SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Adaptive Multi-Loss Learning with Self-Supervised Fine-Tuning for Robust Humerus Segmentation on Radiographs

Anthony Wu, MS, University of California, Irvine; Arya Amirhekmat, MD; Pooya Khosravi, MS; Paramveer Birring; Jason Liang; Maryam Golshan-Momeni, MD; Peter Chang, MD; Roozbeh Houshyar, MD; James Learned, MD

Introduction/Background

Automated humerus segmentation on radiographs is critical for real-time orthopedic AI applications but remains largely unexplored due to anatomical overlap, occlusion, and variability in image quality. While segmentation on CT has been widely studied, axillary shoulder radiographs present unique challenges that demand robust and interpretable solutions. Herein we propose a humerus segmentation framework based on a modified MobileNetV3-Large architecture that runs in real-time.

Methods/Intervention

Ground truth segmentations were annotated on 286 axillary radiographs from the Stanford MURA dataset by a senior orthopedic resident and two medical students. For benchmarking, we perform 5-fold cross validation on binary cross-entropy (BCE), Dice, and focal losses independently, then introduced a dynamic multi-loss formulation incorporating boundary loss to better penalize edge errors. To enhance performance on difficult cases, we implemented two self-supervised strategies during fine-tuning: 1) Hard-sample mining that prioritizes difficult cases such as unclear glenohumeral joint spaces and 2) Pixel-level contrastive loss with pseudo-negatives that directly penalizes false-positive pixels on adjacent bones. Dice score and Hausdorff distance were used as primary evaluation metrics (Table 1), with visualizations of segmentation quality (Fig 1) and per-image performance variability (Fig 2).

Results/Outcome

Baseline BCE loss suffered from class imbalance, resulting in under-segmentation and reduced Dice scores. Focal loss, while designed to address this, showed instability across folds due to the static α parameter, which struggled with the wide variability in humerus-to-background ratios. Dice loss, while better at optimizing overlap, under-penalized edge errors—critical in orthopedic contexts (Fig. 1, Fig. 2). Dynamic incorporation of boundary loss consistently outperformed single-loss formulations, with dynamic dice/boundary loss achieving >0.94 dice (Table 1). Self-supervised hard-sample mining and contrastive learning improved robustness on challenging images and enhanced boundary precision, effectively reducing Hausdorff Distance while maintaining high dice (Table 1, Fig. 1). These enhancements also reduced performance variance, improving generalization on challenging images (Fig. 2).

Conclusion

Combining adaptive loss strategies with self-supervised fine-tuning significantly improves segmentation performance and reliability on axillary shoulder radiographs.

Statement of Impact

This is one of the first systematic studies on humerus segmentation from radiographs. Our approach improves boundary precision and generalization, offering a deployable solution for real-time orthopedic imaging

workflows.

Table 1a

Model Hyperparameters	
Backbone Initial LR	0.0001
Segmentation Head Initial LR	0.01
Validation Evaluation Metric	Dice
Epoch Limit	499
Random State	42
Scheduler Patience	10
Scheduler ____value	0.5
Optimizer	AdamW
Minimum Learning Rate	1e-8
Early Stop Patience	100 Epochs
Loss Weighting Method	Uncertainty based Dynamic Loss Weighting
Fine Tuning Specific Hyperparameters	
Validation evaluation metric	Hausdorff
Epoch Limit	200
Early Stop Patience	50 Epochs
Backbone Initial LR	1e-5
Fine tuning Head LR	1e-6
Minimum Learning Rate	1e-8

Table 1b

Training Augmentations	
Augmentation Type	**Hyperparameters**
Shift	Limit = 0.06 P=0.75
Scale	Limit = 0.1 P=0.75
Rotate	Limit = 15 P=0.75
Horizontal Flip	P=0.5
Elastic Transform	α=120 β=120*0.05 P=0.5
Coarse dropout	Number of holes=[1,4] Hole Height Range=[0.01,0.1] Hole Width Range=[0.01,0.1] P=0.05
Random Brightness Contrast	P=0.3

Table 1c

Humeral Segmentation Results	DICE (↑)		HAUS (↓)	
	Mean	**Std**	**Mean**	**Std**
BCE	0.904	0.026	16.1	3.3
BCE+Boundary	0.916	0.020	14.6	1.9
Focal	0.888	0.039	18.2	6.1
Focal+Boundary	0.916	0.016	14.0	3.0
Dice	0.918	0.015	15.0	1.2
Dice+Boundary	**0.942**	0.005	11.4	2.1
BCE+Dice+Boundary	0.912	0.020	15.8	2.1
Dice + Boundary + Self-supervised Fine Tuning	**0.942**	0.005	**11.3**	2.0

Table 1a. All models were implemented in PyTorch and trained using identical base hyperparameters, except during the fine-tuning stage, which incorporated a self-supervised contrastive loss. Unlike the initial training phase—which monitored the Dice score—the fine-tuning phase optimized directly for the Hausdorff distance. This reflects the objective of the self-supervision strategy, which introduces pseudo-negative labels to explicitly penalize over-segmentation. Since Hausdorff distance is more sensitive to spatial outliers and boundary errors, it serves as a more appropriate optimization target for this task. Table 1b. Summary of training augmentation techniques. All augmentations were implemented using the Albumentations library. Each augmentation was applied probabilistically, with probability values (P) listed per transformation. Parameters for spatial augmentations are defined relative to image dimensions: the limit for shift, scale and hole height/width cutout operations are defined as a fractional range of image width. Rotations are specified in degrees. Augmentations were applied only to the training data within each cross-validation fold. All data were uniformly resized to 256×256 pixels. Table 1c. Summary of final validation evaluation metrics. Evaluation metrics include Dice score and Hausdorff distance (HD). The Dice score quantifies pixel-wise overlap between predicted and ground truth segmentations, with a value of 1 indicating perfect alignment. In contrast, HD measures the maximum Euclidean distance between predicted and actual object boundaries—making it particularly informative for assessing boundary accuracy and shape integrity. Lower HD values indicate better performance; for reference, an HD of X reflects a maximum boundary deviation of X pixels. Models trained with Boundary loss consistently outperformed those trained with a single loss function, yielding higher Dice scores, reduced Dice score variability, and lower mean Hausdorff distances—collectively indicating enhanced accuracy and consistency, particularly at the edges. Among pre–fine-tuning models, the Dice+Boundary configuration achieved the best performance. Self-supervised fine-tuning of this model yielded further reductions in Hausdorff distance, though improvements were modest and Dice scores remained largely unchanged. The three-term BCE+Dice+Boundary model performed intermediately, likely due to competitive rather than synergistic interactions among the loss terms, which may have diluted the boundary-enforcing effect. Bolded values indicate the best-performing model across all metrics.

Figure 1

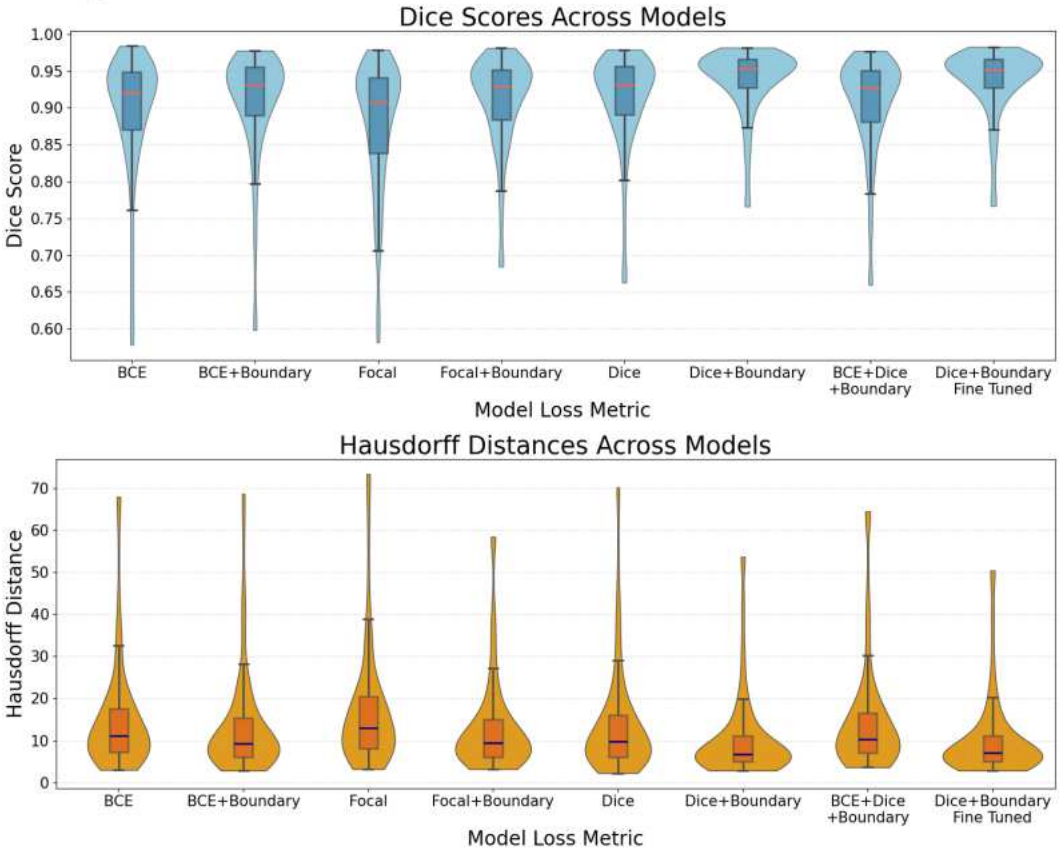

Figure 1. Humerus Segmentation Performance by Loss Function. Violin plots show the distribution of Dice scores and Hausdorff distances for each model variant. For readability, violin plot values are clipped to within three standard deviations of the mean. Adding a Boundary loss consistently improved performance over single-loss models, yielding higher Dice scores, lower Hausdorff distances, and reduced variance. While fine-tuning the Dice+Boundary model had minimal effect on Dice scores, it slightly reduced extreme Hausdorff outliers. The BCE+Dice+Boundary model exhibited intermediate performance across both metrics. The distribution plots suggest that the addition of BCE may have counteracted the synergy between Dice and Boundary losses, leading to Dice and Hausdorff performances to be that of between BCE-only and DICE-only models. Models trained with Focal loss exhibited markedly higher variance in both metrics, likely reflecting the sensitivity of the fixed α parameter to large inter-sample variability in humeral size across radiographs.

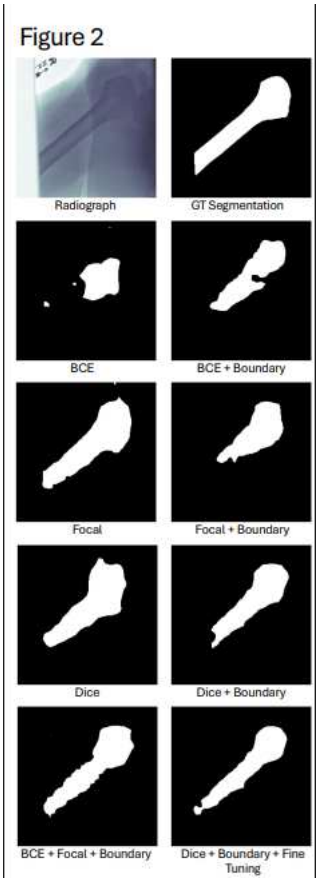

Figure 2. Qualitative comparison of humerus segmentation across models trained with different loss functions. A challenging validation sample was selected based on large discrepancies in Dice score and Hausdorff distance across models. Inference was performed using all saved models from the same fold, which had only seen this sample during validation. This sample exhibits substantial occlusion over the lateral humeral head due to overlying soft tissue—likely subcutaneous fat—resulting in a pronounced intensity discontinuity across the humerus. The baseline BCE model struggled with this variability, segmenting the medial portion of the humeral head but abruptly failing at the lateral intensity drop-off caused by thicker, soft tissue. Incorporating Boundary loss consistently improved shape fidelity of both the humeral shaft and humerus head. The Dice-only model achieved decent shaft segmentation, but adding Boundary loss further refined the shape of both the shaft and the humeral head. Interestingly, while the inclusion of Boundary loss in the Focal loss model significantly refined the humeral head shape in this sample, it came at the cost of degraded shaft segmentation. This trade-off however was not representative of most cases, as models incorporating Boundary loss showed lower variance and better overall performance (Table 1b, Figure 1). The BCE+Dice+Boundary model recovered the humerus shaft, but the contour appears jagged and irregular—suggesting competing optimization pressures between loss terms. Finally, the fine-tuning the Dice+Boundary model with self-supervised contrastive learning yielded a smoothing effect on the articular surface and improved shaft continuity, demonstrating improved generalization and spatial coherence in challenging anatomical contexts.

Keywords

Machine Learning; Humerus Segmentation; Self-supervised Learning; Contrastive Learning; Radiography; Musculoskeletal Radiology