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.

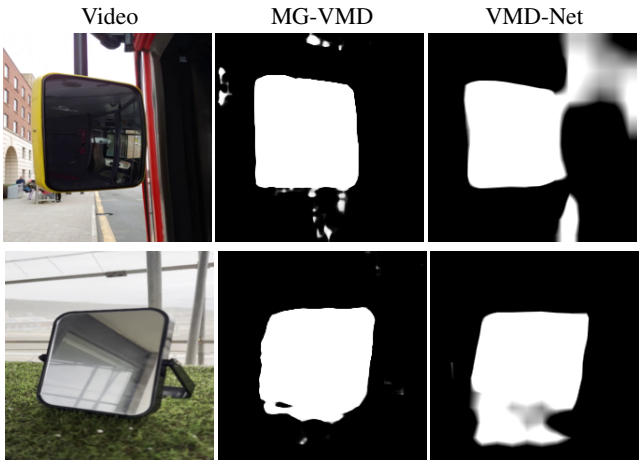

Table 3. Comparison of qualitative results between our proposed method (MG-VMD) and VMD-Net [6] on outdoor scenes featuring mirror regions.

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

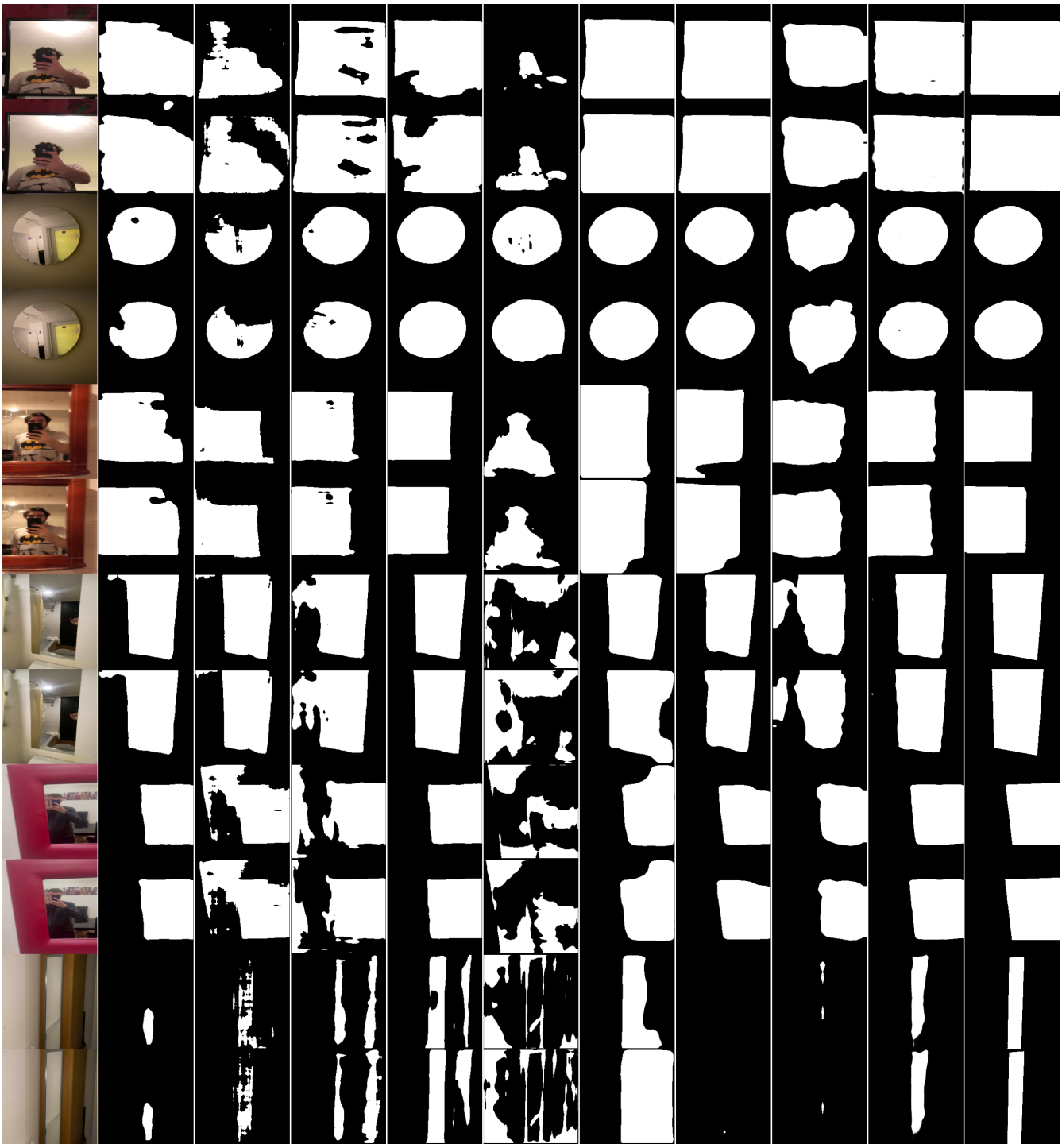

Figure 12. A further comprehensive qualitative results table comparing our proposed method with state-of-the-art video salient object detection, as well as image-based/video mirror detection (namely, FS-Net [2], MGA [4], ALGRF [8], F3Net [3], UFO [7], PMD-Net [5], MirrorNet [9], VMD-Net [6]). The models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