Benchmark of U-Net Architectures in Medical Image Segmentation

Victor Contreras, Diana Castilleja, Ei Ei Nyein Chan, Dr. Dong-Chul Kim

University of Texas Rio Grande Valley

BACKGROUND

Medical image segmentation is a critical task in healthcare applications, enabling precise localization and identification of anatomical structures. Deep learning models, particularly the U-Net architecture, have become the foundation for many segmentation methods. In this study, we evaluate four U-Net-based architectures—TransUNet, SwinUNet, VMUNet, and UNet++—which integrate different advancements such as transformers, attention mechanisms, and nested architectures. These models were benchmarked on two publicly available datasets representing different anatomical regions. We compare their segmentation performance using standard evaluation metrics such as F1 score, Dice coefficient, recall, and precision.

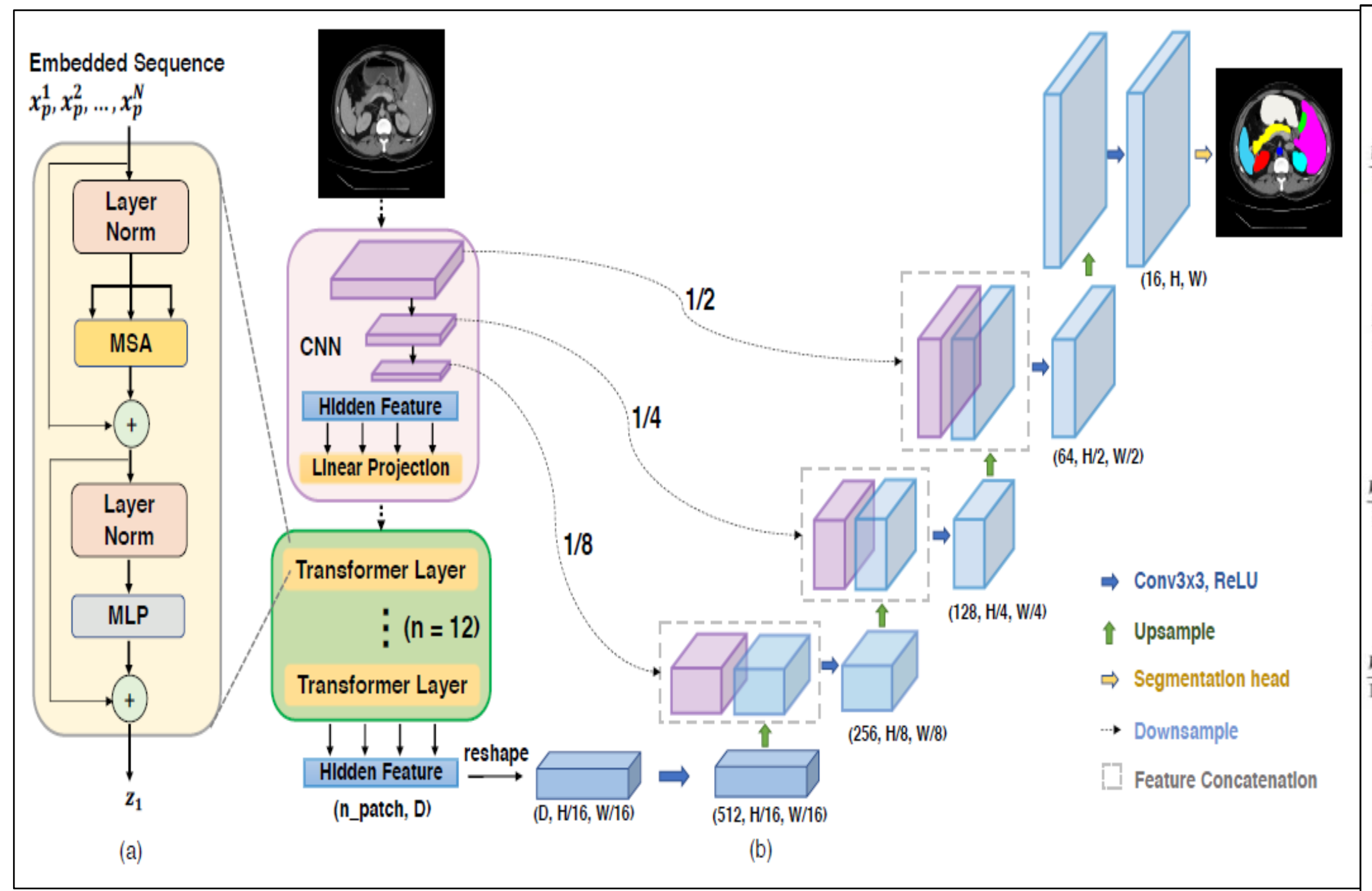
PURPOSE

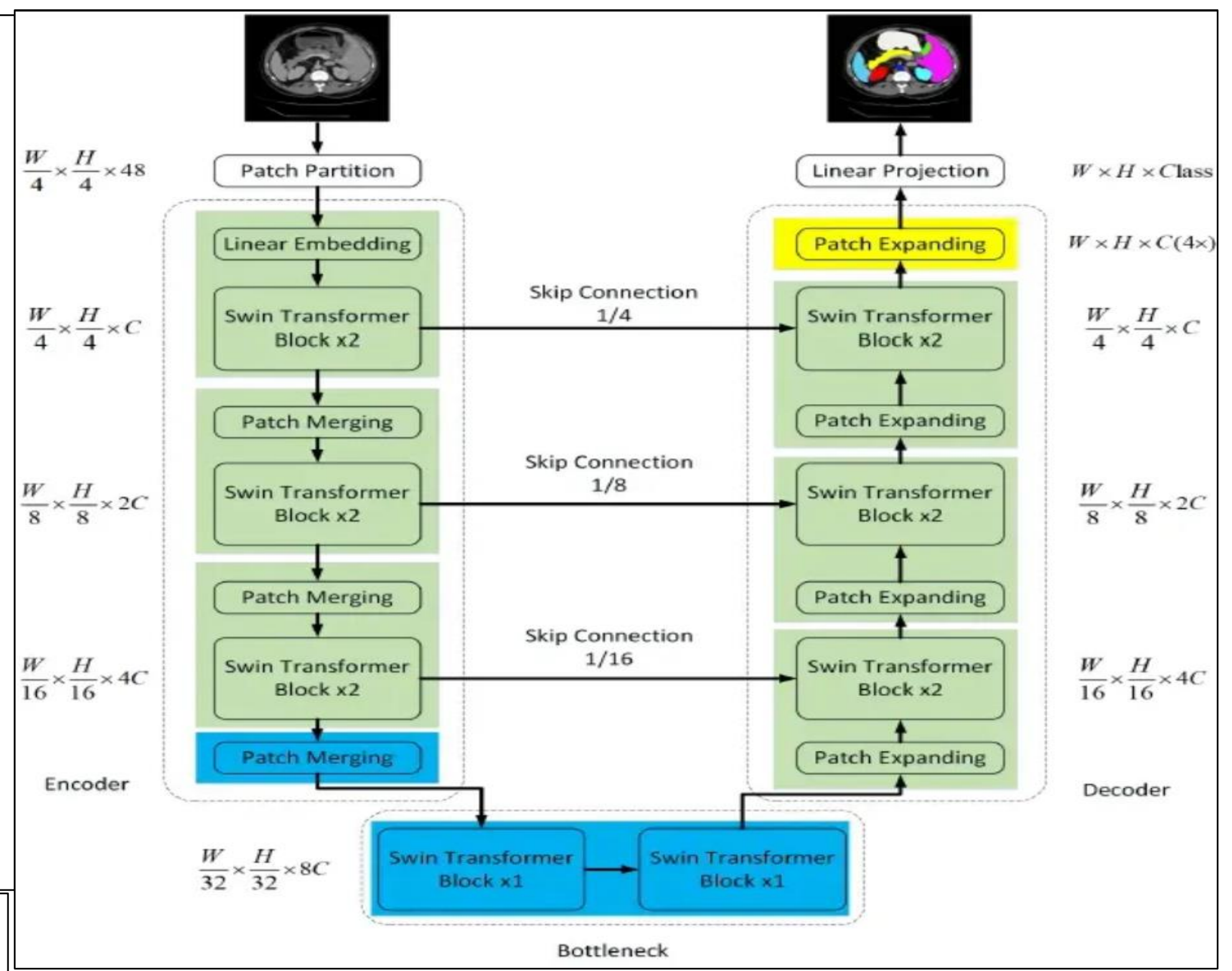
The purpose of this study is to benchmark and compare the performance of four different U-Net-based deep learning architectures—TransUNet, SwinUNet, VMUNet, and UNet++ —to evaluate how various architectural innovations impact segmentation accuracy in medical image tasks. Each model introduces distinct enhancements to the original U-Net framework: TransUNet incorporates transformers for better handling of long-range dependencies, SwinUNet leverages a window-based approach for efficient global context aggregation, VMUNet uses state space models for improved feature representation, and UNet++ introduces broader skip connections and deep, nested architectures for more precise segmentation. This study evaluates these architectures across two publicly available datasets: ISIC2018 for skin lesion segmentation and PolyPGen (from MedSegBench) for polyp segmentation.

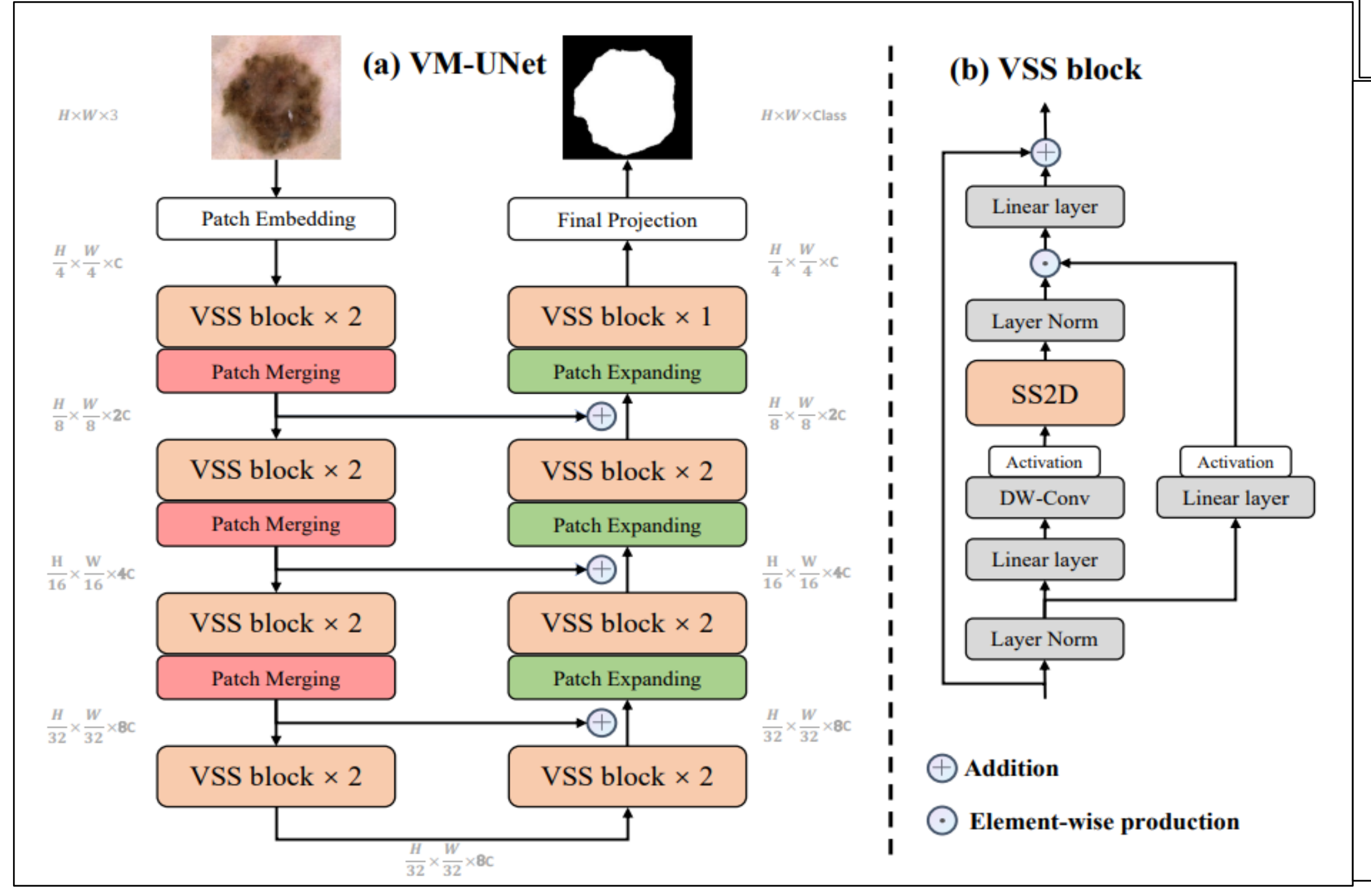
Hyperparameters

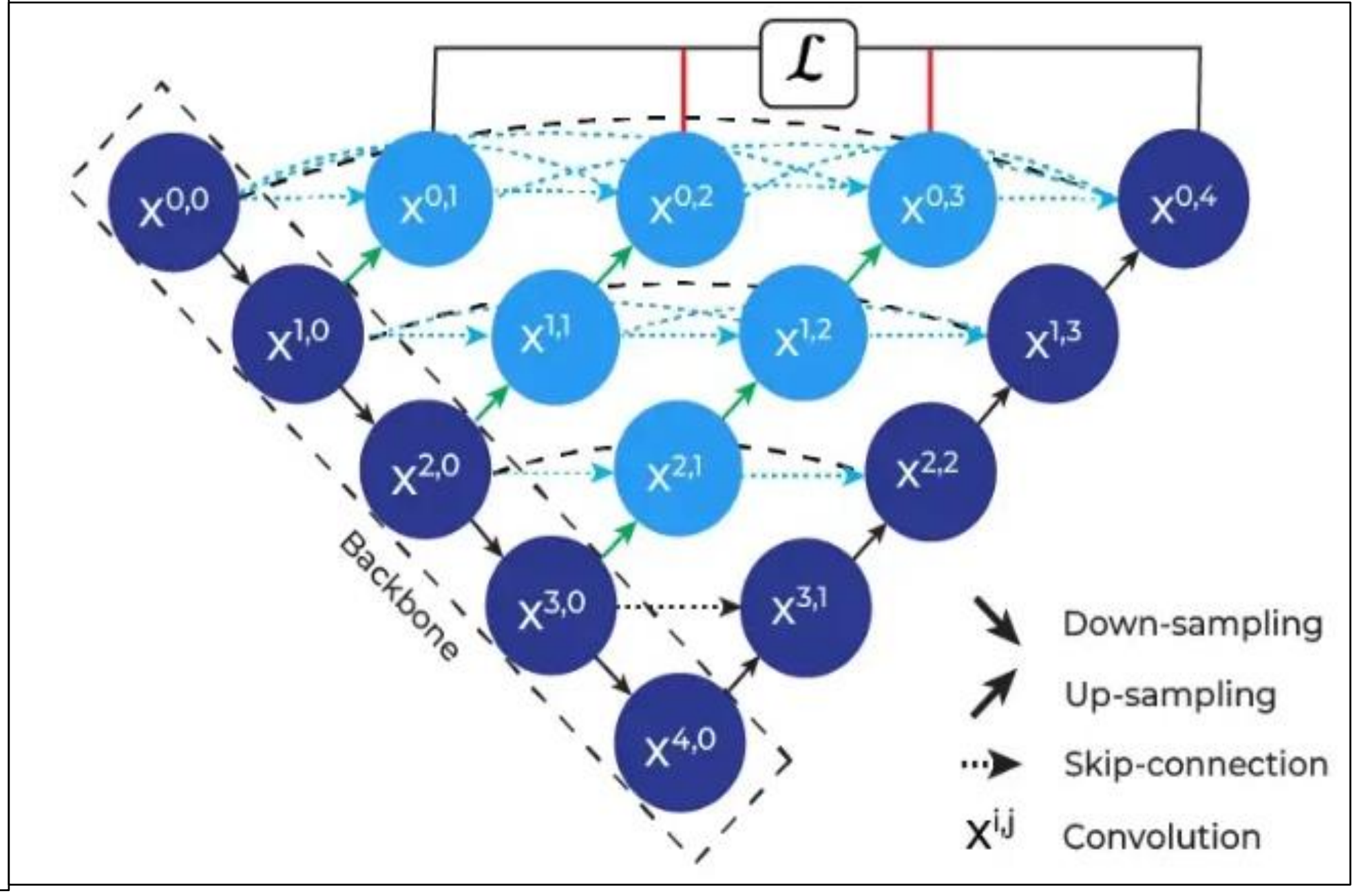
	TransUNet	SwinUNet	VMUnet	UNet++
Optimizer	SGD	SGD	AdamW	Adam
Scheduler	Polynomial Decay	Polynomial Decay	CosineAnnealingLR	ReduceLROnPlateau
Batch Size	24	24	64	16
Loss Function	0.5 * CEL + 0.5 * Dice_Loss	0.4 * CEL + 0.6 * DiceLoss	BCEDiceloss (wb=0.4, wd=0.6)	0.5 * DiceLoss + 0.5 * CEL
Epochs	150	150	300	150

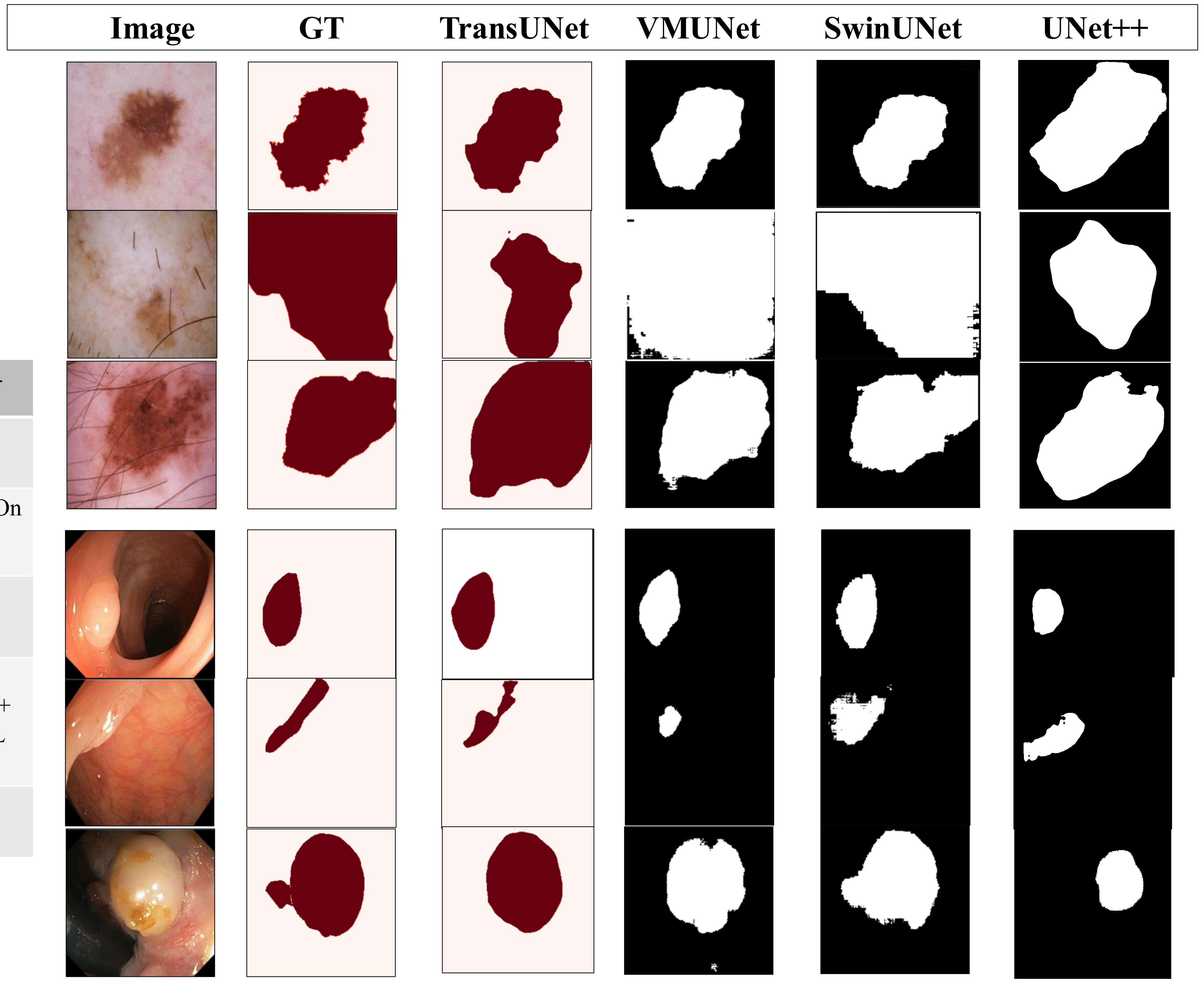
METHODOLOGY AND EXPERIMENTATION




Patch Partition
Linear Embedding
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x2
Patch Merging
Encoder
Swin Transformer Block x1
Swin Transformer Block x1
Bottleneck
Skip Connection 1/4
Skip Connection 1/8
Skip Connection 1/16
Linear Projection
Patch Expanding
Swin Transformer Block x2
Patch Expanding
Swin Transformer Block x2
Patch Expanding
Swin Transformer Block x2
Patch Expanding
Decoder
$\frac{W}{4}\times\frac{H}{4}\times 48$
$\frac{W}{4}\times\frac{H}{4}\times C$
$\frac{W}{8}\times\frac{H}{8}\times 2C$
$\frac{W}{16}\times\frac{H}{16}\times 4C$
$\frac{W}{32}\times\frac{H}{32}\times 8C$
$W\times H\times Class$
$W\times H\times C(4\times)$







RESULTS

ISIC 2018 Dataset

Metrics	Accuracy	Recall	Specificity	F1/ Dice Coefficient	Mean IOU	Precision
TransUNet	0.911	0.837	0.9393	0.838	0.721	0.840
SwinUNet	**0.982**	**0.986**	**0.981**	**0.958**	**0.933**	**0.931**
VMU-Net	0.958	0.898	0.977	0.911	0.838	0.925
Unet++	0.938	0.818	0.969	0.842	0.730	0.875

PolypGen Dataset

Metrics	Accuracy	Recall	Specificity	F1/ Dice Coefficient	Mean IOU	Precision
TransUNet	0.973	0.791	0.990	0.832	0.712	0.877
SwinUNet	**0.979**	0.849	**0.992**	**0.874**	**0.776**	**0.900**
VMU-Net	0.962	**0.859**	0.976	0.840	0.724	0.821
Unet++	0.957	0.592	0.985	0.640	0.478	0.723

CONCLUSIONS AND DISCUSSION

SwinUNet had the best metrics in both the ISIC 2018 dataset and PolyPGen dataset. While all models demonstrate strong performance, a notable limitation observed is their difficulty in accurately segmenting lesions with coloration closely matching the surrounding skin. This suggests a reliance on strong color/contrast cues and highlights an area for future improvement, potentially through advanced feature extraction focusing on texture and boundary details or targeted data augmentation strategies. SwinUNet was the least affected on this limitation.

Additionally, while results are promising, current findings are limited to two datasets, and it remains unclear how each model will perform on other segmentation tasks such as organ segmentation, cell nuclei detection, or multi-class problems.

Future work will involve training and evaluating all four models across the full MedSegBench suite, which contains 35 medical image datasets spanning a wide range of anatomical regions. This broader evaluation will help determine the generalizability, robustness, and scalability of each architecture in diverse clinical scenarios.

BIBLIOGRAPHY

Chen, Jieneng et al.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv.04306, 2021.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. ECCV Workshops

Zhou, Zongwei, et al.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018.

Ruan, J., Xiang, S., Gao, J., Xie, M., Dong, D., Liao, N., Li, Z., Xiong, F., Liu, T., & Fu, Y. (2024). VM-UNet: Vision Mamba UNet for Medical Image Segmentation. arXiv preprint arXiv.02491.