# Video Mirror Detection with the Motion-in-Depth Cue

**Alex Warren[1], Ke Xu[2], Xin Tian[3], Gary K.L. Tam[1], Benjamin W. Wah[4], Rynson W.H. Lau[2]**

[1]Department of Computer Science, Swansea University,
[2]City University of Hong Kong,
[3]Huawei Technologies, and
[4]The Chinese University of Hong Kong

## Abstract

Detecting mirror regions in RGB videos is essential for scene understanding in applications such as scene reconstruction and robotic navigation. Existing video mirror detectors typically rely on cues like inside-outside mirror correspondences and 2D motion inconsistencies. However, these methods often yield noisy or incomplete predictions when confronted with complex real-world video scenes, especially in areas with occlusion or limited visual features and motions. We observe that human perceive and navigate 3D occluded environments with remarkable ease, owing to Motion-in-Depth (MiD) perception. MiD integrates information from visual appearance (image colors and textures), the way objects move around us in 3D space (3D motions), and their relative distance from us (depth) to determine if something is approaching or receding and to support navigation. Motivated by this neuroscience mechanism, we introduce MiD-VMD, the first approach to explicitly model MiD for video mirror detection. MiD-VMD jointly utilizes contrastive 3D motion, depth, and image features through two novel modules based on a combinational QKV transformer architecture. The Motion-in-Depth Attention Learning (MiD-AL) module captures complementary relationships across these modalities with combinatorial attention and enforces a compact encoding to represent global 3D transformations, resulting in more accurate mirror detection and reduced motion artifacts. The Motion-in-Depth Boundary Detection (MiD-BD) module further sharpens mirror boundaries by leveraging cross-modal attention on 3D motion and depth features. Extensive experiments show that MiD-VMD outperforms current SOTAs.

**Code** —
https://github.com/AlexAnthonyWarren1/MiDVMD

## Introduction

Mirrors are abundant in the real world. Successful detection of mirror regions is crucial for reducing errors in many downstream vision tasks such as scene parsing (Huang et al. 2019), 3D reconstruction (Guo et al. 2022; Zeng et al. 2023), and autonomous navigation (Pal, Mondal, and Christensen 2020). However, detecting mirror regions is challenging because mirrors lack a consistent visual appearance and instead reflect their surrounding scenes.



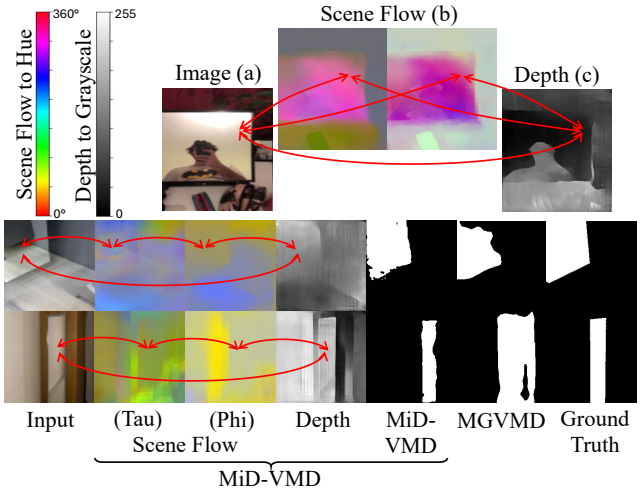


Figure 1: The SOTA video mirror detection MGVMD (Warren et al. 2024) primary models 2D motion, which fails in featureless or weak-motion regions (lower example, 6th column). The upper example illustrates three complementary cues: RGB appearance (a), contrastive 3D motion (b), and depth (c). These complementary cues, shown across columns 1–4 in the lower example, correlate strongly inside and outside mirrors and provide greater robustness than 2D motion alone. This observation aligns with human MiD perception, motivating our MiD-inspired approach.

Yang *et al.* (2019) propose the first image-based mirror detection method by modeling contrasting image features between mirror and non-mirror regions. Two subsequent methods (Mei et al. 2021; Tan et al. 2021) incorporate depth sensors to model contrasted features. Other approaches model appearance correspondence (Lin, Wang, and Lau 2020), semantic correlations (Guan, Lin, and Lau 2022), visual chirality (Tan et al. 2023), and coarse symmetry (Huang et al. 2023) between mirror and non-mirror regions. All these methods, however, address mirror detection only within single RGB or RGB-D images.

Lin *et al.* (2023) propose VMD-Net, the first video mirror detection method, modeling appearance correspondences within and across frames. Yet, such correspondences can be

unreliable: when two similar objects/regions both appear either inside or outside a mirror, VMD-Net may misclassify one as inside and the other outside (Tan et al. 2023). Recently, Warren *et al.* (2024) propose MGVMD, which models inconsistent motion inside and outside mirror regions using 2D optical flow. Although interesting, this method may be prone to noise in flow fields, especially in occluded or complex scenes, leading to over- or under-predictions. It also struggles with featureless regions common in mirror reflections from plain surfaces like walls, floors, and furniture.

In this work we make three key observations motivating our approach. First, humans perceive the world in 3D and handle noisy occluded environments using a well established neuroscience cue called Motion-in-Depth (MiD) (Kim, Angelaki, and DeAngelis 2015; Baker and Bair 2016; Uka and DeAngelis 2006). MiD helps us sense our surroundings and navigate by combining several types of information: 1) what we see with our eyes such as colors textures and shapes (image features); 2) how objects move through space around us (3D motion); and 3) how far away things are (depth). By integrating these clues, the brain robustly understands motion in depth and judges distances even in complex scenes, especially those with mirror reflections and challenging occlusions. Second, although depth and motion features can be high dimensional and noisy, 3D motions in mirror scenes can often be simplified into two low dimensional transformations—one inside and one outside mirror regions—helping reduce noise. Third, as shown in Fig.1 (top row), image depth and motion features complement each other strongly, providing reliable cues for mirror localization. These insights guide us to leverage the Motion-in-Depth cue for robust video mirror detection.

To this end, we propose MiD-VMD, a novel approach to leverage the Motion-in-Depth cue by jointly exploring contrastive 3D motion, depth, and image feature spaces. Our framework incorporates two key modules: Motion-in-Depth Attention Learning (MiD-AL) and Motion-in-Depth Boundary Detection (MiD-BD). MiD-AL leverages the complementary relationships among image, depth, and 3D motion features by mutually and combinatorially attending to all three modalities, promoting comprehensive and redundant information while enforcing low-dimensional embeddings to reduce noise. MiD-BD exploits contrastive depth and 3D motion cues to guide mirror boundary detection and ensure accurate delineation, especially where depth or motion signals are weak.

To our knowledge, this is the first study to explicitly model the Motion-in-Depth, unlike prior works that rely on motion alone, depth alone, or a simple concatenation fusion of both (which we term Motion-and-Depth, M&D) as auxiliary features (Mou et al. 2024). Our method captures correlations among depth-conditioned 3D motion, depth, and image features, leading to significant improvements in mirror detection. Figure 1(a)-(c) illustrates these relationships, with the bottom examples showing MiD-based detection. Our approach (5th column) remains robust even in featureless regions that challenge humans. Importantly, MiD-VMD uses only RGB inputs without explicit depth sensors, further distinguishing it from prior work. Our contributions are:

- We propose the first work to explicitly model the Motion-in-Depth cue and leverage it for video mirror detection through a novel framework, MiD-VMD, which exploits the complementary and contrastive information among depth, 3D motion, and image features.
- Our framework introduces two novel modules: *MiD-AL*, which attentively models correlations among image, depth, and low-dimensional 3D motion features to locate mirror regions; and *MiD-BD*, which uses contrastive depth and 3D motion cues at mirror boundaries to guide accurate boundary learning.
- Extensive experiments demonstrate that MiD-VMD outperforms state-of-the-art video mirror detection methods, is efficient and flexible (requiring only RGB inputs), and is robust across different depth estimation methods.

## Related Work

**Image-based Mirror Detection.** Yang *et al.* (2019) propose the first image-based mirror detection method, based on contextual contrasting features, and the first benchmark dataset for training/evaluation. Several methods have been subsequently proposed for this task. Lin *et al.* (2020) propose a mirror detection method based on detecting appearance correspondences between inside and outside mirror regions. Mei *et al.*(2021) extend the approach in (Yang et al. 2019) by leveraging RGB-D input data for mirror prediction and introducing an RGB-D image dataset for image-based mirror detection. Tan *et al.* (2021) propose to detect 3D mirror planes using a mask RCNN (He et al. 2017) and a mirror normal prediction network. Guan *et al.* (2022) consider that mirrors are typically placed in correlation with certain types of objects, and propose to learn semantic correlation as a cue for mirror detection. Tan *et al.*(2023) propose a mirror detection method based on detecting visual chirality at the pixel level. Huang *et al.* (2023) propose to model the coarse symmetry property of an object and its reflection in the mirror for mirror detection. He *et al.* (2023) propose to detect mirror regions based on the intensity-based low-level and semantics-based high-level features. Xing *et al.* (2025) propose a semi-supervised mirror detection framework with an iterative data engine and dual-scoring approach.

While these image-based methods exhibit good performance, they are not well-suited for our video mirror detection task, as they focus on mirror detection in singular frames and do not consistently perform well.

**Video Mirror Detection (VMD).** Recently, Lin *et al.* (2023) propose the first VMD method, VMD-Net, and the VMD-D dataset. VMD-Net utilizes dual correspondences to correlate objects within and across frames for VMD. Despite the success, object correspondences may not always be found between inside and outside of the mirror. Xu *et al.* (2024) propose a weakly-supervised method to model the feature similarity and contrast in temporal variations, but their method may not always be reliable due to their extremely weak supervision of per-frame zero-one mirror indicators in videos. Warren *et al.* (2024) introduce a VMD method, MGVMD, using 2D optical flow vector fields to detect motion inconsistencies in and around mirror regions. However, this ap-

proach is sensitive to noise from high-dimensional motion. In addition, the reliance on optical flow and motion inconsistencies makes MGVMD less effective when image features are weak or video motions are small.

In this work, we propose a novel approach to directly learn contrasting correlations in depth and 3D motion for VMD. Our approach is flexible (not requiring RGBD inputs), and shown to be robust to different depth estimators.

**Video Salient Object Detection (VSOD)**. VSOD involves identifying prominent objects in videos. Deep-learning based VSOD methods (Jun Wei 2020; Fan et al. 2019; Ji et al. 2021; Li et al. 2019; Liu et al. 2022; Tang, Li, and Xing 2021; Zhao et al. 2024; Zhang et al. 2021; Gu et al. 2020; Wang, Shen, and Shao 2017; Xu et al. 2021) have achieved promising performances by exploiting the strong representation capacity of neural networks, *e.g.*, modelling long-term temporal (Liu et al. 2022), dynamic context (Zhang et al. 2021) multi-level (Gu et al. 2020) features, and preserving spatial continuity with a Mamba-based (He et al. 2025) framework. Despite their success, it is important to note that VSOD methods are not designed or optimized for the VMD task, as mirror regions are not always the most prominent salient objects within a scene and lack distinct visual features of their own. Our experiment shows that our proposed method performs better than these VSOD methods.

**Motion, Depth, Motion-and-Depth (M&D), and Motion-in-Depth (MiD).** Multi-modal features like motion and depth are widely used independently in tasks such as mirror detection (Yang et al. 2019; Warren et al. 2024), 3D reconstruction (Ju et al. 2023), VSOD (Li et al. 2019), and camouflage object detection (Cheng et al. 2022). We term Motion-and-Depth (M&D) the simple concatenation fusion of motion and depth as separate modalities; for example, DCT-Net+ (Mou et al. 2024) uses 2D optical flow and depth maps alongside RGB features. In contrast, MiD—which involves motion in a 3D context, especially along the z-axis—has been studied extensively in neuroscience (Kim, Angelaki, and DeAngelis 2015; Baker and Bair 2016; Uka and DeAngelis 2006). We explicitly model and capture MiD by integrating motion, depth, and image features using a combinatorial Query-Key-Value (QKV) transformer structure. This learns complex cross-modal interactions to directly encode 3D motion cues. We find that a compact MiD encoding strongly contrasts mirror and non-mirror regions. To our knowledge, this is the first work to explicitly model and leverage MiD for improved mirror detection.

## Methodology

MiD-VMD leverages the Motion in Depth (MiD) cue (Kim, Angelaki, and DeAngelis 2015; Baker and Bair 2016; Uka and DeAngelis 2006) for video mirror detection. Human brains perceive their surroundings by combining: 1) what we see (image features); 2) how objects move in 3D space (3D motion); and 3) how far away objects (depth). Integrating these cues enables robust scene understanding even in complex scenes. In mirror detection, this integration is key because mirrors often cause visual ambiguities. Reflections may appear at conflicting depths or show motion patterns inconsistent with the surrounding scene. MiD-VMD leverages MiD for complementary image, motion, and depth cues to resolve ambiguous depth and motion across mirror boundaries, and accurately localize mirrors even in reflected, cluttered or occluded environments. It outperforms existing methods that rely on appearance, 2D motion, or depth alone.

**Overview.** Figure 2 provides an overview of MiD-VMD. It processes three adjacent input images ($I_{N-2}$, $I_{N-1}$, and $I_N$) to estimate depth maps and perform 3D scene flow estimation. Depth maps $D_{N-1}$ and $D_N$ are estimated from the image pairs ($I_{N-2}$, $I_{N-1}$) and ($I_{N-1}$, $I_N$), respectively, using depth estimator like (Lipson, Teed, and Deng 2021). (MiD-VMD is robust to various depth estimators.) These depth maps, along with images $I_{N-1}$ and $I_N$, condition a frozen RAFT-3D (Teed and Deng 2021) model to generate 3D scene flow features, denoted as $Feat_{SF}$. A shared ResNext-101 (Xie et al. 2017) backbone extracts multi-scale image features, $Feat_{N-1}$ and $Feat_N$, from the inputs.

To fully utilize the MiD cue for mirror detection, two novel modules are introduced, both of which take depth-conditioned 3D motion, depth features, and image features as input. The **Motion-in-Depth Attention Learning (MiD-AL)** module learns the complementary correlations between regions inside and outside the mirror by combining these features, with the help of a low-dimensional motion embedding. The **Motion-in-Depth Boundary Detection (MiD-BD)** module uses scene flow and depth features to guide mirror boundary predictions, particularly in challenging areas such as featureless or motionless regions. Finally, the Fusion Refinement Module integrates the outputs of MiD-AL, MiD-BD, and multi-scale image features from the ResNext-101 backbone. It refines the final mirror map predictions, $Pred_{N-1}$ for the previous frame and $Pred_N$ for the current frame, improving the accuracy of video mirror detection.

### The MiD-AL Module

Existing methods rely on depth features from sensors (Mei et al. 2021) or 2D motion features (Warren et al. 2024) to guide the image features, but these can become noisy in low-motion or low-contrast scenes, leading to inaccurate mirror detection. On the other hand, Time-of-Flight (ToF) depth inputs are costly, and standalone depth estimators can be unstable. In this work, we explore the Motion-in-Depth (MiD) cue by examining the relationship between 3D motion and depth features. We first observe a strong complementary correlation between Scene Flow (3D motion), Depth Estimation, and Image Features for mirror localization, which aligns with MiD. Additionally, we observe that 3D motion inside and outside the mirror can be captured by two affine transformations, suggesting that low-dimensional encoding can further reduce noise from scene flow estimators. Depth features further stabilize Scene Flow, enhancing scene understanding through the z-axis (MiD). To this end, we propose the Motion-in-Depth Attention Learning (MiD-AL) module to locate mirror regions by modeling contrastive correlations among 3D motion, depth, and image features.

Figure 3 shows the structure of the MiD-AL module, which takes as input the depth features $Depth_C$ (which are

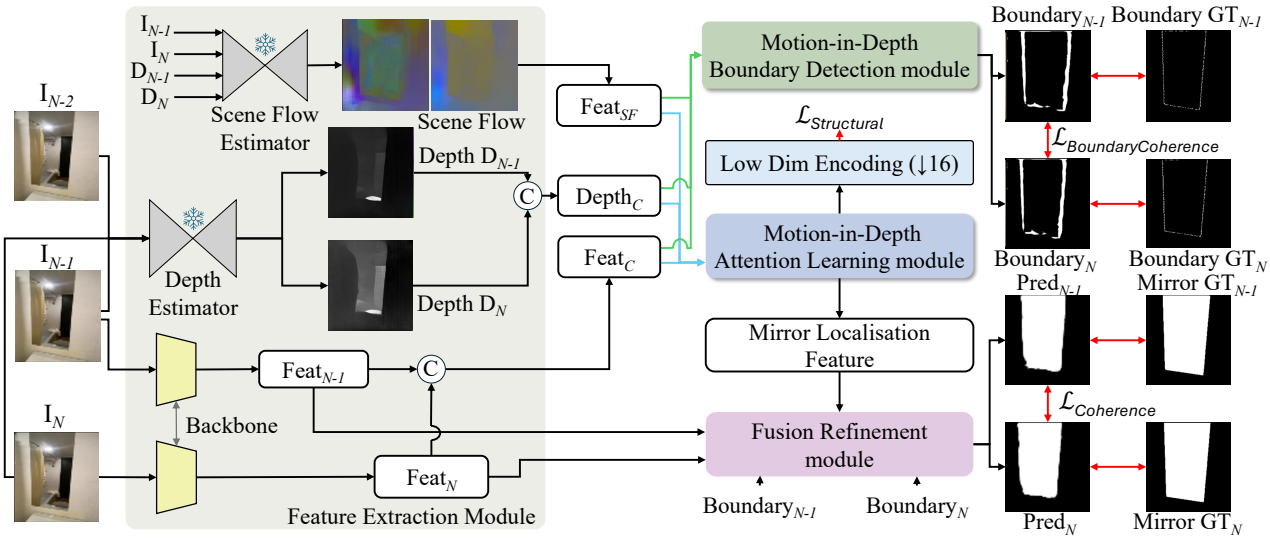


Figure 2: Overview of our MiD-VMD: Given three consecutive input images ($I_{N-2}$, $I_{N-1}$, $I_N$), the Feature Extraction Module—comprising a stereoscopic depth estimator, a scene flow estimator, and an image encoder—produces two depth maps ($D_{N-1}$ and $D_N$), scene flow features ($Feat_{SF}$), and multi-scale image features ($Feat_{N-1}$ and $Feat_N$). The depth maps ($D_{N-1}$ and $D_N$) are then fusion concatenated to ($Depth_C$). The multi-scale image features ($Feat_{N-1}$ and $Feat_N$) are then fusion concatenated to ($Feat_C$). The fusion concatenated depth feature ($Depth_C$), scene flow features ($Feat_{SF}$), and fusion concatenated multi-scale image features ($Feat_C$) are then processed by our Motion-in-Depth Attention Learning module (MiD-AL) to learn correlations in contrastive regions and predict mirror localization features by attentively modeling the Motion-in-Depth cue. Meanwhile, the fusion concatenated multi-scale image features ($Feat_C$), scene flow features ($Feat_{SF}$), and fusion concatenated depth map ($Depth_C$) are fed into our Motion-in-Depth Boundary Detection module (MiD-BD), which extracts and cross-guides mirror boundaries by exploiting depth discrepancies and contrasts within 3D motion at mirror boundaries. Finally, the Fusion Refinement Module combines the mirror localization features from MiD-AL, the Motion-in-Depth guided mirror boundary maps ($Boundary_{N-1}$ and $Boundary_N$) from MiD-BD, and the multi-scale image features ($Feat_{N-1}$ and $Feat_N$) to predict the final mirror maps ($Pred_{N-1}$ and $Pred_N$), respectively.

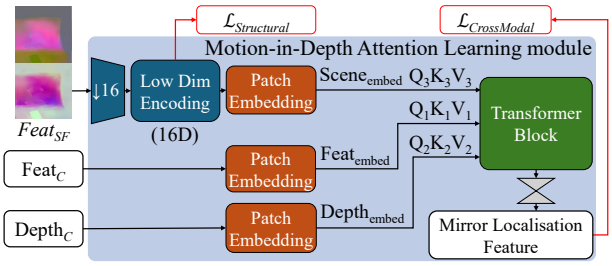


Figure 3: MiD-AL module models correlated contrastive regions in scene flow, depth, and image features to identify mirror locations. Operating in a 16-D space (↓ for dimension reduction), it learns the global 3D affine transformations inside and outside mirrors, reducing noisy predictions.

concatenated depth features from $Depth_{N-1}$ and $Depth_N$), the multi-scale image features $Feat_C$ (which are concatenated image features from $Feat_{N-1}$ and $Feat_N$), and the scene flow features $Feat_{SF}$. To reduce noise in $Feat_{SF}$, we first encode $Feat_{SF}$ into a lower-dimensional space (16-D), as motions inside and outside mirrors are typically described by two 3D global affine transformations. This encoding is crucial for reducing noise from high-dimensional movements, occlusions, and low-contrast regions. The ablation of encoding dimension (Table 3) further supports this. We then apply patch embedding with positional information to each set of input features, creating low-dimensional, learnable representations for 3D scene flow ($Embed_{scene}$), depth ($Embed_{depth}$), and multi-scale image features ($Embed_{image}$). To capture these complementary correlations, we assign $Embed_{scene}$ as $Q_3K_3V_3$, $Embed_{depth}$ as $Q_2K_2V_2$, and $Embed_{image}$ as $Q_1K_1V_1$. This configuration leverages the unique strengths of each set of features: sensitivity to motion for 3D scene flow features, structural stability for depth features, and the multi-scale context for image features. Finally, we use self-attention within a transformer block, leveraging the following combinations of queries (Q) and keys and values (K, V): [($Q_1K_2V_2$), ($Q_1K_3V_3$), ($Q_2K_1V_1$), ($Q_2K_3V_3$), ($Q_3K_1V_1$), ($Q_3K_2V_2$)], to capture a correlated feature space across all three input modalities. The output from this mechanism is subsequently passed through a convolutional layer, producing the mirror localization prediction. This is critical for our Fusion Refinement Module, which enhances the final mirror prediction.

## The MiD-BD Module

As highlighted, depth and 3D motion provide complementary strengths: while depth offers stable priors in low-motion

scenes, 3D motion compensates when depth estimates are unstable, as captured by Motion-in-Depth (MiD) cues for mirror localization. However, occasional weak signals from depth and motion estimators may still introduce inconsistencies along mirror boundaries. These observations motivate the design of our Motion-in-Depth Boundary Detection (MiD-BD) module, which detects mirror boundaries using estimated depth and 3D motion features. To address potential weak signals, we use a feature-guidance approach instead of relying solely on mutual feature attention. We cross-guide scene flow and depth with mirror boundary features from both low-level (textures) and high-level (semantics) image features. This enhances mirror boundary detection.

Figure 4 shows the structure of the MiD-BD module, which takes three inputs: $Feat_C$ (concatenated image features $Feat_{N-1}$ and $Feat_N$), $Feat_{SF}$ (Scene Flow 3D motion), and $Depth_C$ (concatenated depth features $Depth_{N-1}$ and $Depth_N$). We first apply a convolutional boundary feature extraction network to extract boundary features ($Feat_{Boundary}$) from $Feat_C$, ensuring that fine-grained details and contextual information are captured effectively. We then fuse the input depth features $Depth_C$ and scene flow features $Feat_{SF}$ to obtain $Feat_{DepthSF}$. Subsequently, we take advantage of the dynamic weighting capacity of the cross-attention mechanism to guide the extracted boundary features $Feat_{Boundary}$ with contrastive features found within $Feat_{DepthSF}$ ($Depth_C$ and $Feat_{SF}$). We compute attention weights (Beta values) that direct the focus onto relevant $Feat_{DepthSF}$ features, which facilitate the refinement of boundary features and align them more closely with the abrupt depth changes and 3D motion changes observed along mirror boundaries. Finally, the refined Motion-in-Depth boundary features are processed via a convolutional layer to produce the Motion-in-Depth guided mirror boundary maps $Boundary_{N-1}$ and $Boundary_N$. In this way, our module captures and accentuates the sharp boundaries indicative of mirror regions.

Unlike existing methods that use separate edge detection or apply edge preservation losses, our approach leverages 3D motion and depth features to guide mirror boundary learning. By combining temporal features from $Feat_{N-1}$ and $Feat_N$ in a single module call, we reduce memory usage and improve runtime, distinguishing our method from techniques like MGVMD (Warren et al. 2024), which rely on sequential calls. Further, our lightweight MiD-BD focuses on mirror boundaries.

## Fusion Refinement Module

We extend the refinement module in (Lin, Wang, and Lau 2020) to handle multiple input modalities, enhancing the accuracy of mirror predictions $Pred_{N-1}$ and $Pred_N$. Specifically, the Fusion Refinement Module combines $Feat_{N-1}$ and $Boundary_{N-1}$ from MiD-BD with mirror localization features from MiD-AL to predict the mirror map $Pred_{N-1}$. Similarly, it integrates $Feat_N$ and $Boundary_N$ from MiD-BD with MiD-AL features to predict $Pred_N$. Each combined representation is then processed through convolutional layers for per-pixel binary classification.

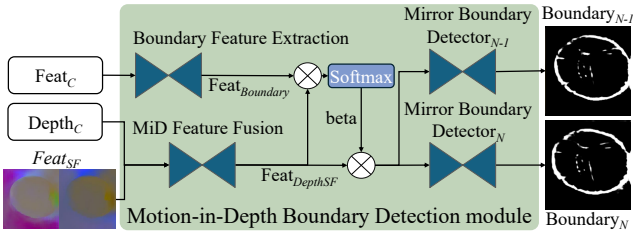


Figure 4: Our MiD-BD module predicts mirror boundaries from depth discrepancies inside and outside mirror regions. $\otimes$ denotes matrix multiplication. It is lightweight and provides boundary cues in featureless mirror regions.

## Loss Function

We train our model with the following loss function:

$$\mathcal{L}_{\text{Mirror}} = \begin{array}{l}(\alpha \cdot \mathcal{L}_{\text{Map}}) + \mathcal{L}_{\text{CrossModal}} + \mathcal{L}_{\text{Boundary}} + \\ \mathcal{L}_{\text{Structural}} + \mathcal{L}_{\text{BoundaryCoherence}} + \mathcal{L}_{\text{Coherence}}\end{array} \quad (1)$$

**Mirror Loss** ($\mathcal{L}_{\text{Map}}$) is a BCE loss that measures the accuracy between the predicted mirror masks $Pred_{N-1}$ and $Pred_N$ and the ground truth mirror masks. We empirically set $\alpha$ to 3 to emphasize the accuracy in detecting mirrors in our video mirror detection task.

**Depth and 3D Motion Prediction Loss** ($\mathcal{L}_{\text{CrossModal}}$) is a BCE loss between the predicted mirror localisation map from the MiD-AL and the ground truth mirror map.

**Mirror Boundary Loss** ($\mathcal{L}_{\text{Boundary}}$) is a BCE loss computed between the detected mirror boundaries by the MiD-BD and the ground truth mirror boundaries.

**Scene Flow Structural Constraint** ($\mathcal{L}_{\text{Structural}}$). We use a covariance loss (Liu, Hu, and Salzmann 2023) to constrain structural coherence and reduce noisy motion in the low-dimensional encoded scene flow. It also facilitates the capture of more meaningful contrastive cues, especially from mirror regions.

**Coherence Loss** ($\mathcal{L}_{\text{Coherence}}$). We use a BCE loss to temporally regulate the mirror predictions $Pred_N$ and $Pred_{N-1}$ to drive temporal consistency in video mirror detection.

**Coherence Boundary Loss** ($\mathcal{L}_{\text{BoundaryCoherence}}$). We use a BCE loss to temporally regulate the mirror boundary predictions $Boundary_{N-1}$ and $Boundary_N$ to drive temporal consistency in video mirror boundary detection.

Ablation studies on loss weighting and full loss analysis can be found in the Supplemental.

# Experiments

**Implementation Details.** Our method is implemented in PyTorch and trained on one NVIDIA RTX3090 GPU. In preprocessing, input RGB images, ground-truth mirror maps, and edge maps are resized to $224 \times 224$. We initialize the ResNext-101 (Xie et al. 2017) backbone from VMD-Net (Lin, Tan, and Lau 2023) pretrained weights. The model is trained for 15 epochs with early stopping using SGD with an initial learning rate of $9e-3$, momentum 0.9, weight decay $5e-4$, thresholding 0.5, batch size 8. An adaptive schedule interpolates the learning rate from $9e-3$ to $3e-3$

| Models | Tasks | $F_\beta\uparrow$ | IoU↑ | Accu↑ | MAE↓ |
|---|---|---|---|---|---|
| F3Net (Jun Wei 2020) | VSD | 0.852 | 0.696 | 0.851 | 0.149 |
| FSNet (Ji et al. 2021) | VSD | 0.831 | 0.710 | 0.853 | 0.147 |
| MGA (Li et al. 2019) | VSD | 0.506 | 0.299 | 0.577 | 0.423 |
| UFO (Su et al. 2023) | VSD | 0.799 | 0.601 | 0.798 | 0.202 |
| Samba (He et al. 2025) | VSD | 0.838 | 0.724 | 0.861 | 0.139 |
| MirrorNet (Yang et al. 2019) | IMD | 0.845 | 0.674 | 0.840 | 0.160 |
| PDNet (Mei et al. 2021) | IMD | 0.824 | 0.722 | 0.857 | 0.143 |
| PMDNet (Lin, Wang, and Lau 2020) | IMD | 0.839 | 0.400 | 0.731 | 0.269 |
| VMDNet (Lin, Tan, and Lau 2023) | VMD | 0.812 | 0.723 | 0.854 | 0.150 |
| MGVMD (Warren et al. 2024) | VMD | 0.869 | 0.725 | 0.873 | 0.128 |
| SAM2 (Ravi et al. 2025) | VOS | 0.743 | 0.651 | 0.801 | 0.199 |
| **Ours** | VMD | **0.884** | **0.746** | **0.889** | **0.112** |

Table 1: Comparison of Video Mirror Detection (VMD), Image-based Mirror Detection (IMD), Video Salient Object Detection (VSD) and Video Object Segmentation (VOS) methods on the MMD dataset (Warren et al. 2024).

over epochs 1–15. We adopt non-overlapping dataset splits from (Warren et al. 2024) and VMD-D (Lin, Tan, and Lau 2023).

**Evaluation Methods and Metrics.** We evaluate our method against 11 state-of-the-art methods, including two video mirror detection methods (*i.e.*, MGVMD (Warren et al. 2024) and VMDNet (Lin, Tan, and Lau 2023)), a large foundational video segmentation method SAM2 (Ravi et al. 2025), five video salient object detection methods (*i.e.*, F3Net (Jun Wei 2020), FSNet (Ji et al. 2021), MGA (Li et al. 2019), UFO (Su et al. 2023), and Samba (He et al. 2025)), and three image-based mirror detection methods (*i.e.*, PDNet (Mei et al. 2021), MirrorNet (Yang et al. 2019), and PMDNet (Lin, Wang, and Lau 2020)). We use their respective pre-trained weights, fine-tune and validate them on the video mirror dataset MMD (Warren et al. 2024)[1]

We use the Intersection over Union (IoU↑) for a geometric interpretation of how predictions overlap with ground truth mirror masks, and the F-beta Score ($F_\beta\uparrow$) for considering both precisions and recalls. We also report the per-pixel accuracy/MAE (Accuracy↑/MAE↓) for a reference.

**Quantitative Comparison.** Table 1 reports the results on the MMD dataset (Warren et al. 2024). We can see that by leveraging 2D motion cues, the state-of-the-art MGVMD method typically outperforms previous mirror detectors in terms of $F_\beta$. However, it still struggles to detect mirrors in complex scenes with clustered or featureless regions, resulting in an IoU similar to those of appearance correspondence-based VMDNet and RGBD-based PDNet. Meanwhile, the results show that fine-tuning large foundational model for video object segmentation (SAM2) may not be optimal to video mirror detection. In contrast, we learn mirror representations by modeling the motion-in-depth cue, which correlates contrastive contextual features in low-dimensional 3D motion, depth, and RGB modalities, yielding consistently better results on all four metrics. We provide speed and parameter analysis of our method against SOTA methods in the Supplemental (our model ranks second in inference time).

**Qualitative Comparison.** We provide visual comparison in Figure 5, where we observe several key points. First, our method shows improved temporal consistency and reduced noise in predictions. This is attributed to the MiD-AL module, which learns mirror representations from the correlation of contrastive features in 3D motion, depth, and RGB modalities, while leveraging a low-dimensional 3D motion prior to minimize noise. Second, our method effectively predicts mirror regions of varying shapes, sizes, and scenes, demonstrating the robustness of our approach by modeling Motion-in-Depth. Additionally, the lightweight MiD-BD module enhances the boundary of the mirror predictions, allowing accurate mirror detection even in featureless or motionless regions, such as reflections from plain walls.

[1]We evaluate on VMD-D (Lin, Tan, and Lau 2023). MiD-VMD surpasses SOTA in Accuracy↑ and MAE↓. However, VMD-D has labeling errors and many small mirrors (Warren et al. 2024), reducing representativeness. Results are discussed in the supplemental.

| Ablated Models | $F_\beta\uparrow$ | IoU↑ | Accuracy↑ | MAE↓ |
|---|---|---|---|---|
| Baseline | 0.806 | 0.681 | 0.856 | 0.144 |
| Baseline + MiD-AL | 0.843 | 0.718 | 0.872 | 0.128 |
| Baseline + MiD-BD | 0.861 | 0.724 | 0.875 | 0.125 |
| Ours | **0.884** | **0.746** | **0.889** | **0.112** |

Table 2: Model ablation study on the MMD dataset.

| Dimensions | FBeta↑ | IoU↑ | Accuracy↑ | MAE↓ |
|---|---|---|---|---|
| 128 (None) | 0.814 | 0.713 | 0.871 | 0.129 |
| 32 | 0.850 | 0.736 | 0.872 | 0.128 |
| 8 | 0.861 | 0.737 | 0.879 | 0.121 |
| **Ours (16)** | **0.884** | **0.746** | **0.889** | **0.112** |

Table 3: Ablation on dimensions of the MiD-AL module.

## Internal Analysis

**Proposed Modules.** We evaluate the effectiveness of our modules on the MMD dataset, as summarized in Table 2. First, we remove both the Motion-in-Depth Attention Learning (MiD-AL) and Motion-in-Depth Boundary Detection (MiD-BD) modules to form the *Baseline* (b), representing performance without multi-modality correlations or contrastive 3D features. We then add each module individually for ablation, denoted as "Baseline+MiD-AL" and "Baseline+MiD-BD". Table 2 demonstrates that both modules improve results over the baseline, highlighting the value of modeling 3D motion, depth, and image features. The full model yields the best performance, with sharper mirror boundaries attributed to contrastive boundary feature learning. Figure 6 provides qualitative comparisons: b+MiD-BD enhances boundary quality over the baseline, while b+MiD-AL further improves mirror localization. Combining both modules delivers the most accurate results overall.

**Dimensions in MiD-AL.** We hypothesize that motion-in-depth cues can be effectively modeled using two global (affine) transformations, one for the motion inside the mirror and one for outside, resulting in a low $\sigma$-dimensional representation ($12 < \sigma \leq 12 + 12 = 24$). To test this hypothesis, we perform an ablation study on the dimensions, using the MMD dataset and the MiD-AL without low-dimensional encoding as baseline (denoted as "128 (None)" in Table 3). Next, we ablate on encoding dimensions of 32, 16 (Ours),

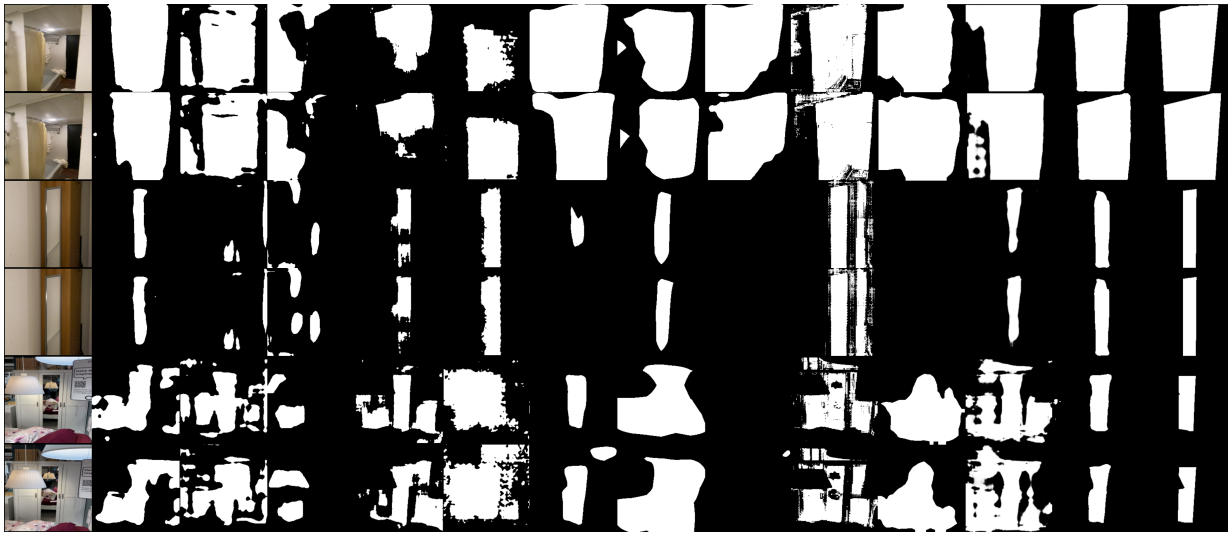


Figure 5: Visual comparisons between our method and state-of-the-art methods trained and validated on the MMD dataset.

| Models | Types | $F_\beta\uparrow$ | IoU↑ | Acc↑ | MAE↓ |
|---|---|---|---|---|---|
| MGVMD | 2D Motion | 0.869 | 0.725 | 0.873 | 0.127 |
| MGVMD+SF | 3D Motion | 0.874 | 0.742 | 0.868 | 0.132 |
| DCTNet+ (Mou et al. 2024) | M&D | 0.827 | 0.729 | 0.861 | 0.140 |
| Ours-Depth Only (No Scene Flow) | Depth | 0.852 | 0.709 | 0.868 | 0.132 |
| **Ours** | **MiD** | **0.884** | **0.746** | **0.889** | **0.112** |

Table 4: Comparisons of MiD to motion, depth, and M&D.

| Models | $F_\beta\uparrow$ | IoU↑ | Acc↑ | MAE↓ |
|---|---|---|---|---|
| Ours (ML-Depth-Pro (Bochkovskii et al. 2025)) | <u>0.879</u> | **0.758** | **0.890** | **0.110** |
| Ours (DepthAnything-V2 (Yang et al. 2024)) | 0.878 | 0.751 | 0.888 | <u>0.112</u> |
| Ours (GA-Net (Zhang et al. 2019)) | 0.857 | 0.727 | 0.874 | 0.126 |
| Ours (Raft-Stereo (Lipson, Teed, and Deng 2021) | **0.884** | 0.746 | <u>0.889</u> | <u>0.112</u> |

Table 5: Robustness of MiD-VMD to depth estimation.

and 8. Table 3 shows that our MiD-AL module with a dimension of 16 delivers the best performance.

**MiD vs. Motion, Depth, and M&D.** Previous works, including those beyond mirror detection, typically rely on either depth (Li et al. 2023) or motion (Li et al. 2019) alone, or use simple motion-depth (M&D) fusion (Mou et al. 2024). To demonstrate the effectiveness of our MiD cue, we further conduct comprehensive ablations. We first adapt MGVMD (Warren et al. 2024) by replacing its 2D optical flow with 3D scene flow to assess the impact of 3D motion. We then compare against DCTNet+ (Mou et al. 2024), a recent SOTA VSOD method that explicitly models 2D motion and depth (M&D). Last, we include a depth-only baseline by reducing RAFT-3D to single-frame depth input. Results in Table 4 show that our method consistently outperforms all alternatives, highlighting the value of explicitly modeling 3D motion (MiD) inspired by human perception.

**Depth Estimation Ablation Study.** Table 5 presents ablation results using ML-Depth-Pro (Bochkovskii et al. 2025), DepthAnything-V2 (Yang et al. 2024), GA-Net (Zhang et al. 2019), and Raft-Stereo (Lipson, Teed, and Deng 2021) depth estimators. While ML-Depth-Pro achieves the best Acc↑ and MAE↓, its inference speed is considerably slower. DepthAnything-V2 and Raft-Stereo show similar performances. This demonstrates the robustness of our model to different depth estimators. Supplemental provides more analysis and timing comparison of our method.


Figure 6: Qualitative results of the ablated models.

## Conclusion

We propose the first method to derive and apply the Motion-in-Depth (MiD) cue without depth sensors for video mirror detection. Our MiD-VMD model introduces two modules: MiD-AL (*Motion-in-Depth Attention Learning*), which reduces noise and exploits correlations among 3D motion, depth, and image features, and MiD-BD (*Motion-in-Depth Boundary Detection*), which improves mirror boundary detection in motionless or featureless regions. MiD-VMD surpasses state-of-the-art methods but is not yet real-time. Future work will focus on accelerating inference via distillation for resource-constrained scenarios such as drone surveying.

## Acknowledgments

Alex is supported by a Swansea GTA Research Scholarship. This project is in part supported by two GRF grants from RGC of Hong Kong (Ref.: 11211223 and 11220724). We gratefully acknowledge support of the HEFCW HERC fund (W21/21HE) for the provision of GPU equipment used in this research. For the purpose of Open Access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript (AAM) version arising from this submission.

## References

Baker, P. M.; and Bair, W. 2016. A Model of Binocular Motion Integration in MT Neurons. *Journal of Neuroscience*.

Bochkovskii, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S. R.; and Koltun, V. 2025. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. arXiv:2410.02073.

Cheng, X.; Xiong, H.; Fan, D.-P.; Zhong, Y.; Harandi, M.; Drummond, T.; and Ge, Z. 2022. Implicit Motion Handling for Video Camouflaged Object Detection. In *CVPR*.

Fan, D.-P.; Wang, W.; Cheng, M.-M.; and Shen, J. 2019. Shifting more attention to video salient object detection. In *CVPR*.

Gu, Y.; Wang, L.; Wang, Z.; Liu, Y.; Cheng, M.-M.; and Lu, S.-P. 2020. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*.

Guan, H.; Lin, J.; and Lau, R. 2022. Learning Semantic Associations for Mirror Detection. In *CVPR*.

Guo, Y.-C.; Kang, D.; Bao, L.; He, Y.; and Zhang, S.-H. 2022. NeRFReN: Neural Radiance Fields With Reflections. In *CVPR*.

He, J.; Fu, K.; Liu, X.; and Zhao, Q. 2025. Samba: A Unified Mamba-based Framework for General Salient Object Detection. In *CVPR*, 25314–25324.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*.

He, R.; Lin, J.; and Lau, R. W. 2023. Efficient Mirror Detection via Multi-level Heterogeneous Learning. In *AAAI*.

Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R. W. H.; and Zuo, W. 2023. Symmetry-Aware Transformer-based Mirror Detection. In *AAAI*.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*.

Ji, G.-P.; Fu, K.; Wu, Z.; Fan, D.-P.; Shen, J.; and Shao, L. 2021. Full-duplex strategy for video object segmentation. In *ICCV*.

Ju, J.; Tseng, C. W.; Bailo, O.; Dikov, G.; and Ghafoorian, M. 2023. DG-Recon: Depth-Guided Neural 3D Scene Reconstruction. In *ICCV*, 18184–18194.

Jun Wei, Q. H., Shuhui Wang. 2020. F3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*.

Kim, H. R.; Angelaki, D. E.; and DeAngelis, G. C. 2015. A Functional Link between MT Neurons and Depth Perception Based on Motion Parallax. *Journal of Neuroscience*.

Li, H.; Chen, G.; Li, G.; and Yu, Y. 2019. Motion guided attention for video salient object detection. In *ICCV*.

Li, J.; Ji, W.; Wang, S.; Li, W.; and cheng, L. 2023. DVSOD: RGB-D Video Salient Object Detection. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS*.

Lin, J.; Tan, X.; and Lau, R. W. 2023. Learning To Detect Mirrors From Videos via Dual Correspondences. In *CVPR*.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive Mirror Detection. In *CVPR*.

Lipson, L.; Teed, Z.; and Deng, J. 2021. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In *3DV*.

Liu, F.; Hu, Y.; and Salzmann, M. 2023. Linear-Covariance Loss for End-to-End Learning of 6D Pose Estimation. In *ICCV*.

Liu, J.; Wang, J.; Wang, W.; and Su, Y. 2022. DS-Net: Dynamic Spatiotemporal Network for Video Salient Object Detection. *arXiv:2012.04886*.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-Aware Mirror Segmentation. In *CVPR*.

Mou, A.; Lu, Y.; He, J.; Min, D.; Fu, K.; and Zhao, Q. 2024. Salient Object Detection in RGB-D Videos. *IEEE TIP*.

Pal, A.; Mondal, S.; and Christensen, H. I. 2020. "Looking at the right stuff"-Guided semantic-gaze for autonomous driving. In *CVPR*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In *ICLR*.

Su, Y.; Deng, J.; Sun, R.; Lin, G.; and Wu, Q. 2023. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *IEEE TMM*.

Tan, J.; Lin, W.; Chang, A. X.; and Savva, M. 2021. Mirror3D: Depth Refinement for Mirror Surfaces. In *CVPR*.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2023. Mirror Detection With the Visual Chirality Cue. *IEEE TPAMI*.

Tang, Y.; Li, Y.; and Xing, G. 2021. Video Salient Object Detection via Adaptive Local-Global Refinement. *arXiv:2104.14360*.

Teed, Z.; and Deng, J. 2021. RAFT-3D: Scene Flow using Rigid-Motion Embeddings. In *CVPR*.

Uka, T.; and DeAngelis, G. C. 2006. Linking Neural Representation to Function in Stereoscopic Depth Perception: Roles of the Middle Temporal Area in Coarse versus Fine Disparity Discrimination. *The Journal of Neuroscience*.

Wang, W.; Shen, J.; and Shao, L. 2017. Video salient object detection via fully convolutional networks. *IEEE TIP*.

Warren, A.; Xu, K.; Lin, J.; Tam, G. K.; and Lau, R. W. 2024. Effective Video Mirror Detection with Inconsistent Motion Cues. In *CVPR*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*.

Xing, Z.; Liu, L.; Yang, Y.; Wang, H.; Ye, T.; Chen, S.; Li, W.; Liu, G.; and Zhu, L. 2025. Detect Any Mirrors: Boosting Learning Reliability on Large-Scale Unlabeled Data with an Iterative Data Engine. In *CVPR*, 25476–25486.

Xu, K.; Siu, T. W.; and Lau, R. W. H. 2024. ZOOM: Learning Video Mirror Detection with Extremely-Weak Supervision. In *AAAI*.

Xu, M.; Fu, P.; Liu, B.; and Li, J. 2021. Multi-stream attention-aware graph convolution network for video salient object detection. *IEEE TIP*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything V2. *arXiv:2406.09414*.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. H. 2019. Where Is My Mirror? In *ICCV*.

Zeng, J.; Bao, C.; Chen, R.; Dong, Z.; Zhang, G.; Bao, H.; and Cui, Z. 2023. Mirror-NeRF: Learning Neural Radiance Fields for Mirrors with Whitted-Style Ray Tracing. In *ACM MM*.

Zhang, F.; Prisacariu, V.; Yang, R.; and Torr, P. 2019. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. In *CVPR*, 185–194.

Zhang, M.; Liu, J.; Wang, Y.; Piao, Y.; Yao, S.; Ji, W.; Li, J.; Lu, H.; and Luo, Z. 2021. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*.

Zhao, X.; Liang, H.; Li, P.; Sun, G.; Zhao, D.; Liang, R.; and He, X. 2024. Motion-aware memory network for fast video salient object detection. *IEEE TIP*.