Anastasia at MEDIQA-MAGIC 2025: A Multi-Approach Segmentation Framework with Extensive Augmentation

Tung Thanh Le[1,2], Tri Minh Ngo[1,2], Khoi Dinh Nguyen[1,2], Trung Hieu Dang[1,2], Trong Hoang Pham[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This study presents our approach for the ImageCLEFmedical 2025 Dermatological Segmentation task, focusing on the impact of data augmentation strategies for medical image segmentation. Our objective is to enhance segmentation accuracy through the combination of diverse augmentation techniques and post-processing refinements. We experimented with 20 augmentation techniques across three categories: geometric, photometric, and noise/artifact. Initially, we trained and evaluated various segmentation architectures, including pure CNNs, ViT-based models, and hybrid designs. Based on validation performance, we selected TransUNet with a hybrid ResNet-50 and ViT-B/16 backbone as the final model. This model was trained using datasets augmented by individual and full transformation strategies. We also experimented with MedSAM as a post-processing refinement, but it was applied only on predictions from the model trained on unaugmented data, resulting in no improvement in score. Experimental results showed that training with the full augmentation suite significantly improved segmentation outcomes, especially in challenging regions. These improvements were consistent across both validation and test datasets. Our findings demonstrate that integrating a hybrid CNN-transformer model with comprehensive augmentation creates a robust pipeline for dermatological image segmentation in clinical settings. Our method ranked first (Top-1) on the final leaderboard of the MEDIQA–MAGIC 2025 Segmentation Subtask.

Keywords

ImageCLEF 2025, Dermatological Segmentation, Data Augmentation, Geometric Transformations, Photometric Adjustments, Noise Artifacts.

1. Introduction

The MEDIQA–MAGIC 2025 challenge at ImageCLEF presents complementary subtasks of dermatological image analysis, including semantic segmentation of problem regions and closed-ended question answering [1]. The DermaVQA-DAS dataset, introduced for these subtasks, comprises patient-generated dermoscopic photographs paired with binary masks and multiple-choice annotations [2].

Accurate delineation of skin lesions is critical for diagnosis and treatment planning but remains challenging due to limited annotated data, high intra-class variability, and imaging artifacts such as hair occlusion and varying illumination [3, 4, 5].

Early encoder–decoder architectures like U-Net and its nested variant U-Net++ exploit multiscale feature fusion via skip connections to capture both local and global context [6]. Attention U-Net further integrates trainable attention gates to focus on relevant regions within the image [7]. The nnU-Net framework automates configuration and hyperparameter tuning of U-Net pipelines, achieving state-of-the-art performance across diverse biomedical segmentation benchmarks [8, 9].

Transformer-based hybrids such as TransUNet combine convolutional encoders with global self-attention to improve boundary delineation [10], while models like Swin UNETR leverage shifted-window attention for efficient multiresolution feature extraction [11]. Foundation models adapted to medical imaging, notably MedSAM, offer promptable zero-shot refinement that sharpens output masks with minimal additional training [5].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
tunglethanh0222@gmail.com (T. T. Le); 23521640@gm.uit.edu.vn (T. M. Ngo); 23520774@gm.uit.edu.vn (K. D. Nguyen); 23521672@gm.uit.edu.vn (T. H. Dang); 23521665@gm.uit.edu.vn (T. H. Pham); thienntb@uit.edu.vn (T. B. Nguyen-Tat)
0000-0002-4809-7126 (T. B. Nguyen-Tat)

Data augmentation remains indispensable for mitigating overfitting on small medical datasets [12]. In addition to classical geometric and photometric transforms, learned augmentation policies such as AutoAugment optimize transformation strategies via reinforcement learning [13], while RandAugment [14] and AugMix [15] have demonstrated strong empirical performance in vision tasks.

Recent advances include self-supervised pre-training of Swin Transformers for 3D medical image analysis, which improves convergence and generalization [16], as well as split-attention U-Net variants that deliver compact models without sacrificing accuracy [17].

In this study, we focus on Subtask 1 (dermatological segmentation) of MEDIQA–MAGIC 2025, proposing a robust pipeline that integrates: (a) comprehensive augmentation across geometric, photometric, and noise/artifact families; (b) a hybrid ResNet-50–ViT-B/16 TransUNet backbone; and (c) post-processing refinement via MedSAM.

2. Background and Related Works

2.1. Dermatological Image Segmentation

Dermatological image segmentation is a crucial but challenging task in computer-aided diagnosis and population-scale screening for skin diseases [18, 19]. Key difficulties arise from the high intra-class variability of lesions, ambiguous boundaries with normal skin, and imaging artifacts such as hair, occlusion, or varying illumination [4, 5]. Public datasets are often small, imbalanced, and costly to annotate at the pixel level, limiting model generalization and causing overfitting when models are deployed on data from diverse clinical settings [18, 19, 5]. These limitations have driven continuous innovation in both dataset design and segmentation architectures for dermatology.

2.2. Data Augmentation in Medical Segmentation

Data augmentation is a central solution for improving generalizability and robustness in medical image segmentation [12]. Geometric transformations (e.g., flipping, rotation, cropping) help address viewpoint and scale variations, while photometric changes (e.g., contrast, brightness, color jitter) account for device- and skin-related differences [20, 12, 6]. The addition of noise and artifact (e.g., Gaussian noise, motion blur, compression artifacts) is essential to simulate real-world imperfections. Studies have shown that systematic, multi-operator augmentation pipelines—notably those implemented in the Albumentations library [20, 12]—can significantly increase segmentation accuracy, even in low-data or high-imbalance regimes. Effective augmentation is now recognized as a critical factor in nearly all state-of-the-art solutions for medical image segmentation [12, 6].

2.3. Deep Learning Architectures for Medical Segmentation

Early breakthroughs in medical segmentation leveraged convolutional neural networks (CNNs), such as U-Net and its extensions [3, 6], which were designed to capture local context and spatial structure. However, CNNs often struggle with long-range dependencies and variable lesion shapes. The introduction of Vision Transformers (ViT) [21] and hybrid models such as TransUNet [10] has significantly advanced the field by combining global attention mechanisms with local feature encoding. More recently, foundation models like MedSAM [5], pretrained on millions of medical images across modalities, have demonstrated strong zero-shot and few-shot generalization for segmentation tasks.

In addition, recent research by Nguyen-Tat et al. has proposed hybrid network architectures and edge-aware attention mechanisms, exemplified by QMaxViT-UNet+ [22], as well as approaches that integrate U-Net, attention mechanisms, and transformers for brain MRI tumor segmentation [23]. Further, these authors have systematically evaluated the effectiveness of preprocessing and deep learning techniques across multiple imaging modalities [24].

Recent benchmarks and challenge results confirm that optimal performance requires both advanced model architectures (e.g., CNN–Transformer hybrids, foundation models) and comprehensive, diverse

Table 1
Dataset splits and file formats for Subtask 1. Each image has four corresponding masks from different annotators.

Split	Number of Images	Number of Masks
Train	2,474	7,448
Validation	157	472
Test	314	944

File format:

- Image files: `.png` or `.jpg`, named as `IMG_{ENCOUNTERID}_{IMAGEID}.png`
- Mask files: `.tiff`, named as `IMG_{ENCOUNTERID}_{IMAGEID}_mask_{ANNOTATOR#}.tiff`

augmentation pipelines [19, 5, 10].

3. Task and Dataset Descriptions

3.1. Task Descriptions

The **2nd MEDIQA–MAGIC 2025** shared task extends the 2024 multimodal dermatology benchmark and targets automatic response generation from combined clinical narratives and dermoscopic photographs.[1] Each encounter consists of (i) a clinical narrative context and (ii) one or more color images of the skin lesion(s).[1] Two complementary subtasks are defined:

1. **Segmentation of dermatological problem regions.** Given the image(s) and clinical history, systems must generate segmentations of the regions of interest that correspond to the described dermatological problem.[3]
2. **Closed-ended question answering.** For the same encounter, systems are given a dermatological query, its accompanying images, and a closed-ended question with multiple-choice options; the objective is to select the single correct answer.[25]

Segmentation performance is assessed with region-overlap metrics such as Intersection-over-Union (Jaccard)[26], whereas the closed-question subtask is evaluated using accuracy and the macro-averaged F_1 score.[27]

In this study, we focus exclusively on Subtask 1—segmentation of dermatological regions of interest.

3.2. Dataset Information

For Subtask 1, we use the **MEDIQA–MAGIC 2025** segmentation dataset [1, 2]. It includes original skin images with binary masks indicating affected regions [2]. The dataset is divided into three disjoint splits, provided by the organisers and adopted without modification. Each image is annotated by four distinct annotators: `ann0`, `ann1`, `ann2`, and `ann3` [28].

4. Methodology

4.1. Augmentation Strategy

State-of-the-art segmentation networks require large, diverse, and balanced datasets to achieve robust performance.[29] However, compiling pixel-level annotations for dermatology is slow, costly, and constrained by ethical considerations. The MEDIQA–MAGIC 2025 training set provides only 2,474 dermoscopic photographs with corresponding binary lesion masks, which is significantly smaller than typical deep learning datasets and exhibits pronounced imbalances across lesion types and anatomical

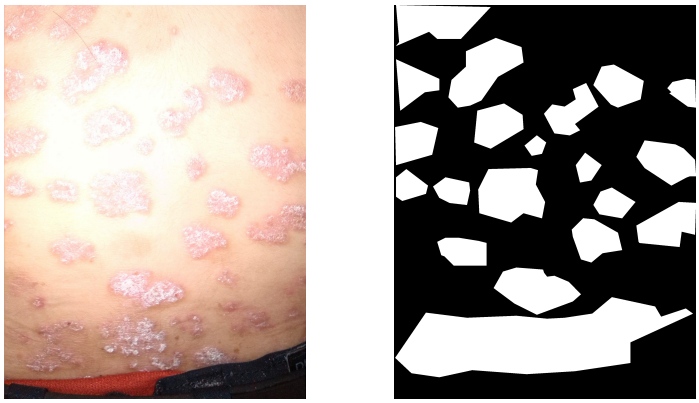

Figure 1: Example data: `IMG_ENC00001_00001.png`, `IMG_ENC00001_00001_mask_ann0.tiff`

sites. To address these challenges, aggressive data augmentation is a cornerstone of our pipeline, strategically designed along three complementary directions—geometric, photometric, and noise & artifact—to enhance model generalization, robustness, and real-world applicability.[12, 30] These directions are chosen for the following reasons [12, 30]:

- **Geometric Augmentations**: Dermatological images are captured from varied angles, scales, and orientations due to differences in imaging equipment, patient positioning, and anatomical locations. Geometric augmentations simulate these variations by altering the spatial configuration of images, enabling the model to learn invariance to viewpoint, scale, and orientation changes. This is critical for ensuring the model can accurately segment lesions regardless of how the image is framed or oriented in real-world scenarios.[12, 30]
- **Photometric Augmentations**: Dermoscopic images are subject to variations in lighting conditions, camera settings, and skin tones, which can significantly affect pixel intensities and colour distributions. Photometric augmentations introduce controlled changes in colour, brightness, and contrast to mimic these real-world variations, ensuring the model remains robust to differences in illumination and imaging hardware. This is particularly important for generalizing across diverse patient populations and clinical settings.[12, 30]
- **Noise & Artifact Augmentations**: Real-world dermoscopic images often contain imperfections such as motion blur, sensor noise, or compression artifacts, which can degrade segmentation performance if the model is not trained to handle them. Noise and artifact augmentations inject realistic acquisition degradations, training the model to ignore irrelevant distortions and focus on the underlying lesion features. This enhances the model's resilience to suboptimal imaging conditions encountered in clinical practice.[12, 30]

To realise these objectives, we employ the **Albumentations** library.[20] A total of **20** augmentation operators are orchestrated and grouped into three semantic families: 18 geometric, 6 photometric, and 6 noise & artifact transforms.[12] This distribution prioritises geometric augmentations with 18 transforms, reflecting the dominant challenge of spatial variability in dermoscopic imaging due to diverse capture angles, scales, and orientations—factors that critically impact lesion segmentation accuracy in clinical settings. The higher number of geometric operators ensures comprehensive coverage of these spatial variations, which are more prevalent and complex than photometric or noise-related issues. In contrast, photometric and noise & artifact directions are allocated 6 transforms each, as their variability can be effectively addressed with a smaller, targeted set of operators, supplemented by composite pipelines that enrich sample diversity without redundancy. This strategic allocation optimises computational efficiency while maximising the model's ability to generalise across real-world imaging conditions.

- **Geometric** (18 transforms): These manipulate spatial configuration to build invariance against viewpoint and scale changes: HorizontalFlip, VerticalFlip, Transpose, CenterCrop, Ran-

domRotate90, RandomSizedCrop, RandomSizedCrop_0.1, PadIfNeeded, ElasticTransform, GridDistortion, OpticalDistortion, AdvancedAugmentation, AdvancedAugmentation_2, CompositeAug, Medium, comprehensive_augmentation, crop_80_comprehensive, and color_transform.

- **Photometric** (6 transforms): Colour and intensity variations mimic different lighting conditions and camera responses: AdvancedAugmentation, AdvancedAugmentation_2, Medium_add_non_spatial_transformations, comprehensive_augmentation, crop_80_comprehensive, and color_transform.
- **Noise & Artefact** (6 transforms): These inject realistic acquisition degradations that the model must learn to ignore: AdditiveNoise, AdvancedAugmentation, AdvancedAugmentation_2, Medium_add_non_spatial_transformations, comprehensive_augmentation, and crop_80_comprehensive.

Composite pipelines: The transforms AdvancedAugmentation, AdvancedAugmentation_2, Medium, Medium_add_non_spatial_transformations, CompositeAug, comprehensive_augmentation, crop_80_comprehensive, and color_transform are higher-level pipelines that compose multiple elementary operations, thereby substantially enriching sample diversity.

While our current pipeline relies on majority voting to aggregate the four annotator masks into a single ground truth, we acknowledge that treating each annotation as a distinct supervision signal—effectively using them as GT augmentations—could provide useful label diversity. To maintain training consistency and evaluation comparability, we did not experiment with this approach in the current study. Nevertheless, we consider it a promising direction for future exploration, particularly in modeling annotator disagreement or uncertainty.[28]

With a robust augmentation strategy in place to address the limitations of the MEDIQA–MAGIC 2025 dataset, the next critical step is selecting an appropriate model architecture that can effectively leverage this enriched data. The choice of model must balance computational efficiency with the ability to capture both local and global contextual features inherent in dermoscopic images, paving the way for the proposed TransUNet-based approach detailed in the following subsection.[10]

4.2. Training Pipeline

The following steps outline our segmentation training strategy based on the MEDIQA–MAGIC 2025 dataset:

1. **Majority Voting Label Aggregation:** Each image is associated with four binary masks from annotators `ann0`–`ann3`. These are aggregated into a single ground-truth mask using pixel-wise majority voting [28, 31].
2. **Data Augmentation by Keyword Category:** Augment the dataset using three families of transformations [12, 20]:
    - `geometric`: flipping, rotation, scaling, etc.
    - `photometric`: brightness, contrast, color jitter, etc.
    - `Noise\&Artifact`: Gaussian noise, blur, synthetic hair occlusion, etc.
3. **Training with Individual Augmentations:** For each augmentation category, we train a TransUNet model with a ResNet50–ViT-B_16 hybrid backbone to assess the impact of each transformation type [10].
4. **Training with Combined Augmentations:** We merge all augmentation categories and train TransUNet models with different backbone configurations [32, 21]:
    - ResNet50–ViT-B_16 (hybrid)
    - ViT-B_16, ViT-B_32

- ViT-L_16, ViT-L_32

The best-performing model is selected based on validation performance [10].

5. **Model Refinement via MedSAM:** We apply the MedSAM framework to refine predictions of the best model. Specifically, we fine-tune four separate MedSAM models with a ViT-B backbone on the training set, each using segmentation masks from a different annotator (ann0–ann3). During inference, bounding boxes predicted by TransUNet are used as prompts, together with the original image, to guide each MedSAM model in generating refined segmentations. [5].

The complete training pipeline is illustrated in Figure 2.

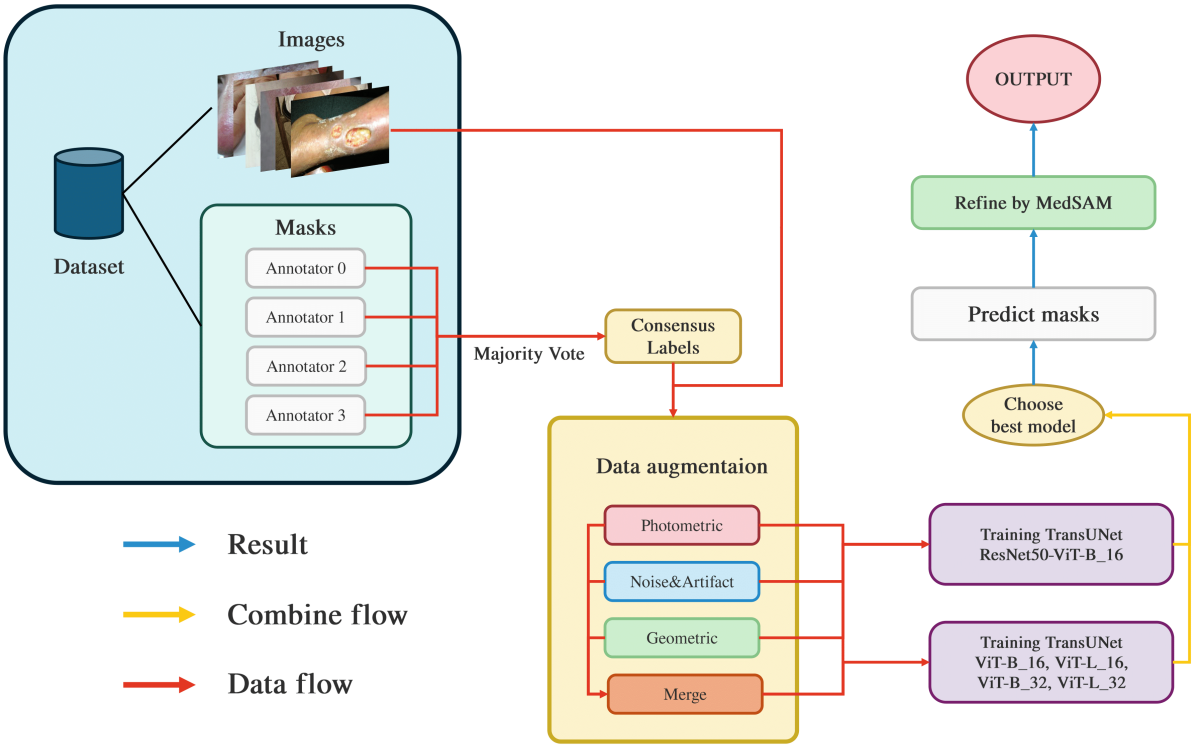

Figure 2: Overview of our segmentation training pipeline. The framework includes label aggregation, augmentation by keyword categories, model training with different backbones, and refinement using MedSAM.

5. Experimental Setup

5.1. Data Preparation

The MEDIQA-MAGIC 2025 dataset, comprising 2,474 training, 157 validation, and 314 test images, is preprocessed to support the segmentation task. Images and masks are normalised to standardise pixel intensity values[33], and multiple masks per image (from more than one annotator) are combined using majority voting[28]. The training set is augmented using the Albumentations library with 20 operators[20], organised into subdirectories (e.g., `AdditiveNoise`, `HorizontalFlip`, `AdvancedAugmentation`) to reflect geometric, photometric, and noise & artefact strategies. Additionally, a test subset of 295 images is extracted from the training set to evaluate the model during training, enhancing real-time performance assessment.

5.2. Experiment Configurations

Several experiments were conducted to evaluate the efficacy of the proposed segmentation framework for the MEDIQA-MAGIC 2025 task, focusing on the integration of various augmentation strategies and model architectures. All experiments were performed using NVIDIA Tesla P100 GPUs to ensure consistent computational performance. The configurations are detailed as follows:

- **TransUnet ResNet50–ViT-B16 with Geometric Augmentation**: The TransUnet ResNet50–ViT-B16 model[32, 21] was specifically trained with the geometric augmentation subset for 10 epochs. This configuration focused on enhancing the model's invariance to spatial variations such as rotations, flips, and crops.
- **TransUnet ResNet50–ViT-B16 with Photometric Augmentation**: The TransUnet ResNet50–ViT-B16 model underwent training with the photometric augmentation subset for 20 epochs. This setup aimed to improve robustness against variations in lighting, colour, and contrast within dermoscopic images.
- **TransUnet ResNet50–ViT-B16 with Noise & Artefact Augmentation**: The TransUnet ResNet50–ViT-B16 model was trained using the noise & artefact augmentation subset for 20 epochs. This configuration targeted the model's resilience to imperfections such as motion blur and sensor noise, common in real-world medical imaging.
- **TransUnet Variants with All Methods**: The TransUnet variants—ResNet50–ViT-B16, ViT-B_16, ViT-B_32, ViT-L_32, and ViT-L_16—were trained using all 20 augmentation methods for 8 epochs.[10] This experiment assessed the performance of different Transformer scales under a unified augmentation strategy to identify the optimal backbone architecture.

Across all configurations, the AdamW optimiser was employed[34] with a learning rate of 1×10^{-5}, a weight decay of 1×10^{-2}, and a batch size of 8, consistent with best practices for deep learning model training. The DiceLoss function (binary mode)[35] was utilised as the loss criterion to optimise segmentation performance. A learning rate scheduler (`ReduceLROnPlateau`) was applied, reducing the learning rate by a factor of 0.5 if the validation loss did not improve for 5 consecutive epochs. Mixed precision training was enabled using a gradient scaler to enhance computational efficiency on GPUs[36]. Additionally, early stopping was implemented with a patience of 10 epochs[37], halting training if the validation loss ceased to improve, ensuring optimal model convergence. These settings ensured a robust evaluation of the proposed methodology across diverse augmentation strategies and architectural paradigms.

5.3. Experimental Results

The performance of the proposed TransUNet segmentation framework was evaluated on the MEDIQA-MAGIC 2025 dataset, comprising 157 dermoscopic image segmentation instances. The primary metrics utilised were the Jaccard Index (IoU)[26] and Dice Coefficient[38], which are standard for assessing segmentation quality. The results are presented in two tables: Table 2 for validation set performance across four augmentation strategies and Table 3 for test set performance of the best model and its refined version.

The validation results indicate that the TransUNet model with the R50-ViT-B16 backbone, trained using all augmentation methods, achieved the highest performance, with a Jaccard Index of 0.6770 and a Dice Coefficient of 0.8074. This highlights the effectiveness of combining a ResNet50 backbone with a Vision Transformer (ViT-B16) and a comprehensive augmentation strategy. Among the R50-ViT-B16 configurations, the "all" augmentation approach significantly outperformed others, while the photometric augmentation yielded the lowest scores (Jaccard: 0.4137, Dice: 0.5853), likely due to insufficient augmentation diversity impacting model generalisation. Comparing ViT variants, models with a patch size of 16 (e.g., ViT-B_16, ViT-L_16) generally outperformed those with a patch size of 32 (e.g., ViT-B_32, ViT-L_32), suggesting that finer patch granularity enhances segmentation accuracy in this context.

Table 2
Validation set performance metrics for different TransUNet configurations. The best results are highlighted in bold.

Model Configuration	Jaccard (IoU)	Dice
TransUNet (R50-ViT-B16, Geometric)	0.6151	0.7617
TransUNet (R50-ViT-B16, Photometric)	0.4137	0.5853
TransUNet (R50-ViT-B16, Noise & Artefact)	0.4480	0.6187
TransUNet (R50-ViT-B16, All)	**0.6770**	**0.8074**
TransUNet (`ViT-B_16`, All)	0.6677	0.8007
TransUNet (`ViT-B_32`, All)	0.5974	0.7479
TransUNet (`ViT-L_16`, All)	0.6729	0.8045
TransUNet (`ViT-L_32`, All)	0.6238	0.7683

Table 3
Test set performance metrics for the best TransUNet configuration and its refined version.

Model Configuration	Jaccard (IoU)	Dice
Best Model (TransUNet (R50-`ViT-B_16`, All))	0.6458	0.7848
Best Model + Refine (TransUNet (R50-`ViT-B_16`, All) + MedSAM fine)	0.6113	0.7587

On the test set, the best model, TransUNet (R50-ViT-B16, All), achieved a Jaccard Index of 0.6458 and a Dice Coefficient of 0.7848, demonstrating robust generalisation to unseen data. However, the refined version of this model, which incorporated MedSAM fine-tuning, showed a slight decrease in performance (Jaccard: 0.6113, Dice: 0.7587). This suggests that the MedSAM fine-tuning approach may require further optimisation to improve test set outcomes, possibly due to overfitting or misalignment with the test distribution. These findings underscore the strength of the proposed TransUNet framework while identifying potential areas for improvement in the refinement process.

6. Conclusion and Future Works

In this paper, we presented a data augmentation pipeline[12, 20] and hybrid model strategy for medical image segmentation in the ImageCLEFmedical 2025 challenge. We began by training multiple segmentation models and selecting the best-performing architecture—TransUNet with a ResNet-50[32] and ViT-B/16[21] hybrid backbone. Training with a comprehensive augmentation set significantly improved segmentation outcomes[10]. The integration of MedSAM for post-processing refinement further elevated segmentation accuracy, particularly in lesion boundaries and varied skin textures[5].

Our results demonstrate the benefits of combining a powerful hybrid model architecture with systematic augmentation. Future directions include: (a) exploring architecture variations such as SwinUNet[11] and nnU-Net[8] under similar augmentation regimes, (b) incorporating learned augmentation policies like AutoAugment[13] for further optimization, (c) extending our approach to high-resolution dermoscopic images and cross-domain generalization[39], and (d) evaluating the performance of other SAM-based or edge-aware refinement methods[40]. (e) using individual annotator masks as separate supervision signals to better model annotation variability. [41]

These enhancements aim to create a more generalizable and scalable framework for medical image segmentation in real-world applications.

Acknowledgments

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D4-2025-04.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] W. Yim, A. Ben Abacha, N. Codella, R. A. Novoa, J. Malvehy, Overview of the mediqa-magic task at imageclef 2025: Multimodal and generative telemedicine in dermatology, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Span, 2025.

[2] W. Yim, Y. Fu, A. Ben Abacha, M. Yetisgen, N. Codella, R. A. Novoa, J. Malvehy, Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images, CoRR (2025).

[3] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[4] Y. Yuan, Y.-C. Lo, Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks, IEEE Journal of Biomedical and Health Informatics 23 (2019) 519–526. doi:10.1109/JBHI.2017.2787487.

[5] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nature Communications 15 (2024) 654. doi:10.1038/s41467-024-44824-z.

[6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2018, pp. 3–11. doi:10.1007/978-3-030-00889-5_1.

[7] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, 2018. URL: https://arxiv.org/abs/1804.03999. arXiv:1804.03999.

[8] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021) 203–211. URL: https://doi.org/10.1038/s41592-020-01008-z. doi:10.1038/s41592-020-01008-z.

[9] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, P. F. Jäger, nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, in: M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J. A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Springer Nature Switzerland, Cham, 2024, pp. 488–498. doi:10.1007/978-3-031-72114-4_47.

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021. URL: https://arxiv.org/abs/2102.04306. arXiv:2102.04306.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2022, pp. 272–284. doi:10.1007/978-3-031-08999-2_22.

[12] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (2019) 60. URL: https://doi.org/10.1186/s40537-019-0197-0. doi:10.1186/s40537-019-0197-0.

[13] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 113–123. doi:10.1109/CVPR.2019.00020.

[14] E. D. Cubuk, B. Zoph, J. Shlens, Q. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 18613–18624. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

[15] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty, 2020. URL: https://arxiv.org/abs/1912.02781. arXiv:1912.02781.

[16] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-supervised pre-training of swin transformers for 3d medical image analysis, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20698–20708. doi:10.1109/CVPR52688.2022.02007.

[17] M. Lee, J. Kim, R. EY Kim, H. G. Kim, S. W. Oh, M. K. Lee, S.-M. Wang, N.-Y. Kim, D. W. Kang, Z. Rieu, J. H. Yong, D. Kim, H. K. Lim, Split-attention u-net: A fully convolutional network for robust multi-label segmentation from brain mri, Brain Sciences 10 (2020). URL: https://www.mdpi.com/2076-3425/10/12/974. doi:10.3390/brainsci10120974.

[18] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific Data 5 (2018) 180161. URL: https://doi.org/10.1038/sdata.2018.161. doi:10.1038/sdata.2018.161.

[19] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, 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), in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 168–172. doi:10.1109/ISBI.2018.8363547.

[20] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2020). URL: https://www.mdpi.com/2078-2489/11/2/125. doi:10.3390/info11020125.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[22] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, Qmaxvit-unet+: A query-based maxvit-unet with edge enhancement for scribble-supervised segmentation of medical images, Computers in Biology and Medicine 187 (2025) 109762. URL: https://www.sciencedirect.com/science/article/pii/S001048252500112X. doi:https://doi.org/10.1016/j.compbiomed.2025.109762.

[23] T. B. Nguyen-Tat, T.-Q. T. Nguyen, H.-N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in mri images: A hybrid approach using unet, attention mechanisms, and transformers, Egyptian Informatics Journal 27 (2024) 100528. URL: https://www.sciencedirect.com/science/article/pii/S1110866524000914. doi:https://doi.org/10.1016/j.eij.2024.100528.

[24] T. B. Nguyen-Tat, T. Q. Hung, P. T. Nam, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, Alexandria Engineering Journal 119 (2025) 558–586. URL: https://www.sciencedirect.com/science/article/pii/S1110016825001176. doi:https://doi.org/10.1016/j.aej.2025.01.090.

[25] W.-w. Yim, A. Ben Abacha, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR-WS.org, Bordeaux, France, 2024. URL: https://ceur-ws.org/Vol-3740/paper-133.pdf.

[26] P. Jaccard, The distribution of the flora in the alpine zone., New Phytologist 11 (1912) 37–50. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x. doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[27] C. J. V. Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, 1979. URL: https://www.bibsonomy.org/bibtex/20bc6015eaa45ecb0c5c9186f68f77562/machinelearning.

[28] A. P. De Rosa, M. Benedetto, S. Tagliaferri, F. Bardozzo, A. D'Ambrosio, A. Bisecco, A. Gallo, M. Cirillo, R. Tagliaferri, F. Esposito, Consensus of algorithms for lesion segmentation in brain mri studies of multiple sclerosis, Scientific Reports 14 (2024) 21348. URL: https://doi.org/10.1038/s41598-024-72649-9. doi:10.1038/s41598-024-72649-9.

[29] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135. doi:https://doi.org/10.1016/j.media.2017.07.005.

[30] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation techniques for deep learning, Pattern Recognition 137 (2023) 109347. URL: https://www.sciencedirect.com/science/article/pii/S0031320323000481. doi:https://doi.org/10.1016/j.patcog.2023.109347.

[31] S. Warfield, K. Zou, W. Wells, Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation, IEEE Transactions on Medical Imaging 23 (2004) 903–921. doi:10.1109/TMI.2004.828354.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[33] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 448–456. URL: https://dl.acm.org/doi/10.5555/3045118.3045167.

[34] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[35] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571. doi:10.1109/3DV.2016.79.

[36] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed precision training, CoRR abs/1710.03740 (2017). URL: http://arxiv.org/abs/1710.03740. arXiv:1710.03740.

[37] L. Prechelt, Early Stopping — But When?, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 53–67. URL: https://doi.org/10.1007/978-3-642-35289-8_5. doi:10.1007/978-3-642-35289-8_5.

[38] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302. URL: http://www.jstor.org/stable/1932409.

[39] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, Z. Xu, Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation, IEEE Transactions on Medical Imaging 39 (2020) 2531–2540. doi:10.1109/TMI.2020.2973595.

[40] Q. Hu, Y. Wei, X. Li, C. Wang, J. Li, Y. Wang, Ea-net: Edge-aware network for brain structure segmentation via decoupled high and low frequency features, Computers in Biology and Medicine 150 (2022) 106139. URL: https://www.sciencedirect.com/science/article/pii/S0010482522008472. doi:https://doi.org/10.1016/j.compbiomed.2022.106139.

[41] L. Zhang, H. Wang, W. Li, Y. Zhang, Y. Gao, Learning from multiple annotators: A survey on models and applications, Pattern Recognition 138 (2023) 109423. URL: https://www.sciencedirect.com/science/article/pii/S0031320323001012. doi:10.1016/j.patcog.2023.109423.