Generalized Implicit Reflection Perception for Mirror Detection

Anonymous submission

Abstract

Mirror detection aims to segment mirror regions based on the differences between entities and imagings. Existing detection methods explicitly utilize the symmetry of mirror imaging or additional cues such as depth to assist in localization. However, in some scenarios, only imagings are observed without corresponding entities, making it difficult to establish explicit symmetric associations. Besides, auxiliary clues inevitably introduces noise and lacks generalization. To address these issues, we formulate an generalized implicit reflection perception (IRP) framework based on visual prompts. Specifically, we introduce a prompt chain generation (PCG) module to perceive mirror region features through chain of thought reasoning, even in the absence of entities. Explicit visual prompts are fixed or slightly updated at the image level. We propose a prompt updating (PU) module based on gate mechanism to achieve feature-level and semantic-aware updates, reducing reflection interference and enhancing the prompt chain. Furthermore, we propose a prompt injection (PI) module to drive the model to localize mirrors. The proposed modules are all plug-and-play. Our method achieves state-of-the-art performance with few computational complexity on four mirror benchmarks, surpassing fully and weakly supervised single-modal, multimodal, and video-level methods. Promising performance is also achieved on seven glass, camouflage and underwater benchmarks. Our code will be available.

Introduction

Unlike conventional low-level structure segmentation tasks such as camouflage object detection (COD), salient object detection (SOD) or semantic segmentation (SS), mirror detection (MD) aims to overcome unique interference from reflection imaging and locate mirror regions. A clear distinction between imagings and entities is of importance for tasks such as object recognition (Wang et al. 2023b), path planning (Tang and Ma 2024), and 3D spatial reconstruction (Wu et al. 2024) in embodied agents.

MD has three main issues: 1) Camouflage: mirrors are typically placed indoors and the reflective property makes them highly similar to the surrounding environment. Relying simply on reflection mechanism to establish associations between entities and imagings may lead to false positives, such as mistaking transparent objects like glass or reflective ceramic walls for mirrors (Figure 1 (b)). Conversely, it may result in false negatives, such as rearview mirrors in cars that only have imagings without corresponding physical entities. 2) Irregular shapes: mirrors exhibit highly irregular shapes due to occlusions, such as bouquets, making it difficult for models to determine through shape matching like circles or rectangles. 3) Large scale variations: mirror regions may occupy the entire image or less than one-tenth, thus accurately and completely localization is challenging.

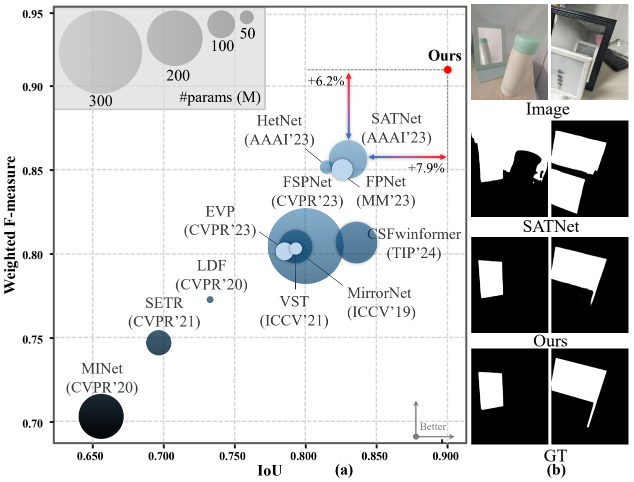

Figure 1: Comparison of our IRP framework with ten SOTA detection methods on weighted F-measure (F_β^w), mean IoU, and parameters using the MSD dataset (Yang et al. 2019). Smaller circle indicates lower parameters.

Previous works design customized modules to explicitly establish reflective and scale awareness, thus addressing the above issues. As shown in Figure 2, we rethink the existing MD frameworks, which can be roughly divided into four categories. a) The naive U-shaped encoder-decoder architecture represented by HetNet (He, Lin, and Lau 2023) captures details and semantic correlations through high-low hierarchical design of several components. b) The siamese encoding architecture represented by SATNet (Huang et al. 2023a) incorporates random image rotations and designs dual-stream networks to explicitly capture symmetric representation consistency. c) The explicit visual prompt-assisted architectures represented by EVP (Liu et al. 2023) and VSCode (Luo et al. 2024) generate frequency or depth rep-

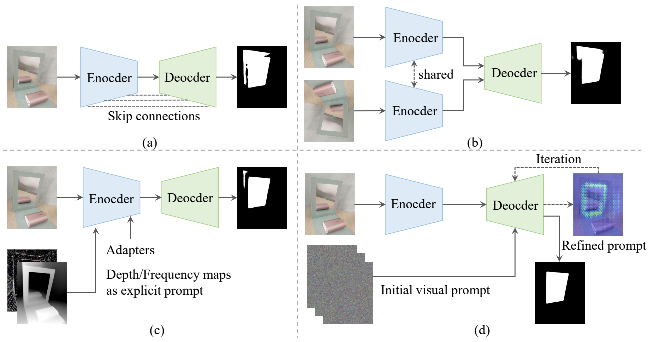

Figure 2: Existing MD and general binary segmentation frameworks. (a) The naive encoding-decoding represented by HetNet; (b) The siamese represented by SATNet; (c) The explicit visual prompt (not specific for MD task) represented by EVP; (d) Our IRP framework.

resentations of inputs and employ adapters in the encoding stage to adapt to different tasks or domains. d) Our proposed IRP framework does not require explicit generation of visual prompts or adapters. Instead, we learn implicit visual prompts at the decoding stage. Both frequency and depth prompts inevitably introduce noise and are suboptimal. In contrast, our learned implicit prompts possess higher-order semantic, effectively avoiding noise interference and enhancing mirror region representation. Furthermore, our approach can generalize to different tasks such as COD. Through implicit learning, we can obtain symmetrical perception from arbitrary angles, while HetNet or SATNet simply establish entities-imagings correlations within a fixed range by randomly rotating 90 degrees (or multiples). As shown in Figure 1 (a), our method significantly outperforms the state-of-the-art (SOTA) methods with extremely low parameters, effectively handling reflective interference, irregular occlusions, and scale modeling.

Based on the above discussion, we propose the IRP framework based on visual prompts to implicitly perceive the context of mirror region. Specifically, inspired by large language models (LLMs) (Wei et al. 2022), we propose the PCG module, which aims to enhance the model's visual reasoning ability through constructing a series of intermediate prompts, *i.e.,* a prompt chain. This is crucial for establishing the associations between entities and imagings, especially when the corresponding entities are missing. However, due to the larger representation space of implicit learning compared to explicit learning. Besedes, imagings can change depending on the environment in which the mirror is located, irregular occlusions and scale changes further expand the boundary of the learning space, making it difficult to learn the crucial prompts. Therefore, we propose the PU module to reduce the exploration path, mitigate noise interference, thus obtain the optimal prompts. We further propose the PI module to inject the prompts into the feature maps to accurately and completely localize the mirror region. Extensive experiments on four mirror benchmarks and seven glass camouflage and underwater benchmarks demonstrate the superiority and robustness of our approach.

In summary, our main contributions are as follows:

- We rethink existing MD frameworks and formulate the IRP framework to implicitly perceive reflections and address occlusion and scale variation issues. To the best of our knowledge, we are the first to model MD task from the implicit visual prompt perspective.
- We propose the PCG module to construct a prompt chain and the PU module to refine the prompts. The proposed PI module injects the prompts into feature maps to accurately and completely localize mirror regions. All the proposed modules are plug-and-play.
- Our approach surpasses various SOTA methods on four MD benchmarks, with low computational complexity. And we validate its generalization on seven datasets involving glass, camouflage and underwater scenarios.

Related Work

Mirror Detection. Mirrors reflect physical entities and create completely identical imagings, which can seriously confuse and impact the understanding and modeling of visual space. Some works attempt to locate mirror regions to eliminate the interference of reflections. (Yang et al. 2019; Lin, Wang, and Lau 2020; Guan, Lin, and Lau 2022) designed several components to establish the associations between entities and imagings. (He, Lin, and Lau 2023; Huang et al. 2023a) utilized rotation strategies to construct mirror symmetry consistency. (Zha et al. 2024) constructed a weakly supervised model based on scribble annotations, *i.e.,*WSMD, and achieved performance comparable to fully supervised methods. (Mei et al. 2021; Tan et al. 2022) leveraged the differences between mirror and non-mirror regions to distinguish, *i.e.,* depth and content distribution. (Lin, Tan, and Lau 2023; Warren et al. 2024) proposed the video-level method based on motion cues. Our method significantly outperforms SOTA single-modal, video-specific approaches and achieves comparable performance to the multi-modal MD method under the fully-supervised setting, while surpassing WSMD under the weakly-supervised setting.

Prompt learning. Prompt learning has been widely applied in natural language processing (Wei et al. 2022), recommendation systems (Li, Zhang, and Chen 2023), and open-world perception (Zhu et al. 2023). (Liu et al. 2023) introduced visual prompts, generated from the high-frequency components of inputs, into low-level structure segmentation tasks. However, the explicit prompt construction based on images may introduce additional noises and lack semantic representations, making it difficult to establish associations between entities and imagings or capture mirror regions contexts. Similarly, (Luo et al. 2024) constructed 2D visual prompts based on mixed datasets and applied to the COD task. But this significantly increases the training cost and lacks generalizability. In contrast, our approach constructs semantic-aware implicit prompts to reduce reflective interference and achieves superior performance across multiple tasks, *e.g.,* COD, with lower computational overhead.

Image Segmentation. Binary segmentation task, such as COD and SOD, although overlaps with some issues of MD,

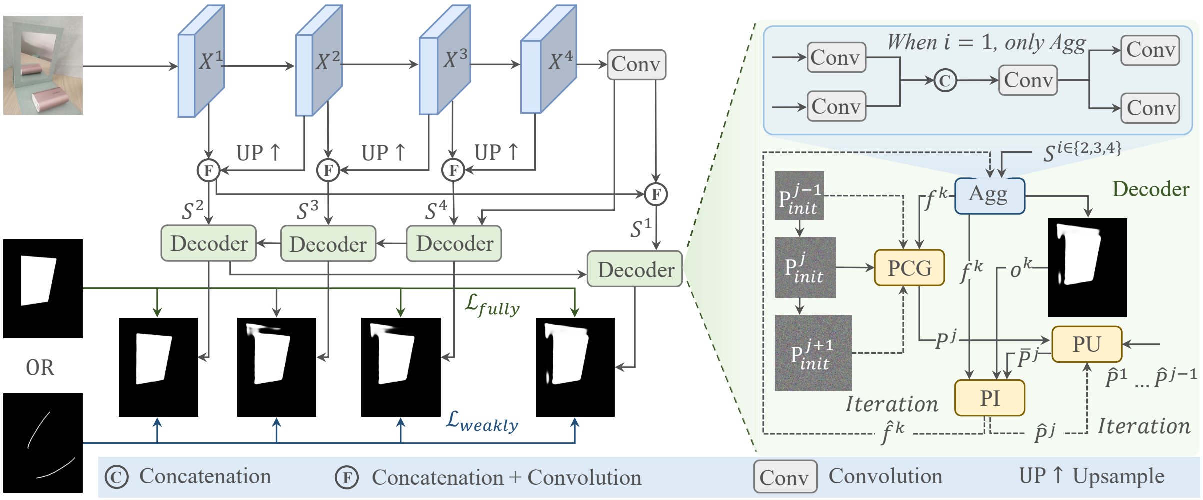

Figure 3: The overview of our IPR framework. We use encoder to extract multi-scale features, continuously upsample and aggregate. Then we apply the PCG module in the decoder to generate a prompt chain to perform association reasoning between entities and imagings. We further utilize the PU module to enhance the coupling among elements in the chain and reduce noise. Finally, we leverage the PI module to inject the refined prompts into the feature maps, which are decoupled using predicted masks, to improve the semantic difference between mirror and non-mirror regions.

e.g., scale changes, is hard to deal with reflections and object occlusions. Taking SOD as an example, if there are only imagings in the images, it may regard them as salient objects. Furthermore, general SS methods represented by SAM (Kirillov et al. 2023) also face these problems. Therefore, it can not directly apply to MD task.

Proposed Method

Overall Architecture

The overview of our IRP framework is shown in Figure 3, which consists of three core components: PCG, PU, and PI modules. For any given image $I \in \mathbb{R}^{3\times H\times W}$, we can obtain 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), where C, H, W denote channels, height and width respectively, $i \in \{0,1,2,3\}$. In the decoding stage, we randomly initialize the prompt parameter and generate a prompt chain $\{P\}^j \in \mathbb{R}^{C_j\times\frac{H}{4^{3-j}}\times\frac{W}{4^{3-j}}}$ through the PCG module, $j \in \{1,2,3\}, C_j = 64\times j$. We further leverage the PU module to update and enhance the prompt chain. For P^1, we fuse P^1 and the channel-shuffled P_s^1 to implicitly learn symmetric consistency. For P^2 and P^3, we fuse the noise-filtered P^1 and P^1, P^2 respectively. Finally, the PI module is used to inject the refined prompts into the feature maps and iterate the prompts.

Prompt Chain Generation Module

Unlike explicit prompt to acquire low-level detailed representations during the encoding stage, we aim to capture the contextual information of the mirror regions. During the decoding stage, we establish the associations and differences between entities and imagings through semantic awareness, thereby providing a unified solution to the problems in MD task, *i.e.,* reflection interference, occlusion, and scale variations. Furthermore, as the implicit space has a broader learning boundary than the explicit space, it is difficult to achieve spatial modeling through a single prompt. Inspired by the chain of thought reasoning in LLMs, we construct a visual prompt chain to search for the optimal prompt through multi-step reasoning. Specifically, we propose the PCG module, with the structural details shown in Figure 4.

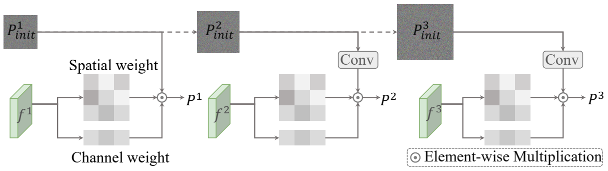

Figure 4: Structure of the PCG module. Generate prompts based on random parameters and feature maps.

We first randomly initialize the prompt learnable parameters $P_{init}^1 \in \mathbb{R}^{64\times\frac{H}{16}\times\frac{W}{16}}$, and construct a chain through transposed convolution (Dumoulin and Visin 2016) upsampling 2× and 4×, which can be represented as:

$$P_{init}^2 = Conv_{3\times3}^T(P_{init}^1), P_{init}^3 = Conv_{3\times3}^T(P_{init}^2) \quad (1)$$

where $Conv_{3\times3}^T$ denotes 3×3 transposed convolution.

We combine P_{init}^j and f^k based on pixel weights to dynamically perceive feature content and accelerate convergence. We apply global average pooling (GAP) and 1×1 convolution to obtain channel weights W_c, and utilize global max pooling (GMP), GAP and 7×7 convolution to obtain spatial weights W_s. Besides, we reshape P_{init}^j to ensure the

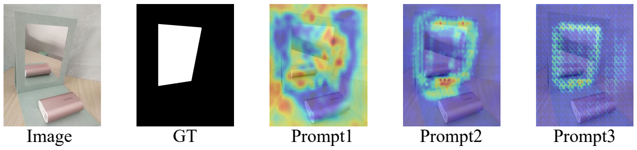

Figure 5: Visualization of different state visual prompt.

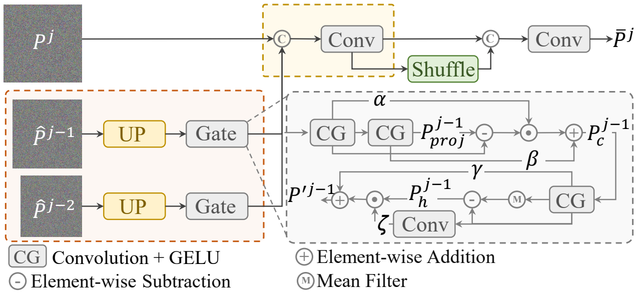

Figure 6: Structure of the PU module. Channel and frequency aggregations are applied as gate to refine prompts.

channel dimensions match. The process is formulated as:

$$W_c = Conv_{1\times1}(\sigma(Conv_{1\times1}(GAP(P^j_{init})))$$
$$W_s = Conv_{7\times7}(Concat(GAP(P^j_{init}); GMP(P^j_{init}))) \quad (2)$$

where $Concat(\cdot;\cdot)$ denotes channel concatenation and σ is the activation function. Thus, we can generate P^j by:

$$P^j = P^j_{init} \odot W_c \odot W_s \quad (3)$$

where $\odot$ denotes element-wise multiplication. The generated prompts share relevant representations and guide the spatial reflection perception.

As shown in Figure 5, we apply the generated prompts to the image, with the prompts gradually approaching the ground truth (GT).

Prompt Updating Module

Although the prompt chain generated by the PCG module has achieved overall reasoning, there is a lack of coupling between the individual prompts, which may lead to the forgetting of critical features. Inspired by dense connection (Huang et al. 2017), when $j > 1$, we fuse the previous $j-1$ prompts, and when $j = 1$, we keep the original prompt. However, the fusion of past prompts is inevitable to introduce noise and redundancy. Therefore, we refine them based on the gate mechanism. We propose the PU module, as shown in Figure 6.

Specifically, when $j = 1$, we leverage channel shuffle (Zhang et al. 2018) on P^1 to implicitly perceive mirror symmetry. We then fuse the original features and generate $\bar{P}^1$ through dimension reduction projection, which can be represented as:

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

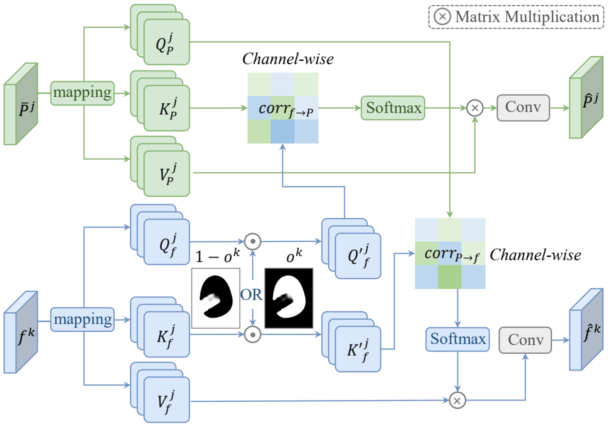

Figure 7: Structure of the PI module. Predicted mask is used to decouple mirror and non-mirror regions.

When $j > 1$, we first upsample the previous $j-1$ prompts to the size of P^j. For $j = 3$, we have:

$$P^2 = UP(P^2), P^1 = UP(P^1) \quad (5)$$

where $UP(\cdot)$ represents upsampling. For P^2, we aggregate and redistribute channel-wise features to enhance semantic knowledge by:

$$P^2_{proj} = \Gamma^1(\Gamma^2(P^2)) \quad (6)$$

$$P^2_c = \alpha \times \Gamma^1(P^2) \odot (\Gamma^1(P^2) - P^2_{proj}) + \beta \times P^2_{proj} \quad (7)$$

where $\Gamma(\cdot)$ denotes the CG module, α and β are learnable parameters.

Image high-frequency components contain detailed representations, *e.g.,* edges and textures, while low-frequency components highlight semantics. Thus we further filter noise in the frequency domain. Inspired by (Pan, Cai, and Zhuang 2022), we can generate high-frequency features by subtracting the low-frequency features obtained using the mean filter from the original features. The process is represented as:

$$P^2_h = \Gamma^3(P^2_c) - Mean(\Gamma^3(P^2_c)) \quad (8)$$

Then we can obtain P'^2 after frequency aggregation by:

$$P'^2 = \zeta \times (Conv(\Gamma^1(P^2_c))) \odot P^2_h + \gamma \times (\Gamma^3(P^2_c)) \quad (9)$$

where ζ and γ are learnable parameters. Similarly, we can generate P'^1. Finally, we fuse P'^1, P'^2 and P^3 and then channel-wise rearrange to generate $\bar{P}^3$. We also obtain $\bar{P}^2$.

Prompt Injection Module

The refined prompts through the PU module can provide guidance to locate mirror regions. To this end, we utilize the PI module to inject the prompts into the feature maps and iterate the prompts through interaction. To further enable prompts to perceive the differences between imagings and entities, we introduce predicted mask from the current level to decouple mirror and non-mirror regions. The detailed structure of the PI module is shown in Figure 7.

Specifically, following (Zamir et al. 2022), we use three sets of 1×1 and 3×3 depth-wise convolutions to encode local features and map to generate the query, key and value, *i.e.,* $\{Q,K,V\}_P^j$ and $\{Q,K,V\}_f^k$, where $k \in \{1,2,3\}$. For f^k, when $k > 1$, we fuse the mirror masks o^k to generate the mirror features by:

$$Q_f^{'k} = Q_f^k \odot o^k, K_f^{'k} = K_f^k \odot o^k \tag{10}$$

when $k = 1$, we have:

$$Q_f^{'k} = Q_f^k \odot (1 - o^k), K_f^{'k} = K_f^k \odot (1 - o^k) \tag{11}$$

Therefore, we can further generate the correlation matrix $corr_{P\to f}$ from prompts to feature maps by:

$$corr_{P\to f} = softmax(\frac{Q_P^j K_f^{'k}}{\tau}) \tag{12}$$

where τ is a learnable scaling factor. Similarly, we can obtain the correlation matrix of feature maps to prompts $corr_{f\to P}$. Unlike previous work (Huang et al. 2023a) that models feature correlations from the spatial dimension, we consider it from the channel perspective based on two reasons: 1) Channels better capture semantic information, which is more aligned with the original intention of implicit prompt. 2) Less computational cost. We observe that channel-based modeling requires at least half the GPU memory compared to spatial modeling, while maintaining comparable performance.

Finally, we can obtain the feature map $\hat{f}^k$ with injected prompts by:

$$\hat{f}^k = Conv(corr_{P\to f} \otimes V_f^k) \tag{13}$$

where $\otimes$ denotes matrix multiplication. Similarly, we can also obtain the updated prompt $\hat{P}^j$.

Loss Function

We apply supervision to all the predicted maps generated by the decoder. For the fully-supervised setting, following (He, Lin, and Lau 2023), we employ weighted binary cross-entropy loss, *i.e.,* $\mathcal{L}_{BCE}^W$ and weighted intersection over union loss, *i.e.,* $\mathcal{L}_{IoU}^W$ to emphasize difficult and critical pixels, which can be expressed as:

$$\mathcal{L}_{fully} = \sum_{i=1}^{4} \mathcal{L}_{BCE}^W + \mathcal{L}_{IoU}^W \tag{14}$$

For the weakly-supervised setting, following (Zha et al. 2024), we utilize partial cross-entropy loss, *i.e.,* $\mathcal{L}_{CE}^P$ and smooth loss, *i.e.,* $\mathcal{L}_S$, which can be represented as:

$$\mathcal{L}_{weakly} = \sum_{i=1}^{4} \mathcal{L}_{CE}^P + \mathcal{L}_S \tag{15}$$

Note that we do not employ edge supervision or contrastive loss. For simplicity, we do not set hyperparameters to balance the losses.

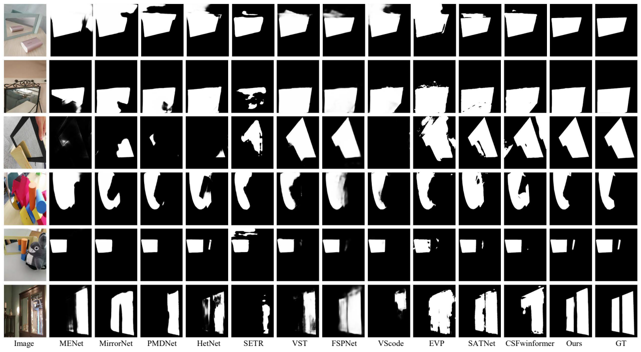

Figure 8: Qualitative comparison on the MD datasets.

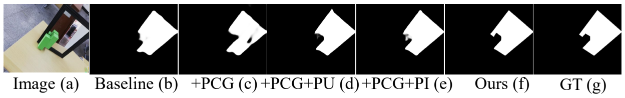

Figure 9: Qualitative ablation of different modules.

Experiments

Datasets. We conduct experiments on eleven datasets. **Mirror datasets:** MSD and PMD (Lin, Wang, and Lau 2020) datasets contain 3,063 and 5,096 training images, and 955 and 571 testing images, respectively. Mirror-RGBD (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. **Glass dataset:** GDD (Mei et al. 2020) contains 2,980 training images and 936 testing images. **Camouflage datasets:** CAMO (Le et al. 2019) has 1,000 and 250 images for training and testing, and COD10K (Fan et al. 2020) includes 3,040 training images and 2,026 testing images. NC4K (Lv et al. 2021) contains 4,121 images, used only for testing. **Underwater datasets:** MAS3K (Li et al. 2020), RMAS (Fu et al. 2023), and UFO120 (Islam, Luo, and Sattar 2020) contain 1,769/2,514/1,500 training images, and 1,141/500/120 testing images, respectively.

Implementation Details. We implement the framework based on PyTorch and conduct all experiments on an NVIDIA A100 GPU. Following (Zha et al. 2024), we utilize the PVT network pretrained on ImageNet as the backbone. For fair comparison, the inputs are resized to 384×384 for both training and testing across all datasets. For training, we set the batch size to 40, the initial learning rate to 1e-4, and use AdamW to optimize with 200 epochs. For testing, we do not employ any post-processing tricks (*e.g.,* CRF) to refine the prediction results.

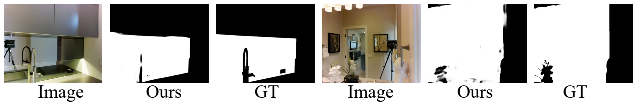

Figure 10: Failure cases.

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											
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
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
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
SAM $_{\text{ICCV'23}}$	G	0.124	–	–	–	0.515	0.052	–	–	–	0.647
DualSAM $_{\text{CVPR'24}}$	G	0.039	0.903	0.932	0.882	0.848	0.034	0.816	0.839	0.705	0.636
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	**0.026**	**0.923**	**0.956**	**0.919**	**0.900**	0.026	**0.846**	**0.902**	**0.771**	**0.728**
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	**0.066**	**0.841**	**0.895**	**0.804**	**0.770**	**0.049**	**0.775**	**0.825**	**0.646**	0.598

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	✗	**0.042**	0.851	**0.906**	0.820	0.773
WSMD$_{\text{AAAI'24}}$	✗	0.088	0.754	0.806	0.655	0.616
Ours*	✗	**0.082**	**0.760**	**0.829**	**0.674**	**0.631**

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
DualSAM	512×512	325.68	159.95
SATNet	512×512	153.00	139.36
CSFwinformer	512×512	139.45	150.54
Ours	384×384	**16.30**	**27.66**
WSMD	352×352	21.39	**26.16**
Ours*	352×352	**13.71**	27.66

Table 3: Model Efficiency Comparison. We compare with five 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	**27.66**	**0.095**	**0.784**	**0.820**	**0.710**	**0.666**

Table 4: Quantitative comparison on VMD dataset.

Evaluation Metrics. We adopt six evaluation metrics: Mean Absolute Error (MAE), S-measure (S_m), mean E-measure (E_m), weighted F-measure (F_β^w) (Fan et al. 2017, 2018), and intersection over union (IoU). The higher the better for the last four. Note that all evaluation data is calculated through the prediction results provided in the original papers or by retraining using the official codes.

Comparison with SOTA Methods

Quantitative Comparison. We select different SOTA methods with different backbones (based on CNN or Transformer) and different settings to compare and validate the superiority of the IPR framework. Specifically, as shown in Table 1, for fully supervised setting, we select five MD methods: MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), HetNet, SATNet, and CSFwinformer (Xie et al. 2024); three SOD methods: MENet (Wang et al. 2023a), 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; and two general segmentation methods: SAM, DualSAM (Zhang et al. 2024), and EVP. Our approach surpasses various types of methods, especially EVP, with gains of 3.8%, 7.8%, 6.0%, 10.8%, and 12.0% respectively on the five metrics of the MSD dataset, and the average surpasses CSFwinformer by around 5.0%, which demonstrates the effectiveness of implicit prompt learning. As shown in Table 2, we select RGB-D SOD method, *i.e.,* PICRNet (Cong et al. 2023a), MD-specific method, *i.e.,* PDNet (Mei et al. 2021), and our method achieve comparable performance to PDNet without using depth modal assistance. For weakly supervised setting, our method also surpassed WSMD. As shown in Table 3, the FLOPs and parameters of the IPR framework are about one-tenth and one-fifth

Methods	Params.	MAE$\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	IoU$\uparrow$
RFENet$_{\text{IJCAI'23}}$	152.65	0.061	0.858	0.913	0.901	0.882
Ours	**27.66**	**0.048**	**0.887**	**0.933**	**0.924**	**0.904**

Table 5: Quantitative comparison on GDD dataset.

Methods	Params.	CAMO			COD10K			NC4K		
		$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$
EVP$_{\text{CVPR'23}}$	64.52	0.846	0.895	0.777	0.843	0.907	0.742	0.874	‡	‡
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	**27.66**	0.854	**0.912**	**0.811**	**0.860**	**0.931**	**0.781**	**0.879**	**0.929**	**0.835**
Methods	**Params.**	**MAS3K**			**RMAS**			**UFO120**		
		$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$
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	**27.66**	**0.903**	**0.865**	**0.948**	**0.882**	**0.838**	**0.956**	**0.865**	**0.870**	**0.916**

Table 6: Quantitative comparison on camouflage and underwater scenarios.

of CSFwinformer, respectively, demonstrating its efficiency. As shown in Tables 4, 5, and 6, the SOTA methods for each task (*e.g.,* VMDNet (Lin, Tan, and Lau 2023) and RFENet (Fan et al. 2023)) are selected to validate the generalization.

Qualitative Comparison. We consider different scenarios and provide visualizations for comparison. As shown in Figure 8, the first two rows represent imagings with and without corresponding entities. The third and fourth rows demonstrate regular and irregular occlusions. Our approach establish associations between entities and imagings as well as context around imagings. The last two rows show scale variations and multiple targets. Our method can enable global modeling to avoid false negative and false positive.

Ablation Study

We validate the effect of proposed modules (Table 7 and Figure 9), mask guidance (Table 8), and number of prompts (Table 9) on the MSD dataset.

Effect of the PCG module. We add the PCG module to the *Baseline*, obtaining improvements of 1.3%, 1.1%, 2.3%, and 4.0% on the four metrics, respectively. The MAE decreases by 1.1%. As shown in Figure 9 (b) and (c), the occlusion problem is alleviated, but it introduces some noises, leading to false detection in the mirror region. Besides, leveraging concatenation strategy proves to be suboptimal as it lacks sufficient semantic and detailed interaction between prompts and feature maps and introduces noise interference. We incorporate the PU and PI modules into the *Baseline+PCG* model to further improve performance.

Effect of the PU module. We incorporate the PU module based on the *Baseline+PCG* model. As shown in Figure 9 (d), false detection is alleviated, and occlusion impact is further mitigated. This indicates that the PU module effectively filters out noise and enhances the inter-prompt correlations, generating a clean and informative chain.

PCG	PU	PI	MAE$\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	IoU$\uparrow$
			0.047	0.875	0.908	0.851	0.819
✓			0.036	0.888	0.919	0.874	0.859
✓	✓		0.030	0.910	0.936	0.895	0.880
✓		✓	0.033	0.899	0.929	0.886	0.869
✓	✓	✓	**0.026**	**0.923**	**0.956**	**0.919**	**0.900**

Table 7: Quantitative ablation of proposed modules. When not using the PI module, we utilize channel concatenation to inject prompts. The first line is the *Baseline* model.

M^1	M^2	M^3	MAE$\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	IoU$\uparrow$
✓			0.029	0.915	**0.958**	0.908	0.892
✓	✓		**0.026**	**0.923**	0.956	**0.919**	**0.900**
✓	✓	✓	0.031	0.908	0.941	0.901	0.885

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

N	0	1	2	3	4	5
S_m	0.875	0.898	0.907	0.923	0.926	0.929
FLOPs	14.96	15.09	15.67	16.30	17.02	17.65

Table 9: The impact of different prompts.

Effect of the PI module. Similarly, by adding the PI module, we surpass the *Baseline+PCG* model but fall short of the *Baseline+PCG+PU* model, indicating that accurate and strongly coupled prompts (chain) are the most crucial, while injection strategy is supportive but also essential. As shown in Figure 9 (e), the detected mirror region is not complete.

Mirror mask. In different stages of the PI module, we fuse feature maps and mirror or non-mirror region masks to enhance perceptual differences. With the increase of mirror masks, performance gradually improves. However, when used in all stages, performance declines sharply. Therefore, it is not enough to only perceive mirror regions; non-mirror regions are also necessary to construct complete context.

Number of prompts. When $N = 0$, there are no PU and PI modules, *i.e., Baseline* model. As N increases, both S_m and FLOPs gradually improve. However, when $N \geq 4$, the growth rate of S_m and FLOPs starts to decrease. Although we can achieve higher performance, it also introduces significant computational burden. Therefore, $N = 3$ is the optimal balance point between performance and efficiency.

Failure Cases Analysis and Broader Impacts

As shown in Figure 10, our method performs poorly when faced with highly irregular occlusions. The proposed approach is not limited to MD task and is applicable to various scenarios, such as camouflage. Please refer to the supplementary material for more details.

Conclusion

We rethink the existing paradigms for MD task and propose the IRP framework. Unlike previous works that rely on explicit rotation or visual prompts, our paradigm allows for effective perception of the representation differences between entities and imagings, leading to accurate and complete localization of mirror regions. Our framework outperforms SOTA methods on eleven benchmarks, with lower computational complexity.

References

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, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

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

Li, L.; Zhang, Y.; and Chen, L. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4): 1–26.

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.

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.

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, 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. 2023a. 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.

Wang, Y. R.; Zhao, Y.; Xu, H.; Eppel, S.; Aspuru-Guzik, A.; Shkurti, F.; and Garg, A. 2023b. Mvtrans: Multi-view perception of transparent objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 3771–3778. IEEE.

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.

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*.

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.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

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, 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.

Zhu, M.; Li, H.; Chen, H.; Fan, C.; Mao, W.; Jing, C.; Liu, Y.; and Shen, C. 2023. Segprompt: Boosting open-world segmentation via category-level prompt learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 999–1008.

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)