Think Twice Before Determining: Towards Reflection-aware Reasoning for Mirror Detection

Anonymous submission

Abstract

Mirror detection 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) only partial or none of entities and imagings are matched; b) smooth surfaces (*e.g.,* tiles) and transparent objects (*e.g.,* glasses) also exhibit reflections. To address these issues, we formulate the Reflection-aware Reasoning network (RRNet) based on visual prompts. Specifically, we design the prompt reasoning (PR) module that generates a chain of thought reasoning to construct complex spatial location and semantic perception. Noise may accumulate gradually through the chain, and crucial clues may also disappear. Therefore, we introduce the prompt denoising (PD) module to filter out noise and enhance the coupling between prompts. Inspired by the frequency differences between mirror and non-mirror regions, we further propose the prompt guidance and updating (PGU) module that decouples features by injecting predicted masks, enabling interaction and updating in the frequency and spatial domain, respectively. Extensive experiments on four mirror benchmarks and two supervision settings demonstrate that our method surpasses state-of-the-art approaches with lower model and computational complexity. Encouraging performance is also achieved on seven benchmarks of glass, camouflage and underwater scenes, showing its generality. Our code will be available.

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 (Balazevic et al. 2024), 3D reconstruction (Wu et al. 2024), visual-language navigation (Li et al. 2024), and other related tasks. Mirrors are regarded as confounding factors in the causal reasoning chain, while MD can remove them to accurately model the spatial and logical relationships between objects.

As shown in Figure 2, existing methods (Guan, Lin, and Lau 2022; Mei et al. 2021; Lin, Wang, and Lau 2020; Yang et al. 2019; Xie et al. 2024) 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 (He, Lin, and Lau 2023; Huang et al. 2023a) utilize rotation strategy to construct consistency perception between imagings and entities from image or feature map level. Similar works (Liu et al. 2023; Zhang et al. 2024) 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

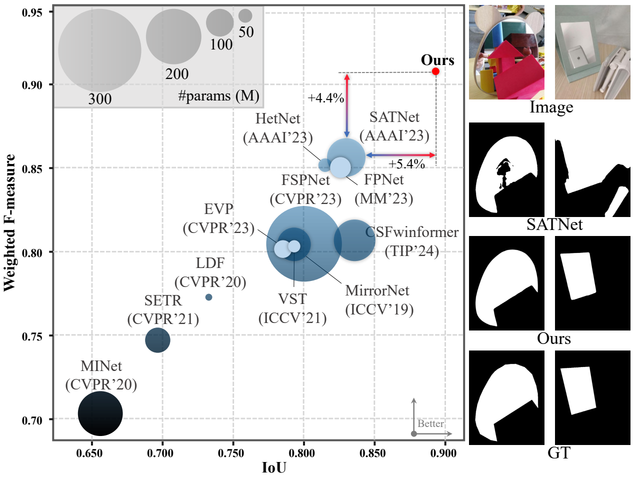

Figure 1: Comparison of our RRNet with different types state-of-the-art (SOTA) detection methods on weighted F-measure (F^w_β), IoU, and parameters using the MSD dataset (Yang et al. 2019). Larger circle indicates more parameters.

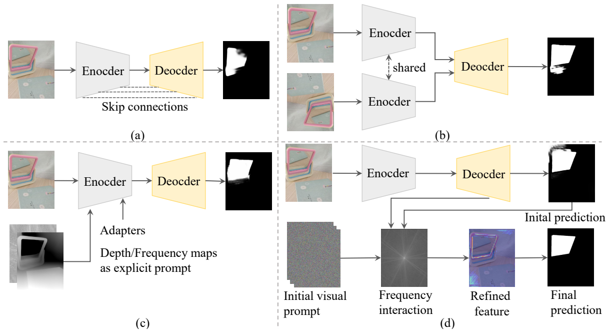

Figure 2: Existing MD and related binary segmentation frameworks. (a) The naive encoding-decoding represented by HetNet (He, Lin, and Lau 2023); (b) The dual-stream represented by SATNet (Huang et al. 2023a); (c) The explicit visual prompt represented by EVP (Liu et al. 2023); (d) Our RRNet.

an efficient and general method that is effective for various scenarios. As shown in Figure 1, our proposed RRNet surpasses various methods with few parameters.

From the perspective of model's output, MD is also a part of binary segmentation aimed at separating target and non-target regions, similar to salient object detection (SOD) and camouflage target detection (COD). However, these methods cannot be directly applied to MD task. SOD methods aim to localize the most salient regions, which may segment both imagings and entities. COD methods are designed to discover camouflaged objects that similar with the environment, *e.g.,* color. Given the similarity between glasses and mirrors, it is possible to detect simultaneously. Moreover, each method is designed with unique components to address specific problems, considering different task settings. The proposed RRNet demonstrates promising performance on one glass and three COD benchmarks, indicating the versatility of our method and plug-and-play modules.

Based on the above, we utilize implicit prompts to establish spatial semantic modeling and combine with frequency perception to capture details, thereby forming the RRNet. Specifically, inspired by thought chain and causal reasoning of large language models (LLMs), we propose the PR module 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 PD module to filter out noise and enhance the coupling between prompts. The learned prompts contain rich semantic knowledge but lack fine-grained perception. Inspired by the idea that low-level features, *e.g.,* edges and textures, have discriminative representations in the frequency domain, and mirrors and non-mirrors exhibit frequency differences, we utilize predicted maps as masks and alternately implement prompt and mirror/non-mirror feature interactions in the frequency domain, then updating in the spatial domain. Thus the prompts possess semantic and detailed knowledge, enabling better localization and refinement of mirror regions.

In summary, our main contributions are as follows:

- Based on visual prompts, we propose the RRNet to incorporate implicit learning of frequency prompts for semantic and detail perception. To the best of our knowledge, we are the first to model MD task from the perspective of frequency prompts.
- We propose the PR module to form a chain of visual thought reasoning, the PD module to reduce noise interference and couple prompts, and the PGU module to utilize prompts for guiding mirror region localization and cross updating.
- Extensive experiments on eleven benchmarks and two task settings demonstrate that our method surpasses SOTA approaches with few model parameters and FLOPs.

Related Work

Mirror Detection. Fully supervised MD methods leverage the unique texture or depth of mirror regions to construct discriminative cues, such as CSFwinformer (Xie et al. 2024). Other approaches focus on establishing context semantic awareness, like PDNet, or building symmetry consistency, such as SATNet. Recently, (Zha et al. 2024) proposed the weakly supervised MD method based on scribble annotations to reduce annotation cost. (Lin, Tan, and Lau 2023) introduced the first MD video dataset and benchmark. Our RRNet outperforms SOTA methods in fully and weakly supervised settings, as well as on the video benchmark.

Prompt Learning. With the development of LLMs (Wei et al. 2022), prompt learning has been widely applied in AIGC (Yang et al. 2024), image enhancement (Wang et al. 2024), and video understanding (Yan et al. 2023). Most relevant to our work is EVP, which utilizes explicit frequency prompts and combines with adapters for various low-level segmentation tasks. However, 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 utilize predicted maps to decouple the feature maps into foreground and background, allowing prompts to alternate in perceiving differences. 3) Unlike directly acquiring frequency from images, which makes it difficult to perceive content dynamically, we leverage feature maps and prompts to interact in the frequency domain and update in the spatial domain.

Frequency Learning. High frequency and low frequency of images represent details and semantics, respectively. Therefore, the characteristic has been applied to various tasks. For example, for object detection, (Wang et al. 2023a) utilize low-frequency invariant and high-frequency variant features to synthesize new images; for image segmentation, (Li et al. 2020b) utilize high-frequency features to refine objects. However, most existing methods focus on images or

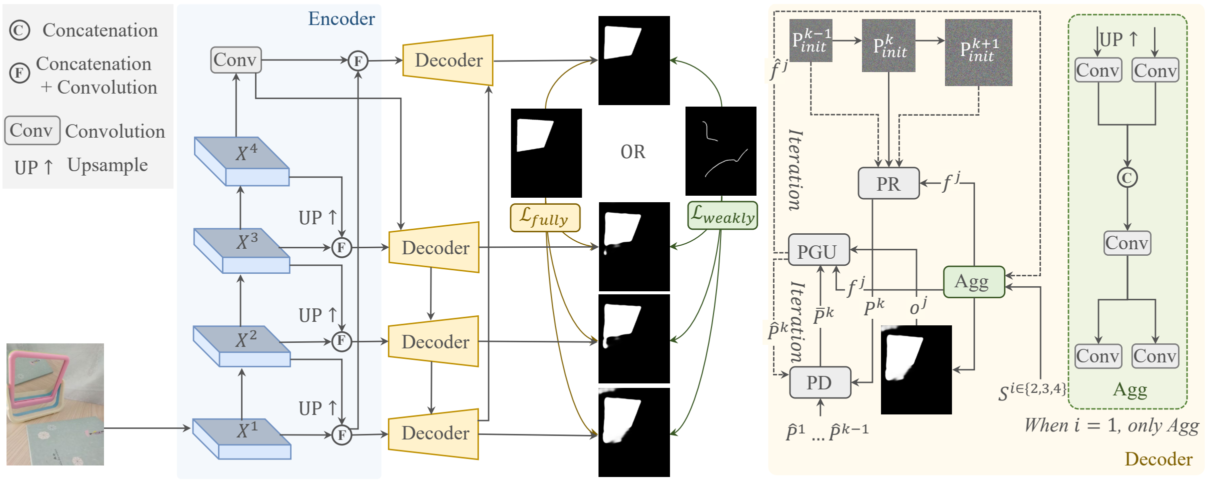

Figure 3: The overview of our RRNet. We utilize the PVT network as the image encoder to obtain multi-scale features and progressively aggregate. We then employ the PR module to construct a visual prompt chain for spatial semantic correlation. Furthermore, we utilize the PD module to avoid noises propagation and crucial clues decay. Finally, we utilize the PGU module to disentangle features and interact with prompts in the frequency domain to compensate for details perception. Considering task differences, we employ different GT as supervision.

features level processing and rarely consider prompts level. Our proposed PGU module models feature and prompt level simultaneously and interacts in the frequency domain.

Proposed Method

Overall Architecture

The overview of our RRNet is shown in Figure 3, which follows an encoder-decoder architecture and incorporates frequency prompts into the decoder. Given an input $I \in \mathbb{R}^{3\times H\times W}$, we generate multi-scale features $X^i \in \mathbb{R}^{C_i\times \frac{H}{4^i}\times\frac{W}{4^i}}$ through the PVT network (Wang et al. 2021) and progressively aggregate adjacent features to obtain S^i, where C, H, W denote channels, height and width respectively, $i \in \{1,2,3,4\}$. In the decoding stage, we fuse S^i and S^{i+1} to generate f^j and predicted maps o^j, where $j = i-1$. We utilize the PR module to generate prompts $P^k \in \mathbb{R}^{C_k\times\frac{H}{4^{3-k}}\times\frac{W}{4^{3-k}}}$ and further update and couple using the PU module, where $k \in \{1,2,3\}, C_k = 64\times k$. Finally, the PI module is employed to inject prompts and perform cross-updating, generating $\hat{f}^j$ and $\hat{P}^k$. Through iterative refinement, we can obtain high-quality outputs. Depending on the specific task setting, different ground truth (GT) are used for supervision.

Prompt Reasoning Module

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 such as glass. Essentially, the goal is to construct effective semantic perception. LLMs utilize chain-of-thought prompting to build logical connections and complex reasoning between texts. Inspired by this, we propose the PR module, 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. The structure details of the PR module are shown in Figure 4.

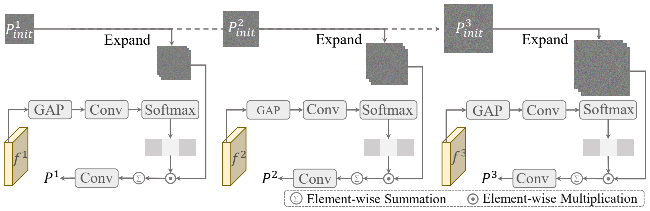

Figure 4: Structure of the PR module. We generate a visual prompt chain of thought reasoning.

Specifically, we randomly initialize the prompt parameters $P^1_{init} \in \mathbb{R}^{C_1\times\frac{H}{16}\times\frac{W}{16}}$ and generate multi-scale prompts by progressive transposed convolution (Dumoulin and Visin 2016), sharing the parameter space, represented as follows:

$$P^2_{init} = \phi^L_{3\times3}(P^1_{init}), P^3_{init} = \phi^L_{3\times3}(P^2_{init}) \qquad (1)$$

where $\phi^L_{3\times3}$ denotes dimension expansion (prompt length, *i.e, L*) after 3x3 transposed convolution. Thus we can obtain $P^2_{init} \in \mathbb{R}^{C_2\times L\times\frac{H}{8}\times\frac{W}{8}}$ and $P^3_{init} \in \mathbb{R}^{C_3\times L\times\frac{H}{4}\times\frac{W}{4}}$.

The pixel weights of prompt can be generated by:

$$W^k_p = softmax(Proj(GAP(f^k))) \qquad (2)$$

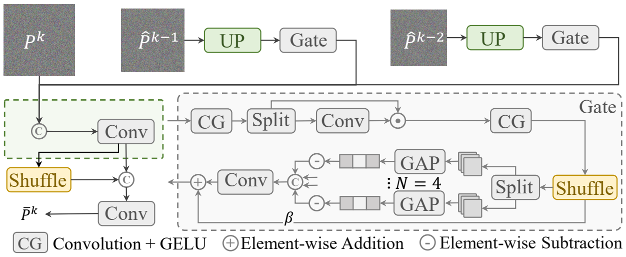

Figure 5: Structure of the PD module. We filter out noise from spatial and channel dimensions and then fusion.

where GAP denotes global average pooling. $Proj$ is a linear mapping layer designed to project dimensions to L. $W_p^k \in \mathbb{R}^L$ can provide priors to avoid slow parameter optimization. We then obtain prompt P^k by:

$$P^k = Conv(\sum_{l=1}^{L} W_p^k \odot P_{init}^k) \tag{3}$$

where $\odot$ and $Conv$ denotes element-wise multiplication and 3×3 convolution, respectively. For simplicity, we omit some tensor adjustments.

Prompt Denoising Module

We still encounter two problems with the prompt chain generated by the PR module: 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 PU module, 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 structure details of the PD module are shown in Figure 6.

Semantic is typically encoded in channels. When $k = 1$, without any preceding prompts, we only leverage channel shuffle (Zhang et al. 2018) on P^k to achieve symmetry consistency implicitly, which is different from explicit perception of SATNet through rotating random angles. We can obtain refined $\bar{P}^1$ by:

$$\bar{P}^1 = Conv(Concat(Shuffle(P^1), P^1)) \tag{4}$$

where $Concat(\cdot;\cdot)$ denotes channel concatenation.

When $k > 1$, we adjust the shapes of the previous $k-1$ prompts to match P^k. When $k = 3$, we have:

$$P^2 = P^2 \mapsto P^3, P^1 = P^1 \mapsto P^3 \tag{5}$$

where $\mapsto$ represents upsampling. Taking P^2 as an example, we filter out the noise in the spatial dimension by:

$$P_{p1}^2, P_{p2}^2 = Split(Conv(P^2)) \tag{6}$$

$$P_s^2 = P_{p2}^2 \odot Conv(P_{p1}^2) \tag{7}$$

where $Split(\cdot)$ denotes channel split.

For channel gate, we apply channel shuffle and split into four groups with different semantics. Inspired by (Pan, Cai, and Zhuang 2022), we 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 can be represented as follows:

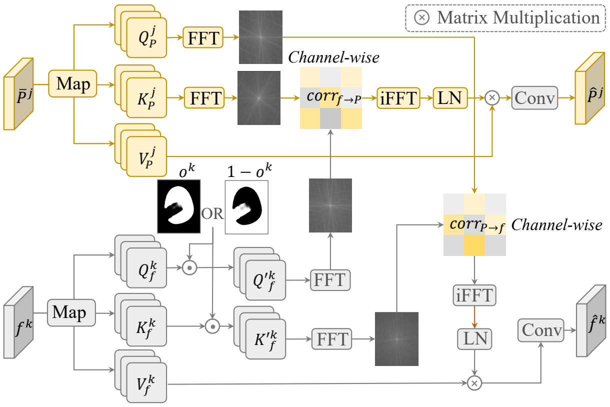

Figure 6: Structure of the PGU module. We utilize output to decouple, with frequency domain interaction and spatial domain updating.

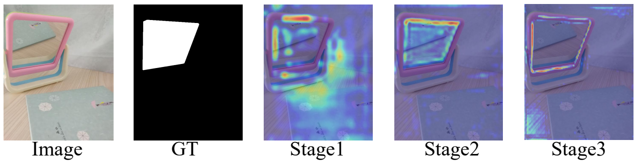
Image GT Stage1 Stage2 Stage3

Figure 7: Visualization of updated feature maps.

$$P_{g^m}^2 = Split(Shuffle(P_s^2)), m \in \{1, 2, 3, 4\} \tag{8}$$

$$P_{g_h^m}^2 = P_{g^m}^2 - Mean(P_{g^m}^2) \tag{9}$$

$$P_g^2 = Concat(P_{g_h^1}^2, ..., P_{g_h^4}^2) \tag{10}$$

$$P'^2 = Conv(Shuffle(P_s^2) + Shuffle(P_s^2)) \tag{11}$$

Similarly, we can generate P'^3. We then fuse P'^1, P'^2 and P^3 and shuffle to generate $\bar{P}^3$. Also, We can obtain $\bar{P}^2$.

As shown in Figure 5, using the prompt chain for step-by-step reasoning, we can accurately locate the mirror region.

Prompt Guidance and Updating Module

Although optimized prompt contains rich semantics, it lacks perception for details. We propose the PGU module, which projects the feature maps and prompts into the Fourier (frequency) domain for interaction to enhance low level representations of mirror regions, such as edges and textures, and update in the spatial domain. To further increase the representation difference between mirror and non-mirror regions, we utilize the predicted maps to disentangle features, allowing prompts to focus on heterogeneous representations at different stages. The detailed structure of the PGU module is shown in Figure 7.

Specifically, We apply 1×1 convolution to $\bar{P}^k$ and f^k for dimension mapping, followed by 3×3 convolution to encode local features. Then, we generate $\{Q, K, V\}_P^k$ through channel split, which is expressed as:

$$\{Q, K, V\}_P^k = Split(Conv(Conv(\bar{P}^k))) \tag{12}$$

Similarly, we can obtain $\{Q, K, V\}_f^k$. Then we use o^k to decouple feature maps. Note that we do not decouple prompts. When $k = 1$, we have:

$$Q_f^{'k} = Q_f^k \odot o^k, K_f^{'k} = K_f^k \odot o^k \tag{13}$$

when $k >= 2$, we have:

$$Q_f^{'k} = Q_f^k \odot (1 - o^k), K_f^{'k} = K_f^k \odot (1 - o^k) \tag{14}$$

Furthermore, we further achieve interaction in the frequency domain. The correlation matrix $corr_{P\to f}$ from prompts to feature maps is represented by:

$$corr_{P\to f} = softmax(\frac{FFT(Q_P^k)FFT(K_f^{'k})}{\tau}) \tag{15}$$

where τ is a learnable scaling factor. FFT denotes fast Fourier transform (Brigham 1988). Similarly, we can generate $corr_{f\to P}$.

Finally, we can obtain updated $\hat{P}^k$ in spatial domain by:

$$\hat{P}^k = Conv(LN(iFFT(corr_{f\to P}) \otimes V_P^k)) \tag{16}$$

where $iFFT$ and LN denote inverse fast Fourier transform and layer normalization (Xiong et al. 2020), respectively. $\otimes$ is matrix multiplication. Also, we can generate updated $\hat{f}^k$.

Loss Function

We apply supervision to o^j. For the fully-supervised setting, following (He, Lin, and Lau 2023), we employ weighted binary cross-entropy loss ($\mathcal{L}_{BCE}^w$) and weighted intersection over union loss ($\mathcal{L}_{IoU}^w$), which is expressed as:

$$\mathcal{L}_{fully} = \sum_{j=1}^{4} \mathcal{L}_{BCE}^w + \mathcal{L}_{IoU}^w \tag{17}$$

For the weakly-supervised setting, following (Zhang et al. 2020), we utilize partial cross-entropy loss ($\mathcal{L}_{CE}^p$) and smooth loss ($\mathcal{L}_S$), which is represented as:

$$\mathcal{L}_{weakly} = \sum_{j=1}^{4} \mathcal{L}_{CE}^p + \mathcal{L}_S \tag{18}$$

Note that we do not leverage edge and contrast losses proposed by (Zha et al. 2024) as additional supervision.

Experiments

Datasets. We conduct experiments on eleven benchmarks. **Mirror scenarios**, *i.e.,* MSD, PMD (Lin, Wang, and Lau 2020), Mirror-RGBD (Mei et al. 2021) and VMD (Lin, Tan, and Lau 2023). For Mirror-RGBD dataset, we do not leverage depth maps. For VMD dataset, we shuffle all video frames to avoid introducing temporal information. **Glass scenario**, *i.e.,* GDD (Mei et al. 2020), consisting of 2980 and 936 images for training and testing, respectively. **Camouflage scenarios**, *i.e.,* CAMO (Le et al. 2019), COD10K (Fan et al. 2020), and NC4K (Lv et al. 2021). We use 4040 training images from CAMO and COD10K, and the remaining along with NC4K are used for testing. **Underwater scenarios**, *i.e.,* MAS3K (Li et al. 2020a), RMAS (Fu et al. 2023), UFO120 (Islam, Luo, and Sattar 2020).

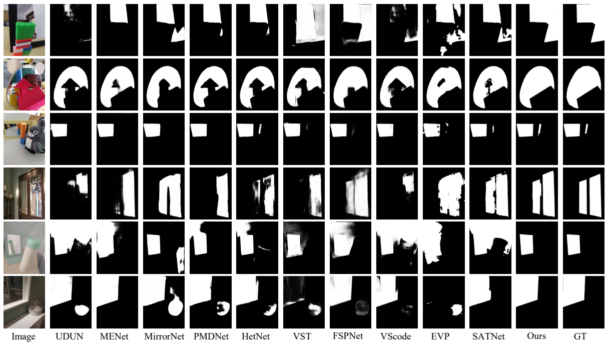

Figure 8: Qualitative comparison on mirror scenes.

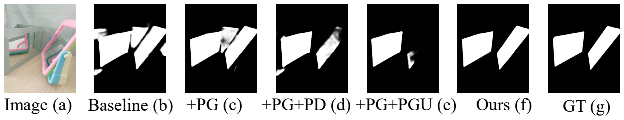

Figure 9: Qualitative ablation of different modules.

Implementation Details. We implement our model via PyTorch and conduct all experiments on an NVIDIA A100. Following (Zha et al. 2024), we employ the PVT network as the image encoder. Following (Lin, Wang, and Lau 2020; Mei et al. 2020; Lv et al. 2021), for all datasets, input are resized to 384×384 for fair comparison. For training, we use the AdamW (Loshchilov and Hutter 2017) as our optimizer to update the model parameters, with 200 epochs, the batch size of 40, and the initial learning rate of 1e-4. For testing, we directly compute generated predictions without adopting any post-processing operations.

Evaluation Metrics. We compute five evaluation metrics, *i.e.,* S-measure (S_m), mean E-measure (E_m), weighted F-measure (F_β^w) (Fan et al. 2017, 2018), intersection over union (IoU) and Mean Absolute Error (MAE). The higher the better for the first four. For MD methods, we directly utilize the provided prediction maps. For other methods, we retrain using the official codes on the MSD and PMD datasets.

Comparison with SOTA Methods

Quantitative Comparison. We consider both fully supervised and weakly supervised settings, different backbone networks, and select various binary and general segmentation methods for comparison. Specifically, as shown in Table

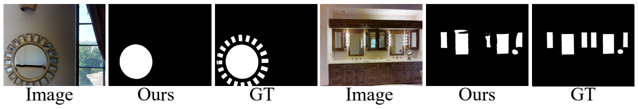

Figure 10: Failure cases.

Methods	Att.	MSD					PMD				
		MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
CNN-based Fully-Supervised											
MINet $_{\text{CVPR'20}}$	S	0.088	0.792	0.819	0.715	0.664	0.038	0.794	0.822	0.667	0.601
LDF $_{\text{CVPR'20}}$	S	0.068	0.821	0.867	0.773	0.729	0.038	0.799	0.833	0.683	0.633
UDUN $_{\text{MM'23}}$	S	0.071	0.815	0.838	0.746	0.713	0.039	0.784	0.794	0.652	0.600
MENet $_{\text{CVPR'23}}$	S	0.054	0.868	0.906	0.829	0.805	0.033	0.826	0.873	0.727	0.680
MirrorNet $_{\text{ICCV'19}}$	M	0.065	0.850	0.891	0.812	0.790	0.043	0.761	0.841	0.663	0.585
PMDNet $_{\text{CVPR'20}}$	M	0.047	0.875	0.908	0.845	0.815	0.032	0.810	0.859	0.716	0.660
HetNet $_{\text{AAAI'23}}$	M	0.043	0.881	0.921	0.854	0.824	0.029	0.828	0.865	0.734	0.690
Transformer-based Fully & Weakly-Supervised											
SETR $_{\text{CVPR'21}}$	S	0.071	0.797	0.840	0.750	0.690	0.035	0.753	0.775	0.633	0.564
VST $_{\text{ICCV'21}}$	S	0.054	0.861	0.901	0.818	0.791	0.036	0.783	0.814	0.639	0.591
VSCode $_{\text{CVPR'24}}$	C	0.077	0.800	0.820	0.721	0.687	0.042	0.787	0.816	0.656	0.607
FSPNet $_{\text{CVPR'23}}$	C	0.057	0.871	0.897	0.818	0.807	0.065	0.743	0.752	0.513	0.530
FPNet $_{\text{MM'23}}$	C	0.042	0.883	0.917	0.849	0.827	0.033	0.823	0.874	0.717	0.673
SAM $_{\text{ICCV'23}}$	G	0.124	–	–	–	0.515	0.052	–	–	–	0.647
EVP $_{\text{CVPR'23}}$	G	0.064	0.845	0.896	0.811	0.780	0.037	0.793	0.861	0.694	0.634
SATNet $_{\text{AAAI'23}}$	M	0.033	0.887	0.916	0.865	0.834	0.025	0.826	0.858	0.739	0.684
CSFwinformer $_{\text{TIP'24}}$	M	0.045	0.875	0.905	0.846	0.821	**0.024**	0.831	0.864	0.756	0.700
Ours	M	**0.027**	**0.917**	**0.950**	**0.909**	**0.888**	0.025	**0.841**	**0.898**	**0.761**	**0.717**
WSMD $_{\text{AAAI'24}}$	W	0.078	0.828	0.878	0.780	0.750	0.051	0.773	0.824	0.630	0.600
Ours*	W	**0.071**	**0.839**	**0.887**	**0.796**	**0.771**	**0.048**	**0.777**	**0.829**	**0.651**	**0.602**

Table 1: Quantitative comparison on MSD and PMD datasets. S, C, G, W, M denote SOD, COD, general segmentation, weakly-supervised MD, fully-supervised MD methods respectively. The best performances are bolded.

Methods	Depth	Mirror-RGBD				
		MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
PICRNet$_{\text{MM'23}}$	✓	0.053	0.815	0.877	0.772	0.718
PDNet$_{\text{CVPR'21}}$	✓	**0.042**	0.856	0.906	0.825	0.778
Ours	✗	0.044	**0.856**	**0.908**	**0.828**	**0.781**
WSMD$_{\text{AAAI'24}}$	✗	0.088	0.754	0.806	0.655	0.616
Ours*	✗	**0.077**	**0.770**	**0.836**	**0.685**	**0.643**

Table 2: Quantitative comparison on Mirror-RGBD dataset.

Methods	Input Size	FLOPs↓	Params.↓
PMDNet	384×384	101.54	147.66
PDNet	416×416	41.16	80.54
SATNet	512×512	153.00	139.36
CSFwinformer	512×512	139.45	150.54
Ours	384×384	**16.53**	**27.69**
WSMD	352×352	21.39	**26.16**
Ours*	352×352	**13.91**	27.69

Table 3: Model Efficiency Comparison. We compare with four MD models on Parameters (M), FLOPs (GMAC).

Methods	Params.	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
VMDNet$_{\text{CVPR'23}}$	62.24	0.105	0.731	0.742	0.623	0.567
Ours	**27.69**	**0.096**	**0.784**	**0.826**	**0.716**	**0.670**

Table 4: Quantitative comparison on VMD dataset.

1, for the fully supervised setting, five MD methods: MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), HetNet (He, Lin, and Lau 2023), SATNet (Huang et al. 2023a), and CSFwinformer (Xie et al. 2024); five SOD methods: MINet (Pang et al. 2020), UDUN (Pei et al. 2023), MENet (Wang et al. 2023b), SETR (Zheng et al. 2021) and VST (Liu et al. 2021); three COD methods: FSPNet (Huang et al. 2023b), FPNet (Cong et al. 2023b), and VSCode (Luo et al. 2024); two general segmentation methods: SAM (Kirillov et al. 2023) and EVP. Our approach surpasses various types of methods, especially EVP, with gains of 3.7%, 7.2%, 5.4%, 9.8%, and 10.8% respectively on the five metrics of the MSD dataset, and the average surpasses SATNet, CSFwinformer by around 4.0%, 5.0%, respectively, which demonstrates the effectiveness of implicit frequency learning. As shown in Table 2, we select two RGB-D SOD methods, *i.e.,* PICRNet (Cong et al. 2023a) and the MD-specific method, *i.e.,* PDNet (Mei et al. 2021). Our method does not utilize depth modality as guidance and achieves comparable performance to PDNet with lower computational costs, significantly surpassing PICRNet. For the weakly supervised setting, our method also surpassed WSMD on three benchmarks. As shown in Table 3, the FLOPs and parameters of the RRNet are about one-tenth and one-fifth of SATNet, respectively, and parameters are about two-thirds of WSMD, demonstrating its efficiency. As shown in Tables 4, 5, and 6, our method also surpasses SOTA approaches, *e.g.,* VMDNet (Lin, Tan, and Lau 2023), RFENet (Fan et al. 2023), DualSAM (Zhang et al. 2024) in video mirror, glass, camouflage and underwater scenes.

Methods	Params.	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
RFENet$_{\text{IJCAI'23}}$	152.65	0.061	0.858	0.913	0.901	0.882
Ours	**27.69**	**0.048**	**0.884**	**0.933**	**0.925**	**0.902**

Table 5: Quantitative comparison on GDD dataset.

Methods	Params.	CAMO			COD10K			NC4K		
		S_m↑	E_m↑	F_β^w↑	S_m↑	E_m↑	F_β^w↑	S_m↑	E_m↑	F_β^w↑
FSPNet$_{\text{CVPR'23}}$	273.79	**0.856**	0.899	0.799	0.851	0.895	0.735	**0.879**	0.915	0.816
VSCode$_{\text{CVPR'24}}$	54.09	0.836	0.892	0.768	0.847	0.913	0.744	0.874	0.920	0.813
Ours	**27.69**	0.853	**0.914**	**0.809**	**0.860**	**0.928**	**0.780**	0.878	**0.929**	**0.831**
Methods	**Params.**	**MAS3K**			**RMAS**			**UFO120**		
		S_m↑	E_m↑	F_β^w↑	S_m↑	E_m↑	F_β^w↑	S_m↑	E_m↑	F_β^w↑
SAM$_{\text{ICCV'23}}$	–	0.763	0.656	0.807	0.697	0.534	0.790	0.768	0.745	0.827
DualSAM$_{\text{CVPR'24}}$	159.95	0.884	0.838	0.933	0.860	0.812	0.944	0.856	0.864	0.914
Ours	**27.69**	**0.903**	**0.865**	**0.949**	**0.876**	**0.833**	**0.951**	**0.867**	**0.870**	**0.920**

Table 6: Quantitative comparison on camouflage and underwater scenes.

Qualitative Comparison. As shown in Figure 8, we provide visual comparisons of different scenarios. The first two rows depict regular and irregular occlusions, the third row represents scale variations, the fourth row illustrates multiple targets, and the last two rows demonstrate scenes with mirrors and tile (or glass). Our method can accurately and completely locate mirror regions.

Ablation Study

We validate the effect proposed modules, mask guidance and frequency interaction on the MSD dataset.

Effect of the proposed modules. As shown in Table 7 and Figure 9, we progressively incorporate components. Specifically, after adding the PR module to the *Baseline* model, all metrics show significant improvements. In Figure 9 (b) and (c), we can establish preliminary perception of differences between entities and imagings, and the false positive is noticeably mitigated. Only by constructing a visual prompt chain using the PR module can we employ the PD and PGU modules for further improvement. Therefore, we separately incorporate the PD and PGU modules based on the *Baseline+PR* model, but the former demonstrates superior improvements, indicating that clean and highly coupled prompts are crucial, and more optimal injection strategy can further highlight. In Figure 9 (d) and (e), the latter still exhibits some missing detections, while the former eliminates this error and is more sensitive to similar texture regions. With the incorporation of all modules, the five metrics improve by 1.8%, 4.0%, 4.5%, 6.0%, and 7.1%, respectively. In Figure 9 (f), we obtain detection results close to the GT, demonstrating the complementarity of the modules.

PR	PD	PGU	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
			0.045	0.877	0.905	0.849	0.817
✓			0.038	0.885	0.917	0.876	0.861
✓	✓		0.029	0.908	0.938	0.900	0.878
✓		✓	0.033	0.901	0.930	0.892	0.870
✓	✓	✓	**0.027**	**0.917**	**0.950**	**0.909**	**0.888**

Table 7: Quantitative ablation of proposed modules. The first line is the *Baseline* model and we utilize channel concatenation to fuse prompts instead of the PR module

M^1	M^2	M^3	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
✓			0.028	**0.920**	0.942	0.904	0.880
✓	✓		**0.027**	0.917	**0.950**	**0.909**	**0.888**
✓	✓	✓	0.029	0.907	0.938	0.901	0.876

Table 8: Quantitative ablation of the PGU module. M^i denotes whether the *i-th* stage fuses the non-mirror mask.

Spa.	Fre.	MAE↓	S_m↑	E_m↑	F_β^w↑	IoU↑
✓		0.029	0.908	0.939	0.898	0.880
	✓	**0.027**	**0.917**	**0.950**	**0.909**	**0.888**

Table 9: Quantitative ablation of the frequency interaction.

N	0	1	2	3	4	5
S_m	0.877	0.895	0.905	0.917	0.921	0.926
FLOPs	14.96	15.11	15.78	16.53	17.17	17.72

Table 10: Quantitative ablation of number of prompts.

Effect of mirror mask. As shown in Table 8, applying mask in all stages is even inferior to using just one, indicating that alternating between mirror and non-mirror regions perception is necessary to capture the semantic differences.

Effect of frequency interaction. As shown in Table 9, performance is better when interacting in frequency domain compared to spatial domain, indicating that obtaining better semantics is not enough, perceiving details is also important.

Number of prompts. As prompts increasing, S_m grows larger. However, when $N>3$, although S_m still increases, it introduces more FLOPs. Therefore, balance between performance and efficiency can be realized when $N = 3$.

Failure Cases Analysis and Broader Impacts

As shown in Figure 10, our method exhibits false negatives in the scenario with some small targets. Our method can be applied to wider scenarios, *e.g.,* all binary segmentation tasks. Moreover, our proposed plug-and-play modules can empower other high-level tasks, such as object detection. Please see our supplementary material for more results.

Conclusion

We compare existing MD paradigms and propose the RRNet. Inspired by LLMs, we introduce the PR module to construct a visual prompt chain, enhancing semantic perception and spatial reasoning in complex spaces. Furthermore, we propose the PD module to avoid noise accumulation and beneficial feature decay, strengthening the coupling between prompts. Finally, we design the PGU module, which leverages the outputs to disentangle the feature maps into mirror and non-mirror regions, injects prompts in the frequency domain to compensate for the lack of detail perception, and then updates in the spatial domain. Extensive experiments on eleven benchmarks and two task settings validate the effectiveness and efficiency of our approach.

References

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

Brigham, E. O. 1988. *The fast Fourier transform and its applications*. Prentice-Hall, Inc.

Cong, R.; Liu, H.; Zhang, C.; Zhang, W.; Zheng, F.; Song, R.; and Kwong, S. 2023a. Point-aware interaction and cnn-induced refinement network for RGB-D salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 406–416.

Cong, R.; Sun, M.; Zhang, S.; Zhou, X.; Zhang, W.; and Zhao, Y. 2023b. Frequency perception network for camouflaged object detection. In *Proceedings of the ACM International Conference on Multimedia*, 1179–1189.

Dumoulin, V.; and Visin, F. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

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, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2777–2787.

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*.

Fu, Z.; Chen, R.; Huang, Y.; Cheng, E.; Ding, X.; and Ma, K.-K. 2023. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*.

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.

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

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

Islam, M. J.; Luo, P.; and Sattar, J. 2020. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv preprint arXiv:2002.01155*.

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.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184: 45–56.

Li, J.; Padmakumar, A.; Sukhatme, G.; and Bansal, M. 2024. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18517–18526.

Li, L.; Rigall, E.; Dong, J.; and Chen, G. 2020a. MAS3K: An open dataset for marine animal segmentation. In *International Symposium on Benchmarking, Measuring and Optimization*, 194–212. Springer.

Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Lin, Z.; Tan, S.; and Tong, Y. 2020b. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 435–452. Springer.

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.

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.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11591–11601.

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.

Pan, Z.; Cai, J.; and Zhuang, B. 2022. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35: 14541–14554.

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

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.

Wang, C.; Pan, J.; Lin, W.; Dong, J.; Wang, W.; and Wu, X.-M. 2024. Selfpromer: Self-prompt dehazing transformers with depth-consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5327–5335.

Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; and Zha, Z.-J. 2023a. Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1064–1073.

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

Wang, Y.; Wang, R.; Fan, X.; Wang, T.; and He, X. 2023b. 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.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21551–21561.

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

Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; and Liu, T. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524–10533. PMLR.

Yan, L.; Han, C.; Xu, Z.; Liu, D.; and Wang, Q. 2023. Prompt Learns Prompt: Exploring Knowledge-Aware Generative Prompt Collaboration For Video Captioning. In *IJCAI*, 1622–1630.

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

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.

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

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12546–12555.

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.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856.

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.

This paper

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (partial)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (yes)

- All assumptions and restrictions are stated clearly and formally. (partial)
- All novel claims are stated formally (e.g., in theorem statements). (partial)
- Proofs of all novel claims are included. (partial)
- Proof sketches or intuitions are given for complex and/or novel results. (partial)
- Appropriate citations to theoretical tools used are given. (partial)
- All theoretical claims are demonstrated empirically to hold. (partial)
- All experimental code used to eliminate or disprove claims is included. (no)

Does this paper rely on one or more datasets? (yes)

- A motivation is given for why the experiments are conducted on the selected datasets (partial)
- All novel datasets introduced in this paper are included in a data appendix. (partial)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (partial)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (partial)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (partial)

Does this paper include computational experiments? (yes)

- Any code required for pre-processing data is included in the appendix. (no).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (no)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (partial)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (partial)
- This paper states the number of algorithm runs used to compute each reported result. (no)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (partial)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (partial)