Spider🕷: A Unified Framework for Context-dependent Concept Segmentation

Xiaoqi Zhao [* 1 2] **Youwei Pang** [* 1 2] **Wei Ji** [* 3] **Baicheng Sheng** [1] **Jiaming Zuo** [2] **Lihe Zhang** [1] **Huchuan Lu** [1]

Abstract

Different from the context-independent (CI) concepts such as human, car, and airplane, context-dependent (CD) concepts require higher visual understanding ability, such as camouflaged object and medical lesion. Despite the rapid advance of many CD understanding tasks in respective branches, the isolated evolution leads to their limited cross-domain generalisation and repetitive technique innovation. Since there is a strong coupling relationship between foreground and background context in CD tasks, existing methods require to train separate models in their focused domains. This restricts their real-world CD concept understanding towards artificial general intelligence (AGI). We propose a unified model with a single set of parameters, Spider, which only needs to be trained once. With the help of the proposed concept filter driven by the image-mask group prompt, Spider is able to understand and distinguish diverse strong context-dependent concepts to accurately capture the Prompter's intention. Without bells and whistles, Spider significantly outperforms the state-of-the-art specialized models in 8 different context-dependent segmentation tasks, including 4 natural scenes (salient, camouflaged, and transparent objects and shadow) and 4 medical lesions (COVID-19, polyp, breast, and skin lesion with color colonoscopy, CT, ultrasound, and dermoscopy modalities). Besides, Spider shows obvious advantages in continuous learning. It can easily complete the training of new tasks by fine-tuning parameters less than 1% and bring a tolerable performance degradation of less than 5% for all old tasks. The source code will be publicly available at Spider-UniCDSeg.

*Equal contribution [1]Dalian University of Technology, China [2]X3000 Inspection Co., Ltd, China [3]Yale University, America. Correspondence to: Lihe Zhang <zhanglihe@dlut.edu.cn>.
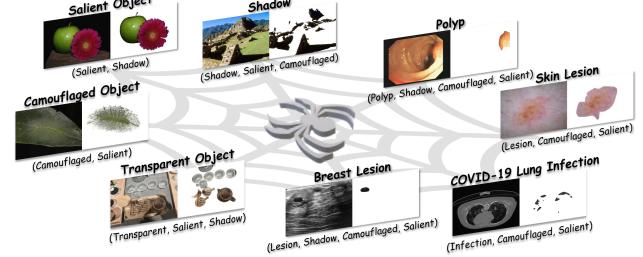
Figure 1. Eight different segmentation tasks with context-dependent concepts are unified into our Spider model. With the interlaced concepts within task domains and class semantic space, Spider can wander to any target of interest.

1. Introduction

In philosophy and cognitive science (Barsalou, 1982; Martial et al., 2018; Lachmann & Van Leeuwen, 2005), concepts usually contain the context-independent (CI) and context-dependent (CD) concepts. For example, semantic segmentation tasks define numerous CI concepts (i.e., multi-granularity semantic classes). The class of an object is fixed no matter what scene it locates in. The context-dependent concepts mean that the target is not cognizable without its environment, such as salient/camouflaged object detection, shadow detection, medical lesion segmentation, as shown in Figure 1. People determine where the target locates mainly according to the surroundings. In this work, we focus on the context-dependent segmentation tasks and expect to build a parameter-unified framework. Existing works explore the in-domain modeling, resulting in repetitive structure design, inefficient data utilisation, and limited multi-domain generalisation. As an alternative, a seminal work of EVP (Liu et al., 2023b) attempts to unify three CD tasks based on low-level structure, but it still requires to train the model one by one for each task and lacks the unification of parameters.

With the development of strong backbones like ConvNeXt (Liu et al., 2022) and ViT(Dosovitskiy et al., 2020), and visual prompt technology (Yan et al., 2023; Potlapalli et al., 2023; Wang et al., 2023b;c), some unification models appear towards the attempt of AGI. In the field of segmentation, generalist models typified by SAM (Kirillov et al., 2023) and SegGPT (Wang et al., 2023c), have shown gratifying versatility. They rely on the visual prompt of a single

pair of image and foreground to understand CI concepts. However, many reports (Tang et al., 2023b; Huang et al., 2023a; Ji et al., 2023b; Chen et al., 2023; Zhou et al., 2023a; Liu et al., 2023c; Ji et al., 2024) show their poor performance on CD concept understanding in terms of both quantitative and qualitative evaluations. It is because the targets in the CD tasks have not the fixed semantic classes, and multiple CD concepts often mingle together in semantic space (Please see Figure 1, in which one target exhibits multiple CD properties and one CD concept covers multiple semantic classes), thus the prompt of a single foreground fails to provide explicit guidance for the segmentation model. To design a parameter-unified context-dependent segmentation model, the first critical issue is how to build a compatible pipeline to understand and distinguish each CD concept. The second one is how to overcome the challenges posed to cross-domain collaborative learning because of the overlapped semantic classes across the foregrounds of different tasks and the large gap among different task domains.

In this paper, we propose a unified CD concept segmentation framework **Spider**, which shares 100% parameters for all tasks based on the idea of dynamic filtering. For structure unification, Spider equips with a segmentation stream and a concept prompt stream. When facing each task, the prompt stream generates a concept filter to act on the tail of the segmentation stream, thereby yielding a unified single-channel output. For parameter unification, the segmentation stream is responsible for learning task-generic representations from cross-domain data through a single set of encoder-decoder parameters. This stream can be substituted to any specialized models.

Previous unified segmentation models (Kirillov et al., 2023; Wang et al., 2023c; Butoi* et al., 2023; Lüddecke & Ecker, 2022) adopt non-local/co-attention or element-wise spatial operations to propagate foreground prompt knowledge to current input features at pixel level, as shown in Figure 2. Because context-dependent targets have uncertain categories, pixel-level feature fusion with foreground prompt embeddings easily produces ambiguities in prompt definitions and is more sensitive to the accuracy of mask annotation. Different from them, we utilize a group of image, foreground and background prompts to comprehensively mine the clues of CD targets and establish high-level image-foreground matching and image-background matching to achieve feature interaction between visual prompt and current input by the concept filter.

Specifically, we utilize the transformer (Vaswani et al., 2017) to establish long-range dependence of concepts and environments within a group of images. Image-group prompts are encoded as the key and value. Both foreground and background mask-group prompts are embedded and encoded as the query. Through multiple cross-attention operations,
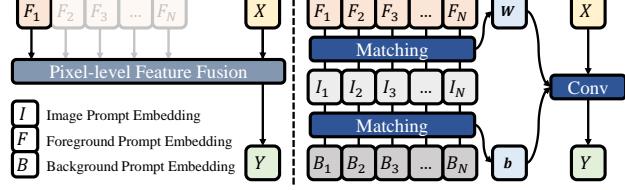

Figure 2. Two types of feature interaction between visual prompts and current image input. The left is used in Universeg (Butoi* et al., 2023) with multiple foreground prompts and others (Wang et al., 2023c; Kirillov et al., 2023; Lüddecke & Ecker, 2022) with a single foreground prompt. The right is ours.

the model can learn the common conceptual expressions in the group of images indicated by their masks. The updated object embeddings and context-aware feature embeddings are used as the weight and bias of concept filter, respectively. With the help of the concept filters, Spider can wander across different task domains and establish the connections among these CD concepts. It supports customized prompts during inference and has the potential of perceiving unseen context-dependent objects. More advantages of the concept filter have been summarized in the Appendix A.1. In addition, we devise a "Balance FP - Unify BP" training strategy to balance different tasks in both forward-propagating batch specification and back-propagating gradient update, thereby guaranteeing the performance in all tasks.

Our main contributions can be summarized as follows:

- We propose a unified model, Spider, which only needs to be trained once and can perform complex context-dependent concept understanding in diverse domains.
- Benefiting from the flexible concept filter, Spider can sensitively perceive the attributes or categories of interest, so that it can be trained on different domains without heavy task-specific heads.
- Spider achieves superior performance in 8 challenging tasks with context-dependent concepts, and it has powerful continuous learning abilities. It can be generalized to new tasks by fine-tuning parameters of less than 1% without any structural modifications, and preserve the performance on old tasks with slight degradation of less than 5%. As the scale and diversity of training data increase, it shows the potential in unseen tasks.

2. Related Work

2.1. Context-dependent Image Segmentation

As shown in Figure 3, the context-independent concept can be well understood without the help of its contexts. While the context-dependent concept is the complete opposite, we have to rely on its surroundings for a clear understanding. In this work, we choose eight representative tasks, including four natural scene tasks and four different modality medical
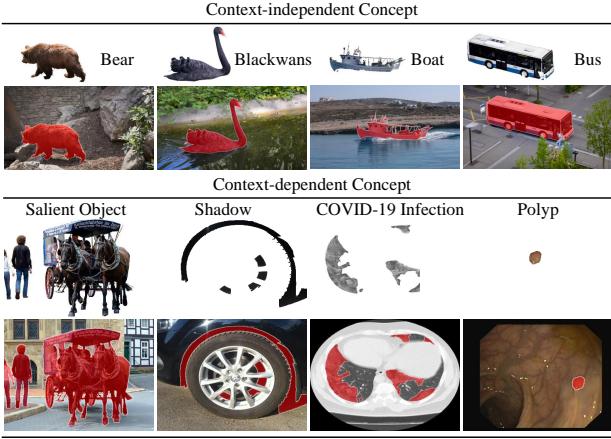
Figure 3. Visual comparison of segmentation objects with context-independent concepts and context-dependent concepts. The odd rows are the pure foregrounds and the even rows are the complete images with the highlighted foregrounds.

tasks, to investigate the unified modeling for the context-dependent concept segmentation. Detailed definition of these context-dependent segmentation tasks can be found in the Appendix A.2. Among these context-dependent segmentation tasks, the U-shaped structure (Ronneberger et al., 2015; Lin et al., 2017) with the encoder-decoder form is the most basic framework. According to the characteristics of different tasks, existing methods mainly focus on four aspects: visual attention (Zhao et al., 2020a; Zhu et al., 2018; Zhang et al., 2020; Kim et al., 2021; Li et al., 2023b; 2021; Ji et al., 2023c), multi-scale feature extraction (Pang et al., 2020b; 2022b; Wu et al., 2019; Takahashi & Mitsufuji, 2021; Zhao et al., 2020c; Li et al., 2023a; Ji et al., 2022; Piao et al., 2019), edge refinement (Sun et al., 2022; Jia et al., 2022; Lin et al., 2020; Zhou et al., 2023b; Lin et al., 2021), and combination of different architectures (Zhang et al., 2021; Xie et al., 2021a; Yang et al., 2021; Liu et al., 2021a; Ji et al., 2021). In this work, we consider the cross-domain learning and design an efficient and unified framework with only one set of parameters and one training session.

2.2. Unified Vision Models

With the development of foundation models (Liu et al., 2021b; Dosovitskiy et al., 2020; Liu et al., 2022; Wang et al., 2023a; Zhang et al., 2022a), solving multiple vision tasks by a single set of model parameters has become an important way to move towards AGI. Previous typical parameter-unified methods (He et al., 2017; Cheng et al., 2022; Vandenhende et al., 2020; Zhao et al., 2022a) are mainly based on multi-task learning. They design multiple task-specific heads/decoders for different tasks, such as object detection and panoptic/instance/semantic segmentation as in (He et al., 2017; Cheng et al., 2022), and SOD and depth/edge estimation as in (Vandenhende et al., 2020; Zhao

et al., 2022a). Because each sample has multiple annotations corresponding to different tasks at the same time, all these studies are performed in-domain and handle different tasks by different heads. However, in the real world, different tasks focus on different objects of interest, data domains, and annotation types. Therefore, the cross-domain learning is a key paradigm to unify model parameters. A simple query-based task formulation is proposed in (Ci et al., 2023) for handling multiple distinctly defined human-centric tasks. Ten instance perception tasks are unified into a prompt-guided object discovery and retrieval paradigm in (Yan et al., 2023). Input-conditioned prompts with the contextual information in (Potlapalli et al., 2023) is designed to guide different image restoration tasks. In (Wang et al., 2023b), the task-specific input-output image pair is used as condition to perform ten different dense prediction tasks. In terms of image segmentation, CLIP-driven universal models (Liu et al., 2023a; Lüddecke & Ecker, 2022) incorporate text embedding to provide the models with different semantic prompts. UniverSeg (Butoi* et al., 2023) employs feature fusion with the query image and example set of image-label pairs to achieve universe medical image segmentation. SAM (Kirillov et al., 2023) designs a powerful segmentation architecture equipped with the reusable image embedding and a orientated prompt branch. In SegGPT (Wang et al., 2023c), the image segmentation is formulated as an in-context coloring problem, which requires a image-mask pair prompt to indicate object segmentation. However, the motivation of these unified methods is oriented to CI concepts. They only focus on the single pair of image-foreground prompt or embedding the prompt knowledge by the pix-level feature fusion. Many works (Tang et al., 2023b; Liu et al., 2023c; Chen et al., 2023; Zhou et al., 2023a; Ji et al., 2023a) report that the current unified/generalist methods are still difficult to handle diverse strong context-dependent segmentation tasks during training and inference. In this work, we propose a simple unified architecture guided by image-mask (foreground and background) group prompts for eight CD concept tasks with multiple modalities.

3. Approach

To furthest share knowledge among various tasks and reduce specialized designs, we attempt to maximize weight sharing. As shown in Figure 4, we directly utilize the general encoder-decoder architecture without any modifications, i.e., a vanilla UNet (Ronneberger et al., 2015) or FPN (Lin et al., 2017) with different backbones. Our core component is the concept filters generated by image and mask prompt streams $\mathcal{S}_i$ and $\mathcal{S}_m$, which are embedded in the final dynamic head to accurately predict different tasks. In this way, all feature extractions and fusion weights absorb multi-domain information and share 100% parameters across all tasks.
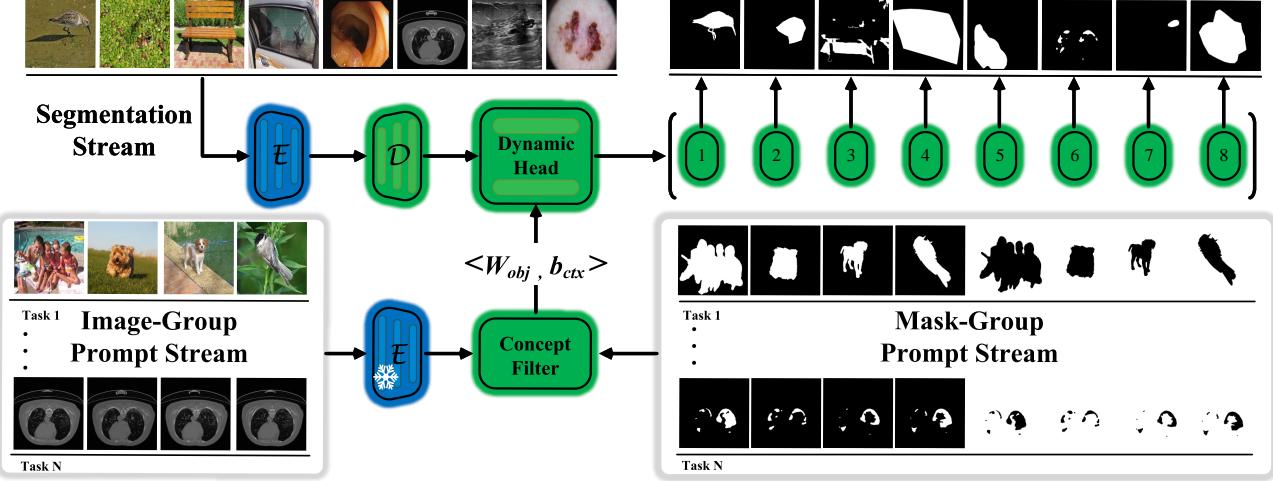
Figure 4. Overall pipeline. It consists of segmentation stream $\mathcal{S}_s$, and image- and mask- group prompt streams $\mathcal{S}_i$ and $\mathcal{S}_m$. $\mathcal{S}_s$ uses the encoder-decoder structure. $\mathcal{S}_i$ is fed into the frozen pre-trained encoder and output the group prompt feature F_{mem} as the key and value of the transformer decoder. $\mathcal{S}_m$ generates the foreground-aware and background-aware queries by masked average pooling on the group prompt features F_{mem}. A series of concept filters $< W_{obj}, b_{ctx} >$ act on the last layer of the decoder to generate dynamic prediction.

3.1. Prompt Generation

The prompt generation strategy is different in the training and inference period. During training, firstly, G pairs of images and masks are randomly selected from each task at each iteration as group prompts. This manner of random combination ensures the performance stability of the concept filter when facing different group prompts in practical applications, and its motivation and effects are similar to those of the masked image modeling (MIM) mechanism (Bao et al., 2021; He et al., 2022; Xie et al., 2022b; Wang et al., 2023c). Next, we pass the image group to the frozen pre-trained encoder **E** to obtain rich high-level semantic features. Finally, the concept filters $< W_{obj}, b_{ctx} >$ derived from the image-group features F_g and mask-group maps M_{fg} and M_{bg}, participate in the convolutional dynamic head for prediction. During inference, to ensure stability and replicability of predictions, we look through all training samples and select 64 representative examples for each task as its fixed group prompt by K-means clustering over their high-level embeddings. Specifically, we first fed all images to the encoder of prompt stream. The extracted high-level feature maps are global average pooled to condense semantic information and reduce the computational complexity in the clustering process. Next, we randomly generate 64 initial clustering centers, iteratively cluster high-level embeddings based on the similarity, and update these centers until convergence. Finally, we select the samples closest to cluster center as the image-group prompts. Quantitative results can be seen in Table 3 and the visualization of clustered group prompt for each concept is presented in the Appendix A.6. It is worth noting that group prompts can be flexibly provided by users and are not limited to these clustered prompts in practical applications.

3.2. Concept Filter

This component is the key. It unifies multiple tasks into a single framework through the idea of conditional filtering. The details are shown in Figure 5. Specifically, we use the learnable projection matrix W_{proj} to transform the deep representations F_g of the image-group prompt and obtain the group prompt feature F_{mem}. The foreground mask group M_{fg} and background mask group M_{bg} corresponding to the targets of interest in the image-group prompt guide to yield foreground descriptor D_{fg} and background descriptor D_{bg} by masked average pooling. This extracts rough representations about foreground and background from F_{mem} specific to the contexts of current task. We further refine the two descriptors by mining foreground/background related semantic cues in the global context from appearance-driven F_{mem}. This process is achieved through multi-head cross-attention (MHCA). In MHCA, D_{fg} and D_{bg} separately act as X, which is further linearly transformed to Q. And F_{mem} serving as Y are mapped to K and V as well. The foreground or background activation map M is computed as:

$$M = \texttt{softmax}(\frac{QK^\top}{d}), \quad (1)$$

where d is a normalization factor. M is exploited to aggregate contextual information from V.

$$\begin{aligned} Z &= X + MVW_Z, \\ X &= Z + \texttt{FFN}(Z), \end{aligned} \quad (2)$$

where W_Z is the learnable weight. After the cascaded FFN, the resulting foreground and background descriptors are taken as object-aware weight W_{obj} and context-aware bias b_{ctx} of concept filter, respectively.
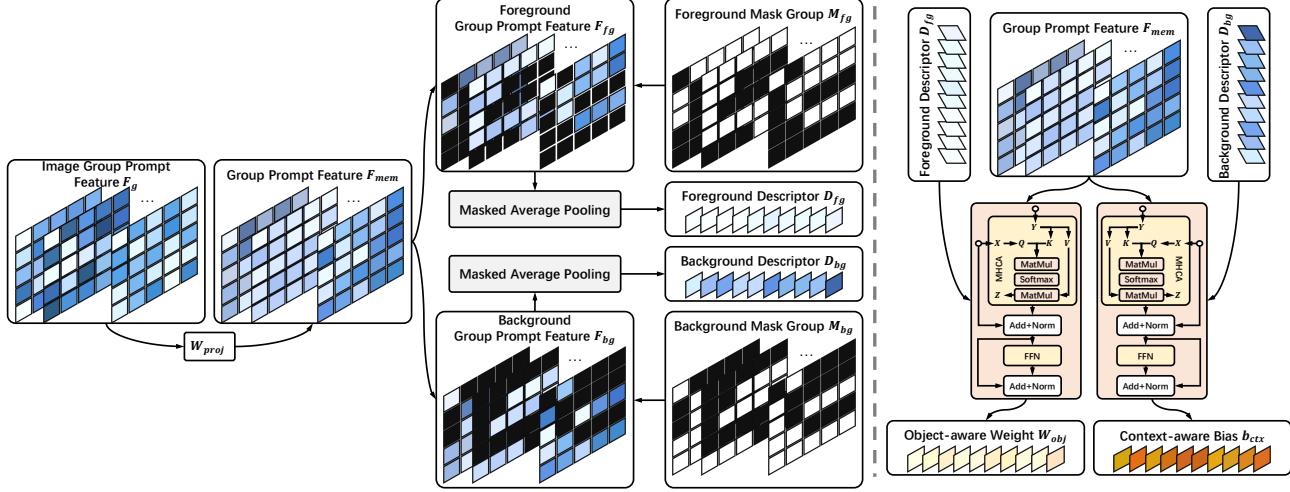
Figure 5. Illustration of generating concept filters.

Algorithm 1 Training and Inference

Training iteration with $N = 16$ and $B = 4$.

Require: A batch $D = \{D_t\}_{t=1}^8$. D_t is N image-mask pairs randomly selected from training set of task t.

1: create image tensor $I \in \mathbb{R}^{8 \times N \times 3 \times H \times W}$ from D
2: create mask tensor $M \in \mathbb{R}^{8 \times N \times 1 \times H \times W}$ from D
3: **for** $i \leftarrow 0, N/B - 1$ **do**
4:    $f \leftarrow \mathbf{D}(\mathbf{E}(I[:, iB : (i+1)B, ...]))$ // generate image feature
5:    $L \leftarrow 0$
6:    **for** $t \leftarrow 1, 8$ **do**
7:       $P_t^I \leftarrow Cat(I[t, 0 : iB, ...], I[t, (i+1)B : N, ...])$ // image group prompt
8:       $P_t^M \leftarrow Cat(M[t, 0 : iB, ...], M[t, (i+1)B : N, ...])$ // mask group prompt
9:       $< W_{obj}, b_{ctx} > \leftarrow \texttt{PromptStream}(P_t^I, P_t^M)$
10:      $P_t \leftarrow \texttt{DHead}(f[t, ...], < W_{obj}, b_{ctx} >)$
11:      $L \leftarrow L + \texttt{Loss}(P_t, M[t, iB : (i+1)B, ...])$
12:    **end for**
13:    $\texttt{backward}(L)$
14: **end for**   // update parameters

Inference with the minibatch of N.

Require: concept filter set $\{< W_{obj}, b_{ctx} >_t\}_{t=1}^8$, input tensor $I \in \mathbb{R}^{N \times 3 \times H \times W}$, task indicator $t \in \{1, 2, ..., 8\}^N$.

Ensure: prediction $Y \in \mathbb{R}^{N \times 1 \times H \times W}$

1: $f \leftarrow \mathbf{D}(\mathbf{E}(I))$
2: **for** $n \leftarrow 1, N$ **do**
3:    $Y[n, ...] \leftarrow \texttt{DHead}(f, < W_{obj}, b_{ctx} >_{t(n)})$
4: **end for**   // traverse all images

3.3. Training and Inference

To simultaneously balance the performance of these tasks in both forward propagation and back propagation during training, we design a **"Balance FP - Unify BP"** strategy. Specifically, we first randomly select N samples for each task, of which B samples are input to the segmentation

Table 1. Data partition in eight tasks, which is widely used by the state-of-the-art specialized methods.

Task	Dataset	#Train	#Test
Salient Object Detection (SOD)	DUTS (Wang et al., 2017)	10,548	5,017
Camouflaged Object Detection (COD)	COD10K (Fan et al., 2020a)	4,040	2,026
Shadow Detection (SD)	SBU (Vicente et al., 2016)	4,085	638
Transparent Object Segmentation (TOS)	Trans10K (Xie et al., 2020)	5,000	4,428
Colon Polyp Segmentation (CPS)	Five datasets (Fan et al., 2020b)	1450	798
COVID-19 Lung Infection (CLI)	COVID-19 data (Fan et al., 2020c)	894	383
Breast Lesion Segmentation (BLS)	BUSI (Al-Dhabyani et al., 2020)	486	161
Skin Lesion Segmentation (SLS)	ISIC18 (Codella et al., 2019)	1,886	808

branch and the rest $N - B$ samples are used as prompts. The samples of all tasks will be concatenated together to separately obtain the input tensor of segmentation stream and prompt stream with a shape of $[8B, C, H, W]$ and $[8(N - B), C, H, W]$. During the forward propagation, the batch normalization (Ioffe & Szegedy, 2015) can make the input distribution of each task closer, which helps the model learn task-shared representations, improving and **balancing** the overall performance. Moreover, we circularly generate eight concept filters in the tail to complete the predictions for the corresponding tasks, avoiding repeated computation caused by full forward propagation. Next, we use the PPA loss (Wei et al., 2020; Fan et al., 2020b; Wang et al., 2023d; He et al., 2023; Xie et al., 2022a; Fan et al., 2020c; Liu et al., 2023b) widely adopted in segmentation tasks to jointly calculate the loss of all samples. During the back propagation, the direction of parameter optimization is **unified** to help Spider obtain better overall performance without favoring a single task. In the inference phase, the input of the segmentation stream supports splicing multiple samples in the batch dimension. We may flexibly assign the concept filters to them for single or multiple concept predictions. Each concept filter receives a group of customized or fixed prompts from the training set as mentioned in Section 3.1. The detailed training and inference process can be found in Algorithm 1.

Table 2. Quantitative comparisons with the unified models and state-of-the-art specialized models on the eight tasks. ↑ and ↓ indicate that the larger scores and the smaller ones are better, respectively. The best scores are highlighted in **red**. Following SegGPT (Wang et al., 2023c), we first adopt ViT-B/L as the encoder of Spider. To facilitate future research comparisons, we further provide the Swin-B/L and ConvNeXt-B/L versions. Our largest version, Spider-ConvNext-L, has the same 1.5G model size as SegGPT.

| Method	Publication	Backbone	Salient DUTS $F_\beta^\omega \uparrow$	$S_m \uparrow$	Camouflaged COD10K $F_\beta^\omega \uparrow$	$S_m \uparrow$	Shadow SBU BER↓	MAE↓	Transparent Trans10K BER↓	MAE↓	Polyp 5 datasets mDice↑	mIoU↑	COVID-19 COVID-19 CT mDice↑	mIoU↑	Breast BUSI mDice↑	mIoU↑	Skin ISIC2018 mDice↑	mIoU↑
Specialized Models																		
MENet (Wang et al., 2023d)	CVPR'23	ResNet-50 (He et al., 2016)	0.8698	0.9028	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PGNet (Xie et al., 2022a)	CVPR'22	Swin-B (Liu et al., 2021b)	0.8736	0.9091	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FEDER (He et al., 2023)	CVPR'23	ResNet-50 (He et al., 2016)	-	-	0.7155	0.8196	-	-	-	-	-	-	-	-	-	-	-	-
FSPNet (Huang et al., 2023b)	CVPR'23	ViT-B16 (Dosovitskiy et al., 2020)	-	-	0.7347	0.8470	-	-	-	-	-	-	-	-	-	-	-	-
SILT (Yang et al., 2023)	ICCV'23	PVTv2-B5 (Wang et al., 2022)	-	-	-	-	0.0402	0.0493	-	-	-	-	-	-	-	-	-	-
SARA (Sun et al., 2023)	CVPR'23	ConvNeXt-L (Liu et al., 2022)	-	-	-	-	0.0429	0.0333	-	-	-	-	-	-	-	-	-	-
EBLNet (He et al., 2021)	ICCV'21	ResNet-50 (He et al., 2016)	-	-	-	-	-	-	0.1383	0.0959	-	-	-	-	-	-	-	-
RFENet (Fan et al., 2023)	IJCAI'23	ResNet-50 (He et al., 2016)	-	-	-	-	-	-	0.1036	0.0767	-	-	-	-	-	-	-	-
LDNet (Zhang et al., 2022b)	MICCAI'22	Res2Net-50 (Gao et al., 2019)	-	-	-	-	-	-	-	-	0.6425	0.7441	-	-	-	-	-	-
WeakPolyp (Wei et al., 2023)	MICCAI'23	PVTv2-B2 (Wang et al., 2022)	-	-	-	-	-	-	-	-	0.7490	0.8066	-	-	-	-	-	-
Inf-Net (Fan et al., 2020c)	TMI'20	Res2Net-50 (Gao et al., 2019)	-	-	-	-	-	-	-	-	-	-	0.4324	0.5285	-	-	-	-
DECOR-Net (Hu et al., 2023)	ISBI'23	Customized Design	-	-	-	-	-	-	-	-	-	-	0.4025	0.6949	-	-	-	-
AAU-net (Chen et al., 2022a)	TMI'22	Customized Design	-	-	-	-	-	-	-	-	-	-	-	-	0.4745	0.6515	-	-
CMUNet (Tang et al., 2023a)	ISBI'23	Customized Design	-	-	-	-	-	-	-	-	-	-	-	-	0.5452	0.8302	-	-
MALUNet (Ruan et al., 2022)	BIBM'22	Customized Design	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.8632	0.8537
EGE-UNet (Ruan et al., 2023)	MICCAI'23	Customized Design	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.8588	0.8498
Unified Models																		
EVP (Liu et al., 2023b)	CVPR'23	MiT-B4 (Xie et al., 2021b)	0.8431	0.9007	0.7262	0.8346	0.0481	0.0312	-	-	-	-	-	-	-	-	-	-
UniverSeg (Butoi* et al., 2023)	ICCV'23	ResNet-101 (He et al., 2016)	-	-	-	-	-	-	-	-	0.5525	0.2610	0.6726	0.3676	0.7749	0.5998	0.7605	0.7082
SegGPT (Wang et al., 2023c)	ICCV'23	ViT-L (Dosovitskiy et al., 2020)	0.3874	0.6283	0.4041	0.6529	0.4640	0.2041	0.4631	0.3064	0.5677	0.7074	0.1309	0.5533	0.3455	0.6033	0.4803	0.4402
Spider	-	ViT-B (Dosovitskiy et al., 2020)	0.8679	0.9074	0.7532	0.8505	0.0440	0.0308	0.0680	0.0550	0.8038	0.8540	0.6913	0.8118	0.8254	0.8607	0.8948	0.8758
Spider	-	ViT-L (Dosovitskiy et al., 2020)	0.8704	0.9102	0.7720	0.8615	0.0429	0.0284	0.0632	0.0545	0.7965	0.8554	0.6915	0.8128	0.8298	0.8599	0.8954	0.8743
Spider	-	Swin-B (Liu et al., 2021b)	0.8688	0.9086	0.7562	0.8527	0.0438	0.0302	0.0673	0.0547	0.8033	0.8561	0.6927	0.8121	0.8297	0.8614	**0.8968**	**0.8767**
Spider	-	Swin-L (Liu et al., 2021b)	0.8729	0.9109	0.7731	0.8620	0.0423	0.0271	**0.0628**	0.0539	0.7975	0.8550	0.6923	0.8121	0.8310	0.8609	0.8961	0.8757
Spider	-	ConvNeXt-B (Liu et al., 2022)	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	**0.0522**	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
Spider	-	ConvNeXt-L (Liu et al., 2022)	**0.8821**	**0.9158**	**0.7893**	**0.8674**	**0.0396**	**0.0265**	0.0636	0.0554	**0.8243**	**0.8664**	**0.6956**	0.8127	**0.8376**	**0.8655**	0.8943	0.8735
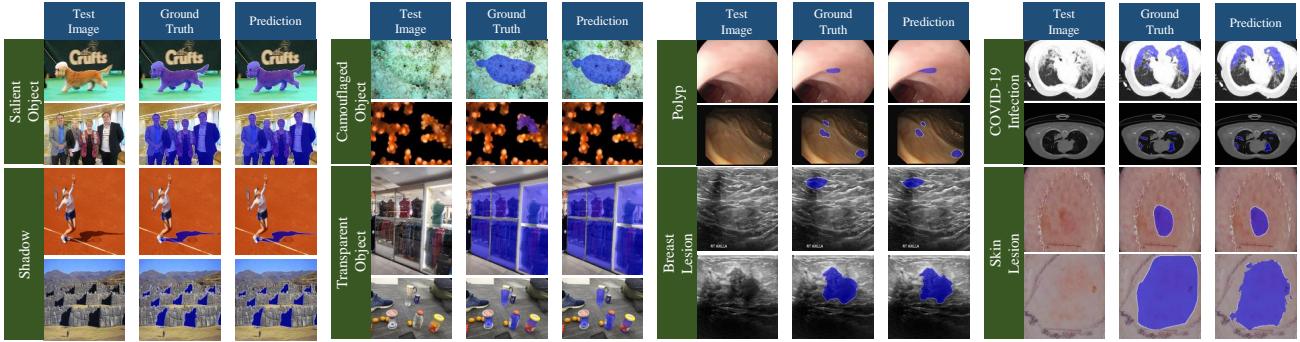
Figure 6. Some visual examples of Spider. Best viewed on screen.

4. Experiments

4.1. Datasets and Metrics

The dataset information is shown in Table 1. We follow the training settings of recent state-of-the-art methods in these tasks and merge all training samples together as our training set. For evaluation, we introduce some widely used metrics, including weighted F-measure (Margolin et al., 2014) (F_β^ω) and S-measure (Fan et al., 2017) (S_m) for SOD and COD, BER (Vicente et al., 2015) and MAE for SD and TOS, and mIoU and mDice for the medical segmentation tasks.

4.2. Implementation Details

We follow many visual unified models (Wang et al., 2023c; Yan et al., 2023; 2022; Ci et al., 2023; Kirillov et al., 2023; Wang et al., 2023b) to adopt a strong backbone as the encoder for covering the rich information from different large-scale datasets, which has become a consensus in current unified modeling field. In this work, we separately adopt the Transformer-based ViT (Dosovitskiy et al., 2020), Swin (Liu et al., 2021b) and CNN-based ConvNeXt (Liu et al., 2022) as the visual encoder to demonstrate the performance of the proposed Spider. All the experiments are implemented on the 8 Tesla A100 GPU for training 50 epochs. The input resolutions of images are resized to 384×384. For each task, the mini-batch sizes of the input and prompt are set to 4 and 12, respectively. We adopt some basic image augmentation techniques to avoid overfitting, including random flipping, rotating and border clipping. The Adam (Kingma & Ba, 2015) optimizer scheduled by "step" with initial learning rate of 0.0001, decay size of 30 and decay rate of 0.9 is
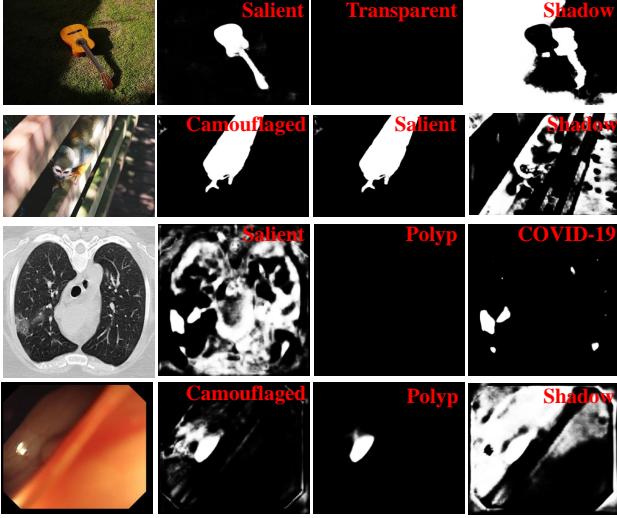
Figure 7. Visual predictions of multiple concepts in an image.

introduced to update model parameters.

4.3. Evaluation

Quantitative Results. We compare the Spider with recent state-of-the-art specialized models and unified models as shown in Table 2. It can be seen that Spider achieves dominant performance on all the tasks and performs better than each specialized model. In particular, it outperform other competitors by more than 30% on the TOS, CLI and BLS tasks. For the unified methods, EVP (Liu et al., 2023b) only unifies three tasks and relays on three sets of adaptor parameters. UniverSeg (Butoi* et al., 2023) only focuses on medical image segmentation tasks. SegGPT (Wang et al., 2023c) and Spider are able to accomplish all tasks with a single set of parameters. Limited by the prompt strategy based on a single image-mask pair, SegGPT cannot show the generalization ability across these tasks involving context-dependent concepts, even if it has been trained on more than 250,000 diverse images.

Qualitative Results. We show some visual results in Figure 6. The detailed group prompts for all tasks and qualitative comparisons with other methods can be found in Appendix A.6 and Appendix A.7. In addition, Spider has multi-concept understanding ability as shown in Figure 7. There are some insightful phenomena. For the monkey (see the 2nd row), Spider predicts salient and camouflaged object segmentation map at the same time. According to the intuitive response of human vision system, zooming out makes the monkey hidden in the surrounding environment, but zooming in makes it slightly stand out. Therefore, the concepts of saliency and camouflage may coexist and even sometimes are manifested in the same object. It is also in line with the research motivation of salient and camouflaged object ranking (Fang et al., 2021; Tian et al., 2022; Lv et al.,
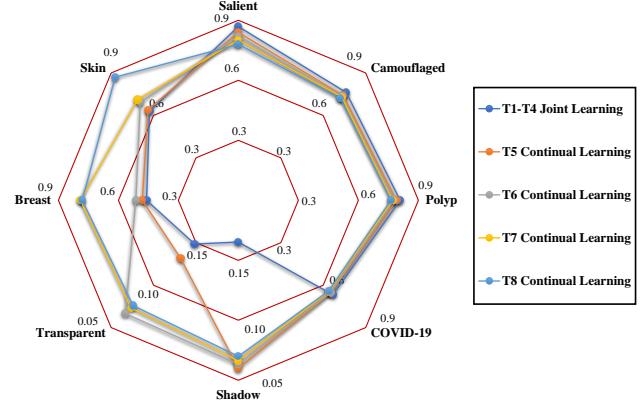

Figure 8. Performance of continuous learning on new tasks. Spider is firstly jointly pretrained on the T1-T4 data, and then continuously fine-tunes it from T5 to T8. T1: Salient Object Detection. T2: Camouflaged Object Detection. T3: Colon Polyp Segmentation. T4: COVID-19 Lung Infection. T5: Shadow Detection. T6: Transparent Object Segmentation. T7: Breast Lesion Segmentation. T8: Skin Lesion Segmentation.

2021; 2023). For the colon image (see the 4th row), we try to elicit the concepts of camouflage, polyp, and shadows. Polyp lesions are usually hidden on the surface of the colon, our camouflaged object prediction effectively perceives polyp, which illustrates that the COD data is possibly beneficial to polyp segmentation. Utilizing a large amount of natural scene data to improve medical lesion segmentation will promote the semi-supervised learning research in medical image field. Finally, we also provide a good shadow detection for the colon image, which reveals a potential application of colonoscope shadow removal for improving the lesion visualization of the medical equipment.

4.4. Ablation Study

In Table 3, all models are based on the ConvNeXt-B (Liu et al., 2022) backbone.

Joint Training vs. Separate Training. We train each model separately on each task with the same number of iterations and architectures as done in joint training. We can observe that the jointly trained models consistently outperform the separately trained ones on all tasks. This indicates that our framework with 100% shared parameters can assimilate rich cross-domain knowledge and well function in specific tasks with the help of the image-mask group prompts.

Concept Filter. Our baseline is the general UNet (Ronneberger et al., 2015) structure without any specified design. The concept filter aims to help the baseline improve scene understanding and task discrimination. We step by step verify the prompts that drive the concept filter. First, image-group prompts have the basic ability to find task commonality from image group, which significantly improves the performance over UNet on all tasks by more than

Table 3. Ablation studies on the eight tasks. All models adopt the ConvNeXt-B as the backbone.

Method	Salient DUTS $F_\beta^\omega\uparrow$	$S_m\uparrow$	Camouflaged COD10K $F_\beta^\omega\uparrow$	$S_m\uparrow$	Shadow SBU BER↓	MAE↓	Transparent Trans10K BER↓	MAE↓	Polyp 5 datasets mDice↑	mIoU↑	COVID-19 COVID-19 CT mDice↑	mIoU↑	Breast BUSI mDice↑	mIoU↑	Skin ISIC2018 mDice↑	mIoU↑
(a) Joint Training vs. Separate Training																
Separate Training	0.8593	0.9012	0.7543	0.8544	0.0476	0.0293	0.0673	0.0576	0.7786	0.8267	0.6367	0.7388	0.7747	0.7826	0.8548	0.8216
Joint Training	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
(b) Concept Filter																
UNet	0.6253	0.6345	0.5346	0.6038	0.1382	0.0846	0.1426	0.1135	0.6144	0.6532	0.3672	0.4112	0.4378	0.4782	0.4889	0.5023
+ Image-Group Prompts	0.7843	0.8346	0.7055	0.7887	0.0534	0.0332	0.0778	0.0685	0.7230	0.7564	0.5732	0.7038	0.7301	0.7901	0.7903	0.7888
+ Mask-Group Prompts (Foreground)	0.8422	0.8907	0.7523	0.8302	0.0473	0.0298	0.0674	0.0581	0.7809	0.8388	0.6509	0.7631	0.7746	0.8316	0.8573	0.8501
+ Mask-Group Prompts (Background)	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
Concept Filter → Addition Fusion	0.6534	0.6497	0.5742	0.6313	0.1108	0.0769	0.1235	0.1096	0.6532	0.6742	0.3809	0.4216	0.5464	0.5589	0.5160	0.5436
(c) Training Strategies in the Unified Framework																
Random FP - Unify BP	0.8608	0.8998	0.7562	0.8573	0.0453	0.0280	0.0655	0.0547	0.8102	0.8607	0.6340	0.7538	0.8008	0.8212	0.8778	0.8645
Balance FP - Separate BP	0.8422	0.8831	0.7383	0.8288	0.0490	0.0299	0.0682	0.0566	0.7979	0.8425	0.6388	0.7612	0.8046	0.8308	0.8508	0.8477
Balance FP - Unify BP	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
(d) Number and Selection of Prompts																
Random Selection (G = 1)	0.7038	0.7134	0.5732	0.6789	0.0789	0.0520	0.1136	0.0971	0.6533	0.7011	0.5620	0.6844	0.6135	0.6346	0.7421	0.7116
Random Selection (G = 4)	0.8091	0.8448	0.6912	0.7802	0.0532	0.0340	0.0707	0.0625	0.7341	0.7788	0.6432	0.7677	0.7790	0.8100	0.8108	0.7979
Random Selection (G = 12)	0.8348	0.8815	0.7298	0.8064	0.0488	0.0310	0.0685	0.0574	0.7736	0.8164	0.6527	0.7809	0.7977	0.8209	0.8402	0.8316
Random Selection (G = 64)	0.8723	0.9101	0.7762	0.8602	0.0444	0.0272	0.0634	0.0525	0.8202	0.8648	0.6910	0.8100	0.8345	0.8630	0.8942	0.8730
Clustering Selection (G = 64)	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733

25%. Then, the foreground features are used as the query of transformer to directly establish the contrast relationship between the object query and the whole image. In this way, " + Mask-Group Prompts (Foreground) " achieves similar performance with the separate training model. Next, the background features are introduced to highlight the importance of the surroundings for concept expression, which achieves 40% performance gain compared to the baseline. Finally, we replace the concept filter with the element-wise addition fusion (keeping similar number of parameters) to show the advantages of the proposed high-level concept matching mechanism.

Training Strategies in the Unified Framework. We conduct the experiments in terms of forward and back propagation, including random data partition and separate gradient update for each task. We can see that "Balance FP - Unify BP" performs the best, which suggests that when training a unified model, all task data should be treated as a whole and each part is equally important. Belittling any one of them will produce negative effect to other tasks.

Number and Selection of Prompts. We evaluate the impact of different number of random prompts in the inference phase. It can be seen that the overall performance is the worst when $G = 1$. As the number increases, the performance is gradually elevated and stabilizes when $G = 64$. Moreover, we select 64 pairs of samples as the group prompts by clustering training data for each task. It can be seen that "Clustering Selection ($G = 64$)" has almost the same performance as "Random Selection". Thus, the strategy of random selection during training indeed makes Spider robust against different group prompts when testing. More experiments of prompts during training and testing

phases can be found in the Appendix A.5.

Continuous Learning & Potential for Unseen Tasks. Figure 8 shows the ability of continuous learning of Spider. First, we jointly train Spider on the four tasks including SOD, COD, CPS, and CLI, to ensure the basic general segmentation capability. Then, the continuous learning is performed on the additional training sets from T5 to T8, where we only fine-tune the last layer of the decoder and the concept filter. The minimal number of trainable parameters ($< 1\%$) drastically accelerates the training process, and alleviates the catastrophic forgetting (Parisi et al., 2019; De Lange et al., 2021). It can be seen that Spider's performance on new tasks is significantly improved, while there is only a negligible performance degradation of no more than 5% on the old tasks. Besides, the performance for SLS is over 0.6 mIoU even if the model is only trained on the T1-T4 data. With the increasing of data scale and diversity, the performance is steadily improved in some unseen tasks. More relevant analysis are be found in the Appendix A.8.

5. Conclusion

We propose Spider, a universal context-dependent concept understanding model, to unify eight segmentation tasks with the proposed group prompt paradigm. Extensive experiments demonstrate the superior performance of the proposed Spider on twelve challenging benchmarks using a single set of parameters. Spider can serve as a solid baseline within the unified cross-domain research. In the future, we will expand Spider to more context-dependent concept understanding tasks, such as industrial defect detection, inharmonious region localization, and defocus blur detection. We are also working on introducing image editing tasks

into the Spider framework, which can simultaneously complete more interesting applications such as shadow detection and removal, salient object detection and camouflageization, inharmonious region localization and harmonization.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. Dataset of breast ultrasound images. *Data in brief*, 28: 104863, 2020.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Barsalou, L. W. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10:82–93, 1982.

Butoi*, V. I., Ortiz*, J. J. G., Ma, T., Sabuncu, M. R., Guttag, J., and Dalca, A. V. Universeg: Universal medical image segmentation. *ICCV*, 2023.

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., and Andre, M. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *BSPC*, 61:102027, 2020.

Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., and Gong, X. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE TMI*, 40:2439–2451, 2021.

Chen, G., Li, L., Dai, Y., Zhang, J., and Yap, M. H. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE TMI*, 2022a.

Chen, G.-P., Li, L., Dai, Y., and Zhang, J.-X. Nu-net: An unpretentious nested u-net for breast tumor segmentation. *arXiv preprint arXiv:2209.07193*, 2022b.

Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., and Zang, Y. Sam fails to segment anything?–sam-adapter: Adapting sam in under-performed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1290–1299, 2022.

Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., and Ouyang, W. Unihcp: A unified model for human-centric perceptions. In *CVPR*, pp. 17840–17852, 2023.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Cong, R., Yang, H., Jiang, Q., Gao, W., Li, H., Wang, C., Zhao, Y., and Kwong, S. Bcs-net: Boundary, context, and semantic for automatic covid-19 lung infection segmentation from ct images. *IEEE TIM*, 71:1–11, 2022.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Dai, D., Dong, C., Xu, S., Yan, Q., Li, Z., Zhang, C., and Luo, N. Ms red: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *MIA*, 75:102293, 2022.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 44:3366–3385, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., and Borji, A. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pp. 4548–4557, 2017.

Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., and Shao, L. Camouflaged object detection. In *CVPR*, pp. 2777–2787, 2020a.

Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., and Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pp. 263–273, 2020b.

Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., and Shao, L. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39: 2626–2637, 2020c.

Fan, K., Wang, C., Wang, Y., Wang, C., Yi, R., and Ma, L. Rfenet: Towards reciprocal feature evolution for glass segmentation. In *IJCAI*, pp. 717–725, 2023.

Fang, H., Zhang, D., Zhang, Y., Chen, M., Li, J., Hu, Y., Cai, D., and He, X. Salient object ranking with position-preserved attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16331–16341, 2021.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43:652–662, 2019.

He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., and Li, X. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pp. 22046–22055, 2023.

He, H., Li, X., Cheng, G., Shi, J., Tong, Y., Meng, G., Prinet, V., and Weng, L. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, pp. 15859–15868, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Hu, J., Yang, Y., Guo, X., Peng, B., Huang, H., and Ma, T. Decor-net: A covid-19 lung infection segmentation network improved by emphasizing low-level features and decorrelating features. In *ISBI*, pp. 1–5, 2023.

Hu, X., Fu, C.-W., Zhu, L., Qin, J., and Heng, P.-A. Direction-aware spatial context features for shadow detection and removal. *IEEE TPAMI*, 42:2795–2808, 2019.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023a.

Huang, Z., Dai, H., Xiang, T.-Z., Wang, S., Chen, H.-X., Qin, J., and Xiong, H. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, pp. 5557–5566, 2023b.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Ji, G.-P., Fan, D.-P., Xu, P., Cheng, M.-M., Zhou, B., and Van Gool, L. Sam struggles in concealed scenes– empirical study on" segment anything". *arXiv preprint arXiv:2304.06022*, 2023a.

Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., and Zheng, Y. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pp. 12341–12351, 2021.

Ji, W., Li, J., Bi, Q., Guo, C., Liu, J., and Cheng, L. Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. *ICLR*, 2022.

Ji, W., Li, J., Bi, Q., Li, W., and Cheng, L. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023b.

Ji, W., Li, J., Bian, C., Zhou, Z., Zhao, J., Yuille, A. L., and Cheng, L. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *CVPR*, pp. 1094–1104, 2023c.

Ji, W., Li, J., Bi, Q., Liu, T., Li, W., and Cheng, L. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, pp. 1–14, 2024.

Jia, Q., Yao, S., Liu, Y., Fan, X., Liu, R., and Luo, Z. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, pp. 4713–4722, 2022.

Jiang, J., Cao, G., Deng, J., Do, T.-T., and Luo, S. Robotic perception of transparent objects: A review. *IEEE TAI*, 2023.

Kim, T., Lee, H., and Kim, D. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM MM*, pp. 2167–2175, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Lachmann, T. and Van Leeuwen, C. Individual pattern representations are context independent, but their collectiverepresentation is context dependent. *The Quarterly Journal of Experimental Psychology Section A*, 58:1265–1294, 2005.

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., and Wang, S. Skin lesion segmentation via generative adversarial networks with dual discriminators. *MIA*, 64:101716, 2020.

Li, J., Ji, W., Bi, Q., Yan, C., Zhang, M., Piao, Y., Lu, H., and Cheng, L. Joint semantic mining for weakly supervised rgb-d salient object detection. *NeurIPS*, pp. 11945–11959, 2021.

Li, J., Ji, W., Wang, S., Li, W., and Cheng, L. Dvsod: Rgb-d video salient object detection. In *NeurIPS*, pp. 8774–8787, 2023a.

Li, J., Ji, W., Zhang, M., Piao, Y., Lu, H., and Cheng, L. Delving into calibrated depth for accurate rgb-d salient object detection. *International Journal of Computer Vision*, 131(4):855–876, 2023b.

Lin, J., Wang, G., and Lau, R. W. Progressive mirror detection. In *CVPR*, pp. 3697–3705, 2020.

Lin, J., He, Z., and Lau, R. W. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pp. 13415–13424, 2021.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017.

Liu, D., Long, C., Zhang, H., Yu, H., Dong, X., and Xiao, C. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR*, pp. 8139–8148, 2020.

Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., and Zhou, Z. Clip-driven universal model for organ segmentation and tumor detection. In *ICCV*, pp. 21152–21164, 2023a.

Liu, N., Zhang, N., Wan, K., Shao, L., and Han, J. Visual saliency transformer. In *ICCV*, pp. 4722–4732, 2021a.

Liu, W., Shen, X., Pun, C.-M., and Cun, X. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pp. 19434–19445, 2023b.

Liu, Y., Xu, S., Zhang, D., and Han, J. Seggpt meets co-saliency scene. *arXiv preprint arXiv:2305.04396*, 2023c.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pp. 10012–10022, 2021b.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *CVPR*, pp. 11976–11986, 2022.

Lüddecke, T. and Ecker, A. Image segmentation using text and image prompts. In *CVPR*, pp. 7086–7096, 2022.

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

Lv, Y., Zhang, J., Dai, Y., Li, A., Barnes, N., and Fan, D.-P. Towards deeper understanding of camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Margolin, R., Zelnik-Manor, L., and Tal, A. How to evaluate foreground maps? In *CVPR*, pp. 248–255, 2014.

Martial, C., Stawarczyk, D., and D'Argembeau, A. Neural correlates of context-independent and context-dependent self-knowledge. *Brain and Cognition*, 125:23–31, 2018.

Paluru, N., Dayal, A., Jenssen, H. B., Sakinis, T., Cenkeramaddi, L. R., Prakash, J., and Yalavarthy, P. K. Anam-net: Anamorphic depth embedding-based lightweight cnn for segmentation of anomalies in covid-19 chest ct images. *IEEE TNNLS*, 32(3):932–946, 2021.

Pang, Y., Zhang, L., Zhao, X., and Lu, H. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pp. 235–252, 2020a.

Pang, Y., Zhao, X., Zhang, L., and Lu, H. Multi-scale interactive network for salient object detection. In *CVPR*, pp. 9413–9422, 2020b.

Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., and Lu, H. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pp. 2160–2170, 2022a.

Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., and Lu, H. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pp. 2160–2170, 2022b.

Pang, Y., Zhao, X., Zhang, L., and Lu, H. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE TIP*, 2023.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Piao, Y., Ji, W., Li, J., Zhang, M., and Lu, H. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pp. 7254–7263, 2019.

Potlapalli, V., Zamir, S. W., Khan, S., and Khan, F. S. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Ruan, J., Xiang, S., Xie, M., Liu, T., and Fu, Y. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *BIBM*, pp. 1150–1156, 2022.

Ruan, J., Xie, M., Gao, J., Liu, T., and Fu, Y. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *MICCAI*, pp. 481–490, 2023.

Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shen, D., and Shi, Y. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*, 2020.

Sun, J., Xu, K., Pang, Y., Zhang, L., Lu, H., Hancke, G., and Lau, R. Adaptive illumination mapping for shadow detection in raw images. In *ICCV*, pp. 12709–12718, 2023.

Sun, Y., Wang, S., Chen, C., and Xiang, T.-Z. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794*, 2022.

Takahashi, N. and Mitsufuji, Y. Densely connected multidilated convolutional networks for dense prediction tasks. In *CVPR*, pp. 993–1002, 2021.

Tang, F., Wang, L., Ning, C., Xian, M., and Ding, J. Cmunet: a strong convmixer-based medical ultrasound image segmentation network. In *ISBI*, pp. 1–5, 2023a.

Tang, L., Xiao, H., and Li, B. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*, 2023b.

Tian, X., Xu, K., Yang, X., Du, L., Yin, B., and Lau, R. W. Bi-directional object-context prioritization learning for saliency ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5882–5891, 2022.

Vandenhende, S., Georgoulis, S., and Van Gool, L. Mtinet: Multi-scale task interaction networks for multi-task learning. In *ECCV*, pp. 527–543, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 30, 2017.

Vicente, T. F. Y., Hoai, M., and Samaras, D. Leave-one-out kernel optimization for shadow detection. In *ICCV*, pp. 3388–3396, 2015.

Vicente, T. F. Y., Hou, L., Yu, C.-P., Hoai, M., and Samaras, D. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pp. 816–832, 2016.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. Learning to detect salient objects with image-level supervision. In *CVPR*, pp. 136–145, 2017.

Wang, M., An, X., Li, Y., Li, N., Hang, W., and Liu, G. Ems-net: Enhanced multi-scale network for polyp segmentation. In *IEEE EMBC*, pp. 2936–2939, 2021.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8:415–424, 2022.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pp. 14408–14419, 2023a.

Wang, X., Wang, W., Cao, Y., Shen, C., and Huang, T. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pp. 6830–6839, 2023b.

Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., and Huang, T. Seggpt: Towards segmenting everything in context. In *ICCV*, pp. 1130–1140, 2023c.

Wang, Y., Wang, R., Fan, X., Wang, T., and He, X. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *CVPR*, pp. 10031–10040, 2023d.

Wei, J., Wang, S., and Huang, Q. F^{3}net: Fusion, feedback and focus for salient object detection. In *AAAI*, pp. 12321–12328, 2020.

Wei, J., Hu, Y., Cui, S., Zhou, S. K., and Li, Z. Weakpolyp: You only look bounding box for polyp segmentation. In *MICCAI*, pp. 757–766, 2023.

Wei, K., Deng, C., Yang, X., and Tao, D. Incremental zero-shot learning. *IEEE Transactions on Cybernetics*, 52(12): 13788–13799, 2021.

Wen, J., He, L., and Zhu, F. Swarm robotics control and communications: Imminent challenges for next generation smart logistics. *IEEE Communications Magazine*, 56:102–107, 2018.

Wu, H., Pan, J., Li, Z., Wen, Z., and Qin, J. Automated skin lesion segmentation via an adaptive dual attention module. *IEEE TMI*, 40:357–370, 2020a.

Wu, T., Tang, S., Zhang, R., Cao, J., and Li, J. Tree-structured kronecker convolutional network for semantic segmentation. In *ICME*, pp. 940–945, 2019.

Wu, Y.-H., Gao, S.-H., Mei, J., Xu, J., Fan, D.-P., Zhao, C.-W., and Cheng, M.-M. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. *arXiv preprint arXiv:2004.07054*, 2020b.

Xie, C., Xia, C., Ma, M., Zhao, Z., Chen, X., and Li, J. Pyramid grafting network for one-stage high resolution saliency detection. In *CVPR*, pp. 11717–11726, 2022a.

Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., and Luo, P. Segmenting transparent objects in the wild. In *ECCV*, pp. 696–711, 2020.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., and Luo, P. Segmenting transparent objects in the wild with transformer. In *IJCAI*, pp. 1194–1200, 2021a.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34: 12077–12090, 2021b.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *CVPR*, pp. 9653–9663, 2022b.

Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., and Lu, H. Towards grand unification of object tracking. In *ECCV*, pp. 733–751, 2022.

Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., and Lu, H. Universal instance perception as object discovery and retrieval. In *CVPR*, pp. 15325–15336, 2023.

Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., and Fan, D.-P. Uncertainty-guided transformer reasoning for camouflaged object detection. In *ICCV*, pp. 4146–4155, 2021.

Yang, H., Wang, T., Hu, X., and Fu, C.-W. Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *ICCV*, pp. 12687–12698, 2023.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022a.

Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., and Yu, Y. Adaptive context selection for polyp segmentation. In *MICCAI*, pp. 253–262, 2020.

Zhang, R., Lai, P., Wan, X., Fan, D.-J., Gao, F., Wu, X.-J., and Li, G. Lesion-aware dynamic kernel for polyp segmentation. In *MICCAI*, pp. 99–109, 2022b.

Zhang, Y., Liu, H., and Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In *MICCAI*, pp. 14–24, 2021.

Zhao, X., Pang, Y., Zhang, L., Lu, H., and Zhang, L. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pp. 35–51, 2020a.

Zhao, X., Zhang, L., Pang, Y., Lu, H., and Zhang, L. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pp. 646–662, 2020b.

Zhao, X., Zhang, L., Pang, Y., Lu, H., and Zhang, L. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pp. 646–662, 2020c.

Zhao, X., Zhang, L., and Lu, H. Automatic polyp segmentation via multi-scale subtraction network. In *MICCAI*, pp. 120–130, 2021.

Zhao, X., Pang, Y., Zhang, L., and Lu, H. Joint learning of salient object detection, depth estimation and contour extraction. *IEEE TIP*, 31:7350–7362, 2022a.

Zhao, X., Pang, Y., Zhang, L., Lu, H., and Ruan, X. Self-supervised pretraining for rgb-d salient object detection. In *AAAI*, pp. 3463–3471, 2022b.

Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., Sun, W., and Lu, H. M2snet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv preprint arXiv:2303.10894*, 2023a.

Zhao, X., Pang, Y., Zhang, L., Lu, H., and Zhang, L. Towards diverse binary segmentation via a simple yet general gated network. *arXiv preprint arXiv:2303.10396*, 2023b.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *CVPR*, pp. 633–641, 2017.

Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., and Gong, C. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023a.

Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., and Shen, D. Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition*, 140:109555, 2023b.

Zhu, L., Deng, Z., Hu, X., Fu, C.-W., Xu, X., Qin, J., and Heng, P.-A. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pp. 121–136, 2018.

Table 4. Performance stability when using low qualities of mask annotations of the prompts during inference.

Mask Prompt	Salient DUTS $F_\beta^\omega \uparrow$	$S_m \uparrow$	Camouflaged COD10K $F_\beta^\omega \uparrow$	$S_m \uparrow$	Shadow SBU BER $\downarrow$	MAE $\downarrow$	Transparent Trans10K BER $\downarrow$	MAE $\downarrow$	Polyp 5 datasets mDice $\uparrow$	mIoU $\uparrow$	COVID-19 COVID-19 CT mDice $\uparrow$	mIoU $\uparrow$	Breast BUSI mDice $\uparrow$	mIoU $\uparrow$	Skin ISIC2018 mDice $\uparrow$	mIoU $\uparrow$
Ground Truth	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
Dilation: Kernel = 3×3	0.8718	0.9100	0.7771	0.8615	0.0446	0.0275	0.0638	0.0525	0.8205	0.8651	0.6922	0.8102	0.8348	0.8630	0.8946	0.8730
Dilation: Kernel = 5×5	0.8718	0.9102	0.7768	0.8617	0.0446	0.0278	0.0639	0.0526	0.8204	0.8650	0.6922	0.8102	0.8345	0.8631	0.8945	0.8728
Dilation: Kernel = 7×7	0.8714	0.9100	0.7764	0.8614	0.0448	0.0280	0.0641	0.0525	0.8204	0.8651	0.6921	0.8103	0.8346	0.8634	0.8944	0.8725
Erosion: Kernel = 3×3	0.8714	0.9098	0.7764	0.8603	0.0448	0.0278	0.0641	0.0530	0.8202	0.8655	0.6917	0.8097	0.8345	0.8627	0.8940	0.8725
Erosion: Kernel = 5×5	0.8708	0.9074	0.7754	0.8589	0.0450	0.0279	0.0645	0.0534	0.8189	0.8643	0.6915	0.8090	0.8338	0.8617	0.8934	0.8721
Erosion: Kernel = 7×7	0.8699	0.9063	0.7732	0.8578	0.0453	0.0284	0.0650	0.0545	0.8158	0.8635	0.6898	0.8068	0.8321	0.8600	0.8922	0.8715

A. Appendix

A.1. Advantages of the Proposed Concept Filter

I) Robustness: The model can use extra background features to regularize representation learning instead of relying solely on foreground features. *II) Learning Efficiency*: Dividing features into foreground and background can make it easier for the model to learn important features. Since different features are represented in the weights and biases, this helps the model converge faster. *III) Interpretability*: Splitting the mask-prompts into foreground and background parts not only helps researchers better understand how the model works, but also increases the model's credibility. *IV) Flexibility*: If there is no obvious background information in some images, the model can leverage the background feature generator to generate meaningful biases. *V) Generalization Ability*: By utilizing foreground and background features separately, the model can better adapt to different data distributions. It can improve the generalization ability of the model, allowing it to handle unseen images. *VI) High Tolerance Rate of Prompt Annotation*: Different from direct pix-level feature fusion, condensing prompt knowledge into high-level expression filters can reduce the model's requirements for mask annotation accuracy of prompts. As shown in Table 4, our performance is stable when faced with varying degrees of dilation and erosion on the prompt mask.

A.2. Definition of Different Context-dependent Image Segmentation Tasks

I) Salient object detection (SOD) (Zhao et al., 2020a; Pang et al., 2020b; Zhao et al., 2020b; Pang et al., 2020a; Zhao et al., 2022b;a; Pang et al., 2023; Zhao et al., 2023b) is often associated with *II) camouflaged object detection* (COD) (Pang et al., 2022a; Zhao et al., 2023b). The former aims at finding visually salient objects, while the latter focuses on hidden objects extremely similar to the surrounding backgrounds. *III) Shadow detection* (SD) is an important research topic. Because the shadow contains rich depth and geometry cues, shadow detection is often applied in image editing tasks, such as shadow removal (Hu et al., 2019) and image synthesis (Liu et al., 2020). In addition, some important details of objects may be hidden in the shadow. It is very necessary to understand the shadow. *IV) Transparent object segmentation* (TOS) is a challenging task due to the properties of reflection and refraction. It can assist indoor smart robots (Jiang et al., 2023) and outdoor unmanned logistics vehicles (Wen et al., 2018) in controlling or avoiding transparent objects. In intelligent diagnosis, fully automatic image segmentation has become an important medical aid. Compared with the organs of fixed shape and appearance, the lesions have strong context-dependent concept, lesion segmentation is more challenging. *V) Colon polyp segmentation* (CPS) (Fan et al., 2020b; Zhao et al., 2021; Wei et al., 2023; Zhang et al., 2022b; Wang et al., 2021) identifies polyps of different sizes hidden on the surface of the intestinal wall. *VI) COVID-19 lung infection* (CLI) (Shan et al., 2020; Cong et al., 2022; Paluru et al., 2021; Wu et al., 2020b; Fan et al., 2020c; Zhao et al., 2023a) captures the infected area from a CT image containing other lung lesions and a large number of anatomical structures and tissue textures. *VII) Breast lesion segmentation* (BLS) (Chen et al., 2022b; Byra et al., 2020; Chen et al., 2021; 2022a; Tang et al., 2023a) needs to overcome the speckle noise in ultrasound images caused by the interaction of scattered sound waves and tissue structures, which reduces image contrast and blurs the lesion edge. *VIII) Skin lesion segmentation* (SLS) (Dai et al., 2022; Lei et al., 2020; Wu et al., 2020a; Ruan et al., 2022; 2023) aims to search dermatofibromas and epidermal cysts from dermoscopic images.

A.3. Challenges of Context-dependent Concept Understanding

For traditional semantic segmentation tasks, labeling data may be relatively easy. As shown in Figure 9 (a), existing CI concept datasets, such as Cityscapes (Cordts et al., 2016) and ADE20K (Zhou et al., 2017), all have multiple concept

(a) CI Concept Segmentation Dataset
(One Image - Multiple Concept Labels)

(b) CD Concept Segmentation Dataset
(One Image - One Concept Label)

(c) Our Multiple Concept Predictions
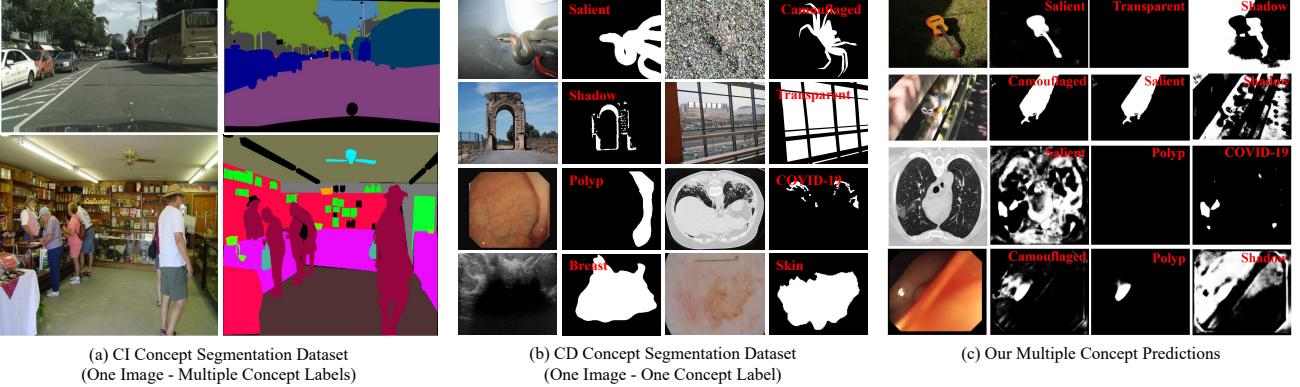
Figure 9. Context-independent concept segmentation datasets vs. Context-dependent concept segmentation datasets.

Table 5. Ablation experiments of the number of prompts in training and testing phases.

| | | Salient | | Camouflaged | | Shadow | | Transparent | | Polyp | | COVID-19 | | Breast | | Skin | |
| | | DUTS | | COD10K | | SBU | | Trans10K | | 5 datasets | | COVID-19 CT | | BUSI | | ISIC2018 | |
#Train	#Test	$F_\beta^\omega \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$S_m \uparrow$	BER $\downarrow$	MAE $\downarrow$	BER $\downarrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	mDice $\uparrow$	mIoU $\uparrow$	mDice $\uparrow$	mIoU $\uparrow$	mDice $\uparrow$	mIoU $\uparrow$
1	64	0.7177	0.7187	0.6145	0.6868	0.0745	0.0502	0.0929	0.0921	0.6621	0.7109	0.5653	0.6758	0.6194	0.6306	0.7544	0.7316
4	64	0.8203	0.8501	0.7134	0.7886	0.0508	0.0331	0.0658	0.0596	0.7545	0.7848	0.6672	0.7707	0.7846	0.8183	0.8178	0.8086
12	64	0.8732	0.9109	0.7779	0.8625	0.0444	0.0272	0.0632	0.0522	0.8211	0.8655	0.6925	0.8106	0.8352	0.8632	0.8949	0.8733
12	1	0.7038	0.7134	0.5732	0.6789	0.0789	0.0520	0.1136	0.0971	0.6533	0.7011	0.5620	0.6844	0.6135	0.6346	0.7421	0.7116
12	4	0.8091	0.8448	0.6912	0.7802	0.0532	0.0340	0.0707	0.0625	0.7341	0.7788	0.6432	0.7677	0.7790	0.8100	0.8108	0.7979
12	12	0.8348	0.8815	0.7298	0.8064	0.0488	0.0310	0.0685	0.0574	0.7736	0.8164	0.6527	0.7809	0.7977	0.8209	0.8402	0.8316
12	64	0.8723	0.9101	0.7762	0.8602	0.0444	0.0272	0.0634	0.0525	0.8202	0.8648	0.6910	0.8100	0.8345	0.8630	0.8942	0.8730
1	1	0.4674	0.5389	0.4375	0.5745	0.2346	0.2541	0.2406	0.2720	0.5935	0.6345	0.3784	0.3990	0.3046	0.4589	0.5846	0.5038

annotations for a single image and do not overlap with each other. Current CI models can well distinguish different concepts. Existing CD concept datasets provide the annotates of single concept, as shown in Figure 9 (b). Actually, multiple CD concepts often co-occur in an object. How to effectively depict the contrast between the foreground and background to highlight the characteristics of each concept is the key to achieving accurate segmentation. Figure 9 (c) shows the multi-concept prediction capabilities obtained by Spider. In addition, concept-shift will produce in some moving objects, which puts higher requirements on the model's ability of concept understanding. In the future, we will study context-dependent concept understanding in video.

A.4. Applications of Context-dependent Concepts Understanding

When multiple context-dependent concepts appear simultaneously in an image, the following applications will occur: ***I) Human-computer interaction and virtual reality***: In human-computer interaction or virtual reality, salient objects in the user interface attract the user's attention. Shadow and transparency effects can be used to create more realistic 3D effects. Camouflage objects can be used to hide or show specific elements, while blurred backgrounds can help users focus on the main content. ***II) Image Editing and Augmented Reality***: Transparent objects can be removed or blurred to make salient objects more prominent and improve the visual effect of the image. ***III) Medical image analysis***: Spider can provide a good shadow detection for the colon image, which reveals a potential application of colonoscope shadow removal for improving the lesion visualization of the medical equipment. ***IV) Military reconnaissance and security inspection***: In the military or security field, salient objects in images may represent important military equipment or potential threats. Camouflaged objects may be used to hide true intentions or devices. ***V) Autonomous driving***: In autonomous driving systems, it required to distinguish the salient objects such as vehicles, pedestrians and the camouflaged obstacles, transparent objects such as glass to ensure the vehicle travels safely. ***VI) Environmental monitoring and urban planning***: In environmental monitoring and urban planning, by identifying multiple context-dependent concpets, we can understand the development of the city, environmental changes, and potential environment problems, etc., providing important basis for urban planning and environmental governance.

Table 6. Five experimental settings on Zero-shot learning (ZSL), Continual Learning (CL) and Incremental Zero-shot learning (IZSL). T1 - T8 are consistent with the expression in Figure 8.

Setting	Finetuning	Test Tasks for ZSL	Test Tasks for CL	Test Tasks for IZSL
S.0	-	T5 - T8	-	-
S.1	S.0 + T5	-	T1 - T5	T6, T7, T8
S.2	S.1 + T6	-	T1 - T6	T7, T8
S.3	S.2 + T7	-	T1 - T7	T8
S.4	S.3 + T8	-	T1 - T8	-

A.5. Number of Prompts

In Table 5, we thoroughly show the impact of different numbers of prompts during training and testing phases. The gap between the best choice (Train: 12, Test: 64) and the worst choice (Train: 1, Test: 1) demonstrates the necessity of group prompts for the model to understand the context-dependent concepts.

A.6. Visualization of Clustered Group Prompts

In Figure 10 - Figure 17, we visualize the clustered group prompts used by each task during inference.

A.7. Qualitative Comparisons

In Figure 18 - Figure 25, we show a qualitative comparison with other methods. We can see that the previous generalist model SegGPT can only distinguish the foreground and background based on the shape cues and cannot truly understand the context-dependent concept.

A.8. Performance Analysis of Spider in Continual/Zero-shot/Incremental Zero-shot learning

As stated in (Wei et al., 2021), **Zero-shot learning (ZSL)** is a hot topic in transfer learning, which handles the issue that some test classes are not included in the training set. **Continual Learning (CL)**, also known as Incremental Learning, Life-long Learning, requires the model to accumulate the knowledge of previous tasks and capture the knowledge of current tasks simultaneously. Catastrophic forgetting is the main reason why the trained model forgets the knowledge of previous tasks when a new task arrives. **Incremental Zero-shot learning (IZSL)** is different from traditional CL and introduces the zero-shot setting. The model trained on the pervious tasks is finetuned on the new task as in CL, but tested on other unseen test sets as in ZSL. In Figure 8, based on the same model pre-trained on the first four tasks T1 - T4, five experimental settings are listed in the Table 6.

ZSL: In S.0, directly tested on four unseen tasks T5 - T8, our performance (Shadow: 0.1732, Transparent: 0.1475, Breast: 0.4895, Skin: 0.6387) is close to or even exceeds that of existing expert models in Table 2, such as EBLNet: 0.1383, AAU-net: 0.4745.

CL: In S.4, our method achieves the gains of -4.7% (average performance of T1 - T4), + 63% (T5), + 59% (T6), + 63% (T7), and + 44% (T8) relative to the results in S.0. Our model has a tolerable performance degradation on old tasks and significant gains on finetuned new tasks.

IZSL: For T6, the model in S.0 has a BER of 0.1475. The model in S.1 has a BER of 0.1260 which achieves a relative improvement of 15%. We can see that a significant performance improvement can be accomplished for T6 when implementing continuous learning once. For T8, mDice scores are 0.6381, 0.6424, and 0.7074 for the models in S.0, S.1, and S.2, respectively. The results in S.1 are almost unchanged with respect to those in S.0, and the results in S.2 show a relative improvement of 11%. We can see that a significant performance improvement can be accomplished for T8 when implementing continuous learning twice.

We found that the data-level correlation between old and new tasks may affect the performance of IZSL. Learning a new task once does not always immediately show improved performance on future tasks. It is important for IZSL to conduct multiple continuous learning processes to accumulate data diversity for model learning.
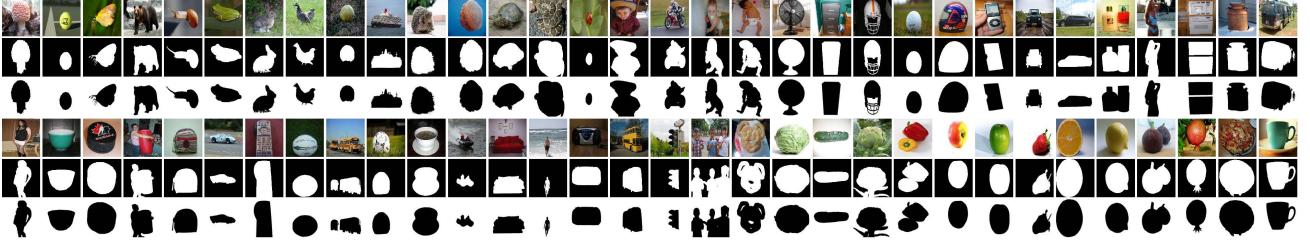
*Figure 10.* Visualization of clustered group prompts for Salient Object Detection.
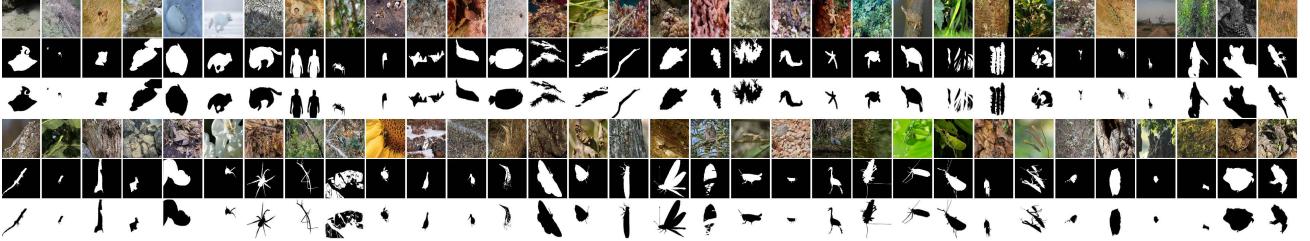
*Figure 11.* Visualization of clustered group prompts for Camouflaged Object Detection.
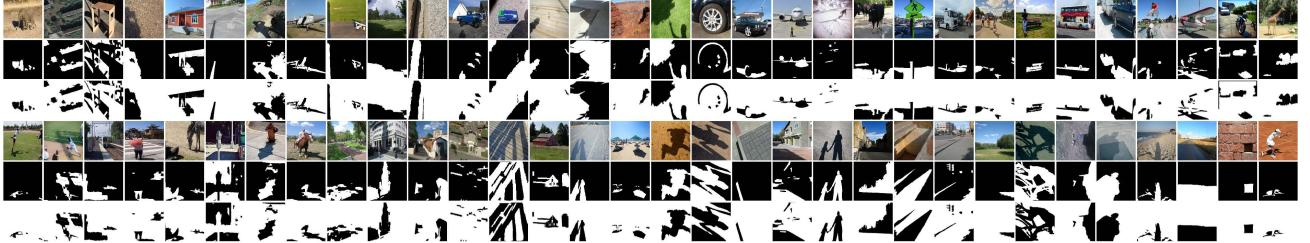
*Figure 12.* Visualization of clustered group prompts for Shadow Detection.

Figure 13. Visualization of clustered group prompts for Transparent Object Segmentation.

A.9. Capability of In-Context Learning

In-context learning usually refers to completing predictions on untrained tasks and samples by providing some prompts to the model. In Figure 26, we separately provide our Spider with image-mask group prompts on video object segmentation and industrial surface defect detection tasks. It can be seen that Spider can capture the specified types of moving objects and defect areas. Therefore, we believe that Spider can have more powerful in-context learning capabilities with the increasing of data scale and diversity.
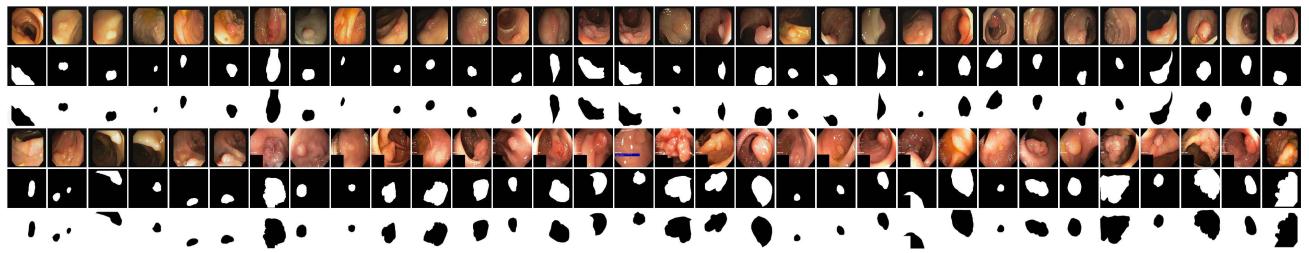
Figure 14. Visualization of clustered group prompts for Polyp Segmentation.
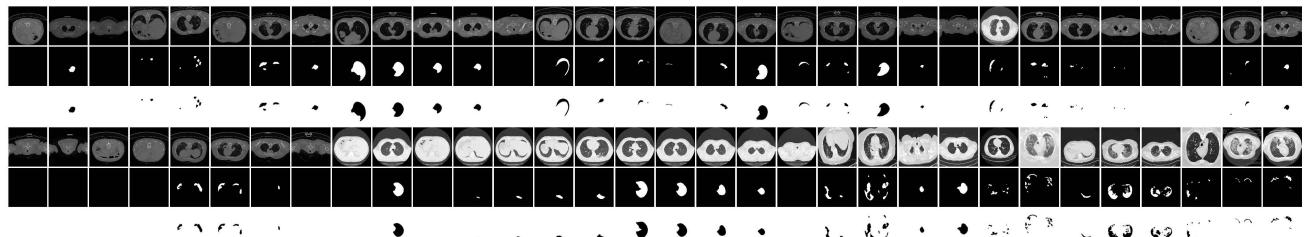
Figure 15. Visualization of clustered group prompts for COVID-19 Lung Infection.
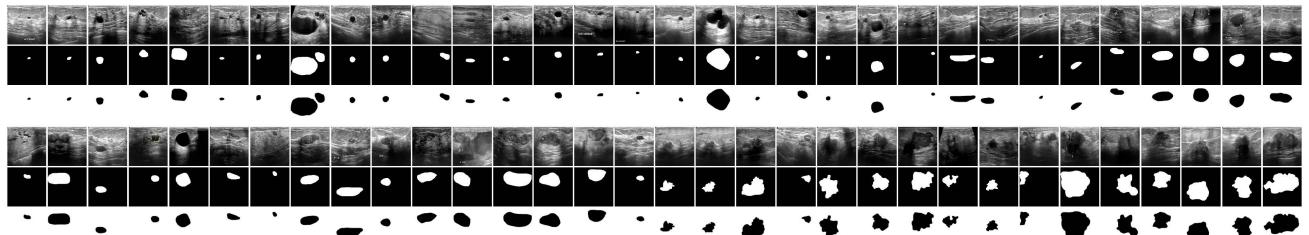
Figure 16. Visualization of clustered group prompts for Breast Lesion Segmentation.
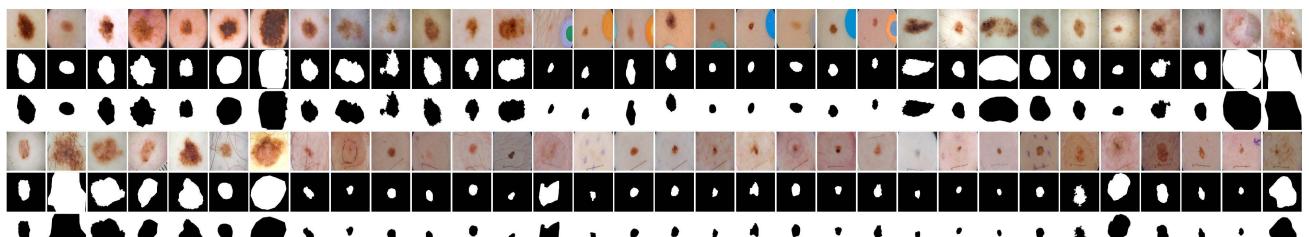
Figure 17. Visualization of clustered group prompts for Skin Lesion Segmentation.
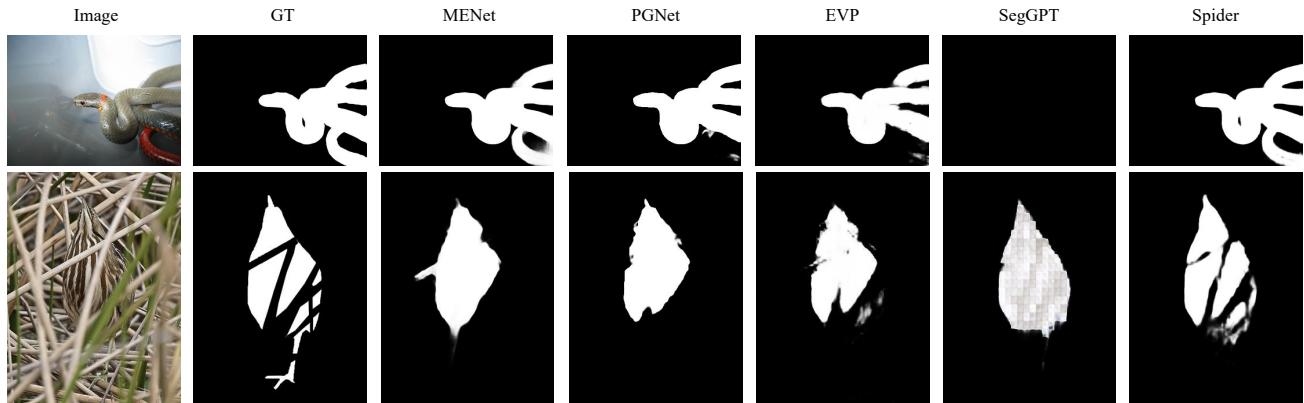
Figure 18. Qualitative comparisons with Salient Object Detection methods. MENet (Wang et al., 2023d) and PGNet (Xie et al., 2022a) are the specialized models.
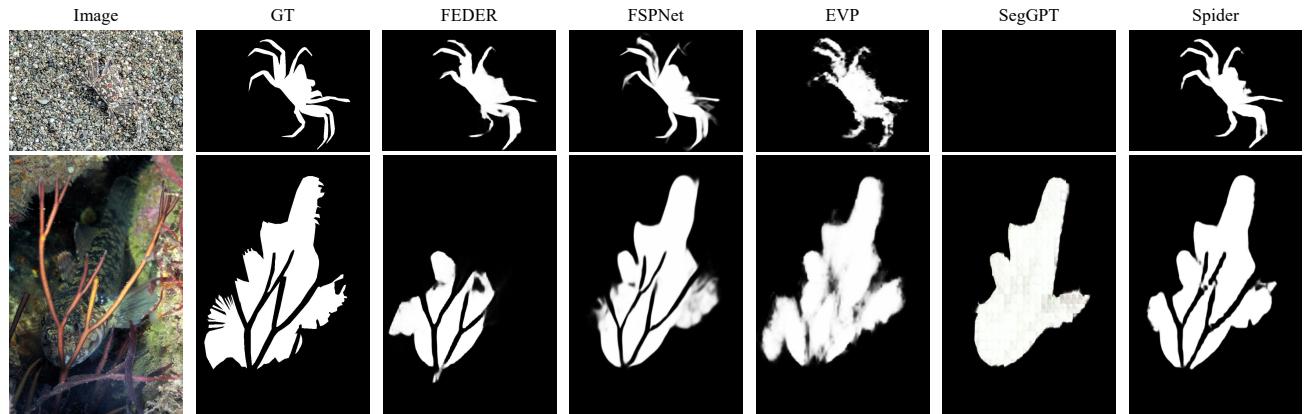

Figure 19. Qualitative comparisons with Camouflaged Object Detection methods. FEDER (He et al., 2023) and FSPNet (Huang et al., 2023b) are the specialized models.
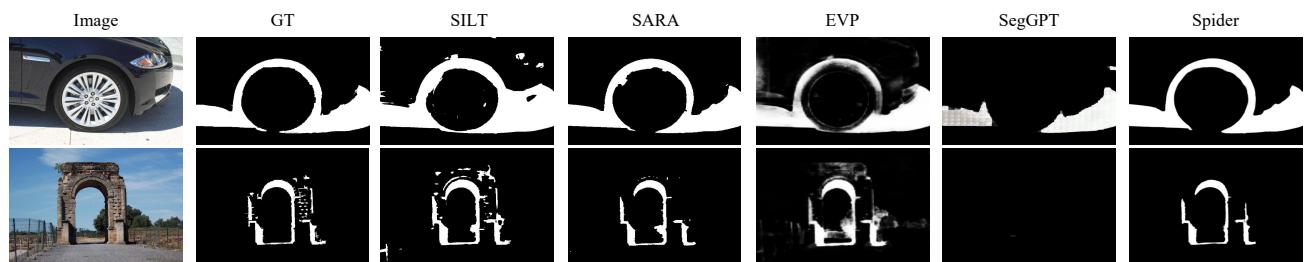
Figure 20. Qualitative comparisons with Shadow Detection methods. SILT (Yang et al., 2023) and SARA (Sun et al., 2023) are the specialized models.
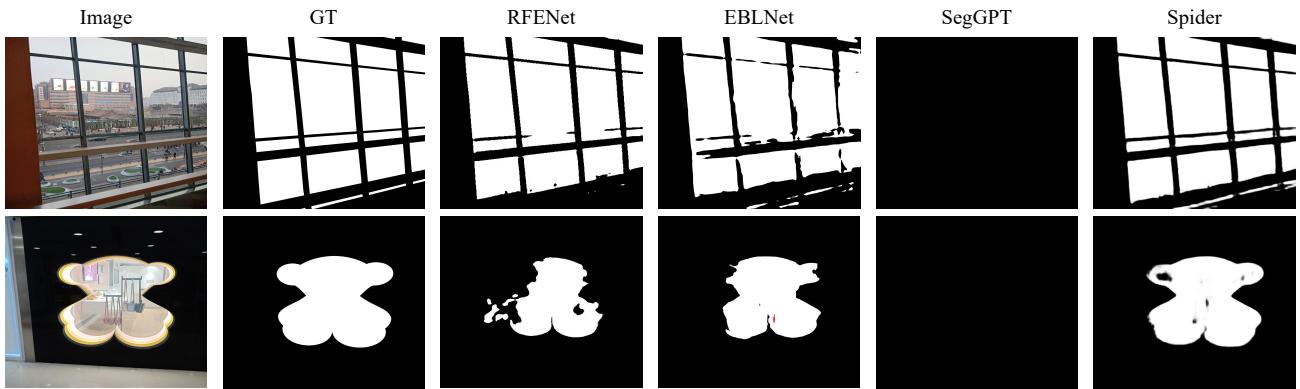
Figure 21. Qualitative comparisons with Transparent Object Segmentation methods. RFENet (Fan et al., 2023) and EBLNet (He et al., 2021) are the specialized models.
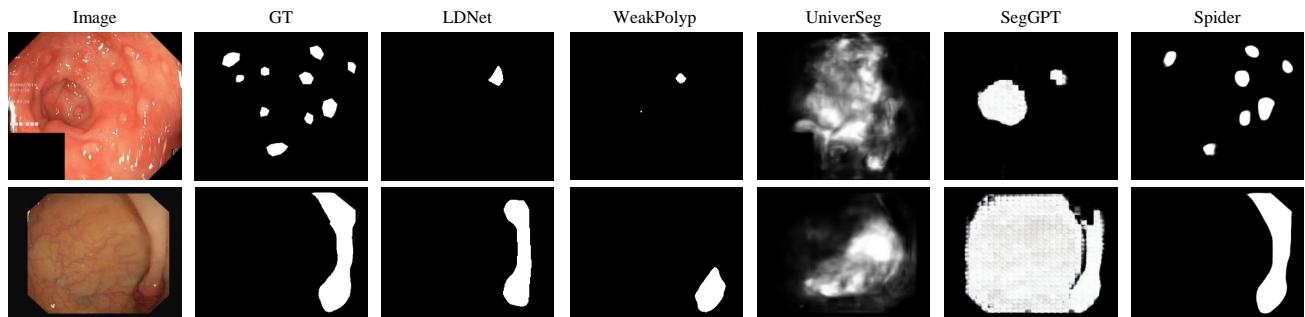
Figure 22. Qualitative comparisons with Polyp Segmentation methods. LDNet (Zhang et al., 2022b) and WeakPolyp (Wei et al., 2023) are the specialized models.
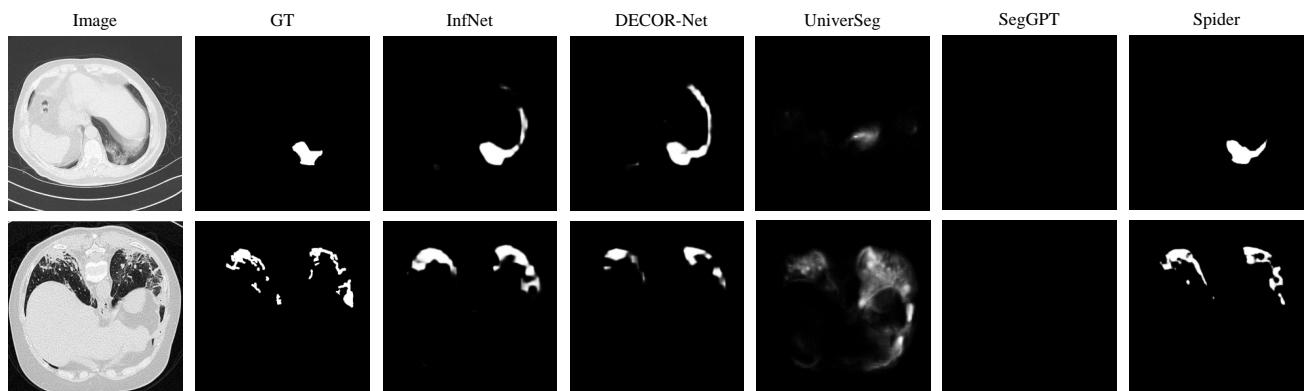

Figure 23. Qualitative comparisons with COVID-19 Lung Infection methods. InfNet (Fan et al., 2020c) and DECOR-Net (Hu et al., 2023) are the specialized models.
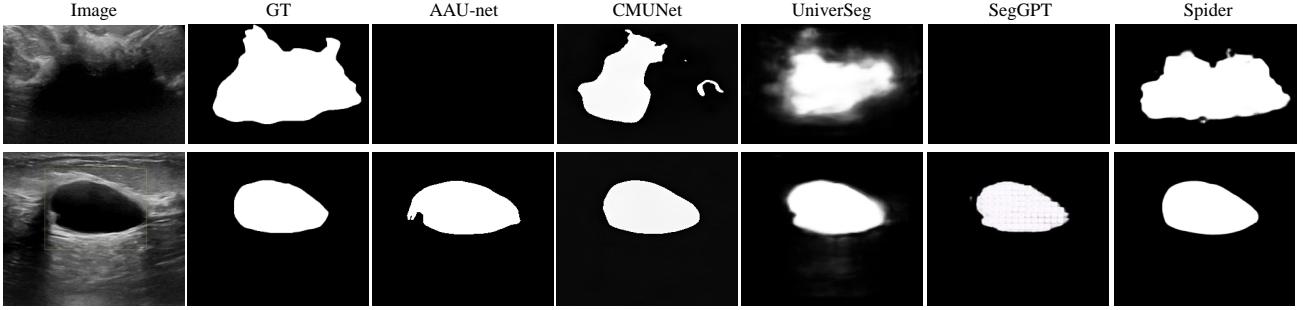
Figure 24. Qualitative comparisons with Breast Lesion Segmentation methods. AAU-net (Chen et al., 2022a) and CMUNet (Tang et al., 2023a) are the specialized models.
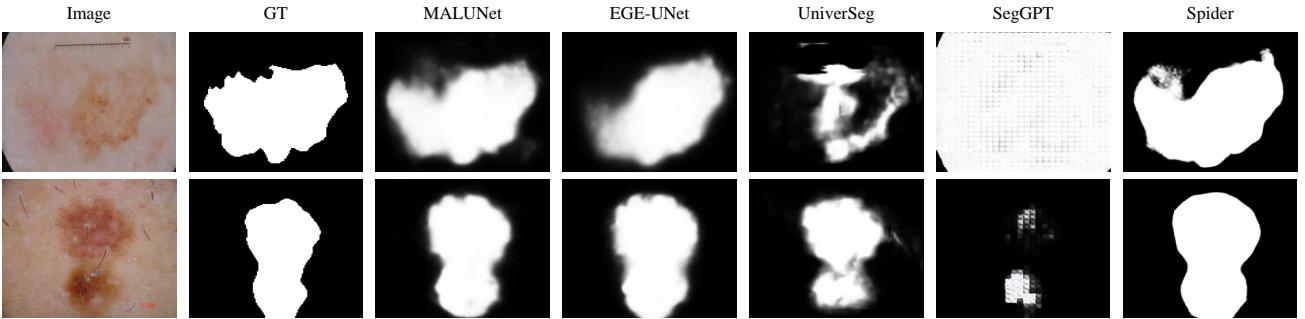

Figure 25. Qualitative comparisons with Skin Lesion Segmentation methods. MALUNet (Ruan et al., 2022) and EGE-UNet (Ruan et al., 2023) are the specialized models.
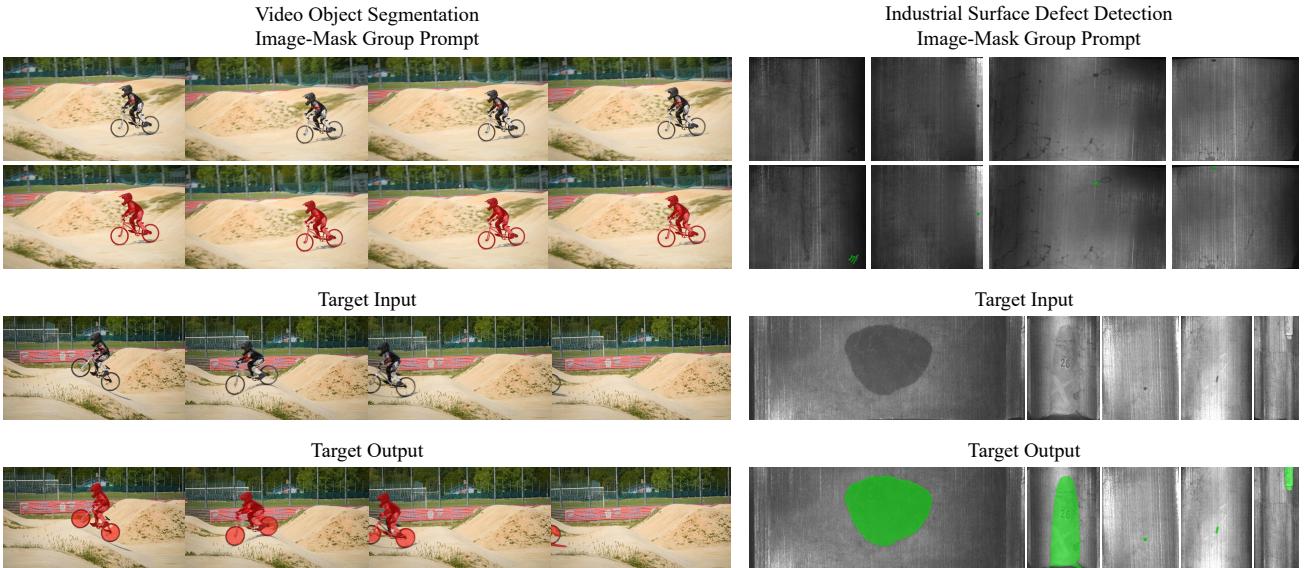

Figure 26. Visual results on the video object segmentation and industrial surface defect detection tasks.