Foundation Models for Image Segmentation: Challenges and Advances

Rafi Ibn Sultan

Department of Computer Science

Wayne State University

hm4013@wayne.edu

Abstract

The field of Natural Language Processing (NLP) has already embraced the era of foundation models, while computer vision is only beginning to tap into this vast reservoir of potential. The swift evolution from traditional Convolutional Neural Network (CNN) and Vision Transformer (ViT)-based models to those built on foundation models marks a significant shift in the landscape of computer vision. This review delves into the impact of Foundation Models, with a particular emphasis on Vision Foundation Models (VFM) and Vision Language Foundation Models (VLM), within the domain of segmentation. It concentrates on the methodologies employed by the Segment Anything Model (SAM), exploring its applications, the limitations it faces, and the potential paths for future development and enhancement in segmentation tasks. In this review, we concentrate on two distinct yet related areas of image segmentation: geographical imagery and medical imagery. Despite their distinct applications in real-world scenarios, these domains share common methodologies for problem-solving. They emphasize the vital role of image segmentation technology across various fields, each presenting unique challenges and demands. The progression in foundation models offers substantial benefits to these areas, highlighting the importance of tailored approaches to meet their specific segmentation needs.

Literature survey. Winter 2024 PhD qualifying exam. Do not distribute.

Contents

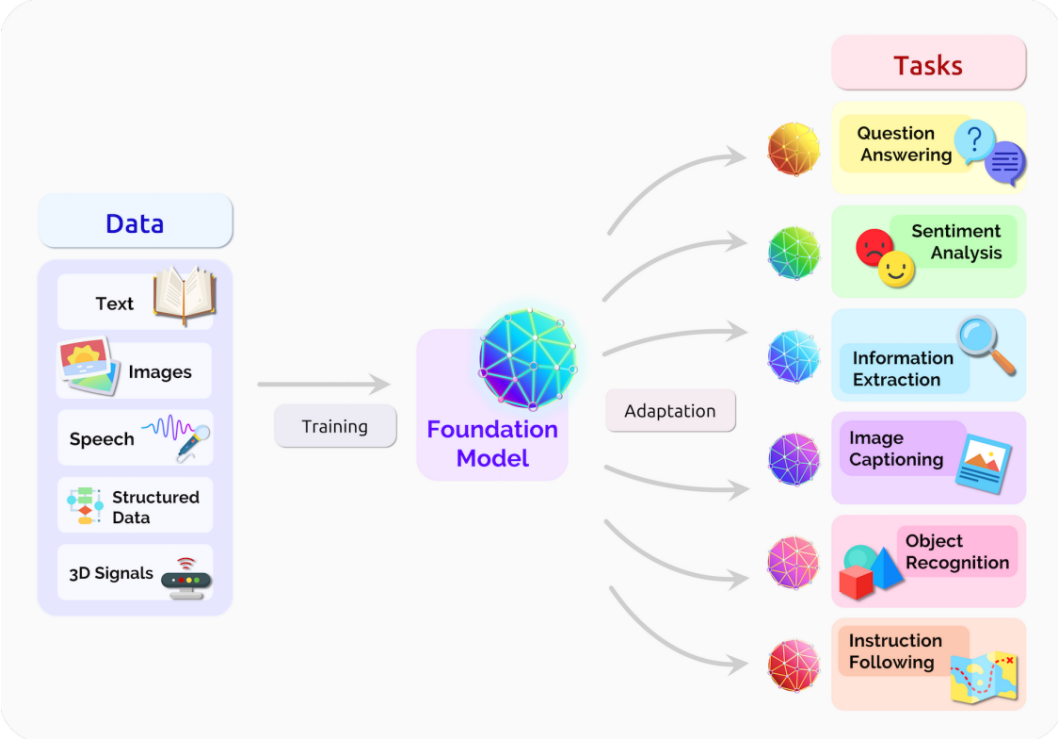


Figure 1: The concept of foundation model in a general sense. Based on the data it has been trained, it can generalize to a variety of downstream tasks [1].

1 Introduction

Large Language Models (LLMs), the precursors to foundation models, have markedly transformed the field of Natural Language Processing (NLP), as emphasized in recent analyses [1]. The shift from developing models from scratch for distinct tasks towards leveraging pre-trained foundation models like GPT series [2] for a multitude of downstream applications is becoming increasingly prevalent [3]. Innovations in generative AI, such as ChatGPT, which is built upon the GPT series, have demonstrated remarkable capabilities on numerous tasks from carefully given prompts. Foundation models, having been trained on extensive and carefully curated datasets, remove the requirement for continuous, task-specific retraining for downstream tasks. Instead, they primarily rely on users providing well-crafted prompts to guide the model's application to specific tasks. Researchers either employ these foundation models directly or employ the learned knowledge to other downstream tasks requiring minimal adjustments. Fig. 1 visually represents the conceptual framework of a foundation model, detailing both its training process and its application across diverse tasks in a generalized manner.

The practice of leveraging techniques and ideas from LLMs for vision-related tasks, leading to the creation of promptable Vision Foundation Models (VFM), is becoming increasingly popular. As research moves forward, researchers are turning to VFMs for a broad spectrum of computer vision challenges, including classification, segmentation, object detection, and captioning. Among these tasks, image segmentation emerges as a notably demanding yet crucial field, now venturing into the realm of leveraging promotable foundation models. The concept revolves around the ability of a single foundation model to accurately segment the region of interest (ROI) across diverse image modalities by simply receiving appropriate prompts, showcasing significant promise for advancing image segmentation. With the emergence of works in both of these domains, the field of foundation models is witnessing the introduction of Vision Language Models (VLMs) as well, a recent application that merges language and vision modalities to tackle computer vision tasks, including segmentation. This development marks a significant step forward in the research domain, offering new possibilities for enhancing task performance by leveraging the combined strengths of both modalities.

Traditionally, image segmentation relied heavily on Convolutional Neural Network (CNN)-based frameworks [4, 3, 5, 6]. Encoder-decoder-based architectures particularly inspired by UNet [7] has become the standard approach for a variety of segmentation tasks across domains, including natural, medical, and remote sensing images. Further, the introduction of Vision Transformers (ViT) [8] brought UNETR [9], borrowing principles from the NLP domain, addressed a significant limitation of CNNs: their difficulty in capturing long-range dependencies within images. These UNETR-based works focus on segmentation using the attention mechanism that is introduced in the NLP domain. Thus, ViT's self-attention mechanisms in UNETR-based models introduced a new horizon for segmentation models, leading to the development of hybrid approaches that combine the strengths of CNNs and ViTs. Beyond building models from the ground up, the field has also explored transfer learning for image segmentation, leveraging pre-trained models to alleviate the computational burden of training for specific tasks.

However, traditional supervised segmentation approaches need a myriad of supervised labels of data and most of the time these approaches aren't scalable at various levels. This may be due to both algorithmic and data challenges: supervised learning algorithms often learn shortcut features to achieve a high performance where the training sets are available whereas it may not generalize well. Transfer learning algorithms can be leveraged to mitigate this issue [10], but it still needs a

fair amount of labeled image data to fine-tune the segmentation models. This can be unscalable and limited by the availability of labeled data sets in most cases.

Vision foundation model-based label-efficient semantic segmentation. The emergence of vision foundation models [11] marks a significant advancement in the evolution of segmentation models, introducing robust zero-shot capabilities and offering versatile prompt-based interactions. The introduction of the Segment Anything Model (SAM) [11] ushers in a new era for promptable VFM in image segmentation. SAM, one of the very first segmentation-based VFM is designed to perform segmentation on a wide range of tasks based on prompts provided, both familiar and novel [12]. This paradigm shift reduces the necessity for task-specific model training, proposing a universal solution for image segmentation across various settings. Since the introduction of SAM, researchers have been utilizing the capabilities of SAM extensively as much as possible. This review aims to thoroughly examine the capabilities of foundation models, particularly the Segment Anything Model, in the context of image segmentation. It will critically evaluate their performance claims and identify opportunities for further improvements in downstream applications.

The structure of this review paper is divided into six main sections. Section 2 provides a comprehensive overview of foundation models, laying the groundwork for further discussion. Following this, the following two sections (Section 3 and Section 4) delve into the current state of research, exploring the challenges, applications, and advancements of foundation models within the realms of geographical and medical imagery segmentation respectively. Then Section 5 discusses the VLMS activities in medical image segmentation. Concluding the paper, Section 6 presents prospective avenues for the evolution of foundation models in the highlighted domains, offering insights into potential future developments and innovations.

2 Foundation Models

Foundation models represent a transformative approach to constructing artificial intelligence systems capable of adapting to a wide array of downstream tasks based on prompts. This methodology relies on training expansive neural networks with vast datasets, frequently employing self-supervised learning methods [1]. Such a strategy enables these models to acquire general representations and skills that are applicable across various domains and applications.

Technique	Traditional Model	Foundation Model
Training	Requires task-specific training	Pre-trained once for numerous tasks
Training method	Lots of human-supervised training	Pre-trained unsupervised or self-supervised training
Interaction	No interaction with the model	Users can interact and provide context
Robustness	Susceptible to change in data	Robust change in data
Adaption	Impossible or hard to adapt to different tasks	Easy to adapt to different tasks sometimes with a couple of samples

Table 1: Traditional models vs foundational models in fundamental principles.

In recent years, there has been significant progress in developing promptable foundation models, which are trained on extensive and diverse datasets. Once trained, these models serve as a versatile base that can be adapted, such as through fine-tuning, to a broad array of downstream tasks related to their initial training [1]. While the core technologies underpinning foundation models, including deep neural networks and self-supervised learning, have existed for some time, the remarkable advancements seen recently, particularly with large language models (LLMs), are largely due to the substantial increase in both data volume and model complexity [13]. For example, models with billions of parameters, like GPT-3 [2], have been effectively deployed for zero/few-shot learning, delivering extraordinary outcomes without the need for extensive task-specific datasets or modifications to the model's parameters.

These foundation models are much different from task-specific traditional models. Table 1 summarizes the fundamental differences between foundation models and traditional models.

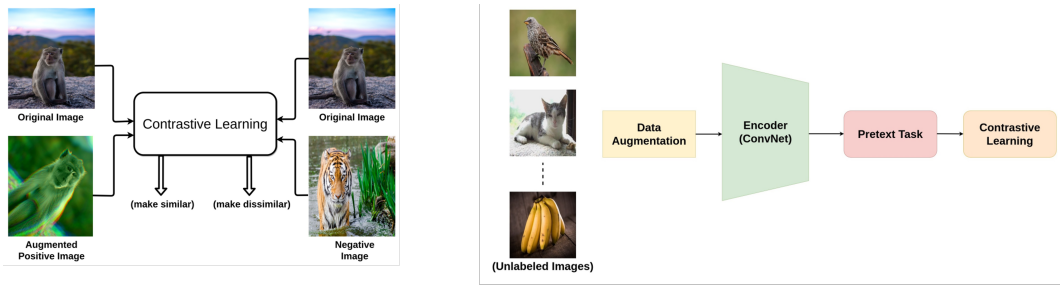
2.1 Pre-training Methods

Foundation models are typically trained on vast datasets to learn a wide range of skills that can be adapted to specific tasks. The pre-training of these models is crucial as it lays the groundwork for their versatility and performance.

2.1.1 Self-Supervised Learning

Self-supervised learning is a subset of unsupervised learning where the model learns to predict part of its input from other parts of its input [14].

Contrastive learning Contrastive Learning (CL) employs a discriminative strategy designed to pull similar samples closer while pushing dissimilar ones apart, as depicted in Fig. 2a. This method utilizes a similarity metric to determine the proximity between two embeddings. In the context of computer vision tasks, a contrastive loss is calculated based on the image features derived from



a) Main principles of CL

b) Pipeline for self-supervised training

Figure 2: Idea and training strategy behind contrastive learning [14].

an encoder network. For example, a training dataset sample is selected alongside a transformed version obtained through data augmentation. As illustrated in Fig. 2b, this augmented variant is treated as a positive sample, whereas other batch or dataset samples (depending on the approach) are viewed as negative samples. The model is then trained to distinguish between positive and negative samples, leveraging a specific pretext task for differentiation. This process enables the model to capture high-quality representations of the data, which are instrumental for knowledge transfer to various downstream applications. The technique of contrasting positive and negative pairs facilitates the learning of distinctive features, thereby improving the model's ability to differentiate between various inputs. Additionally, by autonomously generating labels through the contrasting of positive and negative pairs, the model acquires valuable representations without necessitating labor-intensive and costly manual labeling [15].

Masked image modeling Masked Image Modeling (MIM) [16] is a type of self-supervised learning technique that employs a generative model approach. Its primary objective is to reconstruct masked portions of an image, thereby learning a generalized feature representation across the data distribution. One advantage of MIM is its ability to avoid the inclusion of extraneous information, facilitating the utilization of large datasets. Moreover, MIM focuses on the reconstruction of pixel-level details from the original image. This approach, which does not rely on data augmentation or the comparison of negative pairs, enables the model to develop a robust feature representation. Furthermore, the MIM approach nurtures a deep understanding of the intrinsic patterns and structures within images, promoting a nuanced comprehension of visual contexts. By concentrating on the reconstruction of the image from partially observed data, MIM encourages the model to hone its predictive capabilities, thereby enhancing its ability to discern and interpret complex visual information. Consequently, the model not only learns to fill in the missing pieces but also to appreciate the underlying coherence

and variability within the visual data, laying a solid foundation for advanced visual reasoning tasks. This attribute enhances the foundation model's versatility across various image types as it is exposed to more data during the modeling process. For instance, SAM [11] incorporates this type of self-supervised learning as a component of training its image encoder.

2.1.2 Supervised and Semi-Supervised Learning

Supervised learning is a method that trains models on datasets where each example is associated with a specific label or annotation that denotes the correct outcome. This approach is prevalent across various tasks such as classification, and regression, among others. Within the realm of foundation models, supervised learning plays a crucial role in pre-training models on extensive datasets equipped with annotations. These datasets might comprise labeled images for tasks like object detection or annotated texts for sentiment analysis. The primary benefit of supervised learning lies in its ability to leverage labels provided by humans, facilitating the model's learning process and potentially resulting in more precise representations, especially in scenarios where labeled data are abundant. Despite the effectiveness of these methods, acquiring labeled data is both time-consuming and costly, which can be a huge burden while training the foundation models [17].

Semi-supervised learning falls between supervised and unsupervised learning, leveraging both labeled and unlabeled data during training [18]. This approach is particularly useful when a large amount of unlabeled data is available alongside a smaller set of labeled examples. Models can be pre-trained using unsupervised methods to learn general representations from the unlabeled data and then fine-tuned with supervised learning on the labeled subset to adapt to specific tasks. By utilizing both labeled and unlabeled data, this approach allows foundation models to leverage the abundant unlabeled data to learn generalizable representations, which are then refined with a smaller set of labeled examples for specific tasks [1].

2.1.3 Transfer Learning

While not a pre-training technique in the traditional sense, transfer learning is intimately connected to the concept of foundation models. It entails utilizing a model that has been pre-trained on a substantial dataset and tailoring it for a particular task using a much smaller dataset [19]. This pre-trained model serves as a foundational base, offering a comprehensive set of features that can be fine-tuned or adjusted for new applications. The array of pre-training methodologies plays a crucial

role in shaping foundation models by allowing them to derive knowledge from extensive and diverse datasets, typically through unsupervised or semi-supervised learning approaches. The selection of a pre-training strategy is influenced by factors such as the nature of the data at hand, the model's intended functionalities, and the specific tasks for which it is being customized.

2.1.4 Fine-tuning

Fine-tuning in the context of foundation models is typically pursued under three main scenarios: enhancing a model's performance on a specific task such as open-world object detection, augmenting a model's capability in areas like visual grounding, and adapting a model to address a variety of downstream vision tasks [20]. The premise is that foundation models, having been trained on extensive datasets, possess a keen ability to recognize general features, which can be leveraged to excel in numerous downstream tasks through fine-tuning. This approach is distinct from completely retraining the model, which can be impractical given the colossal size of foundation models, sometimes encompassing billions of parameters [6], making full retraining not only challenging but also inefficient.

Fine-tuning, therefore, becomes a strategic choice, focusing on adjusting a minor portion of the model, such as a part of the decoder, to enhance its performance for specific downstream applications [21]. Additionally, the introduction of Low-rank Adaptation (LoRA) modules [20] represents a method where a minimal parameter module is integrated into the decoder section of a foundation model, with only this module undergoing training to suit a particular downstream task. These methodologies collectively fall under the umbrella of Parameter Efficient Fine-Tuning (PEFT) [22], a concept widely embraced by the foundation model research community for its effectiveness and efficiency in model optimization.

2.1.5 Prompt Generation

For promptable foundation models, the efficacy of the output is significantly influenced by the quality of the input prompt. However, sometimes generating prompts is a complex task or too time-consuming. For instance, within the context of medical imagery segmentation, crafting an effective prompt presents a more complex challenge compared to scenarios involving natural imagery [23]. Often, a level of domain expertise comparable to that of a professional is required to formulate an effective prompt. Furthermore, the presence of low-quality prompts, which can arise from noisy

annotations, has the potential to substantially degrade the accuracy of the segmentation results. Moreover, for geographical image segmentation, crafting effective prompts can be challenging and labor-intensive. Unlike natural images where objects may be localized to specific areas, geographical images often feature elements like roads, rivers, and vehicles dispersed throughout the entire frame. This distribution complicates manual prompting, making it not only difficult but also prone to inaccuracies in many instances.

As a response to these challenges, the development of an auto-prompting mechanism is pursued to create a resilient and adaptable system. This system aims to minimize the impact of prompt quality variability on foundation models' performance, thereby enhancing the reliability and precision of image segmentation outcomes. Numerous approaches have been developed for generating prompts. Two prevalent methods include substituting the prompt encoder in foundation models with a different pre-trained model or employing a secondary model to establish a pipeline dedicated to prompt creation. Therefore, the necessity for automated prompt generation or prompt learning becomes a critical consideration.

2.2 Large Language Models

While NLP-based foundation models are not the central focus of this review, they serve as a crucial foundational element for understanding the broader concept of foundation models. There has been considerable progress in the development of Large Language Models (PLMs), which primarily utilize Transformer architectures as their backbone and are pre-trained on extensive unlabeled text corpora. These models have shown remarkable proficiency across a range of NLP tasks.

Early pivotal PLMs, such as BERT [24], XLNet [25], and the initial models in the GPT series [26], [27, 2] (GPT-1, GPT-2 and GPT-3), have adopted a "pre-training and fine-tuning" approach for tackling downstream tasks. More recent endeavors have revealed that significantly increasing the size of PLMs (to tens of billions of parameters, for instance) not only boosts their performance capacity but also unlocks new, emergent abilities like in-context learning and reasoning capabilities, which are absent in smaller-scale PLMs (such as BERT). These advanced PLMs are often referred to as Large Language Models (LLMs) in scholarly discussions [13]. A noteworthy milestone in the evolution of LLMs is the introduction of ChatGPT based on the GPT series, which has demonstrated exceptional conversational abilities with humans, capturing significant public and academic interest.

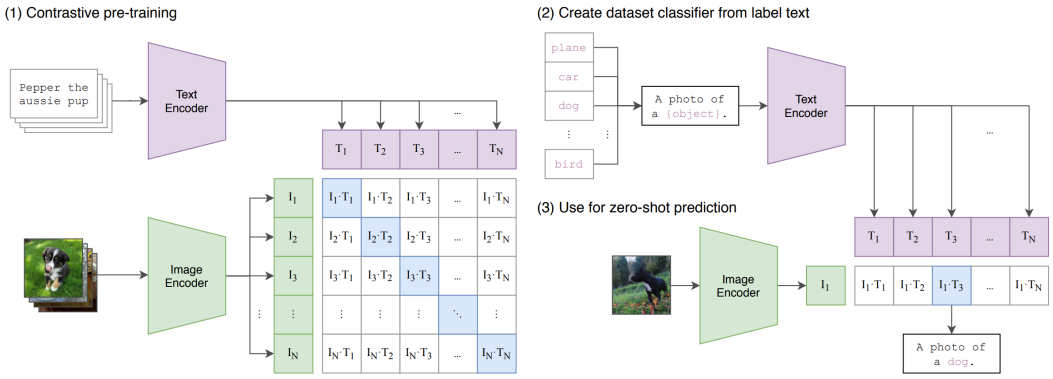


Figure 3: CLIP, trains both an image encoder and a text encoder to accurately identify the matching pairs within a set of (image, text) training examples [28].

2.3 Vision Foundation Models (VFM)

In recent years, the field has seen notable progress in the development of Vision Foundation Models (VFMs) [28, 11, 29]. Initial research efforts [30, 31] for establishing VFMs were concentrated on pre-training Convolutional Neural Networks (CNNs), such as ResNet [32], on the ImageNet dataset [33] for various downstream tasks. However, CNN-based models often encounter difficulties in scaling to extremely large datasets due to their inherent capacity constraints [29]. In response, the Vision Transformer (ViT) model [8] was introduced, utilizing Transformer mechanisms to process visual data by treating images as sequences of patches. The ViT framework, in comparison to its CNN-based equivalents, boasts a significantly greater capacity, enabling it to more efficiently leverage the vast amounts of data available.

Following the introduction of ViT, a wealth of studies have leveraged ViT to create sophisticated VFMs by training on a wealth of data. These VFMs are designed to be promptable, allowing users to interact directly with the model by guiding it towards specific downstream tasks. This approach marks a significant departure from the traditional use of pre-trained models, offering a more flexible and user-directed application methodology. Language-augmented VFMs, such as CLIP [28] are one of the pioneering works in this domain. Focusing on training an image encoder through the alignment of text-image pairs using contrastive learning techniques (illustrated in Fig. 3). On the other hand, vision-only VFMs are trained exclusively on visual data via self-supervised pre-training methods, including techniques like the masked autoencoder (MAE) [16] and Latent Vision Models (LVM) [29]. These promptable VFMs, pre-trained on a broad array of general visual data, have demonstrated strong generalization capabilities and remarkable zero-shot learning performance across various

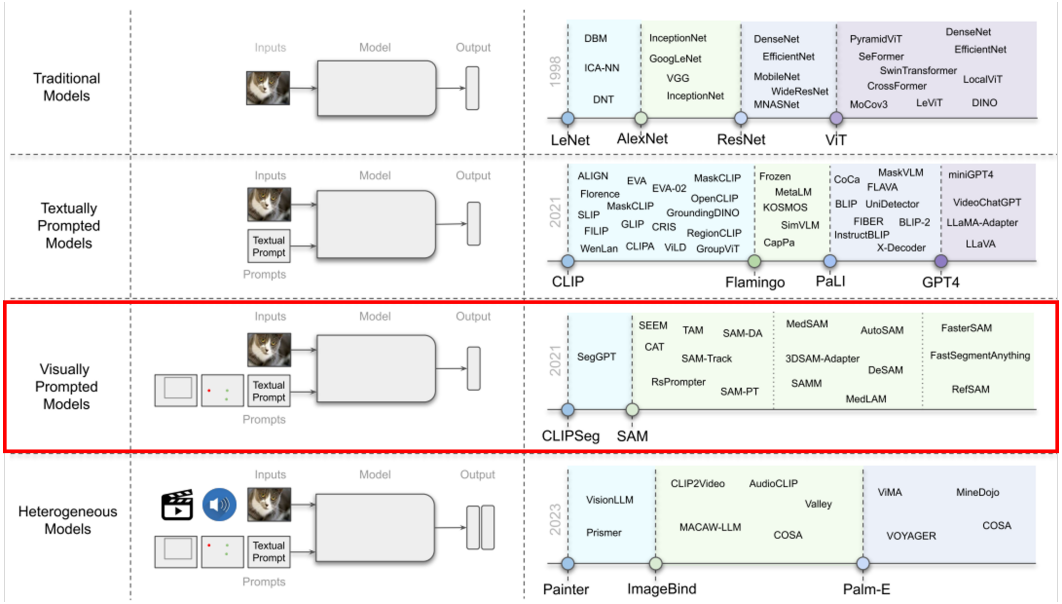


Figure 4: A complete overview of the vision foundation models, the focus being on the visually prompted models [15].

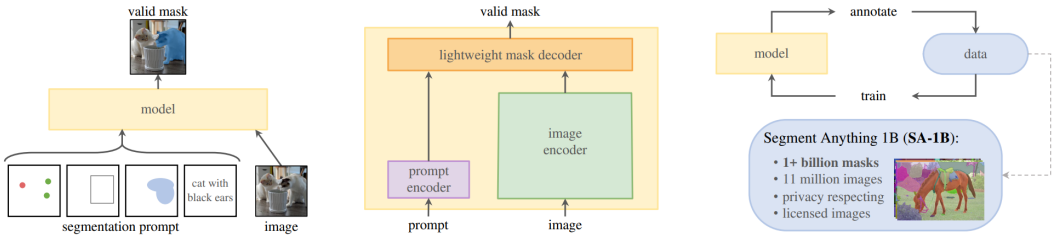
computer vision tasks. Fig. 4 offers a detailed overview of vision foundation models, with the segment concerning visually prompted models emphasized in red.

2.3.1 Segment Anything Model (SAM)

Overview

The Segment Anything Model (SAM) [11] represents a great advancement in the realm of foundation models, specifically designed to enhance segmentation capabilities for objects that have not been previously encountered. This is achieved through the use of prompts provided by users, which can vary widely in their form—ranging from single or multiple points to bounding boxes, or textual descriptions. These prompts enable SAM to generate precise instance segmentation masks for new images without the need for additional training, a capability referred to as zero-shot segmentation. However, it's important to note that the semantic significance of the segmented objects relies heavily on the accuracy and relevance of the prompts used.

SAM is a binary class segmentation model, trained on the expansive SA-1B dataset, comprising 11M high-resolution images and 1.1B high-quality segmentation masks, which is 400 times more than any prior segmentation dataset [11]. This extensive pre-training significantly improves the model's ability to adapt to and segment unseen domain samples. Despite its proven effectiveness in general domains, navigating the complexities of medical imaging with SAM presents unique challenges. Structurally,



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 5: Segment Anything Model (SAM) [11].

the architecture of SAM can be categorized into three main components: image encoder, prompt encoder, and mask decoder, illustrated in Fig. 5.

Image Encoder This module integrates a Masked AutoEncoder (MAE) [16] and a pre-trained Vision Transformer (ViT) [8] to accommodate high-resolution images of various scales. The image encoder's flexibility allows for adjustments from ViT-B to ViT-H or alternative architectures to evaluate the balance between model performance and efficiency. For consistency with the pre-training phase, all images are resized to 1024×1024.

Prompt Encoder It distinguishes between two types of prompts: sparse and dense. Sparse prompts, including points and boxes, leverage ViT's positional encoding merged with learned representations. Text prompts are also considered sparse and are interpreted via a CLIP-pre-trained text encoder. Conversely, dense prompts, such as masks, are processed through convolutional operations to align with the image features identified by the image encoder.

Mask Decoder This component employs the information extracted from the encoders to enable refined predictions. It employs a transformer decoder block and a dynamic prediction head, integrating bidirectional self-attention with cross-attention mechanisms for the simultaneous refinement of prompt-to-image and image-to-prompt embeddings. To ensure efficiency, only two decoder blocks are used for upsampling the embeddings' resolution. An MLP acts as the final layer, transforming output tokens into foreground probabilities for segmentation.

Background

SAM consists of three components in general, outlined in 5: an image encoder (named as $\mathrm{Enc_I}$), a prompt encoder (named as $\mathrm{Enc_P}$), and a mask decoder (named as $\mathrm{Dec_M}$). SAM, as a promptable foundation model, takes an input image, referred to as I, and a set of prompts from the user, as P.

At the beginning, SAM utilizes $\mathrm{Enc_I}$ to extract features from the given image. Later, SAM employs $\mathrm{Enc_P}$ to transform the human-given prompts, having a length of k, into prompt tokens. Specifically:

$$F_I = \mathrm{Enc_I}(I), \quad T_P = \mathrm{Enc_P}(P), \tag{1}$$

From Equation 1, F_I represents the feature embedding of the image where $F_I \in \mathbb{R}^{h \times w \times c}$, h and w express the resolution of the image feature map, and c indicates the feature dimension. Likewise, T_P is the feature embedding of the prompts where $T_P \in \mathbb{R}^{k \times c}$, k is the length of the prompts.

After this, the encoded image and prompts are supplied to the decoder, called $\mathrm{Dec_M}$, which employs attention-based mechanisms for feature interaction. SAM creates the input tokens for the decoder by combining several mask tokens, represented as T_M, with the prompt tokens T_P. These mask tokens play a crucial role in generating the mask output, which is defined as:

$$S = \mathrm{Dec_M}\left(F_I, \mathrm{Concat}(T_M, T_P)\right), \tag{2}$$

where S in Equation 2 represents the output segmentation mask predicted by SAM.

2.4 Promptable Foundation Model

Figure 6: Appropriate masks produced in response to single ambiguous points [11].

Foundation models represent a significant departure from traditional machine learning models, primarily through their use of prompts, allowing for direct interaction between the models and users, a feature absent in traditional models [1]. This interactive capability transforms the way machine learning models are utilized, as seen with LLMs like ChatGPT [2], where users guide the model's tasks through instructions. Similarly, in the realm of VFMs, CLIP [28] undertakes classification tasks based on textual prompts specifying desired classes for image categorization. SAM extends this promptability to image segmentation, standing out as one of the initial models in this category to accept user prompts as guidance for segmentation tasks.

A distinctive feature of the promptable task is its ability to generate a valid segmentation mask from any given segmentation prompt. This prompt acts as the context for the model to work on. A prompt

could be anything that specifies the target for segmentation. A "valid" mask implies that even if the prompt introduces some ambiguity, such as a point on a T-shirt, the resulting mask should still reasonably represent at least one of the plausible objects within the prompt's context—either the human or the T-shirt in this example. A point can produce multiple valid masks, illustrated in Fig. 6.

Prompts in SAM

Sparse Prompts In the sparse prompts category SAM contains three types of prompts, click prompts, bounding box prompts, and text prompts [11]. Click prompts allow users to interact directly with the input image by selecting various points. These selections provide context for the model, distinguishing between foreground points, which should be included in the generated mask, and background points, which fall outside the target mask. Such points are low dimensional, comprising only (x,y) coordinates alongside one of two learned embeddings to denote whether the point is foreground or background.

Bounding box prompts, conversely, offer a user-generated approximation of the object's location. Each box is described through a pair of embeddings: the first combines the positional encoding of the box's top-left corner with a learned embedding for the "top-left corner," while the second follows a similar pattern for the "bottom-right corner." This method provides a straightforward yet effective means for users to guide the model towards the object of interest within the image.

Text prompts serve as user inputs specifying the objects they wish to segment, offering the most intuitive mode of interaction with the model. This method leverages natural language, allowing users to directly communicate with the segmentation model about the specific object they aim to isolate. Unlike click or bounding box prompts, which require users to manually approximate the object's location, text prompts simplify the process by enabling users to "ask" SAM to perform the segmentation. SAM utilizes the text encoder from CLIP [28], capitalizing on the alignment between text and image embeddings established by CLIP, to understand and execute these natural language segmentation requests effectively.

Dense Prompts In SAM, dense prompts function as mask inputs, albeit at a significantly reduced resolution—specifically, four times smaller than the original image size. These dense prompts primarily serve as supplementary inputs that enhance the initial sparse prompts by providing additional context to the model. Acting as a secondary input mechanism, dense prompts on their own have limited effectiveness, a conclusion supported by ablation studies detailed in [34]. This highlights their role in refining and improving the model's performance when combined with other forms of prompts.

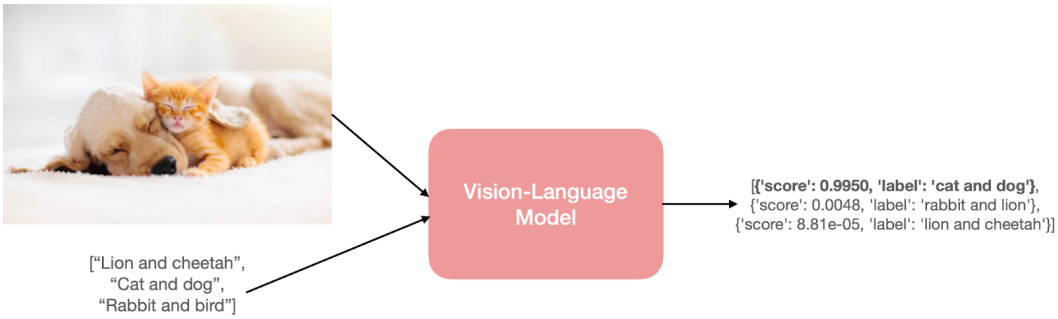


Figure 7: The concept of combining two modalities in VLM, image taken from here.

2.5 Vision Language Foundation Models (VLM)

Vision-language foundation models (VLMs) are a type of foundation model specifically pre-trained to accept and process both visual and language data as inputs. By integrating these two modalities, VLMs are designed to perform a wide array of tasks, leveraging the combined strengths of vision and language understanding [28]. Fig. 7 illustrates a basic concept of how they perform by combining two modalities.

The field of foundation models is currently experiencing a significant upswing in the development of VLMs for a variety of computer vision tasks. Leading this wave is CLIP, which, as previously mentioned, merges image and text modalities to facilitate tasks like image classification and captioning. It has been observed that incorporating information from an additional domain, such as text, tends to significantly enhance the performance of supervised machine learning models. This improvement is particularly notable when the model successfully aligns the contexts of both vision and language [35]. Consequently, more complex tasks, including image segmentation, are beginning to leverage VLMs to boost their effectiveness. Employing both vision and natural language to approach a task mirrors real-world scenarios more closely, making the process more practical and intuitive.

3 Foundation Model: Geographical Image Segmentation

Geographical image segmentation represents one of the more popular applications of semantic segmentation. This process involves a model extracting relevant features from geographical imagery to identify and delineate target objects. This section is divided into two parts, discussing popular works using traditional methods and foundation model-based approaches to geographical image segmentation.

3.1 Traditional Methods

CNN-based Geographical Image Segmentation The use of Fully Convolutional Neural Networks (FCNNs), particularly those based on the UNet architecture [7], has shown promising results [36]. Initially, the application of CNNs, specifically UNet, in geographical image segmentation is examined. Before the advent of foundation models, UNet and encoder-decoder models were the standard for various segmentation tasks. Simple implementations of UNet have demonstrated good results in segmenting geographical images [37]. Furthermore, advanced popular UNet-based models such as D-LinkNet [38], SegNet [39], and others tailored for satellite imagery [40], have been developed to tackle a range of geographical and geospatial segmentation challenges. Additionally, techniques like atrous convolution, neural architecture search, and contrastive learning have been applied [41, 42, 43] to improve CNN-based model performance in geographical segmentation, though fundamental challenges remain. Tile2Net [44] stands out for its work on pedestrian infrastructure segmentation on aerial images. Their methodology, leveraging the Hierarchical Multi-Scale Attention model [45] with HRNet-W48 [46] as the backbone, aimed at segmenting transportation infrastructure components.

In addition to conventional CNN models trained from scratch, there have been mentionable transfer learning-based efforts, exemplified in works like [47, 48], where a model trained from the source task is used to reduce the computational demands for various related downstream tasks. However, it's worth noting that in practice, these researchers often encounter the need to conduct further fine-tuning or retraining of these models to align them with the precise objectives of their respective tasks. This lack of generalizability leads to these source models being re-trained to achieve competitive performance in the downstream task. These transfer learning-based approaches remain task-specific, in contrast to foundation models, which are designed to be more general and not tied to specific tasks. This limitation may be due to both algorithmic and data challenges: (1) supervised learning algorithms often learn shortcut features to achieve high performance in the few urban cities where the training sets are available whereas it may not generalize well to rural areas with diverse geographical environments; (2) transfer learning algorithms can be leveraged to mitigate this issue [10], but it still needs a fair amount of labeled geographical imaginary in each rural area to fine-tune the semantic segmentation models. This can be unscalable and limited by the quality and availability of labeled imaginary data sets. The rise of VFMs [11] represents a big leap in scaling up segmentation models, allowing for powerful zero-shot or few-shot capabilities and flexible prompting. Without any further

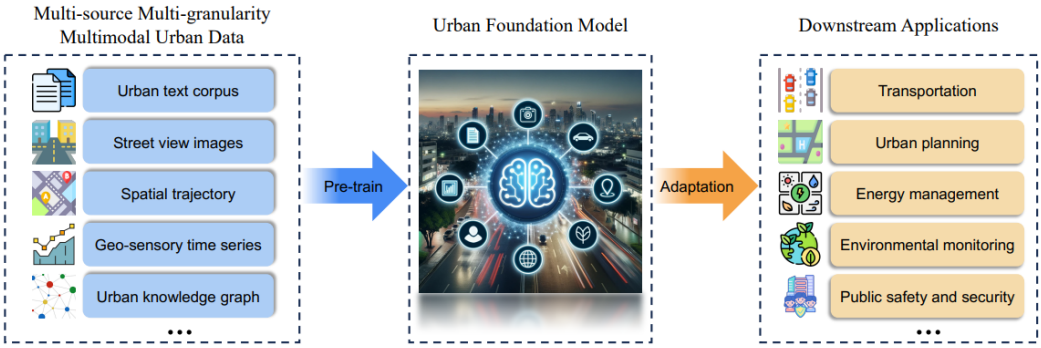


Figure 8: Urban foundation models for undergo pre-training on diverse, multi-layered, and multimodal urban data, making them versatile for numerous downstream urban applications [49].

training, these models can quickly adapt to a new downstream task without the need for re-training all parameters of the model.

3.2 Foundation Model-based Approaches

Geographical and remote sensing image segmentation encompasses a vast field with numerous data types ranging from rural areas to even outer space. Our focus is primarily on urban geographical data, specifically on the application of foundation models to urban datasets in the context of transportation segmentation, network creation, and related areas. These models are leveraged to analyze and interpret complex urban environments, aiding in the development of smart transportation systems, efficient urban planning, and infrastructure management. These urban foundation models are expansive and pre-trained on extensive datasets that include multi-source, multi-granularity, and multimodal urban data. The pre-training phase endows these models with emergent capabilities, enabling them to adapt to a wide array of downstream tasks and domains within urban settings with significant efficacy, illustrated in Fig. 8. Remote Sensing data, particularly in the form of geographical imagery, is often acquired from satellites, aircraft, or drones. This type of imagery is capable of encompassing vast urban expanses within a single image or dataset, making it exceptionally valuable for broad-scale analyses such as land use classification, urban planning, and environmental monitoring.

3.2.1 Pre-training

The extensive use of camera and satellite technologies has led to the accumulation of a substantial amount of visual data in urban environments from aerial perspectives. This data encompasses a wide range of urban elements, including cityscapes, infrastructure, public spaces, and human activities. The abundance of such data has spurred a series of research initiatives aimed at developing large urban

vision models from the ground up. These models are categorized based on the type of pre-training data they employ, which includes on-site urban visual data, remote sensing data, and grid-based meteorological data. This classification reflects the diverse sources of data that can be leveraged to enhance the understanding and analysis of urban environments through advanced vision models.

On-site Urban Visual Data On-site urban visual data, such as street-view imagery [50] and surveillance video [51], originate from ground-level devices like smartphones, cameras, and automotive LiDARs within urban settings, capturing detailed aspects of street scenes, traffic flow, and pedestrian movements. This data type is instrumental for various applications, including real estate valuation [52], identification of congestion hotspots [53], and analysis of urban socioeconomic impacts [54]. Initial research in this domain primarily depended on task-specific supervisory signals, like demographic indicators, to derive visual representations from urban imagery. However, these methods come with the downside of requiring extensive labeling efforts and generally suffer from poor generalizability. To overcome these challenges, more recent studies [54], [5], [55] have pivoted towards self-supervised learning approaches. These methods derive image representations from unlabeled visual content, thereby facilitating diverse predictive applications without the need for explicit labeling. For example, Urban2Vec [5] leverages Tobler's First Law of Geography [56], which suggests that "everything is related to everything else, but near things are more related than distant things," to create a contrastive learning framework. This framework aims to ensure that street-view images that are spatially close to each other share similarities in their latent feature space. Another initiative, KnowCL [54], employs contrastive loss to maximize the mutual information between the representation of a region's street-view image and its associated knowledge graph embedding. Despite these advancements, the majority of research has concentrated on relatively small datasets, leaving the pre-training of foundation models on extensive on-site urban visual data largely unexplored.

Remote Sensing Data Remote Sensing (RS) data, captured from satellites, aircraft, or drones, can encompass wide urban areas in a single image or dataset, making it invaluable for large-scale analyses such as land use classification, urban planning, and environmental monitoring. The success of self-supervised learning (SSL) in computer vision has inspired a wave of research [57, 58] into applying SSL methods to mine insights from unlabeled RS data, with comprehensive reviews of SSL algorithms for RS data available in literature [59]. This burgeoning interest has significantly propelled the development of RS foundation models, garnering attention from both the academic [8, 60] and industrial sectors [3]. Predominantly, RS foundation models adhere to the masked

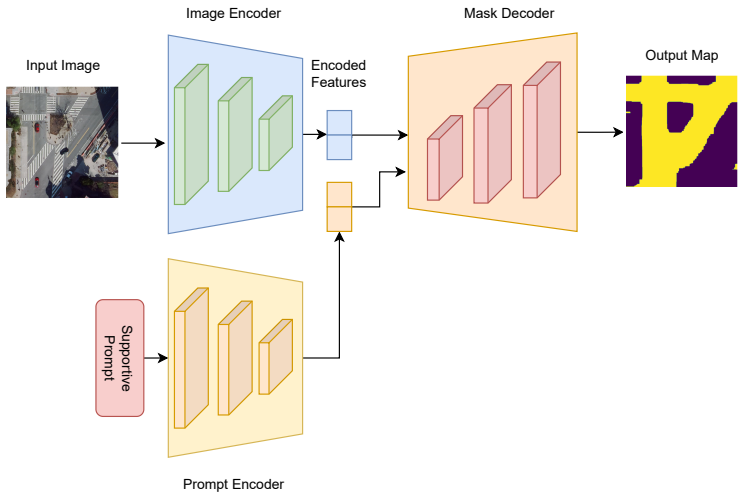


Figure 9: A general approach to using SAM for geographical image segmentation, and road segmentation has been outlined here.

image modeling (MIM) approach [16], which involves encoding a partially masked RS image and subsequently decoding the visible segments to reconstruct the full image. Among the innovations, [60] trained a basic Vision Transformer (ViT) with 100 million parameters to tackle a variety of RS challenges, from urban object detection to broader urban landscape analysis. A novel approach, employing rotated varied-size window attention in place of the standard ViT attention mechanism, was introduced to decrease computational demands for high-resolution RS imagery. ScaleMAE [61] incorporates scale invariance into the conventional ViT framework, using a Laplacian-pyramid network to grasp multi-scale semantic layers in RS imagery. In a more ambitious endeavor, [6] investigates the effects of scaling a ViT model up to one billion parameters within the RS field. Additionally, RingMo [17] utilizes the Swin Transformer [62], a notable ViT variant, as its backbone to adeptly identify dense and small objects frequently missed in RS analyses. RingMo-Sense [63] expands on RingMo's capabilities for spatiotemporal prediction tasks in RS, introducing a multi-branch architecture to learn representations across multiple scales and temporalities, and applying diverse masking techniques for MIM-based pre-training. Contrasting with these models, CSP [64] aims to learn visual representations by using a contrastive pre-training objective that aligns paired locations with their RS images. Moreover, [65] compares different visions of foundation models with CNN-based fine-tuned models for RS geographical images and they conclude that the foundation models fall short compared to the CNN-based fine-tuned models.

Zero or Few-Shot Works SAM [11] represents a breakthrough in the domain, demonstrated through its application in zero-shot and few-shot learning scenarios, addressing geographical image segmen-

tation with remarkable flexibility. Zero-shot learning involves utilizing the unmodified, or vanilla, version of SAM, where segmentation is achieved solely through the use of prompts without any adjustment to the model's parameters. Conversely, few-shot learning entails providing SAM with additional context for a specific application area by using one or more examples, thereby enhancing its understanding and performance on tasks with minimal example input. Fig. 9 showcases a standard methodology for applying SAM to geographical imagery, illustrating all the critical components of SAM actively engaged in segmentation tasks. Specifically, this figure highlights the process of road segmentation. Compared to advancements in SAM applications across various domains, the use of SAM within the realm of geographical imagery remains significantly less developed.

Considering the zero-shot-based approach in geographical images, a notable study, [66] utilized SAM's zero-shot capabilities to label the Remote Sensing Segmentation Dataset as well as existing object detection datasets. [67] developed a mixture of zero-shot and one-shot learning for SAM for segmenting geographical imagery. They utilized text prompts to guide the selection of the one-shot for SAM and implemented a continuous process to segment each object of interest within the scene. In [68], SAM's zero-shot techniques were employed for edge detection. This work showcases SAM's potential in zero-shot learning scenarios, where limited labeled data is available. In addition to these studies, research such as [69] employs SAM and a new loss function to focus more on the boundary information for segmentation tasks on geographical imagery.

3.2.2 Model Tuning

In the context of geographical image segmentation, the scarcity of domain-specific data can hinder the development of large-scale foundation models tailored to this field [49]. To overcome this obstacle, numerous studies have explored the strategy of repurposing existing foundation models, which were initially trained on a broad spectrum of natural imaging data, for geographical segmentation tasks. These approaches allow for the leveraging of the intrinsic capabilities of off-the-shelf foundation models, enabling them to address the unique challenges presented by geographical imagery segmentation.

Model Fine-Tuning As previously highlighted, SAM was primarily trained on natural images, which poses challenges when segmenting geographical images due to the lack of domain-specific knowledge. Consequently, further fine-tuning of SAM is essential to equip it with the necessary expertise for effective segmentation in geographical contexts. For instance, RingMo-SAM [17]

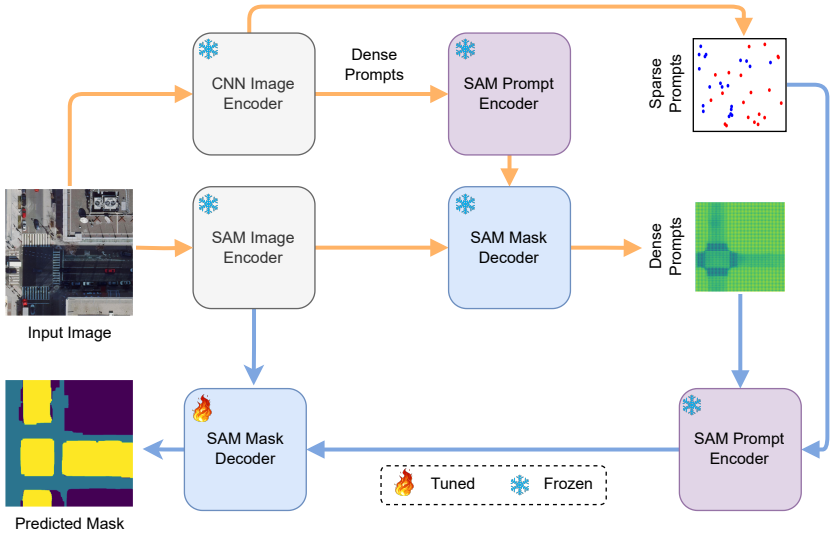


Figure 10: Incorporating both prompt generation and fine-tuning technique from [34].

employs a fine-tuning approach to modify the prompt encoder within SAM for multi-source RS segmentation. Additionally, recent investigations [70], [71] have been conducted into the application of language-augmented VFMs, like CLIP, in the analysis of street scenes. RS-CLIP [72] employs a curriculum learning strategy to enhance the performance of zero-shot classification for remote sensing images with SAM, implementing a multi-stage model fine-tuning process. Specifically, [70] introduces a technique for analyzing street-view imagery utilizing a language-enhanced VFM that has been fine-tuned with a custom loss function and classification layers. This method facilitates the generation of comprehensive textual descriptions from diverse urban settings, associating street view images with geo-tagged textual data. Furthermore, StreetCLIP [71] customizes CLIP for geolocalization tasks related to street-view images by fine-tuning the model with artificial captions via a meta-learning strategy. As the exploration of foundation models within the computer vision sector remains a burgeoning field of study, the adaptation of general-purpose VFMs is currently in its formative stages.

Additionally, [73] explored the use of SAM's semantic segmentation capabilities for planetary geographical mapping by introducing fine-tuning of SAM using knowledge distillation in this specific domain. Notably, they introduced a lightweight student decoder that can be fine-tuned with a limited number of images. SAM-Adapter [74], a mentionable approach of using parameter-efficient fine-tuning (PEFT), where it focuses on adding an adapter to the foundation model and training only the adapter part to improve in various downstream. Extending this approach, in a more recent contribution presented in [75], the authors explored the area of temporal detection in remote sensing images.

SAM-PARSE [76] introduces nearly zero trainable parameters while doing the fine-tuning by directly modifying the parameter space. Finally, GeoSAM [34] delves into the realms of prompt generation and fine-tuning specifically for the segmentation of geographical images, with a particular emphasis on mobility infrastructure segmentation, notably that of pedestrian infrastructure. This research stands out as a mentionable effort, to integrate the entirety of existing technologies associated with the SAM-based approach within its methodology. Fig. 10 visually presents the approach adopted by this work, detailing the strategies employed.

3.2.3 Prompt Generation

The study by [66] introduces SAMRS, a system for Remote Sensing (RS) segmentation, which exploits the zero-shot learning capabilities of SAM. SAMRS constructs six fundamental prompts tailored to the specific traits of RS imagery, selecting the most effective prompt combination through experimental validation for segmentation tasks. RSPrompter [77] advances this methodology by automating the prompt generation process. It creates suitable prompts for SAM, such as point or box embeddings, by examining the encoder's hidden layers. [78] introduces Text2Seg, using Grounding Dino it can automatically generate box prompts for SAM. CS-WSCDNet [79] is another example of prompt generation where it focuses on Change Detection in image pairs of remote sensing images. Additionally, the work by [75] also explores the use of Change Detection, entirely bypassing the need for prompts in SAM and relying solely on the model's remaining components. [80] leverages self-guided few-shot examples generated by SAM which essentially works as automated prompts for remote sensing image segmentation. [81] introduced PerSAM and PerSAM-F where they use self-generated one-shot for auto prompt generation. A very recent work [82] introduces prompts-learning directly in the latent space or the high dimensional space. They argued that they can learn the prompts in the high dimensional space, thus removing the need for human-created prompts. Their implementation encompasses a range of scenarios, including the analysis of on-site urban visual data, notably street scene images.

4 Foundation Model: Medical Image Segmentation

The objective of medical image segmentation is to delineate anatomical or pathological structures within tissues, aiding in computer-assisted diagnostics and advanced surgical procedures [83, 84]. With the advancement of computational capabilities and the expansion of medical data resources,

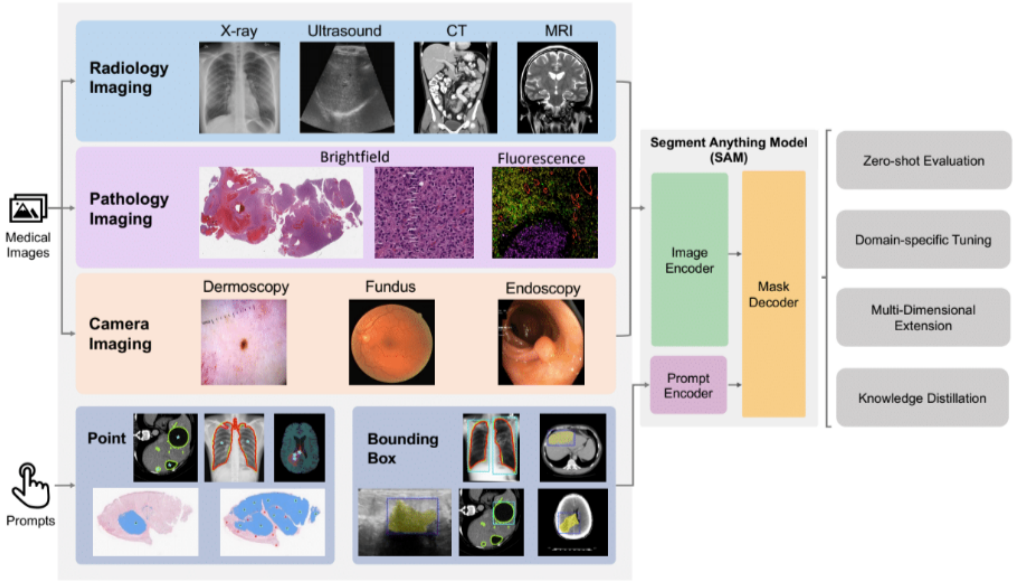


Figure 11: Application of SAM Across Medical Imaging Modalities, and adapting the decoder if needed. [4].

deep learning-based approaches to medical image segmentation have seen substantial improvements in both precision and efficiency over traditional methods [85, 86]. The introduction of the ViT has further propelled medical imaging techniques, with ViT-based methods [87, 88, 89, 90, 91] achieving exemplary results in segmentation tasks. However, these networks are often specialized for specific tasks, and thus, may lack broad applicability. The recent development of SAM opens up possibilities for handling various segmentation challenges under a single, unified framework. SAM's most noteworthy feature is its ability to accurately segment objects from provided points without prior knowledge of the object types or modality, illustrated in Fig. 11. This function is crucial in

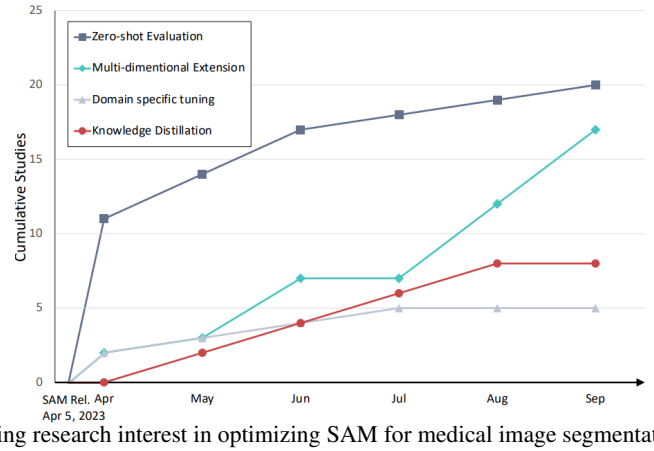


Figure 12: A growing research interest in optimizing SAM for medical image segmentation for the time being of April 5, 2023, to September 2023. Showcasing the extensive research being done in this area using techniques such as outlined [4].

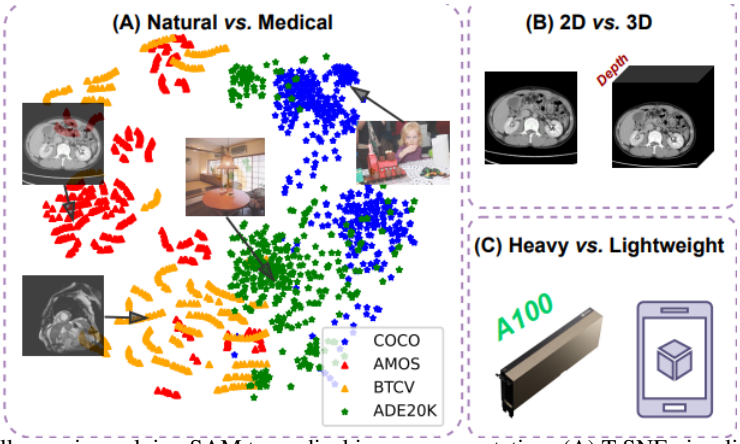


Figure 13: Challenges in applying SAM to medical image segmentation: (A) T-SNE visualization of SAM's encodings, contrasting medical datasets (AMOS, BTCV) with natural datasets (ADE20K, COCO); (B) 2D versus 3D imaging compatibility; (C) Computing demands: heavy versus lightweight [92].

medical imaging, where the precise delineation of various anatomical structures and pathological conditions is essential for accurate diagnosis and treatment planning.

By mirroring the human visual system's versatility in recognizing and segmenting objects, SAM represents a breakthrough in the field of computational image analysis, offering a new horizon of possibilities for researchers and clinicians alike in the nuanced domain of medical imaging. In light of this, researchers have been focusing on adapting SAM for medical image segmentation, identifying and implementing effective strategies to bolster its performance in this domain. This section delves into the widespread integration and innovative applications of foundation models especially SAM in medical imaging, highlighting its significant impact on academic research over the recent year. Fig. 12 showcases the evolving trend of employing the SAM in medical image segmentation, highlighting the recent techniques that have been adopted.

Despite its capabilities, SAM, as noted, has been primarily trained on natural images [11], which poses challenges for zero-shot segmentation on medical images. Furthermore, SAM is originally configured for binary class segmentation of 2D images, which contrasts with the nature of medical imaging examples like CT and MRI scans that often involve 3D images and require multi-class segmentation. The disparities between natural and medical imagery serve as significant obstacles to the adaptation of SAM for medical image segmentation. Fig. 13 visually delineates these differences across three major aspects, illustrating the challenges in integrating SAM within the medical imaging domain. To address these issues, researchers have explored several domain-specific tuning strategies, a summary of these techniques is illustrated in Fig. 14. In the following section, this review focuses on the traditional works that have been done over the years before the emergence of foundation

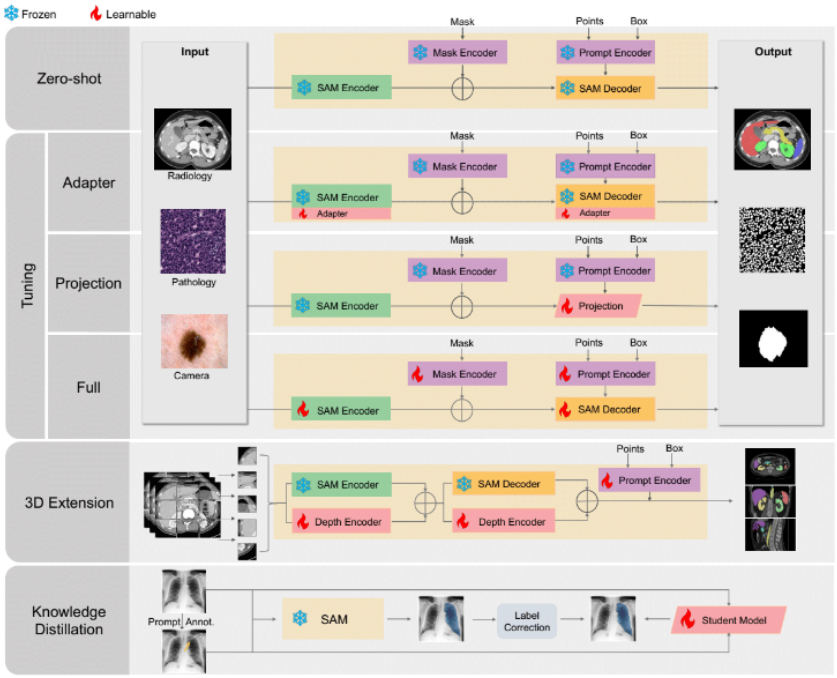


Figure 14: A comprehensive summary of how SAM's adaptation techniques are applied within the medical field. [4].

models, and then gradually the review moves on to the foundation model i.e. SAM-based works in the medical imagery domain.

4.1 Traditional Methods

Before the emergence of SAM, the majority of research in medical imaging was predominantly established on UNet (based on CNN) and UNETR (based on ViT), with some studies leveraging a hybrid approach that combines both. Here, the review mentions several promising contributions of these two algorithms to the field of medical imaging.

UNet-Based One of the most revolutionary segmentation frameworks, UNet, significantly transforming the field of medical image segmentation [93, 94, 95, 96]. CNNs hierarchically process images, capturing both local features and larger structural patterns. However, they occasionally struggle with comprehending the global context, a limitation attributed to their fixed receptive fields.

UNETR-Based After CNN, ViTs have gained prominence for their ability to grasp global image patterns, a marked advancement over the capabilities of CNNs. ViTs are not restricted by fixed receptive fields, thus they excel at capturing both nuanced local details and expansive contextual information. Consequently, the introduction of UNETR-based works has emerged, drawing inspiration from ViTs.

The development of hierarchical transformers, which combine the advantages of both CNNs and transformers, further enhances segmentation proficiency [87, 88, 89, 90, 91].

4.2 Foundation Model-based Approaches

The study from [97] offers an in-depth evaluation of SAM's capability in medical image segmentation, utilizing an extensive collection of 11 datasets that span a wide array of modalities and anatomical structures. Building on this, [98] delves into an analysis of SAM's accuracy in segmentation across 12 public medical imaging datasets. These datasets feature a variety of organs—including the brain, breast, chest, lung, skin, liver, bowel, pancreas, and prostate—alongside a range of imaging techniques such as 2D X-ray, histology, endoscopy, 3D MRI, and CT, covering both normal and lesioned health conditions. [99]'s research further scrutinizes SAM's utility in medical imaging by offering both quantitative and qualitative insights into its zero-shot segmentation capabilities over nine medical image segmentation benchmarks, encompassing diverse imaging modalities like OCT, MRI, and CT, and extending across fields like dermatology, ophthalmology, and radiology.

Further, [100] explores SAM's aptitude for zero-shot generalization across medical images, employing over 12 public datasets that include a variety of organs and modalities. Moreover, [101] assesses the zero-shot performance of SAM 2D in the medical imaging domain across six datasets from four different imaging modalities—X-ray, ultrasound, dermatoscopy, and colonoscopy. This exploration reveals that SAM 2D either matches or exceeds the performance of current state-of-the-art models. Notably, [101] also combines 52 open-source datasets to create a vast medical segmentation dataset, featuring 16 modalities, 68 objects, and 553K slices, and conducts a thorough analysis of various SAM testing strategies on this expansive COSMOS 553K dataset. This wave of research leveraging zero-shot SAM-based approaches heralds a new frontier in the application of advanced techniques with SAM, aiming for superior results in medical imaging. In these mentioned medical image segmentation tasks, the performance of SAM doesn't quite meet the necessary standards for further use, especially in scenarios where very high accuracy is essential. As mentioned earlier, SAM's training data mainly consists of natural images, which often have clear edge details that are quite different from what is typically seen in medical images. As a result, using SAM as it is, without any adjustments or retraining for new and complex medical image segmentation tasks, tends to result in suboptimal outcomes. Therefore, researchers have shifted their focus towards modifying and fine-tuning SAM to improve its effectiveness in these applications.

4.2.1 Model Tuning

Various tuning techniques (illustrated in Fig. 14 have been adopted for SAM in medical imagery.

Projection Tuning This approach involves substituting the existing SAM mask decoder for a novel, task-specific decoder while retaining the remaining elements of SAM unchanged. The intention is to leverage the pre-existing capabilities of the SAM encoder, capitalizing on its extensive understanding of intricate image features such as edges and boundaries, derived from the realm of natural images. As illustrated in Fig. 14, there is a trend in recent studies to replace the projection head with different types of architectures and initiate training from the ground up, employing methods like multi-layer perceptrons (MLPs), convolutions, or vision transformers [102, 103, 104]. This method seeks to preserve the insights gained from the domain of natural images however, its success largely depends on the prompt's quality, particularly when navigating the substantial domain shifts often present in multi-modal images [105, 106].

Adapter Tuning Updating all parameters of the Segment SAM is a resource-intensive and complex process, making widespread application challenging. Consequently, researchers have concentrated on refining a subset of SAM's parameters or incorporating lightweight adapters through various parameter-efficient fine-tuning (PEFT) techniques, rather than overhauling the entire model. For instance, [107] introduces the Medical SAM Adapter (Med-SA), which retains the original SAM parameters unchanged but integrates Low-rank Adaptation (LoRA) modules [20] at specific points within the model. Comprehensive testing across 17 medical image segmentation tasks and 5 modalities demonstrates Med-SA's enhanced performance over both the standard SAM and prior state-of-the-art methods. Similarly, SAMed [104] implements LoRA modules within the pre-trained SAM image encoder and concurrently fine-tunes this along with the prompt encoder and mask decoder on the Synapse multi-organ segmentation dataset. [102] introduces a cost-effective approach for fine-tuning SAM with a limited set of examples, combining an exemplar-guided synthesis module with a LoRA fine-tuning strategy. This method showcases SAM's effective adaptation to the medical field, even with scant labeled data. AdaptiveSAM, proposed in [21], is an adaptive enhancement designed to efficiently tailor SAM to new datasets, enabling text-prompted segmentation within the medical sphere. It utilizes bias-tuning with a significantly reduced count of trainable parameters compared to the original SAM, employing free-form text prompts for object segmentation. Experimental results indicate AdaptiveSAM's superior performance on a variety of medical imaging datasets, including surgery, ultrasound, and X-ray modalities.

To address the considerable domain disparity between natural and medical images, [108] presents SAM-Med2D, a thorough investigation into applying SAM for 2D medical imaging by incorporating learnable adapter layers within the image encoder, fine-tuning the prompt encoder, and updating the mask decoder through interactive training. This effort involved assembling a segmentation dataset of over 4.6 million images and 19.7 million masks. An extensive evaluation across various modalities, anatomical structures, and organs, including tests on 9 MICCAI 2023 challenge datasets, illustrates SAM-Med2D's significantly enhanced performance and generalizability compared to the original SAM framework.

Full Tuning This method involves a significant overhaul rather than mere tweaks, adjusting and refining both the encoder and decoder components of SAM. The aim is to adapt its broad knowledge, initially rooted in natural imagery, to suit the specialized requirements of the medical field. Although this strategy offers the promise of a more profound alignment with the distinct features of medical imagery, it also brings with it the possibility of necessitating more comprehensive training efforts and higher investment in resources [97].

4.2.2 Prompt Generation

Prompts Auto-Generation To facilitate auto-prompting, a direct method involves using a localization framework to generate input prompts for SAM. For segmenting regions of interest (ROI) across varied medical imaging datasets, [109] employs the YOLOv8 model to identify ROI bounding boxes, which serve as input prompts for SAM, enabling fully automated medical image segmentation. MedLSAM [110] introduces a few-shot localization technique that pinpoints 3D bounding boxes surrounding any anatomical structure of interest within 3D medical images. This approach is grounded in the idea that images exhibiting locally similar pixel distributions are likely to represent the same anatomical region across different individuals. From these 3D bounding boxes, 2D projections are generated for each image slice, which then guides SAM in autonomously segmenting the targeted anatomy. Furthermore, [111] develops a one-shot localization and segmentation framework that capitalizes on the relational dynamics with a template image to inform SAM's segmentation tasks. This method employs pre-trained Vision Transformer (ViT)-based foundation models to derive dense features from the template image, enhancing the precision and effectiveness of segmentation by leveraging high-quality, feature-rich inputs.

Learnable Prompts Works such as [112, 92] introduce an innovative approach by training an auxiliary prompt encoder to generate surrogate prompts, thereby circumventing the need for further fine-tuning of SAM. This encoder derives conditional prompts directly from the input image's features, moving beyond traditional prompt methods. As a result, SAM operates in a completely auto-prompting mode, eliminating the requirement for manually crafted prompts. AutoSAM sets new standards in state-of-the-art (SOTA) performance across various medical benchmarks, illustrating its effectiveness in medical image segmentation tasks. The all-in-SAM pipeline [105] leverages the pre-trained SAM to produce pixel-level annotations from weak prompts, which are then utilized to fine-tune SAM based on the strategy outlined in [74]. This approach obviates the need for manual prompts during the inference phase and has been shown to outperform previous SOTA methods in nuclei segmentation, even achieving competitive results when compared to using strongly annotated data.

To tackle the issue of poor prompts negatively affecting mask segmentation in medical imagery, [113] presents the Decoupling Segment Anything Model (DeSAM). This model decouples SAM's mask decoder to perform two distinct sub-tasks: the Prompt-Relevant IoU Module (PRIM) generates mask embeddings from the provided prompts, while the Prompt-Invariant Mask Module (PIMM) merges these embeddings with image embeddings to produce the final segmentation mask. Extensive testing reveals that DeSAM significantly enhances the robustness of SAM's fully automatic mode. [114] unveils SurgicalSAM, a variant that incorporates surgery-specific information into SAM's pre-existing knowledge base to foster better generalization. By employing a lightweight, prototype-based class prompt encoder for fine-tuning and contrastive prototype learning for enhanced class prompting, SurgicalSAM not only achieves SOTA results on two public datasets but also maintains efficiency with a minimal number of adjustable parameters.

Enhancing Reliability Against Prompts with Uncertainty Given SAM's sensitivity to input prompts, accurately estimating uncertainty is crucial to ensure the segmentation results' reliability, especially in the medical imaging domain where accuracy directly impacts clinical decisions. [115] introduces EviPrompt, a novel, training-free prompt-generation method centered on uncertainty estimation. EviPrompt can autonomously generate prompts for SAM in medical image segmentation tasks without requiring clinical expert intervention, needing only a single image-annotation pair for reference. [116] presents a method for enhancing SAM's application in segmenting fundus images through multi-box prompt-triggered uncertainty estimation, utilized as a test-time augmentation

technique. By generating varied predictions from multiple box prompts, assessing the distribution through Monte Carlo simulations, and creating an uncertainty map, this approach provides insights into potential segmentation inaccuracies, thereby increasing SAM's prompt-dependent robustness. Furthermore, [117] introduces URSAM, an uncertainty-rectified SAM framework designed to bolster the auto-prompting process for medical image segmentation. By estimating uncertainty maps and using them to correct possible errors, URSAM significantly enhances segmentation outcomes. Testing on two public 3D medical datasets for the segmentation of 35 organs showed that leveraging uncertainty could lead to improvements, even without manual prompts. Thus, incorporating uncertainty estimation not only helps in pinpointing likely segmentation mistakes but also serves as a crucial tool for clinicians, boosting the dependability of the segmentation process. This integration of uncertainty management into SAM models enhances their adaptability to diverse prompts, providing a path toward more reliable and clinically useful segmentation solutions.

4.2.3 Imaging Modalities Extension

In medical imaging, it's common for protocols to produce images in a three-dimensional (3D) format contrasting to SAM's 2D imaging capabilities. To facilitate 2D to 3D adaptation, the Medical SAM Adapter (Med-SA) [107] employs the Space-Depth Transpose (SDTrans) approach, which utilizes a dual-branch attention mechanism to capture spatial and depth correlations separately. The 3DSAM-adapter, introduced by [118], modifies the SAM architecture to enable volumetric medical image segmentation with only 16.96% of the original model's parameters being tunable. Similarly, [119] presents the modality-agnostic SAM adaptation framework (MASAM), designed for diverse volumetric and video medical datasets. MASAM integrates tunable 3D adapters within each transformer block of the image encoder and fine-tunes them alongside the mask decoder. [120] proposes ProMISe, a prompt-driven model for 3D medical image segmentation, which incorporates lightweight adapters to capture depth-related spatial context without modifying the pre-trained weights. The effectiveness of ProMISe on colon and pancreas tumor segmentation datasets highlights its superior performance over existing state-of-the-art (SOTA) methods. In a different approach, [121] introduces SAM3D, which processes each input slice individually to produce slice embeddings, later decoded by a lightweight 3D decoder for segmentation results.

To accommodate the two-dimensional (2D) design of SAM, axial slices are typically converted to fit within a 2D paradigm, with analysis performed on each slice individually for 3D images.

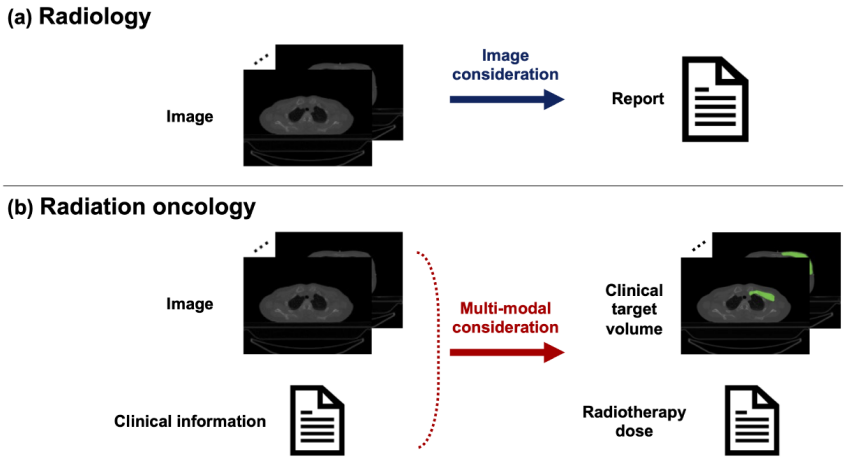


Figure 15: Workflow comparison between Radiology (upper) and Radiation Oncology (lower). Radiology focuses mainly on utilizing imaging data for diagnostic purposes, whereas Radiation Oncology combines imaging data with clinical insights from pathology findings and electronic health records of patients to guide treatment planning [123].

However, this technique does not fully capture the spatial relationships among slices, often leading to inaccuracies in the delineation of organ boundaries. [121] also proposes SAM-Med3D, a model dedicated to volumetric medical image segmentation with a fully trainable 3D SAM-like structure. Trained on a comprehensive 3D dataset containing 21K images and 131K masks across 247 categories, SAM-Med3D's extensive evaluation across 15 public datasets showcases its competitive edge, requiring significantly fewer prompt points than the best fine-tuned SAM variants in the medical field. Drawing inspiration from SAM's architecture, [122] introduces SegVol, an interactive model for volumetric medical image segmentation, particularly tailored for CT volumes. With training on 90k unlabeled and 6k labeled CT volumes, SegVol supports segmentation across over 200 anatomical categories using spatial and textual prompts, surpassing SOTA methods by a significant margin across various segmentation benchmarks.

5 Vision Language Foundation Model: Medical Image Segmentation

Traditionally, AI models have been tailored to process single data modalities, focusing either on visual or textual information. This singular approach stands in stark contrast to the inherently multi-modal method employed by medical practitioners, who rely on a combination of imaging studies and textual electronic medical records to make informed decisions. For example, in radiation oncology, segmenting the target volume poses a greater challenge due to the essential requirement to factor in clinical dimensions that extend beyond the imagery. This includes considerations like the overall stage of cancer, objectives of the treatment, and pathological results, among others [124], illustrated in

Fig. 15. The development of multi-modal AI systems, capable of interpreting and integrating diverse types of data and their interconnections, could significantly enhance the accuracy of diagnoses, tailor treatments to individual patients, and minimize medical errors by offering a holistic view of patient information. While there has been some research in radiology, mainly focusing on classification tasks that involve integrating multi-modality information, very little effort has been directed towards radio-oncology, which deals with segmentation [125]. This review focuses mainly on works with VLMs in medical image segmentation, as there has not been any significant work done in geographical image segmentation.

Fig. 16 presents a comprehensive pipeline as described by [123], illustrating the distinctions between vision-only segmentation and multi-modal segmentation in radio oncology. The figure demonstrates how the lower part of the diagram integrates data from two distinct sources: imagery and textual information. Here, the imagery refers to CT images, while the textual content comprises clinical notes provided by domain experts. This integration showcases the enhanced capability of multi-modal segmentation to leverage a richer context for medical image processing by combining visual data with relevant textual insights.

The integration of image segmentation with dual modalities, such as text and image, has been a significant advancement in enhancing segmentation outcomes. Recent developments in the field have seen the emergence of approaches that combine language capabilities into image segmentation, exemplified by language-driven semantic segmentation [126], open-vocabulary segmentation [127, 128, 129], referring segmentation [130], and reasoning segmentation [131]. This incorporation of language into segmentation processes represents a pivotal change, especially within the medical domain, where the combination of multiple modalities of knowledge is crucial. Notably, LViT [35] and ConTEXTualNet [132] have pioneered the use of text to drive segmentation in chest X-ray radiography, highlighting the growing importance of multimodal approaches in medical imaging analysis.

Further, studies such as those by [123], [124], and [125] have delved into the integration of dual modalities for segmentation tasks for radio oncology, each proposing a distinctive methodology. For example, [123] and [124] implement Large Language Models (LLMs) to analyze clinical text alongside image data. Notably, [124] expands this approach by also focusing on summarizing clinical reports and generating therapeutic plans from clinical notes, integrating segmentation with additional objectives. In contrast, [125] sets itself apart by incorporating a triplet extraction module

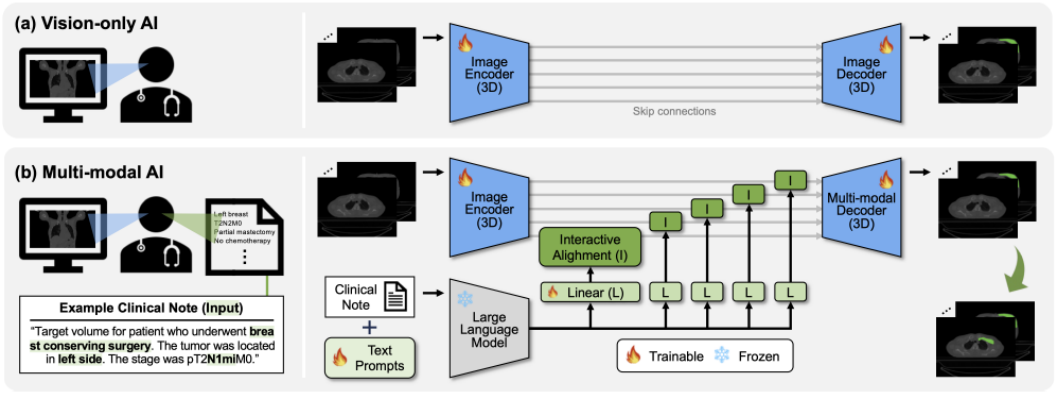


Figure 16: A graphical comparison showcasing the differences between (a) a vision-only model and (b) the newly introduced multi-modal large language model, specifically focusing on their performance in delineating target volumes [123].

that leverages a knowledge base to gather medically pertinent information. This method aims to harness extensive domain-specific knowledge. While the application of a foundation model like SAM has not yet been investigated in this dual-modality context, incorporating SAM could represent a significant advancement. Given SAM's demonstrated capability to integrate text prompts for image segmentation, its application in this domain could enhance the accuracy and efficiency of processing by leveraging both textual and visual inputs simultaneously. This approach could offer a more nuanced and comprehensive analysis, benefiting fields such as medical imaging and diagnosis where integrating multi-modal data is crucial.

6 Future Direction

The transition from traditional model development to the utilization of pre-trained foundation models like SAM has revolutionized the field of image segmentation. By employing user-generated prompts, these models offer a flexible and efficient approach to a wide array of tasks without the need for extensive retraining.

6.1 Geographical Image Segmentation

Compared to its application in natural and medical images, the utilization of foundation models like SAM in geographical imagery lags significantly. The limitations in this area are prominent, highlighting several avenues where future research efforts should be directed to enhance performance and applicability.

Large-Scale Public Datasets A significant hurdle in applying foundation models to geographical image segmentation is the lack of extensive datasets. For instance, [34] highlights the difficulty in obtaining necessary data for their study, such as masks for pedestrian infrastructures like sidewalks and crosswalks, which proved challenging to acquire. The availability of large-scale public datasets encompassing a wide range of class masks would greatly benefit this field. Fortunately, there's also the potential for researchers to employ foundation models like SAM to aid in the generation of these vital datasets [49], offering a promising solution to this data scarcity issue.

Combining Multi-Modality Information The complexity of urban environments, marked by a plethora of data types including geo-text data, street view images, trajectory data, and time series data, underscores the necessity for multimodal foundation models in geographical image segmentation [49]. The integration of these varied data modalities by multimodal foundation models is geared towards attaining a comprehensive grasp of geographical area dynamics. This integrative approach significantly boosts the efficacy of various applications and paves the way toward achieving general intelligence in the domain of urban analysis and planning. Several studies, including [133, 134, 135, 136], have ventured into the intersection of natural language processing and geographical imagery and emphasized the need for combining multi-modality data. Despite these initial explorations, there remains substantial scope for a more integrated and thorough fusion of these two domains, suggesting a promising direction for future research.

6.2 Medical Image Segmentation

Building Large-Scale Medical Datasets The assessment outcomes from multiple investigations [137, 100, 108] across diverse datasets and imaging modalities indicate that the straightforward application of SAM for medical image segmentation fails to achieve optimal results. Addressing this challenge necessitates the creation of expansive medical datasets that encompass a broad spectrum of modalities and segmentation targets. Such datasets are crucial for the development of universal medical segmentation foundation models capable of superior performance and broader applicability.

Accelerating Medical Image Annotation The process of developing segmentation models for medical imagery typically requires domain-specific knowledge to ensure annotations are both reliable and accurate [138, 139], leading to higher costs for annotation compared to those associated with natural images. This challenge is particularly present in the context of 3D volumetric medical data, a common format in the field, where specialists are tasked with the meticulous, slice-by-slice

delineation of objects. Such requirements make the annotation process not only laborious but also exceptionally time-consuming.

Incorporating Scribble and Text Prompts Several empirical studies [100, 140] have demonstrated that box prompts typically outperform point prompts in medical image segmentation due to their ability to provide more precise location information. However, the presence of multiple similar instances near the target of segmentation can make the use of a large bounding box counterproductive, as it may introduce ambiguity and lead to inaccurate segmentation results. Beyond the use of point and box prompts, the incorporation of scribble prompts has gained traction in the field of medical image segmentation [141]. This method proves to be both useful and efficient when integrated with SAM. Merging scribble prompts with either point or box prompts [142] presents a practical and potent approach for addressing targets with non-compact and irregular shapes, such as vessels, intestines, and bones, which are often characterized by their continuity and curvature.

Towards Multi-Modal Medical Images Multi-modal medical imaging is indispensable in clinical settings, offering a comprehensive view of the human body's anatomy, functionality, and pathologies through the integration of complementary information [143]. By adopting SAM to assimilate data from varied input modalities, e.g., CT and MR, there's a significant opportunity to enhance its generalization capability across diverse patient cohorts and imaging techniques, positioning it as an innovative solution for enriching clinical practices.

Incorporating Clinical Information for Radiation Oncology Text prompts have revolutionized the way clinical insights are integrated into the medical image segmentation process. In efforts to incorporate clinical information for radiation oncology, researchers have traditionally relied on a two-model approach: an LLM to interpret clinical data provided by domain experts and a separate segmentation model for image analysis [123, 124]. This methodology, while effective, necessitates extensive training and also fine-tuning the LLM. Text prompts simplify this process by offering a straightforward method to utilize domain-specific knowledge for guiding segmentation tasks, enhancing both the model's comprehension and its precision in medical image analysis. SAM's ability to utilize text information through prompts, as demonstrated in 2.3.1, indicates its potential to streamline this process. Initially, SAM utilized CLIP's text encoder for processing text prompts [11]. However, SAM's architecture allows for compatibility with various text encoders, enabling it to process a wide range of textual information. The integration of SAM in this context eliminates the necessity for additional models. Substituting SAM's prompt encoder with domain-specific text

encoders like medKLIP [125] or ClinicalBert [144], with minimal to no additional training, offers a promising avenue for achieving significant outcomes in radiation oncology segmentation tasks.

Disclaimer

Portions of this report were paraphrased using OpenAI's ChatGPT-4 to assist in correcting and rephrasing content from the original writing. The use of ChatGPT-4 was aimed at enhancing the clarity and conciseness of the presented information, not to generate creative content.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards *et al.*, "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.

[4] H. H. Lee, Y. Gu, T. Zhao, Y. Xu, J. Yang, N. Usuyama, C. Wong, M. Wei, B. A. Landman, Y. Huo *et al.*, "Foundation models for biomedical image segmentation: A survey," *arXiv preprint arXiv:2401.07654*, 2024.

[5] Z. Wang, H. Li, and R. Rajagopal, "Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1013–1020.

[6] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *arXiv preprint arXiv:2304.05215*, 2023.

[7] O. Ronneberger, P. Fischer, and T. 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*. Springer, 2015, pp. 234–241.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] C. Li, H. Bagher-Ebadian, V. Goddla, I. J. Chetty, and D. Zhu, "Focalunetr: A focal transformer for boundary-aware segmentation of ct images," *arXiv preprint arXiv:2210.03189*, 2022.

[10] D. Dais, I. E. Bal, E. Smyrou, and V. Sarhosis, "Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning," *Automation in Construction*, vol. 125, p. 103606, 2021.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[12] C. Zhang, F. D. Puspitasari, S. Zheng, C. Li, Y. Qiao, T. Kang, X. Shan, C. Zhang, C. Qin, F. Rameau *et al.*, "A survey on segment anything model (sam): Vision foundation model meets prompt engineering," *arXiv preprint arXiv:2306.06211*, 2023.

[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[14] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[15] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv preprint arXiv:2307.13721*, 2023.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[17] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[18] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[19] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[21] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel, "Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation," *arXiv preprint arXiv:2308.03726*, 2023.

[22] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[23] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," *arXiv preprint arXiv:2401.03495*, 2024.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[26] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[29] Y. Bai, X. Geng, K. Mangalam, A. Bar, A. Yuille, T. Darrell, J. Malik, and A. A. Efros, "Sequential modeling enables scalable learning for large vision models," *arXiv preprint arXiv:2312.00785*, 2023.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[31] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] R. I. Sultan, C. Li, H. Zhu, P. Khanduri, M. Brocanelli, and D. Zhu, "Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure," *arXiv preprint arXiv:2311.11319*, 2023.

[35] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[36] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in sar satellite images with deep fully convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 12, pp. 1867–1871, 2018.

[37] A. Saha, "Conducting semantic segmentation on landcover satellite imagery through u-net architectures," in *Proceedings of the Future Technologies Conference*. Springer, 2022, pp. 758–764.

[38] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 182–186.

[39] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[40] P. Gudžius, O. Kurasova, V. Darulis, and E. Filatovas, "Deep learning-based object recognition in multispectral satellite imagery for real-time applications," *Machine Vision and Applications*, vol. 32, no. 4, p. 98, 2021.

[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[42] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190.

[43] P. Gudzius, O. Kurasova, V. Darulis, and E. Filatovas, "Automl-based neural architecture search for object recognition in satellite imagery," *Remote Sensing*, vol. 15, no. 1, p. 91, 2022.

[44] M. Hosseini, A. Sevtsuk, F. Miranda, R. M. Cesar Jr, and C. T. Silva, "Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery," *Computers, Environment and Urban Systems*, vol. 101, p. 101950, 2023.

[45] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[46] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[47] J. H. Kim, S. Lee, J. R. Hipp, and D. Ki, "Decoding urban landscapes: Google street view and measurement sensitivity," *Computers, Environment and Urban Systems*, vol. 88, p. 101626, 2021.

[48] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[49] W. Zhang, J. Han, Z. Xu, H. Ni, H. Liu, and H. Xiong, "Towards urban general intelligence: A review and outlook of urban foundation models," *arXiv preprint arXiv:2402.01749*, 2024.

[50] F. Biljecki and K. Ito, "Street view imagery in urban analytics and gis: A review," *Landscape and Urban Planning*, vol. 215, p. 104217, 2021.

[51] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on intelligent transportation systems*, vol. 12, no. 3, pp. 920–939, 2011.

[52] S. Law, B. Paige, and C. Russell, "Take a look around: using street view and satellite images to estimate house prices," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 5, pp. 1–19, 2019.

[53] Y. Shi, F. Lv, X. Wang, C. Xia, S. Li, S. Yang, T. Xi, and G. Zhang, "Open-transmind: A new baseline and benchmark for 1st foundation model challenge of intelligent transportation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6327–6334.

[54] Y. Liu, X. Zhang, J. Ding, Y. Xi, and Y. Li, "Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4150–4160.

[55] T. Li, S. Xin, Y. Xi, S. Tarkoma, P. Hui, and Y. Li, "Predicting multi-level socioeconomic indicators from structural urban imagery," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3282–3291.

[56] H. J. Miller, "Tobler's first law and spatial analysis," *Annals of the association of American geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[57] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190.

[58] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[59] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[60] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022.

[61] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale

geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.

[62] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[63] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang *et al.*, "Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[64] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations," *arXiv preprint arXiv:2305.01118*, 2023.

[65] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu *et al.*, "On the opportunities and challenges of foundation models for geospatial artificial intelligence," *arXiv preprint arXiv:2304.06798*, 2023.

[66] D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang, "Scaling-up remote sensing segmentation dataset with segment anything model," *arXiv preprint arXiv:2305.02034*, 2023.

[67] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior, "The segment anything model (sam) for remote sensing applications: From zero to one shot," *arXiv preprint arXiv:2306.16623*, 2023.

[68] H. Yamagiwa, Y. Takase, H. Kambe, and R. Nakamoto, "Zero-shot edge detection with scesame: Spectral clustering-based ensemble for segment anything model estimation," *arXiv preprint arXiv:2308.13779*, 2023.

[69] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *arXiv preprint arXiv:2312.02464*, 2023.

[70] Y. Zhang, F. Zhang, and N. Chen, "Migratable urban street scene sensing method based on vision language pre-trained model," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102989, 2022.

[71] L. Haas, S. Alberti, and M. Skreta, "Learning generalized zero-shot learners for open-domain image geolocalization," *arXiv preprint arXiv:2302.00275*, 2023.

[72] X. Li, C. Wen, Y. Hu, and N. Zhou, "Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103497, 2023.

[73] S. Julka and M. Granitzer, "Knowledge distillation with segment anything (sam) model for planetary geological mapping," *arXiv preprint arXiv:2305.07586*, 2023.

[74] T. Chen, L. Zhu, C. Ding, R. Cao, S. Zhang, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more," *arXiv preprint arXiv:2304.09148*, 2023.

[75] L. Ding, K. Zhu, D. Peng, H. Tang, and H. Guo, "Adapting segment anything model for change detection in hr remote sensing images," *arXiv preprint arXiv:2309.01429*, 2023.

[76] Z. Peng, Z. Xu, Z. Zeng, X. Yang, and W. Shen, "Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction," *arXiv preprint arXiv:2308.14604*, 2023.

[77] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[78] J. Zhang, Z. Zhou, G. Mai, L. Mu, M. Hu, and S. Li, "Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models," *arXiv preprint arXiv:2304.10597*, 2023.

[79] L. Wang, M. Zhang, and W. Shi, "Cs-wscdnet: Class activation mapping and segment anything model-based framework for weakly supervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[80] X. Qi, Y. Wu, Y. Mao, W. Zhang, and Y. Zhang, "Self-guided few-shot semantic segmentation for remote sensing imagery based on large vision models," *arXiv preprint arXiv:2311.13200*, 2023.

[81] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, P. Gao, and H. Li, "Personalize segment anything model with one shot," *arXiv preprint arXiv:2305.03048*, 2023.

[82] J. Huang, K. Jiang, J. Zhang, H. Qiu, L. Lu, S. Lu, and E. Xing, "Learning to prompt segment anything models," *arXiv preprint arXiv:2401.04651*, 2024.

[83] Z. Liu, H. Wen, Z. Zhu, Q. Li, L. Liu, T. Li, W. Xu, C. Hou, B. Huang, Z. Li *et al.*, "Diagnosis of significant liver fibrosis in patients with chronic hepatitis b using a deep learning-based data integration network," *Hepatology International*, vol. 16, no. 3, pp. 526–536, 2022.

[84] K. Huang, Q. Li, W. Zeng, X. Chen, L. Liu, X. Wan, C. Feng, Z. Li, Z. Liu, and C. Dong, "Ultrasound score combined with liver stiffness measurement by sound touch elastography for staging liver fibrosis in patients with chronic hepatitis b: a clinical prospective study," *Annals of Translational Medicine*, vol. 10, no. 6, 2022.

[85] Y. Chen, C. Zhang, L. Liu, C. Feng, C. Dong, Y. Luo, and X. Wan, "Uscl: pretraining deep ultrasound image diagnosis model through video contrastive representation learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, 2021, pp. 627–637.

[86] L. Gao, R. Zhou, C. Dong, C. Feng, Z. Li, X. Wan, and L. Liu, "Multi-modal active learning for automatic liver fibrosis diagnosis based on ultrasound shear wave elastography," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 410–414.

[87] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[88] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[89] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[90] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 109–119.

[91] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted*

Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, 2021, pp. 171–180.

[92] C. Li, P. Khanduri, Y. Qiang, R. I. Sultan, I. Chetty, and D. Zhu, "Auto-prompting sam for mobile friendly 3d medical image segmentation," *arXiv preprint arXiv:2308.14936*, 2023.

[93] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[94] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[95] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," *arXiv preprint arXiv:2209.15076*, 2022.

[96] H. H. Lee, Q. Liu, S. Bao, Q. Yang, X. Yu, L. Y. Cai, T. Z. Li, Y. Huo, X. Koutsoukos, and B. A. Landman, "Scaling up 3d kernels with bayesian frequency re-parameterization for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 632–641.

[97] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[98] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, "Accuracy of segment-anything model (sam) in medical image segmentation tasks," *arXiv preprint arXiv:2304.09324*, 2023.

[99] P. Shi, J. Qiu, S. M. D. Abaxi, H. Wei, F. P.-W. Lo, and W. Yuan, "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation," *Diagnostics*, vol. 13, no. 11, p. 1947, 2023.

[100] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li, "Sam on medical images: A comprehensive study on three prompt modes," *arXiv preprint arXiv:2305.00035*, 2023.

[101] C. Mattjie, L. Vinicius de Moura, R. Cappelari Ravazio, L. Silveira Kupssinskü, O. Parraga, M. Mussi Delucis, and R. Coelho Barros, "Exploring the zero-shot capabilities of the segment

anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guideline," *arXiv e-prints*, pp. arXiv–2305, 2023.

[102] W. Feng, L. Zhu, and L. Yu, "Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars," *arXiv preprint arXiv:2308.14133*, 2023.

[103] X. Hu, X. Xu, and Y. Shi, "How to efficiently adapt large segmentation model (sam) to medical images," *arXiv preprint arXiv:2306.13731*, 2023.

[104] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[105] C. Cui, R. Deng, Q. Liu, T. Yao, S. Bao, L. W. Remedios, Y. Tang, and Y. Huo, "All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning," *arXiv preprint arXiv:2307.00290*, 2023.

[106] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson *et al.*, "Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging," *arXiv preprint arXiv:2304.04155*, 2023.

[107] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[108] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang *et al.*, "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023.

[109] S. Pandey, K.-F. Chen, and E. B. Dam, "Comprehensive multimodal segmentation in medical imaging: Combining yolov8 with sam and hq-sam models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2592–2598.

[110] W. Lei, X. Wei, X. Zhang, K. Li, and S. Zhang, "Medlsam: Localize and segment anything model for 3d medical images," *arXiv preprint arXiv:2306.14752*, 2023.

[111] D. Anand, V. Singhal, D. D. Shanbhag, S. KS, U. Patil, C. Bhushan, K. Manickam, D. Gui, R. Mullick, A. Gopal *et al.*, "One-shot localization and segmentation of medical images with foundation models," *arXiv preprint arXiv:2310.18642*, 2023.

[112] T. Shaharabany, A. Dahan, R. Giryes, and L. Wolf, "Autosam: Adapting sam to medical images by overloading the prompt encoder," *arXiv preprint arXiv:2306.06370*, 2023.

[113] Y. Gao, W. Xia, D. Hu, and X. Gao, "Desam: Decoupling segment anything model for generalizable medical image segmentation," *arXiv preprint arXiv:2306.00499*, 2023.

[114] W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang, "Surgicalsam: Efficient class promptable surgical instrument segmentation," *arXiv preprint arXiv:2308.08746*, 2023.

[115] Y. Xu, J. Tang, A. Men, and Q. Chen, "Eviprompt: A training-free evidential prompt generation method for segment anything model in medical images," *arXiv preprint arXiv:2311.06400*, 2023.

[116] G. Deng, K. Zou, K. Ren, M. Wang, X. Yuan, S. Ying, and H. Fu, "Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 368–377.

[117] Y. Zhang, S. Hu, C. Jiang, Y. Cheng, and Y. Qi, "Segment anything model with uncertainty rectification for auto-prompting medical image segmentation," *arXiv preprint arXiv:2311.10529*, 2023.

[118] S. Gong, Y. Zhong, W. Ma, J. Li, Z. Wang, J. Zhang, P.-A. Heng, and Q. Dou, "3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation," *arXiv preprint arXiv:2306.13465*, 2023.

[119] C. Chen, J. Miao, D. Wu, Z. Yan, S. Kim, J. Hu, A. Zhong, Z. Liu, L. Sun, X. Li *et al.*, "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation," *arXiv preprint arXiv:2309.08842*, 2023.

[120] H. Li, H. Liu, D. Hu, J. Wang, and I. Oguz, "Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models," *arXiv preprint arXiv:2310.19721*, 2023.

[121] N.-T. Bui, D.-H. Hoang, M.-T. Tran, and N. Le, "Sam3d: Segment anything model in volumetric medical images," *arXiv preprint arXiv:2309.03493*, 2023.

[122] Y. Du, F. Bai, T. Huang, and B. Zhao, "Segvol: Universal and interactive volumetric medical image segmentation," *arXiv preprint arXiv:2311.13385*, 2023.

[123] Y. Oh, S. Park, H. K. Byun, J. S. Kim, and J. C. Ye, "Llm-driven multimodal target volume contouring in radiation oncology," *arXiv preprint arXiv:2311.01908*, 2023.

[124] K. Kim, Y. Oh, S. Park, H. K. Byun, J. S. Kim, Y. B. Kim, and J. C. Ye, "Ro-llama: Generalist llm for radiation oncology via noise augmentation and consistency regularization," *arXiv preprint arXiv:2311.15876*, 2023.

[125] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training," *medRxiv*, pp. 2023–01, 2023.

[126] W. He, S. Jamonnak, L. Gou, and L. Ren, "Clip-s4: Language-guided self-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 207–11 216.

[127] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary universal image segmentation with maskclip," 2023.

[128] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070.

[129] S. Yun, S. H. Park, P. H. Seo, and J. Shin, "Ifseg: Image-free semantic segmentation via vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2967–2977.

[130] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 686–11 695.

[131] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[132] Z. Huemann, J. Hu, and T. Bradshaw, "Contextual net: A multimodal vision-language model for segmentation of pneumothorax," *arXiv preprint arXiv:2303.01615*, 2023.

[133] P. Balsebre, W. Huang, G. Cong, and Y. Li, "City foundation models for learning general purpose representations from openstreetmap," *arXiv e-prints*, pp. arXiv–2310, 2023.

[134] R. Shao, C. Yang, Q. Li, Q. Zhu, Y. Zhang, Y. Li, Y. Liu, Y. Tang, D. Liu, S. Yang *et al.*, "Allspark: a multimodal spatiotemporal general model," *arXiv preprint arXiv:2401.00546*, 2023.

[135] R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang, "Velma: Verbalization embodiment of llm agents for vision and language navigation in street view," *arXiv preprint arXiv:2307.06082*, 2023.

[136] M. M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, "Vision-language models for zero-shot classification of remote sensing images," *Applied Sciences*, vol. 13, no. 22, p. 12462, 2023.

[137] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen *et al.*, "Segment anything model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[138] R. Jiao, Y. Zhang, L. Ding, R. Cai, and J. Zhang, "Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. arxiv," 2022.

[139] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[140] S. Roy, T. Wald, G. Koehler, M. R. Rokuss, N. Disch, J. Holzschuh, D. Zimmerer, and K. H. Maier-Hein, "Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model," *arXiv preprint arXiv:2304.05396*, 2023.

[141] K. Zhang and X. Zhuang, "Zscribbleseg: Zen and the art of scribble supervised medical image segmentation," *arXiv preprint arXiv:2301.04882*, 2023.

[142] H. E. Wong, M. Rakic, J. Guttag, and A. V. Dalca, "Scribbleprompt: Fast and flexible interactive segmentation for any medical image," *arXiv preprint arXiv:2312.07381*, 2023.

[143] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.

[144] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.