

Controllable-LPMoE: Adapting to Challenging Object Segmentation via Dynamic Local Priors from Mixture-of-Experts

Yanguang Sun¹, Jiawei Lian¹, Jian Yang², Lei Luo¹ *

¹PCA Lab, Nanjing University of Science and Technology, Nanjing, China

²PCA Lab, VCIP, College of Computer Science, Nankai University, Tianjin, China

{Sunyg, lianjw}@njust.edu.cn, csjyang@nankai.edu.cn, luoleipitt@gmail.com

Abstract

Large-scale foundation models provide powerful feature representations for downstream object segmentation tasks. However, when adapted to specific tasks through the full-parameter fine-tuning, the enormous parameters being updated often results in significant computational overhead, creating a bottleneck in training efficiency. Although existing methods attempt to fine-tune frozen models by directly embedding trainable prompts, these prompts lack inherent semantic priors, limiting the adaptability of large-scale models. In this paper, we propose a novel dynamic priors-based fine-tuning paradigm with fewer trainable parameters, dubbed Controllable-LPMoE, which adaptively modulates frozen foundation models by dynamically controlling local priors to enhance fine-grained perception for specific segmentation tasks. More specifically, we construct a lightweight dynamic mixed local priors extractor that captures diverse local priors from input images through heterogeneous convolutions while employing a gating network to dynamically output expert priors required for the subsequent fine-tuning. Furthermore, we design a bi-directional interaction adapter that employs cosine-aligned deformable attention and channel-oriented adaptive scale enhancement to interact and restructure between frozen and trainable features, achieving efficient fine-tuning. Extensive experiments validate the superiority of our Controllable-LPMoE approach, demonstrating excellent segmentation performance compared to 31 state-of-the-art (SOTA) methods and adaptability to multiple binary object segmentation tasks.

1. Introduction

Binary object segmentation, as a fundamental task in computer vision, has been widely studied and encompasses multiple directions, including camouflaged object detection (COD) [22, 23, 52, 58, 59], salient object detection (SOD)

[19, 43, 69, 89], polyp segmentation (PS) [16, 68, 85], skin lesion segmentation (SLS) [20, 76], shadow detection (SD) [34, 42, 44], glass detection (GD) [17, 24, 41], among others. Over the past few years, numerous deep learning-based methods [19, 25, 30, 58, 61, 70] have been proposed, contributing to substantial advances in this field.

The efficient extraction and encoding of high-quality features from input images is a critical factor for achieving accurate binary object segmentation tasks. Early research [51, 56, 69, 84] usually utilizes pre-trained convolutional neural networks (*e.g.*, ResNet50 [26] with 25.6M parameters or VGG16 [54] with 14.7M parameters) as feature encoders, which have relatively few parameters and can be adapted to specific tasks through full-parameter fine-tuning. Later, Vision Transformers [12, 45, 67] exhibit strong feature modeling by integrating self-attention and feed-forward networks at each encoding layer. Based on these structures, employing Transformers [30, 43, 46, 52, 58] via full-parameter fine-tuning (as depicted in Fig. 1 (a)) has become the mainstream architecture in binary object segmentation tasks, achieving a significant performance breakthrough. However, during the training process, a series of issues appeared consecutively, the most prominent being a sharp increase in memory consumption and a notable decline in training speed, both resulting from the substantial increase in parameters within Vision Transformers [12, 67]. For example, in the challenging COD task, the ZoomXNet [52] approach adopts PVTv2-b5 [67], which has 82M parameters. The trainable parameters for the FSPNet [30], FSEL [58], and CamoFormer [79] methods are 273.7M, 67.1M, and 71.3M. Similarly, in the VST++ [43] model for the SOD task, 112.2M parameters need to be updated, *etc.*

It is evident that when larger-scale Transformer models (*e.g.*, BEiT-L [1] with 307M parameters or UniPerceiver-L [88] with 302M parameters) with deeper layers and more parameters, which possess stronger modeling capabilities, are used for feature encoding, the feasibility of this fine-tuning paradigm becomes negligible. Recently, the prompt-based fine-tuning paradigm [32] has been proposed in vi-

*Corresponding author.

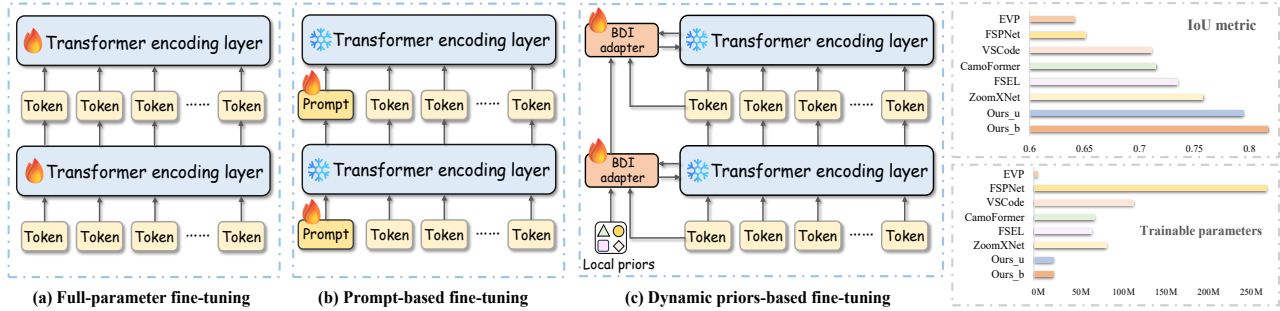


Figure 1. Training paradigms in binary object segmentation tasks: (a) Full-parameter fine-tuning, which updates all model parameters for task adaptation; (b) Prompt-based fine-tuning, which guides learning through trainable prompt embeddings; (c) The proposed dynamic priors-based fine-tuning, which utilizes dynamically controllable local priors to efficiently adapt large-scale models [1, 88]. Additionally, we provide the trainable parameters using different paradigm methods and their IoU accuracy on the challenging COD10K [15] dataset.

sual recognition tasks, which embeds trainable prompts (as shown in Fig. 1(b)) with few parameters into frozen Transformer layers to enable large-scale models to adapt to specific tasks. Inspired by this, EVP [44] and VSCoDe [46] introduce prompt learning into the binary object segmentation task, acquiring task-specific knowledge through different prompts to fine-tune Transformer [45, 75] to segment objects in various scenarios. Although the training efficiency of these models [44, 46] has improved considerably, the segmentation accuracy remains unsatisfactory. As shown in Fig. 1, on the extremely difficult COD10K [15] dataset, the performance of EVP [44] and VSCoDe [46] is considerably lower than that of ZoomXNet [52] and FSEL [58], which adopt full-parameter fine-tuning. The reasons for this can be attributed to two aspects: **1**) it fails to fully leverage the powerful modeling of large-scale models, and **2**) it is closely related to the inherent properties of prompts. In particular, direct-generated prompts lack prior knowledge, making it challenging to refine object details during iterative training. Furthermore, simply embedding trainable prompts fails to adequately incorporate task-specific knowledge into frozen features, influencing the final segmentation performance.

Taking these reasons into account, we propose a novel dynamic priors-based fine-tuning paradigm in this paper, named Controllable-LPMoE. As illustrated in Fig. 1, our method involves few trainable parameters (*i.e.*, 23.4M), yet achieves high segmentation accuracy, benefiting from the efficient fine-tuning of large-scale models. Technically, we construct a lightweight dynamic mixed local priors (DMLP) extractor to generate dynamically controllable local priors with task-specific knowledge through multiple heterogeneous convolutions [4, 11, 18, 28] and a mixture-of-experts (MoE) strategy [35] from input images for subsequent fine-tuning of large-scale foundation models. Moreover, we design a bi-directional interaction (BDI) adapter to facilitate information transfer between the trainable and frozen features, progressively reconstructing their internal informa-

tion through iterative updates of cosine-aligned deformable attention and channel-oriented adaptive scale enhancement components. Ultimately, optimized features not only retain powerful universal representations from large-scale models but also incorporate task-specific knowledge, making them efficiently adaptable to segmentation tasks. Extensive experiments on 18 widely-used benchmark datasets from 6 binary object segmentation tasks demonstrate that our Controllable-LPMoE model consistently outperforms 31 state-of-the-art (SOTA) methods. In summary, the main contributions can be summarized as follows:

- A novel dynamic priors-based fine-tuning paradigm is proposed for adapting large-scale models to binary object segmentation tasks through fewer trainable parameters.
- A lightweight dynamic mixed local priors (DMLP) extractor is designed to dynamically capture various local priors using different convolutions and the MoE strategy.
- An efficient bi-directional interaction (BDI) adapter is introduced to reconstruct the representations of trainable and frozen features, leveraging them through interaction.

2. Related Work

Binary object segmentation tasks. Binary object segmentation is a fundamental research in computer vision, which aims to achieve precise detection and complete segmentation of object regions from input images by constructing a series of frameworks. As a fundamental research, it involves various tasks such as camouflage object detection [21, 30, 33, 52, 81], salient object detection [14, 43, 57, 60, 72], medical image segmentation [16, 68, 80, 85], shadow detection [6, 34, 44], glass detection [17, 25, 41], and so on. Although the categories of objects differ significantly, the design of task-specific architectures is highly similar. The most widely adopted architecture [3, 13, 17, 58, 89] uses a baseline [26, 45, 67] pre-trained on ImageNet [53] to extract initial features, employs the well-designed decoder to generate binary masks, and optimizes the model through a full-

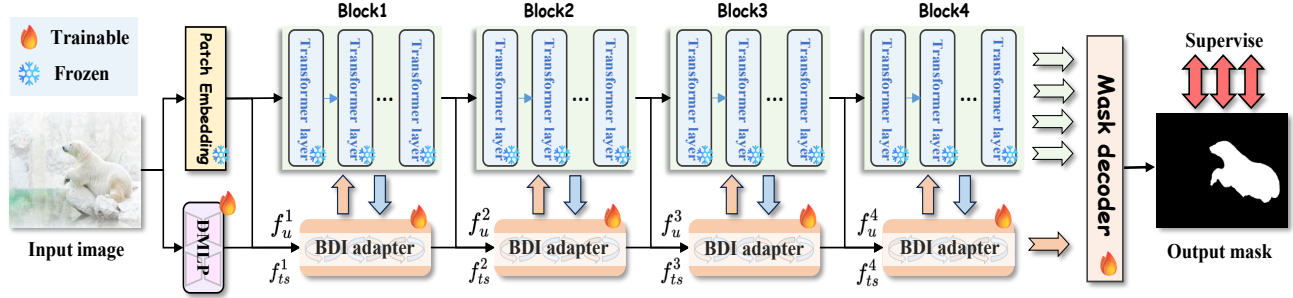


Figure 2. Overview of our Controllable-LPMoE framework. The entire architecture consists of a frozen foundation network [1, 88], a dynamic mixed local priors (DMLP) extractor, four bi-directional interaction (BDI) adapters, and a mask decoder [8]. During training, the proposed Controllable-LPMoE model requires only 23.4M trainable parameters while achieving excellent performance.

parameter fine-tuning paradigm. Excellent performance has been achieved over the past five years [17, 64, 71, 78, 82]. However, with the explosive growth of large-scale models [1, 50, 88] and the presence of hundreds of millions of training parameters, this paradigm faces noticeable limitations, such as computational resource constraints, which limit development in the field. Therefore, establishing an efficient fine-tuning paradigm that better adapts large-scale models and maximizes their advantages in feature modeling is meaningful for the future advancement of these tasks.

Fine-tuning of large-scale models. Deep-level structures offer exceptional feature modeling for large-scale foundation models, but also introduce massive amounts of parameters. Recently, some methods [32, 44, 46] have explored embedding trainable prompts in large-scale models and integrating them into frozen Transformer layers. By updating a subset of parameters, the model adapts to specific visual tasks. To be specific, VPT [32] introduced a small amount of task-specific learnable parameters into the input space for recognition tasks. EVP [44] used features from frozen patch embedding and high-frequency components as prompts to fine-tune SegFormer [75] for low-level structure segmentation. OneTracker [27] designed the CMT Prompter and TTP Transformer layer to adapt the Foundation Tracker to downstream RGB+X tracking tasks. VS-Code [46] exploited 2D prompts to learn the peculiarities across domain and task dimensions for multimodal SOD and COD tasks. Despite the promising performance of task-specific and multi-task models [27, 32, 44, 46], their prompts often lack semantic knowledge and rarely consider the efficient embedding of trainable prompts with frozen structures, which may lead to suboptimal results.

In this paper, we propose an innovative dynamic priors-based fine-tuning paradigm, called Controllable-LPMoE, which introduces a lightweight dynamic mixed local priors (DMLP) extractor to generate dynamic local priors enriched with task-specific knowledge from input images. Additionally, it constructs a cosine-aligned deformable atten-

tion (CDA) for adaptive bi-directional interaction, enabling the efficient fine-tuning of large-scale models to segmentation tasks while utilizing only a few trainable parameters.

3. Methodology

3.1. Overall Architecture

Fig. 2 illustrates the complete framework of the proposed Controllable-LPMoE method, which consists of four parts: (a) BEiT-L [1] / UniPerceiver-L [88] foundation encoding model with frozen parameters. (b) Dynamic mixed local priors (DMLP) extractor. (c) Bi-directional interaction (BDI) adapter. (d) Mask decoder [8]. For an input image \mathcal{I}_c with size $\mathcal{I}_c \in \mathbb{R}^{3 \times H \times W}$, we perform feature encoding in two branches (*i.e.*, task-universal branch, and task-specific branch). The task-universal branch is a large-scale model [1, 88] with frozen parameters that encodes initial features $\{f_u^i\}_{i=1}^5$, which contain powerful universal representations, each with a size of $\frac{H}{16} \times \frac{W}{16}$. The task-specific branch is a lightweight, trainable DMLP extractor that generates task-specific features $\{f_s^i\}_{i=1}^4$, each enriched with local priors for the following fine-tuning. Each feature has a spatial resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$. Furthermore, we integrate attribute information from both branches using the BDI adapter to generate discriminative features, which are then utilized for binary segmentation through a mask decoder [8].

3.2. Dynamic Mixed Local Priors Extractor

The purpose of our DMLP extractor in the task-specific branch is to capture rich local priors and dynamically control their output for subsequent fine-tuning. During the fine-tuning process, these local priors provide task-specific knowledge for segmentation tasks, while the plentiful spatial details they contain help refine the boundary information of objects. Unlike these spatial priors [7, 73], the local priors obtained by our DMLP extractor are dynamic and diverse. Technically, given an input image \mathcal{I}_c in the first stage (as shown in Fig. 3), we employ multiple sets of lightweight heterogeneous convolutions (*i.e.*, depthwise

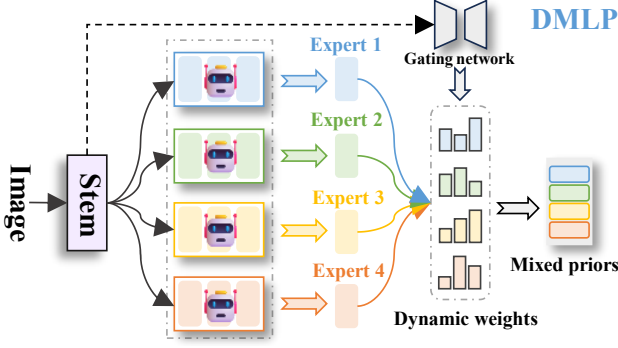


Figure 3. Flowchart of the first stage of our DMLP extractor.

separable convolution [28], atrous convolution [4], asymmetric convolution [11], and wavelet convolution [18]) with different receptive fields to construct four types of local priors $\{E_p^n\}_{n=1}^4$ that contain task-specific knowledge, which can be written as follows:

$$\begin{aligned} E_p^n &= \mathcal{C}_1([l_1^n, l_3^n, l_5^n, l_7^n]), l_1^n = \mathcal{C}_1(\text{stem}(\mathcal{I}_c)), \\ l_{2k+1}^n &= \mathcal{Z}C_{2k+1}^n(l_1^n + l_{2k-1}^n), k = 1, 2, 3, \end{aligned} \quad (1)$$

where $\mathcal{C}_1(\cdot)$, $[\cdot]$, and $\text{stem}(\cdot)$ denote 1×1 convolution, concatenation, and down-sampling operation. l_{2k+1}^n and l_{2k-1}^n are local priors of the n -th type, with different receptive fields. $\mathcal{Z}C_{2k+1}^n(\cdot)$ represents n -th type of lightweight convolution, with a kernel of size $(2k+1) \times (2k+1)$. Considering the diversity [89] of local expert priors $\{E_p^n\}_{n=1}^4$, we propose a dynamic control strategy (DCS) that corrects the proportion of all local priors through dynamic weighting. Specifically, inspired by Mixture of Experts (MoE) [35], we first treat each local prior E_p^n as an expert with different knowledge, and then generate a set of dynamic weights $\{w_i^n\}_{n=1}^4$ through a gating network based on the input feature $\hat{\mathcal{I}}_c$ ($\hat{\mathcal{I}}_c = \text{stem}(\mathcal{I}_c)$), as shown in:

$$w_i^n(\hat{\mathcal{I}}_c) = \text{Softmax}(W_g \hat{\mathcal{I}}_c + \mathbf{b}_g), n = 1, 2, 3, 4, \quad (2)$$

where W_g and \mathbf{b}_g represent the learnable weight matrix and bias vector from the linear layer. “+” denotes the element-wise addition operation. Furthermore, four experts $\{E_p^n\}_{n=1}^4$ with different prior knowledge are integrated using dynamic weights $\{w_i^n\}_{n=1}^4$ to generate the task-specific feature f_s^1 , which incorporates enriched local prior semantics. The process is formulated as follows:

$$f_s^1 = \mathcal{C}_1(\hat{\mathcal{I}}_c + \sum_{n=1}^N w_i^n \otimes E_p^n), N = 4, \quad (3)$$

where “ \otimes ” is the element-wise multiplication. Our dynamic control strategy (DCS) makes all local priors dynamically controllable, and it can continuously adapt and adjust during the fine-tuning process. The proposed DMLP extractor

includes four stages, with each stage using the task-specific feature $\{f_s^{i-1}\}_{i=2}^4$ from the previous stage as input and generating the feature $\{f_s^i\}_{i=2}^4$ through similar operations (*i.e.*, local prior extraction and dynamic integration).

3.3. Bi-directional Interaction Adapter

For large-scale models, the simplest and most straightforward manner for applying them to downstream segmentation tasks is to train and fit them directly through full-parameter fine-tuning. However, existing large-scale foundation models [1, 50, 88] consist of many layers, and their parameters grow exponentially. Updating all parameters through gradient descent and adapting them to segmentation tasks is extremely time-consuming and labor-intensive. Although some existing methods [44, 46] attempt to fine-tune large-scale models with frozen parameters by embedding trainable prompts in binary object segmentation tasks, the lack of semantic priors in the generated prompts hinders their ability to effectively leverage the advantages of large-scale pre-trained models for feature modeling.

Considering the above challenges, we design the BDI adapter that includes cosine-aligned deformable attention (CDA) and channel-oriented adaptive scale enhancement (CASE) to exchange information between frozen and trainable features and iteratively update them. On the one hand, the fine-tuning of large-scale models is guided by dynamic local priors with semantic knowledge; on the other hand, rather than simply embedding prompts, the bi-directional interaction between the two features in the BDI adapter facilitates efficient information transfer, further enhancing the fine-tuning performance of large-scale models.

Input features. The BDI adapter leverages the outputs of the frozen model as task-universal features, while incorporating the abundant local semantic priors from the trainable DMLP extractor as task-specific features. Specifically, we evenly divide the frozen encoding model [1, 88] into four blocks, each containing six encoding layers, and universal features $\{f_u^i\}_{i=1}^5$ ($f_u^i \in \mathbb{R}^{\frac{HW}{16^2} \times D}$) are obtained from the output of the patch embedding and four blocks. Meanwhile, we flatten and concatenate local features $\{f_s^i\}_{i=2}^4$ to generate an initial task-specific feature f_{ts}^1 , that is,

$$f_{ts}^1 = \text{flat}([f_s^2, f_s^3, f_s^4]) \in \mathbb{R}^{(\frac{HW}{8^2} + \frac{HW}{16^2} + \frac{HW}{32^2}) \times D}. \quad (4)$$

Cosine-aligned deformable attention. For the task-universal feature f_u^1 and the task-specific feature f_{ts}^1 , we perform the first knowledge exchange through our CDA mechanism, with the goal of enriching the universal feature with task-specific clues. Technically, we take the normalized feature \tilde{f}_u^1 ($\tilde{f}_u^1 = \text{Norm}(f_u^1)$) as the primary *query*, where $\text{Norm}(\cdot)$ denotes a layerNorm, and the task-specific feature \tilde{f}_{ts}^1 as the auxiliary *value*. To enhance the semantic alignment between *query* and *value*, we incorporate cosine

similarity into the attention weights, enabling the model to focus more on regions with higher relevance, that is,

$$\begin{aligned} \mathcal{A}_\phi^1 &= \text{Softmax}(\Phi(\tilde{f}_u^1)) \otimes \text{Cosine}(\tilde{f}_u^1, \tilde{f}_{ts}^1), \\ \text{Cosine}(\tilde{f}_u^1, \tilde{f}_{ts}^1) &= \text{Softmax}\left(\Phi\left(\frac{\tilde{f}_u^1 \odot \tilde{f}_{ts}^1}{\|\tilde{f}_u^1\| \|\tilde{f}_{ts}^1\|} \otimes \tilde{f}_{ts}^1\right)\right), \end{aligned} \quad (5)$$

where $\Phi(\cdot)$ denotes a linear layer, \odot is the matrix multiplication operation, $\|\cdot\|$ represents the Euclidean norm. By weighting the input feature, our approach dynamically adjusts the attention weight distribution. Subsequently, we update the internal information of the initial input feature, and the formula can be expressed as:

$$\hat{f}_u^1 = f_u^1 + \Phi(\mathcal{A}_\phi^1 \otimes W_v \tilde{f}_{ts}^1 (p_q^1 + \nabla p_{so}^1)) \otimes \Psi_o, \quad (6)$$

where p_q^1 and ∇p_{so}^1 represent the 2-d reference point related to the *query* and the sampling offset [87], while Ψ_o denotes a learnable vector variable initialized to 0, balancing the attention layer's output and the input *query*.

Channel-oriented adaptive scale enhancement. The optimized feature \hat{f}_u^1 is input into the 1-th encoding block to obtain the output feature f_u^2 . Then, we conduct the second knowledge exchange to enhance the expressive ability of the feature f_{ts}^1 . In detail, as opposed to the first exchange, we take the specific feature f_{ts}^1 as the primary *query*, and the output feature f_u^2 as an auxiliary *value* into our CDA component for interactive fusion, that is,

$$\hat{f}_{ts}^1 = f_{ts}^1 + \text{CDA}(\text{Norm}(f_{ts}^1), \text{Norm}(f_u^2)), \quad (7)$$

where $\text{CDA}(\cdot)$ denotes the proposed CDA mechanism. Furthermore, we construct the CASE to strengthen multi-scale information with the channels to generate the task-specific feature f_{ts}^2 for the next stage of interaction. Technically, we first reinterpret the input feature \hat{f}_{ts}^1 by decomposing it into three features and enhancing its linear expression through the depthwise separable convolution [28] with the 3×3 kernel ($\mathcal{DC}_3(\cdot)$), *i.e.*, $(\hat{f}_{ts}^1)_1, (\hat{f}_{ts}^1)_2, (\hat{f}_{ts}^1)_3 = \mathcal{DC}_3(\text{Split}(\text{Norm}(\hat{f}_{ts}^1)))$. Then, we dynamically regulate significant clues within the channel from two perspectives through the channel and reverse attentions [5, 29]. Similarly to our DMLP extractor, we regard the outputs from two perspectives as two experts and adaptively fuse them using dynamic weights w_t^x ($w_t^x = \text{Softmax}(W_g \hat{f}_{ts}^1 + b_g)$) generated by a gating network. The process is as follows:

$$\begin{aligned} f_{ts}^2 &= \hat{f}_{ts}^1 + \text{flat}([\hat{f}_s^2, \hat{f}_s^3, \hat{f}_s^4]), \hat{f}_s^k = \sum_{x=1}^X w_t^x \otimes (E_c^x)_k, \\ (E_c^1)_k, (E_c^2)_k &= \text{CA}((\hat{f}_{ts}^1)_k), \text{RA}((\hat{f}_{ts}^1)_k), k = 2, 3, 4, \end{aligned} \quad (8)$$

where $\text{flat}(\cdot)$ is a flattening operation, $\text{CA}(\cdot)$ and $\text{RA}(\cdot)$ represent the channel [29] and reverse [5] attentions, respectively. Similarly, the obtained features f_{ts}^2 and f_u^2 interactively fuse in the 2-th block to generate the features f_{ts}^3 and f_u^3 . The entire fine-tuning continues until the 4-th block.

3.4. Loss functions

After the interaction is completed, the optimized features of each block are input into a lightweight Transformer-based mask decoder [8], which contains 1.08M parameters, for decoding and output. During the fine-tuning process, we use the binary cross-entropy loss and the Dice coefficient loss to supervise the training of our model, as follows:

$$\mathcal{L}_{all} = \alpha \mathcal{L}_{bce} + \beta \mathcal{L}_{dice}, \quad (9)$$

where α and β represent the hyperparameters set to 5 and 2.

4. Experiment

4.1. Experimental Settings

Datasets. We evaluate our Controllable-LPMoE method on multiple binary object segmentation tasks, including camouflaged object detection (COD), salient object detection (SOD), polyp segmentation (PS), skin lesion segmentation (SLS), shadow detection (SD), and glass detection (GD). For COD, we utilize CAMO-TR[37] and COD10K-TR [15] as joint training datasets and evaluate the accuracy in CHAMELEON [55], CAMO-TE [37], COD10K-TE [15], and NC4K [47] datasets. In SOD, DUTS-TR [66] is employed as the training dataset, while performance is assessed on PASCAL-S [40], ECSSD [77], HKU-IS [39], and DUTS-TE [66]. Regarding PS, we use training images from CVC-ClinicDB [62] and Kvasir [31] to train the model and validate its performance on test images from CVC-300 [2], CVC-ClinicDB [62], and Kvasir [31]. For SLS, we train/test performance in the ISIC17 [10] and ISIC18 [9] datasets, respectively. In SD, we train the model on the training images of SBU [63] and ISTD [65] and evaluate its performance of UCF [86], SBU [63], and ISTD [65]. In addition, we use Trans10k [74] and GDD [48] datasets for both training and testing in the GD task. More details of the datasets are presented in the **supplementary materials**.

Implementation details. All experiments are conducted on four NVIDIA GTX 4090 GPUs, each equipped with 24GB of memory. We utilize the frozen BEiT-L [1] and UniPerceiver-L [88] frameworks, adapting them for binary segmentation tasks through efficient fine-tuning. During fine-tuning, input images are resized to 512×512 , the batch size is set to 4, and the initial learning rate is $5e-5$. The entire training process runs for 80K iterations, with the proposed model optimized using the AdamW optimizer.

Evaluation metrics. We use four evaluation metrics to verify the superiority of our model, including mean Intersection over Union (IoU), mean Dice Coefficient (Dice), weighted F-measure (\mathcal{F}_m^w), and mean absolute error (\mathcal{M}). Better segmentation results are indicated by larger scores for IoU, Dice, and \mathcal{F}_m^w , along with a smaller \mathcal{M} value.

Methods	Pub.	CHAMELEON				CAMO				COD10K				NC4K			
		IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow
PFNet ₂₁ [49]	CVPR	0.751	0.831	0.810	0.033	0.611	0.721	0.695	0.085	0.588	0.697	0.660	0.040	0.670	0.769	0.745	0.053
JSOCOD ₂₁ [38]	CVPR	0.776	0.849	0.833	0.030	0.649	0.750	0.728	0.073	0.612	0.714	0.684	0.035	0.698	0.789	0.771	0.047
ZoomNet ₂₂ [51]	CVPR	0.785	0.856	0.845	0.023	0.675	0.773	0.752	0.066	0.656	0.749	0.729	0.029	0.714	0.800	0.784	0.043
SegMaR ₂₂ [33]	CVPR	0.804	0.871	0.860	0.025	0.675	0.773	0.753	0.071	0.656	0.753	0.724	0.034	-	-	-	-
FDNet ₂₂ [84]	CVPR	0.769	0.855	0.836	0.027	0.702	0.801	0.775	0.063	0.651	0.759	0.730	0.030	0.673	0.774	0.750	0.052
SAM ₂₃ [36]	ICCV	0.560	0.647	0.639	0.081	0.522	0.611	0.606	0.132	0.616	0.698	0.701	0.049	0.615	0.695	0.696	0.078
PopNet ₂₃ [70]	ICCV	0.824	0.887	0.875	0.020	0.666	0.761	0.744	0.077	0.690	0.779	0.757	0.028	0.734	0.817	0.802	0.042
FSPNet ₂₃ [30]	CVPR	0.786	0.858	0.851	0.023	0.721	0.811	0.799	0.050	0.651	0.750	0.735	0.026	0.742	0.825	0.816	0.035
FEDER ₂₃ [21]	CVPR	0.775	0.850	0.834	0.030	0.660	0.763	0.738	0.071	0.640	0.741	0.716	0.032	0.713	0.804	0.789	0.044
EVP ₂₃ [44]	CVPR	0.707	0.799	0.777	0.038	0.674	0.777	0.762	0.067	0.641	0.748	0.726	0.032	-	-	-	-
VSCoDe ₂₄ [46]	CVPR	-	-	-	-	0.757	0.843	0.820	0.046	0.711	0.801	0.780	0.023	0.778	0.854	0.841	0.032
FSEL ₂₄ [58]	ECCV	0.825	0.893	0.877	0.022	0.792	0.872	0.851	0.040	0.735	0.822	0.800	0.021	0.792	0.866	0.853	0.030
CamoFormer ₂₄ [79]	TPAMI	0.805	0.877	0.865	0.022	0.768	0.851	0.831	0.046	0.715	0.805	0.786	0.023	0.784	0.859	0.847	0.030
ZoomXNet ₂₄ [52]	TPAMI	0.829	0.891	0.885	0.018	0.797	0.869	0.857	0.041	0.758	0.839	0.827	0.018	0.799	0.870	0.863	0.028
Ours_u	-	0.856	0.913	0.908	0.016	0.825	0.893	0.875	0.035	0.795	0.866	0.858	0.015	0.826	0.887	0.881	0.024
Ours_b	-	0.863	0.917	0.913	0.015	0.834	0.899	0.883	0.035	0.817	0.883	0.876	0.014	0.842	0.901	0.896	0.022

Table 1. Comparison with state-of-the-art methods on three camouflaged object detection datasets. The top three results are highlighted in orange, teal, and blue. “Ours_u” and “Ours_b” denotes the fine-tuning of different frameworks, *i.e.*, UniPerceiver [88] and BEiT [1].

Methods	Pub.	PASCAL-S				ECSSD				HKU-IS				DUTS-TE			
		IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow	IoU \uparrow	Dice \uparrow	\mathcal{F}_m^w \uparrow	\mathcal{M} \downarrow
MENet ₂₃ [69]	CVPR	0.797	0.865	0.844	0.054	0.881	0.924	0.920	0.031	0.872	0.922	0.917	0.023	0.817	0.880	0.870	0.028
ICON ₂₃ [89]	TPAMI	0.810	0.877	0.853	0.051	0.899	0.939	0.933	0.024	0.882	0.931	0.925	0.022	0.833	0.896	0.882	0.026
GPONet ₂₄ [78]	PR	0.797	0.868	0.845	0.054	0.894	0.937	0.932	0.025	0.869	0.922	0.918	0.023	0.817	0.883	0.872	0.028
MDSAM ₂₄ [19]	MM	0.812	0.876	0.857	0.051	0.913	0.948	0.946	0.021	0.893	0.937	0.935	0.019	0.842	0.889	0.893	0.024
FSEL ₂₄ [58]	ECCV	0.797	0.866	0.838	0.057	0.894	0.936	0.928	0.026	0.866	0.919	0.909	0.027	0.800	0.866	0.847	0.037
VSCoDe ₂₄ [46]	CVPR	0.815	0.878	0.859	0.050	0.910	0.946	0.942	0.021	0.886	0.933	0.930	0.021	0.847	0.904	0.896	0.024
VST ₊₊₂₄ [43]	TPAMI	0.801	0.870	0.846	0.054	0.890	0.934	0.926	0.026	0.867	0.921	0.914	0.025	0.810	0.878	0.866	0.029
Ours_u	-	0.842	0.898	0.882	0.041	0.917	0.947	0.944	0.021	0.909	0.946	0.945	0.016	0.870	0.916	0.900	0.021
Ours_b	-	0.840	0.897	0.879	0.043	0.927	0.955	0.953	0.017	0.906	0.945	0.943	0.017	0.861	0.908	0.900	0.024

Table 2. Comparison with state-of-the-art methods on four salient object detection datasets.

4.2. Comparison with the State-of-the-Art

We compare the performance of the Controllable-LPMoE model against 31 state-of-the-art methods from six different binary object segmentation tasks. In particular, the prediction maps for all competing methods are either provided directly by their respective authors or obtained by training their publicly available open-source code.

Quantitative evaluation. Tables 1-5 present the quantitative results of our model and 31 existing segmentation approaches. From Table 1, the “IoU” metric has improved in four widely utilized COD datasets, increasing by 4.10%, 4.64%, 7.78%, and 5.38% over the recent ZoomXNet [52] method, and by 4.61%, 5.30%, 11.16%, and 6.31% over the recent FSEL [58] method in the highly challenging COD task. For the SOD task in Table 2, compared to the recently proposed VSCoDe [46] model, our Controllable-LPMoE model achieves overall improvements of 21.95%, 23.53%, 31.25%, and 14.29% on four public datasets in terms of the “ \mathcal{M} ” metric. Similarly, the proposed Controllable-LPMoE method demonstrates significant superiority across various metrics in other segmentation tasks, as detailed in Tables 3, 4, and 5. This performance advantage stems from the joint fine-tuning of our DMLP extractor and BDI adapter, which enables the internal features within large-scale models to be

efficiently adapted for binary segmentation tasks.

Qualitative evaluation. Fig. 4 illustrates visual comparison results in various scenarios. As depicted in Figure 4, the proposed Controllable-LPMoE method demonstrates superior segmentation accuracy across different objects, generating predicted maps that not only retain complete object structures, but also exhibit sharp and well-defined edge details. In contrast, some existing methods [52, 58, 68, 79] struggle to achieve this level of precision.

4.3. Ablation Study

To verify the contribution of each key design and the rationale behind its internal structures, we conduct extensive ablation studies based on the UniPerceiver [88] framework.

Effect of each component. In Table 6, we give the quantitative results of each component in the proposed Controllable-LPMoE method. Specifically, the “baseline” (Table 6(a)) includes a UniPerceiver [88] network with frozen parameters and a mask decoder [8]. Table 6(b) validates the effectiveness of our “DMLP” extractor, demonstrating that embedding dynamic local priors with semantic aids in adapting frozen frameworks to binary object segmentation tasks. Furthermore, as shown in Tables 6(c) and (e), the proposed “CDA” component significantly im-

Methods	Pub.	CVC-300				CVC-ClinicDB				Kvasir				ISIC17				ISIC18			
		IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$
PraNet ₂₀ [16]	MICCAI	0.797	0.871	0.843	0.010	0.849	0.899	0.896	0.009	0.840	0.898	0.885	0.030	0.776	0.853	0.830	0.050	0.806	0.881	0.864	0.057
DCRNet ₂₂ [80]	ISBI	0.788	0.856	0.830	0.010	0.844	0.896	0.890	0.010	0.825	0.886	0.868	0.035	0.789	0.866	0.847	0.046	0.802	0.876	0.854	0.059
CFANet ₂₃ [85]	PR	0.827	0.893	0.875	0.008	0.883	0.932	0.924	0.007	0.861	0.915	0.903	0.023	0.793	0.844	0.815	0.051	0.809	0.868	0.846	0.061
LSSNet ₂₄ [68]	MICCAI	0.815	0.884	0.852	0.009	0.875	0.920	0.914	0.010	0.866	0.911	0.895	0.028	0.813	0.881	0.867	0.038	0.824	0.886	0.867	0.054
LBUNet ₂₄ [76]	MICCAI	0.680	0.785	0.734	0.019	0.713	0.797	0.855	0.029	0.748	0.831	0.805	0.048	0.800	0.872	0.864	0.038	0.803	0.879	0.866	0.055
MEGANet ₂₄ [3]	WACV	0.818	0.887	0.863	0.009	0.885	0.930	0.931	0.008	0.859	0.911	0.904	0.026	0.800	0.878	0.864	0.039	0.809	0.885	0.873	0.052
FSEL ₂₄ [58]	ECCV	0.814	0.880	0.856	0.009	0.867	0.914	0.910	0.011	0.852	0.899	0.894	0.027	0.813	0.885	0.871	0.035	0.821	0.885	0.875	0.052
Ours.u	-	0.844	0.910	0.897	0.005	0.887	0.930	0.932	0.006	0.870	0.915	0.914	0.021	0.820	0.890	0.885	0.032	0.827	0.897	0.886	0.045
Ours.b	-	0.839	0.904	0.888	0.006	0.896	0.935	0.939	0.006	0.885	0.930	0.928	0.017	0.814	0.887	0.878	0.034	0.817	0.889	0.878	0.049

Table 3. Comparison with state-of-the-art methods on three polyp segmentation and two skin lesion segmentation datasets.

Methods	Pub.	SBU				UCF				ISTD				Trans10k				GDD			
		IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$	IoU \uparrow	Dice \uparrow	$F_m^w\uparrow$	$M\downarrow$
RMLA ₂₃ [34]	TCSVT	0.798	0.878	0.839	0.032	0.669	0.780	0.713	0.064	0.910	0.947	0.931	0.012	0.888	0.934	0.911	0.044	0.884	0.929	0.910	0.055
SDSAM ₂₃ [6]	TGRS	0.787	0.841	0.816	0.042	0.678	0.746	0.670	0.071	0.891	0.921	0.902	0.023	0.838	0.903	0.860	0.067	0.900	0.937	0.917	0.051
EVP ₂₃ [44]	CVPR	0.815	0.853	0.809	0.038	0.667	0.745	0.672	0.071	0.861	0.886	0.855	0.031	0.889	0.930	0.906	0.046	0.903	0.937	0.921	0.049
FSEL ₂₄ [58]	ECCV	0.835	0.893	0.875	0.028	0.703	0.793	0.745	0.057	0.908	0.945	0.930	0.013	0.896	0.934	0.914	0.043	0.871	0.919	0.897	0.061
Spider ₂₄ [83]	ICML	0.823	0.893	0.868	0.027	-	-	-	-	-	-	-	-	0.892	0.934	0.913	0.043	0.906	0.942	0.924	0.047
Ours.u	-	0.857	0.914	0.909	0.022	0.736	0.831	0.797	0.042	0.941	0.965	0.959	0.008	0.930	0.960	0.947	0.027	0.923	0.952	0.941	0.039
Ours.b	-	0.861	0.917	0.912	0.022	0.736	0.828	0.795	0.043	0.916	0.947	0.935	0.015	0.931	0.961	0.948	0.027	0.922	0.952	0.940	0.037

Table 4. Comparison with recent state-of-the-art methods on three shadow detection datasets.

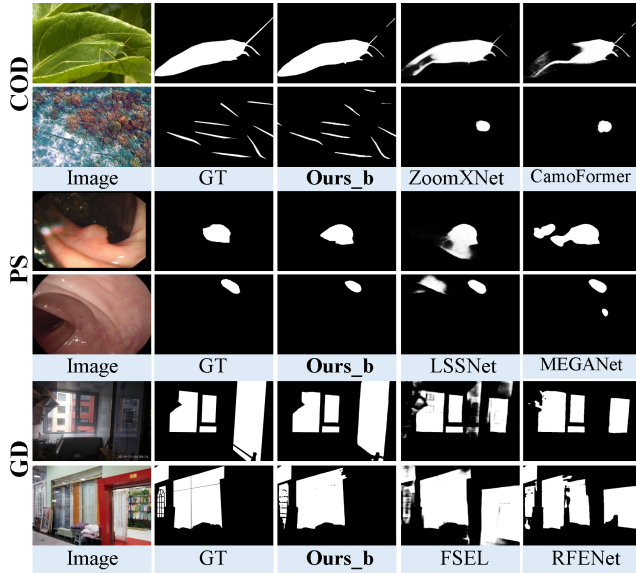
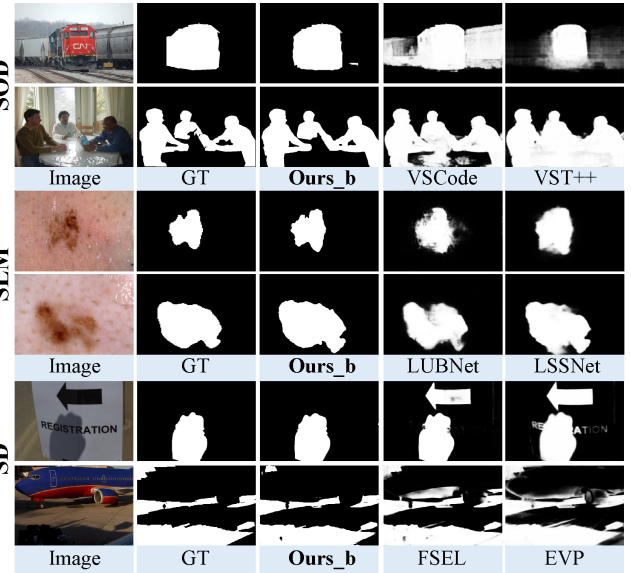


Figure 4. Visual comparison results of our Controllable-LPMoE with multiple state-of-the-art methods on six binary segmentation tasks.

proves segmentation accuracy through the effective interaction between frozen and trainable features. Subsequently, we incorporate the designed “CASE” into the BDI adapter (as shown in Table 6 (d) and (f)), further improving the performance by optimizing the scale information of the task-specific features and adaptively adjusting the significant clues within the channels. Furthermore, in Fig. 5, we present the visual results obtained by gradually adding each component (*i.e.*, DMLP, CDA, and CASE), demonstrating that the predicted map gradually approaches the ground truth (GT). In short, each component is necessary and collectively improves the “baseline” by 27.38%, 17.02%, 21.93%, and 11.32% under the “IoU” metric.

Effect of local priors within the DMLP extractor. Do

Table 5. Comparison with recent state-of-the-art methods on two glass detection datasets.



we really need various local priors? To answer this question, we assess the impact of each local prior in the proposed DMLP extractor (as depicted in Table 7(a)-(d)). These results show that incorporating local priors enhances model performance, benefiting both from the semantic prompts they carry and the rich spatial details they provide. Moreover, we conduct an experimental analysis of the number of experts in Table 7 (e) and (f). Furthermore, we analyze the dynamic control strategy (DCS). Table 7(g) represents the fusion of all local priors using “element-wise addition”, while Table 7(h) denotes the aggregation through a gating network with dynamic weights, with the latter performing better. In conclusion, the design of the proposed DMLP extractor is both well-reasoned and effective.

Num.	Structure Settings				CHAMELEON		CAMO		COD10K		NC4K	
	Base.	DMLP	CDA	CASE	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$
(a)	✓				0.672	0.741	0.705	0.774	0.652	0.729	0.742	0.809
(b)		✓			0.787	0.848	0.781	0.841	0.752	0.820	0.803	0.861
(c)			✓		0.799	0.864	0.803	0.862	0.754	0.824	0.812	0.872
(d)				✓	0.816	0.877	0.804	0.859	0.763	0.832	0.817	0.874
(e)	✓	✓	✓		0.839	0.897	0.821	0.876	0.774	0.841	0.819	0.876
(f)	✓	✓	✓	✓	0.856	0.908	0.825	0.875	0.795	0.858	0.826	0.881

Table 6. Ablation study of individual components in the proposed Controllable-LPMoE framework on challenging COD tasks.

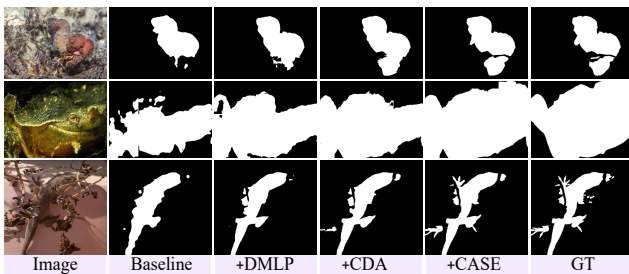


Figure 5. Visual results of the effectiveness of each component.

Num.	BDI Adapter Settings		CHAMELEON		CAMO		COD10K		NC4K	
	$t \rightarrow f$	$f \rightarrow t$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$
(a)	✓		0.796	0.858	0.779	0.840	0.740	0.811	0.796	0.856
(b)		✓	0.787	0.848	0.781	0.841	0.752	0.820	0.803	0.861
(c)	✓	✓	0.799	0.864	0.803	0.862	0.754	0.824	0.812	0.872

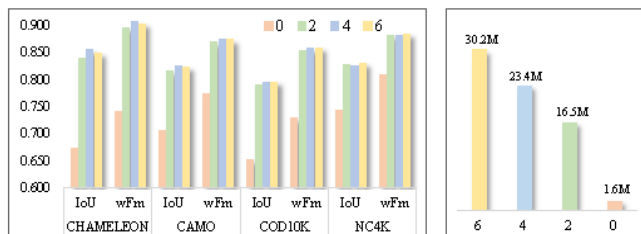
Table 8. Ablation study on the bi-directional interaction architecture of the proposed BDI Adapter. “ f ” and “ t ” denote frozen features and trainable features, respectively.

Effect of bi-directional interaction within the BDI adapter. Bi-directional interaction aims to achieve efficient fine-tuning by leveraging frozen features to enhance the generality of trainable features. Conversely, when the focus shifts to frozen features, frozen features are imbued with task-specific attributes. Table 8 (a) and (b) present the quantitative results for different features as subjects of interaction. Compared to the bi-directional strategy (Table 8 (c)), it is evident that a single interaction performs significantly worse. Furthermore, we conduct an experimental analysis on the impact of the number of interactions. As illustrated in Fig. 6, an increase in the interaction numbers leads to a corresponding increase in the trainable parameters, which in turn enhances performance to a certain extent. To strike a balance between efficiency and performance, we set the number of interactions to 4. These results highlight the effectiveness of the proposed BDI adapter.

Efficiency analysis. In Table 9, we present key metrics of our method under different training paradigms (*i.e.*, full-parameter fine-tuning and dynamic priors-based fine-tuning), including the trainable parameters, the memory required for training, the time consumed per 50 iterations, and the corresponding performance. From Table 9, com-

Num.	DMLP Extractor Settings					CHAMELEON		CAMO		COD10K		NC4K	
	E_1	E_2	E_3	E_4	DCS	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$
(a)	✓					0.741	0.808	0.757	0.821	0.712	0.786	0.786	0.847
(b)		✓				0.741	0.806	0.760	0.820	0.711	0.783	0.784	0.844
(c)			✓			0.730	0.800	0.750	0.812	0.707	0.780	0.781	0.842
(d)				✓		0.734	0.802	0.754	0.816	0.713	0.786	0.786	0.846
(e)	✓	✓				0.755	0.821	0.765	0.832	0.714	0.791	0.790	0.852
(f)	✓	✓	✓			0.767	0.828	0.765	0.830	0.730	0.800	0.794	0.852
(g)	✓	✓		✓		0.757	0.824	0.764	0.825	0.720	0.793	0.790	0.850
(h)	✓	✓	✓	✓	✓	0.787	0.848	0.781	0.841	0.752	0.820	0.803	0.861

Table 7. Ablation study on the internal structure of our DMLP Extractor. “ E_1 ”-“ E_4 ” represent different local prior knowledge.



(a) Performance of different numbers of interactions. (b) Trainable parameters.

Figure 6. Ablation analysis of different interaction numbers. Here, we set the number of interactions to 0, 2, 4, and 6, respectively

Methods	Train.	Train.	Train.	CHAMELEON		CAMO		COD10K		NC4K	
	parameters	memory	times	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$	IoU \uparrow	$\mathcal{F}_m^w\uparrow$
Ours.u \dagger	326.7M	18.63G	0.907	0.861	0.911	0.840	0.888	0.815	0.875	0.838	0.890
Ours.u	23.4M	13.75G	0.819	0.856	0.908	0.825	0.875	0.795	0.858	0.826	0.881
Ours.b \dagger	328M	13.32G	1.032	0.866	0.914	0.851	0.898	0.825	0.882	0.842	0.895
Ours.b	23.4M	8.55G	0.908	0.863	0.913	0.834	0.884	0.817	0.876	0.842	0.896

Table 9. Efficiency analysis for our proposed method, where “ \dagger ” indicates training with the full-parameter fine-tuning strategy.

pared to the full-parameter fine-tuning, the number of parameters required for training is only about 1/14, significantly reducing computational costs. Meanwhile, memory consumption during training decreased by 35.49% and 55.79%, while training speed improved by 10.74% and 13.66%. Although its performance is slightly lower than the full-parameter fine-tuning, the overall performance remains excellent. These results further demonstrate the efficiency of our dynamic priors-based fine-tuning paradigm.

5. Conclusion

In this paper, we propose a novel Controllable-LPMoE method, specifically designed for fine-tuning large-scale models to adapt to binary object segmentation tasks. First, we develop a lightweight DMLP extractor, which generates task-specific features enriched with dynamic local priors, thereby providing more effective support for fine-tuning. Second, we design the BDI adapter, which facilitates efficient interaction between frozen and trainable features to update both types of information. Extensive experiments demonstrate that our method obviously surpasses 31 SOTA models in 18 binary object segmentation datasets.

Acknowledgments

This work was supported in part by the National Science Fund of China (No. 62276135, U24A20330 and 62361166670).

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [2] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *PR*, 45(9):3166–3182, 2012.
- [3] Nhat-Tan Bui, Dinh-Hieu Hoang, Quang-Thuc Nguyen, Minh-Triet Tran, and Ngan Le. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *WACV*, pages 7985–7994, 2024.
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.
- [5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018.
- [6] Xiao-Diao Chen, Wen Wu, Wenya Yang, Hongshuai Qin, Xiantao Wu, and Xiaoyang Mao. Make segment anything model perfect on shadow detection. *TGRS*, 61:1–13, 2023.
- [7] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023.
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022.
- [9] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kaloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the isic. *arXiv preprint arXiv:1902.03368*, 2019.
- [10] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kaloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 isbi. In *ISBI*, pages 168–172, 2018.
- [11] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *ICCV*, pages 1911–1920, 2019.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [13] Songsong Duan, Xi Yang, and Nannan Wang. Multi-label prototype visual spatial search for weakly supervised semantic segmentation. In *CVPR*, pages 30241–30250, 2025.
- [14] Songsong Duan, Xi Yang, Nannan Wang, and Xinbo Gao. Lightweight rgb-d salient object detection from a speed-accuracy tradeoff perspective. *TIP*, 34:2529–2543, 2025.
- [15] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020.
- [16] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273, 2020.
- [17] Ke Fan, Changan Wang, Yabiao Wang, Chengjie Wang, Ran Yi, and Lizhuang Ma. Rfenet: towards reciprocal feature evolution for glass segmentation. In *IJCAI*, pages 717–725, 2023.
- [18] Shahaf E Finder, Roy Amoyal, Eran Treister, and Oren Freifeld. Wavelet convolutions for large receptive fields. In *ECCV*, pages 363–380, 2024.
- [19] Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *ACM MM*, pages 9894–9903, 2024.
- [20] Yilin Guo and Qingling Cai. Bgdiffseg: A fast diffusion model for skin lesion segmentation via boundary enhancement and global recognition guidance. In *MICCAI*, pages 150–159, 2024.
- [21] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pages 22046–22055, 2023.
- [22] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *ICLR*, 2024.
- [23] Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Longxiang Tang, Yulun Zhang, Linghe Kong, and Sina Farsiu. Segment concealed object with incomplete supervision. *TPAMI*, 2025.
- [24] Chunming He, Rihan Zhang, Fengyang Xiao, Chenyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. Run: Reversible unfolding network for concealed object segmentation. *ICML*, 2025.
- [25] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinnet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, pages 15859–15868, 2021.
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [27] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, pages 19079–19091, 2024.
- [28] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

- [29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.
- [30] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, pages 5557–5566, 2023.
- [31] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MM*, pages 451–462, 2020.
- [32] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022.
- [33] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, pages 4713–4722, 2022.
- [34] Leiping Jie and Hui Zhang. Rmlanet: Random multi-level attention network for shadow detection and removal. *TCSVT*, 33(12):7819–7831, 2023.
- [35] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.
- [36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023.
- [37] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranh network for camouflaged object segmentation. *CVIM*, 184:45–56, 2019.
- [38] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, pages 10071–10081, 2021.
- [39] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *CVPR*, pages 5455–5463, 2015.
- [40] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.
- [41] Jiaying Lin, Yuen-Hei Yeung, and Rynson Lau. Exploiting semantic relations for glass surface detection. *NeurIPS*, 35: 22490–22504, 2022.
- [42] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Scotch and soda: A transformer video shadow detection framework. In *CVPR*, pages 10449–10458, 2023.
- [43] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *TPAMI*, 46(11):7300–7316, 2024.
- [44] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pages 19434–19445, 2023.
- [45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.
- [46] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscope: General visual salient and camouflaged object detection with 2d prompt learning. In *CVPR*, pages 17169–17180, 2024.
- [47] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021.
- [48] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don’t hit me! glass detection in real-world scenes. In *CVPR*, pages 3687–3696, 2020.
- [49] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021.
- [50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [51] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.
- [52] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *TPAMI*, 2024.
- [53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.
- [54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [55] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczczyk, Tomasz Depta, Adam Kornacki, and Przemysław Koziel. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.
- [56] Yanguang Sun, Chenxing Xia, Xiuju Gao, Bin Ge, Hanling Zhang, and Kuan-Ching Li. Emcenet: efficient multi-scale context exploration network for salient object detection. In *ICIP*, pages 1066–1070, 2022.
- [57] Yanguang Sun, Chenxing Xia, Xiuju Gao, Hong Yan, Bin Ge, and Kuan-Ching Li. Aggregating dense and attentional multi-scale feature network for salient object detection. *DSP*, 130:103747, 2022.
- [58] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *ECCV*, pages 343–360, 2024.
- [59] Yanguang Sun, Hanyu Xuan, Jian Yang, and Lei Luo. Glconet: Learning multisource perception representation for camouflaged object detection. *TNNLS*, 2024.
- [60] Yanguang Sun, Jian Yang, and Lei Luo. United domain cognition network for salient object detection in optical remote sensing images. *TGRS*, 62:3497579, 2024.

- [61] Yanguang Sun, Jiexi Yan, Jianjun Qian, Chunyan Xu, Jian Yang, and Lei Luo. Dual-perspective united transformer for object segmentation in optical remote sensing images. *arXiv preprint arXiv:2506.21866*, 2025.
- [62] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *TMI*, 35(2):630–644, 2015.
- [63] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pages 816–832, 2016.
- [64] Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-sar: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-sar fusion object detection. *arXiv preprint arXiv:2505.10931*, 2025.
- [65] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018.
- [66] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.
- [67] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*, 8(3):415–424, 2022.
- [68] Wei Wang, Huiying Sun, and Xin Wang. Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In *MICCAI*, pages 446–456. Springer, 2024.
- [69] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xi-angjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *CVPR*, pages 10031–10040, 2023.
- [70] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *ICCV*, pages 1032–1042, 2023.
- [71] Chenxing Xia, Yanguang Sun, Xiuju Gao, Bin Ge, and Song-song Duan. Dminet: dense multi-scale inference network for salient object detection. *TVC*, 38(9):3059–3072, 2022.
- [72] Chenxing Xia, Yanguang Sun, Kuan-Ching Li, Bin Ge, Hanling Zhang, Bo Jiang, and Ji Zhang. Rcnnet: Related context-driven network with hierarchical attention for salient object detection. *ESWA*, 237:121441, 2024.
- [73] Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *CVPR*, pages 5493–5502, 2024.
- [74] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, pages 696–711, 2020.
- [75] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021.
- [76] Jiahao Xu and Lyuyang Tong. Lb-unet: A lightweight boundary-assisted unet for skin lesion segmentation. In *MICCAI*, pages 361–371, 2024.
- [77] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.
- [78] Yugen Yi, Ningyi Zhang, Wei Zhou, Yanjiao Shi, Gengsheng Xie, and Jianzhong Wang. Gponet: A two-stream gated progressive optimization network for salient object detection. *PR*, 150:110330, 2024.
- [79] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *TPAMI*, 2024.
- [80] Zijin Yin, Kongming Liang, Zhanyu Ma, and Jun Guo. Duplex contextual relation network for polyp segmentation. In *ISBI*, pages 1–5, 2022.
- [81] Chenxi Zhang, Qing Zhang, Jiayun Wu, and Youwei Pang. Cgcod: Class-guided camouflaged object detection. *arXiv preprint arXiv:2412.18977*, 2024.
- [82] Yingzhen Zhang, Jimin Dai, Qianliang Wu, Jian Yang, and Lei Luo. Dcnnot: Diffusion-cascaded neural optimal transport for scalable multi-domain image-to-image translation. In *ACM MM*, 2025.
- [83] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: a unified framework for context-dependent concept segmentation. In *ICML*, pages 60906–60926, 2024.
- [84] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR*, pages 4504–4513, 2022.
- [85] Tao Zhou, Yi Zhou, Kelei He, Chen Gong, Jian Yang, Huazhu Fu, and Dinggang Shen. Cross-level feature aggregation network for polyp segmentation. *PR*, 140:109555, 2023.
- [86] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, pages 223–230, 2010.
- [87] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.
- [88] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, pages 16804–16815, 2022.
- [89] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *TPAMI*, 45(03):3738–3752, 2023.