Efficient Fine-Tuning of SAM with LoRA for Nuclei Instance Segmentation: A Cross-Validation Study on the NuInsSeg Dataset

Anirudh Narasimha Bharadwaj

Abstract

This report presents the implementation and evaluation of the Segment Anything Model (SAM) fine-tuned with Low-Rank Adaptation (LoRA) for nuclei instance segmentation on the NuInsSeg dataset. Using a 5-fold cross-validation approach on 665 H&E-stained histological images, the model achieves an average Dice score of 0.8233, Aggregated Jaccard Index (AJI) of 0.6197, and Panoptic Quality (PQ) of 0.5724. Training leverages two Tesla H100 GPUs, with LoRA reducing trainable parameters to 1.04% (6,686,000 out of 643,712,048). Training loss decreases steadily across epochs, but validation metrics show variability, indicating sensitivity to data distribution. Visualizations align well with ground truth, though challenges in segmenting small or overlapping nuclei are evident. The report discusses training dynamics, metric trends, generalizability issues, and proposes improvements such as advanced augmentation and hyperparameter optimization to enhance performance for histopathological applications.

Contents

1 Introduction

Nuclei instance segmentation in H&E-stained histological images is a pivotal task in medical image analysis, aiding in cancer diagnosis and research. The NuInsSeg dataset provides a benchmark for this task, comprising 665 fully annotated samples [**?**]. This assignment implements the Segment Anything Model (SAM) [**?**], fine-tuned with Low-Rank Adaptation (LoRA) [**?**] to perform efficient nuclei instance segmentation, following the experimental setup outlined in the dataset paper.

SAM is a foundation model designed for universal segmentation tasks, but its large parameter count (over 600 million) makes full fine-tuning computationally expensive. LoRA enables parameter-efficient fine-tuning by adapting a small subset of weights, making it suitable for medical imaging tasks with limited computational resources. This report details the methodology, including data preprocessing, model architecture, training setup, and evaluation metrics (Dice, AJI, PQ). Results are analyzed through quantitative metrics, training dynamics, and qualitative visualizations across all folds. The study also explores generalizability challenges and proposes strategies to improve performance, aiming to contribute to reliable nuclei segmentation in clinical settings.

2 Methodology

2.1 Dataset

The NuInsSeg dataset [**?**] comprises 665 H&E-stained histological images, each containing:

- **Input**: 2D RGB images of varying dimensions, typically 1000×1000 pixels.
- **Labels**: Instance segmentation masks where each nucleus is uniquely identified with a distinct integer value (0 for background).

Data preprocessing includes resizing images to 1024×1024 pixels, normalizing pixel values to [0, 1], and applying augmentations: rotations (range 0.3 radians), flips (probability 0.5), and additive noise (probability 0.2). The dataset is split into 5 folds for cross-validation, with a random seed of 42.

2.2 Model Architecture

The Segment Anything Model (SAM) with the ViT-H backbone (checkpoint: `sam_vit_h_4b8939.pth`) is used, consisting of:

- **Image Encoder**: A Vision Transformer (ViT) with 32 blocks, processing 1024×1024 images into 256-dimensional embeddings.
- **Prompt Encoder**: Encodes prompts (points, boxes, or masks) for guided segmentation.
- **Mask Decoder**: A transformer-based decoder that generates segmentation masks from image embeddings and prompts.

LoRA is applied to the attention layers of the image encoder (`qkv` projections across all 32 blocks) and the mask decoder's transformer layers. LoRA parameters are set as:

- Rank (`lora_r`): 16

- Alpha (`lora_alpha`): 32
- Dropout (`lora_dropout`): 0.1

This configuration results in 6,686,000 trainable parameters (1.04% of the total 643,712,048 parameters), significantly reducing computational overhead.

2.3 Training Setup

- **Framework**: PyTorch, executed on Newton Cluster's Tesla H100 GPUs (2 CUDA-enabled GPUs).
- **Cross-Validation**: 5-fold with random shuffling (seed 42).
- **Optimizer**: Adam with a learning rate of 0.0001.
- **Batch Size**: 4, with 60 workers for data loading.
- **Epochs**: 5 per fold.
- **Loss**: Binary Cross-Entropy with logits, optimized for instance segmentation.
- **Validation**: Performed every epoch.

2.4 Evaluation Metrics

- **Dice Coefficient**: Measures overlap between predicted and ground truth masks:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}$$

  where P and G are the predicted and ground truth masks, respectively.

- **Aggregated Jaccard Index (AJI)**: Evaluates instance segmentation by matching predicted and ground truth instances, computing the Jaccard Index over matched pairs, and penalizing unmatched instances.
- **Panoptic Quality (PQ)**: Combines segmentation quality (Dice) and detection quality (matching accuracy):

$$\text{PQ} = \frac{\sum_{\text{matched}} \text{IoU}}{|\text{matched}|} \times \frac{|\text{matched}|}{|\text{matched}| + 0.5|\text{unmatched pred}| + 0.5|\text{unmatched gt}|}$$

3 Results

3.1 Quantitative Results

The model was evaluated across five folds, with the final validation metrics shown in Figures 1, 2, and 3. The metrics are summarized as follows:

- Fold 1: Dice 0.8109, AJI 0.5886, PQ 0.5531
- Fold 2: Dice 0.8162, AJI 0.5944, PQ 0.5488

- Fold 3: Dice 0.8529, AJI 0.6808, PQ 0.6189
- Fold 4: Dice 0.8198, AJI 0.6270, PQ 0.5623
- Fold 5: Dice 0.8169, AJI 0.6364, PQ 0.5789

The average Dice score across folds is 0.8233, AJI is 0.6197, and PQ is 0.5724.

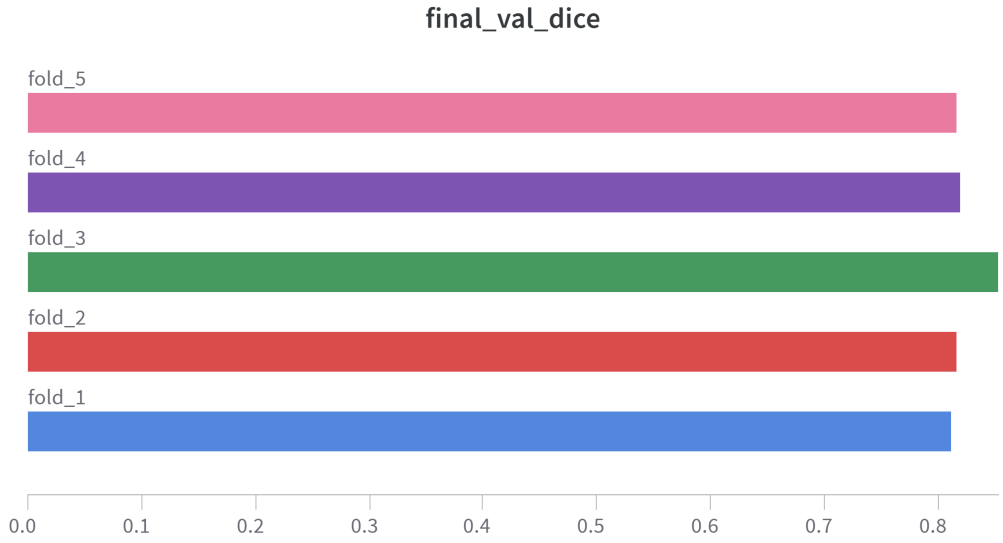

Figure 1: Final validation Dice scores across all folds.

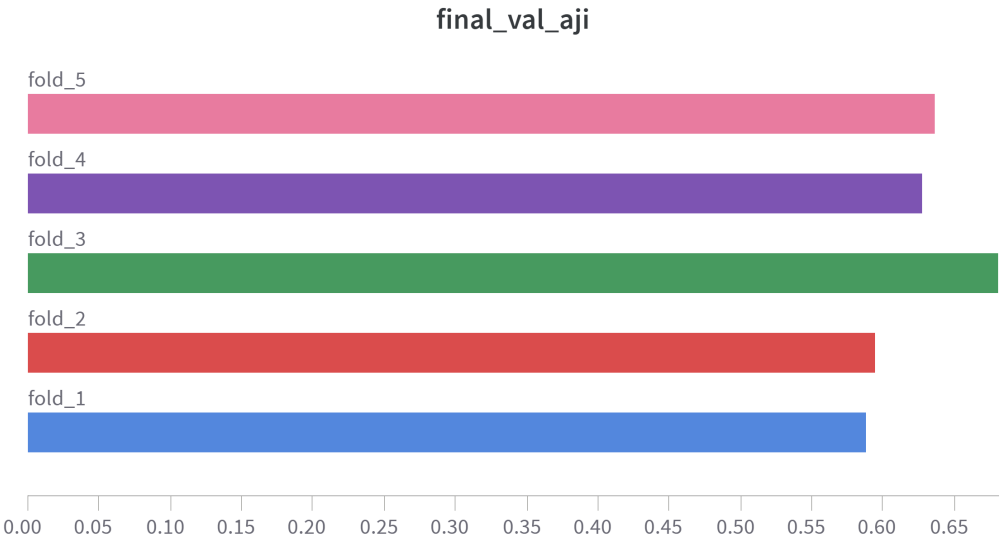

Figure 2: Final validation AJI scores across all folds.

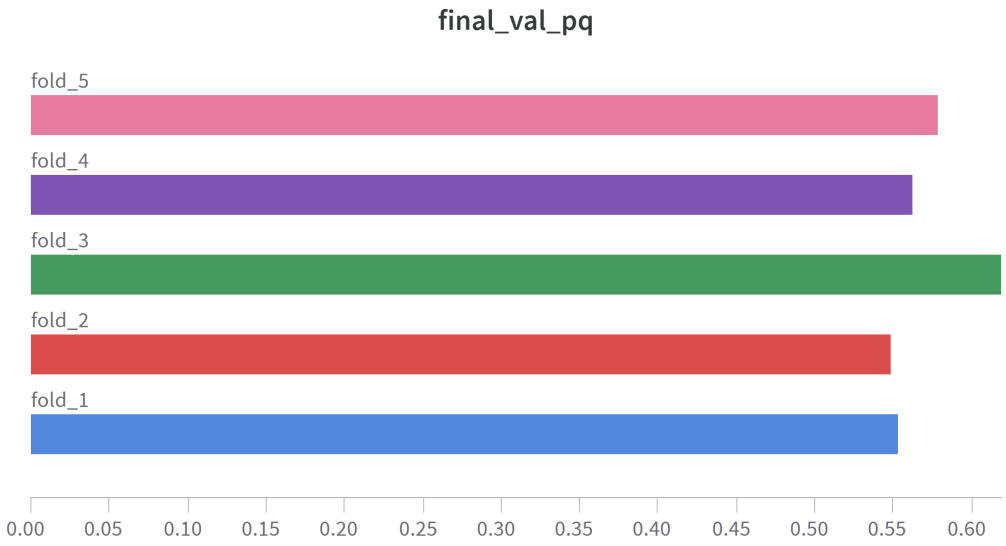

Figure 3: Final validation PQ scores across all folds.

3.2 Qualitative Results

Figure 4 visualizes an initial sample from the dataset, confirming data integrity and label distribution. Figures 5 to 9 compare predictions against ground truth for each fold, showing the input image, ground truth, and predicted segmentation masks.

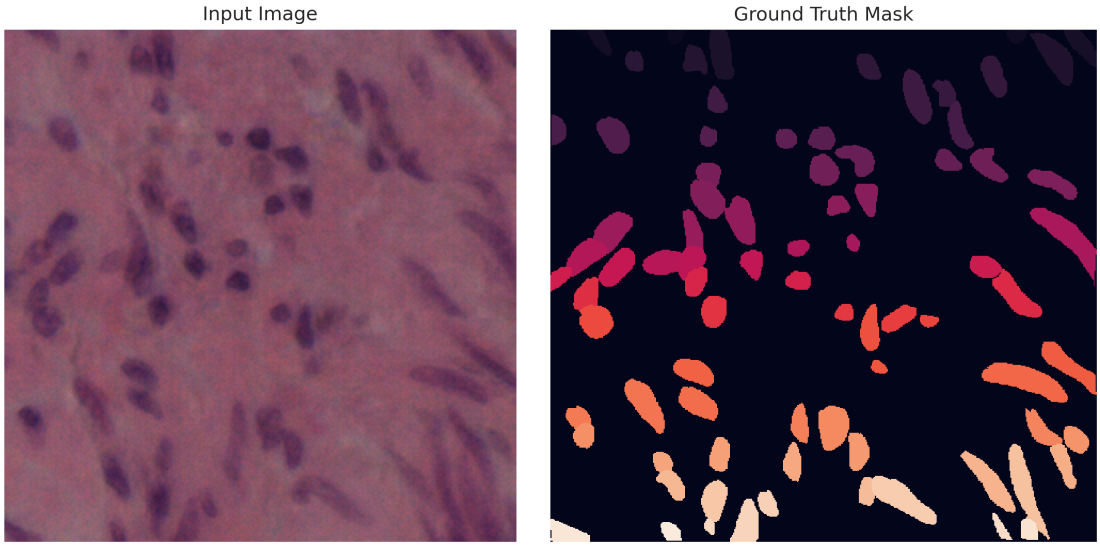

Figure 4: Sample visualization of an H&E-stained image and its ground truth mask.

Figure 5: Prediction vs. ground truth for Fold 1.

Figure 6: Prediction vs. ground truth for Fold 2.

Figure 7: Prediction vs. ground truth for Fold 3.

Figure 8: Prediction vs. ground truth for Fold 4.

Figure 9: Prediction vs. ground truth for Fold 5.

3.3 Training Dynamics

Training loss decreases consistently across folds, starting at approximately 0.0073–0.0078 and stabilizing around 0.0025–0.0034 by epoch 5. For example, in Fold 4, the loss drops from 0.0078 (epoch 1) to 0.0034 (epoch 5), a 56% reduction, indicating effective optimization but suggesting potential early convergence, as the loss plateaus after epoch 3 in most folds.

Validation metrics (Dice, AJI, PQ) across training steps are shown in Figures 10, 11, and 12. The metrics improve over steps but exhibit significant fluctuations. For instance, in Fold 3 (green line), the Dice score peaks at 0.8536 around 2k steps, then slightly dips to 0.8529 by the end. AJI and PQ follow similar trends, with Fold 3 achieving the highest values (AJI: 0.6808, PQ: 0.6189). However, Fold 1 (blue line) shows the lowest performance, with AJI and PQ starting around 0.55 and only reaching 0.5886 and 0.5531, respectively. The variability across folds and steps suggests sensitivity to data distribution and potential overfitting.

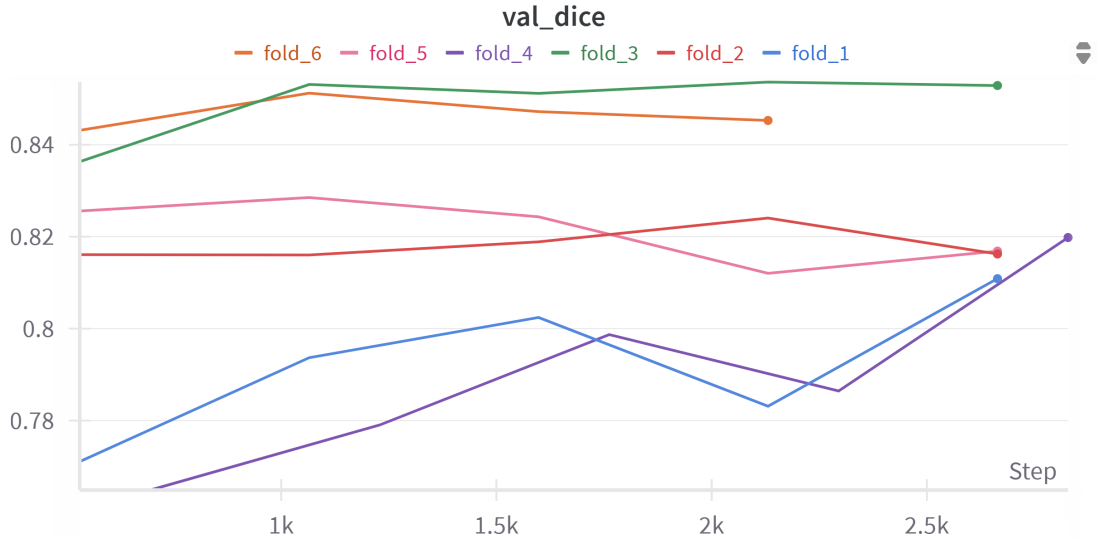

Figure 10: Validation Dice scores across training steps for all folds.

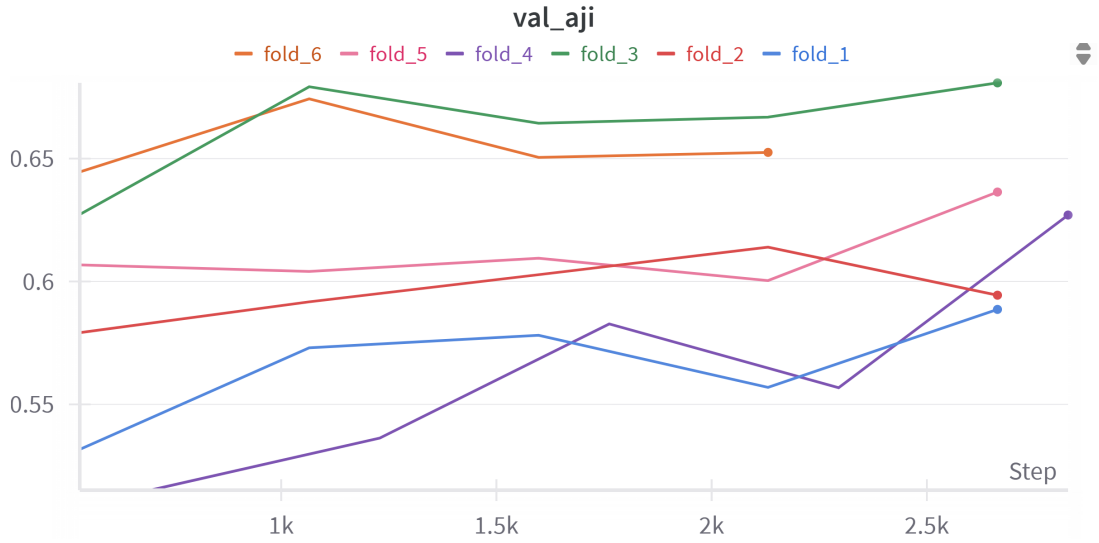

Figure 11: Validation AJI scores across training steps for all folds.

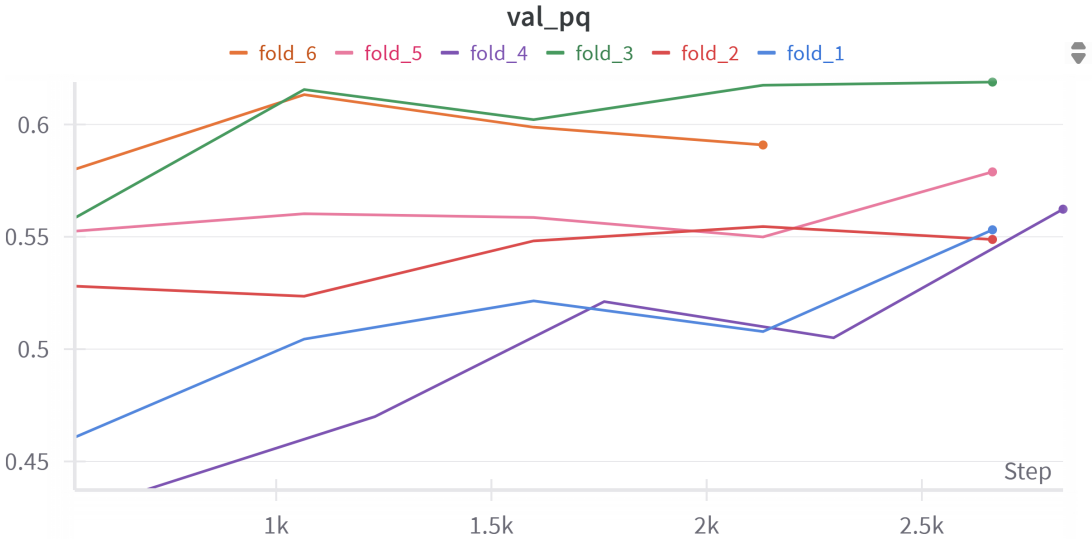

Figure 12: Validation PQ scores across training steps for all folds.

4 Discussion

4.1 Analysis of Loss and Metrics

The steady decline in training loss across all folds demonstrates effective optimization, but the plateauing after 3–4 epochs suggests the model may be reaching a local minimum. The low learning rate (0.0001) ensures stable convergence but may limit the model's ability to escape local minima, as evidenced by the small improvements in validation metrics after epoch 3 in most folds.

Dice scores indicate strong overlap (average 0.8233), but AJI and PQ reveal challenges in instance segmentation. The AJI (average 0.6197) is lower due to difficulties in matching small or overlapping nuclei, as seen in Fold 1 (AJI: 0.5886). PQ (average 0.5724) reflects a balance between segmentation and detection quality, with Fold 3 achieving the highest score (0.6189), likely due to better handling of diverse nuclear shapes. The line plots (Figures 10, 11, 12) show significant fluctuations, particularly in Folds 1, 2, 4, and 5, where metrics drop sharply at certain steps. This instability suggests that the model is sensitive to the learning rate or data distribution.

The bar plots (Figures 1, 2, 3) confirm the variability across folds, with Fold 3 consistently outperforming others, while Fold 1 lags behind, particularly in AJI and PQ. This aligns with the qualitative results, where Fold 1 struggles with small nuclei.

Also to be noted, due to the Cluster going down abruptly, the training had to begin from Fold 4 after a break.

Visualizations (Figures 5 to 9) show good alignment with ground truth, but the model occasionally struggles with small or densely packed nuclei, leading to under-segmentation or merging of instances. This is particularly evident in Fold 1, where the predicted mask misses several small nuclei present in the ground truth.

4.2 Generalizability

The 5-fold cross-validation ensures robust evaluation, but the variability in metrics (e.g., Dice from 0.8109 to 0.8529) suggests limited generalizability. This could stem from:

- **Data Distribution**: Differences in nuclei density, staining variations, or tissue types across folds.

Generalizability to unseen datasets (e.g., different staining protocols or tissue types) may be limited due to the model's sensitivity to training data characteristics.

4.3 Improvements

To enhance performance and generalizability, the following strategies are proposed:

1. **Hyperparameter Tuning**:
    - Increase `max_epochs` (e.g., to 10) to allow further convergence, as the loss plateau suggests undertraining.
    - Experiment with a higher learning rate (e.g., 0.0005) to escape local minima.
    - Tune LoRA rank (e.g., to 32) and alpha (e.g., to 64) to improve model adaptability.
2. **Advanced Augmentation**:
    - Increase `rotation_range` (e.g., to 0.5 radians) and `flip_prob` (e.g., to 0.7) to improve robustness.
    - Introduce color jittering to handle staining variations.
3. **Model Architecture**:
    - Apply LoRA to additional layers (e.g., MLP layers in the image encoder) to capture more features.
4. **Loss Function**:
    - Combine Binary Cross-Entropy with Dice Loss to directly optimize for overlap, addressing segmentation errors.
    - Apply instance-aware weighting to prioritize small or overlapping nuclei.

5 Conclusion

The SAM model, fine-tuned with LoRA, achieves an average Dice score of 0.8233, AJI of 0.6197, and PQ of 0.5724 across 5 folds, demonstrating strong performance in nuclei instance segmentation on the NuInsSeg dataset. Training loss decreases steadily, but validation metric fluctuations indicate challenges in stability and instance matching. Qualitative visualizations show reasonable alignment with ground truth, though improvements are needed for small and overlapping nuclei. Generalizability is limited by data distribution sensitivity. Future work should focus on hyperparameter optimization and advanced augmentation to enhance clinical applicability.

6 References

- Mahbod, Amirreza, et al. "NuInsSeg: A fully annotated dataset for nuclei instance segmentation in H&E-stained histological images." *Scientific Data*, vol. 11, no. 1, 2024, p. 295. https://doi.org/10.1038/s41597-024-03074-7
- Kirillov, Alexander, et al. "Segment Anything." *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026. https://arxiv.org/abs/2304.02643
- Hu, Edward J., et al. "LoRA: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685*, 2021. https://arxiv.org/abs/2106.09685