Computing, Performance and Communication Systems (2026)
Clausius Scientific Press, Canada
DOI: 10.23977/cpcs.2026.100113
ISSN 2371-8889 Vol. 10 Num. 1

U-Net-Based Medical Image Segmentation with Data Augmentation and Loss Function Comparison

Hongru Li[a], Zhitao Wu[b,*]

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China
[a]lihongru415@outlook.com,[b]aswzt@163.com
**Corresponding author*

Keywords: Medical image segmentation; U-Net; skin lesion; Dice loss; Data augmentation

Abstract: Medical image segmentation requires accurate localization despite limited annotations, variable target size, and uncertain boundaries. This paper presents a U-Net study for binary skin lesion segmentation using the labeled ISIC 2018 Task 1 training collection. The experiments compare U-Net with fully convolutional and SegNet baselines, examine binary cross entropy, soft Dice, and combined losses, and test geometric augmentation and input resolution. Lesion and background masks are transformed jointly, and evaluation restores predictions to the original image geometry before thresholding. Dice coefficient, intersection over union, pixel accuracy, inference latency, and total training time provide complementary measures of segmentation quality and computational cost. In the reference configuration, augmented U-Net with a combined loss achieves a mean Dice coefficient of 0.907 and an intersection over union of 0.837, with an inference latency of 17.6 milliseconds per image. The protocol addresses background-dominated accuracy, inconsistent mask interpolation, and leakage between related images through consistent preprocessing and grouped data splitting. Subsequent work will quantify boundary errors, annotation variability, subgroup performance, and external generalization while reporting the additional training and inference costs of higher resolution.

1. Introduction

Segmenting a medical image assigns a spatial label to each pixel. For skin lesions, irregular borders, hair occlusion, low contrast, and variable lesion extent can make that assignment difficult. U-Net combines a contracting path for contextual representation with an expanding path for spatial recovery [1]. Its skip connections make it a relevant starting point when detailed localization must be learned from a limited annotated collection.

Fully convolutional networks provide an established framework for dense prediction [2], and SegNet offers an encoder-decoder design that uses pooling indices during upsampling [3]. Comparing these alternatives can help determine whether a particular U-Net configuration offers a useful balance of overlap accuracy and computation. Identical data partitions, preprocessing, and

threshold rules align the comparison across architectures.

This paper investigates three questions: how the specified U-Net compares with representative dense-prediction baselines, how loss choice interacts with geometric augmentation, and how input resolution affects segmentation quality and latency. The contribution is a reproducible experimental design with explicit metric definitions and a controlled empirical comparison of the resulting segmentation quality and computational cost.

2. Related Theories and Technologies

2.1. Encoder Decoder Segmentation and Skip Connections

The U-Net encoder repeatedly applies convolution and downsampling to extract increasingly contextual features. The decoder upsamples features and concatenates them with corresponding encoder representations, helping recover localization information [1]. The implementation uses four downsampling stages with channel widths 32, 64, 128, and 256, followed by a 512-channel bottleneck. Each stage contains two 3 by 3 convolutions and rectified linear activations. A 1 by 1 convolution and sigmoid output produce foreground probabilities.

Padding preserves spatial dimensions within each convolutional stage, and bilinear upsampling is followed by convolution in the decoder. These implementation choices adapt the original U-Net to the present segmentation task. The FCN baseline uses a VGG-style convolutional encoder and an FCN-8s prediction head [2]. The SegNet baseline uses a corresponding encoder-decoder with pooling-index upsampling [3]. All networks are randomly initialized, and their capacities are considered when interpreting architectural performance.

2.2. Overlap Losses and Class Imbalance

Binary cross entropy averages the negative log probability assigned to the correct foreground or background label. It supplies a pixel-level learning signal, but a small lesion contributes fewer pixels than the surrounding background. Soft Dice loss directly emphasizes foreground overlap. Dice-based optimization is also associated with medical segmentation work such as V-Net [4]. The present study applies this principle in a two-dimensional U-Net for skin lesion segmentation.

For one image, soft Dice is twice the sum of predicted probability times binary label, plus a smoothing constant, divided by the sum of predicted probabilities and labels, plus that constant. The smoothing constant is fixed at 0.000001. Dice loss is one minus this value, averaged over images in the minibatch. Combined loss gives equal weight to binary cross entropy and Dice loss. Equal weighting keeps the combined-loss comparison consistent across experiments.

2.3. Augmentation and Configuration Consistency

Segmentation augmentation applies the same spatial transformation to an image and its mask. Independent transformations invalidate the correspondence between image and supervision. The pipeline uses rotations within 25 degrees, horizontal and vertical flips with probability 0.5, and isotropic scaling between 0.90 and 1.10. Image interpolation is bilinear, while binary masks use nearest-neighbor interpolation. Transformations are followed by padding or center cropping to the working input size.

Self-configuring segmentation research highlights the importance of coordinated preprocessing and training choices [5]. This motivates recording the entire pipeline alongside the architecture. A bottleneck dropout rate of 0.10 is held fixed across the U-Net loss and resolution sweeps, using dropout as a regularization mechanism [6].

3. Experimental Design and Implementation

3.1. Dataset Preparation

The source is the publicly available labeled training collection for ISIC 2018 Task 1 lesion boundary segmentation. Each input is a dermoscopic RGB image paired with a binary lesion mask. An internal training, validation, and test split uses approximately 70%, 15%, and 15% of eligible groups.

Inputs are resized with preserved aspect ratio and padded to a square, with a reference size of 256 by 256 pixels. Pixel values are scaled to the interval from zero to one, and channel means and standard deviations are estimated from training images only. Padded regions are excluded from loss and metric calculations. Spatial augmentation occurs only during training, after the split. The validation and test pipelines remain deterministic.

For evaluation, the padding is removed from each probability map and the map is resized to the original image dimensions using bilinear interpolation. Foreground is then assigned at a fixed probability threshold of 0.5. Ground-truth masks remain at their original dimensions. This procedure prevents the smaller evaluation grid of a low-resolution input from making its boundary task artificially easier.

3.2. Experimental Environment

The implementation uses Python and PyTorch on a single CUDA-capable workstation, and all models were trained, validated, and timed on the same machine.

Total training time includes all fitting and validation over the fixed epoch budget. Model-only inference latency is measured with batch size one, after 50 warm-up passes, over at least 500 timed passes with accelerator synchronization. File access, preprocessing, and restoration to original image size are excluded from that latency.

3.3. Experimental Parameter Setting

Adam provides the optimizer [7], with initial rate 0.0003, momentum coefficients 0.9 and 0.999. Batch size is eight. Each model is trained for 60 epochs, and the checkpoint with highest validation mean Dice is retained. The learning rate is reduced by a factor of 0.5 after eight epochs without a validation-Dice improvement of at least 0.001, with a minimum rate of 0.000001. Training still completes the 60-epoch budget, making total training duration comparable at a fixed budget.

The architecture comparison uses combined loss, geometric augmentation, and 256-pixel inputs for all three networks. The loss comparison is a complete three-by-two design across loss family and augmentation status, using U-Net. The input-size comparison uses 192, 256, and 384 pixels with combined loss and augmentation. These sizes are divisible by the total downsampling factor. Batch size remains eight when memory permits; any reduction is reported alongside the corresponding timing result.

Validation performance determined all configuration choices. After those choices were frozen, every prespecified condition was evaluated once on the internal test set for reporting.

3.4. Experimental Indicators

For one image, let TP, FP, TN, and FN denote foreground true positives, false positives, true negatives, and false negatives. Hard Dice equals twice TP divided by twice TP plus FP plus FN. IoU equals TP divided by TP plus FP plus FN. Pixel accuracy equals TP plus TN divided by the

total number of valid image pixels. All three metrics are computed after restoring original image dimensions and are averaged equally across test images.

If both prediction and reference are empty, Dice and IoU are set to one; an empty prediction against a nonempty reference receives zero. Any such cases must be counted. Mean image-wise Dice and mean image-wise IoU are reported separately because their distributions reflect image difficulty.

Pixel accuracy is secondary because background pixels can dominate it. The tables report point estimates of the overlap metrics. Subsequent evaluations will use seed variability and lesion-group bootstrap intervals to quantify uncertainty. Latency is reported in milliseconds per image and training duration in minutes, with their measurement scopes stated explicitly.

4. Experimental Results and Analysis

4.1. Architecture Comparison at a Fixed Input Size

Table 1 reports the architecture comparison at 256 by 256 pixels.

Table 1. Segmentation architectures under matched preprocessing

Model	Mean Dice	Mean IoU	Pixel accuracy	Latency (ms)	Training (min)
FCN	0.864	0.769	0.947	11.8	31.5
SegNet	0.879	0.791	0.953	13.7	35.0
U-Net	0.907	0.837	0.966	17.6	41.2

U-Net exceeds FCN by 0.043 in mean Dice and 0.068 in mean IoU, with 9.7 additional training minutes, and its latency is 5.8 milliseconds higher. This is an accuracy-cost comparison: the higher-overlap configuration carries a computational price.

Pixel accuracy differs by less than two percentage points across rows, while foreground overlap changes more substantially, which indicates that background pixels can conceal lesion errors.

4.2. Loss Function and Augmentation Interaction

Table 2 evaluates loss choice and augmentation jointly. The six conditions distinguish the effect of augmentation within each loss family. Every row uses U-Net, 256-pixel inputs, and the same 60-epoch training budget.

Table 2. Loss functions with and without geometric augmentation

Loss	Augmentation	Mean Dice	Mean IoU	Training time (min)
Cross entropy	Off	0.874	0.784	35.6
Dice	Off	0.885	0.802	36.2
Combined	Off	0.891	0.811	36.8
Cross entropy	On	0.893	0.814	39.8
Dice	On	0.902	0.829	40.5
Combined	On	0.907	0.837	41.2

Augmentation raises mean Dice by 0.019 with cross entropy, 0.017 with Dice loss, and 0.016 with combined loss. The combined augmented row matches the U-Net row in Table 1. The largest gain from changing the loss alone is smaller than the full difference between the unaugmented cross-entropy row and the augmented combined row. The results therefore separate the contribution of loss selection from the combined effect of loss and augmentation.

The additional training time with augmentation reflects online transformations under an equal epoch budget, and the combined loss adds a smaller computational increment. Boundary-quality

analysis will examine small lesions, faint margins, and hair-obscured regions and quantify uncertainty in paired image-level differences.

4.3. Input Resolution and Inference Efficiency

Table 3 changes input resolution while preserving the reference loss and augmentation. Metrics are still evaluated at original image dimensions, and model-only latency excludes geometric restoration. This maintains a consistent quality definition while exposing the computational effect of the network input.

Table 3. Input resolution and computational tradeoffs

Input size	Mean Dice	Mean IoU	Pixel accuracy	Latency (ms)	Training (min)
192 by 192	0.895	0.819	0.961	11.2	27.6
256 by 256	0.907	0.837	0.966	17.6	41.2
384 by 384	0.912	0.846	0.968	32.9	78.4

Increasing the input from 256 to 384 pixels improves Dice by 0.005 but increases latency by 15.3 milliseconds and training duration by 37.2 minutes. Whether that increment is useful depends on the application. The 192-pixel row offers a lower cost but loses more overlap, plausibly because fine boundaries become less recoverable after downsampling.

Memory consumption complements input-size selection because larger inputs that require smaller batches also change gradient statistics and per-epoch throughput.

5. Research Conclusions and Prospects

5.1. Research Conclusions

This paper presents a binary lesion segmentation study that integrates U-Net, joint image-mask augmentation, and a comparison of cross entropy, Dice, and combined losses. The evaluation prioritizes image-wise Dice and IoU, uses pixel accuracy as a secondary measure, and explicitly accounts for model-only latency and fixed-budget training time. In the reference configuration, mean Dice is 0.907, mean IoU is 0.837, and latency is 17.6 milliseconds, which achieves higher overlap scores than the implemented FCN and SegNet baselines under matched preprocessing.

The results support a concrete configuration decision for the stated workload: the combined loss with augmentation and 256-pixel inputs provides a favorable compromise between overlap quality and computation in this study. Consistent restoration to original geometry, a fixed threshold, and matched partitions were necessary for that comparison. Training efficiency must be evaluated alongside segmentation performance because higher-resolution optimization imposes a substantial additional cost.

5.2. Research Limitations and Prospects

The evaluation uses an internal dermoscopic split. External imaging devices, populations, and modalities require separate evaluation, and more complete patient identifiers will strengthen partition independence.

Future experiments will use external device-specific test sets, report performance by lesion size and acquisition subgroup, and incorporate multiple annotators where possible. Boundary distances, calibration of foreground probabilities, and uncertainty around difficult margins will complement overlap metrics. Resolution will be selected using measured memory, training duration, end-to-end latency, and paired quality differences. Clinical usefulness will be established through a dedicated

validation process beyond the technical segmentation protocol described here.

References

[1] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.

[2] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[3] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[4] Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision. 2016: 565-571.

[5] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature Methods, 2021, 18(2): 203-211.

[6] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[7] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//International Conference on Learning Representations. 2015.