Mirror Detection via Multi-Directional Similarity Perception and Spectral Saliency Enhancement

Zhiwen Shao, Rui Chen, Xuehuai Shi, Bing Liu, Canlin Li, Lizhuang Ma, and Dit-Yan Yeung

***Abstract*—Mirror detection is a challenging task, due to the reflective properties of mirrors. Most existing approaches rely on exploiting the relationship between the content inside the mirror and the surrounding environment to aid in locating mirrors. A typical solution is to utilize contextual contrasted features. However, the discontinuity in content at the edges of mirrors may not always be prominent. To overcome this limitation, we propose a novel mirror detection framework called S^{2}MD including two main modules, multi-directional similarity perception module (MSPM) and spectral saliency enhancement decoder module (SSEDM). Specifically, we employ a backbone network to extract multi-scale global information from images using a dual-path approach. Then, we feed these high-level dual-path features into MSPMs to generate direction-sensitive similarity-consistent features. MSPM utilizes active rotating filters and oriented response pooling to model the similarity relations in different orientations. Moreover, the SSEDM is utilized to enhance the spatial contextual contrasted features using feature spectral residuals and fuse the dual-path features to obtain the final predicted mirror mask. Extensive experiments demonstrate that our method achieves state-of-the-art performance on challenging MSD, PMD, and RGBD-Mirror benchmarks. The code is available at https://github.com/RuiChen-stack/M2SD.**

***Index Terms*—Mirror detection, multi-directional similarity perception, spectral residual**

Manuscript received 9 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62472424, in part by the China Postdoctoral Science Foundation under Grant 2023M732223, in part by the Hong Kong Scholars Program under Grant XJ2023037/HKSP23EG01, and in part by the Research Impact Fund of the Hong Kong Government under Grant R6003-21. It was also supported in part by the National Natural Science Foundation of China under Grants 62402231, 62276266, and 72192821, in part by the Opening Fund of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant VRLAB2024C03, and in part by the Science and Technology Planning Project of Henan Province under Grant 242102211003. (Corresponding authors: Rui Chen, Xuehuai Shi, and Dit-Yan Yeung.)

Z. Shao, R. Chen, and B. Liu are with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China, and also with the Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China. Z. Shao is also with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon 999077, Hong Kong, and also with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: {zhiwen_shao; rui_chen; liubing}@cumt.edu.cn).

X. Shi is with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: xuehuai@njupt.edu.cn).

C. Li is with the School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China (e-mail: li-cl@zzuli.edu.cn).

L. Ma is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: malz@cs.sjtu.edu.cn).

D.-Y. Yeung is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon 999077, Hong Kong (e-mail: dyyeung@cse.ust.hk).

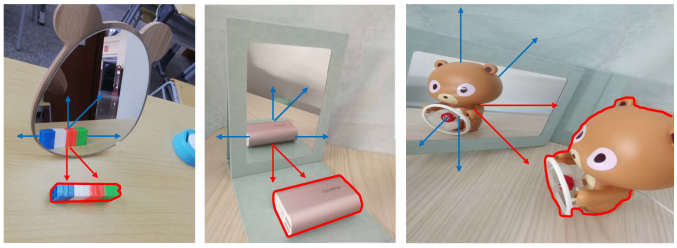

Fig. 1. Illustration of similarities in different directions in example mirror images. Due to the reflection of mirrors, there is a certain similarity between the content in the mirror and the real objects. In the leftmost image, the similarity is mainly observed in the vertical direction, in the middle image, the similarity is predominantly seen along diagonal lines, and in the rightmost image, the two similar teddy bears are roughly aligned in a horizontal direction.

I. Introduction

Mirrors are commonly encountered in everyday life. Due to their reflective nature, many computer vision tasks may mistakenly predict the content within mirrors as real objects, leading to potential performance degradation. Therefore, mirror detection plays a crucial role in computer vision and image processing, which involves segmenting the mirror regions from a given image. An intuitive way is to explore the differences between the content within the mirror and real objects to assist in locating mirrors. However, the diverse appearance variations caused by mirror reflections still pose a significant challenge in mirror detection.

In past five years, different properties of the relationship between the content within mirrors and the content outside of mirrors are taken into consideration. Yang *et al.* [1] observed the discontinuity of content at the edges of mirrors and proposed using contextual contrasted features to distinguish mirrors from the background. However, this method may fail when the content at the mirror edges is similar. Guan *et al.* [2] noticed that people typically place mirrors in specific locations, and thus proposed a method to learn the semantic associations between mirrors and scenes. However, this approach overly relies on the complexity of the scene.

Recently, Huang *et al.* [3] discovered that the content within mirrors exhibits a symmetric relationship with real objects, but it is not entirely symmetrical. This relationship is referred to as loose symmetry. They designed SATNet to perceive such loose symmetry relationship. However, capturing this type of symmetry is challenging, and SATNet only considers symmetry in the horizontal direction. In response to SATNet's limitation, we rethink the concept of symmetry. Symmetry refers to the equal or similar relationship in terms of shape,

size, arrangement, and other aspects of a pattern or object with respect to a point, line, or plane. *The relationship of symmetry is fundamentally a type of similarity relationship*, as illustrated in Fig. 1. To enhance the robustness of learning the relationship between mirror content and real objects, we opt to utilize multi-directional similarity perception instead of loose symmetry relationships.

Based on the aforementioned perspective, we propose a new mirror detection framework called S^2MD. In particular, S^2MD is a dual-path network architecture. We take the original image and its horizontally and vertically flipped counterpart as inputs. To perceive similarity relationships in different directions, we design a new multi-directional similarity perception module (MSPM). It utilizes active rotating filters and oriented response pooling to generate direction-sensitive similarity-consistent features. Moreover, in order to enhance the saliency of contextual contrasted features, we design a spectral saliency enhancement decoder module (SSEDM). It integrates feature spectral residuals with spatial contextual contrasted features to enhance their saliency, and employs dual-path feature fusion and refinement.

The main contributions of this work are threefold:

- We propose a novel multi-directional sensitive feature similarity to model the relationship between the inside and outside content of mirrors, which is integrated into a multi-directional similarity perception module.
- We propose a new spectral saliency enhancement decoder module to enhance the saliency of spatial contextual contrasted features.
- We conduct extensive experiments on three challenging mirror datasets, in which both quantitative and visual results demonstrate the effectiveness of our approach.

II. Related Work

In this section, we review the previous methods that are closely related to our proposed approach, including mirror detection, salient object detection, and local feature matching.

A. Mirror Detection

Mirror detection is a challenging task, as mirrors often lack specific visual features due to their reflective nature. Current methods mostly approach this task by examining the relationship between the reflected content in the mirror and the surrounding objects. Yang *et al.* [1] proposed MirroNet, which is the first method designed for mirror detection. It leverages the discontinuity of content at the edges of mirrors to aid in the detection process. Lin *et al.* [4] developed PMDNet, which progressively extracts relational contextual contrasted features to learn similar relationships. Guan *et al.* [2] proposed SANet, which achieves reliable mirror localization by exploring the semantic association between mirrors and surrounding objects. Tan *et al.* [5] were inspired by visual chirality property and embedded visual chirality cues into detection models to help detect mirror flip. However, these methods still struggle to accurately localize mirror regions, primarily due to the difficulty in capturing such relational properties.

Some other methods take into account additional information that reflects the properties of the mirror. Mei *et al.* [6] introduced depth information into mirror detection. Xie *et al.* [7] proposed a cross-spatial-frequency window transformer (CSFwinformer) to extract spatial and frequency features for mirror texture analysis. These methods have significant limitations on improving model performance. Recently, Huang *et al.* [3] proposed SATNet by capturing the loose symmetry relationship between mirror reflection content and real objects. However, it still has its shortcomings. It only considers loose symmetric relationships in the horizontal direction and is not sensitive to direction. In addition, it uses simple contextual contrasted features and is not suitable for situations where content discontinuity is not obvious.

In contrast with the above methods, regard the relationship between mirror content and real objects as a type of similarity relationship instead of loose symmetry relationships, and improve the performance of mirror detection from the perspectives of multi-directional similarity perception and spectral saliency enhancement.

B. Salient Object Detection

Salient object detection is widely used in various computer vision tasks, with the goal of segmenting the salient object area of an image from the background. Early methods are mostly sparse detection methods [9]–[11]. With the tremendous success of fully convolutional neural network (FCN) [12] in pixel-level semantic segmentation, pixel-level dense detection methods [13]–[17] have emerged and gradually become the mainstream method for salient object detection. For example, Wang *et al.* [18] proposed a new method called multiple enhancement network (MENet), which adopts the boundary sensibility, content integrity, iterative refinement, and frequency decomposition mechanisms of HVS to improve the accuracy and robustness of the model in complex scenes. Wang *et al.* [19] introduced wavelet transform theory into neural networks and proposed a fast and lightweight wavelet neural network (ELWNet) for real-time salient object detection.

Recently, co-salient object detection has become another research hotspot in the field of Salient Object Detection. It aims to detect salient objects across image groups rather than in a single image. Most existing works [20]–[24] use attention mechanisms to explore the consistency between image groups. Fan *et al.* [25] proposed CoEG-Net with a co-attention projection strategy to achieve fast common information learning. Zheng *et al.* [23] proposed a novel memory-aided contrastive consensus learning (MCCL) framework, which is capable of effectively detecting co-salient objects in real time.

While mirror edges may possess some visual saliency, salient object detection methods may not be directly applicable to mirror detection due to the influence of mirror content. In our work, we enhance the saliency of mirrors by employing the spectral residuals.

C. Local Feature Matching

Feature matching is a crucial task in the field of computer vision, which involves finding corresponding feature points

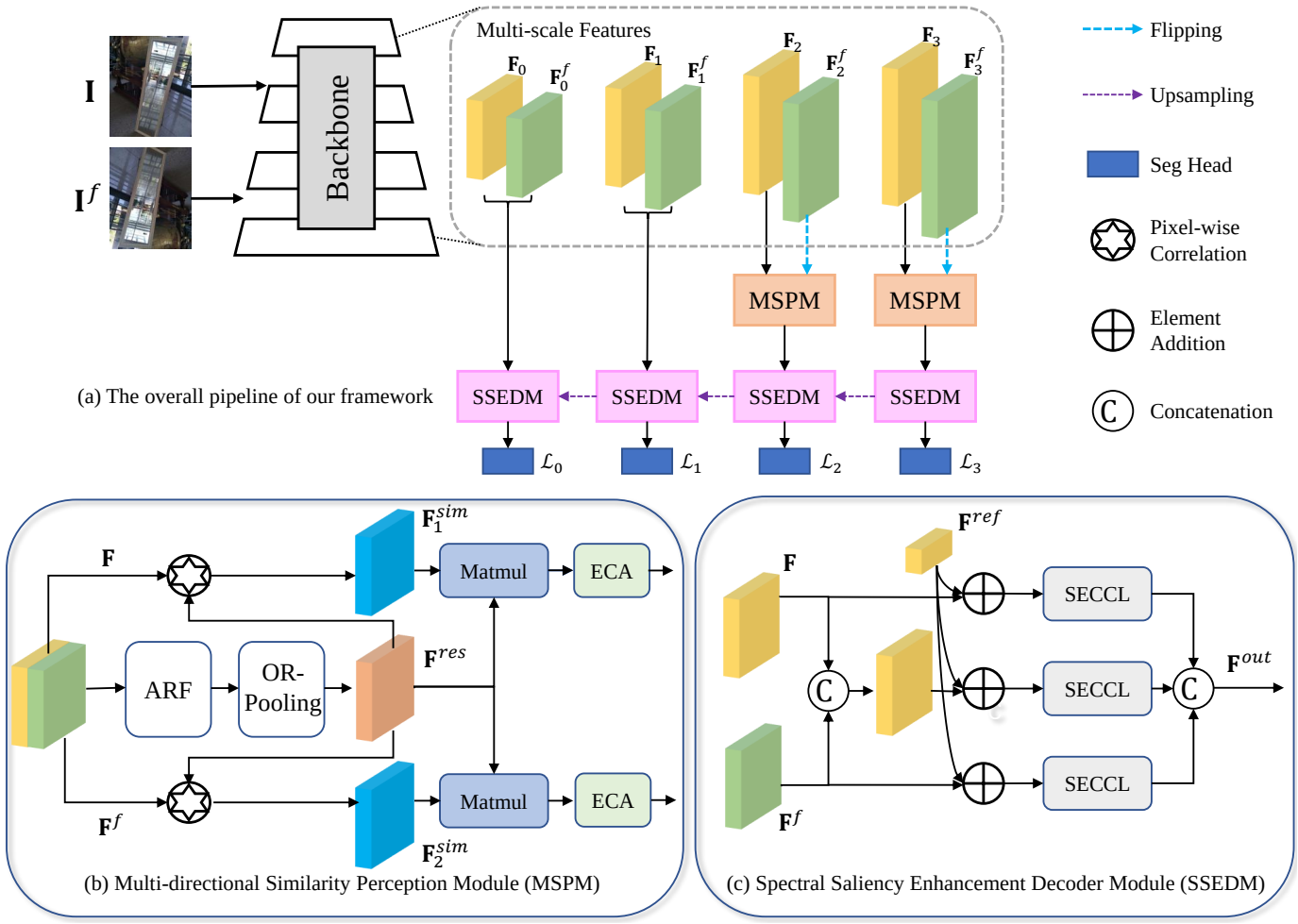

Fig. 2. The architecture of our framework, in which the structures of multi-directional similarity perception module (MSPM) and spectral saliency enhancement decoder module (SSEDM) are shown in (b) and (c), respectively. Given a mirror image, we first generate its horizontally and vertically flipped counterpart, then feed the two images into a weight-sharing backbone [8] to extract multi-scale global information, in which the outputs by the highest two scales are input to MSPMs to model multi-directional similarity. Further, the outputs by all scales are input to SSEDMs to learn contextual contrasted features with saliency enhancement and feature fusion. Finally, the segmentation head following each decoder is used to predict the mirror mask, in which the output by the last segmentation head is treated as the final prediction.

or feature descriptors in different images or visual scenes. The goal of feature matching is to establish associations between similar or corresponding feature points in two or more images, enabling applications such as image alignment, object recognition, and 3D reconstruction. Currently, the mainstream approach is detector-based local feature matching [26]–[29]. This method utilizes feature detectors to find salient points or regions in an image and calculates descriptors for each point. Then, by comparing the descriptors across different images, corresponding feature points or regions can be found, enabling image matching and association. Detector-free local feature matching methods [30]–[33] remove the feature detector phase and directly produce dense descriptors or dense feature matches.

In this paper, we consider the input image and its horizontally and vertically flipped counterpart, and match and fuse the dual features extracted in our framework.

III. Proposed Method

A. Overview

Fig. 2(a) shows the pipeline of our S^{2}MD framework. S^{2}MD adopts a dual-path structure. The use of a dual-path structure is beneficial for enhancing the similarity between features. We take both the original image and the image obtained by horizontally and vertically flip the original image as inputs. We extract multi-scale global information through a weight-sharing backbone, in which the structure of backbone is Swin-S [8].

Specifically, given an input image $\mathbf{I}$ as well as its flipped image $\mathbf{I}^f$, we feed them into the backbone to obtain multi-scale features $\{\mathbf{F}_0, \cdots, \mathbf{F}_3\}$ and corresponding flipped features $\{\mathbf{F}_0^f, \cdots, \mathbf{F}_3^f\}$, respectively. We feed the features from the highest two scales of the dual paths into the multi-directional similarity perception modules (MSPMs) to learn direction-sensitive features and model multi-directional similarity. Then, the paired features at different scales are fed into the spectral saliency enhancement decoder module (SSEDM) to learn con-

textual contrasted features and perform saliency enhancement and feature fusion. Finally, we obtain the prediction masks from each decoder by using the segmentation head. Each prediction mask is compared with the corresponding ground truth at the same scale to calculate the sub-loss, denoted as $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, respectively. The final prediction result of our framework is generated by the last segmentation head.

B. Multi-directional Similarity Perception Module

The MSPM is designed to perceive similar relationships in different directions while considering the enhanced features from dual pathways. First, the feature maps $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$ and $\mathbf{F}^f \in \mathbb{R}^{B\times C\times H\times W}$ are concatenated together to obtain the feature $\mathbf{F}^c \in \mathbb{R}^{B\times 2C\times H\times W}$. Then, active rotating filters (ARFs) [34] are utilized to extract direction-sensitive features.

ARF generates N directional channels of feature maps by actively rotating the canonical filter $N-1$ times. Specifically, denote the canonical filter of ARF as $\mathcal{F}_0 \in \mathbb{R}^{k\times k\times N}$, where k represents the width of the convolutional kernel, and N represents the number of rotations. ARF generates $N-1$ clones of $\mathcal{F}_0$ by rotating it to different angles, so ARF can be seen as a combination of N filters. Let $\mathbf{F}^i(j)$ and $\mathbf{F}^o(j)$ denote the input feature map and the output feature map of the j-th orientation, respectively. The computation process of ARF is defined as

$$\mathbf{F}^o(j) = \sum_{n=0}^{N-1} \mathcal{F}_j(n) * \mathbf{F}^i(n), \quad j = 0, \cdots, N-1, \tag{1}$$

where $\mathcal{F}_j \in \mathbb{R}^{k\times k\times N}$ represents the clone generated by rotating $\mathcal{F}_0$ with j times, and $\mathcal{F}_j(n)$ indicates the n-th orientation channel of $\mathcal{F}_j$. We obtain the orientation-sensitive feature $\mathbf{F}^o \in \mathbb{R}^{B\times N\times 2C\times H\times W}$. With the help of ARF, we can obtain feature maps with orientation channels and model the similarity between features from different directions.

Similarity learning benefits from orientation-sensitive features, as similarity manifests in different directions. Considering similarity itself is orientation-independent, it is expected to extract rotation-invariant feature. We feed $\mathbf{F}^o$ into oriented response pooling (ORPooling) [34] to achieve rotation invariance. The ORPooling achieves rotation invariance by pooling the responses of all N response maps, so we get rotation-invariant feature map $\mathbf{F}^p \in \mathbb{R}^{B\times 2C\times H\times W}$. Then, we consider $\mathbf{F}^p \in \mathbb{R}^{B\times 2C\times H\times W}$ as a feature map with two orientation channels and perform ORPooling again to obtain the final orientation response feature $\mathbf{F}^{res} \in \mathbb{R}^{B\times C\times H\times W}$. In this way, we can align the feature of objects with different orientations. The orientation-invariant feature summarizes salient information across different orientations. Compared to the orientation-sensitive feature $\mathbf{F}^o$, it is more efficient with fewer parameters.

As illustrated in Fig. 2(b), to obtain multi-directional similarity features, we model the relationship between $\mathbf{F}^{res}$ and $\mathbf{F}$, as well as $\mathbf{F}^{res}$ and $\mathbf{F}^f$ separately to generate similarity maps $\mathbf{F}^{sim}$. Specifically, we consider $\mathbf{F}^{res}$ as $\mathbf{K}$ and another feature as $\mathbf{Q}$. After reshaping both $\mathbf{K}$ and $\mathbf{Q}$ to be the size of $\mathbb{R}^{B\times HW\times C}$, the feature similarity map of each pixel is computed as

$$\mathbf{F}^{sim} = \mathbf{Q}\mathbf{K}^\top, \tag{2}$$

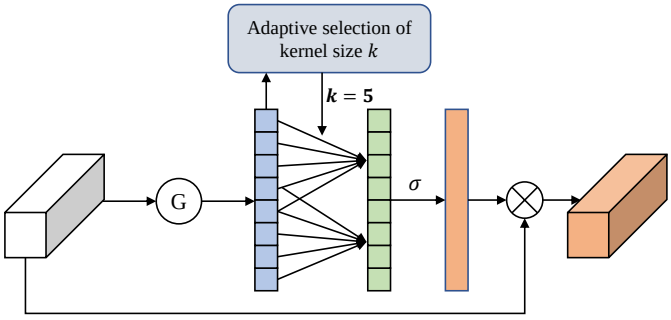

Fig. 3. The architecture of ECA module, where G represents global average pooling, σ represents the sigmoid function, and $\otimes$ represents element-wise product.

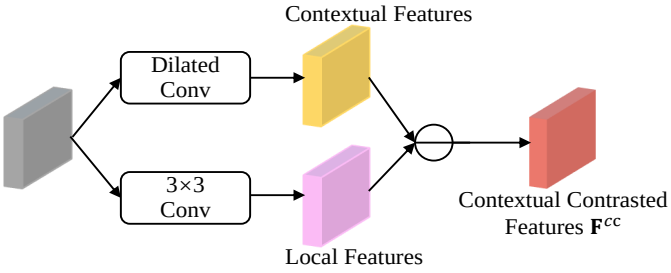

Fig. 4. The structure of CCL module, in which $\mathbf{F}^{cc}$ is obtained by subtracting local features from contextual features.

where $\mathbf{F}^{sim} \in \mathbb{R}^{B\times HW\times HW}$, and $\top$ means transpose. To obtain multi-directional similarity features, we perform matrix multiplication between $\mathbf{F}^{res}$ and $\mathbf{F}^{sim}$.

Due to the weak similarity between the contents in the mirror and the real objects, we employ efficient channel attention (ECA) [35] to enhance the similarity features. The architecture of the ECA moudle is shown in Fig. 3. The enhanced multi-directional similarity features are then obtained as the output of the module.

C. Spectral Saliency Enhancement Decoder Module

Contextual contrasted features are widely applied in mirror detection methods. The decoder used for extracting contextual contrasted features is called the contextual contrasted local (CCL) decoder, as illustrated in Fig. 4. The process of extracting contrastive semantics can be described by the following equation:

$$CCL(\mathbf{F}_i) = \sigma(BN(f_l(\mathbf{F}_i) - f_{ct}(\mathbf{F}_i))), \tag{3}$$

where f_l is the local feature extractor which contains a 3×3 convolution with a dilation rate of 1, BN, and ReLU in turn. f_{ct} is the contextual feature extractor, which consists of dilated convolutions, BN and ReLU. By subtracting the local features from the contextual features, we obtain the contrastive semantics. Due to the reflective nature of mirrors, there is often a significant contrast in color and texture between the mirrored content and its surrounding environment. In [1], this phenomenon is referred to as the content discontinuity at the mirror boundaries. Contextual contrasted features are designed to describe this relationship.

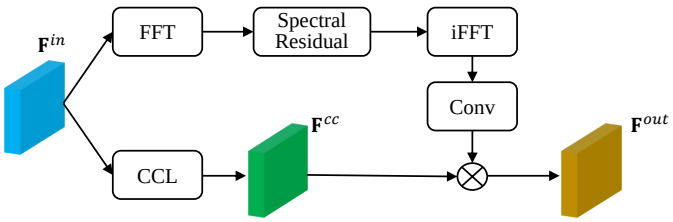

Fig. 5. The architecture of our proposed spectral enhancement contextual contrasted local (SECCL) decoder.

Contextual contrasted features can easily become ineffective when the mirrored content is similar to the surrounding environment. However, human eyes can easily perceive the discontinuity in content. This has led us to reconsider the meaning of contextual contrasted features. Learning contextual contrast is essentially capturing the differences between features at specific locations and those in the neighborhood, which often is an important visual cue that denotes saliency. Salience is highly significant for humans.

Taking the above perspectives into consideration, we propose a novel spectral enhancement contextual contrasted local (SECCL) decoder. Fig. 5 illustrates the architecture of our SECCL. It extracts saliency information from the spectral residual of features and fuse it with the contextual contrasted features to enhance the salience.

For the ensemble of natural images, the amplitude $\mathcal{A}(f)$ of the averaged Fourier spectrum obeys a distribution:

$$E\left\{\mathcal{A}(f)\right\} \propto 1/f. \tag{4}$$

It is discovered that the log spectra of different images exhibit similar trends, though each containing statistical irregularities [36]. These similar trends are considered as redundant information, while the statistical singularities are regarded as saliency information in the images.

In particular, as shown in Fig. 5, we firstly perform a fast Fourier transform (FFT) on the input feature $\mathbf{F}^{in}$ to obtain the amplitude spectrum $\mathcal{A}(f)$. The log spectra can be obtained using the formula:

$$L(f) = log(\mathcal{A}(f)). \tag{5}$$

Then, we can obtain the average spectrum:

$$AL(f) = h_q(f) \cdot L(f), \tag{6}$$

where $h_q(f)$ is a mean filter with a kernel size of q. The spectral residual is defined as

$$R(f) = L(f) - AL(f). \tag{7}$$

Next, we convert the frequency domain features into spatial domain features by using inverse fast Fourier transform (iFFT):

$$S(\mathbf{x}) = \left\|\mathfrak{F}^{-1}(\exp(R(f) + iP(f)))\right\|, \tag{8}$$

where $S(\mathbf{x})$ represents the saliency map, $\mathfrak{F}^{-1}$ represents the iFFT and $P(f)$ represents the phase spectrum. Finally, we perform a simple fusion of the saliency map with the contextual contrasted features to obtain the final output.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS INCLUDING RGB SALIENT OBJECT DETECTION BASED METHODS AND MIRROR DETECTION BASED METHODS ON MSD [1] AND PMD [4] DATASETS. THE BEST RESULTS AND THE SECOND BEST RESULTS ARE SHOWN IN RED AND BLUE, RESPECTIVELY.

Method	MSD [1]			PMD [4]		
	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
CPDNet [38]	57.58	0.743	0.115	60.04	0.733	0.041
MINet [39]	66.39	0.823	0.087	60.83	0.798	0.037
LDF [40]	72.88	0.843	0.068	63.31	0.796	0.037
VST [41]	79.09	0.867	0.052	59.06	0.769	0.035
MirrorNet [1]	78.88	0.856	0.066	58.51	0.741	0.043
PMDNet [4]	81.54	0.892	0.047	66.05	0.792	0.032
SANet [2]	79.85	0.879	0.054	66.84	0.837	0.032
VCNet [5]	80.08	0.898	0.044	64.02	0.815	0.028
SATNet [3]	85.41	0.922	0.033	69.38	0.847	0.025
CSF [7]	82.08	0.896	0.045	70.05	0.838	0.024
S^{2}MD	87.11	0.936	0.032	69.77	0.846	0.024

As shown in Fig. 2(c), our SSEDM is an extension of the SECCL. First, we use the concatenation followed by convolution operation to fuse features $\mathbf{F}$ and $\mathbf{F}^f$, obtaining feature $\mathbf{F}^c$. Denote by $\mathbf{F}_{i+1}^{out}$ the $(i+1)$-th scale output by SSEDM. We upsample $\mathbf{F}_{i+1}^{out}$ as a reference feature $\mathbf{F}^{ref}$, and add it to $\mathbf{F}$, $\mathbf{F}^f$, and $\mathbf{F}^c$ at the i-th scale of SSEDM, respectively. Then, we feed the obtained features into SECCL to learn more salient contextual contrastive semantics.

Finally, we concatenate those three SECCL outputs together to get the output features $\mathbf{F}_i^{out}$ and the corresponding prediction mask $\mathbf{P_i}$, which is given as

$$\mathbf{P}_i = f_{seg}(\mathbf{F}_i^{out}), \tag{9}$$

where f_{seg} is a segmentation head whose output has two channels. The output of the last decoder layer $\mathbf{P}_0$ is adopted as the final prediction result of our network.

As illustrated in Fig. 2(a), our framework contains four scales of SSEDMs and segmentation heads. The full loss function is calculated by considering the difference between $\mathbf{P}_i$ and the ground-truth mask $\mathbf{M}$ at four scales:

$$\mathcal{L} = \sum_{i=0}^{3} w_i \mathcal{L}_i(\mathbf{P}_i, \mathbf{M}), \tag{10}$$

where w_i is the weight of the i-th scale, and the sub-loss $\mathcal{L}_i$ at each scale is a cross-entropy loss [37].

IV. EXPERIMENTS

A. Datasets and Settings

1) Datasets: We conduct experiments on three benchmark datasets: mirror detection dataset (MSD) [1], progressive mirror dataset (PMD) [4], and RGBD-Mirror [6].

- **MSD** is the first mirror dataset, which consists of 3,677 indoor scene images and 341 outdoor scene images. The dataset is divided into 3,063 images for training and 955 images for testing.

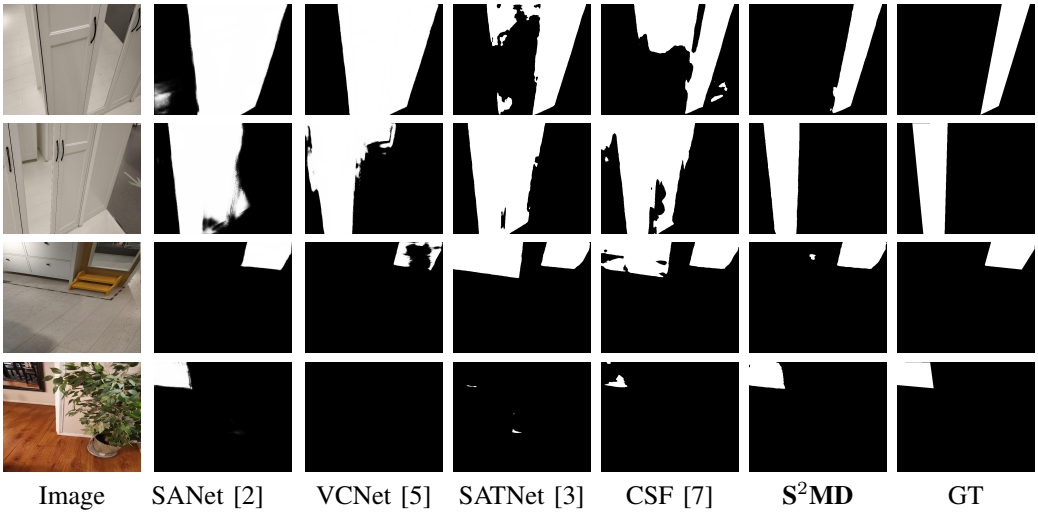

Fig. 6. Visual comparison results on example images from MSD [1] dataset, in which GT denotes ground-truth masks. The first two rows depict scenes where the mirror is similar to the surrounding environment. The last two rows are scenes containing small and inconspicuous mirrors.

- **PMD** is proposed to address the small scale and limited diversity of the MSD dataset. It comprises more diverse scenes by selecting mirror images from six public datasets including ADE20K [42], [43], NYUv2 [44], MINC [45], Pascal-Context [46], SUNRGBD [47], and COCO-stuff [48], in which 5,096 images are used for training and 571 images are used for testing.
- **RGBD-Mirror** is a RGBD-based mirror detection dataset, collected from four existing datasets, including Matterport3D [49], SUNRGBD [47], ScanNet [50], and 2D3DS [51]. It contains 3,049 RGB images and corresponding depth maps. There are 2,000 images for training and 1,049 images for testing.

2) Evaluation Metrics: Similar to previous works [3], [5], we adopt three commonly used dense prediction evaluation metrics: intersection over union (IoU), F-measure F_β, and mean absolute error (MAE) to evaluate the performance of methods.

3) Implementation Details: We implement our network via PyTorch [52], and use the small version of Swin Transformer (namely Swin-S) pretrained on ImageNet-1k [53] as the backbone of our network. The number of rotations N in ARF is set to 8, and the weights w_0, w_1, w_2, and w_3 in Eq. (10) are set to 1.25, 1.25, 1.0, and 1.5, respectively.

Following data augmentation techniques used by previous methods [3], [5], we adopt random resize and crop as well as random horizontal flip to augment training images. For testing, we simply resize input images to 512×512 to evaluate our network. Our network can be trained on a single NVIDIA GeForce RTX 3090/3090Ti GPU, with a batch size set to 4. During training, we use AdamW [54] optimizer and set β_1, β_2, and the weight decay to 0.9, 0.999, and 0.01, respectively. The learning rate is initialized to 3×10^{-5} and decayed by the *poly* strategy with the power of 1.0.

B. Comparison with State-of-the-Art Methods

In this section, we compare our approach against state-of-the-art mirror detection methods under the same evaluation setting. These methods include RGB salient object detection based methods CPDNet [38], MINet [39], and LDF [40], mirror detection based methods MirrorNet [1], PMDNet [4], PDNet [6], SANet [2], VCNet [5], SATNet [3], and CSF [7], and RGBD salient object detection based methods JL-DCF [55], DANet [56], BBSNet [57], VST [41], XMSNet [58], and PopNet [59]. Note that PDNet [6] also implements a version based on depth information. Our S^{2}MD only uses benchmark training images, and do not rely on depth information.

1) Evaluation on MSD and PMD: As shown in Table I, we compare with typical state-of-the-art methods on MSD dataset and PMD dataset, including four RGB salient object detection based methods and six mirror detection based methods. On the MSD dataset, our S^{2}MD outperforms other methods in all evaluation metrics. on the PMD dataset, S^{2}MD achieves competitive performance, with MAE result outperforming other methods. It can be seen that in both datasets, our MAE results reach the highest, indicating that S^{2}MD can accurately distinguish real-world objects and their mirror areas, while some methods are accustomed to detecting similar real areas as mirror areas. Compared to the recent powerful method SATNet using only horizontal direction perception, our S^{2}MD obtains better performance especially for the MSD dataset. This is because S^{2}MD considers similarity relationships from different directions as much as possible, and most MSD images have significant similarity relationships.

Fig. 6 and Fig. 7 show the visualization results of our S^{2}MD and several advanced mirror detection methods on the MSD and PMD datasets, respectively. Note that other methods without code or predicted mirror masks released are not compared. In Fig. 6, the first two rows depict scenes where the mirror is similar to the surrounding environment. In this case, S^{2}MD obtains almost completely correct results, while other methods incorrectly predict similar surrounding areas as mirrors, indicating that S^{2}MD still performs well in challenging scenes without obvious contextual discontinuities. Besides, from the last two rows in Fig. 6 and all the example

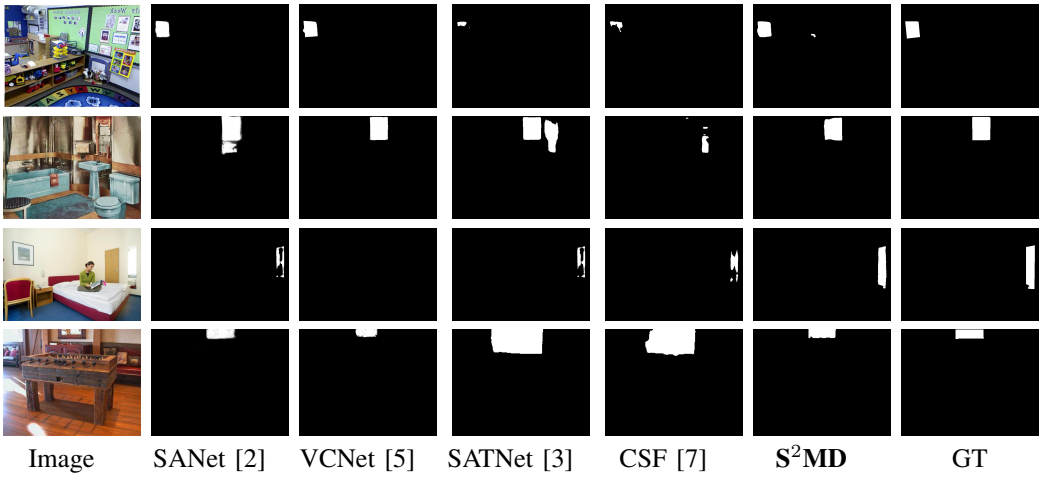

Fig. 7. Visual comparison results on example images from PMD [4] dataset. These input images contain cluttered content, which pose much interference on the detection of mirrors.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS INCLUDING RGBD SALIENT OBJECT DETECTION BASED METHODS AND MIRROR DETECTION BASED METHODS ON RGBD-MIRROR [6] DATASET. "W/ DEPTH" DENOTES THE USE OF DEPTH INFORMATION.

Method	w/ Depth	RGBD-Mirror [6]		
		$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
DANet [56]	✓	67.81	0.835	0.060
JL-DCF [55]	✓	69.65	0.844	0.056
BBSNet [57]	✓	74.33	0.868	0.046
VST [41]	✓	70.20	0.851	0.052
XMSNet [58]	✓	75.37	0.848	0.043
PopNet [59]	✓	78.10	0.883	0.043
PDNet [6]		73.57	-	0.053
PDNet [6]	✓	77.77	0.878	0.041
SANet [2]		74.99	0.873	0.048
VCNet [5]		73.01	0.849	0.052
SATNet [3]		78.42	0.906	0.031
CSF [7]		78.66	0.900	0.031
S²MD		78.60	0.904	0.030

images in Fig. 7, it can be seen that S²MD has advantages in detecting small objects in complex scenes.

2) Evaluation on RGBD-Mirror: Table II presents the results of different methods on the challenging RGBD-Mirror benchmark, in which RGB-D salient object detection methods and mirror detection methods are compared. It can be observed that our S²MD achieves the best performance on the evaluation metric MAE and achieves suboptimal performance on IoU and F_β without using any depth information. Compared to the recent frequency domain based method CSF, S²MD performs better using the proposed spectral saliency enhancement decoder module.

Fig. 8 illustrates the visual results of different methods. It is intuitive that depth maps can provide useful clues for locating mirror areas. However, it may also lead to incorrect segmentation results. For example, all the methods incorrectly predict the right-side region as a mirror for the first example image, especially for the RGBD salient object detection based methods which classify almost the entire high-lighted depth region as a mirror. Besides, in the second row, the region corresponding to the door in the image is easily misclassified as a mirror by methods like VST and VCNet. These regions exhibit significant depth variations and have similar properties to mirrors, such as strong saliency and visual content discontinuity, making them prone to being incorrectly predicted. For the fourth and fifth example images, it can be observed that many methods either incorrectly predict an extra mirror or fail to recognize an additional one. In contrast, our S²MD accurately identifies mirrors by analyzing the interplay of reflections and symmetry within the image. This consistent detection across different scenarios underscores our method's robust capability to pinpoint and interpret reflection characteristics specific to mirrors.

3) Discussion about Model Complexity: We compare our approach with typical methods in terms of the number of parameters, giga floating point operations (GFLOPs), and inference speed in Table III. We find that VCNet has the largest model complexity among these methods, which uses a smaller image input with the most parameters, the highest GFLOPs, and the smallest FPS. In contrast, our S²MD has moderate computational cost with similar GFLOPs to SATNet, and CSF has the fewest parameters and GFLOPs.

Although our S²MD and CSF both utilize frequency domain features, they are significantly different in terms of methodology. CSF mainly considers spatial frequency feature affinities, global feature learning, and cross-modality features fusion. In contrast, S²MD utilizes spectral residuals to enhance the saliency of mirror edges. Based on the experimental results in Table I, CSF achieves an IoU of 82.08, an F_β of 0.896, and an MAE of 0.045 on the MSD dataset. These results are significantly lower than those obtained by our S²MD. On the other two datasets, CSF also works worse than S²MD. These quantitative results show that our method achieves a better balance between the efficiency and the accuracy.

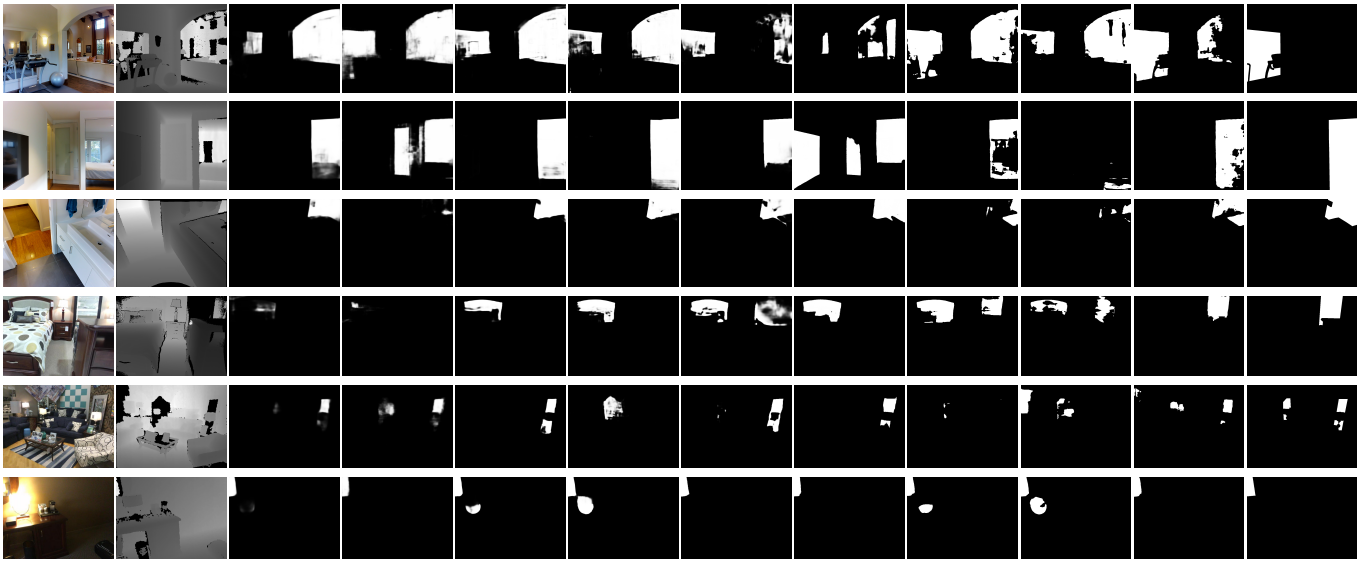

Fig. 8. Visual comparison results on challenging example images from RGBD-Mirror [6] dataset. The first input image exhibits variations in depth, in which early methods fail in this challenging case. The second input image includes symmetric objects inside and outside mirrors. The third input image places mirrors difficult to be detected due to cluttered background. The fourth input image contains both glasses and mirrors.

TABLE III
THE NUMBER OF PARAMETERS (#PARAMS.), GIGA FLOATING POINT OPERATIONS (GFLOPS), AND INFERENCE SPEED (FRAME PER SECOND, FPS) FOR TYPICAL METHODS. THE INFERENCE SPEED IS TESTED ON AN NVIDIA GEFORCE RTX 3090TI GPU.

Method	Backbone	Input Size	#Params.	GFLOPs	FPS
VCNet	ResNeXt101	384×384	333.17M	487.31	13.19
SATNet	Swin-S	512×512	139.36M	153.12	**26.92**
CSF	Swin-S	512×512	**84.75M**	**78.59**	21.91
S²MD	Swin-S	512×512	214.05M	149.23	20.16

TABLE IV
ABLATION STUDY RESULTS ON MSD [1] DATASET. BASELINE USES THE STRUCTURE OF SWIN-S DECODED BY UPERNET. DUAL-PATH DENOTES THE DUAL-PATH SWIN TRANSFORMER. MSPM DENOTES OUR MULTI-DIRECTIONAL SIMILARITY PERCEPTION MODULES AT THE THIRD SCALE. MSPMS DENOTES MSPM AT BOTH THE SECOND SCALE AND THE THIRD SCALE. SSEDM DENOTES OUR SPECTRAL SALIENCY ENHANCEMENT DECODER MODULE. VF DENOTES PERFORMING AN ADDITIONAL VERTICAL FLIP ON THE INPUT IMAGE. SECCL DENOTES OUR SPECTRAL ENHANCEMENT CONTEXTUAL CONTRASTED LOCAL DECODER.

Method	$IoU \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
Baseline	80.46	0.901	0.045
Dual-Path	79.59	0.903	0.044
Dual-Path+MSPM	84.28	0.909	0.040
Dual-Path+MSPMs	84.94	0.921	0.036
Dual-Path+SSEDM	85.26	0.923	0.034
Dual-Path+MSPM+SSEDM	86.23	0.926	0.034
S²MD w/o Vf	86.40	0.924	0.032
S²MD w/o SECCL	86.01	0.926	0.033
S²MD	**87.11**	**0.936**	**0.032**

C. Ablation Study

In this section, we investigate the effectiveness of main components in our framework. Table IV shows the results of different variants of our S²MD on the MSD dataset.

1) Dual-Path Structure: Since we need to model the similarity of local features in two relative directions, we use a dual-path structure to enhance the perception of multi-directional similar features. We take the original image and its diagonally flipped counterpart as inputs to the model to enhance the perception of similarity along diagonal directions. We implement two variants of our S²MD. One is a pure Swin Transformer decoded by UperNet [60], named as Baseline. Another is a dual-path Swin Transformer, where features are trained and supervised separately in two paths, named as Dual-Path. The results in the first two rows of Table IV show that using only the dual-path structure is insufficient for uncovering effective information beneficial for mirror localization.

2) Multi-directional Similarity Perception Module: Low-level semantic features capture the fundamental visual properties of an image, such as edges, textures, colors, and shapes, without involving higher-level semantic understanding. However, the similarity between the content in a mirror and real objects is ambiguous and requires higher-level semantic understanding. To determine which scale is more effective for applying MSPM, we conduct two experiments. In Dual-Path+MSPM, MSPM is applied only to the third scale. Compared to Dual-Path without using MSPM, improvements are obtained in all three evaluation metrics. When applying MSPM to the last two scales, further performance enhancement is achieved by Dual-Path+MSPMs. It can be concluded that applying MSPM to the last both scales yields the best results.

3) Spectral Saliency Enhancement Decoder Module: We verify whether the decoder we designed is suitable for mirror detection by replacing it. In the fifth row of Table IV, we conduct an experiment based on Dual-Path, by replacing the

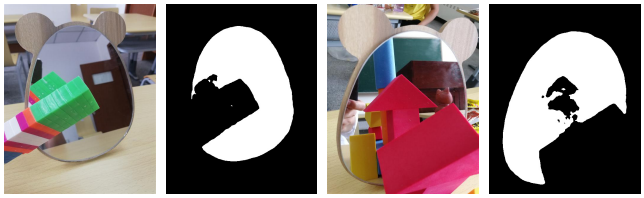

Fig. 9. Failure cases. There are overlaps between mirrored content and real objects in these images.

UperNet decoder with our SSEDM. Compared with the dual-path network using UperNet decoder, the dual-path network using our SSEDM has a gain of 5.6, 0.02, and 0.01 in IoU, F_β, and MAE, respectively. The improvement proves that our decoder module can properly fuse features in the two paths, and is more suitable for the mirror detection task. The improvement in evaluation metrics demonstrates the effectiveness of our SSEDM.

4) Combination of MSPM and SSEDM: To explore the optimal way of combining MSPM and SSEDM, we implement Dual-Path+MSPM+SSEDM and Dual-Path+MSPMs+SSEDM, and find the latter obtains the best performance. Therefore, applying MSPM to the last two scales is still the most reasonable choice when combining with SSEDM. In this case, Dual-Path+MSPMs+SSEDM is our final model S^{2}MD.

5) Vertical Flip: The effectiveness of horizontal flip has been validated by SATNet [3]. To further validate whether using the diagonally flipped image as input is beneficial for perceiving similar features along diagonal directions, we remove the vertical flip and instead input the original image and its horizontally flipped counterpart into the network. The margin between S^{2}MD w/o Vf and S^{2}MD demonstrates the effectiveness of vertical flip.

6) Spectral Enhancement Contextual Contrasted Local Decoder: The comparison of experimental results between S^{2}MD w/o SECCL and S^{2}MD in Table IV demonstrate the effectiveness of SECCL. This is due to our designed SECCL structure with frequency domain saliency information.

D. Limitations

Although our method achieves significant improvements compared to previous works on the MSD dataset, there are a few failure cases on more challenging PMD and RGBD-Mirror datasets, as illustrated in Fig. 9. This demonstrates that our method may fail when the content in the mirror partially overlaps with real objects. Since our method relies on similarity perception, it may struggle to differentiate between highly similar mirrored content and real objects. Besides, as our method leverages the reflective properties of mirrors to identify mirror regions by learning the similarity between mirror contents and real targets, it may excessively rely on this characteristic. For instance, the MSD dataset primarily focuses on indoor scenes, in which mirror regions are quite prevalent, and similarity relationships are abundant. Consequently, our method performs better on this dataset. However, in other complex environments, such similarity relationships may not necessarily exist, and the sizes of mirror regions can vary significantly. Therefore, our method fails to demonstrate significant advancements over previous works in these scenarios.

V. Conclusion

In this paper, we have proposed an innovative module that is sensitive to multiple directions of feature similarity, designed to map the intricate interplay between a mirror's inner and outer elements. This allows for an adaptive response to the varying conditions mirrors encounter across diverse settings, leading to enhanced detection capabilities. Additionally, we have proposed a new spectral saliency enhancement decoder module. This decoder is engineered to amplify the distinctiveness of spatial contextual contrasted features, thereby sharpening the focus on salient aspects within images.

We have compared our proposed framework with state-of-the-art methods on three challenging benchmarks. Both quantitative and visual results indicate that our framework outperforms the previous works. Besides, we have conducted ablation studies which demonstrate that main components in our framework all contribute to mirror detection.

Considering the limitations of our method, in future work we will explore the learning of more discriminative relational features. A potential solution is to separate the features of mirror reflection regions from background features, and then extract the relational features using more powerful framework so as to minimize the influence of regions with similar characteristics. Besides, we will explore describing the reflective properties of mirrors in a more explicit manner rather than directly relying on symmetric similarities to reduce the impact of symmetrical objects on detection results.

References

[1] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. W. Lau, "Where is my mirror?" in *IEEE International Conference on Computer Vision*. IEEE, 2019, pp. 8809–8818.

[2] H. Guan, J. Lin, and R. W. Lau, "Learning semantic associations for mirror detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 5941–5950.

[3] T. Huang, B. Dong, J. Lin, X. Liu, R. W. Lau, and W. Zuo, "Symmetry-aware transformer-based mirror detection," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2023, pp. 935–943.

[4] J. Lin, G. Wang, and R. W. Lau, "Progressive mirror detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 3697–3705.

[5] X. Tan, J. Lin, K. Xu, P. Chen, L. Ma, and R. W. Lau, "Mirror detection with the visual chirality cue," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3492–3504, 2023.

[6] H. Mei, B. Dong, W. Dong, P. Peers, X. Yang, Q. Zhang, and X. Wei, "Depth-aware mirror segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 3044–3053.

[7] Z. Xie, S. Wang, Q. Yu, X. Tan, and Y. Xie, "Csfwinformer: Cross-space-frequency window transformer for mirror detection," *IEEE Transactions on Image Processing*, vol. 33, pp. 1853–1867, 2024.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[9] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *International Journal of Computer Vision*, vol. 115, pp. 330–344, 2015.

[10] J. Kim and V. Pavlovic, "A shape preserving approach for salient object detection using convolutional neural networks," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 609–614.

[11] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 660–668.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3431–3440.

[13] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.

[14] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "Edn: Salient object detection via extremely-downsampled network," *IEEE Transactions on Image Processing*, vol. 31, pp. 3125–3136, 2022.

[15] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8006–8021, 2021.

[16] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 12 321–12 328.

[17] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Ding, Y. Xie, and Z. Li, "Position-aware relation learning for rgb-thermal salient object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2593–2607, 2023.

[18] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2023, pp. 10 031–10 040.

[19] Z. Wang, Y. Zhang, Y. Liu, D. Zhu, S. A. Coleman, and D. Kerr, "Elwnet: An extremely lightweight approach for real-time salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6404–6417, 2023.

[20] Z. Zhu, Z. Zhang, Z. Lin, X. Sun, and M.-M. Cheng, "Co-salient object detection with co-representation purification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8193–8205, 2023.

[21] Y. Ge, Q. Zhang, T.-Z. Xiang, C. Zhang, and H. Bi, "Tcnet: Co-salient object detection via parallel interaction of transformers and cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2600–2615, 2023.

[22] L. Tang, B. Li, S. Kuang, M. Song, and S. Ding, "Re-thinking the relations in co-saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5453–5466, 2022.

[23] P. Zheng, J. Qin, S. Wang, T.-Z. Xiang, and H. Xiong, "Memory-aided contrastive consensus learning for co-salient object detection," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2023, pp. 3687–3695.

[24] Z. Shao, Y. Su, Y. Zhou, F. Meng, H. Zhu, B. Liu, and R. Yao, "Ct-net: arbitrary-shaped text detection via contour transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1815–1826, 2024.

[25] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4339–4354, 2021.

[26] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, pp. 14 254–14 265, 2020.

[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *IEEE Conference on Computer Vision and Pattern Recognition workshops*. IEEE, 2018, pp. 224–236.

[28] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in Neural Information Processing Systems*, pp. 12 405–12 415, 2019.

[29] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *IEEE International Conference on Computer Vision*. IEEE, 2023, pp. 17 627–17 638.

[30] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Extremely dense point correspondences using a learned feature descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 4847–4856.

[31] Y. Di, Y. Liao, H. Zhou, K. Zhu, Y. Zhang, Q. Duan, J. Liu, and M. Lu, "Femip: detector-free feature matching for multimodal images with policy gradient," *Applied Intelligence*, vol. 53, no. 20, pp. 24 068–24 088, 2023.

[32] D. Li and S. Du, "Contextmatcher: Detector-free feature matching with cross-modality context," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[33] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 8922–8931.

[34] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 519–528.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 11 531–11 539.

[36] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[37] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[38] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 3907–3916.

[39] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 9413–9422.

[40] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 13 022–13 031.

[41] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *IEEE International Conference on Computer Vision*. IEEE, 2021, pp. 4722–4732.

[42] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 633–641.

[43] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[44] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[45] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3479–3487.

[46] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 891–898.

[47] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 567–576.

[48] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 1209–1218.

[49] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[50] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5828–5839.

[51] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[55] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 3049–3059.

[56] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-d salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 646–662.
[57] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision*. Springer, 2020, pp. 275–292.
[58] Z. Wu, J. Wang, Z. Zhou, Z. An, Q. Jiang, C. Demonceaux, G. Sun, and R. Timofte, "Object segmentation by mining cross-modal semantics," in *ACM International Conference on Multimedia*. ACM, 2023, pp. 3455–3464.
[59] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, "Source-free depth for object pop-out," in *IEEE International Conference on Computer Vision*. IEEE, 2023, pp. 1032–1042.
[60] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *European Conference on Computer Vision*. Springer, 2018, pp. 418–434.

Zhiwen Shao is currently an Associate Professor with the China University of Mining and Technology, as well as a Postdoctoral Fellow with the Shanghai Jiao Tong University and the Hong Kong University of Science and Technology. He received the B.Eng. degree and the Ph.D. degree in Computer Science and Technology from the Northwestern Polytechnical University and the Shanghai Jiao Tong University in 2015 and 2020, respectively. He has published more than 60 academic papers in popular journals and conferences. His research interests lie in computer vision and affective computing. He has served as an Area Chair for ACM MM 2024, an Associate Editor for TVC, and a Publication Chair for CGI 2023.

Rui Chen is currently a master student at the School of Computer Science and Technology, China University of Mining and Technology, China. He received the B.Eng. degree from the China University of Mining and Technology in 2022. His researches focus mainly on image classification, object detection, and image segmentation.

Xuehuai Shi received the Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. He is currently a Lecturer with the School of Computer Science, Nanjing University of Posts and Telecommunications. His research interests include virtual reality, real-time rendering, and augmented reality.

Bing Liu received the B.S., M.S., and Ph.D. degrees in 2002, 2005, and 2013, respectively, from the China University of Mining and Technology, Xuzhou, China. He is currently an Associate Professor at the School of Computer Science and Technology, China University of Mining and Technology, China. His current research interests include natural language processing, image understanding, and deep learning.

Canlin Li received the B.S. degree from National University of Defense Technology, China in 1998, and the M.S. degree from Zhejiang University, China in 2004, and the Ph.D. degree from Shanghai Jiao Tong University, China in 2010. He is currently an Associate Professor with the School of Computer Science and Technology, Zhengzhou University of Light Industry, China. His research interests include image processing, pattern recognition, artificial intelligence, and visual media computing.

Lizhuang Ma is currently a Distinguished Professor, Ph.D. Tutor, and the Head of the Digital Media and Computer Vision Laboratory at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He is the recipient of the National Science Fund for Distinguished Young Scholars. He received the B.S. and Ph.D. degrees from the Zhejiang University, China in 1985 and 1991, respectively. He has published more than 200 academic research papers in both domestic and international journals. His research interests include computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM.

Dit-Yan Yeung received his B.Eng. degree in Electrical Engineering and MPhil degree in Computer Science from the University of Hong Kong, and Ph.D. degree in Computer Science from the University of Southern California. He started his academic career as an Assistant Professor at the Illinois Institute of Technology in Chicago. He then joined the Hong Kong University of Science and Technology where he is now a Chair Professor at the Department of Computer science and Engineering. His research interests are primarily in computational and statistical approaches to machine learning and artificial intelligence. He is also interested in developing novel machine learning models for various applications particularly in computer vision, education, and recommender systems.