Comparing U-Net and U-Net v2 for ISIC 2017 Skin Lesion Segmentation

Juan Cervantes

Abstract

In this report I present an experimental analysis comparing the performance of the U-Net [1] and U-Net v2 [2] deep learning architectures on the ISIC 2017 Skin Lesion Dataset [3]. The objective was to evaluate whether the novel mechanisms introduced by the authors of U-Net v2 improve segmentation performance in practice. Initially, experimental results revealed that U-Net v2 did not outperform vanilla U-Net with minimal data augmentation. However, when stronger data augmentation strategies were applied, U-Net v2 consistently outperformed vanilla U-Net in both qualitative mask predictions and quantitative metrics. These findings emphasize the importance of data variability and strong regularization in leveraging the novel architectural mechanisms introduced in U-Net v2. In particular, this highlights how transformer based segmentation models, such as U-Net v2 that enhance feature refinement, benefit more when exposed to diverse, augmented training data compared to conventional convolutional architectures.

1. Introduction

Semantic segmentation plays a critical role in medical image analysis, specifically in the detection and delineation of skin lesions. Accurate segmentation of dermoscopic images is essential for early, computer-aided diagnosis and treatment

planning. This task is posed with several challenges due to the visual variability of lesions, lighting inconsistencies, and contrast with surrounding skin. In recent years deep learning based methods, particularly encoder-decoder architectures, have made significant advances in medical image segmentation.

Among these, U-Net has emerged as a foundational architecture for biomedical image segmentation due to its effective use of skip connections and strong performance on limited data [1]. However, as the complexity of segmentation tasks has increased newer models have sought to improve upon U-Net. Recent developments in vision transformers have introduced alternative encoders capable of modeling long-range dependencies. U-Net v2 reimagines the classical U-Net architecture and provides improvement in Dice Similarity Coefficient (DSC) and Intersection Over Union (IoU) model assessment metrics compared to state of the art methods, while preserving memory and computational efficiency. It replaces the traditional convolutional encoder with a Pyramid Vision Transformer (PVT), and incorporates a novel mechanism in between the Encoder and Decoder modules named the Semantics and Detail Infusion (SDI) module. This SDI module refines hierarchical feature maps generated by the Encoder by applying spatial and channel attention mechanisms to the generated features of each level, and sends these refined feature maps to the Decoder with more semantic information and finer details per level of features.

Despite U-Net v2's impressive results in multiple benchmarks, it remains unclear how much of its performance advantage stems from architectural innovations versus experimental advantages such as pretraining or strong data augmentation. In this project I present a controlled empirical comparison between vanilla U-Net and U-Net v2

by progressively tuning hyperparameters and augmentation under identical conditions. I evaluate the practical benefits of U-Net v2's novelties and examine how its performance scales with regularization.

My results demonstrate that while U-Net v2 underperformed under insufficient data augmentation, it consistently surpassed standard U-Net in both IoU and Dice metric scores when trained with stronger regularization strategies, achieving a 2-3% improvement in both IoU and Dice under optimal conditions. This experiment contributes to a clearer understanding of when attention-based segmentation models offer meaningful advantages over convolutional architectures, and under what training procedures these advantages become evident.

2. Related Work

In their paper, the authors of U-Net v2 demonstrate the performance superiority of their architecture over several state of the art methods. Namely, U-Net v2 outperformed the following models in IoU and Dice metrics on the ISIC 2017 Skin Lesion Dataset, EGE-UNet [5], TransFuse [6], MALUNet [7], and of course U-Net [1]. While U-Net v2 consistently outperforms these architectures under fixed training conditions, those results do not fully explore the influence of how model performance varies with different degrees of data augmentation and regularization. To the best of my knowledge, no previous study has conducted a controlled, head to head, experimental comparison between the foundational U-Net and U-Net v2. This gap is what my study aims to fill, and while U-Net v2 showcases impressive performance in the literature this

experimental analysis reveals that its advantage becomes prominent only under certain conditions.

3. Data

The dataset used in this study is the ISIC 2017 dataset for skin lesion segmentation [3]. The dataset includes approximately 2750 dermoscopy images in .jpg format and their ground truth segmentation masks in .png format. The set is broken down into 2000 training images and their masks, 150 validation images and their masks, and 600 test images and their masks.

4. Approach

4.1 Experimental Setup

To ensure a fair comparison, both models were initially trained from scratch without leveraging any pretrained weights, and implemented in a controlled setting wherein data augmentation techniques and hyperparameter tuning strategies were progressively introduced and optimized. Pretraining was only introduced at the end of experimentation for the sake of comparison. The model archetypes were implemented as they are presented in their respective papers, U-Net [1] and U-Net v2 [2], and are trained and validated on the aforementioned ISIC 2017 Skin Lesion Dataset. Experiments were conducted on an NVIDIA A100 GPU in a Hyper Performance Computing (HPC) cluster with PyTorch. Both models are trained on a maximum number of 300 epochs, and on a loss function that combines Binary Cross Entropy with Dice Loss.

The loss function expression is defined as:

$$Total\ Loss = \alpha \cdot BCE_{Loss} + (1 - \alpha) \cdot Dice_{Loss}$$

where α denotes the weight coefficient tuned during experimentation. Multiple values were tested corresponding to an increasing emphasis on Dice Loss (region-level accuracy) over Binary Cross Entropy (pixel-wise classification accuracy). Polynomial learning rate decay was used with a power of 0.9, the Adam optimizer was also used on default parameters, and an initial learning rate of 0.001.

4.2 Data Preparation and Assessment

All images and masks were resized to 256 x 256, and all input images were normalized using ImageNet mean and standard deviation values. Additionally, all ground truth masks were binarized with a standard threshold of 0.5 and converted to single channel tensors for binary segmentation. Augmentations were applied during training to improve generalization, while test-time augmentation was used during evaluation for robustness. Performance was assessed using Dice Coefficient and Intersection over Union scores. These metrics were computed on binarized predictions with a fixed threshold applied to the sigmoid output. For U-Net v2 only the final output from deep-supervision was used for evaluation when it was enabled.

5. Experiments

To evaluate U-Net and U-Net v2, experiments included direct model comparisons, ablation of architectural choices such as deep supervision, loss function tuning, and of course augmentation analysis. The baseline setup of both models initially included training U-Net v2 without its intrinsic deep supervision, and both with limited

data augmentation, consisting only of random horizontal and vertical flips. U-Net used its standard convolutional encoder-decoder structure, while U-Net v2 incorporated a transformer based encoder with its SDI module, as it was presented in the code for its respective paper [4]. Table 1 presents the comparison results of both models on the ISIC 2017 dataset with a varied description per average run. Scores were gathered by averaging 3 runs per experimentation of the model and data.

Run Description	IoU	Dice
U-Net (Baseline, Light Aug)	0.722	0.821
U-Net (Strong Aug)	0.74	0.829
U-Net (Strong Aug, Threshold=0.45)	0.743	0.83
U-Net (Strong Aug, Pretrained: ResNet34)	0.759	0.841
U-Net v2 (No DS, Light Aug)	0.726	0.818
U-Net v2 (DS, Light Aug)	0.735	0.827
U-Net v2 (DS, Strong Aug)	0.746	0.843
U-Net v2 (DS, Strong Aug, Threshold=0.45)	0.753	0.848
U-Net v2 (DS, Strong Aug, Pretrained: PVTv2-B2)	0.774	0.863

Table.1 Experimental comparison between both models and description per average run

5.1 Baseline Runs

The baseline runs with light augmentation and deep supervision disabled show U-Net slightly outperforming U-Net v2, per table 1, achieving marginally higher IoU and Dice scores. This performance gap suggested that U-Net v2 may require stronger regularization and richer data variation to make use of U-Net v2's transformer-based encoder and complex feature refinement mechanisms via the SDI module to realize its potential. This served as motivation for further investigation into the effects of regularization and training strategies.

5.2 Loss Function Tuning

Experimentation with weighted combinations of Binary Cross Entropy and Dice loss were conducted in order to find the optimal balance between pixel-wise accuracy and shape overlap. The tested configurations included (BCE, Dice): (0.5, 0.5), (0.3, 0.7), (0.2, 0.8), and (0.1, 0.9).Results consistently demonstrated that loss function weights of BCE at 0.2 and Dice at 0.8 led to best performance. This configuration yielded smoother and more complete segmentation masks, as evident in figure (E). More balanced configurations (0.5, 0.5) and (0.3, 0.7) consistently resulted in lower overall performance,

5.3 Effect of Increasing Epochs

Subsequent experiments were conducted increasing training duration from 300 to 600 epochs to explore whether extended training would allow U-Net v2 to generalize more effectively. Both models demonstrated improved training loss convergence, but only U-Net v2 benefited from the extended training in terms of validation metrics. This indicated that the additional epochs provided enough learning iterations for the attention mechanisms of U-Net v2 to be utilized effectively. However, in the absence of sufficient data variation, U-Net v2 also became prone to overfitting where its greater capacity also demanded more regularization.

5.4 Ablation of Deep Supervision

Recognizing the underutilization of U-Net v2's deep architecture, ablation was conducted by enabling deep supervision. Deep supervision introduced auxiliary losses from intermediate decoder outputs, thus encouraging the model to learn at multiple scales which resulted in a consistent increase in performance. On average an increase

of ~1%, as shown in Table 1, was observed in both IoU and Dice scores across several runs, indicating deep supervision's effectiveness in stabilizing gradients. Additionally, this gain demonstrates deep supervision's improvement in convergence stability and enhances the model's ability to segment finer structural details. In contrast U-Net, lacking this mechanism, did not benefit from the additional supervision from intermediate hidden layers.

5.5 Effect of Thresholding

A threshold of 0.5 was initially used as is standard with binary classification conventions. However, early improvements were observed in U-Net v2 with a reduced threshold of 0.45, which produced marginal metric gains of ~0.05% in Dice score and ~0.07% in IoU. This helped recover finer lesion boundaries due to U-Net v2's sensitivity in prediction probabilities. Conversely, U-Net showed minimal sensitivity to threshold variation. The optimal results for both models was eventually achieved using the standard 0.5 threshold, specifically when paired with stronger data augmentation and regularization. Results suggest that threshold tuning alone has limited effect unless the model exhibits over or under confidence in probability calibration.

5.6 Test Time Augmentation

To further improve segmentation quality and performance, the effects of Test Time Augmentation (TTA) were evaluated by applying vertical and horizontal flips during inference. In particular, predictions were made on 4 variants of each test image, and the outputs were averaged to form the final mask. TTA provided modest but consistent improvements primarily due to reduced prediction noise and smoother mask boundaries, see figure (E). While TTA was more beneficial for U-Net v2 than U-Net,

these findings underscore the value of TTA as an inference time only regularization that proves useful when evaluation noisy or high resolution data without modifying training strategy.

5.7 Data Augmentation and Regularization Analysis

Data augmentation proved pivotal in enhancing the generalization capabilities of both models, particularly for U-Net v2. While initially both models were trained with light augmentation, limited to vertical and horizontal flips, as augmentation was progressively expanded and diversified U-Net v2 began to leverage its architectural strengths prominently. The progressive introduction of stronger forms of augmentation included random rotations, brightness and contrast shifts, random resized crops, color jitter, sharpness adjustments, and affine transformations. This augmentation pipeline diversified training data and acted as a form of regularization that reduced overfitting. The impact is visually evident in figures (A) and (B), which compare the difference in validation and training loss between strong and light augmentation for both models.

In figure (A), U-Net v2 demonstrates higher initial volatility under heavy augmentation (blue curve), but generalizes better over epochs compared to the light augmentation baseline (purple curve). Similarly, in figure (B), U-Net under heavy augmentation (blue curve) maintains lower validation loss in later epochs demonstrating improved generalization. These trends were also reflected in the final performance metrics as seen in table 1. Under light augmentation U-Net v2 struggled to outperform U-Net. However, under heavy augmentation U-Net v2 marked gains of ~2.3% in IoU and ~1.5% in Dice compared to earlier runs, and achieved approximately a 1.3% better performance over U-Net in both Dice and IoU metric scores, without pretraining. These

improvements affirm the architecture's reliance on data diversity for fully leveraging its transformer based encoder and SDI attention modules.

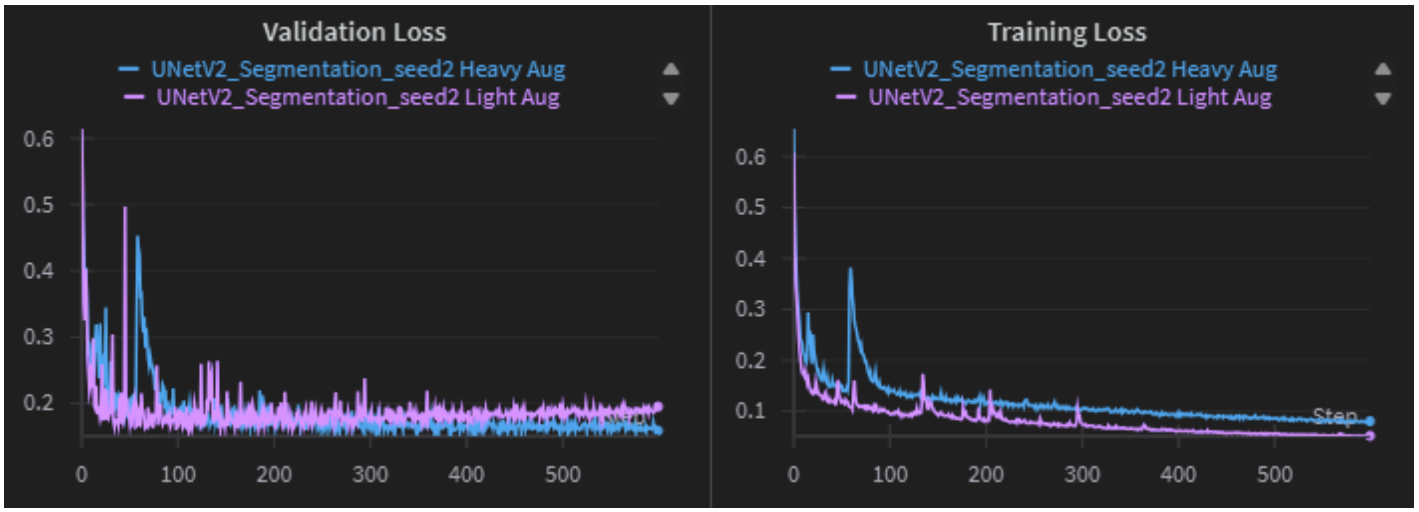

Fig (A). Validation and Training loss comparison in U-Net v2 between strong and limited data augmentation

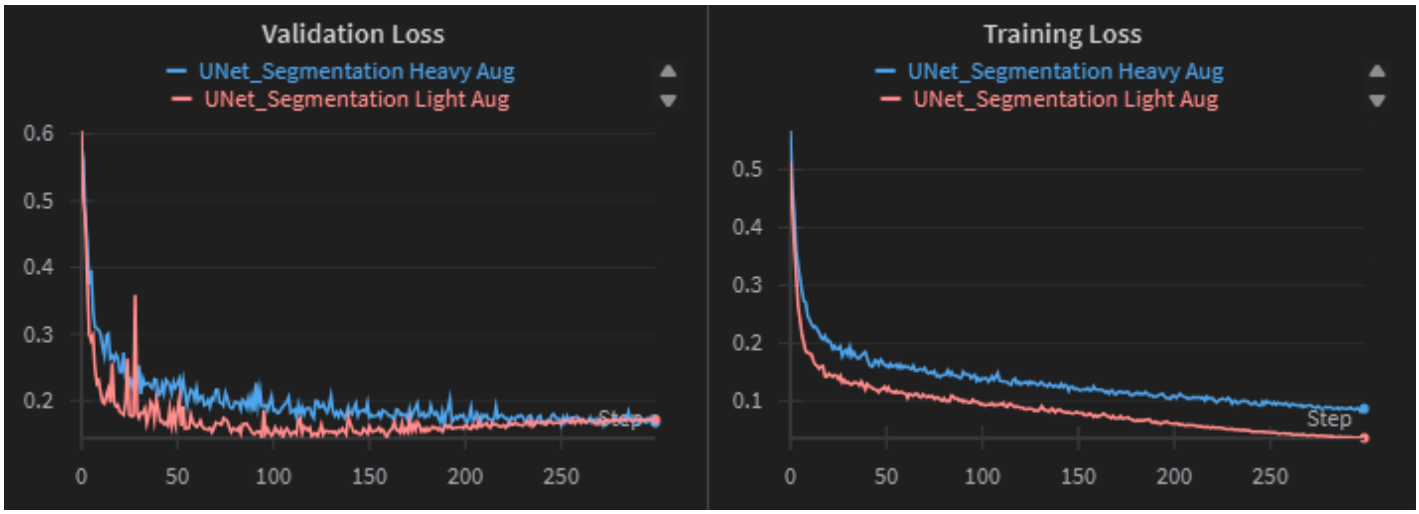

Figure (B). Validation and Training loss comparison in U-Net between strong and limited data augmentation

5.8 Effects of Pretraining

For the sake of comparison, ImageNet pretraining was enabled for both models by incorporating pretrained backbones. Namely, ResNet34 for U-Net and PVT v2 b2 for U-Net v2. As expected, both models yielded increased performance in both IoU and Dice scores, reaching their best performance as seen in table 1. Interestingly, figures (C) and (D) illustrate signs of slight persistent overfitting despite regularization. While pretraining significantly improves model performance and the training and validation loss curves are lower than non-pretrained versions, these results again indicate the need for further tuning. Conclusively, the full benefit of pretraining is best realized when paired with sufficient regularization methods.

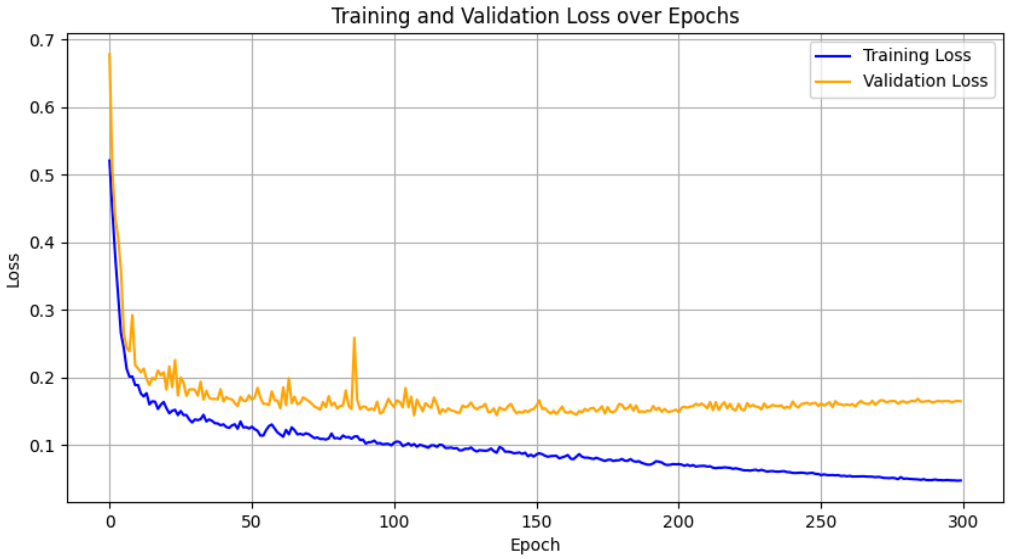

Figure (C). U-Net Pretrained: ResNet34

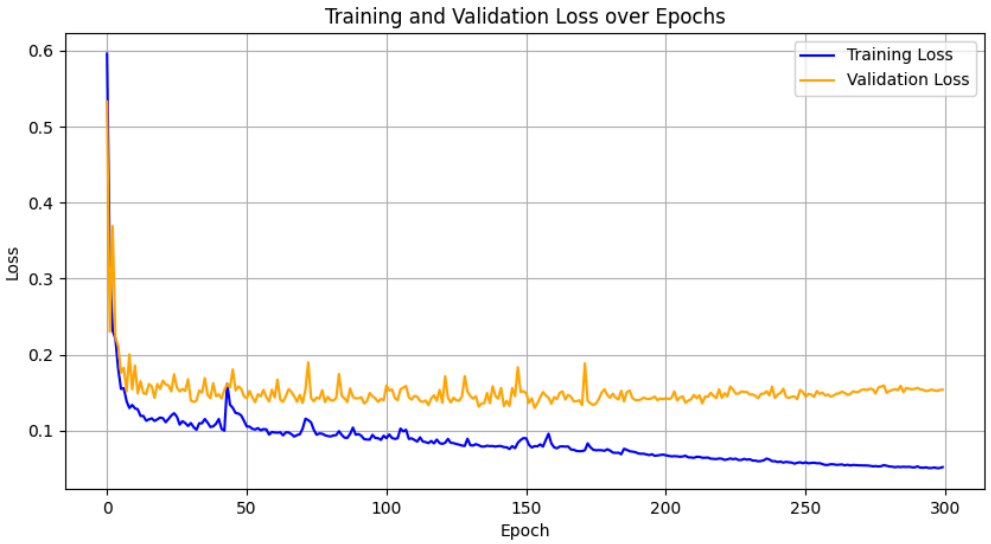

Figure (D). U-Net v2 Pretrained: PVTv2_b2

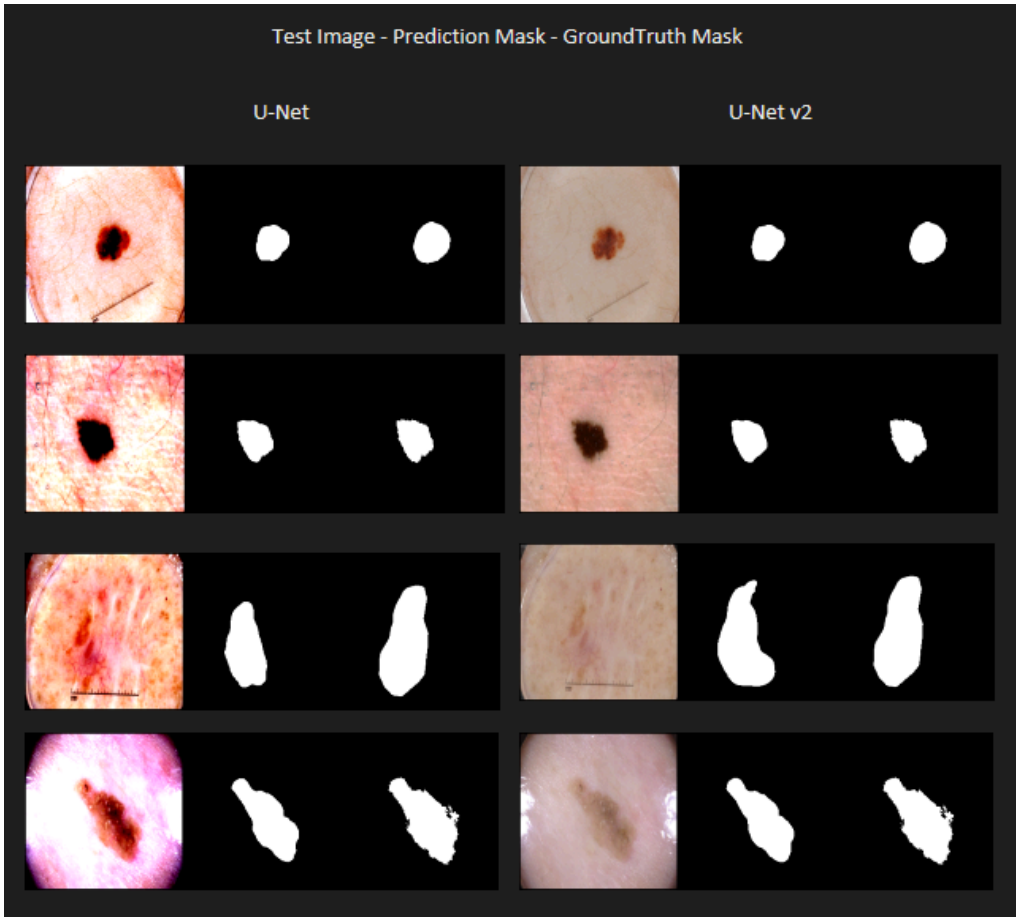

Figure (E). Prediction masks of each model at best performance

Conclusions

This study presents an experimental comparison between vanilla U-Net and U-Net v2 on the ISIC 2017 Skin Lesion dataset. Experiments demonstrated that U-Net v2 does not clearly outperform U-Net when both models are trained with minimal data augmentation. However, with stronger data augmentation and regularization U-Net v2 achieved better IoU and Dice metric scores compared to U-Net. This confirms the importance of training context when leveraging U-Net v2's architectural innovations.

Enabling deep supervision in U-Net v2 improved training stability and increased model performance. Loss function weighting proved pivotal, emphasizing Dice coefficient yielded the most effective learning. Additionally, fine tuning thresholds and incorporating test time augmentation offered marginal and consistent gains in final metric scores. The use of pretrained backbones further improved both models' performance, each achieving their best results yet. Nevertheless, this reintroduced overfitting, thus emphasizing the continued importance of regularization even in a pretrained environment. This study highlights the importance of aligning architectural complexity with dataset variability and size.

References:

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox

“U-Net: Convolutional Networks for Biomedical Image Segmentation”

arXiv:1505.04597v1 18 May 2015

[2] Yaopeng Peng, Milan Sonka, Danny Z. Chen

“U-Net v2: RETHINKING THE SKIP CONNECTIONS OF U-NET FOR MEDICAL IMAGE SEGMENTATION” arXiv:2311.17791v2 [eess.IV] 30 Mar 2024

arXiv:2311.17791v2

[3] Codella N, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza S, Kalloo A, Liopyris K, Mishra N, Kittler H, 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)".

arXiv:1710.05006v3

[4] yaoppeng/U-Net_v2 “Pytorch implementation of U-Net v2: RETHINKING THE SKIP CONNECTIONS OF U-NET FOR MEDICAL IMAGE SEGMENTATION” Github Repository: https://github.com/yaoppeng/U-Net_v2

[5] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu, “EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation”, 2023,

arXiv:2102.08005v2

[6] Yundong Zhang, Huiye Liu, and Qiang Hu, “TransFuse: Fusing Transformers and CNNs for medical image segmentation,” in MICCAI, Proceedings, Part I 24. Springer, 2021, arXiv:2102.08005v2

[7] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu, “MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation,” in BIBM. IEEE, 2022, arXiv:2211.01784v1