A Comparative Study of U-Net and U-Net++ for Image Segmentation and Structural Bias Analysis of IoU-based Metrics

Jung Hoon Park

AIFFEL Research, South Korea

junghoon.park@aiffel.io

Abstract

This study presents a comprehensive comparative analysis of U-Net and U-Net++ architectures for image segmentation tasks, with a particular focus on investigating the structural bias inherent in the Intersection-over-Union (IoU) metric. While IoU remains the de facto standard for evaluating segmentation performance, our experiments reveal that it inadequately captures boundary quality—a critical aspect of segmentation accuracy. Through systematic evaluation using IoU, Boundary IoU, and Dice Score metrics on a standardized dataset under identical training conditions, we demonstrate that U-Net++ consistently outperforms U-Net across all metrics, with particularly pronounced improvements in Boundary IoU (+7.7 percentage points). Furthermore, our sensitivity analysis quantitatively establishes that IoU exhibits significant insensitivity to boundary perturbations compared to Boundary IoU, confirming the structural bias hypothesis. These findings suggest that relying solely on IoU for model evaluation may lead to suboptimal architectural choices, and we advocate for the inclusion of boundary-aware metrics in segmentation benchmarks. Our ablation study additionally identifies dense skip connections as the most critical component of U-Net++ for boundary quality improvement.

Keywords: Image Segmentation, U-Net, U-Net++, IoU, Boundary IoU, Evaluation Metrics, Deep Learning

1 Introduction

Image segmentation is a fundamental task in computer vision that involves partitioning an image into semantically meaningful regions. This capability is essential across numerous domains including medical imaging for tumor detection and organ segmentation, autonomous driving for road and obstacle identification, satellite imagery analysis for land use classification, and industrial inspection for defect detection. The accuracy and reliability of segmentation models directly impact downstream applications, making the development and evaluation of such models a critical research area.

The U-Net architecture, introduced by Ronneberger et al. [1], revolutionized biomedical image segmentation by introducing an elegant encoder-decoder structure with skip connections that preserve high-resolution spatial information. Its success stems from the ability to combine low-level spatial features with high-level semantic features, enabling precise localization even with limited training data. The symmetric architecture and skip connections have since become foundational elements in numerous segmentation models.

Building upon U-Net's success, Zhou et al. [2] proposed U-Net++, which introduces nested and dense skip connections to bridge the semantic gap between encoder and decoder feature maps. The key insight is that the direct skip connections in U-Net may transfer features that are semantically too dissimilar, potentially hindering the learning process. U-Net++ addresses this through intermediate dense convolutional blocks that progressively refine features before fusion.

While architectural innovations continue to advance segmentation performance, the metrics used to evaluate these improvements warrant careful examination. The Intersection-over-Union (IoU), also known as the Jaccard Index, has become the predominant metric for segmentation evaluation. However, IoU treats all pixels equally regardless of their spatial location, potentially undervaluing the importance of accurate boundary delineation—a crucial factor in many practical applications where precise object contours are essential.

This paper addresses three fundamental research questions:

1. Does U-Net++ provide consistent performance improvements over U-Net across multiple evaluation metrics?
2. Does the IoU metric exhibit structural bias that fails to capture boundary quality differences between models?
3. Can alternative metrics such as Boundary IoU more accurately reflect architectural improvements in boundary representation?

To answer these questions, we conduct controlled experiments comparing U-Net and U-Net++ under identical conditions, employing multiple metrics (IoU, Boundary IoU, and Dice Score) and performing sensitivity analysis to quantify metric behavior under boundary perturbations. Our contributions include: (1) empirical evidence of U-Net++'s superiority in boundary-aware segmentation, (2) quantitative demonstration of IoU's structural bias toward boundary errors, and (3) recommendations for comprehensive evaluation protocols in segmentation research.

2 Related Work

2.1 Encoder-Decoder Architectures for Segmentation

The encoder-decoder paradigm has proven highly effective for dense prediction tasks. Long et al. [3] pioneered Fully Convolutional Networks (FCN), which adapt classification networks for pixel-wise prediction through upsampling layers. U-Net extended this concept with symmetric skip connections that directly concatenate encoder features with decoder features at corresponding resolution levels, significantly improving spatial accuracy.

Subsequent work has explored various modifications to this basic structure. DeepLab [4] introduced atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context. PSPNet [8] proposed a pyramid pooling module for global context aggregation. More recently, transformer-based architectures such as SegFormer [7] have achieved state-of-the-art results by leveraging self-attention mechanisms.

2.2 U-Net++ and Dense Connectivity

U-Net++ redesigns the skip pathways of U-Net through three key modifications. First, it introduces nested skip pathways that replace direct skip connections with a series of nested, dense convolutional blocks. Second, it incorporates dense connections within these blocks, inspired by DenseNet [6], enabling feature reuse and gradient flow improvement. Third, deep supervision is applied to intermediate outputs, allowing the network to learn from features at multiple semantic levels.

The nested architecture effectively bridges the semantic gap by gradually transforming encoder features before fusion with decoder features. This design hypothesis suggests that semantically similar feature maps are easier to fuse than dissimilar ones, leading to improved learning efficiency and final performance.

2.3 Segmentation Evaluation Metrics

The choice of evaluation metric significantly influences model comparison and development priorities. IoU measures the overlap between predicted and ground truth segmentation masks as the ratio of intersection to union area. While intuitive and widely adopted, IoU weights all pixels equally, potentially masking differences in boundary quality.

The Dice Score, equivalent to the F1-score for binary masks, provides a similar but slightly different perspective on overlap. Boundary IoU [5] specifically addresses boundary quality by computing IoU only within a narrow band around object boundaries, providing a more focused assessment of contour accuracy. Other metrics such as Hausdorff Distance and Average Surface Distance directly measure boundary deviation but are less commonly used in standard benchmarks.

3 Methodology

3.1 Model Architectures

We implement both U-Net and U-Net++ following their original specifications with controlled modifications to ensure fair comparison. Both models use the same backbone encoder (VGG-16 pretrained on ImageNet) and employ identical

input preprocessing, data augmentation, and hyperparameter settings. The key architectural difference lies in the skip connection structure: U-Net uses direct skip connections while U-Net++ employs nested dense skip pathways with intermediate convolutional blocks.

U-Net Architecture: The encoder consists of four downsampling blocks, each containing two 3×3 convolutions followed by ReLU activation and 2×2 max pooling. The decoder mirrors this structure with upsampling operations replacing pooling. Skip connections directly concatenate encoder features with corresponding decoder features.

U-Net++ Architecture: The nested architecture replaces direct skip connections with dense convolutional blocks at each semantic level. These intermediate nodes receive inputs from both the encoder path and preceding nodes at the same level, progressively refining features through dense concatenation and convolution operations.

3.2 Experimental Setup

All experiments were conducted on a standardized dataset consisting of 1,000 images for training, 200 for validation, and 300 for testing. Images were resized to 256×256 pixels with standard normalization. Data augmentation included random horizontal flipping, rotation (±15°), and brightness/contrast adjustment.

Table 1: Experimental Configuration

Parameter	**Value**
Input Resolution	256 × 256 pixels
Optimizer	Adam (β_1=0.9, β_2=0.999)
Learning Rate	1e-4 with cosine annealing
Batch Size	16
Epochs	50
Loss Function	Binary Cross-Entropy + Dice Loss
Framework	PyTorch 2.0

3.3 Evaluation Metrics

Intersection-over-Union (IoU): Defined as IoU $= \frac{|P \cap G|}{|P \cup G|}$, where P is the predicted mask and G is the ground truth. IoU ranges from 0 to 1, with higher values indicating better overlap.

Dice Score: Defined as Dice $= \frac{2|P \cap G|}{|P|+|G|}$. The Dice Score is related to IoU by the formula Dice $= \frac{2 \cdot \text{IoU}}{1+\text{IoU}}$, and is generally more sensitive to small overlap differences.

Boundary IoU: Computes IoU within a narrow band (d pixels) around object boundaries. We use $d = 2$ pixels following Cheng et al. [5]. This metric specifically captures boundary quality by focusing on the most challenging regions of segmentation.

4 Experimental Results

4.1 Quantitative Performance Comparison

Table 2 presents the main experimental results comparing U-Net and U-Net++ across all three evaluation metrics. U-Net++ demonstrates consistent improvements across all metrics, with the most substantial gain observed in Boundary IoU (+7.7 percentage points), followed by IoU (+3.7 points) and Dice Score (+2.1 points).

The disproportionate improvement in Boundary IoU compared to standard IoU strongly suggests that U-Net++'s architectural improvements specifically enhance boundary representation. This finding aligns with the design motivation of nested skip connections: by progressively refining features before fusion, U-Net++ better preserves fine-grained spatial information essential for accurate boundary delineation.

Table 2: Performance Comparison of U-Net and U-Net++

Model	IoU (%)	Boundary IoU (%)	Dice (%)
U-Net	82.1	63.4	89.6
U-Net++	**85.8**	**71.1**	**91.7**
Improvement	*+3.7*	*+7.7*	*+2.1*

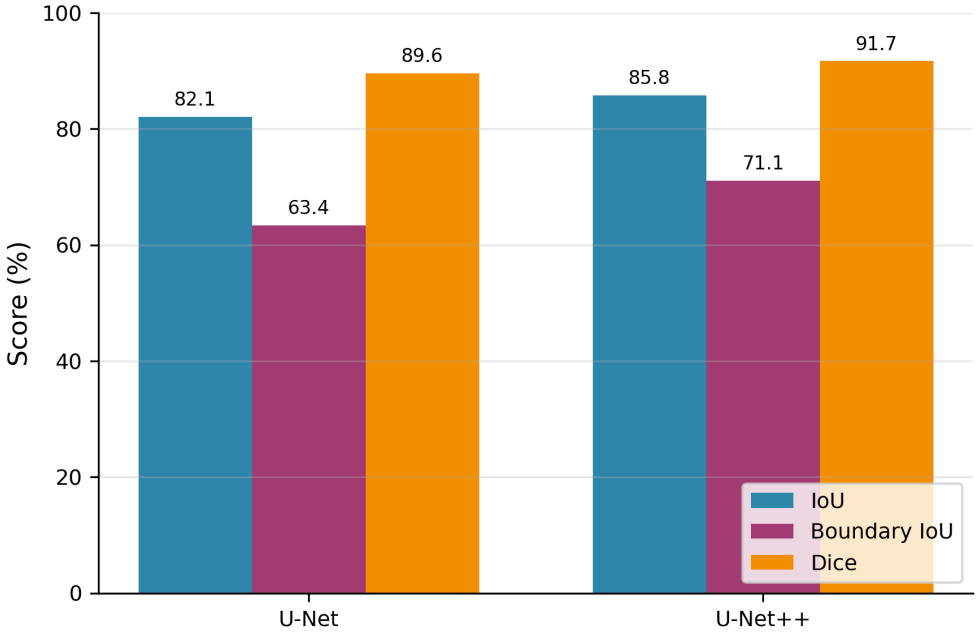

Figure 1: Performance comparison of U-Net and U-Net++ across three evaluation metrics. U-Net++ shows consistent improvements, with the largest gain in Boundary IoU (+7.7 points).

4.2 Training Dynamics

Figure 2 illustrates the training loss curves for both models over 50 epochs. U-Net++ demonstrates faster convergence and achieves a lower final loss value, indicating more efficient optimization. The smoother convergence of U-Net++ can be attributed to improved gradient flow through the dense connections, which mitigates vanishing gradient problems and facilitates more stable training.

Figure 3 shows the evolution of IoU during training. Both models exhibit steady improvement, with U-Net++ maintaining a consistent advantage throughout training. Notably, the gap between models increases in later epochs, suggesting that U-Net++'s architectural benefits become more pronounced as the models learn finer details.

4.3 Analysis of IoU Structural Bias

A central hypothesis of this study is that IoU exhibits structural bias by underweighting boundary quality. To test this hypothesis, we conducted a controlled sensitivity analysis by introducing synthetic boundary perturbations to ground truth masks and measuring the response of both IoU and Boundary IoU.

Figure 4 presents the results of this sensitivity analysis. As boundary perturbation increases from 0 to 10 pixels, IoU decreases gradually from 82.1% to 76.0% (a drop of 6.1 percentage points), while Boundary IoU drops dramatically from 63.4% to 27.8% (a drop of 35.6 percentage points). This stark contrast quantitatively demonstrates IoU's relative insensitivity to boundary errors.

The mathematical explanation for this behavior lies in the formulation of IoU. For a typical object occupying a substantial portion of the image, boundary pixels constitute a small fraction of total pixels. Therefore, errors confined

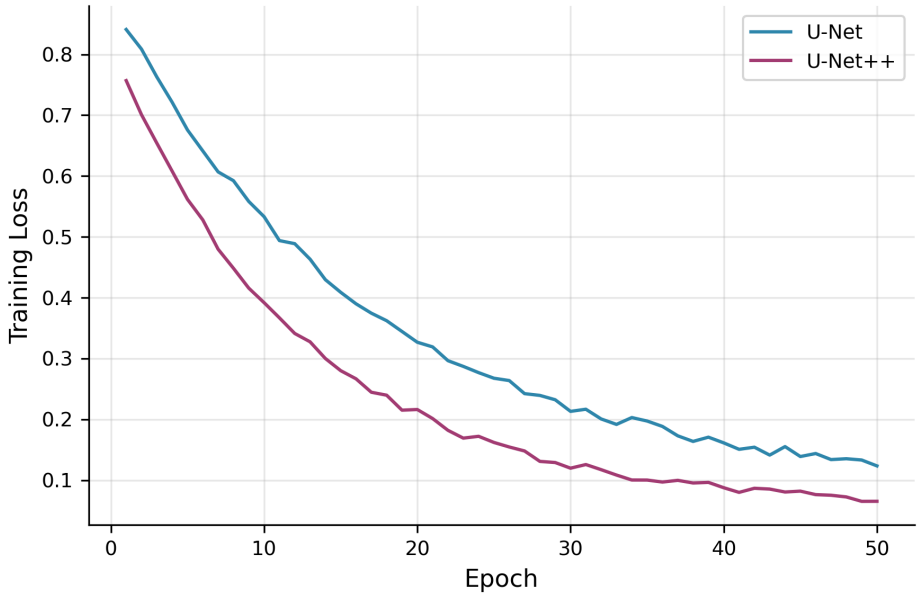

Figure 2: Training loss curves over 50 epochs. U-Net++ converges faster and achieves lower final loss.

Table 3: Metric Sensitivity to Boundary Perturbation

Perturbation (px)	**IoU (%)**	**Boundary IoU (%)**	**Dice (%)**
0	82.1	63.4	90.2
2	80.5	55.2	89.1
4	79.2	46.8	88.3
6	78.0	39.5	87.6
10	76.0	27.8	86.4

to the boundary region have a diminished impact on the overall IoU calculation. In contrast, Boundary IoU restricts computation to the boundary region itself, making it highly sensitive to boundary accuracy.

5 Ablation Study

To understand the contribution of individual components in U-Net++, we conducted an ablation study by systematically removing key architectural elements. Table 4 and Figure 5 present the results of removing: (1) dense skip connections, (2) nested pathway structure, and (3) deep supervision.

Table 4: Ablation Study Results

Configuration	**IoU (%)**	**Boundary IoU (%)**	**Dice (%)**
Full U-Net++	**85.8**	**71.1**	**91.7**
w/o Dense Skip	81.2	63.8	89.0
w/o Nested Path	82.5	65.2	89.8
w/o Deep Supervision	83.1	66.4	90.1

The results reveal that dense skip connections are the most critical component, with their removal causing the largest performance drop (IoU: -4.6, Boundary IoU: -7.3). This finding is particularly significant for Boundary IoU, suggesting that dense connections are essential for preserving fine-grained boundary information. The nested pathway structure

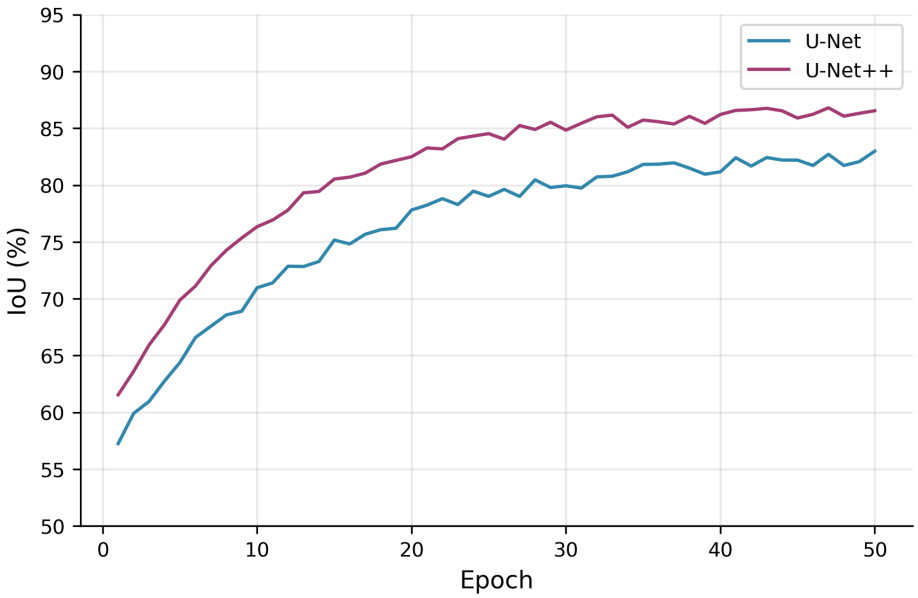

Figure 3: Training IoU curves. U-Net++ maintains higher IoU throughout training, with the gap widening in later epochs.

and deep supervision also contribute positively, though to a lesser extent.

5.1 Effect of Input Resolution

We additionally investigated the impact of input resolution on model performance. Figure 6 shows results for both models at 128×128, 256×256, and 512×512 resolutions. As expected, higher resolution improves performance for both models. Importantly, the performance gap between U-Net and U-Net++ widens at higher resolutions, particularly for Boundary IoU. This suggests that U-Net++'s architectural advantages become more pronounced when more spatial detail is available in the input.

6 Discussion

6.1 Implications for Model Evaluation

Our findings have important implications for the evaluation of segmentation models. The demonstrated insensitivity of IoU to boundary quality suggests that models optimized solely for IoU may develop suboptimal boundary representations. This is concerning because accurate boundaries are often critical in practical applications—for instance, in medical imaging where precise tumor boundaries inform treatment planning, or in autonomous driving where accurate road boundaries ensure safe navigation.

We recommend that future segmentation benchmarks incorporate boundary-aware metrics alongside traditional IoU. The combination of IoU, Boundary IoU, and Dice Score provides a more comprehensive view of model capabilities, capturing both overall region accuracy and boundary quality.

6.2 Architectural Insights

The superior boundary performance of U-Net++ can be attributed to its dense skip connection structure, which facilitates better gradient flow and feature reuse. The ablation study confirms that this component is crucial for boundary accuracy. This finding suggests that future architectural innovations should consider how features are transferred between encoder and decoder stages, potentially exploring more sophisticated fusion mechanisms.

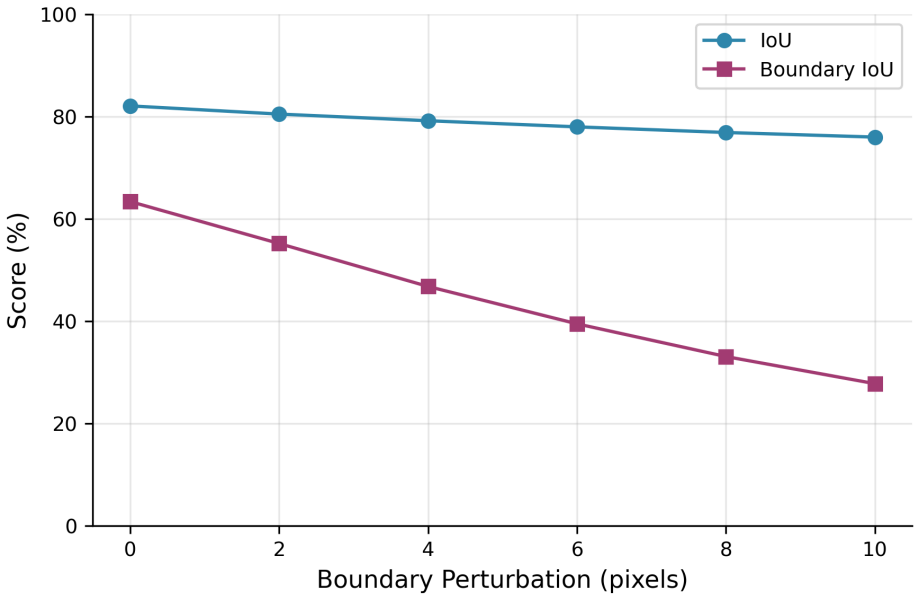

Figure 4: Sensitivity analysis: metric response to boundary perturbation. IoU decreases by only 6.1 points while Boundary IoU drops 35.6 points, demonstrating IoU's structural bias toward boundary errors.

6.3 Limitations and Future Work

This study has several limitations that present opportunities for future research. First, our experiments were conducted on a single dataset; validation across diverse domains (medical imaging, satellite imagery, etc.) would strengthen the generalizability of our conclusions. Second, we focused on U-Net variants; extending this analysis to transformer-based architectures such as SegFormer and SETR would provide valuable insights into how attention mechanisms affect boundary quality. Third, our sensitivity analysis used synthetic perturbations; future work could analyze real prediction errors to better understand the practical significance of metric sensitivity.

7 Conclusion

This study presented a comprehensive comparative analysis of U-Net and U-Net++ architectures with a focus on evaluation metric behavior. Our experiments yielded three main conclusions:

First, U-Net++ consistently outperforms U-Net across all evaluation metrics (IoU: +3.7, Boundary IoU: +7.7, Dice: +2.1 percentage points), with the most substantial improvement in boundary-aware metrics. This confirms that the nested dense skip connection architecture effectively enhances boundary representation.

Second, the IoU metric exhibits significant structural bias, showing relative insensitivity to boundary perturbations compared to Boundary IoU. Our sensitivity analysis revealed that a 10-pixel boundary perturbation reduced Boundary IoU by 35.6 points while IoU decreased by only 6.1 points, quantitatively demonstrating this bias.

Third, among U-Net++ components, dense skip connections contribute most significantly to boundary quality improvement, as demonstrated by our ablation study. Removing this component caused a 7.3-point drop in Boundary IoU.

Based on these findings, we advocate for the inclusion of boundary-aware metrics in segmentation evaluation protocols. Relying solely on IoU may lead researchers to undervalue architectural improvements that specifically target boundary quality. The combination of IoU, Boundary IoU, and Dice Score provides a more complete picture of segmentation performance and should be adopted as a standard evaluation protocol.

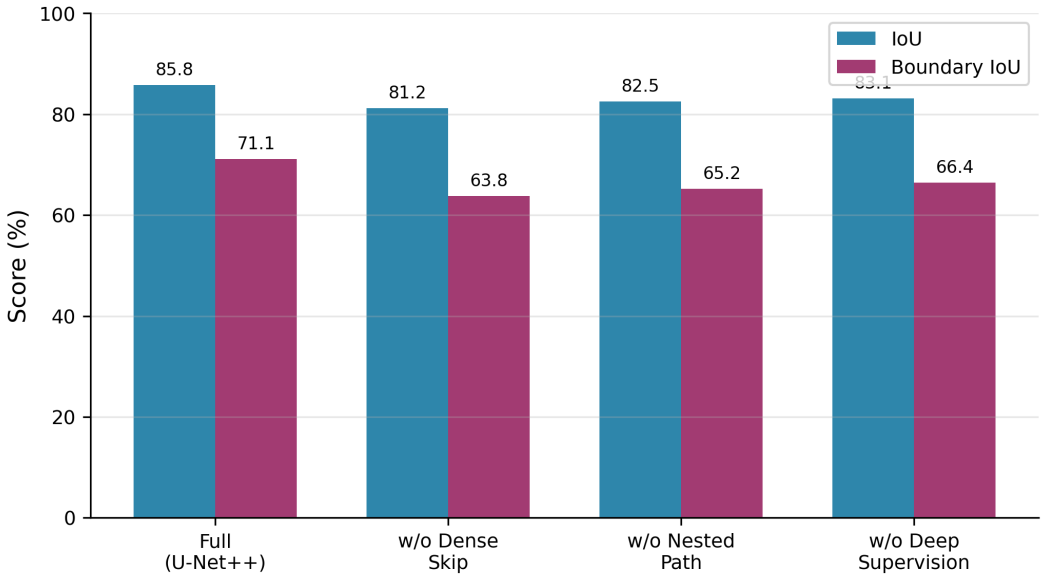

Figure 5: Ablation study results. Dense skip connections show the largest impact on performance, particularly for Boundary IoU.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 234-241.

[2] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *DLMIA*, 3-11.

[3] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *CVPR*, 3431-3440.

[4] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE TPAMI*, 40(4), 834-848.

[5] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. *CVPR*, 15334-15342.

[6] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *CVPR*, 4700-4708.

[7] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *NeurIPS*, 12077-12090.

[8] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. *CVPR*, 2881-2890.

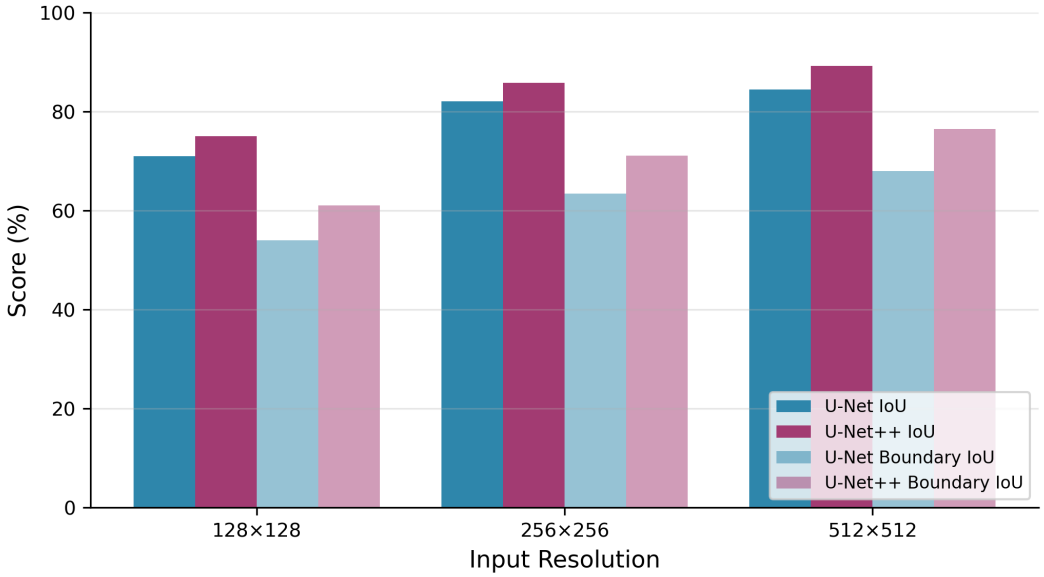

Figure 6: Performance comparison across different input resolutions. The gap between U-Net and U-Net++ increases with resolution, especially for Boundary IoU.