H^{2}Net: Homo- and Heterogeneous Networks for Unified Segmentation

Jinyu Han[1], Changguang Wu[1], Fuming Sun[2], Mengyin Wang[2], Jinhui Tang[3*]
[1]Nanjing University of Science and Technology
[2]Dalian Minzu University
[3]Nanjing Forestry University
{hanjinyu, changguangwu}@njust.edu.cn, {sunfuming, wangmengyin}@dlnu.edu.cn,
tangjh@njfu.edu.cn

Abstract

Unified segmentation aims to consolidate multiple vision tasks into a single model, yet faces two core challenges: learning robust homogeneous features (*e.g.*, shared low- and mid-level cues) to enable cross-domain knowledge transfer, while disentangling heterogeneous features (*e.g.*, task-specific semantic objectives) to avoid negative transfer and preserve task independence. To address these challenges, we propose the Homo- and Heterogeneous Network (H^{2}Net), a unified framework that jointly models shared homogeneous representations and task-specific heterogeneous features. Specifically, H^{2}Net incorporates a Cross-Modal Structure Enhancement Module (CSEM), which integrates auxiliary depth priors via joint frequency–spatial cross-modal attention to strengthen task-agnostic structural representations. In addition, a Task Adapter Pool (TAP) is introduced to model task-specific heterogeneous features by assigning dedicated adapters to individual tasks, enabling task-aware feature modulation and semantic disentanglement within a shared backbone. Extensive experiments on benchmarks spanning eight tasks demonstrate the effectiveness of the proposed approach and its superior performance. Code and results will be available at https://h2net-ijcai26.github.io.

1 Introduction

Unified segmentation aims to develop a single architecture capable of addressing multiple heterogeneous vision tasks within one model, representing a critical step toward general-purpose visual perception. In practice, segmentation tasks span diverse domains, including natural image understanding, remote sensing, and medical imaging. Despite substantial differences in data distributions, visual appearances, and semantic definitions, these tasks share common structural requirements, such as accurate boundary localization and reliable spatial reasoning. For instance, salient object detection [Han *et al.*, 2025b], camouflaged object detection [Sun *et al.*, 2025a] [He *et al.*, 2023], and lesion segmentation [Zhu *et al.*, 2025] all rely on precise structural cues. As a result, a unified segmentation framework not only reduces redundancy across task-specific systems but also exploits shared structural consistency among heterogeneous tasks, thereby promoting cross-task synergy and improving generalization across domains [Zhao *et al.*, 2024] [Wang *et al.*, 2023b].

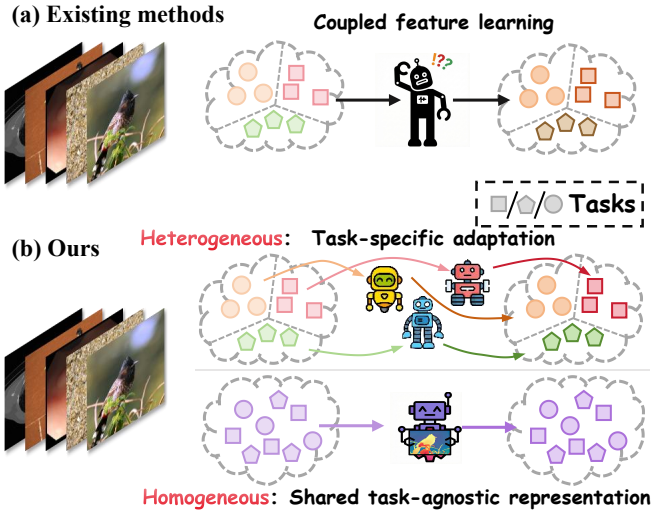

Figure 1: Motivation illustration. (a) Coupled feature modeling in existing methods, where the entanglement of homogeneous and heterogeneous features leads to semantic interference and optimization bias. (b) Our decoupled learning paradigm explicitly disentangles task-agnostic homogeneous features (*e.g.*, structural layouts) from task-specific heterogeneous features to ensure balanced representation learning.

Motivated by this goal, recent studies have explored various architectural paradigms for unified segmentation. For example, SegGPT [Wang *et al.*, 2023c] adopts a prompt-based strategy that leverages visual prompts to enable in-context learning across segmentation tasks, demonstrating strong flexibility under a shared backbone. Furthermore, Spider [Zhao *et al.*, 2024] introduces a concept filtering mechanism to modulate features at the prediction stage, selectively emphasizing task-relevant information. Similarly, VS-Code [Luo *et al.*, 2024] proposes a 2D prompt-based ap-

*Corresponding Author

proach to adapt shared representations to different task requirements. As illustrated in Figure 1(a), existing frameworks predominantly adopt a coupled feature modeling paradigm, in which homogeneous and heterogeneous features are implicitly entangled within a shared representation space. This entanglement merges task-agnostic cues with task-specific semantic objectives, leading to optimization conflicts where dominant gradients from easier or more salient tasks suppress the subtle cues required by more challenging tasks, resulting in semantic interference and biased optimization.

However, they primarily rely on prompt-based mechanisms to handle task differences and largely overlook the decoupling of homogeneous and heterogeneous features during representation learning. To alleviate this issue, as shown in Figure 1(b), we explicitly model homogeneous and heterogeneous features in a decoupled manner. Homogeneous features capture task-agnostic structural cues, such as spatial layouts, enabling the learning of robust and transferable representations across tasks. In contrast, heterogeneous features are learned in task-specific optimization spaces to mitigate inter-task competition and preserve discriminative semantic information. This decoupled design allows task-specific characteristics to be encoded without mutual interference, while fully leveraging a shared homogeneous foundation for unified segmentation.

In this study, we propose H^2Net, a Homo- and Heterogeneous Network that explicitly models homogeneous features and decouples heterogeneous features during the encoding stage. Built upon a pre-trained encoder as a shared backbone, H^2Net restructures task adaptation through a layered design. In the shallow layers, a Cross-Modal Structure Enhancement Module (CSEM) is introduced to integrate auxiliary depth priors [Yang *et al.*, 2024], improving geometric consistency and reinforcing task-agnostic structural representations. Meanwhile, a Task Adapter Pool (TAP) is deployed across all encoding stages to enable task-aware feature modulation: in shallow layers, TAP captures task-related geometric variations under the structural guidance of CSEM, while in deeper layers, TAP further disentangles task-specific semantics, thereby reducing inter-task interference. A shared decoder is then used to generate the final segmentation outputs. Extensive experiments on a wide range of segmentation benchmarks, covering tasks such as SOD, COD, and medical image segmentation, demonstrate the effectiveness of H^2Net as a unified segmentation model.

Our contributions can be summarized as follows:

- We propose the Homo- and Heterogeneous Network H^2Net, a unified segmentation framework that decouples homogeneous and heterogeneous features, enabling accurate localization and robust segmentation across diverse tasks on a single model.
- We introduce a Cross-Modal Structure Enhancement Module (CSEM) to strengthen task-agnostic homogeneous features guided by auxiliary structural priors, resulting in more stable and consistent shared representations.
- We design a Task Adapter Pool (TAP) to disentangle task-specific heterogeneous features by allocating dedicated task-aware adapters within a shared backbone, enabling effective task adaptation while alleviating cross-task interference.
- Extensive experiments on eight benchmarks demonstrate that the H^2Net can effectively handle diverse segmentation tasks, achieving strong performance through joint modeling of homogeneous and heterogeneous representations.

2 Related Work

2.1 Task-specific Segmentation

Task-specific segmentation has been widely studied across diverse vision domains, including salient object detection, shadow detection, remote sensing, and medical image segmentation. These approaches typically rely on specialized architectures or task-tailored mechanisms to handle domain-specific challenges such as low contrast, complex backgrounds, and ambiguous boundaries. For example, Wang et al. [Wang *et al.*, 2023d] enhance salient structures by jointly modeling pixels, regions, and objects, while transformer-based approaches such as [Li *et al.*, 2023] capture long-range contextual dependencies through global self-attention. Sun et al. [Sun *et al.*, 2023] model illumination variations to distinguish shadows from visually similar regions, and Yang et al. [Yang *et al.*, 2023] progressively refine shadow predictions via iterative label tuning. In medical image segmentation, methods such as [Zhu *et al.*, 2025] capture both local and global structural cues to improve robustness. Despite their strong task-specific performance, these approaches rely on specialized designs and optimization objectives, limiting scalability and cross-task knowledge sharing.

2.2 Unified Segmentation and Structural Priors

Unified segmentation aims to integrate multiple segmentation tasks into a shared framework to reduce redundancy and improve generalization. Early approaches mainly adopt multi-task learning strategies, while recent methods explore more flexible task adaptation through prompts or feature modulation, such as SegGPT [Wang *et al.*, 2023c], VSCode [Luo *et al.*, 2024], and Spider [Zhao *et al.*, 2024]. Depth maps have also been introduced as structural priors [Luo *et al.*, 2024], providing geometric cues such as spatial layout and object boundaries; however, the high cost of acquiring real depth data limits their applicability, motivating the use of pseudo-depth generated by monocular depth estimation [Yang *et al.*, 2024] [Wang *et al.*, 2023a] [Changguang *et al.*, 2025]. Despite their flexibility, existing unified segmentation methods generally rely on coupled feature representations and fail to explicitly distinguish cross-task homogeneous representations from task-specific heterogeneous ones, which constrains unified modeling. This limitation motivates our work.

3 Proposed Method

3.1 Overall Architecture

As illustrated in Figure 2, H^2Net is constructed upon a frozen encoder backbone to leverage its robust and generic representation priors. To effectively adapt these representations for

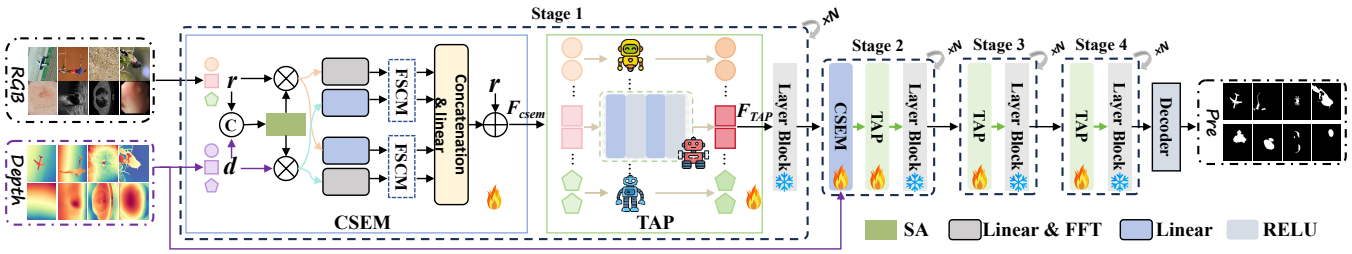

Figure 2: Overview of the proposed H^{2}Net framework. The architecture utilizes a frozen Hiera-L backbone, where the Cross-Modal Structural Enhancement Module (CSEM) is inserted into encoder layers to bridge dual-modal data streams and learn homogeneous features. Subsequently, the Task Adapter Pool (TAP) is embedded to decouple task-specific heterogeneous features, followed by a decoder that generates precise segmentation masks.

unified segmentation, we introduce a Cross-Modal Structure Enhancement Module (CSEM) in the shallow encoder layers. By incorporating auxiliary depth priors, the CSEM calibrates low-level visual responses and reinforces task-agnostic homogeneous features, such as structural layouts and boundary cues, thereby establishing a stable shared representation space. Building on this foundation, the Task Adapter Pool (TAP) allocates task-specific adapters to decouple heterogeneous features, enabling each task to flexibly modulate structural representations according to its intrinsic characteristics. In the deeper layers where high-level semantics dominate, the TAP is applied independently to further disentangle heterogeneous features driven by divergent semantic objectives across tasks. Finally, features from multiple scales are aggregated by a shared decoder to generate the final segmentation predictions.

3.2 Encoder

Our H^{2}Net employs SAM2-Hiera [Ravi *et al.*, 2025] as a fully frozen, shared backbone to maximally preserve its general-purpose representations derived from large-scale pre-training, while effectively mitigating catastrophic forgetting during multi-task optimization. Given that shallow layers primarily encode low-level textures and geometric information, this stage is tasked with not only establishing shared homogeneous features but also accommodating significant structural heterogeneity, thereby compensating for the divergent geometric demands across tasks. As the network deepens and representations become progressively abstract, the deep layers focus exclusively on the semantic space, aiming to disentangle semantic heterogeneity driven by distinct task objectives.

3.3 Cross-Modal Structural Enhancement Module

Homogeneous features in unified segmentation refer to task-agnostic visual cues shared across different tasks, such as structural layouts and boundary continuity. To strengthen this foundation, we introduce monocular depth estimation (MDE) using the pre-trained Depth Anything model [Yang *et al.*, 2024] to generate pseudo-depth maps. Depth cues are less sensitive to appearance variations and task-specific semantics, enabling more consistent geometric alignment across heterogeneous tasks.

To effectively integrate texture-rich RGB features with geometry-aware depth features, we propose the Cross-Modal Structure Enhancement Module (CSEM). As illustrated in Figure 2, CSEM first employs a Spatial Attention (SA) mechanism to pre-align the RGB feature (r) and depth feature (d).

$$\begin{aligned} f_r &= r \times SA(Cat[r, d]), \\ f_d &= d \times SA(Cat[r, d]), \end{aligned} \tag{1}$$

Where $Cat[\cdot]$ represents the concatenation operation, and $\times$ is element-wise multiplication.

Subsequently, to fully exploit the complementary properties of both modalities across different domains, the pre-aligned features enter the Frequency-Spatial Cross-Modal Module (FSCM). In this module, features from the RGB and depth streams explicitly serve as inputs to the frequency domain branches to establish a parallel dual-branch architecture. In the RGB-Frequency branch, RGB features are transformed into the frequency domain to provide global semantic context, interacting with depth features in the spatial domain that offer local geometric constraints. Correspondingly, in the Depth-Frequency branch, depth features are transformed into the frequency domain to capture global structural layouts, interacting with RGB features in the spatial domain that supply fine-grained textural details. This process of constructing frequency-spatial embeddings is described as:

$$\begin{aligned} Q_f^r, K_f^r, V_f^r &= \mathcal{F}(\mathrm{L}(f_r)), q_s^d, k_s^d, v_s^d = \mathrm{L}(f_d), \\ Q_f^d, K_f^d, V_f^d &= \mathcal{F}(\mathrm{L}(f_d)), q_s^r, k_s^r, v_s^r = \mathrm{L}(f_r), \end{aligned} \tag{2}$$

where L is Linear.

The FSCM is designed based on cross-attention mechanisms. Its pivotal role is to leverage the inherent global receptive field of the frequency domain to inject global context into local spatial features, thereby achieving precise synergy between global semantics and local geometry. The attention calculation and fusion process of the FSCM is formulated as follows:

$$\begin{aligned} F_1^{r/d} &= \mathcal{F}^{-1}(Softmax(\frac{Q_f(K_f)^T}{\sqrt{d_k}})) \cdot v_s, \\ F_2^{r/d} &= \mathcal{F}^{-1}(\mathcal{F}(Softmax(\frac{q_s(k_s)^T}{\sqrt{d_k}})) \cdot V_f), \\ F_{csem} &= \mathrm{L}(Cat[F_1^r, F_2^r, F_1^d, F_2^d]), \end{aligned} \tag{3}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fast Fourier Transform and its inverse, respectively.

In summary, CSEM utilizes FSCM to explicitly map the global frequency-domain context of dual-modal features onto local spatial features, thereby facilitating a deep interaction between global semantics and local geometry within the feature space. This process effectively reinforces task-agnostic homogeneous representations that are insensitive to appearance variations.

3.4 Task Adapter Pool

To explicitly decouple heterogeneous features arising from task-specific structural and semantic discrepancies, we introduce the Task Adapter Pool (TAP). Addressing the feature entanglement caused by conflicting inductive biases across tasks, TAP isolates task-specific modulation into dedicated adapter modules while keeping the backbone fully frozen. Specifically, each adapter employs a lightweight bottleneck architecture to model task-specific deviations from the shared representation. It projects the input feature $x \in \mathbb{R}^D$ into a low-dimensional latent space for non-linear transformation before mapping it back to the original dimension, formulated as:

$$\text{Adapter}(x) = \text{L}_{\text{up}}(\text{GELU}(\text{L}_{\text{down}}(x))). \tag{4}$$

Subsequently, the module adopts a residual design where the generated task-specific adjustment is fused with the original input feature x via an addition operation. This design ensures that the adapter focuses on modeling task-specific deviations, namely heterogeneous features, rather than reconstructing the entire representation. To accommodate multiple heterogeneous tasks simultaneously, we establish the Task Adapter Pool, comprising a dedicated set of parameters for each task. During joint multi-task training, data samples are concatenated along the batch dimension to form a mixed batch. The core mechanism of TAP is intra-batch task dispatching, as illustrated in Figure 2, TAP dynamically identifies the source task i of each sample x^i within the mixed batch and routes it exclusively to the corresponding Adapter^i. The generated task-specific residual is then injected into the original feature to yield the modulated output:

$$F^i_{TAP} = x^i + \text{Adapter}^i(x^i). \tag{5}$$

Once this process is applied across all samples, the collection of adapter-modified features, F^i_{TAP}, is reassembled into a single batch. This complete batch is then passed to the subsequent shared and frozen encoder layer block for further processing, yielding the final output:

$$F_{TAP} = Cat[F^i_{TAP}], \quad x_{\text{out}} = \text{LB}(F_{TAP}), \tag{6}$$

where LB is the layer block.

In summary, TAP achieves the explicit disentanglement of heterogeneous features by isolating task-specific modulations into independent parameters. This mechanism effectively resolves feature entanglement caused by conflicting inductive biases, while leveraging joint supervision to ensure that the learned heterogeneous representations remain compatible within the unified decoding framework.

3.5 Decoder

In contrast to the task-specific disentanglement performed in the encoder, the decoder is deliberately designed to be simple and fully shared across all tasks. As illustrated in Figure 2, the decoder consists solely of basic convolution and upsampling operations, without introducing any task-dependent branches or additional adaptation modules. The decoder progressively aggregates multi-scale features from different encoder stages through straightforward concatenation and convolution, and maps the fused representations to dense prediction space. By keeping the decoder lightweight and structurally minimal, we ensure that task-specific reasoning is confined to the encoder, while the decoder functions purely as a unified feature aggregation and prediction head.

Overall, H^{2}Net separates representation learning and prediction by assigning heterogeneous feature disentanglement to the encoder and maintaining a minimal, shared decoder for unified inference.

3.6 Loss Function

To supervise the final prediction P, we employ a hybrid loss function, L_{loss}, which linearly combines the Binary Cross-Entropy (BCE) loss L^{BCE} [De Boer *et al.*, 2005] and the Intersection over Union (IoU) loss L^{IOU} [Shelhamer *et al.*, 2017]:

The hybrid loss is defined as follows:

$$L_{loss} = L^{BCE} + L^{IOU}. \tag{7}$$

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the effectiveness of our unified H^{2}Net, we conducted experiments across eight distinct segmentation tasks. Specifically, we utilized the following datasets: DUTS for salient object detection (SOD), COD10K for camouflaged object detection (COD), SBU for shadow detection (SD), Five-Dataset for colon polyp segmentation (CPS), COVID-19 CT for lung infection segmentation, BUSI for breast lesion segmentation (BLS), ISIC2018 for skin lesion segmentation (SLS), and EORSSD for optical remote sensing image salient object detection (ORSI SOD). Detailed dataset configurations are provided in the supplementary material.

Evaluation Metrics. We adopt standard metrics for each domain: weighted F-measure (F^ω_β) and S-measure (S_α) for SOD, COD, and ORSI-SOD tasks; Balance Error Rate (BER) and Mean Absolute Error (MAE) for the Shadow Detection task; and mean Dice (mDice) and mean IoU (mIoU) for all medical segmentation tasks (Polyp, COVID-19, Breast, and Skin). Notably, we use the same evaluation file as Spider [Zhao *et al.*, 2024].

Implementation Details. Our proposed method is implemented in PyTorch. During both training and testing, all input images and their corresponding ground truths (GT) are resized to 384×384. We employ standard data augmentation techniques, including random horizontal flipping and random cropping. The model utilizes the SAM2 [Ravi *et al.*, 2025]

Methods	Backbone	**Salient** DUTS		**Camouflaged** COD10K		**Skin** ISIC2018		**Shadow** SBU		**Polyp** 5 datasets		**COVID-19** COVID-19 CT		**Breast** BUSI		**ORSI** EORSSD	
		$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑	mIoU ↑	$\mathcal{M} \downarrow$	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$
Specialized Models																	
MENet$_{23}$	ResNet-50	0.8698	0.9028	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Samba$_{25}$	Custom	0.9016	0.9319	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RISNet$_{24}$	PVTv2-B4	-	-	0.7990	0.8730	-	-	-	-	-	-	-	-	-	-	-	-
Camodiffusion$_{25}$	ViT-L	-	-	0.8280	0.8806	-	-	-	-	-	-	-	-	-	-	-	-
C-LPMoE$_{25}$	BEiT-L	-	-	-	-	0.8886	0.8658	-	-	-	-	-	-	-	-	-	-
CCViM$_{25}$	Custom	-	-	-	-	0.8845	0.8724	-	-	-	-	-	-	-	-	-	-
SILT$_{23}$	PVTv2-B5	-	-	-	-	-	-	0.8352	0.0493	-	-	-	-	-	-	-	-
SARA$_{23}$	ConvNeXt-L	-	-	-	-	-	-	0.9078	0.0333	-	-	-	-	-	-	-	-
LSSNet$_{24}$	PVTv2-B2	-	-	-	-	-	-	-	-	0.8329	0.8686	-	-	-	-	-	-
WeakPolyp$_{23}$	PVTv2-B2	-	-	-	-	-	-	-	-	0.7490	0.8066	-	-	-	-	-	-
CAD-Unet$_{25}$	Custom	-	-	-	-	-	-	-	-	-	-	0.7024	0.8145	-	-	-	-
DECOR-Net$_{23}$	Custom	-	-	-	-	-	-	-	-	-	-	0.4025	0.6949	-	-	-	-
AAU-Net$_{22}$	Custom	-	-	-	-	-	-	-	-	-	-	-	-	0.4745	0.6515	-	-
CMUNet$_{23}$	Custom	-	-	-	-	-	-	-	-	-	-	-	-	0.5452	0.8302	-	-
RAMENet$_{25}$	SMT-T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0050	0.9419
GeleNet$_{23}$	Swin-B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0064	0.9376
Unified Models																	
VSCode$_{24}$	Swin-B	0.8912	0.9258	0.7836	0.8688	-	-	-	-	-	-	-	-	-	-	-	-
EVP$_{23}$	MiT-B4	0.8431	0.9007	0.7262	0.8346	-	-	0.8645	0.0489	-	-	-	-	-	-	-	-
SegGPT$_{23}$	ViT-L	0.3874	0.6283	0.4041	0.6529	0.4803	0.4402	0.4305	0.2041	0.5677	0.7074	0.1309	0.5533	0.3455	0.6033	0.0242	0.7904
Spider$_{24}$	ConvNeXt-L	0.8821	0.9158	0.7893	0.8674	0.8943	0.8735	0.8949	0.0265	0.8243	0.8664	0.6956	0.8127	0.8376	0.8655	-	-
H^{2}Net	ConvNeXt-L	0.9011	0.9327	0.7926	0.8792	0.8972	0.8761	0.9171	0.0199	0.8332	0.8701	0.6931	0.8079	0.8395	0.8673	0.0047	0.9425
H^{2}Net	ViT-L	0.9079	0.9370	0.8234	0.8978	0.8976	0.8750	0.9126	0.0207	0.8395	0.8769	0.6866	0.8084	0.8412	0.8682	0.0045	0.9402
H^{2}Net	Hiera-L	0.9043	0.9391	0.7949	0.8913	0.9011	0.8817	0.9248	0.0177	0.8426	0.8789	0.7119	0.8191	0.8480	0.8700	0.0038	0.9429

Table 1: Quantitative comparison with state-of-the-art specialized and unified segmentation methods across 8 tasks. The best results are highlighted in red.

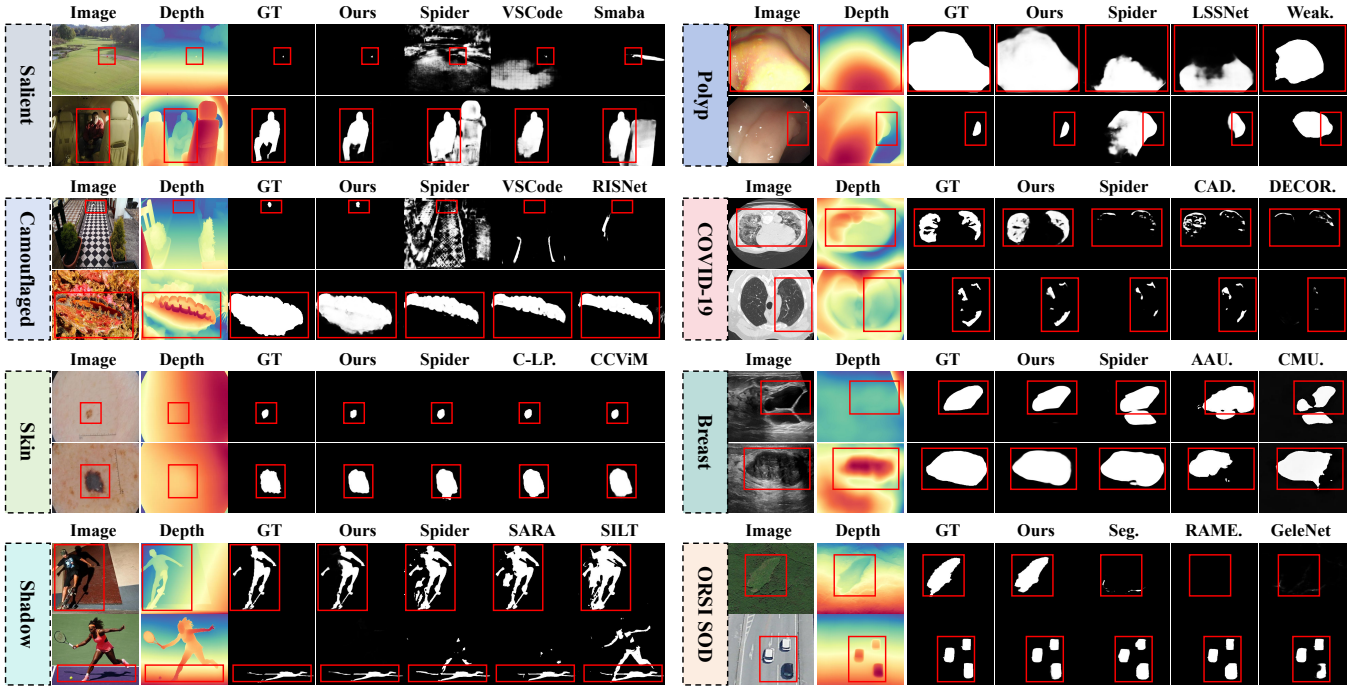

Figure 3: Qualitative comparison with state-of-the-art methods across eight heterogeneous tasks. We compare H^{2}Net against representative task-specific and unified methods (*e.g.*, Spider, SegGPT). Visual results demonstrate that our method consistently produces more accurate segmentation masks with sharper boundaries and fewer false positives compared to competitors, validating its superior generalization capability.

backbone Hiera-L. We use the Adam optimizer with an initial learning rate of 5e-5, which is decayed by a factor of 0.1 every 200 epochs following a step decay schedule. The model is trained for 300 epochs on four NVIDIA RTX 4090 GPUs with a total batch size of 32 (8 per GPU).

4.2 Comparison with SOTA Methods

Quantitative Comparison. We evaluate H^{2}Net against a wide range of state-of-the-art methods, including both domain-specific specialized models and recent unified segmentation frameworks, across eight heterogeneous downstream tasks. These methods include MENet [Wang *et al.*, 2023d], Samba [He *et al.*, 2025], RISNet [Wang *et al.*, 2024a], Camodiffusion [Sun *et al.*, 2025b], C-LPMoE [Sun *et al.*, 2025c], CCViM [Zhu *et al.*, 2025], SILT [Yang *et al.*, 2023], SARA [Sun *et al.*, 2023], LSSNet [Wang *et al.*, 2024b], WeakPolyp [Wei *et al.*, 2023], CAD-Unet [Dang *et al.*, 2025], DECOR-Net [Hu *et al.*, 2023], AAU-Net [Chen *et al.*, 2023], CMUNet [Tang *et al.*, 2023], RAMENet [Han *et al.*, 2025a], GeleNet [Li *et al.*, 2023], VSCode [Luo *et al.*, 2024], EVP [Liu *et al.*, 2023], SegGPT [Wang *et al.*, 2023c], Spider [Zhao *et al.*, 2024]. The quantitative results in Table 1 demonstrate that H^{2}Net achieves consistently strong performance across all benchmarks. Moreover, replacing the backbone from Hiera-L to ConvNeXt-L [Woo *et al.*, 2023] and ViT-L [Oquab *et al.*, 2024] consistently yields strong performance, demonstrating the robustness and generality of H^{2}Net across different backbone architectures.

Compared with existing unified approaches, H^{2}Net exhibits clear advantages in both context-dependent and structure-sensitive tasks. On Salient Object Detection (SOD), H^{2}Net achieves an S_α of 0.9391 on DUTS, surpassing strong unified baselines such as VSCode and Spider. On the challenging Camouflaged Object Detection (COD) benchmark, where subtle structural cues are critical, H^{2}Net attains an F_β^ω of 0.7949 and an S_α of 0.8913 on COD10K, outperforming Spider and demonstrating improved robustness under low-contrast conditions. Moreover, H^{2}Net shows pronounced strengths in other tasks. In Shadow Detection, it achieves the highest mIoU of 0.9248 while reducing MAE to 0.0177, indicating superior structural consistency. In the medical domain, H^{2}Net maintains competitive performance across diverse datasets, including skin, polyp, COVID-19, and breast lesion segmentation, effectively matching or exceeding specialized models without sacrificing architectural generality. H^{2}Net on the ORSI-SOD task, H^{2}Net achieves an S_α of 0.9429 with a remarkably low MAE of 0.0038, outperforming recent specialized approaches.

In summary, these results demonstrate that H^{2}Net successfully bridges the performance gap between unified and task-specific models, delivering strong generalization while maintaining high task-level precision.

Qualitative Comparison. As illustrated in Figure 3, H^{2}Net exhibits superior robustness across heterogeneous scenarios compared to state-of-the-art methods. For instance, in the Camouflaged task (Row 2), where targets share high textural similarity with the background, unified models like Spider often fail to detect the target due to insufficient modeling of homogeneous features. In contrast, H^{2}Net successfully disentangles subtle semantic cues to generate complete predictions. Similarly, in the ORSI remote sensing task (Last Row), while baseline methods such as SegGPT suffer from either missed detections of tiny vehicles or erroneously regarding background artifacts as vehicle parts, our model accurately captures these fine-grained targets without such confusion. In the Shadow detection task (Row 1), competitors like Spider and SARA incorrectly classify the person as the target, resulting in false positives. Conversely, our method yields sharper boundaries and complete target segmentation, verifying its capability to handle heterogeneous features.

In summary, these visual comparisons show that the proposed differential modeling of homogeneous and heterogeneous features effectively reduces feature interference and semantic confusion, leading to accurate segmentation across diverse domains. Moreover, despite the inconsistency of depth priors across tasks, the results indicate that CSEM can effectively align cross-modal features and robustly model homogeneous structural cues. Additional qualitative comparisons under more challenging scenarios are provided in the supplementary material to further validate the robustness of our approach.

4.3 Ablation Studies

To validate the effectiveness and rationale of the key components in our framework, we conducted ablation experiments across datasets from eight different tasks. These experiments were designed to more clearly assess the impact of each component on the performance of various tasks. The evaluation metrics for each task are consistent with those presented in Table 1. The experimental setup follows the configuration of H^{2}Net.

Effectiveness of Key Components. To assess the contribution of the key components in our framework, we conducted ablation experiments with the following configurations: ID 1 uses a frozen encoder, a single adapter, and a decoder as the baseline model; ID 2 adds CSEM to the baseline; ID 3 incorporates TAP into the baseline; ID 4 trains H^{2}Net independently on each task; ID 5 uses the full H^{2}Net model, combining both CSEM and TAP.

Specifically, the results in the Table 2 demonstrate the effectiveness of each component. Adding CSEM to the baseline (ID 2) increases the F_β^ω on DUTS from 0.5835 to 0.8649, indicating that auxiliary depth priors effectively compensate for missing geometric cues and enhance homogeneous representations such as object boundaries and spatial layouts. Incorporating TAP alone (ID 3) further improves F_β^ω to 0.8869, showing that task-specific adapters help disentangle heterogeneous semantics and alleviate negative interference among tasks. This improvement highlights the importance of isolating task-dependent information during representation learning in a unified framework. Moreover, compared with task-wise training (ID 4), the jointly trained H^{2}Net (ID 5) achieves comparable or superior performance, suggesting that the proposed homogeneous–heterogeneous decoupling enables effective unified training without sacrificing task-level accuracy. Together, these results confirm that jointly enhancing shared structures while decoupling task-specific semantics is crucial for robust unified segmentation.

ID	method	**Salient** DUTS		**Camouflaged** COD10K		**Skin** ISIC2018		**Shadow** SBU		**Polyp** 5 datasets		**COVID-19** COVID-19 CT		**Breast** BUSI		**ORSI** EORSSD	
		$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑	mIoU ↑	$\mathcal{M} \downarrow$	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$
Ablation experiments of the H^{2}Net																	
1	Baseline	0.5835	0.7346	0.6348	0.6835	0.7935	0.7947	0.7321	0.0791	0.4279	0.6395	0.5042	0.6961	0.7670	0.7890	0.0138	0.8676
2	w/ CSEM	0.8649	0.8852	0.7434	0.8628	0.8619	0.8489	0.8821	0.0327	0.7963	0.8366	0.6658	0.7724	0.8117	0.8362	0.0072	0.9120
3	w/ TAP	0.8869	0.9126	0.7721	0.8746	0.8810	0.8648	0.8992	0.0243	0.8227	0.8579	0.6901	0.7899	0.8231	0.8540	0.0057	0.9249
4	Specialized	0.9037	0.9356	0.8028	0.9017	0.8954	0.8780	0.9147	0.0206	0.8450	0.8809	0.7024	0.8112	0.8498	0.8715	0.0043	0.9350
5	H^{2}Net	0.9043	0.9391	0.7949	0.8913	0.9011	0.8817	0.9248	0.0177	0.8426	0.8789	0.7119	0.8191	0.8480	0.8700	0.0038	0.9429
Ablation experiments of the CSEM																	
a	Baseline	0.8869	0.9126	0.7721	0.8746	0.8812	0.8648	0.8992	0.0243	0.8225	0.8577	0.6901	0.7899	0.8232	0.8540	0.0057	0.9248
b	w/ MHSA	0.8939	0.9232	0.7812	0.8813	0.8891	0.8716	0.9094	0.0217	0.8307	0.8663	0.6988	0.8016	0.8331	0.8604	0.0049	0.9321
c	w/ CBAM	0.8904	0.9179	0.7767	0.8779	0.8850	0.8682	0.9043	0.0230	0.8260	0.8620	0.6945	0.7957	0.8288	0.8572	0.0053	0.9284
d	w/ SCM	0.8973	0.9285	0.7858	0.8846	0.8937	0.8749	0.9146	0.0203	0.8342	0.8705	0.7032	0.8074	0.8380	0.8636	0.0046	0.9357
e	w/ FCM	0.9008	0.9338	0.7903	0.8880	0.8971	0.8783	0.9197	0.0190	0.8386	0.8742	0.7075	0.8133	0.8437	0.8668	0.0042	0.9397
f	H^{2}Net	0.9043	0.9391	0.7949	0.8913	0.9011	0.8817	0.9248	0.0177	0.8426	0.8789	0.7119	0.8191	0.8480	0.8700	0.0038	0.9429
Ablation experiments of the TAP																	
A	Baseline	0.8649	0.8852	0.7434	0.8628	0.8619	0.8489	0.8821	0.0327	0.7963	0.8366	0.6658	0.7724	0.8117	0.8362	0.0072	0.9120
B	w/ MoE	0.8739	0.8983	0.7783	0.8738	0.8810	0.8633	0.8931	0.0233	0.8079	0.8586	0.6815	0.8089	0.8235	0.8552	0.0057	0.9254
C	w/ SMoE	0.8755	0.9002	0.7744	0.8694	0.8843	0.8695	0.8969	0.0228	0.8091	0.8616	0.6971	0.8152	0.8257	0.8569	0.0053	0.9278
D	H^{2}Net	0.9043	0.9391	0.7949	0.8913	0.9011	0.8817	0.9248	0.0177	0.8426	0.8789	0.7119	0.8191	0.8480	0.8700	0.0038	0.9429

Table 2: Ablation experiments of H^{2}Net.

The visualization results in Figure 4 further provide intuitive evidence for the effectiveness of CSEM and TAP. Observing the F_{CSEM} feature maps reveals that, regardless of the specific task target, the module consistently delineates sharp homogeneous geometric contours within the scene; for instance, in both Shadow Detection and Salient Object Detection tasks, F_{CSEM} identically highlights the structural edges of the person and the shadow, indicating that it extracts task-agnostic universal structural priors. In sharp contrast, the F_{TAP} feature maps display highly differentiated heterogeneous semantic activation patterns: in Skin Lesion Segmentation, the features precisely focus on the lesion area, whereas in Camouflaged Object Detection, the focus shifts entirely to the concealed object. This distinct visual difference confirms that TAP successfully isolates task-specific semantic information. Consequently, the final predictions (Ours) benefit from the fusion of this universal structural skeleton with differentiated feature processing, resulting in segmentation masks that possess both fine-grained boundaries and accurate semantic even in challenging scenarios.

Effectiveness of CSEM. To validate the design rationale of CSEM, we conducted a series of experiments, including: (a) removing the CSEM as the baseline; (b) removing the depth branch and replacing CSEM with a standard multi-head self-attention mechanism; (c) removing the depth branch and replacing CSEM with a convolutional block attention module (CBAM); (d) performing cross-modal attention mechanism only in the spatial domain; (e) performing cross-modal attention mechanism only in the frequency domain; and (f) H^{2}Net, which integrates the frequency-spatial cross-modal attention mechanism.

Specifically, as presented in the Table 2, the quantitative results support our design from three perspectives. First, cross-modal interaction outperforms single-modal self-enhancement. While standard attention modules (b and c) bring marginal gains over the baseline, they consistently lag behind depth-guided method. For instance, on DUTS, H^{2}Net surpasses MHSA by a clear margin (0.9391 vs. 0.9232 in S_α), demonstrating that auxiliary depth priors are essential for modeling robust homogeneous structures beyond RGB information. Second, frequency domain analysis proves critical. The frequency-only method (e) generally outperforms the spatial-only method (d), particularly on structure-sensitive tasks like COD10K (0.7903 vs. 0.7858 in F_β^ω), indicating that frequency components effectively encapsulate low-level structural details (*e.g.*, edges and textures). Finally, spatial-frequency synergy yields superior performance. The full H^{2}Net (f) achieves the best results across all metrics (*e.g.*, boosting mDice to 0.9011 on ISIC2018), confirming that integrating global spatial alignment with frequency-based detail enhancement establishes the most solid foundation for homogeneous features.

Effectiveness of TAP. To validate the design rationale of the Task Adapter Pool (TAP), we conducted a series of experiments, including: (A) removing the TAP module as the baseline; (B) replacing the proposed TAP with a standard Mixture of Experts (MoE); (C) replacing the proposed TAP with a Sparse Mixture of Experts (SMoE [Puigcerver *et al.*, 2024]); and (D) H^{2}Net.

The results are presented in the Table 2. The baseline model (Row A), lacking task adaptation, yields an F_β^ω of 0.8649. This suggests that a strictly shared parameter space faces challenges in effectively modeling the highly heterogeneous features across different tasks. While dynamic routing structures like MoE (Row B) and SMoE (Row C) provide flexibility and improve scores to 0.8739 and 0.8755, their performance gain is relatively limited. This is attributed to the implicit nature of probability-based routing, where differ-

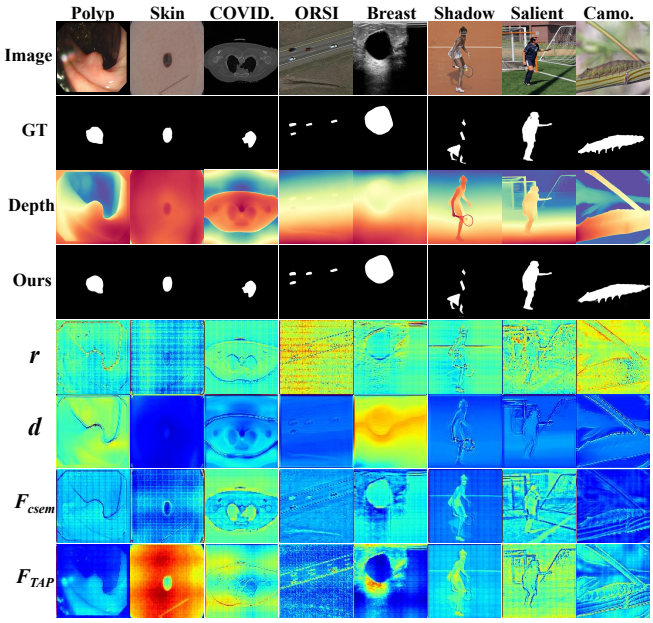

Figure 4: Qualitative results and intermediate feature visualizations across eight heterogeneous segmentation tasks. From top to bottom: input images, ground truths (GT), depth maps, and segmentation predictions of H^{2}Net. The lower rows show intermediate features, where F_{CSEM} highlights structural boundaries and F_{TAP} emphasizes task-specific semantic regions.

ent tasks unintentionally share the same experts. Such overlap introduces interference between heterogeneous features, limiting the degree of expert specialization. In contrast, our H^{2}Net demonstrates a distinct advantage with a superior F_β^ω of 0.9043. By adopting an explicit dispatching strategy, our TAP allocates dedicated adapters to distinct tasks. This design mitigates the risk of feature interference, validating the effectiveness and rationality of TAP.

5 Conclusion

In this paper, we propose H^{2}Net, a unified segmentation framework that introduces a new perspective for unified segmentation from the viewpoint of decoupled representation learning. By separating cross-task homogeneous representations and task-specific heterogeneous semantics during the encoding process, H^{2}Net effectively reduces cross-task interference while enabling efficient knowledge sharing. This design is realized through two complementary components: CSEM, which enhances shared structural representations by leveraging cross-modal structural priors, and TAP, which isolates task-related semantics via task-aware adapters within a shared backbone. Extensive experiments on eight heterogeneous benchmarks validate the effectiveness of this idea, demonstrating that unified training can promote positive cross-task synergy without sacrificing task-level performance. Moreover, H^{2}Net generalizes well across different backbones, indicating strong scalability.

Acknowledgments

This work was supported in part by the Jiangsu Provincial Science and Technology Major Project under Grant BG2024042, by the National Natural Science Foundation of China under Grant 62472067, and by the Joint Funds of Liaoning Science and Technology Program under Grant 2024JH2/102600091.

References

[Changguang *et al.*, 2025] WU Changguang, Jiangxin Dong, Chengjian Li, and Jinhui Tang. Plenodium: Underwater 3d scene reconstruction with plenoptic medium representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Chen *et al.*, 2023] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: An adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging*, 42(5):1289–1300, 2023.

[Dang *et al.*, 2025] Yijie Dang, Weijun Ma, Xiaohu Luo, and Huaizhu Wang. Cad-unet: A capsule network-enhanced unet architecture for accurate segmentation of COVID-19 lung infections from CT images. *Medical Image Anal.*, 103:103583, 2025.

[De Boer *et al.*, 2005] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005.

[Han *et al.*, 2025a] Jinyu Han, Fuming Sun, Yaoyao Hou, Jing Sun, and Haojie Li. Exploring a lightweight and efficient network for salient object detection in orsi. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

[Han *et al.*, 2025b] Jinyu Han, Jing Sun, Fasheng Wang, Fuming Sun, and Haojie Li. Orsidiff: Diffusion model for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

[He *et al.*, 2023] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023.

[He *et al.*, 2025] Jiahao He, Keren Fu, Xiaohong Liu, and Qijun Zhao. Samba: A unified mamba-based framework for general salient object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 25314–25324, June 2025.

[Hu *et al.*, 2023] Jiesi Hu, Yanwu Yang, Xutao Guo, Bo Peng, Hua Huang, and Ting Ma. Decor-net: A covid-19 lung infection segmentation network improved by emphasizing low-level features and decorrelating features. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023.

[Li *et al.*, 2023] Gongyang Li, Zhen Bai, Zhi Liu, Xinpeng Zhang, and Haibin Ling. Salient object detection in optical remote sensing images driven by transformer. *IEEE Transactions on Image Processing*, 32:5257–5269, 2023.

[Liu *et al.*, 2023] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023.

[Luo *et al.*, 2024] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *CVPR*, pages 17169–17180, 2024.

[Oquab *et al.*, 2024] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

[Puigcerver *et al.*, 2024] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *ICLR*. OpenReview.net, 2024.

[Ravi *et al.*, 2025] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[Shelhamer *et al.*, 2017] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[Sun *et al.*, 2023] Jiayu Sun, Ke Xu, Youwei Pang, Lihe Zhang, Huchuan Lu, Gerhard Hancke, and Rynson Lau. Adaptive illumination mapping for shadow detection in raw images. In *IEEE/CVF International Conference on Computer Vision*, pages 12709–12718, 2023.

[Sun *et al.*, 2025a] Fuming Sun, Jinyu Han, Weiyi Wu, Jing Sun, and Mengyin Wang. A unet-like transformer network for camouflaged object detection. *IEEE Transactions on Multimedia*, pages 1–15, 2025.

[Sun *et al.*, 2025b] Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2833–2848, 2025.

[Sun *et al.*, 2025c] Yanguang Sun, Jiawei Lian, Jian Yang, and Lei Luo. Controllable-lpmoe: Adapting to challenging object segmentation via dynamic local priors from mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22327–22337, 2025.

[Tang *et al.*, 2023] Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: A strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023.

[Wang *et al.*, 2023a] Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain, editors, *ACM MM*, pages 3297–3306. ACM, 2023.

[Wang *et al.*, 2023b] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6830–6839. IEEE, 2023.

[Wang *et al.*, 2023c] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1130–1140. IEEE, 2023.

[Wang *et al.*, 2023d] Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10031–10040, June 2023.

[Wang *et al.*, 2024a] Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17201–17211, 2024.

[Wang *et al.*, 2024b] Wei Wang, Huiying Sun, and Xin Wang. Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part VII*, volume 15007 of *Lecture Notes in Computer Science*, pages 446–456. Springer, 2024.

[Wei *et al.*, 2023] Jun Wei, Yiwen Hu, Shuguang Cui, S Kevin Zhou, and Zhen Li. Weakpolyp: You only look bounding box for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 757–766. Springer, 2023.

[Woo *et al.*, 2023] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142, 2023.

[Yang *et al.*, 2023] Han Yang, Tianyu Wang, Xiaowei Hu, and Chi-Wing Fu. SILT: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *IEEE/CVF International Conference on Computer Vision, Paris, France, October 1-6*, pages 12641–12652, 2023.

[Yang *et al.*, 2024] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *NeurIPS*, 2024.

[Zhao *et al.*, 2024] 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 *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[Zhu *et al.*, 2025] Yun Zhu, Dong Zhang, Yi Lin, Yifei Feng, and Jinhui Tang. Merging context clustering with visual state space models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(5):2131–2142, 2025.