Deep Learning for Ovary Segmentation in Pelvic MRI

[Anonymous]

1 Introduction

Deep learning has become the dominant approach for medical image segmentation, enabling automated delineation of anatomical structures and lesions across many imaging modalities. Convolutional neural network (CNN) architectures such as U-Net [1] and its variants have achieved strong results on tasks ranging from brain tumour segmentation to abdominal organ segmentation. However, performance often depends critically on the quality and consistency of the training data as well as on task-specific choices of preprocessing, architecture and loss function.

Segmentation of the ovaries in pelvic MRI is of particular clinical interest in endometriosis and fertility assessment, but it is challenging for both humans and algorithms. Ovaries are small, low-contrast structures surrounded by anatomically complex tissue, and their appearance can be altered by cysts, follicles or pathology. Recent work by Liang *et al.* [2] introduced the UT-EndoMRI dataset, a pelvic MRI dataset with manual segmentations of the uterus, ovaries and endometriomas. They reported lower inter-rater agreement and lower automatic segmentation performance for ovaries than for the uterus, highlighting that ovary segmentation on UT-EndoMRI is intrinsically challenging and that annotation noise is a significant factor.

Loss functions designed for imbalanced segmentation tasks are closely related to our experiments. Dice loss and its variants are widely used as surrogates for overlap metrics, particularly when the foreground occupies a small fraction of the image. The Tversky index generalises Dice by weighting false positives and false negatives differently [3], and the Focal Tversky loss further emphasises hard examples by raising $(1 - \text{Tversky})$ to a power [4]. These losses have been applied to lesion, tumour and organ segmentation where boundary errors are clinically important and background dominates the pixel count.

In this project we focus on automatic ovary segmentation in axial T2-weighted fat-suppressed (T2FS) pelvic MR images. We take as our starting point the ovary segmentation component of the RAovSeg pipeline proposed for UT-EndoMRI by Liang *et al.* [2], which combines an Attention U-Net with sequence-specific preprocessing and a Focal Tversky loss. Rather than treating RAovSeg as a black box, our goal is to reproduce and dissect its key components in an iterative, experiment-driven manner. This fits the broader aim of the course project: to apply deep learning theory to a realistic computer vision problem, to analyse failure modes, and to motivate experiments based on observed issues rather than chasing headline metrics.

Concretely, we address the following questions:

- How much does each component of the RAovSeg-style pipeline (attention, RAovSeg-inspired preprocessing, specialised loss) contribute to ovary segmentation performance compared to a simple U-Net baseline?
- Does transfer learning from a ResNet34 encoder pretrained on natural images provide an additional boost over U-Net-based models trained from scratch?
- How strongly does the mismatch between the visual content of the images (often two visible ovaries) and the ground-truth labels (typically a single annotated ovary) affect the measured performance, and can simple masking strategies mitigate this?

To answer these questions we develop a family of 2D, slice-based ovary segmentation models on T2FS pelvic MR images, all implemented in PyTorch [5]. We start with a baseline U-Net and progressively add attention gates [6], a RAovSeg-inspired contrast preprocessing pipeline [2], the Focal Tversky loss [3, 4], a ResNet34 encoder [7], and an ovary-side masking scheme that zeros out the half of the pelvis opposite to the annotated ovary. Rather than aiming for state-of-the-art performance, we emphasise an iterative, hypothesis-driven exploration of how these components interact, paying particular attention to qualitative failure modes and label ambiguity.

2 Related Work

Deep learning has become the dominant approach for medical image segmentation over the past decade. The U-Net architecture introduced by Ronneberger *et al.* [1] has been particularly influential due to its encoder–decoder design with skip connections, which enables precise localisation while retaining contextual information. Numerous variants have since been proposed, including 3D extensions and architectures tailored to specific organs or modalities.

Attention mechanisms have been incorporated into U-Net-style networks to improve focus on small or ambiguous structures. Oktay *et al.* [6] introduced the Attention U-Net, where learnable attention gates modulate the skip connections based on decoder features, suppressing irrelevant activations from the encoder. This approach has been applied to a range of medical segmentation tasks, including abdominal organ segmentation and lesion detection, and serves as the architectural foundation for the ovary segmentation component of the RAovSeg pipeline.

Liang *et al.* [2] introduced the UT-EndoMRI dataset, which contains pelvic MRI scans of patients with suspected endometriosis. Their work includes manual segmentations of the uterus, ovaries and endometriomas by multiple raters, as well as a two-stage automatic ovary segmentation pipeline called RAovSeg. An important finding is that ovaries are significantly harder to segment than the uterus: inter-rater Dice for ovaries is substantially lower than for the uterus, and automatic methods also achieve lower Dice on ovaries than on larger, more homogeneous structures. This supports our interpretation that ovary segmentation on UT-EndoMRI is inherently challenging and that moderate Dice scores can still be clinically meaningful.

Loss functions designed for imbalanced segmentation tasks are closely related to our experiments. Dice loss and its variants are widely used as surrogates for overlap metrics, particularly when the foreground occupies a small fraction of the image. The Tversky index generalises Dice by weighting false positives and false negatives differently [3], and the Focal Tversky loss further emphasises hard examples by raising $(1 - \text{Tversky})$ to a power [4]. These losses have been applied to lesion, tumour and organ segmentation where boundary errors are clinically important and background dominates the pixel count. In our study we compare a standard binary cross-entropy loss with the Focal Tversky loss to assess its impact on ovary delineation.

Transfer learning from natural-image classification networks is another well-established strategy in medical imaging. Encoders such as ResNet pretrained on ImageNet are often reused as backbones for segmentation networks [7], either frozen or fine-tuned, with reported benefits in data-scarce settings. Several works have shown that such pretrained encoders can improve performance on MRI and CT segmentation tasks, although the gains depend on the domain gap and the amount of annotated data. Our TL-5 model follows this paradigm by combining a ResNet34 encoder with an Attention U-Net decoder, allowing us to quantify the benefit of encoder pretraining in the specific context of ovary segmentation.

Finally, data-centric approaches that adjust the training data or labels instead of the architecture are increasingly recognised as crucial in medical imaging. Examples include relabelling noisy masks, designing task-specific sampling strategies, and generating synthetic training examples. Our ovary-side masking experiment can be viewed in this light: rather than changing the model, we modify the input to better match the single-ovary label definition and study how this affects both performance and failure modes.

3 Data and Preprocessing

3.1 Dataset

We use the public pelvic MRI dataset introduced by Liang *et al.* [2] for endometriosis research, focusing on the T2-weighted fat-suppressed (T2FS) sequences. Each study contains a stack of axial slices and corresponding binary masks for several pelvic organs. In this project we only use the *ovary* masks.

All images are converted to NIfTI format and resampled to a common in-plane resolution. For computational reasons we work with 2D slices rather than full 3D volumes.

3.2 Slice selection and data split

For each patient we construct a list of axial slices for which the ovary mask is non-empty. Slices without any ovary pixels are discarded. Thus, both training and validation are performed on *ovary–positive* slices only, and all reported Dice scores (Sections 4–5) should be interpreted in this setting.

We split the data at the *patient* level to avoid information leakage between training and validation. In practice this yields 56 eligible patients with visible ovaries, of which 44 are assigned to the training set and 12 to the validation set. The same split is reused across all experiments to ensure comparability between model variants.

3.3 Image preprocessing

All images are first cropped around the pelvis to remove empty borders and resized to 256×256 pixels. Intensities are then linearly normalised to zero mean and unit variance per slice. This "plain" normalisation is used for the baseline U-Net (Model 1) and the plain Attention U-Net (Model 2).

For the RAovSeg-inspired models (Models 3–7 and TL-5) we apply an additional preprocessing step that aims to increase ovary contrast relative to surrounding structures, following the description in Liang *et al.* [2]. In our implementation this includes:

- Clipping intensities to a robust range (e.g. 1st–99th percentile) and rescaling to $[0, 1]$.
- A non-linear remapping of intensities that emphasises mid-range values where ovary tissue typically lies.
- Local contrast enhancement (for example using a high-pass or sharpening filter) to highlight edges and small bright structures.

Figure 1 shows a representative slice before and after this RAovSeg-style preprocessing. Unless otherwise stated, all RAovSeg models are trained and evaluated on these preprocessed images.

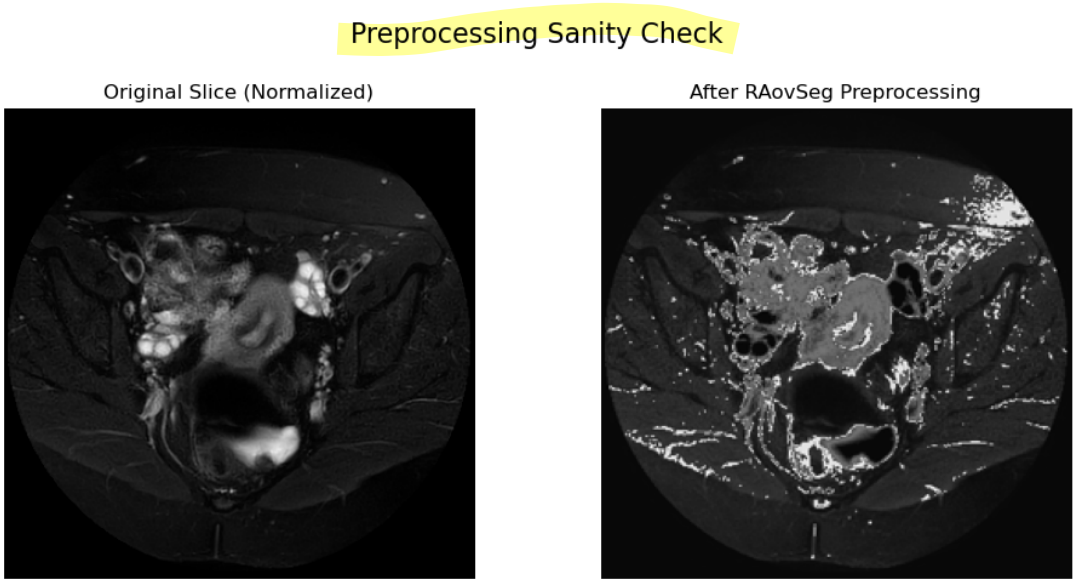

Figure 1: Example T2FS slice before (left) and after (right) RAovSeg-style preprocessing. The preprocessing enhances the contrast of ovary tissue relative to surrounding structures.

3.4 Data augmentation

During training we apply light on-the-fly augmentations to reduce overfitting and encourage spatial invariance. Augmentations include random horizontal and vertical flips, small rotations and translations, and slight intensity jittering. The same augmentations are applied to the image and mask to preserve alignment. No augmentation is used during validation or test-time inference.

3.5 Ovary-side masked variant

As observed during qualitative inspection, many slices contain two visible ovaries even though only one ovary is annotated in the ground truth. To study the impact of this mismatch, we also create an "ovary-side masked" variant of the dataset.

Given a slice and its ovary mask, we compute the centroid of the foreground pixels along the horizontal axis. If the centroid lies in the left half of the image, we treat the left as the "ovary side" and zero out the right half of the image; otherwise we zero out the left half. The mask itself is left unchanged. This produces images in which only the annotated ovary is visible.

Unless explicitly stated, all models are trained and evaluated on the unmasked images. In later experiments (Sections 5–6) we evaluate all models on both the original and ovary-side masked validation sets, and for one model (Model 7) we also use ovary-side masking during training. We treat masked evaluation primarily as a diagnostic tool: in practice, one could imagine a two-pass deployment that segments each half of the image separately, but we do not implement or evaluate such a pipeline here.

4 Methods

4.1 Problem formulation

We treat ovary segmentation as a 2D binary semantic segmentation problem on T2FS MR images. For each slice we have a single-channel image $x \in \mathbb{R}^{1\times H\times W}$ and a binary ground-truth mask $y \in \{0,1\}^{1\times H\times W}$ indicating the annotated ovary. A model f_θ produces a logit map $\hat{z} = f_\theta(x)$, from which we obtain probabilities $\hat{p} = \sigma(\hat{z})$ using the sigmoid function. At test time we threshold at 0.5 to obtain a hard prediction $\hat{y} = \mathbb{1}[\hat{p} > 0.5]$.

Performance is evaluated using the Dice coefficient on ovary–positive validation slices only (Section 3). Given binary masks $\hat{y}$ and y, the Dice score is

$$\mathrm{Dice}(\hat{y}, y) = \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|},$$

with a small ϵ added in the implementation for numerical stability. For each model we report the mean Dice over all validation slices.

4.2 Training setup

All models are trained as 2D slice-based segmenters on the same training split, using the same data loader and slice selection as described in Section 3. For each model we use a fixed number of epochs (20 or 50 depending on the experiment), the Adam optimiser with a constant learning rate, and the same batch size and weight decay. Only the architecture, loss function and preprocessing differ between experiments. For each run we select the checkpoint with the highest validation Dice score and use that model for all subsequent analyses. All models are implemented in PyTorch [5].

4.3 Baseline U-Net and Attention U-Net

Baseline U-Net (Model 1). Our baseline follows a standard 2D U-Net architecture. The encoder consists of four downsampling stages, each containing two 3×3 convolutional layers with ReLU activations and batch normalisation, followed by 2×2 max pooling. The decoder mirrors the encoder with four upsampling stages using transposed convolutions, skip connections from the corresponding encoder features, and `DoubleConv` blocks (two 3×3 convolutions with ReLU and batch normalisation). The final 1×1 convolution maps the last 64 feature channels to a single output logit. Inputs and outputs are single-channel images of size 256×256.

Attention U-Net (Model 2). Model 2 replaces the plain skip connections of the baseline with attention gates, following the Attention U-Net design of Oktay *et al.* [6]. Each skip connection passes through an `AttentionGate`, which takes the decoder feature map (gating signal) and the encoder feature map as input and produces an attention coefficient map that modulates the encoder features before concatenation. The rest of the architecture (encoder/decoder depth and number of channels) is identical to the baseline. Both Models 1 and 2 are trained on plain intensity-normalised T2FS slices.

4.4 RAovSeg-style preprocessing and model variants

The RAovSeg paper proposes a modality-specific preprocessing step that enhances ovary contrast relative to surrounding tissue. We implement a simplified version of this procedure and apply it slice-wise before feeding the images to the network, as described in Section 3.

We then train attention U-Net variants on this RAovSeg-style input:

- **Model 3** (Attn U-Net + RAovSeg, 20 epochs) uses the attention U-Net architecture from Model 2, but with RAovSeg-preprocessed images as input and the same BCE loss as Model 2. This model is trained for 20 epochs.
- **Model 4** (Attn U-Net + RAovSeg, 50 epochs) repeats the same configuration but extends training to 50 epochs to test whether longer training on the RAovSeg input leads to improved performance or overfitting.
- **Model 5** (Attn U-Net + RAovSeg + FTL) keeps the architecture and RAovSeg preprocessing of Model 4 but replaces the BCE loss with Focal Tversky loss (Section 4.7), again trained for 50 epochs.

4.5 Beyond RAovSeg: transfer learning with a ResNet34 encoder

After analysing the main RAovSeg-style components on top of an Attention U-Net, we also wanted to test a more substantial change that goes beyond the original pipeline. To this end, we build a variant in which the encoder is replaced by a ResNet34 backbone pretrained on ImageNet. The decoder retains the Attention U-Net structure with attention gates on the skip connections. We refer to this model as **TL-5** (ResNet34 Attn U-Net + RAovSeg + FTL) to emphasise that it shares the preprocessing and loss configuration with Model 5 but uses a deeper, pretrained encoder.

4.6 Ovary-side masking during training

As discussed in Section 3.5, many slices contain two visible ovaries while only one is annotated. Motivated by the side-swap and double-ovary failure modes observed in our baseline experiments, we design a simple ovary-side masking strategy to explicitly suppress the unannotated side during training and evaluation.

Model 7 (Attn U-Net + RAovSeg + FTL, masked train) uses the same architecture, RAovSeg preprocessing and Focal Tversky loss as Model 5, but the training images are replaced by their ovary-side masked versions. At evaluation time we consider both original and masked validation images to assess how this training regime affects performance under each setting.

4.7 Loss functions

Binary cross-entropy (BCE). For the baseline U-Net (Model 1), the plain Attention U-Net (Model 2) and the RAovSeg Attention U-Nets with BCE (Models 3 and 4), we use the standard binary cross-entropy loss between the predicted probabilities $\hat{p}$ and the ground-truth mask y:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\right].$$

Focal Tversky loss (FTL). For Model 5 and TL-5 we use the Focal Tversky loss, motivated by the small size of the ovary relative to the full image and the asymmetric costs of false positives and false negatives. Given prediction probabilities P and ground truth G, the Tversky index is defined as:

$$\mathrm{TI}(P,G) = \frac{|P \cap G|}{|P \cap G| + \alpha|P \setminus G| + \beta|G \setminus P|},$$

where α and β weight false positives and false negatives. The Focal Tversky loss is then

$$\mathcal{L}_{\mathrm{FTL}} = (1 - \mathrm{TI}(P,G))^{\gamma},$$

with $\gamma > 1$ emphasising hard examples. In our experiments we set $\alpha = 0.7$, $\beta = 0.3$ and $\gamma = \frac{4}{3}$, following the recommendations of Abraham and Khan [4] and empirical tuning on our data.

Model 5. Model 5 shares the architecture and RAovSeg preprocessing of Model 4, but optimises the Focal Tversky loss instead of BCE and is trained for 50 epochs. The goal is to investigate whether a loss tailored to imbalanced segmentation can improve ovary delineation on top of the RAovSeg input.

4.8 Post-processing

Following Liang *et al.* [2], we apply a simple post-processing step based on connected components. Given a predicted binary mask, we identify all connected components and retain only the largest one, removing smaller clusters as likely false positives. This heuristic encodes the assumption that at most one ovary is present in each slice, which is consistent with the single-ovary annotations in the dataset.

We analyse the effect of this post-processing separately, both quantitatively and qualitatively, as it can have a large impact on individual slices despite a relatively small average effect on Dice.

5 Results

In this section we first present quantitative results for all model variants on the pelvic MRI ovary segmentation task, followed by a qualitative analysis of typical success and failure cases. Unless otherwise stated, all Dice scores are computed on ovary–positive validation slices only (Section 3).

5.1 Quantitative evaluation

Table 1 summarises the performance of all trained models. Each model is evaluated twice: (i) on the original validation slices (*orig*), and (ii) on ovary-side masked validation slices (*masked*), where the image half opposite to the annotated ovary is zeroed out (Section 4.6). The last column reports $\Delta\text{Dice} = \text{Dice}_{\text{masked}} - \text{Dice}_{\text{orig}}$.

Table 1: Overview of models and validation Dice scores. All scores are computed on ovary–positive validation slices only. "Orig" denotes the original validation images, while "Masked" denotes ovary-side masked images as described in Section 3.5.

Model	Preproc	Loss	Dice (orig)	Dice (ma
1: U-Net baseline	plain	BCE	0.275	0.39
2: Attention U-Net	plain	BCE	0.247	0.38
3: Attn U-Net + RAovSeg (20 ep)	RAovSeg	BCE	0.306	0.34
4: Attn U-Net + RAovSeg (50 ep)	RAovSeg	BCE	0.264	0.29
5: Attn U-Net + RAovSeg + FTL	RAovSeg	Focal Tversky	0.313	0.34
TL-5: ResNet34 Attn U-Net + RAovSeg + FTL	RAovSeg	Focal Tversky	**0.365**	**0.42**
7: Attn U-Net + RAovSeg + FTL (masked train)	RAovSeg (masked train)	Focal Tversky	0.264	0.35

5.1.1 Overall performance

Across all models, the absolute Dice scores on ovary–positive slices are modest compared to typical organ segmentation benchmarks, but they are consistent with the difficulty of the task and the inter-rater variability reported for the dataset [2]. The RAovSeg-style models (Models 3–5 and TL-5) generally outperform the plain-normalisation models (Models 1–2), and transfer learning with a ResNet34 encoder yields the best overall performance.

5.1.2 Experiment 1: Baseline and problem characterisation (U-Net vs. Attention U-Net)

Here we ask whether adding attention gates to a plain U-Net improves slice-wise ovary segmentation on intensity-normalised T2FS images, and we use this first experiment to characterise the main failure modes of the task.

Comparing Models 1 and 2 reveals that the addition of attention gates to the U-Net does not improve performance under our training regime. On the original validation images, the plain U-Net achieves a Dice of 0.275, while the Attention U-Net achieves 0.247. On ovary-side masked validation images, both models improve (0.392 for the U-Net and 0.381 for the Attention U-Net), but the attention variant still does not surpass the baseline.

Qualitative inspection suggests that both models exhibit similar failure modes: they often either miss the ovary entirely, segment a large region of nearby tissue, or segment both ovaries when two are visible. Attention gates alone do not appear to confer a strong advantage when the input is simply normalised T2FS intensity.

Beyond the raw Dice numbers, this first experiment also helped us characterise the task: the baseline training and validation curves (not shown) already exhibit noisy validation behaviour and large slice-to-slice variability, especially for cases with two visible ovaries. The combination of low Dice, frequent over- and under-segmentation, and side swaps motivated the RAovSeg-inspired preprocessing and loss experiments in Experiments 2–3, as well as the masking experiments introduced later in Section 4.6.

Overall, this comparison shows that attention alone does not meaningfully improve performance over the plain U-Net baseline in our setting.

5.1.3 Experiment 2: Effect of RAovSeg-style preprocessing

Our second question is whether RAovSeg-style contrast preprocessing helps an Attention U-Net focus more reliably on the ovary compared to simple per-slice normalisation.

Introducing the RAovSeg-inspired preprocessing markedly changes the behaviour of the attention U-Net. Comparing Models 2 and 3 (same architecture, different preprocessing) reveals a clear gain: the original-image Dice increases from 0.247 to 0.306. This confirms that enhancing the contrast between suspected ovary tissue and surrounding anatomy helps the network focus on the correct region. However, longer training with RAovSeg (Model 4, 50 epochs) does not lead to further improvements: the Dice decreases to 0.264, and the training curves (Figure 2) indicate overfitting and unstable validation loss.

On masked validation images, the RAovSeg models still benefit from masking but to a lesser extent than the plain models (ΔDice $\approx$ 0.03–0.04 vs. 0.12–0.13). This is consistent with the qualitative impression that RAovSeg encourages more compact, single-blob segmentations that already focus on one ovary, while the plain models tend to segment both ovaries or larger regions.

Thus, RAovSeg-style preprocessing is a key driver of the performance gains we observe over the plain-intensity baselines.

5.1.4 Experiment 3: Focal Tversky loss and longer training

Next, we investigate whether replacing BCE with the Focal Tversky loss and training for more epochs leads to better ovary delineation on top of RAovSeg-preprocessed inputs.

Model 5 combines RAovSeg preprocessing with the Focal Tversky loss and 50 training epochs. Compared to Model 4 (same preprocessing, BCE loss), Model 5 achieves a slightly higher Dice on both original and masked images (0.313 vs. 0.264 on orig; 0.344 vs. 0.295 on masked). The training curves (Figure 3) show smoother convergence and less pronounced overfitting than Model 4, suggesting that Focal Tversky provides a more stable optimisation signal for this imbalanced problem.

However, the improvement is modest in absolute terms, and qualitative failure modes remain similar. Focal Tversky does not resolve issues such as side swaps, missed ovaries or over-segmentation of nearby structures; rather, it slightly refines the alignment of the predicted ovary blob with the ground truth when the model already focuses on the correct region.

These results indicate that Focal Tversky loss mainly stabilises training and yields a modest Dice improvement, but it does not fundamentally alter the dominant failure modes.

5.1.5 Experiment 4: Transfer learning with a ResNet34 encoder

Guided by the limited gains from architectural tweaks within the original RAovSeg-style setup, we then ask if transfer learning from a ResNet34 encoder pretrained on ImageNet can provide a meaningful boost over the best standalone Attention U-Net.

The ResNet34-based attention U-Net (TL-5) achieves the highest Dice scores of all models: 0.365 on original validation images and 0.420 on masked images. This represents an improvement of about 0.05 Dice over the best U-Net variant (Model 5) on original images, and a similar gain on masked images.

The increase is meaningful but not dramatic. Qualitative inspection suggests that TL-5 produces slightly sharper boundaries and fewer gross failures than Model 5, but many of the same error patterns persist (for example, segmenting the wrong ovary when two are visible, or producing fragmented blobs). In other words, transfer learning improves overall accuracy but does not solve the underlying ambiguity in the annotations.

In practice, the ResNet34 encoder delivers the best Dice among our models, but the absolute gain over the best U-Net variant is modest (around 0.05 Dice) and the qualitative failure patterns remain similar.

5.1.6 Experiment 5: Ovary-side masking during training

Finally, we test whether masking out the half of the pelvis opposite to the annotated ovary during training can mitigate the label ambiguity introduced by slices where both ovaries are visible. This ovary-side masking was not part of the original RAovSeg pipeline; we introduce it specifically to target the side-swap and double-ovary failures observed earlier.

Model 7 is trained with ovary-side masked inputs, attempting to align the training signal with the single-ovary ground truth. On the original validation images its Dice (0.264) is comparable to the weaker RAovSeg models, indicating that masking during training does not improve the standard metric. On masked validation images, however, Model 7 benefits strongly (ΔDice = +0.089), reaching 0.354 Dice and approaching the performance of TL-5 in the masked setting.

These results suggest that ovary-side masking is most effective when applied at evaluation time (as a diagnostic or oracle-like setting) rather than as a training-time modification. Training exclusively on masked images may reduce the model's ability to handle the more complex, unmasked clinical setting.

5.2 Training dynamics and post-processing

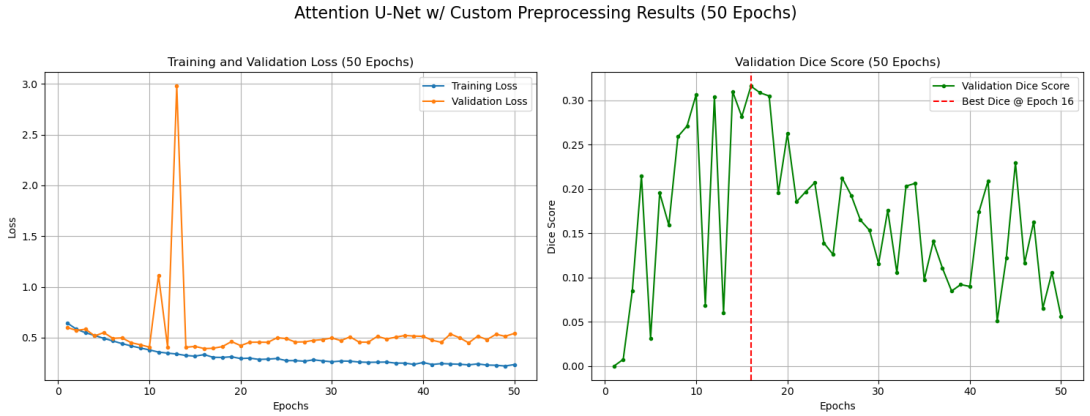

Figure 2: Training and validation loss curves for Model 4 (Attn U-Net + RAovSeg, 50 epochs). The validation loss becomes unstable and begins to increase after around 20 epochs, indicating overfitting despite continued improvements on the training set.

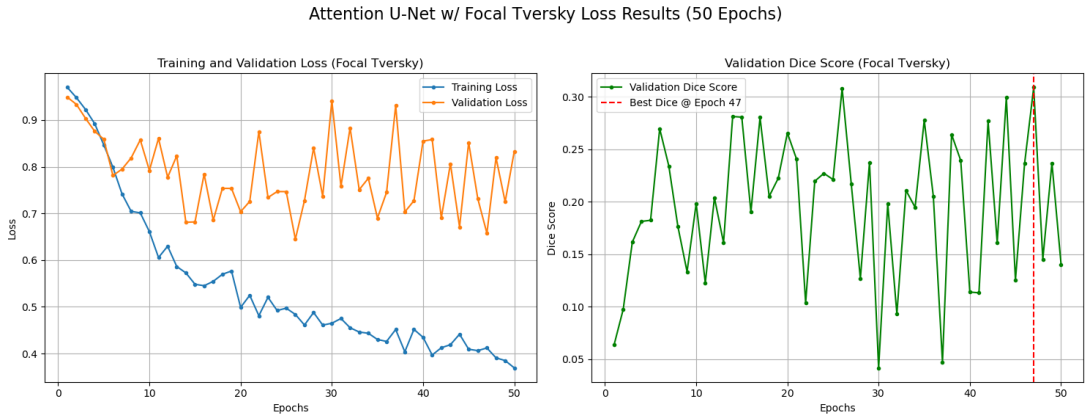

Figure 3: Training and validation loss curves for Model 5 (Attn U-Net + RAovSeg + Focal Tversky loss). Compared to Model 4, the training dynamics are smoother and the validation loss remains more stable, consistent with the modest Dice improvement.

Figure 2 illustrates the training and validation losses for Model 4. Although the training loss continues to decrease throughout the 50 epochs, the validation loss plateaus and becomes noisy after around 20 epochs, and the validation Dice does not improve further. This behaviour is consistent with the quantitative result that Model 4 performs worse than Model 3 despite longer training.

In contrast, Figure 3 shows that Model 5, trained with Focal Tversky loss, exhibits smoother training dynamics and a more stable validation loss. This supports the interpretation that Focal Tversky provides a better-behaved optimisation landscape for this imbalanced segmentation problem.

We also quantify the effect of connected-components post-processing on Model 5. On average, the change in Dice is small (< 1 percentage point) and slightly negative, suggesting that aggressively removing small components can sometimes discard true positives. However, as illustrated in the qualitative examples below, post-processing can have a dramatic impact on individual slices, either by cleaning up scattered false positives or by erroneously removing part of the true ovary.

5.3 Qualitative analysis

To better understand model behaviour, we examine representative slices from the validation set, focusing on both successful and failed cases. Rather than showing large grids of examples, we present a small number of "story" figures that highlight specific phenomena.

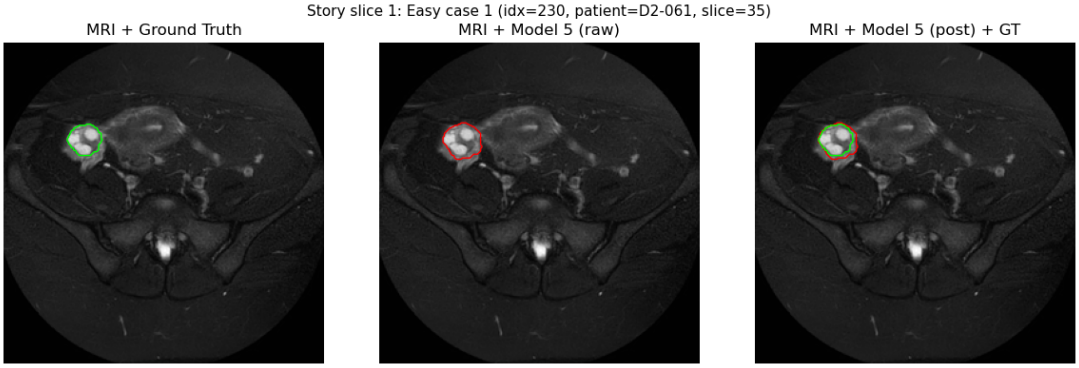

Figure 4: Example of an "easy" case where most models correctly segment the ovary. The RAovSeg-based models produce slightly more compact and better-aligned masks than the plain models.

Figure 4 shows an example where all models produce reasonable segmentations. The ovary is relatively isolated and has good contrast, and the main difference between models lies in the sharpness of the predicted boundaries.

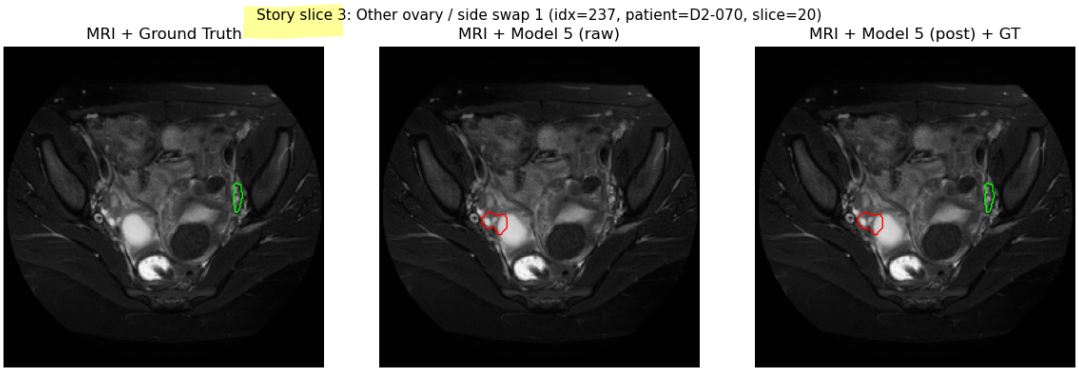

Figure 5: Example of a "side swap" failure where the model segments the contralateral ovary instead of the annotated one. This failure mode is common when both ovaries are visible but only one is annotated.

Figure 5 illustrates a common failure mode we refer to as a *side swap*. The image clearly shows two ovaries, but the ground truth mask covers only one. Several models segment the contralateral ovary instead, producing anatomically plausible masks that nevertheless score very poorly under the single-ovary Dice metric. This mismatch between visual content and label definition is a recurring issue in our experiments and motivates the ovary-side masking analysis.

Figures 6 and 7 show cases where connected-components post-processing either helps or hurts. In the success case, the raw prediction contains a main ovary blob plus several small false positives scattered across the pelvis; post-processing removes these and yields a clean, single-blob mask. In the failure case,

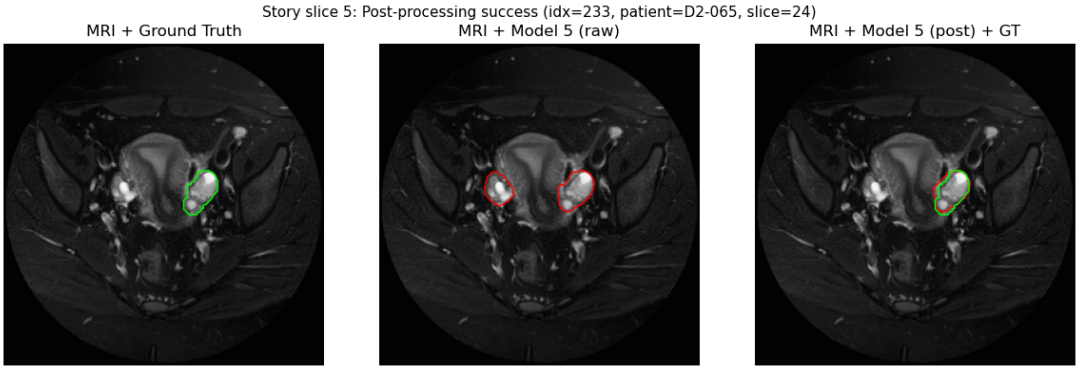

Figure 6: Example where connected-components post-processing improves the prediction by removing small scattered false positives and keeping a single compact ovary blob.

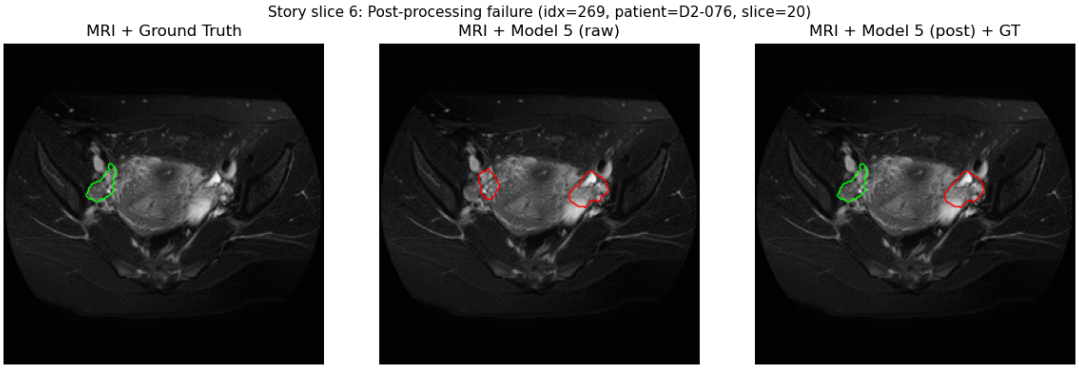

Figure 7: Example where connected-components post-processing hurts performance by discarding part of the true ovary and keeping an incorrect component.

the prediction contains two similarly sized components near the ovaries, and the heuristic of keeping only the largest one discards part of the true ovary, reducing Dice.

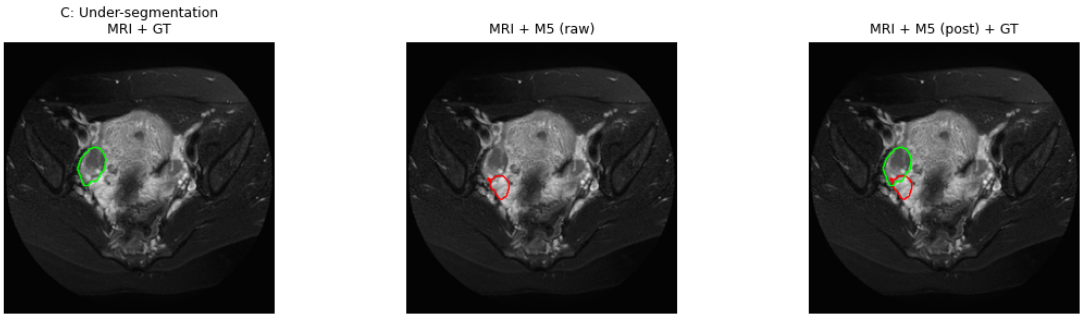

Figure 8: Example of under-segmentation, where the model predicts only a small portion of the ovary.

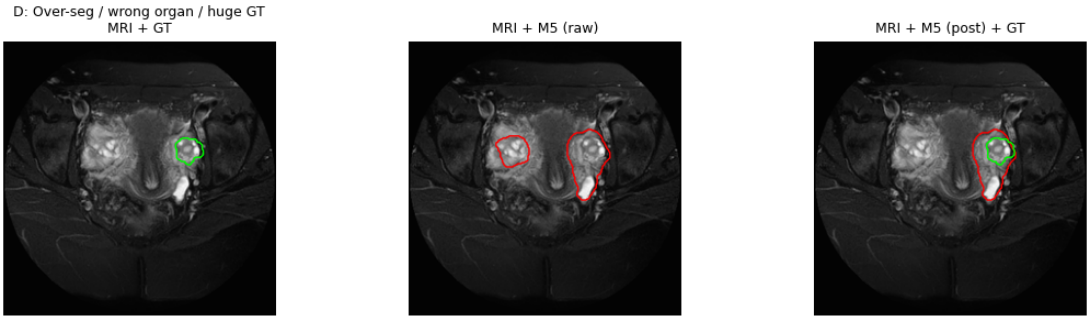

Figure 9: Example of over-segmentation, where the model includes surrounding tissue or both ovaries in the predicted mask.

Finally, Figures 8 and 9 provide examples of under- and over-segmentation. Under-segmentation often occurs when the ovary has low contrast or irregular shape, leading the model to predict a small, conservative blob. Over-segmentation typically arises when the model includes adjacent cysts, vessels or even the entire adnexal region. These failure modes are not unique to any particular architecture or loss; they are driven largely by the inherent ambiguity of the images and labels.

6 Discussion

Our experiments aimed to reproduce and analyse an RAovSeg-style ovary segmentation pipeline on pelvic T2FS MRI, starting from a baseline U-Net and progressively adding architectural changes (attention, transfer learning), preprocessing (RAovSeg-inspired contrast enhancement), a specialised loss (Focal Tversky), and a data-centric ovary-side masking scheme. We can roughly divide our work into two phases: in the first, we reproduced and dissected the main RAovSeg components on top of a 2D U-Net baseline; in the second, guided by the failure modes observed in this first phase, we proposed two simple extensions beyond the original pipeline, namely the ResNet34-based TL-5 model and the ovary-side masking strategy.

6.1 Impact of architectural changes and preprocessing

Across all models, the quantitative differences in Dice score are relatively modest when evaluated on ovary–positive validation slices (Table 1). The baseline U-Net and the plain Attention U-Net perform noticeably worse than the RAovSeg-based models, indicating that simple intensity normalisation is insufficient for robust ovary segmentation.

Putting these numbers in context, Liang *et al.* [2] report an average ovary Dice of roughly 0.48 for both the full RAovSeg pipeline and inter-rater agreement. Our best model (TL-5) reaches a mean Dice of 0.365 on ovary–positive validation slices, which is lower but not wildly inconsistent given that we train only the segmentation component, use a smaller subset of the data, and do not benefit from the dedicated slice selector or 3D aggregation used in the original work. Liang *et al.* also report a Dice of 0.290 for their

full 3D ovary pipeline on Dataset 2, computed over complete volumes including empty slices, whereas our 0.365 Dice is measured on ovary-positive slices only and is therefore not directly comparable.

The RAovSeg-style preprocessing produces a clear improvement over the plain inputs for the same architecture (Model 2 vs. Model 3), and this is supported by the qualitative examples: preprocessed slices exhibit sharper boundaries and higher ovary-to-background contrast, which leads to more compact and localised predictions.

The Focal Tversky loss, introduced in Model 5, provides a small but consistent gain over the BCE-based RAovSeg models. Training is somewhat more stable, and the resulting masks are slightly better aligned with the ground truth. However, the effect size is limited: the main qualitative failure modes remain unchanged and are driven more by annotation ambiguity than by class imbalance alone.

The transfer-learning model TL-5, which replaces the U-Net encoder with a ResNet34 backbone, achieves the highest Dice score on both original and masked validation sets. This suggests that richer encoder features pre-trained on large natural-image datasets can provide some benefit even in a very different domain such as pelvic MRI. At the same time, the modest magnitude of the improvement (roughly 0.05 Dice) and the persistence of the same failure modes highlight that architectural sophistication alone cannot fully overcome label noise and task ambiguity.

6.2 Role of ovary-side masking and label mismatch

The ovary-side masking experiments shed light on the impact of label mismatch between the images and the ground truth. Masking the half of the pelvis opposite to the annotated ovary at evaluation time consistently increases Dice across all models, with particularly large gains for the plain models (Models 1–2). This indicates that a significant portion of the error is due to models segmenting both ovaries or the wrong ovary when two are visible, not just failing to detect any ovary at all.

For the plain models (Models 1–2), this leads to particularly large gains in Dice (up to +0.13), reflecting the fact that these models often segment both ovaries or large contiguous regions in the pelvis. The RAovSeg models and Model 5 also benefit from masking, but to a smaller degree, consistent with the observation that they already produce more compact, single-blob masks. Model 7, which is trained on masked inputs, shows a large relative gain on masked validation images but does not outperform the best unmasked models when evaluated on the original images.

Taken together, these results support the hypothesis that the mismatch between the visual content (two ovaries visible) and the label definition (only one ovary annotated) is a major source of error. The side swap and double-ovary patterns seen in Figures 5, 7, 8 and 9 are penalised heavily by the single-ovary Dice metric, even when the model segments anatomically plausible structures. Masking alleviates this by enforcing the single-ovary assumption in the input, but does not fully resolve the ambiguity. For this reason we view the masked-evaluation scores as a kind of "oracle" upper bound on what a slice-based model could achieve if it were told in advance which side to focus on, rather than as a directly deployable operating point.

6.3 Post-processing as a double-edged sword

The connected-components post-processing step, motivated by the RAovSeg paper, is intuitively appealing: one expects a single ovary per slice, so keeping only the largest component should remove small false positives. Our experiments confirm that this heuristic can improve individual cases, particularly when the model predicts a main ovary blob along with a few small spurious clusters far from the ovary. In such cases, post-processing often turns a low-Dice prediction into a high-Dice one.

However, the same heuristic can be harmful when the model predicts multiple components of similar size near the ovaries. If the true ovary is split across components or if there is a competing bright structure nearby, keeping only the largest component may discard part of the true ovary or select the wrong structure entirely, leading to a substantial drop in Dice. The overall average effect on Dice is small and slightly negative, reflecting this trade-off.

From a deployment perspective, this suggests that post-processing should be applied with caution and ideally tuned or adapted based on additional information, such as the expected size range of the ovary or uncertainty estimates from the model.

6.4 Limitations and future work

Our study has several important limitations:

- **2D slice-based modelling.** We treat each slice independently and do not exploit 3D context across slices. This is a common simplification for limited compute, but in pelvic MRI the spatial continuity of organs across slices could provide valuable cues to disambiguate ovaries from nearby structures.
- **Ovary-positive slice selection.** Both training and validation are restricted to slices with non-empty ovary masks. As a result, the reported Dice scores do not measure how well the models distinguish slices with and without ovaries, and false positives on background slices are not penalised. Moreover, we aggregate Dice over slices rather than per patient, so a few difficult slices can dominate the mean even if most slices for a given subject are segmented reasonably well. A more realistic evaluation would include the full volume and report patient-level metrics.
- **Label ambiguity and noise.** Many slices contain two visible ovaries, yet only one is annotated. Larger qualitative grids (not shown in the main text) confirm that in a non-trivial subset of cases the contralateral ovary is clearly visible but unlabeled, which helps explain the side-swap and double-ovary predictions seen in our story figures. In addition, some masks appear unusually large or irregular. These issues directly impact the Dice scores and limit how well any model can align with the ground truth. We did not attempt to correct or relabel any masks.
- **Limited hyperparameter tuning.** Due to time and compute constraints, we did not perform an extensive search over learning rates, data augmentation strategies, or architectures. It is likely that further tuning could yield modest improvements, especially for the deeper models.
- **Single modality and task.** We only considered axial T2FS images and ovary segmentation. The dataset contains additional sequences (for example T1-weighted) and other organs, which could enable multi-modal or multi-task learning and potentially improve robustness.

Future work could address these limitations by: (1) using 3D or 2.5D architectures that process multiple neighbouring slices jointly; (2) redesigning the evaluation to include empty slices and to explicitly reward correct localisation of both ovaries; (3) incorporating improved or multi-label annotations where both ovaries are segmented; and (4) exploring semi-supervised or self-supervised pretraining on the unlabeled MR volumes. We also did not explore more aggressive optimisation strategies (for example cyclic learning rates or systematic hyperparameter search) or a separate slice-selection classifier as in RAovSeg; these could yield incremental gains but were beyond our computational and time budget.

Overall, our results suggest that data quality and label definitions are at least as important as architectural choices for this task. While RAovSeg-style preprocessing and transfer learning do provide incremental gains, the dominant failure modes have more to do with how the problem is posed than with model capacity alone.

7 Conclusion

We investigated deep learning-based ovary segmentation on pelvic T2FS MRI by reproducing and extending components of the RAovSeg pipeline. Starting from a baseline U-Net, we added attention gates, RAovSeg-inspired contrast preprocessing, the Focal Tversky loss, a ResNet34-based encoder, and a data-centric ovary-side masking scheme. We evaluated each step quantitatively on ovary-positive validation slices and qualitatively using grids of examples (not shown) and carefully selected "story" slices.

Our experiments show that RAovSeg-style preprocessing and transfer learning provide the largest gains over a plain U-Net baseline, while the Focal Tversky loss offers a modest but consistent improvement and smoother training dynamics. Ovary-side masking reveals that a substantial portion of the error arises from label mismatch—particularly when both ovaries are visible but only one is annotated—rather than from sheer model capacity.

From a methodological perspective, the work highlights the importance of understanding dataset quirks and evaluation protocols when interpreting segmentation metrics. Even modest Dice scores can be meaningful when inter- rater agreement is low and labels are ambiguous, as is the case for ovary segmentation in UT-EndoMRI.

From a learning perspective, the project illustrates the value of an iterative, hypothesis-driven workflow: we started from a simple baseline, made targeted changes motivated by the literature, and complemented quantitative metrics with qualitative analysis to understand not only *how well* the models performed but also *where* and *why* they failed. We believe that similar data- and failure-focused analyses

are essential when applying deep learning to medical imaging problems where annotations are costly and often ambiguous.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[2] Xiaomin Liang, Linda A. Alpuin-Radilla, et al. A multi-modal pelvic mri dataset for deep learning-based pelvic organ segmentation in endometriosis. *Scientific Data*, 12(1):1292, 2025.

[3] S. S. M. Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *Machine Learning in Medical Imaging (MLMI)*, volume 10541 of *Lecture Notes in Computer Science*, pages 379–387. Springer, 2017.

[4] 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*, 2019.

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.