GlandSAM: Injecting Morphology Knowledge Into Segment Anything Model for Label-Free Gland Segmentation

Qixiang Zhang, Yi Li, Cheng Xue, Haonan Wang, and Xiaomeng Li, *Member, IEEE*

***Abstract*—This paper presents a label-free gland segmentation, GlandSAM, which achieves comparable performance with supervised methods while no label is required during its training or inference phase. We observe that the Segment Anything model produces sub-optimal results on gland dataset: It either over-segments a gland into many fractions or under-segments the gland regions by confusing many of them with the background, due to the complex morphology of glands and lack of sufficient labels. To address this challenge, our GlandSAM innovatively injects two clues about gland morphology into SAM to guide the segmentation process: (1) *Heterogeneity within glands* and (2) *Similarity with the background*. Initially, we leverage the clues to decompose the intricate glands by selectively extracting a proposal for each gland sub-region of heterogeneous appearances. Then, we inject the morphology clues into SAM in a fine-tuning manner with a novel morphology-aware semantic grouping module that explicitly groups the high-level semantics of gland sub-regions. In this way, our GlandSAM could capture comprehensive knowledge about gland morphology, and produce well-delineated and complete segmentation results. Extensive experiments conducted on the GlaS dataset and the CRAG dataset reveal that GlandSAM outperforms state-of-the-art label-free methods by a significant margin. Notably, our GlandSAM even surpasses several fully-supervised methods that require pixel-wise labels for training, which highlights the remarkable performance and potential of GlandSAM in the realm of gland segmentation.**

Received 17 June 2024; revised 7 September 2024; accepted 1 October 2024. Date of publication 8 October 2024; date of current version 4 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62306254, in part by the Hong Kong Innovation and Technology Fund under Project ITS/030/21, and in part by the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under Grant HZQB-KCZYB-2020083. *(Corresponding author: Xiaomeng Li.)*

Qixiang Zhang, Yi Li, and Haonan Wang are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: ericZhang5915@gmail.com; yili7eli@gmail.com; hwanggr@connect.ust.hk).

Cheng Xue is with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: xchengjlu@gmail.com).

Xiaomeng Li is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China, and also with the HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen 518055, China (e-mail: eexmli@ust.hk).

Digital Object Identifier 10.1109/TMI.2024.3476176

***Index Terms*—Whole slide image, label-free gland segmentation, segment anything model.**

I. Introduction

LABEL-FREE gland segmentation refers to the process of segmenting a whole slide image (WSI) into a glandular region and a non-glandular region without the use of any kind of labels during the training and inference phases. Despite the success of the existing studies on fully supervised gland segmentation [1], [2], [3], [4], [5], [6] and weakly supervised gland segmentation [7], [8], these approaches necessitate pixel-level or weaker labels, *e.g.*, bound-box and patch tag. However, the manual labeling of WSIs remains a considerable challenge due to their extensive scale [9]. Specifically, it usually takes months for a pathology expert to draw pixel-level labels for one WSI at the resolution of $50,000 \times 50,000$ [9], while weaker forms of annotations, *e.g.*, bound box, still cost more than three weeks [8]. To tackle this challenge, in this paper, we propose the first work of label-free gland segmentation, which enables training and inference without relying on any explicit labels.

A simple way to achieve label-free gland segmentation is to adopt prior label-free methods in the field of computer vision [11], [12], [13], [14], [15], [16], [17]. However, adapting these methods to gland segmentation results in poor performance due to the intrigue gland morphology and lack of annotations [10]. Recently, the emergence of the Segment Anything Model (SAM) has demonstrated extraordinary label-free generalizability to different scenarios [11]. With only a few visual prompts, *e.g.*, points, boxes, and scribbles, SAM has achieved robust performance in various segmentation tasks [18], [19], [20], which makes it an attractive choice for developing a label-free gland segmentation method. However, based on our observation, when performing zero-shot gland segmentation with these visual prompts, SAM demonstrates inferior performance: **It either over-segments a gland into numerous fractions or under-segments the gland regions by misclassifying many of them as the background**; see Fig. 1 (b). The reason behind this is the heavy reliance of SAM's zero-shot performance on the inherent connections between pixels of the same object [21], and essentially, grouping similar pixels and separate dissimilar ones. In natural

1558-254X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

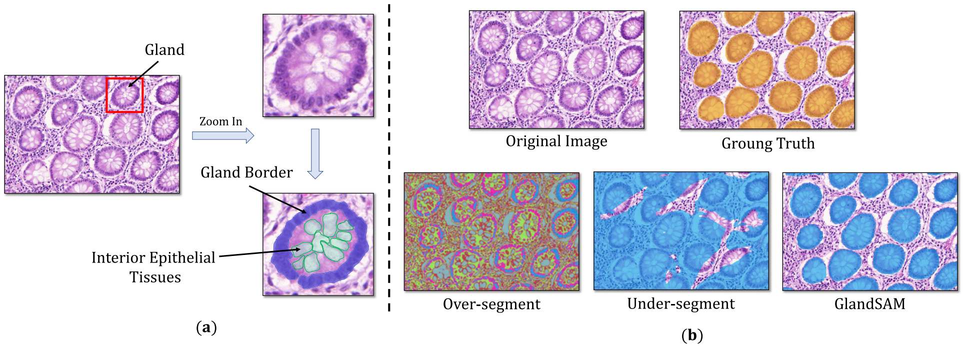

Fig. 1. (a): Example of a gland and its gland border and interior epithelial tissues. (b) Zero-shot Segmentation with point prompts via SAM vs. Our GlandSAM. clay and orange denote the prediction and the ground truth respectively. (This figure incorporates elements from our previous MICCAI conference paper [10]). Over-segmentation refers to the model wrongly segmenting a gland into numerous fractions; Under-segmentation refers to the model misclassifying many glandular regions as background.

images, such implicit object-aware relations are often clearly visible: different parts of a ship, (prows and sterns), share similar properties while being distinct from other objects (people) and the background (water) [22]. With such implicit relation, SAM is able to achieve extraordinary results.

Nevertheless, when it comes to biological tissues in histopathology images, *e.g.*, glands, the situation is different and the above inherent connections no longer hold. The glandular tissues have some unique characteristics: *(1) Heterogeneity within glandular regions;* see Fig. 1(a). Unlike objects in natural images, our segmentation targets, *i.e.*, gland, in histopathology images are composed of different parts, *i.e.*, gland border and interior epithelial tissue, with large variations. The gland borders usually consist of cells with higher gray levels, surrounding the interior epithelial tissues with various color distributions. *(2) Similarity with the background.* The interior epithelial tissues consist of various kinds of cells that may closely resemble those non-glandular tissues in the background. As such, relying solely on these types of prompts can provide only limited information about the morphology of the target objects, SAM methods tend to indiscriminately cluster pixels with similar properties and confuse many gland regions with the background, leading to over-segmentation or under-segment results (see Fig. 1(b)).

To address the above challenge, in this paper, we present GlandSAM, a label-free but accurate and robust method for gland segmentation. **The key insight of our GlandSAM is to inject two clues about gland morphology into SAM to guide the segmentation process**: *(1) Heterogeneity within glands* and *(2) Similarity with the background*. Specifically, our GlandSAM first utilizes the two empirical clues to decompose the intriguing structure of glands by separately selecting proposals for gland sub-regions according to their unique morphological properties. In this way, by resolving the gland with heterogeneous structures into multiple sub-regions, we could accordingly encode the unique morphological properties into respective sub-region proposals. Then, we inject the inherent morphology properties within the gland proposal into SAM in a fine-tuning manner with a morphology-aware semantic grouping (MSG) module, which explicitly groups the semantics of the gland sub-regions to capture the overall morphology information. Finally, we utilize the tuned model to produce well-delineated, complete gland segmentation results (see Fig. 1(b)), while gland proposals serve as visual prompts.

We conducted a thorough evaluation of the performance of our GlandSAM, which involved two categories of comparison targets. Firstly, we compare the segmentation performance of GlandSAM with task-specific supervised and label-free segmentation methods. Secondly, we performed a comparison experiment between our GlandSAM and the recently emerged SAM-based zero-shot segmentation [23] methods. The experimental results on GlaS dataset [2] and the CRAG dataset [4] show that our GlandSAM significantly outperforms the state-of-the-art (SOTA) label-free methods and zero-shot methods by a large margin. Notably, our GlandSAM even outperforms several fully-supervised methods that require pixel-wise labels for training, which further emphasizes the exceptional performance and potential of our GlandSAM.

This work is an extension of our prior conference paper [10], regarding the following highlighted aspects:

(1) The high-level **idea** of GlandSAM in this journal paper builds on our previous work but with significant improvements: (*i*) In MSSG [10], we utilize the unreliable proposal maps extracted via empirical clues to train an extra segmentation model from scratch, which will inevitably incorporate unnecessary noise. In contrast, this journal paper explores an efficient way to leverage SAM's exceptional label-free generalizability for achieving robust label-free gland segmentation. (*ii*) We observe that SAM encounters huge challenges on gland segmentation and highlight primary reasons, *i.e.*, the presence of *heterogeneity within glandular regions* and the *similarity with the background*, and propose to injects gland morphology from the above two clues into SAM to resolve the issues.

(2) The **architecture** of GlandSAM is improved as follows: (*i*) In GlandSAM, we innovatively employ an MSG module to summarize and inject the morphology knowledge hidden inside the proposals into SAM instead of using the noisy proposals to train an extra segmentation model like [10]. (*ii*) We develop a SAM Bypass Adaptation (SAM-BA) mechanism to effectively preserve the injected morphology knowledge while avoiding catastrophic forgetting during the Morphology Knowledge Injection (MKI) phase. The SAM-BA also achieves a time&parameter friendly way to transfer SAM to other domains, providing valuable insights to the

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

community. (*iii*) We innovatively propose to tune the prompt encoder of SAM during MKI which enables utilizing proposal maps as mask prompts to further provide more hints about gland morphology during inference.

(3) We conduct more comprehensive **experiments** to validate the proposed GlandSAM: (*i*) Besides the GlaS dataset used in [10], GlandSAM is further evaluated on the challenging CRAG dataset, while more previous methods are added as counterparts. (*ii*) We achieve 2.11% (4.50%), 2.82% (5.00%), and 7.83% (2.40%) improvement at F1 score, DICE, and mIOU on GlaS (CRAG) dataset, compared with our previous work which is also the current SOTA. (*iii*) We conduct extensive ablation studies on SAM's applications and potential for gland segmentation, which could provide valuable insights to the community.

Our codes are made available at https://github.com/xmed-lab/MSSG

II. Related Works

A. Label-Free Semantic Segmentation

To alleviate the annotation costs, considerable efforts have been extort to design label-free semantic segmentation methods for different tasks on natural images [12], [13], [14], [15], [16], [17] and medical images [24], [25], [26], [27].

1) Label-Free Semantic Segmentation for Natural Image: Prior label-free segmentation methods for natural images can be broadly categorized into coarse-to-fine-grained [13], [14], [15], [28], [29] and end-to-end (E2E) clustering [12], [30], [31]. The former ones typically rely on pre-generated coarse masks, *e.g.*, super-pixel proposals [28], salience masks [15], and self-attention maps [13], [14], [29], as prior, which is not always feasible on gland images. The E2E clustering methods, however, produce under-segment results on gland images by confusing many gland regions with the background [10]. This is due to the fact that E2E clustering relies on the inherent connections between pixels of the same class, as discussed in the Introduction Section, while Glandular tissues demonstrate significant *Heterogeneity within Gland*. As such, the E2E clustering methods tend to indiscriminately cluster pixels with similar properties and confuse many gland regions with the background, leading to under-segment results.

2) Label-Free Semantic Segmentation for Medical Image: Existing label-free segmentation methods have shown promising results in various medical modalities, *e.g.*, magnetic resonance images [32], x-ray images [33] and dermoscopic images [25]. However, directly utilizing these methods to segment glands could lead to over-segment results where a gland is segmented into many fractions rather than being considered as one target [10]. This is because these methods are usually designed to be extremely sensitive to color [25], while gland images present a unique challenge due to their highly dense and complex tissues with intricate color distribution [7].

B. SAM for Medical Image Analysis

Recently, the Segment Anything Model (SAM) was introduced as a groundbreaking foundational model for image segmentation [11] along with multiple concurrent works [11], [34], [35]. SAM introduces the concept of training a large-scale vision transformer using an extremely substantial dataset consisting of 11 million images and 1 billion masks. SAM's most notable feature is its impressive zero-shot segmentation performance via the utilization of diverse visual prompts, *e.g.*, points and bound-boxes (b-box), particularly for previously unseen datasets and tasks [20].

The emergence of the vision foundation model has also provoked the interest of many researchers in the medical image segmentation domain [18], [23], [36], [37]. These studies can be broadly classified into two categories: (1) Designing fine-tuning strategies for the SAM, and (2) evaluating zero-shot generalizability. The former category requires targeted datasets with pixel-level annotations, which are used to transfer SAM into the target domain. For example, [38] proposed to utilize the LoRA strategy to fine-tune the SAM image encoder, prompt encoder, and mask decoder using relatively smaller labeled medical image segmentation datasets. Their proposed SAM-Med achieved state-of-the-art segmentation performance on CT and MRI segmentation tasks. Concurrently, [36] proposed a Medical SAM Adapter that employed several adapters with few parameters to transfer SAM into different medical image segmentation tasks, and achieved considerable improvement on skin lesion segmentation and brain tumor segmentation tasks. Reference [39] applied SAM to the polyp segmentation task using five benchmark datasets under the Everything setting. The results showed that although SAM can accurately segment the polyps in some cases, a large gap exists between SAM and the state-of-the-art methods. Reference [23] assessed the performance of SAM in digital pathology segmentation tasks, including tumor, non-tumor tissue, and cell nuclei segmentation on whole-slide imaging. The experimental results showed that the visual prompts only offer limited clues about the morphology of the target objects. Consequently, relying solely on such prompts, the foundation models, often end up in confusion among classes, where the models primarily focus on low-level features like colors rather than capturing the high-level semantics of the target region and confuse many of them with the background.

III. Methodology

The general pipeline of GlandSAM is depicted in Fig. 2. GlandSAM starts with the Proposal Prompt Mining (PPM) phase, which generates proposal maps for each gland image through empirical clues related to gland morphology. These proposal maps are then utilized to fine-tune the SAM model through Morphology Knowledge Injection (MKI). Finally, the tuned model is employed for segmentation, with the proposal maps serving as visual prompts.

A. Proposal Prompt Mining

During this phase, we utilize empirical clues about gland morphology to extract proposal maps that highlight the gland regions. The two empirical clues could be succinctly summarized as follows: *Each gland exhibits a border region characterized by cells with high gray levels, which encompass interior epithelial tissues similar to the background.* We first

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

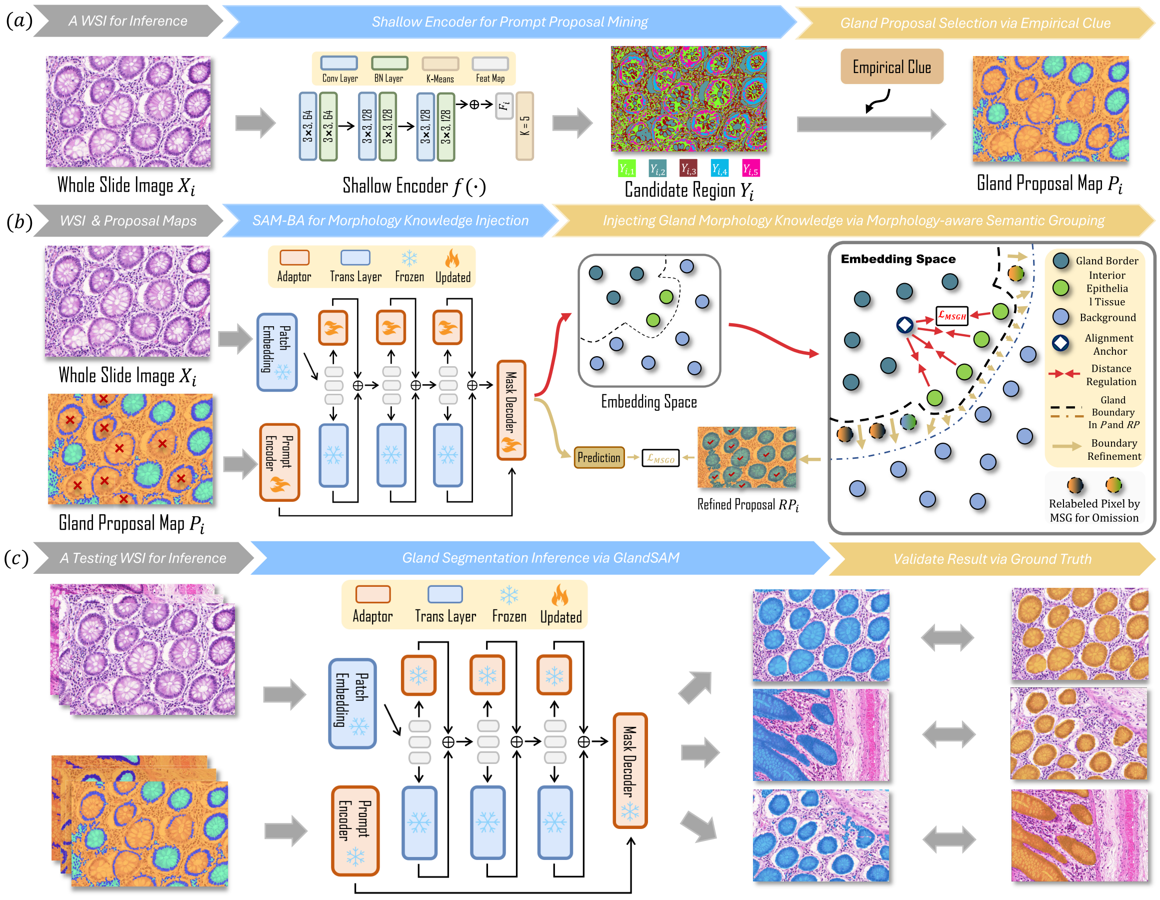

Fig. 2. Overall workflow of our proposed GlandSAM for label-free gland segmentation. (a) Prompt Proposal Mining: We generate proposal maps for each gland sub-region with heterogeneity through empirical clues related to gland morphology. (b) Morphology Knowledge Injection: We inject the gland morphology knowledge hidden inside the proposal maps into SAM in a fine-tuning manner with the MSG module. (c) Gland Segmentation inference via GlandSAM.

train a shallow CNN encoder in a self-supervised manner to divide WSI into several candidate regions, then we utilize the empirical clue to select proposals from these candidates for the gland sub-regions.

Specifically, let the i^{th} input image be denoted as $X_i \in \mathbb{R}^{C\times H\times W}$, where H, W, and C is the height, width, and number of channels. We obtain a feature map $F_i = \|f(X_i)\|_2$, where f is an shallow CNN encoder. We train the encoder in a self-supervised manner. The loss function $\mathcal{L}$ consists of a typical self-supervised Loss $\mathcal{L}_{SS}$, which is the cross-entropy loss between the feature map F_i and the one-hot cluster label $C_i = \arg\max(F_i)$, and a Spatial Continuity Loss $\mathcal{L}_{SC}$, which regularizes the vertical and horizontal variance among pixels within a certain area S to assure the continuity and completeness of the candidate regions. The expressions for the self-supervised Loss $\mathcal{L}_{SS}$ and the Spatial Continuity Loss $\mathcal{L}_{SC}$ are given below:

$$\begin{aligned}\mathcal{L}_{SS}&(F_i[:,h,w], C_i[:,h,w])\\ &= -\sum_{d}^{D} C_i[d,h,w]\cdot \ln F_i[d,h,w] \qquad (1)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{SC}(F_i) = &\sum_{s,h,w}^{S,H-s,W-s} (F_i[:,h+s,w]-F_i[:,h,w])^2\\ &+ (F_i[:,h,w+s]-F_i[:,h,w])^2. \qquad (2)\end{aligned}$$

Then we employ K-means to cluster F_i into 5 candidate regions, denoted as $Y_i = \{y_{i,1} \in \mathbb{R}^{D\times n_0}, y_{i,2} \in \mathbb{R}^{D\times n_2}, \ldots, y_{i,5} \in \mathbb{R}^{D\times n_5}\}$, $n_1 + n_2 + \ldots + n_5$ equals the total number of pixels in the WSI $(H \times W)$.

Proposal Selection via Empirical Clues: The above empirical clue is used to select proposals for gland borders and interior epithelial tissues from the candidate regions Y_i. Particularly, we first select the region with the highest average gray level as the proposal for the *gland borders*. Then, we fill the areas surrounded by the gland border proposals and consider them as the proposal for the *interior epithelial tissues*, while the rest areas of the gland image are regarded as the background. Finally, we obtain the proposal map $P_i \in \mathbb{R}^{3\times H\times W}$, which contains the two proposals for gland sub-regions and one background region.

B. Morphology Knowledge Injection

The proposal maps generated by PPM are then utilized to fine-tune SAM, injecting the implied morphology knowledge. To effectively inject the morphology knowledge, we introduce a morphology-aware semantic grouping (MSG) module during fine-tuning, which summarizes the overall information about glands from their sub-region proposals by explicitly grouping their semantics.

1) Model Design: One simple approach to fine-tune SAM is to unlock all parameters. However, this can lead to three

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

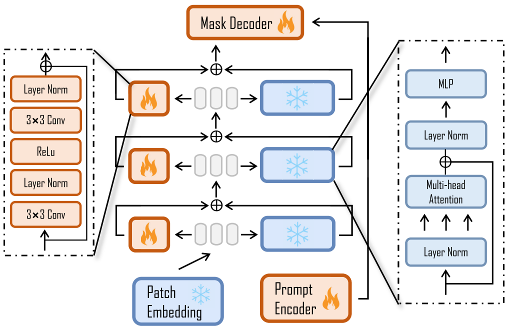

Fig. 3. Detailed model structure of our SAM Bypass Adaption, where SAM is deployed as the base model, while a lightweight two-layer residual block is applied to each Transformer layer as an adaptor.

potential adversarial effects. Firstly, training a large-scale model with unlocked parameters demands excessive computational resources and training time. Secondly, deploying and storing the weight of a new model is cumbersome. Thirdly, adopting this solution without sufficient training samples and accurate labels may lead to catastrophic forgetting [40] and end up with poor performance. To address this issue, we introduce a SAM Bypass Adaption (SAM-BA) to fine-tune SAM in a parameter-efficient manner. Specifically, as shown in Fig. 3, in the SAM-BA, SAM is deployed as a base model, while a lightweight residual block [41] is applied to each transformer layer as an adaptor $Adaptor(\cdot)$ to preserve injected knowledge. During fine-tuning, the parameters of the base model are frozen, while the adaptors' parameters are updated. Each layer output is obtained by fusing the outputs of the adaptors and the Transformer layers of the base model. For instance, the output feature of the i^{th} layer F_i can be obtained as:

$$F_i = \alpha \times B_i(F_{i-1}) + (1-\alpha) \times Adaptor_i(F_{i-1}), \quad (3)$$

where B_i and $Adaptor_i$ denote the i^{th} Transformer layer and adaptor. α is a learnable coefficient parameter.

2) Training Strategy: To inject the gland morphology knowledge, we utilize the sub-region proposal maps highlighting gland sub-regions to fine-tune SAM. Merging the two sub-region proposals as pseudo labels is a simplistic and straightforward approach but it is not optimal in our case. Firstly, the two gland sub-regions exhibit significant heterogeneity in appearance, making it challenging for the model to recognize them as cohesive parts of the same object. Secondly, the PPM phase may produce proposals with inadequate highlighting of many gland regions, particularly the interior epithelial tissues, as shown in Fig. 2 (b) where regions marked with × are **omitted**. As a result, applying pixel-level cross-entropy loss between the gland image and the merged proposal map could introduce unwanted noise into SAM, resulting in sub-optimal performance. As such, we propose two types of Morphology-aware Semantic Grouping (MSG) modules, *i.e.*, MSG for Heterogeneity (MSG-H) and MSG for Omission (MSG-O), to respectively reduce the confusion caused by the two challenges mentioned above. The details of the two MSG modules are described as follows.

Here, we first slice the gland image and its proposal map into patches as inputs. Let the input patch and its corresponding sliced proposal map be denoted as $\hat{X} \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$ and $\hat{P} \in \mathbb{R}^{3\times\hat{H}\times\hat{W}}$. We can obtain the feature embedding map $\hat{F}$ which is derived as $\hat{F} = SAM_{en}(\hat{X})$ and the prediction map $\widetilde{X}$ as $\widetilde{X} = SAM_{de}(F\hat{,} Y)$, where SAM_{en} and SAM_{de} refers to the encoder and mask decoder of SAM respectively.

MSG for Heterogeneity is designed to mitigate the adverse impact of appearance heterogeneity between the gland sub-regions. It regulates the pixel-level feature embeddings of the two sub-regions by explicitly reducing the distance between them in the embedding space. Specifically, according to the proposal map $\hat{P}$, we divide the pixel embeddings in $\hat{F} \in \mathbb{R}^{D\times\hat{H}\times\hat{W}}$ into **G**land border set $|G| = \{g_0, g_1, \ldots, g_{k_g}\}$, **I**nterior epithelial tissue set $|I| = \{i_0, i_1, \ldots, i_{k_i}\}$ and **N**on-glandular, *i.e.*, background, set $|N| = \{n_0, n_1, \ldots, n_{k_n}\}$, where $k_g + k_i + k_n = \hat{H} \times \hat{W}$. Then, we use the average of the pixel embeddings in gland border set $|G|$ as the alignment anchor and pull all pixels of $|I|$ towards the anchor:

$$\mathcal{L}_{MSGH} = \frac{1}{I}\sum_{i\in|I|}\left(i - \frac{1}{G}\sum_{g\in|G|} g\right)^2. \quad (4)$$

MSG for Omission is designed to overcome the problem of partial omission in the proposals. It identifies and relabels the overlooked gland regions in the proposal map and groups them back into the gland semantic category. To achieve this, for each pixel n in the non-glandular, *i.e.*, background, set $|N|$, two similarities are computed with the gland sub-regions $|G|$ and $|I|$ respectively:

$$S_n^G = \frac{1}{|G|}\sum_{g\in|G|}\frac{g}{\|g\|_2}\cdot\frac{n}{\|n\|_2},$$
$$S_n^I = \frac{1}{|I|}\sum_{i\in|I|}\frac{i}{\|i\|_2}\cdot\frac{n}{\|n\|_2}. \quad (5)$$

S_n^G (or S_n^I) represents the similarity between the background pixel n and gland borders (or interior epithelial tissues). If either of them is higher than a preset threshold β (set to 0.7), we consider n as an overlooked pixel of gland borders (or interior epithelial tissues), and relabel n to G (or I). In this way, we could obtain a refined proposal map RP. Finally, we impose a pixel-level cross-entropy loss on the prediction and refined proposal RP to train SAM:

$$\mathcal{L}_{MSGO} = -\sum_{\hat{h},\hat{w}}^{\hat{H},\hat{W}} RP[:,\hat{h},\hat{w}] \cdot \ln \widetilde{X}[:,\hat{h},\hat{w}], \quad (6)$$

The total objective function $\mathcal{L}$ for training the segmentation network can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{MSGO} + \lambda_H \mathcal{L}_{MSGH}, \quad (7)$$

where λ_H (set to 1) is the coefficient.

C. Segmentation With GlandSAM

Finally, we utilize the tuned SAM to perform gland segmentation with proposal maps as visual prompts. Notably,

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

we innovatively employ the Cut-Mix augmentation [42] during segmentation inference. Specifically, we first crop gland images into patches and merge patches from different gland images, and for proposal prompts, We apply the same procedure. In this way, we could create a new gland image that contains mixed regions from other gland images, and present SAM with a wide range of diverse hints derived from different glands, thereby offering a more explicit representation of the pathological structure.

IV. Experiment

A. Datasets

We conduct experiments on three public glandular datasets: Gland Segmentation Challenge (GlaS) dataset [2], Colorectal Adenocarcinoma Gland (CRAG) dataset [4], and Prostate Gland Segmentation (PGlandSeg) dataset [43]. The GlaS dataset comprises 165 histopathology patches stained with H&E extracted from 16 WSIs. We follow the previous works [1], [7], [44], and split the GlaS dataset into a training set with 85 images and a testing set with 80 images. We utilize the images in the training set and their proposals to fine-tune the SAM, and evaluate the performance on the testing set. The CRAG dataset consists of 213 histopathology patches stained with H&E extracted from 38 WSIs. Following previous works [4], we split the CRAG dataset into 85 training images and 80 testing images. Compared with the GlaS dataset, the CRAG dataset contains a higher proportion of irregular malignant glands, making it more challenging. The Prostate Gland Segmentation dataset [43] consists of 1500 histopathology patches stained with H&E obtained from 150 patients, with 18851 glands annotated. Following the official dataset split setting [43], we divide the PGlandSeg into 1000 training images and 500 testing images. Compared with the GlaS and the CRAG datasets, the PGlandSeg dataset contains more glandular structures.

B. Implementation Details

The experiments are conducted on four A100 GPUs. For the PPM, a 3-layer encoder is trained for each training sample. Each convolutional layer employs a 3×3 convolution with a stride of 1 and a padding size of 1. The encoder undergoes training for 50 iterations using an SGD optimizer with a polynomial decay policy and an initial learning rate of 1e-2. During the process of fine-tuning SAM, we utilize the ViT_B version of SAM as our base model and incorporate a lightweight adaptor for each Transformer Layer within the image encoder. Each adaptor is implemented with a three-layer residual block [41]. During the MKI phase, we keep the parameters of the image encoder frozen, while updating the parameters of the adaptors, mask decoder, and prompt encoder. The MKI phase undergoes training for 200 epochs using an AdamW optimizer. In line with prior studies [38], [44], we implemented exponential learning rate decay to ensure the stability of the process. Specifically, we set the initial learning rate to 1e-4 and established a decay period of 100. Note that, for a fair comparison, we utilize the pseudo labels generated by different methods to respectively train a PSPNet [45]) for each method, and compare the performance of the trained PSPNets to evaluate their effectiveness.

C. Comparison With State-of-the-Art Methods

To assess the performance of our GlandSAM, we conduct two kinds of comparative experiments: (1) We compare the segmentation performance with state-of-the-art (SOTA) segmentation methods employing different supervisions, as shown in Table I. (2) We compare GlandSAM with the SAM-based zero-shot segmentation, as presented in Table II.

1) Comparison Results on GlaS Dataset: To begin with, we compare our proposed GlandSAM with many SOTA methods that employ different supervision settings on the GlaS dataset. The quantitative comparison results are illustrated in Table. I. Many unsupervised methods fail on the GlaS dataset due to *(1) Heterogeneity within glandular regions.* and *(2) Similarity with the background.*, and thus obtain limited improvement compared with a randomly initialized network (less than 20%). Our proposed GlandSAM, on the contrary, achieves a much more significant performance advance (+30.65% F1 score, +32.28% Dice, and +38.42% mIOU) with the injection and guidance of gland morphology knowledge. In addition, compared with our previously proposed and published method, *i.e.*, MSSG [10], there is a huge margin of 2.11% at F1 score, 3.82% at Dice, and 7.83% at mIOU. Furthermore, even compared with the fully-supervised segmentation method Unet [46] which requires pixel-level annotations, our completely label-free method can still achieve 2.59%, 1.87%, and 5.21% improvement at F1 score, Dice, and mIOU.

Besides the quantitative results, in Fig. 4, we present the qualitative visualization results of our GlandSAM and its counterpart, *i.e.*, SGSCN [25] and MSSG [10]. The predictions of SGSCN appear to be coarse and inaccurate, especially in the interior epithelial tissues. By incorporating gland morphology knowledge, MSSG achieves a more comprehensive understanding of gland structures, leading to more accurate segmentation results. Furthermore, our GlandSAM model enhances the performance of MSSG by producing even more precise results that closely resemble the ground truth.

2) Comparison Results on CRAG Dataset: To further evaluate the performance of our GlandSAM model, we additionally conduct comparative experiments on the CRAG dataset. The comparison results are also shown in Table. I. Similar to the results on the GlaS dataset, previous label-free segmentation methods achieve limited improvement compared with a randomly initialized model. In contrast, our GlandSAM gained 31.49%, 29.93%, and 20.96% of advancement in F1 score, DICE, and mIOU, with a large margin to its label-free counterparts. Furthermore, even without any annotations from the CRAG dataset, our GlandSAM still outperforms many fully-supervised methods. For example, even with pixel-level annotations, VF-CNN (C8) [47] still lags behind our method with 4.33% at F1 score, 4.06% at Dice, and 8.18% at mIOU. Compared with the SOTA fully-supervised segmentation method, *i.e.*, DSF-CNN, our GlandSAM still narrows the gap to 5.47% and 6.84% at F1 score and DICE.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

TABLE I

Comparison Results on GlaS and CRAG Dataset. Bold and Underline Denote Best and Second-Best Results of the *Label-Free Methods*. † Denotes Our Previous Published MICCAI Conference Paper [10]

Dataset	Method	Backbone	Supervision	F1	DICE	mIOU
GlaS Dataset	Unet [MICCAI'15]	U-Net	Fully	77.78%	79.04%	65.34%
	ResUNet [ITME'18]	U-Net	Fully	78.83%	79.48%	65.95%
	MedT [MICCAI'21]	Transformer	Fully	81.02%	82.08%	69.61%
	LoGo [MICCAI'21]	Transformer	Fully	79.68%	-	67.69%
	UCTransNet [AAAI'22]	Transformer	Fully	-	90.25%	82.24%
	SAM-Adapter [Arxiv'23]	SAM + LoRA	Fully	91.33%	92.74%	85.63%
	Randomly Initialize	PSPNet	None	49.72%	48.63%	32.13%
	DeepCluster [ECCV'18]	PSPNet	None	57.03%	57.32%	40.17%
	IIC [ICCV'19]	PSPNet	None	60.23%	59.48%	42.33%
	PiCIE [CVPR'21]	PSPNet	None	64.98%	65.61%	48.77%
	MaskContrast [CVPR'21]	PSPNet	None	64.98%	65.61%	48.77%
	DINO [ICCV'21]	PSPNet	None	56.93%	57.38%	40.23%
	DSM [CVPR'22]	PSPNet	None	68.18%	66.92%	49.92%
	MaskDistill [ARXIV'22]	PSPNet	None	69.44%	68.35%	51.99%
	STEGO [ICLR'22]	PSPNet	None	63.57%	62.20%	45.14%
	SGSCN [MICCAI'21]	PSPNet	None	67.62%	68.72%	52.16%
	ACSeg [ICCV'23]	PSPNet	None	66.77%	64.11%	49.74%
	MSSG† [MICCAI'23]†	PSPNet	None	78.26%	77.09%	62.72%
	GlandSAM (Ours)	PSPNet	None	**80.37%**	**80.91%**	**70.55%**
CRAG Dataset	Unet [MICCAI 15']	U-Net	Fully	82.70%	84.40%	70.21%
	VF-CNN (C4) [ICCV'17]	RotEqNet	Fully	71.10%	72.10%	57.24%
	VF-CNN (C8) [ICCV'17]	RotEqNet	Fully	74.50%	75.80%	59.77%
	VF-CNN (C12) [ICCV'17]	RotEqNet	Fully	77.60%	78.20%	60.11%
	G-CNN [ICCV'16]	G-CNN	Fully	83.30%	85.60%	-
	Steer G-CNN [CVPR'18]	G-CNN	Fully	81.10%	84.80%	-
	MILDNet [MIA'19]	MILD-Net	Fully	86.90%	88.30%	76.95%
	DSF-CNN [TMI'20]	DSF-CNN	Fully	87.40%	89.10%	-
	TA-Net [WACV'22]	TA-Net	Fully	84.20%	-	-
	SAM-Adapter [Arxiv'23]	SAM + LoRA	Fully	89.42%	90.92%	88.47%
	Randomly Initialize	PSPNet	None	50.44%	52.33%	47.33%
	PiCIE [CVPR'21]	PSPNet	None	67.04%	64.33%	52.06%
	DSM [CVPR'22]	PSPNet	None	67.22%	66.07%	52.28%
	SGSCN [MICCAI'21]	PSPNet	None	69.29%	67.88%	55.31%
	MSSG† [MICCAI'23]	PSPNet	None	77.43%	77.26%	65.89%
	GlandSAM (Ours)	PSPNet	None	**81.93%**	**82.26%**	**68.29%**
PGlandSeg Dataset	Unet [MICCAI'15]	U-Net	Fully	83.72%	84.56%	82.80%
	ResUNet [ITME'18]	U-Net	Fully	88.59%	88.72%	89.33%
	MedT [MICCAI'21]	Transformer	Fully	89.10%	88.20%	82.00%
	LoGo [MICCAI'21]	Transformer	Fully	82.20%	80.92%	83.43%
	UCTransNet [AAAI'22]	Transformer	Fully	88.52%	89.23%	90.11%
	SAM-Adapter [Arxiv'23]	SAM + LoRA	Fully	88.99%	90.33%	89.27%
	Randomly Initialize	PSPNet	None	44.64%	44.77%	45.29%
	DeepCluster [ECCV'18]	PSPNet	None	46.73%	48.92%	47.33%
	SGSCN [MICCAI'21]	PSPNet	None	54.30%	52.27%	56.04%
	MSSG† [MICCAI'23]†	PSPNet	None	70.41%	71.95%	72.88%
	GlandSAM (Ours)	PSPNet	None	**77.60%**	**76.23%**	**77.84%**

Besides the quantitative results, Fig. 5 shows the predictions of our GlandSAM and its counterparts, *i.e.*, SGSCN [25] and MSSG [10], on the CRAG dataset. As can be seen in Fig. 5, even without any kind of labels, GlandSAM can still present smooth and accurate predictions, proving its robustness.

3) Comparison Results on PGlandSeg Dataset: To additionally evaluate the generalizability of our GlandSAM model, we additionally conduct comparative experiments on a much larger PGlandSeg dataset. The comparison results are also shown in Table. I. Similar to the outcomes observed on the GlaS and CRAG datasets, prior label-free segmentation techniques show only marginal enhancements when contrasted with a randomly initialized model. In contrast, our GlandSAM gained 32.96%, 31.46%, and 32.55% of advancement in F1 score, DICE, and mIOU, with a large margin to its label-free counterparts.

D. Comparison With SAM-Based Zero-Shot Segmentation

The most notable feature of SAM lies in its remarkable zero-shot generalizability, which has piqued the interest of numerous researchers exploring its application in various medical domains [23], [37]. In this study, we perform a comparative analysis with zero-shot segmentation pipelines. The SAM-based zero-shot segmentation requires several visual prompts, *e.g.*, points, bound-boxes (b-boxes), and scribbles, which are usually obtained from an expert pathologist, *e.g.*, clicking in the target regions, drawing b-boxes or scribbles. To stimulate the plausible prompting strategies, we evaluate the zero-shot segmentation performance under the following settings:

- We randomly select several points from the ground-truth mask, including at least one central point for each gland.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

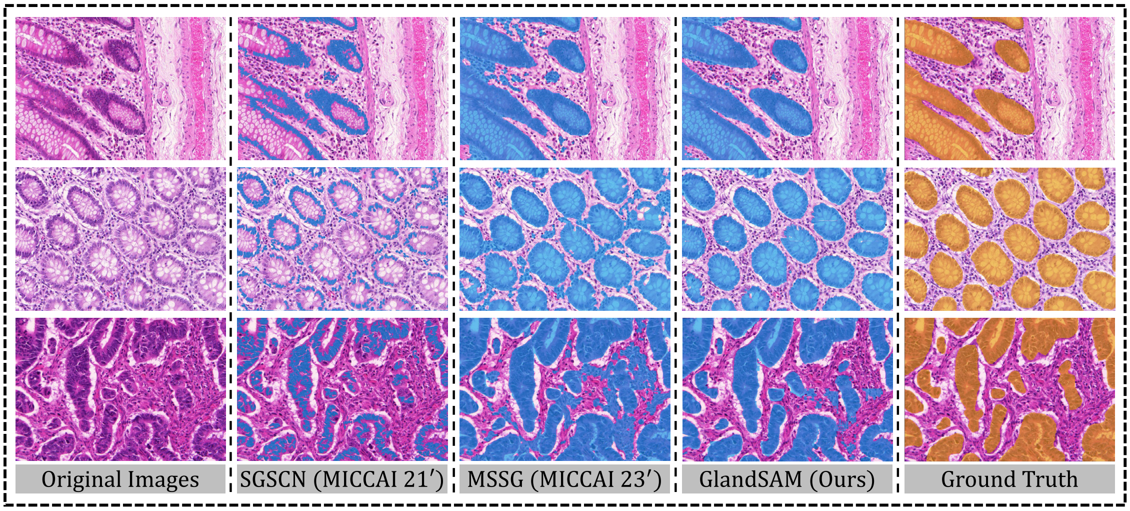

Fig. 4. Visualization of predictions on GlaS dataset. Blue denotes predicted glandular regions of SGSCN [25], MSSG [10], and our GlandSAM. Orange denotes ground truth regions of glandular tissues.

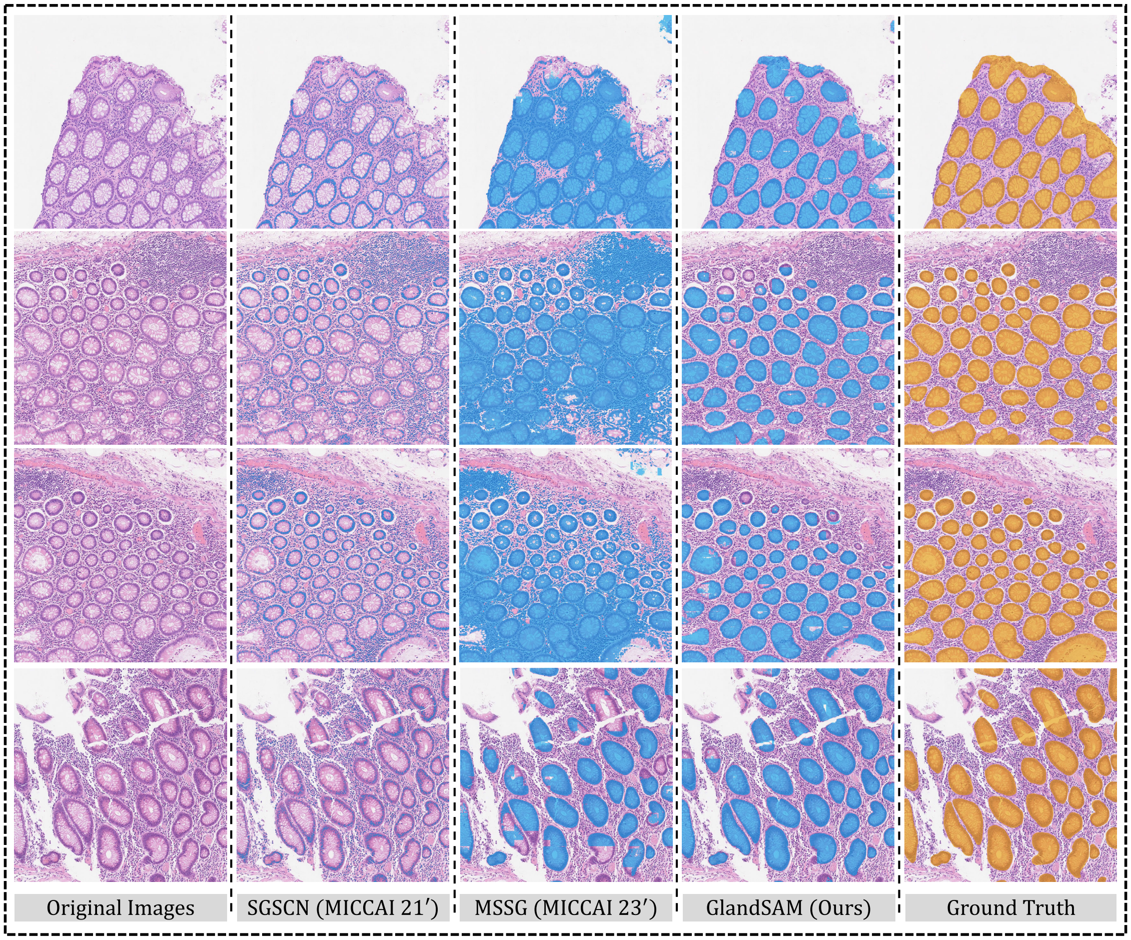

Fig. 5. Visualization of predictions on CRAG dataset. Blue denotes predicted glandular regions of SGSCN [25], MSSG [10], and our GlandSAM. Orange denotes ground truth regions of glandular tissues.

- We utilized the instance mask to draw b-boxes for glands.
- We draw scribbles across different sub-regions of glands for the glandular regions according to the ground truth.

Table. II and Fig. 6 illustrate the quantitative and qualitative comparison results on the GlaS dataset. As depicted in Table II, when provided with less than 20 point prompts, SAM exhibits poor performance, with F1 scores lower than 60%. This is primarily due to the limited information provided by the point prompts, which fail to encompass the comprehensive characteristics of glandular tissues. Consequently, SAM tends to focus on specific gland sub-regions, disregarding the intricate morphology and interplay between different regions

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

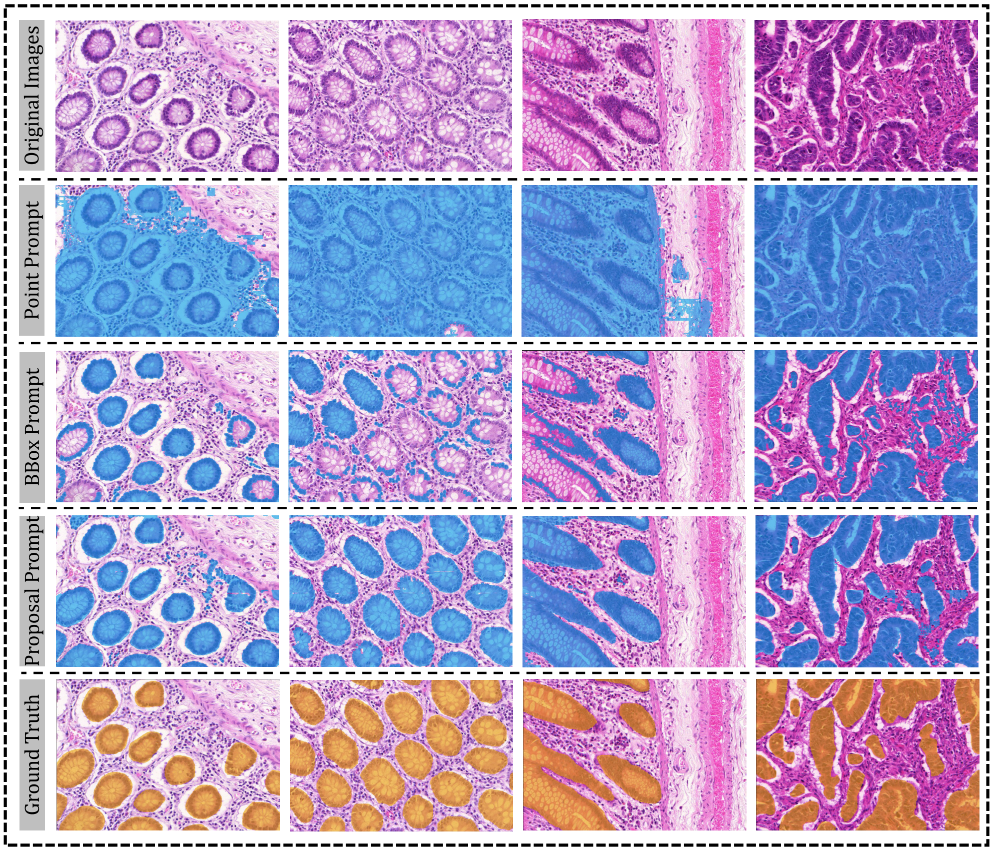

Fig. 6. Visualization of the predictions from the GlandSAM and the zero-shot segmentation with SAM using different kinds of prompts.

TABLE II
COMPARISON RESULTS WITH SAM-BASED ZERO-SHOT SEGMENTATION ON GLAS DATASET, WHERE BOLD AND UNDERLINE DENOTES THE BEST AND THE SECOND-BEST RESULTS

Method	Prompt Type	Prompt Num	F1	DICE	mIOU
GlandSAM	Proposal Prompt	5 Proposal	77.76%	80.24%	67.00%
	Proposal Prompt	10 Proposal	77.25%	80.01%	66.68%
	Proposal Prompt	15 Proposal	77.96%	79.79%	68.37%
	Proposal Prompt	20 Proposal	**80.37%**	**80.91%**	**70.55%**
Zero-shot	Point Prompt	1 Point	55.72%	55.60%	38.94%
	Point Prompt	5 Point	54.88%	52.79%	40.63%
	Point Prompt	10 Point	59.86%	57.24%	42.72%
	Point Prompt	20 Point	65.70%	65.86%	49.10%
	Point Prompt	16^2 Point	70.89%	73.04%	64.21%
	Point Prompt	64^2 Point	71.48%	73.59%	64.77%
	Point Prompt	128^2 Point	75.66%	78.30%	67.23%
	B-Box Prompt	1 b-box	67.94%	70.33%	59.88%
	B-Box Prompt	5 b-box	72.69%	73.00%	62.78%
	B-Box Prompt	10 b-box	74.55%	76.03%	64.60%
	Scribble Prompt	1 Scribble	54.77%	56.67%	46.33%
	Scribble Prompt	5 Scribble	57.56%	60.11%	48.71%
	Scribble Prompt	10 Scribble	59.20%	63.89%	50.66%

and the background. With an increase in the number of point prompts to 16^2, SAM demonstrates significant improvement, but it still falls short compared to our GlandSAM. Moreover, it is worth noting that such a high number of points is impractical in real-world scenarios due to its labor-intensive nature. Secondly, we also perform experiments using scribble prompts. Scribble prompts outline a broader area of glandular tissue, offering more comprehensive information about glands compared to isolated points. Consequently, SAM with scribble prompts achieves a higher performance, *i.e.*, about a 5% gap at F1 score. However, the scribble prompts still cannot cover all areas of glands leading to the overlook of some morphological characteristics. Compared with the above two kinds of prompts, b-boxes provide a more explicit spatial constraint for segmentation by highlighting the entire region of glands. Consequently, they demonstrate the highest zero-shot segmentation performance. Specifically, when the number of the b-box increase to 10, the mIOU rises to 64.60%. However, the limitation of this kind of b-box prompts is also obvious. On the one hand, the segmentation performance still lags behind many task-specific label-free methods, *e.g.*, MSSG [10], due to the unique characteristics of gland morphology and the large domain gap.

In contrast, our GlandSAM demonstrates significantly improved segmentation performance. Specifically, when provided with 20 proposals, GlandSAM achieves a notable improvement of 5.82% in F1 score, 4.88% in DICE, and 5.95% in mIOU compared to SAM with bounding box prompts. Furthermore, even when compared to SAM with a substantial number of point prompts (specifically, 128^2), which would be extremely labor-intensive in real-world scenarios, our GlandSAM still outperforms it with improvements of 4.71% in F1 score, 2.61% in DICE, and 3.32% in mIOU. Moreover, in Fig. 5, we showcase the qualitative visualization results of our GlandSAM model in comparison to zero-shot segmentation methods, where GlandSAM yields much more accurate, complete, and visually appealing results that closely resemble ground truth.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

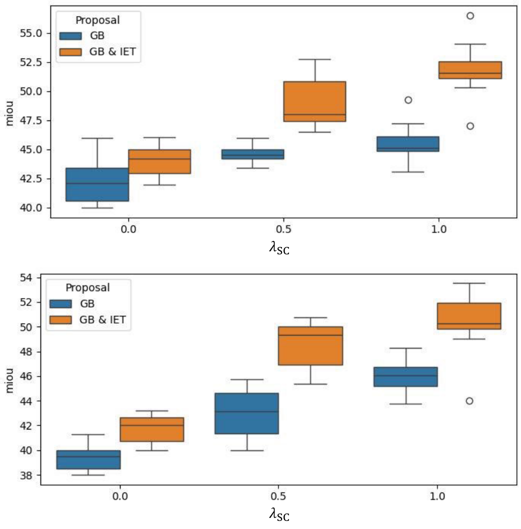

Fig. 7. Ablation studies of the spatial continuity loss during the PPM stage on the GlaS dataset (upper row) and CRAG dataset (bottom row). "GB" denotes the **G**land **B**order proposal, "IET" denotes the **I**nterior **E**pithelial **T**issue proposal.

TABLE III
Ablation Study on Fine-Tuning Strategy

Method	Fine-tuning Strategy MSG-H	 MSG-O	mIOU	Improvement
Baseline	✗	✗	64.72%	-
Setting 1	✔	✗	68.64%	+3.92%
Setting 2	✗	✔	65.66%	+0.94%
Final Version	✔	✔	70.55%	+5.83%

E. Ablation Studies

To validate the effectiveness of each module in GlandSAM, we conduct comprehensive ablation studies, which are organized as follows: (1) We evaluate GlandSAM with varying numbers of prompts, shown in Table. II and Fig.6. (2) We assess the quality of proposals after the PPM phase, as depicted in Fig.7. (3) We evaluate the efficacy of the two morphology-aware semantic grouping modules during MKI phase, as illustrated in Table. III.

1) Studies on the Prompt Number: We begin by evaluating the segmentation performance of GlandSAM with varying numbers of proposal prompts. To achieve this, we merge the number of images and their corresponding proposals into a single image and proposal map. It is important to note that as the number of proposal prompts increases, the computational resources and inference time required also increase. The results are shown in Table. II. As can be seen, with the number of proposal prompts increasing, the segmentation performance of our GlandSAM advances slightly but steadily. It is also worth noting that, even with only 5 proposal prompts, our GlandSAM still beats all zero-shot segmentation methods using SAM.

2) Studies on Proposal Prompt Mining: To verify the effectiveness of each component used in the Proposal Prompt Mining (PPM) stage, we conduct a few ablation studies and list the performance of our CNN encoder. Specifically, we compare our shallow CNN encoder $f_i(\cdot, \theta_i)$ with the same CNN but without training with Spatial Continuity Loss L_{SC} (λ_{SC} set to 0). Fig. 7 reveals the results on two datasets. From the quantitative results, we observe that $\mathcal{L}_{SC}$ can not only improve the mIOU of gland proposals but also improve the robustness of the CNN encoder. From the visualization of the initial proposals (Fig. 8), we can observe that the major reason for the improvement is that L_{SC} can help the model generate more intact gland borders, leading to more accurate extraction of interior epithelial tissue proposals.

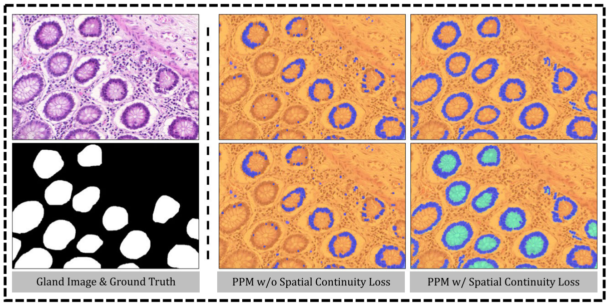
Gland Image & Ground Truth
PPM w/o Spatial Continuity Loss
PPM w/ Spatial Continuity Loss

Fig. 8. Visualization of the proposal maps after the PPM stage. The upper row shows the extracted **gland border (GB)** proposal only, and the bottom row shows both **GB** and **interior epithelial tissue (IET)** proposal. With Spatial Continuity Loss, the gland border proposal shows better completeness, which leads to better extraction of IET proposal.

TABLE IV
Ablation Study on Hyper-Parameter α in Equation (3)

α Setting	mIOU	Improvement
0.0 (Baseline)	66.47%	-
0.3 (Setting 1)	68.93%	+2.46%
0.5 (Final Version)	70.55%	+4.08%
0.8 (Setting 2)	68.79%	+2.32%

3) Studies on Morphology Knowledge Injection: To further study the contribution of the two MSG modules during the morphology knowledge injection stage, we list the performance on the GlaS dataset with different fine-tuning settings in Table. III. We first fine-tune the SAM with only pixel-level cross-entropy loss as the baseline. Then, we progressively add the two MSG modules to respectively verify their effectiveness. As shown in Table. III, we gain 3.92% and 0.94% of mIOU with the involvement of MSG for Heterogeneity and MSG for Omission respectively. Furthermore, when involving both modules, our network can finally achieve 70.55% at mIOU, surpassing the baseline of over 5.83%. Moreover, to determine appropriate values for hyper-parameter α in equation (3), we implement extensive experiments with the different values of α on the GlaS dataset. Experimental results are given in Tab. IV. We can observe that the mIOU achieves the highest values when α is set to 0.5, reaching 70.55%, which surpasses the baseline (α is set to 0) by 4.08%.

F. Studies on CutMix Augmentation During Inference

Since the resolution of whole slide images is quite big, we usually need to crop them into patches before feeding them into deep models. During segmentation inference, we employ Cut-Mix augmentation upon the prompt proposals by mixing patches of proposal maps into a new prompt proposal. This step intends to increase the diversity of the prompts. In Tab. V, we show the ablations to demonstrate the contribution

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ABLATION STUDY ON CUT-MIX AUGMENTATION DURING THE SEGMENTATION INFERENCE

α Setting	mIOU	Improvement
w/o. Inference CutMix Aug.	68.92%	-
w/. Inference CutMix Aug.	70.55%	+1.63%

TABLE VI
ABLATION STUDY ON INSTANCE-LEVEL GLAND SEGMENTATION ON THE GLAS DATASET

Method	Supervision	Object F1	Object Dice	Object Hausdorf
DCAN	Instance Mask	91.2%	89.7%	45.42
MILD-Net	Instance Mask	91.4%	91.3%	41.54
MPCNN	Instance Mask	89.1%	88.2%	57.41
Unet	Semantic Mask	83.1%	84.4%	77.62
MedT	Semantic Mask	83.7%	84.5%	80.54
DeepCluster	None	33.6%	37.4%	132.9
SGSCN	None	63.3%	64.8%	117.3
GlandSAM (Ours)	None	80.3%	79.5%	78.44

of cut-mix augmentation (+1.63%) during segmentation inference.

G. Studies on Instance-Level Segmentation Performance

Despite the semantic-level segmentation, we also demonstrate the performance from the instance-level segmentation on the GlaS dataset; see Table. VI. In order to obtain the instance-level prediction, we first utilize the `openCV` library to perform convolutional erosion (*i.e.*, `cv2.erode`) for 5 iterations to disconnect different glandular objects that are linked together. Then, we used `torchvision.ops` library to repurpose the eroded mask into bound-boxes. Finally, we apply the generated bound-boxes on the original semantic masks, and consider the intersection of the semantic mask and each bound-box as the mask for each gland instance. As shown in Table. VI, even without any annotations including mask and bound-box contained in the instance mask, our GlandSAM could still lower the gap compared with the fully-supervised methods to 3.4%, 5.0%, and 2.1 at Object F1 score, Object DICE and Object Hausdorf. Furthermore, compared with its unsupervised counterpart, our GlandSAM achieves a huge gap, *i.e.*, over 15% at Object F1 score.

V. DISCUSSION

The high cost of annotation has been a significant barrier to the development of DL-based gland segmentation methods. Therefore, it is highly desirable to design a label-free method that **eliminates the need for any annotations**. In recent times, the Segment Anything Model (SAM) has garnered significant attention due to its remarkable generalizability across a wide range of domains. Several researchers have reported the exceptional performance and adaptability of SAM in various domains. However, the typical pipeline of applying SAM, *i.e.*, zero-shot segmentation with visual prompts, results in limited performance, due to the presence of *heterogeneity within glandular regions* and the *similarity with the background*. To this end, we introduce a novel GlandSAM, which utilizes empirical clues to generate initial proposals for gland sub-regions. Subsequently, we employ a morphology-specific grouping (MSG) mechanism to inject morphology knowledge into SAM. This innovative approach enables the integration of morphology knowledge, which is obtained without any supervision, into SAM, resulting in significant improvements in gland datasets. We verify the superiority of our GlandSAM on three public benchmarks. Quantitative results are offered in Table I-V, explicitly showing the superiority of the proposed method, while Fig. 4-8 intuitively illustrate the improvement.

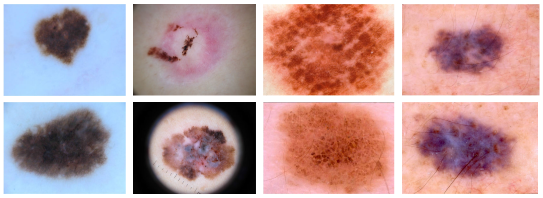

Fig. 9. Samples from ISIC Challenge 2017 dataset, showcasing the Heterogeneity across the skin lesions.

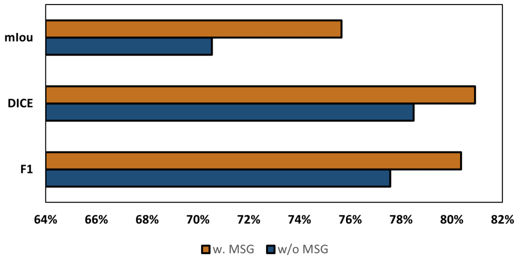

Fig. 10. Performance improvement with MSG on skin lesion dataset.

Although our GlandSAM is designed specifically for the gland segmentation tasks, the idea of semantic grouping, *i.e.*, injecting the semantic knowledge extracted from medical morphology clues into the large foundation model, can also be applied to segment multiple meaningful pathological targets, *e.g.*, cells and nuclei, whose morphological features can be quantified and provide valuable information regarding tumor aggressiveness grading [48], diagnosis [49], [50], staging [51], and prognosis [52]. Moreover, the idea of semantic grouping can be applied to various medical domains besides the histopathology image, addressing a common issue: *heterogeneity within class*, where different parts or objects of one same class may differ from each other. As shown in Fig. 9 which takes skin lesion segmentation as an example, the variation across different kinds of skin lesions is quite obvious. Herein, we conduct experiments about the effectiveness of MSG modules on the ISIC Challenge 2017 dataset [53], and list the performance in Fig. 10. According to Fig.10, with the MSG module, we achieve 2.79%, 2.42%, and 5.10% improvement in F1 score, DICE, and mIOU respectively, proving its effectiveness and potential to apply to other medical fields.

Despite the progress, our GlandSAM still has some limitations. First, visually noticeable clues about object morphology are not always available and may require certain biological knowledge. In the future, we will try to follow a more generalized method to produce initial proposals. For example, attention maps from DINO [29] could be potential. The culprit of its bad performance lies in the indiscriminate aggregation of attention maps from different attention heads. Considering the literature on the multi-head self-attention mechanism, we plan

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
COMPARISON EXPERIMENT RESULTS WITH SELF-ATTENTION METHODS (DENOTED WITH †) ON THE GLAS DATASET

Method	Supervision	Evaluation Metric		
		F1	Dice	mIOU
PiCIE	None	56.21%	56.94%	39.81%
DINO†	None	56.93%	57.38%	40.23%
GlandSAM	None	80.37%	80.91%	70.55%

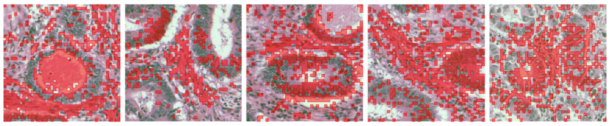

Fig. 11. Visualization of self-attention maps on gland patches.

to exploit different attention maps as proposals of different sub-regions, which seems more reasonable and general for different segmentation tasks. Furthermore, as we look ahead, our intention is to extend the application of our methods to gigapixel whole slide images, which encompass a diverse range of tissues. We aim to utilize these methods for various histopathology image segmentation tasks, including tumor and nuclei segmentation.

VI. Conclusion

In this study, we propose a novel GlandSAM for label-free gland segmentation. GlandSAM innovatively exploits empirical clues about gland morphology to extract meaningful morphological knowledge, which is then injected into SAM using a morphology-aware semantic grouping module. As a result, SAM learns comprehensive information about glands and produces well-defined and complete glandular regions. Our GlandSAM, even without any supervision, could achieve comparable or even better results than fully supervised methods on three glandular benchmarks. Moreover, the idea of our morphology-aware semantic grouping is a potential solution to address the common *heterogeneity within class* which could further benefit other medical image segmentation tasks.

References

[1] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2021, pp. 36–46.

[2] K. Sirinukunwattana et al., “Gland segmentation in colon histology images: The glas challenge contest,” *Med. Image Anal.*, vol. 35, pp. 489–502, Jan. 2017.

[3] H. Chen, X. Qi, L. Yu, and P.-A. Heng, “DCAN: Deep contour-aware networks for accurate gland segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2487–2496.

[4] S. Graham et al., “MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images,” *Med. Image Anal.*, vol. 52, pp. 199–211, Feb. 2019.

[5] Y. Xu et al., “Gland instance segmentation using deep multichannel neural networks,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2901–2912, Dec. 2017.

[6] H. Wang, M. Xian, and A. Vakanski, “TA-net: Topology-aware network for gland segmentation,” in *Proc. WACV*, 2022, pp. 1556–1564.

[7] Y. Li, Y. Yu, Y. Zou, T. Xiang, and X. Li, “Online easy example mining for weakly-supervised gland segmentation from histology images,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2022, pp. 578–587.

[8] L. Yang et al., “BoxNet: Deep learning based biomedical image segmentation using boxes only annotation,” 2018, *arXiv:1806.00593*.

[9] C. L. Srinidhi, O. Ciga, and A. L. Martel, “Deep neural network models for computational histopathology: A survey,” *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.

[10] Q. Zhang, Y. Li, C. Xue, and X. Li, “Morphology-inspired unsupervised gland segmentation via selective semantic grouping,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist.*, 2023, pp. 281–291.

[11] J. Chen, Z. Yang, and L. Zhang. (2023). *Semantic Segment Anything*. [Online]. Available: https://github.com/fudan-zvg/Semantic-Segment-Anything

[12] J. Hyun Cho, U. Mall, K. Bala, and B. Hariharan, “PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16789–16799.

[13] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, “Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8354–8365.

[14] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. Freeman, “Unsupervised semantic segmentation by distilling feature correspondences,” in *Proc. ICLR*, 2021, pp. 1–20.

[15] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, “Unsupervised semantic segmentation by contrasting object mask proposals,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10032–10042.

[16] Z. Deng and Y. Luo, “Learning neural eigenfunctions for unsupervised semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 551–561.

[17] K. Li et al., “ACSeg: Adaptive conceptualization for unsupervised semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7162–7172.

[18] Y. Huang et al., “Segment anything model for medical images?” *Med. Image Anal.*, vol. 2, May 2023, Art. no. 103061.

[19] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: An experimental study,” *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102918.

[20] C. Zhang et al., “A comprehensive survey on segment anything model for vision and beyond,” 2023, *arXiv:2305.08196*.

[21] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, “Segment anything is not always perfect: An investigation of SAM on different real-world applications,” 2023, *arXiv:2304.05750*.

[22] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng, “Instance-aware self-supervised learning for nuclei segmentation,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2020.

[23] R. Deng et al., “Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging,” 2023, *arXiv:2304.04155*.

[24] J. Chen and E. C. Frey, “Medical image segmentation via unsupervised convolutional neural network,” 2020, *arXiv:2001.10155*.

[25] E. Ahn, D. Feng, and J. Kim, “A spatial guided self-supervised clustering network for medical image segmentation,” in *Medical Image Computing and Computer Assisted Intervention—-MICCAI*. Cham, Switzerland: Springer, 2021.

[26] W. Bai et al., “Self-supervised learning for cardiac MR image segmentation by anatomical position prediction,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2019.

[27] I. Aganj, M. G. Harisinghani, R. Weissleder, and B. Fischl, “Unsupervised medical image segmentation based on the local center of mass,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–20, Aug. 2018.

[28] J.-J. Hwang et al., “SegSort: Segmentation by discriminative sorting of segments,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7333–7343.

[29] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[30] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–26.

[31] X. Ji, A. Vedaldi, and J. Henriques, “Invariant information clustering for unsupervised image classification and segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9864–9873.

[32] L. Liu, A. I. Aviles-Rivero, and C.-B. Schönlieb, “Contrastive registration for unsupervised medical image segmentation,” 2020, *arXiv:2011.08894*.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.

[33] Q. Huang et al., “A chan-vese model based on the Markov chain for unsupervised medical image segmentation,” *Tsinghua Sci. Technol.*, vol. 26, no. 6, pp. 833–844, Dec. 2021.

[34] X. Zou et al., “Segment everything everywhere all at once,” 2023, *arXiv:2304.06718*.

[35] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, “SegGPT: Segmenting everything in context,” 2023, *arXiv:2304.03284*.

[36] J. Wu et al., “Medical SAM adapter: Adapting segment anything model for medical image segmentation,” 2023, *arXiv:2304.12620*.

[37] C. Mattjie et al., “Zero-shot performance of the segment anything model (SAM) in 2D medical imaging: A comprehensive evaluation and practical guidelines,” 2023, *arXiv:2305.00109*.

[38] K. Zhang and D. Liu, “Customized segment anything model for medical image segmentation,” 2023, *arXiv:2304.13785*.

[39] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong, “Can SAM segment polyps?” 2023, *arXiv:2304.07583*.

[40] J. O. Zhang, A. Sax, A. Zamir, L. Guibas, and J. Malik, “Side-tuning: A baseline for network adaptation via additive side networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 698–714.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, “CutMix: Regularization strategy to train strong classifiers with localizable features,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[43] M. Salvi et al., “A hybrid deep learning approach for gland segmentation in prostate histopathological images,” *Artif. Intell. Med.*, vol. 115, May 2021, Art. no. 102076.

[44] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, “UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1–22.

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[46] O. Ronneberger et al., “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, pp. 234–241.

[47] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, “Rotation equivariant vector field networks,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5058–5067.

[48] M. Fleming, S. Ravula, and S. F. Tatishchev, “Colorectal carcinoma: Pathologic aspects,” *J. Gastrointestinal Oncol.*, vol. 3, no. 3, p. 153, 2012.

[49] H.-W. Cho et al., “Pancreatic tumors: Emphasis on CT findings and pathologic classification,” *Korean J. Radiol.*, vol. 12, no. 6, p. 731, 2011.

[50] H. Zhang, C.-Y. Yu, and B. Singer, “Cell and tumor classification using gene expression data: Construction of forests,” *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 7, pp. 4168–4172, Apr. 2003.

[51] D. W. Scott et al., “High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements with diffuse large B-cell lymphoma morphology,” *Blood*, vol. 131, no. 18, pp. 2060–2064, May 2018.

[52] R. Boyar Cetinkaya, B. Aagnes, T. Å. Myklebust, and E. Thiis-Evensen, “Survival in neuroendocrine neoplasms; a report from a large Norwegian population-based study,” *Int. J. Cancer*, vol. 142, no. 6, pp. 1139–1147, Mar. 2018.

[53] N. C. F. Codella et al., “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),” in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on March 05,2025 at 02:25:40 UTC from IEEE Xplore. Restrictions apply.