Optimising SAM For Medical Image Segmentaion

Tapera Chikumbu
tpchikumbu@gmail.com
University of Cape Town
Cape Town, South Africa

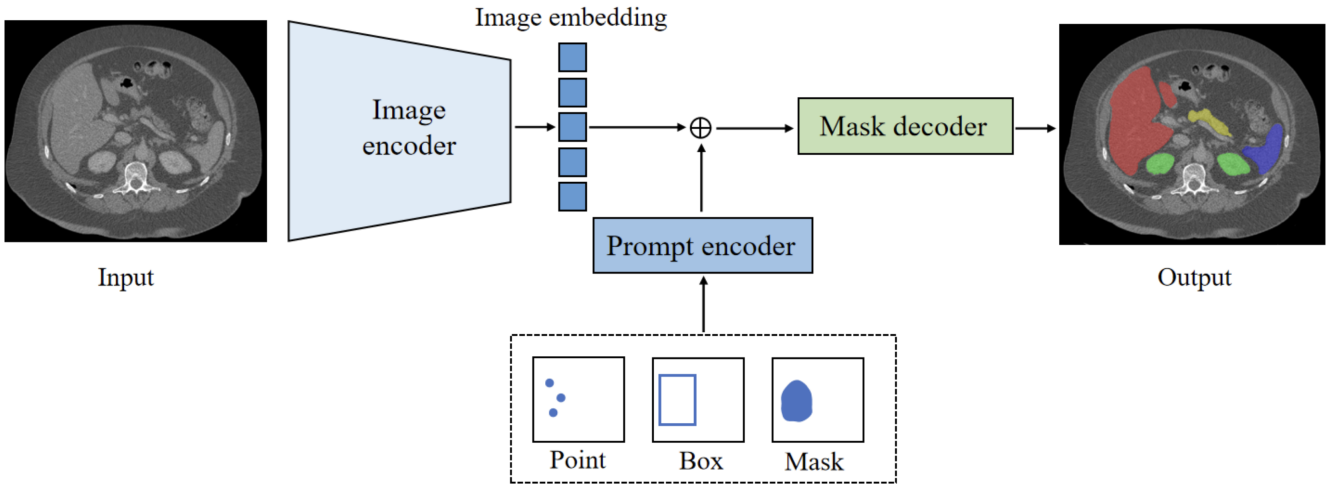

Figure 1: Architecture of SAM

ABSTRACT

Segment Anything Model (SAM) is a foundation model for natural image segmentation. Its novel architecture and large training dataset have helped it become a hot topic in computer vision. However, research shows the standard version of SAM is less accurate when applied to other image domains. Medical images for example tend to have lower colour contrast, more complex object shapes and less available training data. These and other factors limit the accuracy of SAM when applied to other models. This review explores attempts made to address the aforementioned challenges SAM faces with medical images. A brief look is also taken at how SAM overcomes similar challenges in other image domains. Parameter Efficient Fine-Tuning, prompt engineering and custom dataset curation were found to be the most successful improvement strategies across the domains considered.

KEYWORDS

Image Processing, Computer Vision, Machine learning, Segment Anything Model

Figure 2: Comparison of original image to semantic masks generated by SAM

1 INTRODUCTION

Image segmentation is the selection of semantically related pixels in an image. Similar to how a photographer can focus on different objects when taking a picture. An extension of this is semantic segmentation. Seen in Figure 2, the selected groups of pixels are labelled according to their properties [14]. In 2023, Meta released Segment Anything Model (SAM) [9]. It has since become a common benchmark for assessing natural image segmentation. SAM offers four different prompting methods. Its vast training dataset also allows it to have good zero-shot transfer when applied to unseen data.

However, SAM is not well suited to segmenting medical images. When tested against other segmentation models like U-NET and

UCTransNet [10, 20] SAM appears less impressive. Medical images have different object shapes, resolutions, dimensions and colour contrasts that what SAM is trained on. This limits the amount of correct inference that SAM can draw from medical images. Several attempts have been made to improve the performance of SAM on medical images. Approaches include fine-tuning model parameters and changing prompting methods. This paper reviews variations of SAM designed to improve its performance in under-performed scenes with a focus on medical image segmentation. Part of this includes identifying the strengths of different optimisations. With this knowledge, a plan is made to improve the performance of SAM on a subset of medical images.

2 UNDERSTANDING SAM

SAM is a promptable foundation model designed for segmenting natural images [9] . Its main objective is zero-shot learning. This entails transferring to new tasks with different distributions while maintaining performance. Key to this are its transformer architecture and large dataset. In the version first proposed by Vaswani et al. [18], transformers replace the recurrent layers of traditional neural networks with feed-forward self-attention layers. As a result, training times and layer complexity are reduced. Transformers also use global computations that offer more parallelism. Further work by Carion et al. [1] proved the benefits of using a transformer architecture for image processing tasks. SAM expands on this by using a pre-trained vision transformer (ViT). This consists of an image encoder, prompt encoder and mask decoder. Three different encoders come bundled with SAM. The largest, ViT-H, offers the most dense neural network with 636 million parameters. ViT-L and ViT-B have 308 million and 91 million parameters respectively. The larger encoders offer better accuracy but result in slower inference time.

Upon release, SAM was accompanied by its large, diverse training dataset, SA-1B. Included in the dataset are >1 billion segmentation masks from >11 million high-resolution images [9]. When generating masks for new images, SAM offers point and bounded box prompting. These allow for different levels of specificity and can be adjusted to match the segmentation task.

3 IMPROVING SAM

Contrary to its name, SAM does not have the same performance across all types of images. There are many domains where SAM is shown to be less efficient than specified segmentation models. Some examples include complex image shapes, shadow detection and medical segmentation. The term "under-performed scenes" was used by prior researchers [2, 15] to refer to these tasks that SAM struggles with. Various experiments have been performed on improving SAM's accuracy in such scenes. Of particular interest are improvements in the domain of medical images. However, to better understand all possible improvement methods, experiments in other domains were also studied. 1 gives a summary of those reviewed.

3.1 Medical

3.1.1 MedSAM. [10] Regarded as one of the most popular attempts at using SAM to the medical field. During development, it was intended to serve as a foundation model for medical image segmentation. As such, several specialised models have been developed from MedSAM and it is used as a benchmark for many more models [7, 8, 19]. It achieves this using a collection of over a million masked pairs from different anatomical structures and modalities. Having such large amounts of input data improves the zero shot transfer while sacrificing the amount of time required to train it. Unlike SAM with three different prompting options, MedSAM is limited to only bounding box prompts. These allow the segmentation of 3D images as a collection of 2D slices.

3.1.2 SegmentAnyBone. [6] specialises in skeletal segmentation of Magnetic Resonance Imaging (MRI). Its 2D branch is build upon the architecture proposed by Wu et al. [19] where Adapter blocks are added to the the image encoder and mask decoder. These additions allow efficient parameter tuning while keeping over 90% of the models parameters fixed. The model also introduces hybrid prompting. Experiments tested a dynamic prompt generation algorithm. This changes the model task from predicting corresponding mask regions to generating all ground truth masks for an image. MRI images offers a depth dimension when recording data. By viewing 3D images as collections of 2D slices, SegmentAnyBone is able to make predictions on volumetric data [6]. An additional Attention branch is added to improve performance on 3D images. This allows information from neighbouring slices to be considered during segmentation.

3.1.3 ClickSAM. [7] Another domain specific medical image model. It improves the performance of SAM on ultrasound breast images. Opts for point prompting to increase the accuracy of generated masks. User supplied clicks are passed to the mask decoder and used to generate further clicks according to the Centroidal Voronoi Tessellation algorithm [4]. These improve the accuracy of the final mask generated by providing feedback on the accuracy of each point within the mask.

3.1.4 SlideSAM. [12] Intended for the segmentation of both 2D and 3D medical images. This model makes semantic inferences using adjacent slices of a volume. These additional slices add more detail in multiple dimensions for a better segmentation result. A sliding window captures three adjacent slices at a go. They are then processed by the same encoding architecture used by SAM [9]. A hybrid approach to prompt generation is also explored in SlideSAM. Point prompts and bounding boxes are sampled with equal probability from the ground truth masks. Artificial noise is then added to the prompts for a more accurate representation of human input.

3.1.5 Med-SA. [19] Designed to improve on MedSAM [10], this model introduces a Space-Depth Transpose (SD-Trans) technique. With this, the input spatial dimension is transposed to the depth dimension. Doing so allows the same self-attention blocks to process different dimensional information given different input. In addition, its Hyper-Prompting Adapter (HyP-Adpt) facillitates deep

Variation	Category	Methods
MedSAM	Medical foundation model	Large training dataset Custom prompting
SegmentAnyBOne	MRI	Adapter blocks, 3D Attention branch, Hybrid prompting
ClickSAM	Ultrasound imaging	Custom prompting
SlideSAM	Medical	Multi-slide inference, Custom prompting
Med-SA	Medical foundation model	Adapter blocks
SimAda	Adaption framework	Adapter blocks

Table 1: Summary of reviewed experiments

prompt adaptations. Inspired by hyper networks, prompt information is concatenated and reduced as prompt embedding. This allows hyper-prompting on the less parameterised adapter level. This way, the model is can easily accommodate different modalities and downstream tasks.

3.1.6 SimAda. [15] Offers a more general approach to model adaptation. SimAda is proposed as a unified framework for adapting any kind of transformer-based model. Allows easy fine-tuning of model parameters through the introduction of trainable adapters. This lets the model representation better match a dataset's feature distribution while maintaining simplicity. Also introduces the possibility of parallel and mixed adapter designs. At their core, all transformer networks contain multi-head self-attention layers and fully connected feed-forward networks (FFN) [18]. By adding adapter blocks in different parts of the network, four different variations of SimAda are developed for testing on underperformed scenes.

4 DISCUSSION

Several metrics are used for assessing segmentation accuracy. The most common ones seen were variations of the Dice loss and Intersection Over Union (IOU). The Sørensen-Dice coefficient [3] measures the similarity of two samples. This is done by comparing the intersection of said samples to the sum of the individual samples. The samples considered in image segmentation are the ground truth masks and the predicted masks for the model. Dice loss is the corresponding difference function, given by 1 - Dice coefficient.

$$\begin{aligned} Dice &= \frac{2|A \cap B|}{|A| + |B|} \\ L_{\text{Dice}} &= 1 - \frac{2|A \cap B|}{|A| + |B|} \end{aligned} \tag{1}$$

IOU is calculated with the Jaccard index [13]. It as another measure that compares the intersection of samples to their union. When applied to binary classification tasks, it is written in terms of the number of true positives (TP), false negatives (FN) and false positives (FP).

$$\text{IOU} = \frac{TP}{TP + FP + FN} \tag{2}$$

All variations of SAM need accurate data to train on. The amount of annotated medical images available is thoroughly dwarved by the SA-1B dataset of natural images. ColonDB [15] has a total of 380 images. SegmentAnyBone [6] only had 271 images in its dataset. WORD [12] has only a 150 chest volumes. Even if combined, all the medical image datasets considered in the models mentioned so far are not a fraction the size of SA-1B with over 10 million images.

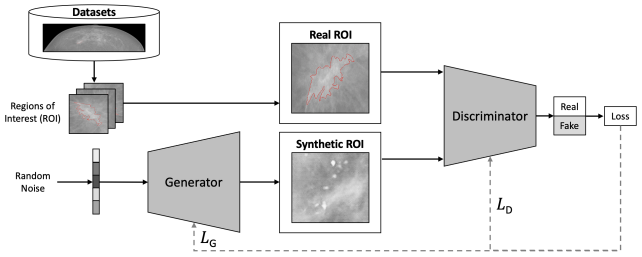

Figure 3: GAN showing creation of additional synthetic data

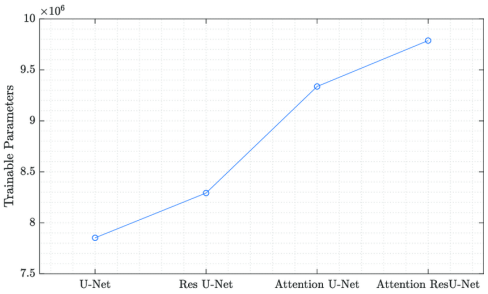

Figure 4: Number of trainable parameters in variations of U-Net

With such a large size difference, it would be difficult for SAM to get the same performance on medical images that it has with natural images. The use of synthetically generated images has the potential to solve this issue [11]. A Generative Adversarial Network (GAN) can be applied to generate new images from existing ones. A general GAN architecture is shown in Figure 3. This generates more data for model training and evaluation. As a result, models become less prone to over-fitting their training data.

The encoders available in SAM all have large parameter counts. ViT-B, which is the smallest available encoder, has nearly 100 million trainable parameters. For comparison, other segmentation models like U-Net and Attention U-Net have under 10 million trainable parameters [17]. Figure4 from that article illustrates this point .

Fine-tuning so many parameters during model training is too resource intensive to be feasible [5]. Most models studied instead perform Parameter Efficient Fine-Tuning (PEFT) [6, 12, 15, 19]. Adapter blocks are added to different parts of the encoder pipeline. These few blocks are fine-tuned while the remaining parameters in the

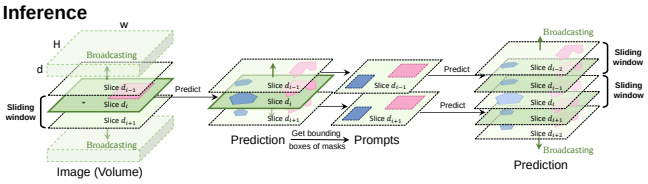

Figure 5: Sliding window technique for 3D segmentation

model are held constant. In this way, less resources are required to adapt the model to a new task.

Using adapters with a GELU activation function for polyp segmentation [2], the mean dice coefficient of SAM went from 77.8% in SAM to 85.0%. In the same experiment, mean IoU increased from 0.707 to 0.776. Different activation functions were used with various success. SlideSAM [12] and SimAda [15] for example, introduce a low rank adapter (LoRA) layers to their architecture. Hyp-Adpt [19] contains a ReLU activation that allowed Med-SA to outperform all other models it was compared to. Med-SA BBox 0.5 was the lowest overall variation of the model and still achieved a mean Dice of 87.6% on the BTCV dataset. The finer-tuned MedSAM BBox 0.5 had a significantly lower mean Dice score of 69.2%.

Similar comparisons can be drawn when assessing the performance of SimAda [15] on the ISIC and ColonDB medical datasets. Instead of Dice or IoU metrics, experiment performance was measured using the mean absolute error (MAE). All variations of SimAda recorded lower errors that SAM on the datasets when both used the ViT-B architecture. The smallest improvements across all variations were 33.2% on ISIC and 5.5% on ColonDB. In both instances, the worst performing variation was the one using a LoRA structure.

SAM was designed for segmenting 2D images. It is unable to directly work with higher dimensional images. However, experiments show SAM can be adapted to segment 3D images [6, 12, 19]. Models typically do this by slicing the 3D structure and applying the segmentation model to a smaller volume of the image.

This can be performed slice by slice with significant improvements over other neural networks. The mean dice coefficient of SegmentAnyBone was 86.87% [6]. This is a vast improvement over the closest 3D model, nnUnet (3D) which only scored a dice of 73.32%. Ontop of image slicing, SegmentAnyBone incorporates a 3D attention branch that resizes and downsamples the selected volume to a lower resolution. When added to very deep neural networks, such branches help the model decide which branch to focus on for making predictions [16]. Doing so allows branches in the multi-head attention blocks of the encoders and decoders to have better performance on 3D images.

However, some contextual information can be lost when considering each slice separately. This leads to less accurate segments with high levels of noise between slices. A viable solution to this is a sliding window, shown in Figure 5. Adjacent slices are added to the embedded input when performing segmentation [12]. By considering three slices at a time, this method returns dice scores above 80% after a single prompt in both the WORD and CHAOS datasets. For comparison, SAM only achieved a Dice of 74.98% on the WORD dataset after 40 prompts.

The different prompting methods offered by SAM help it serve as a foundation model. Another optimisation seen in specialised models is choosing a prompting method that better suits the domain of interest. ClickSAM [7] creates additional prompts using a system of positive and negative clicks. These help it blow passed SAM in performance when applied to the Dataset of Breast Ultrasound Images (BUSI). The 0.8074 IoU MedSAM records on a malignant tumor is irrelevant when compared to ClickSAM's 0.9439 on the same image. In addition, this prompting system results in faster segmentation over MedSAM.

SlideSAM [12] uses a hybrid prompting approach. It combines the generality of bounded boxes with the precision of point prompting. This combination of prompts compliments the sliding window approach, helping it reach the levels of performance mentioned earlier. Figure 6 further illustrates the level of improvement SAM achieves over other models.

5 CONCLUSIONS

Though SAM has underwhelming performance when used for medical image segmentation, its adaptability allows it to be easily optimised for the task. A key factor to this is the availability of good training data. For accurate semantic segmentation of images, SAM must be trained on high quality, well labelled data. There is less data available for medical images than natural images. The use of synthetic images is a possible avenue for improving SAMs performance that can be used during experiments

The large parameter sizes of the available encoders are not a challenge when Adapter blocks are introduced to the transformer architecture. With the majority of other parameters kept constant, a small fraction can be freely tuned and adjusted to better help the model understand the prevalent features of the dataset.

Despite all seeming to boost performance, there is no universal adapter structure that always offers the best improvement. This motivates the use of adaptation frameworks like SimAda [15] that allow different adapter structures to be easily tested. These abstract the design process so more time can be spent training and evaluating models. It's therefore expected that an adaptation framework be used to fine tune SAM instead of manually adding the adapter blocks.

Performance increases are also possible through prompt engineering. Just as there is no adapter structure that always outshines the rest, there is also no prompting method that guarantees the best segmentation results. Mask and point prompts offer greater precision when dealing with complex shapes like legions and tumors. However the generality of bounded box prompts makes grouping related pixels easier. Combinations of these and other prompting methods can all lead to the creation of better segmentation masks. Several prompting options are to be studied and tested when deciding the final ones to be used for the chosen image domain.

3D images can be transformed into 2D vectors that SAM is able to process. This process is faster than comparable CNN methods for segmenting image volumes. It also offers greater accuracy making it a state of the art technique. The sliding window technique seen in SlideSAM [12] is of key interest. Considering larger numbers of adjacent small volumes at once for better segmentation accuracy is a method that can be further tested in experiments with 3D images.

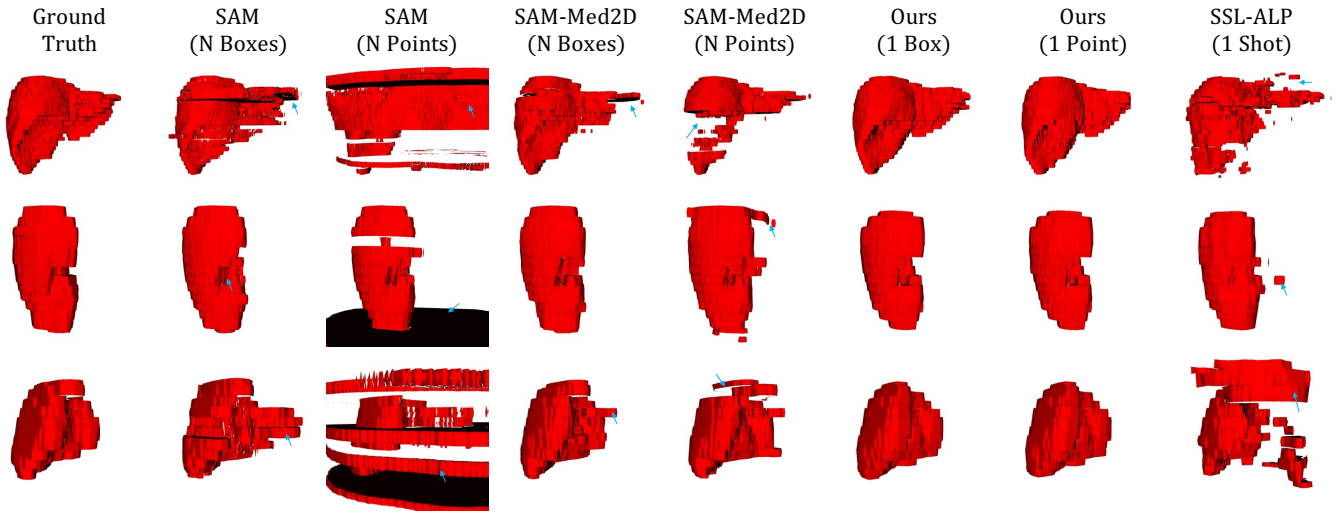

Figure 6: Visual comparison of SlideSAM against other models on the CHAOS dataset

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872 [cs.CV]

[2] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. 2023. SAM Fails to Segment Anything? – SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. arXiv:2304.09148 [cs.CV]

[3] Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 3 (1945), 297–302. http://www.jstor.org/stable/1932409

[4] Qiang Du, Vance Faber, and Max Gunzburger. 1999. Centroidal Voronoi Tessellations: Applications and Algorithms. *SIAM Rev.* 41, 4 (1999), 637–676. https://doi.org/10.1137/S0036144599352836 arXiv:https://doi.org/10.1137/S0036144599352836

[5] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the Effectiveness of Parameter-Efficient Fine-Tuning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 11 (Jun. 2023), 12799–12807. https://doi.org/10.1609/aaai.v37i11.26505

[6] Hanxue Gu, Roy Colglazier, Haoyu Dong, Jikai Zhang, Yaqian Chen, Zafer Yildiz, Yuwen Chen, Lin Li, Jichen Yang, Jay Willhite, Alex M. Meyer, Brian Guo, Yashvi Atul Shah, Emily Luo, Shipra Rajput, Sally Kuehn, Clark Bulleit, Kevin A. Wu, Jisoo Lee, Brandon Ramirez, Darui Lu, Jay M. Levin, and Maciej A. Mazurowski. 2024. SegmentAnyBone: A Universal Model that Segments Any Bone at Any Location on MRI. arXiv:2401.12974 [eess.IV]

[7] Aimee Guo, Grace Fei, Hemanth Pasupuleti, and Jing Wang. 2024. ClickSAM: Fine-tuning Segment Anything Model using click prompts for ultrasound image segmentation. arXiv:2402.05902 [cs.CV]

[8] Bozhen Hu, Bin Gao, Cheng Tan, Tongle Wu, and Stan Z. Li. 2023. Segment Anything in Defect Detection. arXiv:2311.10245 [cs.CV]

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]

[10] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2023. Segment Anything in Medical Images. arXiv:2304.12306 [eess.IV]

[11] Richard Osuala, Grzegorz Skorupko, Noussair Lazrak, Lidia Garrucho, Eloy García, Smriti Joshi, Socayna Jouide, Michael Rutherford, Fred Prior, Kaisar Kushibar, Oliver Díaz, and Karim Lekadir. 2023. medigan: a Python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging* 10, 06 (Feb. 2023). https://doi.org/10.1117/1.jmi.10.6.061403

[12] Quan Quan, Fenghe Tang, Zikang Xu, Heqin Zhu, and S. Kevin Zhou. 2023. Slide-SAM: Medical SAM Meets Sliding Window. arXiv:2311.10121 [cs.CV]

[13] Raimundo Real and Juan M. Vargas. 1996. The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology* 45, 3 (1996), 380–385. http://www.jstor.org/stable/2413572

[14] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. 2006. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

[15] Yiran Song, Qianyu Zhou, Xuequan Lu, Zhiwen Shao, and Lizhuang Ma. 2024. SimAda: A Simple Unified Framework for Adapting Segment Anything Model in Underperformed Scenes. arXiv:2401.17803 [cs.CV]

[16] Marijn Stollenga, Jonathan Masci, Faustino Gomez, and Juergen Schmidhuber. 2014. Deep Networks with Internal Selective Attention through Feedback Connections. arXiv:1407.3068 [cs.CV]

[17] Anastasios Temenos, Nikolaos Temenos, Anastasios Doulamis, and Nikolaos Doulamis. 2022. On the Exploration of Automatic Building Extraction from RGB Satellite Images Using Deep Learning Architectures Based on U-Net. *Technologies* 10 (01 2022), 19. https://doi.org/10.3390/technologies10010019

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. 2023. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv:2304.12620 [cs.CV]

[20] Yichi Zhang and Rushi Jiao. 2023. Towards Segment Anything Model (SAM) for Medical Image Segmentation: A Survey. arXiv:2305.03678 [eess.IV]