Seeing Beyond Illusion: Generalized and Efficient Mirror Detection

Mingfeng Zha[1], Guoqing Wang[1]*, Tianyu Li[1], Wei Dong[2], Peng Wang[1], Yang Yang[1]

[1]University of Electronic Science and Technology of China
[2]Xi'an University of Architecture and Technology
zhamf1116@gmail.com

Abstract

Reflective imaging enables the mirror imagings and physical entities to possess identical attributes, *e.g.,* color and shape. Current mirror detection (MD) methods primarily rely on designing functional components to establish the correlation and disparities between the imagings and entities, thereby identifying the mirror regions. However, the exploration of extended scenes with dynamic content changes is rarely investigated. Therefore, we propose the MirrorSAM designed for MD based on the Segment Anything Model (SAM). Specifically, due to the varying reflections produced by mirrors in different positions and the complex visual space that interferes with localization, we design the hierarchical mixture of direction experts (HMDE) in the low-rank space to reduce biases towards entities in SAM and dynamically adjust experts based on the input scene. We observe differences in depth between mirrors and adjacent areas, and propose the depth token calibration (DTC), which introduces a learnable depth token to generate the depth map and serve as an error correction factor. We further formulate the selective pixel-prototype contrastive (SPPC) loss, selecting partially confusable samples to promote the decoupling of mirror and non-mirror representations. Extensive experiments conducted on four mirror benchmarks and two settings demonstrate that our approach surpasses state-of-the-art methods with few trainable parameters and FLOPs. We further extend to four transparent surface benchmarks to validate generalization.

Project — https://winter-flow.github.io/project/MirrorSAM

Introduction

Existing segmentation and detection works have made remarkable progress in physical entity perception, but lack exploration into virtual imaging. When objects with nearly identical attributes coexist in the same visual space without proper distinction, it can severely hinder downstream tasks, *e.g.,* path planning (Tang and Ma 2024) and 3D reconstruction (Liu et al. 2024). Mirror detection (MD) aims to differentiate between mirror and non-mirror regions, providing prior guidance for accurate scene understanding.

Existing MD methods typically rely on static designs, *e.g,* context comparison (Guan, Lin, and Lau 2022), mirror symmetry characteristics (He, Lin, and Lau 2023), or low-level visual differences (Xie et al. 2024) (*e.g.,* texture), which lack adaptability to dynamic scenarios. Besides, MD encounters several challenges: 1) Reflection interference, requiring strategies to handle complete/partial correspondences between entities and imagings; 2) Distinguishing mirrors from other smooth and reflective objects, *e.g.,* tiles and glasses; 3) Deformation and occlusion, where variations in mirror regions caused by shooting angles or obstructions make it difficult to rely on predefined shape priors for discrimination. A common approach to address these issues involves combining multiple specialized components in a certain manner to expand the representation space, posing computational complexity and coordination challenges among components. Furthermore, mirrors can reflect any entity, but existing datasets only capture a subset of possibilities, lacking adequate prior knowledge. Therefore, we propose the MirrorSAM based on SAM (Kirillov et al. 2023), which incorporates direction-aware expert groups for fine-tuning, depth calibration, and partial pixel-prototype contrastive learning. This raises three key questions: *1) Why introduce expert learning and direction-aware mechanism? 2) Why introduce depth calibration, and how does it differ from depth priors? 3) Why not use full-pixel contrastive learning?*

We answer the first question. SAM trained on numerous images containing physical entities, excels at capturing entity representations but remains insensitive to virtual imagings. Initially, we utilize LoRA (Hu et al. 2021) with minimal learnable parameters to fine-tune the image encoder for the MD task, yet the performance is not superior. To broaden the perceptual space, we introduce mixture of experts (MoE) (Jacobs et al. 1991) that dynamically select subsets of highly responsive experts based on input to uniformly choose solutions. Each expert is responsible for partial representations and scenes, collaborating with others to avoid conflicts and redundancies that arise from directly combining multiple components. This approach not only reduces computational complexity but also simplifies the structure of experts, minimizing internal uncertainties. In specific scenarios, the relationships between all corresponding entities and imagings exhibit a certain directionality. In other words, different entity-imaging regions can implicitly achieve consistency through directions acting as bridges. Unlike previous works (Huang et al. 2023; He, Lin, and Lau 2023) that establish symmetry by fixed-angle rotations, *e.g.,* 90 degrees or

*Corresponding author.

multiples, we propose a direction-aware mechanism based on coordinate systems and kernel learning. This allows for arbitrary-angle consistency, significantly expanding the potential solution space and enhancing flexibility.

We answer the second question. Based on the physical principle of planar mirror imaging (Born and Wolf 2013), the distances from an object to the mirror surface and its reflection are equal. Consequently, the depth within the mirror region is significantly greater than that of adjacent non-mirror regions, and accurate depth differences can effectively aid in localization. Unlike multimodal segmentation frameworks (Mei et al. 2021), which directly use depth maps as additional inputs and progressively fuse them with RGB features, we propose a more efficient approach. Specifically, we introduce an extra token to generate a depth map at the output end and utilize an error map for calibration. Considering: 1) In real-world scenarios, obtaining ⟨Depth, RGB⟩ pairs is often impractical; 2) Dual-stream networks incur high training and deployment costs. Our single-stream framework achieves high-quality predictions without requiring paired testing data.

We answer the third question. Pixel-to-pixel contrastive learning focuses on local regions and lacks contextual understanding, which may cause local optima and high computational complexity of $\mathcal{O}(N^2)$. Prototypes capture the global characteristics of mirror and non-mirror regions, providing robust references for pixel-level differentiation with a reduced complexity of $\mathcal{O}(N)$. Simple samples contribute little to training, and hard samples may introduce noises or outliers, thus excessive focus risks overfitting. To balance these issues, we prioritize semi-hard samples, which encourage the model to refine the attention on critical regions. Additionally, our approach restricts contrastive learning to within-sample, *i.e.,* non-generalized intra-batch contrasts.

Technically, we introduce the HMDE to bridge the representation gap between physical entities and mirrored imagings. The HMDE dynamically adjusts the structure based on scene context, enabling omni-perception. Furthermore, we propose the DTC, which differs from leveraging output tokens for generating detection maps. Instead, depth tokens are utilized for generating depth maps and guiding the model to focus on error-prone areas. To disentangle representations, we formulate the SPPC loss to emphasize pixel and prototype differences, especially for confusable samples. Extensive experiments conducted on eight benchmarks demonstrate the effect of the proposed approach.

In summary, our main contributions are as follows:

- We formulate the MirrorSAM based on expert learning and hierarchical modeling for perceiving mirror regions.
- We propose three customized components, the HMDE to automatically adapt to various scenes and imagings content changes, the DTC as the constraint to encourage focusing on depth error regions, refine prediction results, and the SPPC loss to selectively learn decoupled and compacted representations.
- Extensive experiments on four mirror and four transparent surface benchmarks validate the superiority of the proposed method and the effect of components.

Related Work

Mirror Detection. Mirrors reflect physical entities and create identical yet illusory imagings, which can seriously confuse and impact the understanding and modeling of visual space. For the fully-supervised setting, Yang *et al.* (Yang et al. 2019) pioneered the first MD method, MirrorNet, which leverages the correlation between internal and external features of mirrors. Lin *et al.* (Lin, Wang, and Lau 2020) designed the PMDNet, which compares mirror features with context for correspondence and incorporates edge information. Guan *et al.* (Guan, Lin, and Lau 2022) constructed semantic associations among objects based on graph representation. These methods share the common motivation, *i.e.,* establishing the associations between entities and imagings. Huang *et al.* (Huang et al. 2023) developed the dual-stream network based on Transformer to exploit the symmetry property of mirrors. He *et al.* (He, Lin, and Lau 2023) presented the HetNet, which combines low-level and high-level features in a heterogeneous manner. Both aim to utilize rotation strategies to construct mirror symmetry consistency. Other works (Mei et al. 2021; Tan et al. 2022; Zha et al. 2024a; Xie et al. 2024) leveraged the structure differences between mirror and non-mirror regions to distinguish, *i.e.,* depth, content distribution, and frequency. For the data-efficient setting, Zha *et al.* (Zha et al. 2024b) introduced the first weakly supervised dataset and model based on scribble annotations, *i.e.,* WSMD, achieving performance comparable to fully supervised methods. Lin *et al.* (Lin and Lau 2023) formulated the self-supervised pre-training strategy. For the video-level setting, Lin *et al.* (Lin, Tan, and Lau 2023) proposed the first video-level dataset and model, *i,e,* VMD. Warren *et al.* (Warren et al. 2024) improved based on inconsistent motion clues. Xu *et al.* (Xu, Siu, and Lau 2024) focused on extremely-weak supervision.

SAM for Downstream Tasks. (Chen et al. 2023) introduced several adapters containing two-layer MLPs. (Yu et al. 2024) integrated traditional segmentation frameworks with SAM and directly encoded depth maps as cues. (Zhang et al. 2024) designed a dual-stream multi-scale guidance architecture for underwater animal scenes. However, MD-specific SAM has been scarcely explored. Our MirrorSAM explores and proves the potential.

Proposed Method

Overall Architecture

As illustrated in Figure 1, the MirrorSAM framework builds upon the basic structure of SAM. The HMDE is integrated into the image encoder, and the DTC is incorporated into the pixel decoder. Using the last layer features of the decoder and the predicted maps, we generate prototypes and compute the SPPC loss. Additionally, the primary loss function is adapted based on the specific supervision setting.

Hierarchical Mixture of Direction Experts

In previous works (Zhang et al. 2024; Cheng et al. 2024), LoRA is directly applied to the image encoder to modulate features for specific tasks. Considering the complexity and ambiguity of the MD visual space, we introduce

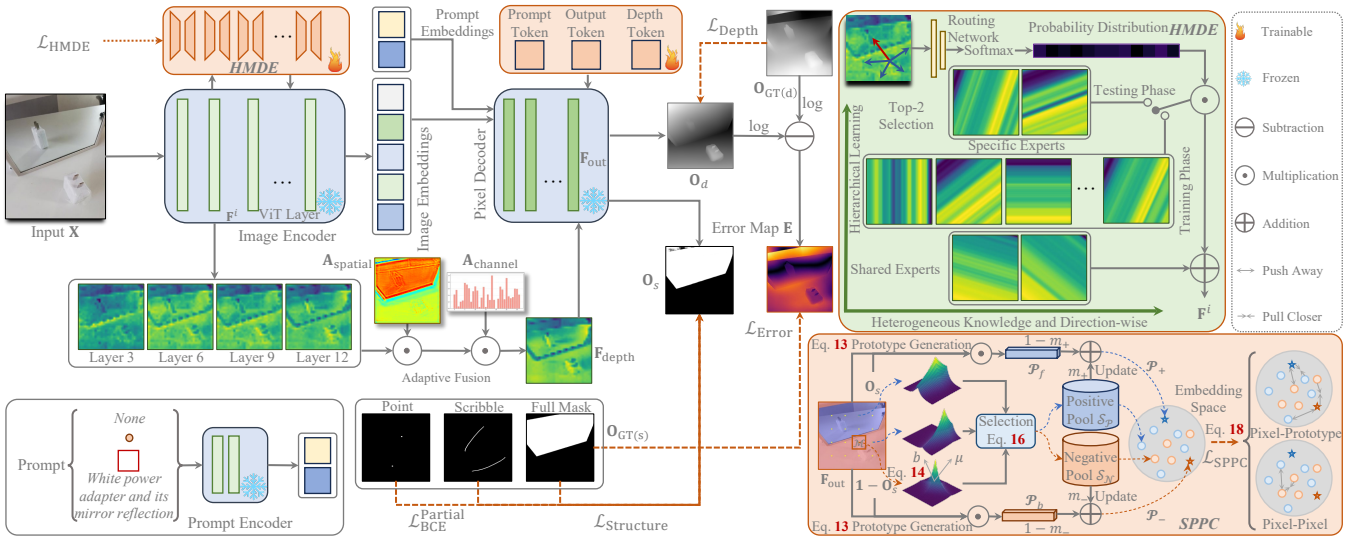

Figure 1: The framework of our MirrorSAM. Given any input image, we utilize the image encoder to extract features, where the HMDE modulates features at each stage. We adaptively fuse features from multiple scales. Subsequently, the depth token, prompt token, and output token are fed to the decoder to obtain depth estimation maps and prediction maps. For the prediction maps, we apply the SPPC loss as a constraint. For the depth maps, we apply two losses, one for supervision and the other for calibration. When weak (scribble or point) supervision is applied, we only adjust the output side, *i.e.,* loss function.

expert learning and directional correlation. We expand expert groups within each LoRA to increase representation capacity and specialization, enabling better handling of diverse scenes and imaging content variations. When imagings and corresponding entities align perfectly, we can establish global semantic awareness based on directional consistency. For partial matches, directional signals are learned from paired samples and propagated to unmatched ones to achieve complementarity.

Specifically, given input image $\mathbf{X} \in \mathbb{R}^{3\times H_0\times W_0}$, we utilize the encoder to obtain multi-scale features $\mathbf{F}^i \in \mathbb{R}^{C_i\times H_i\times W_i}$ (channel, height, and width) and optimize parameters through the HMDE. We define the angle θ of each directional convolution kernel $\mathbf{K}_\theta(m,n) \in \mathbb{R}^{C_i\times k_h\times k_w}$, and calculate the transformed direction by applying the rotation matrix to the standard convolution $\mathbf{K}(m',n')$,

$$\begin{bmatrix} m' \\ n' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} m-c_m \\ n-c_n \end{bmatrix} \tag{1}$$

where m' and n' are the rotated coordinates, c_m and c_n are the coordinates of the center point of the kernel.

We then obtain feature updates specific to the direction,

$$\mathbf{F}^i_\theta(c',p,q) := \sum_{c=1}^{C_i}\sum_{m=0}^{k_h-1}\sum_{n=0}^{k_w-1} \mathbf{F}^i(c,p+m,q+n)\cdot\mathbf{K}_\theta(c',c,m,n) \tag{2}$$

where (p,q) represents the pixel position. Based on the directional kernel, we further extend to the MoE paradigm. MoE comprises two core elements: the routing network $\mathcal{G}$ and the experts $\mathcal{E}$. The routing network assigns weights to the experts, while the experts handle specific representations. Intuitively, imagings and entities features could be allocated to different experts, enabling heterogeneous processing and decoupling. For $\mathbf{F}^i$, we have,

$$\mathcal{G}(\mathbf{F}^i) = \texttt{Softmax}\left(\mathcal{F}(\texttt{GAP}(\mathbf{F}^i)) + \mathbb{N}\right) \tag{3}$$

where $\mathcal{F}$, GAP, and $\mathbb{N}$ denote convolution, global average pooling, and Gaussian distribution noise, respectively. Noise is used to mitigate biases and is only applied during the training phase, being removed during the testing phase. We set learnable θ for each expert and integrate features from several directions. However, assigning weights and activations to all experts by $\mathcal{G}$ may weaken communication between the experts and be inefficient. We divide the expert group into shared and specific experts. The former perceive general reflection imaging and associated directions, while the latter, anchored to the former, construct consistency and customized constraints,

$$\mathbf{F}^i := \sum_{k=1}^{N_s}\mathcal{E}^k(\mathbf{F}^i) + \sum_{k=N_s+1}^{N_e}\mathcal{G}^k(\mathbf{F}^i)\cdot\mathcal{E}^k(\mathbf{F}^i;\theta^k) + \mathbf{F}^i \tag{4}$$

where N_s and N_e represent the number of shared experts and total experts, respectively. During the testing phase, to ensure high efficiency, we activate the Top-2 weighted experts and ignore the remaining. Compared to LoRA, despite the additional learning cost introduced, the performance improvement is significant.

Ideally, the workload of all experts is $\frac{1}{N_e}$. To avoid overload of partial experts and idleness of others, we reduce the excessive reliance of the gate network on input based on mutual information (MI) (Kraskov, Stögbauer, and Grassberger 2004) $\mathcal{I}$,

$$\mathcal{I}(\mathbf{F};\mathcal{G}(\mathbf{F})) = \mathcal{H}(\mathcal{G}(\mathbf{F})) - \mathcal{H}(\mathcal{G}(\mathbf{F})|\mathbf{F}) \tag{5}$$

where a smaller $\mathcal{I}$ indicates a more balanced load, and the former and latter represent the entropy

and conditional entropy of expert assignments, respectively. Our objective is to minimize $\mathcal{I}$. For $\mathcal{G}(\mathbf{x}^n) = [g^1(\mathbf{x}^n), g^2(\mathbf{x}^n), \ldots, g^{N_e-N_s}(\mathbf{x}^n)]$, we let $\sum_{k=N_s+1}^{N_e} g^k(\mathbf{x}^n) = 1$, where $g^k(\mathbf{x}^n)$ represents the probability of assigning the *n-th* sample in the batch B to the *k-th* expert. We reformulate Eq. 5 as,

$$\begin{aligned}\mathcal{I}(\mathbf{F};\mathcal{G}(\mathbf{F})) = &- \sum_{k=N_s+1}^{N_e} \mathbb{E}_{\mathbf{x}}[g^k(\mathbf{x})] \log \mathbb{E}_{\mathbf{x}}[g^k(\mathbf{x})] \\ &+ \frac{1}{B}\sum_{n=1}^{B}\sum_{k=N_s+1}^{N_e} g^k(\mathbf{x}^n)\log g^k(\mathbf{x}^n)\end{aligned} \tag{6}$$

To enable experts to learn heterogeneous knowledge, we leverage game theory (Fudenberg 1991) to cultivate the *competitive relationship* among experts, where each expert aims to maximize the uniqueness of their own representations, *i.e.,* mutual opposition. We implement by $\mathcal{L}_{\text{comp}}$,

$$\mathcal{L}_{\text{comp}} = \sum\nolimits_{i=1}^{N_e}\sum\nolimits_{j=i+1}^{N_e} \mathbb{E}_{\mathbf{x}}\left[\texttt{sim}(\mathcal{E}^i(\mathbf{x}), \mathcal{E}^j(\mathbf{x}))\right] \tag{7}$$

where $\texttt{sim}$ is the cosine similarity. The constraint loss $\mathcal{L}_{\text{HMDE}}$ for the HMDE can be defined as,

$$\mathcal{L}_{\text{HMDE}} = \lambda_{\text{MI}}\mathcal{I} + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}} \tag{8}$$

where λ_{MI} and λ_{comp} are balance parameters.

Depth Token Calibration

Modeling the spatial geometry of the scenes through depth differences estimation facilitates discrimination. The most direct approach is to incorporate paired depth maps or those estimated by pretrained models as the prior and explicitly inject the RGB modality to form the multi-branch architecture for input (or output). Considering error accumulation and computational costs, we introduce a depth token for depth map generation, formulating two heads for output-level calibration. Unlike segmentation branches that encode only the semantic-rich features of the encoder's final layer for decoding, we select features at multiple scales to provide rich details and semantics for geometry perception.

Technically, due to differences in high-level and low-level representations, we choose L scales (in practice, $L = \{3, 6, 9, 12\}$) and generate spatial and channel weight distributions, *i.e.,* $\mathbf{A}_{\text{spatial}}$ and $\mathbf{A}_{\text{channel}}$ by,

$$\mathbf{A}_{\text{spatial}}, \mathbf{A}_{\text{channel}} = \sigma(\mathcal{F}(\sum\nolimits_{l=1}^{L}\mathcal{F}(\mathbf{F}^l))) \tag{9}$$

where σ denotes the sigmoid function. Note that the convolution parameter settings for generating $\mathbf{A}_{\text{spatial}}$ and $\mathbf{A}_{\text{channel}}$ are different. We further partition and allocate the overall weights to each sub-item for adaptive fusion,

$$\mathbf{F}_{\text{depth}} = \sum\nolimits_{l=1}^{L}\mathbf{F}^l \cdot \mathbf{A}^l_{\text{spatial}} \cdot \mathbf{A}^l_{\text{channel}} \tag{10}$$

We fuse the learnable depth token and features $\mathbf{F}_{\text{out}}$ to derive estimated depth maps $\mathbf{O}_d$. We further utilize the depth loss $\mathcal{L}_{\text{Depth}}$ (*i.e.,* Huber loss, robust to noises) to optimize and generate error maps $\mathbf{E}$,

$$\mathbf{E} = \| \log \mathbf{O}_d - \log \mathbf{O}_{\text{GT(d)}} \| \tag{11}$$

where $\|\cdot\|$ denotes Min-Max normalization. To enhance the contrast of neighboring regions while ensuring convergence stability, we leverage a logarithmic transformation. Directly applying noise-carrying $\mathbf{E}$ to the segmentation maps $\mathbf{O}_{\text{GT(s)}}$ may disrupt the correct regions. We formulate $\mathcal{L}_{\text{Error}}$ to promote focusing on error regions through implicit calibration,

$$\mathcal{L}_{\text{Error}} = \frac{-\sum_{i=1}^{H_0}\sum_{j=1}^{W_0}\mathbf{E}^{ij} \times \mathbf{O}^{ij}_{\text{GT(s)}} \log(\mathbf{O}^{ij}_s)}{\sum_{i=1}^{H_0}\sum_{j=1}^{W_0}\mathbf{E}^{ij} + \epsilon} \tag{12}$$

where ϵ is an extremely small value.

Selective Pixel-Prototype Contrastive Loss

Unlike prior works that establish robust semantic knowledge by contrasting across images, considering that: 1) Due to reflective imaging, even the same mirror placed in different scenes or at different positions within the same scene, or with adjusted spatial relationships of entities, presents different contents, indicating that mirrors are highly scene-dependent and sensitive to spatial arrangements; 2) The limited scale and diversity of existing mirror datasets make it challenging to learn universal representations; 3) Mirror and non-mirror regions within individual samples exhibit certain distributional distinctiveness. Considering the above, we formulate the hierarchical aggregation from pixel-to-prototype and pixel-to-pixel at the intra-sample level. The similarity of content inside and outside the mirror results cause pixels exhibiting ambiguity, thus we aim to select samples that are easily confused for particular attention.

In detail, we leverage the final layer feature $\mathbf{F}_{\text{out}}$ of the decoder along with the foreground and background segmentation map $\mathbf{O}_s, 1-\mathbf{O}_s$ to generate foreground and background prototypes, *i.e.,* $\mathcal{P}_f$ and $\mathcal{P}_b$ as base reference,

$$\mathcal{P}_f = \frac{\sum_{i,j}\mathbf{O}^{ij}_s \cdot \mathbf{F}^{ij}_{\text{out}}}{\sum_{i,j}\mathbf{O}^{ij}_s},\ \mathcal{P}_b = \frac{\sum_{i,j}(1-\mathbf{O}^{ij}_s) \cdot \mathbf{F}^{ij}_{\text{out}}}{\sum_{i,j}(1-\mathbf{O}^{ij}_s)} \tag{13}$$

For the pixel feature $\mathbf{f}^{ij}$ at position (i, j), we establish its Laplacian distribution to accelerate computation and improve noise resistance (compared to the Gaussian distribution),

$$\mu^{ij} = \mathbf{f}^{ij},\ b^{ij} = \frac{1}{|\mathcal{M}(i,j)|}\sum_{(p,q)\in\mathcal{M}(i,j)}|\mathbf{f}^{pq} - \mu^{ij}| \tag{14}$$

where μ and b represent the location and scale parameters, respectively, $|\mathcal{M}(i,j)|$ represents the number of neighboring pixels (*e.g.,* 3×3) for the pixel $\mathbf{P}^{ij}$ (*i.e.,* dynamic region generation, as opposed to fixed partitioning at the patch level). We further calculate the distance between pixel distributions, *e.g,* negative pairs,

$$\mathcal{D}(\mathbf{P}^{ij}, \mathbf{P}^{kl}) = \| \sum_{c=1}^{C}\left(\log\frac{b^{kl,c}}{b^{ij,c}} + \frac{|\mathbf{f}^{ij,c} - \mathbf{f}^{kl,c}|}{b^{kl,c}}\right) \| \tag{15}$$

For each anchor pixel, confusable samples are selected according to the rule,

$$\mathcal{S}_{\mathcal{N}} = \{\mathbf{P}^{kl} \mid \mathcal{D}^{mn} < \mathcal{D}^{kl} < (1+\alpha)\cdot\mathcal{D}^{mn}\} \tag{16}$$

where $\mathcal{D}^{mn} = \mathcal{D}(\mathbf{P}^{ij}, \mathbf{P}^{mn}), \mathcal{D}^{kl} = \mathcal{D}(\mathbf{P}^{ij}, \mathbf{P}^{kl})$, $\mathbf{P}^{mn}$ represent positive sample pixels, and α is the interval parameter. The inherent uncertainty within pixels is difficult to measure through distance. We aim to assign higher weights

to high-uncertainty pixels. We define the differential entropy $\mathcal{H}_{\mathcal{M}}$ of region $\mathcal{M}$ as,

$$\mathcal{H}_{\mathcal{M}} = 1 + \ln(2b) \tag{17}$$

Thus, the pixel weight w for $\mathcal{S}$ is generated by normalizing $\mathcal{H}$. Similarly, we can obtain the semi-hard positive sample pool $\mathcal{S}_{\mathcal{P}}$. We derive the SPPC loss $\mathcal{L}_{\text{SPPC}}$,

$$\mathcal{L}_{\text{SPPC}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{|\mathcal{S}_{\mathcal{P}}|}\log\left[w^k \cdot \texttt{sim}'(\mathcal{S}_{\mathcal{P}}^{k(i)}, \mathcal{P}_{+}^{(i)}) \frac{1}{\texttt{sim}'(\mathcal{S}_{\mathcal{P}}^{k(i)}, \mathcal{P}_{-}^{(i)}) + \sum_{j=1}^{|\mathcal{S}_{\mathcal{N}}|} w^j \cdot \texttt{sim}'(\mathcal{S}_{\mathcal{P}}^{k(i)}, \mathcal{S}_{\mathcal{N}}^{j(i)})}\right] \tag{18}$$

where $\texttt{sim}' = \exp(\texttt{sim})/\tau$, τ is a temperature coefficient. Note that $\mathcal{P}_+$ and $\mathcal{P}_-$ (not $\mathcal{P}_f$ and $\mathcal{P}_b$) represent the positive and negative prototypes, updated from $\mathcal{S}_{\mathcal{P}}^K$ and $\mathcal{S}_{\mathcal{N}}^K$ based on momentum m. For the fully-supervised setting, following (He, Lin, and Lau 2023), we leverage $\mathcal{L}_{\text{Structure}}$ as the primary loss,

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Structure}} + \sum_{\mathcal{L}\in\{\mathcal{L}_{\text{SPPC}}, \mathcal{L}_{\text{Depth}}, \mathcal{L}_{\text{Error}}, \mathcal{L}_{\text{HMDE}}\}} \lambda_{\mathcal{L}}\mathcal{L} \tag{19}$$

where $\lambda_{\mathcal{L}}$ denotes weight parameter. For weakly-supervised setting, following (Zha et al. 2024b), we switch the primary loss to $\mathcal{L}_{\text{BCE}}^{\text{Partial}}$. And we adjust $\mathcal{L}_{\text{SPPC}}$ to $\mathcal{L}_{\text{SPPC}}^{\text{Partial}}$.

Experiments

Datasets. We conduct experiments on eight datasets. **Mirror datasets:** MSD (Yang et al. 2019) and PMD (Lin, Wang, and Lau 2020) datasets contain 3,063 and 5,096 training images, 955 and 571 testing images, respectively. MirrorD (Mei et al. 2021) contains 2,000 training images and 1,049 testing images, and is accompanied by depth maps. VMD (Lin, Tan, and Lau 2023) has 143 (7,835 images) and 126 (7,152 images) videos for training and testing. **Transparent surface datasets:** GDD (Mei et al. 2020) contains 2,980 training images and 936 testing images. GSD (Lin, He, and Lau 2021) consists of 3,202 training pairs and 810 testing pairs, with a larger coverage of regions and contrast distributions. GlassD (Lin, Yeung, and Lau 2022) comprises 2,400 training images and 609 testing images with paired depth maps. Trans10K (Xie et al. 2020) consists of 10,428 images with three categories: things, stuff, and background. Images are divided into 5,000, 1,000, and 4,428 images for training, validation, and testing, respectively.

Implementation Details. We implement the model and conduct experiments on an A100 GPU via Pytorch. To reduce memory consumption and align with the settings in (Xie et al. 2024; Huang et al. 2023), we adjust the input size from the default 1024×1024 to 512×512 and initialize the parameters with pre-trained SAM-B. We utilize the AdamW optimizer to update the parameters, with the initial learning rate and the weight decay both set to 0.001. The batch size is 16, and the total epochs is 200. During the training phase, we freeze the base components, *e.g.,* image encoder, and optimize the DTC and the HMDE. During the testing phase, we do not employ any post-processing operations, *e.g.,* CRF. If depth labels are not provided, we leverage (Yang et al. 2024b) for generation.

Evaluation Metrics. We adopt seven evaluation metrics: S-measure (S_m) (Fan et al. 2017), mean E-measure (E_m) (Fan et al. 2018), weighted F-measure, (F_β^w), maximum F-measure(F_β) (Margolin, Zelnik-Manor, and Tal 2014), Intersection over union (IoU), Mean Absolute Error (MAE), and Balance Error Rate (BER). Note that the higher the better for the first five. To evaluate efficiency, we leverage model trainable parameters and computational complexity.

Comparison with SOTA Methods

Quantitative Comparison. In Table 1, our approach outperforms various types of methods, *e.g.,* DualSAM, with gains of 1.4%, 3.3%, 2.6%, 4.2%, and 6.5% respectively on the five metrics of the MSD dataset, and the average surpasses CSFW by around 5.0%, which demonstrates the effectiveness of expert adaptation and differential learning. In Table 3, our method achieves comparable performance without complex modality fusion and temporal signal. In Table 2, the trainable parameters and FLOPs of MirrorSAM are about one-fifteenth and one-thirty-fifth of CSFW, respectively, highlighting the efficiency.

Qualitative Comparison. We consider different scenarios and provide visualizations for comparison. In Figure 2, the first line illustrates smooth surface interference, *i.e.,* tile. The second and third rows depict incomplete correspondence and irregular occlusion. The fourth row pertains to situations with only imagings and no corresponding physical entities. The last row demonstrates complex multi-object perception. Our approach effectively handles intricate visual spatial variations and reflection interferences.

Ablation Study and Discussion

We validate the effect of the proposed components (Table 4, Figure 3) and crucial hyperparameter settings (Figure 4) on the MSD and PMD datasets.

Expert Modeling. Directly applying the vanilla SAM to MD yields unsatisfactory results, but incorporating LoRA fine-tuning (Table 4 Line 9) significantly improves performance, highlighting the necessity of domain knowledge adaptation. Introducing the HMDE to establish dynamic perceptual scene changes and perceive visual spatial content and geometric relationships in arbitrary directions further enhances the effect, with the two elements (MoE and angle) proving complementary. We further boost through $\mathcal{L}_{\text{HMDE}}$-controlled diverse experts.

Depth Injection. We explore three strategies: 1) Fusing depth map and RGB image as input; 2) Using the depth map as the dense prompt during the training phase; 3) Employing for the full phase, yet the performance is inferior to calibrating at the output end. We attribute this to the noise and inaccuracies in depth maps generated by pre-trained depth estimation models (or low-quality paired), which can degrade original representations when explicitly encoded and interacted with. In contrast, implicit depth learning through gradient optimization facilitates calibration and enhances feature quality. Moreover, not using depth maps during the testing phase is more aligned with practical deployment.

Contrast Settings. Pixel-, prototype-, and pixel-prototype-level contrastive learning are all inferior to the SPPC. We analyze: 1) The insufficient number of prototypes, even for the

Methods	Attr.	Transformer Encoder	Foundation Model	Prompt	MSD					PMD				
					MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
UDUN (Pei et al. 2023) MM'23	S	✗	✗	✗	0.071	0.815	0.838	0.746	0.713	0.039	0.784	0.794	0.652	0.600
MENet (Wang et al. 2023) CVPR'23	S	✗	✗	✗	0.054	0.868	0.906	0.829	0.805	0.033	0.826	0.873	0.727	0.680
MirrorNet (Yang et al. 2019) ICCV'19	M	✗	✗	✗	0.065	0.850	0.891	0.812	0.790	0.043	0.761	0.841	0.663	0.585
PMDNet (Lin, Wang, and Lau 2020) CVPR'20	M	✗	✗	✗	0.047	0.875	0.908	0.845	0.815	0.032	0.810	0.859	0.716	0.660
SANet (Guan, Lin, and Lau 2022) CVPR'22	M	✗	✗	✗	0.054	0.862	0.898	0.829	0.798	0.071	0.808	0.839	0.721	0.668
VCNet (Tan et al. 2022) TPAMI'22	M	✗	✗	✗	0.044	‡	‡	‡	0.854	0.028	‡	‡	‡	0.694
HetNet (He, Lin, and Lau 2023) AAAI'23	M	✗	✗	✗	0.043	0.881	0.921	0.854	0.824	0.029	0.828	0.865	0.734	0.690
SETR (Zheng et al. 2021) CVPR'21	S	✓	✗	✗	0.071	0.797	0.840	0.750	0.690	0.035	0.753	0.775	0.633	0.564
VST (Liu et al. 2021) ICCV'21	S	✓	✗	✗	0.054	0.861	0.901	0.818	0.791	0.036	0.783	0.814	0.639	0.591
DSAM (Yu et al. 2024) MM'24	C	✓	✓	✓	0.037	0.888	0.919	0.871	0.832	0.031	0.800	0.839	0.745	0.666
VSCode (Luo et al. 2024) CVPR'24	C	✓	✗	✓	0.077	0.800	0.820	0.721	0.687	0.042	0.787	0.816	0.656	0.607
FSEL (Sun et al. 2025) ECCV'24	C	✓	✗	✗	0.043	0.868	0.918	0.859	0.814	0.038	0.803	0.844	0.737	0.671
SAM (Kirillov et al. 2023) ICCV'23	G	✓	✓	✗	0.108	0.755	0.768	0.689	0.624	0.063	0.688	0.711	0.597	0.616
EVP (Liu et al. 2023) CVPR'23	G	✓	✗	✓	0.064	0.845	0.896	0.811	0.780	0.037	0.793	0.861	0.694	0.634
LISA (Lai et al. 2024) CVPR'24	G	✓	✓	✓	0.061	0.822	0.859	0.799	0.771	0.044	0.811	0.825	0.678	0.660
DualSAM (Zhang et al. 2024) CVPR'24	O	✓	✓	✓	0.039	0.903	0.932	0.882	0.848	0.034	0.816	0.839	0.705	0.636
HSAM (Cheng et al. 2024) CVPR'24	O	✓	✓	✓	0.042	0.878	0.909	0.865	0.833	0.032	0.808	0.833	0.721	0.619
SATNet (Huang et al. 2023) AAAI'23	M	✓	✗	✗	0.033	0.887	0.916	0.865	0.834	0.025	0.826	0.858	0.739	0.684
CSFW(Xie et al. 2024) TIP'24	M	✓	✗	✗	0.045	0.875	0.905	0.846	0.821	0.024	0.831	0.864	0.756	0.700
Ours	M	✓	✓	✓	**0.025**	**0.936**	**0.958**	**0.924**	**0.913**	**0.024**	**0.867**	**0.907**	**0.788**	**0.759**

Table 1: Quantitative comparison on MSD and PMD datasets. S, C, G, O, M denote salience detection, camouflage detection, general segmentation, other related tasks segmentation, and MD methods, respectively. Foundation models include *e.g.,* LLMs and SAM. Best performance in **bold**, second in underline. ‡ represents data is unavailable. ↑ indicates higher values are better, while ↓ indicates the opposite.

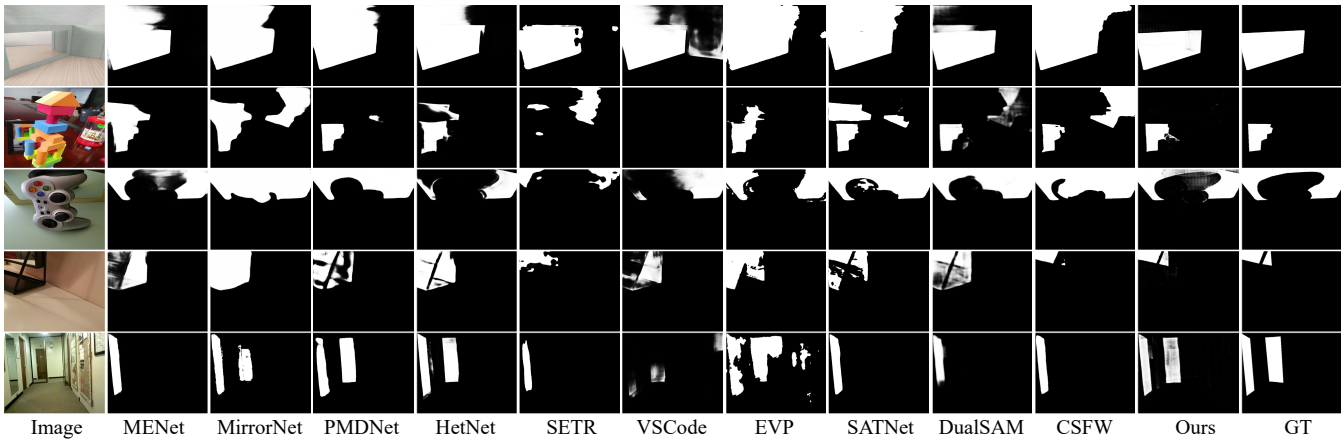

Figure 2: Qualitative comparison on MD scenarios. Best viewed by zooming in.

	SATNet	CSFW	DualSAM	WSMD	Ours
FLOPs↓ (GMAC)	153.00	139.45	39.51	21.39	4.21
Params.↓ (M)	139.36	150.54	74.30	26.16	10.23

Table 2: Quantitative comparison of model efficiency.

Methods	Depth	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
VSCode (Luo et al. 2024)	✓	0.045	0.839	0.892	0.801	0.763
Ours	✓	**0.033**	**0.879**	**0.925**	**0.840**	**0.820**
Methods	Video	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
MGVMD (Warren et al. 2024)	✓	0.103	0.729	0.738	0.627	0.569
Ours	✗	**0.099**	**0.761**	**0.775**	**0.655**	**0.600**

Table 3: Comparison on MirrorD and VMD datasets.

entire batch, is not enough for contrast; 2) Pixels are easily affected by noise interference; 3) Pixel-prototype treats all samples equally. The selection and discrimination of semi-hard positive/negative samples is crucial.

SAM Prompt Strategies. In the baseline, prompts are automatically generated. We further explore: 1) Randomly shifting several pixels based on GT to generate points and boxes prompt; 2) Leveraging Qwen (Yang et al. 2024a) to generate image descriptions, which are then encoded into embedding vectors using CLIP (Radford et al. 2021) and merged with visual features. We observe that box prompts yield better performance, possibly due to providing more visual priors and avoiding semantic conflicts with text representations.

Crucial Hyperparameters. We observe that performance gradually improves as $1 < N_e \leq 7$, but decreases when $7 < N_e$. We analyze that while each expert specializes in part of the data space, excessive experts may cause conflicts and redundancies. When $10\% < R_s \leq 60\%$, performance gradually improves, but declines when $60\% < R_s$. This in-

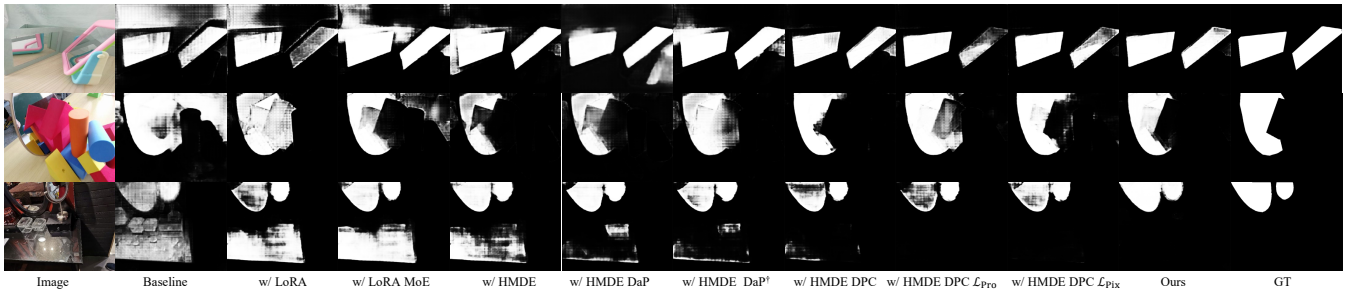

Figure 3: Qualitative ablation of proposed components. From top to bottom, the scenes include cross-imaging, complex correspondence and occlusion, and smooth tabletop reflection interference. Partial variants correspond to Table 4.

Components/Strategies			MSD					PMD				
			MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
HMDE	DPC	SPPC	0.108	0.755	0.768	0.689	0.624	0.063	0.688	0.711	0.597	0.616
✓	✗	✗	0.035	0.878	0.897	0.855	0.836	0.038	0.801	0.829	0.713	0.688
✓	✓	✗	0.028	0.908	0.933	0.894	0.889	0.029	0.839	0.872	0.756	0.723
✓	✗	✓	0.032	0.895	0.918	0.888	0.875	0.032	0.848	0.861	0.745	0.736
✗	✓	✓	0.055	0.803	0.811	0.744	0.685	0.051	0.746	0.766	0.651	0.645
✓	✓	✓	**0.025**	**0.936**	**0.958**	**0.924**	**0.913**	**0.024**	**0.867**	**0.907**	**0.788**	**0.759**
MoE	Angle	$\mathcal{L}_{\text{HMDE}}$	0.043	0.846	0.856	0.801	0.783	0.044	0.773	0.785	0.678	0.661
✓	✗	✗	0.039	0.857	0.874	0.824	0.801	0.042	0.784	0.796	0.689	0.672
✓	✓	✗	0.037	0.868	0.885	0.839	0.820	0.039	0.792	0.813	0.695	0.680
✓	✓	✓	**0.035**	**0.878**	**0.897**	**0.855**	**0.836**	**0.038**	**0.801**	**0.829**	**0.713**	**0.688**
DaI	DaP	DaP†	0.035	0.878	0.897	0.855	0.836	0.038	0.801	0.829	0.713	0.688
✓	✗	✗	0.033	0.885	0.883	0.866	0.848	0.035	0.815	0.838	0.725	0.695
✗	✓	✗	0.033	0.888	**0.911**	0.869	**0.860**	0.034	0.812	0.841	**0.735**	0.690
✗	✗	✓	**0.030**	**0.895**	0.904	**0.875**	0.855	**0.032**	**0.820**	**0.847**	0.729	**0.699**
$\mathcal{L}_{\text{Pro}}$	$\mathcal{L}_{\text{Pix}}$	$\mathcal{L}_{\text{Pix}\leftrightarrow\text{Pro}}$	0.028	0.908	0.933	0.894	0.889	0.029	0.839	0.872	0.756	0.723
✓	✗	✗	0.030	0.913	0.938	0.899	0.895	0.029	0.835	0.884	0.767	0.731
✗	✓	✗	0.028	0.915	**0.945**	0.906	**0.902**	0.027	**0.845**	0.880	0.769	**0.738**
✗	✗	✓	**0.027**	**0.922**	0.940	**0.908**	0.899	**0.026**	0.841	**0.888**	**0.773**	0.730
Ponit	Box	Text	0.025	0.936	0.958	0.924	0.913	0.024	0.867	0.907	0.788	0.759
✓	✗	✗	0.023	0.942	0.955	0.932	0.918	0.022	0.875	0.916	0.797	0.769
✗	✓	✗	**0.021**	**0.945**	0.961	**0.944**	**0.928**	**0.022**	**0.896**	**0.935**	**0.821**	**0.786**
✗	✗	✓	0.022	0.941	**0.964**	0.940	0.925	0.022	0.888	0.928	0.808	0.780

Table 4: Quantitative ablation of components and strategies. DaI, DaP, DaP†, $\mathcal{L}_{\text{Pro}}$, $\mathcal{L}_{\text{Pix}}$, and $\mathcal{L}_{\text{Pix}\leftrightarrow\text{Pro}}$ denote depth as input, depth as prompt (training only), depth as prompt (full phase), prototype-to-prototype, pixel-to-pixel, and pixel-to-prototype contrastive learning, respectively.

dicates that the benefit from simple samples is limited, and increasing proportion of confusable samples can enhance the discriminative ability. However, too many hard samples may cause matching errors and degrade performance.

Weakly-supervised Setting. In Table 5, we evaluate under the weakly supervised setting based on the scribble and point annotations provided by (Zha et al. 2025c) on the MSD and PMD datasets. Scribbles and points provide the basic structure and location of the foreground and background, respectively. Our method is superior and exhibits a larger gap under the point setting. Beyond the critical components, we attribute to: 1) Different from WSMD, which adopts Canny(Canny 1986)-generated edge maps as supervision, SAM provides more accurate and fine-grained knowledge priors. 2) Unlike WSMD, which establishes prototype contrasts between images and prototype enhancements in PFA, we consider the association of confusable pixels and dynamic prototypes within a single sample. We argue that the representations of mirror and non-mirror regions across images do not possess strong semantic correlations; thus, enforcing alignment may disrupt the original features.

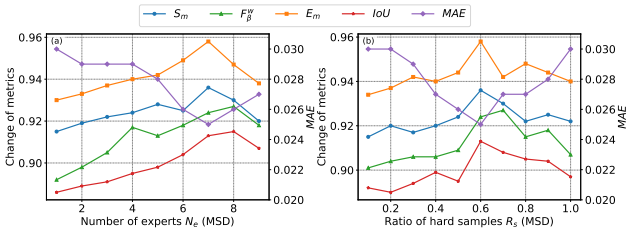

Figure 4: Quantitative ablation of hyperparameter settings.

Methods	Scribble-MSD					Scribble-PMD				
	MAE	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE	S_m↑	E_m↑	F_β^w↑	IoU↑
PFA (Chan et al. 2024)	0.075	0.815	0.886	0.771	0.762	0.057	0.760	0.805	0.642	0.585
WSMD	0.078	0.828	0.878	0.780	0.750	**0.051**	0.773	**0.824**	0.630	0.600
Ours	**0.070**	**0.843**	**0.903**	**0.799**	**0.778**	0.053	**0.784**	0.816	**0.665**	**0.618**
Methods	Point-MSD					Point-PMD				
	MAE	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE	S_m↑	E_m↑	F_β^w↑	IoU↑
PFA (Chan et al. 2024)	0.155	0.673	0.755	0.573	0.573	0.095	0.689	0.751	0.563	0.486
WSMD	0.168	0.696	0.742	0.586	0.560	0.091	0.706	0.760	0.550	0.501
Ours	**0.147**	**0.719**	**0.768**	**0.613**	**0.584**	**0.088**	**0.716**	**0.777**	**0.583**	**0.519**

Table 5: Weakly-supervised setting.

Methods	GDD				GSD			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
RFENet (Fan et al. 2023)	0.874	0.929	0.062	5.79	0.836	0.904	0.049	6.24
CMCM (Lin et al. 2025)	0.883	0.933	0.059	5.65	0.849	0.912	0.050	6.02
Ours	**0.887**	**0.945**	**0.051**	**5.08**	**0.871**	**0.927**	**0.048**	**5.55**
Methods	GlassD				Trans10K			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
RFENet (Fan et al. 2023)	0.699	0.825	0.046	11.42	**0.912**	‡	0.043	3.68
CMCM (Lin et al. 2025)	0.742	0.853	0.043	9.33	0.899	0.878	0.046	3.55
Ours	**0.764**	**0.875**	**0.038**	**8.28**	0.903	**0.895**	**0.042**	**3.13**

Table 6: Comparison on transparent surface datasets.

Broader Impacts. In Table 6, we achieve promising performance as well. We attribute depth disparity and reliance on orientation modeling. And we will explore more scenes (Zha et al. 2025b), and multimodal learning (Zha et al. 2025a).

Acknowledgments

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

References

Born, M.; and Wolf, E. 2013. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chan, G.; Zhang, P.; Dong, H.; Ji, S.; and Chen, B. 2024. Scribble-Supervised Semantic Segmentation with Prototype-based Feature Augmentation. In *Forty-first International Conference on Machine Learning*.

Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2023. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. *arXiv preprint arXiv:2304.09148*.

Cheng, Z.; Wei, Q.; Zhu, H.; Wang, Y.; Qu, L.; Shao, W.; and Zhou, Y. 2024. Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3511–3522.

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

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, K.; Wang, C.; Wang, Y.; Wang, C.; Yi, R.; and Ma, L. 2023. Rfenet: Towards reciprocal feature evolution for glass segmentation. *arXiv preprint arXiv:2307.06099*.

Fudenberg, D. 1991. *Game theory*. MIT press.

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

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

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

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

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6): 066138.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Lin, J.; He, Z.; and Lau, R. W. 2021. Rich context aggregation with reflection prior for glass surface detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13415–13424.

Lin, J.; and Lau, R. W. 2023. Self-supervised pre-training for mirror detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12227–12236.

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

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

Lin, J.; Yeung, Y. H.; and Lau, R. W. 2022. Depth-aware glass surface detection with cross-modal context mining. *arXiv preprint arXiv:2206.11250*.

Lin, J.; Yeung, Y.-H.; Ye, S.; and Lau, R. W. 2025. Leveraging RGB-D Data with Cross-Modal Context Mining for Glass Surface Detection. *AAAI*.

Liu, J.; Tang, X.; Cheng, F.; Yang, R.; Li, Z.; Liu, J.; Huang, Y.; Lin, J.; Liu, S.; Wu, X.; et al. 2024. MirrorGaussian: Reflecting 3D Gaussians for Reconstructing Mirror Reflections. In *ECCV*.

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

Liu, W.; Shen, X.; Pun, C.-M.; and Cun, X. 2023. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19434–19445.

Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.-P.; Khan, F.; and Han, J. 2024. 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*, 17169–17180.

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

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

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

Pei, J.; Zhou, Z.; Jin, Y.; Tang, H.; and Heng, P.-A. 2023. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2139–2147.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sun, Y.; Xu, C.; Yang, J.; Xuan, H.; and Luo, L. 2025. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, 343–360. Springer.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2022. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3492–3504.

Tang, J.; and Ma, H. 2024. Large-Scale Multi-Robot Coverage Path Planning via Local Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17567–17574.

Wang, Y.; Wang, R.; Fan, X.; Wang, T.; and He, X. 2023. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10031–10040.

Warren, A.; Xu, K.; Lin, J.; Tam, G. K.; and Lau, R. W. 2024. Effective Video Mirror Detection with Inconsistent Motion Cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17244–17252.

Xie, E.; Wang, W.; Wang, W.; Ding, M.; Shen, C.; and Luo, P. 2020. Segmenting transparent objects in the wild. In *Proceedings of the European Conference on Computer Vision*, 696–711.

Xie, Z.; Wang, S.; Yu, Q.; Tan, X.; and Xie, Y. 2024. CS-Fwinformer: Cross-Space-Frequency Window Transformer for Mirror Detection. *IEEE Transactions on Image Processing*.

Xu, K.; Siu, T. W.; and Lau, R. W. 2024. ZOOM: Learning Video Mirror Detection with Extremely-Weak Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6315–6323.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

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

Yu, Z.; Zhang, X.; Zhao, L.; Bin, Y.; and Xiao, G. 2024. Exploring Deeper! Segment Anything Model with Depth Perception for Camouflaged Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4322–4330.

Zha, M.; Fu, F.; Pei, Y.; Wang, G.; Li, T.; Tang, X.; Yang, Y.; and Shen, H. T. 2024a. Dual Domain Perception and Progressive Refinement for Mirror Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zha, M.; Li, T.; Wang, G.; Wang, P.; Wu, Y.; Yang, Y.; and Shen, H. T. 2025a. Implicit Counterfactual Learning for Audio-Visual Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22349–22360.

Zha, M.; Pei, Y.; Wang, G.; Li, T.; Yang, Y.; Qian, W.; and Shen, H. T. 2024b. Weakly-Supervised Mirror Detection via Scribble Annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6953–6961.

Zha, M.; Wang, G.; Pei, Y.; Li, T.; Tang, X.; Li, C.; Yang, Y.; and Shen, H. T. 2025b. Heterogeneous Experts and Hierarchical Perception for Underwater Salient Object Detection. *IEEE Transactions on Image Processing*.

Zha, M.; Wang, G.; Pei, Y.; Li, T.; Tang, X.; Ma, J.; Yang, Y.; and Shen, H. T. 2025c. Think Twice Before Determining: Towards Scene-aware Visual Reasoning for Mirror Detection.

Zhang, P.; Yan, T.; Liu, Y.; and Lu, H. 2024. 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*, 2578–2587.

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