Visual Prompt-guided Reasoning for Generalizable Mirror Detection

Mingfeng Zha, Guoqing Wang, *Member, IEEE,* Yunqiang Pei, Tianyu Li, Xiongxin Tang
Jiayi Ma, *Senior Member, IEEE,* Yang Yang, *Senior Member, IEEE,* and Heng Tao Shen, *Fellow, IEEE*

Abstract—**Mirror detection (MD) aims to overcome interference caused by reflections and locate mirror regions. Existing methods focus on designing components to explicitly establish the associations between physical entities and corresponding imagings, or utilizing rotation to construct symmetric consistency. We observe that: a) incomplete and incorrect correspondence between entities and imagings; b) other physical materials (*e.g.,* glass) exhibit characteristics partially similar to mirrors, causing confusion when they co-occur; c) complex interfering factors (*e.g.,* occlusion) and reflection mechanisms may expand vector space several times over. To address these issues in a unified manner, we formulate the scene-aware visual reasoning network (SVRNet) based on visual prompts. Specifically, we construct the prototype-guided prompt chain reasoning (PPCR) that generates a mixed chain of thought reasoning based on maximal difference heterogeneous prototypes to construct comprehensive spatial location and semantic perception. Noise may accumulate gradually through the chain, and crucial clues may also disappear. Therefore, we design the prompt evolution (PE) to filter out noise and enhance the coupling between prompts. We further develop the mixture of prompt injection expert (MPIE) to dynamically select the optimal injection strategy in the low-rank space based on specific scene. Due to reflection interference and random parameter space introducing potential ambiguity, we formulate the three-way evidence-aware (TEA) loss to quantify the uncertainty, thereby providing reliable predictions. To leverage historical knowledge and further disentangle representations, we propose the frequency prototype contrastive (FPC) loss for learning more generalizable features across images. Finally, we relabel 25,828 images and formulate the first point-supervised MD framework. Extensive experiments conducted on four mirror benchmarks under three settings demonstrate that our method surpasses state-of-the-art approaches. Promising results are also achieved on six related benchmarks, showing its generality.**

Index Terms—**Mirror detection, prompt learning, visual reasoning, mixture of experts, uncertainty quantification.**

This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091, the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

Mingfeng Zha, Guoqing Wang, Yunqiang Pei, Tianyu Li and Yang Yang are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. (Email: zhamf1116@gmail.com)

Xiongxin Tang is with the Institute of Software, Chinese Academy of Science, Beijing 100190, China

Jiayi Ma is with the School of Electronic Information, Wuhan University, Wuhan, 430072, China.

Heng Tao Shen is with the School of Computer Science and Technology, Tongji University, Shanghai 201804, China, with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and with the Peng Cheng Laboratory, Shenzhen 518066, China.

Corresponding author: Guoqing Wang. Email: gqwang0420@uestc.edu.cn.

Project page: https://winter-flow.github.io/project/SVRNet

Manuscript received December 19, 2024; revised December 19, 2024.

I. Introduction

Unlike common segmentation tasks such as semantic segmentation (SS) that aim to segment physical entities, mirror detection (MD) focuses on distinguishing imagings and physical entities, ultimately identifying the reflection area, *i.e.,* mirror region. Misidentifying imagings as physical objects may significantly impact scene understanding [1], 3D reconstruction [2], visual-language navigation [3], and other related tasks. Mirrors can be regarded as confounding factors in the causal reasoning chains, while MD can remove to accurately model the spatial and logical relationships between objects.

As shown in Figure 1, existing methods [4]–[8] primarily adopt an encoding-decoding framework with several components to explicitly establish the correlations and differences between mirror and non-mirror regions, such as depth and frequency (texture). Some methods [9], [10] utilize rotation strategy to construct consistency perception between imagings and entities from image or feature map level. Similar works [11], [12] employ large-scale visual foundation models with several adapters for fine-tuning. The foundation models are trained on large datasets consisting of images with entities, which may introduce biases to imagings. Only incorporating adapters is insufficient to address reflection interference. Furthermore, the above methods have limited applicability. Due to shooting angles, some scenes only exist partial correspondences, or even only have imagings. Smooth walls, tabletops, or glass surfaces can also produce reflections and form imagings, although the optical imaging characteristics differ from mirrors, leading to confusion. MD task also faces similar challenges as the SS task, *e.g.,* occlusion and scale variations. The most straightforward approach is to design additional modules, but introduces extra parameters and computational overhead. Therefore, it is necessary to design an efficient and general method that is effective for various scenarios. We utilize prompts learned from latent (random) space to establish spatial semantic modeling and combine with uncertainty-aware mixture-of-experts (MoE) to inject prompts, thereby formulating the SVRNet. Hence, we wonder three questions. *1) Why choose to learn prompts implicitly rather than explicitly? 2) Why introduce uncertainty estimation? 3) Why inject prompts based on MoE?*

We answer the first question. First, explicit prompts, *e.g.*, image frequency and depth map, are not easily obtained and often require additional tools like depth cameras and task-specific pre-trained models. Conversely, implicit prompt learning starts from the given data, capturing features from

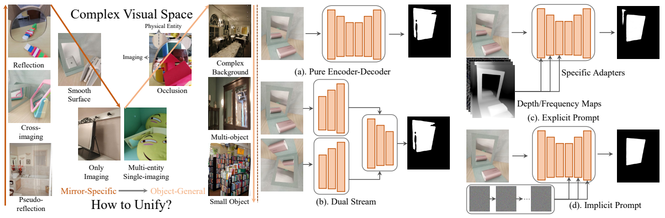

Fig. 1. **Motivation and Framework Comparison.** Our method aims to dynamically and uniformly address the complex visual space of MD across different scenarios, avoiding the stacking of singular functional components. (a) The naive encoding-decoding represented by HetNet [9]; (b) The dual-stream represented by SATNet [10]; (c) The explicit visual prompt represented by Spider [13] (not specific for MD task); (d) Our SVRNet.

the random parameter space. *Second*, explicit prompts struggle to adapt effectively to content variations, focusing more on low-level structures, while implicit prompts construct semantic awareness, thus being more suitable for exploring high-dimensional complex visual spaces. *Third*, implicit strategy can establish associations between arbitrary structures of heterogeneous or homogeneous nodes, such as chains and trees, unlike the former's fixed clues.

We answer the second question. First, learning from latent space and lack of sufficient priors, the optimization process in parameter tuning may carry inherent noise, leading to inadequate modeling of fine-grained regions. While it is possible to design purification components, they can only filter out more obvious noise. *Second*, due to reflective interference, prompts may introduce ambiguity, resulting in mismatches or even misguidance during fusion. Therefore, statistical analysis, *i.e.*, uncertainty estimation, of expert injection is required to constrain. We aim for the model to pay more attention to high uncertainty regions of interest.

We answer the third question. First, based on MoE, we can scale the network parameters arbitrarily within the same computational budget, thus expanding the border of representation space for various tasks. *Second*, The complexity of the prompts varies according to the characteristics of the input. The simpler the visual spatial relationships, the fewer the prompts required, and the more straightforward the interaction mode. Fixed processing units struggle to effectively adapt to such variations; instead, this work can be delegated to some experts, who collaborate to optimize and determine the best solution. *Third*, unlike previous works defining models at the dataset or task level, we partition each image into multiple subspaces, dynamically selecting the most suitable one or more experts to explore the most significant representations and correlations in different pixels or regions. This enables us to customize image-level segmentation model, facilitating mapping from reflection-confused to reflection-aware space.

Technically, inspired by thought chain and causal reasoning of large language models (LLMs), we propose the prompt chain reasoning (PPCR) to construct a visual prompt chain that establishes semantic associations and achieves preliminary localization through hierarchical inference. Unlike explicit prompts that leverage prior knowledge of images, we randomly initialize and learn in the high-dimensional space, which can be easily disturbed by initial noise and reflections, resulting in a convoluted and slow learning path towards the optimal point. Moreover, noise may propagate along the prompt chain, potentially weakening crucial clues. Therefore, we introduce the prompt evolution (PE) to filter out noise and enhance the coupling between prompts. We then propose the mixture of prompt injection expert (MPIE) to dynamically select the optimal injection strategy in the low-rank space based on specific scene. We decouple prompts into high- and low-frequency and extend the interaction modes from spatial, channel, and frequency domains to further enhance the heterogeneity of experts. To improve discriminative perception, we crossly provide semantic and position priors to the experts, *i.e.,* coarse-grained prediction maps and uncertainty maps. Due to reflection interference causing potential ambiguity in prompts and features, we propose the three-way evidence-aware (TEA) loss to quantify the uncertainty, thereby providing reliable predictions. To leverage historical knowledge and further disentangle representations, we divide the prototypes into four classes, *i.e.,* mirror high and low-frequency prototypes and non-mirror high and low-frequency prototypes, thus formulating the frequency prototype contrastive (FPC) loss. By bringing similar prototypes closer and pushing dissimilar prototypes apart, we aim to learn more general representations across images. In addition, we expand the existing scribble-based MD datasets and construct the first point annotation dataset and framework.

In summary, our main contributions are as follows:

- Incorporating implicit learning of visual prompts, we propose the SVRNet to establish semantic reasoning for complex spatial perception. To the best of our knowledge, we are the first to model MD task from the perspective of prompt learning.
- We propose the PPCR to form the mixed chain of visual thought reasoning, the PE to reduce noise interference and couple prompts, the MPIE to customize experts for selecting appropriate injection strategies, the TEA loss to quantify fusion uncertainty, and the FPC loss to improve consistency and robustness.
- We relabel 7,835 images using scribbles and construct the first weakly supervised MD dataset based on point annotations, which contains 17,993 images.
- We formulate the first point-supervised MD network that efficiently detects mirror regions. By simply changing the loss function, our method can seamlessly switch between full supervision and scribble, point supervision, to some extent achieving a unified framework.
- Extensive experiments on four benchmarks and three task settings demonstrate that our method surpasses state-of-the-art approaches. Our approach also shows promising results on six related benchmarks.

II. Related Work

A. Mirror Detection

Mirrors reflect physical entities and create completely identical imagings, which can seriously confuse and impact the understanding and modeling of visual space. For the fully-supervised setting, Yang *et al.* [7] proposed the first MD method, called MirrorNet, which explores the correlation between internal and external features of mirrors. Lin *et al.* [6] introduced the PMDNet, which compares mirror features with context for correspondence and incorporates edge information. Guan *et al.* [4] constructed semantic associations among objects based on graph representation. The motivations of these work are similar, *i.e.,* establishing the associations between entities and imagings. Huang *et al.* [10] built a dual-stream network based on Transformer to explore the symmetry property of mirrors. He *et al.* [9] presented the HetNet, which explores low-level and high-level features in heterogeneous manner. Both aim to utilize rotation strategies to construct mirror symmetry consistency. Some works [5], [8], [14], [15] leveraged the differences between mirror and non-mirror regions to distinguish, *i.e.,* depth, content distribution and frequency. For the data-efficient setting, Zha *et al.* [16] constructed the first weakly supervised dataset and model based on scribble annotations, *i.e.,* S-Mirror and SMD, and achieved performance comparable to fully supervised methods. Lin et al. [17] designed a self-supervised pre-training strategy. For the video-level setting, Lin *et al.* [18] proposed the first video-level dataset and model, *i,e,* VMD. Warren *et al.* [19] improved it based on inconsistent motion clues. Xu *et al.* [20] focused on extremely-weak supervision.

Unlike the above works, our approach can effectively adapt to more benchmarks and tasks settings, alleviating various issues without stacking complex components or branches.

B. Visual Prompt Learning

With the development of LLMs [21], prompt learning has been widely applied in AIGC [22], image enhancement [23], and video understanding [24]. Liu *et al.* [11] introduced visual prompts, generated from the high-frequency components of inputs, into low-level structure segmentation tasks. However, the explicit prompt construction based on images may introduce additional noises and lack semantic representations, making it difficult to establish associations between entities and imagings or capture mirror regions contexts. Similarly, Luo *et al.* [25] constructed 2D visual prompts based on mixed datasets and applied to the COD task. But this significantly increases the training cost and lacks generalizability. We have three distinct differences. 1) We construct an efficient visual prompt chain based on implicit learning to perceive spatial semantics, achieving a many-to-one relationship where multiple semantic prompts perceive an image, as opposed to EVP where a low-level prompt corresponds to one image. 2) We decouple prompts into high- and low- frequency part to provide more fine-grained guidance for features. Similarly, we also decouple features based on the prediction maps to induce perceptual differences in prompts. 3) Unlike directly acquiring frequency from images, which makes it difficult to perceive content dynamically, we leverage feature maps and prompts to interact in mixed domains.

III. Proposed Method

A. Motivation and Overview

We aim to enable implicit prompt learning to perceive complex spaces, dynamically equip experts based on contextual variations, and achieve model customization at the image level. The overview of our SVRNet is shown in Figure 2, which follows an encoder-decoder architecture and incorporates visual prompts into the decoder. Given any input $\mathbf{I} \in \mathbb{R}^{3\times H^0\times W^0}$, we generate multi-scale features $\mathbf{X}^k \in \mathbb{R}^{C^k\times H^k\times W^k}$ and progressively aggregate adjacent features to obtain $\mathbf{A}^k$, where C^k, H^k and W^k denote k-th stage channels, height and width, respectively. In the decoding stage, we generate content weights based on $\mathbf{A}^k$. We utilize the PPCR to generate foreground and background prompt pools and then formulate the mixed prompt chain $\mathbf{p}$. The PE is designed for filtering out harmful information and purifying prompts. The MPIE is employed to dynamically inject prompts into decoder features, *i.e.*, $\mathbf{F}^k$. The FPC and TEA losses are used to constrain the model to learn more effectively. Through iterative refinement, we can obtain high-quality outputs. We offer three settings to meet diverse task needs.

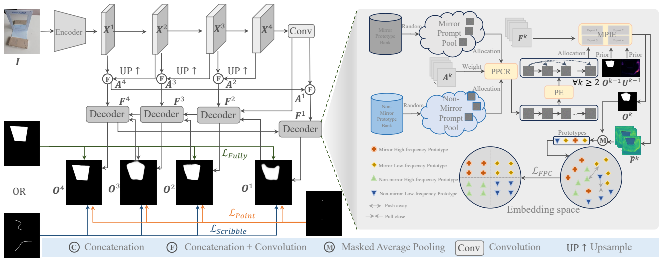

Fig. 2. The overview of our SVRNet. We utilize the image encoder to obtain multi-scale features and progressively aggregate. We employ the PPCR to construct two hierarchical visual prompt pools for spatial semantic correlation. Furthermore, we utilize the PE to avoid noises propagation and crucial clues decay. We then utilize the MPIE to disentangle prompts and interact with features in the low rank space via heterogeneous experts. Finally, we quantify uncertainty using the TEA loss to provide reliable results and the FPC loss to ensure compact homogeneous representations. Considering setting differences, we employ different ground truth (GT) as supervision.

B. Prototype-guided Prompt Chain Reasoning

Imagings and surrounding entities share similar attributes, causing four challenges: 1) Establishing connections between imagings and corresponding entities; 2) Discriminating between imagings and non-corresponding entities; 3) Failure of association criteria when there are only partial or no corresponding entities; 4) Confusion caused by similar textures and reflective surfaces. Essentially, the goal is to construct effective semantic perception. LLMs utilize chain-of-thought (CoT) prompting to build logical connections and complex reasoning between texts. Inspired by this, we propose the PPCR, forming a visual prompt chain that enables step-by-step reasoning for scene perception and establishes semantic connections between objects, thereby addressing the above issues in a unified manner. Moreover, we utilize critical representations of feature maps to generate pixel weights for prompts, enabling dynamic perception and providing priors to accelerate optimization. Random initialization can lead to optimization difficulties, so we use historical features prototypes as priors. The details of the PPCR are shown in Figure 3.

We apply masked average pooling separately on feature $\hat{\mathbf{F}}$ (output of the MPIE), foreground prediction map $\mathbf{O}$, and back-

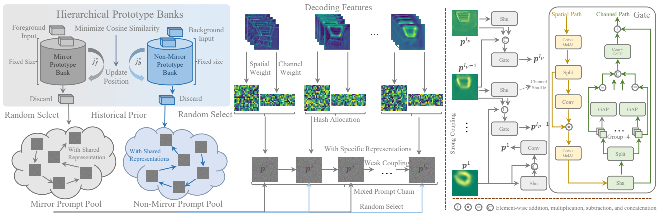

Fig. 3. Structure of the PPCR (Left) and PE (Right). We construct mirror and non-mirror prototype bank based on prototypes with maximal heterogeneity to provide historical priors for prompts and then build prompt pools. We further randomly select two types of prompts to compose the mixed prompt chain (weak connection). We utilize feature weights as current priors and employ the PE to enhance the coupling between prompts (strong connection).

ground prediction map $1 - \mathbf{O}$ to obtain foreground prototypes $\mathbf{v}_f$ and background prototypes $\mathbf{v}_b$:

$$\mathbf{v}_f = \frac{\sum_{h=1,w=1}^{H,W} \mathbf{O}^{hw} \cdot \hat{\mathbf{F}}^{hw}}{\sum_{h=1,w=1}^{H,W} \mathbf{O}^{hw}}, \mathbf{v}_b = \frac{\sum_{h=1,w=1}^{H,W} (1 - \mathbf{O}^{ij}) \cdot \hat{\mathbf{F}}^{hw}}{\sum_{h=1,w=1}^{H,W} 1 - \mathbf{O}^{hw}} \tag{1}$$

where i, j represent the current pixel position indices and $(\cdot)$ denotes element-wise multiplication. We then transfer them into the foreground and background prototype banks respectively, denoted as $\mathbf{M}_f, \mathbf{M}_b \in \mathbb{R}^{N_v \times D}$, each containing N_v prototypes $v_f, v_b \in \mathbb{R}^D$. In theory, the larger the bank, the more references for representation, and the more the model can benefit. Considering: 1) the balance between storage space and training/inference cost; 2) representation redundancy, we update the banks based on cosine similarity metrics to ensure the size remains constant and the heterogeneity of elements. Therefore, for $\mathbf{M}_f$, the objective is to find the index j_f^* such that removing $\mathbf{v}_f^j$ from the memory bank and adding the new element $\mathbf{v}_u$ minimizes the average cosine similarity among all pairs of elements in the updated bank $\mathbf{M}_f^j = (\mathbf{M}_f \setminus \{\mathbf{v}_f^j\}) \cup \{\mathbf{v}_u\}$. We obtain j_f^* by:

$$j_f^* = \underset{j \in \{1,2,\ldots,N_v\}}{\arg\min} \left(\frac{1}{N_v(N_v - 1)} \sum_{1 \le i < k \le N_v} \frac{\mathbf{v}_f^i \mathbf{v}_f^k}{\|\mathbf{v}_f^i\| \|\mathbf{v}_f^k\|} \bigg|_{\mathbf{M}_f^j} \right) \tag{2}$$

where $\|\cdot\|$ represents Euclidean norm. Similarly, we can obtain the updated position j_b^* of $\mathbf{M}_b$. Note that when the prototype bank is not full, elements can be added directly without using the update mechanism.

We then initialize the prompt parameters $\mathbf{p}^0$ based on standard Gaussian distribution and generate prompt pool by progressive convolution, sharing the parameter space. Utilizing historical knowledge to provide priors can accelerate learning. The most direct approach to integrating hierarchical prototypes into prompts $\mathbf{p}^i$ is by aggregating all elements. However, foreground and background prototypes exhibit distinctiveness, incorporating into the same prompt may cause conflicts. We first offer a suboptimal solution, *i.e.,* randomly selecting and introducing an indicator function δ where position i is 0 or 1 to ensure the prototypes are mutually exclusive.

$$\mathbf{p}^i := \mathbf{p}^i + \delta^i \mathbf{v}_f + (1 - \delta^i)\mathbf{v}_b, \mathbf{p}^{i+1} = \mathcal{F}(\mathbf{p}^i), \forall i \ge 1 \tag{3}$$

where $\mathcal{F}(\cdot)$ denotes convolution. Heterogeneity in prompts can be achieved within the current or few batches. With the updating of input data, the lack of memory in the indicator function may lead to forgetting previous allocation results, thereby resulting in heterogeneous fusion. To address this issue, we instead establish the hierarchical prompt pools corresponding to the prototype banks. We further generate the pixel weights of prompts to dynamically perceive content by:

$$\mathbf{W}_c^k = \mathsf{Softmax}(\mathsf{Proj}(\mathsf{GAP}(\mathbf{A}^k))), \mathbf{W}_s^k = \mathcal{F}(\sigma(\mathbf{A}^k)) \tag{4}$$

where $\mathsf{GAP}, \mathsf{Proj}$ denote global average pooling and linear mapping, respectively. σ is sigmoid function. $\mathbf{W}_s^k, \mathbf{W}_c^k$ provide specific content guidance. When the number of prompts N_p exceeds the length of decoding stages L_f, one-to-one correspondence cannot be achieved. Therefore, we employ the hash allocation strategy:

$$\mathbf{p}^{I_p} := \mathcal{F}(\mathbf{W}_s^{\mathsf{Allocation}(I_p, L_f)} \cdot \mathbf{p}^{I_p} \cdot \mathbf{W}_c^{\mathsf{Allocation}(I_p, L_f)}),$$
$$\mathsf{Allocation}(I_p, L_f) = \begin{cases} L_f & \text{if } I_p \bmod L_f = 0 \\ I_p \bmod L_f & \text{otherwise} \end{cases} \tag{5}$$

where I_p denote the index of prompts. We then randomly select several prompts from the mirror and non-mirror prompt pools to form the mixed prompt chain.

C. *Prompt Evolution*

We still encounter two problems with the prompt chain generated by the PPCR. 1) Noise (error) propagation, where noise is introduced at each step and transmitted to the next step; 2) Critical features may decay or be overshadowed by noise. Therefore, we propose the PE, which employs spatial and channel gate mechanisms to filter out harmful information and integrates all previous prompts to update the current prompt, enhancing the correlation between prompts and avoiding feature forgetting. The details of the PE are shown in Figure 3.

Semantic is typically encoded in channels. When $I_p = 1$, without any preceding prompts, we only leverage channel shuffle on $\mathbf{p}^1$ to achieve symmetry consistency implicitly, which is different from explicit perception of SATNet through rotating random angles. We can obtain refined $\mathbf{p}^1$ by:

$$\mathbf{p}^1 := \mathcal{F}(\mathsf{Cat}(\mathsf{Shuffle}(\mathbf{p}^1), \mathbf{p}^1)) \tag{6}$$

where $\mathsf{Cat}(\cdot;\cdot)$ denotes channel concatenation.

When $I_p > 1$, we fuse previous $I_p - 1$ prompts. We have:

$$\mathbf{p}^{I_p} := \mathcal{F}(\mathsf{Cat}(\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^{I_p}, \mathsf{Shuffle}(\mathbf{p}^{I_p}))) \tag{7}$$

For $\mathbf{p}^{I_p}$, we filter out the noise in the spatial dimension by:

$$\mathbf{p}_{p1}^{I_p}, \mathbf{p}_{p2}^{I_p} = \mathsf{Split}(\mathcal{F}(\mathbf{p}^{I_p})), \ \mathbf{p}^{I_p} := \mathbf{p}_{p2}^{I_p} \cdot \mathcal{F}(\mathbf{p}_{p1}^{I_p}) \tag{8}$$

For channel gate, we apply shuffle and split into four groups with different semantics. We then use the difference between each original group features and the mean-filtered to obtain the high-frequency components, aiming to filter out channel-wise frequency domain noise. The process is:

$$\hat{\mathbf{p}}_g^{I_p} = \mathsf{Split}(\mathsf{Shuffle}(\mathbf{p}^{I_p}))_g - \mathsf{GAP}(\mathsf{Split}(\mathsf{Shuffle}(\mathbf{p}^{I_p}))_g) \tag{9}$$

$$\mathbf{p}^{I_p} := \mathcal{F}(\mathsf{Cat}(\hat{\mathbf{p}}_1^{I_p}, ..., \hat{\mathbf{p}}_g^{I_p})_{g=1}^4) + \mathbf{p}^{I_p} \tag{10}$$

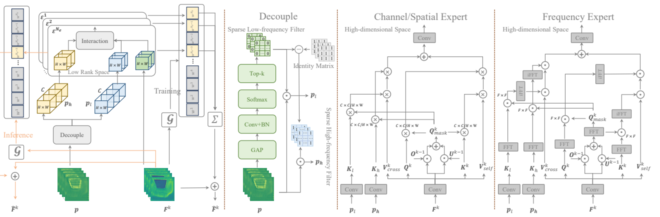

Fig. 4. Structure of the MPIE. We decouple the prompts into high-pass and low-pass parts and leverage MoE strategy to dynamically guide selection. By equipping experts with varying ranks and interaction spaces, we expand the selection possibilities.

D. Mixture of Prompt Injection Expert

Different scenarios determine varying visual reasoning complexities and pathways, thus the injection strategy should be tailored to the scenario. Therefore, we delegate this task to experts to choose the appropriate approach. Traditional MoE maintains feature dimensions unchanged for routing and linear combination. Although this ensures information integrity, the computational overhead contradicts our objectives. We transition to the most informative low-rank space for expert selection, achieving the balance between performance and efficiency. By increasing the number of experts, we can partition subspaces more finely, but this may lead to homogeneity. We thus dynamically adjust ranks and interaction modes to expand expert diversity. In addition, corresponding to the mixed prompt chain, we decompose prompts into high- and low- pass, alternately integrating features with semantics and focusing priors. When the number of prompts is greater than the number of features, we employ the hash allocation strategy in Eq 5. The details of the MPIE are shown in Figure 4.

Traditional image processing techniques, such as Fourier transforms, are challenging to dynamically separate features and are susceptible to sparse content deviations. We utilize learnable high-pass and low-pass filters, and introduce a DySparse$^\kappa$ selection operator to enhance the kernel elements with more informative content, thereby generating high and low-frequency prompts. The DySparse$^\kappa$ operator is:

$$\mathsf{DySparse}^\kappa(\mathsf{Att}^{-c_j}, t^\kappa) = \begin{cases} \mathsf{Att}^{-c_j} & \text{if } \mathsf{Att}^{-c_j} \geq t^\kappa \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where t^κ is the κ-th largest value in the c_j-th column of Att. The low- and high- frequency filters $\mathbf{f}^l$ and $\mathbf{f}^h$ are learned by:

$$\mathbf{f}_l = \mathsf{DySparse}^\kappa(\mathsf{Softmax}(\mathsf{BN}(\mathcal{F}(\mathsf{GAP}(\mathbf{p}))))),\ \mathbf{f}_h = \mathbf{I} - \mathbf{f}_l \tag{12}$$

where BN and $\mathbf{I}$ denotes Batch Normalization and identity matrix, respectively[1]. We apply $\mathbf{f}_l$, $\mathbf{f}_h$ to each group g of $\mathbf{p}$ to obtain low- and high- frequency components $\mathbf{p}_l$ and $\mathbf{p}_h$:

$$\mathbf{p}_l := \mathsf{Cat}(\mathbf{f}^1_l\mathbf{p}^1, ..., \mathbf{f}^g_l\mathbf{p}^g)^4_g,\ \mathbf{p}_h := \mathsf{Cat}(\mathbf{f}^1_h\mathbf{p}^1, ..., \mathbf{f}^g_h\mathbf{p}^g)^4_g \tag{13}$$

Injecting prompts into features through single component is the most straightforward approach, but it lacks diversity and dynamism, failing to adjust the optimal strategy based on scenarios. An alternative solution is to cascade multiple components, yet this introduces excessive computational complexity. Therefore, we introduce MoE, assigning different types and numbers of experts to perceive scene changes and generating gate values through a routing network as the weights for linear combinations of different experts. For the routing network $\mathcal{G}(\cdot)$, we refrain from excessive design, reducing it to a low-rank dimension through GAP and feature mapping:

$$\mathcal{G}(\mathbf{F}^k) = \mathsf{Softmax}(\mathsf{Proj}(\mathsf{GAP}(\mathbf{F}^k)) + \epsilon) \tag{14}$$

where ϵ represent Gaussian noise for training. Preserving the rank of features retains the maximum information content. However, manually determining the optimal rank is hard, and compressing the representation space into a single type lacks diversity. Therefore, we apply dynamic low-rank decomposition aim at reducing the feature dimension C to different ranks R ($R < C$) and ensuring the efficiency and heterogeneity of the experts.

$$\mathbf{p}_l := \mathcal{F}^{C\to R}_l(\mathbf{p}_l),\ \mathbf{p}_h := \mathcal{F}^{C\to R}_h(\mathbf{p}_h),\ \mathbf{F}^k := \mathcal{F}^{C\to R}(\mathbf{F}^k) \tag{15}$$

Potential noise in the prompts lead to insufficient or suppressed fine-grained representations (often in high uncertainty regions). Features suffer from similar issues, thus uncertainty estimation affects the fusion effect. We collect evidence from prompts and features based on the subjective logic theory [26], establishing connections on the Dirichlet distribution $\mathcal{D}$. Technically, for $\forall i, j \in \mathbb{F}$, *i.e.*,$\{\mathbf{p}_h, \mathbf{p}_l, \mathbf{F}^k\}$, category m, we utilize Softplus to obtain the evidence map $\mathbf{e}^m_i$ and compute the belief masses $\mathbf{b}^m_i$, Dirichlet strength $\mathbf{S}_i$ and uncertainty map $\mathbf{u}_i$ by:

$$\mathbf{b}^m_i = \frac{\mathbf{e}^m_i}{\mathbf{S}_i}, \quad \mathbf{u}_i = \frac{2}{\mathbf{S}_i} \quad \text{and} \quad \mathbf{S}_i = \sum_{m=0}^{1} \mathbf{e}^m_i + 1 \tag{16}$$

The belief masses of evidence may conflict. Therefore, we quantify consistency $\mathbf{c}_{ij}$ of different evidences by:

$$\mathbf{c}_{ij} = \sum_{m=0}^{1} \mathbf{b}^m_i \cdot \mathbf{b}^{1-m}_j \tag{17}$$

We then fuse $\mathbf{b}^m$ based on the Dempster-Shafer theory [27]:

$$\mathbf{b}^m = \frac{\sum_{i,j\in\mathbb{F}}\sum_{j\neq i}(\mathbf{b}^m_i \cdot \mathbf{u}_j) + \sum_{i,j\in\mathbb{F}}\mathbf{b}^m_i \cdot \prod_{j\neq i}\mathbf{u}_j}{1 - \sum_{i,j\in\mathbb{F}}\sum_{j\neq i}\mathbf{c}_{ij}} \tag{18}$$

Thus we have fused uncertainty map $\mathbf{U}$ by:

$$\mathbf{U} = \frac{\prod_{i\in\mathbb{F}}\mathbf{u}_i}{1 - \sum_{i,j\in\mathbb{F}}\sum_{j\neq i}\mathbf{c}_{ij}} \tag{19}$$

The foundation of MPIE lies in the experts $\mathcal{E}$. Specifically, we apply 1×1 convolution to $\mathbf{p}_l$, $\mathbf{p}_h$ for dimension mapping, followed by 3×3 convolution to encode local features, generating $\mathbf{K}_l$ and $\mathbf{K}_h$. For $\mathbf{F}^k$, we generate query, key and value through channel split, which is expressed as:

$$\{\mathbf{Q}, \mathbf{Q}_{mask}, \mathbf{K}, \mathbf{V}_{self}, \mathbf{V}_{cross}\}^k = \mathsf{Split}(\mathcal{F}(\mathcal{F}(\mathbf{F}^k))) \tag{20}$$

Then we use $\mathbf{O}^{k-1}$ to decouple feature maps and $\mathbf{U}^{k-1}$ to introduce uncertainty. For $\forall k > 2,\ k\%2 \neq 0$, we have:

$$\mathbf{Q}^k_{mask} := \mathcal{F}(\mathbf{Q}^k_{mask} \cdot (1 - \mathbf{O}^{k-1})) + \mathcal{F}(\mathbf{Q}^k_{mask} \cdot \mathbf{U}^{k-1}) \tag{21}$$

Otherwise, we have:

$$\mathbf{Q}^k_{mask} := \mathcal{F}(\mathbf{Q}^k_{mask} \cdot \mathbf{O}^{k-1}) + \mathcal{F}(\mathbf{Q}^k_{mask} \cdot \mathbf{U}^{k-1}) \tag{22}$$

[1] For simplicity, we denote the prompts without indices in this section.

Furthermore, we achieve interaction at three spaces, *i.e.,* channel ($C \times C$), spatial ($H \times W$), and frequency ($F \times F$). Thus the correlation matrix $\mathbf{cm}_{p_h \to f}$ from $\mathbf{K}_h$ to $\mathbf{Q}^k$ is :

$$\mathbf{cm}_{p_h \to f} = \mathsf{Softmax}(\frac{\mathbf{Q}^k_{mask}\mathbf{K}_h}{\tau_{p_h \to f}}) \in \{\mathbb{R}^{C \times C} \vee \mathbb{R}^{H \times W} \vee \mathbb{R}^{F \times F}\} \tag{23}$$

where $\tau_{p_h \to f}$ is a learnable scaling factor. To balance diversity and complexity among N_e experts, we randomly select an interaction space for each expert. Similarly, we can generate $\mathbf{K}_l$ to $\mathbf{Q}^k_{mask}$ pair $\mathbf{cm}_{p_l \to f}$, $\mathbf{Q}^k_{mask}$ to $\mathbf{Q}^k$ pair $\mathbf{cm}_{m \to f}$ and $\mathbf{Q}^k_{mask}$ to $\mathbf{K}^k$ pair $\mathbf{cm}_{f \to f}$. We obtain the updated features $\mathbf{F}^k$ of spatial or channel expert by:

$$\begin{aligned}\mathbf{F}^k = \mathcal{F}(&\mathcal{F}^{R \to C}(\mathbf{cm}_{p_h \to f} \otimes \mathbf{V}^k_{cross} \otimes \mathbf{cm}_{p_l \to f}) \\ &+ \mathcal{F}^{R \to C}(\mathbf{cm}_{f \to f} \otimes \mathbf{V}^k_{self} \otimes \mathbf{cm}_{m \to f}))\end{aligned} \tag{24}$$

where $\otimes$ is matrix multiplication. $\mathcal{F}^{R \to C}$ is used to restore features from the low-rank dimension back to the original dimension. Operations in the frequency space are similar. The final output of MPIE is:

$$\hat{\mathbf{F}}^k = \mathbf{F}^k + \sum_{i=0}^{N_e} \mathcal{G}(\mathbf{F}^k) \cdot \mathcal{E}^i(\mathbf{F}^k, \mathbf{p}_l, \mathbf{p}_h) \tag{25}$$

During the training phase, we leverage all expert knowledge. In the testing phase, we select the top-k experts to ensure efficiency. In Figure 8, we illustrate the performance and efficiency variations for different number of expert.

E. Loss Function

The label of frequency prototype $\mathbf{v}$ is defined as y. The similarity $\mathcal{S}^{ij}$ between prototype $\mathbf{v}^i$ and prototype $\mathbf{v}^j$ is $\frac{\mathbf{v}^i \mathbf{v}^j}{\tau \|\mathbf{v}^i\| \|\mathbf{v}^j\|}$, where τ is a temperature coefficient. For $\mathcal{S}^{ij}$, the condition for positive pairs is $\mathrm{y}^i = \mathrm{y}^j$ where $i \neq j$, forming $\mathcal{P}^i$ and the condition for negative pairs is $\mathrm{y}^i \neq \mathrm{y}^j$, forming $\mathcal{N}^i$. For any batch containing N samples, we have:

$$\begin{aligned}\mathcal{L}_{\mathrm{FPC}} = &-\frac{1}{N}\sum_{i=1}^{N}\log(\frac{\sum_{j \in \mathcal{P}^i}\exp(\mathcal{S}^{ij})}{\sum_{j=1}^{N}\exp(\mathcal{S}^{ij}) - \exp(\mathcal{S}^{ii})}) \\ &-\frac{1}{N}\sum_{i=1}^{N}\log(1 - \frac{\sum_{j \in \mathcal{N}^i}\exp(\mathcal{S}^{ij})}{\sum_{j=1}^{N}\exp(\mathcal{S}^{ij}) - \exp(\mathcal{S}^{ii})})\end{aligned} \tag{26}$$

The former loss is for positive pairs, and the latter is for negative pairs. To quantify the fusion uncertainty, we first generate concentration parameter $\boldsymbol{\alpha}^m$:

$$\boldsymbol{\alpha}^m = (\mathbf{b}^m \cdot \mathbf{S}) + 1, \ \hat{\boldsymbol{\alpha}}^m = \mathcal{Y}^m + (1 - \mathcal{Y}^m)\boldsymbol{\alpha}^m \tag{27}$$

where $\mathcal{Y}^m$ represent GT for class-m. We use Kullback-Leibler (KL) divergence to keep evidence for the negative label as 0. We have fused strength $\mathbf{S} = \frac{2}{\mathbf{U}}$. We then formulate the $\mathcal{L}_{\mathrm{TEA}}$:

$$\mathcal{L}_{\mathrm{TEA}} = \sum_{l=1}^{L_f}\sum_{m=0}^{1}\mathcal{Y}^m(\Gamma(\mathbf{S}) - \Gamma(\boldsymbol{\alpha}^m)) + \mathsf{KL}[\mathcal{D}(\mathbf{O}|\hat{\boldsymbol{\alpha}}^m)\|\mathcal{D}(\mathbf{O}|1)] \tag{28}$$

where $\Gamma(\cdot)$ represent the digamma function. We apply supervision to prediction $\mathbf{O}$ for all stages. For the fully-supervised setting, we employ intersection over union loss ($\mathcal{L}_{\mathrm{IoU}}$), thus we formulate $\mathcal{L}_{\mathrm{Fully}}$ by:

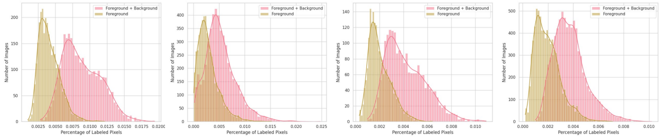

Fig. 5. Percentage of labeled pixels on four mirror datasets. From left to right: MSD, PMD, Mirror-RGBD, and VMD.

$$\mathcal{L}_{\mathrm{Fully}} = \sum_{l=1}^{L_f}\mathcal{L}_{\mathrm{TEA}} + \mathcal{L}_{\mathrm{IoU}} + \lambda\mathcal{L}_{\mathrm{FPC}} \tag{29}$$

Point supervision can be regarded as a special case of scribble supervision. Therefore, for the weakly-supervised setting, we utilize partial TEA loss ($\mathcal{L}^{\mathrm{p}}_{\mathrm{TEA}}$) and smooth loss ($\mathcal{L}_{\mathrm{S}}$), thus we formulate $\mathcal{L}_{\mathrm{weakly}}$ by:

$$\mathcal{L}_{\mathrm{Weakly}} = \sum_{l=1}^{L_f}\mathcal{L}^{\mathrm{p}}_{\mathrm{TEA}} + \mathcal{L}_{\mathrm{S}} + \lambda\mathcal{L}_{\mathrm{FPC}} \tag{30}$$

where we empirically set $\lambda = 0.1$. Note that we do not leverage edge loss like [16] as additional supervision.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets for Fully-Supervised. MSD [7] and PMD [6] datasets contain 3,063 and 5,096 training images, 955 and 571 testing images, respectively. Mirror-RGBD [5] contains 2,000 training images and 1,049 testing images, accompanied by depth maps. VMD [18] has 143 (7,835 images) and 126 (7,152 images) videos for training and testing. We shuffle all video frames to avoid introducing temporal information. To validate the generalization, we further conduct experiments in six related benchmarks, *i.e.,* GDD [56], [57] (glass), ISTD [58] (shadow), DUTS [59] (salience), COD10K [60] (camouflage), MAS3K [61] (underwater), ORSI4199 [62] (remote sensing).

2) Datasets for Weakly-Supervised. Based on S-Mirror [16], we relabel the training images of VMD with scribble. As shown in Figure 5, the majority of the annotation ratios are below 0.01, with a comparable ratio between foreground and background. We further relabel the training images of MSD, PMD, Mirror-RGBD, and VMD via point.

3) Evaluation Metrics. Follow [15], [16], we compute five evaluation metrics, *i.e.,* S-measure (S_m), mean E-measure (E_m), weighted F-measure (F^w_β), intersection over union (IoU) and Mean Absolute Error (MAE). The higher the better for the first four. For fair comparison, we utilize the provided prediction results (or pretrained model weights), or retrain using the official codes.

4) Annotation strategy. For scribble annotation, our labeling process is divided into two steps: initial labeling and calibration. We first collect training images of MSD dataset, which mainly contains indoor scenes and is easy to find mirror regions, taking about 5s to label an image. Similarly, we label the training images of PMD dataset, which is collected from six public datasets with a wide range of scenes, leading to interference during the labeling process and prolonging the time,

TABLE I

Quantitative comparison on four mirror datasets. S, M, C, U, O, WS, WP, Att. D., and V. represent denote SOD, MD, COD, unified segmentation, other binary segmentation, weakly supervised (scribble), weakly supervised (point), attribution, depth and video, respectively. The best performances are bolded and the second are underlined.

Methods	Att.	D.	V.	MSD					PMD					Mirror-RGBD					VMD				
				MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
CNN-based Fully-Supervised Setting																							
MINet [28] CVPR'20	S	✗	✗	0.088	0.792	0.819	0.715	0.664	0.038	0.794	0.822	0.667	0.601	0.063	0.803	0.866	0.746	0.695	0.148	0.628	0.660	0.485	0.459
LDF [29] CVPR'20	S	✗	✗	0.068	0.821	0.867	0.773	0.729	0.038	0.799	0.833	0.683	0.633	0.060	0.805	0.870	0.744	0.698	0.146	0.645	0.682	0.523	0.475
MENet [30] CVPR'23	S	✗	✗	0.054	0.868	0.906	0.829	0.805	0.033	0.826	0.873	0.727	0.680	0.052	0.811	0.868	0.784	0.722	0.133	0.703	0.709	0.589	0.510
MirrorNet [7] ICCV'19	M	✗	✗	0.065	0.850	0.891	0.812	0.790	0.043	0.761	0.841	0.663	0.585	0.055	0.786	0.819	0.760	0.695	0.145	0.675	0.750	0.564	0.505
PMDNet [6] CVPR'20	M	✗	✗	0.047	0.875	0.908	0.845	0.815	0.032	0.810	0.859	0.716	0.660	0.050	0.826	0.873	0.791	0.731	0.128	0.709	0.732	0.601	0.532
SANet [4] CVPR'22	M	✗	✗	0.054	0.862	0.898	0.829	0.798	0.071	0.808	0.839	0.721	0.668	0.048	0.834	0.887	0.800	0.750	0.132	0.724	0.744	0.599	0.518
VCNet [14] TPAMI'22	M	✗	✗	0.044	‡	‡	‡	0.854	0.028	‡	‡	‡	0.694	0.052	‡	‡	‡	0.730	0.123	0.716	0.740	0.606	0.539
HetNet [9] AAAI'23	M	✗	✗	0.043	0.881	0.921	0.854	0.824	0.029	0.828	0.865	0.734	0.690	0.048	0.840	0.892	0.805	0.751	0.118	0.730	0.739	0.610	0.544
GateNet [31] IJCV'24	U	✗	✗	0.053	0.872	0.907	0.829	0.811	0.048	0.785	0.829	0.649	0.621	0.059	0.795	0.	0.769	0.711	0.153	0.653	0.683	0.496	0.429
PDNet [5] CVPR'21	M	✓	✗	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.042	0.856	0.906	0.825	0.778	‡	‡	‡	‡	‡
PopNet [32] ICCV'23	C	✓	✗	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.039	0.847	0.915	0.835	0.780	‡	‡	‡	‡	‡
VMDNet [18] CVPR'23	M	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.105	0.731	0.742	0.623	0.567
MGVMD [19] CVPR'24	M	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.100	0.742	0.756	0.639	0.591
VGSDNet [33] AAAI'24	O	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.102	0.740	0.748	0.633	0.589
Transformer-based Fully-Supervised Setting																							
SETR [34] CVPR'21	S	✗	✗	0.071	0.797	0.840	0.750	0.690	0.035	0.753	0.775	0.633	0.564	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
VST [35] ICCV'21	S	✗	✗	0.054	0.861	0.901	0.818	0.791	0.036	0.783	0.814	0.639	0.591	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
FSPNet [36] CVPR'23	C	✗	✗	0.057	0.871	0.897	0.818	0.807	0.065	0.743	0.752	0.513	0.530	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
FPNet [37] MM'23	C	✗	✗	0.042	0.883	0.917	0.849	0.827	0.033	0.823	0.874	0.717	0.673	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
VSCode [25] CVPR'24	C	✗	✗	0.077	0.800	0.820	0.721	0.687	0.042	0.787	0.816	0.656	0.607	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
Mask2Former [38] CVPR'22	U	✗	✗	0.065	0.826	0.850	0.760	0.736	0.042	0.802	0.836	0.697	0.650	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
SAM [39] ICCV'23	U	✗	✗	0.124	‡	‡	‡	0.515	0.052	‡	‡	‡	0.647	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
EVP [11] CVPR'23	U	✗	✗	0.064	0.845	0.896	0.811	0.780	0.037	0.793	0.861	0.694	0.634	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
Spider [13] ICML'24	U	✗	✗	0.041	0.898	0.932	0.871	0.856	0.028	0.847	0.889	0.750	0.717	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
DualSAM [12] CVPR'24	O	✗	✗	0.039	0.903	0.932	0.882	0.848	0.034	0.816	0.839	0.705	0.636	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
GhostingNet [40] TPAMI'24	O	✗	✗	0.064	0.863	0.897	0.830	0.811	0.038	0.802	0.856	0.685	0.642	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
SATNet [10] AAAI'23	M	✗	✗	0.033	0.887	0.916	0.865	0.834	0.025	0.826	0.858	0.739	0.684	0.034	0.857	0.900	0.829	0.772	0.101	0.728	0.763	0.653	0.609
CSFwinformer [8] TIP'24	M	✗	✗	0.045	0.875	0.905	0.846	0.821	0.024	0.831	0.864	0.756	0.700	**0.031**	0.864	0.908	0.836	0.786	0.102	0.738	0.755	0.642	0.600
DPRNet [15] TCSVT'24	M	✗	✗	0.033	0.904	0.934	0.888	0.866	0.026	0.844	0.894	0.766	0.721	0.047	0.845	0.899	0.811	0.761	0.100	0.749	0.777	0.675	0.628
PICRNet [41] MM'23	S	✓	✗	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.053	0.815	0.877	0.772	0.718	‡	‡	‡	‡	‡
XMSNet [42] MM'23	S	✓	✗	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.033	**0.872**	0.909	0.840	0.775	‡	‡	‡	‡	‡
CPNet [43] IJCV'24	S	✓	✗	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.035	0.870	0.913	**0.844**	0.788	‡	‡	‡	‡	‡
SLTNet [44] CVPR'22	C	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.102	0.766	0.783	0.669	0.652
S2Net [45] CVPR'23	O	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.098	0.755	0.778	0.672	0.646
SLANet [46] ICME'24	M	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	0.099	‡	‡	‡	0.634
TBGDiff [47] MM'24	U	✗	✓	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	**0.096**	0.763	0.795	0.689	0.655
Ours	M	✗	✗	**0.024**	**0.926**	**0.959**	**0.922**	**0.902**	**0.022**	**0.872**	**0.922**	**0.810**	**0.770**	0.034	0.865	**0.923**	0.842	**0.800**	0.097	**0.782**	**0.822**	**0.710**	**0.668**
Weakly-Supervised Setting																							
SS [48] CVPR'20	WS	✗	✗	0.158	0.681	0.747	0.567	0.527	0.055	0.726	0.790	0.571	0.513	0.127	0.654	0.722	0.537	0.444	0.199	0.636	0.668	0.487	0.458
SCWS [49] AAAI'21	WS	✗	✗	0.121	0.770	0.814	0.678	0.659	0.059	0.759	0.807	0.599	0.579	0.118	0.690	0.743	0.547	0.498	0.189	0.644	0.684	0.498	0.465
TEL-S [50] CVPR'22	WS	✗	✗	0.113	0.790	0.828	0.726	0.702	0.065	0.746	0.791	0.559	0.563	0.100	0.733	0.785	0.615	0.576	0.184	0.631	0.685	0.480	0.440
SAM-S [39] ICCV'23	WS	✗	✗	0.124	0.766	0.809	0.686	0.663	0.083	0.720	0.759	0.496	0.537	0.110	0.716	0.756	0.577	0.545	0.195	0.628	0.666	0.468	0.432
SCOD [51] AAAI'23	WS	✗	✗	0.092	0.786	0.851	0.728	0.685	0.055	0.764	0.819	0.609	0.586	0.106	0.698	0.762	0.581	0.518	0.180	0.645	0.685	0.500	0.460
SMD [16] AAAI'24	WS	✗	✗	0.078	0.828	0.878	0.780	0.750	0.051	0.773	0.824	0.630	**0.600**	0.088	0.754	0.806	0.655	0.616	0.160	0.664	0.711	0.525	0.479
Ours-S	WS	✗	✗	**0.063**	**0.846**	**0.895**	**0.805**	**0.782**	**0.047**	**0.778**	**0.836**	**0.650**	0.597	**0.074**	**0.773**	**0.834**	**0.692**	**0.654**	**0.150**	**0.689**	**0.723**	**0.552**	**0.516**
SS-P [48] CVPR'20	WP	✗	✗	0.248	0.610	0.628	0.472	0.457	0.143	0.636	0.660	0.390	0.405	0.187	0.566	0.627	0.350	0.339	0.207	0.586	0.650	0.424	0.380
SCWS-P [49] AAAI'21	WP	✗	✗	0.211	0.613	0.696	0.492	0.460	0.118	0.666	0.701	0.458	0.441	0.161	0.622	0.656	0.432	0.399	0.195	0.604	0.648	0.441	0.394
TEL-P [50] CVPR'22	WP	✗	✗	0.196	0.650	0.710	0.530	0.502	0.123	0.670	0.702	0.474	0.448	0.190	0.588	0.607	0.387	0.366	0.191	0.611	0.678	0.466	0.410
SAM-P [39] ICCV'23	WP	✗	✗	0.220	0.635	0.674	0.513	0.491	0.137	0.637	0.664	0.414	0.398	0.181	0.583	0.631	0.368	0.350	0.210	0.586	0.660	0.434	0.385
SCOD-P [51] AAAI'23	WP	✗	✗	0.183	0.687	0.720	0.573	0.553	0.109	0.687	0.733	0.513	0.476	0.145	0.650	0.688	0.479	0.443	0.184	0.614	0.684	0.469	0.418
SMD-P [16] AAAI'24	WP	✗	✗	0.168	0.696	0.742	0.586	0.560	0.091	0.706	0.760	0.550	0.501	0.135	0.670	0.713	0.516	0.465	0.175	0.621	0.690	0.473	0.416
Ours-P	WP	✗	✗	**0.141**	**0.725**	**0.775**	**0.628**	**0.605**	**0.081**	**0.718**	**0.782**	**0.572**	**0.517**	**0.121**	**0.691**	**0.755**	**0.550**	**0.508**	**0.163**	**0.667**	**0.702**	**0.516**	**0.476**

which is about 9s for an image. Mirror-RGBD dataset contains 202 scenes, which contain many high-resolution images and interfering objects, thus taking about 8s to label an image. Due to the clear similarities between frames, the annotations within the same segment in the VMD dataset have shorter durations compared to those between different segments. During the annotation process, we follow the guideline of simplicity and accuracy, using few scribbles and accurately labeling the foregrounds and backgrounds. After completing the initial labeling, we double-check to ensure that we do not miss any mirror region and there are no mislabeling and complex scribbles. For point annotation, we employ the same strategy. The main difference is that scribble annotation takes two to three times longer than point annotation.

B. Implementation Details

We implement our model via PyTorch and conduct all experiments on NVIDIA A100 GPUs. Following [6], [13], [16], [30], [56], [63], for mirror, glass, camouflage, salience datasets, input are resized to 384×384 for fair comparison. Following [12], [64], for shadow, underwater datasets, input are resized to 512×512. Following [65], 352×352 are applied for remote sensing dataset. For training, we use the AdamW as our optimizer to update the model parameters, with 200 epochs, the batch size of 40, and the initial learning rate of 1e-4. To prevent overfitting, we employ data augmentation, *e.g.*, random rotation. For testing, we compute generated predictions without adopting any post-processing operations.

C. Comparison on MD benchmarks

We consider both fully supervised and weakly supervised settings, different backbone networks, and select various BS and SS methods for comparisons.

1) Quantitative Comparison. As shown in Table I, for the fully-supervised settings, our method surpasses the second best by 0.9%, 2.2%, 2.5%, 3.4%, 3.6% and -0.2%, 2.8%, 2.8%, 4.4%, 4.9% on the MSD and PMD benchmarks, respectively. Represented by DualSAM, based on vision foundation model *i.e.,* SAM, theoretically leveraging strong pre-training priors and parameter fine-tuning should yield promising results. However, it fails to achieve performance comparable to MD-specific models in practice. We attribute it to insufficient modeling of reflection perception and biases towards entities. In other words, its inability to effectively discriminate between

TABLE II
QUANTITATIVE ABLATION OF CRUCIAL COMPONENTS AND BACKBONES OF OUR SVRNET. C1, C2, C3, C4 AND C5 INDICATE THE PPCR, PE, MPIE, FPC LOSS, AND TEA LOSS, RESPECTIVELY. B1, B2, B3, B4, B5 REPRESENT PVT-V2B2 [52], SWIN-B [53], SWIN-S [53], RESNETXT101 [54], RESNET50 [55], RESPECTIVELY. PPCR IS THE FOUNDATION FOR OTHER COMPONENTS, IMPLYING THAT THEY CANNOT WORK WITHOUT THE PPCR. THE FIRST LINE IS THE *Baseline* MODEL. WHEN THE MPIE IS NOT AVAILABLE, WE UTILIZE CHANNEL CONCATENATION AS AN ALTERNATIVE.

Method	Backbone					Component					MSD					PMD				
	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
I	✓										0.047	0.872	0.884	0.843	0.808	0.033	0.798	0.852	0.726	0.675
II	✓					✓					0.040	0.888	0.907	0.874	0.839	0.029	0.820	0.874	0.743	0.695
III	✓					✓	✓				0.032	0.907	0.930	0.893	0.864	0.026	0.839	0.890	0.761	0.725
IV	✓					✓		✓			0.035	0.895	0.924	0.882	0.852	0.026	0.832	0.885	0.764	0.718
V	✓					✓	✓	✓			0.028	0.911	0.945	0.913	0.888	0.024	0.858	0.905	0.788	0.747
VI	✓					✓	✓	✓	✓		0.025	0.923	0.953	0.920	0.896	0.023	0.865	0.914	0.800	0.759
VII	✓					✓	✓	✓	✓	✓	0.024	0.926	0.959	0.922	0.902	0.022	0.872	0.922	0.810	0.770
VIII		✓				✓	✓	✓	✓	✓	0.025	0.930	0.955	0.927	0.895	0.020	0.865	0.927	0.814	0.765
IX			✓			✓	✓	✓	✓	✓	0.026	0.924	0.952	0.929	0.899	0.020	0.873	0.918	0.807	0.773
X				✓		✓	✓	✓	✓	✓	0.034	0.902	0.943	0.891	0.876	0.025	0.844	0.891	0.770	0.735
XI					✓	✓	✓	✓	✓	✓	0.032	0.897	0.932	0.880	0.870	0.027	0.839	0.884	0.762	0.728

TABLE III
QUANTITATIVE COMPARISON ON COMPUTATION AND MODEL COMPLEXITY. WE COMPARE WITH FOUR MD MODELS ON PARAMETERS (M), FLOPS (GMAC).

Methods	Input Size	Backbone	FLOPs↓	Params.↓
PMDNet	384×384	ResNeXt101	101.54	147.66
SATNet	512×512	Swin-S	153.00	139.36
CSFwinformer	512×512	Swin-B	139.45	150.54
DPRNet	384×384	PVT-v2	30.47	31.19
Ours-R101	384×384	ResNeXt101	39.58	45.25
Ours-S	384×384	Swin-S	31.49	51.73
Ours-B	384×384	Swin-B	48.60	89.80
Ours-P	384×384	PVT-v2	**15.15**	**27.24**

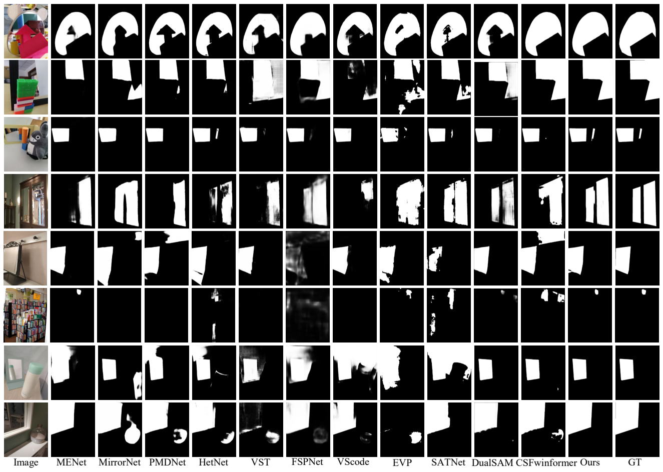

Fig. 6. Qualitative comparison on the mirror scenes under fully supervised.

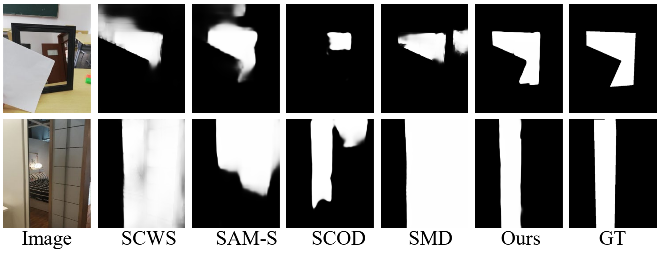

Fig. 7. Qualitative comparison on the mirror scenes under weakly supervised.

mirror and entity disparities. Compared to EVP, VSCode, and Spider, which are also based on visual prompts, our method's advantage is evident, demonstrating that implicitly learning semantic prompts and constructing prompt chains for visual reasoning is superior to explicitly utilizing independent prompts. In addition, the inclusion of MoRE further enhances scene change perception. Unlike FPNet, CSFwinformer, and DPRNet, which consider foreground-background differences from features perspective, our method has two main distinctions. 1) Motivation, we decouple prompts into high- and low- frequency to provide finer-grained references for features; 2) Technically, we design more optimal content-based kernel decomposition component instead of using traditional image transforms, as elaborated in Table V.

Without the assistance of depth maps and temporal signals, our method still outperforms the second best by -0.3%, -0.7%, 1.0%, -0.2%, 1.2% and -0.1%, 1.9%, 2.7%, 2.1%, 1.3% on the Mirror-RGBD and VMD benchmarks, respectively. Depth differences can preliminarily locate mirror regions, our SVRNet, despite not utilizing depth maps, surpasses MD- and depth-specific model *i.e.,* PDNet, SOD- and RGBX-specific model *i.e.,* XMSNet, indicating that efficient visual reasoning can avoid complex prior fusion and achieve better localization effects to some extent. Our method processes video frames independently but achieves performance comparable to TBGDiff that fuses text-to-image diffusion model and sequence signals, showing the additional scalability potential.

In the scribble-supervised setting, our method surpasses the second best by an average of 1.1%, 1.7%, 1.7%, 2.7%, and 2.6% on four benchmarks, respectively. Similarly, in the point-supervised setting, the improvements are 1.6%, 2.7%, 2.7%, 3.5%, and 4.1%, indicating that the proposed components are not limited to fully supervised settings.

2) Qualitative Comparison. As show in Figure 6, the first two rows represent imagings with and without corresponding entities. The third and fourth rows demonstrate regular and irregular occlusions. Our approach establish associations between entities and imagings as well as context around

TABLE IV
QUANTITATIVE ABLATION ON THE COMPONENTS OF THE PPCR AND THE PE.

Model	Choice		Prototype				Weight		MSD					PMD				
	CoT	Single	Foreground	Background	All	Half	Spatial	Channel	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I		✓					✓		0.028	0.900	0.928	0.895	0.864	0.028	0.837	0.888	0.763	0.735
II	✓						✓		0.027	0.911	0.940	0.907	0.877	0.026	0.851	0.903	0.779	0.750
III	✓		✓		✓		✓		0.026	0.917	0.939	0.915	0.889	0.025	0.849	0.910	0.786	0.755
IV	✓			✓	✓		✓		0.027	0.914	0.940	0.913	0.884	0.026	0.847	0.907	0.790	0.759
V	✓		✓	✓	✓		✓		0.029	0.908	0.933	0.905	0.879	0.027	0.839	0.901	0.781	0.751
VI	✓		✓	✓		✓	✓		0.026	0.922	0.949	0.921	0.896	0.024	0.856	0.916	0.798	0.768
VII	✓		✓	✓		✓		✓	0.025	0.920	0.953	0.917	0.898	0.024	0.859	0.912	0.803	0.765
VIII	✓		✓	✓		✓	✓	✓	0.024	0.926	0.959	0.922	0.902	0.022	0.872	0.922	0.810	0.770

Model	Gate.		Choice				Choice		MSD					PMD				
	Spatial	Channel	Cascade	Residual	Dense	Shuffle	All	Group	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
IX	✓		✓						0.026	0.905	0.945	0.904	0.886	0.025	0.846	0.892	0.785	0.742
X		✓	✓						0.027	0.908	0.943	0.909	0.884	0.024	0.848	0.894	0.787	0.737
XI	✓	✓	✓						0.027	0.913	0.947	0.914	0.888	0.025	0.854	0.898	0.792	0.749
XII	✓	✓		✓					0.025	0.920	0.953	0.920	0.895	0.023	0.859	0.908	0.798	0.756
XIII	✓	✓			✓				0.025	0.922	0.955	0.922	0.900	0.022	0.862	0.912	0.799	0.762
XIV	✓	✓				✓	✓		0.024	0.925	0.957	0.928	0.900	0.022	0.866	0.915	0.803	0.763
XV	✓	✓				✓		✓	0.024	0.926	0.959	0.922	0.902	0.022	0.872	0.922	0.810	0.770

TABLE V
QUANTITATIVE ABLATION ON THE COMPONENTS OF THE MPIE.

Model	Choice		Connection				Expert			MSD					PMD				
	MoE	Fixed	Add	Multiply	Attention	Ours	Spatial	Channel	Frequency	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I		✓	✓							0.026	0.893	0.930	0.893	0.882	0.028	0.826	0.882	0.780	0.730
II	✓		✓							0.027	0.907	0.941	0.905	0.892	0.027	0.836	0.891	0.789	0.744
III	✓			✓						0.027	0.909	0.938	0.903	0.889	0.026	0.838	0.895	0.786	0.740
IV	✓				✓					0.026	0.915	0.944	0.905	0.894	0.026	0.846	0.901	0.791	0.747
V	✓					✓	✓			0.026	0.919	0.948	0.912	0.895	0.025	0.850	0.905	0.791	0.750
VI	✓					✓		✓		0.024	0.917	0.950	0.908	0.898	0.025	0.848	0.908	0.793	0.746
VII	✓					✓			✓	0.026	0.913	0.943	0.908	0.892	0.026	0.852	0.902	0.795	0.749
VIII	✓					✓	✓	✓		0.024	0.924	0.955	0.919	0.900	0.023	0.856	0.911	0.795	0.757
IX	✓					✓	✓		✓	0.025	0.920	0.951	0.921	0.897	0.023	0.859	0.909	0.797	0.753
X	✓					✓		✓	✓	0.025	0.923	0.950	0.917	0.902	0.022	0.863	0.908	0.801	0.758
XI	✓					✓	✓	✓	✓	0.024	0.926	0.959	0.922	0.902	0.022	0.872	0.922	0.810	0.770

Model	Choice		Space				Guidance			MSD					PMD				
	Sparse	Full	Fourier	Wavelet	Laplace	Ours	Same	Alter	Uncertainty	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
XII			✓							0.026	0.911	0.947	0.907	0.899	0.024	0.858	0.899	0.795	0.759
XIII				✓						0.027	0.918	0.951	0.910	0.896	0.027	0.864	0.905	0.797	0.755
XIV					✓					0.026	0.914	0.958	0.912	0.890	0.025	0.861	0.903	0.801	0.761
XV		✓				✓				0.026	0.917	0.952	0.916	0.895	0.024	0.866	0.908	0.799	0.758
XVI	✓					✓				0.025	0.915	0.955	0.919	0.897	0.024	0.868	0.911	0.795	0.762
XVII	✓					✓	✓			0.025	0.920	0.957	0.915	0.897	0.024	0.867	0.910	0.802	0.762
XVIII	✓					✓		✓		0.025	0.923	0.955	0.918	0.898	0.023	0.865	0.914	0.804	0.765
XIX	✓					✓		✓	✓	0.024	0.926	0.959	0.922	0.902	0.022	0.872	0.922	0.810	0.770

imagings. The fifth and sixth rows show scale variations and multiple targets. Our method can enable global modeling to avoid false negative. The last two rows demonstrate scenes with mirrors and tile (or glass). Our approach perceives semantic differences, mitigating false positives. In Figure 7, despite using scribble as supervision, our approach allows achieving detection results close to the GT.

3) Why SVRNet can work in the fully- and weakly- supervised settings? The several components we proposed exhibit promising robustness under various loss function supervisions, in other words, the perception of mirror regions is universal.

4) Why SVRNet can work in the video setting? Apart from the effective processing of individual frames by our components, the FPC loss can establish inter-frame associations to some extent, capturing temporal information.

D. Ablation Study

We validate the effect of proposed components, critical hyperparameters on the MSD and PMD datasets.

1) Effect of the crucial components. As shown in Table II, comparing Models I and II, directly incorporating the PPCR leads to an overall performance improvement of approximately 2%. Combining the PE to filter out noise in prompts and enhance coupling doubles this improvement. However, when combined with MPIE, although it enhances the injection strategy, the improvement is not as significant as when combined with PE. Additionally, the fusion of PE and MPIE results in better performance than using either one alone, indicating their complementary characteristics. In other words, high-quality prompts are fundamental as they enable better understanding of the semantic and structural aspects of input scenarios, while better injection methods can expand the model's representational boundaries. Furthermore, we enhance the model's focus on high uncertainty samples and regions by employing the TEA loss, facilitating the generation of reliable predictions. The FPC loss further enhances the performance, validating its effect in establishing universal representations across images. In Figure 9, as each component is introduced, issues such as reflective interference, size variations, only imagings with no corresponding entities or partial correspondence, and irregular regions are gradually alleviated. And the mirror region transitions from high uncertainty to low uncertainty. We observe that

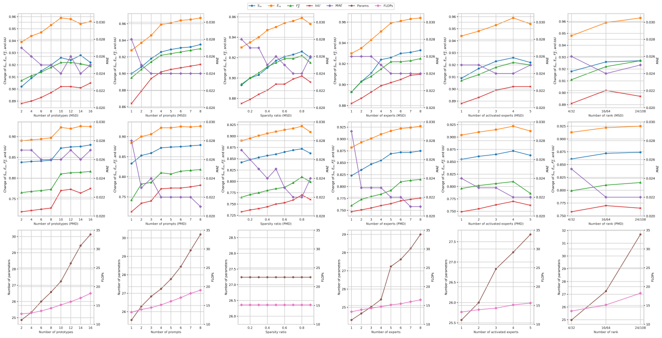

Fig. 8. Quantitative ablation on the number of prototypes $\mathbf{v}$, prompts $\mathbf{p}$, sparsity of frequency decomposition, the number of experts n and activated experts k, the minimum and maximum rank combination, and the corresponding computation and model complexity.

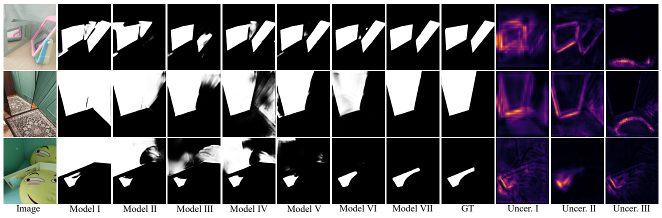

Fig. 9. Qualitative ablation of different component combinations. The model names cmespond to Table II. Uncer. denotes uncertainty map.

Ucer. III show high uncertainty towards non-mirror regions, leading us to infer that the model in this stage pays more attention to the background regions under the guidance of the background mask.

TABLE VI
QUANTITATIVE ABLATION ON RANK R.

Model	Range		Space				MSD				
	Min	Max	All	Random	Linear	Exponential	MAE↓	S_m↑	E_m↑	F^w_β↑	IoU↑
I	✓		✓				0.030	0.883	0.900	0.873	0.866
II		✓	✓				0.023	0.930	0.950	0.928	0.909
III	✓	✓		✓			0.026	0.917	0.947	0.902	0.891
IV	✓	✓			✓		0.027	0.919	0.944	0.909	0.893
V	✓	✓				✓	0.024	0.926	0.959	0.922	0.902

2) Effect of backbone network and model analysis. As shown in Table II and Table III, we use input size of 384×384 and employ three Transformer-based models, *i.e.,* PVT-v2, Swin-S, Swin-B, and CNN-based model, ResNeXt101 as backbone networks. The computational and model complexity of SVR-Net based on PVT-v2 are smaller than the other four versions, with superior performance. On the MSD dataset, using Swin-S and Swin-B as backbones, our method surpasses SATNet and CSFwinformer by 0.7%, 3.7%, 3.6%, 6.4%, 6.5%, and 1.0%, 5.5%, 5.0%, 8.1%, 7.4% across five metrics. When utilizing ResNext101, we outperform PMDNet by 1.3%, 2.7%, 3.5%, 4.6%, and 6.1%. With PVT-v2 as the backbone, our method surpass DPRNet by 0.9%, 2.2%, 2.5%, 3.4%, 3.6%. Despite the introduction of several components, particularly the MPIE based on the MoE architecture, our approach maintains high performance and efficiency.

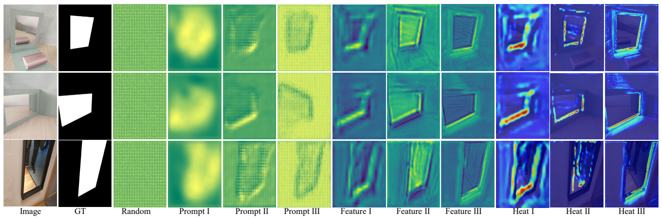

Fig. 10. Qualitative ablation on different state visual prompts, feature maps and heat maps. For convenience, we select three prompts.

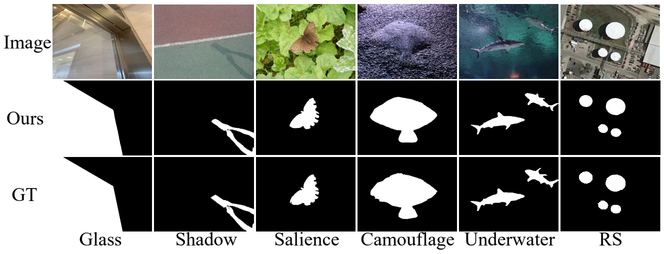

Fig. 11. Qualitative results on other scenes.

3) Effect of components of the PPCR. As shown in Table IV, the performance gap between Models I and II is significant. We analyze that this difference may be due to the fact that independent prompts can only convey limited information, making it difficult to cover complex contexts and susceptible to variations in input. In contrast, the CoT format can progressively reason through perceiving complex visual spaces, and by adjusting the content and quantity of prompts, it can exhibit better robustness and adaptability to different scenarios. In Figure 10, we can transition from the cluttered attention space to focused one by utilizing a chain consisting of only three prompts. Compared to Models III-VI, using either the foreground or background prototype alone to provide priors for prompts is not as effective as combining the two. The foreground-background prototypes can enhance the representations of mirror regions while suppressing interference from non-mirror regions. However, integrating both prototypes into same prompt is less effective than using single prototype, as our analysis suggests prototype conflicts leading to information decay. Therefore, we achieve optimal results by inclusively incorporating the prototypes into different prompt pools. According to Models VI-VIII, the spatial and channel weights of features are complementary.

4) Effect of components of the PE. As shown in Table IV, corresponding to the spatial and channel weights of features generated by the PPCR, in the PE, filtering noise based on spatial and channel information concurrently yields the best results. The classical CoT involves cascading several prompts, but long-range modeling is constrained, resulting in suboptimal performance. In contrast, residual connections can selectively preserve critical features and propagate along the chain. Dense connections further improve performance. And leveraging channel shuffle to mimic the transformation of mirror symmetry in semantic space brings additional benefits. Contrasting Models XIV and XV, dividing the semantic representation into several sub-regions through grouping mechanism helps preserve local features and achieve multi-path prompts purification.

TABLE VII
QUANTITATIVE COMPARISON ON GLASS, SHADOW, SALIENCE, CAMOUFLAGE, UNDERWATER AND REMOTE SENSING BENCHMARKS.

Methods	GDD					ISTD					DUTS				
	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
CSFwinformer [8] TIP'24	0.042	‡	‡	‡	0.912	0.026	‡	‡	‡	0.833	0.027	0.914	0.945	0.880	0.818
CamoFormer [66] TPAMI'24	0.043	0.907	0.942	0.933	0.915	0.015	0.949	0.967	0.926	0.927	0.025	0.915	0.947	0.882	0.820
Spider [13] ICML'24	0.044	0.903	0.938	0.929	0.910	0.019	0.935	0.964	0.882	0.911	0.029	0.909	0.942	0.869	0.800
VSCode [25] CVPR'24	0.046	0.901	0.931	0.924	0.909	0.020	0.929	0.951	0.897	0.908	0.026	0.918	0.948	0.883	0.816
DualSAM [12] CVPR'24	0.048	0.890	0.925	0.922	0.900	0.014	0.944	0.969	0.925	0.902	0.028	0.916	0.951	0.885	0.815
Ours	**0.039**	**0.918**	**0.952**	**0.942**	**0.924**	**0.010**	**0.965**	0.970	**0.944**	**0.940**	**0.023**	**0.927**	**0.963**	**0.901**	**0.842**

Methods	COD10K					MAS3K					ORSI4199				
	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
CSFwinformer [8] TIP'24	0.027	0.849	0.895	0.732	0.680	0.026	0.869	0.827	0.924	0.777	0.030	0.869	0.933	0.845	0.763
CamoFormer [66] TPAMI'24	0.023	0.869	0.931	0.786	0.726	0.024	0.888	0.840	0.930	0.788	0.029	0.883	0.943	0.862	0.782
Spider [13] ICML'24	0.027	0.853	0.918	0.756	0.701	0.025	0.875	0.844	0.925	0.792	0.029	0.878	0.940	0.857	0.758
VSCode [25] CVPR'24	0.028	0.847	0.913	0.744	0.688	0.026	0.862	0.814	0.920	0.769	0.028	0.877	0.945	0.855	0.762
DualSAM [12] CVPR'24	0.028	0.837	0.902	0.746	0.690	0.023	0.884	0.838	0.933	0.789	0.030	0.855	0.928	0.843	0.753
Ours	**0.021**	**0.879**	**0.944**	**0.798**	**0.743**	**0.020**	**0.903**	**0.865**	**0.949**	**0.807**	0.028	**0.894**	**0.954**	**0.874**	**0.799**

5) Effect of components of the MPIE. As shown in Table V, after introducing MoE, even without the additional design of complex experts, the performance of Models II surpasses Model I through simple fusion operations, demonstrating the necessity of dynamic routing experts based on input. Comparing Models II-IV with Models V-XI, more fine-grained representations, involving low- and high-frequency prompts and more diverse interactive spaces, *i.e.,* spatial, channel, and frequency, signifies that sub-space partitioning and new space generation can achieve better injection effects. For frequency decomposition, in comparison between Models XV and XVI, employing all kernel elements is suboptimal. Instead, the sparse mechanism can filter out low-information or even noisy elements, where less is more desirable. Unlike Fourier, Wavelet, and Laplace transforms that use fixed basis functions, which may lead to poor performance when the features vary significantly, our dynamic kernel can learn and adjust kernel parameters based on input data, making it more sensitive to the scene changes *i.e.,* Models XII-XIV and XVI. For different stages, alternately employing foreground-background masks for guidance enhances the perception of feature distinctiveness, proving more effective than using single type of mask. Due to the inherent noise and the tendency to overlook regions of high uncertainty during the optimization process, using uncertainty estimation as guidance can provide focus position. In Figure 10, we present uncertainty maps generated/injected at different stages. *Why choose Evidential Deep Learning (EDL) over Monte Carlo Sampling (MCS) to generate uncertainty maps?* 1) Unlike MCS, which estimates uncertainty by multiple forward passes, EDL achieves higher computational efficiency during inference by directly outputting uncertainty measures (such as evidence values); 2) SVRNet has several critical hyperparameters, such as the number of prompts; using MCS would re-determine the number of sampling points, constraining the model's generalization capability.

For hyperparameters, we analyze on the MSD dataset, and similarly on the PMD dataset.

6) Effect of the number of prototypes N_v. As shown Figure 8 column 1, when $N_v = 0$, no historical prior information is provided for prompts. When $N_v \leq 6$, the prototypes are initialized and gradually accumulated, with small increases in performance. When $6 < N_v \leq 10$, as the prototype quality improves and heterogeneity enhances, the increase becomes significantly larger. However, when $10 < N_v$, the prototype update magnitude is limited, and the performance increase slows down. Besides, the model and computational complexity show overall linear growth. To achieve a balance between efficiency and performance, we set $N_v = 10$.

7) Effect of the number of prompts N_p. In Figure 8 column 2, as N_p increases, the performance gradually improves. When $4 < N_p$, the results remain essentially unchanged, indicating the state of redundancy where the introduction of more prompts yields minimal benefits. Combining with the complexity of computation and model, we set $N_p = 4$. In Figure 10, we observe that visual prompts progressively approach the objects, providing guidance.

8) Effect of the sparsity of frequency decomposition κ. In Figure 8 column 3, when selecting global kernel elements, $\kappa = 100\%$. As κ ranges from 1% to 85%, the performance of the five indicators improves gradually. We analyze that when quite few key elements are selected, other equally important elements are also filtered out, which is unfavorable for decomposition. When $\kappa = 85\%$, the decomposition effect is optimal, corresponding to the best performance. When κ exceeds 85%, noise and low-quality elements are also included, which interferes with the decomposition process.

9) Effect of the number of candidate experts N_e and activated expert N_{ae}. In Figure 8 column 4 and 5, $N_e = 1$ represents the fixed path. For $1 < N_e \leq 6$, as N_e increases, the expandable path grows larger, and performance improvement is significant. When $6 < N_e$, the improvement slows down, indicating that the overall semantic spatial partitioning is mostly completed, and the benefits of finer-grained partitioning are limited. To achieve better balance, we choose $N_e = 6$ to compare different numbers of active experts. Performance improves when $1 < N_{ae} \leq 4$, then decline, highlighting the importance of expert number and sparsity. We set N_{ae} = 4. In

10) Effect of the number of rank N_r. As shown in Table 8, we first conduct three sets of experiments. 1) All experts use the minimum rank, *i.e.,* 4; 2) All experts use the maximum rank, *i.e.,* 108; 3) Ranks are arbitrarily selected from minimum to maximum. The best performance in the third set indicates that rank variation among experts can enhance heterogeneity, *i.e.,* diversity. We further explore rank allocation strategies: 1) Linear growth; 2) Exponential growth; 3) Random selection, with the second approach yielding better performance. We analyze that arithmetic progression is inferior to exponential growth in reflecting rank differences, while randomness lacks stability. In Figure 8 column 6, we also provide the best rank combinations, *i.e,* 16 & 64.

E. Broader Impacts

1) Quantitative comparison: In Table VII, our SVRNet achieves promising performance on six different task benchmarks and five evaluation metrics. Different from VSCode, which apply additional data like depth maps to train, SVRNet only utilizes training images corresponding to each benchmark, achieving an IoU exceeding 5.5% on COD10K. EVP and Spider, based on prompt learning, concentrate on general low level structure segmentation. However, the former directly leverages high-frequency features of images as visual prompts, resulting in subpar performance for non-natural images such as those in remote sensing that are prone to complex background interference, similar to CSFwinformer. The latter introduces macro-level task prompts but lacks micro-level image-specific prompts associations, making it challenging to handle objects with rich details. ICON and GateNet, due to relatively simple model structures, have limited representation space. FSPNet and CamoFormer are designed for COD and perform well in underwater scenarios where objects exhibit camouflage properties. Similarly, DualSAM excels in COD scenes.

2) Qualitative results: As shown in Figure 11, the first row demonstrates transparent object scenarios. The second row illustrates scenes with shadow deformation and lighting variations. The third row showcases salient objects in complex natural backgrounds. The fourth row displays highly camouflaged targets. The fifth row depicts low-contrast underwater scenes. The sixth row presents complex remote sensing scenarios. Our method achieves accurate localization and comprehensive detection driven by the characteristics of different tasks.

3) Why SVRNet can work on various tasks? The core of SVRNet lies in scene-aware visual reasoning, enabling it to equip different lengths and types of reasoning chains and hybrid experts based on the characteristics of inputs, achieving customization. When dealing with glass, natural, and camouflage scenes, our model pays more attention to context awareness and differential modeling, such as depth and frequency. In remote sensing scenarios, the target scales are typically smaller and easily affected by complex background interference, hence requiring more visual prompts and experts. In shadow and underwater scenes, the model focus more on low-level visual features, such as light intensity and color.

V. Conclusion

We rethink the existing MD paradigms and introduce visual prompts. By incorporating hierarchical prototype banks and prompt pools to construct mixed prompt chains for visual spatial reasoning. We expect the prompts to effectively guide localization, and we design coupled purification and statistical uncertainty estimation. In response to the scene awareness of the prompt chain, we also propose MPIE for custom injection strategies for different inputs. We formulate the FPC loss to decouple and obtain highly generalized representations. To enrich the MD community, we relabel 25,828 images for weakly supervised research. Our method achieves promising performance on three task settings and ten datasets. Numerous ablation experiments validate the effect of each component.

References

[1] I. Balazevic, D. Steiner, N. Parthasarathy, R. Arandjelović, and O. Henaff, "Towards in-context scene understanding," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole *et al.*, "Reconfusion: 3d reconstruction with diffusion priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 551–21 561.

[3] J. Li, A. Padmakumar, G. Sukhatme, and M. Bansal, "Vln-video: Utilizing driving videos for outdoor vision-and-language navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 517–18 526.

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

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

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

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

[8] 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*, 2024.

[9] R. He, J. Lin, and R. W. Lau, "Efficient mirror detection via multi-level heterogeneous learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 790–798.

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

[11] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 434–19 445.

[12] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic animals and where to find them: Segment any marine animal with dual sam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2578–2587.

[13] X. Zhao, Y. Pang, W. Ji, B. Sheng, J. Zuo, L. Zhang, and H. Lu, "Spider: A unified framework for context-dependent concept understanding," *arXiv preprint arXiv:2405.01002*, 2024.

[14] 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, 2022.

[15] M. Zha, F. Fu, Y. Pei, G. Wang, T. Li, X. Tang, Y. Yang, and H. T. Shen, "Dual domain perception and progressive refinement for mirror detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[16] M. Zha, Y. Pei, G. Wang, T. Li, Y. Yang, W. Qian, and H. T. Shen, "Weakly-supervised mirror detection via scribble annotations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6953–6961.

[17] J. Lin and R. W. Lau, "Self-supervised pre-training for mirror detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 227–12 236.

[18] J. Lin, X. Tan, and R. W. Lau, "Learning to detect mirrors from videos via dual correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9109–9118.

[19] A. Warren, K. Xu, J. Lin, G. K. Tam, and R. W. Lau, "Effective video mirror detection with inconsistent motion cues," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 244–17 252.

[20] K. Xu, T. W. Siu, and R. W. Lau, "Zoom: Learning video mirror detection with extremely-weak supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6315–6323.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] F. Yang, S. Yang, M. A. Butt, J. van de Weijer *et al.*, "Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] C. Wang, J. Pan, W. Lin, J. Dong, W. Wang, and X.-M. Wu, "Selfpromer: Self-prompt dehazing transformers with depth-consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5327–5335.

[24] L. Yan, C. Han, Z. Xu, D. Liu, and Q. Wang, "Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning." in *IJCAI*, 2023, pp. 1622–1630.

[25] Z. Luo, N. Liu, W. Zhao, X. Yang, D. Zhang, D.-P. Fan, F. Khan, and J. Han, “Vscode: General visual salient and camouflaged object detection with 2d prompt learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 169–17 180.

[26] A. Jsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[27] A. P. Dempster, “A generalization of bayesian inference,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[28] Y. Pang, X. Zhao, L. Zhang, and H. Lu, “Multi-scale interactive network for salient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[29] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, “Label decoupling framework for salient object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 025–13 034.

[30] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, “Pixels, regions, and objects: Multiple enhancement for salient object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 031–10 040.

[31] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, “Towards diverse binary segmentation via a simple yet general gated network,” *International Journal of Computer Vision*, pp. 1–78, 2024.

[32] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1032–1042.

[33] F. Liu, Y. Liu, J. Lin, K. Xu, and R. W. Lau, “Multi-view dynamic reflection prior for video glass surface detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3594–3602.

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

[35] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, “Visual saliency transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4722–4732.

[36] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, “Feature shrinkage pyramid for camouflaged object detection with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5557–5566.

[37] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, “Frequency perception network for camouflaged object detection,” in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 1179–1189.

[38] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[40] T. Yan, J. Gao, K. Xu, X. Zhu, H. Huang, H. Li, B. Wah, and R. W. Lau, “Ghostingnet: A novel approach for glass surface detection with ghosting cues,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[41] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, “Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 406–416.

[42] 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 *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3455–3464.

[43] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, “Cross-modal fusion and progressive decoding network for rgb-d salient object detection,” *International Journal of Computer Vision*, pp. 1–19, 2024.

[44] X. Cheng, H. Xiong, D.-P. Fan, Y. Zhong, M. Harandi, T. Drummond, and Z. Ge, “Implicit motion handling for video camouflaged object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 864–13 873.

[45] L. Liu, J. Prost, L. Zhu, N. Papadakis, P. Liò, C.-B. Schönlieb, and A. I. Aviles-Rivero, “Scotch and soda: A transformer video shadow detection framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 449–10 458.

[46] M. Xu, J. Wu, Y. Lai, and Z. Ji, “Fusion of short-term and long-term attention for video mirror detection,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–9.

[47] H. Zhou, H. Wang, T. Ye, Z. Xing, J. Ma, P. Li, Q. Wang, and L. Zhu, “Timeline and boundary guided diffusion network for video shadow detection,” *arXiv preprint arXiv:2408.11785*, 2024.

[48] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, “Weakly-supervised salient object detection via scribble annotations,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 546–12 555.

[49] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, “Structure-consistent weakly supervised salient object detection with local saliency coherence,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3234–3242.

[50] Z. Liang, T. Wang, X. Zhang, J. Sun, and J. Shen, “Tree energy loss: Towards sparsely annotated semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 907–16 916.

[51] R. He, Q. Dong, J. Lin, and R. W. Lau, “Weakly-supervised camouflaged object detection with scribble annotations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 781–789.

[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pvt v2: Improved baselines with pyramid vision transformer,” *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[53] 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,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[54] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. Lau, “Don’t hit me! glass detection in real-world scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3687–3696.

[57] H. Mei, X. Yang, L. Yu, Q. Zhang, X. Wei, and R. W. Lau, “Large-field contextual feature learning for glass detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3329–3346, 2022.

[58] J. Wang, X. Li, and J. Yang, “Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[59] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to detect salient objects with image-level supervision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[60] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, “Concealed object detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6024–6042, 2021.

[61] L. Li, E. Rigall, J. Dong, and G. Chen, “Mas3k: An open dataset for marine animal segmentation,” in *International Symposium on Benchmarking, Measuring and Optimization*. Springer, 2020, pp. 194–212.

[62] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, “Orsi salient object detection via multiscale joint region and boundary model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[63] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, “Simultaneously localize, segment and rank the camouflaged objects,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 591–11 601.

[64] T. Yan, Z. Wan, X. Deng, P. Zhang, Y. Liu, and H. Lu, “Mas-sam: Segment any marine animal with aggregated features,” *arXiv preprint arXiv:2404.15700*, 2024.

[65] G. Li, Z. Bai, Z. Liu, X. Zhang, and H. Ling, “Salient object detection in optical remote sensing images driven by transformer,” *IEEE Transactions on Image Processing*, 2023.

[66] B. Yin, X. Zhang, D.-P. Fan, S. Jiao, M.-M. Cheng, L. Van Gool, and Q. Hou, “Camoformer: Masked separable attention for camouflaged object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.