DenseSAM: Semantic Enhance SAM For Efficient Dense Object Segmentation

Linyun Zhou[1] , **Jiacong Hu**[1] , **Shengxuming Zhang**[1,2] , **Xiangtong Du**[3] ,
Mingli Song[1] , **Xiuming Zhang**[4*] and **Zunlei Feng**[1,2,5*]
[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Xuzhou Medical University
[4]The First Affiliated Hospital, College of Medicine, Zhejiang University
[5]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

Abstract

Dense object segmentation is essential for various applications, particularly in pathology image and remote sensing image analysis. However, distinguishing numerous similar and densely packed objects in this task presents significant challenges. Several methods, including CNN- and ViT-based approaches, have been proposed to tackle these issues. Yet, models trained on limited datasets exhibit limited generalization ability. The Segment Anything Model (SAM) has recently achieved significant progress in zero-shot segmentation but relies heavily on precise positional guidance. However, providing numerous accurate location prompts in dense scenarios is time-consuming. To overcome this limitation, we conducted an in-depth exploration of the SAM mechanism and found that its strong generalization ability stems from the encoder's edge detection capability, which is semantically independent, making location prompts essential for segmentation. This insight inspired the development of DenseSAM, which replaces location prompts with semantic guidance for automatic segmentation in dense scenarios. Specifically, it uses local details to weaken the edges of background objects, leverages global context to enhance intra-class feature similarity, while further increasing contrast with the background, and integrates a dual-head decoding process to enable lightweight automatic semantic segmentation. Extensive experiments on pathology images demonstrate that DenseSAM delivers remarkable performance with minimal training parameters, providing a cost-effective and efficient solution. Moreover, experiments on remote sensing images further validate its excellent scalability, making DenseSAM suitable for various dense object segmentation domains. The code is available at https://github.com/imAzhou/DenseSAM.

*Corresponding Author. Email: zunleifeng@zju.edu.cn

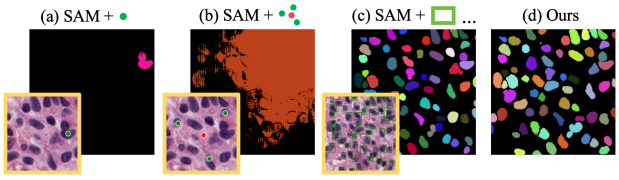

Figure 1: From left to right: SAM output mask prompted with one point, multiple points and all boxes respectively, the rightmost is our output without location prompt.

1 Introduction

Dense object segmentation is a critical visual task with broad applications in specialized fields such as cell segmentation in pathology image analysis. This task involves numerous similar and closely packed objects, posing challenges in adhesion and differentiation. Numerous methods have been proposed to tackle these challenges, including CNN-based [Tian *et al.*, 2020; Yoo *et al.*, 2019] and ViT-based [Prangemeier *et al.*, 2020; Lin *et al.*, 2024]. Although these methods have shown remarkable performance, they require extensive training resources and exhibit limited generalization ability.

Recently, the advent of the Segment Anything Model (SAM) [Kirillov *et al.*, 2023] has marked groundbreaking progress in zero-shot image segmentation. Leveraging a vast number of parameters and pre-training on the SA-1B segmentation dataset, SAM excels at accurately segmenting objects across diverse domains. However, despite its powerful capabilities, SAM remains significantly limited in dense object segmentation due to its heavy reliance on numerous accurate manual location prompts to achieve satisfactory results.

As shown in Fig.1 (a), when a single positive point is provided, two adjacent cells are segmented together. Furthermore, using a few positive and negative points still results in large areas of adhesion (b). Consequently, satisfactory results are achieved only when bounding box prompts are provided for each object (c). However, locating all objects in dense scenarios is inherently time-consuming and labor-intensive. This raises two key questions: *Why does SAM heavily rely on precise location prompts, and how can we achieve automatic and accurate segmentation of dense objects without relying on location prompts?*

To answer the first question, we delved deeply into the me-

chanics of SAM. First, we hypothesize that SAM's encoder detects all objects in an image but relies on location prompts as query tokens to employ the attention mechanism for locating image tokens corresponding to objects near the specified positions. Based on this hypothesis, we adopted two metrics to analyze the image tokens: (1) a local metric that calculates the variance of image tokens, and (2) a global metric that measures the cosine similarity between image tokens. As shown in Fig. 2(b), the image tokens output by the SAM encoder exhibit higher variance at the object edges and high similarity between the selected object token (green star) and other objects, including those with different semantics (green arrow), which validates the above hypothesis. After cross-attention in the SAM decoder (c), the variance of image tokens within the selected object becomes consistent, and the similarity with distant object tokens decreases. Therefore, we infer that SAM accurately identifies object edges but relies on precise location to compute attention between query and image tokens. More specifically, SAM's strong generalization ability stems from its encoder, while its decoder depends on positional cues to locate objects detected by the encoder, rather than on the semantic relationships between objects.

To address the second question, semantic guidance is proposed as a replacement for location prompts to achieve automatic segmentation. Specifically, we observe that in dense object segmentation, these numerous closely packed objects often share similar and homogeneous semantic features. This observation inspired us to leverage SAM's strong generalization capability in edge detection and replace location-based prompts with semantic guidance, enabling automatic and precise segmentation. Thus, an Efficient Semantic Injection (ESI) module is proposed to achieve this goal. Specifically, image tokens from the hidden layers of SAM's image encoder are utilized for local information extraction and global context contrastive learning. By using a local workflow to attenuate background object edges and a global workflow to enhance intra-class feature similarity while increasing contrast with the background, these semantic features are integrated into the decoding process. The metric results for both local and global analysis on the ESI output image tokens are shown in Fig. 2(d). As illustrated, the distribution of image tokens reveals a clear spatial correspondence between foreground and background semantic information. Additionally, a dual-head decoding structure is designed to simultaneously output objects and their boundaries, significantly improving the distinction between adjacent objects. By utilizing semantic guidance, along with the efficient dual-head structure, we enable lightweight automatic semantic segmentation with only a few trainable parameters.

Our contributions are summarized as follows: (1) We propose an automated, cost-efficient segmentation framework, DenseSAM, which is built upon a thorough analysis of SAM, demonstrating that its generalization ability arises from its precise edge detection. (2) A novel Efficient Semantic Injection (ESI) module is proposed that integrates local and global perspectives to enhance foreground-background distinction. (3) Extensive experiments showing DenseSAM's state-of-the-art performance with minimal parameters and strong generalization to new tasks.

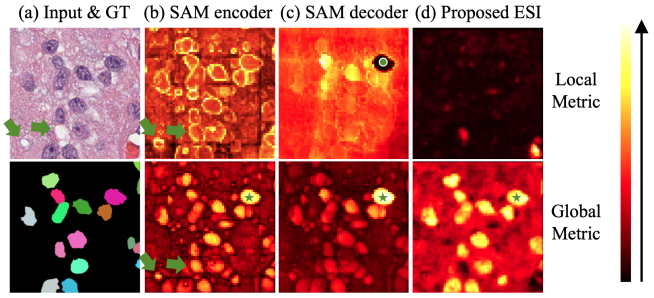

Figure 2: Heatmap in local and global metric. For ESI, brighter regions in local indicate background areas with higher variance, while darker regions in global reflect lower similarity to cell objects.

2 Related Work

Dense Object Segmentation. Dense object segmentation is a challenging task that deals with identifying and separating individual objects within a dense scenarios [Tu *et al.*, 2008; Li *et al.*, 2018; Lou *et al.*, 2012]. For specialized domain like pathology and remote sensing image analysis, its main challenges include not only segmenting numerous closely packed objects but also addressing the frequent occlusions, partial coverage, and objects adhesion, which significantly complicate the segmentation process. Prevailing approaches often enhance task-oriented segmentation by training models with extensive domain-specific samples to overcome these challenges. Depending on the model structure used, they can be roughly categorized as CNN-based, ViT-based, and a combination of both. CNN-based methods excel at capturing fine-grained local and multi-scale features through their hierarchical convolutional structures, making them ideal for distinguishing closely packed objects in dense scenarios [Tian *et al.*, 2020; Yoo *et al.*, 2019]. In contrast, ViT-based methods can handle larger receptive fields and effectively capture global contextual information, making them well-suited for managing complex scenes with significant contextual relationships [Wang *et al.*, 2024; Prangemeier *et al.*, 2020]. Building on the complementary strengths of CNNs and ViTs, various approaches have emerged that integrate both into a single framework to achieve enhanced performance [Li *et al.*, 2023; Hu *et al.*, 2023].

Even though these expert methods have shown promising results in specific fields, training them from scratch demands substantial resources. Moreover, they exhibit poor generalization across different domains, such as a model tailored for cell segmentation is unsuitable for building extraction.

Segment Anything Model and It's Variants. The Segment Anything Model (SAM) [Kirillov *et al.*, 2023] is an advanced universal segmentation model, which consists of an image encoder, a prompt encoder, and a mask decoder. Trained on 1.1 billion masks and 11 million images, SAM demonstrates impressive zero-shot capabilities and has inspired a surge of research into its potential across various downstream tasks. Depending on which components of SAM are tuned, we classify those variants into four types: a). Tuning whole SAM, b). Tuning minimal parameters, c). Training auxiliary network, d). Training auxiliary network without lo-

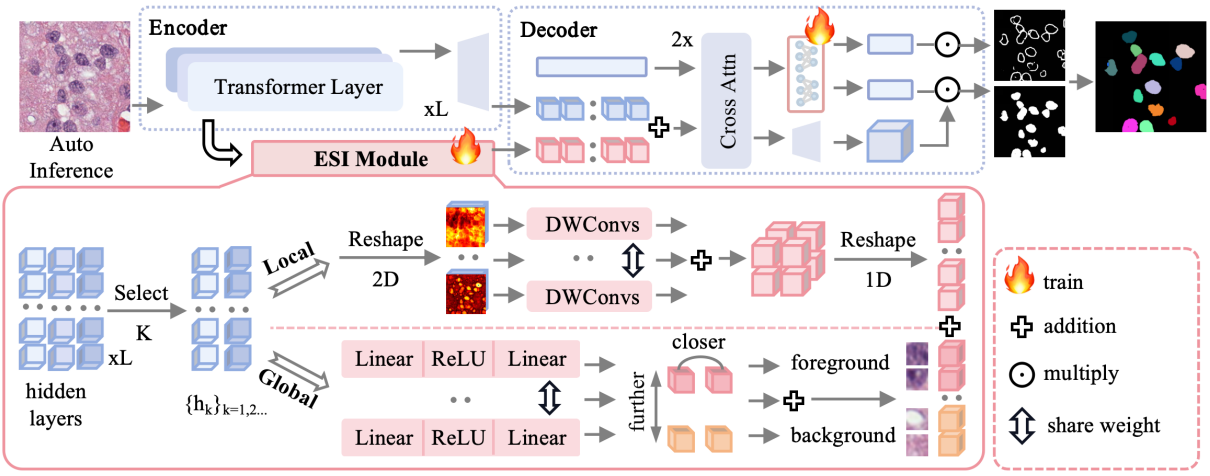

Figure 3: The proposed DenseSAM framework. The newly added ESI module injects semantic information into the decoding process, replacing SAM's location prompts to enable automatic segmentation.

cation prompts. For example, MedSAM [Ma *et al.*, 2024] fine-tuning the entire SAM on 1.57 million medical images for promptable medical segmentation. To enable efficient fine-tuning without relying on large datasets, Med-SA [Wu *et al.*, 2023] fine-tunes minimal additional parameters in the SAM image encoder for efficient adaptation to medical imaging, while SAM-PARSER [Peng *et al.*, 2024] and CAT-SAM [Xiao *et al.*, 2024] use similar techniques for various domains, including medical imaging, remote sensing, and natural images. More advanced approaches [Hu *et al.*, 2024c; Zhang *et al.*, 2024b] use or train auxiliary networks (such as language models or expert models) to generate location prompts automatically, replacing the time-consuming manual intervention. Apart from using location prompts, some methods [Zhang *et al.*, 2024a; Chen *et al.*, 2024] eliminate them altogether by training auxiliary networks to generate masks directly, while either training or freezing the image encoder.

We propose a new type that achieves automatic segmentation by training only a few semantic parameters, offering a cost-effective and efficient solution for adapting SAM to dense object segmentation.

3 Why Dose SAM Rely on Location Prompts?

Inspired by studies analyzing the working mechanisms of visual models [Selvaraju *et al.*, 2017; Hu *et al.*, 2024b; Hu *et al.*, 2024a], in this section, we aim to understand the internal mechanism of SAM and investigate its reliance on precise location prompts. Based on SAM's architecture, which consists of a heavyweight image encoder with 640M parameters and a lightweight mask decoder and prompt encoder with a combined 4M parameters, we hypothesize that SAM's encoder detects all objects in an image but requires location prompts as query tokens to identify image tokens corresponding to objects near the specified positions. To validate this hypothesis, two analytical metrics are introduced to examine the distributional changes of image tokens within the encoder and decoder: (1) a local metric that calculates the variance of image tokens and (2) a global metric that measures the cosine similarity between image tokens.

First, the feature extraction mechanism within the SAM image encoder was examined, focusing on its reliance on a series of attention operations applied to flattened one-dimensional image patch embeddings. This process can be represented by the following equation:

$$E_{l+1} = f'_{proj}(f_{attn}(f_{proj}(E_l))), l \in \{1, ..., L\}, \quad (1)$$

where E_l represents the encoder layer's output image tokens with dimensions (h, w, c), h, w and c denote the height, width, and channel, respectively. The total number of encoder layers is denoted by L. The projection function, f_{proj}, maps the flattened image tokens $(h * w, c)$ before it is fed into the attention operation f_{attn}. After computing attention, the image tokens are reshaped back into two-dimensional space and projected by f'_{proj}, returning E_l to (h, w, c).

We visualized the layer's output E_l using heatmaps for analysis. Keeping h and w constant, we calculated the variance along the channel c to generate the local metric heatmap in Fig.2, revealing that image tokens near object edges exhibit greater variability in their channel features. Then, we calculated the cosine similarity between a specific token (the green star in Fig.2) and all other tokens, generating a similarity matrix of shape (h, w), which forms the global metric heatmap displayed in Fig.2. As observed, tokens within the same object exhibit the highest similarity, while tokens from other objects, regardless of distance or semantic class, also display high similarity, as indicated by the green arrows in Fig.2. Thus, it is concluded that *the SAM encoder extracts the edges of all objects and preserves object-centric similarity, irrespective of their semantics.*

Next, the image token update process during SAM's decoding phase was analyzed using the same statistical approach. In this phase, image tokens are decoded alongside the location prompts, as demonstrated below:

$$R_{k+1}, Q_{k+1} = f_{twa}(R_k, Q_k), k \in \{0, 1\}, \quad (2)$$

$$M = f_{decode}(R_2, Q_2), \quad (3)$$

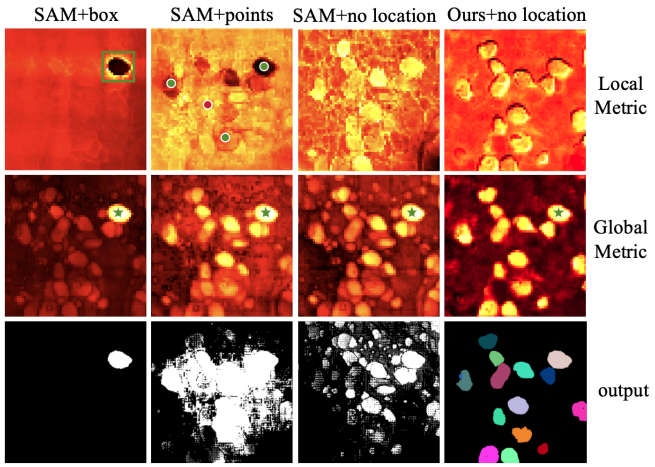

Figure 4: Analysis of typical cases in SAM and our decoder using local and global metrics.

Here, R_k and Q_k represent the one-dimensional image tokens and the query tokens containing location prompts in the SAM mask decoder, respectively. The core module of the mask decoder, f_{twa}, refers to the Two-Way Attention block, which performs k rounds of cross-attention to update the feature values of R_k and Q_k. The updated R_2 and Q_2 are then passed into the decoding function f_{decode}, where R_2 is reshaped into a two-dimensional space and upsampled, while Q_2 undergoes processing through a set of MLPs for channel mapping. The final output masks M are generated by combining the upsampled R_2 and the mapped Q_2.

The same local and global analyses are performed on the updated R_2 under one-point location prompts. As shown in Fig.2, by incorporating the location of the green star into the query tokens, the variance of image tokens belonging to the green star's object becomes more consistent after two iterations of cross-attention between the query tokens and image tokens, while the cosine similarity with distant tokens decreases. More typical cases on the updated R_2 are provided in Fig.4. The analysis results further confirm our hypothesis that *the lightweight SAM decoder relies heavily on precise and unambiguous location prompts to identify corresponding objects, and without such prompts, SAM performs significantly deteriorates with densely packed homogeneous objects.*

The above analysis inspired us to explore whether semantic guidance could replace location prompts for dense objects with homogeneous semantics to achieve a lightweight, automatic, and efficient segmentation tool.

4 How To Enhance SAM for Dense Objects?

Since SAM's strong generalization ability stems from its encoder, while its decoder relies on location prompts to segment objects, the presence of numerous semantically similar objects in dense object scenes inspired us to replace location prompts with semantic guidance in the decoder.

To enable semantics-guided automatic segmentation, an Efficient Semantic Injection (ESI) module is proposed, which includes both local and global workflows. As shown in Fig.3, the image tokens from the hidden layers of the SAM encoder are leveraged for local details extraction and global context comparison, given its capability to detect object edges, as analyzed in Section 3. In the local workflow, after reconstructing the image tokens into a two-dimensional (2D) space, a depth-wise convolutional layer is incorporated to capture local information, as demonstrated to be beneficial in previous works [Yuan *et al.*, 2021; Li *et al.*, 2021]. The process is expressed by the following formula:

$$R_k^{local} = f_{local}(\mathcal{T}_{reshape}(E_k)), k \in \{1, ..., K\}, \quad (4)$$

$$f_{local}(x) = Conv(ReLU(DWConv(x))), \quad (5)$$

where f_{local} denotes the local module to process image tokens E_k. For different layer's output E_k, the parameters of the f_{local} module are shared, where K is a hyper-parameter.

In the global workflow, supervised contrastive learning [Khosla *et al.*, 2020] is employed to enhance the similarity within the same semantic category and amplify the distinction between different categories. The process is expressed by the following formula:

$$R_k^{global} = f_{global}(E_k), k \in \{1, ..., K\}, \quad (6)$$

$$f_{global}(x) = Proj(ReLU(Proj(x))), \quad (7)$$

where f_{global} denotes the global module, and the weights are shared, similar to those in f_{local}.

In summary, the semantic injection process of ESI can be expressed as follows:

$$R_{ESI} = Sum(Flatten(R_k^{local}), R_k^{global}), \quad (8)$$

$$R_2^{new} = R_2^{old} + R_{ESI}. \quad (9)$$

Subsequently, the R_2^{new} is fed into the decoder for cross-attention with the query tokens. The analysis results on both local and global metrics validate the effectiveness of the proposed ESI module. As shown in Fig.2, the output R_{ESI} increases the variance of background object tokens and the similarity between foreground tokens increases, while their similarity with background tokens is significantly reduced.

To refine the distinction between overlapping objects, we have innovatively designed a Dual-Head (DH) structure, enabling the simultaneous output of both object and its boundaries. Specifically, without location prompts, we incorporate semantic embeddings into the query tokens and use two sets of MLPs for mapping the query tokens. The process can be expressed as following:

$$M_b, M_o = f_{decode}^{DH}(R_2^{new}, Q_2^{'}), \quad (10)$$

where $Q_2^{'}$ denotes the query tokens incorporated with newly added semantic embedding, which is processed by f_{decode}^{DH}, the proposed Dual-Head decoding structure that includes two sets of MLPs. The output results, M_b and M_o, represent the masks of objects and their boundaries, respectively. A marker-controlled watershed algorithm is applied to post-process these two masks, as in Al-Net [Zhao *et al.*, 2021], yielding the final instance segmentation results.

The modules we proposed require only about 3M trainable parameters, making lightweight automatic semantic segmentation feasible. In the following experimental section, we will demonstrate the effectiveness of the proposed DenseSAM on six datasets and conduct an ablation study to substantiate the validity of each component.

Type	Method	Year	Trainable Params	CoNIC			CPM17			MoNuSeg		
				Dice	AJI	PQ	Dice	AJI	PQ	Dice	AJI	PQ
NonSAM-based	PseudoEdgeNet	2019	43.43M	44.98	25.76	6.74	72.46	49.42	45.10	56.68	29.03	31.86
	Scribble2Label	2020	33.82M	67.04	48.62	45.84	70.82	50.26	45.99	73.44	54.20	51.78
	C2FNet	2020	11.50M	33.18	13.25	4.00	64.39	42.11	27.06	63.96	40.51	34.12
	SSL-Net	2020	-	65.98	43.04	44.08	70.28	49.4	42.68	72.51	51.69	49.97
	SC-Net	2023	81.23M	66.72	47.13	42.59	73.73	51.69	49.66	74.41	56.20	53.19
	BoNuS	2024	52.93M	70.33	52.13	46.30	75.13	54.54	49.91	<u>76.73</u>	<u>60.73</u>	<u>55.43</u>
SAM-based	Med-SA$_{+point}$	2023	13.00M	70.60*	52.98*	48.69*	77.28*	51.23*	49.67*	74.11*	47.62*	47.37*
	Med-SA	2023	13.00M	41.75*	38.97*	34.12*	71.35*	42.08*	39.16*	71.30*	33.27*	35.69*
	UN-SAM	2024	319.09M	<u>71.58*</u>	<u>55.62*</u>	<u>51.97*</u>	<u>78.99*</u>	<u>60.74*</u>	<u>51.96*</u>	75.31*	58.02*	49.13*
	DenseSAM (ours)		3.11M	**78.19**	**63.76**	**61.17**	**84.14**	**71.84**	**66.87**	**80.41**	**66.50**	**61.44**

Table 1: Comparison with SOTA methods on CoNIC, CPM17 and MoNuSeg datasets. **Bold** indicates the best and <u>underlined</u> denotes the second-best. "-" indicates missing data in original paper, while "*" denotes reproduced results. All values are in %.

5 Experiments

5.1 Experimental Setup

Datasets. To evaluate the performance of the proposed DenseSAM in dense segmentation of pathology images, we used three commonly used pathology datasets. Additionally, to demonstrate the applicability of DenseSAM in other dense segmentation domains, we also used three common remote sensing datasets, showcasing DenseSAM's robust dense segmentation capability. The six datasets are as follows:

Pathology datasets: CPM17 [Vu *et al.*, 2019] includes 32 / 32 pathology images for train / test, with sizes of 500 × 500 or 600 × 600 pixels. We resize each image to 1024 × 1024 pixels, and then crop it into 512 × 512 pixel patches with no overlap. CoNIC [Graham *et al.*, 2021b] consists of 4981 image patches, each sized 256 × 256. Following BoNuS [Lin *et al.*, 2024] we randomly split all images into 7:1:2 ratio, resulting in 3486 / 997 / 498 for train / validation / test. MoNuSeg [Kumar *et al.*, 2017] comprises 44 images with each of size 1000 × 1000 pixels. The dataset has 30 / 14 images for train / test. We randomly split the train subset into 24 / 6 images for train / validation. Each image is resized to 1024 × 1024, and then crop to 256 × 256 without overlap.

Remote sensing datasets: WHU Building [Ji *et al.*, 2018] has a ground resolution of 0.3 meters and an image size of 512 × 512 pixels. It contains 4736 / 1036 / 2416 images for train / validation /test. Inria Building [Maggiori *et al.*, 2017] contains 360 images collected from five cities at a 30cm resolution. We process it consistent with UANet [Li *et al.*, 2024] and crop them in to 512 × 512 pixels, resulting in 9737 / 1942 images for train / validation. Massachusetts Building [Mnih, 2013] owns 151 aerial images with spatial resolution 1 meters and an image size of 1500 × 1500 pixels. we crop the images into 500 × 500 pixels, get 1233 / 36 / 90 images for train / validation / test.

Metrics. For semantic segmentation, we adopt three pixel-level metrics: Intersection over Union (IoU), F1 score (F1), and precision (Pre.). We calculate the mean value of the background and foreground in these metrics, consistent with UANet [Li *et al.*, 2024]. For instance segmentation, three object-level evaluation metrics is utilized: object-level Dice coefficient (Dice), aggregated Jaccard index (AJI), and panoptic quality (PQ), align with BoNuS [Lin *et al.*, 2024].

Implementation Details. By default, we use the ViT-H type for the image encoder, unless otherwise specified. For the loss function, we use a linear combination of Dice loss, BCE loss and contrastive loss. We adopt Adam optimizer and training ranges from 10 to 30 epochs.

5.2 Quantitative Results

In this section, we compare the DenseSAM against several NonSAM-based and SAM-based methods on pathology and remote sensing datasets. To ensure fairness, we reference results from the original papers or reproduced SAM-based methods using open-source code when results were missing.

Pathology Instance Segmentation. To validate the effective of the proposed DenseSAM, we compare three instance metrics on the CoNIC, CPM17 and MoNuSeg datasets. As shown in Table 1, DenseSAM outperforms expert networks including PseudoEdgeNet [Yoo *et al.*, 2019], Scribble2Label [Lee and Jeong, 2020], C2FNet [Tian *et al.*, 2020], SSL-Net [Xie *et al.*, 2020], SC-Net [Lin *et al.*, 2023], and BoNuS [Lin *et al.*, 2024]. These expert networks are specializing in cell segmentation tasks which training parameters range from 10 to 100 million, far beyond of DenseSAM.

For SAM-based methods, Med-SA [Wu *et al.*, 2023] and UN-SAM [Chen *et al.*, 2024] is selected for comparison. Table 1 shows that Med-SA with point prompts performed rather poorly on AJI metric, which is the primary metric for evaluating whether adherent cells are separated. Moreover, when the point prompts are removed, Med-SA's performance significantly deteriorates across all metrics. UN-SAM trains a large number of parameters to eliminate the need for location prompts. However, compared to UN-SAM, DenseSAM achieves more impressive performance across all metrics with fewer trainable parameters.

It is worth mentioning that CoNIC is extracted from the Lizard [Graham *et al.*, 2021a] dataset, which includes around fifty thousand nuclei from 16 different centers, representing

Type	Method	Year	Trainable Params	WHU			Inria			Massachusetts		
				IoU	F1	Pre.	IoU	F1	Pre.	IoU	F1	Pre.
NonSAM-based	UNet	2015	24.71M	85.92	92.39	91.78	74.40	85.32	86.39	68.48	81.47	80.99
	Uniformer	2023	79.64M	90.55	95.04	95.01	84.37	91.52	91.84	73.80	84.92	87.60
	FSAU-Net	2024	28.27M	91.73	93.67	93.60	80.43	90.78	90.71	-	-	-
	UANet	2024	15.60M	92.15	95.91	95.96	83.08	90.76	92.04	76.41	86.63	87.94
	RSBuilding	2024	98.70M	92.15	95.88	95.93	82.68	90.52	91.40	-	-	-
	GLGFF-Net	2024	136.07M	91.30	95.45	95.01	84.94	91.85	91.59	75.33	85.93	85.03
SAM-based	SAM-PARSER$_{+box}$	2024	3.96M	81.80	88.40	-	-	-	-	-	-	-
	CAT-SAM$_{+box}$	2024	1.90M	93.60	96.69*	96.70*	83.29*	90.64*	90.09*	80.75*	88.90*	89.38*
	RSAM-Seg	2024	350.29M	92.83*	96.22*	96.38*	84.24	83.37	83.90	80.16*	88.51*	89.30*
	DenseSAM (ours)		2.97M	**95.23**	**97.53**	**97.64**	**85.79**	**92.18**	**92.03**	**82.03**	**89.77**	**89.50**

Table 2: Comparison with SOTA methods on remote sensing datasets. **Bold** indicates the best and underlined denotes the second-best. "-" indicates missing data in original paper, while "*" denotes reproduced results. All values are in %.

	SAM decoder	Our decoder	- Local	- Global	- DH
Params	3.63M	3.11M	0.48M	0.28M	0.14M
Dice	74.25	84.01	80.77	74.37	69.60
AJI	57.56	71.62	66.84	59.79	52.43
PQ	39.30	66.12	60.46	55.75	23.67

Table 3: Ablation study on the CPM17 dataset. The SAM decoder is retrained as the baseline. The proposed ESI module (including local and global workflows) and Dual-Head (DH) are sequentially removed to verify their effectiveness.

diverse cell segmentation domains. The proposed DenseSAM demonstrated superior performance compared to expert models on the CoNIC.

Remote Sensing Semantic Segmentation. Table 2 shows the comparison results on three remote sensing datasets. UNet [Ronneberger *et al.*, 2015] and Uniformer [Li *et al.*, 2023] are classic segmentation networks used in general domains, and FSAU-Net [Hu *et al.*, 2023], UANet [Li *et al.*, 2024], RSBuilding [Wang *et al.*, 2024] and GLGFF-Net [Fu *et al.*, 2024] represent SOTA expert networks for building extraction in recent years. Table 2 shows that DenseSAM, which leverages SAM's powerful capability while training only 0.6% of its parameters, surpasses domain-specific models that require training a large number of parameters.

For SAM-based methods, SAM-PARSER [Peng *et al.*, 2024] and CAT-SAM [Xiao *et al.*, 2024] require training a small number of additional parameters but still relying on manual location prompts. RSAM-Seg [Zhang *et al.*, 2024a] eliminates the need for location prompts but incurs significant training costs. From Table 2, it can be observed that DenseSAM surpasses these SAM-based methods without requiring manual prompts or incurring significant training costs.

5.3 Qualitative Results

We visualized representative results of DenseSAM across six datasets. Fig. 6 shown that DenseSAM can identify most dense objects without location prompts, even for very small areas, such as the small building in the Massachusetts. However, the proposed method has limitations. In Fig. 6, cases of poor performance, such as undifferentiated adhesive cells, background misidentified as cells, and shadow-obscured buildings, are highlighted with red circles.

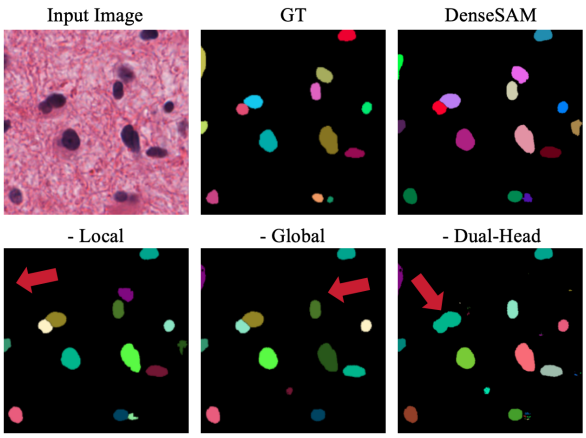

Figure 5: Visualization results of the ablation study on CPM17 dataset, with each module removed progressively.

5.4 Ablation Study

In this section, we conduct ablation studies on CPM17. By sequentially removing each of the proposed modules, the experimental results clearly demonstrate their effectiveness. As shown in Table 3, SAM exhibits strong object detection capabilities. However, even with full retraining, its decoder still delivers subpar performance on instance segmentation metrics such as AJI and PQ. DenseSAM leverages SAM's powerful edge detection and object-centric similarity capabilities to achieve SOTA performance with minimal training cost. The ablation visualization in Fig.5 further demonstrates the effectiveness of the proposed modules. Red arrows indicate issues caused by removing specific modules, including missed cell detection and failure to separate adherent cells.

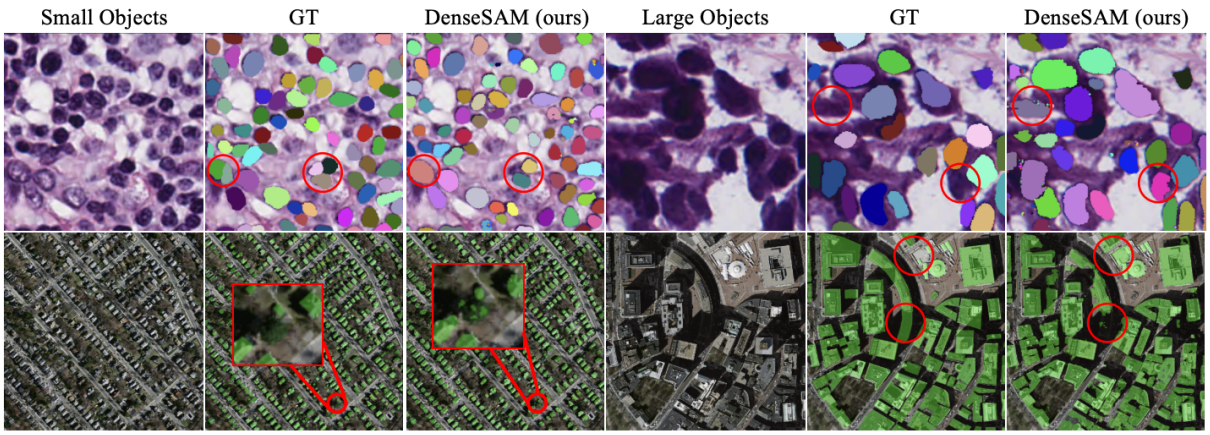

Figure 6: Representative results of DenseSAM in MoNuSeg and Massachusetts Building datasets. The left three columns shows the GT and prediction results for small objects, while the right three shows for large. Poor segmentation cases are highlighted with red circles.

Figure 7: Visualization of SAM encoder analysis results in both local and global metrics for natural images (sampled from MSCOCO).

6 Limitations and Future Work

The proposed DenseSAM, while performing well across multiple datasets for automatic dense object segmentation, has some limitations that warrant further exploration. Firstly, DenseSAM relies on pixel-level annotated masks for supervised training. Given that homogeneous dense objects often follow unified prior patterns, future research could investigate the incorporation of domain-specific knowledge to mitigate this dependency. Secondly, DenseSAM is limited to segmenting semantic masks for specific classes and does not handle multi-class segmentation tasks. Future research could focus on addressing this limitation.

Beyond specialized domains, the pattern analysis and potential applications of SAM in natural images are worth exploring. The analysis results of MSCOCO [Lin *et al.*, 2014] samples are shown in Fig. 7. As shown, SAM demonstrates strong edge detection ability in natural images, but object-centric similarity patterns appear only in the densely packed fruits in the fourth column. These characteristics highlight the need for further investigation into maximizing SAM's capabilities with minimal training costs.

7 Conclusion

In this paper, we present a novel, cost-effective, and efficient solution for automatic dense object segmentation. Using the introduced local and global metrics, we thoroughly analyzed SAM's mechanism and discovered that its strong generalization ability stems from its precise edge detection capabilities. Moreover, SAM exhibits robust object-centric feature similarity in dense scenes, both of which are independent of object semantics. This inspired us to replace positional cues with semantic guidance for dense object segmentation, particularly in scenarios involving numerous similar and homogeneous objects. Therefore, we propose DenseSAM, which utilizes the ESI module to effectively inject semantic information into the decoding process, replacing the time-consuming and labor-intensive manual location prompts. Through local and global workflows, the ESI module effectively distinguishes between foreground and background objects. Equipped with a dual-head structure, DenseSAM enhances the differentiation of overlapping object edges, enabling lightweight and automatic instance segmentation. Extensive experiments across six datasets for pathology and remote sensing image analysis demonstrate that DenseSAM achieves state-of-the-art performance in both semantic segmentation and instance segmentation metrics, with only $\sim$3M trainable parameters, making it an economical and practical tool for dense object segmentation. In future research, it is anticipated that the in-depth exploration of the SAM mechanism and the efficient, low-cost semantic injection method in DenseSAM will inspire more innovative ideas.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62376248), the Huadong Medicine Joint Fund of the Zhejiang Provincial Natural Science Foundation of China (LHDMZ25H160002), the Zhejiang Province Health Major Science and Technology Program of National Health Commission Scientific Research Fund (No. WKJ-ZJ-2426) and Information Technology Center, ZheJiang University.

References

[Chen *et al.*, 2024] Zhen Chen, Qing Xu, Xinyu Liu, and Yixuan Yuan. Un-sam: Universal prompt-free segmentation for generalized nuclei images. *arXiv preprint arXiv:2402.16663*, 2024.

[Fu *et al.*, 2024] Wei Fu, Kai Xie, and Leyuan Fang. Complementarity-aware local-global feature fusion network for building extraction in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Graham *et al.*, 2021a] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 684–693, 2021.

[Graham *et al.*, 2021b] Simon Graham, Mostafa Jahanifar, Quoc Dang Vu, Giorgos Hadjigeorghiou, Thomas Leech, David Snead, Shan E Ahmed Raza, Fayyaz Minhas, and Nasir Rajpoot. Conic: Colon nuclei identification and counting challenge 2022. *arXiv preprint arXiv:2111.14485*, 2021.

[Hu *et al.*, 2023] Minghong Hu, Jiatian Li, Yunfei Zhao, Mei Lu, and Wen Li. Fsau-net: a network for extracting buildings from remote sensing imagery using feature self-attention. *International Journal of Remote Sensing*, 44(5):1643–1664, 2023.

[Hu *et al.*, 2024a] Jiacong Hu, Hao Chen, Kejia Chen, Yang Gao, Jingwen Ye, Xingen Wang, Mingli Song, and Zunlei Feng. Transformer doctor: Diagnosing and treating vision transformers. *Advances in Neural Information Processing Systems*, 37:54026–54053, 2024.

[Hu *et al.*, 2024b] Jiacong Hu, Jing Gao, Jingwen Ye, Yang Gao, Xingen Wang, Zunlei Feng, and Mingli Song. Model lego: Creating models like disassembling and assembling building blocks. *Advances in Neural Information Processing Systems*, 37:127711–127738, 2024.

[Hu *et al.*, 2024c] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 11, pages 12511–12518, 2024.

[Ji *et al.*, 2018] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Kumar *et al.*, 2017] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

[Lee and Jeong, 2020] Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 14–23. Springer, 2020.

[Li *et al.*, 2018] Qingpeng Li, Yunhong Wang, Qingjie Liu, and Wei Wang. Hough transform guided deep feature extraction for dense building detection in remote sensing images. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1872–1876. IEEE, 2018.

[Li *et al.*, 2021] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.

[Li *et al.*, 2023] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023.

[Li *et al.*, 2024] Jiepan Li, Wei He, Weinan Cao, Liangpei Zhang, and Hongyan Zhang. Uanet: An uncertainty-aware network for building extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2023] Yi Lin, Zhiyong Qu, Hao Chen, Zhongke Gao, Yuexiang Li, Lili Xia, Kai Ma, Yefeng Zheng, and Kwang-Ting Cheng. Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. *Medical Image Analysis*, 89:102933, 2023.

[Lin *et al.*, 2024] Yi Lin, Zeyu Wang, Dong Zhang, Kwang-Ting Cheng, and Hao Chen. Bonus: Boundary mining for nuclei segmentation with partial point labels. *IEEE Transactions on Medical Imaging*, 2024.

[Lou *et al.*, 2012] Xinghua Lou, Ullrich Koethe, Jochen Wittbrodt, and Fred A Hamprecht. Learning to segment dense

cell nuclei with shape prior. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1012–1018. IEEE, 2012.

[Ma *et al.*, 2024] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[Maggiori *et al.*, 2017] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017.

[Mnih, 2013] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[Peng *et al.*, 2024] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 5, pages 4515–4523, 2024.

[Prangemeier *et al.*, 2020] Tim Prangemeier, Christoph Reich, and Heinz Koeppl. Attention-based transformers for instance segmentation of cells in microstructures. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 700–707. IEEE, 2020.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Tian *et al.*, 2020] Kuan Tian, Jun Zhang, Haocheng Shen, Kezhou Yan, Pei Dong, Jianhua Yao, Shannon Che, Pifu Luo, and Xiao Han. Weakly-supervised nucleus segmentation based on point annotations: A coarse-to-fine self-stimulated learning strategy. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 299–308. Springer, 2020.

[Tu *et al.*, 2008] Peter Tu, Thomas Sebastian, Gianfranco Doretto, Nils Krahnstoever, Jens Rittscher, and Ting Yu. Unified crowd segmentation. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pages 691–704. Springer, 2008.

[Vu *et al.*, 2019] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7:433738, 2019.

[Wang *et al.*, 2024] Mingze Wang, Keyan Chen, Lili Su, Cilin Yan, Sheng Xu, Haotian Zhang, Pengcheng Yuan, Xiaolong Jiang, and Baochang Zhang. Rsbuilding: Towards general remote sensing image building extraction and change detection with foundation model, 2024.

[Wu *et al.*, 2023] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[Xiao *et al.*, 2024] Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Ruijie Ren, Xiaoqin Zhang, and Shijian Lu. Catsam: Conditional tuning network for few-shot adaptation of segmentation anything model. *arXiv preprint arXiv:2402.03631*, 2024.

[Xie *et al.*, 2020] Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. Instance-aware self-supervised learning for nuclei segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 341–350. Springer, 2020.

[Yoo *et al.*, 2019] Inwan Yoo, Donggeun Yoo, and Kyunghyun Paeng. Pseudoedgenet: Nuclei segmentation only with point annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 731–739. Springer, 2019.

[Yuan *et al.*, 2021] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34:7281–7293, 2021.

[Zhang *et al.*, 2024a] Jie Zhang, Xubing Yang, Rui Jiang, Wei Shao, and Li Zhang. Rsam-seg: A sam-based approach with prior knowledge integration for remote sensing image semantic segmentation. *arXiv preprint arXiv:2402.19004*, 2024.

[Zhang *et al.*, 2024b] Xin Zhang, Yu Liu, Yuming Lin, Qingmin Liao, and Yong Li. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 20, pages 22520–22528, 2024.

[Zhao *et al.*, 2021] Jing Zhao, Yong-Jun He, Si-Qi Zhao, Jin-Jie Huang, and Wang-Meng Zuo. Al-net: Attention learning network based on multi-task learning for cervical nucleus segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2693–2702, 2021.