MVMD: A Multi-View Approach for Enhanced Mirror Detection

Yidan Shen*, Yu Wen*, Chen Zhang, Xin Fu, Renjie Hu
University of Houston
{yshen20, ywen8}@uh.edu, chenzh0220@outlook.com, {xfu8, rhu7}@central.uh.edu

Abstract

In 3D reconstruction, mirrors introduce significant challenges by creating distorted and fragmented spaces, resulting in inaccurate and unreliable 3D models. As 3D reconstruction typically relies on multi-view images to capture different perspectives of a scene, detecting and labeling mirrors in multi-view images before reconstruction can effectively address this issue. However, existing methods focus solely on single-image detection, overlooking the rich information provided by multi-view setups. To overcome this limitation, we propose MVMD, a novel Multi-View Mirror Detection method, along with the first database specifically designed for mirror detection in multi-view scenes.

The design of MVMD is grounded in the inherent associations between objects seen from different views and those reflected inside and outside of mirrors. These relationships are learned through cross- and self-attention mechanisms. MVMD consists of three key blocks: the Inter-Views Block tracks the shifts of objects within mirrors caused by changes in viewpoint; the Intra-View Block detects object reflections inside mirrors; and the Refinement Block sharpens mirror boundaries and enhances detected details.

Experimental results show that our method improves accuracy by up to 2.6% and IoU by up to 11.1%, compared to single-image mirror detection techniques. This substantial improvement makes MVMD particularly effective for computer vision tasks, especially in enhancing the accuracy of 3D reconstruction in mirror-dense environments. Code and data are available at: https://github.com/mvmdwacv25.

1. Introduction

Mirrors pose a major challenge to 3D scene reconstruction, as they mislead algorithms into interpreting reflections as real objects, resulting in incorrect depth estimates and a distorted spatial structure. This misinterpretation complicates the accurate reconstruction of a scene, with reflections often appearing as extensions of the real environment, leading to errors in object recognition and scene analysis. Traditional 3D reconstruction methods and state-of-the-art algorithms, including NeRF [19], 3DGS [10], and Multi-view Stereo Vision (MVS) models (e.g., COLMAP [20]), all struggle with mirror-related issues. Mirrors introduce problems such as phantom objects [18] and distorted geometries [18,28], making it challenging for these techniques to accurately differentiate between real objects and mirror reflections, resulting in poor reconstruction quality in mirror-rich areas.

Identifying mirrors before 3D reconstruction can significantly enhance reconstruction performance. However, current mirror detection algorithms [7, 14, 16, 17, 27, 30] are primarily designed for single-view scenarios, limiting their effectiveness in multi-view 3D reconstruction, where consistent mirror identification from different angles is crucial. Although video-based methods [13,22,25] incorporate temporal information, they often lack true multi-view perspectives due to limited camera angles, which restricts their applicability in capturing diverse spatial details required for accurate 3D mirror detection. Specifically, single-view and video methods lack the rich spatial information inherent in multi-view setups, hindering their ability to accurately detect mirrors and differentiate reflections from real objects. As a result, relying on these methods can lead to incomplete or distorted reconstructions [16]. While depth maps can sometimes improve single-view mirror detection by providing additional guidance [16, 30], they are costly to obtain and are often unavailable in many 3D reconstruction tasks. Furthermore, due to limitations in network architectures optimized for single inputs, most existing algorithms cannot effectively process multi-view stereo (MVS) images or video sequences with varying perspectives [13, 22, 25], despite their common use in 3D scene reconstruction. Therefore, designing a network that can handle multi-view inputs using only RGB images is highly desirable, as it aligns well with the data characteristics of 3D reconstruction and addresses the unique challenges of mirror detection in such scenarios.

However, developing a Multi-View Mirror Detection method using only RGB images presents several challenges.

*Co-first authors

First, there is a significant lack of multi-view datasets that include mirrors, which severely limits the ability to train models effectively. Second, different viewpoints result in varied visual appearances of a scene, especially in the presence of mirrors. As the viewpoint shifts, the scene inside the mirror changes at a different rate compared to the outside. This discrepancy makes it challenging for algorithms to distinguish between changes caused by actual viewpoint shifts and those caused by mirror reflections. This is because both can produce similar visual alterations across different views. This ambiguity complicates the detection process and may lead to inaccurate scene interpretations. Finally, objects such as windows and doors can exhibit depth discontinuities in multi-view inputs similar to those caused by mirrors, further complicating the detection process [30].

To address these challenges, we propose a mirror detection method designed for multi-view RGB image inputs without relying on explicit depth information. Our approach comprises three main components. The Inter-Views Block targets changes in different views caused by mirror reflections, distinguishing them from those due to actual viewpoint shifts. It employs cross-attention and self-attention mechanisms to capture reflection movements across views. The Intra-view Block focuses on objects with depth discontinuities by conducting cross-attention between the image and its mirror-flipped version, capturing the relationships between objects inside and outside mirrors. Lastly, the Refinement Block improves the final prediction using an edge-enhancement network. To train our network, we developed a new multi-view dataset featuring 98 scenes and a total of 3,181 images. This dataset is the first to include mirrors in multi-view scenes, offering a diverse representation of mirrors across various scenarios. It serves as a valuable resource for both training and evaluating mirror detection models. Our experiments demonstrate that our method significantly outperforms state-of-the-art techniques in terms of both efficiency and accuracy. To summarize, our key contributions are as follows:

- As no existing datasets meet the needs of Multi-View Mirror Detection, we have created a pioneering dataset designed to fulfill the requirements, which includes a wide range of scenes and provides an extensive collection of multi-view images.
- Unlike conventional methods that rely on single images or video inputs, we introduce a multi-input architecture that leverages three inputs to enhance mirror detection in multi-view applications.
- We implement an Inter-views Block that employs cross-attention and self-attention mechanisms to target changes caused by mirror reflections across different views, distinguishing these from those due to actual viewpoint shifts.
- To identify and correlate objects inside and outside a mirror, we design an Intra-view Block to address the challenge of reflection symmetry by leveraging the observation that mirror reflections are flipped versions of real objects.
- Extensive evaluations demonstrate that our model significantly improves IOU by up to 11.1% compared to existing state-of-the-art methods, particularly in complex scenes with diverse object geometries and occlusions caused by objects in front of mirrors.

2. Related Work

2.1. Semantic Segmentation

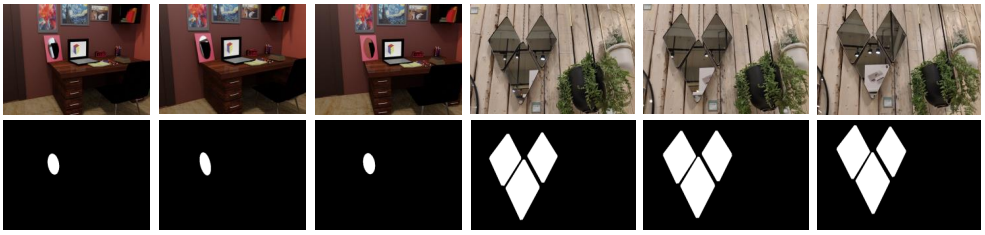
Semantic segmentation involves classifying each pixel in an image into a specific semantic category. Traditional methods for semantic segmentation have been applied to mirror detection tasks, aiming to accurately identify mirror regions versus other objects in the scene. Recent advancements [1, 2, 4, 6, 9, 11] often rely on single RGB images, which lack the comprehensive contextual understanding of the entire scene. Other methods incorporate depth information [3, 31] as an additional supervisory signal to enhance results; however, this does not apply to 3D reconstruction datasets, which often lack depth information. Moreover, many recent approaches utilize complex networks, such as fully convolution networks (FCNs) [15] and pyramid modules [1, 29], which greatly impact network efficiency. Despite these efforts, treating mirrors as an additional semantic category with existing semantic segmentation methods often yields unsatisfactory results because these methods struggle to accurately classify the complex and varied content reflected within the mirror.

2.2. Mirror detection

Mirror detection approaches can be broadly categorized into single-image and video-based methods, each addressing unique challenges and use cases. Single-image methods rely on static images for mirror detection, employing techniques such as edge detection and feature matching. In contrast, video-based methods utilize sequences of frames to leverage temporal information, incorporating techniques like optical flow and motion tracking.

2.2.1 Single Image-based Mirror Detection

Due to their simplicity, single-image-based methods have become popular in applications where computational efficiency is crucial. Yang et al. [27] first utilized contextual contrast cues in a single RGB image for mirror segmentation. Mei et al. [16] leveraged ground truth depth to predict mirrors in a single image but relied heavily on salient discontinuities in the depth map. Similarly, Zhou et al. [30]



(a) Illustrative examples in Multi-View Mirror Detection (MVMD) dataset.

(b) Spatial distribution of mirrors.

Figure 1. Visualization and analysis of the Multi-View Mirror Detection (MVMD) dataset.

employed a student model trained with distilled knowledge and multi-modal feature fusion from the single RGB image and its corresponding depth map, also depending on salient discontinuities. Lin et al. [13] detected mirrors by contrasting contextual relationships inside and outside their boundaries but required salient boundaries and corresponding objects. Guan et al. [7] first acquired class-specific knowledge and then used semantic associations between mirrors and their surrounding objects to locate mirrors in a single image. Tan et al. [21] embedded visual charity features into the detection model, depending on the characteristics of objects inside the mirror. He et al. [8] used a Multi-Level Heterogeneous Contrasted Module and a Reflection Semantic Logical Module for accurate mirror detection. In spite of the improvements, most existing methods use complex pyramid structures that are computationally intensive and impractical for real-time applications. Moreover, these methods often struggle to distinguish mirrors from other objects such as windows, doors, and art frames. It may also perform poorly with small mirrors due to insufficient information. In contrast, our proposed method utilizes three images as input, overcoming the limitations of single-image approaches and addressing these issues without using complex structures.

2.2.2 Video-based Mirror Detection

Single-image-based methods cannot fully address the challenges in dynamic and complex environments, leading to the development of video-based methods that leverage temporal information for consistency between frames. Lin et al. [13] use intra-frame and inter-frame correspondences from three video frames to predict mirror masks. However, their approach may fail when intra-frame correspondences are absent (e.g. when objects visible in the mirror are not present outside the mirror). Warren et al. [22] enhance predictions by leveraging motion inconsistencies in optical flow fields. However, their method faces challenges in specific scenarios. It struggles when mirror reflections move at speeds comparable to other objects. Additionally, scenes with depth discontinuities, such as windows and doors, also present issues due to similar motion inconsistencies. Xu et al. [25] introduce temporal consistency to learn from weak frame-level indicators for detecting mirrors in motion. Nonetheless, this method performs poorly in static mirror scenes, which hinders its application in 3D reconstruction.

3. MVMD Dataset

To advance the field of Multi-View Mirror Detection and support the training and testing of our proposed MVMD network, we present the MVMD (Multi-View Mirror Detection) dataset, filling a critical gap as the first dataset tailored for multi-view mirror detection. This comprehensive dataset comprises 98 diverse scenes and 3,181 images, specifically designed to address the challenges associated with mirror detection in multi-view environments.

Specifically, we create the MVMD dataset using Blender 4.0.1 [5], a photorealistic 3D creation suite, and incorporate selected real-world scenarios from VMD [13] and Mirror-NeRF [28] to ensure a diverse and realistic representation of scenes and mirrors suitable for real-world applications. We applied data augmentation techniques such as cropping and rotation to the original images to expand the dataset. Specifically, the MVMD dataset includes 1,559 images generated using Blender, with a resolution of 640×480, 1,192 images from the VMD-D dataset at a resolution of 1280×720, and 430 images from the Mirror-NeRF dataset, with resolutions of 800×800 and 400×300. Within the dataset, images from VMD-D and Mirror-NeRF were selected to ensure that inter-view angular separations exceed 0.8 degrees for adequate view variation and that each scene includes at least three multi-view images to satisfy MVMD's input requirements. This diverse dataset provides a solid foundation for developing and evaluating Multi-View Mirror Detection methods. Sample images from the MVMD dataset are shown in Fig. 1a. For additional dataset examples, please refer to Appendix A.3.

To capture real-world diversity in mirror detection, the dataset features mirrors positioned across various regions of the images, ranging from the center to the edges. The spatial distribution of mirrors is illustrated in Fig. 1b. Additionally, the dataset includes mirrors of different shapes and sizes, ranging from common forms like round, ellipsoid,

and rectangular to less common shapes such as irregular polygons and asymmetrical forms. This variety enhances the dataset's robustness, supporting comprehensive training and testing of models for real-world applications.

4. Methodology

4.1. Key Insights for Multi-View Mirror Detection

Understanding the core behavior of mirrors from varying viewpoints is crucial for designing an effective mirror detection network. This deep understanding reveals several key insights that guide the development of our proposed method, including dynamic reflections, object correspondence, and distinct edges. These insights directly inform the design of each component of our network, ensuring that our approach effectively addresses the unique challenges of mirror detection in multi-view settings.

First, as the viewpoint shifts, reflections in mirrors dynamically alter the perceived positions, orientations, and visibility of objects within the scene. Unlike objects outside the mirror, which maintain a consistent spatial relationship with the camera, reflections within the mirror behave unpredictably. The mirror's transformative effect can cause these reflections to flip and distort the scene in ways that differ from the direct view. This variability in reflections provides valuable information for detecting mirrors and guides the design of our Inter-Views Block, which is tailored to capture these dynamic differences as the viewpoint shifts.

Second, objects can appear both in actual space and as reflections on a mirror's surface, maintaining their physical properties and spatial relationships. This correspondence can be leveraged to identify the presence of mirrors. To utilize this, our Intra-View Block is specifically designed to identify and exploit these relationships, enhancing mirror detection based on object correspondences.

Finally, mirrors often exhibit distinct edges or frames that serve as clear indicators of their presence, providing additional cues for accurate detection. Our Refinement Block focuses on these features to improve detection accuracy by emphasizing the unique edges or frames of the mirrors.

4.2. Architecture Overview

Based on the discussion in Section 4.1, we construct the overall MVMD network, integrating the Inter-Views, Intra-View, and Refinement Blocks, as illustrated in Fig. 2.

Specifically, the MVMD network processes three input images, I_1, I_2, and I_3, captured from different viewpoints. The angle between the viewpoints of I_1 and I_2 is small, while that between I_1 and I_3 is larger. Each RGB image, with dimensions (H $\times$ W), is first fed into a pre-trained ResNeXt-101 backbone network [24] to extract multi-scale features. The high-level features (from the 5th scale of the ResNeXt network) of I_1, I_2, and I_3 are then processed by the Inter-Views Block to identify potential mirror locations by analyzing information from these different viewpoints. Subsequently, the Intra-View Block uses the low-level feature (from the 2nd scale of the ResNeXt network) Fl_1 from I_1 and the output of the Inter-Views Block, F_{Inter}, to identify possible mirror regions by examining the relationships between objects inside and outside the mirror. The outputs of both blocks, F_{Inter} and F_{Intra}, are combined and passed through a decoder to generate an initial mask P_i. Finally, this initial mask P_i, along with the low-level feature Fl_1, is processed through the Refinement Block. This block uses edge information to produce the final mask P_f. In the following sections, we will provide a detailed description of each block's functionality and how they contribute to the overall network architecture.

4.3. Inter-views Block

As illustrated in Fig. 2 (blue block), the Inter-Views Block is designed to capture and analyze differences in objects within the mirror area between two images as the viewpoint changes. It employs cross-attention mechanisms to detect these differences using high-level features, while self-attention provides a comprehensive understanding of potential mirror regions.

To detect mirror reflections across different views, cross-attention is applied between the feature sets $[Fh_1, Fh_2]$ and $[Fh_1, Fh_3]$. This dual-view approach increases the likelihood of identifying reflection changes, addressing cases where a single view may not capture significant variations. The cross-attention mechanism specifically targets changes due to mirror reflections rather than direct viewpoint shifts. Each cross-attention operation is followed by a self-attention step, which focuses on interpreting the entire image and distinguishing between mirror and non-mirror areas. The attention mechanism is formulated as:

$$F_{ATT} = \text{cat}\left(\mathcal{S}(\mathcal{C}\left(Fh_1, Fh_2\right)), \mathcal{S}(\mathcal{C}\left(Fh_1, Fh_3\right))\right), \quad (1)$$

where $\mathcal{C}(.,.)$ represents cross-attention, $\mathcal{S}(.)$ denotes self-attention, cat$(.,.)$ indicates concatenation along feature channels, and F_{ATT} is the resulting output.

To enhance feature representation, we use a channel attention mechanism inspired by [23]. This involves max and average pooling to obtain channel descriptors, which are processed through a shared MLP. The weighted features are then integrated using a convolution network with Batch Normalization and ReLU activation. The Inter-Views Block's process is summarized by:

$$F_{Inter} = \mathcal{CONV}\left(\mathcal{CA}\left(F_{ATT}\right)\right), \quad (2)$$

where $\mathcal{CA}$ denotes channel attention, $\mathcal{CONV}$ represents convolution, and F_{Inter} is the output of the Inter-View Block. The input and output of the Inter-View Block have dimensions $(H \times W)/16$ and 256 channels.

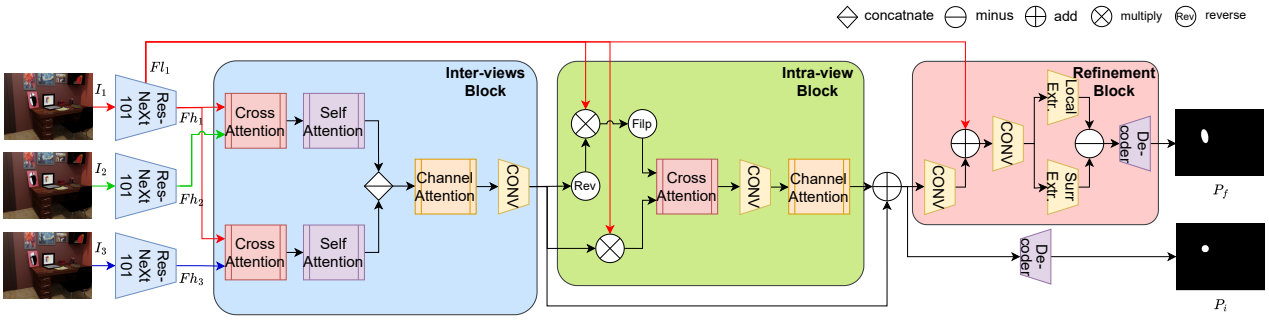


Figure 2. Overall architecture of the MVMD network. The network consists of three blocks: the Inter-Views Block, the Intra-View Block, and the Refinement Block. Information flow is depicted using colored arrows: red for I_1, green for I_2, and blue for I_3. The combined information flow is indicated by black arrows.

4.4. Intra-view Block

As illustrated in Fig. 2 (green block), the output from the Inter-Views Block is integrated with the target image information in the Intra-View Block to capture relationships between corresponding objects inside and outside the mirror. Unlike existing methods that rely only on flipped versions or mask guidance, our approach incorporates cross-attention mechanisms between the mirror and non-mirror areas of the target image, utilizing horizontally flipped versions for non-mirror regions. This design enables the network to learn correspondences accurately and efficiently.

In the Intra-View Block, the output from the Inter-Views Block, F_{Inter}, is multiplied by the target image's low-level feature, Fl_1, to isolate the mirror area:

$$F_{Mir} = Fl_1 \times F_{Inter}, \tag{3}$$

where F_{Mir} represents the feature map highlighting the mirror area. Conversely, to isolate the non-mirror area, F_{Inter} is subtracted from 1 and multiplied with Fl_1, followed by a horizontal flip:

$$F_{NoMir} = \mathcal{FLIP}(Fl_1 \times (1 - F_{Inter})), \tag{4}$$

where F_{NoMir} denotes the flipped feature map of the non-mirror area, and $\mathcal{FLIP}$ represents the horizontal flip operation.

These two feature maps, F_{Mir} and F_{NoMir}, are then input into a cross-attention operation to learn the association between objects inside and outside the mirror. The cross-attention ensures accurate correspondence by aligning both mirror and non-mirror features to the same orientation. Additionally, channel attention is employed to guide the network's focus toward crucial information. The process of the Intra-View Block can be summarized as follows:

$$F_{Intra} = \mathcal{CA}\left(\mathcal{CONV}(\mathcal{C}(F_{Mir}, F_{NoMir}))\right), \tag{5}$$

where $\mathcal{CA}$ denotes channel attention, $\mathcal{CONV}$ represents the convolution operation, and $\mathcal{C}$ stands for cross-attention. The input features Fl_1 and F_{Inter} have dimensions $(H \times W)/8$ and $(H \times W)/16$, respectively, with 256 channels. The output, F_{Intra}, retains dimensions $(H \times W)/16$ and 256 channels.

4.5. Refinement Block

By directly integrating the outputs from the Inter-Views and Intra-View Blocks, we generate an initial mirror mask that effectively covers the mirror areas. The process can be formulated as:

$$P_c = \mathcal{D}(F_{Intra} + F_{Inter}), \tag{6}$$

where $\mathcal{D}$ is the decoder that produces the initial mask P_i.

Although we have obtained an initial mask that effectively covers the mirror areas and aligns with the intended design, it still suffers from rough edges and noise, which limits its applicability in scenarios requiring high precision. To address these issues, we design the Refinement Block for edge enhancement and noise reduction. It processes the combined features from the Inter-Views Block (F_{Inter}) and the Intra-View Block (F_{Intra}), along with the low-level features Fl_1 of the target image, as follows:

$$F_f = \mathcal{CONV}(F_{Intra} + F_{Inter}) + Fl_1. \tag{7}$$

The edge-enhancement network applies two parallel convolution layers to the combined input F_f: one for local feature extraction and another for surrounding feature extraction. The local extractor focuses on fine-grained details within small regions using a 3×3 kernel and a dilation rate of 1. In contrast, the surrounding extractor captures broader contextual information using a 5×5 kernel and a dilation rate of 2. The outputs of these two extractors are then subtracted to isolate edges by emphasizing the differences between local features and broader context. A decoder with convolution layers processes the result to produce the final mask P_f. This process is mathematically formulated as:

$$P_f = \mathcal{D}\left(\mathcal{E}_{\mathcal{L}}(\mathcal{CONV}(F_f)) - \mathcal{E}_{\mathcal{S}}(\mathcal{CONV}(F_f))\right), \tag{8}$$

where $\mathcal{E}_{\mathcal{L}}$ denotes the local feature extractor, $\mathcal{E}_{\mathcal{S}}$ denotes the surrounding feature extractor, and $\mathcal{D}$ represents the decoder that produces the final binary mask P_f. The block takes feature maps with dimensions $(H \times W)/8$ and $(H \times W)/16$, each with 256 channels, and outputs a binary mask of size $H \times W$.

4.6. Loss Function

We use Mean Squared Error (MSE) as the loss function to train the MVMD network because it effectively measures the pixel-wise discrepancies between the ground truth masks G and the predicted masks P_i of mirrors. The loss function has two main components. The first component, $L(P_c, G)$, measures the error of the initial mask P_i and helps train the Inter-Views Block and Intra-View Block. The second component, $L(P_f, G)$, evaluates the error of the final mask P_f and is crucial for training the entire network.

To ensure the initial mask effectively guides the Inter-Views and Intra-View Blocks and that the final mask loss properly supervises the overall network performance, we assign a weight factor of 2 to the loss associated with the final mask. This weighting balances the contributions of both loss components while avoiding excessive penalization of the initial mask. Thus, the total loss is computed as:

$$L_{\text{Total}} = L(P_c, G) + 2 \times L(P_f, G), \tag{9}$$

where $L(.,.)$ represents the MSE calculation and G is the ground truth mirror mask for the target image I_1.

5. Experiment

5.1. Dataset and Implementation

From the proposed MVMD dataset, we use 2,798 images from 85 scenes (87%) for training and 383 images from 13 scenes (13%) for testing and validation. The dataset is split using random sampling to ensure that the model is evaluated in diverse environments, thereby simulating real-world performance. We apply the hold-out validation strategy to ensure an unbiased evaluation. For uniform processing, all images and their corresponding masks are resized to 640 $\times$ 480.

The MVMD network is implemented in PyTorch and trained on a cluster with an A100 GPU. We use ResNeXt-101 [24], pre-trained on ImageNet, as the backbone network to extract image features. The training utilizes the Adam optimizer with a batch size of 2, momentum of 0.9, weight decay of 5×10^{-4}, and a base learning rate of 0.0001. We apply cosine learning rate decay with a 3-epoch warm-up period to adjust the learning rate over 20 epochs.

5.2. Results Analysis

Detection Accuracy Given the absence of established methods for Multi-View Mirror Detection, we evaluate our approach by comparing it to six state-of-the-art techniques from related fields: VMD [13] for video mirror detection, MirrorNet [27], PMD [14], PDNet [16], and SANet [7] for single-image mirror detection, and GlassNet [12] for glass surface detection. All methods are re-trained on our dataset. For PDNet, which requires ground truth depth, we generate depth using the state-of-the-art single-image depth estimation method [26]. For single-image mirror detection, we treat each view as an independent image for prediction. As shown in Tab. 1, our method outperforms all competitors by a significant margin across all four metrics, with a notable 11.1% improvement in IOU and a 2.6% increase in accuracy, leveraging angle-diverse images for enhanced mirror detection.

Method	IOU	MAE	Accuracy	NMSE
VMD [13]	0.8808	0.0150	0.9850	0.1658
MirrorNet [27]	0.8794	0.0177	0.9823	0.1707
PDNet [16]	0.8181	0.0184	0.9816	0.2276
SANet [7]	0.7903	0.0362	0.9638	0.3469
PMD [14]	0.8823	0.0165	0.9835	0.1574
GlassNet [12]	0.8603	0.0163	0.9837	0.1688
Ours	**0.9019**	**0.0106**	**0.9894**	**0.1183**

Table 1. Comparison of state-of-the-art methods. The best results are highlighted in bold for reference.

Network Size Tab. 2 provides a quantitative assessment of MVMD compared to several state-of-the-art methods, which includes metrics such as parameter count, memory usage, and F_β score to illustrate the computational complexity and resource demands of each method. Our network, which relies heavily on attention mechanisms rather than extensive convolution layers, demonstrates a 35% improvement in both parameter efficiency and memory usage. Moreover, our robust architecture and multi-view image utilization contribute to a 5.9% enhancement in the F_β score, reflecting improved performance and scalability.

Visual Comparison Fig. 3 visually contrasts the performance of our method with selected state-of-the-art approaches from related fields. Our method consistently outperforms others in mirror detection. Existing techniques often misidentify mirror-like objects, such as wall frames or holes, and fail when objects are positioned directly in front of the mirror. They also struggle when the reflected scene closely resembles the surrounding environment. As illustrated in Fig. 3, MVMD uniquely identifies details that other methods miss: it correctly detects the white bulbs in the right corner of the mirror in the first image, the books in front of the mirror in the fifth image, and the small mirror on the right side of the shelf in the seventh image, due to the effectiveness of our Inter-View and Intra-View Blocks. Additionally, MVMD successfully identifies the small angled section in the right corner of the eighth image, a feature

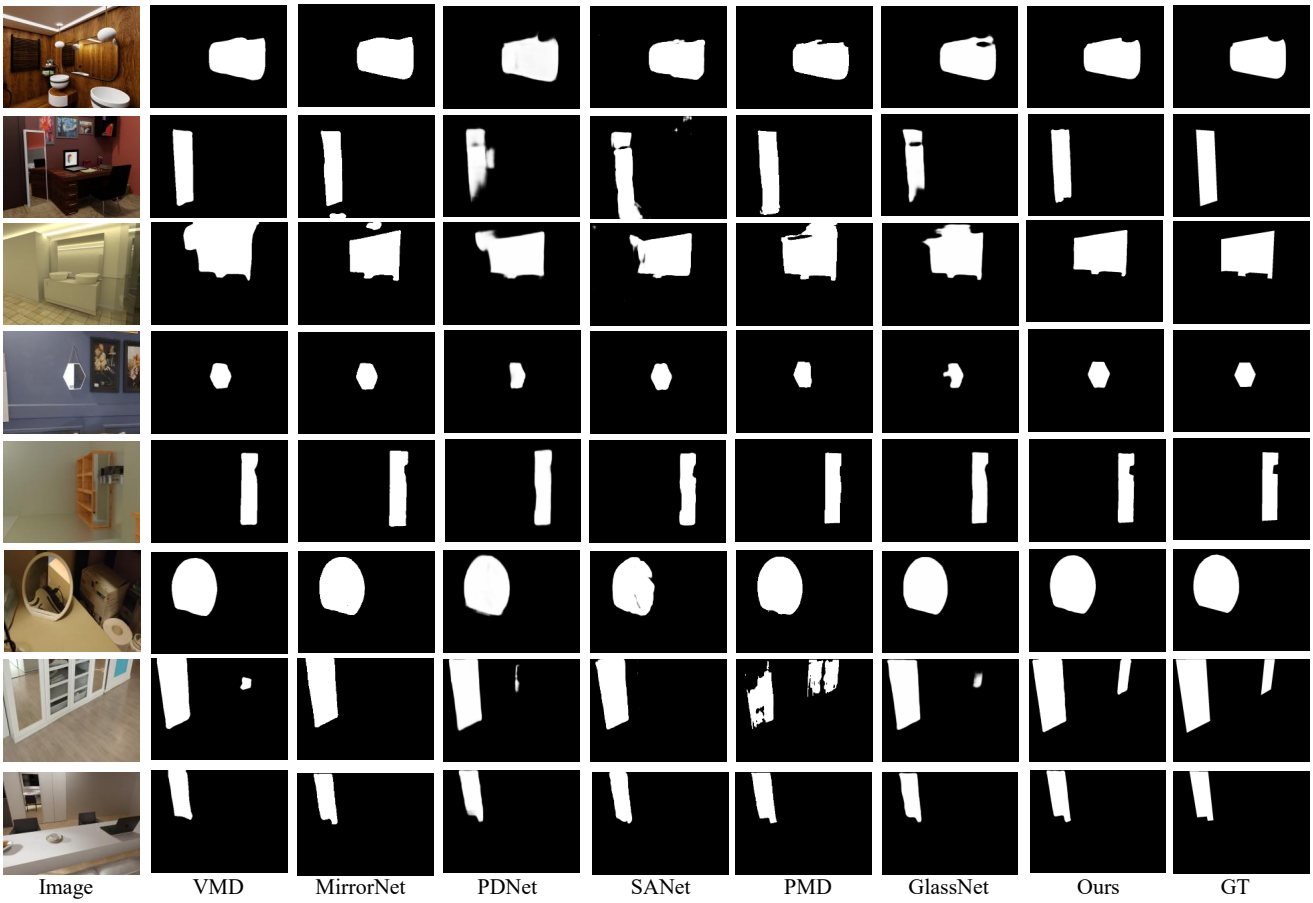


Figure 3. Visual comparisons of our method with state-of-the-art methods.

highlighted by the Refinement network. Additional results for the MVMD method can be found in Appendix A.2.

5.3. Ablation Study

To validate our MVMD model, we conducted a series of ablation experiments. For input variations, we tested the model with only two images, as well as with three images captured from the same scene in a random order. For structural variations, we considered the Inter-Views Block and the initial mask decoder as the baseline configuration. We then assessed the model's performance by removing either the Intra-View Block or the Refinement Block, resulting in configurations that included the baseline plus one of these additional blocks.

Effect of Suitable Inputs As illustrated in Tab. 3, using only two images often fails to capture significant changes in the mirror reflections, leading to less accurate predictions. Furthermore, employing three randomly selected images from the same scene may result in the network failing to recognize the mirror area if the chosen views are not representative. This issue arises because differing perspectives of the same object can make it challenging for the network to learn object correspondences across images. In contrast, selecting three strategically chosen views provides a more comprehensive perspective, enhancing prediction accuracy and minimizing the risk of missing critical information.

Impact of Individual Blocks As shown in Tab. 3, the Intra-view Block plays a crucial role in learning object correspondences within the image, significantly enhancing the network's ability to predict mirror regions. Additionally, the Refinement Block markedly improves the clarity of mirror edges, especially in scenarios where objects are positioned in front of the mirror. This demonstrates the Refinement Block's essential role in enhancing edge precision and detail, making it a critical component for achieving high accuracy in mirror detection tasks.

5.4. Attention visualization

To gain a deeper understanding of our MVMD module, we visualize the attention heat maps. For each attention map, we compute the maximum value across all channels for each pixel, where yellow indicates high attention and

	Parameter	Memory Usage	F_β
MirrorNet [27]	121.77M	464.50	0.9279
PDNet [16]	80.54M	307.24	0.8855
SANet [7]	103.07M	403.73	0.8659
PMD [14]	147.66M	563.28	0.9259
GlassNet [12]	201.72M	769.50	0.9190
ours	**71.68M**	**283.06**	**0.9442**

Table 2. Comparative analysis of computational complexity for MVMD. The best results are marked in bold for reference.

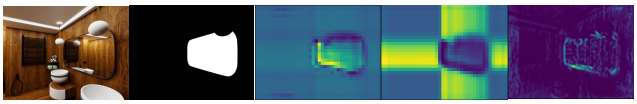


Figure 4. Visualization of attention heat maps. 'Inter Cross ATTN' represents the heat map for cross-attention in the Inter-Views Block, 'Inter Self ATTN' for self-attention in the same block, and 'Intra Cross ATTN' for cross-attention in the Intra-View Block.

dark blue indicates low attention.

We first examine the cross-attention heat map between I_1 and I_2, and the self-attention heat map that follows the cross-attention in the Inter-Views Block. As shown in Fig. 4, the cross-attention highlights the left side of the mirror, where there is a noticeable difference in objects. The self-attention then focuses on the areas outside the mirror, further enhancing the learned information. For the Intra-View Block, the cross-attention heat map emphasizes the light bulb in the left corner and the wash basin at the bottom left, which correspond to objects outside the mirror and the mirror edge. For additional visualizations of attention and intermediate results, please refer to Appendix A.1.

6. limitation and discussion

In cases where objects within the mirror, such as walls or ceilings, lack distinctive features, the network may struggle to differentiate between various camera poses. This challenge arises because the absence of distinctive characteristics limits the network's ability to detect variations between different views. Additionally, when reflections in the mirror closely resemble real-world objects outside the mirror, the network may struggle to distinguish between the mirror and its surroundings, as illustrated in Fig. 5.

Moreover, our method assumes that the images in each scene are ordered according to the sequence of camera movement, with the camera typically moving in a consistent direction. Consequently, the angular difference between the first and second images is usually smaller than that between the first and third images. If the images are not in the correct order, pre-processing is necessary to align them with our dataset's assumptions.

Method	IOU	MAE	Accuracy	NMSE
baseline	0.8597	0.0153	0.9847	0.1779
b + refine	0.8924	0.0127	0.9873	0.1364
b + intra	0.8826	0.0123	0.9877	0.1411
I_1 + I_2	0.9000	0.0108	0.9892	0.1208
rand 3 img	0.9000	0.0108	0.9892	0.1208
full	0.9019	0.0106	0.9894	0.1183

Table 3. Quantitative comparison of ablation study results. 'b' denotes the baseline configuration, while 'rand 3 img' refers to using three images captured from the same scene in random order.

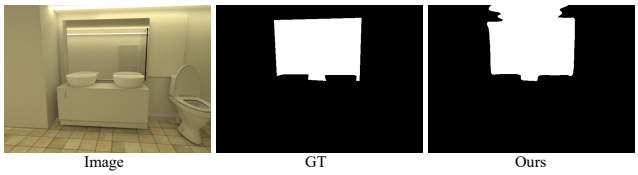


Figure 5. Example of failure cases due to similar reflections between mirror and surroundings.

7. Conclusion

In this paper, we introduce an innovative mirror detection method that utilizes multi-view images without requiring depth information. This technique is particularly effective for 3D reconstruction tasks and has the potential to enhance methods such as NeRF and 3D Gaussian Splatting. The MVMD network is designed with Inter-Views, Intra-View, and Refinement Blocks, each exploiting distinct characteristics of mirrors to enhance detection accuracy. Additionally, we provide a valuable Multi-View Mirror Detection (MVMD) dataset comprising 3,181 images from diverse scenes, which facilitates the training and evaluation of mirror detection models.

Our experimental results highlight the exceptional performance of MVMD, consistently surpassing state-of-the-art methods across various metrics. The method excels in distinguishing mirrors from similar objects and accurately detecting mirrors in challenging real-world conditions.

In summary, MVMD marks a significant advancement in Multi-View Mirror Detection, presenting a robust new tool for researchers and practitioners in computer vision and 3D reconstruction. By addressing the complex issue of mirror detection in multi-view scenarios, this work enhances the accuracy and reliability of 3D scene understanding.

8. Acknowledgment

This research is partially supported by NSF grants CCF-2130688, and CNS-2107057.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L Yuille, et al. Deeplab: Semantic image segmentation with deep convolutional nets. *Atrous convolution, and fully connected CRFs*, 40(4):834–848, 2017. 2

[2] 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 *European conference on computer vision*, pages 561–577. Springer, 2020. 2

[3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3029–3037, 2017. 2

[4] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2393–2402, 2018. 2

[5] Blender Foundation. Blender - a 3d modelling and rendering software. https://www.blender.org/. 3

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2

[7] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5941–5950, June 2022. 1, 3, 6, 8

[8] Ruozhen He, Jiaying Lin, and Rynson W. H. Lau. Efficient mirror detection via multi-level heterogeneous learning, 2022. 3

[9] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 2

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 1

[11] Zhuoying Li, Bohua Wan, Cong Mu, Ruzhang Zhao, Shushan Qiu, and Chao Yan. Ad-aligning: Emulating human-like generalization for cognitive domain adaptation in deep learning. In *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*, pages 794–798, 2024. 2

[12] Jiaying Lin, Zebang He, and Rynson W.H. Lau. Rich context aggregation with reflection prior for glass surface detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13410–13419, 2021. 6, 8

[13] Jiaying Lin, Xin Tan, and Rynson W.H. Lau. Learning to detect mirrors from videos via dual correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9109–9118, June 2023. 1, 3, 6

[14] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020. 1, 6, 8

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[16] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3044–3053, June 2021. 1, 2, 6, 8

[17] Haiyang Mei, Letian Yu, Ke Xu, Yang Wang, Xin Yang, Xiaopeng Wei, and Rynson W. H. Lau. Mirror segmentation via semantic-aware contextual contrasted feature learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19:1 – 22, 2022. 1

[18] Jiarui Meng, Haijie Li, Yanmin Wu, Qiankun Gao, Shuzhou Yang, Jian Zhang, and Siwei Ma. Mirror-3dgs: Incorporating mirror reflections into 3d gaussian splatting. *arXiv preprint arXiv:2404.01168*, 2024. 1

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[20] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[21] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson W.H. Lau. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3492–3504, 2023. 3

[22] Alex Warren, Ke Xu, Jiaying Lin, Gary K.L. Tam, and Rynson W.H. Lau. Effective video mirror detection with inconsistent motion cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17244–17252, June 2024. 1, 3

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018. 4

[24] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 4, 6

[25] 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. 1, 3

[26] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 6

[27] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 6, 8

[28] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4606–4615, 2023. 1, 3

[29] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[30] 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

[31] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing*, 16:677–687, 2022. 2