AI-SAM: Automatic and Interactive Segment Anything Model

Yimu Pan, Sitao Zhang, Alison D. Gernand, Jeffery A. Goldstein, James Z. Wang

Abstract Semantic segmentation is a core task in computer vision, traditionally approached as either automatic or interactive. Interactive approaches, exemplified by the Segment Anything Model, have shown promise as pre-trained models, but current adaptation strategies tend to favor either automatic or interactive methods. Interactive approaches rely on user prompts, whereas automatic methods bypass interactive promptability entirely. We introduce the Automatic and Interactive Segment Anything Model (AI-SAM), a novel paradigm that addresses these limitations. At its core is the Automatic and Interactive Prompter (AI-Prompter), which automatically generates initial prompts while allowing user input. AI-SAM achieves state-of-the-art performance in both medical and non-medical applications and offers flexibility to further enhance results through user interaction. Code is available at https://github.com/ymp5078/AI-SAM.

Key words: Semantic segmentation, interactive model, medical applications.

1 Introduction

Pre-trained foundation models have gained prominence in various domains due to their effectiveness in tasks with limited annotations. In image segmentation, interactive models have attracted attention for their dual roles, as pre-training models and data annotation tools. A prime example in this category is the Segment Any-

Yimu Pan (✉) · Sitao Zhang · Alison D. Gernand
The Pennsylvania State University, University Park.
e-mail: ymp5078@psu.edu; sitao.zhang@psu.edu; adg14@psu.edu

Jeffery A. Goldstein
Northwestern University, Chicago. e-mail: ja.goldstein@northwestern.edu

James Z. Wang
The Pennsylvania State University, University Park. e-mail: jwang@psu.edu

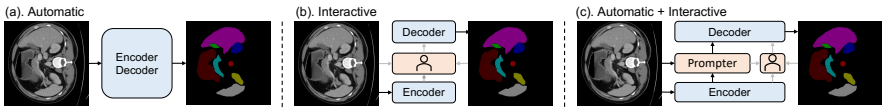


Fig. 1: Comparison of three approaches: (a) the automatic approach, (b) the interactive approach, and (c) a hybrid approach that combines elements of both automatic and interactive approaches. Black arrows indicate the automatic data flow, whereas grey arrows denote human intervention.

thing model (SAM) [23], which is acclaimed for its interactive capabilities. SAM is trained on the large-scale SA-1B dataset, which includes labels based on input prompts (e.g., points, bounding boxes, and text) along with corresponding output segmentation masks. Recent studies [31, 29, 19] have demonstrated the transferability of such models to diverse tasks. However, a limitation of SAMs is their inadequate understanding of class semantic granularity, often resulting in subpar performance in tasks with significant domain shifts unless prior adaptation is conducted. Further, they are trained with ambiguous targets where a single prompt may yield multiple masks, which may not correspond to the semantics required for downstream tasks (i.e., varying class semantic granularity).

To address these issues, some researchers have attempted to delineate semantic granularity by fine-tuning SAM on downstream datasets using synthetic prompts, a process termed "interactive adaptation." Others have used SAM merely as an initialization step for automatic semantic segmentation tasks. Despite these efforts, current approaches often compromise SAM's intrinsic capabilities by failing to achieve an optimal balance between automatic and interactive adaptation. Interactive adaptation may underperform in fully automatic settings, necessitating human intervention to finalize the segmentation map regardless of task complexity. Furthermore, while there are many possible prompts for interactive methods, their differences have not been comprehensively studied. On the other hand, automatic adaptation eliminates SAM's intrinsic promptability, rendering the adapted model unsuitable for interactive annotation or intervention. Notably, SAM's strengths are most evident in scenarios with limited or biased data, where iterative model refinement is common. In limited data settings, engineers iteratively label additional data in response to ongoing model performance. An integrated model that combines both automatic and interactive capabilities could significantly expedite this process, autonomously labeling simpler samples while gradually reducing human input for complex cases. Similarly, in bias data scenarios such as [10, 50], user intervention is critical for generating accurate results. The distinctions between approaches are illustrated in Fig. 1

In light of the limitations of current SAM adaptation methods and the potential of a hybrid automatic-interactive model, key research questions arise: (i) How can a model seamlessly combine automatic and interactive capabilities? (ii) What constitutes an effective prompt for such models?

To address these questions, we introduce a novel automatic and interactive segmentation paradigm. Within this framework, we analyze the characteristics of effec-

tive prompts and propose a new adaptation method: the Automatic and Interactive Segment Anything Model (AI-SAM). AI-SAM preserves the inherent promptability of SAM while achieving state-of-the-art (SOTA) performance in downstream automatic and interactive segmentation tasks. AI-SAM is designed to complement and enhance current SAM-based automatic adaptation models. Upon generating an initial automatic segmentation result, AI-SAM can incorporate additional user feedback (e.g., class labels, points, bounding boxes) to refine the outcome.

Our **main contributions** are as follows:

- *New segmentation paradigm:* We propose a new automatic and interactive segmentation paradigm and introduce the first model within this paradigm.
- *Functionality enhancement:* We design the first automated point prompt generation module and its corresponding specialized loss functions.
- *Empirical validation:* We performed comprehensive quantitative and qualitative experiments on various datasets to evaluation the effectiveness of our method.

2 Related Works

Automatic Semantic Segmentation. The field of semantic segmentation has experienced growth and transformation. Earlier approaches [28, 34, 47, 16, 1, 5, 6, 7, 8, 53, 51] primarily relied on convolutional neural networks. Subsequent innovations incorporated attention modules [48, 18, 54]. Reflecting broader trends in computer vision, recent developments have pivoted towards transformer-based models [52, 44, 36, 12, 11], often leveraging large-scale pre-trained models [42, 43, 37, 17]. These methods are fully automated in nature.

Adaptation of Interactive Segmentation Models. Interactive segmentation models [26, 25, 23] generate segmentation maps based on user-provided prompts. These models are designed to be generalized to a wide range of tasks. Unlike traditional automatic segmentation models, which are constrained to predefined classes, interactive models, in theory, can handle any class. SAM [23] has emerged as a prominent interactive model due to its exceptional performance across multiple domains. Subsequent research has focused on enhancing SAM's interactive segmentation capability in specific domains through fine-tuning with synthetic prompts and segmentation pairs [29, 14, 22, 15]. Additionally, other studies have extended SAM to automatic segmentation, either through workarounds [24, 27] or automatic adaptation methods [49, 9, 13, 30, 3, 45, 20, 35]. However, a gap remains in effectively adapting interactive segmentation models to an automatic segmentation setting while retaining their inherent promptability, which is essential for maintaining flexibility and usability.

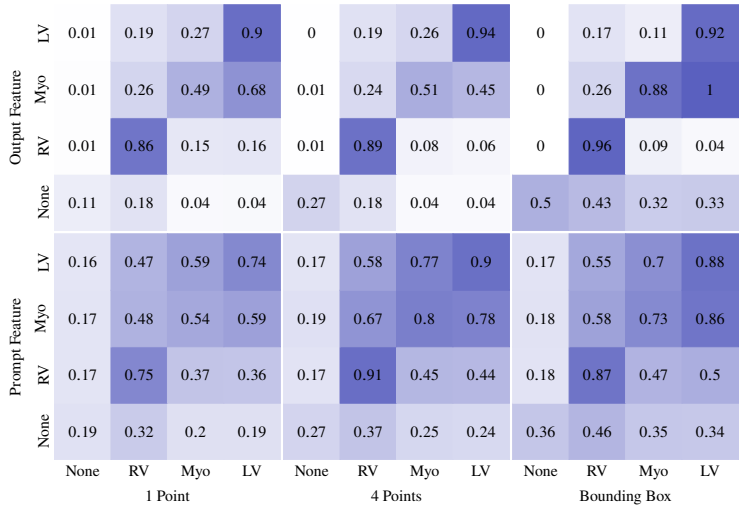


Fig. 2: The Output Confusion Metrices (top row) and the Prompt Confusion Metrices (bottom row) generated using SAM feature on the ACDC dataset. Values are normalized to a 0–1 scale and averaged over the entire dataset.

3 Preliminary

Despite the variety of prompts that SAM accepts, their respective abilities to generate effective segmentation results remain underexplored. Given that the efficacy of interactive segmentation methods is intrinsically linked to prompt quality, it is essential to systematically analyze various prompts. In this section, we establish a theoretical framework for evaluating the error behaviors of different prompt types and demonstrate an example application of the framework to SAM. Our focus is on bounding box and point prompts, as they are the most commonly utilized visual prompts in this context. The goal of this section is to evaluate different prompts and identify the most suitable type for our setting.

Confusion Matrix as the Framework. Visual prompts facilitate the model's differentiation of intended and unintended behaviors, functioning analogously to classifiers. For a prompt to be effective, it must clearly separate class semantics. Inspired by the conventional confusion matrix used in classification, we introduce the concept of a **Prompt Confusion Matrix** (PCM). In the PCM, True Semantic Similarity (TSS) occupies the diagonal, representing correct semantic interpretation, while False Semantic Similarity (FSS) resides off-diagonal, indicating semantic misinterpretations. Further, because the precision of the generated segmentation also requires evaluation in order to compare the effect of prompts on the output, we quantify it with the **Output Confusion Matrix** (OCM). Here, True Output Similarity (TOS) is placed on the diagonal and False Output Similarity (FOS) on the off-diagonal. A functional PCM correlates with the OCM, under the assumption that clearer semantic separation via

prompts produces better segmentation results. Both PCM and OCM function like conventional confusion matrices, where high diagonal values and low off-diagonal values are desirable. While other metrics derived from the confusion matrix might be simpler for benchmarking, they often overlook certain error patterns. Thus, our proposed PCM and OCM are more suitable as evaluation tools. Various implementations of PCM and OCM based on different architectures still adhere to the core concept of the confusion matrix.

Implementation of PCM and OCM. If a prompt separates image features into distinct groups effectively, differences between these group features should be evident in the PCM. We define the image feature X_i of class i and prompt features P_j of class j, where X_i comprises all encoded image patches with location information within the segmentation mask of class i, and P_j is an aggregated representation of the prompt. Each point prompt P_j is represented by its nearest image patch, and each bounding box prompt by averaging the image patches within the box; this averaging approach is adopted because SAM represents an entire box as a single prompt. We define semantic similarity $s_{i,j} = \text{mean}_{x_i \in X_i}(\text{sim}(x_i, p_{x_i}^{\max}))$ as the mean of the highest similarities between each image feature x_i and its most similar prompt feature $p_{x_i}^{\max} = \max_{p_j \in P_j}(\text{sim}(x_i, p_j))$, where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. Additionally, output similarity is defined as the overlap ratio between the segmentation masks of two classes, closely mirroring the segmentation model's performance. Due to the significant impact of different implementations on results, our implementation should not serve as a metric to quantify prompt effectiveness; it is solely used to reveal the error behaviors of different prompt types. Based on our implementation, we propose the following propositions: (1) PCM correlates with OCM. (2) For the same model, adding the same type of prompts cannot decrease TSS/FSS.

We empirically validate Proposition 1 with the Automated Cardiac Diagnosis Challenge (ACDC) dataset. Point prompts are randomly sampled from the ground truth segmentation map, and the box prompts are the tightest bounding boxes around the ground truth segmentation map. From Fig. 2, we observed a correlation between PCM and OCM, supporting Proposition 1. Proposition 2 is also evident since adding points enhances TSS but not FSS, as shown by the PCM's transition from one to four points in Fig. 2. Proposition 2 is proved by contradiction: assuming adding a prompt $\hat{p_j}$ decreases TSS/FSS. To change the TSS/FSS for any x_i, $\max_{p_j \in P_j}(\text{sim}(x_i, p_j)) = \text{sim}(x_i, \hat{p_j}) > \text{sim}(x_i, p_j)$, which would actually raise TSS/FSS, contradicting the assumption.

Based on Proposition 1, achieving effective segmentation results requires high TSS and low FSS. Proposition 2 suggests that while we cannot reduce FSS by adding prompts, we can enhance TSS. Therefore, point prompts, with their lower FSS, appear to be a more viable prompt type. Additionally, we note that the prompt semantics for class Myo are often similar to those for class LV, leading to confusion in the OCM—particularly with box prompts. Interestingly, this observation aligns with SAM's prompt design, where either only point prompts are used or background point prompts are allowed in addition to the box prompt to reduce confusion; however, box prompts are never used in isolation. Based on these insights, our method will exclusively use point prompts for simplicity.

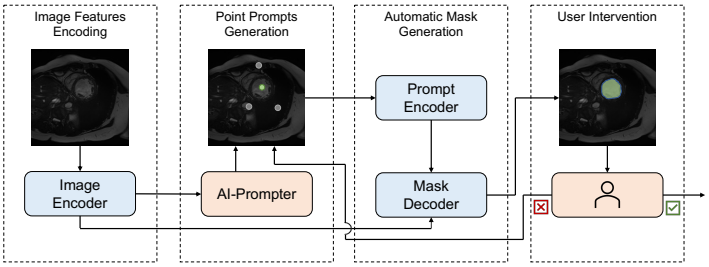


Fig. 3: The pipeline of AI-SAM, showing the interaction between its modules. Users can review segmentation results and adjust prompts as needed. The green point represents the foreground point, while the gray points indicate the background points.

4 Method

We introduce AI-SAM, which comprises an Automatic and Interactive Prompter (AI-Prompter) for automatic prompt generation, heuristic-based prompt loss functions, a classifier to exclude prompts from non-existing classes, and a SAM-based interactive segmentation model. The AI-Prompter, steered by the prompt loss functions, is trained to generate a set of easily modifiable and correctable point prompts for a given image and target class. We detail the overall architecture of AI-SAM and the properties that guide the AI-Prompter training process in the following subsections. Details of our custom classifier, which leverages encoded image features, are provided in the Appendix.

Automatic and Interactive Segmentation. Fig. 1 illustrates the differences among common segmentation approaches. Automatic segmentation models autonomously produce segmentation masks for predefined classes based solely on an input image. By contrast, interactive segmentation models require both an input image and user-provided prompts to generate the corresponding segmentation mask. The distinction primarily lies in how they handle class semantics: automatic segmentation integrates class semantics into mask generation, whereas interactive segmentation relies solely on the semantics introduced through user prompts, regardless of inherent class semantics. Therefore, a model that integrates both automatic and interactive features should bridge class semantics with prompt semantics. The essence of the AI-SAM is its ability to simulate the human prompting process for each predefined class during automatic training, thereby preserving the model's awareness of prompts. The AI-Prompter, crafted to emulate human-like prompt generation, processes image features and generates point prompts for the mask decoder during training. At inference time, AI-SAM independently generates prompts and corresponding segmentation masks for predefined classes, while also accommodating user adjustments—such as adding or removing points—to refine segmentation quality. AI-SAM is designed for end-to-end training, utilizing ground truth segmentation masks as targets. Its architecture is depicted in Fig. 3.

Automatic and Interactive Prompter. Despite the fact that point prompts have been shown to cause less confusion, as demonstrated in Sec. 3, their potential for interactive adaptation remains largely untapped, possibly due to two main issues. First, there is no clear consensus on what constitutes an "optimal" point. Unlike bounding boxes, where the standard is the tightest box encompassing the object, no comparable consensus exists for point prompts. Prior research in different domains [40, 38, 10] has typically regressed points using a Gaussian distribution around the true location, but such ground truth points are not available in our context, requiring a data-driven approach to infer the optimal points. Furthermore, this method lacks contextual richness, making the learning process challenging.

To overcome this, we draw inspiration from the heatmap representation [39] in the field of pose estimation and propose the concept of a "generalized point." Given a point embedding p_i on an image (i.e., the positional embedding), we define generalized point representation as the weighted sum of these positional embeddings, represented as $P = [p_0, ...p_I]$ using weights $W = [w_0, ..., w_I]$, where I is the number of positional embeddings in an image. That is, $P^{\text{g}} = W^\top P$. As the base interactive model provides the representation for positional embeddings, our task is reduced to modeling the generation of W. To differentiate from the traditional encoding of points, we refer to the conventional representation as a one-hot point, which is a special case of the generalized point with W being one-hot.

We thus propose the AI-Prompter, which processes image feature X and produces weights $W = \text{AIPrompter}(X, c)$, where c denotes a specific class. The differentiable nature of P^{g} allows it to be fed directly into the prompt encoder as $P^{\text{g}} = \text{AIPrompter}(X, c)^\top P$, facilitating the application of any automatic adaptation method. The entire model is trained end-to-end, using only the ground truth segmentation masks as targets. The AI-Prompter utilizes a customized architecture described in the Appendix.

Prompt Heuristic Loss. As both an automatic and interactive model, AI-SAM is designed with a focus on usability as well as performance. There is no assurance that the generated generalized points will be accurately positioned within the object of interest. Neural networks may exploit shortcuts, relying on class representation instead of precise point prompts to adapt automatically. If the generated points are linked to class representation without correctly attending to the object's location, usability is entirely compromised. Therefore, ensuring that produced points are correctly associated with the target object location is vital.

Moreover, although the proposed generalized point representation simplifies training, modifying these points can pose challenges due to the disparity between one-hot points and generalized points. Users typically find modifying one-hot points intuitive, as it involves directly clicking on an image. However, modifying a generalized point, which may correspond to multiple parts of the image, can be as complex as altering the segmentation mask itself. Hence, producing generalized points that closely resemble the simplicity and directness of one-hot points is equally crucial.

To address these issues, we introduce two heuristics based on the annotators' intuitions: (1) P^{g} should be situated within the target object. (2) P^{g} should closely resemble a one-hot point prompt. These heuristics lead to the definition of two loss

functions, the point correctness loss $\mathcal{L}^{\text{pc}} = \text{mean}_c \mathcal{L}_c^{\text{pc}}$ and the point sharpness loss $\mathcal{L}^{\text{ps}} = \text{mean}_c \mathcal{L}_c^{\text{ps}}$, as follow:

$$\mathcal{L}_c^{\text{pc}} = 1 - \frac{\mathbf{1}_c^\top W + \gamma}{\mathbf{1}^\top W + \gamma}, \quad \mathcal{L}_c^{\text{ps}} = 1 - \frac{\max(\mathbf{1}_c \odot W) + \gamma}{\mathbf{1}_c^\top W + \gamma}, \tag{1}$$

where $\mathbf{1}_c$ is an indicator function with its ith element being 1 if the image feature x_i belongs to class c and 0 otherwise, $\mathbf{1}$ is a matrix of ones the same dimension as W, $\odot$ represents element-wise multiplication, and γ is a hyper-parameter. $\mathcal{L}_c^{\text{pc}}$ penalizes the difference between the total weight $\mathbf{1}^\top W$ and the correct weight $\mathbf{1}_c^\top W$, thus encourage correct P^{g}. $\mathcal{L}_c^{\text{ps}}$ penalizes the difference between the correct weight $\mathbf{1}_c^\top W$ and the maximum correct weight $\max(\mathbf{1}_c \odot W)$, thus encourage one-hot points. The combination of $\mathcal{L}^{\text{pc}}$ and $\mathcal{L}^{\text{ps}}$ ensures the final P^{g} is both inside the object and closely aligned with a one-hot point.

Without regularization, the aforementioned loss would cause all points to converge to the same optimal point, making it impossible to control the number of points. As a generalized point can encapsulate information from many locations, multiple points might seem redundant. However, to enhance usability, the generalized points are encouraged to be one-hot, thereby precluding them from representing information from multiple locations. More importantly, representing all prompt information in one generalized point would require users to modify the entire class prompt simultaneously; distributing the information across multiple prompts allows for more fine-grained control. Therefore, we need a heuristic for diverse point prompt.

Based on Proposition 1, which states that adding points can improve performance, we aim to encourage the model to produce multiple points with high TSS and low FSS. To achieve these goals, we introduce the prompt diversity loss

$$\mathcal{L}^{\text{pd}} = \beta^{\text{in}} \mathcal{L}^{\text{in}} + \beta^{\text{out}} \mathcal{L}^{\text{out}}, \tag{2}$$

consisting of an inter-class diversity loss $\mathcal{L}^{\text{in}} = \text{mean}_c \mathcal{L}_c^{\text{in}}$ and an intra-class diversity loss $\mathcal{L}^{\text{out}} = \text{mean}_n \mathcal{L}_n^{\text{out}}$. We utilize $\hat{P}^{\text{g}} = W^\top (P + X)$ as the point feature, considering both image features and positional embeddings as important for distinguishing optimal points. These loss components are defined as follows:

$$\mathcal{L}_c^{\text{in}} = -\log \frac{\exp(1/\tau)}{\sum_n \exp(\text{sim}(\hat{p}^{\text{g}}_{n^+}, w_n)/\tau)}, \ \mathcal{L}_n^{\text{out}} = -\log \frac{\exp(1/\tau)}{\sum_c \exp(\text{sim}(\hat{p}^{\text{g}}_{c^+}, w_c)/\tau)}, \tag{3}$$

where c^+ and n^+ denote the class index and prompt index of the anchor point, respectively. $\mathcal{L}^{\text{in}}$ penalizes similarity among points within the same class, contributing to TSS, as similar points do not improve TSS. Conversely, $\mathcal{L}^{\text{in}}$ penalizes similarity between points of different classes, contributing to FSS, as incorrect points can significantly increase FSS. The final prompt heuristic loss is defined as:

$$\mathcal{L}^{\text{ph}} = \alpha^{\text{pc}} \mathcal{L}^{\text{pc}} + \alpha^{\text{ps}} \mathcal{L}^{\text{ps}} + \alpha^{\text{pd}} \mathcal{L}^{\text{pd}}, \tag{4}$$

where α^{pc}, α^{ps}, and α^{pd} represent the weights assigned to each loss components.

	Method	Synapse										ACDC			
		DICE ↑	HD95↓	Aorta	GB	KL	KR	Liver	PC	SP	SM	DICE ↑	RV	Myo	LV
Automatic	R50+UNet [4]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92	87.55	87.10	80.63	94.92
	R50+AttnUNet [4]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95	86.75	87.58	79.20	93.47
	TransUNet [4]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	89.71	88.86	84.53	95.73
	SwinUNet [2]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	90.00	88.55	85.62	95.83
	MT-UNet [41]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81	90.43	86.64	89.04	95.62
	MISSFormer [21]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81	90.86	89.55	88.04	94.99
	CASTformer [46]	82.55	22.73	**89.05**	67.48	86.05	82.17	**95.61**	67.49	91.00	81.55	-	-	-	-
	PVT-CASCADE [32]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69	91.46	88.9	89.97	95.50
	TransCASCADE [32]	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52	91.63	89.14	**90.25**	95.50
	Parallel MERIT [33]	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15	**92.32**	**90.87**	90.00	**96.08**
	Cascaded MERIT [33]	**84.90**	13.22	87.71	74.40	**87.79**	84.85	95.26	71.81	**92.01**	85.38	91.85	90.23	89.53	95.80
	SAMed_h [49]	84.30	16.02	87.81	**74.72**	85.76	81.52	95.76	70.63	90.46	**87.77**	-	-	-	-
	AI-SAM	84.21	**12.11**	88.89	74.53	86.56	**85.01**	96.30	**72.84**	90.32	79.24	92.06	90.18	89.94	96.05
Interactive	SAM* gt box	90.16	3.27	92.47	**90.82**	91.40	90.71	92.44	76.87	95.88	**90.70**	79.56	88.15	57.05	93.49
	MedSAM* gt box	88.82	2.41	90.15	82.97	90.95	89.76	95.74	76.33	94.55	90.14	67.95	92.22	16.50	95.14
	AI-SAM gt label	87.56	10.14	91.51	83.78	89.48	87.34	96.51	76.96	94.62	80.28	93.02	92.58	90.21	96.26
	AI-SAM 1 rd pt	87.91	6.78	90.26	83.70	89.85	88.49	96.52	77.16	95.32	81.97	93.04	92.58	90.26	96.28
	AI-SAM gt box	**90.66**	**1.73**	**95.03**	85.20	**93.40**	**92.13**	**96.76**	**81.79**	**96.27**	84.73	**93.89**	**94.13**	**90.95**	**96.59**

Table 1: Comparison of AI-SAM to SOTA methods on multiple datasets. Results of previous work produced by us are noted with an *. All the metrics are detailed in the Appendix. ↑: higher is better, ↓: lower is better. Best results are highlighted in bold. gt box: the tightest bounding box from the ground truth segmentation. 1 rd pt: one randomly sampled point from the ground truth segmentation.

5 Experiments

To validate our approach, we conducted extensive experiments using multiple public-domain datasets and compared AI-SAM with other state-of-the-art (SOTA) models.
Dataset. The Synapse Multi-organ Dataset[1] contains 30 abdominal CT scans labeled for 8 organs: aorta, gallbladder (GB), left kidney (KL), right kidney (KR), liver, pancreas (PC), spleen (SP), and stomach (SM). Following [4, 32], 12 scans are used for testing and the rest for training. The Automated Cardiac Diagnosis Challenge (ACDC)[2] dataset includes 100 cardiac MRI scans labeled with left ventricle (LV), right ventricle (RV), and myocardium (Myo), with 20 scans for testing and the remainder for training. Prepossessing steps and evaluation metrics for each dataset follow the corresponding prior work and are detailed in the Appendix.

5.1 AI-SAM in Medical Image Segmentation

We evaluate our model in automatic semantic segmentation tasks and integrate AI-Prompter with other SAM-based automatic adaptation methods, demonstrating its ability to enhance these methods without performance loss. We also assess interactive segmentation capabilities by incorporating additional guidance during inference.

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789/

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

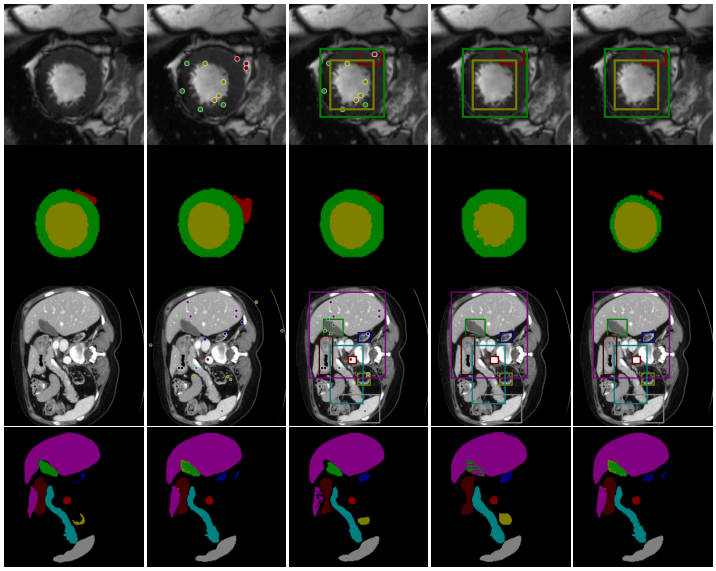
An ablation study validates key assumptions about the AI-Prompter, while qualitative results highlight the effectiveness of AI-Prompter and AI-SAM. Due to space constraints, model implementation details and results on non-medical images are provided in the Appendix. In the automatic setting, only the image is used as input, whereas the interactive setting includes synthesized user inputs generated from the ground truth segmentation map to guide the model.

Automatic Medical Image Segmentation. As shown in Table 1, in the automatic setting, our method achieves SOTA performance on both datasets. Additionally, we include another SAM-based adaptation method, SAM_d, which performs comparably to our approach despite its lack of promptability. These results underscore the effectiveness of the automatic and interactive adaptation paradigm, even when only the automatic function is active.

Interactive Medical Image Segmentation. Our method shows significant performance improvement in interactive use. Providing class labels for each image notably boosts performance (Table 1), indicating that many errors stem from classifier misclassifications, which may underestimate AI-Prompter's accuracy in automatic settings. Users can interact with AI-SAM, similar to the original SAM, by providing points and bounding boxes. To assess additional point prompts, we randomly sample points from the ground truth mask, observing modest gains. However, in real-world use, human-provided points would likely yield greater improvements, especially when correcting segmentation errors. For user-provided bounding boxes, we use the tightest box around the segmentation mask. Since this bounding box is consistent across models, we compare MedSAM (current SOTA), SAM, and AI-SAM. AI-SAM uniquely constrains AI-Prompter's learned weights W to ensure point prompts fall within the box. As shown in Table 1, AI-SAM achieves SOTA performance in interactive evaluation, despite MedSAM's access to a much larger training set.

Qualitative Evaluation of AI-SAM. We visualize the generated points and segmentation masks before and after adding a bounding box. We focus on cases where AI-SAM performs poorly for at least one class in the automatic setting to show how interactive use improves the results. From the medical image segmentation examples in Fig.4, we observe that the red class in ACDC and the yellow class in Synapse are missed in the automatic setting but correctly identified when points are refined by the bounding box. However, providing the same bounding box to SAM or MedSAM introduces more errors. SAM's class-agnostic prompts often segment unintended objects, especially when smaller classes or unintended semantics (from pre-training data) are within the bounding box. SAM also tends to favor larger regions, as seen in the bottom row of Fig.4, where the smaller part (pink) is ignored.

Interactive adaptation mitigates this issue, as shown in MedSAM's GT Box result (pink), but can introduce new errors (yellow). Notably, performance for the Myo (green) class in ACDC remains low for both SAM and MedSAM. Fig.4 shows frequent overlap between Myo (green) and LV (yellow) bounding boxes, with Myo often misclassified as LV. This can be explained using the proposed SCM: in the last SCM of Table2, the bounding box prompt for Myo (second row) has higher FSS than LV (first row), making Myo prone to misclassification as LV.



(a) Ground Truth (b) AI-SAM Automatic (c) AI-SAM GT Box (d) SAM GT Box (e) MedSAM GT Box

Fig. 4: Qualitative results of the AI-Prompter on ACDC using 4 points. The odd rows are the generated prompts, and the even rows are the segmentation maps. Samples were selected where AI-SAM's automatic mode performed poorly on certain classes to demonstrate how the interactive version rectifies these errors. The images are zoomed in for improved visibility. The color palette is in the Appendix.

# Points	1	2	4	8	16
DICE	92.17	92.25	92.06	92.05	91.77

Table 2: Ablation study on the number of point prompts per class. The original SAM uses up to 16 points, equivalent to 4 points per class in this setting.

5.2 Ablation Study

Number of Points. To investigate how the number of points affects the model's performance, we conducted experiments using different numbers of points on the ACDC dataset. As shown in Table 2, the performance remains relatively consistent. However, as we use too many points, the performance begins to degrade. Thus, the number of points is has minimal effect if we use a reasonable number of points.

Prompt Heuristic Loss. Our loss design aims to align generated prompts with the intuitions of usability (Eq. 1) while maintaining strong model performance. To evaluate the effectiveness of the Prompt Heuristic Loss design, we employ both qualitative and quantitative criteria for assessment.

Point Type	P^g	P	Δ
AI-SAM	92.06	92.06	0.0
w/o $\mathcal{L}^{pd}$	92.12	92.12	0.0
w/o $\mathcal{L}^{pd}$, $\mathcal{L}^{ps}$	91.99	91.98	0.01
w/o $\mathcal{L}^{ph}$	91.75	50.43	41.32

Table 3: Ablation Study on ACDC using AI-Prompter with 4 points. Mean DICE scores (%) are reported. Δ is the change in performance when switching point type from generalized point P^g to one-hot point P.

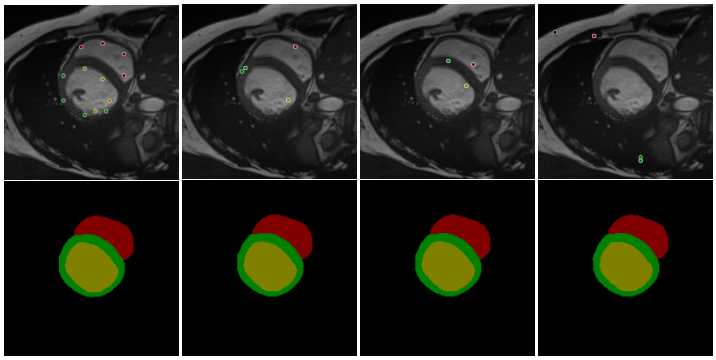
In qualitative evaluation, we visually inspect the generated prompts to ensure that the points are located on the object of interest and that the correct number of points are generated. For quantitative evaluation, we assess the impact of the proposed losses on model performance by conducting ablation studies. Specifically, we investigate the effects of removing individual components of the loss design and transitioning from generalized points to one-hot points.

Our ablation studies include the following scenarios: removing the diversity loss $\mathcal{L}^{pd}$, removing both the diversity loss and sharpness loss $\mathcal{L}^{ps}$, and removing the entire prompt heuristic loss $\mathcal{L}^{ph}$. The results presented in Table 3 highlight the significance of these loss functions. Notably, removing the diversity loss $\mathcal{L}^{pd}$ does not significantly affect the model's ability to utilize one-hot points. However, eliminating the entire prompt heuristic loss $\mathcal{L}^{ph}$ leads to a substantial drop in performance (i.e., 41.32), underscoring the critical role of the prompt heuristic loss for usability.

When comparing the results of AI-SAM with and without $\mathcal{L}^{ph}$ (92.12 vs. the results in Fig. 2), it is likely that the model produces fewer points as its performance falls within that range. To further examine the effect of the diversity loss, we visualized the generated points in Fig. 5. When the complete prompt heuristic loss is applied, we obtain four points for each class, and these points align with human intuition. However, upon removing the diversity loss, we observe that almost all points converge to the same location, which is consistent with our intuition in Eq. 1 that encourages points to be similar, as supported by the qualitative results in Fig. 5b.

Furthermore, if we additionally eliminate the point sparsity loss, the model not only exhibits behavior similar to that when the diversity loss is removed (as shown in Fig. 5c) but also performs worse. Furthermore, we start to see differences between generalized points and one-hot points (as indicated in Table 3). Lastly, when the entire prompt heuristic loss is removed, the model not only produces fewer than four points for each class but also generates many incorrect points (as shown in Fig. 5d), which aligns with our intuition that a neural network can learn class representation instead point locations without proper supervision. These ablation results collectively underscore the effectiveness of each proposed loss in serving its intended purpose.

Prompt Quality. While we have demonstrated that adding a point sampled from the ground truth segmentation mask can enhance model performance (as shown in Table 1) and visualized the points generated by AI-Prompter (as shown in Fig. 5), the superiority of the AI-Prompter-generated points over those sampled from the



(a) AI-Prompter (b) w/o $\mathcal{L}^{pd}$ (c) w/o $\mathcal{L}^{pd}\,\mathcal{L}^{ps}$ (d) w/o $\mathcal{L}^{ph}$

Fig. 5: Qualitative results of the AI-Prompter on ACDC using 4 points. The first row is the generated prompt and the second row is the segmentation map. Images are zoomed in for improved visibility. The color palette is in the Appendix.

ground truth segmentation mask remains to be established. Although the proposed PCM could potentially assess prompt quality, our current implementation, as outlined in Sec. 3, does not fulfill this purpose. To empirically validate the quality of the points generated by AI-Prompter, we conducted an experiment where we replaced the AI-Prompter-generated points with points sampled from the ground truth on ACDC. The result was a significant decrease in the DICE score, from 92.06 to 64.05, indicating the effectiveness of AI-Prompter in generating high-quality prompts.

6 Discussion and Conclusion

The limitations and broader implications of our work are discussed in the Appendix. We introduced AI-SAM, a novel paradigm that bridges the gap between automatic and interactive segmentation. We analyzed different prompt types and proposed a method to generate effective prompts. This unified framework not only offers a new approach to segmentation but also holds significant promise in real-time medical imaging applications. We anticipate that AI-SAM will inspire further advancements in the field.

Acknowledgements Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under award number R01EB030130. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work used computing resources at the National Center for Supercomputing Applications through allocation IRI180002 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants Nos. 2138259, 2138286, 2138307, 2137603, and 2138296.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-UNet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 205–218. Springer (2022)
3. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Medical Image Analysis **98**, 103310 (2024)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 801–818 (2018)
9. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: SAM-Adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3367–3375 (2023)
10. Chen, Y., Zhang, Z., Wu, C., Davaasuren, D., Goldstein, J.A., Gernand, A.D., Wang, J.Z.: AI-PLAX: AI-based placental assessment and examination using photos. Computerized Medical Imaging and Graphics **84**, 101744 (2020)
11. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1290–1299 (2022)
13. Cui, C., Deng, R., Liu, Q., Yao, T., Bao, S., Remedios, L.W., Landman, B.A., Tang, Y., Huo, Y.: All-in-SAM: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. In: Journal of Physics: Conference Series, vol. 2722(1), p. 012012. IOP Publishing (2024)
14. Dai, H., Ma, C., Liu, Z., Li, Y., Shu, P., Wei, X., Zhao, L., Wu, Z., Zhu, D., Liu, W., et al.: SAMAug: Point prompt augmentation for segment anything model. arXiv preprint arXiv:2307.01187 (2023)
15. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al.: Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
16. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9716–9725 (2021)
17. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19358–19369 (2023)

18. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
19. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7545–7556 (2023)
20. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (SAM) to medical images. arXiv preprint arXiv:2306.13731 (2023)
21. Huang, X., Deng, Z., Li, D., Yuan, X.: MISSformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
22. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36** (2024)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
24. Lei, W., Xu, W., Li, K., Zhang, X., Zhang, S.: MedLSAM: Localize and segment anything model for 3D medical images. Medical Image Analysis **99**, 103370 (2025)
25. Lin, J., Chen, J., Yang, K., Roitberg, A., Li, S., Li, Z., Li, S.: AdaptiveClick: Clicks-aware transformer with adaptive focal loss for interactive image segmentation. arXiv preprint arXiv:2305.04276 (2023)
26. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: SimpleClick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22290–22300 (2023)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
29. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
30. Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M.: AdaptiveSAM: Towards efficient tuning of SAM for surgical scene segmentation. In: Annual Conference on Medical Image Understanding and Analysis, pp. 187–201. Springer (2024)
31. Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., Dong, Y., Lam, K., Lo, F.P.W., Xiao, B., et al.: Large ai models in health informatics: Applications, challenges, and the future. IEEE Journal of Biomedical and Health Informatics (2023)
32. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6222–6231 (2023)
33. Rahman, M.M., Marculescu, R.: Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In: Medical Imaging with Deep Learning (2023)
34. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
35. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: AutoSAM: Adapting SAM to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
36. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021)
37. Su, W., Zhu, X., Tao, C., Lu, L., Li, B., Huang, G., Qiao, Y., Wang, X., Zhou, J., Dai, J.: Towards all-in-one pre-training via maximizing multi-modal mutual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15888–15899 (2023)

38. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision, pp. 529–545 (2018)
39. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Advances in Neural Information Processing Systems **27** (2014)
40. Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
41. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2390–2394. IEEE (2022)
42. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: ONE-PEACE: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
43. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: InternImage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14408–14419 (2023)
44. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
45. Xiong, X., Wang, C., Li, W., Li, G.: Mammo-SAM: Adapting foundation segment anything model for automatic breast mass segmentation in whole mammograms. In: International Workshop on Machine Learning in Medical Imaging, pp. 176–185. Springer (2023)
46. You, C., Zhao, R., Liu, F., Dong, S., Chinchali, S., Topcu, U., Staib, L., Duncan, J.: Class-aware adversarial transformers for medical image segmentation. Advances in Neural Information Processing Systems **35**, 29582–29596 (2022)
47. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 325–341 (2018)
48. Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E.: ACFnet: Attentional class feature network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6798–6807 (2019)
49. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
50. Zhang, Z., Davaasuren, D., Wu, C., Goldstein, J.A., Gernand, A.D., Wang, J.Z.: Multi-region saliency-aware learning for cross-domain placenta image segmentation. Pattern Recognition Letters **140**, 165–171 (2020)
51. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
52. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
53. Zhu, L., Ji, D., Zhu, S., Gan, W., Wu, W., Yan, J.: Learning statistical texture for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12537–12546 (2021)
54. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 593–602 (2019)