ZOOM: Learning Video Mirror Detection With Extremely-Weak Supervision

Ke Xu[*], **Tsun Wai Siu**[*], **Rynson W.H. Lau**[†]

Department of Computer Science, City University of Hong Kong

Abstract

Mirror detection is an active research topic in computer vision. However, all existing mirror detectors learn mirror representations from large-scale pixel-wise datasets, which are tedious and expensive to obtain. Although weakly-supervised learning has been widely explored in related topics, we note that popular weak supervision signals (*e.g.*, bounding boxes, scribbles, points) still require some efforts from the user to locate the target objects, with a strong assumption that the images to annotate always contain the target objects. Such an assumption may result in the over-segmentation of mirrors. Our key idea of this work is that the existence of mirrors over a time period may serve as a weak supervision to train a mirror detector, for two reasons. First, if a network can predict the existence of mirrors, it can essentially locate the mirrors. Second, we observe that the reflected contents of a mirror tend to be similar to those in adjacent frames, but exhibit considerable contrast to regions in far-away frames (*e.g.*, non-mirror frames). In this paper, we propose ZOOM, the first method to learn robust mirror representations from extremely-weak annotations of per-frame ZerO-One Mirror indicators in videos. The key insight of ZOOM is to model the similarity and contrast (between mirror and non-mirror regions) in temporal variations to locate and segment the mirrors. To this end, we propose a novel fusion strategy to leverage temporal consistency information for mirror localization, and a novel temporal similarity-contrast modeling module for mirror segmentation. We construct a new video mirror dataset for training and evaluation. Experimental results under new and standard metrics show that ZOOM performs favorably against existing fully-supervised mirror detection methods.

Introduction

Mirrors are made to reflect objects in the surroundings for different purposes (*e.g.*, monitoring traffic situations, checking dressings, and decorating rooms). However, such reflected contents of mirrors may fail existing computer vision models in various tasks, *e.g.*, depth estimation (Tan et al. 2021), lane detection (Feng et al. 2022), and scene parsing (Zhou et al. 2017; Xie et al. 2023). Hence, it is essential to design effective and robust mirror detectors.

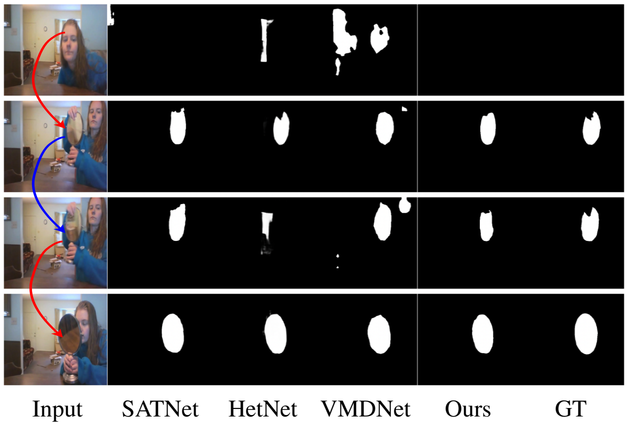

Figure 1: We propose ZOOM, which learns to detect mirrors with extremely weak supervision, *i.e.*, the zero-one mirror indicators. The key insight of ZOOM is to model the similarity (blue arrow) and contrast (red arrows) in temporal variations for mirror detection. It achieves promising results against fully-supervised mirror detectors.

In recent years, a few deep methods are proposed to train mirror detectors in a fully-supervised manner with large amount of mirror images and annotations. Yang *et al.* (2019) propose the first deep network to learn contextual contrasted features for detecting mirrors in single RGB images. Two other methods (Mei et al. 2021; Tan et al. 2021) extend the modeling of contrasted features by incorporating depth information. Other methods exploit appearance correspondences (Lin, Wang, and Lau 2020; Lin, Tan, and Lau 2023), semantic relationships (Guan, Lin, and Lau 2022), visual chiral discrepancy (Tan et al. 2023), coarse symmetry relations (Huang et al. 2023), and intensity-based low-level and semantics-based high-level features (He, Lin, and Lau 2023), for distinguishing the real and reflected contents. Despite the success, these methods typically require tedious pixel-level labeling of large amounts of training data. They may also suffer from the over-detection limitation, as their fully-supervised learning schemes implicitly assume that test images always contain mirrors (*e.g.*, Figure 1 first row).

Weakly-supervised learning is a straightforward and possible solution to reduce labeling efforts. We note that there are four main types of weak supervision signals widely studied in related computer vision tasks, *i.e.*, bounding boxes (Dai, He, and Sun 2015), scribbles (Zhang et al. 2020;

[*] These authors contributed equally.

[†] Rynson Lau is the corresponding author.

Yu et al. 2021; He et al. 2023), points (Gao et al. 2022b; Kim et al. 2023), and class labels (Araslanov and Roth 2020; Qin et al. 2022; Li, Xie, and Lin 2018; Piao et al. 2021; Wang et al. 2017). However, while the class labels may not work well as mirrors may reflect objects of different classes, bounding boxes, scribbles, and points still require annotators to pay considerable effort to recognize/locate the targets.

In this paper, we aim to answer the question of *whether it is possible to learn robust mirror representations with minimum supervision.* Our key observation is that the temporal existence of mirrors can be used as weak supervision to train a mirror detector. The reasons are two-fold. First, when a network learns to predict the existence of mirrors, it essentially learns to locate the mirrors. Second, due to the relative motions between mirrors and cameras, we observe that the reflected contents of a mirror tend to be similar to those in adjacent frames, but exhibit considerable contrast to those in far-away frames, *e.g.*, non-mirror frames (Figure 1 1st column). Such temporal information can be modeled for mirror segmentation. The first observation inspires us to explore the knowledge of mirror presence/absence to train a mirror detector, which is much cheaper than existing weak labels as localization is no longer required by the annotators. The second observation inspires us to model the temporal variations in similarity and contrast to segment the mirror regions.

Inspired by the above observations, in this paper, we propose ZOOM, the first method that learns robust mirror representations from extremely-weak annotations of per-frame ZerO-One Mirror indicators in videos. ZOOM has two main novelties. First, we propose a novel Class Activation Maps (CAM) (Zhou et al. 2016) based fusion strategy to leverage temporal consistency information for robust mirror localization. Second, we propose a novel temporal similarity-contrast modeling module to model the similarity of mirror regions of adjacent frames and the contrast between mirror and non-mirror regions of distant frames for mirror segmentation. To facilitate the learning process, we construct a new video mirror dataset for the training and evaluation of ZOOM. This dataset does not assume that mirrors always exist. To summarize, this work has four main contributions:

- We propose ZOOM, the first method that learns from extremely-weak annotations of frame-level ZerO-One Mirror indicators for video mirror detection.
- We propose a novel fusion strategy to localize mirrors by introducing temporal consistency information, and a novel temporal similarity-contrast modeling module to segment mirrors by modeling the feature similarity of mirror regions in adjacent frames and the feature contrast of mirror/non-mirror regions in distant frames.
- We construct a video mirror dataset, which covers diverse daily life scenes for training and evaluation. The key feature of our dataset is that it does not assume the presence of mirrors.
- Experiments with new and standard metrics on our and existing datasets show that ZOOM achieves promising results against existing fully-supervised mirror detectors.

Related Work

Deep Mirror Detection. Yang *et al.* (2019) propose the first deep network to segment mirrors in single images via contextual contrasted feature modeling. Later, a few methods propose to model the appearance correspondences (Lin, Wang, and Lau 2020), semantic associations (Guan, Lin, and Lau 2022), visual chiralty (Tan et al. 2023), and coarse symmetry relations (Huang et al. 2023) between real and reflected contents. He *et al.* (2023) propose an efficient mirror detector by learning intensity-based contrast and semantic features. Two other methods (Mei et al. 2021; Tan et al. 2021) extend the contrast modeling of RGB images by incorporating depth information. Most recently, a concurrent work by Lin *et al.* (2023) models inter- and intra-frame appearance correspondences for video mirror detection.

While all existing mirror detection methods are fully-supervised, this paper presents a novel method to train a video mirror detector with 0/1 mirror indicators.

Video Salient Object Detection (VSOD). Aiming at the segmentation of visually distinctive (*i.e.*, salient) objects from an input video, the majority of VSOD methods (Fan et al. 2019; Gu et al. 2020; Zhang et al. 2021; Li et al. 2019; Chen et al. 2021; Song et al. 2018) is fully-supervised, which focus on the modeling of dynamic visual contrasts (in contrast to static ones from single images, *e.g.*, (Tu et al. 2016; Hu et al. 2018)). To alleviate the annotation costs, scribbles (Zhao et al. 2021) and points (Gao et al. 2022a) are exploited as weak supervision signals for VSOD.

However, VSOD methods detecting mirrors well, as the reflected contents of mirrors are not always salient.

Weak Supervision Signals. Bounding boxes (Dai, He, and Sun 2015; Liang et al. 2022), scribbles (Zhang et al. 2020; Yu et al. 2021; He et al. 2023), points (Yang et al. 2018; Gao et al. 2022b; Kim et al. 2023), and class labels (Araslanov and Roth 2020; Qin et al. 2022; Li, Xie, and Lin 2018; Liu et al. 2023; Piao et al. 2021; Wang et al. 2017; Kweon et al. 2021; Tian et al. 2020, 2022) are popular weak supervision signals for various vision tasks. However, bounding boxes, scribbles, and points still require the efforts of annotators to provide location information.

In this paper, we explore a weaker and more challenging supervision signal, the zero or one mirror indicator. Our supervision signal is similar to class labels in that they both do not have explicit location information. However, methods using class labels typically leverage strong semantics (*e.g.*, certain shapes and appearances of a specific class), which may not work well on mirrors due to the changes of reflected contents in mirrors.

Dataset

To facilitate weakly-supervised training and evaluation, we first construct a video mirror detection dataset, which contains 200 videos ($12,490$ frames). Figure 2 shows some examples in our dataset.[1] We discuss the details below.

Video Collection. To make our dataset represent daily-life scenes, we collect 140 videos from two public datasets:

[1] https://drive.google.com/drive/folders/199OpHuHkmbY4ib5TJKV_m7rxN1JgmHWI?usp=sharing.

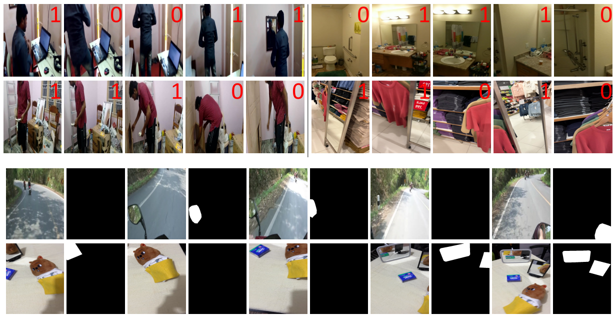

Figure 2: Our dataset examples. Upper two rows show four training video clips (with corresponding mirror indicators marked in red). Bottom two rows show two test video clips (with corresponding ground truth mirror maps).

70 videos from the Charades (Sigurdsson et al. 2016) and 70 videos from the Charades-Ego (Sigurdsson et al. 2018), which record daily indoor activities. We capture 60 videos by ourselves using smartphones. We trim each video to have a duration of $5 \sim 8$ seconds at 10 FPS. The total duration of our videos is $1,252$ seconds.

Dataset Annotation. We randomly split our dataset into a training set of 150 videos (9,398 images) and a test set of 50 videos (3,092 images). We assign frame-level binary mirror indicators to the training set for training our method and annotate pixel-level mirror masks for the test set for performance evaluation. In addition, we uniformly sample $\sim 20\%$ frames from the training set and annotate pixel-level mirror masks for them, in order to collect dataset statistics and fine-tune existing methods.

Contrast Distribution. Figure 3 shows the color contrasts (χ^2 distance of RGB histograms) between the mirror regions and non-mirror regions. We include the distributions of MSD (Yang et al. 2019), PMD (Lin, Wang, and Lau 2020) and VMD (Lin, Tan, and Lau 2023) for reference, to which our dataset has similar color contrast distributions.

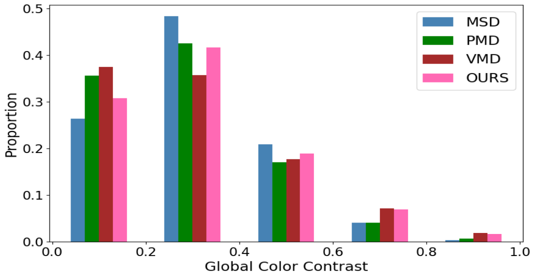

Figure 3: Color contrast distributions.

Area Distribution. Figure 4 shows the mirror area distributions of our training and test sets, respectively. It shows that after the dataset split, the distributions of training and test sets are still aligned well. It also shows that our dataset contains mirrors of different area ratios, while small mirrors being the most make our dataset challenging.

Temporal & Spatial Distribution. We analyze both temporal and spatial existences of mirrors in our dataset in Figure 5. To analyze the temporal existences of mirrors, we use the relative time (frame index) of the mirror disappearing in a video. Figure 5 (left) shows that in our dataset, mirrors may move out of the camera across the whole video duration, although there is a relatively higher chance that mirrors disappear at the end of a video. To analyze the spatial existences of mirrors, Figure 5 (right) shows the probability map, which indicates how likely each pixel belongs to a mirror. We can see that mirrors tend to occupy the majority of the image except for the bottom parts, as mirrors tend to be placed around human eyesight.

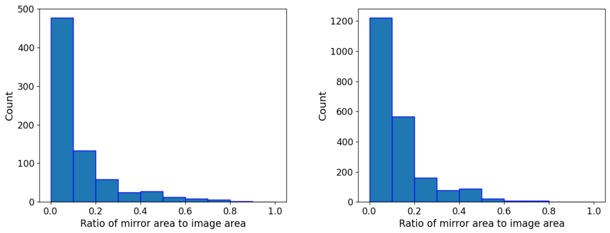

Figure 4: Mirror area distributions of our training (left) and test (right) sets.

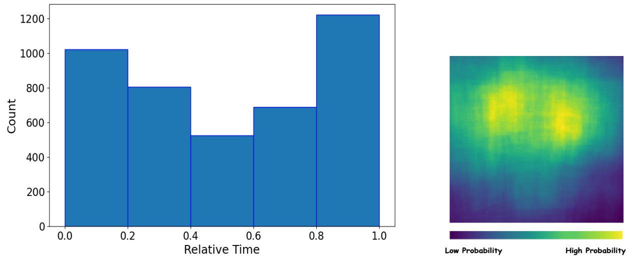

Figure 5: Temporal (left) and spatial (right) distributions.

Proposed Method

Besides requiring expensive pixel-wise labels for training, we note that existing mirror detection methods typically assume the existence of mirrors, which often result in the mirror over-segmentation. Our key idea of this work is to exploit the temporal presence of mirrors as weak supervision to train a mirror detector, as learning to predict the mirror presence essentially locates the mirrors. Besides, we observe that the reflected contents of a mirror tend to be similar to those in adjacent frames, but exhibit considerable contrast to regions in far-away non-mirror frames. Such temporal knowledge further helps the mirror detection.

To this end, we propose ZOOM, to learn robust video mirror representations from the extremely-weak supervision of per-frame zero-one mirror indicators. Formally, given a collection of videos of N frames $\mathcal{Y} = \{y_1, ..., y_N\}$ and their corresponding mirror indicators $\mathcal{S} = \{s_1, ..., s_N\} \in \{0, 1\}$ as supervision, ZOOM is a deep function Z_θ to be trained to produce the mirror maps $\hat{\mathcal{M}} = \{\hat{m}_1, ..., \hat{m}_N\}$ as:

$$\hat{\mathcal{M}} = Z_\theta(\mathcal{Y}), \tag{1}$$

where learnable parameter θ contains two groups of parameters $\{\theta^c, \theta^s\}$ for the classification and segmentation, respectively. Figure 6 illustrates the overview of training ZOOM.

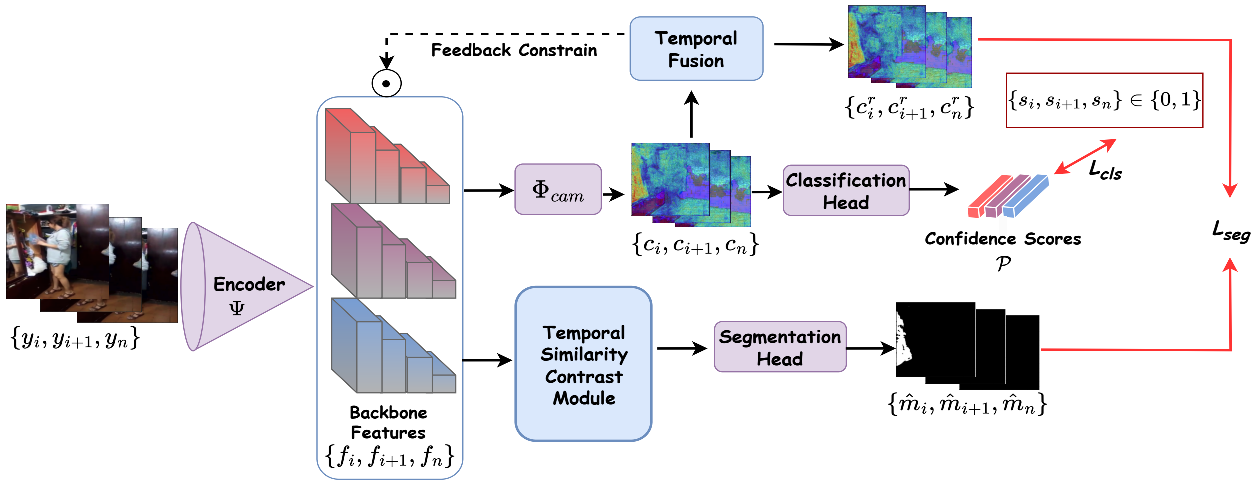

Figure 6: Overview of ZOOM. Given input frames y_i, y_{i+1}, y_n, we use an encoder Ψ to extract backbone features f_i, f_{i+1}, f_n, respectively. We attach a classification network to predict the mirror presence, which produces mirror localization maps c_i, c_{i+1}, c_n via Φ_{cam}. We perform a temporal fusion process to exploit temporal consistency to obtain refined maps c_i^r, c_{i+1}^r, c_n^r as pseudo ground truth. We propose a temporal similarity-contrast modelling module to model the temporal varying similarity/contrast between mirror/non-mirror regions for mirror segmentation.

Localizing the Mirrors

Unlike existing weak supervision signals, our zero-one mirror indicators are extremely-weak (*i.e.*, without mirror location information). Hence, while training ZOOM to locate mirrors via a classification of mirror existences, we consider the CAM techniques (Zhou et al. 2016; Wang et al. 2020; Selvaraju et al. 2017; Jiang et al. 2021) to generate the localization maps for input frames.

Locating Mirrors In Single Frames. We note that CAM methods are typically leveraged in post-processing steps to generate pseudo ground truth maps. As our goal is to generate mirror localization maps with their corresponding features, both of which are used to facilitate the mirror localization and segmentation. Hence, to determine where a neural network focuses on when recognizing the mirror in one frame, we modify the last two layers (*i.e.*, the global average pooling (GAP) and the fully connected (FC) layer) of CAM (Zhou et al. 2016).

Specifically, given input frames $\mathcal{Y} = \{y_i, y_{i+1}, y_j\}$ (of which y_i and y_{i+1} are two adjacent mirror frames and y_j is a non-mirror frame of the same video), we first extract their corresponding multi-scale deep features $\mathcal{F} = \{f_i, f_{i+1}, f_j\}$ using an encoder Ψ (a pre-trained ResNext backbone (Xie et al. 2017)). We then generate localization maps $\mathcal{C} = \{c_i, c_{i+1}, c_j\}$ and corresponding confidence scores $\mathcal{P} = \{p_i, p_{i+1}, p_j\}$ as:

$$\mathcal{C} = \Phi_{cam}(\mathcal{F}), \mathcal{P} = Cls(\mathcal{C}), \tag{2}$$

in which Φ_{cam} is a mapping function consisting of three convolution layers for reducing and aligning the channel dimension of features $\mathcal{F}$, and predicting the localization maps $\mathcal{C}$, respectively. $\mathcal{C}$ is then fed to a 1×1 convolutional classification head (Cls) to produce the confidence scores $\mathcal{P}$ for calculating the binary cross-entropy loss.

Temporal Fusion. The localization maps obtained from the classification are coarse (it tends to focus on the most discriminative pixels) and noisy (it may identify both mirror and non-mirror pixels). We consider a temporal similarity fusion strategy, as:

$$\hat{c} = c_i + c_{i+1} - c_j, \tag{3}$$

which tends to aggregate more confident pixels and suppress the noisy ones. We then leverage a feedback strategy to enhance the backbone features $\mathcal{F}$ using $\hat{c}$. Specifically, we first conduct min-max normalization to constrain activation values to $[0, 1]$. We then feedback the fused localization map to the backbone features as:

$$\mathcal{F}^c = \mathcal{F} \cdot \hat{c}, \tag{4}$$

where $\cdot$ is element-wise multiplication. We omit simple convolutions to align feature dimensions for simplicity.

Pseudo-label Generation. After obtaining the enhanced mirror-aware backbone features $\mathcal{F}^c$, we perform another classification process to obtain localization maps as:

$$\mathcal{C}^r = \Phi_{cam}(\mathcal{F}^c). \tag{5}$$

We do not directly use the fused maps $\hat{c}$ as pseudo-labels as they may not be accurate due to initially incorrect classification.

Segmenting the Mirrors

Our design is based on the observation that the reflected contents of a mirror tend to be consistent to those in adjacent frames, but exhibit certain contrast to regions in distant frames. To this end, we model the feature similarity between mirror regions in adjacent frames, and the feature contrast between mirror and non-mirror regions in distant frames, for mirror segmentation.

Temporal Similarity Modeling. As shown in Figure 7 (upper part), given the input backbone features $\mathcal{F} = \{f_i, f_{i+1}, f_j\}$, we first apply three groups of convolutions to reduce and align their feature dimensions, respectively, to produce $\{f'_i, f'_{i+1}, f'_j\} \in \mathbb{R}^{H \times C \times W}$, where H, C, W represents the

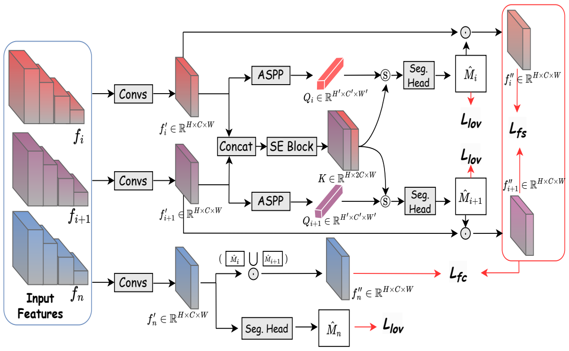
Convs
Convs
Convs
f_i
f_{i+1}
f_n
Input Features
$f'_i \in \mathbb{R}^{H\times C\times W}$
$f'_{i+1} \in \mathbb{R}^{H\times C\times W}$
$f'_n \in \mathbb{R}^{H\times C\times W}$
ASPP
ASPP
Concat
SE Block
$Q_i \in \mathbb{R}^{H'\times C'\times W'}$
$Q_{i+1} \in \mathbb{R}^{H'\times C'\times W'}$
$K \in \mathbb{R}^{H\times 2C\times W}$
Seg. Head
Seg. Head
Seg. Head
$\hat{M}_i$
$\hat{M}_{i+1}$
$\hat{M}_n$
$(\hat{M}_i \bigcup \hat{M}_{i+1})$
L_{lov}
L_{lov}
L_{lov}
L_{fs}
L_{fc}
$f''_i \in \mathbb{R}^{H\times C\times W}$
$f''_{i+1} \in \mathbb{R}^{H\times C\times W}$
$f''_n \in \mathbb{R}^{H\times C\times W}$

Figure 7: Proposed temporal similarity-contrast modeling module. It aims to enhance feature similarity between mirror regions in adjacent frames, and maximize feature contrast between mirror/non-mirror regions in distant frames.

feature height, channel numbers, and width, respectively. Note that f'_i and f'_{i+1} are features extracted from adjacent mirror frames while f'_j are features from the distant non-mirror frame. We first model the feature similarity between f'_i and f'_{i+1}. Specifically, we first concatenate f'_i and f'_{i+1} and then apply the SE block (Hu, Shen, and Sun 2018) to adjust the channel-wise feature consistency to produce $K \in \mathbb{R}^{H\times C\times W}$. We then apply the ASPP module (Chen et al. 2017) to encode f'_i and f'_{i+1} into two compact embeddings Q_i and $Q_{i+1} \in \mathbb{R}^{H'\times C'\times W'}$, respectively. We then compute the cosine similarity betwen Q_i and K to obtain the confidence map U_i (Wang et al. 2022), and we obtain U_{i+1} similarly (denoted as Ⓢ in Figure 7). Finally, we use U_i and U_{i+1} to re-weight K and fed them into the segmentation head to produce the mirror maps $\hat{M}_i$ and $\hat{M}_{i+1}$, respectively. To enhance the network to capture the content similarity in mirror regions, we minimize the feature similarity loss as:

$$\mathcal{L}_{fs} = \|\hat{M}_i \cdot f'_i - \hat{M}_{i+1} \cdot f'_{i+1}\|_2. \qquad (6)$$

Temporal Contrast Modeling. As shown in Figure 7 (lower part), we construct an additional branch to model the temporal contrast information. To maximize the feature discrepancy between the mirror region of i-th (or $i+1$-th) frame and the non-mirror region of n-th frame, we minimize the feature contrast loss as:

$$\mathcal{L}_{fc} = -\|(\hat{M}_i \bigcup \hat{M}_{i+1}) \cdot f'_n - \hat{M}_i \cdot f'_i - \hat{M}_{i+1} \cdot f'_{i+1}\|_2, \quad (7)$$

where $\hat{M}_i$ and $\hat{M}_{i+1}$ are mirror regions of frame i and $i+1$, respectively, while $\hat{M}_i \bigcup \hat{M}_{i+1}) \cdot f'_n$ extracts the corresponding non-mirror features in frame n. As all frames are from the same video clip, f''_n also represents the non-mirror regions in frame i and $i+1$ to some extents. By modelling such contrast, Eq. 7 helps the model to segment the mirrors more accurately, and also suppress background noisy predictions with the all-zero supervision and the parameter-sharing of the segmentation head.

Training and Inference

Loss Function. We use the binary cross entropy loss ($\mathcal{L}_{bce}$) to supervise the mirror localization process. In addition to the $\mathcal{L}_{fs}$ and $\mathcal{L}_{fc}$, we use the Lovász-Softmax loss (Berman, Triki, and Blaschko 2018) ($\mathcal{L}_{lov}$) for the mirror segmentation process. The whole loss function can be written as:

$$\mathcal{L} = \sum \mathcal{L}_{bce}(\mathcal{Y}, \mathcal{S}) + \mathcal{L}_{lov}(\hat{\mathcal{M}}, \mathcal{C}^r) + \mathcal{L}_{fs} + \mathcal{L}_{fc}. \quad (8)$$

In practice, we first train the classification branch, in which we forward the classification process twice and backpropagate it once to produce pseudo-labels. We then train the segmentation branch and freeze the classification-related parameters. Note that for training we assume that f_n is extracted from the non-mirror frames. However, in inference, if one frame is classified as non-mirrors, the corresponding segmentation process is not performed. We use two frames for inference.

Implementation Details. We have implemented the proposed model under Pytorch (Paszke et al. 2017), and tested it on a PC with an i7 4GHz CPU and a GTX4090 GPU. We use ResNext-101 (Xie et al. 2017) pre-trained on ImageNet (Deng et al. 2009) to initialize our encoder network, while other network parameters are initialized using the truncated normal initializer (with the randomness seed set to 2333). For loss minimization, we adopt the AdamW optimizer (Loshchilov and Hutter 2019). The base learning rate, batch size, and the number of training epochs are $2e^{-4}$, 8, and 120, respectively, while the learning rate is reduced by 10 at the $90th$ epoch. Input frames are resized to 352×352. No post-processing techniques are used to refine our results.

Results

Experimental Setups

Evaluation Methods. We compare our method to **seven** fully-supervised mirror detection methods with publically available codes, including six image-based methods (*i.e.*, MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), SANet (Guan, Lin, and Lau 2022), VCNet (Tan et al. 2023), SAT (Huang et al. 2023), and HetNet (He, Lin, and Lau 2023)) and one most recent video-based method (VMDNet (Lin, Tan, and Lau 2023)).

Evaluation Datasets. We report the mirror detection performance on the proposed video mirror detection dataset and the existing video dataset (Lin, Tan, and Lau 2023). We finetune all competing methods using our training data when they are evaluated on our test set. We follow the experimental setups of VMDNet (Lin, Tan, and Lau 2023) for experiments on their dataset.

Evaluation Metrics. We follow previous methods to use the intersection over union (IoU), mean absolute error (MAE), and F-measure (F_β, β is set to 0.3 as suggested in (Achanta et al. 2009)) to evaluate the mirror detection performance.

Note that our dataset may have video frames that do not contain mirrors. To evaluate mirror detection performance on such frames, we first propose to compute the true negative rate (TNR) as:

$$TNR = TN/(TN + FP), \qquad (9)$$

where TN represents the number of true negative pixels (*i.e.*, correctly-detected non-mirror pixels) and FP represents the number of false positive pixels (*i.e.*, falsely-detected non-mirror pixels). This measures to what extent a

Table 1: Quantitative comparison on the proposed dataset. "Model†" represents models finetuned on our dataset (in fully-supervised ways). Best and second best results are marked in **bold** and <u>underlined</u>, respectively, for reference.

Methods	$w/$ **Mirror**			w/o **Mirror**		
	IoU↑	F_β ↑	MAE↓	TNR↑	ER↓	MAE↓
MirrorNet	0.468	0.564	0.203	0.641	0.941	0.357
MirrorNet†	0.572	0.758	0.093	0.843	1.000	0.164
PMDNet	0.585	0.728	0.074	0.932	0.855	0.068
PMDNet†	0.647	0.807	0.062	0.967	0.966	0.082
SANet	0.541	0.760	0.167	0.753	1.000	0.351
SANet†	0.633	0.819	0.083	0.925	0.999	0.164
VCNet	0.612	0.755	0.069	0.899	0.932	0.104
VCNet†	<u>0.660</u>	<u>0.825</u>	0.256	0.971	1.000	0.264
HetNet	0.576	0.739	0.068	0.927	0.834	0.075
HetNet†	0.635	0.798	<u>0.046</u>	<u>0.983</u>	0.808	<u>0.017</u>
SAT	0.569	0.757	0.081	0.882	0.516	0.118
SAT†	**0.724**	**0.858**	**0.040**	0.936	0.518	0.064
VMDNet	0.320	0.664	0.103	0.972	<u>0.238</u>	0.028
VMDNet†	0.449	0.616	0.095	0.910	0.893	0.090
Ours	0.513	0.774	0.070	**0.994**	**0.091**	**0.012**

detector performs correctly in detecting non-mirror regions, and a higher TNR indicates a better performance. We also measure the error rate (ER) by computing the ratio between the number of non-mirror frames detected to have mirrors and the number of total non-mirror frames. This measures how often a detector may detect mirrors that do not exist, and a lower ER indicates a better performance. We also compute the MAE for reference.

Comparing to State-of-the-arts

Quantitative Results. Table 1 reports the comparisons on our test set between ZOOM (weakly-supervised) and existing methods (pre-trained and finetuned in fully-supervised ways). Note that frames are tested following the chronological order but we separate frames into two groups, *i.e.*, frames with and without mirrors, for evaluation and discussion. Accordingly, several observations can be made. First, we can see that all existing methods tend to achieve better performance when they are finetuned on our dataset. This is understandable as the domain discrepancy exists. Second, comparing to image-based methods, the video-based VMDNet (Lin, Tan, and Lau 2023) shows relatively lower generalization ability. This is due to that it relies on modeling the temporal appearance correspondence, which is a strong assumption to hold in daily life scenes. Third, our method achieves the best performance on the non-mirror frames under all three metrics, which shows that our method tends to make less wrong predictions. Last, although our model is trained with extremely-weak supervisions, it still produces promising results comparing to fully-supervised methods on the mirror frames, which verifies the effectiveness of ZOOM. Table 2 further reports the comparisons on the VMD test set (Lin, Tan, and Lau 2023), which shows that ZOOM performs favorably against existing mirror detectors.

Qualitative Results. Figure 8 shows the visual comparisons between between ZOOM and state-of-the-art mirror detectors on our test set. Although we note that sometimes ZOOM may not produce pixel-wisely accurate mirror maps, it generally locates the mirrors well in these challenging cases.

Table 2: Quantitative results on the VMD dataset (Lin, Tan, and Lau 2023). Best results are marked in **bold** for reference.

Methods	F/W	IoU↑	F_β ↑	MAE↓
MirrorNet	F	0.505	0.681	0.145
PMDNet	F	0.532	0.749	0.128
VCNet	F	0.539	0.749	0.123
HetNet	F	**0.567**	0.751	0.120
SAT	F	0.318	0.564	0.334
VMDNet	F	**0.567**	**0.787**	**0.105**
Ours	W	0.294	0.448	0.387

Table 3: Ablation study. The upper, middle, and bottom parts compare different ablated versions of the localization, segmentation, and loss functions, respectively.

Methods	IoU↑	F_β ↑	TNR↑
Φ_{cam} *Only*	0.276	0.420	0.733
$\mathcal{C}^r \rightarrow \hat{\mathcal{C}}$	0.334	0.571	0.896
Our $\mathcal{C}^r$	0.473	0.668	0.990
f_i *Only*	0.318	0.465	0.801
$f_i + f_{i+1}$	0.422	0.658	0.935
$f_i + f_n$	0.392	0.633	0.907
$w/o\ \mathcal{L}_{fs}$	0.474	0.647	0.950
$w/o\ \mathcal{L}_{fc}$	0.492	0.695	0.928
$\mathcal{L}_{lov} \rightarrow \mathcal{L}_{proj}$	**0.520**	0.724	0.961
Ours	0.513	**0.774**	**0.994**

Internal Analysis

Ablation Study. We report ablation results in Table 3. We first analyze the qualities of pseudo labels, by removing the proposed temporal fusion and apply Φ_{cam} to individual frames to generate pseudo maps (denoted as "Φ_{cam} *Only*"). We then add the temporal fusion (Eq. 3) but exclude the feedback constraint (Eq. 4) (denoted as "$\mathcal{C}^r \rightarrow \hat{\mathcal{C}}$"). Last, we report the quality of pseudo labels generated by the proposed approach (denoted as "*Our* $\mathcal{C}^r$"). The upper part of Table 3 verifies the effectiveness of the temporal fusion for mirror localization and pseudo labels generation.

Next, we evaluate the components of the proposed temporal similarity-contrast modeling module. We first adapt our method to process the single frame only (denoted as "f_i *Only*"). We then investigate the effectiveness of modeling temporal coherence and contrast separately (denoted as "$f_i + f_{i+1}$" and "$f_i + f_n$", respectively). The middle three rows of Table 3 show that modeling either the similarity or contrast temporally improves the performance, while modeling both of them results in the best results (Ours). Moreover, we find that the performance of removing the SE block degrades from (IoU/F_β: 0.513/0.774) to (IoU/F_β: 0.433/0.682) while replacing the ASPP module with simple pooling operations yields to (IoU/F_β: 0.470/0.687).

Last, we analyze the loss terms in Eq. 8. Specifically, we first remove the feature similarity term $\mathcal{L}_{fs}$ and the feature contrast term $\mathcal{L}_{fc}$ separately (denoted as "$w/o\ \mathcal{L}_{fs}$" and

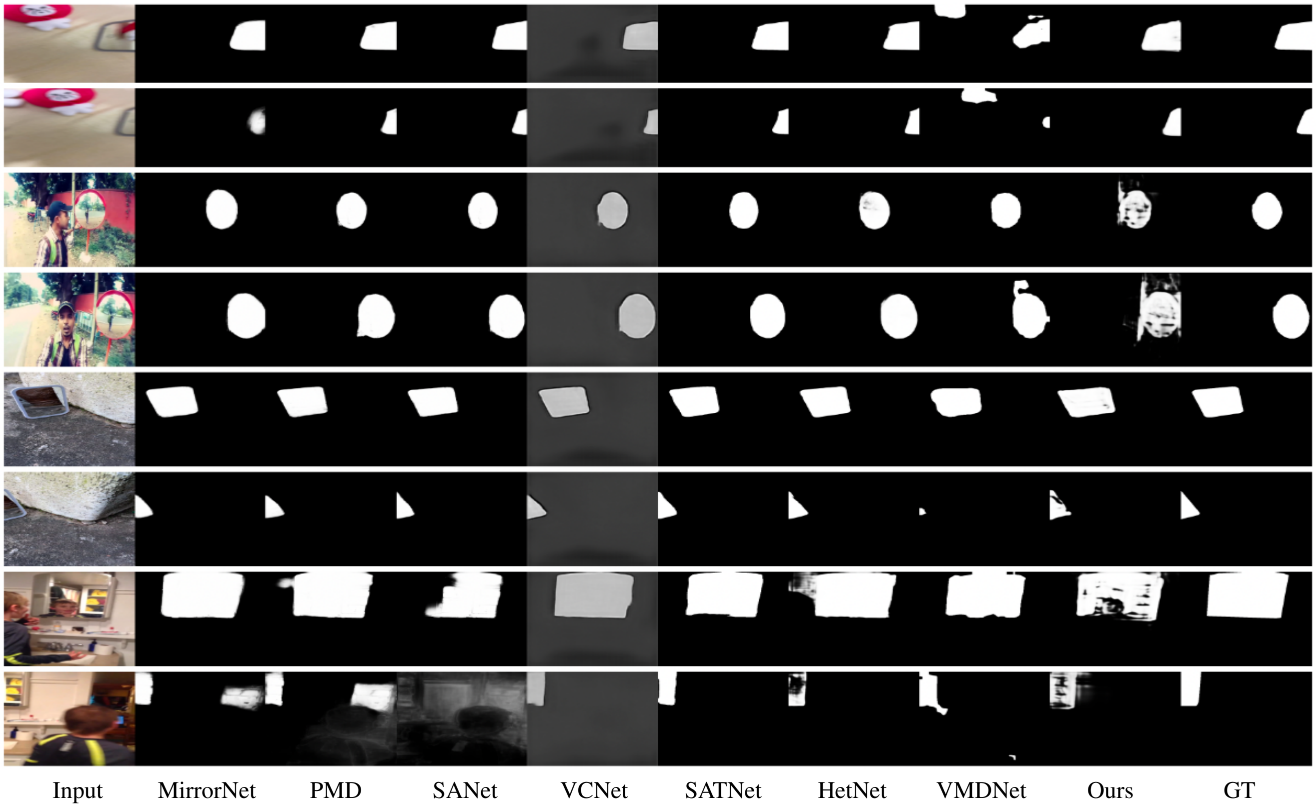

Figure 8: Visual comparison between ZOOM (weakly-supervised) and existing mirror detectors (fully-supervised).

Table 4: Models' efficiency in terms of model size and inference time for reference.

Methods	Reso.	Num. of Param.	FPS
MirrorNet	384×384	121.77	7.82
PMD	384×384	147.66	7.41
SANet	384×384	104.80	8.53
HetNet	352×352	49.92	49.23
VMDNet	384×384	62.24	17.06
Ours	352×352	57.26	19.02

"$w/o\ \mathcal{L}_{fc}$", respectively). We also replace the $\mathcal{L}_{lov}$ with the box loss proposed in $\mathcal{L}_{proj}$ (Tian et al. 2021) ("denoted as $\mathcal{L}_{lov} \rightarrow \mathcal{L}_{proj}$"). The bottom four rows of Table 3 shows the degraded performance when $\mathcal{L}_{fs}$ and $\mathcal{L}_{fc}$ are removed. Besides, while the projection loss may result in higher IoU, it degrades the performance of other metrics, as it tends to produce false positive errors. Generally, Table 3 verifies the effectiveness of our designs.

Model Efficiency. Table 4 compares the model size (in terms of number of parameters) and inference time (in terms of FPS), between ZOOM and existing methods. We can see that ZOOM performs at a reasonable computational cost.

Limitations. Our model does have some limitations. As shown in Figure 9(a), we note that our method may fail when the mirror is small (far away) across the whole video, which makes ZOOM hard to segment it accurately. ZOOM may also fail to mis-classify the presence of mirrors when the target scene contains mirror-like objects (*e.g.*, glossy objects with reflections/highlights in Figure 9(d)). Nonetheless, ZOOM can serve as a new baseline, to shed the light on exploring cheap labels for mirror detection in the wild.

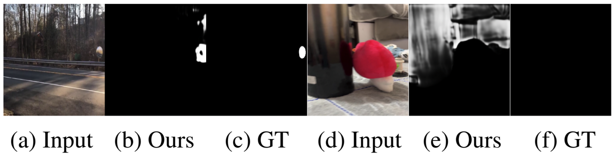

Figure 9: Failure cases. Our method may fail (a) when the mirror is small across the whole video, or (d) when there exists distracting mirror-like objects.

Conclusion

In this paper, we have proposed a novel approach, named ZOOM, to learn an effective mirror detector from extremely-weak annotations of per-frame ZerO-One Mirror indicators in videos. ZOOM leverages CAMs with a novel fusion strategy to model temporal consistency information for mirror localization. It also includes a novel temporal similarity-contrast modeling module for mirror segmentation. For training and evaluation of ZOOM, we have constructed a new video mirror dataset. We have conducted experiments on the proposed dataset as well as the existing mirror dataset under new and standard metrics. Results show that ZOOM performs favorably against existing fully-supervised mirror detection methods.

Acknowledgements

This project is in part supported by a GRF grant from the Research Grants Council of Hong Kong (No.: 11211223) and an SRG grant from City University of Hong Kong (No.: 7005674).

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *CVPR*.

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *CVPR*.

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*.

Chen, C.; Wang, G.; Peng, C.; Fang, Y.; Zhang, D.; and Qin, H. 2021. Exploring Rich and Efficient Spatial Temporal Interactions for Real-Time Video Salient Object Detection. *IEEE TIP*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587.

Dai, J.; He, K.; and Sun, J. 2015. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *ICCV*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

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

Feng, Z.; Guo, S.; Tan, X.; Xu, K.; Wang, M.; and Ma, L. 2022. Rethinking Efficient Lane Detection via Curve Modeling. In *CVPR*.

Gao, S.; Xing, H.; Zhang, W.; Wang, Y.; Guo, Q.; and Zhang, W. 2022a. Weakly Supervised Video Salient Object Detection via Point Supervision. In *ACM MM*.

Gao, S.; Zhang, W.; Wang, Y.; Guo, Q.; Zhang, C.; He, Y.; and Zhang, W. 2022b. Weakly-Supervised Salient Object Detection Using Point Supervison. In *AAAI*.

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

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

He, R.; Dong, Q.; Lin, J.; and Lau, R. W. 2023. Weakly-Supervised Camouflaged Object Detection with Scribble Annotations. In *AAAI*.

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

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*.

Hu, X.; Zhu, L.; Qin, J.; Fu, C.-W.; and Heng, P.-A. 2018. Recurrently aggregating deep features for salient object detection. In *AAAI*.

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

Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE TIP*.

Kim, B.; Jeong, J.; Han, D.; and Hwang, S. J. 2023. The Devil is in the Points: Weakly Semi-Supervised Instance Segmentation via Point-Guided Mask Representation. In *CVPR*.

Kweon, H.; Yoon, S.-H.; Kim, H.; Park, D.; and Yoon, K.-J. 2021. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *ICCV*.

Li, G.; Xie, Y.; and Lin, L. 2018. Weakly supervised salient object detection using image labels. In *AAAI*.

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

Liang, Z.; Wang, P.; Xu, K.; Zhang, P.; and Lau, R. W. 2022. Weakly-supervised salient object detection on light fields. *IEEE TIP*.

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

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

Liu, F.; Liu, Y.; Kong, Y.; Xu, K.; Zhang, L.; Yin, B.; Hancke, G.; and Lau, R. 2023. Referring Image Segmentation Using Text Supervision. In *ICCV*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

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

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NeurIPS Workshop*.

Piao, Y.; Wang, J.; Zhang, M.; and Lu, H. 2021. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *ICCV*.

Qin, J.; Wu, J.; Xiao, X.; Li, L.; and Wang, X. 2022. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *AAAI*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos. In *CVPR*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Song, H.; Wang, W.; Zhao, S.; Shen, J.; and Lam, K.-M. 2018. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*.

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

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

Tian, X.; Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. 2020. Weakly-supervised salient instance detection. In *BMVC*.

Tian, X.; Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. 2022. Learning to detect instance-level salient objects using complementary image labels. *IJCV*.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. Boxinst: High-performance instance segmentation with box annotations. In *CVPR*.

Tu, W.-C.; He, S.; Yang, Q.; and Chien, S.-Y. 2016. Real-time salient object detection with a minimum spanning tree. In *CVPR*.

Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *CVPR workshops*.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *CVPR*.

Wang, X.; Yu, Z.; De Mello, S.; Kautz, J.; Anandkumar, A.; Shen, C.; and Alvarez, J. M. 2022. Freesolo: Learning to segment objects without annotations. In *CVPR*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*.

Xie, Z.; Wang, S.; Xu, K.; Zhang, Z.; Tan, X.; Xie, Y.; and Ma, L. 2023. Boosting Night-time Scene Parsing with Learnable Frequency. *IEEE TIP*.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. 2019. Where is my mirror? In *ICCV*.

Yang, X.; Xu, K.; Chen, S.; He, S.; Yin, B. Y.; and Lau, R. 2018. Active matting.

Yu, S.; Zhang, B.; Xiao, J.; and Lim, E. G. 2021. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *AAAI*.

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020. Weakly-supervised salient object detection via scribble annotations. In *CVPR*.

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

Zhao, W.; Zhang, J.; Li, L.; Barnes, N.; Liu, N.; and Han, J. 2021. Weakly supervised video salient object detection. In *CVPR*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *CVPR*.