GS - TransUNet: Integrated 2D Gaussian Splatting and Transformer UNet for Accurate Skin Lesion Analysis

Anand Kumar[a], Kavinder Roghit Kanthen[a], and Josna John[a]

[a]University of California, San Diego

ABSTRACT

We can achieve fast and consistent early skin cancer detection with recent developments in computer vision and deep learning techniques. However, the existing skin lesion segmentation and classification prediction models run independently, thus missing potential efficiencies from their integrated execution. To unify skin lesion analysis, our paper presents the Gaussian Splatting - Transformer UNet (GS - TransUNet), a novel approach that synergistically combines 2D Gaussian splatting with the Transformer UNet architecture for automated skin cancer diagnosis. Our unified deep learning model efficiently delivers dual-function skin lesion classification and segmentation for clinical diagnosis. Evaluated on ISIC-2017 and PH2 datasets, our network demonstrates superior performance compared to existing state-of-the-art models across multiple metrics through 5-fold cross-validation. Our findings illustrate significant advancements in the precision of segmentation and classification. This integration sets new benchmarks in the field and highlights the potential for further research into multi-task medical image analysis methodologies, promising enhancements in automated diagnostic systems.

Keywords: Skin lesion analysis, Gaussian Splatting, Vision Transformer, Dermoscopy

1 INTRODUCTION

Skin cancer has emerged as one of the most critical challenges in public health, with melanoma—the deadliest form—accounting for approximately 75% of all skin cancer-related deaths.[1] Early detection and accurate diagnosis are paramount in mitigating mortality rates and improving patient outcomes. Over the past decades, medical imaging and artificial intelligence advancements have facilitated automated skin cancer diagnosis systems, demonstrating performances on par with expert dermatologists.[2,3] However, these systems often simplify the task to binary classification, neglecting the crucial role of lesion segmentation. In practice, segmentation provides essential information about the lesion's asymmetry, border irregularities, intensity, and size, which are indispensable for effective diagnosis.[4]

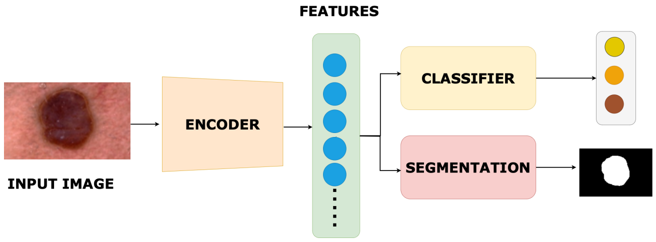


Figure 1: A simplified overview. The flowchart shows how our dual-task approach works for skin lesion classification and segmentation in a unified manner.

Recognizing this gap, we propose Gaussian Splatting-Transformer UNet (GS - TransUNet), a novel framework designed to optimize lesion segmentation and classification tasks jointly. By integrating Convolutional Neural Networks (CNNs) and Vision Transformers (ViT),[5] GS - TransUNet captures both local details (lesion texture and color) and global context (lesion shape and spatial relationships), as shown in Fig.1. Unlike previous models,[6]

Correspondence: ank029@ucsd.edu

GS - TransUNet introduces 2D Gaussian Splatting, a fully differentiable method for generating precise binary masks of elliptical lesions, improving segmentation accuracy even in cases without ground truth masks. This integration enhances segmentation consistency and ensures robust classification performance by focusing on the lesion region and ignoring artifacts such as hair, scales, and tapes.

Building upon the foundational architecture of UNet,[7] GS - TransUNet addresses UNet's inherent limitation of modeling long-range dependencies by incorporating Transformer layers into the encoder. These layers leverage Multi-Head Self-Attention (MSA) to establish correlations over long distances, enabling the segmentation of expansive or irregular regions commonly observed in dermoscopic images. The framework also introduces innovative loss strategies, including Dual Task Consistency (DTC),[8] to ensure alignment between segmentation and classification outputs, further strengthening the network's robustness. We validate GS - TransUNet on the ISIC-2017[9] and PH2[10] datasets, demonstrating its efficacy in both segmentation and classification. Our results highlight significant improvements in metrics such as the DICE coefficient, Jaccard index, precision, recall, and classification accuracy, surpassing state-of-the-art models. Moreover, the framework's computational efficiency and consistency losses ensure practical applicability for real-world dermatological diagnosis.

To summarize, our framework distinguishes itself from existing approaches through these advancements:

1. GS - TransUNet introduces a dual-task model that jointly optimizes segmentation and classification tasks. By leveraging Gaussian splatting, the model seamlessly integrates these two tasks, enabling mutual reinforcement and improving overall performance.

2. We introduce a novel approach that generates segmentation masks through two parallel networks using Gaussian splatting and signed distance fields. This dual-path design ensures that the generated masks are consistent and robust to noise, addressing common challenges in medical image segmentation.

3. GS - TransUNet achieved a 2.5% improvement in accuracy, setting new benchmarks in classification and segmentation tasks.

1.1 Related Works

Skin cancer analysis has had various advancements over the years, which have been influential in designing our model's architecture. We highlight and compare other approaches with ours in the following section.

1.1.1 Skin Lesion Segmentation

The evolution from thresholding and active contour models [11–13] to deep neural networks has revolutionized skin lesion segmentation, as seen in the 19-layer fully convolutional network employing Jaccard distance loss[14] and star shape priors for global structure preservation.[15] With inspiration from,[16] GS - TransUNet advances segmentation by introducing 2D Gaussian splatting and dual-task learning, ensuring robust and consistent mask generation without reliance on specific priors or loss functions.

1.1.2 Skin Image Classification

The advent of deep neural networks has significantly enhanced skin image classification, transitioning from manual feature engineering[17] to automated methods like attention residual learning.[18] GS - TransUNet employs Vision Transformers (ViT) to enhance classification by capturing long-range dependencies and global context in skin images.

1.1.3 Multi-task Learning

Initial multi-task learning approaches like MB-DCNN[19] and MT-TransUNet[20] showcased the benefits of integrating segmentation and classification tasks but often relied on separate or token-based methods. GS - TransUNet advances multi-task learning with an end-to-end framework that enhances task interdependence through dual-task consistency and robust mask generation.

1.1.4 Advanced Vision Transformers in Dermatology

Vision Transformers (ViTs),[5] originally developed for natural language processing, have redefined the analysis of dermatological images by capturing long-range dependencies and global context. GS - TransUNet builds on ViTs within the Transformer UNet architecture, leveraging their strengths for dermatological image analysis.

1.1.5 Task Regularization for Enhanced Learning

Consistency regularization has proven effective in semi-supervised learning contexts, as demonstrated by MT-TransUNet,[8,21] which aligns segmentation and classification through dual-task and attended region consistency losses. GS - TransUNet enhances this strategy with 2D Gaussian splatting and dual-task consistency loss, achieving robust performance without heavy dependence on large labeled datasets.

The development of GS - TransUNet model stands as a testament to the confluence of deep learning innovations, transformer architectures, and consistency regularization strategies in the domain of skin lesion analysis and pushing the boundaries for obtaining robust and reliable results.

2 METHODS

2.1 Model Architecture

The GS - TransUNet model as shown in Fig. 2combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers to capture local and global features from skin lesion images. The architecture integrates a Transformer UNet structure, which effectively models long-range dependencies, a key requirement for handling the complex patterns found in skin lesions. This is achieved by employing the Vision Transformer as the encoder and utilizing UNet-style skip connections to preserve spatial information during upsampling.

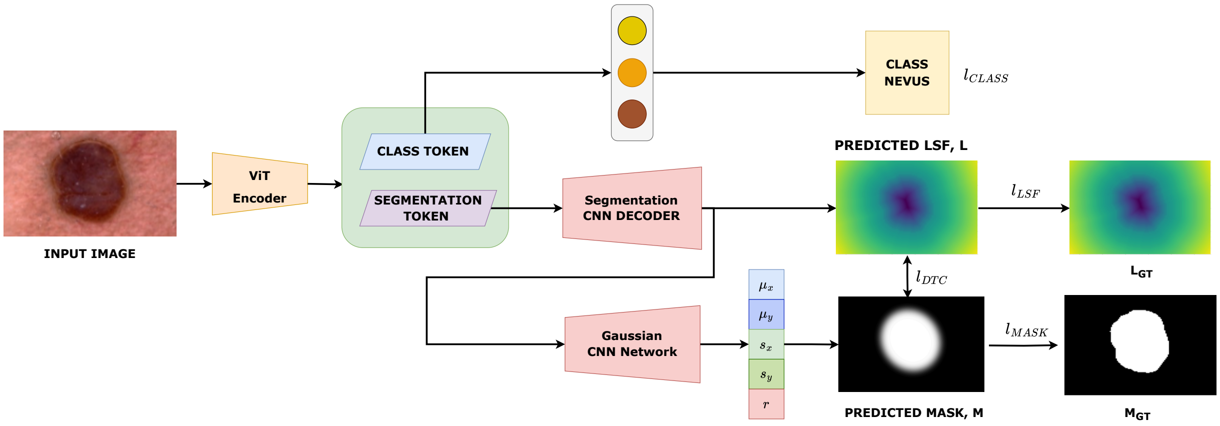


Figure 2: Architecture of the GS - TransUNet model, combining Vision Transformer with UNet for enhanced segmentation and classification.The input image is passed through a pre-trained vision transformer encoder to obtain the class(global) and segment(local) tokens. These tokens are used for simultaneous classification and segmentation. The classification is done using an MLP, and the segmentation mask is predicted using two parallel networks: (i) CNN Decoder to obtain level set function and (ii) CNN Gaussian Network to obtain binary mask. More explanation in Appendix A.

The input image I is first downsampled by a factor of 4 using a ResNet50 backbone [22] to create a feature map F of size $H' \times W' \times C'$. These features are split into patches and passed through the Vision Transformer layers[5] to generate embedding tokens, which are then used to predict class labels and generate segmentation masks.

The input features are converted into patches of size $P \times P$ and then mapped to a sequence of linear embeddings. These embeddings form the input tokens for the Transformer:

$$\mathbf{X} = \text{Reshape}\left(\text{Conv}\left(I\right)\right)$$
$$\mathbf{Z} = \mathbf{AX}$$

where $\mathbf{A}$ is a learnable linear map that projects the input patch into a D-dimensional embedding space.

The Transformer layer primarily comprises Multihead Self-Attention (MSA)[23] and a Feed-Forward Network (FFN), with Layer Normalization (LN) applied after each operation with detailed architecture given in Appendix A. The outputs of each layer are given by:

$$\mathbf{Z_i} = \text{MSA}(\text{LN}(\mathbf{Z_{i-1}})) + \mathbf{Z_{i-1}}$$
$$\mathbf{Z_i} = \text{FFN}(\text{LN}(\mathbf{Z_i})) + \mathbf{Z_i}$$

The tokens obtained after the transformer layers are split into two: the *classification*(global) token, which is the first index of the tokens ($\mathbf{Z_i}[0,:]$) and *segmentation*(local) tokens, the remaining tokens ($\mathbf{Z_i}[1:,:]$).

2.2 2D Gaussian Splatting

A novel 2D Gaussian Splatting method as shown in Fig. 3 is used to generate segmentation masks, focusing on creating consistent and accurate representations of elliptical lesion boundaries. This method utilizes Gaussian functions to model the shape and scale of lesions, improving the precision of boundary delineation compared to traditional deconvolution techniques. This splatting method is particularly effective in highlighting the asymmetry and irregular borders characteristic of melanoma.

To model the lesions, we use a Gaussian function for a pixel(x, y) in the image, $G(x,y) = \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right)$, where $\Sigma = R\begin{bmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{bmatrix} R^T$ and $R = \begin{bmatrix} \cos r & -\sin r \\ \sin r & \cos r \end{bmatrix}$ is the rotation matrix for angle r. The input consists of 6 features: center coordinates (μ_x and μ_y), size of the splat (s_x and s_y) and rotation (r) in radians.

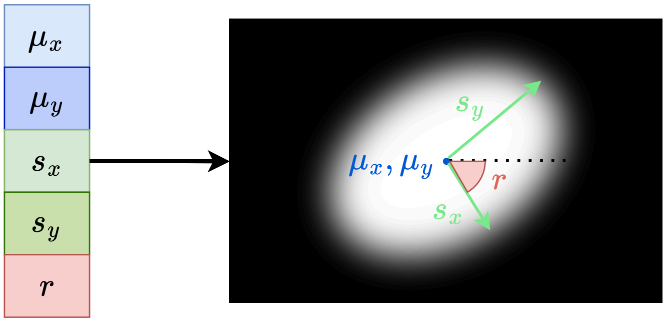


Figure 3: Illustrative diagram for generating binary masks using 2D Gaussian Splatting. The Gaussian Splats are generated using 6 features: $\mu_x, \mu_y, s_x, s_y, r$

This Gaussian is "splatted" onto the image grid to form the segmentation mask M. The elements of the covariance matrix are adjusted based on the scale and rotation parameters derived from the network.

2.3 Loss Functions and Training Strategy

Our model employs several loss functions to enhance segmentation and classification accuracy. Focal Loss[24] addresses class imbalance by emphasizing difficult samples, defined as:

$$\text{Focal Loss}(p, y) = -\sum_{i=1}^{K} \left(y_i \log(p_i)(1-p_i)^\gamma \alpha_i + (1-y_i)\log(1-p_i)p_i^\gamma(1-\alpha_i)\right),$$

where α and γ are hyperparameters, leading to the classification loss $l_{\text{CLASS}} = \text{FocalLoss}(\hat{z}, z_{GT})$. Dice Loss[25] enhances segmentation by maximizing the overlap between predicted and true masks:

$$\text{Dice Loss}(X, Y) = 1 - \frac{2|X \cap Y|}{|X| + |Y|}.$$

L_2 Loss aligns predicted masks with ground truth when available:

$$l_{MASK} = L_2(M, M_{GT}), \quad l_{LSF} = L_2(L, L_{GT}).$$

The level set function (L) encourages dual task consistency by representing distances from the boundary. It is derived from the binary mask using:

$$T(x) = \begin{cases} -\inf_{y \in \partial S} ||x - y||_2, & x \in S_{\text{in}} \\ 0, & x \in \partial S \\ +\inf_{y \in \partial S} ||x - y||_2, & x \in S_{\text{out}} \end{cases} = \text{sign}(x) \cdot \inf_{y \in \partial S} ||x - y||_2$$

Each pixel in L is valued by its minimum distance to the boundary ∂S, with directionality indicating whether it is inside $(-)$ or outside $(+)$ the boundary.

When segmentation ground truth is available, L and the binary mask M are aligned with the ground truth using L2 loss:

$$l_{MASK} = L_2(M, M_{GT}), \quad l_{LSF} = L_2(L, L_{GT})$$

The Dual-Task Consistency Loss ensures consistency between binary masks and level set functions by converting the level set function (L) back to a binary mask (M') and comparing it with the generated binary mask (M):

$$M' = \text{Sigmoid}(k \cdot L), \quad l_{DTC} = L_2(M', M),$$

where k is typically set to 1500.

Our overall loss function combines these elements, optimizing segmentation and classification with the equation:

$$l_{\text{TOTAL}} = l_{\text{CLASS}} + \lambda_M \cdot l_{MASK} + \lambda_L \cdot l_{LSF} + \lambda_{DTC} \cdot l_{DTC} + \lambda_{Dice} \cdot \text{Dice Loss}(M, M_{GT}),$$

where λ_M, λ_L, λ_{DTC}, and λ_{Dice} are weighting factors balancing each loss term's contribution. The training strategy uses a batch size of 8 with data augmentation (flipping, rotation, noise) to improve generalization. These loss functions guide the model to learn effectively from both labeled and unlabeled data, enhancing robustness and accuracy.

3 EXPERIMENTS

3.1 Datasets and Evaluation Metrics

The model was evaluated on the ISIC-2017 and PH2 datasets, which provide a comprehensive set of dermoscopic images with corresponding segmentation masks and classification labels. The evaluation metrics include the Jaccard index (JA), Dice coefficient (DI), pixel-wise accuracy (pixel-AC), sensitivity (pixel-SE), specificity (pixel-SP), and classification accuracy (AC).

3.1.1 ISIC-2017 Dataset

The ISIC-2017 dataset contains 2,750 dermoscopic images with annotations for training and testing lesion segmentation and classification algorithms.[9] This dataset presents a variety of lesion types, providing a robust foundation for model validation.

3.1.2 PH2 Dataset

The PH2 dataset consists of 200 dermoscopic images with detailed segmentation masks and classification labels.[10] It includes images of melanomas, common nevi, and atypical nevi, offering a diverse sample for evaluating model performance.

Models	Melanoma Classification					Keratosis Classification					Average
	M_AC	M_AUC	M_AP	M_SE	M_SP	K_AC	K_AUC	K_AP	K_SE	K_SP	Accuracy
RegNet[28]	0.860	0.812	0.596	0.400	0.975	0.940	0.987	0.957	0.904	0.953	0.806
EffNet[29]	0.860	0.861	0.740	0.400	0.975	0.926	0.979	0.946	0.833	0.962	0.813
ViT[5]	0.886	0.898	0.752	0.600	0.958	0.960	0.968	0.943	0.714	0.972	0.813
SDL[18]	0.876	-	-	-	-	0.933	-	-	-	-	0.814
MB-DCNN[19]	0.856	0.892	0.703	0.443	0.963	0.914	0.931	0.884	0.872	0.823	0.807
MT-TransUNet[6]	0.873	0.895	0.678	0.466	0.975	0.926	0.959	0.901	0.881	0.944	0.826
GS - TransUNet w/o Focal Loss(Ours)	0.880	0.905	0.749	0.600	0.950	0.926	0.954	0.939	0.761	0.991	0.826
GS - TransUNet w Focal Loss(Ours)	0.886	0.889	0.756	0.567	0.966	0.940	0.969	0.953	0.880	0.969	0.846

Table 1: Melanoma and Keratosis Classification Results for different networks with the best highlighted using red and the second best using yellow for each metric baseline networks. AC, AUC, AP, SE, and SP refer to Accuracy, Area Under Curve, Average Precision, Sensitivity, and Specificity respectively.

3.2 Implementation Details

We use a batch size of 8, split into two mini-batches of size 4, trained on l_{SEG} and $l_{NON-SEG}$ respectively, enhancing robustness to samples without segmentation masks. Networks are pre-trained on ImageNet1K.[26] Weights are updated with the Adam optimizer[27] using default momentum and weight decay. Models are trained for 80 epochs with an initial learning rate of 10^{-5}, which decreases on plateau. The transformer-based networks use 16×16 image patches with ResNet50[22] as the backbone and 4 transformer layers. We set $\lambda_M = 0.25$, $\lambda_L = 0.5$, $\lambda_{Dice} = 0.5$, and exponentially increase λ_{DTC} over epochs. Focal loss parameters γ and α are set to 2 and 0.25.

Model performance is evaluated using 5-fold cross-validation, where the dataset is split into 5 equal parts, and the model is trained from scratch 5 different times by keeping one of the parts as the testset and the rest 4 as trainset. During training, images are augmented with vertical and horizontal flips, resizing, center-cropping, rotation, and Gaussian noise, each with a 0.5 probability. Input images are sized 224×224. Each image undergoes three rounds of augmentation for testing, and results are averaged.

Our models are implemented using PyTorch and Sklearn, trained with CUDA on an RTX 3090 GPU with 24 GB RAM.

3.3 Performance Analysis

The GS - TransUNet model demonstrates substantial improvements over state-of-the-art and baseline methods. For baseline methods, we used EffNet (Efficient Net),[29] RegNet (Regulated Net),[28] and ViT (Vision Transformer)[5] models pre-trained on ImageNet.[26] We also performed classification using traditional statistical models such as Support Vector Machines (SVMs),[30] AdaBoost,[31] Extreme Gradient Boosting (XGBoost),[32] etc. and their results are given in Appendix B. The state-of-the-art methods include MT-TransUNet (Multi-Task Transformer UNet),[6] MB-DCNN (Mutual Bootstrapping Deep Convolutional Neural Networks)[19] and SDL (Synergic Deep Learning).[18] For melanoma classification, the model achieved an accuracy of 88.6% with an AUC of 0.889, while keratosis classification reached an accuracy of 94.0% with an AUC of 0.969 as shown in Table 1. These results highlight the model's efficacy in distinguishing between skin lesions, a crucial capability for accurate diagnosis.

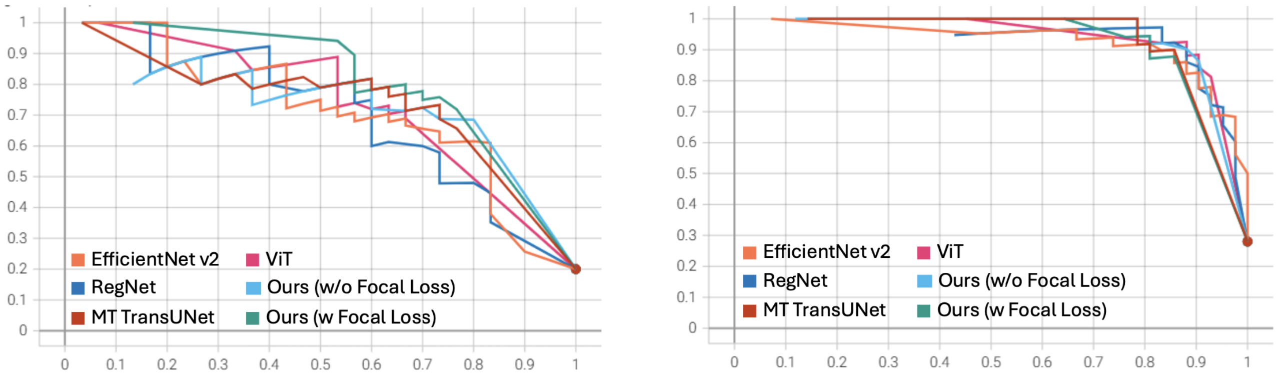
Table 1 provides a detailed comparison of classification metrics across different models, illustrating GS - TransUNet's superior performance. The model's segmentation results, presented in Table 2, further emphasize its ability to generate precise and consistent lesion boundaries, crucial for reliable diagnosis where our model outperforms other state-of-the-art methods such as MB-DCNN and MT-TransUNet on pixel-wise accuracy and pixel-wise specificity resulting fewer false positives.

The Precision-Recall (PR) curves for both melanoma Fig. 4a and keratosis classification Fig. 4b were instrumental in visualizing the model's efficacy. Our model with focal loss sustained higher precision across an extended

Methods	Jaccard (JA)	DICE Score	Pixel-AC	Pixel-SE	Pixel-SP
MB-DCNN[19]	0.595	0.730	0.914	0.839	0.921
MT-TransUNet[6]	0.606	0.733	0.928	0.797	0.930
GS - TransUNet w/o Focal Loss(Ours)	0.580	0.712	0.932	0.692	0.973
GS - TransUNet w Focal Loss(Ours)	0.554	0.689	0.934	0.630	0.973

Table 2: Segmentation Results for different networks with the best highlighted using red and the second best using yellow for each metric.

range of recall values than its counterpart without focal loss and other models. These curves emphasized the model's capability to maintain a high true positive rate with minimal false positives, particularly beneficial in medical contexts where the stakes of misdiagnosis are high.



(a) Melanoma

(b) Keratosis

Figure 4: Precision-Recall Curves for Melanoma and Keratosis cases

3.4 Output Visualization

Figs. 5 and 6 display qualitative results from the GS - TransUNet model. The segmentation outputs clearly delineate lesion boundaries, with our model ensuring robust handling of artifacts like hair and shadows, improving both segmentation and classification outputs. However, challenges remain in handling irregularly shaped lesions, as the Gaussian splatting method inherently favors elliptical structures, leading to oversimplified masks. These failure cases illustrate areas for further improvement, such as handling complex lesion shapes or varying lighting conditions. Despite these limitations, GS - TransUNet demonstrates superior performance and generalization, with future improvements needed to address these edge cases.

4 DISCUSSION AND CONCLUSION

In this paper, we present GS - TransUNet, a unified approach for skin cancer classification and segmentation by utilising ViT and Gaussian Splats. GS - TransUNet challenges the norm by implementing a parallel mask rendering technique that uses the probabilistic properties of 2D Gaussian splatting to enhance segmentation mask prediction. The model's architecture leverages the long-range dependencies handled by transformers and the detailed localization afforded by Gaussian splatting, allowing for refined interpretation of dermoscopic images. Through this dual-task method, GS - TransUNet model represents a significant advancement in automated dermatological diagnosis. The model achieves higher accuracy and computational efficiency by integrating segmentation and classification tasks than traditional methods. This innovative approach sets new benchmarks in

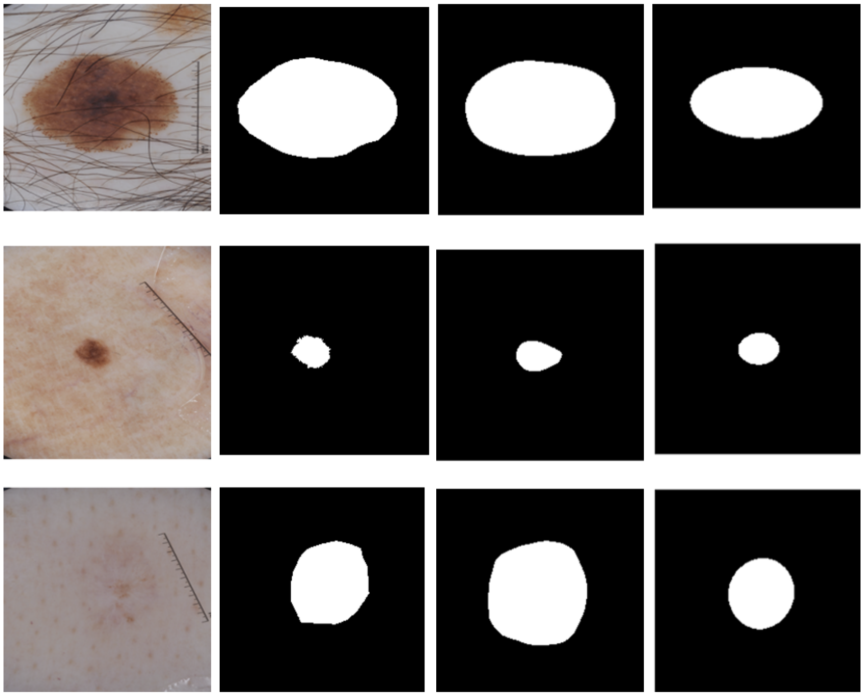

(a) Skin Image (b) Ground Truth (c) MT-TransUNet (d) Ours

Figure 5: Qualitative results of our network (GS - TransUNet) and MT-TransUNet with each row showing a different test case.

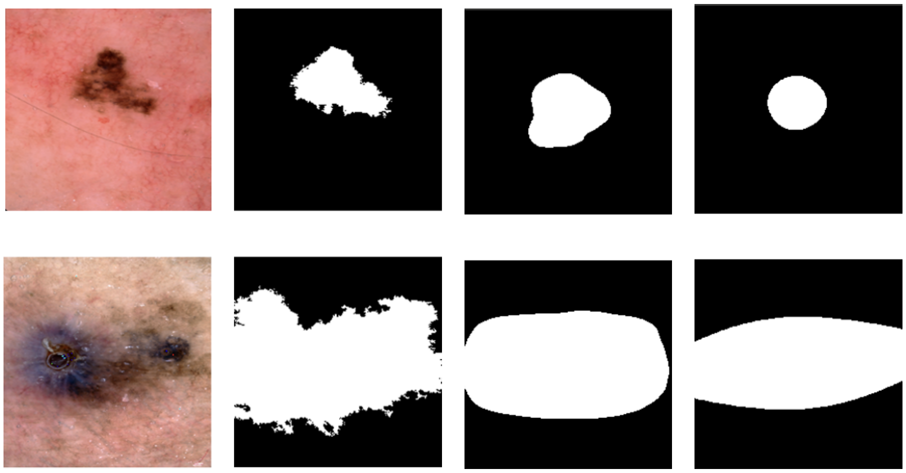

(a) Skin Image (b) Ground Truth (c) MT-TransUNet (d) Ours

Figure 6: Failure Cases of our network (GS - TransUNet) and MT-TransUNet for different test cases.

the field and underscores the potential of integrating deep learning techniques for improved patient outcomes in skin cancer detection.

Since Gaussian Splats generate solely elliptical masks, our technique will fail when the ground truth mask is not spherical. However, that is not true for skin cancer segmentation, as most skin blobs are circular. While our current work focuses on developing a dual-task approach to skin cancer analysis, future work can focus on generalizing our approach for other scenarios and potentially integrating 3D Gaussian Splats in MRI segmentation.

REFERENCES

[1] Li, H., He, X., Zhou, F., Yu, Z., Ni, D., Chen, S., Wang, T., and Lei, B., "Dense deconvolutional network for skin lesion segmentation," *IEEE journal of biomedical and health informatics* **23**(2), 527–537 (2018).

[2] Pham, C. T., Luong, M. C., Hoang, D. V., and Doucet, A., "Ai outperformed every dermatologist: Improved dermoscopic melanoma diagnosis through customizing batch logic and loss function in an optimized deep cnn architecture," (2020).

[3] Reiter, O., Rotemberg, V., Kose, K., and Halpern, A. C., "Artificial intelligence in skin cancer," *Curr. Dermatol. Rep.* **8**, 133–140 (Sept. 2019).

[4] Celebi, M. E., Kingravi, H. A., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., Moss, R. H., Malters, J. M., Grichnik, J. M., Marghoob, A. A., Rabinovitz, H. S., and Menzies, S. W., "Border detection in dermoscopy images using statistical region merging," *Skin Res. Technol.* **14**, 347–353 (Aug. 2008).

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020).

[6] Chen, J., Chen, J., Zhou, Z., Li, B., Yuille, A., and Lu, Y., "Mt-transunet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification," (2021).

[7] Ronneberger, O., Fischer, P., and 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*], 234–241, Springer (2015).

[8] Luo, X., Chen, J., Song, T., and Wang, G., "Semi-supervised medical image segmentation through dual-task consistency," in [*Proceedings of the AAAI conference on artificial intelligence*], **35**(10), 8801–8809 (2021).

[9] Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., and Halpern, A., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," (2018).

[10] Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., and Rozeira, J., "Ph 2-a dermoscopic image database for research and benchmarking," in [*2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*], 5437–5440, IEEE (2013).

[11] Hemalatha, R., Thamizhvani, T., Dhivya, A. J. A., Joseph, J. E., Babu, B., and Chandrasekaran, R., "Active contour based segmentation techniques for medical image analysis," *Medical and Biological Image Analysis* **4**(17), 2 (2018).

[12] Ravichandran, K. and Ananthi, B., "Color skin segmentation using k-means cluster," *International Journal of Computational and Applied Mathematics* **4**(2), 153–158 (2009).

[13] Yogarajah, P., Condell, J., Curran, K., Cheddad, A., and McKevitt, P., "A dynamic threshold approach for skin segmentation in color images," in [*2010 IEEE International Conference on Image Processing*], 2225–2228, IEEE (2010).

[14] Nasr-Esfahani, E., Rafiei, S., Jafari, M. H., Karimi, N., Wrobel, J. S., Samavi, S., and Soroushmehr, S. R., "Dense pooling layers in fully convolutional network for skin lesion segmentation," *Computerized Medical Imaging and Graphics* **78**, 101658 (2019).

[15] Mirikharaji, Z. and Hamarneh, G., "Star shape prior in fully convolutional networks for skin lesion segmentation," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*], 737–745, Springer (2018).

[16] Zhou, X., Wang, D., and Krähenbühl, P., "Objects as points," *arXiv preprint arXiv:1904.07850* (2019).

[17] Hagerty, J. R., Stanley, R. J., Almubarak, H. A., Lama, N., Kasmi, R., Guo, P., Drugge, R. J., Rabinovitz, H. S., Oliviero, M., and Stoecker, W. V., "Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images," *IEEE journal of biomedical and health informatics* **23**(4), 1385–1391 (2019).

[18] Zhang, J., Xie, Y., Wu, Q., and Xia, Y., "Skin lesion classification in dermoscopy images using synergic deep learning," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*], 12–20, Springer (2018).

[19] Xie, Y., Zhang, J., Xia, Y., and Shen, C., "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE transactions on medical imaging* **39**(7), 2482–2493 (2020).

[20] Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P.-A., "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE transactions on medical imaging* **36**(4), 994–1004 (2016).

[21] Zamir, A. R., Sax, A., Cheerla, N., Suri, R., Cao, Z., Malik, J., and Guibas, L. J., "Robust learning through cross-task consistency," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 11197–11206 (2020).

[22] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems* **30** (2017).

[24] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of the IEEE international conference on computer vision*], 2980–2988 (2017).

[25] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* , 240–248 (2017).

[26] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[27] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," (2017).

[28] Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., and Xu, Z., "Regnet: Self-regulated network for image classification," (2021).

[29] Tan, M. and Le, Q. V., "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR* **abs/1905.11946** (2019).

[30] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B., "Support vector machines," *IEEE Intelligent Systems and their applications* **13**(4), 18–28 (1998).

[31] Schapire, R. E., "Explaining adaboost," in [*Empirical inference: festschrift in honor of vladimir N. Vapnik*], 37–52, Springer (2013).

[32] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," in [*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*], 785–794 (2016).

APPENDIX

A Encoder and Decoder Architectures

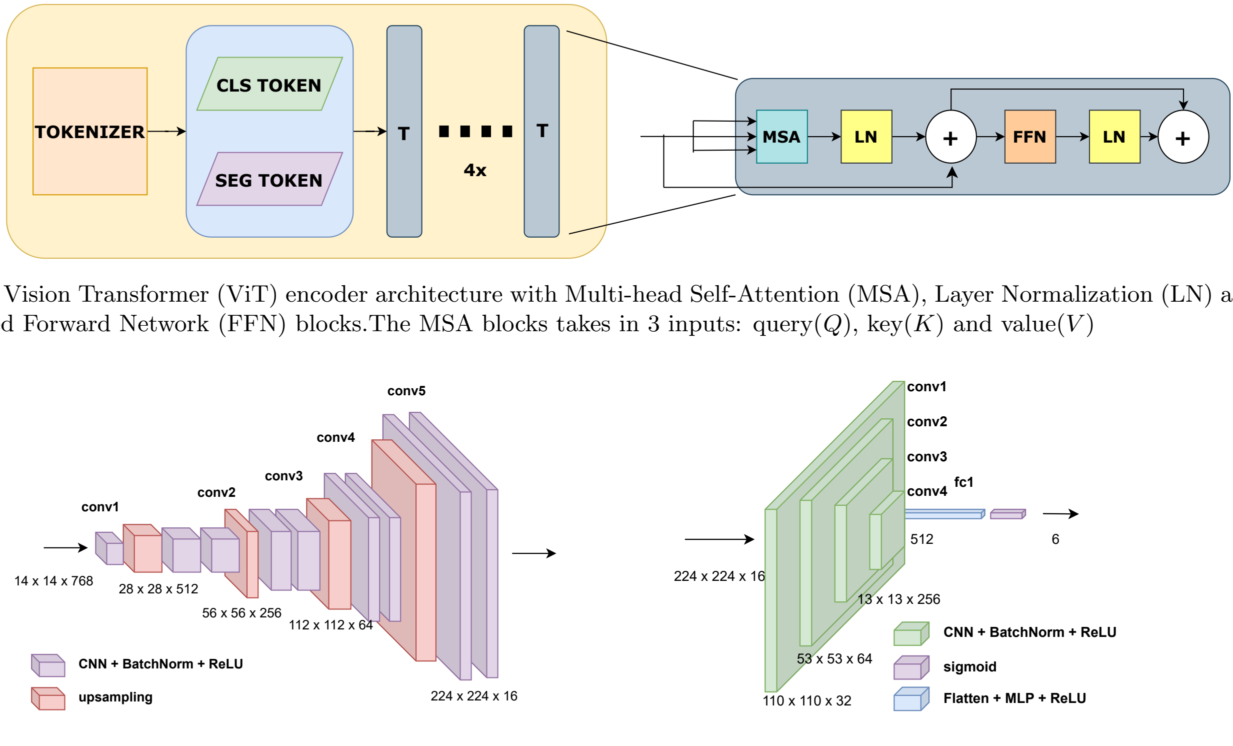
The detailed architecture for ViT encoder is given in Fig. 7a where the the image is passed through the tokenizer to obtain segmentation (local) tokens and the class (global) token is a learned embedding. The tokenizer generates a token(feature vector) for a given patch of 14 × 14 image incase of our ViT-B/14 architecture. The tokens are concatenated along with sinusoidal position embeddings and passed through the a sequence of 4 transformer blocks to get the feature representation for a given image as shown in Fig. 1.

For obtaining the segmentation features, we employ a CNN decoder as shown in Fig. 7b, to upsample the segmentation tokens back into the input resolution of 224 × 224. These features are used to get the level set function through a single CNN block of stride 1 and output channels 1 and the Gaussian Splat features using the CNN Network shown in Fig. 7c to get feature vector of size 6 consisting of position, scale and rotation values.

B Analysis on Statistical Models

We perform skin cancer classification using SVM,[30] AdaBoost,[31] XGBoost[32] and Logistic Regression and report the top 6 best performing models in Table 3. We either directly pass the image data into the classifier or perform dimensionality reduction using Principal Component Analysis(PCA), Linear Discriminant Analysis(LDA), or pre-trained neural networks such as RegNet,[28] EffNet[29] and ViT.[5] Using these features, the classifiers are trained on the training set and their performance is evaluated on the testset similar to other methods in the section 3.3.

The performance of statistical methods are not up to the mark and they fail especially for melanoma classification as they overfit for the more frequently occurring keratosis class. Therefore, these models are neither reliable nor robust for critical skin cancer diagnosis.



(a) Vision Transformer (ViT) encoder architecture with Multi-head Self-Attention (MSA), Layer Normalization (LN) and Feed Forward Network (FFN) blocks.The MSA blocks takes in 3 inputs: query(Q), key(K) and value(V)

(b) CNN Decoder for Segmentation Features from the segmentation tokens of transformer.

(c) CNN Network for Gaussian Splat features from the segmentation features.

Figure 7: Architecture of the ViT encoder and CNN decoder for segmentation features and Gaussian Splat features.

Models	**Melanoma Classification**					**Keratosis Classification**					**Average**
	M_AC	**M_AUC**	**M_AP**	**M_SE**	**M_SP**	**K_AC**	**K_AUC**	**K_AP**	**K_SE**	**K_SP**	**Accuracy**
SVM	0.800	0.657	0.295	0.000	1.000	0.767	0.834	0.713	0.167	1.000	0.567
AdaBoost	0.787	0.589	0.268	0.000	0.983	0.773	0.731	0.592	0.262	0.972	0.573
XGBoost + PCA	0.820	0.699	0.381	0.167	0.983	0.767	0.866	0.698	0.333	0.935	0.613
SVM + ViT	0.813	0.800	0.538	0.100	0.992	0.853	0.931	0.867	0.500	0.991	0.673
XGBoost + EffNet	0.820	0.779	0.518	0.233	0.967	0.840	0.898	0.813	0.500	0.972	0.687
Logistic_Reg + ViT	0.827	0.809	0.567	0.367	0.942	0.840	0.916	0.832	0.762	0.870	0.713

Table 3: Melanoma and Keratosis Classification Results for statistical models. AC, AUC, AP, SE, and SP refer to Accuracy, Area Under Curve, Average Precision, Sensitivity, and Specificity respectively.