OVARIAN SEGMENTATION IN T2-WEIGHTED FAT-SATURATED MRI USING ATTENTION U-NET AND UNILATERAL MASKING

Esther Helga Klemenzardóttir
Bioinformatics Research Centre, Aarhus University
8000 Aarhus C, Denmark
202007451@post.au.dk

Anders Lydig Kristensen
Bioinformatics Research Centre, Aarhus University
8000 Aarhus C, Denmark
201808522@post.au.dk

December 5, 2025

ABSTRACT

We investigated the segmentation of ovarian tissue in T2-Weighted Fat-Saturated MRI using the UT-EndoMRI dataset ($N = 37$ patients, yielding $S = 177$ annotated 2D slices). The main challenge addressed is "Single-Label Ambiguity," where patients possess bilateral anatomy (two ovaries) but the ground truth masks annotate only one, creating contradictory supervision signals. To fix this, we implemented a Unilateral Masking strategy that geometrically isolates the target anatomy. We compared three approaches: a Standard U-Net, an Attention U-Net, and a Transfer Learning model using a pre-trained ResNet34 encoder. Our experiments reveal that the Attention U-Net combined with Unilateral Masked data achieves a mean volumetric Dice score of 0.5185, outperforming the baseline and exceeding reported inter-rater reliability (≈ 0.48). In contrast, we found that Transfer Learning resulted in severe overfitting due to the small dataset size, and intensity-based preprocessing (RAovSeg) failed because the bi-modal intensity of fluid-filled follicles overlaps with background noise. These findings demonstrate that for this specific domain, resolving geometric data conflicts is more effective than increasing model complexity.

Keywords ovarian segmentation · T2-weighted MRI · Attention U-Net · medical image segmentation · deep learning

1 Introduction

Endometriosis affects approximately 190 million women, necessitating precise ovarian segmentation from MRI for surgical planning. However, this task remains notoriously difficult; inter-rater agreement is significantly lower (Dice ≈ 0.48) than for other pelvic organs, and current automated baselines struggle to exceed this "human ceiling" [1].

Motivated by the limited success of existing pipelines on the UT-EndoMRI dataset, we aim to systematically investigate why standard Deep Learning models fail to confidently localize ovarian tissue in T2-Weighted Fat-Saturated sequences [1, 2]. Our objective is to start with a standard U-Net to diagnose specific failure modes—whether anatomical, textural, or signal-based—and subsequently develop targeted interventions to address them.

2 Related Work

Since its introduction, the U-Net architecture [3] has become the de facto standard for biomedical image segmentation due to its ability to localize structures effectively from sparse data. However, the segmentation of healthy or endometriotic ovaries presents a significantly higher challenge than other pelvic organs. Liang and Giancardo [2] shows that while inter-rater agreement for the uterus is relatively high (Dice ≈ 0.73), agreement for ovaries drops significantly (Dice ≈ 0.48). This difference highlights the inherent ambiguity of ovarian boundaries in MRI, establishing a "human performance ceiling" that is much lower than typical segmentation tasks.

While recent works such as Wang et al. [4] have achieved high accuracy in segmenting Epithelial Ovarian Cancer (EOC) on T2-weighted images, these tasks generally involve large, distinct tumor masses. In contrast, segmenting anatomical ovaries in patients with endometriosis—where the organ may be deformed, small, or obscured by lesions—remains an open challenge requiring specialized architectural and data engineering interventions.

To address the difficulty of detecting small target structures among heterogeneous background tissue, Oktay et al. [5] introduced Attention U-Net. By integrating Attention Gates (AGs) into the skip connections, the network learns to suppress feature responses in irrelevant background regions (e.g., abdominal fat) while boosting the signal of the target organ. This mechanism is particularly relevant for T2-weighted Fat Saturated (T2FS) sequences, where hyperintense

blood vessels, fat or fluid pockets often mimic the intensity profile of ovarian follicles, acting as distractors for standard CNNs.

A unique challenge in our dataset is the presence of bilateral anatomy (two ovaries) paired with single-label ground truth. Training on such data typically confuses models, as valid features on the non-annotated side are penalized as false positives. To resolve this, we adopt a Unilateral Masking strategy similar to Cardenas et al. [6], who utilized "unilateral input data" to segment lymph node clinical target volumes. By masking the unannotated side of the MRI slice, we align the input signal with the available ground truth without compromising the anatomical validity of the segmentation task.

The most direct precedent for our work is the RAovSeg pipeline proposed by Liang and Giancardo [2] for the UT-EndoMRI dataset. Their approach utilizes a multi-stage pipeline involving a ResNet classifier to select slices followed by an Attention U-Net for segmentation. Importantly, their method relies on a specific intensity-based preprocessing step that inverts high-intensity pixels and boosts mid-intensity ranges, based on the assumption that ovaries appear as hypointense (grey) structures relative to hyperintense (white) fat. We investigate whether this intensity assumption holds for T2FS sequences, where fluid-filled ovarian follicles often exhibit hyperintense signals that overlap with the intensity range of the background noise.

3 Methods

The source code for this project is available at: `https://github.com/Lydig/dlvr-project`. The data can be found at: `https://zenodo.org/records/15750762`.

3.1 Dataset and Preprocessing

We used the UT-EndoMRI D2 Dataset [2], consisting of Pelvic MRI scans. We specifically selected the T2-weighted Fat Saturated (T2FS) sequences, as they provide high contrast for fluid-filled structures like ovarian follicles.

Cohort Selection: The raw dataset contains 73 patients with varying pathologies. To establish a feasible scope for semantic segmentation, we applied strict inclusion criteria: patients needed **(i)** data completeness, meaning matching T2FS volumes and standard ovary masks (`_ov`); **(ii)** absence of pathology such as endometriomas (`_em`) or cysts (`_cy`), which were excluded to focus on standard ovarian tissue segmentation; and **(iii)** successful quality control, requiring removal of subjects with gross label mismatches or cases where the ovary crossed the anatomical midline. After filtering, the final cohort included $N = 37$ patients.

Image Metadata and Dimensionality: The raw MRI volumes typically have dimensions of $480 \times 480 \times 32$. The voxel spacing is anisotropic ($0.5 \times 0.5 \times 5.0$ mm). Since the Z-axis resolution is $10\times$ lower than the in-plane resolution, treating the data as 3D volumetric samples is suboptimal. Therefore, we treat the problem as 2D slice-wise segmentation [1]. We extracted all 2D slices containing positive ovary annotations, resulting in a total of 177 positive slices ($S = 177$).

Experimental Preprocessing Techniques: While the final model utilized standard min-max normalization, we experimented with two advanced preprocessing techniques to address the intensity ambiguity of T2FS images (see Figure 1).

RAovSeg Intensity Transformation: Following the methodology of Liang et al. [1], we implemented a non-linear intensity mapping designed to separate ovarian tissue from fat. The transformation assumes ovaries occupy a mid-intensity range and applies the following logic to the normalized image I:

$$I_{new}(x) = \begin{cases} 1.0 & \text{if } 0.22 \leq I(x) \leq 0.30 \\ 1.0 - I(x) & \text{if } I(x) > 0.5 \\ I(x) & \text{otherwise} \end{cases} \tag{1}$$

This functions by inverting high-intensity pixels (suppressing fat) and boosting mid-range pixels (highlighting solid ovarian tissue), which is hypothesized to make the ovaries stand out more.

CLAHE (Contrast Limited Adaptive Histogram Equalization): To enhance local texture features rather than global intensity, we applied CLAHE [7]. This algorithm partitions the image into contextual tiles and applies histogram equalization locally. We utilized a clip limit of 0.03 to prevent the amplification of noise in homogeneous regions.

3.2 Stratified Data Splitting

Given the small dataset size ($N = 37$, $S = 177$) and high variance in ovarian visibility, a random split risks creating "empty" folds with little information. We implemented a stratified split based on information density. Patients were

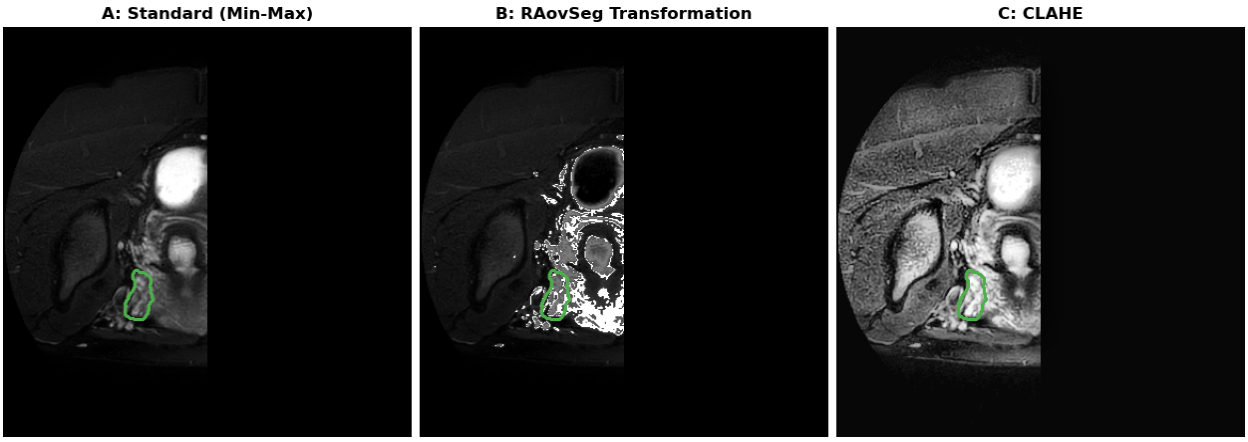

Figure 1: Visualizing preprocessing effects on Patient D2-051 (Slice 21). The green contour denotes the ground truth ovary. **(A) Standard:** Min-max normalization preserves the original intensity distribution. **(B) RAovSeg:** The transformation successfully suppresses the hyperintense abdominal feature (top center), turning it black. In this specific slice, the ovarian tissue falls within the target mid-intensity range and is preserved. **(C) CLAHE:** Local contrast enhancement was used to shift the learning signal away from absolute pixel intensity and towards recognizable textures, shapes, and edges.

sorted by their total number of positive ovary slices and assigned to 5 folds in a round-robin fashion. This ensures that every fold contains an approximately equal number of positive samples (mean ≈ 4.8 slices per patient).

The Validation Set was used exclusively for model selection, hyperparameter tuning, and qualitative assessment to guide architectural decisions. The Test Set was strictly held out and accessed only for the final performance reporting, ensuring that no design decisions were biased by test data.

3.3 Data Engineering: The Unilateral Masking Method

A critical issue identified during exploratory analysis was SLA. Patients typically possess two ovaries, but the ground truth masks in this dataset annotate only one. Training a model on full slices creates a contradictory loss signal, as the model is penalized for correctly detecting the non-annotated ovary on the opposite side.

To resolve this, we engineered a unilaterally masked dataset. We looked at every MRI volume as having a left and a right hemisphere (defined by the midline). For training and inference, the hemisphere not containing the ground truth mask was strictly blacked out. We manually audited all 177 slices to verify that no anatomical structures were accidentally clipped. This Unilateral Masking procedure ensures that the input data aligns strictly with the single-label ground truth. Retaining the contradictory loss signal would compromise the reliability and interpretability of all downstream experiments. Performance differences could otherwise arise from supervisory inconsistency rather than meaningful variation in model behavior.

3.4 Network Architectures

We evaluated three variations of the U-Net architecture to handle the challenges of low-contrast ovarian tissue.

Baseline U-Net: We utilized the standard U-Net architecture [3] with a depth of 4. U-Net was chosen as the baseline because its encoder-decoder structure with skip connections allows for precise localization by combining high-level semantic features with low-level spatial details, which is critical for medical image segmentation where data is sparse.

The architecture consists of contracting (encoder) and expanding (decoder) paths. Each block in the encoder applies two consecutive 3×3 convolutions, each followed by batch normalization and a ReLU activation function, followed by 2×2 Max Pooling. The decoder utilizes 2×2 transposed convolutions for upsampling, concatenating features with the corresponding encoder level, followed by the same double convolution block. The final layer is a 1×1 convolution mapping the 64 feature channels to the binary output class.

Attention U-Net: To address the issue of the model being "distracted" by bright background tissue (e.g., abdominal fat), we integrated Attention Gates (AGs) into the skip connections [5]. AGs filter the features propagated from the encoder

using the decoder signal as a gating coefficient. This allows the model to suppress irrelevant background regions while boosting the feature response of small target structures before concatenation.

Transfer Learning (ResNet34): We hypothesized that the limited dataset size might prevent the model from learning complex textural filters from scratch. To test this, we replaced the standard encoder with a ResNet34 backbone pre-trained on ImageNet [8]. The input layer was modified to accept single-channel MRI inputs by repeating the weights across the channel dimension. We employed a two-phase training strategy: the encoder was frozen for the first 10 epochs to allow the decoder to learn initial weights without propagating large, unstable gradients into the pre-trained encoder. After 10 epochs, the full network was unfrozen for fine-tuning [9].

3.5 Training Configuration

Through a preliminary hyperparameter search, we established a set of parameters that produced stable convergence for the baseline model. These parameters were kept constant for all subsequent experiments to ensure comparability. All models were implemented in PyTorch and trained on a single NVIDIA GTX 1070 GPU. All models were optimized using AdamW (Weight Decay $1e^{-5}$) with a fixed learning rate of $1e^{-4}$ and a batch size of 12. Training proceeded for a maximum of 100 epochs, utilizing early stopping with a patience of 30 epochs. To mitigate overfitting on the small cohort, we applied on-the-fly data augmentation—specifically random rotations ($\pm 25^{\circ}$) and translations ($\pm 10\%$)—strictly to the training set.

Loss Functions: Our primary loss function was a composite of Binary Cross Entropy (BCE) and Dice Loss. The BCE component provides pixel-wise supervision, while the Dice component directly optimizes the overlap metric [10].

To address class imbalance, we implemented the Focal Tversky Loss [11]. We set a $\alpha = 0.7$ and $\beta = 0.3$ to penalize false positives more heavily than false negatives, and applied a focusing parameter of $\gamma = 1.33$ to down-weight easy background examples.

3.6 Evaluation Methodology

Our evaluation process was driven by a tight feedback loop between quantitative metrics and qualitative inspection. **i. Slice-wise Dice (Training):** Calculated on the validation set after every epoch. This metric was used to monitor convergence and trigger Early Stopping. **ii. Qualitative Inspection:** We manually inspected predictions on the validation set slice-by-slice. These observations (e.g., identifying specific false positive or false negative patterns) were the primary driver for our experiment choices. **iii. 3D Volumetric Dice (Reporting):** For the final evaluation, we reconstruct the full 3D volume from the 2D slice predictions of the test set and calculate the Dice score on the volume. This better reflects deployment conditions, where the clinically relevant question is how accurately the entire ovary is segmented rather than performance on individual 2D slices.

3.7 Final Robustness: 5-Fold Cross-Validation

To ensure statistical validity, our best performing model (based on quantitative and qualitative analysis) was evaluated using a stratified 5-fold cross-validation. We implemented two inference-time optimizations:

A. Dynamic Threshold Tuning: Instead of a fixed threshold (0.5), we performed a grid search ($\tau \in [0.3, 0.7]$) on the validation set of each fold to find the optimal decision boundary, which was then applied to the test set.

B. Post-Processing: We applied a Keep Largest Component filter. Since the Unilateral Masking method guarantees a maximum of one object per image, we label connected components in the 2D prediction mask and remove all but the largest one. This effectively removes scattered noise.

4 Results

4.1 Development Experiments and Model Selection

Our experimental process followed two distinct phases. First, we performed an iterative optimization (Exp 01 $\rightarrow$ Exp 03) to establish a viable configuration, sequentially addressing the SLA and sensitivity to small targets. In the second phase (Exp 04 $\rightarrow$ Exp 07), we used Exp 03 as the fixed baseline to test specific hypothesis-driven solutions to issues identified from evaluating the models. This resulted in the Attention U-Net on Unilateral Masking data (Exp 03) becoming our best performing model. Table 1 summarizes the quantitative performance of all seven development experiments. The Best Val 2D Dice represents the peak slice-wise performance used for model checkpointing, while the Test 3D Dice represents the clinical volumetric evaluation on the held-out test set (Fold 0 in this stratification).

Exp	Architecture	Data	Preprocessing	Loss	Best Val 2D Dice	Test 3D Dice (Mean ± SD)
01	U-Net	Original	Min-Max	Dice+BCE	0.5220	0.2518 ± 0.1330
02	U-Net	Unilateral	Min-Max	Dice+BCE	0.6601	0.4946 ± 0.1184
03	**Att. U-Net**	**Unilateral**	**Min-Max**	**Dice+BCE**	**0.6261**	$\mathbf{0.5041 \pm 0.1394}$
04	Att. U-Net	Unilateral	RAovSeg	Dice+BCE	0.4140	0.4403 ± 0.1320
05	ResNet34-UNet	Unilateral	Min-Max	Dice+BCE	0.3522	0.4943 ± 0.0995
06	Att. U-Net	Unilateral	CLAHE	Dice+BCE	0.5454	0.4060 ± 0.1702
07	Att. U-Net	Unilateral	Min-Max	Focal Tversky	0.5622	0.3578 ± 0.0868

Table 1: Quantitative results of all development experiments. Best slice-wise Dice is used for model checkpointing; Test 3D Dice reflects final volumetric performance on the held-out set.

The most significant quantitative improvement occurred between Exp 01 and Exp 02, where applying Unilateral Masking and aligning the input data with the single-label ground truth nearly doubled the Test 3D Dice ($0.25 \rightarrow 0.49$). Conversely, interventions aiming to use intensity (Exp 04) or local contrast (Exp 06) degraded performance, reducing the Test Dice to 0.44 and 0.40 respectively.

4.2 Training Dynamics

An analysis of the training curves reveals distinct behavioral differences between the approaches (Figure 2).

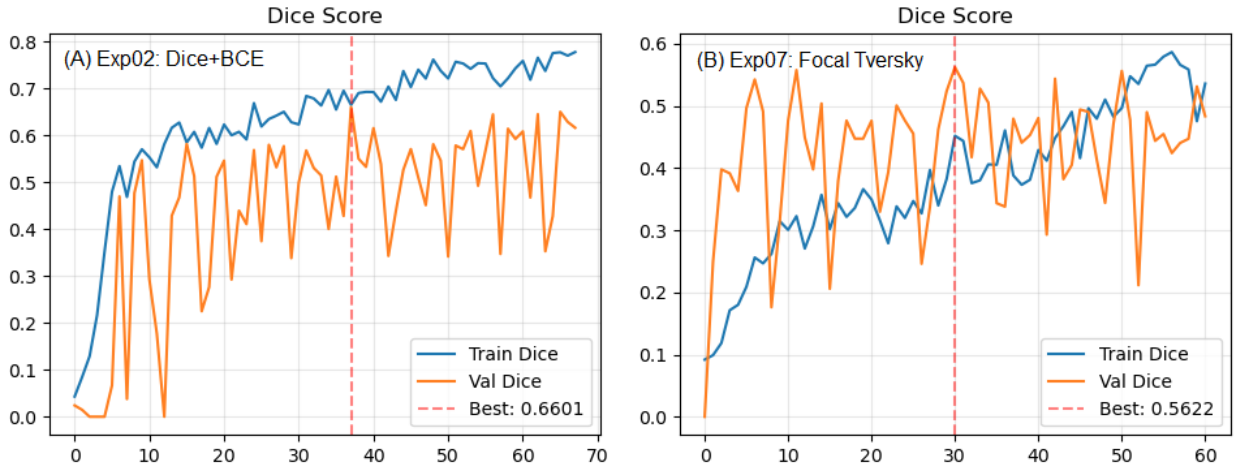

Figure 2: Training dynamics comparison. **(A) Exp 02 (Unilateral Masking Baseline):** The Validation Dice (orange) tracks the Training Dice (blue) closely, reaching a peak of ~ 0.66, indicating convergence and good generalization. **(B) Exp 07 (Focal Tversky):** Despite a stable Training curve (blue), the Validation Dice (orange) oscillates violently and fails to improve beyond 0.40. This confirms that the Focal loss function destabilized the optimization on this dataset.

In the Baseline experiments (Exp 02, 03), the validation Dice curve tracks the training curve closely, indicating effective generalization. However, in the Transfer Learning experiment (Exp 05), we observed severe overfitting: the Training Dice plateaus between 0.70 and 0.80, while the Validation Dice plateaued at 0.35, suggesting the pre-trained encoder memorized the small dataset rather than learning generalizable features. Furthermore, the Focal Tversky experiment (Exp 07) displayed jagged, unstable loss curves throughout training, confirming that the optimization landscape was too volatile for the complex loss function to navigate given the small sample size.

4.3 Qualitative Analysis

Visual inspection of the validation predictions provided key insights into the failure modes of the different approaches.

As shown in Figure 3A, Exp 01 frequently predicted the unlabelled ovary on the opposite hemisphere, confirming the "confusing signal" hypothesis. Exp 02 resolved this, but introduced a new issue: distinguishing ovaries from other hyperintense tissues (fat, vessels). Figure 3B illustrates the failure mode of intensity-based preprocessing. The RAovSeg transformation relies on ovaries falling within a specific mid-intensity range. While this successfully enhanced "grey" ovarian tissue, our T2FS sequences frequently contain hyperintense (bright white) fluid-filled follicles. The

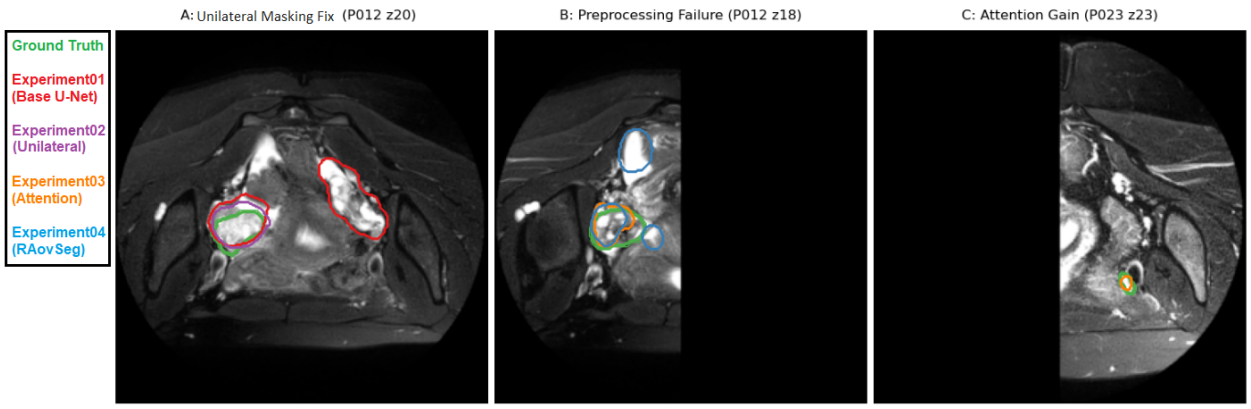

Figure 3: Qualitative comparison of experimental outcomes. **(A) Unilateral Masking Fix:** The Baseline model (Red) identifies two ovaries. The model trained with Unilateral Masking (Purple) correctly adheres to the single-label ground truth (Green), successfully eliminating the false positive ovary. For illustration purpose the unilateral masking is not shown on this picture, but is shown on B and C. **(B) Preprocessing Failure:** The RAovSeg model (Blue) segments the same ovary as the the Attention U-Net (Orange), but finds false positives in the bright areas. The RAovSeg model technically sees a different image, where these intense white areas are inverted into black by the RAovSeg preprocessing, however this is not shown on this figure. **(C) Sensitivity Gain:** The Standard U-Net (Purple - absent) misses the small anatomical target, whereas the Attention U-Net (Orange) successfully identifies it.

preprocessing inadvertently inverted these bright target regions, suppressing the signal alongside the fat it was intended to remove. This confirms that ovaries in this dataset exhibit a bi-modal intensity distribution that simple thresholding cannot resolve. Finally, while Exp 03 did not yield a massive quantitative leap, Figure 3C demonstrates an example of its superior sensitivity in detecting small, low-contrast ovaries that the Standard U-Net missed.

4.4 Final Robustness: 5-Fold Cross-Validation

Based on the development phase, the Attention U-Net on Unilateral Masked data was selected as the best performing architecture. To obtain a statistically stable performance estimate, we evaluated this configuration using Stratified 5-Fold Cross-Validation with inference-time optimization.

Fold	Optimal Threshold (τ)	Post-Processing	Test 3D Dice
0	0.7	On	0.5085
1	0.5	On	0.4787
2	0.6	On	0.4294
3	0.3	On	0.6887
4	0.7	On	0.4895
Mean	-	-	**0.5185**
SD	-	-	**0.0832**

Table 2: Results of the Stratified 5-Fold Cross-Validation. The optimization loop selected the "Keep Largest Component" post-processing step for all 5 folds. The high standard deviation (SD) and variance in optimal thresholds (τ) reflect the heterogeneity of the patient cohort.

The final mean 3D Dice score was 0.5185. Notably, the post-hoc optimization loop selected the "Keep Largest Component" filter for 100% of the folds (5/5), empirically confirming that noise reduction was universally beneficial. However, the optimal probability threshold τ varied significantly between folds (ranging from 0.3 to 0.7), indicating that the model's confidence calibration is highly dependent on the specific patient batch. Fold 3 achieved a significantly higher score (0.6887) than the others, suggesting that this subset contained patients with higher contrast or clearer anatomical boundaries.

5 Discussion

Our results demonstrate that a specialized U-Net architecture, combined with motivated data engineering, can achieve competitive segmentation of ovarian tissue in T2FS MRI sequences. Our final model achieved a mean 3D Dice score of

0.5185 for this challenging dataset. The discussion evaluates the mechanisms that enabled the best performing model to succeed and highlights the limitations of other proposed explanations. Table 1 summarizes experiment outcomes.

5.1 Clinical Validity of the Methodology

A potential critique of the Unilateral Masking method is that masking half the input image simplifies the task artificially. We argue that this is a necessary and clinically valid intervention to address the SLA of the dataset. By physically blacking out the non-annotated hemisphere, we aligned the input signal with the ground truth without altering the spatial dimensions (256×256) or the anatomy of the target hemisphere. In a deployment scenario, this approach translates to a simple pipeline: the full patient MRI is split into left and right volumes, processed independently in parallel, and stitched back together. This treats each ovary as an independent diagnostic question, which is anatomically appropriate.

Furthermore, our model was trained exclusively on slices containing ovarian tissue. Consequently, the model assumes that a target exists within the input. This setup mirrors the segmentation stage ("AttUSeg") of the RAovSeg pipeline proposed by Liang et al. [1]. A complete clinical system would require an upstream classifier (equivalent to "ResClass") to filter out empty slices before passing them to our segmentation network. Without this prior detection step, our model is prone to predicting false positives in empty slices containing only fat or vessels.

5.2 The "Intensity Trap" and Texture Learning

Our experiments revealed that pixel intensity is a confounding variable in T2FS sequences. We hypothesized that Preprocessing (RAovSeg/CLAHE) or Transfer Learning (ResNet34) would help the model distinguish ovaries from background based on texture. Both hypotheses failed, but for different reasons.

A. Preprocessing Failure: The RAovSeg intensity transformation assumes that ovaries occupy a specific mid-intensity range (hypointense relative to fat). While this is true for solid tissue, our T2FS dataset reveals a bi-modal distribution. As shown in Figure 4, the mid intensity assumption holds for Patient D2-051 (Top Row), where the ovary intensity (green histogram) falls below the inversion cutoff (red dashed line). However, for Patient D2-024 (Bottom Row), the fluid-filled follicles appear hyperintense. The histogram shows the ovary signal pushing past the 0.5 threshold, making it indistinguishable from fat. Consequently, the preprocessing logic inverts the ovary itself, turning the bright target black and effectively erasing the signal. This proves that global intensity thresholding is fundamentally incompatible with T2FS sequences due to histogram overlap.

B. Transfer Learning Failure: We introduced a pre-trained ResNet34 encoder to force the model to learn textural features rather than intensity. However, this resulted in severe overfitting. As seen in Experiment 05, the Training Dice reached ≈ 0.75 while the Validation Dice stagnated at ≈ 0.31. This massive generalization gap indicates that the dataset ($N = 37$, $S = 177$) is too small to fine-tune a deep, high-capacity encoder. Instead of learning generalizable texture filters, the ResNet memorized the specific noise patterns of the training patients.

5.3 Training Dynamics and Loss Formulation

The comparison between Standard Dice+BCE (Exp 03) and Focal Tversky Loss (Exp 07) highlights the difficulty of navigating the loss landscape with sparse data. While the Training Loss for Exp 07 decreased smoothly, the Validation metrics oscillated wildly. This suggests that the optimizer successfully minimized the Focal Loss on the training set, but the features it learned did not generalize. The Focal parameter γ, designed to focus on "hard" examples, likely forced the model to overfit to ambiguous boundary noise in the training data, interpreting artifacts as signal. In contrast, the standard Dice+BCE loss provided a smoother optimization landscape, allowing the Attention U-Net to converge to a more stable, albeit less aggressive, minima.

5.4 Benchmarking and Limitations

Our final 3D Dice score of 0.5185 compares favorably to the baseline results reported in the literature. Liang et al. [1] reported a Dice score of 0.290 for their full RAovSeg pipeline and 0.272 for nnU-Net on their dataset. While direct comparison is difficult due to differences in dataset composition and exclusion criteria, our result suggests that the Attention U-Net combined with Unilateral Masking data engineering is a highly effective strategy for this domain.

However, the high standard deviation ($\sigma \approx 0.08$) and the wide variance in optimal thresholds ($\tau \in [0.3, 0.7]$) across folds indicate that our solution is not yet universally generalizable. The model performs well on "typical" patients but degrades on patients with low contrast or complex pathologies. This fragility is a direct consequence of data sparsity; $N = 37$ ($S = 177$) is insufficient to capture the full variance of human anatomy.

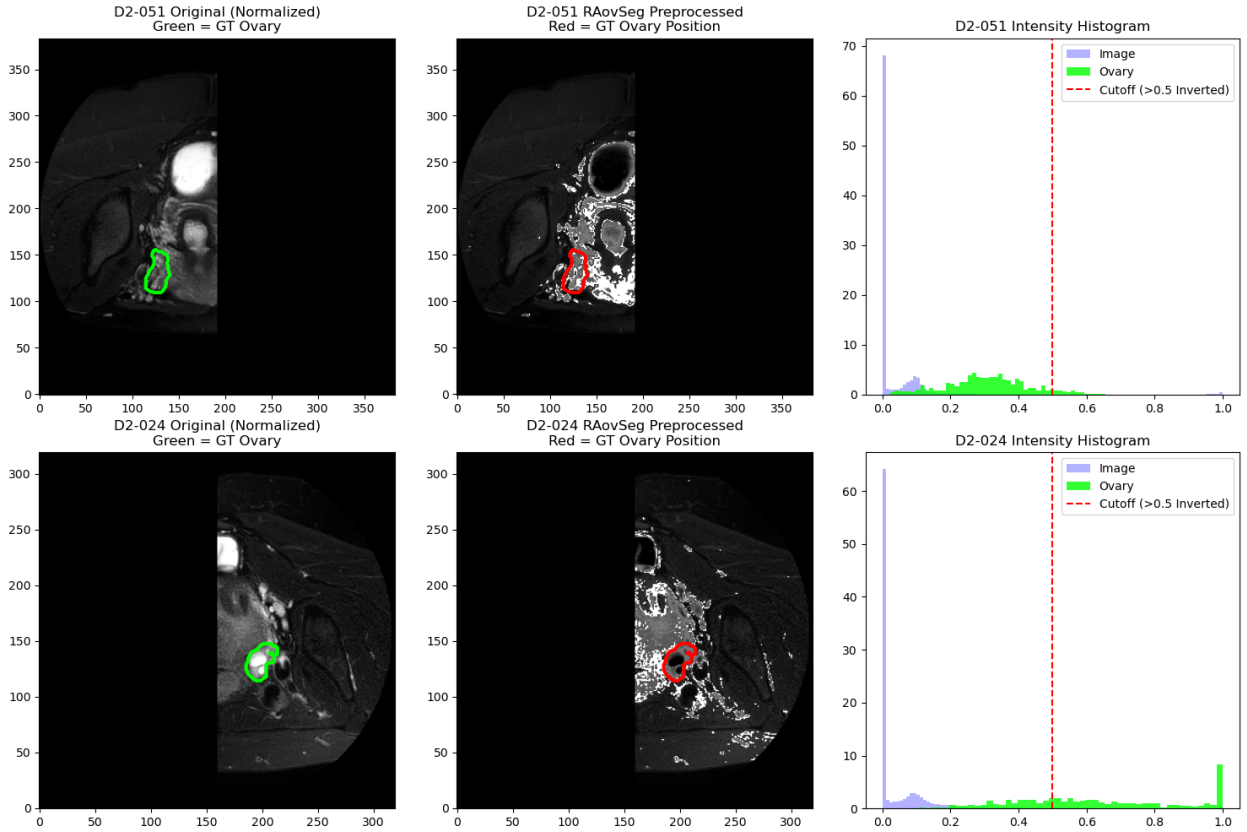

Figure 4: Analysis of the "Intensity Trap." **Top Row (Success Case):** The ovary in Patient D2-051 is grey. The histogram (Green) stays to the left of the Red Cutoff line (0.5). The preprocessing correctly enhances it while suppressing the background. **Bottom Row (Failure Case):** The ovary in Patient D2-024 contains bright fluid. The histogram (Green) shifts to the right, overlapping with the background noise. The preprocessing logic interprets the ovary as "fat" and inverts it (turns it black), destroying the training signal.

6 Conclusion and Future Work

By identifying the SLA in the dataset and addressing it through Unilaterally Masking our data, we achieved a mean volumetric Dice score of 0.5185. This result surpasses the baseline inter-rater agreement reported in the literature (≈ 0.48) [1], validating our hypothesis that precise data alignment is the most significant factor for model performance in this domain.

Our experiments showed that while complex architectural interventions (Transfer Learning) and loss formulations (Focal Tversky) offer theoretical advantages, they failed to generalize on this small, heterogeneous cohort ($N = 37$, $S = 177$). The model struggled to navigate the bi-modal intensity distribution of T2FS images, where ovaries can appear either hypointense or hyperintense. Ultimately, a streamlined Attention U-Net combined with careful post-processing proved to be the most effective strategy, highlighting the importance of domain-specific geometric constraints over algorithmic complexity when data is sparse.

Future research should prioritize two architectural shifts to address the limitations of 2D slice processing. First, implementing 3D architectures (e.g., V-Net) would leverage volumetric context along the Z-axis, suppressing the spatially inconsistent false positives observed in our results. Second, we recommend replacing Unilateral Masking with physical cropping of the target hemisphere. Unlike our current method, which downsamples the full field of view, physical cropping would allow the model to process the anatomy at a higher effective resolution. This would likely improve the detection of small ovarian follicles that are currently lost during resizing, without requiring larger input images. Finally, applying these methods to a larger cohort ($N > 100$) would allow for a re-evaluation of Transfer Learning, potentially alongside CLAHE, to resolve intensity ambiguity without the overfitting risks encountered in this study.

References

[1] Xiaomin Liang, Linda A Alpuing Radilla, Kamand Khalaj, Haaniya Dawoodally, Chinmay Mokashi, Xiaoming Guan, Kirk E Roberts, Sunil A Sheth, Varaha S Tammisetti, and Luca Giancardo. A multi-modal pelvic mri dataset for deep learning-based pelvic organ segmentation in endometriosis. *Scientific Data*, 12(1292), 2025. doi:10.1038/s41597-025-05623-3.

[2] Xiaomin Liang and Luca Giancardo. Uthealth - endometriosis mri dataset (ut-endomri), 2024. URL https://doi.org/10.5281/zenodo.15750762.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[4] Y Wang et al. Deep learning-based segmentation of epithelial ovarian cancer on t2-weighted magnetic resonance images. *Insights into Imaging*, 2023. PMC10006162.

[5] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[6] Carlos E Cardenas, Beth M Beadle, Adam S Garden, Heath D Skinner, Jinzhong Yang, Dong Joo Rhee, Rachel E McCarroll, Tucker J Netherton, Sweet Ping Ng Gay, Lifei Zhang, and Laurence E Court. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. *International Journal of Radiation Oncology*Biology*Physics*, 109(3):801–812, 2021.

[7] Akshansh Mishra. Contrast limited adaptive histogram equalization (clahe) approach for enhancement of the microstructures of friction stir welded joints. *arXiv preprint arXiv:2109.00886*, 2021.

[8] Le Peng, Hengyue Liang, Gaoxiang Luo, Taihui Li, and Ju Sun. Rethinking transfer learning for medical image classification. *arXiv preprint arXiv:2106.05152*, 2021. URL https://arxiv.org/abs/2106.05152.

[9] Davood Karimi, Simon K. Warfield, and Ali Gholipour. Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks. *arXiv preprint arXiv:2006.00356*, 2020. doi:10.48550/arXiv.2006.00356. URL https://arxiv.org/abs/2006.00356.

[10] Vishal Rajput. Robustness of different loss functions and their impact on network's learning capability. arXiv preprint, 2021. Computer Science Department, KU Leuven, Belgium.

[11] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *arXiv preprint arXiv:1810.07842*, 2018. doi:10.48550/arXiv.1810.07842. URL https://arxiv.org/abs/1810.07842.