

OBBSeg: Irregular Lesion Segmentation under Oriented Bounding Box Annotations

Jun Wei¹, Xinchang Liu¹, Yu Liu¹, Chuhua Yang¹, Shuhui Wang², and Hui Huang^{1*}

¹ Guangdong Provincial Key Laboratory of Visual Media and Multidimensional Intelligence, CSSE, Shenzhen University, China

² Institute of Computing Technology, Chinese Academy of Sciences, China
weijun@szu.edu.cn, hhzhiyan@gmail.com

Abstract. Pixel-level annotation remains a major bottleneck in medical image segmentation, making weak supervision an attractive yet under-constrained alternative. We propose **OBBSeg**, an intermediate supervision paradigm guided by **Oriented Bounding Boxes (OBBs)** that bridges the gap between full and weak supervision. By jointly encoding spatial extent and orientation, OBBs provide compact geometric supervision that better aligns with elongated or anisotropic lesions, reducing the ambiguity of coarse box annotations. To mitigate the inherent rectangular bias of OBBs, we introduce a **Mask-to-OBB loss**, a differentiable formulation that enforces geometric consistency between predicted masks and OBB regions. Furthermore, we incorporate **prompt-driven semantic guidance** through two complementary modules—**PAFE** and **DBFE**—which enhance foreground representation and suppress background interference. Extensive experiments on **13 datasets** across **5 imaging modalities** show that OBBSeg not only outperforms existing weakly supervised methods but also achieves performance comparable to fully supervised approaches, demonstrating its potential for **efficient and scalable medical image segmentation**. The code is available at <https://github.com/StarLxc3/OBBSeg>.

Keywords: Medical image segmentation, Weak supervision, Oriented Bounding box

1 Introduction

Accurate segmentation of medical images plays a critical role in computer-aided diagnosis and treatment planning. By delineating anatomical structures and pathological regions, segmentation models enable reliable disease screening, quantitative analysis, and therapy assessment.

Despite recent advances, medical image segmentation still faces a fundamental trade-off between **annotation accuracy and efficiency**. Fully supervised models, such as U-Net [34] and its variants [47], achieve strong performance

* Corresponding author

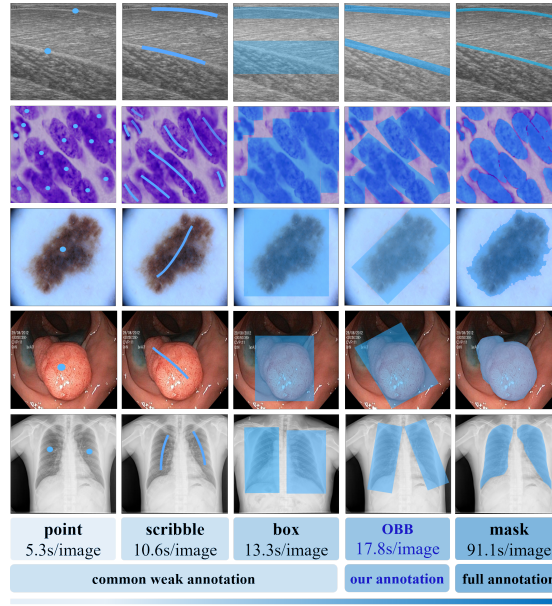


Fig. 1: Visualization and annotation cost (seconds per image) comparison among different annotation types (*e.g.*, point, scribble, box, OBB, and mask). OBB annotation achieves a favorable balance between labeling efficiency and target coverage, providing geometry-aware and orientation-sensitive supervision with minimal annotation cost.

with pixel-level annotations, but dense mask labeling is expensive and time-consuming, especially for lesions with irregular or ambiguous boundaries. To alleviate this burden, weakly supervised segmentation methods employ coarse annotations such as points [2], scribbles [24], or bounding boxes [28]. These annotations significantly reduce labeling effort but provide limited geometric constraints. In particular, conventional bounding boxes fail to encode orientation, making them inadequate for **elongated or anisotropic lesions**.

To bridge the gap between full and weak supervision, we propose an intermediate paradigm, termed **Oriented Bounding Box Guided Segmentation (OBBSeg)**, combining the efficiency of weak supervision with the geometric fidelity of full supervision. Fig. 1 illustrates the proposed Oriented Bounding Box (OBB) annotations, which provide a compact, orientation-aware annotation that better captures lesion geometry than conventional bounding boxes. By jointly encoding spatial extent and orientation, OBBs tightly fit elongated or anisotropic lesions (see row 1), providing stronger geometric constraints and reducing ambiguous regions (see rows 3-5). Moreover, they alleviate the overlap between closely packed lesions that commonly occurs with conventional box annotations (see row 2).

Furthermore, OBB annotation remains **lightweight**, requiring only four corner points per instance. To evaluate its annotation cost, we conducted a human

annotation study using our training set, where three independent annotators labeled 1,616 OBBs across 1,450 images. The average annotation time was **17.8 s per OBB**, only slightly longer than standard bounding boxes (13.3 s per image) and substantially shorter than dense mask annotation (91.1 s per image). These results show that OBB annotation provides richer geometric information with only a modest increase in annotation effort, making it an effective compromise between annotation efficiency and geometric precision.

Although OBB annotations capture target extent and orientation, their rectangular geometry may introduce shape bias during segmentation learning. To address this limitation, we propose the **Mask-to-OBB (M2O) loss**, a supervision strategy that bridges coarse OBB annotations and fine-grained segmentation masks. M2O projects predicted masks into the OBB space through rotation and projection operations, supervising only spatial extent and orientation while excluding explicit boundary information. This design effectively mitigates rectangular shape bias without iterative refinement. Implemented entirely with matrix operations, M2O is end-to-end **differentiable, plug-and-play transferable** to other segmentation tasks, and incurs **no additional inference cost**. As shown in Table 4, M2O consistently outperforms direct OBB supervision, demonstrating the effectiveness of the proposed supervision strategy.

Although OBB supervision reduces annotation cost and provides geometric priors, it remains limited in modeling complex lesion appearance and cross-domain variation. To address this, we propose a unified OBBSeg framework that integrates OBB-based geometric constraints with prompt-guided semantic learning. Different from existing prompt-based methods that introduce prompts only at the decoder stage, OBBSeg injects prompt information into the encoder to explicitly guide feature learning with semantic priors.

In this design, OBBs provide structural cues of spatial extent and orientation, while prompts progressively refine semantic representations and enhance boundary awareness during feature extraction. To further strengthen weakly supervised learning, we introduce two complementary modules: **Prompt-assisted Foreground Enhancer (PAFE)**, which highlights lesion regions using prompt-derived masks while suppressing background interference, and the **Differential-based Foreground Enhancer (DBFE)**, which improves feature discriminability by explicitly enhancing foreground-background contrast.

Overall, OBBSeg jointly leverages geometric constraints, prompt-driven semantics, and differential feature enhancement to improve segmentation accuracy and generalization under weak supervision. Our contributions are fourfold:

1. We propose **OBBSeg**, a new segmentation paradigm that leverages oriented bounding boxes to bridge the gap between weakly and fully supervised learning, providing geometry-aware supervision with low annotation cost.
2. We introduce the **M2O loss**, a differentiable and noise-free supervision formulation that removes the rectangular bias of box annotations and enables geometry-consistent learning.

3. We integrate **prompt-guided learning** with OBB supervision and design two complementary modules, **PAFE** and **DBFE**, to enhance foreground representation and improve segmentation robustness.
4. Extensive experiments on **13 medical datasets** across **5 imaging modalities** demonstrate that OBBSeg outperforms existing weakly supervised methods and achieves performance approaching fully supervised models.

2 Related Work

Medical Image Segmentation. Medical image segmentation aims to delineate anatomical and pathological structures from modalities such as X-ray [19], MRI [27], and ultrasound [16]. Fully supervised models, including U-Net [34], U-Net++ [47], and PraNet [14], have achieved strong performance across various medical imaging tasks. However, these methods rely heavily on pixel-level annotations, which are expensive and time-consuming to obtain. To reduce annotation costs, weakly supervised segmentation methods have been increasingly explored. In this work, we adopt oriented bounding boxes as supervision, which provide geometry-aware annotations that better capture object extent and orientation while maintaining low labeling cost.

Weakly Supervised Learning. Weakly supervised segmentation aims to reduce annotation effort by using coarse labels such as points [2], boxes [28], scribbles [24], and circles [23]. [28] proposed a box-supervised approach with Background-Aware Pooling and Noise-Aware Loss, achieving competitive results on the PASCAL VOC 2012 benchmark [13]. WeakPolyp [41] introduced a projection-backprojection strategy together with Scale Consistency Loss to mitigate box-shape bias, significantly improving polyp segmentation performance. Zhang *et al.* [45] proposed a Class-Driven Scribble Promotion network that integrates scribble annotations with image-level pseudo-labels and uncertainty-aware optimization. Despite these advances, weak supervision often struggles to capture accurate geometric structures due to the coarse nature of the annotations. In contrast, our method introduces the Mask-to-OBB loss to provide geometry-consistent supervision while eliminating the rectangular bias of box annotations.

Prompt-based Image Segmentation. Prompt-based segmentation leverages auxiliary guidance, such as points, boxes, or textual cues, to improve model adaptability and generalization. Representative models such as SAM [22] and SAM2 [33] encode prompts into geometric embeddings to guide segmentation. However, most existing approaches introduce prompts only at the decoder stage, limiting their influence on feature representation. In contrast, our method incorporates prompts into the encoder to enhance foreground representations and suppress background noise, while further utilizing them for supervision, enabling end-to-end prompt-guided segmentation.

3 Proposed Methodology

Fig. 2(a) presents the overall architecture of the proposed OBBSeg framework, which leverages prompts as auxiliary cues and oriented bounding boxes as weak

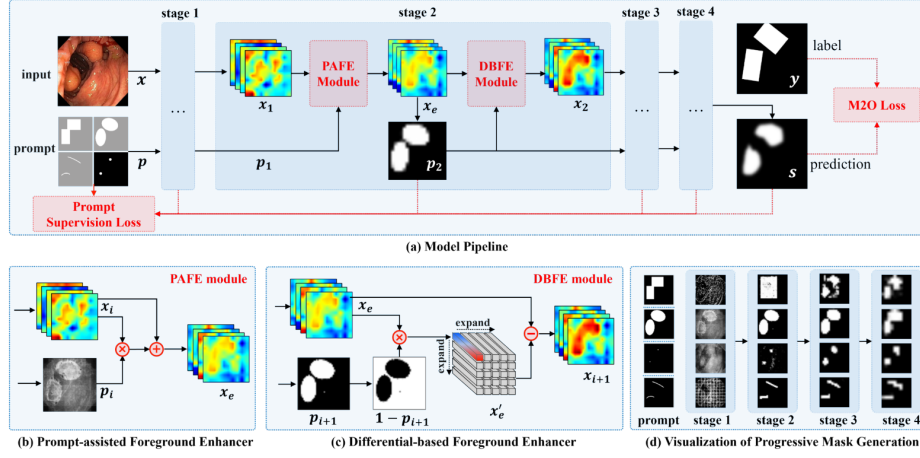


Fig. 2: Overview of the proposed OBBSeg framework. (a) Overall architecture. (b) The PAFE module enhances lesion-focused features via prompt guidance. (c) The DBFE module improves foreground-background discrimination. (d) Progressive mask generation from prompt inputs.

supervision. Built upon a ViT [12] backbone, OBBSeg is organized into four stages. In each stage, the PAFE module exploits prompt cues to suppress background noise, while the DBFE module enhances foreground-background contrast to highlight discriminative regions. At the supervision level, OBBSeg adopts a dual-strategy design. Prompt supervision provides hierarchical guidance to refine coarse prompts, while the proposed M2O loss aligns soft mask predictions with OBB annotations to alleviate box-induced shape bias. Through progressive feature refinement and noise-robust supervision, OBBSeg produces accurate segmentation masks with minimal annotation cost.

3.1 Prompt-assisted Foreground Enhancer

Fig. 2(b) illustrates the Prompt-assisted Foreground Enhancer (PAFE), designed to suppress background noise under weak supervision. Unlike conventional attention mechanisms that lack explicit guidance, PAFE utilizes prompt cues to guide the network toward lesion-relevant regions, improving semantic consistency and interpretability. Given an input image x , the backbone extracts a hierarchy of feature maps $\{x_i \mid i \in (0, 1, 2, 3)\}$. For each stage i , a prompt-derived binary mask p_i is generated and resized to match the spatial resolution of x_i . Foreground-aware filtering is achieved by element-wise multiplication ($\odot$) between p_i and x_i . Since prompts (e.g., points or scribbles) may only partially cover the target region, a residual connection adds the original feature map back to preserve information, producing the enhanced feature representation x_e :

$$x_e = x_i \odot p_i + x_i. \quad (1)$$

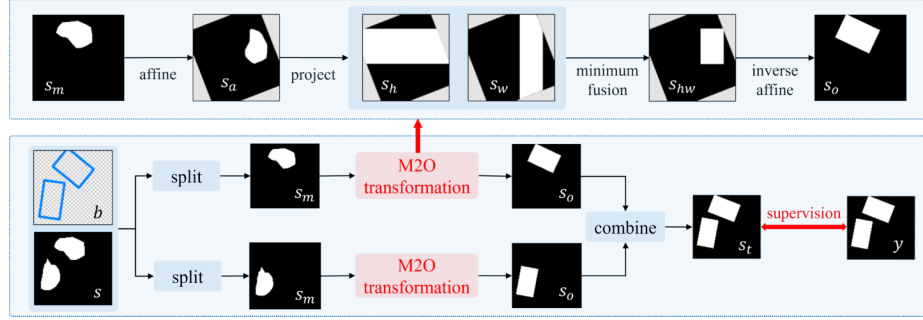


Fig. 3: Overview of the proposed Mask-to-OBB (M2O) loss. The predicted mask s is projected into the OBB space to preserve spatial extent and orientation while suppressing fine-grained shape details. The projected representation is aligned with the OBB label y , enabling geometry-consistent learning from OBB annotations.

To obtain the prompt mask p_i , we directly leverage user-provided prompts (e.g., points or scribbles) as partial supervision signals. Supervision is enforced only on confidently labeled pixels within the prompt regions, providing reliable and noise-free guidance. This partial supervision encourages the model to progressively learn object-aware masks aligned with the target’s location and shape, as illustrated in Fig. 2(d).

3.2 Differential-based Foreground Enhancer

Medical image segmentation often suffers from high visual similarity between foreground and background, which hinders discriminative feature learning, especially under limited training data. To address this issue, we propose the Differential-based Foreground Enhancer (DBFE), which explicitly amplifies foreground-background contrast to facilitate feature discrimination. The overall architecture is shown in Fig. 2(c).

Given the prompt-derived binary mask p_{i+1} highlighting the foreground region, its complement $(1 - p_{i+1})$ represents the background. We first extract background cues by applying the background mask to the input feature x_e via element-wise multiplication followed by average pooling, producing a background representation x'_e . To enhance feature contrast, a differential factor is computed by subtracting x'_e from x_e and applying the foreground mask p_{i+1} through element-wise multiplication ($\odot$), which emphasizes discriminative foreground information. Following a residual design, this differential factor is added back to the original feature to preserve contextual information, yielding the enhanced feature x_{i+1} .

$$\begin{aligned} x'_e &= \text{pooling}((1 - p_{i+1}) \odot x_e) \\ x_{i+1} &= (x_e - x'_e) \odot p_{i+1} + x_e \end{aligned} \quad (2)$$

3.3 Mask-to-OBB Loss

Direct supervision with OBBs introduces rectangular shape bias, misleading the model toward box-like predictions. To address this, we propose Mask-to-OBB (M2O) loss, which removes shape bias by aligning predictions with OBBs in a geometry-aware way. Specifically, we extract instance masks, rectify them via affine transformation, and eliminate shape details through axis-wise projection and minimum fusion. The resulting masks retain spatial extent without contour bias, and are inversely transformed back for pixel-wise loss computation. This enables effective learning from coarse OBB annotations without shape-induced errors.

Split. Given the predicted mask s and a set of OBB masks $\{b_1, b_2, \dots, b_M\}$, we first decompose s into instance-level sub-masks, each corresponding to an individual object. Specifically, for each OBB mask b_m , the sub-mask is extracted by element-wise multiplication $s_m = s \odot b_m$.

Affine Transformation. Given an OBB mask b_m , we extract its rotation angle θ_m and construct the affine transformation matrix R_m . This matrix rotates the sub-mask s_m by θ_m clockwise, resulting in the rectified prediction s_a which is aligned to axis orientation.

$$s_a = \text{Warp}(s_m, R_m) \quad (3)$$

Projection. We perform max-projection on s_a along the horizontal and vertical axes, yielding the maximum probability vectors: s_h and s_v . This step completely eliminates the shape and contour information of s_a , retaining only its spatial location and dimension information.

$$\begin{aligned} s_h &= \max(s_a, \text{axis} = 1) \in [0, 1]^{1 \times W} \\ s_w &= \max(s_a, \text{axis} = 0) \in [0, 1]^{H \times 1} \end{aligned} \quad (4)$$

Minimum Fusion. To reconstruct the spatial mask, we back-project the horizontal and vertical projection vectors by replicating s_h across width and s_w across height, yielding two 2D masks $s_h \in [0, 1]^{H \times W}$ and $s_w \in [0, 1]^{H \times W}$. These masks are then fused via element-wise minimum operation to produce the final aligned prediction s_{hw} .

$$\begin{aligned} s_h &= \text{repeat}(s_h, H, \text{axis} = 1) \in [0, 1]^{H \times W} \\ s_w &= \text{repeat}(s_w, W, \text{axis} = 0) \in [0, 1]^{H \times W} \\ s_{hw} &= \min(s_h, s_w) \in [0, 1]^{H \times W} \end{aligned} \quad (5)$$

Inverse Affine Transformation. We apply the inverse transformation matrix R_m^{-1} to s_{hw} , rotating it counterclockwise by θ_m back to an axis-aligned state. This produces the transformed predicted mask s_o .

$$s_o = \text{Warp}(s_{hw}, R_m^{-1}) \quad (6)$$

Combination. For each image, we repeat the above transformation process for all OBBs, generating M rectified prediction masks $s_o^1, s_o^2, \dots, s_o^M$. We then

aggregate them via pixel-wise maximum fusion to obtain the final transformed mask s_t :

$$s_t = \max(s_o^1, s_o^2, \dots, s_o^M) \in [0, 1]^{H \times W} \quad (7)$$

Supervision. The transformed prediction s_t produced by the M2O transformation enables direct computation of Binary Cross-Entropy loss $\mathcal{L}_{\text{BCE}}$ and Dice loss $\mathcal{L}_{\text{Dice}}$ against the box-shaped OBB ground truth y . The combined loss $\mathcal{L}_{\text{M2O}}$ is defined as Eq. 8. Definitions of $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{Dice}}$ are provided in the supplementary material.

$$\mathcal{L}_{\text{M2O}} = \mathcal{L}_{\text{BCE}}(s_t, y) + \mathcal{L}_{\text{Dice}}(s_t, y) \quad (8)$$

3.4 Prompt Supervision Loss

To compensate for the limited supervision from OBB annotations and better exploit prompt masks, we introduce the Prompt Supervision Loss $\mathcal{L}_{\text{PS}}$. This loss serves two purposes: (1) guiding early-stage feature learning to refine intermediate predictions p_i , and (2) strengthening supervision to complement OBB-based learning. Specifically, $\mathcal{L}_{\text{PS}}$ enforces hierarchical consistency between the intermediate predictions and the final prediction with respect to the prompt mask:

$$\mathcal{L}_{\text{PS}} = \mathcal{L}(s, p) + \sum_{i=1}^4 \mathcal{L}(p_i, p). \quad (9)$$

To support different prompt types (*e.g.*, box, point, scribble, circle), we adopt adaptive supervision functions $\mathcal{L}$ tailored to each prompt. This results in $\mathcal{L}_{\text{PS}}^{\text{box}}$, $\mathcal{L}_{\text{PS}}^{\text{scribble}}$, $\mathcal{L}_{\text{PS}}^{\text{point}}$, and $\mathcal{L}_{\text{PS}}^{\text{circle}}$. Detailed implementations for each prompt type are provided in the supplementary material.

3.5 Total Loss

Following WeakPolyp, we adopt the Scale Consistency Loss $\mathcal{L}_{\text{SC}}$ to alleviate the supervision deficiency of OBB annotations by enforcing prediction consistency across different input scales. Combined with the proposed $\mathcal{L}_{\text{M2O}}$ and $\mathcal{L}_{\text{PS}}$, the overall loss function is defined as Eq. 10, where λ is a balancing hyper-parameter. Detailed analysis of λ is provided in the supplementary material.

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{M2O}} + \lambda \mathcal{L}_{\text{PS}} + \mathcal{L}_{\text{SC}} \quad (10)$$

4 Experimental Results

Datasets. We evaluate OBBSeg on **13 public datasets across five imaging modalities**: colonoscopy (Kvasir [20], ClinicDB [3], ColonDB [37], ETIS [36], Endo [40], SUN-SEG [21]), dermoscopy (ISIC-2017 [9], ISIC-2018 [10, 39]), MRI (NCI-ISBI [4]), CT (Synapse [17]), and ultrasound (TN3K [16], DDTI [30], BUSI [1]). Detailed descriptions are provided in the supplementary material.

Table 1: Comparison of OBBSeg with fully supervised models on polyp datasets. Dice (%) is reported. "-" indicates missing results. **Avg.** denotes the weighted average across datasets and serves as the primary indicator. Best results are in bold.

Methods	Year	Supervision	Five Standard Polyp Datasets (FSPD)					SUN-SEG		
			ClinicDB	ColonDB	ETIS	Kvasir	Endo	Avg.	Easy	Hard
U-Net [34]	2015 _{MICCAI}	full	82.3	51.2	39.8	81.8	71.0	56.1	-	-
U-Net++ [47]	2018 _{DLMIA}	full	79.4	48.3	40.1	82.1	70.7	54.6	-	-
PraNet [14]	2020 _{MICCAI}	full	89.9	70.9	62.8	89.8	87.1	74.0	68.9	66.0
SANet [42]	2021 _{MICCAI}	full	91.6	75.3	75.0	90.4	88.8	79.4	69.3	69.4
MEGANet [5]	2024 _{WACV}	full	93.8	79.3	73.9	91.3	89.9	81.4	61.2	69.3
CASCADE [31]	2023 _{WACV}	full	94.3	82.5	80.1	92.6	90.5	84.7	-	-
FLA-Net [25]	2023 _{MICCAI}	full	88.5	-	-	-	87.4	88.0	85.6	85.8
MS-TFAL [11]	2023 _{MICCAI}	full	91.1	-	-	-	89.1	90.1	85.9	86.2
LGRNet [44]	2024 _{MICCAI}	full	93.3	-	-	-	91.6	92.5	87.5	87.6
EMCAD [32]	2024 _{CVPR}	full	95.2	92.3	92.3	92.8	-	92.6	-	-
MedSAM2(Box) [26]	2025	full+prompt	94.6	91.7	93.2	94.4	93.7	92.8	-	-
WeakPolyp [41]	2023 _{MICCAI}	weak	85.2	74.5	71.1	86.1	84.8	76.7	79.2	80.7
WeakPolyp+M2O	-	weak	89.2	75.9	72.4	90.0	90.0	79.0	81.1	81.0
SAM2(Point)	2024	full+prompt	85.3	86.4	79.0	89.0	83.9	84.6	87.6	85.7
SAM2(Box)	2024	full+prompt	93.0	92.3	91.0	89.7	92.4	91.7	93.1	92.6
SAM3(Point)	2025	full+prompt	86.7	82.0	77.8	80.7	84.6	81.4	74.5	72.4
SAM3(Box)	2025	full+prompt	93.4	92.4	91.5	89.9	94.6	92.1	94.2	93.4
SAM2+OBBSeg(Point)	Ours	weak+prompt	91.9	87.3	86.9	91.2	92.1	88.5	90.3	89.7
SAM2+OBBSeg(Scribble)	Ours	weak+prompt	92.4	89.2	90.4	94.2	92.6	90.6	92.1	92.0
SAM2+OBBSeg(Box)	Ours	weak+prompt	95.1	93.1	92.7	95.7	94.3	93.6	94.9	94.5
SAM2+OBBSeg(Circle)	Ours	weak+prompt	95.1	93.8	92.9	95.7	94.8	94.0	95.1	94.8

Table 2: Comparison of OBBSeg with fully supervised models on skin lesion, nodule and tumor datasets.

Methods	Year	Supervision	Skin Lesion			Nodule & Tumor			
			ISIC2017	ISIC2018	Avg.	TN3K	DDTI	BUSI	Avg.
U-Net [34]	2015 _{MICCAI}	full	83.0	86.7	85.3	78.1	63.6	74.0	75.4
U-Net++ [47]	2018 _{DLMIA}	full	83.0	87.5	85.8	63.8	65.1	74.8	65.6
AttnUNet [29]	-	full	83.7	87.1	85.8	79.9	-	74.5	79.0
MEGANet [5]	2024 _{CVPR}	full	61.2	69.3	66.3	51.4	61.0	45.5	51.9
PraNet [14]	2020 _{MICCAI}	full	83.0	88.6	86.5	81.1	81.4	78.6	80.8
SANet [42]	2021 _{MICCAI}	full	84.8	88.1	86.9	83.8	83.6	80.1	83.2
TransUNet [6]	-	full	85.0	89.1	87.6	81.8	71.5	-	80.0
CASCADE [31]	2023 _{WACV}	full	85.5	90.4	88.6	-	-	79.2	79.2
EMCAD [32]	2024 _{CVPR}	full	86.0	91.0	89.1	-	-	80.3	80.3
KnowSAM [18]	2025 _{TMI}	semi	-	87.2	87.2	81.2	65.0	-	78.4
WeakPolyp [41]	2023 _{MICCAI}	weak	84.7	87.8	86.6	77.8	79.2	75.7	77.6
WeakPolyp+M2O	-	weak	85.2	89.3	87.8	78.8	80.0	77.7	78.8
SAM2(Point)	2024	full+prompt	83.7	87.0	85.7	79.1	85.4	86.2	81.1
SAM2(Box)	2024	full+prompt	87.1	90.2	89.0	81.3	90.2	89.5	83.8
SAM3(Point)	2025	full+prompt	76.4	83.6	80.7	58.2	58.8	72.4	60.4
SAM3(Box)	2025	full+prompt	90.2	95.6	93.5	87.4	88.9	88.6	87.8
SAM2+OBBSeg(Point)	Ours	weak+prompt	86.0	89.3	88.0	84.3	86.9	87.0	85.1
SAM2+OBBSeg(Scribble)	Ours	weak+prompt	90.9	92.5	91.9	86.6	89.1	89.9	86.7
SAM2+OBBSeg(Box)	Ours	weak+prompt	93.1	95.0	94.3	88.8	93.3	92.7	89.0
SAM2+OBBSeg(Circle)	Ours	weak+prompt	93.7	94.8	94.4	91.2	94.3	93.2	92.0

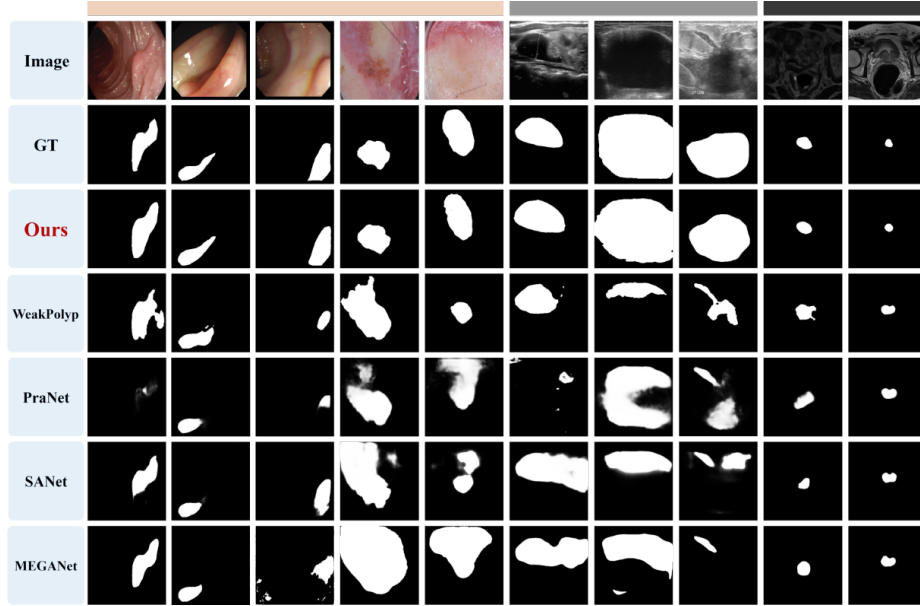


Fig. 4: Visualization comparison across three modalities.

Competing Methods. We compare OBBSeg with 20 representative methods, including U-Net [34], U-Net++ [47], AttnUNet [29], TransUNet [6], PraNet [14], FLA-Net [25], MS-TFAL [11], SANet [42], LGRNet [44], MEGANet [5], CASCADE [31], EMCAD [32], WeakPolyp [41], KnowSAM [18], BoxInst [38], BoxTeacher [8], PointSup [7], AGMM [43], GazeMedSeg [46], and MedSAM2 [26].

Implementation Details. We use two backbones: Res2Net50 [15] and the SAM2 encoder [33, 35]. When using SAM2, we initialize the encoder with its pretrained weights and keep them frozen during training, and optimize only the OBBSeg-specific modules. Input images are resized to 352×352 . Following WeakPolyp [41], OBB annotations are generated from ground truth masks via random rotations and translations (Table 9 and Table 10) to simulate real annotations. In the promptable setting, prompts are also simulated from ground truth masks following the standard protocol in promptable segmentation. These masks are used only offline to derive sparse or coarse prompt cues, and are not provided as dense training targets or loss supervision.

Training is performed on two RTX 4090 GPUs using AdamW ($\text{lr}=1e^{-4}$, weight decay= $5e^{-4}$) with a batch size of 16 for 16 epochs. For datasets without official splits, we adopt an 80/20 train-test split. During inference, OBBSeg can operate with prompts (SAM2+OBBSeg) or in a prompt-free manner, enabling both interactive and automatic segmentation.

Evaluation Metrics. Model performance is evaluated using the Dice coefficient and the 95th percentile Hausdorff Distance (HD95). Dice measures overlap,

while HD95 measures boundary accuracy. Higher Dice and lower HD95 indicate better performance. See the supplementary material for detailed definitions.

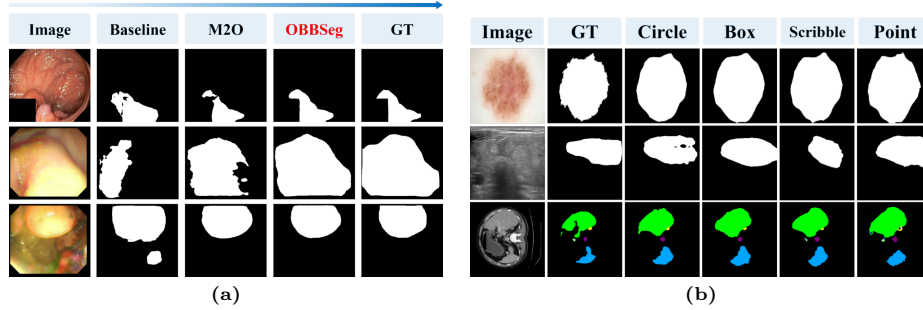


Fig. 5: (a) Impact of the proposed modules with the SAM2 backbone. (b) Visualization of representative failure cases and limitations.

Table 3: Comparison with weakly supervised models.

Methods	Supervision	Kvasir	NCI-ISBI
BoxInst [38]	Box	65.7	73.78
AGMM [43]	Scribble	67.2	72.7
PointSup [7]	Point	73.0	73.4
BoxTeacher [8]	Box	73.3	75.6
AGMM [43]	Point	75.5	73.8
GazeMedSeg [46]	Gaze	77.8	77.6
WeakPolyp+M2O	OBB	90.0	85.1
OBBSeg(Ours)	OBB+Point	91.2	86.0
OBBSeg(Ours)	OBB+Scribble	94.2	90.8
OBBSeg(Ours)	OBB+Box	95.7	93.3
OBBSeg(Ours)	OBB+Circle	95.7	94.1

Quantitative Comparison. Tables 1 and 2 report OBBSeg results on diverse segmentation tasks, using the weighted average Dice score (“Avg.”) based on dataset sizes. Two baselines, WeakPolyp and SAM2, are considered. WeakPolyp uses Res2Net [15] as the backbone and does not require prompts. By simply adding the M2O loss without other modifications, we achieve a consistent **1%–2%** improvement across all datasets, demonstrating the effectiveness of M2O. For SAM2, we integrate the full OBBSeg framework (M2O + Prompt) and evaluate different prompt types. This leads to substantial performance gains, approaching, and on several datasets exceeding, the performance of fully supervised methods across datasets. These results demonstrate the effectiveness of combining weak supervision with prompt guidance, highlighting its potential for large-scale applications. Table 3 further compares OBBSeg with existing weakly supervised methods. WeakPolyp+M2O already outperforms prior approaches by

Table 4: Ablation study of $\mathcal{L}_{M2O}$, $\mathcal{L}_{PS}$, PAFE, and DBFE on five polyp datasets. "Base" denotes the baseline using the SAM2 encoder trained under OBB supervision, rather than the original fully supervised SAM2.

Base	$\mathcal{L}_{M2O}$	$\mathcal{L}_{PS}$	DBFE	PAFE	without prompt			
✓					77.2			
✓	✓				80.5			
✓	✓		✓		83.3			
✓	✓			✓	81.3			
✓	✓		✓	✓	83.6			
Base	$\mathcal{L}_{M2O}$	$\mathcal{L}_{PS}$	DBFE	PAFE	Circle	with prompt Box	Scribble	Point
✓	✓	✓			82.5	82.4	78.5	78.2
✓	✓	✓	✓		91.8	91.5	86.4	81.7
✓	✓	✓	✓	✓	94.0	93.6	90.6	88.5

Table 5: Detailed segmentation results for all organs. Dice (in %) is used for evaluation.

Prompts	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach	Avg.
Point	85.8	86.3	84.6	85.0	86.3	80.4	82.9	79.4	83.8
Box	90.5	81.3	82.6	91.4	85.5	85.2	87.6	88.2	86.5
Circle	91.7	82.6	80.1	90.9	89.0	84.7	87.2	89.0	86.9
Scribble	88.8	88.9	83.5	89.3	92.0	84.8	87.5	91.1	88.2

a large margin. Building upon this, OBBSeg with prompt guidance further improves performance, achieving state-of-the-art results on both the polyp (Kvasir) and prostate (NCI-ISBI) datasets.

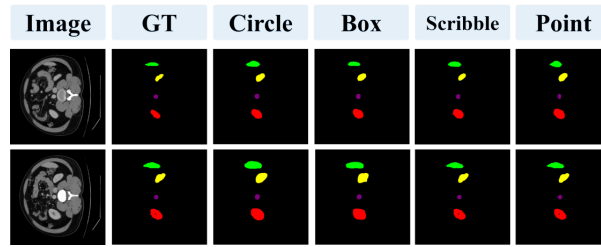


Fig. 6: Visualization of multi-organ segmentation on Synapse dataset.

Multi-organ Segmentation. We evaluate OBBSeg on the 8-class Synapse dataset [17], achieving average Dice of **83.8%**, **86.5%**, **86.9%**, **88.2%** across Point, Box, Circle, Scribble prompts in Table 5, respectively, demonstrating its effectiveness for multi-class segmentation.

Visualization Comparison. Fig. 4 shows qualitative comparisons across three modalities. Despite using only coarse OBB supervision, our model (row 3) accurately captures lesion boundaries. Especially for elongated or tilted struc-

tures (column 1, 2, 3), OBBSeg not only outperforms the weakly supervised WeakPolyp but also achieves visually competitive results with fully supervised models(row 5, 6, 7), highlighting OBBSeg’s consistent superiority across diverse modalities.

Ablation Study. Table 4 and Fig. 5(a) present the contribution of each module. M2O provides clear gains and establishes robust supervision with reduced box bias. When prompts are available, the prompt supervision loss yields the largest improvement, with circle prompts achieving the best performance. DBFE and PAFE are complementary modules that benefit from accurate foreground cues provided by prompts. DBFE relies on foreground-background contrast estimation; however, noisy early predictions limit its effectiveness without prompt guidance. Prompt supervision supplies reliable foreground cues, enabling DBFE to better distinguish features and further improve performance.

Boundary Segmentation Analysis. Table 6 reports boundary segmentation performance (HD95) under box and OBB annotations. OBB supervision achieves notably lower HD95 values, indicating more precise and smoother boundaries, confirming that OBBs provide stronger geometric constraints and better boundary localization.

Performance under Equivalent Annotation Budget. To further validate the efficiency of Oriented Bounding Box (OBB) supervision, we conduct a comparative analysis under a fixed annotation budget in Table 7. Based on the average annotation time per instance, we subsample the OBB training set such that its cumulative annotation cost aligns with that of the full axis-aligned bounding box (AABB) training set.

Elongated Lesion Segmentation. Table 8 presents segmentation results for elongated lesions. Incorporating OBB and M2O loss improves performance, with particularly strong gains for slender lesions (+5.3%), demonstrating that OBBs better capture the shape of elongated structures.

Table 6: Comparison of boundary segmentation using the HD95 index under box and OBB annotations (lower is better).

	Methods	Five Standard Polyp Datasets					SUN-SEG	
		Clinic	Colon	ETIS	Kvasir	Endo	Easy	Hard
Box	Point	16.5	24.5	21.3	23.6	6.7	17.8	19.1
	Scribble	13.1	17.0	10.1	15.9	8.6	15.1	15.7
	Box	16.0	16.1	10.9	22.5	5.0	18.4	18.6
	Circle	15.0	12.1	10.6	18.4	4.6	16.9	17.3
OBB	Point	9.2	16.6	13.0	16.7	5.7	12.0	12.6
	Scribble	7.7	11.2	6.2	9.8	6.0	11.2	11.0
	Box	4.7	5.9	5.8	6.8	4.3	7.1	7.0
	Circle	5.2	5.4	4.5	8.2	3.5	6.9	6.6

OBB Sensitivity Analysis. We evaluated the sensitivity of OBB annotation to rotational and positional errors. Table 9 shows that varying OBB angle by $\pm 30^\circ$ causes only a minor performance drop ($< 2\%$). Table 10 shows that shift-

Table 7: Performance comparison under the same total annotation budget.

Method	+Point	+Scribble	+Box	+Circle
Axis-aligned Box	86.5%	89.2%	84.7%	93.2%
OBB (Ours)	87.8%	90.9%	93.5%	93.7%

Table 8: M2O loss for elongated (aspect ratio > 2) lesions

Methods	FSPD		SUN-SEG	
	Avg.	Elongated	Avg.	Elongated
WeakPolyp [41]	76.7	65.9	79.8	76.2
WeakPolyp+M2O	79.0	71.2	81.0	81.5
–	(+2.3)	(+5.3)	(+1.2)	(+5.3)

Table 9: OBB angle sensitivity. Max/Min values in each row are in red; Δ Dice indicates the maximum performance drop.

Angle	0°	±5°	±10°	±15°	±20°	±25°	±30°	Δ Dice
Point	88.5	88.4	87.8	87.6	87.5	87.6	86.6	-1.9
Scribble	90.6	89.9	89.9	89.5	89.9	89.1	88.6	-2.0
Box	93.6	93.2	93.1	93.2	93.2	93.2	93.2	-0.5
Circle	94.0	93.2	93.0	92.8	92.3	92.6	91.7	-2.3

Table 10: Ablation study on OBB padding sensitivity; Δ Dice indicates the performance drop.

Padding	0	0~5	0~10	0~15	0~20	Δ Dice
WeakPolyp [41]	78.5	75.3	71.7	69.0	66.3	-12.2
Point	88.5	87.1	87.1	84.3	81.3	-7.2
Scribble	90.6	90.3	89.0	86.5	85.1	-5.5
Box	93.6	92.9	91.5	88.3	88.4	-5.3
Circle	94.0	92.6	91.4	87.8	87.5	-6.5

Table 11: Robustness of human OBB annotations under comparable deviations. Deviation is measured by $1 - \text{IoU}(B_{\text{ref}}, B_{\text{ann}})$ (%); Δ Dice indicates the performance drop.

Method	Setting	Deviation	Δ Dice
WeakPolyp	Box (padding=0~20)	19.74	-12.2
WeakPolyp+M2O	Human OBB Annotation	21.20	-1.5

ing the padding by up to 15 pixels has negligible effect, demonstrating strong tolerance to annotation errors.

We further evaluate the robustness of human OBB annotations introduced in Sec. 1. As shown in Table 11, human OBBs have a deviation of 21.20%, which is comparable to the strongest Box perturbation in Table 10 (Padding 0~20, deviation 19.74). Under a similar perturbation level, WeakPolyp suffers a substantial drop of 12.2% ($78.5 \rightarrow 66.3$), whereas WeakPolyp+M2O with human OBBs decreases by only 1.5% ($79.0 \rightarrow 77.5 \pm 0.32$). This suggests that the proposed OBB supervision remains stable under realistic annotation variations.

Model Size and Inference Time. OBBSeg can be integrated with different backbones, and the model size and inference time therefore depend on the selected architecture. Using SAM2 as the backbone, the model size is approximately 900 MB, with per-image inference times of 8.4 ms (Point), 12.3 ms (Scribble), 12.1 ms (Box), and 8.2 ms (Circle).

Failure Cases and Limitations. Due to the coarse nature of weak supervision, the annotations lack fine-grained boundary details. In addition, the projection mechanism in the M2O loss provides limited supervision within the OBB region. As shown in Fig. 5(b), the method may produce suboptimal segmentation for lesions with complex contours or indistinct boundaries. Furthermore, the projection-based design may struggle with highly complex tree-like anatomical structures (*e.g.*, vascular networks) where topology plays an important role. The current framework is also designed for static 2D images and does not explicitly model temporal dependencies in medical videos. Extending the framework to native 3D medical volumes remains an important direction for future work.

5 Conclusion

We introduced OBBSeg, a new supervision paradigm that combines oriented bounding boxes with prompt-guided learning to provide stronger geometric constraints than conventional weak labels. The proposed Mask-to-OBB loss and hierarchical prompt supervision alleviate box-induced bias and significantly improve segmentation performance. Extensive experiments show that OBBSeg achieves accuracy approaching fully supervised methods across diverse datasets. Nevertheless, segmentation quality is still limited by the coarse nature of OBB annotations. Future work will explore adaptive label refinement and hybrid supervision to further enhance scalability and generalization.

Acknowledgments

This work was supported in part by ICFCRT (W2441020), Guangdong Basic and Applied Basic Research Foundation (2023B1515120026), Shenzhen Peacock Program (KQTD20210811090044003), Young Faculty Startup Fund of Shenzhen University, and Scientific Development Fund from Guangdong Provincial Key Laboratory of Visual Media and Multidimensional Intelligence.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data in Brief* **28**, 104863 (2020)
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What’s the point: Semantic segmentation with point supervision. In: *ECCV*. pp. 549–565. Springer (2016)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilar-iño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG* **43**, 99–111 (2015)
4. Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K.: Nci-isbi 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive* (2015). <https://doi.org/10.7937/K9/TCIA.2015.ZFOVLOPV>
5. Bui, N.T., Hoang, D.H., Nguyen, Q.T., Tran, M.T., Le, N.: MEGANET: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In: *WACV*. pp. 7985–7994 (2024)
6. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *MIA* p. 103280 (2024)
7. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-Supervised Instance Segmentation. In: *CVPR*. pp. 2617–2626 (2022)
8. Cheng, T., Wang, X., Chen, S., Zhang, Q., Liu, W.: Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In: *CVPR*. pp. 3145–3154 (2023)
9. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: *ISBI*. pp. 168–172 (2018)
10. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: *ISBI*. pp. 168–172 (2018)
11. Cui, B., Zhang, M., Xu, M., Wang, A., Yuan, W., Ren, H.: Rectifying noisy labels with sequential prior: Multi-scale temporal feature affinity learning for robust video segmentation. In: *MICCAI*. pp. 90–100. Springer (2023)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR* (2021)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *IJCV* **88**(2), 303–338 (2010)
14. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: *MICCAI*. pp. 263–273. Springer (2020)
15. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. *TPAMI* **43**(2), 652–662 (2019)
16. Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., Li, G.: Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: *ISBI*. pp. 257–261. IEEE (2021)

17. harrigr: Segmentation outside the cranial vault challenge (2015). <https://doi.org/10.7303/SYN3193805>, <https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn3193805&type=ENTITY>
18. Huang, K., Zhou, T., Fu, H., Zhang, Y., Zhou, Y., Gong, C., Liang, D.: Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *TMI* (2025)
19. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. *TMI* **33**(2), 233–245 (2013)
20. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: *International Conference on Multimedia Modeling*. pp. 451–462. Springer (2019)
21. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research* **19**(6), 531–549 (2022)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: *ICCV*. pp. 4015–4026 (October 2023)
23. Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J.: Weakly supervised mitosis detection in breast histopathology images using concentric loss. *MIA* **53**, 165–178 (2019)
24. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: *CVPR*. pp. 3159–3167 (2016)
25. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: *MICCAI*. pp. 497–507. Springer (2023)
26. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600* (2025)
27. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *TMI* **34**(10), 1993–2024 (2014)
28. Oh, Y., Kim, B., Ham, B.: Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In: *CVPR*. pp. 6913–6922 (2021)
29. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. *arXiv Preprint arXiv:1804.03999* (2018)
30. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: *International Symposium on Medical Information Processing and Analysis*. vol. 9287, pp. 188–193. SPIE (2015)
31. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: *WACV*. pp. 6222–6231 (2023)
32. Rahman, M.M., Munir, M., Marculescu, R.: EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: *CVPR*. pp. 11769–11779 (2024)
33. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollar, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: *ICLR*. vol. 2025, pp. 28085–28128 (2025)
34. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *MICCAI*. pp. 234–241. Springer (2015)

35. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: ICML. pp. 29441–29454. PMLR (2023)
36. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery* **9**(2), 283–293 (2014)
37. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. *TMI* **35**(2), 630–644 (2015)
38. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: CVPR. pp. 5443–5452 (2021)
39. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, 180161 (2018). <https://doi.org/10.1038/sdata.2018.161>
40. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering* **2017**(1), 4037190 (2017)
41. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: MICCAI. pp. 757–766. Springer (2023)
42. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI. pp. 699–708. Springer (2021)
43. Wu, L., Zhong, Z., Fang, L., He, X., Liu, Q., Ma, J., Chen, H.: Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In: CVPR. pp. 15454–15464 (2023)
44. Xu, H., Yang, Y., Aviles-Rivero, A.I., Yang, G., Qin, J., Zhu, L.: Lgrnet: Local-global reciprocal network for uterine fibroid segmentation in ultrasound videos. In: MICCAI. pp. 667–677. Springer (2024)
45. Zhang, X., Zhu, L., He, H., Jin, L., Lu, Y.: Scribble hides class: Promoting scribble-based weakly-supervised semantic segmentation with its class label. In: AAAI. pp. 7332–7340 (2024)
46. Zhong, Y., Tang, C., Yang, Y., Qi, R., Zhou, K., Gong, Y., Heng, P.A., Hsiao, J.H., Dou, Q.: Weakly-supervised medical image segmentation with gaze annotations. In: MICCAI. pp. 530–540. Springer (2024)
47. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *TMI* (2019)