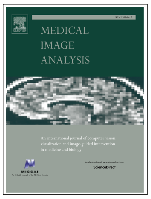
Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media



# Robust fine-grained echocardiographic view classification with supervised contrastive learning

Preshen Naidoo [a,*], Patricia Fernandes [a], Nasim Dadashi Serej [a], Catherine C Stowell [b], Charlotte H Manisty [c], Darrel P Francis [b], Massoud Zolgharni [a,b]

[a] *Translational Healthcare Research Centre, University of West London, UK*
[b] *National Heart and Lung Institute, Imperial College London, UK*
[c] *Barts Heart Centre and University College London, UK*

ARTICLE INFO

*Keywords:*
Echocardiography
Contrastive learning
View classification
Observer variability

ABSTRACT

Accurate classification of echocardiographic views is fundamental for automated cardiac analysis. However, clinical practice relies on a large, heterogeneous set of fine-grained acquisitions that introduce substantial inter-observer variability. Existing studies have primarily focused on limited view sets, often collapsing specialised views into broad categories, which limits their clinical relevance. We address this limitation by introducing TTE47, the first publicly available benchmark comprising 47 clinically meaningful views annotated independently by three experts. This dataset enables the rigorous quantification of inter-observer agreement and establishes a foundation for reproducible, clinically relevant evaluation. To tackle the dual challenges of subtle inter-class distinctions and structured label variability, we propose a novel supervised contrastive learning framework incorporating a tailored loss function. Our method outperforms cross-entropy and standard supervised contrastive baselines, achieving leading performance among evaluated methods on TTE47 and surpassing prior work on TMED-2 without dataset-specific pretraining, using a model pretrained on TTE47. Beyond accuracy, we introduce clustering-based metrics, Detection Rate and Label Recovery Precision, that measure semantic coherence and the model's ability to resist annotation variability. Results show that the learned feature space aligns more strongly with underlying anatomical structure than with any single annotator's style, enabling resilience to label shifts and maintaining robustness comparable to human-level disagreement. By integrating multi-expert evaluation, robust representation learning, and interpretable feature-space analysis, this work establishes a scalable and clinically relevant framework for fine-grained echo view classification. Our findings highlight the potential of contrastive pretraining to standardise interpretation, mitigate subjectivity, and enhance the reliability of AI-assisted echocardiography in diverse clinical settings.

## 1. Introduction

Cardiovascular disease remains a leading cause of global mortality, and echocardiography (echo) is a key non-invasive imaging modality for diagnosing and monitoring cardiac conditions (Steffner et al., 2024). Among echo techniques, transthoracic echocardiography is the most widely used. Professional societies recommend acquiring a standard set of cardiac views for every examination, with additional views obtained as clinically indicated. In a typical echo study, a sonographer manually manipulates a handheld transducer on the patient's chest, adjusting its angle and position to capture the required anatomical perspectives of the heart.

Accurate identification of echocardiographic views is a critical prerequisite for any automated analysis pipeline, as each standard view provides a distinct anatomical perspective and is associated with specific diagnostic measurements (Madani et al., 2018; Li et al., 2024). Misclassification of views can lead to erroneous interpretations on downstream analyses (Wegner et al., 2022). For example, confusing the apical two-, three-, and four-chamber views corrupts global longitudinal strain and biplane Simpson volumes/ejection fraction, as these planes are not interchangeable. However, this task is non-trivial as many views differ only subtly in appearance, and ambiguous or intermediate views arising from variations in probe orientation, patient-specific anatomy, or intentional acquisition of nonstandard perspectives further complicate classification

[*] Corresponding author.
*E-mail addresses:* preshen.naidoo@uwl.ac.uk (P. Naidoo), patricia.fernandes@uwl.ac.uk (P. Fernandes), nasim.dadashiserej@uwl.ac.uk (N. Dadashi Serej), c.stowell@imperial.ac.uk (C.C Stowell), c.manisty@ucl.ac.uk (C.H Manisty), d.francis@imperial.ac.uk (D.P Francis), massoud.zolgharni@uwl.ac.uk (M. Zolgharni).

https://doi.org/10.1016/j.media.2026.104006
Received 21 August 2025; Received in revised form 23 February 2026; Accepted 23 February 2026
Available online 26 February 2026

(Jeon et al., 2023; Letnes et al., 2021). Unlike other imaging modalities, echo is susceptible to such variability, making standardisation difficult (Naser et al., 2024).

Manual interpretation remains the clinical standard but is labour-intensive, expertise-dependent, and prone to intra- and inter-observer variability, leading to inconsistencies in both view identification and subsequent diagnoses (Liao et al., 2020; Azarmehr et al., 2021). This subjectivity, combined with the high volume of echocardiographic studies performed daily, creates significant workload pressures and challenges for consistency at scale. The complexity of echo images, often depicting overlapping cardiac structures, further adds to the difficulty. For example, apical two- and four-chamber views both include the left atrium and ventricle, yet differ in the presence of right-sided chambers; even minor changes in transducer angle can produce ambiguous representations between such views (Gearhart et al., 2022; Jeon et al., 2023).

To address these challenges, there is a growing need for automated, accurate, and scalable view classification methods (Zhang et al., 2018). A comprehensive transthoracic echocardiographic examination encompasses a broad range of views (Mitchell et al., 2018; Wegner et al., 2022), and certain abnormalities are only visible or best assessed from specialised views (we provide a detailed description of each view including its clinical use in Appendix A). Deep learning approaches have shown strong potential in this domain but have often been evaluated on limited subsets of views, limiting their clinical utility (Ghorbani et al., 2020; Zamzmi et al., 2022). Existing models tend to perform well on visually distinct views but degrade in accuracy when faced with a larger and more diverse set of classes (Madani et al., 2018; Jeon et al., 2023). As the number of views increases, distinctions become increasingly nuanced, contributing to annotation complexity and model confusion (Howard et al., 2020).

Contrastive learning has emerged as a powerful strategy for learning high-quality representations across diverse domains (Holste et al., 2024). Although extensively explored in self-supervised settings, its application in supervised contexts, particularly for echo view classification, remains limited. In this domain, observer variability introduces structured label noise due to expert subjectivity, which can degrade model performance (Howard et al., 2020; Azarmehr et al., 2021). To address this significant challenge, we propose a novel supervised contrastive learning framework designed to improve robustness to such variability, thereby enabling accurate fine-grained classification.

## 1.1. Related work

### 1.1.1. View classification

Previous studies have shown that convolutional neural networks (CNNs), often pretrained on ImageNet, can achieve near expert-level accuracy when classifying a limited set of common echocardiographic views. For example, Madani et al. (2018) trained a CNN to classify 15 transthoracic echo views, achieving 91.7% accuracy on single frames and 97.8% using multi-frame inputs across 12 views. Naser et al. (2024) applied 2D and 3D ResNet-18 models to classify 9 views (including an “other” category), achieving 96.8% and 96.6% accuracy on internal and external test sets, respectively. Li et al. (2024) trained a multi-task CNN on over 170,000 images spanning 6 views plus “other,” attaining 97.8% accuracy. In contrast-enhanced studies, Zhu et al. (2022) used EchoVNet to classify 9 views, achieving 99.1% (non-contrast) and 99.5% (contrast) accuracy. These works highlight that for constrained sets of common views, CNNs can reach high performance.

Some studies have extended classification to broader sets. Jeon et al. (2023) expanded view classification to 25 standard views and 6 additional categories by dividing the task into modality-specific groups (15 Doppler, 3 M-mode, 13 B-mode), each handled by a separate deep learning model, achieving accuracies of 98.3%, 99.5%, and 96.6%, respectively. Gearhart et al. (2022) developed a pediatric echocardiography model, classifying 27 pediatric view types with 90.3% accuracy, while Zhang et al. (2018) proposed a fully automated pipeline that reached 84% accuracy across 23 views, though performance degraded at larger view scales.

Overall, prior studies demonstrate that deep learning achieves strong accuracy when distinguishing among a small number of views, but performance typically declines as the number of classes increases and distinctions become more subtle (Ostvik et al., 2019; Howard et al., 2020). Azarmehr et al. (2021) report that their 1-cell-DARTS model achieves an 8% absolute performance gain when the target label set is reduced from 14 to 5 views, with prominent confusion observed between apical five-chamber and the LV-focused apical four-chamber. Likewise, Howard et al. (2020) note that distinguishing apical two-chamber from apical three-chamber is a common source of error, reflecting the adjacency of these imaging planes.

To our knowledge, this work is the first to classify all 47 views routinely acquired in a comprehensive transthoracic echocardiography examination, including both core views and secondary variants (e.g., zoomed or measurement-focused acquisitions), without collapsing them into catch-all “other” categories. This fine-grained approach captures the full complexity of clinical practice and enables more detailed analysis for both workflow automation and interpretation.

Importantly, clinical guidelines prescribe not only core views but also specialised and focused variants, such as right-ventricle–focused apical two-chamber or modified apical four-chamber views for atrial septum assessment (Mitchell et al., 2018; Otto, 2018). These finer subdivisions have been shown to provide critical diagnostic information (Khamis et al., 2017; Vaseli et al., 2019; Jeon et al., 2023), particularly when pathology is localised to specific regions only visible in modified planes (Naser et al., 2024). However, scaling to many views introduces challenges: subtle probe angulations can cause transitions between related views (e.g., apical four- vs. five-chamber), complicating annotation and increasing inter-observer variability. Thus, while higher granularity enhances diagnostic specificity, it also demands models that are robust to annotation inconsistency and capable of capturing the intrinsic semantic structure of echocardiographic images.

### 1.1.2. Contrastive learning

In recent years, batch-based methods in contrastive learning have become a leading approach in self-supervised representation learning, enabling models to learn discriminative features by pulling positive pairs closer and pushing many negative pairs apart in the feature space (He et al., 2019; Chen et al., 2020; Zbontar et al., 2021). By training the network to increase similarity between positive pairs while decreasing similarity to negative examples, the resulting feature embeddings capture a meaningful semantic structure, enhancing inter-class separability and intra-class similarities.

In contrastive learning, the loss function plays a central role in shaping discriminative representations, supported by the combined effects of augmentations, sampling strategy, and network architecture. van den Oord et al. (2019) proposed InfoNCE, a loss function for contrastive representation learning, which generalises the noise contrastive estimation (NCE) framework (Gutmann and Hyvärinen, 2010) by replacing binary classification with a softmax over multiple negatives and has become the foundation for many modern contrastive learning methods.

Chen et al. (2020) extend InfoNCE by introducing the NT-Xent loss, which applies embedding $l_2$-normalisation, temperature scaling, and symmetric loss computation by evaluating the loss in both directions of each positive pair, improving representation learning in the SimCLR framework. Khosla et al. (2020) extend this by proposing SupCon, a supervised variant of the NT-Xent loss that leverages label information to define positive pairs, enabling more effective use of class-level supervision in contrastive learning.

While contrastive learning has been extensively studied in the context of self-supervised learning, significantly fewer studies have focused on exploring contrastive objectives for supervised settings, particularly for echo view classification.

#### 1.1.3. Sensitivity of contrastive objectives to label uncertainty

While contrastive learning is a powerful approach for representation learning, a fundamental challenge is mitigating the effects of false negatives and false positives, which can lead to suboptimal representations that hinder model performance (Huynh et al., 2022; Xu et al., 2024). False positives occur when pairs of examples are incorrectly considered similar despite belonging to different classes, while false negatives arise when samples from the same class are mistakenly treated as dissimilar; often due to noisy labels. Such noise can result from observer variability, particularly in domains like medical imaging where labelling is inherently subjective (Dedieu et al., 2024; Zhang et al., 2025).

Unlike self-supervised methods such as SimCLR, which utilise a single positive per anchor (typically an augmented view of the same image), supervised contrastive learning (SCL) incorporates multiple positives per anchor, essentially all other samples of the same class in a batch (Khosla et al., 2020). SCL relies on the assumption of correct labels to construct positive pairs, and when mislabelled examples cause false positive pairings, the contrastive training signal becomes misleading, which in turn corrupts the learned representation (Li et al., 2022).

Graf et al. (2021) showed that as label noise increases, the number of training iterations required for SCL to fit the data grows super-linearly, much faster than for a cross-entropy model, indicating that SCL struggles to learn effectively in the presence of noisy labels. Such noise introduces false relationships that distort the learned feature space.

While extensive research has studied the impact of false negatives in self-supervised contrastive learning (Huynh et al., 2022; Chen et al., 2022; Xu et al., 2024; Auh et al., 2024), studies focusing on false positives in supervised contrastive learning remain limited, particularly in medical imaging where label noise and observer variability further complicate the identification of reliable positive pairs.

#### 1.1.4. Strategies for handling false positives and false negatives

Several studies have introduced threshold-based strategies to mitigate the effects of false negatives in contrastive learning, particularly in self-supervised scenarios. Such thresholds can be defined based on cosine similarity, predetermined values, top-$k$ similarity rankings, or loss-based confidence metrics.

For false negatives, common in self-supervised settings, Huynh et al. (2022) proposed filtering the most similar negatives using a combination of a top-$k$ ranking and a fixed similarity threshold. Similarly, Xu et al. (2024) introduced both absolute and relative similarity thresholds for identifying negatives that may semantically align with the anchor. Chen et al. (2022) applied pseudo-label confidence thresholds derived from clustering structures in the embedding space, while Auh et al. (2024) dynamically adjusted cosine similarity thresholds and temperatures during training to refine false negative detection.

In the context of supervised or semi-supervised learning, where false positives are more prevalent, thresholds are also widely used. In medical imaging, several studies leveraged thresholds on likelihood ratios or loss values to correct noisy labels (Huang et al., 2022a; Jiang et al., 2023; Shi et al., 2024). Similarly, Li et al. (2022) used class-wise dynamic thresholds to select confident sample pairs based on both cross-entropy loss and representation similarity. Other studies proposed feature similarity-based thresholds to identify reliable positive pairs or cluster-consistent samples (Guo et al., 2025; Guan et al., 2024). Recently, (Zhang et al., 2025) introduced a dual-branch framework for classifying samples into clean, hard, or noisy partitions using a fixed soft-label confidence threshold.

Despite these studies, many thresholding methods can be difficult to tune and are highly dependent on dataset characteristics, noise types, and training dynamics. Guo et al. (2025) observe that optimal threshold values vary across datasets, highlighting the need for dataset-specific tuning. Li et al. (2022) adjust their parameters which define their thresholds according to the dataset and the type of noise. Multiple studies highlight the difficulty in estimating the precise noise rate of noisy labels or pairs, which is often required to set an accurate threshold (Li et al., 2022; Jiang et al., 2023). Other studies show that performance can be highly sensitive to threshold values and associated hyperparameters, sometimes requiring dynamic adjustment during training (Auh et al., 2024; Huynh et al., 2022; Zhang et al., 2025). Guan et al. (2024) highlight that inaccurate similarity thresholds may result in the exclusion of true positives or the inclusion of false positives.

In the context of view classification involving 47 echocardiographic views, semantically similar classes often exhibit high similarity scores. This increased similarity between closely related views presents a significant challenge for threshold-based strategies, as distinguishing true positives from false ones becomes more difficult.

### 1.2. Main contributions

Expanding the number of echo views enhances clinical insight by capturing more comprehensive anatomical information, but also increases interpretation variability, leading to annotation noise often manifesting as false positives. These weaken the contrastive objective by pulling semantically dissimilar representations closer together, degrading inter-class separability. We evaluate contrastive learning strategies that mitigate label noise and expert disagreement while preserving high discriminability among subtly distinct views.

The main contributions of this work are outlined as follows:

- **Fine-grained classification of all echo views:** We present the first study to classify all 47 views routinely acquired in a comprehensive transthoracic echocardiography examination. This includes both core views and secondary variants, such as zoomed or measurement-focused acquisitions, without aggregating them into catch-all “other” categories. This fine-grained approach captures the full complexity of real-world diagnostic practice, providing a more detailed and clinically relevant analysis.
- **TTE47 Benchmark:** We introduce *TTE47*, the first publicly available dataset for fine-grained transthoracic echocardiography view classification, distinguished by comprehensive multi-expert annotations. The fully annotated test set contains more than 5000 images, each independently labelled by three expert clinicians, enabling rigorous multi-observer evaluation under realistic clinical conditions. This dataset provides a common benchmark for the echocardiography community to compare classifier performance using a standardised reference set, and can be adapted for models targeting fewer classes by merging corresponding families of views for evaluation.
- **Contrastive Learning Framework:** We develop and evaluate supervised contrastive learning strategies tailored for noisy multi-class echo image classification. Our proposed framework, incorporating a tailored objective, improves robustness under label ambiguity, outperforming current supervised contrastive and cross-entropy baselines, especially on difficult or under-represented views.
- **EchoFine Pretrained and Fine-Tuned Models:** We release both pretrained and fine-tuned model weights. These resources provide the research community with high-quality, task-specific initialisation. This not only ensures full reproducibility of our results but also offers a strong foundation for advancing echocardiographic analysis in both research and clinical settings.
- **Representation Space Evaluation:** We introduce a principled clustering-based analysis to assess the structure and robustness of the learned representation space. Semantic alignment is quantified using two novel metrics, Detection Rate (DR) and Label Recovery Precision (LRP), which measure the model’s ability to identify corrupted samples and recover the original class structure under synthetic label noise. These metrics provide insights into the stability and interpretability of the learned features beyond conventional accuracy measures.
- **Observer Variability Analysis:** We perform a detailed inter-observer analysis using triple-annotated test data to quantify expert disagreement and assess model alignment. Results show that

the model generalises beyond any single annotator, approximating consensus-driven labelling behaviour. This highlights that expert agreement correlates with model confidence and reliability, thereby demonstrating the importance of evaluating under multi-expert conditions.

- **Robustness to Annotator Identity:** We evaluate the model's resilience to changes in annotator identity by simulating inter-expert variation in the training set, based on disagreement patterns observed among three independent experts in the test set. This analysis assesses whether model performance remains stable if the training set were annotated by a different expert, reflecting a critical requirement for clinical deployment.
- **Cross-Dataset Generalisation:** We demonstrate the generalisability of our model through zero-shot inference on an independent, publicly available dataset (TMED-2), and further show that fine-tuning surpasses prior state-of-the-art performance. This highlights the robustness and transferability of representations learned from fine-grained echo views to external, coarse-grained domains.

## 2. Methods

### *2.1. Datasets*

**TTE47:** A random sample of real-world echocardiographic studies was extracted from the Imperial College Healthcare NHS Trust database, yielding a total of 91,139 images. Ethical approval for this study was obtained from the Health Research Authority (Integrated Research Application System [IRAS] identifier: 243023). Only studies with complete demographic information were included (Table 1).

In this hospital routine transthoracic echo exams are predominantly non-contrast, while contrast is mainly used in stress echocardiography for qualitative assessment. Scans were acquired on a broad range of ultrasound systems: Philips (iE33, Affiniti 70C, EPIQ 7C, Affiniti 50G, CX50), GE (Vivid i, Vivid q), and Siemens, providing device diversity. The dataset comprises randomly sampled studies from this hospital's echocardiography archive and reflects the full clinical spectrum of patients undergoing transthoracic echocardiography. It includes cases with diverse cardiac pathologies such as dilated cardiomyopathy, ischemic cardiomyopathy, valve stenosis, regurgitation, effusions, as well as examinations without identified heart disease.

Each image was manually annotated by a cardiologist ("Expert-1") using our web-based platform (unityimaging.net), and classified into one of 47 predefined categories, as detailed in the Tables A.14–A.19. The dataset was partitioned into training (76,589), validation (9,103), and test (5,447) subsets, as summarized in Table 2. Echocardiograms were obtained across 19,169 studies, and there was no overlap of patients or studies across the training, validation, and test sets.

To evaluate inter-observer variability, two additional clinical experts independently annotated the test set. These experts were blinded to both the original labels and to each other's annotations. In contrast to Expert-1, who was required to assign a single label to each image, Experts 2 and 3 were given the option to select 'not sure' for ambiguous cases. The test set size was selected to be sufficiently large for a reliable evaluation while remaining manageable given the intensive manual effort required for multi-expert annotation across 47 distinct views.

The test subset of 5447 images and corresponding labels from all three clinicians is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license. The release of this dataset received a favorable opinion from the South Central – Oxford C Research Ethics Committee (Integrated Research Application System ID: 279328, REC reference: 20/SC/0386). Due to data governance restrictions, the training set cannot be made publicly available, however, we provide both pretrained and fine-tuned model weights.

**Table 1**
Demographic characteristics of the TTE47 dataset. The study includes a broad distribution of adult patients, with representation across age groups, BMI categories, and sexes. Data were acquired using a diverse set of scanner manufacturers and models, reflecting real-world clinical heterogeneity.

| Characteristic | Category | Proportion (%) |
|---|---|---|
| Age (years) | 18–30 | 2.3 |
| | 31–50 | 12.7 |
| | 51–70 | 27.9 |
| | 71 + | 57.2 |
| Sex | Female | 50.5 |
| | Male | 48.6 |
| | Other | 0.8 |
| BMI (kg/m$^2$) | 18–24.9 (Normal) | 36.0 |
| | 25–29.9 (Overweight) | 34.9 |
| | 30–34.9 (Obesity I) | 18.0 |
| | 35 + (Obesity II + ) | 11.1 |

**Table 2**
Distribution of echo view classes across the training, validation, and test sets, indicating the number of samples per view in each subset.

| View | Training | Validation | Testing | Total |
|---|---|---|---|---|
| a2ch-full | 2931 | 349 | 209 | 3489 |
| a2ch-la | 1936 | 230 | 138 | 2304 |
| a2ch-lv | 2987 | 355 | 213 | 3555 |
| a3ch-full | 3561 | 424 | 254 | 4239 |
| a3ch-la | 1622 | 193 | 115 | 1930 |
| a3ch-lv | 1926 | 229 | 137 | 2292 |
| a3ch-outflow | 637 | 75 | 45 | 757 |
| a4ch-full | 3353 | 399 | 239 | 3991 |
| a4ch-ias | 1319 | 157 | 94 | 1570 |
| a4ch-la | 2733 | 325 | 195 | 3253 |
| a4ch-lv | 2495 | 297 | 178 | 2970 |
| a4ch-ra | 1713 | 204 | 122 | 2039 |
| a4ch-rv | 1236 | 147 | 88 | 1471 |
| a5ch-full | 753 | 89 | 53 | 895 |
| a5ch-outflow | 786 | 93 | 56 | 935 |
| apex | 723 | 86 | 51 | 860 |
| doppler-ao | 1489 | 177 | 106 | 1772 |
| doppler-av | 2587 | 308 | 184 | 3079 |
| doppler-mv | 1701 | 202 | 121 | 2024 |
| doppler-pv | 1776 | 211 | 126 | 2113 |
| doppler-tissue-lateral | 1120 | 133 | 80 | 1333 |
| doppler-tissue-rv | 780 | 92 | 55 | 927 |
| doppler-tissue-septal | 1034 | 123 | 73 | 1230 |
| doppler-tv | 1902 | 226 | 135 | 2263 |
| mmode-a4ch-rv | 1592 | 189 | 113 | 1894 |
| mmode-ivc | 1159 | 138 | 82 | 1379 |
| mmode-plax-av | 1374 | 163 | 98 | 1635 |
| mmode-plax-lv | 492 | 58 | 35 | 585 |
| mmode-plax-mitral | 768 | 91 | 54 | 913 |
| plax-full-la | 1123 | 133 | 80 | 1336 |
| plax-full-lv | 1799 | 214 | 128 | 2141 |
| plax-full-mv | 1307 | 155 | 93 | 1555 |
| plax-full-out | 1509 | 179 | 107 | 1795 |
| plax-full-rv-ao | 961 | 114 | 68 | 1143 |
| plax-tv | 1688 | 201 | 120 | 2009 |
| plax-valves-av | 2000 | 238 | 142 | 2380 |
| plax-valves-mv | 1841 | 219 | 131 | 2191 |
| psax-all | 2097 | 249 | 149 | 2495 |
| psax-tv | 1335 | 159 | 95 | 1589 |
| psax-av | 2125 | 253 | 151 | 2529 |
| psax-pv | 1868 | 222 | 133 | 2223 |
| psax-lv-base | 1437 | 171 | 102 | 1710 |
| psax-lv-mid | 1669 | 198 | 119 | 1986 |
| psax-lv-apex | 1627 | 193 | 116 | 1936 |
| subcostal-heart | 1212 | 144 | 86 | 1442 |
| subcostal-ivc | 849 | 101 | 60 | 1010 |
| suprasternal | 1657 | 197 | 118 | 1972 |
| **Total** | **76589** | **9103** | **5447** | **91139** |

**TMED2:** To assess the generalisability of our framework beyond the TTE47 domain, we evaluate it on the publicly available TMED2 dataset (Huang et al., 2022b), which provides annotations for echo view classi-

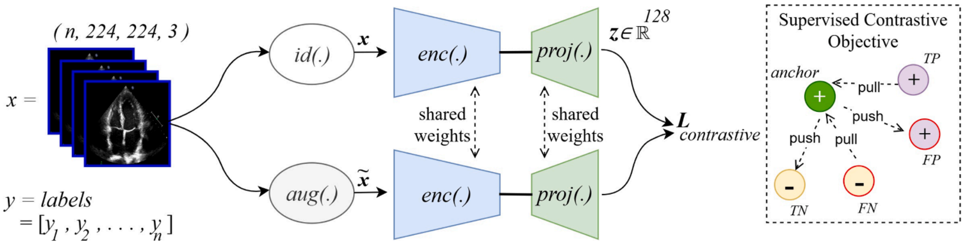


**Fig. 1.** Overview of the proposed contrastive learning framework (EchoFine). Augmented and original images are processed through encoders and projectors with shared weights. A supervised contrastive loss, less sensitive to false positives, is used to optimise the representations.

fication. Following the official DEV479 split for training, validation, and testing, we use the predefined view categories and report performance as the mean classification accuracy across the three predefined folds.

### 2.2. *Multi-expert annotations*

In the test set, 553 cases showed complete disagreement among the three expert annotators, reflecting the inherent ambiguity of certain echocardiographic views. These cases pose a challenge for performance evaluation, as no definitive ground truth can be established.

To provide a more comprehensive assessment, we extend evaluation beyond a single test set metric by analysing meaningful subsets of the test set and complementary performance criteria. This approach offers a more nuanced view of the model's practical utility in real-world scenarios, where inter-observer variability is common.

The evaluation framework consists of:

1. **Primary Annotator Set ($E_1$):** The full original test set ($n = 5{,}447$) annotated by Expert-1.
2. **Consensus Set ($CS$):** Images where all three experts independently assigned the same label ($n = 3{,}327$), representing the highest-confidence labels.
3. **Union-Agreement Evaluation ($UA$):** A performance metric in which a model prediction is considered correct if it matches the label assigned by *any* of the three experts, providing a broader measure of agreement.

*Let* $E_1$, $E_2$, and $E_3$ denote the annotations from Experts 1, 2, and 3, respectively. The consensus set is defined as:

$$CS \subseteq E_1 \cap E_2 \cap E_3 \quad \text{where} \quad \text{label}_{E_1}(x) = \text{label}_{E_2}(x) = \text{label}_{E_3}(x).$$

The Union-Agreement evaluation is defined as:

$$UA = \left\{ x \in E_1 \,:\, \hat{y}(x) \in \left\{ \text{label}_{E_1}(x), \text{label}_{E_2}(x), \text{label}_{E_3}(x) \right\} \right\}.$$

This stratified framework captures both inter-observer consistency (via subsets such as $CS$) and broader agreement measures (via $UA$), which are particularly relevant in clinical practice where minor differences between expert opinions are common.

### 2.3. *Representation learning framework*

Fig. 1 illustrates the core components of the proposed framework, which is conceptually inspired by prior work (Chen et al., 2020; Khosla et al., 2020), but specifically adapted for fine-grained echo view classification. Spatial augmentation, denoted as $aug(\cdot)$, applies mild random transformations, including rotations within ±10°, zooming with a factor of 0.2, and translations up to 10% of the frame size.

We form asymmetric pairs by applying the identity mapping $\text{id}(\cdot)$ to one view. The encoder, denoted $enc(\cdot)$, maps each image to a compact latent representation in $\mathbb{R}^{2048}$. This is followed by a non-linear projection head, $proj(\cdot)$, composed of two fully connected layers, each with 128 units. The first layer employs a ReLU activation function, resulting in final projections in $\mathbb{R}^{128}$.

As the diversity of echo views increases, so does the similarity among many view types, making the contrastive objective more challenging. One of the ways we address this in the proposed framework is by applying an asymmetric augmentation strategy: one input in each contrastive pair remains unaltered, while the other undergoes mild augmentations. This design encourages the model to capture clinically meaningful variability that reflects real-world acquisition differences, rather than overfitting to unrealistic transformations. The second component is a specialised supervised contrastive loss, detailed in Section 2.4.

Notably, some augmentation techniques, such as cropping, may inadvertently generate images that resemble other valid views (Chartsias et al., 2021). By limiting augmentation severity, the framework promotes discrimination between subtle but anatomically valid variations, enhancing robustness in the presence of closely related view classes.

### 2.4. *Contrastive learning objectives*

For this framework, we investigate tailored variants of the supervised contrastive loss (SupCon), each modifying the objective to mitigate false positives through distinct formulation strategies. For clarity and alignment, we adopt the same variable naming conventions used in prior studies.

The supervised contrastive objective is defined over a multiviewed mini-batch constructed from $N$ original image-label pairs $\{(x_k, y_k)\}_{k=1}^{N}$. Each image $x_k$ yields two views: one unaltered and one mildly augmented. This results in a set of $2N$ inputs $\{(\tilde{x}_\ell, \tilde{y}_\ell)\}_{\ell=1}^{2N}$, where $\tilde{x}_{2k-1} = x_k$ and $\tilde{x}_{2k} = \text{aug}(x_k)$, with $\tilde{y}_{2k-1} = \tilde{y}_{2k} = y_k$ for $k = 1, \ldots, N$.

Each sample is processed through a shared encoder and projection head to produce a $d$-dimensional embedding:

$$\mathbf{z}_\ell = \text{Proj}(\text{Enc}(\tilde{x}_\ell)) \in \mathbb{R}^d$$

To ensure that similarity scores are bounded and comparable, the output embeddings are $\ell_2$-normalized such that $\|\mathbf{z}_\ell\|_2 = 1$ for all $\ell$.

For a given anchor index $i \in \{1, \ldots, 2N\}$, let $P(i)$ denote the set of indices corresponding to other samples in the batch that share the same label:

$$P(i) = \{p \in A(i) \,:\, \tilde{y}_p = \tilde{y}_i\}$$

where $A(i) = \{1, \ldots, 2N\} \setminus \{i\}$ includes all other samples in the multiviewed batch (positives and negatives, excluding i). The temperature scaling factor is $\tau > 0$.

The total loss over a batch of anchors:

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{2N} \mathcal{L}_i$$

where $\mathcal{L}_i$ is one of the following contrastive objectives.

#### 2.4.1. *Supervised contrastive loss (SupCon)*

The standard supervised contrastive loss (Khosla et al., 2020) is given by:

$$\mathcal{L}_i^{\text{sup}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \left( \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i^\top \mathbf{z}_a / \tau)} \right) \tag{1}$$

In the SupCon formulation, the denominator aggregates similarities between the anchor and all other samples in the batch, excluding the anchor itself. This set inherently includes both false positives and false negatives. While all non-anchor samples contribute to the denominator uniformly, false positives in the numerator directly influence the optimisation objective by encouraging alignment between semantically dissimilar embeddings. As such, erroneous positives in the numerator may distort the learned representation space more severely than their presence in the denominator. Therefore, our focus is on mitigating the impact of false positives within the numerator.

### 2.4.2. Drop contrastive loss (DropCon)

To improve robustness to false positives without imposing fixed weights or similarity thresholds, we propose a variant, termed **DropCon**, which discards the $k$ least similar positives from the numerator summation. This method is motivated by the observation that in real-world datasets, particularly in fine-grained view classification, some false positives may exhibit high similarity to true positives, especially during early training stages.

$$\mathcal{L}_i^{\text{drop}} = -\frac{1}{|P(i)| - k} \sum_{p \in P(i) \setminus D(i)} \log \left( \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i^\top \mathbf{z}_a / \tau)} \right) \tag{2}$$

Here, $D(i) \subseteq P(i)$ is the subset of the $k$ least similar positives for anchor $i$, identified by sorting the dot products $\mathbf{z}_i^\top \mathbf{z}_p$ (embeddings are $\ell_2$-normalised). The remaining positives in $P(i) \setminus D(i)$ are used to compute the contrastive loss. When the number of positives satisfies $|P(i)| \leq \upsilon$ for some constant $\upsilon > k$, DropCon defaults to the standard SupCon loss.

Top-$k$ and DropCon both rank positives by similarity, but they differ in a way that matters for fine-grained regimes. Top-$k$ retains only the $k$ most similar positives and discards the rest; when intra-class similarities are uniformly high (as in our 47-view setting), this can inadvertently remove many true positives and makes performance sensitive to the choice of $k$ (which is further complicated by batch-to-batch variation in $|P(i)|$). By contrast, DropCon trims only the $k$ least similar positives, the lowest-confidence tail, while preserving the diversity of the remaining positives and suppressing likely mislabelled pairs. This tail trimming is typically less sensitive to $k$ for small $k$ relative to $|P(i)|$, and we revert to SupCon when $|P(i)| \leq \upsilon$, avoiding brittle behaviour in sparse-positive batches. In our study, where observer variability is common due to subtle anatomical differences and expert bias, it may be more reliable to discard a few lower-similarity positives than to select a fixed number of presumed confident pairs.

### 2.4.3. Log-of-sum loss (LogSum)

$$\mathcal{L}_i^{\text{logsum}} = -\log \left( \frac{\sum_{p \in P(i)} \exp(\mathbf{z}_i^\top \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i^\top \mathbf{z}_a / \tau)} \right) \tag{3}$$

This formulation aggregates the contributions of all positive pairs before applying the logarithm. Unlike the standard SupCon loss, which averages individual log-probabilities, the LogSum variant prioritizes the overall strength of alignment with the positive set. Diluting the false positives amongst the true positives may improve robustness to noisy or weakly aligned positives by reducing the impact of false pairs, which correspond to mislabelled or ambiguous samples.

## 2.5. Synthesising annotation variability

Because the training set was annotated by a single expert, it does not capture the variability observed in the triple-annotated test set. To assess how the model would perform if the training labels reflected the style of a different annotator, or the broader variability across a pool of experts, we simulated inter-expert variation in the training data, using patterns of disagreement observed among the three annotators in the test set. Crucially, this procedure does not use test-set statistics and therefore does not constitute data leakage. Instead, we exploit only the qualitative observation that expert disagreements occur predominantly within anatomically related families (e.g., apical variants among themselves) rather than across unrelated families (e.g., apical vs. PLAX).

**Table 3**
Defined view groupings used for synthetic annotation variability simulation.

| Group | Included View Classes |
|---|---|
| Apical | a2ch-full, a2ch-la, a2ch-lv, apex,<br>a3ch-full, a3ch-la, a3ch-lv, a3ch-outflow,<br>a4ch-full, a4ch-la, a4ch-lv, a4ch-ias, a4ch-ra, a4ch-rv,<br>a5ch-full, a5ch-outflow |
| PLAX | plax-full-out, plax-full-lv, plax-full-la, plax-full-rv-ao,<br>plax-full-mv, plax-valves-av, plax-valves-mv, plax-tv |
| PSAX | psax-all, psax-tv, psax-av, psax-pv,<br>psax-lv-base, psax-lv-mid, psax-lv-apex |
| M-mode | mmode-a4ch-rv, mmode-ivc, mmode-plax-mitral,<br>mmode-plax-av, mmode-plax-lv |
| Doppler | doppler-ao, doppler-av, doppler-mv, doppler-pv,<br>doppler-tv, doppler-tissue-lateral, doppler-tissue-rv, doppler-tissue-septal |

We defined clinically meaningful groupings of related views based on anatomical proximity and common misclassification patterns (Table 3). Examples include the apical group (a2ch-full, a2ch-lv, a2ch-la) and the parasternal long-axis (PLAX) group (plax-full-lv, plax-valves-av, plax-full-out), among others. Views within a group are closely related and more likely to be interpreted differently by different experts.

To simulate this variability, we randomly selected a proportion of samples (from the training split) within each class and reassigned their labels to a different class within the same group, excluding the original label. For example, a sample labelled as a4ch-full could be reassigned to a4ch-lv or a4ch-la, reflecting plausible alternative interpretations by another expert.

This label-variation process was applied (on the training split) at multiple levels, 0%, 10%, 20%, 30%, 40%, and 50%, with the proportion calculated per class to preserve the overall dataset distribution. This procedure does not assume that any label is incorrect; rather, it reproduces realistic patterns of expert disagreement to study the model's robustness to annotation variability. Here, 0% variation corresponds to the original, unaltered training labels, while, for example, a 10% variation level indicates that the original training label was synthetically altered for 10% of samples within each class.

## 2.6. Clustering-based analysis of learned representations

After synthesising inter-expert variability as described in Section 2.5, we conduct contrastive pretraining using the full training split, which includes the specified proportion of samples that have their labels synthetically altered while the remaining samples retain their original annotations. For example, at a 20% variation level, 80% of the training data remains unaltered and 20% is modified. We subsequently examine whether these samples are still attracted to their underlying semantic clusters despite having their labels intentionally modified.

To assess the structure and robustness of the learned feature space under simulated inter-expert variability, we perform an unsupervised clustering analysis on training embeddings obtained from the pretrained encoder. The aim is to evaluate whether semantically similar samples remain grouped together despite the presence of labels that have been intentionally varied to mimic realistic patterns of expert disagreement (as detailed in Section 2.5).

Given a training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$, where $y_i$ may have been altered due to simulated inter-expert variability, we extract embeddings using an encoder $enc(\cdot)$ defined earlier, yielding representations $z_i = enc(x_i) \in \mathbb{R}^{2048}$. These embeddings are then L2-normalised and reduced in dimensionality using Principal Component Analysis to obtain $\tilde{z}_i \in \mathbb{R}^{200}$ prior to clustering.

We apply MiniBatch $k$-Means with $k = 47$ to the reduced embeddings $\{\tilde{z}_i\}$, assigning each sample to a cluster:

$$\hat{c}_i = \arg\min_j \|\tilde{z}_i - \mu_j\|_2^2,$$

where $\mu_j$ is the centroid of cluster $j$.

Within each cluster $C_j$, the dominant label $\hat{y}_j$ is determined as the most frequent class label among its members:

$$\hat{y}_j = \arg\max_{y\in\mathcal{Y}} \text{count}(y \in C_j).$$

A sample is flagged as a *label-cluster mismatch* if its label (possibly altered) $y_i$ does not match the cluster's dominant label $\hat{y}_{\hat{c}_i}$.

Since the original labels $y_i^{\text{orig}}$ are known (due to the controlled variation injection), we quantify the network's robustness using two metrics:

*2.6.0.1. Detection Rate (DR).* The proportion of training samples with simulated inter-expert variation that are correctly identified as label–cluster mismatches:

$$\text{DR} = \frac{|\hat{\mathcal{N}} \cap \mathcal{N}|}{|\mathcal{N}|},$$

where $\mathcal{N} = \{i : y_i \neq y_i^{\text{orig}}\}$ is the set of altered samples, and $\hat{\mathcal{N}} = \{i : y_i \neq \hat{y}_{\hat{c}_i}\}$ is the set of all label-cluster mismatches identified through clustering.

*2.6.0.2. Label Recovery Precision (LRP).* Among the samples flagged as mismatches, the proportion for which the cluster's dominant label matches the original (pre-variation) class:

$$\text{LRP} = \frac{|C|}{|\mathcal{N}|}, \quad \text{where } C = \{i \in \hat{\mathcal{N}} \cap \mathcal{N} : \hat{y}_{\hat{c}_i} = y_i^{\text{orig}}\}.$$

DR quantifies the proportion of label–cluster mismatches originating from simulated inter-expert variation, while LRP measures the proportion of these that align with the original (pre-variation) class cluster. This clustering-based analysis provides insight into the semantic consistency of the learned feature space: a high DR indicates the model can separate samples with altered labels from the rest of the data, while a high LRP indicates the model can preserve or recover the underlying class semantics.

It is important to note that, in real-world multi-expert settings, such mismatches can arise from both natural variability in interpretation and genuine annotation errors. Robustness to either source of discrepancy is valuable, as it enables the model to maintain stable and discriminative representations regardless of the origin of the inconsistency.

To the best of our knowledge, the evaluation metrics introduced in this work, DR and LRP, are novel contributions to the analysis of representation learning under annotation variability. Unlike prior approaches that detect problematic labels using loss dynamics, agreement heuristics, or centroid distances, our method applies unsupervised clustering to learned embeddings and defines two explicit, quantitative metrics. These metrics offer a principled and interpretable framework for assessing both the semantic integrity and robustness of the representation space, and have not previously been reported in the context of contrastive learning or medical image analysis.

### 2.7. Implementation details

All images were preprocessed by resizing to 224 × 224x3 pixels and normalising pixel intensities to the [0,1] range to ensure consistency across inputs. For a fair evaluation, the data splits (train, val, test) used in all experiments remain fixed. In addition, both the selected training images and their corresponding simulated inter-expert variation labels were kept identical across all variation levels to ensure consistent evaluation across methods. All training experiments were conducted on an NVIDIA RTX 4090 GPU (24GB VRAM).

All models are initialised with ImageNet-pretrained weights prior to contrastive pretraining. During pretraining, the Adam optimiser is used with cosine weight decay, incorporating a warm-up phase to a peak learning rate of 1e-3, followed by cosine decay with a final decay rate of 1e-5. A batch size of 64 is used for all contrastive pretraining runs, constrained by available GPU memory. For SCL, we use the validation set with early stopping. We use a large shuffle buffer to improve class mixing across batches, which helps distribute rarer classes across training steps.

For downstream evaluation, the projection head is discarded, and a 47-class softmax classification layer is appended to the encoder. Fine-tuning is performed end-to-end using the Adam optimiser with a learning rate of 1e-4, employing sparse categorical cross-entropy loss and early stopping criteria. Fine-tuning also uses a batch size of 64. This framework is implemented in Tensorflow (v2.16).

## 3. Results and discussion

To establish a baseline, we first assessed several widely used CNN and transformer-based architectures trained directly on the downstream task without contrastive pretraining. As shown in Table 4, each model was initialised with ImageNet weights and trained to classify 47 fine-grained echo views. These results illustrate the performance ceiling when models are trained solely with cross-entropy loss, where each image is treated independently, without leveraging relational structure through contrastive representation learning.

The results in Table 4 are consistent with prior studies that attempted a broader range of echocardiographic views for classification (Zhang et al., 2018; Gearhart et al., 2022). Models trained solely with cross-entropy loss exhibit similar accuracy limits when classifying a large number of views, highlighting the challenges of fine-grained classification and the need for more robust methods to address subtle distinctions and inter-class ambiguities. Given their strong accuracy and comparatively light computational footprints under our GPU constraints, we adopted Xception, ConvNeXt-T, SwinV2-T, and EfficientNetV2 as backbones for subsequent experiments, enabling us to isolate the effect of the learning objectives while ensuring fair and consistent comparisons across architectures. Given our limited GPU memory, these models enables larger contrastive batches; heavier backbones would force smaller batch sizes ($\leq 16$), thereby reducing the number of positives and negatives per anchor and weakening the supervised contrastive signal. With 47 views, larger batches increase within-batch class diversity, strengthening the positive–negative contrasts that drive contrastive learning.

Before turning to the contrastive learning results, Section 3.1 serves as an important precursor: it quantifies inter-observer variability on the triple-annotated test set, establishing realistic clinical bounds and revealing anatomy-consistent disagreement patterns that frame the remainder of the analysis.

The performance of the proposed framework is established jointly in Section 3.2 *and* Section 3.3. In Section 3.2, supervised contrastive pretraining (especially using *LogSum*) with downstream fine-tuning sustains higher accuracy than non-contrastive baselines under increasing simulated annotator variation. Complementarily, Section 3.3 assesses the *representation quality* of the pretrained encoders before fine-tuning: clustering-based DR/LRP metrics and t-SNE visualisations indicate that *LogSum* preserves semantic structure and resists label shifts by aligning altered samples with their original class clusters. This representation-space analysis is crucial because differences between contrastive objectives can be attenuated after fine-tuning; examining the embeddings (before fine-tuning) directly makes these method-level distinctions clearer.

Subsequent sections broaden the evaluation scope. Section 3.4 examines performance across expert–agreement subsets (CS, UA, and E1) to disentangle model error from label uncertainty. Section 3.5 studies how accuracy varies with label resolution by regrouping predictions from 47 to 20 and 7 classes. Finally, Section 3.6 assesses cross-dataset generalisation on TMED-2, including zero-shot transfer and fine-tuning results.

**Table 4**
Baseline performance of widely used neural network architectures on the test set for 47-class echocardiographic view classification. These results provide reference metrics prior to applying contrastive representation learning methods. Results are reported as $\mu \pm \sigma$ across two independent runs.

| Model (Params) | Accuracy | F1 Score | Precision | Recall |
|---|---|---|---|---|
| ResNet50 (25.6M) | 85.53 ± 0.26 | 85.24 ± 0.24 | 86.36 ± 0.10 | 85.46 ± 0.19 |
| ResNet101 (44.7M) | 86.74 ± 0.18 | 86.17 ± 0.42 | 87.02 ± 0.36 | 86.13 ± 0.11 |
| DenseNet121 (8.1M) | 87.81 ± 0.52 | 87.37 ± 0.38 | 87.78 ± 0.92 | 87.72 ± 0.07 |
| SwinTransformerV2-T (29.3M) | 89.14 ± 0.12 | 88.81 ± 0.00 | 88.90 ± 0.06 | 89.28 ± 0.11 |
| ConvNeXt-B (89M) | 89.50 ± 0.26 | 88.99 ± 0.02 | 89.63 ± 0.63 | 88.96 ± 0.27 |
| ConvNeXt-T (29.5M) | 89.60 ± 0.45 | 89.32 ± 0.47 | 89.45 ± 0.53 | 89.75 ± 0.46 |
| ViT-S (21.6M) | 89.64 ± 0.50 | 89.59 ± 0.46 | 89.67 ± 0.66 | 89.94 ± 0.24 |
| ViT-B (85.7M) | 89.65 ± 0.01 | 89.95 ± 0.12 | 89.57 ± 0.15 | 89.42 ± 0.03 |
| EfficientNetV2 (21.6M) | 89.72 ± 0.88 | 89.77 ± 1.26 | 89.29 ± 1.04 | 89.26 ± 1.25 |
| Xception (22.9M) | 90.55 ± 0.06 | **90.39 ± 0.01** | **90.74 ± 0.05** | **90.65 ± 0.36** |
| SwinTransformerV2-B (87.9M) | **90.64 ± 0.07** | 90.32 ± 0.01 | 90.74 ± 0.06 | 90.33 ± 0.03 |

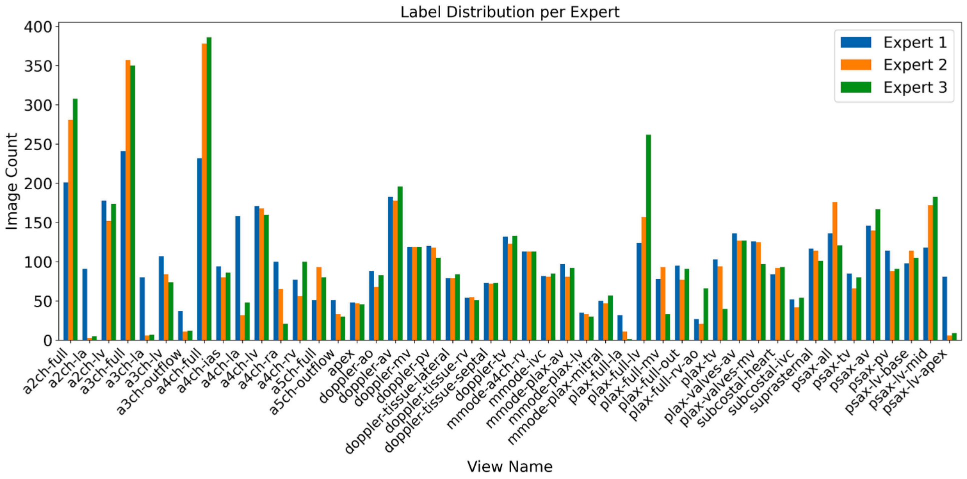


**Fig. 2.** Distribution of annotations across 47 echocardiographic view classes in the test subset for three independent experts. Each bar shows the number of images assigned to a given view by each expert. The plot reveals differences in labelling practices and interpretation.

### 3.1. Inter-observer analysis

Of the 5447 multi-expert annotated samples in the test set, 3327 received unanimous agreement across all three experts, 553 exhibited complete disagreement, and the remaining 1567 showed partial agreement, with two experts assigning the same label. This corresponds to a lower bound of approximately 10.2% full inter-observer variability, rising to 28.8% when partial disagreements are included. These figures provide a realistic threshold for annotation variability inherent in clinical practice, against which model robustness can be compared. While we synthesise variability levels of up to 50% in this study to stress-test model behaviour, the 10–30% range is relevant in practice, as it closely reflects the level of disagreement observed among human experts.

Fig. 2 shows the distribution of annotations across 47 echo view classes for all three independent experts. Common views such as *a2ch-full*, *a3ch-full*, and *a4ch-full* have relatively high coverage across experts, though some variation in frequency remains. Certain views show strong consistency in both presence and count; for example, *mmode-a4ch-rv* is annotated 113 times by all three experts, while others, such as *doppler-tissue-rv*, appear infrequently but consistently. Notably, substantial disagreement is observed for finer-grained classes such as *a3ch-la*, which is labelled 80 times by Expert-1 but fewer than 10 times by the others. These patterns reflect variations in labelling practices, class familiarity, and subjective interpretation.

Cohen's Kappa scores, presented in Table 5, provide a quantitative measure of inter-observer agreement using a fixed subset of 4882 test images for which all three expert annotations were available. The agreement between Expert-2 and Expert-3 ($\kappa = 0.817$) is higher than their individual agreement with Expert-1 ($\kappa = 0.764$ and $\kappa = 0.716$, respectively), suggesting a closer alignment in annotation style between the former pair.

**Table 5**
Pairwise Cohen's Kappa ($\kappa$) between the three experts. Values indicate moderate-to-substantial agreement.

| Comparison | Kappa |
|---|---|
| Expert-1 vs Expert-2 | 0.76 |
| Expert-1 vs Expert-3 | 0.72 |
| Expert-2 vs Expert-3 | 0.82 |

These results demonstrate that even among experienced clinicians, a natural level of variability exists when interpreting fine-grained echocardiographic views. Such inherent variability establishes an important baseline that any automated system must be able to accommodate in order to achieve robust and clinically reliable performance.

Fig. 3 presents the pairwise confusion matrix between Expert-2 and Expert-3, with additional matrices for the other expert pairs (Expert-1 vs. Expert-2 and Expert-1 vs. Expert-3) provided in Appendix E. The matrix shows a strong concentration of counts along the diagonal, indicating high overall agreement between the two experts, particularly for common or distinctive views such as *a2ch-full*, *a5ch-full*, *mmode-a4ch-rv*, *mmode-plax-lv*, and *subcostal-ivc*.

Off-diagonal entries reveal systematic areas of disagreement, especially within related subclasses such as the apical views (*a2ch-full* vs.

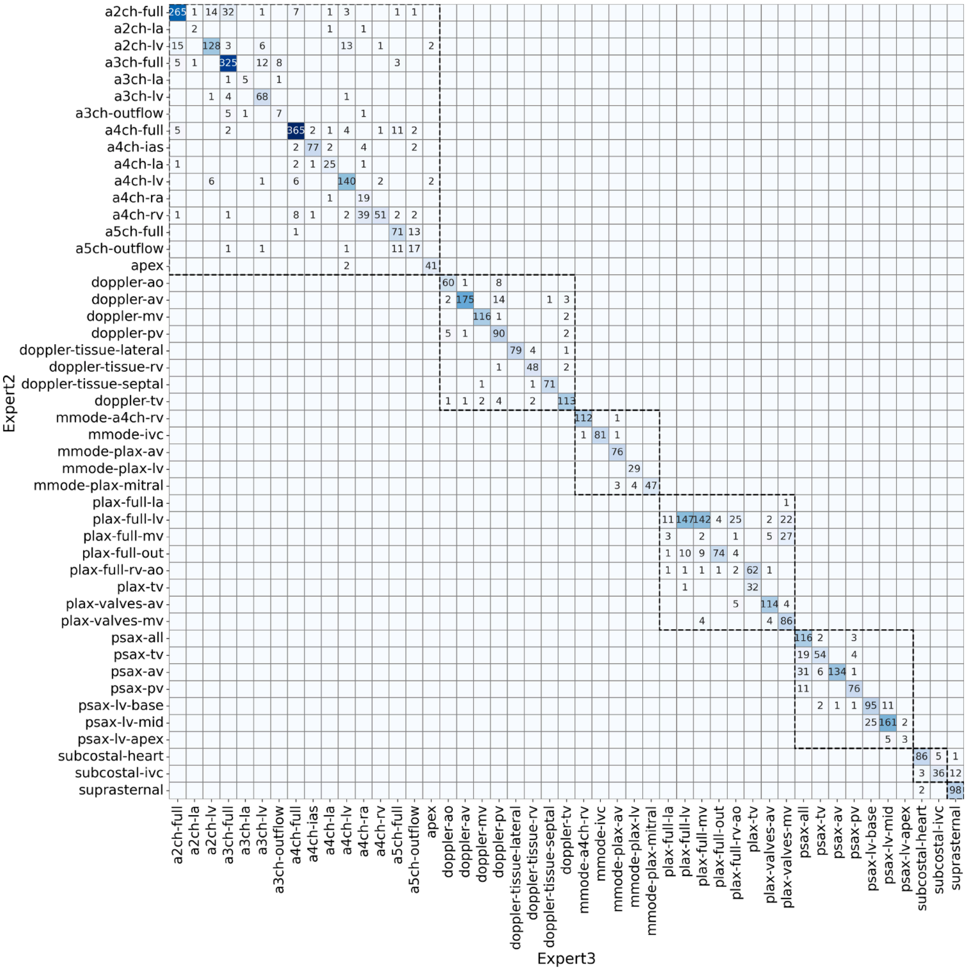


**Fig. 3.** Confusion matrix between Expert-2 and Expert-3. Strong diagonal dominance reflects high agreement, while off-diagonal clusters highlight systematic ambiguity within related subclasses.

*a2ch-la*), parasternal long-axis variants (*plax-full-lv*, *plax-full-mv*, *plax-full-out*), and parasternal short-axis categories (*psax-av*, *psax-mv*, *psax-lv-base*). These disagreements are not random but clustered within anatomically related groups, reflecting structured ambiguity where small changes in probe angle or acquisition focus can yield different but reasonable labels. This pattern emphasises the inherent variability of expert interpretation in fine-grained echocardiographic view classification and highlights the importance of designing models robust to such structured disagreement.

### 3.2. Downstream evaluation

To evaluate the efficacy of contrastive pretraining under varying levels of simulated inter-expert variation, we measured downstream classification accuracy across models fine-tuned on training sets altered at different variation levels (0–50% as described in Section 2.5). For each method, we report the best test performance at each variation level, selected from exhaustive sweeps over $\tau$, $k$, and $\upsilon$ (Appendix B). This procedure mitigates the influence of hyperparameter sensitivity and stress-tests robustness; although Table 6 reports only a single best value per method and level, the full accuracy landscape is provided in Appendix B.

As shown in Table 6, models trained from random initialisation (RandInit) degrade substantially under increasing simulated inter-expert variation across all architectures. For example, Xception declines from 87.98% at 0% variation to 68.01% at 50%, while ConvNeXt-T drops from 73.22% to 50.06% and Swin-T from 70.08% to 42.56%. This drop is expected, as the training labels are progressively altered away from Expert-1's annotations, increasing divergence in labelling style and potentially causing confusion for the model when learning consistent class boundaries. ImageNet initialisation consistently improves robustness, retaining higher accuracy at 50% variation (e.g., 82.73% for Xception, 78.47% for ConvNeXt-T, 76.71% for Swin-T, and 72.54% for EfficientNetV2), but still exhibits clear sensitivity as label variation increases.

In contrast, models pre-trained with supervised contrastive learning demonstrated substantially greater robustness across all architectures. The standard SupCon objective consistently outperformed both non-contrastive baselines across variation levels, reaching 94.05% at 0% variation for Xception and remaining at 87.67% at 50% variation. Similar trends were observed for the other backbones, with performance decreasing from 92.42% to 85.06% for ConvNeXt-T, from 92.14% to 85.35% for Swin-T, and from 93.81% to 87.23% for EfficientNetV2-S as variation increased from 0% to 50%. These results indicate that supervised contrastive pretraining improves resilience to label variation beyond what can be achieved by either random initialisation or ImageNet initialisation alone.

Other contrastive loss variants yielded additional performance gains, though the improvements were modest. At 0% variation, DropCon achieved the strongest results for Xception (94.25%) and EfficientNetV2-S (94.02%), while LogSum was marginally best for ConvNeXt-T (92.73%) and Swin-T (92.58%). As annotation variation increased, LogSum emerged as the strongest or co-leading objective across most settings. In particular, for ConvNeXt-T, LogSum remained the top-performing method across the entire variation range, decreasing from 92.73% at 0% variation to 86.66% at 50%. For Swin-T and

**Table 6**
Downstream classification accuracy on the test set across different levels of simulated inter-expert variation in the training data. RandInit refers to models trained from random initialisation without contrastive pretraining. ImageNet refers to models initialised with ImageNet-pretrained weights but without contrastive learning. The remaining methods use contrastive pretraining (EchoFine framework) and present results after downstream fine-tuning.

| Architecture | Method | Variation Level (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 |
| Xception | RandInit | 87.98 ± 0.26 | 81.20 ± 1.34 | 80.46 ± 0.16 | 75.17 ± 0.05 | 72.69 ± 1.86 | 68.01 ± 0.14 |
| | ImageNet | 90.55 ± 0.06 | 87.45 ± 0.16 | 88.30 ± 0.34 | 86.86 ± 0.01 | 83.77 ± 0.59 | 82.73 ± 0.10 |
| | SupCon | 94.05 ± 0.04 | 93.47 ± 0.23 | 92.64 ± 0.24 | 91.85 ± 0.28 | 90.02 ± 0.49 | 87.67 ± 0.38 |
| | DropCon | **94.25 ± 0.35** | 93.38 ± 0.01 | 92.79 ± 0.28 | 91.59 ± 0.24 | 90.15 ± 0.49 | 88.10 ± 0.06 |
| | LogSum | 94.16 ± 0.09 | **93.67 ± 0.20** | **93.28 ± 0.02** | **92.28 ± 0.06** | **90.63 ± 0.28** | **88.36 ± 0.31** |
| ConvNeXt-T | RandInit | 73.22 ± 0.28 | 69.22 ± 0.17 | 63.77 ± 1.50 | 62.06 ± 0.21 | 57.55 ± 0.84 | 50.06 ± 0.47 |
| | ImageNet | 89.60 ± 0.45 | 87.42 ± 0.15 | 85.84 ± 0.10 | 83.96 ± 0.03 | 81.46 ± 1.45 | 78.47 ± 0.40 |
| | SupCon | 92.42 ± 0.15 | 91.89 ± 0.26 | 90.52 ± 0.16 | 89.79 ± 0.20 | 88.21 ± 0.09 | 85.06 ± 0.17 |
| | DropCon | 92.69 ± 0.11 | 91.81 ± 0.13 | 90.66 ± 0.40 | 89.86 ± 0.01 | 88.87 ± 0.01 | 85.79 ± 0.31 |
| | LogSum | **92.73 ± 0.42** | **92.56 ± 0.08** | **91.95 ± 0.21** | **91.07 ± 0.32** | **89.57 ± 0.00** | **86.66 ± 0.39** |
| SwinTransV2-T | RandInit | 70.08 ± 0.04 | 61.13 ± 0.04 | 56.06 ± 0.38 | 51.73 ± 0.98 | 48.55 ± 0.34 | 42.56 ± 0.84 |
| | ImageNet | 89.14 ± 0.12 | 88.20 ± 0.50 | 85.44 ± 1.54 | 85.62 ± 0.12 | 83.26 ± 0.11 | 76.71 ± 0.05 |
| | SupCon | 92.14 ± 0.50 | 91.18 ± 0.23 | 90.72 ± 0.12 | 89.79 ± 0.02 | 88.09 ± 0.38 | 85.35 ± 0.40 |
| | DropCon | 92.50 ± 0.23 | 91.44 ± 0.09 | 90.41 ± 0.01 | 90.09 ± 0.39 | 88.25 ± 0.29 | **86.40 ± 0.29** |
| | LogSum | **92.58 ± 0.29** | **92.11 ± 0.09** | **91.79 ± 0.04** | **91.00 ± 0.22** | **89.27 ± 0.52** | 85.88 ± 0.11 |
| EfficientNetV2 | RandInit | 76.63 ± 0.09 | 77.30 ± 0.47 | 76.46 ± 0.46 | 69.29 ± 0.24 | 69.25 ± 1.36 | 69.25 ± 1.14 |
| | ImageNet | 89.72 ± 0.88 | 87.01 ± 0.69 | 84.41 ± 0.35 | 78.70 ± 1.80 | 78.80 ± 1.56 | 72.54 ± 1.85 |
| | SupCon | 93.81 ± 0.02 | 93.19 ± 0.15 | 92.70 ± 0.21 | 91.50 ± 0.51 | 89.54 ± 0.64 | 87.23 ± 0.30 |
| | DropCon | **94.02 ± 0.00** | 93.23 ± 0.23 | 92.62 ± 0.04 | 92.07 ± 0.22 | 90.26 ± 0.03 | **88.53 ± 0.04** |
| | LogSum | 93.62 ± 0.01 | **93.70 ± 0.09** | **92.90 ± 0.21** | **92.18 ± 0.29** | **90.78 ± 0.13** | 88.47 ± 0.20 |

EfficientNetV2-S, LogSum led across intermediate variation levels, while DropCon became slightly stronger at the highest variation level (50%), reaching 86.40% and 88.53%, respectively.

These results support the hypothesis that contrastive objectives not only facilitate stronger representation learning for fine-grained echocardiographic view classification, but also act as an implicit regulariser under annotation variability, partially mitigating the influence of noisy supervision while preserving discriminative structure in the feature space.

The performance trends across variation levels indicate that while all contrastive methods offer substantial gains under clean conditions (0%), their advantages are amplified when training labels deviate from the evaluation annotator's style. In particular, the LogSum variant shows greater resilience, likely due to its logarithmic attenuation of uncertain similarities, diluting false positives in a way that reduces sensitivity to weak or noisy positives. This highlights the relevance of contrastive representation learning for real-world medical data, where inter-expert variability is common, and demonstrates its role in improving robustness to annotator identity; a key requirement for deployment across institutions and operators.

We performed selection-aware statistical significance testing to assess whether the observed performance differences between contrastive objectives are statistically reliable. For each architecture and variation level, methods were compared using a paired, non-parametric permutation test on per-image correctness, with the best configuration for each method re-selected within each resample to match the model-selection protocol used in Table 6 (best $\tau$, $k$, and $\upsilon$ per variation level, averaged over multiple runs). Holm's correction was applied across the six variation levels per architecture. LogSum significantly outperformed SupCon and DropCon across most architecture–variation settings (Holm-adjusted permutation test, $p < 0.05$ in the majority of cases), whereas differences between SupCon and DropCon were rarely significant.

Our observed robustness of contrastive learning methods under increasing simulated inter-expert variation aligns with prior theoretical findings. Graf et al. (2021) demonstrated that supervised contrastive loss implicitly regularises model training, enabling robustness against noisy or randomly assigned labels. This property is particularly beneficial for echo view classification, where inter-observer variability can introduce structured differences in labelling. Khosla et al. (2020) further showed that the gradient dynamics of the supervised contrastive objective naturally emphasise harder positive and negative samples, enhancing discriminability. In the context of echocardiography, this can improve separation between visually similar classes, such as those within the apical view group. Together, these findings suggest that supervised contrastive learning promotes tighter intra-class clustering and stronger inter-class separation in the learned representation space Khosla et al. (2020), Graf et al. (2021).

As Xception achieved the strongest overall performance across variation levels in the previous analysis, the remainder of this section focuses on results obtained with the Xception backbone. To further assess robustness to annotator identity, we computed Cohen's $\kappa$ between model predictions and each of the three experts at increasing levels of simulated inter-expert variation (Table 7). The baseline model trained from random initialisation showed rapidly declining agreement with Expert-1 (from $\kappa = 0.86$ at 0% variation to $\kappa = 0.65$ at 50%) and even lower agreement with Experts 2 and 3 (dropping to $\kappa = 0.55$ and $\kappa = 0.53$, respectively). This reflects the model's reliance on the single training annotator's style and its limited ability to generalise when labels deviate.

By contrast, our best-performing model (EchoFine with an Xception backbone pretrained using the LogSum contrastive objective from Table 6) maintained much stronger agreement across all experts. Agreement with Expert-1 remained very high ($\kappa = 0.95$ at 0% and $\kappa = 0.88$ at 50%), while agreement with Expert-2 and Expert-3 stabilised around $\kappa = 0.76$ and $\kappa = 0.71$, respectively, even under high variation. Because the training data are annotated by a single expert, stronger alignment between the model and that expert is expected. The role of multi-expert annotations here is therefore not to remove this bias, but to contextualise model–expert agreement relative to expert–expert variability and to assess whether model behaviour remains within clinically realistic disagreement bounds.

These values are close to the observed inter-expert agreement ($\kappa = 0.76$ and $\kappa = 0.72$ with Expert-1; $\kappa = 0.82$ between Experts 2 and 3) from Table 5, suggesting that the model does not simply overfit to the labelling style of the primary annotator, but instead captures consensus-driven annotation boundaries.

Complementing Table 7, Table 8 reports the modified Williams Index (mWI), the ratio of mean model-expert to mean

**Table 7**
Cohen's $\kappa$ between model predictions and each expert across different levels of simulated inter-expert variation in the training data. RandInit refers to models trained from random initialisation. LogSum refers to our best supervised contrastive framework (EchoFine + LogSum).

| Model | Expert | Variation Levels (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 |
| RandInit | Expert-1 | 0.86 | 0.84 | 0.82 | 0.81 | 0.71 | 0.65 |
| | Expert-2 | 0.69 | 0.69 | 0.69 | 0.68 | 0.62 | 0.55 |
| | Expert-3 | 0.66 | 0.65 | 0.65 | 0.64 | 0.59 | 0.53 |
| LogSum | Expert-1 | **0.95** | **0.95** | **0.95** | **0.94** | **0.92** | **0.88** |
| | Expert-2 | **0.76** | **0.75** | **0.76** | **0.75** | **0.74** | **0.72** |
| | Expert-3 | **0.71** | **0.71** | **0.71** | **0.70** | **0.70** | **0.68** |

**Table 8**
Modified Williams Index (mWI) and mean Model–Expert Cohen's $\kappa$ across simulated inter-expert variation levels. The mean Expert–Expert $\kappa$ was 0.77 for both *RandInit* and *EchoFine* + *LogSum* across all variation levels.

| Model | Metric | Variation Levels (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 |
| RandInit | Mean Model–Expert $\kappa$ | 0.74 | 0.73 | 0.72 | 0.71 | 0.64 | 0.58 |
| | mWI | 0.96 | 0.95 | 0.94 | 0.93 | 0.84 | 0.75 |
| LogSum | Mean Model–Expert $\kappa$ | 0.81 | 0.80 | 0.80 | 0.80 | 0.79 | 0.76 |
| | mWI | **1.05** | **1.05** | **1.05** | **1.04** | **1.03** | 0.99 |

expert-expert Cohen's $\kappa$ computed at each variation level. In our setup, the modified Williams Index (mWI) is computed from Cohen's $\kappa$: let $\overline{\kappa}_{\text{M–E}} = \frac{1}{3}\left[\kappa(\hat{y}, y^{(1)}) + \kappa(\hat{y}, y^{(2)}) + \kappa(\hat{y}, y^{(3)})\right]$ and $\overline{\kappa}_{\text{E–E}} = \frac{1}{3}\left[\kappa(y^{(1)}, y^{(2)}) + \kappa(y^{(1)}, y^{(3)}) + \kappa(y^{(2)}, y^{(3)})\right]$, then $\text{mWI} = \overline{\kappa}_{\text{M–E}}/\overline{\kappa}_{\text{E–E}}$. Interpretation: $\text{mWI} \approx 1$ indicates the model agrees with experts about as much as experts agree among themselves; $\text{mWI} > 1$ indicates the model is more consistent with experts than experts are with each other; $\text{mWI} < 1$ indicates the model lags behind expert–expert agreement.

Under RandInit, the mean model-expert $\kappa$ declines from 0.74 (0%) to 0.58 (50%), and mWI falls from 0.96 to 0.75, indicating the model increasingly underperforms relative to the expert-expert agreement baseline as simulated annotator variability rises. In contrast, our framework (model with EchoFine + LogSum) maintains substantially higher agreement with experts (0.81→0.76 from 0% to 50%) and preserves mWI $\geq 1$ through 40% variation (1.05 at 0–20%, 1.04 at 30%, 1.03 at 40%), remaining near parity even at 50% (0.99). Given that expert-expert $\kappa$ is stable at $\approx 0.77$ across conditions, these results corroborate our earlier observations: the proposed contrastive framework maintains expert-level consistency over a wide range of simulated annotator variability, whereas training without contrastive pretraining rapidly falls below the expert-expert agreement boundary.

Confusion matrices of the best-performing model overall (EchoFine with LogSum trained at 0% variation level i.e, the original training labels) and representative prediction examples are presented in Appendices E and F, respectively.

### 3.3. Robustness to annotation variability

To explain the performance differences between contrastive learning objectives, we analyse the structure of the learned feature space produced during contrastive pretraining. This representation-level analysis provides empirical evidence of how different objectives organise embeddings under increasing annotation variability, beyond what is observable from downstream accuracy alone. We focus on the best-performing configurations identified in Table 6, using contrastively pretrained Xception backbones, and examine their embedding behaviour in detail.

We evaluated the consistency and robustness of the learned feature representations under varying levels of simulated inter-expert variation. Embeddings obtained from the contrastive pretraining phase (before fine-tuning) were clustered using the unsupervised procedure described in Section 2.6, and we analysed how well altered samples remained aligned with their original semantic clusters. Table 9 summarises detection and recovery outcomes across three methods as the proportion of simulated variation in the training set increased from 0% to 50%.

At 10% variation, the framework using the LogSum objective achieved the highest Detection Rate (97.3%), correctly identifying nearly all altered labels as cluster–label mismatches. It also achieved the highest Label Recovery Precision (87.1%), indicating that most altered samples were pulled toward their original class clusters: for example, at 10% variation, of the 7284 synthetically altered samples, 7090 were detected and 6336 were correctly recovered. This trend persisted across higher variation levels, with LogSum consistently outperforming other objectives in LRP and often in DR, suggesting superior ability to preserve semantic coherence in the feature space.

t-SNE visualisations in Fig. 4 and examples in Fig. 5 further support these findings. Even without simulated variation (Fig. 4(b)), we observed over 3000 cluster–label mismatches across methods (Table 9, 0% row), likely reflecting inherent intra-observer variability in the single-expert (Expert-1) training annotations, where subtle view distinctions may be inconsistently labelled (see Appendix D).

When 10% simulated variation was introduced (Fig. 4(c)), the global cluster structure remained largely intact. Notably, many altered points gravitated toward their correct semantic cluster rather than the altered label, indicating that the model favoured the intrinsic structure of the data over the supervision signal. This behaviour highlights the potential of contrastive pretraining to detect and resist both simulated and naturally occurring annotation variability.

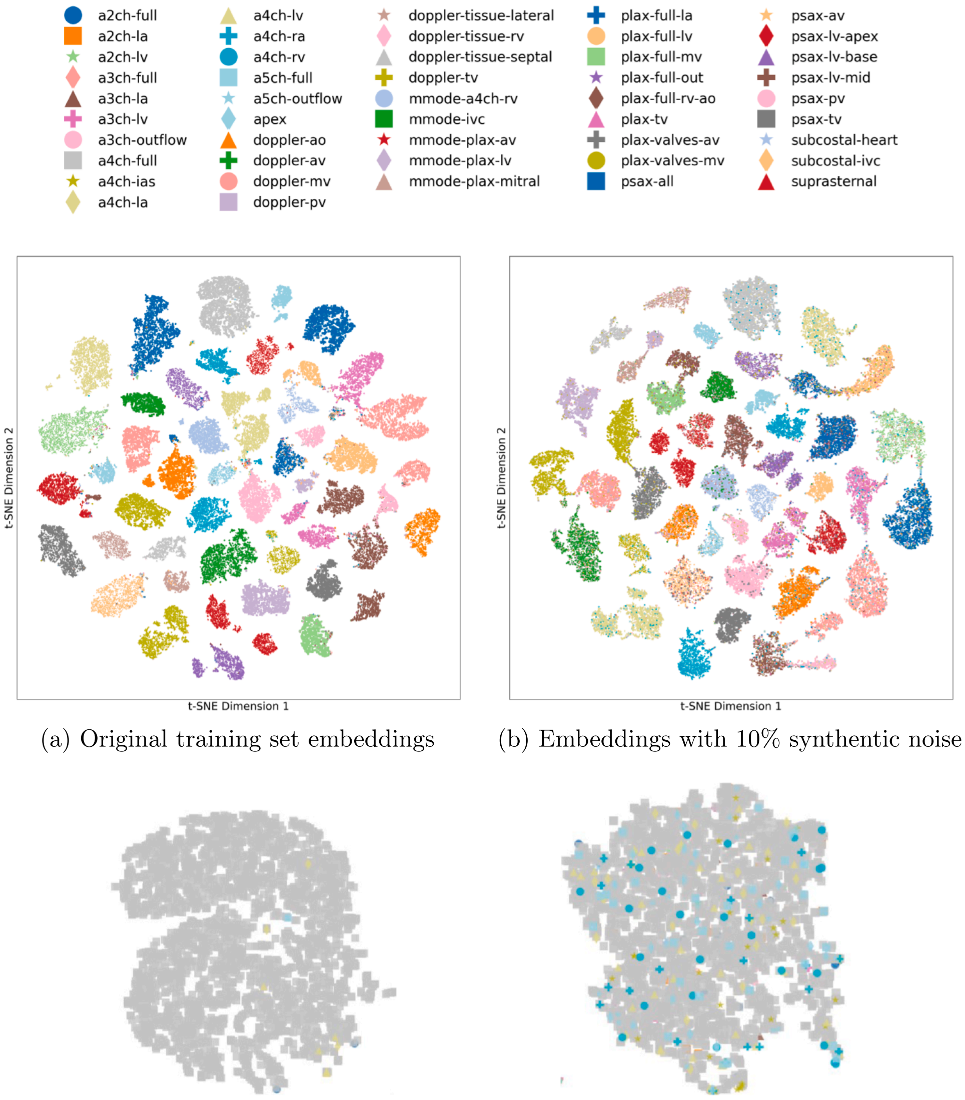
For example, Fig. 4(e) shows that samples originally belonging to the a4ch-full view, despite having their labels synthetically altered, continue to cluster around the primary a4ch-full embedding region. Rather than dispersing toward clusters corresponding to their altered labels, these samples remain anchored to their underlying semantic group. This behaviour indicates that the learned representations are driven more strongly by intrinsic anatomical structure than by noisy supervision.

### 3.4. Performance across expert agreement subsets

To further assess the robustness of the learned representations, we evaluated model performance under increasing levels of simulated annotation variability across multiple expert-defined subsets (Table 10). While accuracy naturally declines as variability increases, the extent of this decline differs by subset. The *Consensus Set* (CS), consisting of samples with unanimous expert agreement, shows the highest and most stable performance, maintaining 92.3% accuracy even at 50% variability. This suggests that strong inter-expert agreement correlates with high-confidence, semantically coherent samples that are more resilient to annotation uncertainty.

The *Union-Agreement* (UA) metric also consistently outperforms the standard Primary Annotator set ($E_1$), indicating that incorporating diverse expert perspectives can enhance robustness. These findings suggest that expert agreement serves as a useful proxy for annotation confidence and helps identify samples where the model is most reliable, highlighting the importance of structured evaluation subsets when assessing performance under conditions of label variability. Note that higher performance on CS and UA reflects greater label confidence and evaluation tolerance, and is method-agnostic rather than model-specific.

Overall, these results reinforce the inherent difficulty of fine-grained classification in the presence of real-world annotation differences, while demonstrating that the model learns stable and clinically meaningful representations that generalise across annotator styles and maintain reliability even when confronted with uncertain or inconsistent labels.



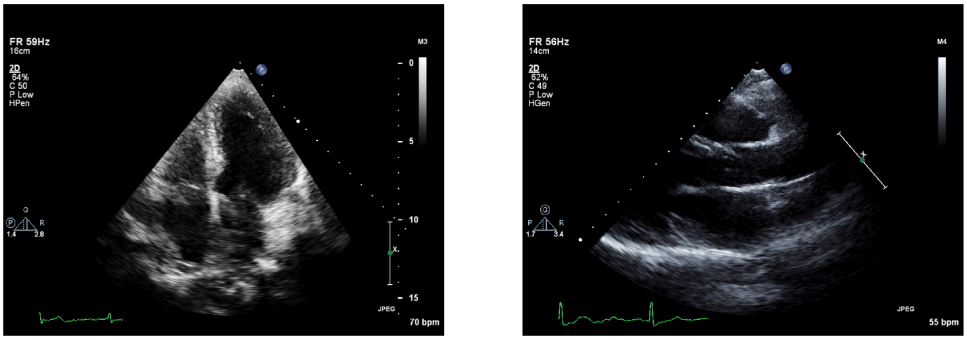
(a) Original training set embeddings

(b) Embeddings with 10% synthentic noise

(c) Zoomed cluster of (a)

(d) Zoomed cluster of (b)

**Fig. 4.** t-SNE visualisation of the learned feature space from contrastive pretraining. (a) Embeddings obtained after pretraining with the original training set, showing well-separated clusters. (b) Embeddings after pretraining with 10% simulated inter-expert variation in the training labels. Much of the original structure is retained, and many altered samples are pulled toward clusters matching their original semantic class, indicating that the model's representation space can resist annotation variability and preserve semantic consistency.



(a) Apical Mismatch

(b) Parasternal Mismatch

**Fig. 5.** Visual examples of label-cluster mismatches identified through unsupervised clustering. (**a**) An image synthetically mislabelled as *a4ch-lv* is correctly grouped with the *a4ch-full* cluster, illustrating semantic resilience despite altered labels. (**b**) A view labelled as *plax-full-la* is assigned to the *plax-full-out* cluster, demonstrating the model's ability to preserve structure among closely related parasternal views.

**Table 9**
Clustering-based detection and recovery statistics across contrastive objectives. $\mathcal{N}$ denotes the set of samples with simulated inter-expert variation, $\hat{\mathcal{N}}$ the label-cluster mismatches, and $\mathcal{C}$ the correctly recoverable samples. DR quantifies the proportion mismatches originating from simulated variation, while LRP indicates the proportion of these recoverable samples that align with their original (pre-variation) class cluster.

| **Synthesised Variation % (#)** | **Method** | $\lvert\hat{\mathcal{N}}\rvert$ | $\lvert\hat{\mathcal{N}} \cap \mathcal{N}\rvert$ | $\lvert\mathcal{C}\rvert$ | **DR (%)** | **LRP (%)** |
|---|---|---|---|---|---|---|
| 0 ($\lvert\mathcal{N}\rvert = 0$) | SupCon | 3632 | – | – | – | – |
| | DropCon | 3151 | – | – | – | – |
| | LogSum | 3026 | – | – | – | – |
| 10 ($\lvert\mathcal{N}\rvert = 7284$) | SupCon | 10,638 | 7036 | 6226 | 96.6 | 85.5 |
| | DropCon | 11,840 | 6970 | 6072 | 95.7 | 83.4 |
| | LogSum | 10,019 | 7090 | **6336** | **97.3** | **87.1** |
| 20 ($\lvert\mathcal{N}\rvert = 14572$) | SupCon | 21,040 | 14,231 | 12,279 | 97.7 | 84.3 |
| | DropCon | 17,939 | 13,964 | 12,264 | 95.8 | 84.2 |
| | LogSum | 18,089 | 14,197 | **12321** | **97.4** | **84.6** |
| 30 ($\lvert\mathcal{N}\rvert = 21859$) | SupCon | 22,161 | 20,225 | 17,379 | 92.5 | 79.5 |
| | DropCon | 25,554 | 21,210 | **18477** | **97.0** | **84.5** |
| | LogSum | 24,225 | 21,199 | 17,976 | 97.0 | 82.2 |
| 40 ($\lvert\mathcal{N}\rvert = 29147$) | SupCon | 32,153 | 28,105 | 24,006 | 96.4 | 82.4 |
| | DropCon | 33,052 | 28,191 | 24,006 | 96.7 | 82.4 |
| | LogSum | 32,105 | 28,287 | **24249** | **97.0** | **83.2** |
| 50 ($\lvert\mathcal{N}\rvert = 36434$) | SupCon | 40,966 | 35,054 | 28,264 | 96.2 | 77.6 |
| | DropCon | 43,484 | 34,920 | 26,146 | 95.8 | 71.8 |
| | LogSum | 40,274 | 35,324 | **28674** | **96.9** | **78.7** |

**Table 10**
Accuracy of the EchoFine (Xception, LogSum) method across increasing levels of simulated annotation variability, evaluated on three subsets: the primary annotator set ($E_1$), the consensus set (CS), and the union-agreement evaluation (UA).

| Subset | Variation Levels (%) | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 30 | 40 | 50 |
| Primary Annotator ($E_1$) | 94.2 | 93.7 | 93.3 | 92.3 | 90.6 | 88.4 |
| Consensus Set (CS) | 97.7 | 97.4 | 97.1 | 96.3 | 94.9 | 92.3 |
| Union-Agreement (UA) | 95.7 | 95.6 | 95.2 | 94.0 | 92.8 | 90.4 |

**Table 11**
Inference-only classification accuracy evaluated at three levels of view granularity on the test set. Coarser accuracies were computed by grouping model predictions and ground truth labels from 47 fine-grained views into 20 intermediate and 7 high-level view categories. Accuracy improves with coarser groupings, reflecting increased tolerance to intra-group view variation.

| Granularity Level | Number of Classes | Accuracy |
|---|---|---|
| Fine-Grained | 47 | 94.2 |
| Intermediate | 20 | 97.0 |
| Coarse | 7 | 98.8 |

### 3.5. Granularity of echo views

Using the best-performing model after downstream fine-tuning on the original training set, EchoFine (Xception, LogSum), we conducted inference-only experiments to evaluate how classification performance varies across different levels of view granularity.

Classification performance improved progressively with coarser levels of view granularity, as shown in Table 11. At the fine-grained level involving 47 distinct echocardiographic views, the model achieved an accuracy of 94.2% (from Section 3.2). When these views were grouped into 20 intermediate categories based on anatomical and functional similarity, accuracy increased to 97.0%. Further consolidation into 7 broad view classes yielded 98.8% accuracy. The granularity analysis re-maps the same fine-grained predictions into 20 and 7 class categories to illustrate how label resolution affects accuracy and to contextualise task difficulty.

This trend reflects the model's greater tolerance to intra-class variation at coarser levels, where semantically related views are aggregated, thereby reducing the penalty for visually subtle distinctions. These results highlight the inherent challenge of fine-grained classification and demonstrate that performance is sensitive to the resolution at which view labels are defined. Full details of the hierarchical granularity levels are provided in Appendix C.

### 3.6. Cross-dataset generalisation

Given the scarcity of publicly available datasets for echocardiographic view classification, and the complete absence of datasets annotated at a fine-grained level, it is not currently possible to directly benchmark our approach against comparable methods for 47-class classification. Instead, we assess generalisation using the TMED-2 dataset (Huang et al., 2022b), one of the few publicly accessible resources for echo view classification. TMED-2 contains only five classes (A2C, A4C, PLAX, PSAX, other), and most prior work evaluates performance on the four defined views excluding the “other” category. Results are compared against recent state-of-the-art methods, as presented in Table 12.

Despite being trained on a distinct dataset with 47 fine-grained labels, models pretrained using the EchoFine framework achieve strong zero-shot performance on TMED-2. In particular, using the LogSum objective with an Xception backbone attains 90.0% top-1 accuracy in a zero-shot setting, demonstrating that contrastive representations learned from fine-grained supervision generalise well to new, coarse-grained domains. When fine-tuned on TMED-2, EchoFine models pretrained contrastively on TTE47 consistently surpass prior state-of-the-art results (without any TMED-2–specific pretraining), achieving mean accuracies of approximately 98% across objectives and folds, indicating that EchoFine provides a strong and transferable initialisation for downstream tasks. For the setting where both pretraining and fine-tuning are

**Table 12**
Reported accuracy on the TMED-2 dataset for echo view classification using a consistent subset of four anatomical views (PLAX, PSAX, A4CH, A2CH).

| Study / Setting | Objective | Accuracy (%) |
|---|---|---|
| Fix-A-Step (Huang et al., 2023) | – | 94.1 |
| SPEMix (Ma et al., 2024) | – | 97.3 |
| Benchmarking SSL (Huang et al., 2024a) | – | 95.1 |
| InterLUDE (Huang et al., 2024b) | – | 96.8 |
| Reported ResNet50 (Kim et al., 2025) | – | 83.7 |
| EchoFM (Kim et al., 2025) | – | 94.2 |
| EchoFine (pretrained and fine-tuned on TTE47; zero-shot inference on TMED2) | LogSum | **90.0** ± 0.5 |
| | SupCon | 89.1 ± 0.4 |
| | DropCon | 88.0 ± 0.6 |
| EchoFine (pretrained on TTE47; fine-tuned on TMED2) | LogSum | 98.1 ± 0.1 |
| | SupCon | 97.6 ± 0.2 |
| | DropCon | **98.2** ± 0.1 |
| EchoFine (pretrained and fine-tuned on TMED2) | LogSum | 97.9 ± 0.3 |
| | SupCon | 97.7 ± 0.2 |
| | DropCon | **98.0** ± 0.1 |

performed on TMED-2, performance remains similarly high, with mean accuracy approaching 98%.

When pretrained contrastively on the 47 fine-grained views of the TTE47 dataset and then fine-tuned on TMED-2, the observed gains arise because the richer, more granular supervision forces the encoder to learn semantically aligned, anatomy-aware features that transfer across datasets. Fine-grained contrastive pretraining with the EchoFine framework builds invariances to acquisition differences while sharpening boundaries between closely related views, yielding more generalisable initial weights for downstream tasks. These findings highlight both the robustness and transferability of representations learned through supervised contrastive pretraining on fine-grained echocardiographic views.

## 4. Conclusion

This study presents a comprehensive investigation of echocardiographic view classification at full clinical granularity, spanning 47 clinically meaningful views acquired under real-world conditions. We introduce *TTE47*, the first publicly available benchmark pairing fine-grained view categories with independent multi-expert annotations, enabling rigorous quantification of inter-observer variability and establishing a foundation for reproducible, clinically relevant evaluation.

To address the dual challenges of subtle inter-class distinctions and structured label noise from observer variability, we propose a supervised contrastive framework. This approach improves inter-class separability and robustness to both synthetic and naturally occurring annotation variability, outperforming supervised contrastive and cross-entropy baselines. We further introduce two clustering-based metrics: Detection Rate and Label Recovery Precision, which provide principled measures of semantic coherence and resilience in the learned feature space, extending evaluation beyond standard accuracy metrics.

Our method defines an initial performance benchmark on *TTE47* and achieves state-of-the-art performance on TMED-2, surpassing prior results without dataset-specific pretraining. Using a model pretrained on TTE47 and fine-tuned on TMED-2, we demonstrate that fine-grained contrastive representations transfer effectively across domains and to coarser view sets. Notably, strong DR and LRP scores under 10–30% simulated inter-expert variation fall within the range of observed human disagreement, suggesting robustness comparable to expert-level variability.

Inter-observer variation is an inherent feature of clinical practice: some differences reflect genuine subjectivity, while others arise from inconsistency or error. Our clustering analysis shows that the learned feature space aligns more closely with underlying anatomical structure than with any single annotator's style, enabling the model to resist label shifts. This prevents overfitting to one expert and ensures semantically meaningful representations that generalise across annotation styles and institutions.

Although accuracy naturally declines as training labels deviate from the evaluation style, our framework preserves stable and discriminative class boundaries and maintains high agreement across experts. By leveraging representation consistency as an unsupervised signal, it actively resists variability and provides a foundation for future noise-aware fine-tuning strategies, such as using corrected cluster assignments as pseudo-labels.

Finally, while increased view granularity yields richer diagnostic information, it also amplifies annotation complexity and observer variability. Our findings show that contrastive pretraining preserves semantic structure even under these conditions, enabling reliable classification across 47 fine-grained views. By demonstrating resilience to annotation variability at this scale, the proposed framework provides a pathway toward clinically relevant, high-resolution view classification that better reflects the realities of comprehensive echocardiographic practice.

## CRediT authorship contribution statement

**Preshen Naidoo:** Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Conceptualization; **Patricia Fernandes:** Writing – review & editing; **Nasim Dadashi Serej:** Writing – review & editing, Supervision; **Catherine C Stowell:** Data curation; **Charlotte H Manisty:** Writing – review & editing; **Darrel P Francis:** Data curation, Writing - review & editing; **Massoud Zolgharni:** Data curation, Project administration, Supervision, Writing - review & editing.

## Data availability

The TTE47 dataset is publicly available at thrive-centre.com/datasets/TTE47. Pre-trained and fine-tuned models are provided in the EchoForge model library. The source code for training and evaluation is available on the github EchoFine repository.

## Funding

This work was supported in part by the British Heart Foundation, United Kingdom (grant no. RG/F/22/110059). P. Naidoo is supported by the Vice Chancellor's Scholarship at the University of West London.

## Declaration of competing interest

Charlotte H Manisty reports a relationship with MyCardium AI that includes: board membership and equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research and open-access release has been conducted under: Imperial College London and University of West London, United Kingdom [IRAS: 279328, REC:20/SC/0386].

## Appendix A. Echocardiography view definitions

The clinical uses listed for each echocardiographic view are grounded in established echocardiography guidelines and standard textbooks, which define diagnostic measurements primarily at the level of view families rather than for every individual acquisition variant (Mitchell et al., 2018; Otto, 2018). In addition, the selection of views and associated clinical use cases was informed by collaboration with cardiologists at University College London and Imperial College London, ensuring consistency with real-world clinical practice.

**Table A.13**
Apical echocardiographic views with representative images and descriptions.

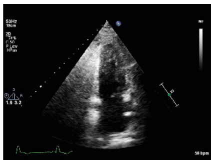
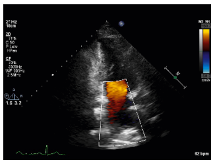
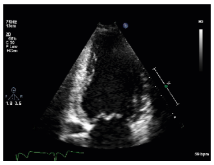
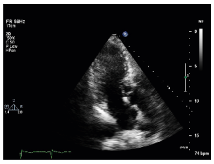
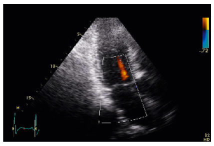
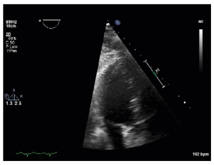
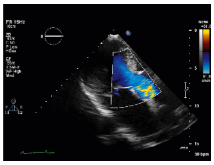
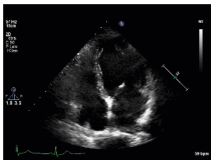
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **a2ch-full** | Apical 2-chamber, covering the depth of the whole of the LV and LA. | LV volumes (EDV, ESV), LV ejection fraction (biplane) |  |
| **a2ch-la** | Apical 2-chamber that either has Doppler around the MV or LA, or is plain 2D but not covering the whole depth of the LV. | LA volume |  |
| **a2ch-lv** | Apical 2-chamber, covering the whole depth of the LV, but not the whole depth of the LA. | Wall motion abnormalities (anterior, inferior, apical segments) |  |
| **a3ch-full** | Apical 3-chamber, covering the whole depth of the LV and the LA. | LV global longitudinal strain. |  |
| **a3ch-la** | Apical 3-chamber, that either has Doppler around the MV or LA, or is plain 2D but not covering the whole depth of the LV. | Mitral regurgitation jet via color doppler |  |
| **a3ch-lv** | Apical 3-chamber, covering the whole depth of the LV but not the whole depth of the LA. | Wall motion abnormalities (anteroseptal, inferolateral, apex) |  |
| **a3ch-outflow** | Apical 3-chamber, with either colour that is more looking at LVOT than the MV or grayscale but limited to the LVOT area. | LVOT VTI (PW Doppler) |  |
| **a4ch-full** | Apical 4-chamber. Needs to show the full depth of all four chambers without specifically focussing on LV or RV. | LV volumes (EDV, ESV), LV ejection fraction (biplane) |  |

**Table A.14**
Apical echocardiographic views with representative images and descriptions.

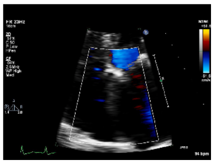
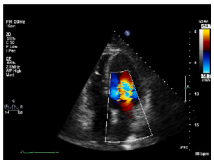
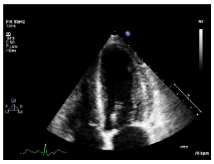
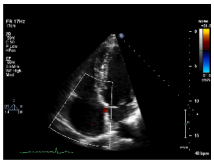
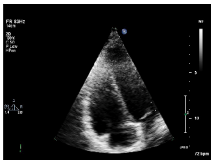
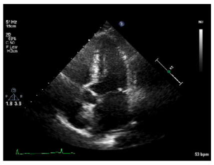
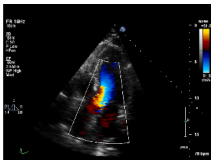
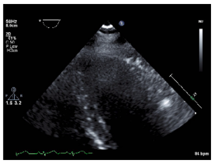
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **a4ch-ias** | Apical 4-chamber, focused on the inter-atrial septum. | Inter-atrium shunt (ASD/PFO) |  |
| **a4ch-la** | Apical 4-chamber, that either has Doppler around the MV or LA, or is plain 2D but covering the whole depth of LA and not whole depth of the LV. | Mitral regurgitation jet via color doppler |  |
| **a4ch-lv** | Apical 4-chamber, covering the whole depth of the LV and not the whole depth of the LA. | Wall motion assessment (inferoseptal and anterolateral walls) |  |
| **a4ch-ra** | Apical 4-chamber, focused on the right atrium, or with colour doppler on the Tricuspid valve or Right Atrium. | RA area |  |
| **a4ch-rv** | Apical 4-chamber, focused on the right ventricle. | RV basal, mid, length diameter |  |
| **a5ch-full** | Full apical 5-chamber with depth covering from apex to back of atria, without colour on LVOT. | LVOT VTI (PW Doppler), Aortic gradient ( CW Doppler) |  |
| **a5ch-outflow** | Apical 5-chamber including the LVOT but eirther not covering the full depth from apex of the heart to the back of the atria, or with colour on the LVOT | Aortic regurgitation or stenosis via color doppler. |  |
| **apex** | Apical window with insufficient depth to reach the mitral ring. | LV apical thrombus |  |

**Table A.15**
M-Mode echocardiographic views with representative images and descriptions.

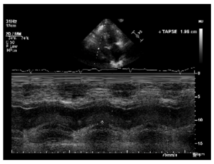
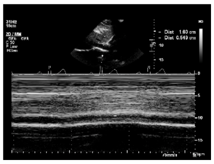
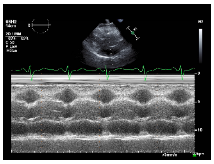
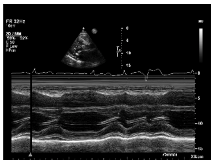
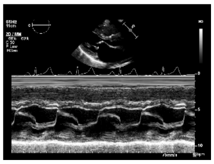
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **mmode-a4ch-rv** | M-mode for measuring TAPSE in apical 4-chamber view. | RV systolic dysfunction |  |
| **mmode-ivc** | M-mode of the inferior vena cava. | IVC diameter (inspiration / expiration) |  |
| **mmode-plax-av** | M-mode of the aortic valve in PLAX view. | Aortic root diameter, Aortic root dilation |  |
| **mmode-plax-lv** | M-mode focusing on LV walls in PLAX. | LVEDD, LVESD, IVS, posterior wall thickness, FS, LV hypertrophy |  |
| **mmode-plax-mitral** | M-mode focusing on mitral valve in PLAX. | EPSS, E-F slope, Mitral stenosis, LV systolic dysfunction |  |

**Table A.16**
PLAX echocardiographic views with representative images and descriptions.

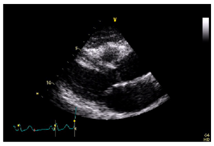
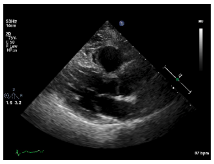
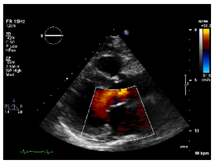
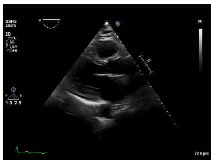
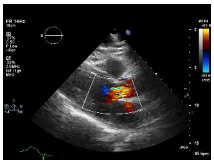
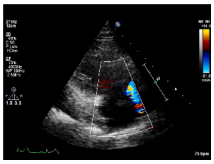
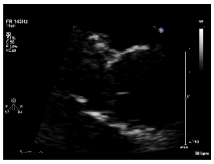
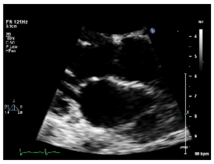
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **plax-full-la** | Parasternal long-axis, with the imaging sector covering the whole left atrium, intended for LA measurements. | LA diameter |  |
| **plax-full-lv** | Parasternal long-axis, with the imaging sector covering the LV but not the whole LA. | Linear measurements: LVEDD, LVESD, IVS thickness, Posterior wall thickness |  |
| **plax-full-mv** | Parasternal long-axis, centered on the mitral valve. This is signified by either so much zoom that parts of the LA and LV are lost from view or colour doppler over the Mitral Valve. | Mitral regurgitation jet via color Doppler |  |
| **plax-full-out** | Parasternal long-axis, zoomed out. | pleural and pericardial effusion |  |
| **plax-full-rv-ao** | Parasternal long-axis, where either the image has been zoomed or otherwise positioned to focus on the right ventricle and aorta or there is colour doppler in the region of the RV, aortic valve, or aorta. | RVOT diameter, Perimembranous VSD |  |
| **plax-tv** | Parasternal inflow view including the tricuspid valve. | Tricuspid regurgitation jet via color Doppler |  |
| **plax-valves-av** | Parasternal long-axis, focussed on the aortic valve. | LVOT, Aortic root diameter |  |
| **plax-valves-mv** | Parasternal long-axis, focused on the mitral valve. | Mitral valve abnormalities (e.g., prolapse or rheumatic changes) |  |

**Table A.17**
PSAX echocardiographic views with representative images and descriptions.

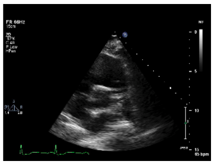
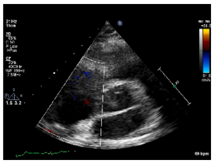
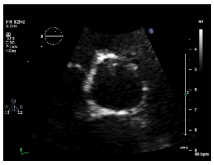
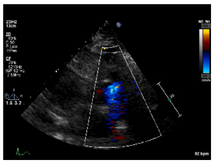
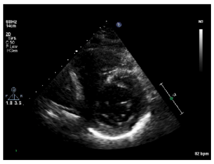
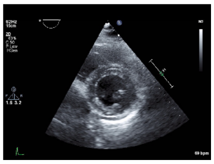
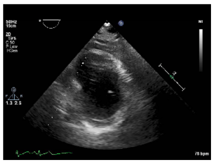
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **psax-all** | Parasternal short-axis, valve level, with the field of view covering the positions of the tricuspid, aortic and pulmonary valves (even if they are not clearly seen). | PV, AV, and TV stenosis/regurgitation |  |
| **psax-tv** | Parasternal short-axis, focused on tricuspid valve. | Tricuspid regurgitant jet velocity (CW Doppler) |  |
| **psax-av** | Parasternal short-axis, with field of view focused on aortic valve. | Aortic valve cusp morphology (tricuspid, bicuspid, unicuspid) |  |
| **psax-pv** | Parasternal short-axis, focused on pulmonary valve. | Pulmonary regurgitant jet velocity (CW Doppler) |  |
| **psax-lv-base** | Parasternal short-axis, left ventricle base level. | Mitral valve morphology ( stenosis and regurgitation), Regional wall motion abnormalities (basal) |  |
| **psax-lv-mid** | Parasternal short-axis, left ventricle mid-level. | Regional wall motion abnormalities (mid-level hypokinesia, akinesia, dyskinesia), LV hypertrophy |  |
| **psax-lv-apex** | Parasternal short-axis, left ventricle apex level. | Wall motion analysis (Apical hypokinesia, akinesia) |  |

**Table A.18**
Doppler echocardiographic views with representative images and descriptions.

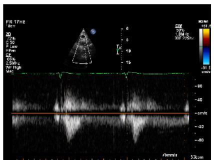
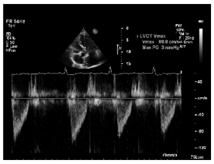
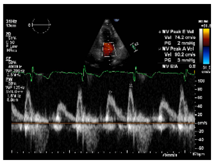
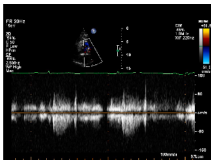
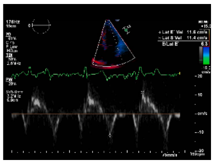
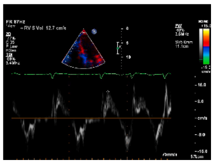
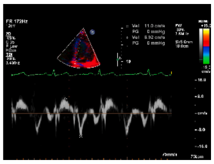
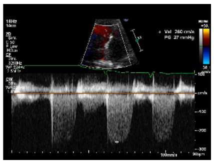
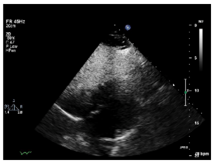
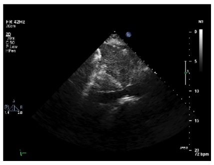
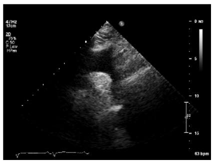
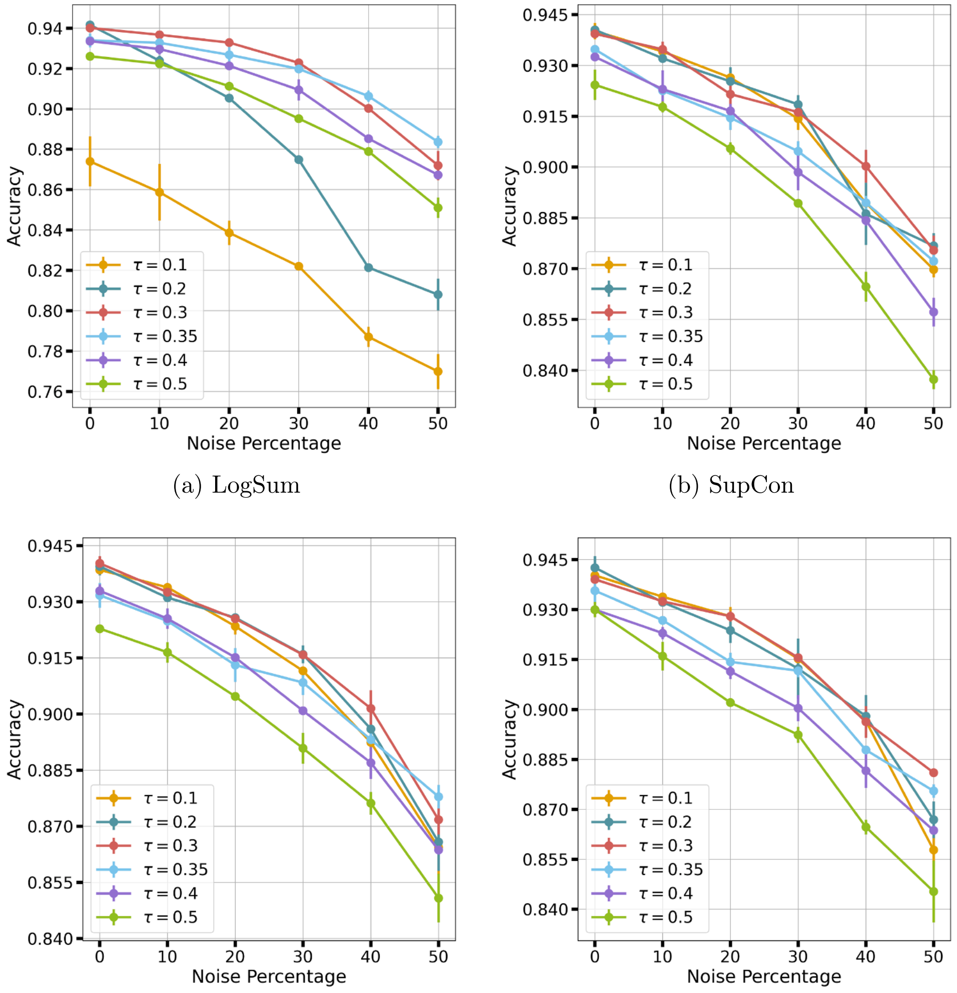
| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **doppler-ao** | Spectral Doppler of the descending aorta. | CW Doppler peak velocity |  |
| **doppler-av** | Spectral Doppler of the aortic valve. | LVOT VTI (PW Doppler) |  |
| **doppler-mv** | Spectral Doppler of the mitral valve. | Mitral inflow PW Doppler (E, A peak velocities, E/A ratio, deceleration time) |  |
| **doppler-pv** | Spectral Doppler of the pulmonary valve. | Peak velocity / pressure gradient via CW doppler |  |
| **doppler-tissue-lateral** | Tissue Doppler of the left ventricle lateral wall. | TDI of lateral mitral annulus (E' and A' velocities) |  |
| **doppler-tissue-rv** | Tissue Doppler of the right ventricle free wall. | S' (systolic annular velocity) for RV systolic function |  |
| **doppler-tissue-septal** | Tissue Doppler of the left ventricle septal wall. | TDI of septal mitral annulus (E' and A' velocities) |  |
| **doppler-tv** | Spectral Doppler of the tricuspid valve. | Peak velocity / pressure gradient via CW Doppler |  |

**Table A.19**
Subcostal and suprasternal echocardiographic views with representative images and descriptions.

| View Name | Description | Clinical Use | Image |
|---|---|---|---|
| **subcostal-heart** | Subcostal window, focused on the heart. | Pericardial effusion |  |
| **subcostal-ivc** | Subcostal window, focused on the inferior vena canva. | IVC diameter and collapsibility, Elevated right atrial pressure |  |
| **suprasternal** | Suprasternal view. | Coarctation of aorta |  |

## Appendix B. Ablation studies

For clarity, ablation plots are shown for the Xception backbone only; similar trends were observed across ConvNeXt-T, SwinTransformerV2-T, and EfficientNet-V2, as reflected in the results reported in Table 6.



(a) LogSum

(b) SupCon

(c) DropCon($k$=1,$v$=2)

(d) DropCon($k$=1,$v$=3)

**Fig. B.6.** Effect of temperature ($\tau$) and parameters parameters ($k$, $v$) across increasing annotation variability for EchoFine with an Xception backbone evaluated under different contrastive objective functions and configurations.

## Appendix C. Hierarchical grouping of echo views

**Table C.20**
Hierarchical grouping of echocardiographic views from fine-grained to coarse granularity.

| Level 3 (Coarse)<br>*7 Views* | Level 2 (Intermediate)<br>*20 Views* | Level 1 (Fine-Grained Views)<br>*47 Views* |
|---|---|---|
| Apical | a2ch | a2ch-full, a2ch-la, a2ch-lv |
| | a3ch | a3ch-full, a3ch-la, a3ch-lv, a3ch-outflow |
| | a4ch | a4ch-full, a4ch-la, a4ch-lv, a4ch-ias, a4ch-ra, a4ch-rv |
| | a5ch | a5ch-full, a5ch-outflow |
| | apex | apex |
| PLAX | PLAX | plax-full-out, plax-full-mv, plax-full-lv, plax-full-la, plax-full-rv-ao, plax-valves-av, plax-valves-mv, plax-tv |
| PSAX | PSAX | psax-all, psax-lv-base, psax-lv-mid, psax-lv-apex, psax-av, psax-tv, psax-pv |
| Subcostal | subcostal | subcostal-ivc, subcostal-heart |
| Suprasternal | suprasternal | suprasternal |
| M-mode | mmode-a4ch-rv | mmode-a4ch-rv |
| | mmode-ivc | mmode-ivc |
| | mmode-plax | mmode-plax-mitral, mmode-plax-av, mmode-plax-lv |
| Doppler | doppler-ao | doppler-ao |
| | doppler-av | doppler-av |
| | doppler-mv | doppler-mv |
| | doppler-tv | doppler-tv |
| | doppler-pv | doppler-pv |
| | doppler-tissue-lat | doppler-tissue-lat |
| | doppler-tissue-rv | doppler-tissue-rv |
| | doppler-tissue-septal | doppler-tissue-septal |

## Appendix D. Representation–label discrepancies in training embeddings

Although the training set was annotated by a single expert, clustering of the learned embeddings revealed consistent representation–label mismatches, suggesting latent ambiguity or inconsistencies in the labelling process. While these do not constitute formal intra-observer variability, they provide insight into annotation uncertainty and overlapping class boundaries in clinical echo interpretation.

To probe this further, we applied K-Means clustering to embeddings extracted from the pretrained encoder (without synthetic variation) and compared expert-provided labels against the dominant class within each cluster. This analysis identified cases where the semantic structure of the representation space diverged from the assigned class label, highlighting potential label–representation discrepancies.

Fig. D.7 shows the ten most frequent mismatches, many involving semantically adjacent views, such as *a3ch-outflow* vs. *a3ch-lv* or *plax-full-la* vs. *plax-full-out*. These confusions likely reflect anatomical overlap and intra-class variability. Mapping fine-grained views to their corresponding coarse 7-view categories confirmed that most mismatches occur within the same anatomical group. The group-wise confusion matrix (Fig. D.8) shows this effect, while Table D.21 quantifies it: 2,534 of 3027 mismatches (83.7%) remained within the same 7-view group.

Class-level analysis (Table D.22) revealed that views such as *a3ch-outflow*, *mmode-plax-lv*, and *subcostal-heart* were most frequently involved in mismatches, consistent with their greater acquisition variability and subtle distinctions.

Overall, these findings suggest that while the model exhibits ambiguity at the fine-grained level, it preserves the hierarchical organization of echocardiographic views, respecting broader anatomical categories even when class boundaries are less distinct. This capacity to maintain high-level semantic structure provides resilience to fine-grained inconsistencies in expert labelling and demonstrates the clinical plausibility of the learned representation space.

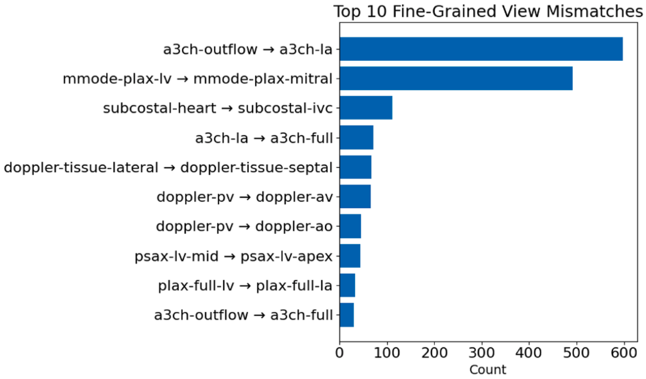


**Fig. D.7.** Top 10 most frequent fine-grained view mismatches, identified by clustering training embeddings. These confusions reflect potential annotation ambiguity or intra-class visual similarity.

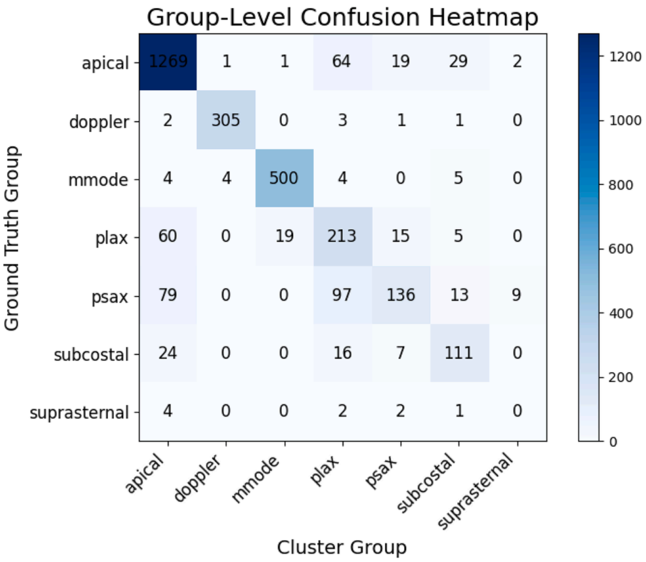


**Fig. D.8.** Group-wise confusion matrix derived from cluster-label mismatches. Most inconsistencies occur within the same coarse view category, such as Apical or PLAX, indicating semantic consistency at a higher granularity.

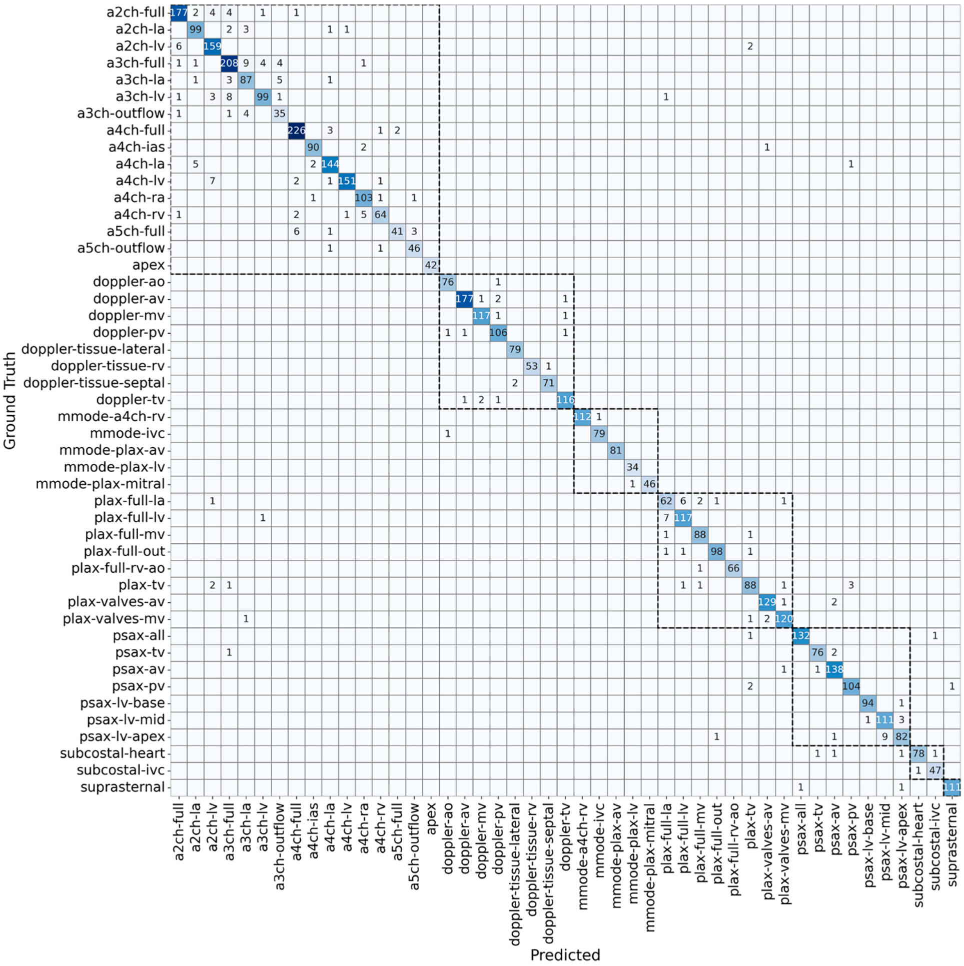
**Table D.21**
Distribution of cluster-label mismatches occurring within the same coarse (7-view) group vs. across different groups. The majority of mismatches are intra-group, suggesting strong structural alignment between learned embeddings and anatomical view categories.

| Mismatch Type | Count | Proportion (%) |
|---|---|---|
| Intra-group (same 7-view) | 2534 | 83.7 |
| Inter-group (different 7-view) | 493 | 16.3 |

**Table D.22**
Top 10 most frequently mismatched ground truth classes in the training set, based on dominant cluster disagreement. These classes exhibit the highest label-representation inconsistency and may reflect views that are visually ambiguous or difficult to label precisely.

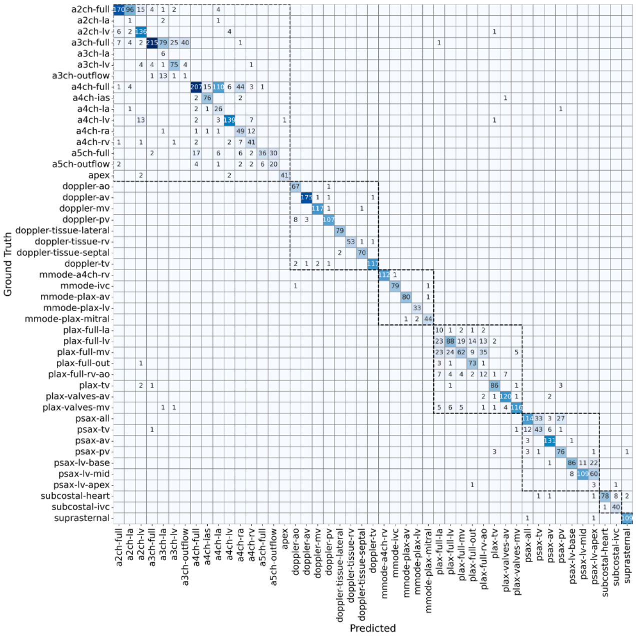
| View Class | Mismatch Count |
|---|---|
| a3ch-outflow | 637 |
| mmode-plax-lv | 492 |
| subcostal-heart | 141 |
| doppler-pv | 137 |
| a3ch-la | 109 |
| doppler-tissue-lateral | 86 |
| a3ch-lv | 86 |
| a2ch-lv | 80 |
| psax-pv | 77 |
| psax-lv-apex | 73 |

## Appendix E. Additional confusion matrices for expert and model

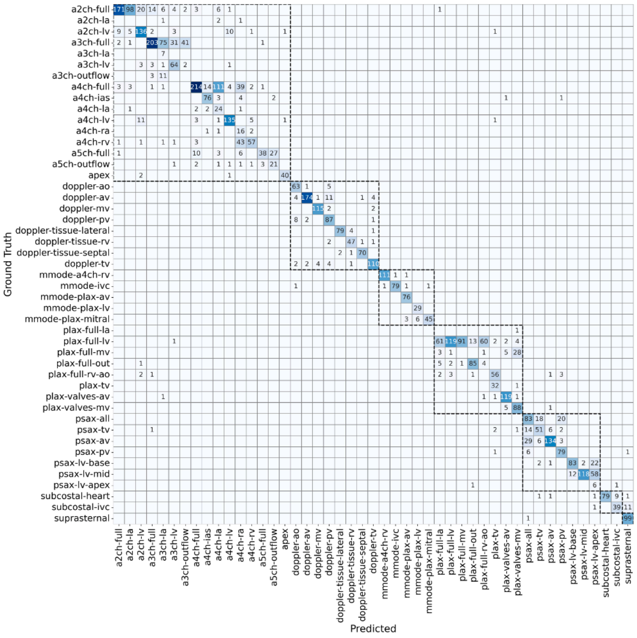


| |a2ch-full|a2ch-la|a2ch-lv|a3ch-full|a3ch-la|a3ch-lv|a3ch-outflow|a4ch-full|a4ch-ias|a4ch-la|a4ch-lv|a4ch-ra|a4ch-rv|a5ch-full|a5ch-outflow|apex|doppler-ao|doppler-av|doppler-mv|doppler-pv|doppler-tissue-lateral|doppler-tissue-rv|doppler-tissue-septal|doppler-tv|mmode-a4ch-rv|mmode-ivc|mmode-plax-av|mmode-plax-lv|mmode-plax-mitral|plax-full-la|plax-full-lv|plax-full-mv|plax-full-out|plax-full-rv-ao|plax-tv|plax-valves-av|plax-valves-mv|psax-all|psax-tv|psax-av|psax-pv|psax-lv-base|psax-lv-mid|psax-lv-apex|subcostal-heart|subcostal-ivc|suprasternal|
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|a2ch-full|177|2|4|4| |1| |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a2ch-la| |99| |2|3| | | | |1|1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a2ch-lv|6| |159| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |2| | | | | | | | | | | |
|a3ch-full|1|1| |208|9|4|4| | | | |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a3ch-la| |1| |3|87| |5| | |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a3ch-lv|1| |3|8| |99|1| | | | | | | | | | | | | | | | | | | | | | |1| | | | | | | | | | | | | | | | | |
|a3ch-outflow|1| | |1|4| |35| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a4ch-full| | | | | | | |226| |3| | |1|2| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a4ch-ias| | | | | | | | |90| | |2| | | | | | | | | | | | | | | | | | | | | | | |1| | | | | | | | | | | |
|a4ch-la| |5| | | | | | |2|144| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| | | | | | |
|a4ch-lv| | |7| | | | |2| |1|151| |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a4ch-ra| | | | | | | | |1| | |103|1| |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a4ch-rv|1| | | | | | |2| | |1|5|64| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a5ch-full| | | | | | | |6| |1| | | |41|3| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|a5ch-outflow| | | | | | | | | |1| | |1| |46| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|apex| | | | | | | | | | | | | | | |42| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|doppler-ao| | | | | | | | | | | | | | | | |76| | |1| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|doppler-av| | | | | | | | | | | | | | | | | |177|1|2| | | |1| | | | | | | | | | | | | | | | | | | | | | | |
|doppler-mv| | | | | | | | | | | | | | | | | | |117|1| | | |1| | | | | | | | | | | | | | | | | | | | | | | |
|doppler-pv| | | | | | | | | | | | | | | | |1|1| |106| | | |1| | | | | | | | | | | | | | | | | | | | | | | |
|doppler-tissue-lateral| | | | | | | | | | | | | | | | | | | | |79| | | | | | | | | | | | | | | | | | | | | | | | | | |
|doppler-tissue-rv| | | | | | | | | | | | | | | | | | | | | |53|1| | | | | | | | | | | | | | | | | | | | | | | | |
|doppler-tissue-septal| | | | | | | | | | | | | | | | | | | | |2| |71| | | | | | | | | | | | | | | | | | | | | | | | |
|doppler-tv| | | | | | | | | | | | | | | | | |1|2|1| | | |116| | | | | | | | | | | | | | | | | | | | | | | |
|mmode-a4ch-rv| | | | | | | | | | | | | | | | | | | | | | | | |112|1| | | | | | | | | | | | | | | | | | | | | |
|mmode-ivc| | | | | | | | | | | | | | | | |1| | | | | | | | |79| | | | | | | | | | | | | | | | | | | | | |
|mmode-plax-av| | | | | | | | | | | | | | | | | | | | | | | | | | |81| | | | | | | | | | | | | | | | | | | | |
|mmode-plax-lv| | | | | | | | | | | | | | | | | | | | | | | | | | | |34| | | | | | | | | | | | | | | | | | | |
|mmode-plax-mitral| | | | | | | | | | | | | | | | | | | | | | | | | | | |1|46| | | | | | | | | | | | | | | | | | |
|plax-full-la| | |1| | | | | | | | | | | | | | | | | | | | | | | | | | |62|6|2|1| | | |1| | | | | | | | | | |
|plax-full-lv| | | | | |1| | | | | | | | | | | | | | | | | | | | | | | |7|117| | | | | | | | | | | | | | | | |
|plax-full-mv| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| |88| | |1| | | | | | | | | | | | |
|plax-full-out| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1|1| |98| |1| | | | | | | | | | | | |
|plax-full-rv-ao| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| |66| | | | | | | | | | | | | |
|plax-tv| | |2|1| | | | | | | | | | | | | | | | | | | | | | | | | | |1|1| | |88| |1| | | |3| | | | | | |
|plax-valves-av| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |129|1| | |2| | | | | | | |
|plax-valves-mv| | | | |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1|2|120| | | | | | | | | | |
|psax-all| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| | |132| | | | | | | |1| |
|psax-tv| | | |1| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |76|2| | | | | | | |
|psax-av| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| |1|138| | | | | | | |
|psax-pv| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |2| | | | | |104| | | | | |1|
|psax-lv-base| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |94| |1| | | |
|psax-lv-mid| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1|111|3| | | |
|psax-lv-apex| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| | | | | | |1| | |9|82| | | |
|subcostal-heart| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1|1| | | |1|78|1| |
|subcostal-ivc| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1|47| |
|suprasternal| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |1| | | | | |1| | |111|

**Fig. E.9.** Comparison of model and expert agreement using a confusion matrix computed on a fixed subset of 4882 test images with complete annotations from all experts (excluding samples with “not sure”). The matrix shows Model vs. Expert 1, with strong diagonal alignment reflecting high consistency across clinically meaningful views.

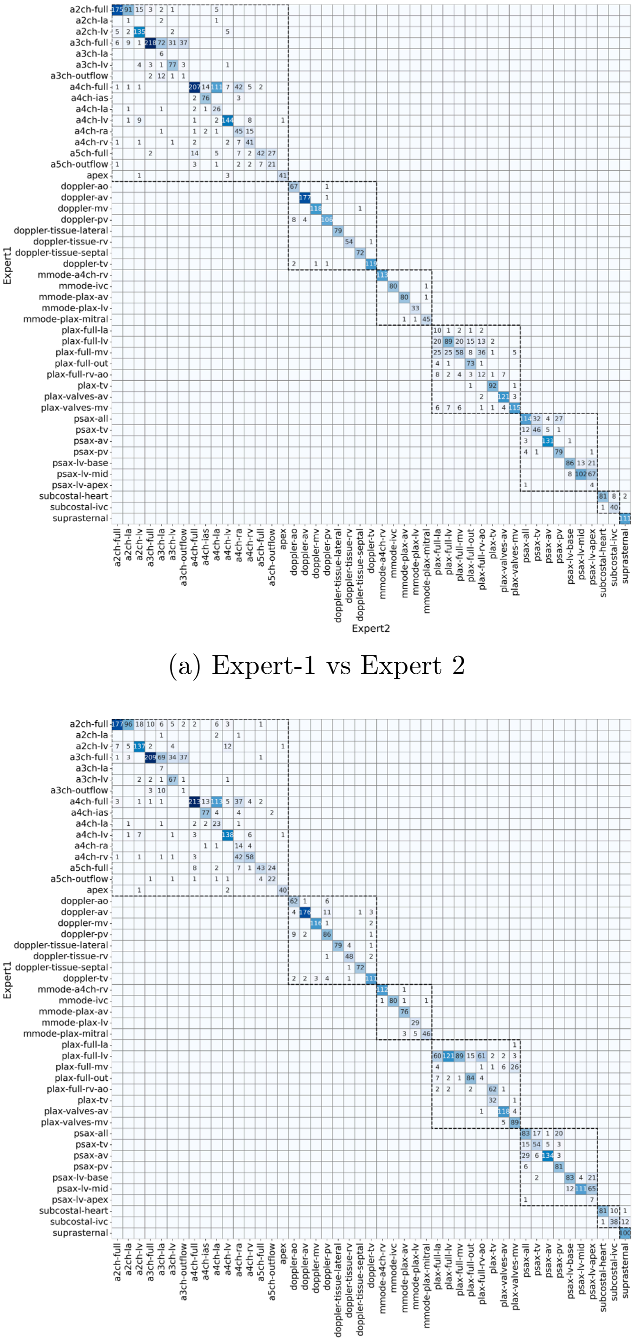


(a) Model vs Expert 2



(b) Model vs Expert 3

**Fig. E.10.** Confusion matrices comparing model predictions with Expert 2 and Expert 3 on a fixed subset of 4882 test images with complete multi-expert annotations. High diagonal concentration across both matrices indicates strong alignment between the model and each expert, even in the presence of label sparsity or fine-grained view distinctions.

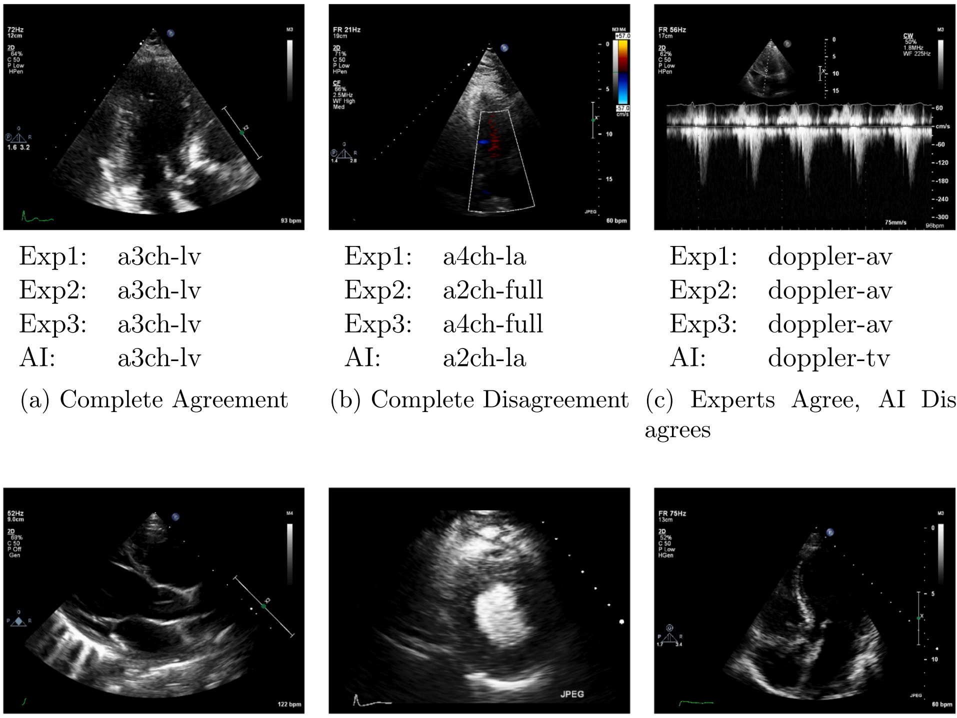


(a) Expert-1 vs Expert 2

(b) Expert-1 vs Expert 3

**Fig. E.11.** Confusion matrices comparing Expert-1 against Expert 2 and Expert 3 on the same fixed subset of 4882 test images. While overall agreement remains high for common view classes, off-diagonal inconsistencies are more pronounced than in the model comparisons, particularly for rare or visually similar views, highlighting inter-observer variability.

## Appendix F. Illustrative examples of inter-observer variability and model predictions



| | |
|---|---|
| Exp1: | a3ch-lv |
| Exp2: | a3ch-lv |
| Exp3: | a3ch-lv |
| AI: | a3ch-lv |

(a) Complete Agreement

| | |
|---|---|
| Exp1: | a4ch-la |
| Exp2: | a2ch-full |
| Exp3: | a4ch-full |
| AI: | a2ch-la |

(b) Complete Disagreement

| | |
|---|---|
| Exp1: | doppler-av |
| Exp2: | doppler-av |
| Exp3: | doppler-av |
| AI: | doppler-tv |

(c) Experts Agree, AI Disagrees

| | |
|---|---|
| Exp1: | psax-lv-mid |
| Exp2: | psax-lv-apex |
| Exp3: | psax-av |
| AI: | psax-lv-mid |

(d) AI = Exp1

| | |
|---|---|
| Exp1: | psax-lv-apex |
| Exp2: | psax-lv-mid |
| Exp3: | psax-lv-mid |
| AI: | psax-lv-mid |

(e) AI = Exp 2-3

| | |
|---|---|
| Exp1: | a5ch-full |
| Exp2: | a5ch-full |
| Exp3: | a4ch-full |
| AI: | a4ch-full |

(f) AI = Exp3

**Fig. F.12.** Cases illustrating inter-observer variability and model alignment on challenging echo views. Each subfigure includes labels assigned by three human experts and the model, highlighting cases of agreement, disagreement, and partial consensus. This analysis highlights both the inherent subjectivity in fine-grained cardiac view interpretation and the model's potential to approximate expert reasoning under uncertainty.

## References

Auh, J., Cho, C., Kim, S.-t., 2024. Improved contrastive learning model via identification of false-negatives in self-supervised learning. ETRI J. 46 (6), 1020–1029. https://doi.org/10.4218/etrij.2023-0285

Azarmehr, N., Ye, X., Howard, J., Lane, E., Labs, R., Shun-Shin, M.J., Cole, G.D., Bidaut, L., Francis, D., Zolgharni, M., 2021. Neural architecture search of echocardiography view classifiers. J. Med. Imaging 8. 10.1117/1.jmi.8.3.034002.

Chartsias, A., Gao, S., Mumith, A., Oliveira, J., Bhatia, K., Kainz, B., Beqiri, A., 2021. Contrastive learning for view classification of echocardiograms. Lect. Note. Comput. Sci. 12967 LNCS, 149–158. https://doi.org/10.1007/978-3-030-87583-1{_}15

Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. A simple framework for contrastive learning of visual representations. CoRR abs/2002.05709. https://arxiv.org/abs/2002.05709.

Chen, T.-S., Hung, W.-C., Tseng, H.-Y., Chien, S.-Y., Yang, M.-H., 2022. Incremental false negative detection for contrastive learning. In: International Conference on Learning Representations. https://openreview.net/forum?id=dDjSKKA5TP1.

Dedieu, L., Nerrienet, N., Nivaggioli, A., Simmat, C., Clavel, M., Gauthier, A., Sockeel, S., Peyret, R., 2024. Contrastive-Based Deep Embeddings for Label Noise-Resilient Histopathology Image Classification. https://arxiv.org/abs/2404.07605.

Gearhart, A., Goto, S., Deo, R.C., Powell, A.J., 2022. An automated view classification model for pediatric echocardiography using artificial intelligence. J. Am. Soc. Echocardiogr. 35 (12), 1238–1246. https://doi.org/10.1016/J.ECHO.2022.08.009

Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y., 2020. Deep learning interpretation of echocardiograms. npj Digit. Med. 3 (1), 1–10. https://doi.org/10.1038/s41746-019-0216-8

Graf, F., Hofer, C., Niethammer, M., Kwitt, R., 2021. Dissecting supervised contrastive learning. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 3821–3830. https://proceedings.mlr.press/v139/graf21a.html.

Guan, J., Liu, J., Huang, S., Yang, Y., 2024. ECLB: Efficient contrastive learning on bi-level for noisy labels. Knowl. Base. Syst. 300, 112128. https://doi.org/10.1016/j.knosys.2024.112128

Guo, Y., Bai, L., Yang, X., Liang, J., 2025. Improving image contrastive clustering through self-Learning pairwise constraints. IEEE Trans. Neural Netw. Learn. Syst. 36 (1), 328–340. https://doi.org/10.1109/TNNLS.2023.3329494

Gutmann, M., Hyvärinen, A., 2010. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Teh, Y.W., Titterington, M. (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 297–304. https://proceedings.mlr.press/v9/gutmann10a.html.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2019. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , https://doi.org/10.1109/CVPR42600.2020.00975

Holste, G., Oikonomou, E.K., Mortazavi, B.J., Wang, Z., Khera, R., 2024. Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning. Commun. Med. 4 (1), 1–10. https://doi.org/10.1038/s43856-024-00538-3

Howard, J.P., Tan, J., Shun-Shin, M.J., Mahdi, D., Nowbar, A.N., Arnold, A.D., Ahmad, Y., McCartney, P., Zolgharni, M., Linton, N. W.F., Sutaria, N., Rana, B., Mayet, J., Rueckert, D., Cole, G.D., Francis, D.P., 2020. Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. J. Med. Artifi. Intell. 3 (March), 4. https://doi.org/10.21037/JMAI.2019.10.03

Huang, B., Lin, Y., Xu, C., 2022a. Contrastive label correction for noisy label learning. Inf. Sci. (N.Y.) 611, 173–184. https://doi.org/10.1016/j.ins.2022.08.060

Huang, Z., Jiang, R., Aeron, S., Hughes, M.C., 2024a. Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Huang, Z., Long, G., Wessler, B., Hughes, M.C., 2022b. TMED 2: A dataset for semi-supervised classification of echocardiograms. In: DataPerf Workshop at ICML. https://tmed.cs.tufts.edu/papers/HuangEtAl_TMED2_DataPerf_2022.pdf.

Huang, Z., Sidhom, M.-J., Wessler, B.S., Hughes, M.C., 2023. Fix-A-Step: Semi-supervised Learning from Uncurated Unlabeled Data. https://arxiv.org/abs/2208.11870.

Huang, Z., Yu, X., Zhu, D., Hughes, M.C., 2024b. InterLUDE: interactions between labeled and unlabeled data to enhance semi-supervised learning. In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org.

Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M., 2022. Boosting contrastive self-supervised learning with false negative cancellation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 986–996. https://doi.org/10.1109/WACV51458.2022.00106

Jeon, J., ha, S., Yoon, Y., Kim, J., Jeong, H., Jeong, D., Jang, Y., Hong, Y., Chang, H.-J., 2023. Echocardiographic View Classification with Integrated Out-of-Distribution Detection for Enhanced Automatic Echocardiographic Analysis. arXiv:2308.16483.

Jiang, H., Gao, M., Hu, Y., Ren, Q., Xie, Z., Liu, J., 2023. Label-noise-tolerant medical image classification via self-attention and self-supervised learning. arXiv:2306.09718.

Khamis, H., Zurakhov, G., Azar, V., Raz, A., Friedman, Z., Adam, D., 2017. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Med. Image Anal. 36, 15–21. https://doi.org/10.1016/j.media.2016.10.007

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA.

Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q., 2025. EchoFM: foundation model for generalizable echocardiogram analysis. IEEE Trans. Med. Imaging , 1. https://doi.org/10.1109/TMI.2025.3580713

Letnes, J.M., Eriksen-Volnes, T., Nes, B., Wisløff, U., Salvesen, O., Dalen, H., 2021. Variability of echocardiographic measures of left ventricular diastolic function. The HUNT study. Echocardiography 38 (6), 901–908. https://doi.org/10.1111/ECHO.15073

Li, S., Xia, X., Ge, S., Liu, T., 2022. Selective-supervised contrastive learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 316–325.

Li, X., Zhang, H., Yue, J., Yin, L., Li, W., Ding, G., Peng, B., Xie, S., 2024. A multi-task deep learning approach for real-time view classification and quality assessment of echocardiographic images. Sci. Rep. 14 (1), 1–12. https://doi.org/10.1038/s41598-024-71530-z

Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2020. On modelling label uncertainty in deep neural networks: automatic estimation of intra- observer variability in 2D echocardiography quality assessment. IEEE Trans. Med. Imaging 39 (6), 1868–1883. https://doi.org/10.1109/TMI.2019.2959209

Ma, S., Zhang, Y., Li, D., Sun, Y., Qiu, Z., Wei, L., Dong, S., 2024. SPEMix: A lightweight method via superclass pseudo-label and efficient mixup for echocardiogram view classification. Front. Artif. Intell. 7, 1467218. https://doi.org/10.3389/FRAI.2024.1467218/BIBTEX

Madani, A., Arnaout, R., Mofrad, M., Arnaout, R., 2018. Fast and accurate view classification of echocardiograms using deep learning. npj Digit. Med. 1 (1), 1–8. https://doi.org/10.1038/s41746-017-0013-1

Mitchell, C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M.C., Horton, K., Ogunyankin, K.O., Palma, R.A., Velazquez, E.J., Madison, W., 2018. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of Echocardiography. J. Am. Soc. Echocardiogr. 32, 1–64. https://doi.org/10.1016/j.echo.2018.06.004

Naser, J.A., Lee, E., Pislaru, S.V., Tsaban, G., Malins, J.G., Jackson, J.I., Anisuzzaman, D.M., Rostami, B., Lopez-Jimenez, F., Friedman, P.A., Kane, G.C., Pellikka, P.A., Attia, Z.I., 2024. Artificial intelligence-based classification of echocardiographic views. Eur. Heart J. - Digit. Health 5 (3), 260–269. https://doi.org/10.1093/ehjdh/ztae015

van den Oord, A., Li, Y., Vinyals, O., 2019. Representation Learning with Contrastive Predictive Coding. https://arxiv.org/abs/1807.03748.

Ostvik, A., Smistad, E., Aase, S.A., Haugen, B.O., Lovstakken, L., 2019. Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. Ultras. Med. Biol. 45 (2), 374–384. https://doi.org/10.1016/J.ULTRASMEDBIO.2018.07.024

Otto, C.M., 2018. Textbook of Clinical Echocardiography. Elsevier/Saunders.

Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., Wu, J., 2024. A survey of label-noise deep learning for medical image analysis. Med. Image Anal. 95, 103166. https://doi.org/10.1016/j.media.2024.103166

Steffner, K.R., Christensen, M., Gill, G., Bowdish, M., Rhee, J., Kumaresan, A., He, B., Zou, J., Ouyang, D., 2024. Deep learning for transesophageal echocardiography view classification. Sci. Rep. 14 (1), 1–10. https://doi.org/10.1038/s41598-023-50735-8

Vaseli, H., Liao, Z., Abdi, A.H., Girgis, H., Behnami, D., Luong, C., Dezaki, F.T., Dhungel, N., Rohling, R., Gin, K., Abolmaesumi, P., Tsang, T., 2019. Designing lightweight deep learning models for echocardiography view classification. In: Fei, B., Linte, C.A. (Eds.), Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling. SPIE, p. 109510F. https://doi.org/10.1117/12.2512913

Wegner, F.K., Benesch Vidal, M.L., Niehues, P., Willy, K., Radke, R.M., Garthe, P.D., Eckardt, L., Baumgartner, H., Diller, G.-P., Orwat, S., 2022. Accuracy of deep learning echocardiographic view classification in patients with congenital or structural heart disease: importance of specific datasets. J. Clin. Med. 11 (3). https://doi.org/10.3390/jcm11030690

Xu, L., Xie, H., Wang, F.L., Tao, X., Wang, W., Li, Q., 2024. Contrastive sentence representation learning with adaptive false negative cancellation. Inform. Fusion 102, 102065. https://doi.org/10.1016/j.inffus.2023.102065

Zamzmi, G., Oguguo, T., Rajaraman, S., Antani, S., 2022. Open world active learning for echocardiography view classification. Proc. SPIE Int. Soc. Opt. Eng. 12033, 20. https://doi.org/10.1117/12.2612578

Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow twins: self-supervised learning via redundancy reduction. CoRR abs/2103.03230. https://arxiv.org/abs/2103.03230.

Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C.R., Fleischmann, K.E., Melisko, M., Qasim, A., Shah, S.J., Bajcsy, R., Deo, R.C., 2018. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation 138 (16), 1623–1635. https://doi.org/10.1161/CIRCULATIONAHA.118.034338/SUPPL{_}FILE/OCT

Zhang, S., Chu, S., Qiang, Y., Zhao, J., Wang, Y., Wei, X., 2025. Combating medical label noise through more precise partition-correction and progressive hard-enhanced learning. Comput. Method. Program. Biomed. 265, 108734. https://doi.org/10.1016/j.cmpb.2025.108734

Zhu, Y., Ma, J., Zhang, Z., Zhang, Y., Zhu, S., Liu, M., Zhang, Z., Wu, C., Yang, X., Cheng, J., Ni, D., Xie, M., Xue, W., Zhang, L., 2022. Automatic view classification of contrast and non-contrast echocardiography. Front. Cardiovascul. Med. Volume 9 - 2022. https://doi.org/10.3389/fcvm.2022.989091