Weakly-Supervised Mirror Detection via Scribble Annotations

**Mingfeng Zha[1], Yunqiang Pei[1], Guoqing Wang[1]*, Tianyu Li[1],
Yang Yang[1], Wenbin Qian[2], Heng Tao Shen[1]**

[1]University of Electronic Science and Technology of China
[2]Jiangxi Agricultural University
zhamf1116@gmail.com, simon1059770342@foxmail.com, gqwang0420@uestc.edu.cn, cosmos.yu@hotmail.com,
yang.yang@uestc.edu.cn, qianwenbin1027@126.com, shenhengtao@hotmail.com

Abstract

Mirror detection is of great significance for avoiding false recognition of reflected objects in computer vision tasks. Existing mirror detection frameworks usually follow a supervised setting, which relies heavily on high quality labels and suffers from poor generalization. To resolve this, we instead propose the first weakly-supervised mirror detection framework and also provide the first scribble-based mirror dataset. Specifically, we relabel 10,158 images, most of which have a labeled pixel ratio of less than 0.01 and take only about 8 seconds to label. Considering that the mirror regions usually show great scale variation, and also irregular and occluded, thus leading to issues of incomplete or over detection, we propose a local-global feature enhancement (LGFE) module to fully capture the context and details. Moreover, it is difficult to obtain basic mirror structure using scribble annotation, and the distinction between foreground (mirror) and background (non-mirror) features is not emphasized caused by mirror reflections. Therefore, we propose a foreground-aware mask attention (FAMA), integrating mirror edges and semantic features to complete mirror regions and suppressing the influence of backgrounds. Finally, to improve the robustness of the network, we propose a prototype contrast loss (PCL) to learn more general foreground features across images. Extensive experiments show that our network outperforms relevant state-of-the-art weakly supervised methods, and even some fully supervised methods. The dataset and codes are available at https://github.com/winter-flow/WSMD.

Introduction

Mirrors are commonly used in everyday, but their reflective properties can disrupt tasks such as image enhancement (Wu et al. 2023) (Wang et al. 2021a) (Wang, Sun, and Sowmya 2019) (Wang, Sun, and Sowmya 2021), segmentation (Jain et al. 2021), and visual language navigation (An et al. 2021), making the study of mirror detection (MD) an important topic. Current research on MD utilizes pixel-level labels as supervised signals to train models. However, obtaining dense pixel labels is expensive. In this paper, we propose a weakly-supervised MD method. In the weakly supervised learning paradigm, there are four types of supervised signals: image-level, point-level, scribble-level, and box-level. We provide scribble annotation and use it to formulate our framework because it directly gives the location of mirror regions and offers flexibility in handling complex scenes. Therefore, we relabel 10,158 images, including 3,063 from MSD dataset (Yang et al. 2019), 5,095 from PMD dataset (Lin, Wang, and Lau 2020), 2,000 from Mirror-RGBD dataset (Mei et al. 2021) and name the new dataset S-Mirror. The labeling time for each image slightly differs as the varying scene complexity of these datasets, averaging around 5s, 6s, and 8s, respectively. As shown in Figure 2, the percentage of labeled pixels is less than 0.01 for most images, significantly lower than full annotation and relevant weak annotation works (about half of (He et al. 2023)).

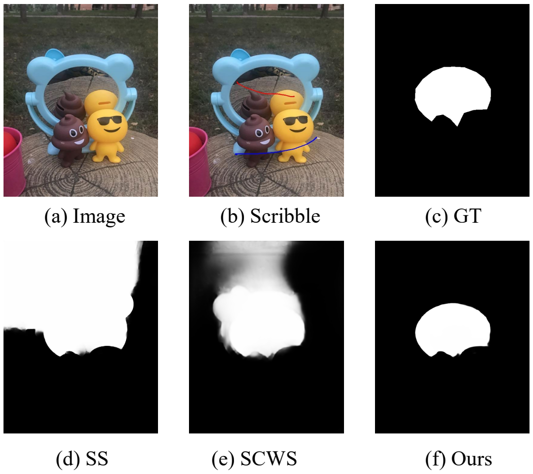

Figure 1: (a) Original image. (b) Scribbled image. (c) Ground-truth pixel-level annotations. (d) (Zhang et al. 2020) and (e) (Yu et al. 2021) are weakly supervised SOD models. (f) is our detection result.

Compared to traditional image detection tasks, MD shows some task-specific challenges: a) the scale of mirror regions varies greatly with some occupying more than half of the image and some occupying less than one tenth; b) many of the mirror regions are irregular and subject to occlusion; c) the diverse imagings and the varying surroundings of mirrors cause high noise as reflective property, thus making it a crucial task to distinguish between imagings (re-

[*]Corresponding author

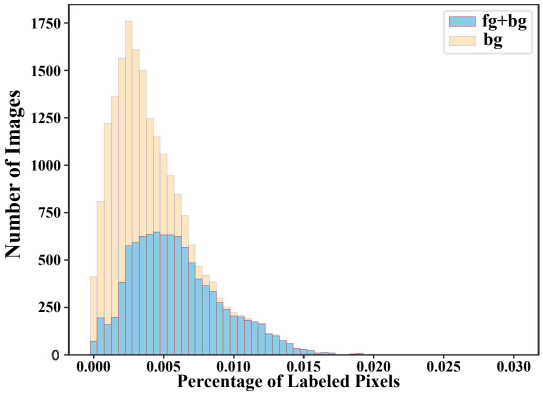

Figure 2: Percentage of labeled pixels in S-Mirror dataset. fg and bg denote the foreground and background, respectively, *i.e.,* the red and blue scribbles in Figure 1.

flective objects) and entities (objects outside the mirrors). See the supplementary material for some examples showing the above cases, which make it a great challenge to formulate a weakly-supervised MD framework, and there are only few related weakly supervised works, *i.e.*, scribble-based salient object detection (SOD) and camouflage object detection (COD). However, these methods cannot be directly applied to MD tasks due to the following reasons: 1) logical and physical associations between imagings and entities are not established; 2) mirror regions are not as salient as entities; 3) most camouflage objects have a single form, while mirror regions are diverse as reflection.

To resolve this, we for the first time formulate a weakly-supervised MD framework utilizing scribble-based supervision. As shown in Figure 1, our method achieves promising results. We propose a local-global feature enhancement (LGFE) module with both global context understanding (*e.g.,* establishing logical and physical associations between imagings and entities, mirror scale variation perception) and local details enhancement (*e.g.,* edges, textures, colors) to improve long- and short-distance dependence sensitivity. Moreover, scribble is difficult to represent the underlying structure information. The foreground feature representation is not salient and distinctive enough as reflection interference. Therefore, we propose a foreground-aware mask attention (FAMA), fusing the initial prediction foreground mask and edge mask for semantic and boundary awareness to refine the mirror mask. Furthermore, to improve the robustness of the network, we propose to mine the prototype features of various foreground and background features and formulate it as a novel prototype contrast loss (PCL), which aims at pulling the foreground prototypes closer, pushing the foreground and background prototypes away, thus producing more generalizable image feature representations.

In summary, our main contributions are as follows:

- We propose the first weakly supervised MD dataset based on scribble annotations. Compared to pixel-level annotations, quickly and flexibly annotating few pixels allows us to obtain the location and partial structure information of the foreground and background regions.
- We propose the first weakly supervised MD network that efficiently detects mirror regions with only simple scribble annotations and mirror edges as supervision signals.
- We formulate a local-global feature enhancement module (LGFE) and a foreground-aware mask attention (FAMA) to mitigate scale variation, occlusion, irregularity, and reflection interference. Additionally, we design a prototype contrast loss (PCL) to leverage inter-image information for improving network robustness.
- Extensive experiments on three mirror datasets show that our network outperforms relevant state-of-the-art methods on all evaluation metrics and achieves performance comparable to fully supervised approaches.

Related Works

Salient Object Detection. SOD aims to discover salient regions in images and has achieved significant progress. Ma *et al.* (Ma, Xia, and Li 2021) proposed aggregating adjacent feature layers to reduce interference. In recent years, some weakly supervised SOD works have also emerged. Zhang *et al.* (Zhang et al. 2020) proposed the first SOD method based on scribble annotations, which greatly reducing image annotation workload while achieving good performance. Yu *et al.* (Yu et al. 2021) proposed an end-to-end detection network based on structure consistency. Gao *et al.* (Gao et al. 2022) first proposed a multi-round training detection method based on point annotations. In addition, there are also similar works. For example, He *et al.* (He et al. 2023) first proposed a COD method based on scribble annotations, designing multiple functions to guide and constrain the model.

Mirror Detection. MD aims to detect mirror regions in images. Currently, there are many fully-supervised detection methods proposed. Yang *et al.* (Yang et al. 2019) first introduced the task and proposed MirrorNet, which explores feature differences inside and outside mirrors. Lin *et al.* (Lin, Wang, and Lau 2020) proposed a progressive detection approach, exploring local feature similarity. Guan *et al.* (Guan, Lin, and Lau 2022) discovered potential feature correlations from a semantic association perspective. In addition, some works attempt to explore characteristics of mirrors. Mei *et al.* (Mei et al. 2021) incorporated depth information because the depth of mirror regions can differ significantly from their surroundings. Huang *et al.* (Huang et al. 2023) designed a dual-stream network based on Swin Transformer (Liu et al. 2021b), using symmetry invariance. Some works also consider the constraints of practical application scenarios. For example, He *et al.* (He, Lin, and Lau 2023) designed a efficient network by selectively processing structures based on the differences between low-level and high-level features.

Methodology

Overview

The overall framework of our method is shown in Figure 3. It consists of four important parts, *i.e.*, edge generation (EG) module (four 1×1 convolutions), CFM (Dong et al. 2021), local-global feature enhancement module (LGFE), foreground-aware mask attention (FAMA) and

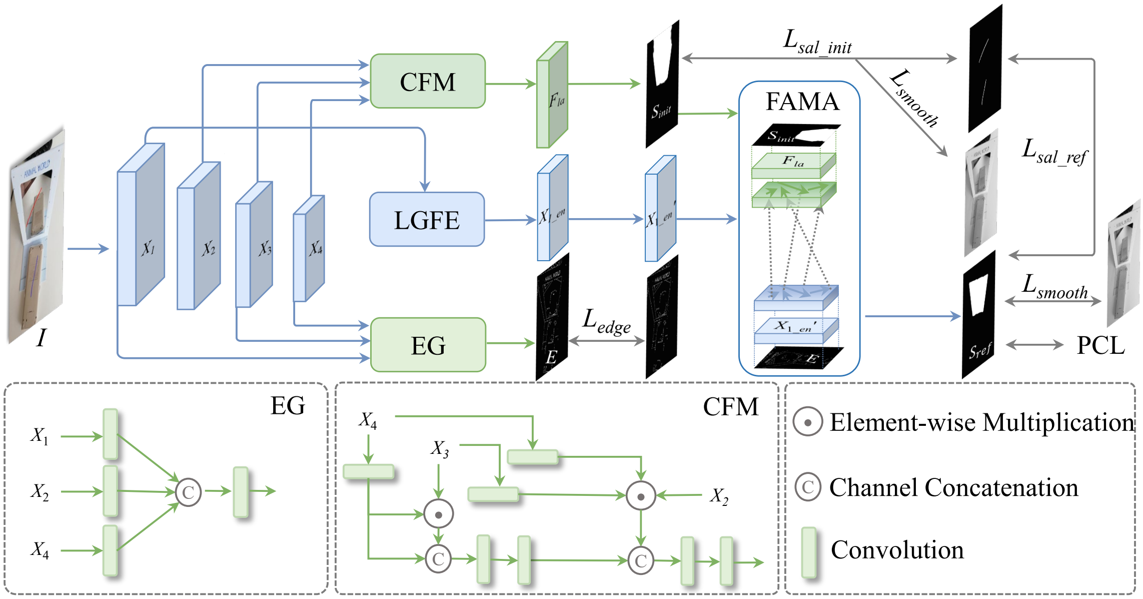

Figure 3: The overall structure of our proposed method. We first use PVT network (Wang et al. 2021b) as the backbone to extract multi-scale long-range dependency feature maps. We then utilize EG module to generate edge maps and LGFE module to enhance low-level feature maps. We progressively decode features using CFM and apply FAMA to fuse semantic and edge features. Finally, we use saliency maps, edges, and auxiliary PCL loss as the entire loss function to supervise model training.

prototype contrast learning loss (PCL). We first feed an image $I \in \mathbb{R}^{3\times H\times W}$ to generate multi-scale feature maps $X_i \in \mathbb{R}^{C_i\times\frac{H}{4^i}\times\frac{W}{4^i}}$, where $i \in \{1,2,3,4\}$, $C_i \in \{64,128,320,512\}$, H and W denote height and width respectively. We then feed low-level features X_1 and X_2, along with high-level feature X_4 into EG to produce edge map E. We also feed X_1 into LGFE to obtain the enhanced feature map X_{1_en}, which combines context and details information. Next, the initial prediction map S_{init} is decoded by progressively fusing X_2, X_3, and X_4. We integrate CFM's final feature map F_{la} with S_{init}, $X_{1\_en}^{'}$ (after adjustment based on X_{1_en}), and E jointly into FAMA. Through the semantic and edge-aware fusion, we generate refined prediction map S_{ref}. Furthermore, we design PCL as an auxiliary loss to enhance the model's robustness.

Local-global Feature Enhancement Module

We found that mirror regions can be highly variable in scale, irregular in shape, and prone to occlusion. Although the feature maps X_i generated from PVT network contain long-range dependencies and rich contextual semantics, they lack local information construction. In addition, in this paper, we introduce object edges as auxiliary supervised signals, which may introduce interference, particularly in weakly supervised scenarios. Therefore, enhancing local useful features and suppressing background information (*e.g.,* noisy texture, edges) is essential to retain details of mirror regions. To achieve this, we propose a local-global feature enhancement (LGEF) module to process X_1, as shown in Figure 4.

To illustrate, we create a duplicate of X_1 and name it X_{1_loc} to handle local features. For X_1, we employ Squeeze-and-Excitation (SE) Attention (Hu, Shen, and Sun 2018) and Convolutional Block Attention Module (CBAM) (Woo et al. 2018) to obtain the channel attention map ca_1 and spatial attention map sa_1, respectively,

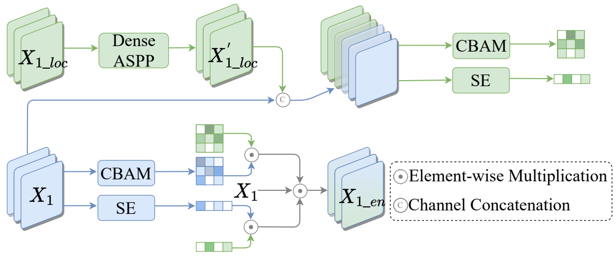

Figure 4: Structure of Local Global Feature Enhancement (LGFE) module. We first use DenseASPP on X_{1_loc} to obtain local features at different scales, and then use CBAM and SE on the fused feature maps to acquire spatial and channel attention, respectively. A similar process is performed on X_1. Finally, we fuse X_1 with four attentions.

$$ca_1 = SE(X_1), sa_1 = CBAM(X_1) \tag{1}$$

For X_{1_loc}, DenseASPP (Zhang et al. 2020) is first applied to extract local features using various dilation rates, generating the feature map $X_{1\_loc}^{'}$ that contains local perceptions. To further integrate contextual and local information while suppressing noise interference, we concatenate X_1 and $X_{1\_loc}^{'}$ along the channel axis and use 1×1 convolution to reduce channels by half. Subsequently, SE and CBAM are employed to obtain channel attention map ca_2

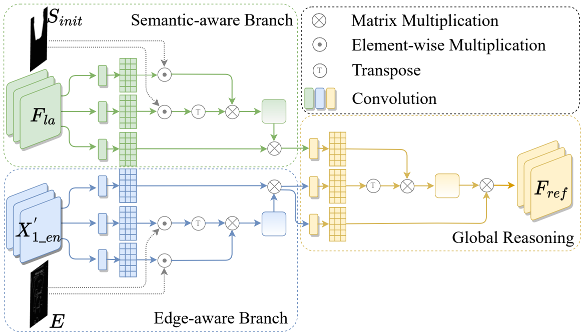

Figure 5: Structure of Foreground-Aware Mask Attention (FAMA). We first input F_{la}, S_{init} and X'_{1_en}, E to the semantic and edge-aware branches, respectively, and then to the cross-attention for fusion.

and spatial attention map sa_2. The process is formulated as:

$$\begin{aligned} ca_2 = SE(Conv_{1\times1}(concat(X_1, X_{1\_loc}{}'))), \\ sa_2 = CBAM(Conv_{1\times1}(concat(X_1, X_{1\_loc}{}'))) \end{aligned} \tag{2}$$

Finally, we fuse X_1 with the four attentions to generate the enhanced X_{1_en},

$$X_{1\_en} = (ca_1 \odot ca_2) \odot X_1 \odot (sa_1 \odot sa_2)) \tag{3}$$

where $\odot$ is element-wise multiplication. X_{1_en} can provide a robust foundation for the subsequent refinement of the prediction map.

Foreground-aware Mask Attention

Mirror regions are susceptible to interference from complex imagings and extra-mirror entities, resulting in less distinctive features from surroundings. Besides, weak annotations do not contain complete semantic regions, making it difficult to predict the object structure completely. To this end, we propose a foreground-aware mask attention (FAMA) that fuses foreground feature representation and edge guidance to obtain more complete mirror structure, as shown in Figure 5. Specifically, FAMA is divided into two branches: semantic-aware branch and edge-aware branch. The semantic-aware branch enhances the detection of mirror regions by incorporating a foreground mask prior, while the edge-aware branch refines the structure information by integrating edge maps. These two branches interact with each other to improve the overall detection quality.

The core module of FAMA is based on multi-Dconv head transposed attention (MDTA) (Zamir et al. 2022), an efficient improved self-attention (SA) (Vaswani et al. 2017), which can be expressed as:

$$MDTA(Q, K, V) = softmax(\frac{QK^T}{\alpha})V \tag{4}$$

The generation of Q, K, and V is similar to SA, with the difference that MDTA uses a 3×3 depth-wise convolution (Sandler et al. 2018) to encode local features. And MDTA explores global feature dependencies from the channel dimension rather than spatial. α is a learnable scaling parameter that allows the gradient to remain stable during training.

For the semantic-aware branch, the input $F_{la} \in \mathbb{R}^{32\times\frac{H}{8}\times\frac{W}{8}}$ is processed by 3x3 and 1x1 convolutions to generate the query, key, and value matrices. To compute the associations of the mirror region features, we perform element-wise multiplication of $S_{init} \in \mathbb{R}^{1\times\frac{H}{8}\times\frac{W}{8}}$ with the query and key matrices to obtain Q'_f and K'_f, while keeping the value matrix V_f unchanged. The subsequent operations are the same as those in MDTA. This process is written as:

$$F_{seg} = softmax(\frac{Q'_f K'^T_f}{\alpha})V_f \tag{5}$$

Similarly, for the edge-aware branch, we use the two inputs $X'_{1\_en} \in \mathbb{R}^{32\times\frac{H}{8}\times\frac{W}{8}}$ (adjust the size of $X_{1\_en} \in \mathbb{R}^{64\times\frac{H}{4}\times\frac{W}{4}}$ sequentially using 1×1 and 3×3 convolutions.) and $E \in \mathbb{R}^{1\times\frac{H}{8}\times\frac{W}{8}}$ to obtain Q'_e, K'_e, and V_e. The edge map E can be generated by:

$$E = EG(X_1, X_2, X_4) \tag{6}$$

Then we can obtain F_{edge} fused with edge priors,

$$F_{edge} = softmax(\frac{Q'_e K'^T_e}{\alpha})V_e \tag{7}$$

The features processed by these two branches possess semantic and edge contextual associations, respectively. To enrich the mirror region with more complex underlying structure features, we design a global reasoning module. Specifically, the semantic feature F_{seg} and the edge feature F_{edge} undergo the same convolutional processing to generate Q_s, K_{edge} and V_{edge}. The subsequent operations are the same as MDTA, generating F_{ref},

$$F_{ref} = softmax(\frac{Q_s K^T_{edge}}{\alpha})V_{edge} \tag{8}$$

Finally, we can obtain the refined prediction map $S_{ref} \in \mathbb{R}^{1\times\frac{H}{8}\times\frac{W}{8}}$ by compressing the channels of F_{refine} to 1 using a 1×1 convolution.

Prototype Contrast Loss

The semantic representation of mirror (forground) and non-mirror (background) regions in images differs, leading to closer feature distances for mirror regions and far distances between mirror and non-mirror regions in high-dimension feature space. Considering these, we design the PCL to learn more robust and essential feature representations.

In particular, we use $F_{sal} \in \mathbb{R}^{64\times\frac{HW}{64}}$ (Similar operations (Zhang et al. 2020) are performed based on S_{ref}, further fuse edge features and merge dimensions to generate) and $S_{ref} \in \mathbb{R}^{1\times\frac{HW}{64}}$ (After width, height expansion and dimensions merging) to generate foreground prototype feature $P_f \in \mathbb{R}^{1\times64}$, while background prototype feature $P_b \in \mathbb{R}^{1\times64}$ is generated using the background mask $1 - S_{ref} \in \mathbb{R}^{1\times\frac{HW}{64}}$ instead S_{ref}. So We have:

$$P_f = S_{ref} \otimes F^T_{sal}, P_b = (1 - S_{ref}) \otimes F^T_{sal} \tag{9}$$

Next, we use cosine similarity to calculate the distance sim between the two prototypes, and subsequently compute the negative sample (foreground and background prototypes pair) loss function. The sim is written as:

$$sim = \frac{P_f \cdot P_b}{\parallel P_f \parallel \times \parallel P_b \parallel} \tag{10}$$

where $\cdot$ represents dot product, $\parallel \cdot \parallel$represents l2 norm. If there are n samples, these sims will form a list.

Different samples have different inital similarity, and we tend to focus on negative samples with lower similarity and positive samples with higher similarity. To achieve this, we perform weighted calculations. The weight for the i-th element in the sim list can be expressed as:

$$w_i = \frac{e^{sim_i}}{\sum_j^n e^{sim_j}} \tag{11}$$

The weight list and the sim list can be multiplied correspondingly to get the weighted sim w_{sim}. Finally, the negative sample loss can be written as:

$$\mathcal{L}_- = -\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} log(1 - w_{sim}) \tag{12}$$

Similarly, we can get positive sample loss by calculating the distance between foreground features, writing:

$$\mathcal{L}_+ = -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} I_{[i \neq j]} log(w_{sim}) \tag{13}$$

where the function I represents 1 when i and j are not equal, and 0 otherwise.

Loss Function

Inspired by (Zhang et al. 2020), we adpot four functions to supervise the model training. Partial cross entropy (PCE) is used for the initial and refined saliency maps, *i.e.,* $\mathcal{L}_{sal\_init}$ and $\mathcal{L}_{sal\_ref}$. Smooth loss (SL) is employed to align the mirror region with image structure, *i.e.,* $\mathcal{L}_{smooth}$ (using the input grayscale map). Cross entropy (CE) is applied to the edge detection network, *i.e.,* $\mathcal{L}_{edge}$. Finally, PCL is utilized to reinforce foreground and background feature learning. The entire loss function can be defined as:

$$\begin{aligned} \mathcal{L}_{final} = PCE(S_{init}, mask) + PCE(S_{ref}, mask) \\ +SL(S_{init}, gray) + SL(S_{ref}, gray) \\ +\alpha CE(E, gt) + \beta(\mathcal{L}_- + \mathcal{L}_+) \end{aligned} \tag{14}$$

where mask denotes the product of the foreground and full scribble masks, gt is generated by the canny edge detector (Canny 1986). The performance may be better if a more advanced edge detection method is used, for example, RCF (Liu et al. 2017). α and β are hyperparameters.

Experiments

Datasets. We collect training images from MSD, PMD, and Mirror-RGBD datasets, totaling 10,158 images, and relabel them as the training set of S-Mirror dataset. Models are evaluated using the testing sets of the above three datasets.

Implementation Details. We implement our network using PyTorch and conduct experiments on an A100 GPU. Specifically, We use PVT network pretrained on ImageNet as the backbone to accelerate convergence. Various data augmentation methods are employed, such as random rotation, horizontal and vertical flipping. All images are resized to 352×352. During the training phase, the batch size is 16, the initial learning rate is 1e-4, the decay rate is 0.9, Adam is used as the optimizer, and the epoch is 150. We first train our model on MSD dataset and then use the trained model weights as initial weights for further training on PMD and Mirror-RGBD dataset. No post-processing strategies are used during the testing phase.

Evaluation Metrics. We use five evaluation metrics: S-measure (S_m) (Fan et al. 2017), mean E-measure (E_m) (Fan et al. 2018), weighted F-measure (F_β^w) (Margolin, Zelnik-Manor, and Tal 2014), Mean Absolute Error (MAE), and Intersection over union (IoU).

Comparison with State-of-the-arts

To demonstrate the superiority of our method, we first compare it with several state-of-the-art models on RGB-based MSD and PMD dataset. As shown in Table 1, we select eight SOD models, namely CPDNet (Wu, Su, and Huang 2019), MINet (Pang et al. 2020b), LDFNet (Wei et al. 2020), VST (Liu et al. 2021a), R3Net (Deng et al. 2018), EGNet (Zhao et al. 2019), PoolNet (Liu et al. 2019), SETR (Zheng et al. 2021), four MD models, namely MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), HetNet (He, Lin, and Lau 2023), SATNet (Huang et al. 2023), and three related weakly supervised models, namely SS (Zhang et al. 2020), SCWS (Yu et al. 2021), WSCOD (He et al. 2023). Our method outperforms all the weakly supervised models and achieves comparable performance to fully supervised SOD and MD models. More evaluations regarding the robustness of our method utilizing Precion-Recall and F-Measure curves are provided in the supplementary material. We also select some representative samples for visual comparison. As shown in Figure 6, the first row demonstrates scene where the mirror region is occluded, our method can effectively establish logical and physical associations of objects, distinguish between occlusion and mirror area. In the second row, there is significant mirror reflection, our method can accurately tell whether it is a imaging or an entity, achieving complete detection. The third and fourth rows show scenes with large scale mirror variations, Our method can capture long and short-range dependencies, obtaining accurate mirror regions.

We also compare our method with seven RGBD SOD models, namely A2dele (Piao et al. 2020), HDFNet (Pang et al. 2020a), S2MA (Liu, Zhang, and Han 2020), JL-DCF (Fu et al. 2020), DANet (Zhao et al. 2020), BBSTNet (Fan et al. 2020), VST (Liu et al. 2021a), and two MD models, namely PDNet (using depth information) and SATNet, as well as RGB-based SS, SCWS, and WSCOD on Mirror-RGBD dataset. As shown in Table 2, our method also outperforms all the related weakly supervised detection methods and reduces the gap with fully supervised methods. We select several examples for comparison. As shown in Figure 7,

Methods	Sup.	MSD					PMD				
		S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓
CPDNet	F	0.725	0.770	0.625	0.576	0.116	0.779	0.817	0.651	0.600	0.041
MINet	F	0.792	0.819	0.715	0.664	0.088	0.794	0.822	0.667	0.601	0.038
LDF	F	0.821	0.867	0.773	0.729	0.068	0.799	0.833	0.683	0.633	0.038
VST	F	0.861	0.901	0.818	0.791	0.054	0.783	0.814	0.639	0.591	0.036
R3Net	F	0.723	0.743	0.615	0.554	0.111	0.720	0.756	0.561	0.496	0.045
EGNet	F	0.771	0.776	0.668	0.630	0.096	0.617	0.593	0.362	0.210	0.088
PoolNet	F	0.804	0.831	0.717	0.691	0.094	0.588	0.532	0.313	0.192	0.089
SETR	F	0.797	0.840	0.750	0.690	0.071	0.753	0.775	0.633	0.564	0.035
MirrorNet	F	0.850	0.891	0.812	0.790	0.065	0.761	0.841	0.663	0.585	0.043
PMDNet	F	0.875	0.908	0.845	0.815	0.047	0.810	0.859	0.716	0.660	0.032
HetNet	F	0.881	0.921	0.854	0.824	0.043	0.828	0.865	0.734	0.690	0.029
SATNet	F	0.887	0.916	0.865	0.834	0.033	0.826	0.858	0.739	0.684	0.025
SS	W	0.681	0.747	0.567	0.527	0.158	0.726	0.790	0.571	0.513	0.055
SCWS	W	0.770	0.814	0.678	0.659	0.121	0.759	0.807	0.599	0.579	0.059
WSCOD	W	0.786	0.851	0.728	0.685	0.092	0.764	0.819	0.609	0.586	0.055
Ours	W	**0.828**	**0.878**	**0.780**	**0.750**	**0.078**	**0.773**	**0.824**	**0.630**	**0.600**	**0.051**

Table 1: Quantitative comparison on MSD and PMD datasets with five evaluation metrics. F, W denote fully supervised and weakly supervised, respectively. The best weakly supervised performances are bolded.

Methods	Mirror-RGBD				
	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓
A2dele	0.641	0.730	0.505	0.428	0.120
HDFNet	0.671	0.663	0.521	0.447	0.095
S2MA	0.765	0.797	0.646	0.609	0.075
JL-DCF	0.815	0.861	0.750	0.696	0.057
DANet	0.800	0.842	0.728	0.678	0.063
BBSTNet	0.840	0.881	0.786	0.743	0.048
VST	0.815	0.859	0.751	0.702	0.054
PDNet	0.856	0.906	0.825	0.778	0.042
SATNet	0.857	0.901	0.829	0.784	0.031
SS	0.654	0.722	0.537	0.444	0.127
SCWS	0.690	0.743	0.547	0.498	0.118
WSCOD	0.698	0.762	0.581	0.518	0.106
Ours	**0.754**	**0.806**	**0.655**	**0.616**	**0.088**

Table 2: Quantitative comparison on Mirror-RGBD dataset with five evaluation metrics. The best weakly supervised performances are bolded.

the first row demonstrates that our method can exploit context and obtain complete detection results when the mirror region is similar to the surroundings and has a large scale. In the second row, the mirror region has a small scale, causing A2dele to even miss, but our method can determine. The third row shows that our method can establish the relationship between multiple objects. Although our method does not use depth information, it still performs well.

To verify the lightness of our model, we compare it with related weakly supervised models. As shown in Table 3, our method is also efficient.

Methods	Input Size	Params.	FLOPs
SS	352×352	16.80	70.85
SCWS	352×352	63.54	53.80
WSCOD	352×352	32.65	14.27
Ours	352×352	26.16	21.39

Table 3: Model Efficiency Comparison. We compare with three related weakly supervised models on Parameters (M), FLOPs (GMAC).

Ablation Study

We conduct ablation experiments on MSD dataset, as shown in Table 4. We also select a representative image to visualize the ablation process, as shown in Figure 8.

Effect of LGFE. Based on the Baseline, we enhance X_1 by adding an LDEF module to obtain richer feature representations with more semantic and detailed information. As a result, we achieve improvements of 1.7%, 2.4%, 2.8%, 2.4%, 1.5% on the S_m, E_m, F_β^w, IoU, and MAE metrics, respectively. LGFE module can establish global and local dependencies, effectively distinguishing between imagings and objects. The visualization results show that after adding a LGFE module, the non-mirror region is significantly reduced without affecting the mirror area.

Effect of FAMA. We evaluate the performance of FAMA on both the Baseline and "Baseline+LGFE" network. Compared to the Baseline, we observe improvements of 2.5%, 3.4%, 4.6%, 4.0%, and 2.1% on the five metrics, respectively. After adding a LGFE module, the performance is further enhanced, demonstrating the complementarity of the two modules. The visualization results also show that the addition of FAMA effectively integrated edge features, reduce

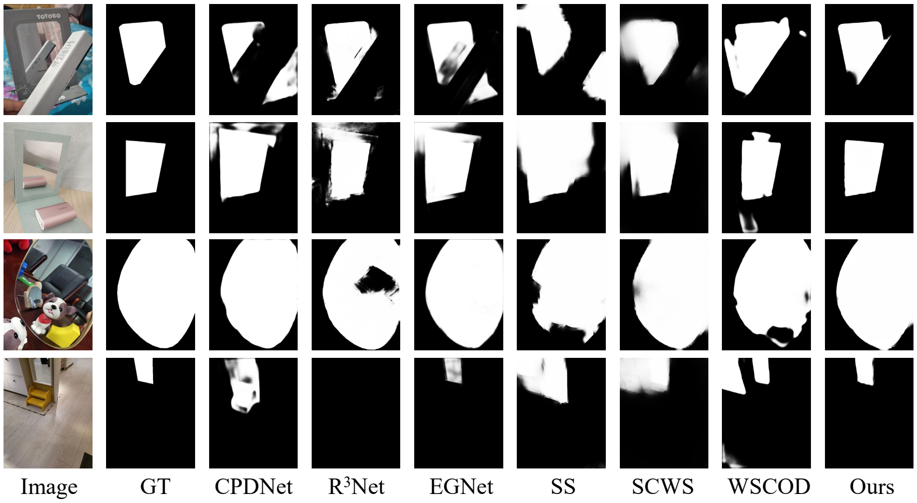

Figure 6: Qualitative comparison on MSD and PMD datasets. Occlusion, mirror reflection, large-scale and small-scale scenes are shown from top to bottom.

Method	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓
B	0.793	0.833	0.717	0.695	0.106
B+I1	0.810	0.857	0.745	0.719	0.091
B+I2	0.818	0.867	0.763	0.735	0.085
B+I3	0.799	0.845	0.725	0.701	0.098
B+I1+I2	0.820	0.873	0.772	0.740	0.081
Ours	**0.828**	**0.878**	**0.780**	**0.750**	**0.078**

Table 4: Results of ablation study on MSD dataset. B, I1, I2, and I3 indicate Baseline, LGFE, FAMA, and PCL, respectively. Based on our good baseline and added incrementally, the proposed method reaches the best performances (bolded data).

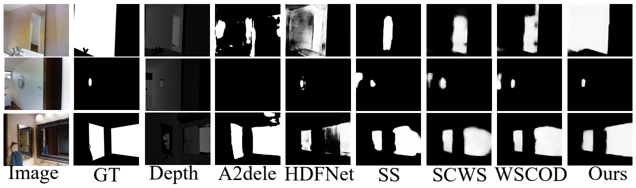

Figure 7: Qualitative comparison on Mirror-RGBD dataset, showing large-scale&similar to surroundings, small-scale and multi-objects scenes from top to bottom.

mirror interference, leading to more accurate foreground detection. Although the "Baseline+LGFE+FAMA" network achieves promising detection results, very close to the GT, it suffers from the issue of excessive de-interference.

Effect of PCL. Similar to evaluating FAMA, we introduce PCL as an auxiliary loss to the Baseline and "Baseline+LGFE+FAMA" network. If the obtained mirror features are not accurate enough, foreground and background prototypes may contain noise, resulting in little improvements. On the contrary, with the addition of LGFE module and FAMA, the mirror regions become more complete, leading to significant improvements. The visualization results after adding PCL to the Baseline show that the model mistakenly identifies the lower right area of the image as a mirror, despite greatly reducing the misidentified area on the left. Based on the "Baseline+LGFE+FAMA" network, PCL can fully utilize the high-quality foreground features among images to alleviate over-detection.

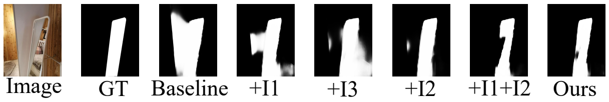

Figure 8: Visualization results of ablation study. Baseline and stage models suffer from over- or under-detection. Our method achieves more accurate detection.

Conclusion

In this paper, we propose the first scribble-based weakly supervised MD dataset, requiring less than 0.01 of pixel annotation and offering a simple and flexible process. Using the relabeled dataset, we propose a novel MD framework with three carefully designed components. Firstly, we propose a local-global feature enhancement (LGFE) module to tackle problems such as scale variation, irregularity, and occlusion of mirror region, thereby improving the representation quality for fine details. Secondly, we design a foreground-aware mask attention (FAMA) by combining foreground semantics and edge features, which promotes the expansion and completeness of scribble regions while reducing interference from mirror imaging. Finally, we formulate a prototype contrast loss (PCL) to learn the similarity of foreground-background semantic features between images, enabling more robust feature representations. Extensive experiments show that our method surpasses state-of-the-art weakly supervised approaches, achieving performance comparable to fully supervised learning while being lightweight.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Sichuan Science and Technology Program under grant 2022YFG0032, and the China Academy of Space Technology (CAST) Innovation Program.

References

An, D.; Qi, Y.; Huang, Y.; Wu, Q.; Wang, L.; and Tan, T. 2021. Neighbor-view enhanced model for vision and language navigation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5101–5109.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 679–698.

Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; and Heng, P.-A. 2018. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 684–690.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*.

Fan, D.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 698–704.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 4548–4557.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 2075–2089.

Fu, K.; Fan, D.-P.; Ji, G.-P.; and Zhao, Q. 2020. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 3052–3062.

Gao, S.; Zhang, W.; Wang, Y.; Guo, Q.; Zhang, C.; He, Y.; and Zhang, W. 2022. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 670–678.

Guan, H.; Lin, J.; and Lau, R. W. 2022. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5941–5950.

He, R.; Dong, Q.; Lin, J.; and Lau, R. W. 2023. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 781–789.

He, R.; Lin, J.; and Lau, R. W. 2023. Efficient Mirror Detection via Multi-Level Heterogeneous Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 790–798.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R. W.; and Zuo, W. 2023. Symmetry-Aware Transformer-based Mirror Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 935–943.

Jain, J.; Singh, A.; Orlov, N.; Huang, Z.; Li, J.; Walton, S.; and Shi, H. 2021. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3697–3705.

Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3917–3926.

Liu, N.; Zhang, N.; and Han, J. 2020. Learning selective self-mutual attention for RGB-D saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13756–13765.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021a. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4722–4732.

Liu, Y.; Cheng, M.-M.; Hu, X.; Wang, K.; and Bai, X. 2017. Richer convolutional features for edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3000–3009.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Ma, M.; Xia, C.; and Li, J. 2021. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2311–2318.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3044–3053.

Pang, Y.; Zhang, L.; Zhao, X.; and Lu, H. 2020a. Hierarchical dynamic filtering network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 235–252.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020b. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9413–9422.

Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; and Lu, H. 2020. A2dele: Adaptive and attentive depth distiller for efficient

RGB-D salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9060–9069.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, G.; Sun, C.; and Sowmya, A. 2019. Erl-net: Entangled representation learning for single image de-raining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5644–5652.

Wang, G.; Sun, C.; and Sowmya, A. 2021. Context-enhanced representation learning for single image deraining. *International Journal of Computer Vision*, 1650–1674.

Wang, G.; Yang, Y.; Xu, X.; Li, J.; and Shen, H. 2021a. Enhanced context encoding for single image raindrop removal. *Science China Technological Sciences*, 2640–2650.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021b. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13025–13034.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1662–1671.

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3907–3916.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. 2019. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8809–8818.

Yu, S.; Zhang, B.; Xiao, J.; and Lim, E. G. 2021. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3234–3242.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12546–12555.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8779–8788.

Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; and Zhang, L. 2020. A single stream network for robust and real-time RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 646–662.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 6881–6890.