shadow detection method SCOTCH [14]). You can find more visual results in Section 1.1 of the supplementary material. The first row presents segmentation examples of small mirrors. In this example, only DVMDNet accurately segments the mirror region on the bathroom wall thanks to considering both depth and temporal information. DVMDNet can also accurately handle large mirrors (row 2) by considering depth information. We can see that PDNet and DVMDNet both exploit depth information not treat the iron ring as a mirror. Benefiting from their extracted textual features, CSFwinformer also detects the mirror correctly. For the multiple mirrors (row 3), only DVMDNet can accurately handle it by considering triple consistency and global content discontinuity over the RGB and depth. The example in the 4th row presents a challenging case with similar boundaries and appearance. Although the glass door on the cabinet appears similar to the oval makeup mirror on the table, DVMDNet can still accurately distinguish between them.

Method	IoU↑	F_β ↑	Accuracy↑	MAE↓
Baseline	0.6873	0.8258	0.9389	0.0589
Baseline + TC	0.7283	0.8478	0.9463	0.0526
Baseline + TD	0.7122	0.8508	0.9455	0.0533
DVMDNet *w/o* D	0.7024	0.8388	0.9452	0.0547
DVMDNet	**0.7423**	**0.8581**	**0.9490**	**0.0509**

Table 3. Ablation study results, trained and tested on the proposed DVMD-D. "Baseline" denotes our network without all proposed modules. "TC" is the triple consistency module. "TD" is the temporal discontinuity module.

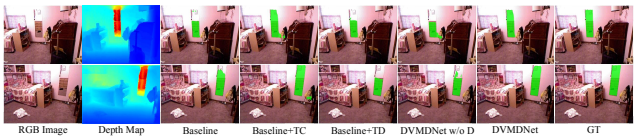


Figure 8. Visual ablation comparison of various DVMDNet variants.

We can see that although PDNet utilizes the depth cue, it still cannot detect it correctly without considering temporal information.

5.3. Ablation Study

We conducted an ablation study to validate the effectiveness of each key component in DVMDNet. Our findings are summarized in Tab. 3 and Fig. 8.

Benefits of Depth Cues. We want to explore the benefit of the depth cue for video mirror segmentation. To do this, we conduct following experiment: Retrain DVMDNet from scratch without including the depth branch and test it without depth information (Tab. 3 5th row and Fig. 8 6th column); Compared to the original DVMDNet (Tab. 3 5th row and Fig. 8 7th column), it can not achieve the same quality. Notably, DVMDNet without depth information (also included in Tab. 2) outperforms all relevant state-of-the-art methods.

Effectiveness of the Triple Consistency and Temporal Discontinuity Modules. Tab. 3 illustrates the effectiveness of each component in our model. As shown in the last row, our final proposed network, which includes the TC and TD modules, outperforms all other baselines across all metrics. It is noteworthy that the base model achieves competitive results, further demonstrating the critical importance of depth information for video mirror segmentation. We can see that adding the TC module achieves better results than "Baseline" indicating that it significantly benefits the mirror detection task from the prospect of triple consistency. Furthermore, adding the TD on "Baseline+TC" (i.g., "DVMDNet") further improves performance. Fig. 8 provides a visual example of the component analysis. We can see that the TC module helps the base model predict more mirror regions, and the TD module helps improve the performance of "Baseline+TC" by removing the overpredicted region.

6. Conclusion

In this paper, we proposed a method for detecting mirrors in RGB-D videos. The method includes two novel modules: the TC module and the TD module. Additionally, we constructed a challenging large-scale benchmark with diverse scenes. Our method is not without limitations. You can find a discussion in Section 2 of the supplementary material.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009. 7

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 7

[3] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *CVPR*, 2021. 7, 8

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 3

[5] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. 1

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 7

[7] Huankang Guan, Jiaying Lin, and Rynson W H Lau. Learning Semantic Associations for Mirror Detection. In *CVPR*, pages 5941–5950, 2022. 1, 7, 8

[8] Ruozhen He, Jiaying Lin, and Rynson WH Lau. Efficient mirror detection via multi-level heterogeneous learning. In *AAAI*, volume 37, pages 790–798, 2023. 1, 2, 7, 8

[9] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson WH Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection. In *AAAI*, volume 37, pages 935–943, 2023. 1, 7, 8

[10] Jiaying Lin and Rynson WH Lau. Self-supervised pre-training for mirror detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12227–12236, 2023. 1

[11] Jiaying Lin and Xin Tan. Learning to detect mirrors from videos via dual correspondences. In *CVPR*, 2023. 1, 2, 3, 4, 6, 7, 8

[12] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *CVPR*, 2020. 1, 2, 7, 8

[13] Fang Liu, Yuhao Liu, Jiaying Lin, Ke Xu, and Rynson WH Lau. Multi-view dynamic reflection prior for video glass surface detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3594–3602, 2024. 7, 8

[14] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Scotch and soda: A transformer video shadow detection framework. In *CVPR*, 2023. 7, 8

[15] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, 2017. 7

[16] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-Aware Mirror Segmentation. In *CVPR*, pages 3043–3052, 2021. 1, 2, 3, 6, 7, 8

[17] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, pages 3687–3696, 2020. 7, 8

[18] Daehee Park and Yong Hwa Park. Identifying Reflected Images from Object Detector in Indoor Environment Utilizing Depth Information. *IEEE Robotics and Automation Letters*, 6(2):635–642, 2021. 1, 2

[19] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7

[20] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. Springer, 2012. 3

[21] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[22] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson W H Lau. Mirror detection with the visual chirality cue. *TPAMI*, 2023. 1

[23] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022. 5

[24] Alex Warren, Ke Xu, Jiaying Lin, Gary KL Tam, and Rynson WH Lau. Effective video mirror detection with inconsistent motion cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17244–17252, 2024. 2

[25] Thomas Whelan, Michael Goesele, Steven J Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, Richard A Newcombe, M Goesele, et al. Reconstructing scenes with mirror and glass surfaces. *ACM Trans. Graph.*, 37(4):102–1, 2018. 1

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 7

[27] Jing Wu and Ze Ji. Seeing the unseen: Locating objects from reflections. *Lecture Notes in Computer Science*, 10965 LNAI:221–233, 2018. 1

[28] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 4

[29] Zhifeng Xie, Sen Wang, Qiucheng Yu, Xin Tan, and Yuan Xie. Csfwinformer: Cross-space-frequency window transformer for mirror detection. *IEEE Transactions on Image Processing*, 2024. 1, 8

[30] Ke Xu, Tsun Wai Siu, and Rynson WH Lau. Zoom: Learning video mirror detection with extremely-weak supervision.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6315–6323, 2024. 2

[31] Mingchen Xu, Jing Wu, Yukun Lai, and Ze Ji. Fusion of short-term and long-term attention for video mirror detection, 2024. 2, 3, 6

[32] 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. 1, 2, 6, 7, 8

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 7

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 7

[35] Wujie Zhou, Yuqi Cai, Xiena Dong, Fangfang Qiang, and Weiwei Qiu. Adrnet-s*: Asymmetric depth registration network via contrastive knowledge distillation for rgb-d mirror segmentation. *Information Fusion*, 108:102392, 2024. 1, 2

[36] Wujie Zhou, Yuqi Cai, and Fangfang Qiang. Morphology-guided network via knowledge distillation for rgb-d mirror segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 1, 2

Supplementary Material of RGB-D Video Mirror Detection

Mingchen Xu[1] **Peter Herbert**[1] **Yu-Kun Lai**[1] **Ze Ji**[2] **Jing Wu**[1]
[1]School of Computer Science and Informatics, Cardiff University
[2]School of Engineering, Cardiff University
{xum35, herbertp1, laiy4, jiz1, wuj11}@cardiff.ac.uk

1. Experiments

1.1. Comparison to SOTA Techniques

Fig. 2 provides a visual comparison of the results produced by our method with those obtained from two prior mirror segmentation methods (*i.e.*, PDNet [3] and VMDNet [1]) as well as the best approach from each of the three other categories (*i.e.*, glass detection method GDNet [4], single-image mirror detection CSFwinformer [5], video shadow detection method SCOTCH [2]). The first two rows present segmentation examples of small mirrors. In this example, only DVMDNet accurately segments the mirror region on the bathroom wall thanks to considering both depth and temporal information. DVMDNet can also accurately handle large mirrors (rows 3-4) by considering depth information. We can see that PDNet and DVMDNet both exploit depth information not treat the iron ring as a mirror. Benefiting from their extracted textual features, CSFwinformer also detects the mirror correctly. For the multiple mirrors (rows 5-6), only DVMDNet can accurately handle it by considering triple consistency and global content discontinuity over the RGB and depth. The example in the 7th and 8th rows presents a challenging case with similar boundaries and appearance. Although the glass door on the cabinet appears similar to the oval makeup mirror on the table, DVMDNet can still accurately distinguish between them. We can see that although PDNet utilizes the depth cue, it still can not detect it correctly without considering temporal information.

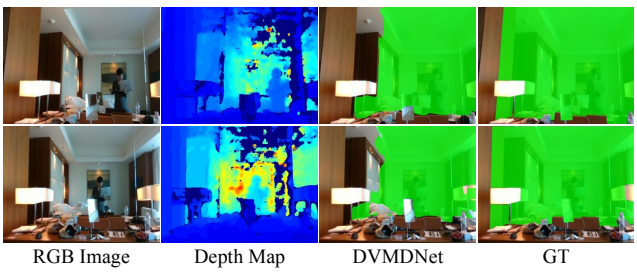


Figure 1. An example of a failure case is a mirror that covers almost the entire image.

2. Conclusion

Our method is not without limitations. Fig. 1 shows that our method fails when the mirror covers almost the entire image. In this scenario, the local consistency between the inside and outside of the mirror is difficult to extract, which leads to errors in global consistency based on local consistency. In addition, the content discontinuities between the inside and outside of the mirror over RGB and depth are also hard to quantify. It is a challenging situation even for human perception, and leaves an interesting direction for future research.

References

[1] Jiaying Lin and Xin Tan. Learning to detect mirrors from videos via dual correspondences. In *CVPR*, 2023. 1

[2] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Scotch and soda: A transformer video shadow detection framework. In *CVPR*, 2023. 1

[3] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-Aware Mirror Segmentation. In *CVPR*, pages 3043–3052, 2021. 1

[4] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, pages 3687–3696, 2020. 1

[5] Zhifeng Xie, Sen Wang, Qiucheng Yu, Xin Tan, and Yuan Xie. Csfwinformer: Cross-space-frequency window transformer for mirror detection. *IEEE Transactions on Image Processing*, 2024. 1

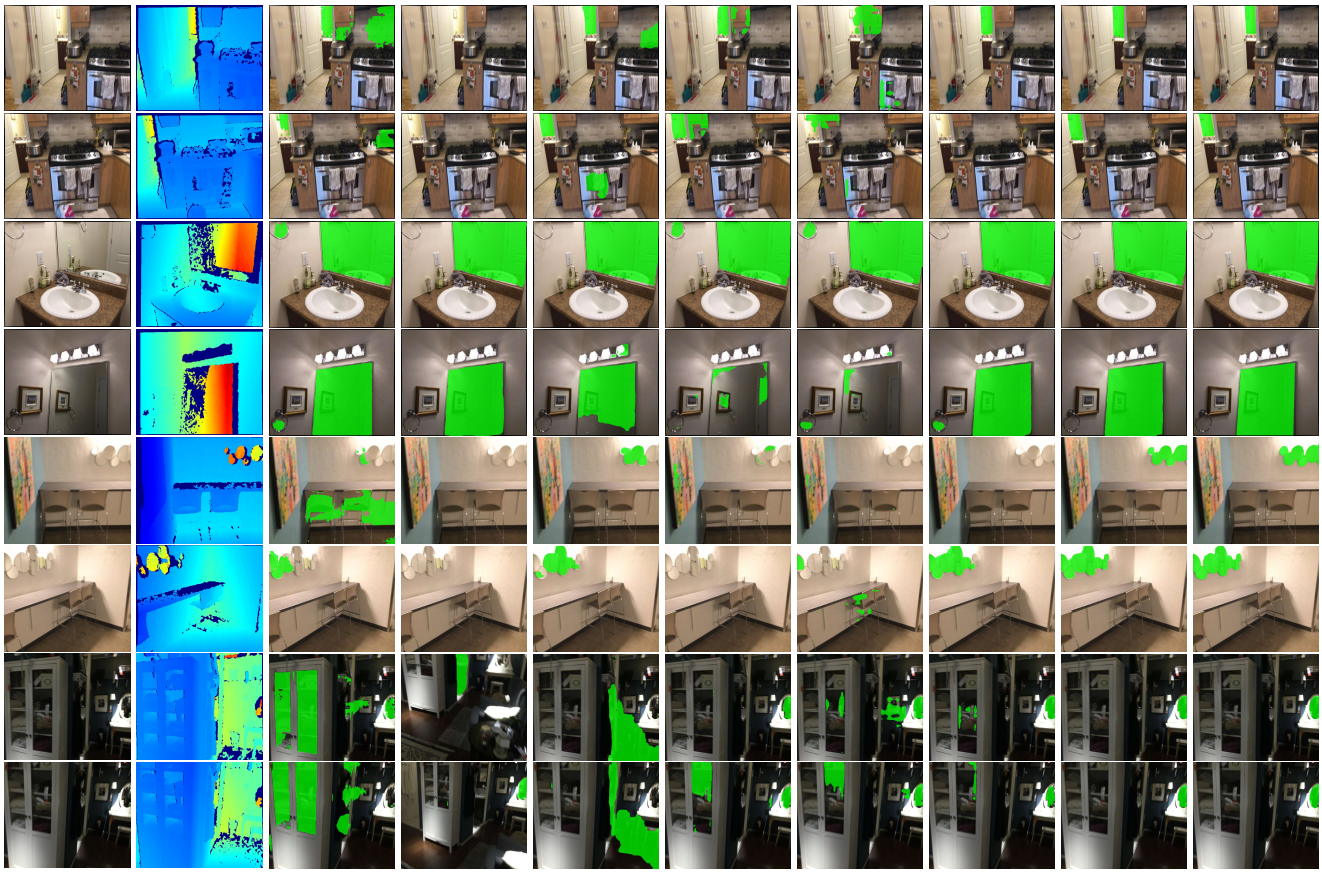


Figure 2. Visual comparison of DVMDNet with state-of-the-art segmentation methods retrained on the DVMD-D demonstrates that DVMDNet outperforms the competing methods on scenes with small mirrors (rows 1-2), large mirrors (rows 3-4), multiple mirrors (rows 5-6), and challenging scenes featuring similar boundaries and appearances (rows 7-8).