Learning to Detect Mirrors from Videos via Dual Correspondences

Jiaying Lin[1]* Xin Tan[2]* Rynson W.H. Lau[1]†
[1]City University of Hong Kong [2]East China Normal University
jiayinlin5-c@my.cityu.edu.hk, xtan@cs.ecnu.edu.cn, Rynson.Lau@cityu.edu.hk

Abstract

Detecting mirrors from static images has received significant research interest recently. However, detecting mirrors over dynamic scenes is still under-explored due to the lack of a high-quality dataset and an effective method for video mirror detection (VMD). To the best of our knowledge, this is the first work to address the VMD problem from a deep-learning-based perspective. Our observation is that there are often correspondences between the contents inside (reflected) and outside (real) of a mirror, but such correspondences may not always appear in every frame, e.g.*, due to the change of camera pose. This inspires us to propose a video mirror detection method, named VMD-Net, that can tolerate spatially missing correspondences by considering the mirror correspondences at both the intra-frame level as well as inter-frame level via a dual correspondence module that looks over multiple frames spatially and temporally for correlating correspondences. We further propose a first large-scale dataset for VMD (named VMD-D), which contains 14,987 image frames from 269 videos with corresponding manually annotated masks. Experimental results show that the proposed method outperforms SOTA methods from relevant fields. To enable real-time VMD, our method efficiently utilizes the backbone features by removing the redundant multi-level module design and gets rid of post-processing of the output maps commonly used in existing methods, making it very efficient and practical for real-time video-based applications. Code, dataset, and models are available at https://jiaying.link/cvpr2023-vmd/*

1. Introduction

Mirrors appear everywhere. They can adversely affect the performance of computer vision tasks (*e.g.*, depth estimation [35], vision-and-language navigation [2], semantic segmentation [49]), due to their fundamental property that they reflect objects from their surroundings. Thus, it is necessary to build a robust computer vision model that can distinguish mirrors from their surrounding objects correctly.

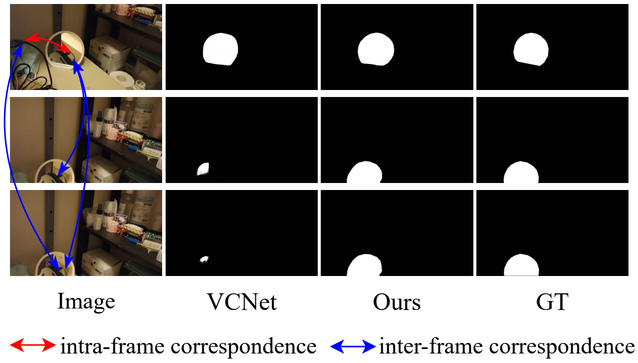


Figure 1. Although state-of-the-art single-image mirror detection method VCNet [36] performs well on a single image (*e.g.*, the first row) by using implicitly intra-frame correspondence, it may fail when the intra-frame cue is weak or even absent in some video frames (*e.g.*, the second and third rows). The lack in exploiting inter-frame information causes the current mirror detection methods to produce inaccurate and inconsistent results when applied to VMD. In contrast, our method can perform well in both situations by utilizing the proposed dual correspondence module to exploit intra-frame (spatial) and inter-frame (temporal) correspondences.

Existing single-image mirror detection methods exploit different cues, such as context contrast [42], explicity correspondences [22], semantics association [14], and chirality and implicit correspondences [36], to detect mirrors from single RGB input images. Despite these recent efforts being put into the mirror detection problem, none of them focuses on detecting mirrors from videos. However, a lot of real-world computer vision applications are video-based (*e.g.*, robotic navigation, autonomous driving, and surveillance), rather than image-based. Hence, solving the video mirror detection (VMD) problem can benefit these applications.

In this paper, we aim to address the VMD problem. There are two major challenges with this problem. First, to the best of our knowledge, there are no existing large-scale datasets that can be used for training and evaluation on the VMD problem. Second, existing mirror detection methods,

[*] Joint first authors.
[†] Corresponding author.

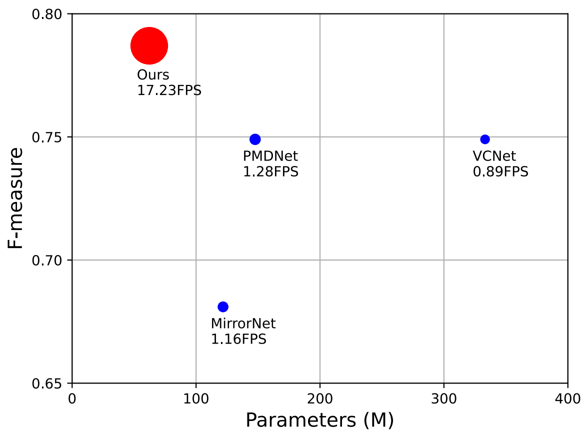


Figure 2. Quantitative comparison on the performance and efficiency between existing mirror detection methods and our method for VMD. All models are trained/tested on the proposed VMD-D dataset, under a single RTX 3090 GPU. Our model has ∼5 times smaller network parameters and runs ∼18× faster than the state-of-the-art image-based mirror detection method, VCNet [36], and still outperforms it by a large margin.

which are all developed for the image-based mirror detection task, are all based on static cues. None of them take advantages of the dynamic nature of videos in the VMD problem. Figure 1 shows that the current state-of-the-art mirror detection method, VCNet [36], may fail when correspondences are missing in some challenging frames (*e.g.*, second and third rows) due to, for example, the change of camera pose, even though it may perform well in some easy cases (*e.g.*, the first row). Besides, as the image-based mirror detection task is already very challenging, existing methods for this task often adopt heavy network design and time-consuming post-processing techniques [19] to improve their results. Figure 2 shows that existing image-based mirror detection models run at about 1fps, even on one of the latest GPUs, which cannot support real-time VMD. All these drawbacks motivate us to develop a large-scale dataset and an effective/efficient method for video mirror detection.

In this paper, we address the VMD problem in two ways. First, we construct the first large-scale video mirror detection benchmark dataset (VMD-D). It contains 14,987 image frames in 269 videos, coming from diverse scenes. The constructed VMD-D dataset provides large-scale and high-diversity data for training and evaluation on the VMD problem. Second, we propose an effective and efficient method, called VMD-Net, for the VMD problem. The proposed method exploits multi-frame correspondences at both intra-frame (spatial) and inter-frame (temporal) levels. Compared with state-of-the-art image-based mirror detection methods, which typically adopt heavy pipelines, our method uses a light-weight network architecture without the need for any post-processing techniques. As a result, our method is efficient for real-time applications. In particular, our method has ∼5 times fewer network parameters and runs ∼18× faster than the latest state-of-the-art image-based mirror detection method, VCNet [36]. We conduct comprehensive experiments to demonstrate the effectiveness and efficiency of our proposed method. Experimental results show that our method outperforms state-of-the-art methods from relevant tasks on the proposed large-scale VMD-D dataset.

Our key contributions can be summarized as follows:

- We construct the first large-scale video mirror detection dataset, called VMD-D. The new dataset contains 14,988 image frames from 269 videos with precise annotated masks.

- We propose a novel network, called VMD-Net, to exploit both intra-frame and inter-frame correspondences via a dual correspondence (DC) module. This DC module allows VMD-Net to tolerate occassionally missing correspondences in the temporal dimension.

- Extensive evaluations show that our method outperforms existing state-of-the-art methods from relevant tasks on our proposed VMD-D dataset.

2. Related Work

Image-based Mirror Detection. Recently, Yang *et al.* [42] propose the first mirror detection dataset and the first mirror detection network to detect mirrors by modeling contextual contrasted information. Lin *et al.* [22] then propose a more diverse and larger mirror detection benchmark, and a correspondence-aware method for mirror detection, which correlates the contents inside and outside of the mirror. Mei *et al.* [27] further consider using RGB-D data for detecting mirrors, and construct the first RGB-D dataset for mirror detection. Tan *et al.* [35] address the depth refinement problem on mirror surfaces by proposing another RGB-D mirror dataset and a detect-and-refine method for mirror detection. Most recently, Tan *et al.* [36] propose an image-based method for mirror detection based on the chirality cue and implicit correspondences.

Despite their success, none of these methods focuses on the VMD problem. They also have high computational costs, thus not practical for real-time applications. In this paper, we aim to address the VMD problem by constructing a benchmark dataset and proposing an efficient/effective method for the problem.

Video Object Segmentation Video object segmentation (VOS) aims to segment the target object(s) from the input videos. Currently, it can be categorized into two main individual tasks: the unsupervised video object segmentation (UVOS) [32, 38–40, 43, 45] and the semi-supervised video

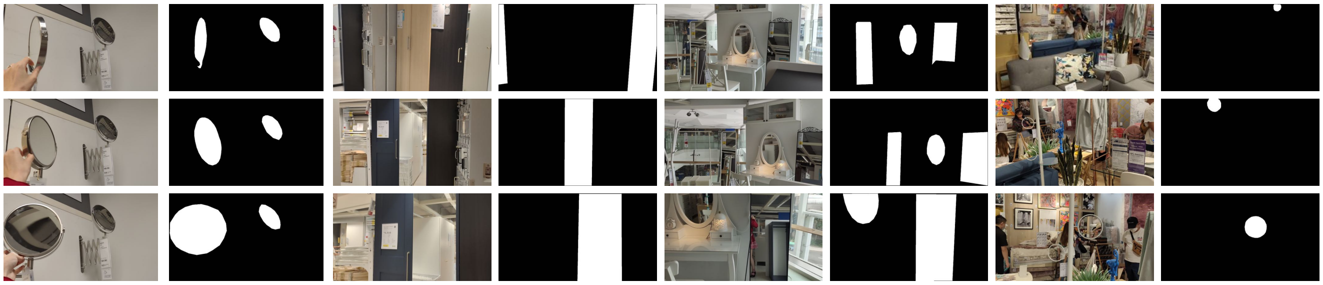


Figure 3. Snapshots of the proposed Video Mirror Detection dataset, VMD-D, with pixel-level annotations.

object segmentation (SVOS) [7, 8, 23–25]. The main difference between UVOS and SVOS is that UVOS does not require annotation for the first frame of the input video during the inference phase, while SVOS needs to do it. Thus, UVOS is more related to the video mirror detection problem since we do not use the first-frame ground truth mirror label to infer the location of mirrors in the rest of the video frames. UVOS focuses on automatically detecting the primary foreground object without using the ground truth mask of the first frame during testing. Early UVOS methods use handcrafted features [30, 34] to detect foreground objects in the input videos. Recent CNN-based UVOS models [38–40] focus on exploring different cues like memory [38], graph neural network [39] and visual attention [40] for better performance. The latest UVOS methods [32, 45] take optical flow as auxiliary information to help segment videos, thanks to the rapid evolution on optical flow estimation [37].

However, UVOS methods cannot directly address the video mirror detection problem, since mirrors are not always the primary foreground objects while having changing appearance patterns during motion.

Video Salient Object Detection Video salient object detection (VSOD) is related to the UVOS task. It aims to automatically detect the most visually distinctive objects from an input video without indicating where the salient object is in the first frame. Unlike the methods for single-image salient object detection [29, 48], VSOD methods require building a model with a thorough understanding of dynamic attention. Most earlier VSOD methods detect the salient objects using handcrafted features [4, 15, 16]. With the help of deep neural networks, recent VSOD methods achieve great progress. Fan *et al.* [11] take attention shift into account. Gu *et al.* [12] expolit self-attention models to detect salient objects from videos efficiently. Zhang *et al.* [46] proposes a dynamic strategy of context fusion for VSOD.

Similar to UVOS, the methods for video salient object detection are not directly applicable to video mirror detection since mirrors are not always distinctive.

3. Video Mirror Detection Dataset (VMD-D)

To facilitate research on the video mirror detection problem, we first contribute a large-scale video mirror detection dataset (named VMD-D). It consists of 269 videos in 14,988 image frames with corresponding precise annotations from diverse scenes. Figure 3 shows some example video frames in our proposed VMD-D dataset. The details of our VMD are discussed below.

3.1. Dataset Construction

To construct the first large-scale video mirror detection dataset, we use a smartphone to collect high-resolution videos (*e.g.*, 1920×1080 resolution) with mirrors in daily-life scenes. Following a common practice [6] used to construct datasets for video-based problems, we manually trim the collected videos to make sure that each video frame has at least one mirror region. After that, we obtain 269 video clips and then randomly split them into a training set with 143 videos (containing 7,835 images) and a test set with 126 videos (containing 7,152 images). We then label pixel-level mirror masks by annotators. The total duration of our videos is 502 seconds for covering long-temporal scenarios. The frame rate is 30 fps for all collected videos.

3.2. Dataset Analysis

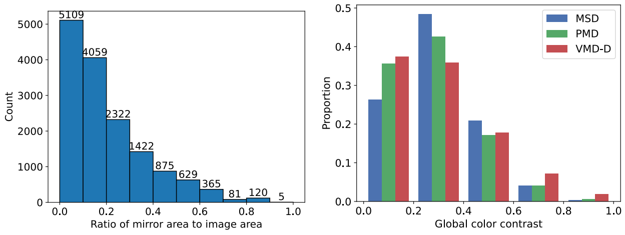
Table 1 shows the comparison of different datasets from the relevant areas, including image mirror detection [22, 42] (top group), video object segmentation [20, 28, 33] (second group), and video shadow detection [6] (third group), and ours. Our dataset reflects good video quality (*i.e.*, high-resolution videos) and quantity (*i.e.*, a large number of annotated frames), and is practical for being the first step for large-scale video mirror detection.

Area Distribution. Figure 4(a) shows the ratio of mirror area over the image area (mirror area distribution). We can see that our dataset contains mirrors covering a wide range of area ratios. We also note that it consists of a lot of small mirrors (< 0.1), which makes our dataset very challenging.

Contrast Distribution. We analyze the contrasts between the mirror regions and non-mirror regions by computing χ^2

Dataset	#Videos	#Annos.	#Time (s).	#FPS	Max Reso.
MSD [42]	-	4,018	-	-	640×512
PMD [22]	-	6,461	-	-	4000×3000
FBMS [28]	59	720	-	-	640×480
STV2 [20]	14	947	-	-	640×360
DAVIS [33]	50	3455	144	24	1920×1080
ViSha [6]	120	11,685	390	30	1280×720
Ours	269	15,066	502	30	1920×1080

Table 1. Comparison of different video datasets for relevant tasks.



(a) mirror area distribution

(b) color contrast distribution

Figure 4. Statistics of our VMD dataset.

distance between their RGB histograms. We also compare the distribution between MSD [42] and PMD [22], as shown in Figure 4(b). In general, VMD-D has more images with extremely low color contrasts (< 0.1) compared with existing mirror datasets MSD and PMD. This indicates our images are collected from very diverse scenes, making them more challenging to detect.

4. Method

4.1. Overview

We observe that there are often spatial correspondences between the contents inside (reflected) and outside (real) of a mirror. However, such spatial correspondences may not always appear in every frame. The lack in exploiting temporal correspondence may cause current image-based mirror detection methods to fail in the VMD task. Hence, in this paper, we propose a method that is able to tolerate temporally missing correspondences by considering mirror correspondences at both the intra-frame level (spatial) as well as the inter-frame level (temporal).

Figure 5 shows the overall structure of our proposed method VMD-Net. The key idea of our network design is to leverage the intra-frame and the inter-frame correspondences for video mirror detection. Our VMD-Net takes three images from the same video clips as inputs. The first two images I_t and I_{t+1} are from adjacent video frames, while the third image I_n is randomly selected from other frames. We first apply a shared backbone network ResNext-101 [41] to extract multi-scale features from the input images. Unlike existing single-image mirror detection methods, which make full usage of the features at all stages (*i.e.*, from the 1^{st} to the 5^{th} scales), for each input image i, our VMD-Net only utilizes the low-level features at the 2^{nd} scale (denoted as $\mathbf{F}^i_{low}$) and the high-level features at the 5^{th} scale. Following the design of DeepLabV3 [5], the high-level features from the backbone network are then fed to an atrous spatial pyramid pooling module to obtain enhanced semantic features (denoted as $\mathbf{F}^i_{high}$). Taking image I_t as an example, we then assign a dual correspondence module to both low-level features $\mathbf{F}^t_{low}$ and high-level features $\mathbf{F}^t_{high}$ to produce the intermediate map P_t and the final output M_t. Similarly, the dual correspondence module takes the low-level features and high-level features from I_{t+1}, I_n to produce the intermediate maps P_{t+1}, P_n and the final outputs M_{t+1}, M_n.

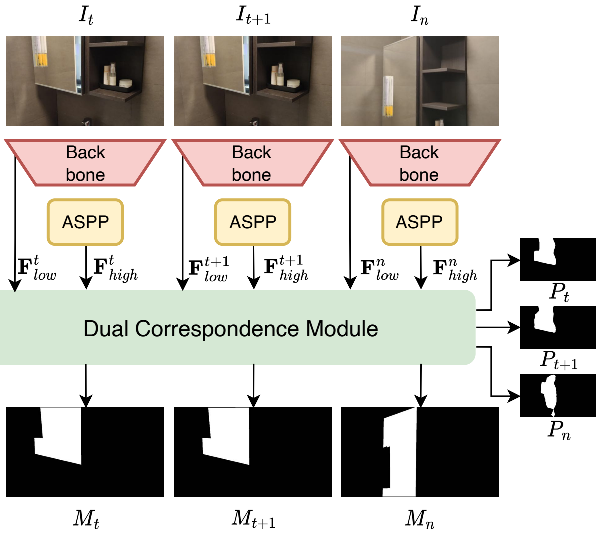


Figure 5. The framework of our proposed method. I_t, I_{t+1}, I_n are input images. In particular, I_t and I_{t+1} are two adjacent video frames, and I_n is a video frame randomly selected from the same video of I_t. P_t, P_{t+1}, P_n are intermediate maps and the final outputs maps are M_t, M_{t+1}, M_n for the corresponding input images.

4.2. Dual Correspondence (DC) Module

Figure 6 shows the structure of the proposed DC module. It consists of two stages. The first stage of the DC module aims to learn **intra-frame correspondences** for all input images and the **short-term inter-frame correspondences** inside mirrors in the adjacent input image I_t and I_{t+1}. The second stage of the DC module focuses on extracting **long-term inter-frame correspondences** across all input images to enhance the learning of temporal correspondence. This enables our VMD-Net to exploit correspondences at both intra-frame and inter-frame levels at different tempo-

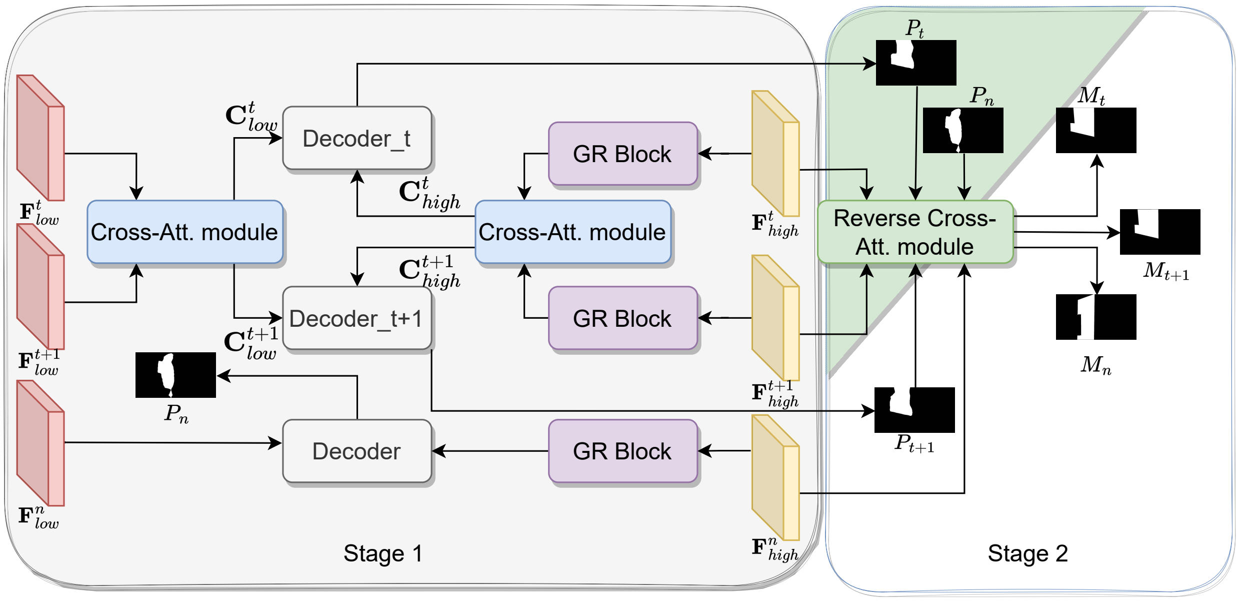


Figure 6. Our DC module. It consists of two stages. The first stage in the DC module aims to extract intra-frame and short-term temporal correspondences from two adjacent video frames I_t, I_{t+1}, while the second stage in the DC module focuses on learning long-term temporal correspondences from all input frames I_t, I_{t+1}, I_n.

ral scales.

Stage 1. Our DC module takes the low-level features $\mathbf{F}_{low}$ and high-level features $\mathbf{F}_{high}$ as input. In the first stage, we first extract the intra-frame correspondences from the high-level features $\mathbf{F}_{high}$ of each image by the global relation (GR) block proposed in [22]. The GR block can effectively and efficiently extract the correspondence between the content inside and outside of the mirrors by modeling spatial corresponding relations in a single image. The high-level correspondence-aware features from the randomly selected frame I_n are directly concatenated with the corresponding low-level features $\mathbf{F}^n_{low}$ in the decoder to output an intermediate map P_n, while the high-level correspondence-aware features extracted from the two adjacent frames I_t and I_{t+1} are forwarded to a cross-attention (Cross-Att.) module to learn short-term temporal correspondences. To compute short-term temporal correspondences between I_t and I_{t+1}, we also feed their low-level features $\mathbf{F}^t_{low}, \mathbf{F}^{t+1}_{low}$ to another cross-attention module.

Our cross-attention module is inspired by [17], but we extend it from self-attention (single-input and single-out features) to cross-attention with two-input and two-output features and also enhance its ability to model spatial correspondences. Our module first applies two convolutional layers with 1×1 filters to the two input features $\mathbf{F}_t$ and $\mathbf{F}_{t+1}$ and further generates a spatially enhanced affinity matrix $A^{se} \in \mathbb{R}^{(2H+2W-1)\times(W\times H)}$. Different from the original affinity matrix $A \in \mathbb{R}^{(H+W-1)\times(W\times H)}$ in [17], which is computed by extracting local contextual information in the horizontal and vertical directions of the input features, our spatially enhanced affinity operation also computes global contextual information by taking the diagonal items of the input features into our design. Formally, our cross-attention process can be described as:

$$K = Conv_{1\times1}(\mathbf{F}_t); Q = Conv_{1\times1}(\mathbf{F}_{t+1}), \quad (1)$$

$$A^{se} = \mathbb{A}^{se}(K, Q), \quad (2)$$

$$V = Conv_{1\times1}(\mathbf{F}_t), \quad (3)$$

$$\mathbf{C}^t = K + \omega_t \sum_i^{(2H+2W-1)\times(W\times H)} A^{se} V, \quad (4)$$

$$\mathbf{C}^{t+1} = Q + \omega_{t+1} \sum_i^{(2H+2W-1)\times(W\times H)} A^{se} V, \quad (5)$$

where $\mathbf{C}$ is the output of the cross-attention module. ω_t and ω_{t+1} are learnable parameters. $\mathbb{A}^{se}$ is our spatially enhanced affinity operation. $\mathbf{C}^t$ and $\mathbf{C}^{t+1}$ are the output correspondence-aware features. We apply the cross-attention module to both $\mathbf{F}^t_{low}, \mathbf{F}^{t+1}_{low}$ and $\mathbf{F}^t_{high}, \mathbf{F}^{t+1}_{high}$ to obtain correspondence-aware features $\mathbf{C}^t_{low}, \mathbf{C}^{t+1}_{low}$ and $\mathbf{C}^t_{high}, \mathbf{C}^{t+1}_{high}$. We then concatenate the correspondence featuresfrom the same input image but at different levels. The concatenated features $[\mathbf{C}_{low}; \mathbf{C}_{high}]$ are forwarded to a decoder to obtain an intermediate prediction P_t for I_t. Similarly, we obtain the intermediate predictions P_{t+1} and P_n for I_{t+1} and I_n.

Stage 2. The second stage of our DC module takes all intermediate prediction maps P_t, P_{t+1}, P_n and high-level features $\mathbf{F}_{high}^{t}, \mathbf{F}_{high}^{t+1}, \mathbf{F}_{high}^{n}$ as inputs. The input features and prediction maps are forwarded to a reverse cross-attention module. This module aims at explicitly exploiting the correspondence between the contents inside and outside of the mirrors in different frames for long-range temporal correspondences. In particular, we need to explicitly model the correspondences between input frame I_t and the randomly selected frame I_n. To do this, we first multiply F_{high} by their corresponding intermediate prediction map P, which are normalized by a sigmoid function. We then reverse P to obtain the reversed prediction map $\hat{P}$, which indicates the potential non-mirror regions. We also notice that the mirrors will potentially flip the real object horizontally in its content. Thus, we conduct a horizontal flip to the input non-mirror features to model the potential relation of mirror reflection. Similarly, we compute temporal correspondences between I_t and I_{t+1} with this strategy. In the second stage of the DC module, we also use our cross-attention module to extract the inter-frame correspondences. The process of our reverse cross-attention module can be formulated as follows:

$$\Omega_n, \_ = \mathbb{CA}(P_n \mathbf{F}_{high}^{n}, (\hat{P}_t \mathbf{F}_{high}^{t})^{\top}), \tag{6}$$

$$\Omega_t, \_ = \mathbb{CA}(P_t \mathbf{F}_{high}^{t}, (\hat{P}_{t+1} \mathbf{F}_{high}^{t+1})^{\top}), \tag{7}$$

$$\Omega_{t+1}, \_ = \mathbb{CA}(P_{t+1} \mathbf{F}_{high}^{t+1}, (\hat{P}_t \mathbf{F}_{high}^{t})^{\top}), \tag{8}$$

where $\mathbb{CA}$ is our cross-attention module. $^{\top}$ is a horizontal flip operation. Ω is the final output feature. Note that we omit the second output of the cross-attention modules used in the 2^{nd} stage for convenience. We then forward the output features to individual decoders to obtain the final output predictions M_t, M_{t+1}, M_n for the input images I_t, I_{t+1}, I_n, respectively.

4.3. Loss Functions

We adopt the Lovász-Softmax loss function [3] to supervise our network training. The final loss function is:

$$\mathcal{L} = \sum_{i}^{i \in \{t, t+1, n\}} \mathcal{L}_h(P_i, G_i) + \mathcal{L}_h(M_i, G_i), \tag{9}$$

where $\mathcal{L}_h(\cdot, \cdot)$ denotes the lovasz-hinge loss. P_i, M_i, G_i are the intermediate output from the DC module, the final output of our network, and the ground truth label of image I_i, respectively.

5. Experiments

5.1. Experimental Settings and Evaluation Metrics

We have implemented VMD-Net in Pytorch [31] and trained it on a PC with an RTX3090 GPU card. During training, we resize the input images to 384×384. We use ResNext-101 [41] pre-trained on ImageNet [9] as our backbone network to extract image features. We adopt Adam [18] as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The base learning rate, batch size, and the number of training epochs are 0.0001, 8, and 15, respectively. We use the cosine learning rate decay with 3-epoch warm-up period to adjust the learning rate during training. Note that we do **NOT** apply any post-processing techniques to refine our predictions.

We employ four popular metrics, intersection over union (IoU), pixel accuracy, F-measure (F_β) [1], and mean absolute error (MAE) to evaluate the performance of tested methods quantitatively. F_β is defined as:

$$F_\beta = \frac{1 + \beta^2 (Precision \times Recall)}{\beta^2 Precision + Recall},$$

where β is set to 0.3 as suggested in [1]. It evaluates the overall performance between precision and recall.

5.2. Comparison to the State-of-the-art Methods

Due to the lack of methods for video mirror detection, we compare our method with 14 state-of-the-art methods from relevant fields, including GateNet [48] and MINet [29] for salient object detection; PCSA [13] for video salient object detection; DeepLabV3 [5], PSPNet [47] and OCRNet [44] for semantic segmentation; TVSD [6], STICT [26] and ScCor [10] for video shadow detection; HFAN [32] for video object segmentation; GlassNet [21] for glass surface detection; MirrorNet [42], PMDNet [22] and VCNet [36] for single-image mirror detection. We train and test all baseline methods on our video mirror detection dataset VMD-D using their released codes on the same platform. Table 2 shows the quantitative results. Our method achieves the best performance with a large margin on all four metrics.

Figure 7 visually compares the results of our method with the selected state-of-the-art methods from relevant fields. We can see that when compared with image-based mirror detection methods PMDNet [22] and VCNet [36], which produce inaccurate and incoherent results both in a short-term temporal interval (*e.g.*, the 1st and 2nd rows; the 3rd and the 4th rows) and a long-term temporal interval (*e.g.*, the 5th and 6th rows; the 7th and 8th rows), our method can precisely predict the mirror regions by exploiting inter-frame and intra-frame correspondences in both situations. We attribute the superior performances of our method to the design of the DC module, which enables the modeling of spatial, short-range, and long-range temporal correspondences.

In particular, we also compare the efficiency of our method with most existing mirror detection methods. Table 3 shows the quantitative results in the number of network parameters and FPS. Our model has significantly fewer network parameters and runs much faster than the

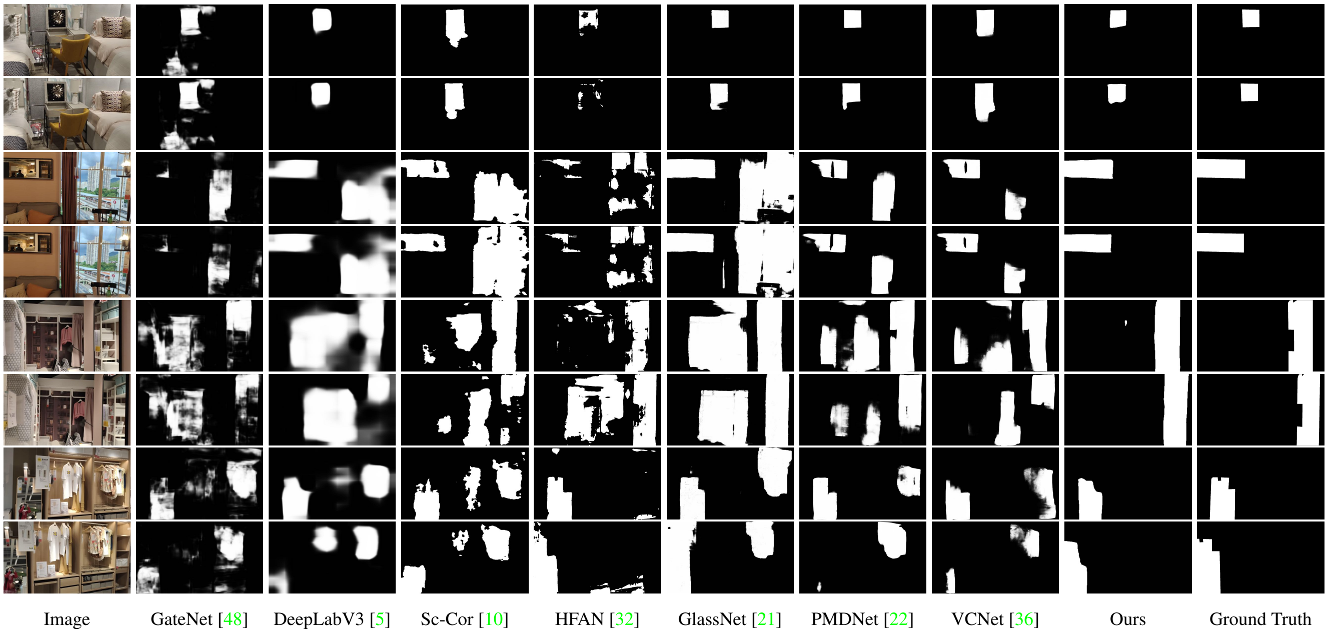


Figure 7. Visual comparison between the proposed VMD-Net and selected state-of-the-art methods from relevant fields.

Methods	IoU↑	Accuracy↑	F_β↑	MAE↓
GateNet [48]	0.429	0.851	0.665	0.153
MINet [29]	0.412	0.854	0.676	0.148
PCSA [13]	0.193	0.803	0.464	0.198
DeepLabV3 [5]	0.481	0.846	0.681	0.157
PSPNet [47]	0.464	0.850	0.665	0.152
OCRNet [44]	0.394	0.786	0.640	0.175
TVSD [6]	0.480	0.875	0.746	0.125
STICT [26]	0.164	0.809	0.530	0.198
Sc-Cor [10]	0.512	0.863	0.696	0.137
HFAN [32]	0.459	0.876	0.706	0.124
GlassNet [21]	0.552	0.864	0.718	0.137
MirrorNet [42]	0.505	0.855	0.681	0.145
PMDNet [22]	0.532	0.872	0.749	0.128
VCNet [36]	0.539	0.877	0.749	0.123
Ours	**0.567**	**0.895**	**0.787**	**0.105**

Table 2. Quantitative comparison between the proposed VMD-Net and 14 state-of-the-art methods from relevant fields. The best results are shown in bold.

state-of-the-art image-based mirror detection method, VCNet [36]. Specifically, our model has ∼5 times fewer network parameters and runs ∼18× faster than VCNet [36]. Furthermore, our model outperforms VCNet [36] by a large margin. These results demonstrate the superior practicality and suitability of our model for the VMD problem.

Methods	Params. (M) ↓	FPS ↑	F_β ↑
MirrorNet [42]	121.77	1.16	0.681
PMDNet [22]	147.66	1.28	0.749
VCNet [36]	333.17	0.89	0.749
Ours	**62.24**	**17.06**	**0.787**

Table 3. Quantitative comparison on the performance and efficiency between existing mirror detection methods and our method for VMD. All models are trained/tested on the proposed VMD-D dataset, on a single RTX 3090 GPU.

5.3. Ablation Study

We have conducted ablation experiments to verify the effectiveness of our design. First, we replace the proposed DC module with a convolution layer and restrict the network to take a single image as input to form the image-based baseline (“I-Base”) for comparison. We then modify I-Base to take three images as input as our video-based baseline (“V-Base”). Our method (“Ours”) is V-Base with the full DC module (*i.e.*, DC with its first stage and its second stage). To test the effectiveness of each stage in the proposed DC module, we also construct two ablated models, which remove the first stage (“Ours w/o DC 1^{st}”) or the second stage (“Ours w/o DC 2^{nd}”) in the proposed DC module. Note that for “Ours w/o DC 1^{st}” ablated model, since the second stage of the DC module requires intermediate prediction maps P_t, P_{t+1}, P_n to compute the reverse attention, we directly concatenate the low-level features $\mathbf{F}_{low}$ with high-

Ablation	IoU↑	Accuracy↑	F_β↑	MAE↓
I-Base	0.424	0.854	0.686	0.146
V-Base	0.446	0.866	0.719	0.134
Ours w/o DC 1^{st}	0.524	0.876	0.743	0.124
Ours w/o DC 2^{nd}	0.533	0.878	0.736	0.122
Ours (I_{same} only)	0.525	0.871	0.740	0.123
Ours (I_{adj} only)	0.539	0.876	0.745	0.124
Ours	**0.567**	**0.895**	**0.787**	**0.105**

Table 4. Ablation Study. "Ours (I_{same} only)" / "Ours (I_{adj} only)" is the ablation model that takes three same/adjacent image frames as input, respectively. DC 1^{st} and DC 2^{nd} are the first and the second stage of the DC module, respectively.

level features $\mathbf{F}_{high}$, and then forward them to a decoder to predict P_t, P_{t+1}, P_n without computing intra-frame correspondences and short-term inter-frame correspondences.

Table 4 shows the performances of different ablated models. As shown in the last row, our final proposed network with the full DC module outperforms other baselines on all metrics. In particular, even the ablated models with only a single-stage DC module (3^{rd} row and 4^{th} row) can perform better than the baseline networks without the full DC module ("I-Base" and "V-Base"). This indicates the importance of correspondence learning for the VMD task. Also, the effective combination of the two stages in the DC module ("Ours") can boost the overall performance compared with the single-stage DC module. This shows that learning "dual" correspondences outperforms learning "sole" correspondences in our network.

To further investigate the role of different types of correspondences in our network, we replace the input images (*i.e.*, two adjacent frames and one randomly selected frame) with three adjacent frames (I_{adj} only) or three video frames from the same frame (I_{same} only). The "I_{adj} only" model focuses on short-term temporal correspondences while the "I_{same} only" model can only learn intra-frame correspondences, due to their limited input information. The results are listed in the last three rows and show that our method, which exploits intra-frame as well as both short-term and long-term temporal correspondences, outperforms these ablated models. Figure 8 shows the visual results of different ablated models. Our method with sufficient learning of spatial and temporal correspondences can locate the mirrors precisely in both frames. In particular, the results in the last row show that our method can still detect the mirrors by exploiting long-range temporal correspondences even though intra-frame correspondences are missing in the image.

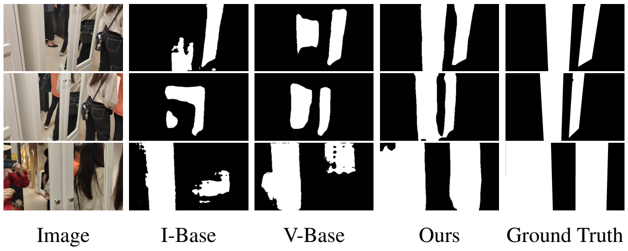


Figure 8. Visual comparison of different ablated models.

6. Conclucsion

In this paper, we have investigated the video mirror detection (VMD) problem. To the best of our knowledge, we are the first to address the VMD problem from a deep-learning-based perspective. We have constructed the first large-scale video mirror dataset (VMD-D). It contains 14,988 image frames from 269 videos with corresponding masks. We have also proposed a novel network, called VMD-Net, to leverage both intra-frame and inter-frame correspondences for video mirror detection. Experimental results show that our VMD-Net outperforms state-of-the-art methods from relevant tasks on our VMD-D dataset.

Our method does have limitations. Figure 9 shows that our method tends to produce results with coarse boundaries. Unlike existing image-based mirror detection methods [22, 36], which often explicitly adopt edge supervision in their network training, our method does not leverage such auxiliary information since we do not plan to focus this incremental cue in this first work for VMD. As a future work, we are working on improving our method by leveraging additional information like boundaries to help detect mirrors more precisely in videos.

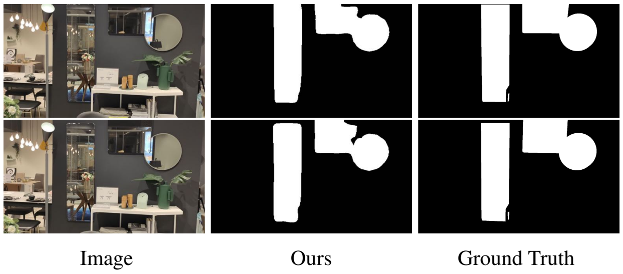


Figure 9. Failure cases. Our method may not be good at boundary extraction, due to the lack of explicit supervision of edges.

Acknowledgments

Xin Tan is sponsored by Shanghai Sailing Program (23YF1410500). This work is in part supported by two SRG grants from City University of Hong Kong (Ref: 7005674 and 7005843).

References

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

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 1

[3] 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*, pages 4413–4421, 2018. 6

[4] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE Transactions on Image Processing*, 26(7):3156–3170, 2017. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4, 6, 7

[6] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *CVPR*, pages 2715–2724, 2021. 3, 4, 6, 7

[7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, pages 640–658. Springer, 2022. 3

[8] Suhwan Cho, Heansung Lee, Minhyeok Lee, Chaewon Park, Sungjun Jang, Minjung Kim, and Sangyoun Lee. Tackling background distraction in video object segmentation. In *ECCV*, pages 446–462. Springer, 2022. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 6

[10] Xinpeng Ding, Jingwen Yang, Xiaowei Hu, and Xiaomeng Li. Learning shadow correspondence for video shadow detection. In *ECCV*, pages 705–722. Springer, 2022. 6, 7

[11] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 3

[12] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, 2020. 3

[13] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, 2020. 6, 7

[14] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *CVPR*, pages 5941–5950, June 2022. 1

[15] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*. IEEE, 2008. 3

[16] Fang Guo, Wenguan Wang, Jianbing Shen, Ling Shao, Jian Yang, Dacheng Tao, and Yuan Yan Tang. Video saliency detection using object proposals. *IEEE transactions on cybernetics*, 48(11):3159–3170, 2017. 3

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, 2019. 5

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011. 2

[20] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013. 3, 4

[21] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pages 13415–13424, 2021. 6, 7

[22] Jiaying Lin, Guodong Wang, and Rynson W. H. Lau. Progressive mirror detection. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[23] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *CVPR*, pages 1362–1372, 2022. 3

[24] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, pages 648–665. Springer, 2022. 3

[25] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *European Conference on Computer Vision*, pages 468–486. Springer, 2022. 3

[26] Xiao Lu, Yihong Cao, Sheng Liu, Chengjiang Long, Zipei Chen, Xuanyu Zhou, Yimin Yang, and Chunxia Xiao. Video shadow detection via spatio-temporal interpolation consistency training. In *CVPR*, pages 3116–3125, 2022. 6, 7

[27] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, pages 3044–3053, 2021. 2

[28] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013. 3, 4

[29] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 3, 6, 7

[30] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, pages 1777–1784, 2013. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,

Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[32] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *ECCV*, pages 596–613. Springer, 2022. 2, 3, 6, 7

[33] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 3, 4

[34] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, pages 3227–3234, 2015. 3

[35] Jiaqi Tan, Weijie Lin, Angel X Chang, and Manolis Savva. Mirror3d: Depth refinement for mirror surfaces. In *CVPR*, pages 15990–15999, 2021. 1, 2

[36] Xin Tan, Jiaying Lin, Ke Xu, Chen Pan, Lizhuang Ma, and Rynson W. H. Lau. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 6, 7, 8

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 3

[38] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, pages 4481–4490, 2017. 2, 3

[39] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, pages 9236–9245, 2019. 2, 3

[40] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, pages 3064–3074, 2019. 2, 3

[41] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 4, 6

[42] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019. 1, 2, 3, 4, 6, 7

[43] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, pages 931–940, 2019. 2

[44] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, pages 173–190. Springer, 2020. 6, 7

[45] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *ICCV*, pages 8781–8790, 2021. 2, 3

[46] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563, 2021. 3

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 6, 7

[48] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51. Springer, 2020. 3, 6, 7

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1