SSR-SAM: Retrieval-Style Segment Anything Model for Semi-Supervised Ultra-High-Resolution Image Segmentation

Shijie Li[1], Yiming Chen[1], Zhineng Chen[2], Kai Hu[3], Xieping Gao[4]*
[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Institute of Trustworthy Embodied AI, Fudan University
[3]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University
[4]Hunan Provincial Key Laboratory of Al and International Communication , Hunan Normal University
{lisj19, zhinchen}@fudan.edu.cn, ymchen23@m.fudan.edu.cn, kaihu@xtu.edu.cn, xpgao@hunnu.edu.cn

Abstract

Accurate segmentation of ultra-high-resolution (UHR) images, which often exceed tens of millions of pixels, is critically important in domains such as remote sensing and biomedical imaging. However, acquiring pixel-level annotations for such high-resolution images is prohibitively expensive and labor-intensive. While semi-supervised semantic segmentation can significantly reduce the annotation burden, its extension to UHR images holds great potential for addressing the unique challenges posed by sparse supervision. To this end, we propose SSR-SAM, a retrieval-style semi-supervised segmentation framework tailored for UHR images. Leveraging the promptable paradigm of the Segment Anything Model (SAM), SSR-SAM treats locally annotated regions as prompts to retrieve semantically consistent pixels across the entire image. Building upon this retrieval-style segmentation paradigm, we further introduce prompt-level perturbation, a novel trail to deploy consistency regularization for semi-supervised segmentation. It encourages the model to learn consistency across predictions guided by diverse visual-semantic prompts, thereby enhancing generalization on unlabeled data. We evaluate SSR-SAM on three UHR datasets: Inria Aerial, BCSS, and URUR. Experimental results show that SSR-SAM achieves clear performance gains over the labeled-only supervision, with average mIoU improvements of 4.9%, 4.15%, and 2.5%, respectively. Additionally, SSR-SAM possesses zero-shot segmentation capability, exhibiting potential for general retrieval-style segmentation tasks.

Code — https://github.com/LSJ-i/SSR-SAM

Introduction

Semantic segmentation has witnessed significant progress with the rise of deep learning techniques (Long, Shelhamer, and Darrell 2015; Chen et al. 2017; Ravi et al. 2024; Xie et al. 2025). Among its key sub-fields, Ultra-High-Resolution (UHR) image segmentation has gained attention in applications like remote sensing and digital pathology (Chen et al. 2019a; Li et al. 2021; Shen et al. 2022; Shan et al. 2021; Sun et al. 2023; Luo et al. 2023). UHR segmentation aims to parse extremely large images containing dense or large-scale objects such as buildings, roads, or tumor regions, and the immense size and complexity of UHR data pose unique challenges. To tackle these challenges, previous works have explored cascade architectures (Cheng, et al. 2020), multi-scale feature fusion (Madabhushi and Lee 2016; Guo et al. 2022), and collaborative designs (Chen et al. 2019b), yielding improvements in accuracy and efficiency (Xu et al. 2021; Ji et al. 2023; Wang et al. 2025; Liu et al. 2024b; Shan and Wang 2022; He, Yang, and Qi 2021; Lai et al. 2021; Yang et al. 2023). However, these methods remain highly dependent on dense annotations, which are prohibitively expensive for UHR images. Noting that UHR images often exhibit semantic sparsity and structural redundancy, such as repeated or large-scale instances across megapixels, this opens opportunities to exploit semantic consistency for efficient learning under limited supervision. Motivated by this, we explore semi-supervised semantic segmentation (SSS) for UHR images to reduce annotation cost while preserving segmentation quality.

The field of SSS has progressed rapidly, primarily built on self-training and consistency regularization (Grandvalet and Bengio 2004; Pham et al. 2021; Xie et al. 2019). Self-training uses reliable pseudo-masks, while consistency regularization enforces stable predictions under input perturbations. Based on these principles, numerous SSS methods have been proposed (Zhong et al. 2021; Alonso et al. 2021; Kwon and Kwak 2022; Liu et al. 2021), achieving

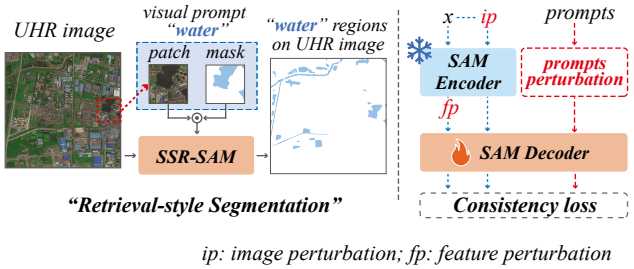


Figure 1: **Left**: Retrieval-style segment model utilizes local patch and its category-specific mask as prompt to retrieve the semantically similar pixels on other regions. **Right**: Prompt-level perturbation introduces another trail to deploy consistency regularization within SSS.

*Corresponding Author

promising performance with limited labels. More recently, the integration of pre-trained foundation models such as CLIP (Radford et al. 2021) and DINO v2 (Oquab et al. 2023) has further advanced SSS (Hoyer et al. 2025; Yang, Zhao, and Zhao 2024). Despite this trend, the Segment Anything Model (SAM) (Kirillov et al. 2023), a foundation model designed for universal segmentation, remains largely underexplored in SSS, especially for UHR data. Given SAM's powerful promptable segmentation across domains (Chen, et al. 2024; Wu et al. 2023), incorporating it into SSS frameworks has strong potential for improving generalization.

As an interactive segmentation model, SAM generates masks based on user-defined prompts (e.g., points, boxes, or text). Despite its strong generalization capability, it is class-agnostic and lacks the semantic specificity required for semantic segmentation. To address this limitation, recent studies have introduced learnable prompts or aligned prompts with external detectors (Zhang et al. 2023; Li, et al. 2023; Ren et al. 2024; Liu et al. 2024a). While effective to some extent, these approaches often tie specific categories to fixed prompt embeddings, restricting flexibility and expressiveness. In contrast, we consider prompt diversity a crucial factor in shaping segmentation outcomes, and leveraging this diversity offers a new perspective for integrating SAM into SSS frameworks.

Motivated by the above observations, we propose a retrieval-style SAM for semi-supervised segmentation on UHR images, termed SSR-SAM. As Figure 1 shows, SSR-SAM firstly constructs retrieval-style segmentation. It utilizes objects with known semantics to generate visual-semantic prompts and uses these prompts to retrieve semantically similar pixels across the entire image. This retrieval-style paradigm enables the reproduction of masks from sparse annotations to a much broader unknown extent, supporting efficient and scalable semantic segmentation on UHR images. Building upon the retrieval-style segmentation, we further propose prompt perturbation to enhance semi-supervised training. By deploying consistency regularization on these perturbed prompts and the original prompts, SSR-SAM enhances the generalization and robustness of retrieval-style segmentation on UHR images, promoting the sparse annotation driven accurate segmentation.

To enable SSR-SAM, we firstly propose a Mask-induced Semantic Prompt Generator (MSPG) to access diverse visual-semantic prompts. MSPG utilizes the annotations of labeled patches to generate category-specific masked images. These category-specific masked patches are then fed into a pre-trained ViT (Dosovitskiy, et al. 2021), and the outcome [CLS] tokens are served as the visual-semantic prompts. Once obtained, these prompts are leveraged not only during supervised training but also as the foundation for consistency regularization. Specifically, during training on labeled images, all visual-semantic prompts are fed into SSR-SAM to retrieve corresponding category outcomes. For unlabeled images, we apply prompt perturbation, where the teacher receives the full set of semantic prompts and the student is guided by a randomly sampled subset of the full set. Unlike traditional image- or feature-level perturbations, prompt perturbation introduces variation solely at the prompt level, encouraging the student to remain consistent with the teacher's predictions despite reduced visual-semantic prompts. When combined with a self-training pipeline, this design enforces prompt-level consistency regularization and promotes robust learning from unlabeled data.

Empowered by MSPG and prompt perturbation driven consistency regularization, SSR-SAM forms a retrieval-style semi-supervised semantic segmentation framework tailored for UHR images. We evaluate SSR-SAM alongside state-of-the-art SSS methods on several representative UHR image benchmarks, including Inria Aerial, URUR, and BCSS. Experimental results demonstrate the effectiveness of SSR-SAM, with clear mIoU improvements observed on multiple datasets.

Furthermore, SSR-SAM inherently supports open-set segmentation. By using any known masked object as the query, it can retrieve semantically related regions across different images. We validate this zero-shot prediction capability on the LEVIR-CD dataset, demonstrating the model's potential in open-world retrieval-style segmentation.

Our main contributions can be summarized as follows:

- We propose a retrieval-style segmentation framework based on SAM. A mask-induced semantic prompts generator is proposed to obtain semantic prompts for retrieval-style segmentation.
- Based on the retrieval-style segmentation, we introduce prompts perturbation, a new trail to deploy consistency regularization under the semi-supervised segmentation.
- Experiments on Inria Aerial, BCSS, and URUR confirm the effectiveness of SSR-SAM under semi-supervised conditions. And its zero-shot capability is validated on LEVIR-CD, highlighting its potential for open-world segmentation.

Related Work

In this section, we provide a brief review of recent works related to UHR image segmentation, SSS, and promptable segmentation.

UHR Image Segmentation

UHR segmentation plays a vital role in domains like remote sensing (Kotaridis and Lazaridou 2021; Yuan, Shi, and Gu 2021) and digital pathology (Wu et al. 2015; Kumar et al. 2017), where fine-grained analysis of megapixel-scale images is essential. Recent methods such as CascadePSP (Cheng, et al. 2020), SGNet (Wang et al. 2025), and ISDNet (Guo et al. 2022) improve segmentation quality and efficiency through cascaded refinement, contextual integration, and multi-resolution processing, respectively. Transformer-based models (Dosovitskiy, et al. 2021; Xie, et al. 2021) further enhance global context modeling. However, semi-supervised solutions tailored for UHR segmentation remain largely underexplored to date.

Semi-Supervised Semantic Segmentation

SSS has achieved strong performance by combining self-training with consistency regularization (Chen et al. 2021;

Yu et al. 2019; Bai et al. 2023). CutMix-based methods (Yun et al. 2019; Chen et al. 2021), FixMatch-style augmentations (Sohn et al. 2020; Yang et al. 2023), and teacher-student frameworks (Tarvainen and Valpola 2017; Berthelot et al. 2019) dominate this field. Recent approaches like CorrMatch (Sun et al. 2024) and AllSpark (Wang et al. 2024) explore correlation-aware learning and labeled-feature-driven supervision, respectively. Foundation-model-based methods, e.g., SemiVL (Hoyer et al. 2025), further improve SSS by incorporating vision-language priors. Despite these advances, their application to UHR segmentation is rarely addressed.

Promptable Segmentation

Promptable segmentation enables flexible mask generation via interactive inputs (e.g., points, boxes, text). SAM (Kirillov et al. 2023) exemplifies this direction with strong generalization but lacks semantic labeling. Extensions such as Semantic-SAM (Li, et al. 2023) and Grounded SAM (Ren et al. 2024) incorporate language models or detection to provide semantic prompts. Recent works explore SAM in specialized domains including medical (Zhang et al. 2023; Ma, et al. 2024) and remote sensing images (Wang, et al. 2024; Osco, et al. 2023), adapting its capabilities for fine-grained, domain-specific segmentation. However, its integration into semi-supervised UHR segmentation remains unexplored.

Methodology

SSR-SAM employs SAM as the backbone, followed by the Mask-Induced Semantic Prompt to implement retrieval-style segmentation. Moreover, inspired by UniMatch (Yang et al. 2023), we develop our semantic prompt perturbation strategy to enhance consistency regularization.

Mask-induced Semantic Prompt Generation

Our goal is to develop a retrieval-style segmentation framework for UHR images by leveraging promptable segmentation. The Segment Anything Model (SAM) and its enhanced variants (Ravi et al. 2024) provide a natural foundation, exhibiting strong performance across diverse real-world scenarios and serving as an ideal backbone for our approach.

SAM employs a three-stage architecture: (1) an image encoder to extract global features, (2) a prompt encoder to map various prompt types (e.g., points, boxes, text) into a shared embedding space, and (3) a mask decoder that treats these prompt embeddings as queries to retrieve the corresponding masks from image features. SSR-SAM adopts this paradigm (Figure 2a) but extends it to capture visual–semantic associations between similar objects—an essential capability missing from vanilla SAM.

To address this, we introduce the Mask-induced Semantic Prompt Generator (MSPG), a module that extracts class-aware semantic prompts from selected object regions. As depicted in Figure 2b, MSPG produces rich semantic embeddings by applying class-specific masking to local patches and feeds them into a pre-trained encoder, thereby embedding explicit visual–semantic context into each prompt.

Let x^l be a labeled image patch and m_c be one of its binary masks for the category c (e.g., a satellite image patch and its associated "Water" region). Each pixel in m_c takes a value of either 0 or 1. By performing an element-wise product, we obtain a masked image that preserves only the foreground of the target category from C categories:

$$x_c^m = x^l \odot m_c, \quad c = 0, 1, \ldots, C-1 \tag{1}$$

We feed x_c^m into a pre-trained ViT, and extract the output `[CLS]` token as the semantic prompt s_c, which serves as a compact category-specific embedding:

$$s_c = \mathcal{M}(x_c^m) \tag{2}$$

Here, $\mathcal{M}(\cdot)$ denotes the Mask-induced Semantic Prompt Generator (MSPG). Applying this process to multiple category-specific masks of x^l yields a collection of semantic tokens. These tokens form a set of visual-semantic prompts $\{s_c^l\}$, which will be used to guide our retrieval-style segmentation framework.

Since the number of prompts per class can vary—hindering parallel decoding in SAM—we introduce a learnable query embedding. For each category c, the MSPG yields a variable-sized set of semantic tokens $\mathrm{S}_c = \{ s_{c,1}, s_{c,2}, \ldots, s_{c,n_c} \}$,, where n denotes the number of instances of category c in the images. We introduce a learnable query vector $q_c \in \mathbb{R}^d$. The semantic tokens are projected to keys and values and a cross-attention operation compresses the variable-sized token set into a fixed-length prompt:

$$K_c = W_k \mathrm{S_c}, V_c = W_v \mathrm{S_c} \tag{3}$$

$$s_c^{\mathrm{com}} = \mathrm{softmax}\left(\frac{q_c K_c^\top}{\sqrt{d}} \right) V_c. \tag{4}$$

where $W_k, W_v \in \mathbb{R}^{d \times d}$ are learnable matrices and $K_i, V_i \in \mathbb{R}^{n \times d}$. The resulting $s_c^{\mathrm{com}} \in \mathbb{R}^d$ is used as the final category-level semantic prompt. This transformation is integrated directly into the decoder; therefore, in subsequent definitions, we do not state it explicitly and treat it as an inherent capability of the decoder.

With this setup, the retrieval-style segmentation pipeline can be formulated as:

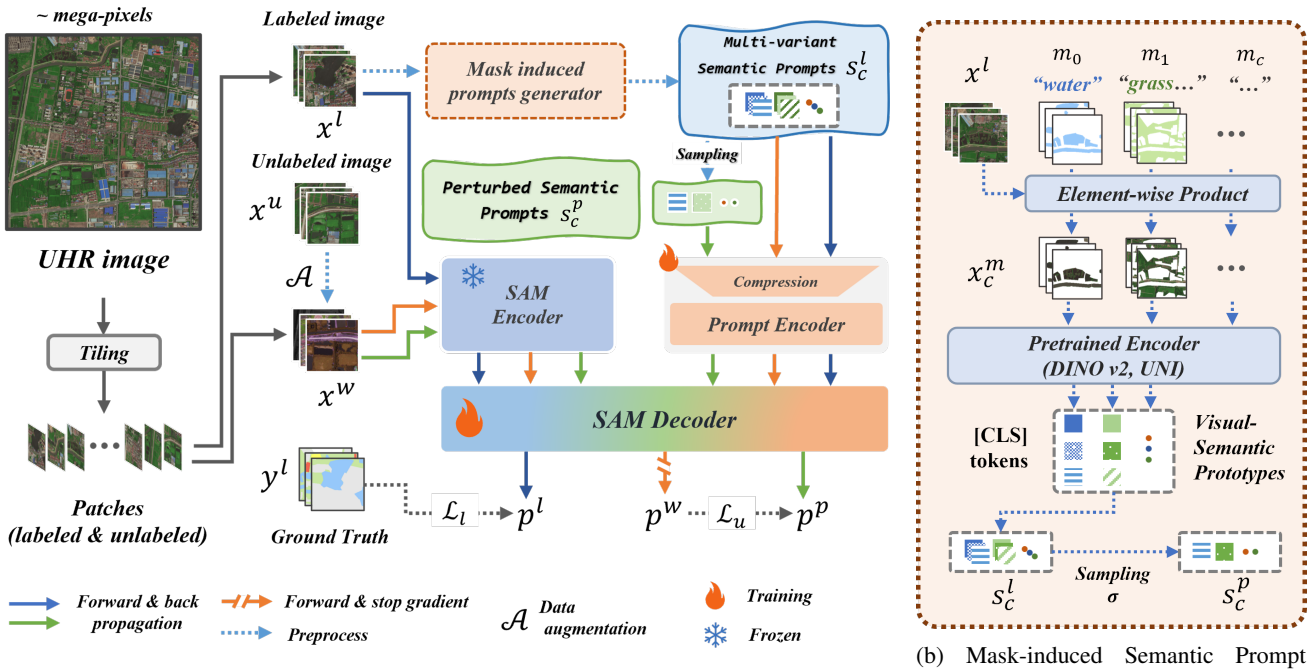
$$e = g(x), \quad p = h(e, s_c) \tag{5}$$

Here, $g(\cdot)$ and $h(\cdot, \cdot)$ denote the image encoder and mask decoder of SAM, respectively. The vector e corresponds to the image features extracted from an image patch x, and s_c is the visual-semantic prompt of category c.

Furthermore, SSR-SAM naturally enables zero-shot inference. By providing any image along with category-specific masks, the model can retrieve semantically similar regions across the image, without knowing its exact category. We validate this zero-shot capability on the LEVIR-CD dataset, and additional implementation details are provided in the Appendix.

SSS based on UniMatch

Based on the established retrieval-style segmentation pipeline, we describe the formulation of SSS based on UniMatch (Yang et al. 2023), which serves as our UHR SSS



(a) Framework of SSR-SAM.

(b) Mask-induced Semantic Prompt Generator.

Figure 2: (a) The framework of SSR-SAM. It deploys semi-supervised segmentation by capture invariance between semantic prompts of labeled image and their randomly sampled ones. The image- and feature-level perturbation is omitted. (b) The detail structure of our proposed mask-induced semantic prompts generator. Labeled images' masks are leveraged to produce multi-variant prompts for retrieval promptable segmentation and semi-supervised training.

baseline. SSS typically involves two data partitions: a small labeled set $\mathcal{D}^l = \{(x_i^l, y_i^l)\}$ and a large unlabeled set $\mathcal{D}^u = \{x_i^u\}$. On the labeled subset, training is conducted via standard supervised learning. Given a mini-batch of N^l labeled samples, the supervised loss is defined as:

$$\mathcal{L}^l = \frac{1}{N^l} \sum_{i=1}^{N^l} \mathrm{H}\left(p_i^l, y_i^l\right) \tag{6}$$

where p^l is the model's prediction, and $\mathrm{H}(\cdot)$ denotes the cross-entropy loss.

To exploit unlabeled data, UniMatch employs weak-to-strong consistency regularization, inspired by FixMatch (Sohn et al. 2020) and other consistency-based approaches (Xie et al. 2020; Berthelot et al. 2019, 2020). Each unlabeled image x^u undergoes two types of augmentations: a weak augmentation $\mathcal{A}^w$ (e.g., cropping, flipping) and a strong augmentation $\mathcal{A}^s$ (e.g., CutMix (Yun et al. 2019), color jittering). The model is trained to align predictions from the strongly augmented view with pseudo-labels generated from the weak view. We follow this paradigm and construct our retrieval-style framework:

$$x^w = \mathcal{A}^w(x^u),\ x^s = \mathcal{A}^s(x^w) \tag{7}$$

$$e^w = g(x^w),\ p^w = h(e^w, \{s_c^l\}), \tag{8}$$

$$e^s = g(x^s),\ p^s = h(e^s, \{s_c^l\}) \tag{9}$$

Additionally, UniMatch generates two independent strong augmentations x^{s_1} and x^{s_2} from x^w via non-deterministic augmentation $\mathcal{A}^s$:

$$x^{s_1} = \mathcal{A}^s(x^w),\ x^{s_2} = \mathcal{A}^s(x^w) \tag{10}$$

These multiple views are used to enforce consistency with the pseudo-labels from x^w.

Beyond image-level augmentations, UniMatch introduces feature-level perturbation to further enhance invariance. The weakly augmented input x^w is encoded to feature, then perturbed using a feature-level operator $\mathcal{F}$ (e.g., Dropout). We also adopt this strategy as:

$$p^f = h(\mathcal{F}(e^w), \{s_c^l\}) \tag{11}$$

In conclusion, the final unsupervised loss is:

$$\mathcal{L}^u = \frac{1}{2N^u} \sum_{i=1}^{N^u} \mathbb{1}(\max(p_i^w) \geq \tau) \cdot \left(\mathrm{H}(p_i^f, p_i^w) + \frac{1}{2}\left(\mathrm{H}(p_i^{s_1}, p_i^w) + \mathrm{H}(p_i^{s_2}, p_i^w)\right)\right) \tag{12}$$

Here, N^u is the number of unlabeled samples in a batch, and τ is a confidence threshold to filter unreliable pseudo-labels. In our implementation, we retain both image-level and feature-level perturbations for regularizing consistency over unlabeled samples.

Visual-Semantic Prompt Perturbation

Most existing SSS methods are built upon conventional architectures (e.g., ResNet + DeepLabv3+) and closed-set

datasets. However, recent advances push the field toward more effective and practical paradigms, such as incorporating foundation models (Hoyer et al. 2025; Yang, Zhao, and Zhao 2024) and tackling open-set tasks in domains.

Following this trend, SSR-SAM leverages SAM as its backbone and integrates the Mask-induced Semantic Prompt Generator (MSPG) to enable retrieval-style segmentation. To further enhance consistency regularization in SSS, we propose a novel strategy called visual-semantic prompt perturbation. As discussed in Section , modern SSS frameworks are primarily driven by two components: self-training and consistency regularization. While SSR-SAM implements self-training through pseudo-mask generation, traditional consistency regularization methods mainly focus on perturbations at the image or feature level. However, these strategies do not fully exploit the unique architecture of prompt-based models.

In SSR-SAM, predictions are directly influenced by semantic prompts in an independent embedding space. Thanks to MSPG, multiple semantic prompts from labeled data can be extracted within each class. These class-wise prompt embeddings naturally contain intra-class variation, which can be harnessed as a novel source of perturbation. We refer to this as visual-semantic prompt perturbation, which complements existing consistency regularization methods.

Formally, let $\{s_c^l\}$ denote all semantic prompts extracted from labeled data, which have C categories and each category contains arbitrary prototypes. For each unlabeled image, we sample a fixed number k of prompts per class (with replacement) to form perturbed prompts $\{s_c^p\}$. The full set $\{s_c^l\}$ is used by the teacher network, while the sampled subset $\{s_c^p\}$ is used by the student network.

Following the notation introduced earlier, the prediction for the perturbed prompts is:

$$\{s_c^p\} = \sigma(\{s_c^l\}, k), \quad p^p = h(e^w, \{s_c^p\}) \tag{13}$$

where $\sigma(\cdot)$ denotes random sampling with replacement and k is the sample size. Similar to Equation 12, combining all perturbations at image, feature, and prompts levels, the loss function of unlabeled images for SSR-SAM can be formulated as:

$$\begin{aligned}\mathcal{L}^u = \frac{1}{N^u} \sum \mathbb{1}(\max(p^w) \geq \tau) \cdot \\ \big(\alpha \cdot \mathrm{H}(p^f, p^w) + \beta \cdot \mathrm{H}(p^s, p^w) + \gamma \cdot \mathrm{H}(p^p, p^w)\big)\end{aligned} \tag{14}$$

where α, β, and γ are set to balance the effects of multiple perturbations, and we report the ablation of their combination in the Appendix. For simplicity, the dual views of strong augmentation, x^{s_1} and x^{s_2}, are denoted collectively, and their contributions are treated equally. Following the co-training paradigm, the overall training objective of SSR-SAM is defined as:

$$\mathcal{L} = \mathcal{L}^l + \lambda \mathcal{L}^u \tag{15}$$

where the λ is set to balance the effect of unlabeled data, and the default set is 0.5.

To summarize, traditional consistency regularization focuses on perturbations of image-level and feature-level representations (x and e). In contrast, our method pioneers a new perspective by introducing perturbations in the prompt space, which play a critical role in promptable segmentation. Additionally, pseudo-labels (i.e., m_c^u) from unlabeled images can be used to generate auxiliary semantic prompts during training. The procedure for obtaining these prompts is detailed in the Appendix, and their impact on prompt perturbation is evaluated in Figure 3.

Fully-Sup Method		Net	mIOU (%)			
WSDNet	[CVPR'23]	R18	75.2			
GPWFormer	[IJCAI'23]	R18	76.5			
UANet	[TGRS'24]	VGG16	83.1			
Semi-Sup Method (labeled images size)			2% (63)	4% (126)	8% (252)	16% (504)
AugSeg	[CVPR'23]	R101	76.8	77.6	79.1	79.5
UniMatch	[CVPR'23]	R101	75.4	78.4	81.0	82.6
CorrMatch	[CVPR'24]	R101	73.4	77.1	79.6	81.1
SemiVL	[ECCV'24]	CLIP	80.6	81.4	81.4	81.2
UniMatch v2	[TPAMI'25]	DINO	84.4	85.4	86.0	86.1
Labeled Only		SAM	79.5	80.8	82.1	82.2
SSR-SAM		SAM	**84.7**	**85.8**	**86.7**	**87.0**

Table 1: Comparison of segmentation performance on Inria Aerial dataset under varying supervision ratios.

Fully-Sup Method		Net	mIOU (%)			
ADS-UNet	[ESA'23]	UNet	61.1			
DETisSeg	[BSPC'24]	R50	62.9			
SAM-Path	[MedAGI'23]	SAM	66.0			
Semi-Sup Method (labeled images size)			1/64 (106)	1/32 (213)	1/16 (426)	1/8 (853)
CRCFP	[MedIA'24]	R101	-	-	-	47.2
AugSeg	[CVPR'23]	R101	40.1	43.7	45.7	46.1
UniMatch	[CVPR'23]	R101	42.9	45.3	46.4	46.9
CorrMatch	[CVPR'24]	R101	44.3	44.1	45.9	46.4
SemiVL	[ECCV'24]	CLIP	50.6	51.9	52.2	50.8
UniMatch v2	[TPAMI'25]	DINO	47.8	48.6	50.9	48.2
Labeled Only		SAM	36.1	43.4	45.6	50.7
SSR-SAM		SAM	39.1	47.0	52.1	54.2
SSR-SAM		UNI	**55.8**	**54.8**	**57.5**	**57.7**

Table 2: Comparison of segmentation performance on BCSS dataset under varying supervision ratios.

Experiments

Experimental Setup

Dataset: To evaluate the effectiveness of SSR-SAM on real-world UHR images like remote sensing and medical images, we benchmark SSS on Inria Aerial (Maggiori et al. 2017), URUR (Ji et al. 2023) and BCSS (Amgad et al. 2019). Inria Aerial has 180 UHR images. Each image contains 5000 × 5000 pixels and is annotated with a binary mask for building/non-building areas. It covers diverse urban landscapes, ranging from dense metropolitan districts to alpine resorts. For this dataset, we follow the protocol

Fully-Sup Method		Net	mIOU (%)			
ISDNet	[CVPR'22]	R18	45.8			
WSDNet	[CVPR'23]	R18	46.9			
Semi-Sup Method (labeled images size)			2% (1078)	4% (2157)	8% (4314)	16% (8628)
AugSeg	[CVPR'23]	R101	40.2	43.8	44.5	42.4
UniMatch	[CVPR'23]	R101	40.4	43.3	42.5	42.5
CorrMatch	[CVPR'24]	R101	42.7	40.4	40.7	42.3
SemiVL	[ECCV'24]	CLIP	44.9	46.1	45.8	46.1
UniMatch v2	[TPAMI'25]	DINO	45.1	45.5	46.3	46.4
Labeled Only		SAM	39.3	41.6	42.6	42.1
SSR-SAM		SAM	43.0	43.7	44.3	44.7

Table 3: Comparison of segmentation performance on URUR dataset under varying supervision ratios.

as (Ji et al. 2023; Chen et al. 2019b) by splitting images into training, validation and testing sets with 126, 27, and 27 images. The URUR dataset contains 3008 UHR images with size of 5120×5120, captured from 63 cities. The training, validation and testing set includes 2157, 280 and 571 images, respectively, with an approximate ratio of 7:1:2. All the images are exhaustively manually annotated with pixel-level categories including 8 classes: building, farmland, greenhouse, woodland, bareland, water, road and others. The BCSS dataset has over 20,000 semantic segmentation annotations of tissue regions sampled from 151 H&E stained breast cancer images at 40× magnification from TCGA-BRCA (Lingle et al. 2016). The annotations include 21 classes. We keep the major 4 classes: Tumor, Stroma, Inflammatory and Necrosis. The rest are grouped into the 'others' class. For URUR and BCSS, we use their official training, validation, and test splits. Furthermore, we use several images from the test set of LEVIR-CD (Chen and Shi 2020) to evaluate the zero-shot capability of SSR-SAM.

Implementation details: We first construct training data by tiling the UHR images into non-overlapping 1024×1024 (on Aerial and URUR) or 512 × 512 patches (on BCSS). Under this setup, UHR training images of Aerial, BCSS and URUR are tiled into 3150, 6784 and 53925 patches, respectively. For the prompt encoder and mask decoder, we also keep the same architecture as the vanilla SAM, except for adding a prompt compression block in front of the prompt encoder. We benchmark semi-supervised segmentation with incremental fractions from 2% to 16% on aerial images and 1/64 to 1/8 on pathological images and the detailed sizes are shown in Table 1-3. The results of only using the labeled data for fully supervised training is annotated as "Label only". Parameters of the encoder are frozen, the multivariant prompt compression, prompt encoder and mask decoder of SAM are trainable. AdamW (Loshchilov and Hutter 2018) optimizer with an initial learning rate of 2×10^{-4}, a cosine annealing without warm restart scheme (Loshchilov and Hutter 2016) for learning rate scheduling was used to update the model weights. All models including reproduced methods are trained for 100 epochs. Color transformations and CutMix are used to form $\mathcal{A}^s$. A raw image is resized between 0.5 and 2.0, cropped, and flipped to obtain its weakly augmented sample. We adopt a channel dropout of 50% probability for feature perturbation. The confidence threshold τ is set as 0.9 in all experiments.

Perturbations		Training size		
$p^s + p^f$	p^p	63	126	252
		79.5	80.8	82.1
	✓	83.1	85.1	86.2
✓		83.3	85.2	86.3
✓	✓	**84.7**	**85.8**	**86.7**

Table 4: Ablation study on the effectiveness of deploying different perturbations on Inria Aerial.

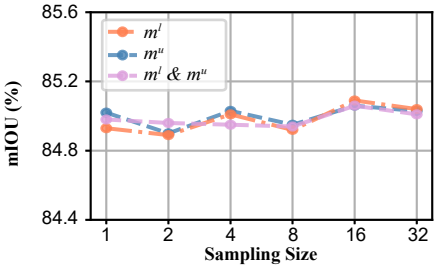


Figure 3: Comparison of different prompt perturbation strategies and sampling size on Inria Aerial (126 labeled).

Comparison with State-of-the-art Methods

We select five representative methods, i.e., AugSeg (Zhao et al. 2023), UniMatch (Yang et al. 2023), CorrMatch (Sun et al. 2024), SemiVL (Hoyer et al. 2025) and UniMatch v2 (Yang, Zhao, and Zhao 2024), to benchmark SSS on the mentioned *real-world* datasets. While AugSeg and UniMatch are classical frameworks based on ResNet & DeepLabv3+, CorrMatch is the current state-of-the-art with the same backbone and decoder. SemiVL introduces CLIP (Radford et al. 2021) into SSS, and UniMatch V2 updates the image encoder to DINO v2.

Inria Aerial. As shown in Table 1, our proposed SSR-SAM significantly outperforms several fully supervised semantic segmentation baselines even under extremely limited annotations (2% and 4%). Within the semi-supervised setting, we observe a consistent performance hierarchy favoring models built upon large pretrained vision backbones. For example, methods based on ResNet-101 are generally outperformed by those using vision-language pretrained models. This trend aligns with prior findings that leveraging strong visual priors significantly benefits segmentation under limited supervision. Notably, SSR-SAM achieves the highest mIoU across all labeled image ratios, respectively. These results highlight the effectiveness of our method in low-label regimes. Lastly, comparing SSR-SAM with its supervised-only counterpart, referred to as the "Labeled Only" baseline, it outperforms the baseline by 5.2%, 5.0%, 4.6%, and 4.8%, respectively. This consistent improvement confirms that our prompt perturbation contributes directly to

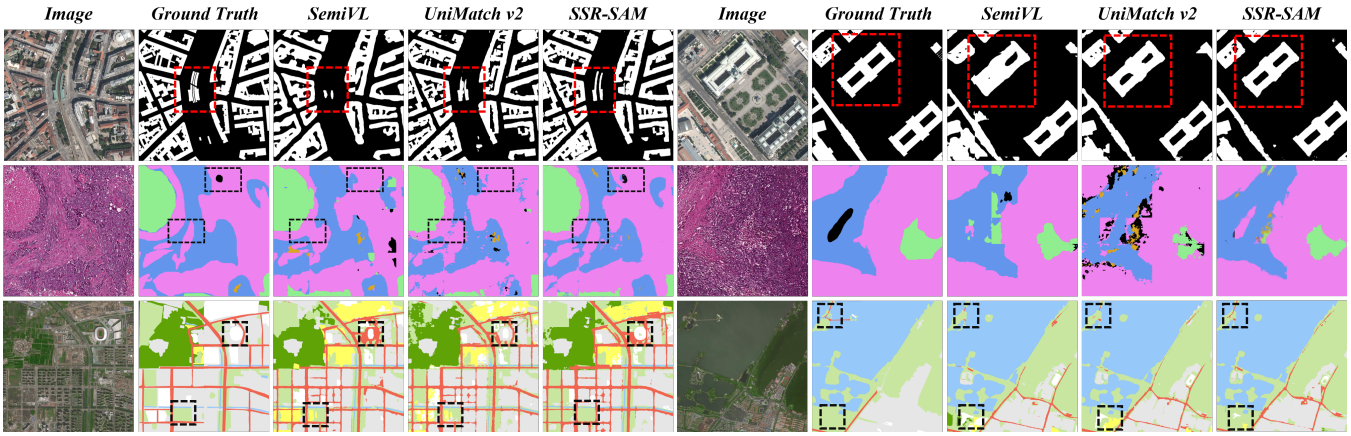


Figure 4: Qualitative comparisons with state-of-the-art methods SemiVL (Hoyer et al. 2025) and UniMatch v2 (Yang, Zhao, and Zhao 2024). The first to last rows are images from Inria Aerial, BCSS and URUR, respectively. All images are cropped and scaled into an appropriate region and size for visualization.

the performance gains and generalization robustness.

BCSS. As shown in Table 2, SSR-SAM with the UNI encoder consistently outperforms both fully supervised and semi-supervised baselines across all annotation ratios on the BCSS dataset. With only 1/8 of labels, SSR-SAM (with UNI) achieves 57.7% mIoU—approaching the fully supervised methods, despite using just a fraction of the annotations. Against existing semi-supervised methods, SSR-SAM (UNI) consistently leads: under extreme sparsity (1/64), it attains 55.8% mIoU versus 40.1%–44.3% for ResNet-based approaches (AugSeg, UniMatch, CorrMatch) and 47.8%–50.8% for CLIP/DINO-based methods (UniMatch v2, SemiVL), confirming SAM's strong transferability and advantages of introducing it to SSS. Comparing to the "Labeled Only" retrieval baseline, SSR-SAM's semi-supervised modules yield substantial gains: using SAM as the backbone, prompt perturbation and MSPG boost 1/8 performance from 50.7% to 54.2% (+3.5), and switching to the UNI encoder further raises it to 57.7% (+7.0), highlighting the effectiveness of prompt-based retrieval and consistency regularization, as well as SSR-SAM's flexibility to integrate any frozen pretrained backbone while delivering strong semi-supervised performance.

URUR. As reported in Table 3, SSR-SAM (SAM) significantly narrows the gap to fully supervised baselines using only partial labels. Against traditional semi-supervised methods, SSR-SAM (SAM) leads clearly: at 2% labels it attains 43.0% mIoU versus 40.2%–42.7% for AugSeg, UniMatch, and CorrMatch, confirming its improvements under sparse supervision. While CLIP- and DINO-based methods (SemiVL: 44.9%, UniMatch v2: 45.1% at 2%) benefit from strong visual priors, SSR-SAM's frozen SAM encoder combined with prompt perturbation still yields competitive results. Compared to the "Labeled Only" baseline, our semi-supervised design consistently boosts performance by 3.7, 2.1, 1.7, and 2.6 points at 2%, 4%, 8%, and 16% labels, respectively—validating that prompt-level consistency is an effective enhancer in varying scenarios.

Ablation Study and Analysis

Effectiveness of semantic prompt perturbation. In Table 4, using prompt perturbation alone (p^p) achieves comparable performance to employing image- and feature-level perturbations ($p^s + p^f$). Combining both regularization terms yields the best results.

Comparison of prompt perturbation strategies. As described in Section , we can also generate chaotic semantic prompts m^u from pseudo-masks of x^u. We conduct an ablation by feeding the student network separately with sampled m^l, m^u, and their combination ($m^l \,\&\, m^u$), while the teacher always receives the raw m^l. The results are shown in Figure 2. We observe that sampling from m^l, m^u, or both contributes equally to performance, with mIoU ranging only between 84.9% and 85.1%.

Qualitative comparisons. Figure 4 presents qualitative comparisons with state-of-the-art methods SemiVL and UniMatch v2. SSR-SAM consistently achieves finer boundary delineation, as highlighted by the bounding boxes.

Zero-shot capability. As reported in the Supplementary Materials, SSR-SAM demonstrates zero-shot segmentation ability on aerial images.

Conclusion

In this paper, we have presented SSR-SAM, a novel retrieval-style framework that leverages the SAM for semi-supervised segmentation on UHR images. By treating sparsely annotated patches as semantic prompts, SSR-SAM effectively propagates label information to semantically similar regions across predominantly unlabeled data. We further introduce prompt perturbation as a dedicated consistency regularizer at the prompt level. Through extensive evaluation on multiple real-world UHR benchmarks, we demonstrate that SSR-SAM not only achieves state-of-the-art performance with minimal annotations but also maintains strong generalization across diverse scenarios. In future work, we aim to extend SSR-SAM to broader domains.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 32341012, 62372170, 62172103).

References

Alonso, I.; et al. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *ICCV*, 8219–8228.

Amgad, M.; et al. 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467.

Bai, Y.; et al. 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In *CVPR*, 11514–11524.

Berthelot, D.; et al. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*.

Berthelot, D.; et al. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. arXiv:1911.09785.

Chen, H.; and Shi, Z. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote. Sens.*, 12: 1662.

Chen, K.; ; et al. 2024. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE TGRS*.

Chen, L.-C.; et al. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4): 834–848.

Chen, W.; et al. 2019a. Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-High Resolution Images. In *CVPR*.

Chen, W.; et al. 2019b. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *CVPR*, 8924–8933.

Chen, X.; et al. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2613–2622.

Cheng, H. K.; ; et al. 2020. CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement. In *CVPR*.

Dosovitskiy, A.; ; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. *NeurIPS*.

Guo, S.; et al. 2022. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *CVPR*, 4361–4370.

He, R.; Yang, J.; and Qi, X. 2021. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, 6930–6940.

Hoyer, L.; et al. 2025. Semivl: Semi-supervised semantic segmentation with vision-language guidance. In *ECCV*, 257–275. Springer.

Ji, D.; et al. 2023. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *CVPR*, 23621–23630.

Kirillov, A.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Kotaridis, I.; and Lazaridou, M. 2021. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173: 309–322.

Kumar, N.; et al. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7): 1550–1560.

Kwon, D.; and Kwak, S. 2022. Semi-supervised semantic segmentation with error localization network. In *CVPR*, 9957–9967.

Lai, X.; et al. 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 1205–1214.

Li, F.; ; et al. 2023. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.

Li, Q.; et al. 2021. From Contexts to Locality: Ultra-High Resolution Image Segmentation via Locality-Aware Contextual Correlation. In *ICCV*, 7252–7261.

Lingle, W.; et al. 2016. The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA).

Liu, S.; et al. 2021. Bootstrapping semantic segmentation with regional contrast. In *ICLR*.

Liu, S.; et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*.

Liu, W.; et al. 2024b. Ultra-high Resolution Image Segmentation via Locality-aware Context Fusion and Alternating Local Enhancement. *IJCV*, 132(11): 5030–5047.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam.

Luo, Y.; Chen, Z.; Zhou, S.; Hu, K.; and Gao, X. 2023. Self-distillation augmented masked autoencoders for histopathological image understanding. In *BIBM*, 1343–1349.

Ma, J.; ; et al. 2024. Segment anything in medical images. *Nat. Commun.*, 15(1): 654.

Madabhushi, A.; and Lee, G. 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33: 170–175.

Maggiori, E.; et al. 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE IGARSS*, 3226–3229.

Oquab, M.; et al. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Osco, L. P.; ; et al. 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *JAG*, 124: 103540.

Pham, H.; et al. 2021. Meta Pseudo Labels. In *CVPR*.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ravi, N.; et al. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Ren, T.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Shan, L.; and Wang, W. 2022. Mbnet: A multi-resolution branch network for semantic segmentation of ultra-high resolution images. In *ICASSP*, 2589–2593.

Shan, L.; et al. 2021. Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images. In *ICPR*, 1460–1466. IEEE.

Shen, T.; et al. 2022. High Quality Segmentation for Ultra High-Resolution Images. In *CVPR*, 1310–1319.

Sohn, K.; et al. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*.

Sun, B.; et al. 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *CVPR*, 3097–3107.

Sun, K.; Chen, Z.; Wang, G.; Liu, J.; Ye, X.; and Jiang, Y.-G. 2023. Bi-directional feature fusion generative adversarial network for ultra-high resolution pathological image virtual re-staining. In *CVPR*, 3904–3913.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*.

Wang, D.; ; et al. 2024. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In *NeurIPS*.

Wang, H.; et al. 2024. AllSpark: Reborn Labeled Features from Unlabeled in Transformer for Semi-Supervised Semantic Segmentation. In *CVPR*, 3627–3636.

Wang, S.; et al. 2025. Toward Real Ultra Image Segmentation: Leveraging Surrounding Context to Cultivate General Segmentation Model. *NeurIPS*, 37: 129227–129249.

Wu, G.; et al. 2015. Histological image segmentation using fast mean shift clustering method. *Biomedical engineering online*, 14: 1–12.

Wu, J.; et al. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Xie, E.; ; et al. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*.

Xie, E.; Lyu, J.; Wu, D.; Shen, H.; and Zhou, Y. 2025. Char-SAM: Turning Segment Anything Model into Scene Text Segmentation Annotator with Character-level Visual Prompts. In *ICASSP*, 1–5.

Xie, Q.; et al. 2019. Self-training with Noisy Student improves ImageNet classification. *Cornell University - arXiv,Cornell University - arXiv*.

Xie, Q.; et al. 2020. Unsupervised data augmentation for consistency training. In *NeurIPS*.

Xu, Z.; et al. 2021. Efficient transformer for remote sensing image segmentation. *Remote Sensing*, 13(18): 3585.

Yang, L.; Zhao, Z.; and Zhao, H. 2024. UniMatch V2: Pushing the Limit of Semi-Supervised Semantic Segmentation. *arXiv:2410.10777*.

Yang, L.; et al. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 7236–7246.

Yu, L.; et al. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer.

Yuan, X.; Shi, J.; and Gu, L. 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169: 114417.

Yun, S.; et al. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.

Zhang, J.; et al. 2023. Sam-path: A segment anything model for semantic segmentation in digital pathology. In *MICCAI*, 161–170. Springer.

Zhao, Z.; et al. 2023. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *CVPR*, 11350–11359.

Zhong, Y.; et al. 2021. Pixel contrastive-consistent semi-supervised semantic segmentation. In *ICCV*, 7273–7282.