Effective Video Mirror Detection with Inconsistent Motion Cues

Alex Warren[1], Ke Xu[2], Jiaying Lin[2], Gary K.L. Tam[1], Rynson W.H. Lau[1,2]
Department of Computer Science, Swansea University[1] and City University of Hong Kong[2]
alex.warren@swansea.ac.uk, kkangwing@gmail.com, jiayinlin5-c@my.cityu.edu.hk
k.l.tam@swansea.ac.uk, Rynson.Lau@cityu.edu.hk

Abstract

Image-based mirror detection has recently undergone rapid research due to its significance in applications such as robotic navigation, semantic segmentation and scene reconstruction. Recently, VMD-Net was proposed as the first video mirror detection technique, by modeling dual correspondences between the inside and outside of the mirror both spatially and temporally. However, this approach is not reliable, as correspondences can occur completely inside or outside of the mirrors. In addition, the proposed dataset VMD-D contains many small mirrors, limiting its applicability to real-world scenarios. To address these problems, we developed a more challenging dataset that includes mirrors of various shapes and sizes at different locations of the frames, providing a better reflection of real-world scenarios. Next, we observed that the motions between the inside and outside of the mirror are often inconsistent. For instance, when moving in front of a mirror, the motion inside the mirror is often much smaller than the motion outside due to increased depth perception. With these observations, we propose modeling inconsistent motion cues to detect mirrors, and a new network with two novel modules. The Motion Attention Module (MAM) explicitly models inconsistent motions around mirrors via optical flow, and the Motion-Guided Edge Detection Module (MEDM) uses motions to guide mirror edge feature learning. Experimental results on our proposed dataset show that our method outperforms state-of-the-arts. The code and dataset are available at https://github.com/AlexAnthonyWarren/MG-VMD.

1. Introduction

Mirrors and reflective surfaces are abundant in our surroundings. Successful detection of mirrors underpins many vision applications such as autonomous drone navigation and vehicle vision systems [16]. When a scene contains mirrored regions, mirror identification is critical to reducing errors in applications such as salient object detection [25, 26], semantic segmentation [16, 21], scene reconstruction[3, 4] and NeRF modeling [6, 31].

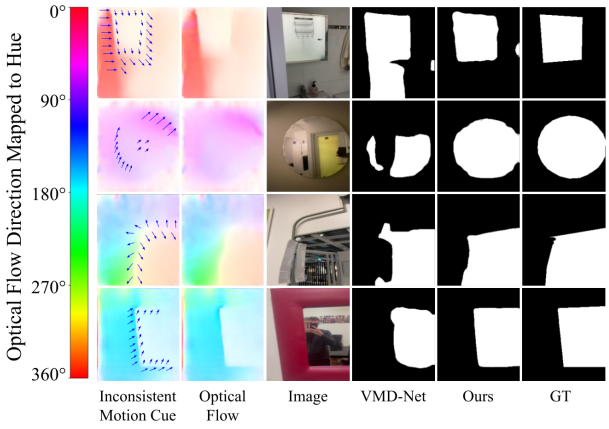

Figure 1. We propose to model motion inconsistency as a cue for mirror detection. The 1^{st} column shows the inconsistent motions (depicted with arrows). The locations of these cues align well with the mirror regions. The 2^{nd} column shows the optical flow computed between frame N (3^{rd} column), and frame N-1 from respective videos in our new dataset. Our method (5^{th} column) predicts more reliable and consistent mirror regions, outperforming VMD-Net [12] (4^{th} column). The 6^{th} column is the ground truth.

Mirror detection involves binary classification of individual pixels in RGB images to determine if they are inside or outside of the mirror. It is a challenging research as mirrors typically do not have their own visual appearance. They instead reflect objects within their surroundings. This visual attribute makes mirrors difficult to detect.

There are a few image-based and one video-based mirror detection methods in the literature. Yang et al. [30] first introduced a mirror detection method using contrast and semantics. Lin et al. [11] introduced a PMD-Net that utilizes edge detection and leverages contextual contrasting objects inside and outside of mirrored regions to detect mirrored regions. Tan et al. [22] detect mirrors using Visual Chirality Cue [13] that encodes mirror-like symmetrical properties at the pixel level. Mei et al. [15] factor depth estimation into their mirror detection process and show good results

in RGB and RGB-D mirror datasets. Recently, Lin et al. [12] introduced the first video mirror detection method using inter-frame and intra-frame dual correspondences.

Despite these exciting results, we observe research gaps and limitations in these works, especially in video mirror detection. The state-of-the-art VMD-Net [12] models dual correspondences inside and outside of mirrors to aid in the detection of mirror regions in videos. However, the approach is not reliable as correspondences can appear completely inside or outside of mirrors as discussed in [22]. In addition, the proposed VMD-D dataset contains primarily small mirrors, with small inter-frame motion, mirror coverage and content variation. This limits its applicability to mirror detection in real-world scenarios.

To address these issues, we first developed a new, more challenging dataset, Mirror Motion Dataset (MMD). The dataset contains 37 (≈9-second) videos with substantial motion, each accompanied by manual mirror annotations. To address the small mirror issue in VMD-D, we capture a high variance of mirror sizes and shapes at different spatial locations, generalizing to real-world scenarios. We further include challenging examples, including low-contrast environments, multi-/large/full-mirror scenes and under different lighting conditions. Next, we observe that the motions between the inside and outside of mirrors are often inconsistent. For instance, when moving in front of a mirror, the motion inside is often smaller than outside due to the increased depth perception. The inconsistent motion cues (depicted by arrows in the 1^{st} column in Figure 1) also apply in challenging scenarios like low-contrast and multi-mirror scenes. These observations motivate us to ask the question: *Can we apply the motion inconsistency cue to improve mirror detection in challenging complex video scenes?*

Neuroscience studies suggest that humans often rely on dynamic perceptual cues to identify mirrored regions in daily life. Tamura et al. [18] [19] show that humans are capable of distinguishing mirrors in dynamic scenes involving rotational movement. [20] highlights that rotational, parallax, forward and backward motions are exploitable motion inconsistencies for mirror detection. To our knowledge, such motion inconsistency cues have not been utilized before for the video mirror detection task.

To this end, we propose a novel network to model motion inconsistency for mirror detection, with two novel modules. First, *the Motion Attention Module (MAM)* predicts mirror regions by exploiting inconsistent motion cues obtained from optical flow fields. Second, *the Motion-guided Edge Detection Module (MEDM)* further uses motion to guide mirror edge detection. Experimental results show that the proposed network outperforms relevant state-of-the-art methods. To summarise, our main contributions include:

- We propose a novel deep-learning video mirror detection method based on the motion inconsistency cue.
- Our network includes two novel modules. Motion Attention Module (MAM) attentively predicts mirror regions using motion inconsistency in the optical flow fields. Motion-guided Edge Detection Module (MEDM) uses the motions to guide mirror edge feature extraction.
- We further present a new video dataset (MMD). Compared to VMD-D, our dataset contains mirrors of various shapes and size at different locations of the images. It also covers various scenes, *e.g.*, low-light scenes, and multi-/large/full mirror scenes. It is a more challenging dataset, more suitable for real-life scenarios.
- Extensive evaluations show that the proposed model outperforms existing methods on video mirror detection.

2. Related Works

Video Mirror Detection. VMD-Net [12] introduced the first video mirror detection method. It employs dual correspondences across both inter-frame and intra-frame to predict mirrored regions. They further presented a video mirror dataset for this purpose. However, despite its demonstrated success, we observe two limitations. First, the idea of dual correspondences is not reliable as correspondences can occur completely inside or outside of the mirrors [22]. Second, their dataset contains an excess of small mirrors, which does not generalize well to real-world scenarios.

This paper aims to facilitate video mirror detection in the wild from two perspectives. First, we propose to model the inconsistent motion cues to learn more robust mirror representations. Second, we introduce a new benchmark dataset comprising videos with more complex scenarios.

Image-Based Mirror Detection. [30] pioneered mirror detection by utilizing contextual contrasting features and introducing a mirror detection dataset. [15] refined detection using RGB-D data and presented the first RGB-D mirror dataset. [11] introduced a method leveraging contextual contrasts to relate mirror interior and exterior contents, along with an expanded image-based mirror dataset. [22] incorporated chirality cues for mirror detection. Although these work achieved good results, they mainly focus on single-image mirror detection and lack the capacity to identify mirrors within temporal videos.

Video Salient Object Detection. Video Salient Object Detection (VSOD) seeks to identify the most visually prominent object in videos. Early VSOD studies [27] [28] utilized manually crafted features to detect these salient objects. In more recent research [9] [8] [9] [14] [23], the focus has shifted to deep learning models, which have demonstrated promising results in VSOD. These techniques however are not optimized for the video mirror detection task. This is because mirrored regions may not always represent the most visually significant objects within a scene. Our technique aims to identify mirrors in real-life scenarios.

Optical Flow Estimation. Optical flow estimation fo-

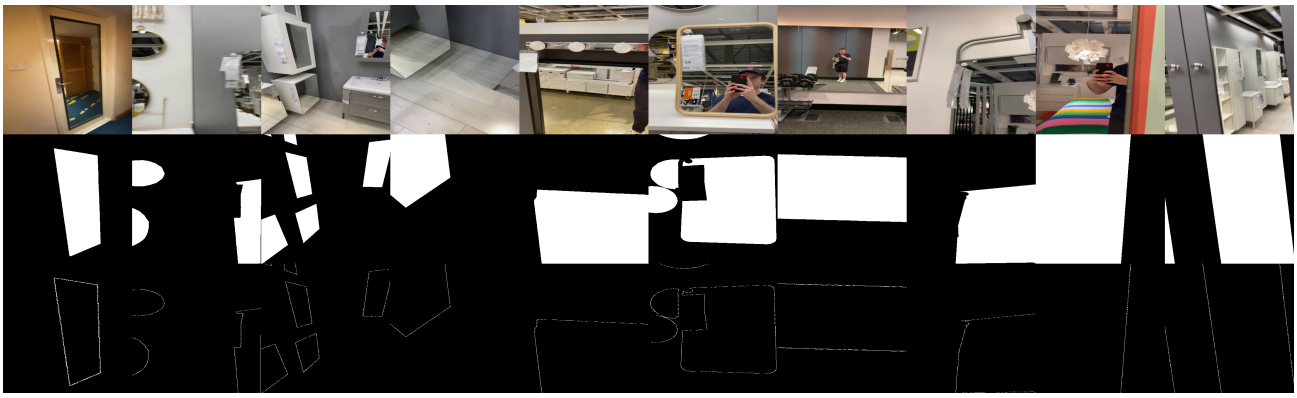

Figure 2. Snapshots of our proposed Motion Mirror Dataset (MMD), with pixel-level mirror and edge annotations.

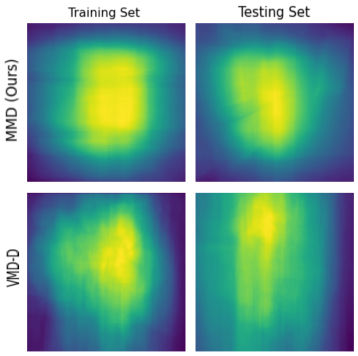

Figure 3. Cumulative Distributions.

Figure 4. Mirror Sizes.

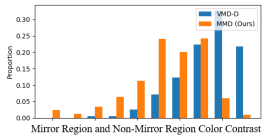

Figure 5. Color Contrast.

cuses on extracting per-pixel motion from video sequences. Early efforts [5] used hand-crafted features of image brightness between consecutive frames. In recent studies, convolutional deep learning and transformer-based approaches have made significant improvements [24] [7]. In our context, optical flow estimation is not inherently applicable to video mirror detection and cannot serve as a direct detection method. This study aims to leverage deep-learning optical flow as a way to model inconsistent motion cues, which enables a robust end-to-end video mirror detection framework.

3. Proposed Dataset

To create our Motion Mirror Dataset (MMD), we first recruited six volunteering participants (final year university students) to capture 37 nine-second smartphone videos from their surroundings (e.g., living rooms, bathrooms). Participants were instructed to maintain consistent motion around mirror regions while recording. We recorded videos under various lighting conditions, including daily light and low-light settings. The dataset was subsequently divided into non-overlapping training (4,653 samples) and testing (5,074 samples) sets. To accomplish this, we randomly split the 37 videos into the two sets (training and testing). Specifically 18 videos (4,653 samples) are put into the training set and 19 videos (5,074 samples) are put into the testing set. Each frame was meticulously annotated with pixel-level mirror masks. We applied a depth-first search algorithm to the pixel-level mirror masks to extract edge masks, which we then manually verified for ground truth accuracy. All videos in the dataset have a frame rate of 30 fps.

Dataset	SSIM ↑
VMD-D	0.7950544833184201
Ours (MotionMirrorDataset)	0.9033545507233643

Table 1. SSIM of MMD and VMD-D. A higher SSIM indicates a more reasonable split of dataset.

Dataset	Areas not covered by mirrors ↓
(Ours) MMD Training	0.000061
(Ours) MMD Testing	0.000015
VMD-D Training	0.00038147
VMD-D Testing	0.00024414

Table 2. Mirror Coverage. A smaller non-coverage rate indicates a more even spatial distribution of mirrors.

Statistic Analysis. We compare our proposed MMD with VMD-D [12] in terms of Cumulative Distribution and Structural Similarity between training and testing sets, Mirror Size Distribution, and Area Coverage.

Figure 3 shows the cumulative distribution of all mirrors in both training and testing sets. Compared to VMD-D, our MMD dataset contains more intricate scenarios. VMD-D [12] however is highly biased with majority of mirror regions fall in the center in the train/test sets whilst there are nearly none along the edge (darker color on left/right). The mirror regions of our proposed dataset (MMD) are spread out across the entire heatmap where mirrors are also found at the image edges. We further compute the Structural Similarity Index (SSIM) on the cumulative area distributions (Figure 3) directly to compare the structure between the whole training and testing sets. The higher SSIM of our MMD in Table 1 shows that the training and testing sets are similar in structure and fairer for evaluation purpose.

Figure 4 shows the size distribution of mirrors in both MMD and VMD-D datasets. VMD-D contains a lot of

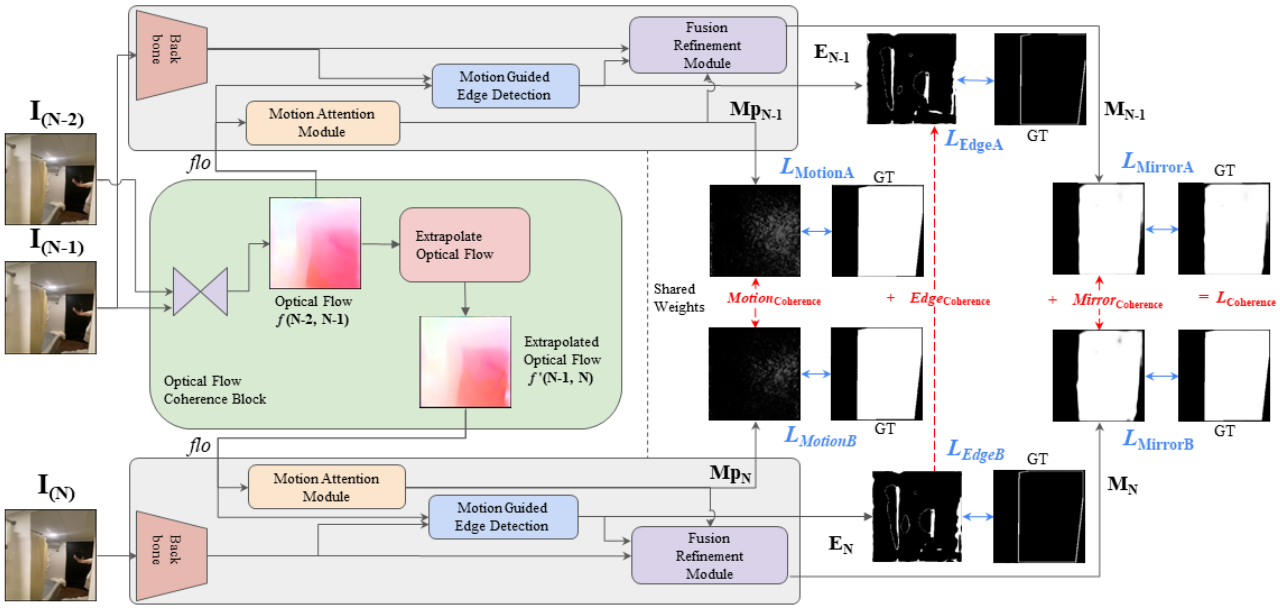

Figure 6. Method overview. Our MG-VMD network takes three adjacent frames (I_{N-2}, I_{N-1}, I_N) as inputs and extracts their multi-scale features via a shared ResNext-101 [29] backbone and DeepLabV3[2]. We apply the optical flow coherence block to process I_{N-2} and I_{N-1} to compute an optical flow map for I_{N-1}, followed by an optical flow vector field extrapolation to estimate the optical flow map for I_N. These two optical flow maps are fed into our proposed MAM and MEDM modules to guide the mirror localization and mirror-edge detection processes, respectively. The proposed Motion Attention Module (MAM) learns to locate mirrors from optical flow fields, exploiting motion cues, attentively modeling motion inconsistency. The proposed Motion-guided Edge Detection Module (MEDM) learns to extract and refine mirror boundary edge features. Finally, a fusion refinement module is used to predict the mirror maps by fusing the predicted mirror boundary maps of MEDM, the mirror localization maps of MAM, and the extracted multi-scale image backbone features.

small mirrors with few/no large mirrors, which does not represent real-world scenarios. In contrast, our MMD dataset has a more reasonable distribution, containing a variety of small, medium and large sized mirrors, and therefore is more challenging than VMD-D.

Table 2 shows the areas (ratio to the whole image) where there is no mirror coverage within the two datasets. The smaller ratios show that MMD contains mirrors that are more evenly spatially distributed across the whole image. MMD is more challenging and will introduce less spatial bias in the model learning.

We further compare the color contrast between mirror and non-mirror regions in all images between MMD and VMD-D. The histogram in Figure 5 shows that MMD contains more images with lower color contrast between mirror and non-mirror regions. Lower color contrast implies a decrease in the amount of contextual information available for a model to leverage, making it more challenging to detect mirrors. The 4th column of Figure 2 shows one example. These show that MMD is more challenging than VMD-D.

4. Proposed Method

Our proposed method, Motion-guided Video Mirror Detection (MG-VMD), aims to enhance mirror detection by modelling motion inconsistency cues. Drawing inspiration from our own observations and neuroscience studies, we hypothesise that modeling motion inconsistency cues can improve the accuracy of mirror detection, especially in dynamic video scenarios.

The core idea behind MG-VMD is to utilize optical flow to model motion inconsistency and use it to guide the extraction of multi-scale image and edge features, allowing for the temporal detection of mirrored regions in videos. To achieve this, we introduce two key modules: the Motion Attention Module (MAM) and the Motion-guided Edge Detection Module (MEDM).

Specifically, the Motion Attention Module (MAM) is designed to explicitly predict mirror locations by modelling motion inconsistency cues from the underlying optical flow fields. This module enhances video mirror prediction by providing spatial contextual information through the multi-scale image features and motion cues. It informs the subsequent stages of the model about the location of mirror regions in the video.

The Motion-guided Edge Detection Module (MEDM) is dedicated to predicting mirror boundaries, contributing to the enhancement of overall mirror region predictions. This module takes advantage of image features and employs mo-

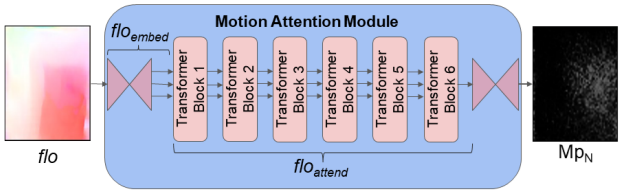

Figure 7. The proposed MAM (Motion Attention Module) aims to model the inconsistent motion cues to localize the mirrors.

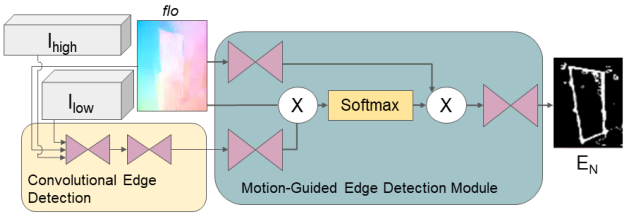

Figure 8. Our proposed MEDM (Motion-Guided Edge Detection Module) aims to detect the mirror boundaries based on the captured motion cues in the optical flow vector fields.

tion inconsistency cues to guide the detection of features at the edges, specifically tailored for mirror detection.

The Fusion Refinement Module, inspired from PMD-Net [11], further takes the MAM predicted mirror locations and the MEDM refined mirror boundaries to predict mirrors. The integration plays a pivotal role in improving mirror prediction by alleviating errors arising from the limited capability of edge detection or inconsistent motion cues alone. It overcomes the limitations of individual modules and provides spatial and temporal context for accurate mirror prediction.

In Figure 6 (Architecture Overview), the MG-VMD processes adjacent images (I_{N-2}, I_{N-1}, I_N) as inputs. Optical flow field extrapolation is employed to enhance coherency and address temporal errors between adjacent branches (N-1 and N). The model utilizes a shared ResNext-101 [29] backbone and DeepLabV3 [2] to extract multi-scale image features. The outputs include predicted mirror binary maps (M_{N-1} and M_N), predicted mirror-boundary edge maps (E_{N-1} and E_N), and mirror motion predictions (Mp_{N-1} and Mp_N). The collaboration of MAM, MEDM, and the Fusion Refinement Module ensures accurate and robust video mirror detection by considering both spatial and temporal aspects of mirror cues.

4.1. Optical Flow Coherence Block

Our architecture leverages a pre-trained optical flow model. Similar to previous work [10], our optical flow block utilizes frozen weights. However, optical flow may become unreliable in scenarios with low contrast in image details or when images become saturated. To achieve a simpler design while obtaining more stable optical flow, we introduce our Optical Flow Coherence Block, inspired by the work in [24] for motion estimation. This block extracts motion and extends it using linear extrapolation. The extrapolated optical flow motion, denoted as flo, is then utilized in our subsequent motion-guided modules, MAM and MEDM. We empirically find that this simple design outperforms alternatives (refer to Section 6 of the Supplemental) whilst enhancing coherency.

Specifically, the block calculates motion vectors $f(N-2, N-1)$ between frames I_{N-2} and I_{N-1}. These vectors are then extrapolated by one frame to estimate motion $f'(N-1, N)$ between I_{N-1} and I_N. By using extrapolation, we connect the motion information from the frame at N-2 to N-1, and then to N. This connection helps maintain a smooth and consistent flow of motion information, reducing errors from neighboring frames and sudden changes in motion, especially around mirror regions.

4.2. Motion Attention Module

Figure 7 shows the architecture of the Motion Attention Module (MAM). This module focuses on direct mirror prediction from the optical flow fields. It leverages self-attention to capture and refine better features, to model motion inconsistency cue and to inform mirror locations. The output is then fed to the Fusion Refinement Module for the final mirror refinement.

Specifically, we initiate the MAM by embedding patches with positional information to create flo_{embed}. Subsequently, we guide flo_{embed} through six transformer blocks. Each block incorporates self-attention, treating flo_{embed} as Q, K, and V in the self-attention mechanism, with output denoted as flo_{attend}. To enhance its representation, flo_{attend} undergoes batch normalization, followed by processing through a feed-forward design.

The output of the self-attention mechanism within one Transformer $\text{Block}_{\text{Layer N}}$ is then passed to the subsequent Transformer $\text{Block}_{\text{Layer N+1}}$. Following this sequence, the output from Transformer $\text{Block}_{\text{Layer 6}}$ is channeled through a convolutional layer. The intentional use of six sequential transformer blocks serves a dual purpose: it equips the model to comprehend long-range dependencies, crucial for temporal relationships, and enables the capture of intricate spatial dependencies. This enhanced capability is particularly valuable for predicting spatial locations of the attended mirrors, denoted as Mp_{N-1} and Mp_N.

4.3. Motion-guided Edge Detection Module

Figure 8 shows the architecture of the Motion-guided Edge Detection Module (MEDM). Unlike previous mirror edge detection in [11, 22] that only exploits image features, we also leverage the inconsistent motion cues captured in the optical flow field to guide the mirror edge detection process.

MEDM processes three inputs: I_{low} (a 2nd-level multi-scale image feature), I_{high} (a 5th-layer multi-scale image

feature), and *flo* (the optical flow field for motion guidance). The process begins with convolutional layers that extract and predict edge features—referred to as E_N or E_{N-1}—from I_{low}, I_{high}, and *flo*.

The cross-attention mechanism involves matrix multiplication between the transformed *flo* and E_N or E_{N-1}. A Softmax activation determines the cross-attended beta value between *flo* and E_N or E_{N-1}. This beta value, representing the attention weight, is then used to focus on *flo*, determining the contribution of each element in the optical flow field to the final cross-attended motion-guided edge feature. We specifically employ a cross-attention mechanism in this step to guide the extracted edge feature with the optical flow field. The motion-guided edge feature undergoes a convolutional pass, producing the output: the motion-guided edge feature map, E_N or E_{N-1}, respectively.

4.4. Fusion Refinement Module

We refine the motion-guided edge detection features from MEDM and the motion-based mirror predictions from MAM. This refining step is crucial to address limitations in each module (MEDM, MAM) and ensure accurate mirror prediction with both spatial and temporal context.

The process involves fusion-refinement, where model information is combined, refined, and fed into a convolutional binary classification head. This produces per-pixel binary classification maps, named M_{N-1} and M_N, using two instances of the Fusion Refinement Module. Our approach is inspired by PMD-Net's [11] refinement strategy, adapted here to improve multiple modalities and guide mirror prediction for M_{N-1} and M_N.

4.5. Loss Function

To train the model, we use the loss function ($\mathcal{L}_{\text{Final}}$):

$$\begin{aligned}
\mathcal{L}_A &= (\mathcal{L}_{\text{MirrorA}} \cdot \alpha) + (\mathcal{L}_{\text{MotionA}} \cdot \zeta) + (\mathcal{L}_{\text{EdgeA}} \cdot \gamma) \\
\mathcal{L}_B &= (\mathcal{L}_{\text{MirrorB}} \cdot \alpha) + (\mathcal{L}_{\text{MotionB}} \cdot \zeta) + (\mathcal{L}_{\text{EdgeB}} \cdot \gamma) \\
\mathcal{L}_{\text{Coherence}} &= (\text{Mirror}_{\text{Coherence}} \\
&\quad + \text{Edge}_{\text{Coherence}} + \text{Motion}_{\text{Coherence}}) \cdot \delta \\
\mathcal{L}_{\text{Final}} &= \mathcal{L}_{\text{Coherence}} + \mathcal{L}_A + \mathcal{L}_B
\end{aligned}$$

We discuss the loss component below:

Temporal Loss ($\mathcal{L}_{\text{Coherence}}$): To make predictions across consecutive frames more consistent over time, we use BCELossWithLogits (binary cross-entropy) between M_{N-1} and M_N, Mp_{N-1} and Mp_N, as well as E_{N-1} and E_N. We set δ to 1, calling this loss $\mathcal{L}_{\text{Coherence}}$. This loss is aimed at reducing temporal inaccuracies in edge prediction, motion mirror location, and mirror prediction between frames in a video. It ensures that predictions from our model's combined features remain consistent over time, addressing a current limitation in video mirror detection.

Mirror Binary Losses ($\mathcal{L}_{\text{MirrorA}}$ and $\mathcal{L}_{\text{MirrorB}}$): We use two BCELossWithLogits loss functions to predict mirrored regions in M_{N-1} ($\mathcal{L}_{\text{MirrorA}}$) and M_N ($\mathcal{L}_{\text{MirrorB}}$). These losses are then compared with the actual ground truth mirror maps. We set α to 3 for these losses, emphasizing the importance of accurately detecting mirrors in the model.

Motion Mirror Losses ($\mathcal{L}_{\text{MotionA}}$ and $\mathcal{L}_{\text{MotionB}}$): Two additional BCELossWithLogits loss functions are used to predict mirror regions from motion-guided image features Mp_{N-1} ($\mathcal{L}_{\text{MotionA}}$) and Mp_N ($\mathcal{L}_{\text{MotionB}}$). These losses are then compared with the ground-truth mirror maps. We set ζ to 1 for both of these losses.

Mirror Edge Losses ($\mathcal{L}_{\text{EdgeA}}$ and $\mathcal{L}_{\text{EdgeB}}$): Two BCELossWithLogits loss functions are utilized for predicting mirror boundaries in E_{N-1} ($\mathcal{L}_{\text{EdgeA}}$) and E_N ($\mathcal{L}_{\text{EdgeB}}$). These losses are compared against the ground-truth mirror edge maps. We set $\gamma = 2$ for each of these losses, emphasizing the precision of mirror edge prediction.

5. Experiments

5.1. Setups and Metrics

We implemented MG-VMD using PyTorch, trained and validated it against comparative models on a NVIDIA RTX 3090 GPU. In the data preprocessing phase, all images, ground-truth mirror maps, and edge maps were resized to a resolution of 224 × 224. Our model's ResNext-101 backbone and DeepLabV3 parameters were initialized using the pretrained ResNext-101 backbone and pretrained DeepLabV3 weights from VMD-Net [12].

We utilized a Stochastic Gradient Descent Optimizer with a starting learning rate of 9e-3, a momentum of 1.9, and a weight decay of 5e-4. Then, we applied an adaptive learning rate to the optimizer, interpolating the learning rate from 9e-3 to 3e-3 across epochs 1 to 15. The model underwent 15 epochs of training with a batch size of 8.

For validation, we employed four quantitative metrics: Mean Absolute Error (MAE↓), F-measure (F_β↑), Accuracy↑, and Intersection over Union (IoU↑). F-measure (F_β↑) is computed as follows:

$$F_\beta = \frac{1 + \beta^2(precision * recall)}{\beta^2 + recall}, \tag{1}$$

where β is set to 0.3, as suggested by previous research [11] [1]. The F-measure (F_β↑) assesses the harmonic mean between precision and recall.

5.2. Experimental Results

In Table 3 and Figure 9, we evaluate our proposed technique against 8 state-of-the-art methods in Video Salient Object Detection, and Image-Based/Video Mirror Detection. Using their respective pre-trained weights, we fine-tune and validate these results on our proposed MMD dataset..

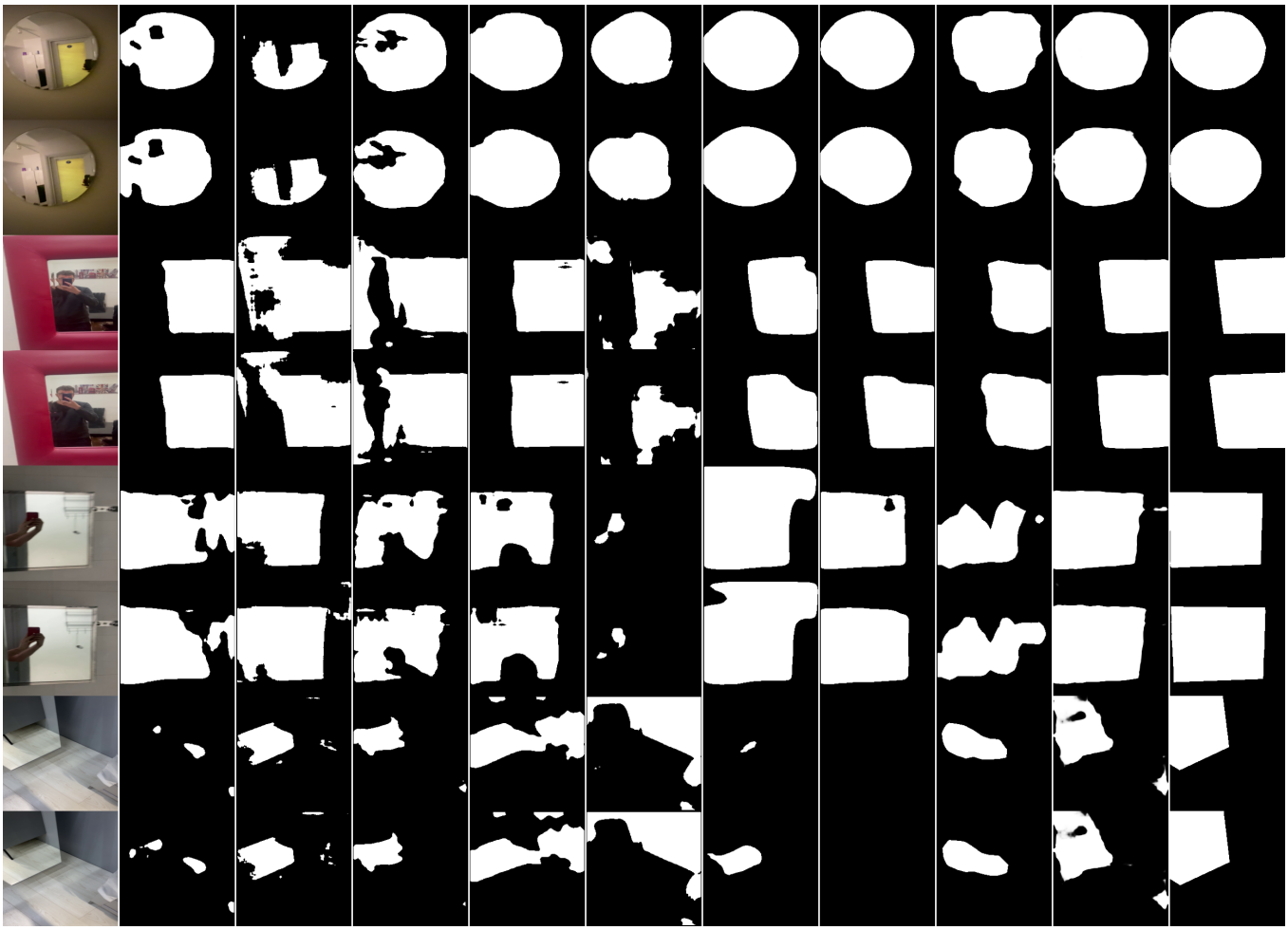

Figure 9. Qualitative results table comparing our proposed method against state-of-the-art Video Salient Object Detection, and Image-Based/Video Mirror Detection FS-Net [8], MGA [10], ALGRF [23], F3Net [9], UFO[17], MirrorNet[30], PMD-Net [11], VMD-Net[12] trained and validated on our proposed MMD dataset.

5.3. Comparison Against State-of-the-art

Our MG-VMD achieves better performance in Mean Absolute Error (MAE↓), F-measure (F_β↑), Accuracy↑, and Intersection over Union (IoU↑) when compared to state-of-the-art Video Salient Object Detection, Image-Based Mirror Detection and Video Mirror Detection methods fine-tuned and validated on our proposed Motion Mirror Dataset. We attribute the better performance to the exploiting of motion inconsistency cues across adjacent frames through our novel modules: Motion-Guided Edge Detection Module (MEDM) and Motion Attention Module (MAM). Figure 9 shows that our method is able to detect mirrors well in complex (*e.g.*, low contrast) scenarios, outperforming state-of-the-arts.

We also evaluate the generalizability of our model. In the Supplemental, we show results comparing MG-VMD and VMD-Net trained on our MMD dataset and tested on VMD-D [12]. MG-VMD outperforms VMD-Net, showing that our model has slightly better/close generalization performances.

5.4. Ablation

We conducted an ablation study to assess the effectiveness of the proposed modules in MG-VMD. Initially, we removed both the Motion Attention Module and the Motion-guided Edge Detection Module, resulting in the baseline with only the Refinement module, which incorporates convolutional layers. This configuration is referred to as the "baseline" for comparative analysis. The rationale behind evaluating the baseline's performance without the Optical Flow Coherence Block, Motion-Guided Edge Detection Module, and Motion Attention Module was to quantitatively gauge the efficacy of our deep-learning approach without leveraging inconsistent motion cues.

Table 4 shows the performance of the ablated models:

Models	Accuracy↑	MAE↓	F_β↑	IoU↑
MGA	0.566037	0.433963	0.489176	0.281285
UFO	0.796797	0.203204	0.797447	0.598641
FSNet	0.853287	0.146713	0.837732	0.707848
ALGRF	0.723772	0.276229	0.692829	0.502814
F3Net	0.838863	0.161137	0.847321	0.670618
PMD-Net	0.740909	0.259091	0.847211	0.424085
MirrorNet	0.835227	0.164772	0.838987	0.666242
VMDNet	0.854415	0.145585	0.812347	0.722634
Ours	0.872532	0.127468	0.869419	0.725130

Table 3. Quantitative results table comparing our proposed method against state-of-the-art Video Salient Object Detection, and Image-Based/Video Mirror Detection FS-Net [8], MGA [10], ALGRF [23], F3Net [9], UFO [17], PMD-Net [11], MirrorNet[30], VMD-Net[12] methods trained and validated on our proposed MMD dataset. Red, Blue and Green indicate the best, second and third best performances, respectively.

Models	Accuracy↑	MAE↓	F_β↑	IoU↑
baseline	0.857519	0.142481	0.837578	0.700855
b + MEDM	0.858176	0.141824	0.847853	0.707350
b + MAM	0.866218	0.133782	0.841614	0.713005
Ours	0.872532	0.127468	0.869419	0.725130

Table 4. Ablation Results: Qualitative results table comparing our proposed method (MG-VMD (Ours)) against ablated versions of our model: b + MAM, b + MEDM, and baseline (b), trained and validated on our proposed MMD dataset. Red, Blue and Green indicate the best, second and third-best performances, respectively.

"baseline (b)", "b + MEDM", "b + MAM", and "MG-VMD (Ours)". Both MAM and MEDM exploit motion inconsistency cues, demonstrating better performance compared to the baseline. The full model combines the complementary outputs from MAM and MEDM, yielding the best results.

Our baseline outperforms state-of-the-art VMD-Net [12] for two reasons. First, our model is fine-tuned on VMD-Net's [12] ResNext-101 backbone and DeepLabV3 weights (Section 5). Second, VMD-Net [12] assumes dual correspondence, but in real-life datasets, dual correspondence does not always occur. This shows the usefulness of our MMD dataset.

Table 4 shows that MEDM and MAM each contributes to the model individually. Without MEDM, mirror edges become noisy (*e.g.*, Figure 10 Column 4 Rows 2 and 4). Without MAM, MEDM may mistakenly detect object edges inside mirrors (*e.g.*, cabinet edges in Figure 10 Column 3 Row 2), leading to inaccurate mirror detection. (Best viewed by zooming in.) Both modules use motion features, complementing each other and boost performance (Figure 10 Column 5, and Table 4).

6. Conclusion

This paper introduces a novel deep-learning approach to tackle video mirror detection by leveraging inconsistent motion cues within and outside mirrored regions. We also present a benchmark video mirror dataset featuring consistent inter-frame motion across long videos. Our experimental results demonstrate that our proposed MG-VMD deep-learning method outperforms state-of-the-arts in Video Mirror Detection, Image-Based Mirror Detection, and Video Salient Object Detection. Furthermore, we conclude that our proposed Motion Mirror Dataset (MMD) is challenging compared to state-of-the-art Video Mirror Datasets.

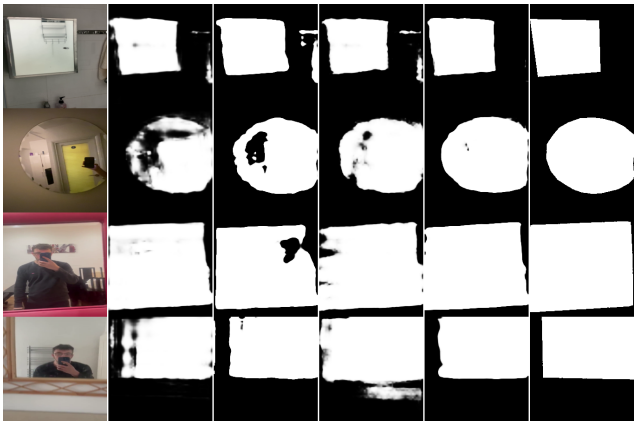

Figure 10. Qualitative results of the ablated models.

Figure 11. Failure cases. Our Method may not be good in scenarios containing both mirror regions and windows, due to windows irregular motion compared to a scene.

Our method is not without limitation. Figure 11 shows a scenario with both mirror and window in a scene. We found that similar to mirror regions, windows can have different motions from their surroundings. Thus, our method tends to over-predict window regions as mirrors. We plan to improve our design to mitigate this in the future.

Acknowledgements: Alex is supported by a Swansea GTA Research Scholarship. This project is in part supported by a GRF grant from the Research Grants Council of Hong Kong (Ref.: 11211223). We gratefully acknowledge the support of the HEFCW HERC fund (W21/21HE) for the provision of GPU equipment used in this research.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 6

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 4, 5

[3] Zheng Dong, Ke Xu, Ziheng Duan, Hujun Bao, Weiwei Xu, and Rynson W. H. Lau. Geometry-aware two-scale pifu representation for human reconstruction. In *NeurIPS*, 2022. 1

[4] Zheng Dong, Ke Xu, Yaoan Gao, Qilin Sun, Hujun Bao, Weiwei Xu, and Rynson W. H. Lau. Sailor: Synergizing radiance and occupancy fields for live human performance capture. *ACM Trans. Graph.*, 2023. 1

[5] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, 2003. 3

[6] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *CVPR*, 2022. 1

[7] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 3

[8] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, 2021. 2, 7, 8

[9] Qingming Huang Jun Wei, Shuhui Wang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020. 2, 7, 8

[10] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, 2019. 5, 7, 8

[11] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *CVPR*, 2020. 1, 2, 5, 6, 7, 8

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

[13] Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *CVPR*, 2020. 1

[14] Jing Liu, Jiaxiang Wang, Weikang Wang, and Yuting Su. Ds-net: Dynamic spatiotemporal network for video salient object detection. *arXiv:2012.04886*, 2022. 2

[15] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021. 1, 2

[16] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *CVPR*, 2020. 1

[17] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Trans. Multimedia*, 2023. 7, 8

[18] Hideki Tamura, Maki Tsukuda, Hiroshi Higashi, and Shigeki Nakauchi. Perceptual segregation between mirror and glass material under natural and unnatural illumination. *Journal of Vision*, 2016. 2

[19] Hideki Tamura, Hiroshi Higashi, and Shigeki Nakauchi. Multiple cues for visual perception of mirror and glass materials. *Journal of Vision*, 2017. 2

[20] Hideki Tamura, Hiroshi Higashi, and Shigeki Nakauchi. Dynamic visual cues for differentiating mirror and glass. *Scientific Reports*, 2018. 2

[21] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 2021. 1

[22] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson W.H. Lau. Mirror detection with the visual chirality cue. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1, 2, 5

[23] Yi Tang, Yuanman Li, and Guoliang Xing. Video salient object detection via adaptive local-global refinement. *arXiv:2104.14360*, 2021. 2, 7, 8

[24] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3, 5

[25] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *CVPR*, 2022. 1

[26] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson Lau. Learning to detect instance-level salient objects using complementary image labels. *Int. J. Comput. Vis.*, 2022. 1

[27] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. *arXiv:1704.05737*, 2017. 2

[28] Wenguan Wang, Xiankai Lu, Jianbing Shen, David Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 2

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 4, 5

[30] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W. H. Lau. Where is my mirror? In *ICCV*, 2019. 1, 2, 7, 8

[31] 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 *ACM MM*, 2023. 1

Effective Video Mirror Detection with Inconsistent Motion Cues
Supplementary Material

Alex Warren[1], Ke Xu[2], Jiaying Lin[2], Gary K.L. Tam[1], Rynson W.H. Lau[1,2]
Department of Computer Science, Swansea University[1] and City University of Hong Kong[2]
alex.warren@swansea.ac.uk, kkangwing@gmail.com, jiayinlin5-c@my.cityu.edu.hk
k.l.tam@swansea.ac.uk, Rynson.Lau@cityu.edu.hk

This supplementary material provides additional details and comparisons of our implementations. These include:

- A detailed description of our data collection process.
- Further statistical analysis of the dataset.
- Quantitative results and analysis of the generalization capabilities of our method and the state-of-the-art models, trained on MMD (Ours) and tested on VMD-D [6].
- Additional qualitative results showcasing the performance of our method, and
- A video highlighting the results of the ablation study.
- Evaluation on Outdoor Video
- Design Choice of Optical Flow Coherence Block

1. Motion Mirror Dataset

1.1. Dataset Creation

We embarked on the development of a comprehensive video mirror dataset tailored for real-world applicability. This dataset comprises thirty-seven 9-second videos (sequence of images without audio) intentionally capturing diverse lighting and environmental conditions. An essential consideration was incorporating consistent motion in these videos to simulate scenarios akin to those encountered in drone footage.

To compile the dataset, we recruited six individuals who voluntarily participated in the data collection process. These participants received explicit instructions to use contemporary mobile phone cameras, ensuring a minimum video resolution of 1080p (1920x1080px). They were directed to record mirror videos in various locations, including homes (specifically bathrooms, living rooms, bedrooms, and hallways), department stores, gyms, and within cars. Additionally, participants were guided to record in different lighting conditions, spanning daytime, natural and unnatural lighting, night-time, and darker environments. The dataset encompasses a wide range of features and conditions, making it relevant for video mirror detection in diverse real-world scenarios.

The manual annotation process was carefully executed. Every third frame of the videos underwent annotation, and interpolation between frames was performed using the method outlined in [1]. Furthermore, a depth-first-search algorithm was computed on each frame to obtain edge features. Each annotated frame, including mirror and edge annotations, underwent manual verification to ensure it attained ground truth quality.

2. Dataset Comparison

Previously, Jiaying Lin *et al*. [6] proposed the VMD-D dataset, which consists of 269 (≈1.9s) videos. These videos contain mostly stationary and abruptly moving scenes from department stores. Our proposed dataset MMD consists of videos that are 4.8x longer, containing examples from diverse scenes (*e.g*., kitchens, hallways, living/bath/bedrooms), with various mirror occlusions, multi-mirrors, and lower contrast between mirror/non-mirror.

Table 1 compares some statistical information between the proposed MMD dataset and the existing VMD-D dataset.

Dataset	FPS	Total Frames	Mean Video Length		Videos	
			Frames	Seconds	Training	Testing
VMD-D	30	14987	55.71	1.86	143	126
MMD (Ours)	30	9727	262.77	8.75	18	19

Table 1. Statistical analysis table comparing our proposed dataset (MMD) with the existing dataset (VMD-D) [6].

3. Generalization on VMD-D dataset [6]

We further compare the generalizing performance of our proposed model, MG-VMD, with VMD-Net [6]. We train both models on our dataset (MMD) and test on the VMD-D [6] dataset. Table 2 shows that the generalization performances of the two models are very close, with ours slightly better on Accuracy and MAE.

Model	Accuracy↑	MAE↓	F_β↑
Ours	0.666342	0.333658	0.385283
VMD-Net	0.654699	0.3453	0.38978

Table 2. Quantitative results table comparing our proposed method with the state-of-the-art VMD-Net [6]. The two models are trained on our proposed dataset (MMD) and tested on the VMD-D [6] dataset directly. The results show that the generalization ability of the two models are similar. Red and Blue indicate the best and second-best performances.

4. Further Qualitative Results on MMD

Figure 12 further presents more qualitative results, comparing our proposed model with eight state-of-the-art methods from Video Salient Object Detection and Image-Based/Video Mirror Detection.

When evaluating on the video mirror detection task, our method demonstrates better temporal consistency compared to VMD-Net [6]. This is evident in the 3rd - 8th rows, and 11th to 12th rows. This enhanced temporal consistency can be attributed to our model leveraging inconsistent motion cues, and the use of the optical flow coherence block and coherence loss. Single-frame mirror detection method (PMD-Net [5]) occasionally fails to detect mirrors in these videos, specifically in the 11th row.

In addition, our method exhibits greater robustness and stability in identifying mirror regions and their boundaries compared to video salient object detection methods.

4.1. Qualitative Video Showcase

Along with this supplementary material, we present a video titled "`MG-VMD_qualitative_MMD.avi`", demonstrating the qualitative performance of our video mirror detection method under the ablation study. The video comprises six rows, with the last four each representing a different stage of our method, with the inclusion of the respective Motion-guided Edge Detection Module (MEDM) and Motion Attention Module (MAM):

1. Current Frame
2. Ground Truth
3. Baseline
4. Baseline + MEDM
5. Baseline + MAM
6. MG-VMD (combining MEDM and MAM)

The samples featured in the video are drawn from our proposed MMD dataset. The final exported video maintains a frame rate of 30 fps, and each prediction map has a resolution of (224×224px).

5. Evaluation on Outdoor Video

Our proposed method primarily focuses on indoor scenes containing mirrors. We note that mirrors are more commonly found indoors. Indoor mirrors are specifically useful for indoor 3D reconstruction and robotic navigation. However, in Table 3, we show that our method performs better than VMD-Net on handling outdoor scenes containing mirror regions.

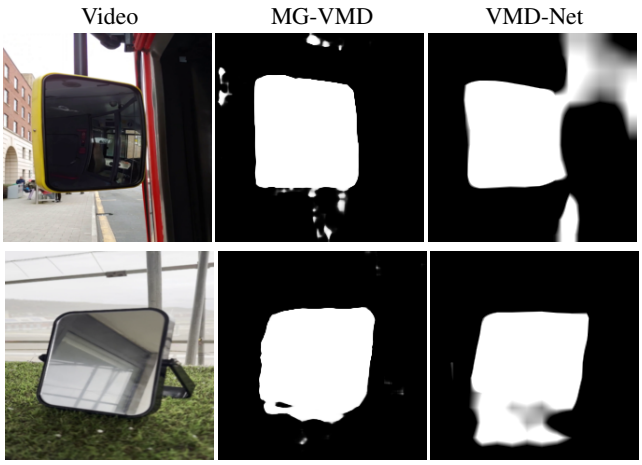

Table 3. Comparison of qualitative results between our proposed method (MG-VMD) and VMD-Net [6] on outdoor scenes featuring mirror regions.

6. Justification of Design Choice of Optical Flow Coherence Block

Models	Accuracy↑	MAE↓	F_β↑	IoU↑
N-1 to N-2	0.833256	0.166744	0.76463	0.632305
No Flow Extrapolation	0.846201	0.153799	0.840227	0.681888
MG-VMD (Ours)	0.872532	0.127468	0.869419	0.72513

Table 4. Qualitative results table comparing our proposed method, against different ways of modelling motion coherence, trained and validated on our proposed MMD dataset. Red, Blue and Green indicate the best, second and third best performances, respectively.

We offer empirical justification for our design choices within the Optical Flow Coherence block through quantitative evaluation of various motion-driven designs. Table 4 demonstrates that our current linear extrapolation design outperforms alternative methods in modeling cohesive motion for video mirror detection. In Table 4 Row 1, we note that the performance of using optical flow from N-1 to N-2 ("N-1 to N-2") to penalize the inconsistency between mirror detection results performs worse than our linear extrapolation method. In addition, in Table 4 Row 2, we show that our linear extrapolation performs better than a dual inference of the frozen deep learning optical flow estimation for each branch (N-2 to N-1) and (N-1 to N) ("No Flow Extrapolation"). We find that such dual inference accumulates errors between optical flow vector fields. Overall, we observe that both of these methods suffer from pixel-level errors in optical flow vector fields, attributed to low contrast

between mirror and non-mirror regions in our dataset, as well as frozen weights in the optical flow. Our simple linear extrapolation design, however, stabilizes flows, reduces pixel-level errors, and ensures temporal consistency within our model.

References

[1] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Eur. Conf. Comput. Vis.*, 2022. 1

[2] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Int. Conf. Comput. Vis.*, 2021. 4

[3] Qingming Huang Jun Wei, Shuhui Wang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020. 4

[4] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Int. Conf. Comput. Vis.*, 2019. 4

[5] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 4

[6] Jiaying Lin, Xin Tan, and Rynson W.H. Lau. Learning to detect mirrors from videos via dual correspondences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 2, 4

[7] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Trans. Multimedia*, 2023. 4

[8] Yi Tang, Yuanman Li, and Guoliang Xing. Video salient object detection via adaptive local-global refinement, 2021. 4

[9] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W. H. Lau. Where is my mirror? In *Int. Conf. Comput. Vis.*, 2019. 4

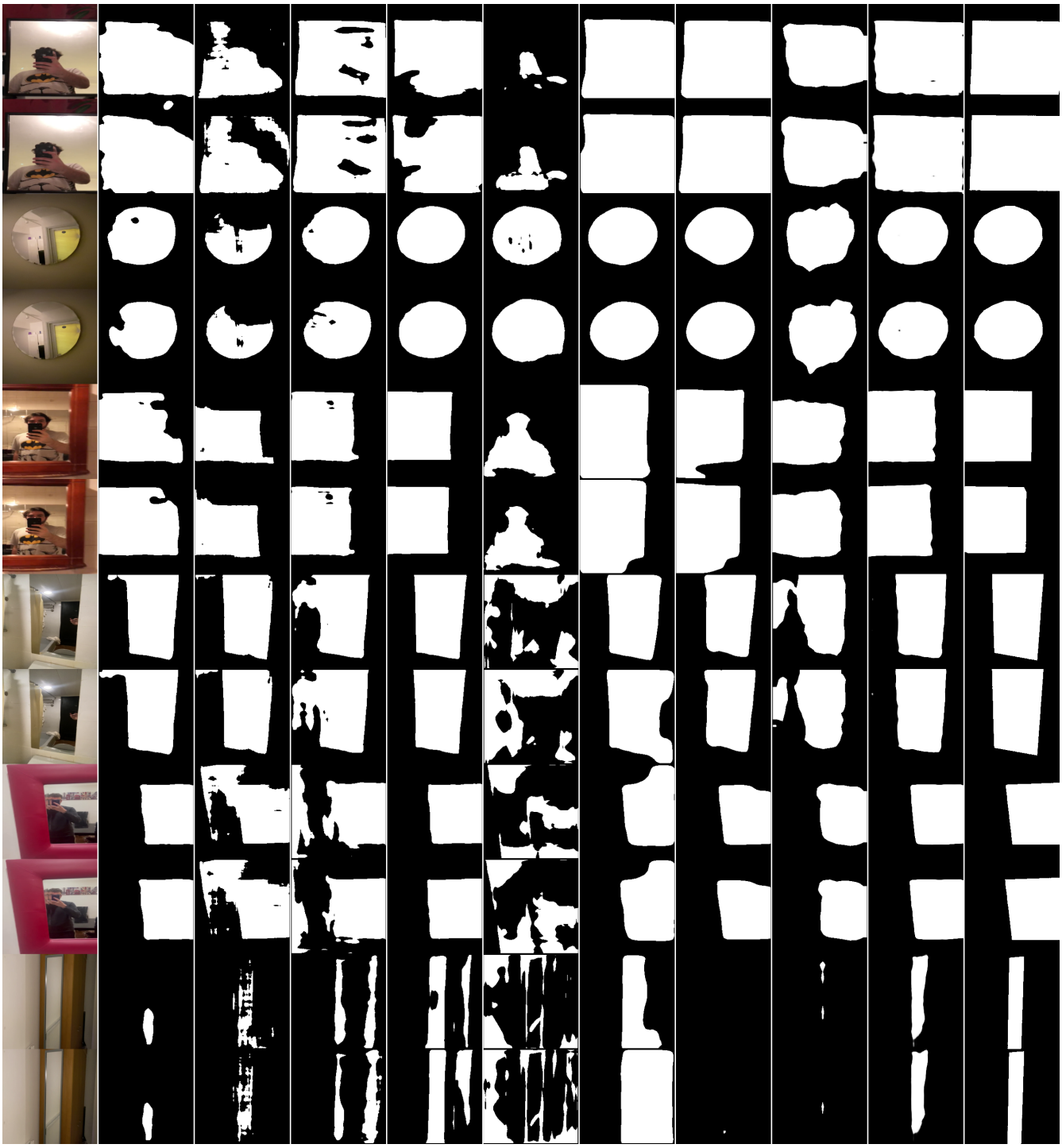

Figure 12. A further comprehensive qualitative results table comparing our proposed method with state-of-the-art video salient object detection, as well as image-based/video mirror detection (namely, FS-Net [2], MGA [4], ALGRF [8], F3Net [3], UFO [7], PMD-Net [5], MirrorNet [9], VMD-Net [6]). The models were trained and validated on our proposed MMD dataset.