Classification-to-Segmentation: Class Activation Mapping for Zero-Shot Skin Lesion Segmentation

Anonymous Authors

Anonymous Institution

Abstract. Zero-shot skin lesion segmentation allows for efficient processing times, without reliance on laborious data gathering for training supervised segmentation models. However, while foundation models such as Segment Anything (SAM) have shown strong ability to segment skin lesions without prior domain-knowledge, they require manual prompt generation. This paper introduces an automated prompt generation method for Medical Segment Anything Model (MedSAM) that uses class activation maps as weak localizations to produce the required bounding box and coordinate point prompts needed for zero-shot segmentation. This classification-to-segmentation method reduces the time required for clinicians to assess a skin lesion, while maintaining a human-in-the-loop approach by providing an initial assessment mask which can be analyzed and refined. Zero-shot and finetuned classification-to-segmentation performance with an array of established CAM-based explanation methods is assessed for both CNN and vision transformer models from ISIC 2017. With examples such as MobileNet-v2 demonstrating favourable zero-shot generalization, actionable segmentation results are produced. The Code and Demo will be made available.

Keywords: Zero-shot Segmentation · Classification-to-Segmentation · Class Activation Mapping · Segment Anything · Explainable AI

1 Introduction

The use of segmentation in skin lesion imaging is critical for the identification of skin cancers, other skin conditions and the monitoring of their progression. It is instrumental for informing a clinician's decisions relative to a patient's treatment plan based on the features and structures presented within the regions of interest (ROIs) in an image. However, manually segmenting ROIs from medical images is a time-consuming process that is prone to imprecision [16]. Deep learning (DL) approaches to automated medical image segmentation have shown the ability to produce accurate image and pixel level ROI identification for a broad array of diseases and imaging methodologies such as X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and ultrasound [24]. Besides drastically reducing the time taken to assess images [23], automated segmentation models can be implemented in a human-in-the-loop approach allowing clinicians to further prompt the model, and approve or refine its outcome. Predominant approaches to DL-based segmentation include supervised learning strategies with

variants of the UNet architecture [19], such as transformer hybrid ones [4]. However, supervised learning strategies remain fundamentally bottle-necked by their requirement for large amounts of labeled data.

To reduce, or eliminate, the reliance on large annotated datasets for DL segmentation, research on the applications of unsupervised, few-shot (learning from only a few labeled examples) and zero-shot (generalizing to unseen classes without labeled examples) learning has increased in recent years [14]. For example, Meta's large-scale pre-trained Segment Anything Model (SAM) can perform zero-shot segmentation of unseen classes from a broad array of domains without the requirement for image- or pixel-level labels for the new domain [7,26]. SAM accepts bounding box, sample points and mask prompts that can be used to indicate foreground and background regions, aiding in the precision of identified ROIs. Despite its remarkable generalization capabilities, the lack of domain-specific knowledge still results in a decrease in performance in comparison to supervised models [26]. MedSAM [10], a SAM variant fine-tuned on over 1.5 million medical images, is able to reduce the performance gap to supervised models. In general, it has been shown that the strategy for prompting SAM and its variants is integral to performance [11].

In this paper, we investigate how high-level localizations can be produced from the prior knowledge of skin lesion classification models to automatically generate initial segmentation masks using a zero-shot strategy, prior to human-in-the-loop refinement. Unlike existing methods, which have predominantly used prompt strategies such as a manual selection of points/bounding boxes from users, or ground-truth masks [12,11], to increase the performance of zero-shot segmentation we propose the use of image-level classification labels, which are far less time-intensive to collect, and Class Activation Mapping (CAM) methods to produce heatmap-based localizations [20]. By examining the relation between activations of produced feature maps and their importance for prediction of the target class, CAM methods produce a visualization of the relative importance of regions of the input image to the prediction as a heatmap. Our approach regards these heatmaps as noisy masks that can be used for the automated generation of bounding box and sample point prompts to MedSAM, demonstrating the classification-to-segmentation approach for a range of benchmark CNN and transformer models, as well as common CAM variants.

The main contributions of this paper are:

- a zero-shot classification-to-segmentation approach for skin lesion analysis using class activation mapping to generate noisy masks from image-level labels, used to produce bounding boxes and sample points for MedSAM;
- use of image-level classification labels, which are easier and less time-consuming to obtain than pixel-wise annotations;
- a comprehensive assessment of the classification-to-segmentation approach across various CNN and transformer-based models, along with multiple CAM variants;
- promotion of responsible AI by providing clinicians with automatically generated initial segmentation masks to support decision-making.

The remainder of the paper is structured as follows: Section 2 reviews relevant literature and existing approaches related to zero-shot segmentation and classification-to-segmentation. Section 3 details the proposed methodology, including detail on the classification-to-segmentation prompting strategies used. Section 4 presents the experimental setup, dataset details and evaluation metrics. Section 5 presents the results and associated analysis.

2 Related Work

Implementations of Medical Image segmentation commonly use the UNet architecture in supervised-learning tasks [14,19]. Notable works such as RUNet [6], R2U-Net [2] and more recently transformer-based variants, such as TransUNet [4] and UNETR [5], have been utilized for a broad array of tasks including cardiac, lung and skin segmentation. While such supervised-learning implementations are an active area of research and still dominate domain-specific segmentation benchmarks, the emergence of foundation models has allowed a broad range of implementations to be developed that showcase the advantages of not relying on domain-specific data collection [13]. The two learning paradigms which have seen a lot of attention in recent years are unsupervised and zero-shot applications, both with the intrinsic aim of reducing or removing the requirement of expensive domain-data collection. The prominence of large scale foundation models with the ability to adapt to diverse domains for zero-shot segmentation has generated a large number of works, with many enhanced and domain-specific implementations of SAM [26]. Additionally, while not a foundation model, the Unsupervised Universal Image Segmentation (U2Seg) model proposed by Niu et al. has demonstrated strong generalization ability and surpasses previous benchmarks of task specific methods [15]. The comparison and adoption of one of these paradigms is not universal, and requires consideration of domain-specific requirements. For medical image segmentation, this includes the requirements that would be necessary for integrating automated tools into clinical practice.

Unsupervised skin lesion segmentation has developed in recent years, motivated in part by the scarcity of annotated publicly available data [9,14]. Segmentation implementations have traditionally focused on either instance or semantic segmentation, where the individual objects or pixels of an image respectively are assigned a category and associated mask [25]. The recent work of Niu et al. has shown the ability to perform both of these tasks, termed panoptic segmentation, for a broad range of classes from the COCO dataset [15]. While skin lesion segmentation has traditionally incorporated unsupervised learning techniques from manually designed features [9], U2Seg could automate and remove the dependency of these manual constraints. In the medical domain, the implementation of automated DL solutions, despite strong performance presented in the literature, is low from a lack of trust in black-box models [8]. Given that the two main areas clinicians and patients refer to for trust in DL implementations are the task context and the ability to create a mutual understanding of the decision reasoning [21], it is unlikely that clinicians will trust such unsupervised methods.

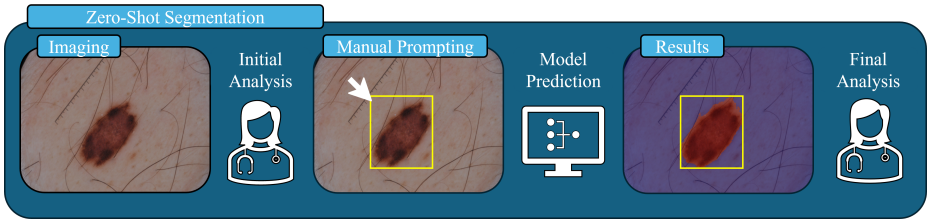

Fig. 1. Schematic of zero-shot vs unsupervised learning paradigms for skin lesion segmentation within clinical practice.

While zero-shot learning strategies also lack domain-knowledge, in contrast to unsupervised methods they require users to generate prompts, such as text, bounding boxes or coordinate points from which the masks are predicted [18]. As shown in Figure 1, the interaction allows the clinician to specify and refine the mask for a broader range of applications that would be required in clinical practice. Kirillov et al. developed SAM as a foundation model for natural image segmentation that can produce valid masks from coordinate, bounding box, mask and text prompts [7]. The model, consisting of a ViT-based image-encoder and transformer-decoder structure, acts to embed and perform cross-attention between the image and prompts, mapping them to a set of N foreground masks. Adapting a model from natural to medical images for zero-shot segmentation is a difficult task from the inherent differences in their structure and composition, with medical images commonly being low contrast, greyscale and 2D or 3D [23]. To improve the performance of SAM for the medical domain, numerous works in recent years have proposed alterations of the architecture, finetuning or optimizations [26].

3 Methodology

To increase the efficiency of zero-shot skin-lesion segmentation, which typically requires manual prompts, this paper demonstrates a methodology to incorporate techniques from responsible AI research to enable automated prompt generation for foundation models. Post-hoc prediction explanation visualization methods have seen increased use across domains, including medical imaging where responsible AI research has been prominent. Examples include Class Activation Mapping (CAM) methods, such as Grad-CAM [20] originally developed for CNNs, which calculates the gradient of a predicted class logit with respect to the activations of the final layer feature map, producing a heat-map of important regions of the input image to the output classification. As shown in Figure 2, a broad array of CAM-based localization methods have been developed. To use these localizations, produced from image-level classification labels, for guiding zero-shot segmentation models such as SAM requires treating these localization heatmaps

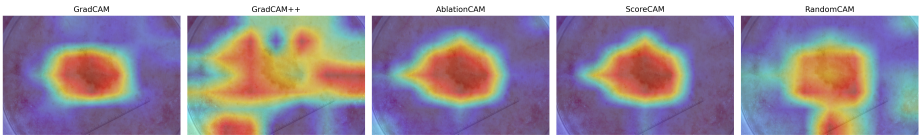

Fig. 2. Heatmaps examples from CAM methods applied to swin finetuned for ISIC 2017 melanoma classification [20,3,17,22].

as weak or noisy segmentation masks, which can be refined to produce a combination of mask, bounding box and coordinate point prompts. The subsequent sections outlines the process of automatically generating these prompts from classifiers fine-tuned on binary skin-lesion classification and additionally the complete zero-shot scenario. The presented zero-shot classification-to-segmentation is the most extreme example of non-data reliance, in which localizations from classifiers are produced with no domain training.

3.1 CAM-based Prompt Generation

The proposed method shown in Algorithm 1 outlines the automatic generation of bounding boxes and coordinate point prompts from an image with CAM. A preprocessed input image is first passed through an ImageNet pretrained or ISIC 2017 finetuned model to obtain the prediction logits. The heatmap is then extracted, and to localize the salient region, a bounding box is computed by applying a percentile intensity threshold to the heatmap and selecting the minimum and maximum coordinates from the remaining points. If the number of high-activation coordinates is insufficient to define a bounding region reliably, the algorithm falls back to a probabilistic sampling method. This involves flattening and amplifying the heatmap to emphasize stronger activations, followed by sampling based on normalized probabilities. Additionally the resulting indices are mapped back to 2D coordinates, producing a set of spatial prompts that reflect the model's attention. The final output consists of the bounding box and set of coordinate prompts that together capture the most informative regions of the input image to be useful as prompts to a foundation model.

3.2 Classification-to-Segmentation

As shown in Figure 3, the proposed methodology enables improving the efficiency of zero-shot segmentation by providing a clinician with an initial segmented assessment of the image prior to any manual instructions for refining the lesion mask. By automatically providing a mask from a classifier with skin-lesion domain knowledge, while maintaining a human-in-the-loop approach, the classification-to-segmentation approach demonstrates how responsible AI implementations can be used to generate actionable results. Thereby supporting clinicians with efficient tools rather than naively replacing them.

Algorithm 1 CAM-based Automatic BBox and Point Prompt Generation

Require: $image\_path$, $model$, $CAM\_extractor$, $percentile$, num_points, $amplification\_factor$

$image \leftarrow$ `LoadImage`$(image\_path)$
$input\_tensor \leftarrow$ `Preprocess`$(image)$
$model \leftarrow$ `SetEvalMode`$()$, $output \leftarrow model(input\_tensor)$
$predicted\_class \leftarrow \arg\max(output)$
$activation\_map \leftarrow CAM\_extractor(input\_tensor,\ predicted\_class)$
$heatmap \leftarrow$ `NormalizeResize`$(activation\_map[0])$
$threshold \leftarrow$ `Percentile`$(heatmap,\ percentile)$, $mask \leftarrow heatmap \geq threshold$
$coordinates \leftarrow$ `IndicesWhere`$(mask)$
if $|coordinates| \geq 4$ **then**
    $[y_{min}, x_{min}], [y_{max}, x_{max}] \leftarrow \min\,/\,\max(coordinates)$
else
    $flattened \leftarrow$ `Flatten`$(heatmap)^{amplification\_factor} + \varepsilon$
    $valid\_indices \leftarrow$ `NonZero`$(flattened)$
    $probabilities \leftarrow$ `Normalize`$(flattened[valid\_indices])$
    $sampled \leftarrow$ `Sample`$(valid\_indices,\ num\_points,\ prob)$
    $x \leftarrow sampled \bmod width$, $y \leftarrow \lfloor sampled/width \rfloor$
    $[x_{min}, y_{min}], [x_{max}, y_{max}] \leftarrow \min\,/\,\max(x,\ y)$
end if
$coordinate\_prompts \leftarrow$ `SamplePoints`$(heatmap,\ num\_points,\ amplification\_factor)$
return $[x_{min}, y_{min}, x_{max}, y_{max}]$, $coordinate\_prompts$

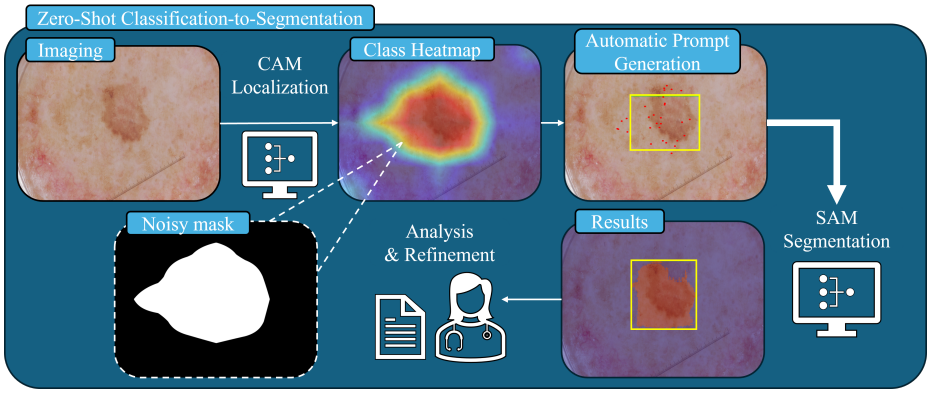

Fig. 3. Schematic of zero-shot classification-to-segmentation for skin lesions.

4 Experiments

4.1 Experimental Setup

The data used in this paper to finetune a range of models for skin lesion classification is the ISIC 2017 challenge dataset[1], consisting of 2000 training and 600 test images for both binary classification and segmentation tasks, the dataset allows for experiments of how to localizations of classifiers can be used to support segmentation. To evaluate the segmentation performance, standard metrics available from the literature (precision, recall, F1, IoU and dice score) were used.

Experiments were conducted on a high-performance computing (HPC) cluster. Each compute node was equipped with 64 Intel CPU cores (2 × Intel Xeon Gold 6430, 32 cores per processor, 2.1 GHz), 512 GB of RAM, and 4 NVIDIA L40 GPUs with 48 GB of VRAM each, connected via PCIe. The HPC environment provided access to the NVIDIA HPC SDK which was used to enable CUDA 12.0 support. Across all experiments, the memory footprint remained below 15 GB of system RAM per process, well within the available node resources.

4.2 CAM Segmentation Baseline

The first experiments act as a baseline for the localization abilities of the common CAM-methods: GradCAM, GradCAM++, AblationCAM, ScoreCAM. To provide a baseline for the CAM-methods, RandomCAM is used which generates heatmaps with random uniform values in the range [-1, 1]. To justify the usefulness of CAM localizations, their performance should be greater than those randomly produced from RandomCAM. In order to get the broadest analysis on producing zero-shot segmentation results from classifiers, a range of state-of-the-art models were chosen: ResNet50, MobileNet-v2, EfficientNet, ViT and Swin. The mix of CNN and Transformer based computer vision models allows for additional analysis of the relative localization ability of these architectures and the associated CAM-methods, which were developed for CNNs.

For zero-shot classification, each CAM-method is applied to the chosen models, loaded with pretrained weights from ImageNet, for each image in the test set, producing a heatmap. From which each step in Algorithm 1 up to step 7 is applied, resulting in a threshold mask that is converted into a binary mask that is compared to the ground truth mask to determine the metric values. For finetuned classification, the same experimental procedure is applied except all models were first trained on ISIC 2017 for a maximum of 500 epochs and a patience of 20.

4.3 Classification-to-Segmentation

The main component of the experiments involve evaluating the application of Algorithm 1 to produce the prompts needed as input to a foundation model, MedSAM, for zero-shot skin lesion segmentation. This classification-to-segmentation

[1] ISIC dataset publicly available at: https://challenge.isic-archive.com/data/#2017

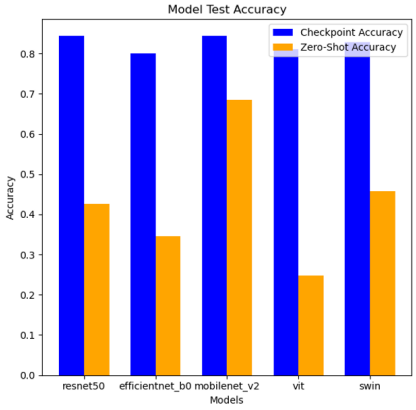

Fig. 4. Zero-shot and finetuned test accuracy for models on ISIC 2017.

approach is broken down into two main experiments: the individual method and the best method. The individual method applies each model and CAM-method, storing the associated results and is important to understand the relative performance of each permutation. The best method is an approach which reflects the challenges from both the non-homogeneous nature of skin-lesion images and the lack of universal localization ability from available CAM-methods. This method therefore produces the initial mask for each CAM-method and returns the best resultant mask for evaluation, simulating providing a clinician with multiple mask options and selecting the best. Both experiment methods are conducted for zero-shot and finetuned classifiers, whereby the resultant bounding box and coordinate point prompts from Algorithm 1 are used to prompt MedSAM, from which the output mask is evaluated with the metrics outlined.

5 Results

Figure 4 shows the test accuracy of the chosen models on the ISIC 2017 dataset for both zero-shot and finetuned. As expected, finetuned models achieved higher accuracy across the board, all over 80%. In contrast, zero-shot performance varied significantly. MobileNet-V2 showed the strongest performance for both cases, with a drop in performance of 17.2%. Meanwhile, the next closest model, swin, saw a performance drop of 47.6%. These results highlight the gain from finetuning and the variability in generalization ability across architectures in zero-shot classification settings. These results additionally indicate that if intuitively the classification ability of these models is correctly linked to the ability to localize salient regions of the skin lesion images, that MobileNet-v2 should have the strongest performance in both zero-shot and finetuned prompt generation conditions.

Table 1. Mean metrics of CAM methods for noisy mask generation applied to a finetuned Swin model.

Method	Precision	Recall	F1	IoU	Dice
GradCAM	0.4873	0.3998	0.3454	0.2356	0.3454
GradCAM++	0.2537	0.1650	0.1547	0.0960	0.1547
AblationCAM	0.2940	0.3287	0.2677	0.1802	0.2677
ScoreCAM	0.3645	0.3013	0.2630	0.1768	0.2630
RandomCAM	0.2864	0.3289	0.2254	0.1482	0.2254

Table 1 shows the mean performance metrics for each CAM method applied to a finetuned Swin Transformer model. These metrics reflect the quality of the noisy mask generation, which is necessary for prompting in the later segmentation tasks. GradCAM achieved the highest performance across all metrics, with an IoU of 0.236 and dice score of 0.345, indicating relatively better localization ability. In contrast, GradCAM++ yielded the lowest scores, suggesting reduced effectiveness in generating meaningful activation maps in this context. AblationCAM and ScoreCAM showed comparable results, though both underperformed compared to GradCAM. Interestingly, RandomCAM outperformed GradCAM++ in most metrics, showcasing a worse than random performance an emphasizing the sensitivity of certain methods to model architecture and training. These results highlight substantial variability across CAM methods and therefore the importance of method selection when using CAMs for prompt generation.

The evaluation of the individual combinations of classification model and CAM method to produce prompts for MedSAM is shown in Figure 5 for both zero-shot and finetuned classifiers. The heatmaps, representing the mean IoU values for each combination, demonstrate the relative performance increase of MedSAM when prompts are created from finetuned classifiers with domain-knowledge. This result supports the premise, and base intuition, that domain training increases the ability of a model to localize known classes. The highest individual result for zero-shot classification-to-segmentation is ScoreCAM applied to MobileNet-v2, with an IoU of 0.184. While for the finetuned case, MobileNet-v2 with AblationCAM achieved an IoU of 0.282, the highest individual CAM-method performance across the CAM-baseline, zero-shot and finetuned MedSAM experiments.

Table 2 reports the mean segmentation performance of MedSAM prompted by the best CAM-method for each zero-shot and fine-tuned classifier. Conforming with previous results, Fine-tuned MobileNet-V2 achieved the highest performance overall, with an IoU of 0.366 and a dice score of 0.510. Swin ranked second, with an IoU of 0.357 and dice of 0.498. In the zero-shot setting, MobileNet-V2, as expected, led with an IoU of 0.337 and dice of 0.470, followed by Swin. Compared to the ablation baselines (SAM-base: IoU 0.672, MedSAM-base: IoU

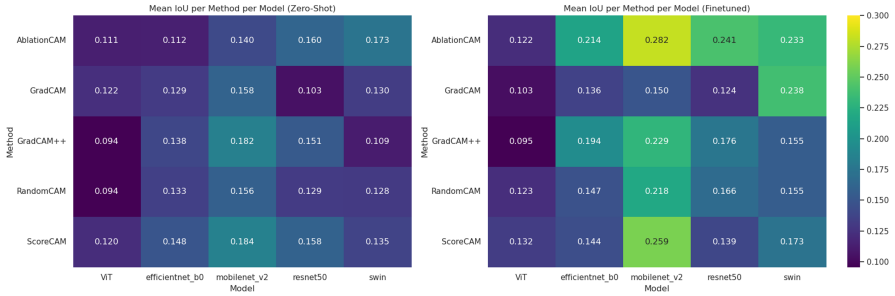

Fig. 5. Comparison of mean IoU for CAM methods vs models for zero-shot (left) and finetuned (right) classifiers.

0.671), all CAM-prompt methods performed below the manual prompting methods from users or bounding boxes using the ground truth, but fine-tuned CAMs were shown to be able to produce actionable segmentation masks. These results highlight the potential of CAM-guided prompting for zero-shot segmentation, particularly when classifiers are fine-tuned.

By inspecting the distribution of IoU values for each model, shown in Figure 6, the improvement in performance from adaptively selecting the most appropriate CAM-method is clear from the broad increase in the frequency of higher IoU values and fewer instances of lower or zero values. For example, the highest performing model, MobileNet-v2, ∼400 test samples produced segmentation masks with an IoU of zero for the zero-shot case, indicating total failure in those cases. In contrast, the fine-tuned MobileNet-V2 reduced this number dramatically to ∼5 samples. This represents a ∼98.75% reduction in complete segmentation failures.

As can be seen from the examples shown in Figure 7, the success, or failure, of the proposed method is based on three main areas: case difficulty, localization quality and specificity of the ground truth mask. The inherent variability of skin lesions and their proportions within imaging is challenging, as CAM-methods rely on discriminating between regions of an image in comparison to humans which can imply an object, even if it takes up a majority of the image. Secondly, the truthfulness and quality of localizations from CAM to the true ROI has to be correct for a given model. Current state-of-the-art explanation methods are not globally truthful and reflect the inherent fallibility of black-box models, often providing evidence of counterfactual predictions, especially when this coincides with correct predictions. While this represents the reality of imperfect classifiers, it presents a current limitation in the ability to localize an image instance from a model's global class-understanding. Finally, the variability of the task itself is challenging from the 'fuzzy' and pixel-specific ground truth styles. Whereby some examples outline the broad area in which the lesion is found, while others have fine-grained pixel-wise annotations.

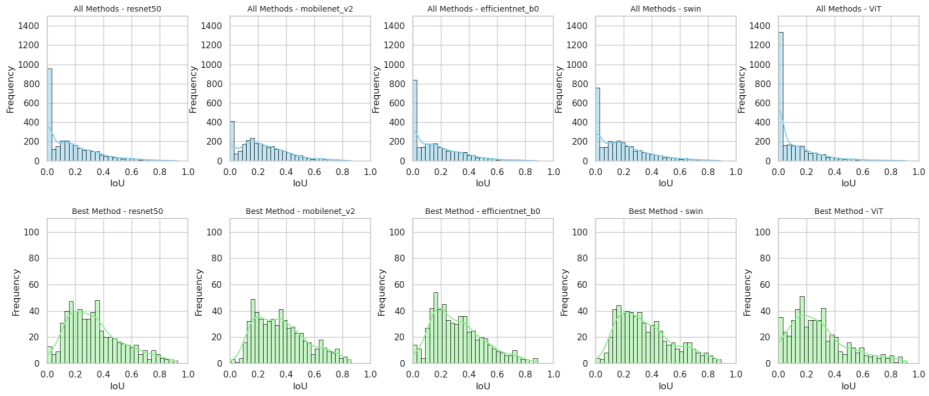

Fig. 6. Distribution of IoU values for all methods and best selected method for the test-set for finetuned classifiers.

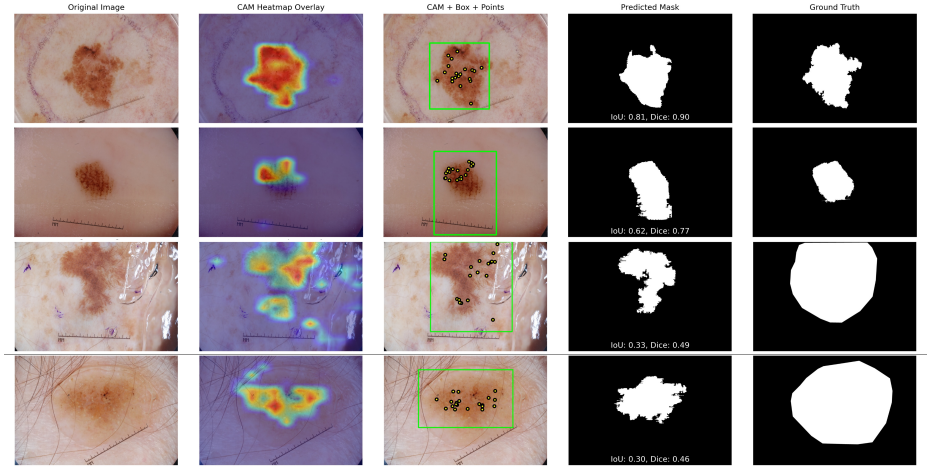

Fig. 7. Example of four test set results representing a range of qualities

Table 2. Segmentation performance of MedSAM guided by the best CAM method prompts from zero-shot and finetuned classifiers. Highlighted rows show the best (blue) and second best (orange) results, respectively.

	Model	Precision	Recall	F1	IoU	Dice
Ablation	SAM-base	-	-	-	0.672	0.805
	MedSAM-base	-	-	-	0.671	0.782
Zero-shot	EfficientNet	0.719	0.384	0.414	0.290	0.414
	MobileNet_v2	0.740	0.468	0.470	0.337	0.470
	ResNet50	0.723	0.399	0.427	0.300	0.427
	ViT	0.723	0.345	0.384	0.269	0.384
	Swin	0.743	0.427	0.446	0.315	0.446
Finetuned	EfficientNet	0.713	0.479	0.457	0.320	0.457
	MobileNet_v2	**0.764**	**0.547**	**0.510**	**0.366**	**0.510**
	ResNet50	0.739	0.496	0.475	0.337	0.475
	ViT	0.750	0.357	0.409	0.285	0.409
	Swin	**0.750**	**0.534**	**0.498**	**0.357**	**0.498**

Base SAM and MedSAM results for ISIC 2017 are from [1].

6 Conclusion

In this work, we adopted the use of Class Activation Mapping (CAM) methods for guiding zero-shot skin lesion segmentation with MedSAM. Our results demonstrate that fine-tuned classification models improved CAM-based bounding box and coordinate prompts, allowing for automated segmentation with MedSAM, thus improving on the common requirement for manual labelling. Among the tested architectures, MobileNet-V2 consistently achieved the highest performance in both zero-shot and fine-tuned settings, highlighting its suitability in zero-shot scenarios.

While CAM methods provided reasonable localization prompts, their performance varied across architectures – especially with transformer-based models like ViT and Swin. This suggests that standard CAM techniques may not fully capture the spatial reasoning capabilities of transformers. Future work will explore alternative explainability approaches tailored to transformer architectures. Techniques such as attention rollout or Layer-wise Relevance Propagation (LRP) may offer improved localization fidelity and more meaningful activation maps. Integrating these methods could enhance the quality of automated segmentation prompts and further reduce failure cases, especially in zero-shot scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data and Code Availability. The data used, ISIC 2017, is publicly available at: https://challenge.isic-archive.com/data/#2017. The code implementation and associated demo for this paper will be made available.

References

1. Al Muaitah, A., Deriche, M.: A modified sam-based skin cancer segmentation pipeline: Skin-sa model. In: 2023 24th International Arab Conference on Information Technology (ACIT). pp. 1–4. IEEE (2023)
2. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent Residual U-Net for Medical Image Segmentation. Journal of medical imaging **6**(1), 014006–014006 (2019)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers. Medical Image Analysis **97**, 103280 (2024)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
6. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A.: Left-ventricle Quantification using Residual U-Net. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9. pp. 371–380. Springer (2019)
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992–4003. IEEE, Paris, France (Oct 2023)
8. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy Clinical AI Solutions: A Unified Review of Uncertainty Quantification in Deep Learning Models for Medical Image Analysis. Artificial Intelligence in Medicine **150**, 102830 (2024)
9. Li, X., Peng, B., Hu, J., Ma, C., Yang, D., Xie, Z.: USL-Net: Uncertainty Self-Learning Network for Unsupervised Skin Lesion Segmentation. Biomedical Signal Processing and Control **89**, 105769 (2024)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images. Nature Communications **15**(1), 654 (Jan 2024)
11. Mattjie, C., De Moura, L.V., Ravazio, R., Kupssinskü, L., Parraga, O., Delucis, M.M., Barros, R.C.: Zero-Shot Performance of the Segment Anything Model (SAM) in 2D Medical Imaging: A Comprehensive Evaluation and Practical Guidelines. In: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). pp. 108–112. IEEE, Dayton, OH, USA (Dec 2023)
12. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment Anything Model for Medical Image Analysis: An Experimental Study. Medical Image Analysis **89**, 102918 (Oct 2023)
13. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3523–3542 (2021)

14. Mirikharaji, Z., Abhishek, K., Bissoto, A., Barata, C., Avila, S., Valle, E., Celebi, M.E., Hamarneh, G.: A Survey on Deep Learning for Skin Lesion Segmentation. Medical Image Analysis **88**, 102863 (Aug 2023)
15. Niu, D., Wang, X., Han, X., Lian, L., Herzig, R., Darrell, T.: Unsupervised universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22744–22754 (2024)
16. Preim, B., Botha, C.: Chapter 4 - Image Analysis for Medical Visualization. In: Preim, B., Botha, C. (eds.) Visual Computing for Medicine (Second Edition), pp. 111–175. Morgan Kaufmann, Boston, second edition edn. (2014)
17. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 983–991 (2020)
18. Ren, W., Tang, Y., Sun, Q., Zhao, C., Han, Q.L.: Visual Semantic Segmentation Based on Few/Zero-Shot Learning: An Overview. IEEE/CAA Journal of Automatica Sinica (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
21. Steerling, E., Siira, E., Nilsen, P., Svedberg, P., Nygren, J.: Implementing AI in Healthcare—The Relevance of Trust: A Scoping Review. Frontiers in health services **3**, 1211150 (2023)
22. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
23. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET image processing **16**(5), 1243–1267 (2022)
24. Xiao, H., Li, L., Liu, Q., Zhu, X., Zhang, Q.: Transformers in Medical Image Segmentation: A Review. Biomedical Signal Processing and Control **84**, 104791 (2023)
25. Yin, C., Tang, J., Yuan, T., Xu, Z., Wang, Y.: Bridging the Gap Between Semantic Segmentation and Instance Segmentation. IEEE Transactions on Multimedia **24**, 4183–4196 (2021)
26. Zhang, Y., Shen, Z., Jiao, R.: Segment Anything Model for Medical Image Segmentation: Current Applications and Future Directions. Computers in Biology and Medicine **171**, 108238 (Mar 2024)