Journal of Medical Imaging & Case Reports

https://doi.org/10.17756/micr.2025-116

Case Series

Open Access

Accurate Segmentation of Pigmented Skin Lesions Using Grounded-segment-anything

Takashi Nagaoka*

Department of Computational Systems Biology, Faculty of Biology-Oriented Science and Technology, Kindai University, Japan

***Correspondence to:**
Takashi Nagaoka
Department of Computational Systems Biology,
Faculty of Biology-Oriented Science and Technology,
Kindai University,
Japan.
E-mail: nagaoka@waka.kindai.ac.jp

Received: February 27, 2025
Accepted: March 25, 2025
Published: March 28, 2025

Citation: Nagaoka T. 2025. Accurate Segmentation of Pigmented Skin Lesions Using Grounded-segment-anything. *J Med Imaging Case Rep* 9(1): 20-25.

Published by United Scientific Group

Abstract

Melanoma is a highly aggressive form of skin cancer, and accurately segmenting pigmented lesions is crucial for early diagnosis. However, manual annotation is time-consuming, subjective, and requires expertise from dermatologists. To address these challenges, we propose a fully unsupervised segmentation approach using grounded-segment-anything model (Grounded-SAM), which operates without the need for annotated training data. Grounded-SAM combines Grounding-DINO for object detection with the SAM for segmentation. We evaluated the method using 2,594 dermoscopic images from the 2018 international skin imaging collaboration (ISIC) dataset and compared its performance with Otsu's thresholding method, using the Jaccard index as the primary evaluation metric. The proposed method achieved an average Jaccard coefficient of 0.509, which significantly outperformed Otsu's method (0.386, $p < 0.05$). Moreover, for images in which Grounding-DINO successfully detected lesion bounding boxes, segmentation accuracy improved further, yielding a coefficient of 0.588. Despite these improvements, the method exhibited limitations in cases with low-contrast lesions, small lesion areas, and significant vignetting. Our findings demonstrate that Grounded-SAM enables the effective segmentation of pigmented skin lesions without requiring annotated datasets, providing a promising solution for the reduction of annotation costs in AI-assisted melanoma detection. Future research should focus on prompt selection and integrating hybrid annotation strategies to enhance clinical applicability.

Keywords

Skin lesion segmentation, Deep learning, Grounded-segment-anything, Unsupervised segmentation, Zero-shot learning, Melanoma detection, Dermoscopy

Introduction

Melanoma is a highly aggressive form of skin cancer that poses significant diagnostic challenges, particularly in its early stages. Early detection is crucial, as melanoma is readily treatable if promptly diagnosed and treated. However, distinguishing early-stage melanoma from benign lesions is challenging due to their visual similarities. Traditionally, melanoma diagnosis has relied on the expertise of dermatologists, making it a subjective process. This reliance on visual inspection generally results in inconsistencies in diagnostic accuracy, with experienced dermatologists achieving higher success rates than less experienced practitioners [1].

Dermoscopy has emerged as a valuable tool for improving the accuracy of melanoma diagnosis. This non-invasive method magnifies pigmented skin lesions, enabling the visualization of microscopic structures that are otherwise invisible to

the naked eye. The detailed images produced by dermoscopy provide critical information to distinguish benign from malignant lesions. However, even with dermoscopy, the subjective nature of visual assessment remains a significant challenge. Artificial intelligence (AI) assisted evaluation systems have been introduced to address this issue through objective and reproducible assessments of dermoscopic images.

Recent advancements in AI and deep learning have significantly improved the accuracy of melanoma diagnosis. AI models can analyze vast amounts of dermoscopic data, identify patterns, and make predictions with accuracy comparable to or exceeding that of dermatologists. Esteva et al. demonstrated that deep convolutional neural networks (CNNs) achieved a dermatologist-level classification accuracy for skin cancer, with an area under the curve score of 0.91 [2]. Additionally, large-scale dermoscopic datasets, such as the ISIC and HAM10000 [3], have facilitated the development of AI-driven diagnostic tools. However, these methods typically require extensive annotated training data, which poses a major challenge in real-world applications.

The effectiveness of AI-based melanoma detection is heavily reliant on the quality and quantity of training data. High-quality lesion segmentation has been shown to improve melanoma classification accuracy. Maron et al. [4] reported that models trained on precisely segmented lesion boundaries achieved up to 8.9% higher classification performance than those trained on raw images [4]. Similarly, Dzieniszewska et al. [5] demonstrated that enhancing lesion segmentation quality through self-training improved skin cancer detection accuracy, achieving a classification accuracy of up to 88.0% [5]. However, obtaining large-scale annotated segmentation data is costly and time-consuming, as expert annotation often takes several minutes per image [6]. These limitations highlight the need for segmentation models that can operate without relying on extensive labeled training data.

To address this issue, this study explores the use of Grounded-SAM for lesion segmentation [7]. Grounded-SAM integrates two powerful AI models: Grounding DINO for object detection [8] and SAM for segmentation [9]. The key advantage of this approach is that it eliminates the need for additional labeled training data, making it highly scalable for clinical applications. Grounding-DINO utilizes text prompts to identify and localize objects, minimizing the dependence on extensive manual annotation. Kirillov et al. [9] demonstrated that the SAM achieves state-of-the-art segmentation across various domains, including medical imaging, without requiring task-specific training [9]. Ma et al. further validated its applicability in medical imaging, showing its effectiveness in segmenting anatomical structures in chest X-rays and MRI scans [10].

The success of text-prompt-driven segmentation depends on the careful selection of prompts. Previous research has shown that optimized prompt selection enhances object detection accuracy [11]. In this study, we empirically evaluate the effectiveness of various text prompts and select "pigmented lesions or vascular lesions" as the most suitable descriptor for lesion localization.

We apply Grounded-SAM to the segmentation of pigmented skin lesions and evaluate its performance using the ISIC 2018 dataset [12]. Specifically, we compare its segmentation accuracy against Otsu's thresholding method [13], using the Jaccard index as the primary metric for evaluation. By demonstrating the feasibility of a model that does not require manually annotated training data, this research aims to contribute to the automation of lesion segmentation and the reduction of annotation costs in AI-assisted melanoma detection.

Method

This study utilized 2,594 dermoscopic images from task 1 of the ISIC 2018 challenge to evaluate lesion segmentation accuracy. Each image was paired with lesion masks annotated by dermatologists, which served as the ground truth for performance assessment. This study aimed to improve segmentation accuracy through a combination of advanced preprocessing techniques and AI-based segmentation models.

To ensure high-quality input images for segmentation, several preprocessing steps were applied. Hair artifacts, which can obscure lesion structures and adversely affect segmentation accuracy, were removed using the black hat filtering technique, a morphological operation known for its effectiveness in hair removal from dermoscopic images [14]. This step enhances lesion visibility and improves segmentation accuracy.

The original RGB images were converted to grayscale, and a morphological operation using a 17 × 17 pixel structuring element was applied to isolate dark hair regions. These regions were then inpainted using morphological reconstruction, removing hair effectively while preserving the underlying lesion details. Additionally, image boundaries were cropped by five pixels from each side to remove peripheral artifacts that could interfere with segmentation.

To maintain consistent cross-model input dimensions, all images were resized to 224×224 pixels, ensuring compatibility with deep learning architectures. Furthermore, contrast enhancement was performed using contrast-limited adaptive histogram equalization (CLAHE) [15], which was applied to the L-channel in the lab color space to enhance lesion visibility while preventing excessive enhancement in uniformly pigmented areas.

To perform segmentation, a two-step AI-based pipeline was adopted, leveraging the strengths of both the object detection and semantic segmentation models. The Grounding-DINO model, configured with the Swin transformer architecture [16], was used as the first stage. Grounding-DINO is a zero-shot object detection model that localizes objects within an image based on text prompts rather than predefined training datasets. This approach reduces the need for extensive annotated datasets, a major limitation in medical AI applications, and allows for flexible adaptation to new lesion types. A series of candidate text prompts were tested, and "pigmented lesions or vascular lesions" was selected as the most effective descriptor for lesion localization based on empirical evaluations.

Once the lesions were localized, the SAM was employed

for precise semantic segmentation. Specifically, the Vision Transformer [17] encoder variant was used, as it is optimized for handling high-resolution images with fine-grained details, making it particularly well-suited for medical imaging applications. The SAM produces multiple segmentation masks for each detected region, from which the mask with the highest confidence score was selected. The final segmentation masks were binarized, assigning a pixel value of 255 (white) to the segmented lesion regions and 0 (black) to the background.

To provide a baseline comparison, Otsu's binarization method was applied to the preprocessed images. This global thresholding technique was chosen because it represents a widely used traditional segmentation method in dermatology and provides insight into the added value of AI-based approaches. Fan et al. [18] analyzed the application of Otsu's method for lesion segmentation in dermoscopy images and reported that while it performs well in certain cases, it struggles with lesions that have indistinct boundaries and heterogeneous textures [18]. This highlights the necessity of AI-based segmentation methods that can adapt to complex lesion structures.

Segmentation accuracy was evaluated using the Jaccard index, which measures the degree of overlap between predicted and ground truth masks. The statistical significance of the segmentation performance differences was assessed using the Wilcoxon signed-rank test, with a significance threshold of $p < 0.05$. The entire pipeline was implemented using Python, with OpenCV for image processing, PyTorch for deep learning model inference, and Supervision for annotation handling. To optimize computational efficiency, all model inferences were conducted in a GPU-enabled environment.

Result

Figure 1 provides examples of the semantic segmentation results obtained in this study. Panel A shows the original dermoscopic image, while panel B presents the mask image provided by a dermatologist. Panel C depicts the preprocessed dermoscopic image, where CLAHE was applied to improve contrast, and the black hat filter was used to remove hair artifacts. The instance segmentation result generated by Grounding-DINO is shown in panel D, with a purple bounding box indicating the area identified as "pigmented lesions or vascular lesions." Finally, panel E displays the semantic segmentation output produced by the SAM, which closely matches the dermatologist-provided mask, demonstrating the ability of the model to accurately segment lesions using zero-shot learning.

The effectiveness of the Grounded-SAM approach was evaluated using the Jaccard index, which quantifies the overlap between the predicted segmentation masks and the ground truth annotations. A higher Jaccard index indicates greater segmentation accuracy. Figure 2A shows the distribution of Jaccard coefficients for all the images in the dataset. The proposed method achieved an average Jaccard coefficient of 0.509 ± 0.305 (median: 0.628), which was significantly higher than the 0.386 ± 0.216 (median: 0.379) obtained using Otsu's binarization method. A Wilcoxon signed-rank test confirmed that this difference was statistically significant ($p < 0.05$).

Among the 2,594 images, 2,244 cases were successfully detected by Grounding-DINO, meaning that a bounding box was generated around the lesion. For this subset, the average Jaccard coefficient improved, yielding 0.588 ± 0.247 (median: 0.662), as shown in figure 2B. The increase in segmentation accuracy for this subset suggests that accurate object detection prior to segmentation plays a crucial role in achieving high segmentation performance. The Wilcoxon signed-rank test also confirmed that the improvement in the Jaccard index for this subset was statistically significant compared to Otsu's method ($p < 0.05$).

Despite the overall improvement, 551 images exhibited a Jaccard coefficient of less than 0.1, indicating segmentation failures. These cases were predominantly characterized by light pigmentation, small lesions, or significant vignetting, which

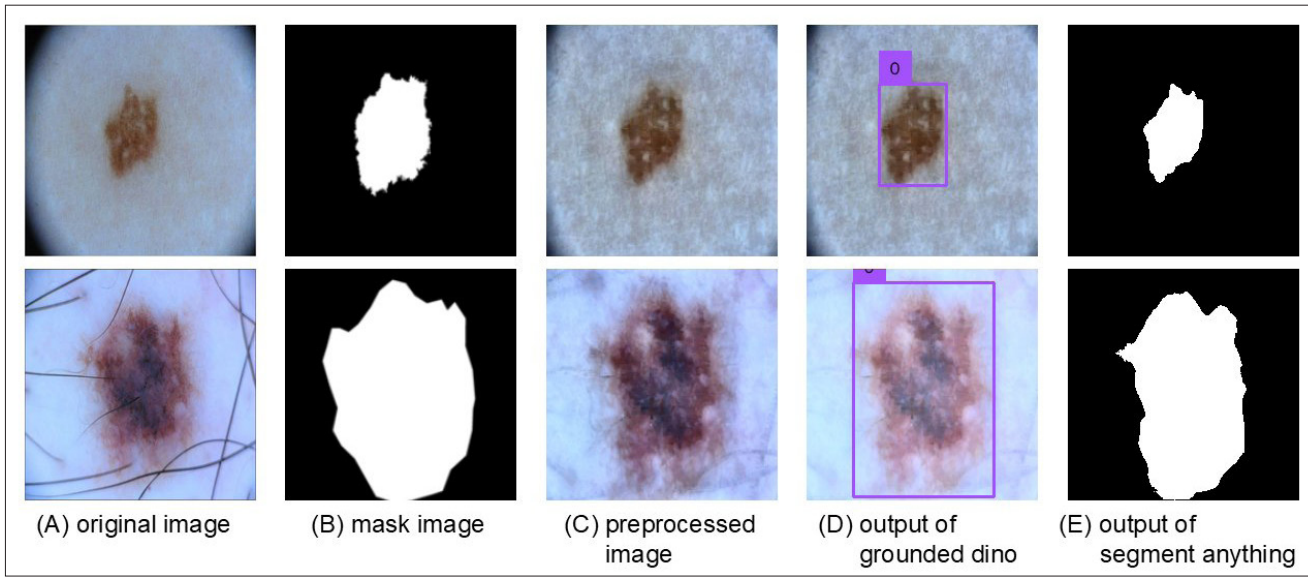

Figure 1: Examples of semantic segmentation in this study. (**A**) Original dermoscopy image, (**B**) Masked image provided by a dermatologist, (**C**) Pre-processed dermoscopy image, (**D**) Output of instance segmentation by Grounding-DINO and (**E**) Output of semantic segmentation by segment anything.

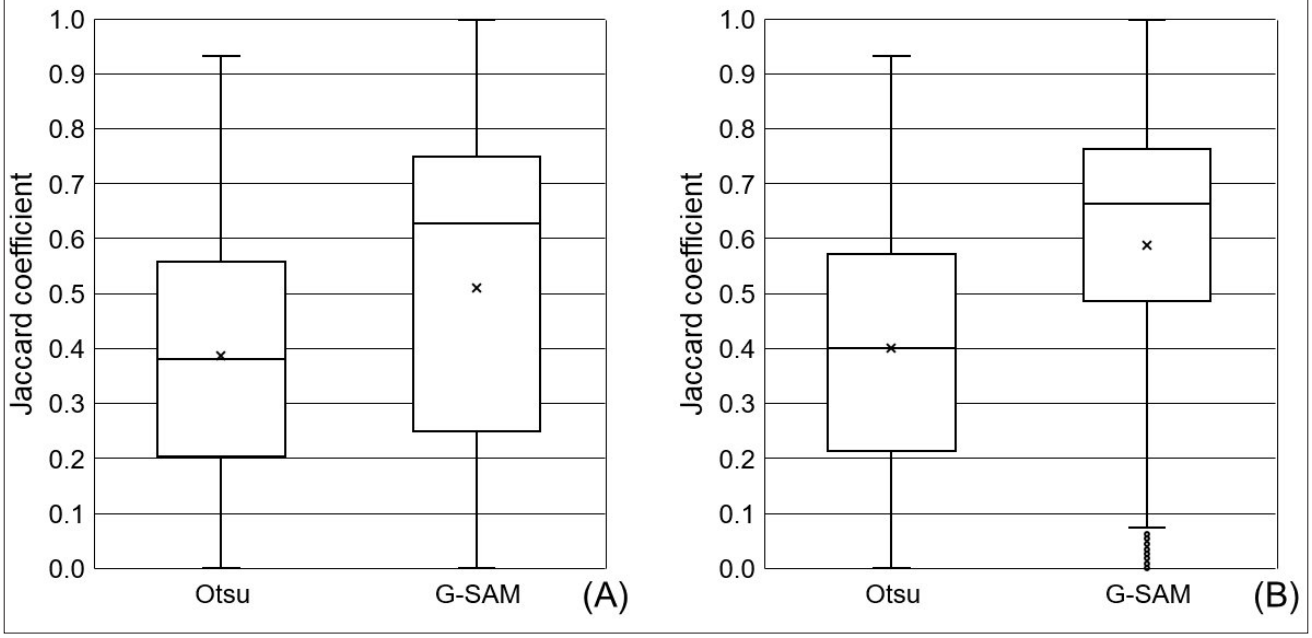

Figure 2: Comparison of Jaccard coefficients for lesion segmentation using Otsu's binarization and Grounded-SAM. (**A**) Box plot showing the Jaccard coefficient distribution across all images, and (**B**) Box plot for a subset of images where Grounding-DINO successfully detected lesion bounding boxes. In both cases, Grounded-SAM significantly outperformed Otsu's binarization, as confirmed by the Wilcoxon signed-rank test ($p < 0.05$).

likely contributed to the failure of the model in detecting lesion boundaries. Further analysis is needed to improve performance in these challenging cases, potentially by refining the text prompt used in Grounding-DINO or incorporating additional preprocessing techniques.

The improved performance of the Grounded-SAM technique, compared to Otsu's binarization, highlights its robustness in handling complex lesion structures. The ability to generate accurate segmentation without requiring extensive manual annotation makes it a promising approach for automated melanoma diagnosis.

Discussion

The results of this study demonstrated that Grounded-SAM achieved a higher average Jaccard coefficient than Otsu's binarization, indicating its superior performance in pigmented skin lesion segmentation. However, rather than presenting a dramatic improvement over existing methods, the key finding here is that Grounded-SAM achieved this performance without the need for manually annotated training data. While traditional supervised segmentation models depend on large-scale datasets with expert annotations, this method leverages pre-trained models to operate in a zero-shot manner. The fact that Grounded-SAM outperformed Otsu's binarization, another unsupervised segmentation method, suggests that large-scale vision models can provide a viable alternative when labeled datasets are scarce or costly to obtain. At the same time, our results highlight the remaining challenges of this approach, particularly in cases with light pigmentation, small lesions, or substantial vignetting effects, warranting further refinements for clinical applicability.

One of the major advantages of Grounded-SAM is its ability to process images without the need for manually annotated datasets. Training deep learning-based segmentation models typically requires high-quality annotations, which are often difficult and expensive to obtain. Previous studies have shown that segmentation accuracy strongly depends on the quantity and quality of labeled data. For example, Maron et al. [4] demonstrated that precise lesion segmentation improved melanoma classification accuracy by up to 8.9%, increasing balanced accuracy from 66.7% - 75.6% [4]. While supervised deep learning models have achieved high accuracy in melanoma segmentation, they rely heavily on extensive ground truth annotations. In contrast, Grounded-SAM overcomes this limitation by leveraging large-scale pre-trained vision models, offering a promising alternative for medical applications.

Despite these advantages, Grounded-SAM presented limitations in this study. In particular, segmentation failures occurred more frequently in images with low-contrast lesions, small lesion areas, or pronounced vignetting effects. These findings suggest that while the model can generalize across a wide range of lesion types, it struggles with lesion identification in cases characterized by low contrast with the background. Similar issues have been noted in previous studies, emphasizing the importance of contrast enhancement and adaptive preprocessing techniques in improving dermoscopic image segmentation [19]. Our study incorporated CLAHE and black hat filtering to enhance lesion visibility, but further improvements will be needed to accommodate more subtle lesion features.

Another key finding was that the choice of text prompts in Grounding-DINO significantly influenced segmentation accuracy. In this study, the prompt "pigmented lesions or vascular lesions" allowed the model to accurately detect target areas while minimizing false positives. Previous research has

suggested that optimized prompt selection can improve object detection performance [11]. However, our study did not conduct an extensive exploration of alternative prompt configurations, which could have yielded better results. Future research should systematically evaluate various prompts to further refine the performance of Grounded-SAM.

Our findings suggest that while Grounded-SAM serves as a viable unsupervised segmentation method for skin lesions, it is important to recognize that improved segmentation alone does not automatically translate into improved melanoma diagnosis. Melanoma diagnosis involves assessing multiple dermoscopic criteria, including asymmetry, border irregularity, color variation, and specific structural patterns [20]. Future work should investigate the integration of Grounded-SAM with feature extraction and classification models to evaluate its utility in melanoma evaluation.

Additionally, while this study utilized a fully unsupervised approach, semi-supervised methods incorporating minimal annotations could also be explored. For instance, object detection models such as You Only Look Once have demonstrated high accuracy in detecting lesions with minimal labeled data [21]. Combining such weakly supervised detection methods with Grounded-SAM could potentially enhance segmentation performance while minimizing the need for extensive manual annotations.

In conclusion, Grounded-SAM presents a fully unsupervised approach to skin lesion segmentation. While it does not achieve perfect segmentation, its adaptability and ease of use highlight its potential value. Further improvements in preprocessing techniques, prompt optimization, and model refinement will be essential for broader clinical application.

Conclusion

This study investigated the feasibility of using the Grounded-SAM framework for the segmentation of pigmented skin lesions in dermoscopic images without the need for manually annotated training data. By combining the Grounding-DINO model for object detection with the SAM for segmentation, we achieved an average Jaccard coefficient of 0.509 across the dataset. This performance was superior to Otsu's binarization, a traditional unsupervised segmentation method, which obtained a Jaccard coefficient of 0.386. When lesion detection by Grounding-DINO was successful, segmentation accuracy improved further, yielding an average Jaccard coefficient of 0.588.

These results demonstrate that, despite requiring no task-specific training, the Grounded-SAM approach can achieve segmentation performance comparable to that of some supervised methods. This is particularly notable given that deep learning-based segmentation typically depends on large-scale annotated datasets, which are costly and time-consuming to produce. The ability to perform lesion segmentation without such datasets suggests the potential for reducing annotation burdens while maintaining a reasonable level of accuracy.

However, this study also highlighted some challenges. Segmentation accuracy varies across lesion types, with failures occurring more frequently in cases with light pigmentation, small lesions, or strong vignetting effects. Additionally, we used a single text prompt ("pigmented lesions or vascular lesions") to guide Grounding-DINO Since prompt selection is a critical factor influencing segmentation outcomes, future research should systematically explore a range of prompts to determine their impact on performance.

Furthermore, while our approach was entirely unsupervised, there is potential to incorporate minimal human supervision, such as bounding box annotations, to refine lesion localization. Future studies could investigate hybrid approaches that integrate weak supervision methods, such as the You Only Look Once-based bounding box detection, to enhance the robustness of the segmentation process.

In summary, this study demonstrates that the Grounded-SAM framework provides a promising, fully unsupervised approach to skin lesion segmentation. While its segmentation accuracy does not yet match that of fully supervised deep learning models, its ability to operate without annotated training data makes it a valuable alternative for applications where large-scale annotation is infeasible. Further optimization, particularly in text prompt selection and potential hybrid annotation strategies, could improve its utility for clinical and research applications in dermatology.

Acknowledgements

This work was funded by the Japan Society for the Promotion of Science (JSPS) under grant numbers 20K12698 and 24K15777.

Conflict of Interest

None.

Ethical Approval Statement

This study utilized publicly available ISIC 2018 data and did not involve human or animal subjects, which would have required ethical approval.

References

1. Ko CJ, Braverman I, Sidlow R, Lowenstein EJ. 2019. Visual perception, cognition, and error in dermatologic diagnosis: key cognitive principles. *J Am Acad Dermatol* 81(6): 1227-1234. https://doi.org/10.1016/j.jaad.2018.10.082
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639): 115-118. https://doi.org/10.1038/nature21056
3. Tschandl P, Rosendahl C, Kittler H. 2018. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 5: 180161. https://doi.org/10.1038/sdata.2018.161
4. Maron RC, Hekler A, Krieghoff-Henning E, Schmitt M, Schlager JG, et al. 2021. Reducing the impact of confounding factors on skin cancer classification via image segmentation: technical model study. *J Med Internet Res* 23(3): e21695. https://doi.org/10.2196/21695
5. Dzieniszewska A, Garbat P, Piramidowicz R. 2024. Improving skin lesion segmentation with self-training. *Cancers* 16(6): 1120. https://doi.org/10.3390/cancers16061120
6. Ferreira PM, Mendonça T, Rozeira J, Rocha P. 2012. An annotation

tool for dermoscopic image segmentation. In Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications, pp 1-6.

7. Ren T, Liu S, Zeng A, Lin J, Li K, et al. 2024. Grounded sam: assembling open-world models for diverse visual tasks. *arXiv* 2024: 1-11. https://doi.org/10.48550/arXiv.2401.14159
8. Liu S, Zeng Z, Ren T, Li F, Zhang H, et al. 2024. Grounding DINO: marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pp 38-55.
9. Kirillov A, Mintun E, Ravi N, Mao H, Rolland A, et al. 2023. Segment anything. In International Conference on Computer Vision, pp 4015-4026.
10. Deng R, Cui C, Liu Q, Yao T, Remedios LW, et al. 2023. Segment anything model (sam) for digital pathology: assess zero-shot segmentation on whole slide imaging. *arXiv* preprint. https://doi.org/10.48550/arXiv.2304.04155
11. He W, Chen W, Chen B, Yang S, Xie D, et al. 2023. Unsupervised prompt tuning for text-driven object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2651-2661.
12. Codella NCF, Rotemberg V, Tschandl P, Celebi ME, Dusza SW, et al. 2019. Skin lesion analysis toward melanoma detection: a challenge at the international skin imaging collaboration (ISIC) 2018.
13. Otsu N. 1979. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9(1): 62-66. https://doi.org/10.1109/TSMC.1979.4310076
14. Khan AH, Iskandar DNF, Al-Asad JF, El-Nakla S. 2021. Classification of skin lesion with hair and artifacts removal using black-hat morphology and total variation. *Int J Com Dig Sys* 10(1): 597-604. http://dx.doi.org/10.12785/ijcds/100157
15. Zuiderveld KJ. 1994. Contrast limited adaptive histogram equalization. *Graphics Gems* 4(1): 474-485.
16. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, et al. 2021. Swin transformer: hierarchical vision transformer using shifted windows. In Proceedings IEEE/CVF International Conference on Computer Vision, pp 10012-1023.
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2020. An image is worth 16 × 16 words: transformers for image recognition at scale.
18. Fan H, Xie F, Li Y, Jiang Z, Liu J, et al. 2017. Automatic segmentation of dermoscopy images using saliency combined with Otsu threshold. *Comput Biol Med* 85: 75-85. https://doi.org/10.1016/j.compbiomed.2017.03.025
19. Schaefer G, Rajab MI, Celebi ME, Iyatomi H. 2011. Colour and contrast enhancement for improved skin lesion segmentation. *Comput Med Imaging Graph* 35(2): 99-104. https://doi.org/10.1016/j.compmedimag.2010.08.004
20. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, et al. 1994. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol* 30(4): 551-559. https://doi.org/10.1016/s0190-9622(94)70061-3
21. Redmon J, Divvala S, Girshick R, Farhadi A. 2016. You only look once: unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 779-788.