A Deep Learning Approach to Image Segmentation: Utilizing Personalize SAM Model

Bantwal Vaibhav Mallya
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, United States
bmallya3@gatech.edu

Karthikay Gundepudi
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, United States
kgundepudi3@gatech.edu

Chunyuan Deng
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, United States
cdeng73@gatech.edu

***Abstract*—The Segment Anything Model (SAM) has emerged as a robust framework in segmentation, offering remarkable versatility as a promptable framework. However, customizing SAM for specific visual concepts without manual prompting remains unexplored. This paper introduces a training-free approach, Personalize-SAM (PerSAM), for personalized object segmentation. PerSAM utilizes just one-shot data—a single image with reference masks—to empower personalized segmentation capabilities. Through target-semantic prompting techniques, PerSAM efficiently tailors SAM to individual users without extensive training.**

Keywords: Personalize-SAM, Prompting, SAM, Segmentation.

I. Introduction

Driven by the increasing availability of large-scale datasets and computational resources, recent advancements in models across language, vision, and related domains have been remarkable. These models, particularly in segmentation tasks like Segment Anything Model (SAM), showcase exceptional capabilities and functionalities, often incorporating human feedback to improve performance. While SAM [1] introduces a promptable segmentation framework, facilitating the segmentation of diverse objects in visual contexts, it encounters limitations when it comes to segmenting specific visual concepts without manual prompting.

Consider scenarios where we aim to isolate cats or dogs from cluttered photo albums or locate a particular object within a room. Traditionally, accomplishing such tasks using Vanilla SAM [2] or similar models like Mask R-CNN [3] would require intensive manual effort and time. Recognizing this challenge, we explore PerSAM [4], a training-free personalization approach tailored for SAM.

PerSAM operates on just one-shot data—a reference image and a rough mask of the targeted concept—to efficiently customize SAM for object segmentation. Unlike traditional approaches that rely on extensive training or manual intervention, PerSAM streamlines the personalization process, making it accessible and efficient. By leveraging the available data, PerSAM empowers SAM to accurately segment specific visual concepts with minimal user input.

The key innovation of PerSAM lies in its ability to adapt SAM to individual user needs without the need for extensive training or manual prompting efforts. Through techniques such as target-guided attention and target-semantic prompting, PerSAM enables SAM to focus on foreground regions and incorporate high-level semantics for precise segmentation.

In summary, PerSAM offers a practical and effective solution for customizing SAM for personalized object segmentation. By leveraging one-shot data, PerSAM streamlines the segmentation process, making it more accessible and efficient for various applications, from isolating objects in images to locating specific items within complex visual contexts. In Section II, we explore similar studies that were conducted in this same domain. Section III explores our methodology. The Results and Comparisons are included in Section IV. Lastly we have conclusions in Section V and Project Contributions in Section VI.

II. Literature Review

Over the years, segmentation models have evolved significantly, each contributing to the advancement of computer vision tasks. In this literature review, we discuss the chronological growth of our exploration, comparison, and final selection of the PerSAM model for our research.

Kirillov et al. [1] presented the Segment Anything model, which introduced a promptable segmentation framework capable of segmenting diverse objects in visual contexts. While versatile, SAM lacks the ability to segment specific visual concepts without manual prompting.

Kavur et al. [2] discussed ensemble methods, demonstrating that combining multiple vanilla-style deep learning models can improve segmentation performance. By leveraging the diversity of individual models, ensembles enhance segmentation accuracy and robustness.

He et al. [3] introduced Mask R-CNN, a versatile instance segmentation model capable of predicting segmentation masks for each instance in an image. Although initially designed for object detection, Mask R-CNN's capabilities extend to semantic segmentation tasks.

Finally, Zhang et al. [4] addressed this limitations, by proposing PerSAM, a method to personalize the Segment Anything Model with one-shot data. By leveraging available data efficiently, PerSAM enables customization of SAM for specific visual concepts, expanding its applicability to a wider range of scenarios.

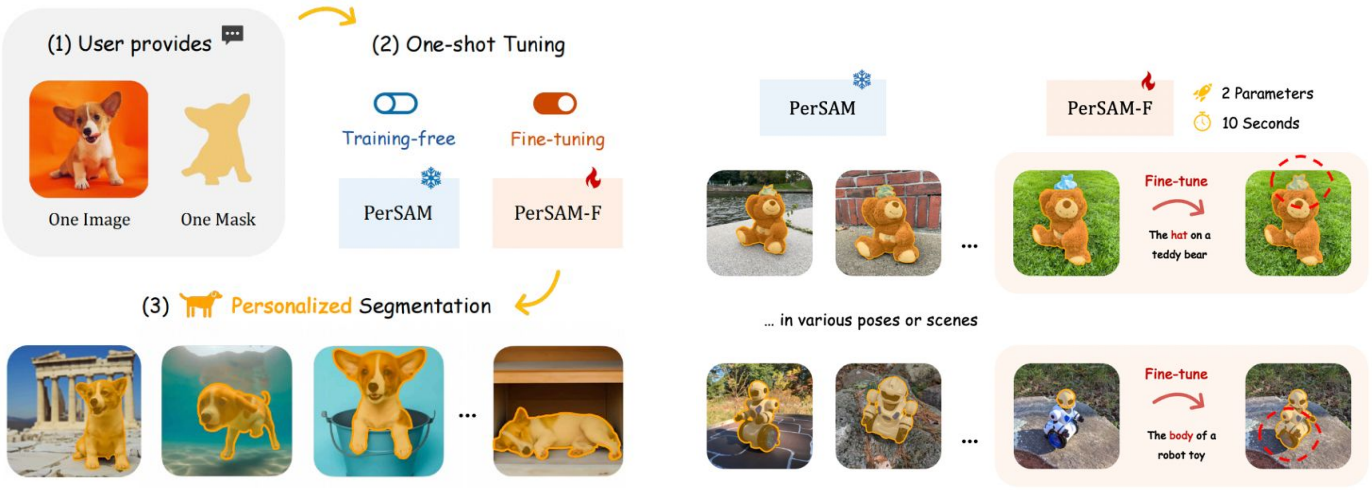

Fig. 1. Cycle of Personalize SAM

Ronneberger et al. [5] proposed the U-Net architecture, which laid the foundation for many subsequent segmentation models. Its encoder-decoder structure with skip connections revolutionized biomedical image segmentation.

Through a chronological exploration and comparison of various segmentation models, we ultimately selected PerSAM as the most suitable approach for our research, demonstrating its effectiveness in personalized object segmentation.

III. Methods

Our journey towards achieving personalized object segmentation began with an exploration of the U-Net architecture, a renowned framework for biomedical image segmentation. However, we encountered several limitations with U-Net, such as the tendency to produce overly smooth segmentations, especially around object boundaries, and the requirement for large amounts of annotated data for training. These drawbacks led us to seek an alternative solution that could better address the specific challenges of our datasets.

Subsequently, we turned our attention to the Personalize Segment Anything Model with One Shot (PerSAM) approach proposed by Zhang et al. [4]. PerSAM offers a promising methodology for customizing segmentation models with minimal data requirements, making it an attractive option for our research objectives. The key innovation of PerSAM lies in its ability to personalize the Segment Anything Model (SAM) using only one-shot data, comprising a reference image and a corresponding mask of the target concept.

The methodology of PerSAM involves injecting high-level target semantics into SAM through training-free techniques. By leveraging the available one-shot data efficiently, PerSAM empowers SAM to segment specific visual concepts without the need for manual prompting. This personalized approach enhances the versatility of SAM, enabling it to adapt to a wider range of scenarios and datasets.

Mathematically, the target-guided attention mechanism in PerSAM can be represented as follows:

$$\text{Attention}(x_i, x_j) = \frac{\exp(\theta(x_i) \cdot \phi(x_j))}{\sum_k \exp(\theta(x_i) \cdot \phi(x_k))}$$

where x_i and x_j represent the input values, θ and ϕ are learnable parameters, and " $\cdot$ " denotes the dot product operation.

Additionally, PerSAM incorporates target-semantic prompting to provide SAM with high-level semantic information about the target object, further enhancing its segmentation capabilities. This can be expressed through the following formula:

$$h_t = \text{ReLU}(W_t \cdot [h_{t-1}, x_t])$$

where h_t represents the hidden state at time t, W_t is a learnable weight matrix, $[h_{t-1}, x_t]$ denotes the concatenation of the previous hidden state and the current input feature, and ReLU is the rectified linear activation function.

The methodology of PerSAM involves injecting high-level target semantics into SAM through training-free techniques. By leveraging the available one-shot data efficiently, PerSAM empowers SAM to segment specific visual concepts without the need for manual prompting. This personalized approach enhances the versatility of SAM, enabling it to adapt to a wider range of scenarios and datasets.

One of the core principles underlying PerSAM is the concept of target-guided attention, which directs SAM's attention towards foreground target regions for intensive feature aggregation. Additionally, PerSAM incorporates target-semantic prompting to provide SAM with high-level semantic information about the target object, further enhancing its segmentation capabilities.

Overall, the methodology of PerSAM represents a significant advancement in the field of personalized object segmentation. By leveraging one-shot data and integrating high-level

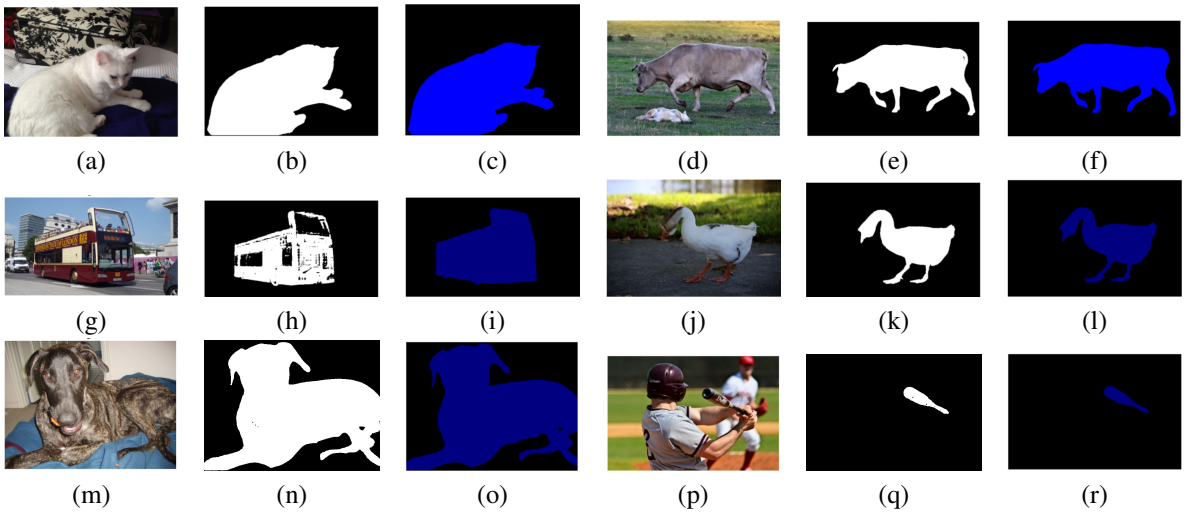

Fig. 2. (a) Original cat image; (b) Prompt cat mask; (c) Predicted cat mask; (d) Original cow image; (e) Prompt cow mask; (f) Predicted cow mask; (g) Original bus image; (h) Prompt bus mask; (i) Predicted bus mask; (j) Original bird image; (k) Prompt bird mask; (l) Predicted bird mask; (m) Original dog image; (n) Prompt dog mask; (o) Predicted dog mask; (p) Original baseball bat image; (q) Prompt baseball bat mask; (r) Predicted baseball bat mask;

TABLE I
MEAN IOU SCORES FOR PERSAM MODEL FOR EACH STUDENT ACROSS EVERY DATASET

Object Category	Student 1	Student 2	Student 3	Object Category	Student 1	Student 2	Student 3
Baseball Bat	0.6347	0.6123	0.6238	Dog	0.8847	0.8265	0.8579
Bird	0.6035	0.6581	0.6649	Dolphin Above	0.7233	0.7981	0.7685
Breast	0.6924	0.7134	0.7256	Dolphin Below	0.8696	0.8378	0.8167
Bus	0.7989	0.8136	0.8294	Polyp	0.8361	0.8289	0.8214
Cat	0.9123	0.8757	0.8091	Salt Dome	0.8435	0.8206	0.8553
Chalk Group	0.6681	0.6932	0.6753	Skin	0.5439	0.6837	0.5169
Clock	0.8745	0.7801	0.7378	Stop Sign	0.8741	0.8654	0.8769
Cow	0.8199	0.8207	0.8184	Tie	0.6849	0.6378	0.7235

target semantics into SAM, PerSAM offers a practical and effective solution for customizing segmentation models to suit specific visual concepts and datasets.

IV. RESULTS

In this study, we evaluated the performance of the Personalized Segment Anything Model (PerSAM) across various object categories using Intersection over Union (IoU) scores. Prior to implementing the PerSAM model, the experiment involved three students segmenting datasets containing images of different objects using the SAM prompts. The segmentation masks generated by the PerSAM model were compared with the segmentation masks provided by each student using SAM prompts.

Table I presents the average Intersection over Union (IoU) scores achieved by the PerSAM model for each student across all datasets. The results reveal varying levels of performance across different object categories. Generally, the PerSAM model demonstrates robust performance, achieving high IoU scores for several object categories. However, there are instances of variability in performance between students and object categories. Figure 2 shows the original RGB image, prompt mask and predicted mask of various object categories.

The PerSAM model consistently performs well across a majority of object categories, with IoU scores ranging from moderate to high. Notably, categories such as Bus, Cat, Cow, Dog, Dolphin Below, Polyp, Salt Dome and Stop Sign exhibit high ($\geq$ 0.8) IoU scores, indicating accurate segmentation by the PerSAM model in these categories. In datasets like Baseball Bat, Bird, Breast, Chalk Group, Dolphin Above and Tie, the model, has performed moderately well exhibiting mean IoU scores between 0.6 and 0.8. On the other hand, the Clock and Skin object categories demonstrate more variability in performance. While the PerSAM model performs moderately well for the Clock category, there is significant variability in performance between students. Additionally, the Skin category shows relatively poorer performance compared to other categories, with one student achieving notably low IoU scores. These variations in performance could be attributed to factors such as the complexity of the objects, the diversity and quality of training data, model hyper parameters, and differences in student expertise and understanding of the segmentation task.

V. Conclusions

In our study, we utilized the Personalized Segment Anything Model (PerSAM), an existing framework, and adapted it to personalize segmentation for specific visual concepts using only one-shot data. By leveraging PerSAM's capabilities, we tailored it to our dataset, effectively segmenting it in one shot with the provided mask. This approach extends PerSAM's applicability to various scenarios without introducing new techniques or modifications.

Our methodology involved a thorough examination of PerSAM's architecture and functionality to ensure alignment with our dataset and segmentation objectives. Using one-shot data minimized the data acquisition burden and streamlined segmentation, making our approach practical and scalable for diverse scenarios and datasets.

In this study, we conducted an evaluation of PerSAM for semantic segmentation tasks across various object categories. The performance of the PerSAM model was assessed through the comparison of predicted segmentation masks with ground truth masks provided by the Segment Anything Model (SAM) prompts. Our analysis revealed valuable insights into the effectiveness of the PerSAM model in segmenting diverse objects across different datasets.

The experimental results demonstrated that the PerSAM model exhibits robust performance across a wide range of object categories. In particular, the model achieved high Intersection over Union (IoU) scores for object categories such as Baseball Bat, Bird, Breast, Bus, Cat, Chalk Group, Cow, Dog, Dolphin Above, Dolphin Below, Polyp, Salt Dome, Stop Sign, and Tie. These findings underscore the capability of the PerSAM model to accurately segment objects with varying shapes, textures, and complexities.

While the overall performance of the PerSAM model was commendable, there were instances of variability in performance observed across different object categories and between individual students. Object categories such as Clock and Skin demonstrated more variability in performance, suggesting potential areas for further optimization and improvement.

The variability in performance could be attributed to factors such as the complexity of the objects, the diversity and quality of training data, model hyperparameters, and differences in student expertise. Addressing these factors through targeted model refinement, dataset augmentation, and continuous training efforts could potentially enhance the performance of the PerSAM model and improve its applicability across a broader range of object categories and segmentation tasks.

VI. Project Contributions

In our collaborative effort, all team members contributed to the project's success by engaging in the necessary grunt work. We collectively explored segmentation models, studying, comparing, and evaluating different architectures while focusing on understanding their intricacies and performance benchmarks. Each team member played a crucial role in the meticulous tasks of manual mask separation, studying various architectures, delving into training methodologies, and critically analyzing model performances. Throughout the project, we actively participated in regular meetings and brainstorming sessions, fostering collaboration and idea exchange. After identifying a set of papers and articles relevant to our research, we convened regularly to discuss findings, share insights, and brainstorm potential approaches. Together, our teamwork and collective efforts led us to adopt PerSAM as the foundation of our approach, ensuring a successful project execution from research to implementation.

VII. References

[1] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[2] Kavur, A. Emre, Ludmila I. Kuncheva, and M. Alper Selver. "Basic ensembles of vanilla-style deep learning models improve liver segmentation from ct images." Convolutional neural networks for medical image processing applications. CRC Press, 2022. 52-74.

[3] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[4] Zhang, Renrui, et al. "Personalize segment anything model with one shot." arXiv preprint arXiv:2305.03048 (2023).

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