Anomaly Detection in Fetal Echocardiography via Cross-modal Translation and Region-Discriminated Error Calibration

Qianye Yang[1], Yingyu Yang[1], Yipei Wang[5], Yipeng Hu[5], Can Peng[1], Elena D'Alberti[2], Bojana Salovic[2], Amaya Iriondo-Coysh[6], Beverly Tsai-Goodman[7], Julene Carvalho[3,4], Fabricio da Silva Costa[8], Makrina Savvidou[6], Olga Patey[2,3,4], Aris T. Papageorghiou[2], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
qianye.yang@eng.ox.ac.uk
[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK
[3] Brompton Centre for Fetal Cardiology, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK
[4] Fetal Medicine Unit, St Georges University Hospital NHS Foundation Trust, London, UK
[5] Hawkes Institute, Department of Medical Physics and Biomedical Engineering, University College London, London, UK
[6] Chelsea and Westminster Hospital NHS Foundation Trust, London, UK
[7] University of Birmingham, Birmingham, UK
[8] Maternal Fetal Medicine Unit, Gold Coast University Hospital, Southport, Australia

Abstract. Congenital heart disease (CHD) is the most common birth defect. Many CHD phenotypes are individually rare, which limits labeled data and motivates anomaly detection for multi-modal fetal echocardiography screening. Color flow Doppler (CFD) provides critical hemodynamic cues in fetal echocardiography, yet its color overlays are often spatially sparse and temporally intermittent, and vary substantially across scanners and acquisition settings. As a result, reconstruction-error-based anomaly detection can be poorly calibrated because the reconstruction-error may be dominated by acquisition shift rather than pathology. Our key innovation in this paper is to propose a bi-level optimization scheme that couples a video translator (inner model) with an error estimator (outer model). The inner and outer tasks are interdependent: the error estimator predicts clip-wise reconstruction-error estimation to re-weight the cross-modal translation model training, while the translation errors produced in the inner model are used to supervise the outer error estimator. At inference, we derive a region-discriminated anomaly score by standardizing observed translation errors with clip-conditioned predicted error statistics and aggregating in-/out-region deviations into a participant-level score. The framework was validated on a large-scale, multi-site clinical cohort of 440 participants, covering six CHD types. Results show that the proposed method outperforms standard reconstruction-

based baselines, achieving $AUROC_{all}$=0.723 and Spec@Sens75%=0.497 with all improvements being statistically significant ($p < 0.01$). Code will be released at *https://github.com/QianyeYang/XCalibAD*.

Keywords: Congenital Heart Disease · Anomaly detection · Cross-modal translation · Bi-level optimization

1 Introduction

Congenital heart disease (CHD) is the most common congenital anomaly, with a birth prevalence of roughly 8-9 per 1,000 live births worldwide [16]. Despite this overall prevalence, many clinically important CHD subtypes are individually rare, limiting the availability of labeled data per phenotype and making fully supervised learning challenging in real-world screening settings. This motivates anomaly detection approaches that learn patterns of normal fetal echocardiography and flag deviations without requiring exhaustive labels for every CHD subtype.

Most anomaly detection methods in medical imaging follow a one-class paradigm, where models trained on normal data flag anomalies via reconstruction error [4, 15]. In free-hand ultrasound, acquisition-induced appearance variation and inter-scanner domain shift can dominate anomaly-indicative residuals, resulting in poorly calibrated scores that might not be specific to pathology. This is further exacerbated in fetal echocardiography: B-mode captures anatomy whereas color flow Doppler (CFD) encodes hemodynamics [2], yet CFD is spatially sparse and highly sensitive to acquisition settings and scanner variability [3]. Cross-modal B-mode to CFD translation is thus appealing, but existing work mainly focuses on synthesis rather than reconstruction-error-based anomaly detection in fetal echocardiography video [13, 5, 1].

In this paper, we formulate CHD triage as anomaly detection driven by a cross-modal translation task together with reconstruction-error calibration. Our key hypothesis is that, for healthy fetal hearts, the CFD appearance is largely predictable from the underlying anatomy and motion observable in B-mode, comparatively much less from clip-dependent variability caused by acquisition conditions. Therefore, instead of treating raw residuals as anomalies, we aim to estimate the expected translation error for each input clip and then score anomalies by how much the observed error exceeds this input-conditioned expectation. This turns reconstruction error into a calibrated, region-discriminated deviation score that is robust to Doppler sparsity and inter-scanner appearance/domain shift. To operationalize this, we couple a cross-modal translator with an error estimator in a bi-level framework that learns clip-conditioned error statistics, enabling calibrated anomaly scoring under different acquisition conditions.

Our main contributions are: (i) a cross-modal B-mode to CFD anomaly detection framework that combines video translation with video-clip-conditioned error calibration for robust reconstruction-error scoring; (ii) a bi-level optimization scheme that couples translation learning with error calibration to better handle

heterogeneous normal acquisition patterns; (iii) extensive validation on a real-world multi-site clinical cohort with multiple scans per participant, demonstrating statistically significant and consistent improvements over reconstruction-based baselines and ablations across threshold-free and operating-point metrics. To our knowledge, this is the first CHD-oriented anomaly detection study in fetal echocardiography that leverages paired B-mode and CFD ultrasound data.

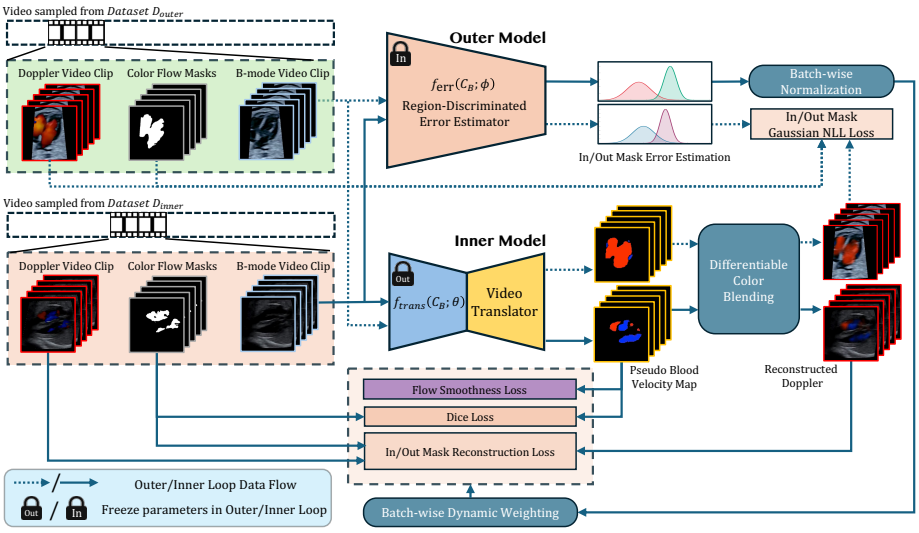

Fig. 1. Overview of the proposed bi-level framework. The outer model estimates in/out-mask reconstruction error statistics to produce batch-wise weights, while the inner model reconstructs Doppler from B-mode; the two networks are trained by alternating updates on $\mathcal{D}_{outer}$ and $\mathcal{D}_{inner}$, which are only from healthy cases.

2 Methods

An overview of the proposed method is shown in Fig. 1. Denote the set of N participants. Each participant $i \in \mathcal{P}$ has multiple paired B-mode/Doppler scan videos indexed by $\mathcal{S}_i$. The full dataset is written as $\mathcal{D} = \cup_{i\in\mathcal{P}}\{(V_B^{(i,s)}, V_D^{(i,s)}, V_M^{(i,s)}) : s \in \mathcal{S}_i\}$, where $V_B^{(i,s)}$ and $V_D^{(i,s)}$ are spatially aligned B-mode and Doppler videos, and $V_M^{(i,s)}$ is a binary video mask indicating the color-Doppler active region. $V_M^{(i,s)}$ is computed from the frame-wise difference between Doppler and B-mode via a deterministic operator $\mathcal{G}(\cdot)$, which denotes a simple image pre-processing pipeline that includes intensity thresholding followed by morphological cleanup to retain the valid color-overlay region $V_M^{(i,s)} = \mathcal{G}(|V_D^{(i,s)} - V_B^{(i,s)}|)$.

The whole dataset $\mathcal{D}$ is randomly split into three non-overlapping subsets at the participant-level, making $\mathcal{D} = \mathcal{D}_{inner} \cup \mathcal{D}_{outer} \cup \mathcal{D}_{test}$, where $\mathcal{D}_{inner}$ and $\mathcal{D}_{outer}$ are used to train the inner and outer models. During training and inference, both the inner and outer models operate on a temporally contiguous clip of fixed length L. For a start time index t, define the temporal window $\mathcal{T}(t, L) = \{t, t+1, \ldots, t+L-1\}$. The sampled clips $C(t)$ are extracted from the full videos over the temporal window $\mathcal{T}(t, L)$. Each training instance is then a triplet $x^{(i,s)}(t) = (C_B^{(i,s)}(t), C_D^{(i,s)}(t), C_M^{(i,s)}(t))$.

2.1 Cross-modal Translation

The Cross-modal translator is a deep neural network $f_{trans}(C_B; \theta) : \mathcal{C} \to \mathcal{U}$ with learnable parameters θ, which maps an input B-mode clip to a Doppler-like reconstruction through an explicit blood flow representation. Given a B-mode clip $C_B \in [0,1]^{3\times L\times H\times W}$, the translator predicts a single-channel flow map $u = \tanh(f_\theta(C_B)) \in [-1,1]^{1\times L\times H\times W}$, where the $\tanh(\cdot)$ activation enforces a bounded range and stabilizes subsequent color overlay rendering. The flow magnitude $m = |u|^\gamma \in [0,1]^{1\times L\times H\times W}$ encodes the strength of the color overlay, while the sign of u determines the blood flow direction (representing in red/blue) via a smooth gate $p = (\tanh(ku)+1)/2$, where k represents color gate sharpness. Using fixed endpoints $\mathbf{c}_r = (1.0, 0.2, 0)$ and $\mathbf{c}_b = (0, 0.2, 1.0)$, we form a differentiable color field $\mathbf{c}(u) = p\mathbf{c}_r + (1-p)\mathbf{c}_b \in [0,1]^{3\times L\times H\times W}$ and synthesize the Doppler-like output by alpha blending $\hat{C}_D = m\mathbf{c}(u) + (1-m)C_B$. The translator is optimized using a masked reconstruction loss within the ground-truth Doppler color region, together with a soft Dice loss aligning m to the binary color mask and a spatiotemporal smoothness regularizer on u based on the squared ℓ_2 norm of gradients along time and spatial dimensions. We define the in-mask reconstruction loss as $\mathcal{L}_{\text{in}} = \|C_M \odot (\hat{C}_D - C_D)\|_1 / (\|C_M\|_1 + \varepsilon)$, and the out-of-mask consistency loss as $\mathcal{L}_{\text{out}} = \|(1-C_M) \odot (\hat{C}_D - C_B)\|_1 / (\|1-C_M\|_1 + \varepsilon)$, where $\odot$ denotes element-wise multiplication and ε is a small constant. To stabilize learning of the color-overlay support, we apply a soft Dice loss between the magnitude map m and the binary mask C_M, given by $\mathcal{L}_{\text{dice}} = 1 - (2\langle m, C_M\rangle + \varepsilon)/(\|m\|_1 + \|C_M\|_1 + \varepsilon)$. We further regularize the flow map with a spatiotemporal smoothness term $\mathcal{L}_{\text{smooth}} = \|\nabla u\|_2^2$, where ∇ denotes finite-difference gradients over (L, H, W). The overall objective for the translator is as follows with λ_{out}, λ_{dice} and λ_{smooth} controls the weights for individual loss terms.

$$\mathcal{L}_{\text{trans}} = \mathcal{L}_{\text{in}} + \lambda_{\text{out}}\mathcal{L}_{\text{out}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} \tag{1}$$

During bi-level training, each clip-wise translator loss $\mathcal{L}_{\text{trans}}$ is further reweighted by a sample difficulty weight predicted by the outer model, yielding a weighted inner objective. Details will be introduced in Sec. 2.3.

2.2 Region-Discriminative Error Estimation

The region-discriminative error estimator is defined as $f_{\text{err}}(C_B; \phi)$ with learnable parameters ϕ, which predicts the expected translation error of the in-

ner model conditioned on the input B-mode clip. Given C_B, we first compute two observed errors from the current translator output $\hat{C}_D$: an in-mask error $e_{\text{in}} = \|C_M \odot (\hat{C}_D - C_D)\|_1/(\|C_M\|_1 + \varepsilon)$ and an out-of-mask error $e_{\text{out}} = \|(1 - C_M) \odot (\hat{C}_D - C_B)\|_1/(\|1 - C_M\|_1 + \varepsilon)$. We supervise the estimator on the log-errors $y_{\text{in}} = \log(e_{\text{in}} + \varepsilon)$ and $y_{\text{out}} = \log(e_{\text{out}} + \varepsilon)$. The estimator outputs clip-wise Gaussian parameters $(\mu_{\text{in}}, \sigma_{\text{in}})$ and $(\mu_{\text{out}}, \sigma_{\text{out}})$, modeling input-dependent variation in the error estimates. The error estimator is trained by minimizing the Gaussian negative log-likelihood of $(y_{\text{in}}, y_{\text{out}})$ under the predicted distributions, encouraging accurate, input-conditioned error calibration. The overall objective for the error estimator is

$$\mathcal{L}_{\text{err}} = \mathcal{L}_{\text{nll}}(y_{\text{in}}; \mu_{\text{in}}, \sigma_{\text{in}}) + \mathcal{L}_{\text{nll}}(y_{\text{out}}; \mu_{\text{out}}, \sigma_{\text{out}}), \tag{2}$$

where $\mathcal{L}_{\text{nll}}(y; \mu, \sigma) = 0.5 \cdot ((y-\mu)/\sigma)^2 + \log(\sigma)$ and σ is constrained to be positive via a softplus activation: $\text{softplus}(r) = \log(1 + \exp(r))$.

2.3 Bi-level Optimisation of the Networks

We optimise the inner translator parameters θ and the outer error-estimator parameters ϕ in a bi-level manner, where the outer model predicts clip-wise translation difficulty and uses it to re-weight the inner objective. Let $x = (C_B, C_D, C_M)$ denote a training instance sampled from a temporally contiguous window, and recall that $\mathcal{D} = \mathcal{D}_{inner} \cup \mathcal{D}_{outer} \cup \mathcal{D}_{test}$ is split at the participant level.

Given a mini-batch $\mathcal{B} = \{x_i\}_{i=1}^{B}$, the outer model produces difficulty scores from the predicted error statistics. In our implementation we use the predicted in-mask mean log-error as reconstruction difficulty, i.e., $d_i = \mu_{\text{in}}(C_{B,i}; \phi)$, and convert it to within-batch weights $w_i = \exp(d_i/\eta) \;/\; \sum_{j=1}^{B} \exp(d_j/\eta)$, where η is a temperature controlling the sharpness of the weighting. The weighted inner objective is then defined as

$$\mathcal{L}_{inner}(\theta; \phi) = \sum_{i=1}^{B} w_i \, \mathcal{L}_{\text{trans}}(x_i; \theta), \tag{3}$$

We consider sample with bigger reconstruction errors are hard cases which could benefit from higher weighting during updating θ.

The outer model is trained to predict the translation error distribution of the current inner model. Specifically, for samples drawn from $\mathcal{D}_{outer}$ we compute the observed log-errors $(y_{\text{in}}, y_{\text{out}})$ from the translator outputs and minimise the negative log-likelihood objective $\mathcal{L}_{\text{err}}(\phi; \theta)$ defined in Sec. 2.2. We denote the outer-model objective by $\mathcal{L}_{\text{outer}}$, which corresponds to the negative log-likelihood loss (i.e., $\mathcal{L}_{\text{outer}} \equiv \mathcal{L}_{\text{err}}$). The resulting bi-level problem can be written as

$$\begin{aligned} \phi^* &= \arg\min_{\phi} \; \mathbb{E}_{x \sim \mathcal{D}_{\text{outer}}} \big[\mathcal{L}_{\text{outer}}(x; \phi, \theta^*(\phi)) \big], \\ \text{s.t.} \;\; \theta^*(\phi) &= \arg\min_{\theta} \; \mathbb{E}_{x \sim \mathcal{D}_{\text{inner}}} \big[\mathcal{L}_{\text{inner}}(x; \theta, \phi) \big]. \end{aligned} \tag{4}$$

Directly differentiating through $\theta^*(\phi)$ introduces second-order terms. Following a standard first-order approximation [6, 10], we ignore the implicit dependence of θ^* on ϕ when updating the outer model, i.e., $\nabla_\phi \mathcal{L}_{\text{outer}}(\phi; \theta^*(\phi)) \approx \nabla_\phi \mathcal{L}_{\text{outer}}(\phi; \theta)$. Optimization is performed by alternating updates: we first update ϕ on $\mathcal{D}_{outer}$ using error targets computed from the current inner model while holding θ fixed, and then update θ on $\mathcal{D}_{inner}$ by minimizing a difficulty-weighted translator objective, where the weights w_i are treated as constants with respect to θ.

2.4 Anomaly Detection

At test time, given a clip $x(t) = (C_B(t), C_D(t), C_M(t))$ from $\mathcal{D}_{test}$, the inner translator produces a Doppler-like reconstruction $\hat{C}_D(t) = f_{trans}(C_B(t); \theta)$. We then compute the observed in-mask and out-of-mask reconstruction errors $e_{\text{in}}(t)$ and $e_{\text{out}}(t)$ (as defined in Sec. 2.2) and their log-transforms $y_{\text{in}}(t) = \log(e_{\text{in}}(t) + \varepsilon)$ and $y_{\text{out}}(t) = \log(e_{\text{out}}(t) + \varepsilon)$. Conditioned on the same input clip, the outer model outputs clip-wise error statistics $f_{\text{err}}(C_B(t); \phi) = (\mu_{\text{in}}(t), \sigma_{\text{in}}(t), \mu_{\text{out}}(t), \sigma_{\text{out}}(t))$. We define region-discriminated standardized deviations as

$$s_{\text{in}}(t) = \big(y_{\text{in}}(t) - \mu_{\text{in}}(t)\big) \,/\, \sigma_{\text{in}}(t), \qquad s_{\text{out}}(t) = \big(y_{\text{out}}(t) - \mu_{\text{out}}(t)\big) \,/\, \sigma_{\text{out}}(t), \quad (5)$$

and compute the clip-level anomaly score by a convex combination

$$s(t) = \alpha\, s_{\text{in}}(t) + (1 - \alpha)\, s_{\text{out}}(t), \quad (6)$$

where $\alpha \in [0, 1]$ controls the relative contribution of the Doppler color region versus the background. Larger values of $s(t)$ indicate that the observed translation errors are higher than expected for the given input after accounting for input-conditioned difficulty.

For each clip $x(t)$, we compute $s(t)$ as in Eq. 6 and report anomaly decisions at the participant level. Specifically, for a participant i, we extract clips from all available scan videos using a sliding window, compute the corresponding set of clip scores $\{s_k^{(i)}\}$, and define the participant-level anomaly score as the mean of the top 30% clip scores.

3 Experiments

Dataset Curation and Preprocessing We curated a paired B-mode/CFD dataset of 934 second-trimester participants with 3,800 spatially aligned scan videos from four medical centres: John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK; St George's University Hospitals NHS Foundation Trust, London, UK; Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK; and Gold Coast University Hospital, Southport, Australia. $\mathcal{D}_{inner}$ and $\mathcal{D}_{outer}$ include 247 healthy participants each (889 and 857 videos), and $\mathcal{D}_{test}$ includes 129 normal

(412 videos) and 311 CHD participants (1,642 videos). The CHD cohort covers six diagnoses: hypoplastic left heart syndrome (HLHS, $n = 22$), coarctation of the aorta (CoA, $n = 71$), persistent left superior vena cava (LSVC, $n = 68$), complete atrioventricular septal defect (CAVSD, $n = 27$), tetralogy of Fallot (TOF, $n = 44$), and ventricular septal defect (VSD, $n = 79$). Exams were acquired on a small set of ultrasound platforms dominated by Hitachi ARIETTA 850/70 and GE Voluson E10. For preprocessing, we computed an axis-aligned bounding square enclosing the CFD color-overlay sector, applied the same square crop to paired B-mode/CFD frames to preserve correspondence, resized to 256×256, and normalized intensities to $[0, 1]$.

Implementation Details Models were implemented in PyTorch [11] and trained on an NVIDIA Quadro RTX 6000 (24 GB) with batch size 4. We used a 3D UNet [12] and a 3D ResNet18 [7] as the inner and outer backbones, optimized with Adam [8] (lr 1×10^{-4}) for 600 epochs using clip length $L = 24$. For differentiable Doppler rendering we set $\gamma = 1.0$ and $k = 1.0$, and used within-batch weighting temperature $\eta = 1.0$. Loss weights were $\lambda_{\text{out}} = 0.2$, $\lambda_{\text{dice}} = 1.0$, and $\lambda_{\text{smooth}} = 0.2$, and anomaly scoring used $\alpha = 0.8$ in Eq. 6. Hyperparameters were selected empirically in preliminary experiments.

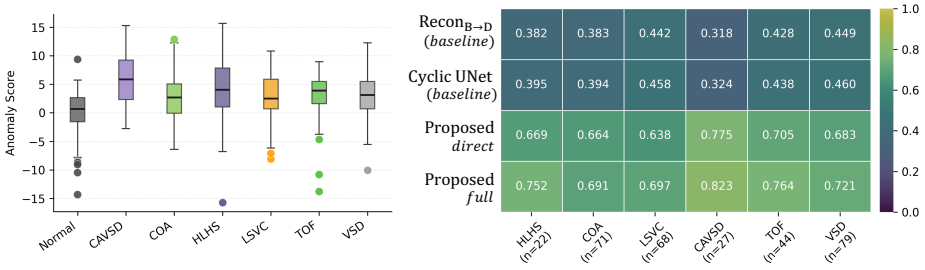

Fig. 2. Participant-level results by CHD type. Left: anomaly-score distributions (normal vs. six CHD types). Right: one-vs-normal AUROC heatmap comparing baselines and the ablation (**Proposed-direct**).

Table 1. Participant-level CHD anomaly detection performance on the test set. Mean±std is from 100 bootstrap resamples; random AUPRC=0.707.

Method	$AUROC_{all}$	$AUPRC_{all}$	Spec@Sens=75%
Recon$_{B\to D}$ [14]	0.416 ± 0.031	0.613 ± 0.033	0.192 ± 0.041
Cyclic UNet [9]	0.427 ± 0.032	0.621 ± 0.034	0.201 ± 0.041
Proposed-direct	0.674 ± 0.028	0.807 ± 0.026	0.481 ± 0.061
Proposed-full	**0.723 ± 0.026**	**0.858 ± 0.020**	**0.497 ± 0.073**

Comparison and Ablation Studies We compared the proposed method with reconstruction-based baselines trained on normal participants only. **Recon**$_{B\to D}$ scores anomalies using the B-mode→CFD reconstruction error [14]. **Cyclic UNet** employs a cyclic architecture and scores anomalies by the final B-mode error (B-mode→CFD→B-mode) [9]. We further report an ablation, **Proposed-direct**, which directly predicts CFD appearance without the explicit blood-flow-map representation and differentiable rendering used in our full model. Finally, **Proposed-full** denotes our full approach. Results are summarised in Table 1, with CHD-type-wise AUROC visualised in Fig. 2.

Evaluation Metrics All metrics are computed at the participant level. For participant i, we aggregate clip scores $\{s_k^{(i)}\}$ by averaging the top fraction $\rho = 0.3$, yielding $S_i = \text{mean}(\text{Top}_\rho(\{s_k^{(i)}\}))$. We report threshold-free AUROC$_{all}$ (all CHD vs. normal) and AUPRC$_{all}$. To analyse phenotype-specific behaviour, we compute type-wise AUROC by comparing each CHD type against normal participants (Fig. 2). For assistive triage, we additionally report Spec@Sens = 75% as a clinically motivated triage operating point to balance sensitivity and false-positive burden under variable Doppler acquisition.

4 Results and Discussion

CHD Anomaly Detection Table 1 summarises participant-level performance. **Proposed-full** consistently outperformed reconstruction-based baselines, achieving the highest overall AUROC and the best assistive-triage operating-point performance. **Proposed-direct** substantially improved over baselines but remained below **Proposed-full**, indicating additional benefit from explicit flow-map learning beyond direct Doppler prediction. Fig. 2 shows CHD-type-wise AUROC: both proposed variants improved AUROC for all reported types, with the largest gains for phenotypes with more distinctive hemodynamic deviations, while baselines stayed near chance for most types. Fig. 2 also shows anomaly-score distributions for **Proposed-full**, where CHD groups shift to higher scores than normal, supporting a unified triage decision rule. All improvements observed are statistically significant (p-value< 0.01; paired two-sided t-tests on bootstrapped resamples).

Case Studies Fig. 3 presents two qualitative examples to demonstrate the interpretability of the proposed region-discriminated anomaly score. In the HLHS case (Case 1), the ground-truth CFD shows little or no flow within the left ventricle (LV), whereas the reconstruction hallucinates a strong LV colour overlay with intensity comparable to the right ventricle (RV) (yellow arrows), yielding a prominent out-of-mask error signal. In the VSD case (Case 2), the ground-truth CFD exhibits a color jet crossing the interventricular septum (yellow circles), but the reconstruction fails to reproduce this abnormal region and instead follows a normal-like pattern, which elevates the in-mask error signal.

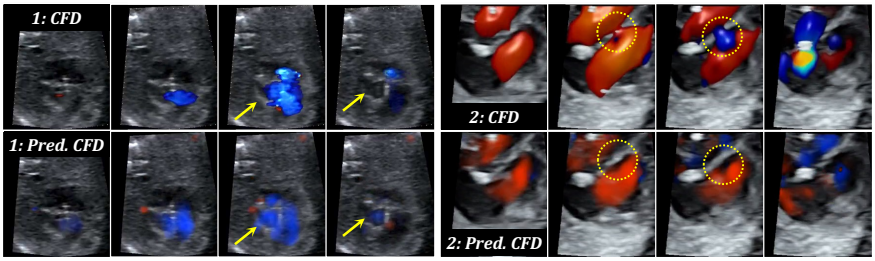

Fig. 3. Qualitative case studies demonstrating region-discriminated interpretability. Each case shows a temporal sequence over half cardiac cycle (top to bottom: ground-truth CFD, reconstructed Doppler).

5 Conclusion

We propose a bi-level optimized CHD anomaly detector for fetal echocardiography using cross-modal translation with region-calibrated errors. On a multi-site cohort, it significantly outperforms reconstruction baselines and yields interpretable error patterns. Future work will expand view coverage and Doppler diversity, and reduce reliance on Doppler data at inference to support broader clinical scenarios.

Acknowledgments. This work was supported by the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography) and partly supported by the UKRI grant EP/X040186/1 (Turing AI Fellowship). Co-authors J. Alison Noble and Aris Papageorghiou were supported by the Oxford Partnership Comprehensive Biomedical Research Centre with funding from the NIHR Biomedical Research Centre (BRC) funding scheme.

Disclosure of Interests. The authors have no competing interests.

References

1. Bhatnagar, N.T.R., Jain, S., Lau, C., Liu, T.Y., Gambini, L., Arnaout, R.: Generative deep learning for foundational video translation in ultrasound. arXiv preprint arXiv:2511.03255 (2025)
2. Carvalho, J., Axt-Fliedner, R., Chaoui, R., Copel, J., Cuneo, B., Goff, D., Gordin Kopylov, L., Hecher, K., Lee, W., Moon-Grady, A., et al.: Isuog practice guidelines (updated): fetal cardiac screening. Ultrasound Obstet Gynecol **61**(6), 788–803 (2023)
3. Day, T.G., Kainz, B., Hajnal, J., Razavi, R., Simpson, J.M.: Artificial intelligence, fetal echocardiography, and congenital heart disease (2021)
4. Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Deep learning for medical anomaly detection–a survey. ACM Computing Surveys (CSUR) **54**(7), 1–37 (2021)

5. Guo, J., Bu, R., Shen, W., Feng, T.: Towards robust multimodal ultrasound classification for liver tumor diagnosis: A generative approach to modality missingness. Computer Methods and Programs in Biomedicine **265**, 108759 (2025)
6. He, C., Ye, H., Shen, L., Zhang, T.: Milenas: Efficient neural architecture search via mixed-level reformulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11993–12002 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Liang, Z., Anthony, H., Wagner, F., Kamnitsas, K.: Modality cycles with masked conditional diffusion for unsupervised anomaly segmentation in mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 168–181. Springer (2023)
10. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS. vol. 32 (2019)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
13. Roshanitabrizi, P., Guo, P., Aharonyan, A.A., Brown, K., Broudy, T.G., Parida, A., Tapp, A., Jiang, Z., Tompsett, A., Rwebembera, J., et al.: Synthesis of pathological dual-channel color doppler echocardiograms for equitable diagnosis of heart diseases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–599. Springer (2025)
14. Sakurada, M., Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA'14). pp. 4–11. Association for Computing Machinery (2014). https://doi.org/10.1145/2689746.2689747
15. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)
16. Van Der Linde, D., Konings, E.E., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J., Roos-Hesselink, J.W.: Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. Journal of the American College of Cardiology **58**(21), 2241–2247 (2011)