FUSION OF SHORT-TERM AND LONG-TERM ATTENTION FOR VIDEO MIRROR DETECTION

Mingchen Xu, Yu-Kun Lai, Ze Ji and Jing Wu

Cardiff University, Cardiff, United Kingdom

***Abstract*–Techniques for detecting mirrors from static images have witnessed rapid growth in recent years. However, these methods detect mirrors from single input images. Detecting mirrors from video requires further consideration of temporal consistency between frames. We observe that humans can recognize mirror candidates, from just one or two frames, based on their appearance (e.g. shape, color). However, to ensure that the candidate is indeed a mirror (not a picture or a window), we often need to observe more frames for a global view. This observation motivates us to detect mirrors by fusing appearance features extracted from a short-term attention module and context information extracted from a long-term attention module. To evaluate the performance, we build a challenging benchmark dataset of 19,255 frames from 281 videos. Experimental results demonstrate that our method achieves state-of-the-art performance on the benchmark dataset.**

***Index Terms*— mirror detection, information fusion, short-term attention, long-term attention, benchmark**

1. INTRODUCTION

Mirrors are commonly seen in environments. More and more attention has been drawn to mirror detection in computer vision. It is because, on the one hand, detecting mirrors can benefit scene-understanding tasks. The reflection of the mirror can provide hints for locating objects [1] with 3D information [2]. reconstructing human pose [2], and reconstructing scenes [3]. On the other hand, ignoring mirrors may affect the performance of some computer vision tasks. For example, a service robot may treat reflected objects as real ones. Therefore, it is important for computer vision systems to be able to detect and segment mirrors from input images.

Research on detecting mirrors from static images has witnessed rapid growth in recent years. Existing methods exploit context contrast [6], reflection relation [7], semantic relation [8], depth information [9, 10, 11], visual chirality [12] and symmetry relation [13] to detect mirrors. However, these methods detect mirrors from single input images. To detect mirrors from videos requires further consideration of temporal consistency between frames. Recently, Lin *et al.* [5] proposed the first video mirror detection model, VMDNet, which extracts correspondence between the mirrors and the surroundings at both the intra-frame and the inter-frame levels. However, this method relies on the extraction of the correspondence and may fail when the correspondence cannot be established. For example, the top two rows in Fig. 1 show the same mirror hanging on the wall. However, the VMDNet is confused by the different mirror reflections in the two frames, and cannot detect the mirror correctly. Moreover, the VMDNet will predict other objects as mirrors since it separately considers correspondences at the short-term and long-term levels. For example, the bottom two rows in Fig.1 shows that the VMDNet fails to distinguish the painting and mirror, as correspondences for both of them are extracted.

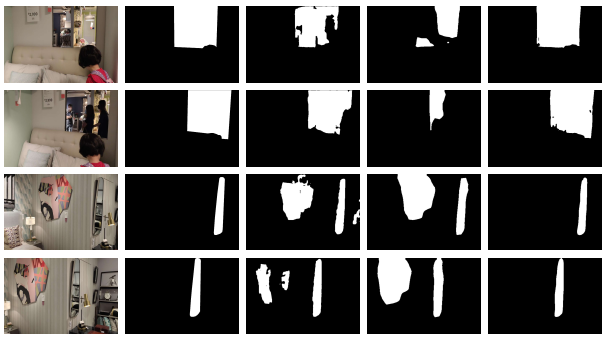

Fig. 1. Two normal scenarios where existing methods [4] [5] fail. HetNet[4] is a single-image mirror detection method, and VMDNet [5] is designed for video mirror detection. Compared to HetNet and VMDNet, our method can detect the mirror regions correctly by fusing short-term information and long-term information.

To address the above problems, we propose a novel approach to detect mirrors in videos. We observe that humans can recognize the appearance (e.g., shape, color) of candidate mirrors from just one or two frames. However, to make sure that the candidate is indeed a mirror (not a picture or window), we often need to see more frames to have a global view. This observation motivates us to extract appearance features at a micro (short video clips) and to extract context features at a macro view (long video clips) , and then combine them to predict the mirror. Our approach is different from VMDNet, which utilizes long-term information as an auxiliary task in

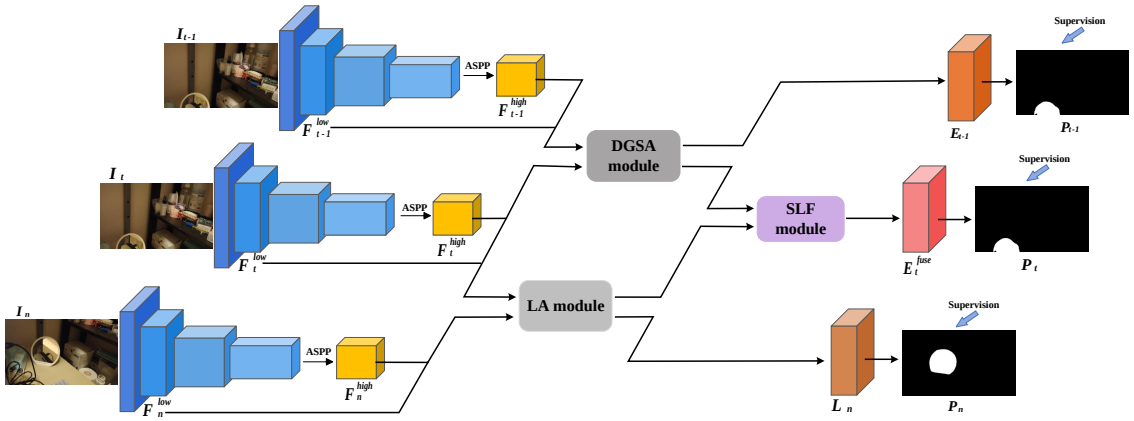

Fig. 2. The architecture of our proposed model. We first feed three frames from the same video to the backbone feature extractor, then the DGSA module to extract appearance features from adjacent frames, and an LA module to extract context features from long video clips parallelly. Second, the SLF module fuses short-term attention and long-term attention to finalize the mirror region.

the first stage, and separately considers short-term and long-term information in the second stage. Our approach tries to combine short-term information and long-term information to better predict the mirror map.

Our method consists of three modules: 1) a Dual Gated Short-term Attention (DGSA) module to extract appearance features from adjacent frames; 2) a Long-term Attention (LA) module to extract context features from long video clips to obtain position information of mirrors; and 3) a Short-Long Fusion (SLF) module to fuse appearance features and context features to finalize the mirror region.

To evaluate the performance of video mirror detection, we also construct a challenging benchmark dataset that includes a variety of scenes from real living and working environments. Most of our data are from two public datasets: NYUv2 [14], ScanNet [15], and others are captured by ourselves. Our dataset has 19,255 frames from 281 videos with pixel-wise annotation. Compared to the first video mirror dataset (VMD) proposed by [5], of which 95% are collected from furniture stores, our dataset covers about 20 scene types (e.g., gym, lift, kitchen) which largely increases the diversity of data.

Our contributions can be summarized as:

- We propose a novel transformer network for video mirror detection. It consists of three modules (DGSA module, LA module, and SLF module) to support the extraction and fusion of short-term and long-term attention to improve video mirror detection.
- We construct a challenging benchmark dataset that contains 19,255 frames from 281 videos and pixel-wise annotations from a variety of real-world scenes.
- We have conducted extensive experiments on both the VMD dataset and our dataset to demonstrate that our method achieves state-of-the-art performance.

2. RELATED WORK

2.1. Mirror Detection

In recent years, many works [6, 7, 8, 12, 4, 9, 10, 11], are proposed to detect mirrors from single images. They exploit context contrast [6], reflection relation [7], semantic relation [8], depth information [9, 10, 11], visual chirality [12] and symmetry relation [13] to detect mirrors. Although the single-image mirror detection model achieves reliable results, their performance on video is not good because of insufficient exploitation of temporal information. Recently, Lin *et al.* [5] propose the first video detection network, named VMDNet. It focuses on extracting mirror correspondence at intra-frame and inter-frame levels.

3. METHOD

3.1. Overall Structure

Fig. 2 shows the architecture of the proposed FusionFormer. To enable extraction and fusion of short-term and long-term attention, our model first takes three frames from the same video as input. Two are adjacent frames and the third I_n is a random other frame. Then, we employ the Mix Transformer [16] as the encoder, which can encode long-range dependencies. Adhere to the approach in [5], we utilize the second scale for the low-level features (F_i^{low}) and the fifth scale after the atrous spatial pyramid (ASPP) for the high-level features F_i^{high}. After obtaining features from three input frames, we

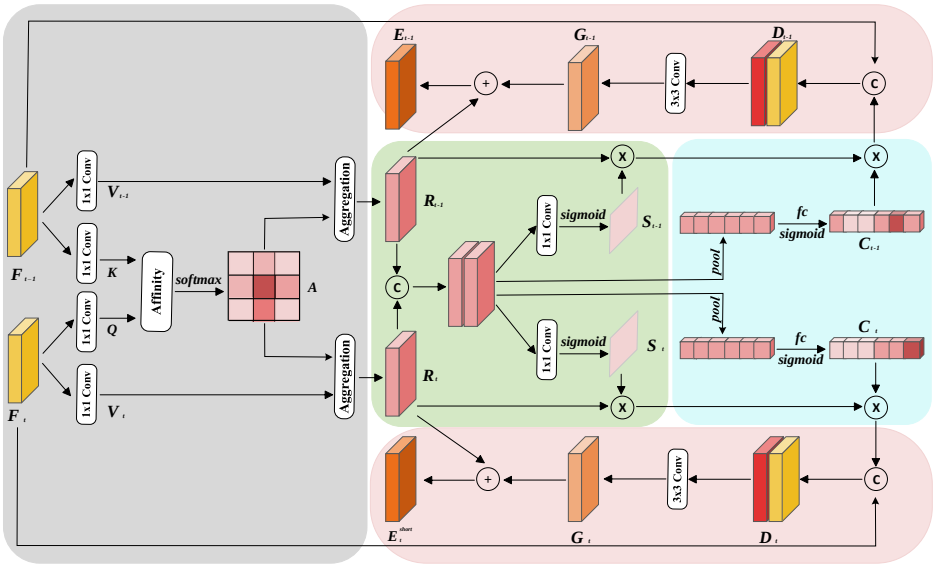

Fig. 3. The schematic illustration of Dual Gated Short-term Attention (DGSA) module. The grey part represents the short-term attention (SA) block. Pink parts represent the fusion blocks. The green and blue parts represent the spatial-wise gate (SG) block and the channel-wise gate (CG) block, respectively.

assign the Dual Gated Short-term Attention (DGSA) module on the low-level $F^{low}_{i\in\{t-1,t\}}$ and high-level features $F^{high}_{i\in\{t-1,t\}}$ to extract appearance features from adjacent frames, and the Long-term Attention (LA) module on low-level features $F^{low}_{i\in\{t,n\}}$ and high-level features $F_{i\in\{t,n\}^{high}}$ to extract context features from long video clips at the same time. Finally, the Short-Long Fusion (SLF) module combines appearance features and context features selectively to produce the final mirror prediction P_t.

3.2. Dual Gated Short-term Attention Module

The DGSA module aims to extract appearance features from the short-term information. It is inspired by the cross attention module proposed in [5], which can extract correspondences between the content inside and outside of the mirror at the intra-frame level and the inter-frame level. However, occlusions, appearance changes, etc., may affect the correspondence extraction. Therefore, we propose to weigh the mirror correspondence features differently.

Fig.3 shows the schematic illustration of the DGSA. We use $F_{i\in\{t-1,t\}}$ to denote $F^{low}_{i\in\{t-1,t\}}$ or $F^{high}_{i\in\{t-1,t\}}$ to be visual clear. Our DGSA module consists of four blocks: a short-term attention (SA) block, a spatial-wise gate (SG) block, a channel-wise gate (CG) block, and two fusion blocks. Given the input features $F_{i\in\{t-1,t\}}$, we first use the SA block to extract short-term correspondence features (denotes R_{t-1}, R_t):

$$R_{t-1} = \omega_{t-1} \sum_{i}^{(2H+2W-1)\times(W\times H)} AV_{t-1}, \quad (1)$$

$$R_t = \omega_t \sum_{i}^{(2H+2W-1)\times(W\times H)} AV_t, \quad (2)$$

where $\mathbf{A}$ is the correspondence attention map. ω_{t-1} and ω_t are the learnable parameters. Then, we concatenate the R_{t-1}, R_t and feed them to the SG block and the CG block to produce spatial gated mask S_{t-1}, S_t and channel gated mask C_{t-1}, C_t.

In the fusion block, we first use dual gated attention features D_{t-1}, D_t to refine the original features F_{t-1}, F_t, and then fuse the refined features G_{t-1}, G_t with the correspondence features R_{t-1}, R_t to obtain enhanced dual gated short-term attention features E_{t-1}, E_t. The fusion block process can be formulated as:

$$E_{t-1} = R_{t-1} + Conv_{3\times3}(Cconcat(F_{t-1}, D_{t-1})), \quad (3)$$

$$E_t = R_t + Conv_{3\times3}(Cconcat(F_t, D_t)), \quad (4)$$

3.3. Short-long Fusion Module

The SLF module is designed to fuse short-term features with long-term features to further focus on the mirror with a global view. The reason to take long-term features into account is that we notice the mirror frequently appears throughout the whole video. Here, we utilize the LA module, instead of DGSA, to obtain the long-term relation features because we find that the dual gated mechanism may be confused by the mirror appearance changes in long video clips. LA module

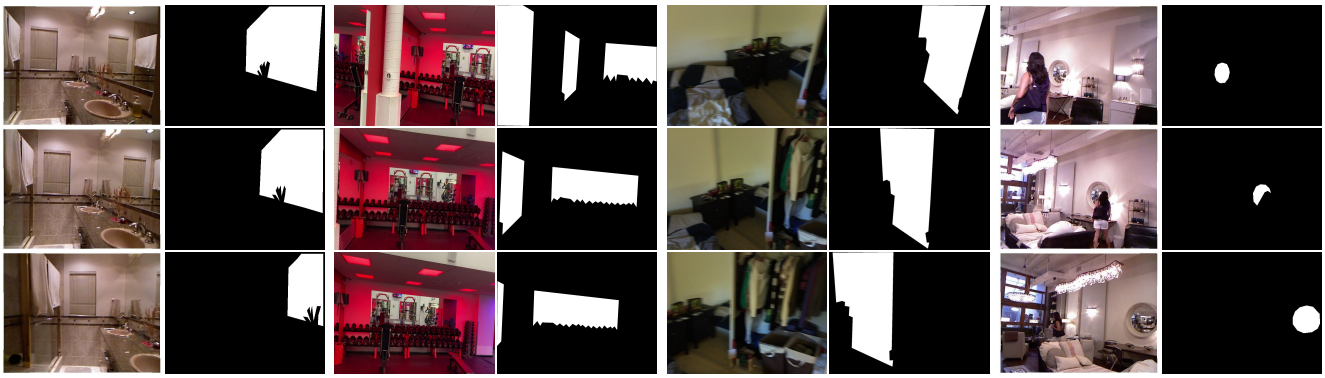

Fig. 4. Videos in our ViMirr dataset show high diversity and low similarity. They cover lots of daily scenes.

follows the design of the cross attention module [5]. The difference is that we are using it to extract the long-term correspondence, not the short-term correspondence.

Fig. 5 shows the architecture of the SLF module. We weight the enhanced short-term attention features E_t^{short} with the long-term attention features R_t^{long}. In this way, the correspondence features E_t^{fuse} are extracted, which encodes both the appearance of the mirror from the short-term features and the position of the mirror from the long-term features.

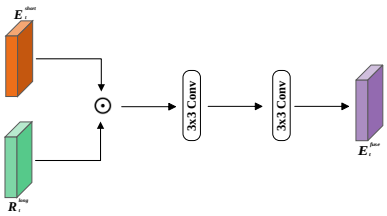

Fig. 5. Details of the Short-long Fusion (SLF) module.

3.4. Loss function

Following [17], we adopt the binary cross-entropy (BCE) and the Lovasz-hinge loss to supervise the training of the mirror maps:

$$\mathcal{L} = \sum_{i}^{i\in\{t-1,t,n\}} \mathcal{L}_{hinge}(P_i, G_i) + \mathcal{L}_{bce}(P_i, G_i), \tag{5}$$

where $\mathcal{L}_{hinge}$, $\mathcal{L}_{bce}$ are the lovasz-hinge loss and the binary cross-entropy (BCE) loss. P_i and G_i are the final predicted map and the ground truth of frames.

4. EXPERIMENTS

4.1. Datasets

Recently, Lin *et al.* proposed the first video mirror dataset (VMD) in [5], which contains 14,987 frames from 269 videos with corresponding annotated masks. However, we notice that their data were mostly collected from similar scenes. In particular, more than 95% of their data are collected from furniture stores (e.g. IKEA). This limits the diversity of the data, and will affect the generalization of the model to other scenes. Following [7], we use SSIM [21] to study the similarity of videos in VMD. As the frames are similar in the same video, we use the first frame in each video to calculate the similarity score. Our ViMirr obtains 39.54% similarity score, which is much lower than the 51.21% of the VMD dataset. The details of the similarity score calculation are given in Section 1.2 of the supplemental material.

To address the limitations of VMD dataset, we construct the ViMirr dataset, which has 19,255 frames from 276 videos. Fig. 4 shows some example video frames in ViMirr. To cover diverse realistic scenes, we studied five existing widely used datasets (*i.e.* Matterport3D [22], NYUv2 [14], ScanNet [15], DAVIS [23] and YouTube-VOS [24]), and manually selected 78 videos from NYUv2 and 126 videos from ScanNet, which contain mirrors in the videos. The indices of the videos selected are provided in Section 1.1 of the supplemental material. Moreover, we captured 13 videos to provide more popular scenarios (e.g., gym, lift) in daily life. Some examples we captured are given in Section 1.3 of the supplemental material. For both the collected and captured videos, we then manually annotated the mirrors in each frame. Example annotations can be seen in Fig. 4.

4.2. Implementation Details

The model was implemented in PyTorch [25] and trained on a PC with an NVIDIA RTX 4090 GPU card. During training, the images are resized to 512×512. We use Mix Transformer (MiT) [16] pre-trained on ADE20K [26, 27] dataset as the backbone network to extract image features. We adopt AdamW [28] with a weight decay of 5×10^{-4} as the optimizer. The base learning rate, batch size, and the number of training epochs are 0.00001, 5, and 15, respectively.

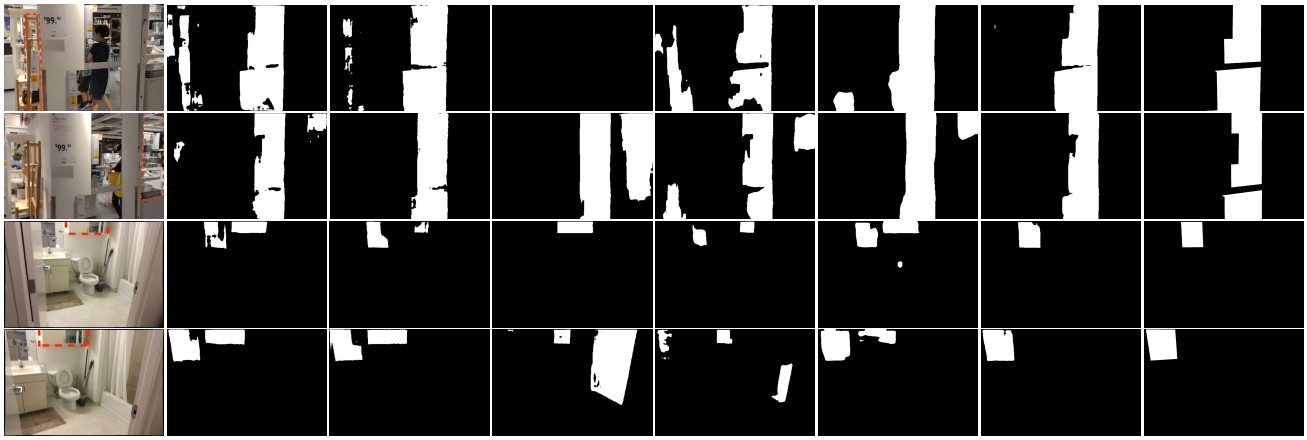

Fig. 6. Visual results of our model, compared with relevant state-of-the-art methods.

Table 1. Quantitative results on the VMD dataset (second column) and our benchmark (third column). The best results are shown in bold.

Method	VMD				ViMirr			
	IoU↑	F_β ↑	Accuracy↑	MAE↓	IoU↑	F_β ↑	Accuracy↑	MAE↓
TVSD [18]	0.3060	0.5343	0.8160	0.1839	0.1413	0.3394	0.8605	0.1394
SCOTCH [19]	0.5949	0.7281	0.8766	0.1233	0.6289	0.7596	0.9299	0.0702
GDNet [20]	0.5576	0.7335	0.8820	0.1179	0.5335	0.7118	0.9200	0.0799
MirrorNet [6]	0.5417	0.7506	0.8787	0.1211	0.4671	0.7015	0.9055	0.0944
PMD [7]	0.5309	0.7823	0.8771	0.1229	0.5258	0.7417	0.9233	0.0766
SANet [8]	0.4908	0.7202	0.8755	0.1245	0.4479	0.6286	0.9024	0.0975
HetNet [4]	0.5145	0.7547	0.8726	0.1274	0.4544	0.6825	0.9063	0.0935
VMD [5]	0.5673	0.7873	0.8950	0.1052	0.4224	0.7001	0.9096	0.0903
Ours	**0.6343**	**0.8104**	**0.9004**	**0.0995**	**0.6455**	**0.8261**	**0.9515**	**0.0484**

4.3. Evaluation Metrics and Comparison with the State-of-the-arts Methods

We compare our method with state-of-the-art methods from four relevant fields: TVSD [18] and SCOTCH [19] for video shadow detection , GDNet [20] for single-image glass detection, MirrorNet [6], PMD [7], SANet [8] and HeNet [4] for single-image mirror detection and VMDNet [5] for video mirror detection, and the metrics we use are: intersection over union (IoU), F-measure (F_β), pixel-accuracy, and mean absolute error (MAE).

Table 1 shows the quantitive results on the VMD dataset and the proposed ViMirr dataset. Our method achieves the best performance on all four metrics. Fig. 6 shows the visual comparisons. We can see that the image sequences contain some regions (e.g. wood shelf or the door-like area of the first two rows and cabin in the third and fourth rows where red dotted lines circles) that look like mirrors. VMDNet tends to detect these regions as mirrors, while our method can differentiate them well.

Table 2. Ablation study results, trained and tested on the VMD dataset. "Baseline" denotes our network without all proposed modules. "CA" is the cross attention module proposed in [5]. "DGSA" is our dual-gated short-term attention module. "SLF" is our short-long fusion module.

Method	IoU↑	F_β ↑	MAE↓	Accuracy↑
Baseline	0.6075	0.7676	0.1056	0.8943
+CA	0.6126	0.7919	0.1054	0.8946
+DGSA	0.6147	0.8017	0.1045	0.8954
+DGSA+SLF	0.6343	0.8104	0.0995	0.9004

4.4. Ablation study

We carried out ablation studies to demonstrate the effectiveness of our model. The last row in Table 2 shows that our final model with DGSA module and SLF module outperforms other baselines on all four metrics. The CA module brings improvements of baseline which shows the effectiveness of the spatial and temporal correspondence features. Compared with it, the DGSA module further improves the baseline, es-

pecially on F_β by filtering the dual correspondence features. By fusing long-term correspondence features, the SLF module significantly benefits the mirror video detection tasks from a global view. A visual example of the ablation study is provided in Section 2 of the supplemental material.

5. CONCLUSION

In this paper, we have proposed a transformer network for Video Mirror Detection. It detects mirrors by fusing appearance features extracted from a short-term module and context information extracted from a long-term attention module. In addition, we construct a challenging benchmark that includes 19,255 frames from 281 videos covering a variety of daily scenes. Our experimental results demonstrate that the proposed model achieves state-of-the-art performance on both the VMD dataset and the benchmark.

6. REFERENCES

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

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

[3] Thomas Whelan, Michael Goesele, Steven J Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, Richard A Newcombe, M Goesele, et al., "Reconstructing scenes with mirror and glass surfaces.," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 102–1, 2018.

[4] Ruozhen He, Jiaying Lin, and Rynson WH Lau, "Efficient mirror detection via multi-level heterogeneous learning," in *AAAI*, 2023, vol. 37, pp. 790–798.

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

[6] "Where Is my mirror?," in *ICCV*, 2019, pp. 8809–8818.

[7] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau, "Progressive mirror detection," in *CVPR*, 2020.

[8] Huankang Guan, Jiaying Lin, and Rynson W H Lau, "Learning Semantic Associations for Mirror Detection," in *CVPR*, 2022, pp. 5941–5950.

[9] Daehee Park and Yong Hwa Park, "Identifying Reflected Images from Object Detector in Indoor Environment Utilizing Depth Information," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 635–642, 2021.

[10] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei, "Depth-Aware Mirror Segmentation," in *CVPR*, 2021, pp. 3043–3052.

[11] Jiaqi Tan, Weijie Lin, Angel X. Chang, and Manolis Savva, "Mirror3D: Depth refinement for mirror surfaces," in *CVPR*, 2021, pp. 15985–15994.

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

[13] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson WH Lau, and Wangmeng Zuo, "Symmetry-aware transformer-based mirror detection," in *AAAI*, 2023, vol. 37, pp. 935–943.

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

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

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

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

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

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

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

[21] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[23] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[24] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *ECCV*.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin,

Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[26] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017, pp. 633–641.

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

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