Applications, Deployment Challenges, and Research Gaps of the Segment Anything Model Family: A Review

Guanyu Chen

School of Computer Science, Guangdong University of Technology, Guangzhou, China
1260098198@qq.com

Abstract Vision foundation models have developed rapidly, and the Segment Anything Model (SAM) family is a representative line of work in promptable segmentation. This paper reviews the development from the original SAM to SAM 2 and SAM 3, and compares two domain-adapted approaches: SAM-Med2D for medical imaging and SUGFW for industrial defect detection. Based on the literature reviewed, we summarize practical challenges at the technical, data, and deployment levels, and identify open questions on lightweight deployment, few-shot adaptation, multimodal prompting, and long-horizon video segmentation. The evidence suggests that the SAM family is a flexible general-purpose framework, but using it reliably in specialized fields still needs domain adaptation, representative validation data, and careful engineering for deployment.

Keywords Segment Anything Model; promptable segmentation; domain adaptation; medical image segmentation; industrial defect detection

1. Introduction

1.1 Research Background and Significance

Traditional segmentation models, such as U-Net [1] and Mask R-CNN [2], are trained in a fully supervised way for one fixed task. When the categories or the scene change, the model has to be trained again, and pixel-level labels are expensive to get, especially in medicine and manufacturing. In 2023, Meta released the original SAM [3], which introduced promptable segmentation through points, boxes, and masks. SAM 2 [4] extended this idea to video through memory-based temporal modeling, and the recent SAM 3 [5] further explores concept-level prompting and open-vocabulary segmentation.

However, SAM was trained on SA-1B, a dataset of natural images. Medical and industrial images look very different from natural photos in pixel values, textures, and object shapes, so the performance of zero-shot SAM often degrades substantially in these fields [6,7]. To deal with this domain gap, many adapted versions have been proposed. This paper uses two representative cases: SAM-Med2D [6], which represents data-intensive medical adaptation, and SUGFW [7], which represents industrial adaptation under a very small labeling budget.

1.2 Research Status and Review Scope

Current research broadly follows two directions. One line tracks the general-purpose evolution of the SAM family, from promptable image segmentation to memory-based video segmentation and concept-level prompting. The other line focuses on domain adaptation, where researchers modify the data, the prompt strategy, or the trainable modules. The two cases above were chosen because they stand for two typical adaptation routes: using large domain data, and saving labels with active learning. Existing studies often focus on a single SAM version, a single technical issue, or one domain, so this paper compares the model evolution, the adaptation strategies, and the deployment constraints together. The practical goal is to clarify when SAM can be transferred directly, when domain adaptation is required, and what deployment trade-offs remain.

1.3 Paper Structure

The rest of this paper is organized as follows. Section 2 introduces the theoretical background and related work. Section 3 describes the three SAM generations and their typical applications. Section 4 discusses deployment problems at the technical, data, and engineering levels. Section 5 summarizes research gaps and future directions, and Section 6 concludes the paper.

2. Theoretical Foundations and Literature Review

2.1 Core Concepts of Visual Segmentation

Visual segmentation divides an image into meaningful regions. Semantic segmentation labels every pixel with a category, instance segmentation separates different object instances, and panoptic segmentation combines the two. Traditional methods are trained on a fixed dataset with manual masks, so they generalize poorly to new scenes and depend on expensive labels. Prompt-based interaction offers a different way: the user guides a pretrained model at inference time with points, boxes, masks, or text. SAM is a representative segmentation foundation model because it combines large-scale pretraining with such a promptable interface [3].

2.2 Core Theories Supporting the SAM Family

2.2.1 Transformer-Based Image Segmentation

SAM is built on the Vision Transformer (ViT) [8], which cuts an image into patches and uses self-attention [9] to model the relations between them. Unlike a CNN, where the receptive field grows slowly layer by layer, a Transformer lets every position collect information from the whole image in one step. This helps the model understand long-range context and objects of very different sizes. Because the encoder learns general rather than category-specific features, these features contribute to the transferable representation used by SAM. The large-scale pretraining and the promptable design also play important roles in its zero-shot ability.

2.2.2 Interactive Prompt Learning

SAM has three parts: an image encoder, a prompt encoder, and a mask decoder [3]. The prompt encoder turns points, boxes, and masks into feature vectors, and the mask decoder combines them with the image features. For an ambiguous prompt such as a single click, the decoder predicts several candidate masks, and the user can add positive or negative clicks to fix mistakes. This turns segmentation into an interactive process, and it also makes SAM a useful labeling tool, because correcting a mask can reduce annotation effort compared with drawing one from scratch.

2.2.3 Temporal Modeling for Video Segmentation

In videos, the same object should look consistent across frames, but it may deform, be occluded, or leave the view. SAM 2 keeps a memory of the target: a memory bank stores features and predictions of past frames, and a memory attention module lets the current frame use this memory before decoding the mask [4]. The mask is then propagated through the video as a sequence called a masklet, so the user only needs to give prompts on one or a few frames.

2.3 Relation between SAM Design and Traditional Segmentation Theories

2.3.1 Breakthrough of the Promptable Paradigm over Traditional Frameworks

The biggest difference between SAM and traditional networks is how the task is defined. Traditional models have a closed category set fixed before training. SAM instead treats segmentation as a prompt-guided process, and the number of possible objects is not limited by a category list. This can reduce the need for class-specific retraining in domains close to the pretraining distribution, and ambiguities that are hard to describe in training labels can be solved interactively at inference time.

2.3.2 Academic Debates: Universal Models versus Task-Specific Models

Whether a universal model can replace task-specific ones is still an open question. Zero-shot tests in medical imaging show that SAM often fails on CT, MR, and ultrasound images [6,10], while fine-tuned SAM-based models can reach strong results. SAM also returns masks without category labels, so tasks like driving scene parsing still need an extra classification module. In addition, for factories with only a few fixed defect types, a small task-specific model may be cheaper and faster. These findings suggest that foundation and task-specific models may be complementary.

2.4 Survey of Related Research

Existing studies on SAM can be grouped into two classes. The first class follows the general-purpose evolution of the family, from image segmentation to video and concept-level segmentation; Section 3 discusses this line in detail. The second class studies domain adaptation. In medicine, SAM-Med2D fine-tuned SAM on about 4.6 million medical images with parameter-efficient adapters [6]. SAM2-UNet inserts adapters into the Hiera encoder of SAM 2 and uses a U-shaped decoder, improving several natural and medical segmentation tasks with low training cost [11]. In industry, SUGFW uses the uncertainty of SAM's zero-shot predictions to select samples for active learning [7].

2.5 Critical Summary of the Literature

The literature reviewed here has three recurring limitations. First, studies often cover only one SAM version or one domain, so cross-generation and cross-domain comparison is difficult. Second, technical, data, and engineering challenges are usually discussed separately, although they interact strongly in deployment. Third, most adaptation studies follow the same pattern of data collection and fine-tuning, while the mechanism of transferable domain knowledge remains unclear. These gaps motivate the review framework of this paper: model evolution, domain adaptation, deployment challenges, and future directions.

ISSN 3106-7018 (Print) ISSN 3106-5791 (Online)

3. Iterative Evolution of the SAM Family and Typical Applications

3.1 The Original SAM: Architecture, Capability and Native Limitations

The first SAM was trained on 11 million natural images and more than one billion masks from the SA-1B dataset [3]. Its main contribution is zero-shot segmentation: with a simple prompt, users can segment almost any object in a natural image without task-specific training. Meta also released a data engine in which the model helps label new data and the new data improves the model again. However, the first SAM only works on static images. Its performance drops sharply on professional images such as medical scans [6], and it handles small targets such as tiny lesions or thin scratches poorly.

3.2 SAM 2: Unified Image and Video Segmentation

SAM 2 defines the Promptable Visual Segmentation (PVS) task, which handles images and videos in one framework [4]. The user can give clicks, boxes, or masks on any frame, and the model returns a mask immediately and propagates it through the video. Its architecture has four parts: a Hiera image encoder [12] pre-trained with masked auto-encoding [13], a memory attention module, a prompt encoder, and a mask decoder. Each frame is encoded only once in a streaming way, so the model can run in near real time.

To train SAM 2, Meta built a model-in-the-loop data engine and released the SA-V dataset, which is much larger than earlier video datasets such as YouTube-VOS [4,14]. In experiments, SAM 2 needs about three times fewer user interactions than earlier video methods, and on images it is both more accurate and about six times faster than the first SAM [4]. Remaining problems include target loss after heavy occlusion, mask drift, and the lack of category understanding.

3.3 SAM 3: From Geometric Prompts to Concept Prompts

SAM 3 moves from segmenting single objects to segmenting concepts [5]. It defines the Promptable Concept Segmentation (PCS) task. The user gives a short noun phrase, such as "yellow school bus", or positive and negative example boxes. The model must then detect, segment, and track all matching objects in images and videos. SAM 3 consists of a DETR-style detector [15] and a SAM 2-style tracker that share one vision-language backbone [5,16]. A key design is the presence token, which separates judging whether the concept exists from locating the objects, and improves open-vocabulary detection. Its data engine combines human annotators, AI annotators, and AI verifiers, and produced the SA-Co benchmark with 207K concepts on 120K images and 1.7K videos [5].

On benchmarks, the paper reports a zero-shot mask AP of 48.8 on LVIS, compared with 38.5 for the best previously reported system [5,17]. On SA-Co/Gold, measured with the cgF1 metric, SAM 3 scores 54.1, more than double the strongest baseline OWLv2* (24.6), and reaches 74% of the estimated human performance [5]. These gains come with higher cost: the model has about 848 million parameters, and the reported speed of 30 ms per image was measured on an H200 GPU [5], which is far from edge devices.

3.4 Typical Vertical Applications of the SAM Series

3.4.1 Medical Image Segmentation: SAM-Med2D

Medical images have many modalities, complex structures, and very few labels, so using SAM directly gives poor results [6,10]. SAM-Med2D is a representative large-scale study of adapting SAM to 2D medical images [6]. The authors collected about 4.6 million images and 19.7 million masks, covering ten imaging types and 31 organs. They froze the image encoder, inserted adapters into each Transformer block, and fine-tuned the prompt encoder and mask decoder with point, box, and mask prompts together, while earlier works usually supported only one prompt type [18]. On nine unseen MICCAI 2023 challenge datasets, SAM-Med2D clearly outperformed the original SAM and generalized better [6]. This case shows that large domain data plus full-prompt fine-tuning is a practical medical route, but it needs millions of labels.

3.4.2 Industrial Defect Detection: SUGFW Active Learning Framework

Industrial quality inspection faces the opposite situation: defect types are diverse, but the labeling budget is very small. SUGFW uses SAM as a feature extractor rather than the final segmenter [7]. It measures the uncertainty of SAM's masks under test-time augmentation, uses the uncertainty values to weight regional features, and selects the most representative samples for human labeling with a greedy strategy. The small labeled set is then used to fine-tune SAM with adapters. On the NEU-Seg steel surface defect dataset, SUGFW beat existing methods by at least 4 percentage points in mean IoU, with only 3%, 10%, or 20% of the data labeled. With 20% labeling, it reached an IoU of 0.8423, close to the 0.8599 of full labeling [7].

3.4.3 Autonomous Driving: A Potential Application

Autonomous driving needs category labels and real-time speed at the same time, which conflicts with SAM's interactive and label-free design. A recent survey on image segmentation for intelligent transportation systems points out that foundation-model-based segmentation is promising for road scene understanding, but real-time performance and safety-critical reliability remain the main barriers [19]. At present, the evidence for SAM in driving scenes mainly comes from survey papers rather than primary experimental studies. In practice, current work mostly uses SAM offline, for example as a labeling tool to build driving datasets faster.

3.4.4 Other Vertical Fields

SAM has also been tried in agriculture and robotics. In agriculture, Carraro et al. evaluated zero-shot SAM on crop and field imagery, and found that it works without training but is sensitive to prompt choice and field conditions [20]. In robotics, GraspSAM extends SAM with adapters for prompt-driven grasp detection and reports strong results on several grasping benchmarks [21]. Most of these works follow the same routine: collect a domain dataset, fine-tune SAM with adapters, and test on their own benchmarks. Careful comparisons across fields and models are still rare.

4. Multi-Dimensional Dilemmas in the Practical Deployment of SAM

4.1 Technical-Level Dilemmas

4.1.1 Weak Cross-Domain Generalization

Multiple studies report a clear performance drop when the target domain is very different from natural images [6,7,10]. This is not surprising, because the pretraining data of SAM is dominated by natural photos. For reliable use in a specialized field, domain data, proper prompting, or some adaptation is usually needed.

4.1.2 Insufficient Accuracy for Fine-Grained and Small Targets

SAM can also struggle with small, thin, or low-contrast targets, such as tiny lesions, small polyps, scratches, or pinholes [6,7]. One possible contributing factor is the loss of fine spatial detail during down-sampling, but dataset quality, prompt design, and task-specific factors also matter, so this should not be read as a confirmed root cause.

4.1.3 Unstable Temporal Tracking in Long Videos

For long videos, wrong predictions can enter the memory of SAM 2 and cause mask drift or target loss after heavy occlusion or re-appearance [4]. SAM 3 improves this by re-prompting the tracker with confident detections from time to time [5], but the trade-off between memory length, identity consistency, and computation cost is still open.

4.2 Data-Level Dilemmas

4.2.1 Scarcity of Professional Annotated Data

Fine-tuning SAM for a specialized field may need very large annotated datasets, as the millions of masks used by SAM-Med2D show [6]. Professional labels are expensive: medical annotation needs clinical expertise, and industrial defects are rare and imbalanced [7]. Pretraining and benchmark distributions are also uneven: SA-1B is dominated by general visual content, and strong results on one benchmark do not guarantee robustness across hospitals, factories, or devices [3,6].

4.2.2 High Cost of Interactive Annotation

Interactive prompts make each label cheaper than drawing masks by hand, but the total human cost is still high at large scale. In the data engines of SAM 2 and SAM 3, annotators still need to check and correct predictions frame by frame and object by object [4,5]. The low annotation cost of SAM only becomes real after the model has been adapted to the target field.

4.3 Engineering Deployment Dilemmas

4.3.1 Large Model Parameter Scale

The engineering cost is measurable. The original SAM with a ViT-H encoder has about 636 million parameters, and SAM 3 has about 848 million [3,5]. SAM 3 processes an image with over one hundred objects in 30 ms, but only on an H200 data-center GPU [5]; on CPUs and edge devices, such models are hard to run in real time. Lightweight variants cut the image encoder to only a few million parameters [22,23], but usually with some accuracy loss.

4.3.2 Slow Real-Time Inference on Terminals

Many practical applications, such as autonomous driving and production-line inspection, require low-latency segmentation [19]. The SAM family was designed for broad capability rather than minimal edge latency, and memory-based video processing adds further cost [4]. Efficient backbones and lightweight variants reduce part of this gap [22,23], but deployment still requires careful measurement of latency, memory use, and accuracy on the target hardware.

4.3.3 Limited System-Level Deployment Evaluation

Finally, many reviewed studies emphasize accuracy over latency, memory use, and update cost. They often load a pretrained SAM, collect a domain dataset, add an adapter, and report accuracy on their own benchmark. System-level evaluation on real hardware is still limited. Engineering research on compression, quantization, and system optimization is therefore as important as algorithm research for the SAM family.

5. Current Research Gaps and Future Development Trends

5.1 Existing Core Research Gaps

Three gaps stand out in the literature reviewed here. First, systematic analysis of when and why prompting fails is still limited: few works compare prompt types, error patterns, and cross-domain conditions in a controlled way [3,4,5]. Second, parameter-efficient training is often confused with efficient inference: adapters lower training cost but do not reduce the computational burden of the frozen backbone [24,25], and genuinely efficient variants are still limited [22,23]. Third, there is no reusable few-shot adaptation system for vertical industries: SAM-Med2D relies on millions of labels [6], while SUGFW saves labels but stays domain-specific [7].

ISSN 3106-7018 (Print) ISSN 3106-5791 (Online)

5.2 Future Development Trends

Four directions appear particularly relevant. The first is lightweight deployment, combining quantization, distillation, efficient backbones, and parameter-efficient fine-tuning for phones, embedded cameras, and vehicles [22,23,24,25]. A second direction is few-shot and zero-shot domain adaptation, for example by combining parameter-efficient fine-tuning with active learning, as SUGFW does [7]. The third is stable long-horizon video segmentation, with hierarchical memory designs and jointly trained detector-tracker systems [5]. The fourth direction is multimodal prompt fusion. SAM 3 already accepts text phrases and image examples [5], and future systems may support richer multimodal prompts.

6. Conclusions and Outlook

6.1 Main Conclusions

This paper reviewed the development of the SAM family and the problems of using it in real applications. The original SAM brought promptable segmentation to images, SAM 2 extended it to videos with a memory design, and SAM 3 moved to concept-level segmentation with a detector-tracker design. Two cases show different adaptation routes: SAM-Med2D uses large domain data and full-prompt fine-tuning, while SUGFW reuses SAM's features to select samples under a very small labeling budget. The main bottlenecks are weak cross-domain generalization, scarce professional labels, and high engineering cost.

6.2 Limitations of This Review

This review is based on literature reading and case comparison, without our own experiments, so the performance conclusions come from the cited papers. It is a narrative review rather than a systematic review, and the selected cases illustrate different adaptation routes rather than an exhaustive comparison of all SAM variants. In addition, this paper does not make cross-dataset quantitative comparisons, so it cannot judge which adaptation strategy is best overall. Other universal segmentation models, such as SEEM [26] and SegGPT [27], and fields such as remote sensing, are not discussed in detail.

6.3 Outlook

In the near term, combining parameter-efficient fine-tuning with active learning is a promising direction for data-scarce settings. In the longer term, connecting promptable segmentation with multimodal large models may move the field from segment anything towards understand and segment anything, but this still requires systematic evaluation across domains.

References

[1] Ronneberger, O., Fischer, P., Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany.

[2] He, K., Gkioxari, G., Dollár, P., et al. (2017) Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy.

[3] Kirillov, A., Mintun, E., Ravi, N., et al. (2023) Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France.

[4] Ravi, N., Gabeur, V., Hu, Y.-T., et al. (2025) SAM 2: Segment Anything in Images and Videos. Proceedings of the International Conference on Learning Representations (ICLR), Singapore.

[5] Carion, N., Gustafson, L., Hu, Y.-T., et al. (2026) SAM 3: Segment Anything with Concepts. Proceedings of the International Conference on Learning Representations (ICLR).

[6] Cheng, J., Ye, J., Deng, Z., et al. (2023) SAM-Med2D. arXiv preprint arXiv:2308.16184.

[7] Ma, X., Fu, J., Wang, L., et al. (2025) Application of Active Learning to Defect Detection Guided by SAM Feature. Electro-Mechanical Engineering, 41(3), 80-86.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations (ICLR), virtual.

[9] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, USA.

[10] Mazurowski, M. A., Dong, H., Gu, H., et al. (2023) Segment Anything Model for Medical Image Analysis: An Experimental Study. Medical Image Analysis, 89, 102918.

[11] Xiong, X., Wu, Z., Tan, S., et al. (2026) SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation. Visual Intelligence, 4, 2.

[12] Ryali, C., Hu, Y.-T., Bolya, D., et al. (2023) Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. Proceedings of the International Conference on Machine Learning (ICML), Honolulu, USA.

[13] He, K., Chen, X., Xie, S., et al. (2022) Masked Autoencoders Are Scalable Vision Learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA.

[14] Xu, N., Yang, L., Fan, Y., et al. (2018) YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany.

[15] Carion, N., Massa, F., Synnaeve, G., et al. (2020) End-to-End Object Detection with Transformers. Proceedings of the European Conference on Computer Vision (ECCV), virtual.

[16] Radford, A., Kim, J. W., Hallacy, C., et al. (2021) Learning Transferable Visual Models from Natural Language Supervision. Proceedings of the International Conference on Machine Learning (ICML), virtual.

[17] Gupta, A., Dollár, P., Girshick, R. (2019) LVIS: A Dataset for Large Vocabulary Instance Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA.

[18] Ma, J., He, Y., Li, F., et al. (2024) Segment Anything in

Medical Images. Nature Communications, 15, 654.
[19] Akter, S., Shihab, I. F., Sharma, A. (2026) Image Segmentation with Large Language Models: A Survey with Perspectives for Intelligent Transportation Systems. IEEE Transactions on Intelligent Transportation Systems, early access, doi: 10.1109/TITS.2026.3695429.
[20] Carraro, A., Sozzi, M., Marinello, F. (2023) The Segment Anything Model (SAM) for Accelerating the Smart Farming Revolution. Smart Agricultural Technology, 6, 100367.
[21] Noh, S., Kim, J., Nam, D., et al. (2025) GraspSAM: When Segment Anything Model Meets Grasp Detection. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Atlanta, USA, 14023-14029.
[22] Shu, H., Li, W., Tang, Y., et al. (2025) TinySAM: Pushing the Envelope for Efficient Segment Anything Model. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Philadelphia, USA.
[23] Xiong, Y., Varadarajan, B., Wu, L., et al. (2024) EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA.
[24] Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019) Parameter-Efficient Transfer Learning for NLP. Proceedings of the International Conference on Machine Learning (ICML), Long Beach, USA.
[25] Hu, E. J., Shen, Y., Wallis, P., et al. (2022) LoRA: Low-Rank Adaptation of Large Language Models. Proceedings of the International Conference on Learning Representations (ICLR), virtual.
[26] Zou, X., Yang, J., Zhang, H., et al. (2023) Segment Everything Everywhere All at Once. Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA.
[27] Wang, X., Zhang, X., Cao, Y., et al. (2023) SegGPT: Towards Segmenting Everything in Context. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France.