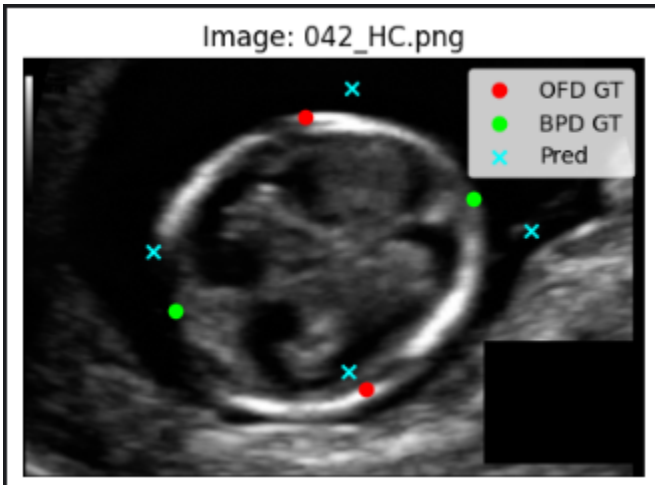
Janeshvar Sivakumar

Landmark-Based Biometry Identification

Given axial fetal ultrasound images, the goal is to locate four landmarks that define two clinically meaningful diameters (BPD and OFD) so that simple geometric measurements, such as diameters and derived head circumference, can be produced robustly. I chose to focus on a heatmap-based approach after experimenting with direct coordinate regression. While coordinate regression worked decently, it suffered from two major issues: limited data amplified its sensitivity to noise, and its inherent weakness in preserving spatial context made precise localisation unstable. Heatmaps, on the other hand, naturally represent uncertainty, allow the network to express multimodal beliefs (which is particularly important for symmetric anatomical structures), and integrate cleanly with convolutional backbones that preserve spatial structure.

My initial assumptions were simple and explicit:
(a) the ground-truth CSV correctly contained the two OFD and two BPD coordinates per image,
(b) training a single UNet-like heatmap network to predict four heatmaps would be sufficient as a baseline, and
(c) standard per-image normalization combined with modest geometric augmentations would provide adequate robustness.

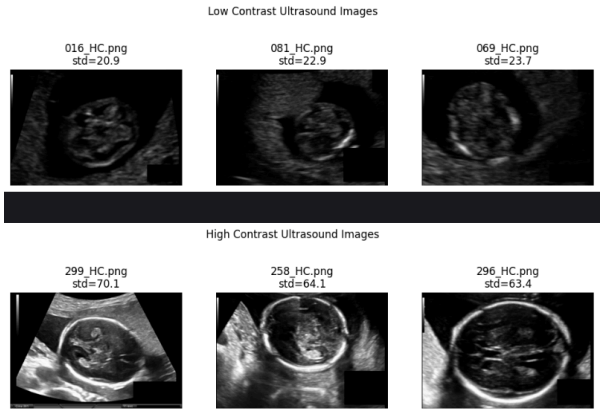
As development progressed, I treated these assumptions as hypotheses to verify rather than fixed truths.



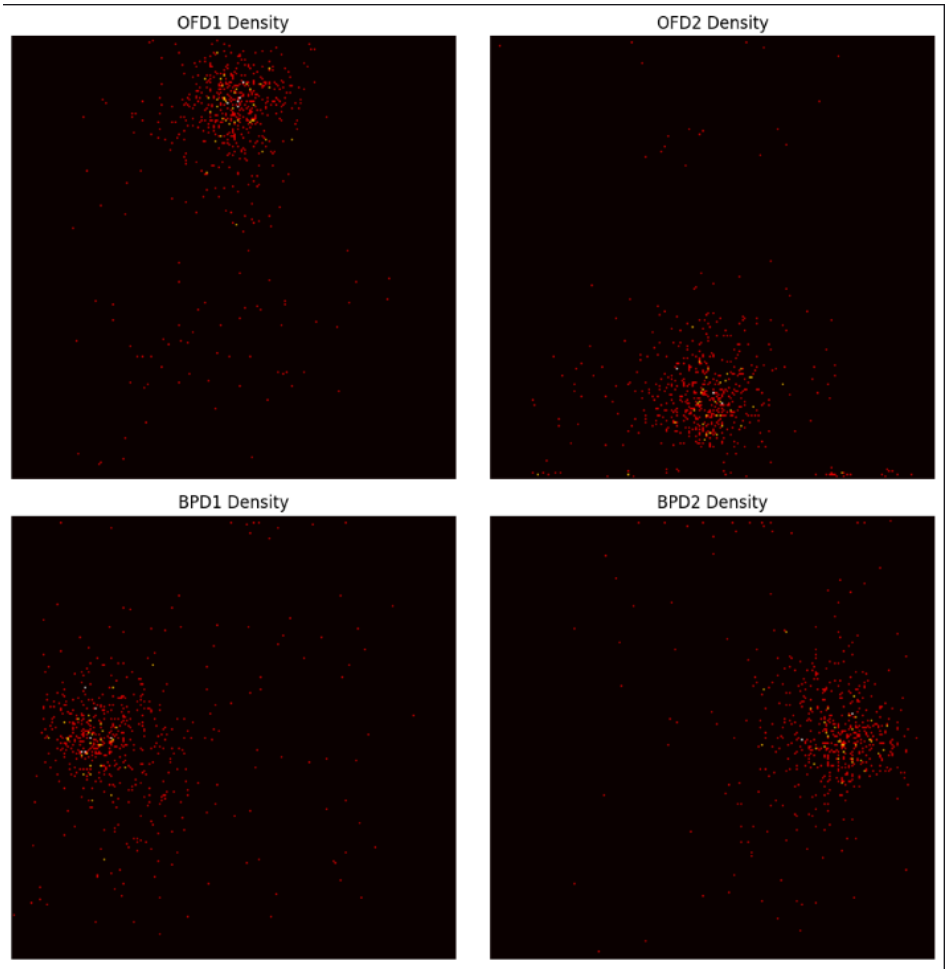
Result of Coordinate Regression

Data exploration and preprocessing

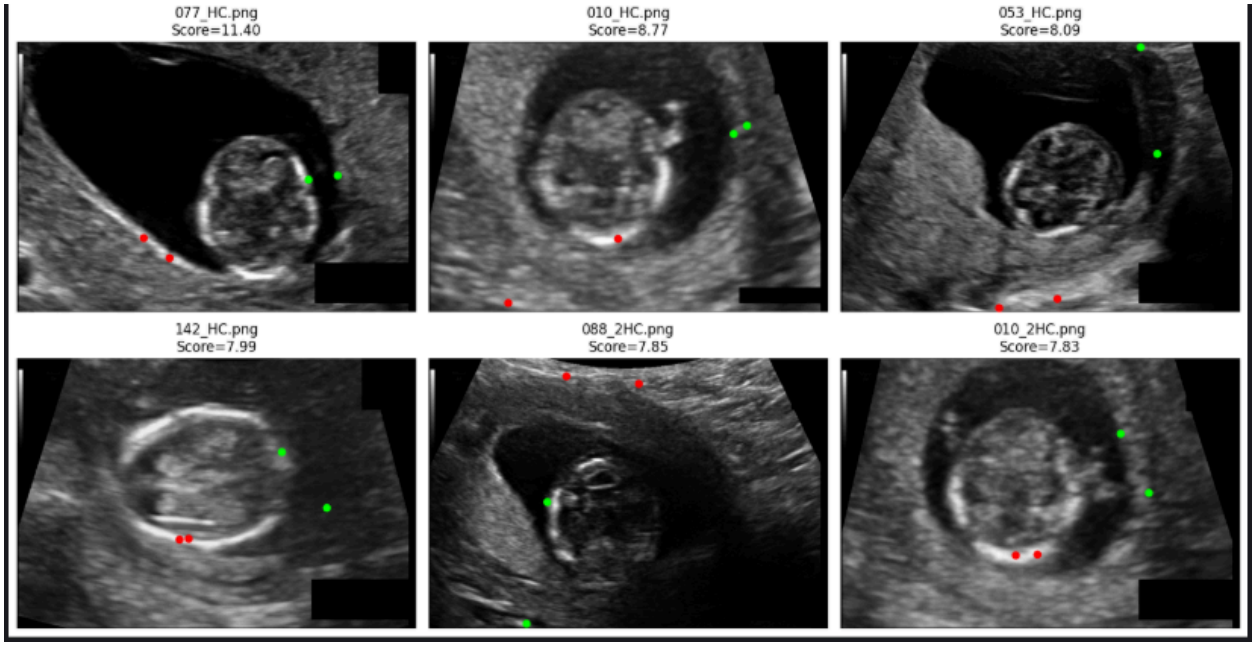
The first part of the work focused on data hygiene. The image corpus is heterogeneous: most images are around 540 × 800 pixels, but several other resolutions are present. Mean intensity and contrast vary widely across images, so I normalized images on a per-image basis using z-score normalization rather than a global min–max scheme. This stabilized the input distribution and made both training behaviour and visual inspection more meaningful.



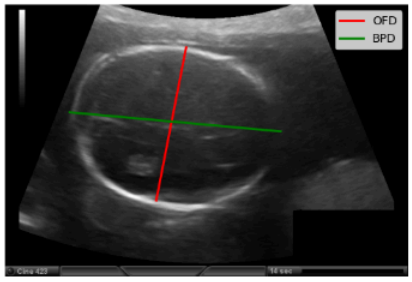
While plotting landmark density maps, I noticed a large number of spatial outliers.



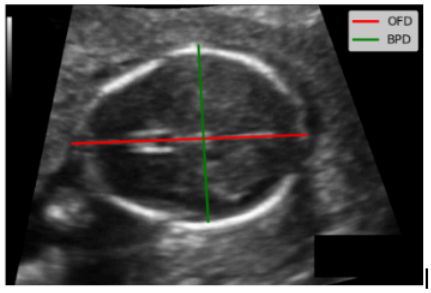
This led to the suspicion that some dataset labels were incorrect rather than simply noisy. To investigate this systematically, I represented each image's eight-landmark vector and used a Mahalanobis-style distance to flag anomalous samples.



Visual inspection of these flagged cases revealed a mixture of legitimate anatomical extremes and genuinely erroneous annotations. Instead of aggressively pruning all flagged samples, I retained a conservative "core" subset, removing only the most obvious failures. This maintained anatomical diversity while reducing the most harmful label noise.



Before correction



After Correction

A critical discovery occurred during visualization: the CSV semantics were reversed. In the raw CSV, the fields labeled as OFD corresponded to what is anatomically the BPD (horizontal), and vice versa. This was not obvious from loss curves or naive metrics, but it became immediately clear when ground-truth

lines were plotted directly on images. I fixed this mapping centrally in the dataset loader, ensuring all downstream training and evaluation used the corrected semantics.

Before prototyping any specific model, I surveyed existing literature and came across **BiometryNet**. I found its evaluation methodology particularly useful and adopted its measurement-level metrics, as well as drawing inspiration from its **HRNet-based** architecture as a conceptual baseline. However, given time constraints and the need to first stabilize preprocessing and label semantics, I proceeded with a UNet-based baseline and explored improvements from there.

Early modeling attempts and failures

The first working model was a compact HRNet that took a single-channel normalized image and produced four heatmaps. These heatmaps were supervised using an MSE loss, and the model was trained with modest geometric augmentation (rotation ±10°, small scaling, translation, and horizontal flips).

Early quantitative indicators looked acceptable, but visual inspection revealed important failure modes that the numbers did not capture. Two patterns stood out:

Collapsed landmarks: when the model had low confidence, it sometimes produced heatmaps that collapsed pairs of landmarks toward the midline.
Multimodal heatmaps: for the OFD pair, the network occasionally produced two strong symmetric peaks rather than a single clean peak.

These observations made it clear that pointwise losses and hard argmax decoding were insufficient. The model needed to be able to express uncertainty and respect geometry, and the evaluation needed to measure geometry rather than only pointwise distances.

In the initial stages, BPD heatmaps consistently performed better than OFD heatmaps. This imbalance motivated the introduction of channel-wise weights in the loss function to increase the importance of the harder landmarks. Applying these weights led to measurable improvements, particularly for OFD localisation.

A predicted heatmap should be interpreted as a belief distribution rather than a decisive coordinate. Two symmetric peaks on a BPD or OFD channel do not necessarily indicate an error; instead, the model is expressing that multiple positions are plausible given the available evidence. Hard argmax decoding arbitrarily selects one peak, leading to noisy and unstable coordinate predictions.

To address this, I adopted soft-argmax decoding, where the expected coordinate of the heatmap is computed by interpreting it as a probability distribution. This produces stable, subpixel coordinates that reflect multimodality without arbitrary snapping. It also enables uncertainty estimation through measures such as heatmap entropy or spatial variance. These uncertainty estimates can then be correlated with prediction error, which is valuable for deployment scenarios where low-confidence predictions should be flagged or rejected.

Geometry-aware constraints

Because BPD and OFD are geometrically linked, they should be near orthogonal in a correctly acquired axial plane and share a common center. I treated this as soft prior knowledge rather than a hard rule. I implemented an angle-consistency regularizer that penalizes deviations of the predicted angle between the BPD and OFD vectors from 90°. Predicted heatmaps were decoded to continuous coordinates using soft-argmax, the two diameter vectors were computed, and a small squared penalty on (angle − 90°) was added to the loss.

Although this approach is theoretically sound, time constraints limited the extent to which it could be fully tuned and validated. As a result, this constraint did not lead to substantial quantitative improvements within the available timeframe, but it remains a promising direction.

Evaluation philosophy

Clinicians ultimately care about derived biometry, such as BPD and OFD lengths, rather than whether individual landmark coordinates are off by a few pixels. Accordingly, I defined two levels of evaluation:

- Landmark-level: normalized mean error (NME), normalized by a head-size reference (OFD), along with per-landmark distances to identify which points are harder to localize.
- Measurement-level: BPD and OFD lengths computed from predicted and ground-truth landmarks, reporting Mean Absolute Error (L1), median L1, bias (signed mean error), and Bland–Altman 95% limits of agreement. Bland–Altman analysis is particularly informative because it exposes systematic bias and agreement ranges in a clinically familiar format.

Uncertainty-aware evaluation was also incorporated. Heatmap entropy or spatial variance was correlated with absolute error, enabling the construction of risk–coverage curves. In deployment, this supports conservative decision rules such as reporting measurements only when model uncertainty falls below a threshold.

Hypotheses explored

At a minimum, three directions were explored.

Hypothesis 1 : Baseline heatmap UNet
Motivation: A compact UNet is a reliable baseline for heatmap prediction and preserves spatial context.
I trained a single-head UNet to predict four heatmaps using a channel-wise weighted MSE loss.
This produced reasonable heatmaps on many images but failed in edge cases. In particular, symmetric endpoints led to multimodal heatmaps where hard decoding destabilized predictions. This baseline was necessary but insufficient for clinical-ready output.

Hypothesis 2 : Weighted heatmap loss and uncertainty weighting

Motivation: Some landmarks are inherently harder, and training can be dominated by easier ones. Additionally, low-quality images should contribute less to training.

I applied per-channel weights in the heatmap loss and experimented with downweighting borderline samples identified during EDA. I also explored weighting losses based on heatmap sharpness.

Per-landmark weighting improved focus on harder endpoints and reduced the influence of noisy labels. However, weighting alone did not resolve geometric inconsistencies.

Hypothesis 3 : Architectural variations

Motivation: Predicting BPD and OFD jointly may introduce task interference.

I explored architectural ideas such as multi-head decoders, where a shared encoder feeds two separate heads: one for BPD and one for OFD. This allows task-specific refinement while maintaining shared anatomical context. I also considered shared-backbone architectures with shallow task-specific refinements. While these were not fully implemented within the project timeframe, they represent a clear and logical next step.

	arch	stage	prep	aug	landmark	bpd_l1	ofd_l1
0	high_res	Stage1	P0	A0	26.3405	21.2477	42.2049
1	multi_head	Stage3	P1	A1	14.3956	6.1054	15.0157
2	single_head	Stage1	P0	A0	24.5663	27.8604	33.8511

Best performing variants of the three approaches

Limitations

Dataset size and label quality: Even after EDA and pruning, clean data remains limited and label noise persists.

Low-quality images: Severe low-contrast or high-noise images remain challenging; augmentation cannot recover missing information.

Generalization: The dataset originates from a single acquisition distribution, and robustness across devices and operators is untested.

Metric sensitivity: Landmark-level metrics can be misleading for diameters, motivating a focus on measurement-level evaluation.

Future work

Measurement-Space Evaluation. I would shift from landmark errors to diameter accuracy—reporting predicted BPD/OFD in mm with confidence intervals and quantifying clinical relevance as percentage error

Geometric Constraints as Learned Losses. I would enforce anatomical rules (BPD ≤ OFD, center alignment, angle constraints) as differentiable losses during training, then ablate each to show individual impact on measurement stability.

Cross-Device Generalization. I would systematically test on different ultrasound machines and operators, identify which image properties degrade performance, and propose targeted augmentation strategies for each failure mode.

Confidence Calibration & Clinical Triage. I would validate that model uncertainty accurately reflects prediction error through calibration curves, then define automated triage rules to flag uncertain cases for clinician review.

Multi-Head Architecture.I would implement a shared encoder with separate BPD/OFD decoder heads to reduce task interference while preserving anatomical context sharing, measuring improvement in measurement consistency.

Explainability for Clinical Trust. I would generate attention maps, create a failure mode atlas of 10 hardest cases, and produce calibration visualizations that clinicians can use to assess model reliability.

Better quality documentation. Given the time constraints and my ongoing college work (final year thesis) I was unable to formalise this document, for example, using LaTeX.

Key takeaways

• Visualize early and often. Metrics alone led me astray early on. Plotting predicted heatmaps and GT lines exposed the swapped labels and the multimodal beliefs that heatmaps encode.
• Preserve uncertainty. If a model can show it is uncertain, you can use that to make safer downstream decisions; forcing a single point estimate hides important information.
• Simple anatomy-informed priors are powerful. A weak angle constraint and a midpoint consistency check reduced many plausible but incorrect geometries.
• Label semantics matter. The simple mistake of swapped labels consumed a lot of debugging time; always confirm label-image consistency visually before drawing conclusions from loss curves.
• One change at a time. I structured experiments sequentially so that I could attribute effects to a single change (data, loss, architecture, constraint). That discipline made conclusions defensible even when results were mixed.

Janeshvar Sivakumar

I wanted to elaborate on this in more detail outside the report

Preprocessing and augmentation strategy

Before committing to any single model architecture, I spent a significant amount of time iterating on preprocessing and augmentation choices, as these stages fundamentally shape what the network can and cannot learn. Rather than treating preprocessing as a fixed step, I treated it as a sequence of controlled experiments, gradually narrowing down what actually helped under the constraints of limited data and time.

Preprocessing experiments

I experimented with preprocessing at multiple stages and levels of aggressiveness. The earliest versions used only simple intensity scaling to [0, 1], which quickly proved insufficient. Due to the large variability in acquisition conditions, this led to unstable training behavior and inconsistent convergence.

I then explored several alternatives:

- Global min–max normalization across the dataset
- Per-image min–max normalization
- Per-image z-score normalization

Among these, per-image z-score normalization consistently produced the most stable training dynamics. It reduced sensitivity to global brightness differences and made the model less dependent on absolute intensity values, which are particularly unreliable in ultrasound. Importantly, this choice also improved interpretability during visualization, as predicted heatmaps aligned more consistently with anatomical structures across images.

I also experimented with when preprocessing should occur relative to augmentation. Applying normalization before geometric augmentation led to subtle artifacts after interpolation, whereas

performing normalization after augmentation and resizing produced more consistent inputs. This ordering was therefore fixed as part of the final pipeline.

Augmentation experiments

Augmentation played a critical role in improving generalization, especially given the limited dataset size. However, aggressive augmentation quickly degraded performance by producing anatomically implausible samples. I therefore adopted a staged approach.

I experimented with:

- Rotation at multiple ranges (±5°, ±10°, ±20°)
- Scaling and translation with different magnitudes
- Horizontal flipping
- Combined affine transformations versus isolated transforms

Larger rotations (±20°) and strong scaling often harmed performance, particularly for OFD localization, by distorting anatomical relationships. Through iterative experimentation, I settled on modest affine augmentation: rotations within ±10°, small translations and scaling, and horizontal flips applied probabilistically. This struck a balance between improving robustness and preserving anatomical plausibility.

A key design choice was to apply augmentation **before heatmap generation**, ensuring that landmarks and images underwent the same geometric transformation. This avoided inconsistencies between image content and supervisory signals and was critical for stable learning.

Three-stage architecture exploration under time constraints

Given the limited time available, I adopted a structured, three-stage experimental strategy rather than exploring many architectures in parallel.

Stage 1 : Establishing a stable baseline
The first goal was not accuracy, but stability. I implemented a simple UNet-based heatmap model with minimal preprocessing and no geometry constraints. This stage helped surface foundational issues such as label semantics errors, decoding instability, and preprocessing weaknesses. Importantly, this stage validated the decision to use heatmaps over direct coordinate regression.

Stage 2 : Improving performance through loss design and preprocessing
Once the baseline was stable, I focused on improving performance without changing the core architecture. This included:

- Channel-wise weighted heatmap losses to address imbalance between BPD and OFD difficulty
- Refining preprocessing and augmentation choices
- Introducing soft-argmax decoding to stabilize coordinate extraction

This stage produced the most reliable gains relative to implementation effort and helped isolate the impact of data handling versus model capacity.

Stage 3 : Architectural refinement and geometry awareness
Only after preprocessing and loss behavior were well understood did I explore architectural and geometric extensions. This included:

- Geometry-aware angle consistency constraints
- Conceptual exploration of multi-head architectures
- Shared-backbone designs with task-specific refinement

Due to time constraints, not all architectural variants were fully trained and evaluated. However, this staged approach ensured that architectural complexity was introduced only after simpler explanations for failure modes had been ruled out.

Changing preprocessing, augmentation, loss functions, and architecture simultaneously makes it difficult to attribute improvements or failures to specific causes. By fixing preprocessing and augmentation first, then refining loss design, and only then considering architectural changes, I was able to reason clearly about why certain behaviors emerged and which interventions were genuinely beneficial.