Hierarchical Class Incremental Learning of Anatomical Structures in Fetal Echocardiography Videos

Arijit Patra and J. Alison Noble

***Abstract*— This paper proposes an ultrasound video interpretation algorithm that enables novel classes or instances to be added over time, without significantly affecting prediction abilities on prior representations. The motivating application is fetal echocardiography in mid-trimester scans. In this application, a sonographer may acquire multiple video clips of the heart at different points in the full scan. The goal is to make a complete inference of the health of the fetal heart from those multiple clips.**

To address this scenario, we propose to use an incremental learning approach to build a hierarchical network model that allows for a parallel inclusion of previously unseen anatomical classes without requiring prior data distributions. Super classes are obtained by coarse classification followed by fine classification to allow the model to self-organize anatomical structures in a sequence of categories through a modular architecture.

We show that this approach can be adapted with new variable data distributions without significantly affecting previously learned representations. Two extreme situations of new data addition are considered; (1) new class data is available over time with volume and distribution similar to prior available classes, and (2) imbalanced datasets arrive over future time to be learned in a few-shot setting. In either case, availability of data from prior classes is not assumed. Evolution of the learning process is validated using incremental accuracies of fine classification over novel classes and compared to results from an end-to-end transfer learning-derived model fine-tuned on a clinical dataset annotated by experienced sonographers. The modularization of subsequent learning reduces the depreciation in future accuracies over old tasks from 6.75% to 1.10% using balanced increments. The depreciation is reduced from 6.95% to 1.89% with imbalanced data distributions in future increments, while retaining competitive classification accuracies in new additions of fine classes with parameter operations in the same order of magnitude in all stages in both cases.

***Index Terms*— Fetal ultrasound, Hierarchical models, Incremental learning, Similarity learning**

Manusript submitted for review on April 30, 2019. The authors acknowledge EPSRC grant EP/M013774/1 (Seebibyte) and ERC Advanced Grant 694581 (PULSE).

Arijit Patra and J. Alison Noble are with the Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, UK (e-mail: arijit.patra@eng.ox.ac.uk; Alison.noble@eng.ox.ac.uk).

I. Introduction

FETAL ultrasound (US) is widely employed in pre-natal healthcare diagnosis and pregnancy monitoring worldwide due to its non-invasive, non-ionizing nature and non-trivial issues with other modalities like MRI which are often restricted to 20 weeks gestational age and above and are not suitable for large scale screenings in low and middle income countries health systems. However, US-based diagnostic solutions have well-known limitations; operator variability in acquisition and interpretation, cost of equipment, and a shortage of highly-skilled sonographers. This led researchers to explore medical image analysis solutions for automated standardized view plane-finding and biometry.

Congenital heart disease is a major cause of global infant mortality, estimated at 8 in 1000 live births [20]. Fetal heart anomalies are often misdiagnosed due to a lack of equipment and expertise causing fetal heart US (echocardiography) to be left out from 20-week pregnancy scan assessment. Analysis of fetal echocardiography video is a challenging task, even for experienced sonographers, due to potentially indistinct appearance of multiple moving anatomical structures in a small area, small size of objects-of-interest, unpredictable fetal movements, speckle, shadowing and other imaging artefacts. Recently, deep learning algorithms have shown significant success in domains like image, speech and video understanding. Adaptation of deep learning methods, such as convolutional and recurrent networks, have achieved state-of-the-art results for some medical image analysis tasks [6].

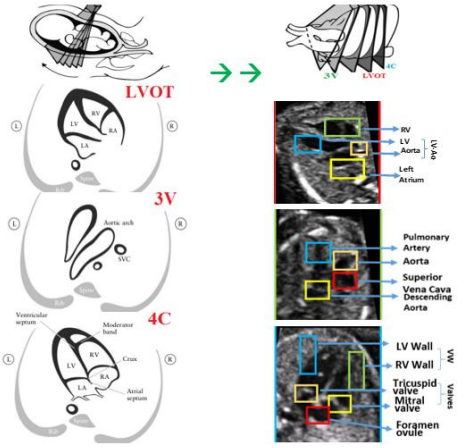

Fig. 1. The ISUOG scanning protocols (*top*) [21] for LVOT, 3V and 4C view assessments and anatomical structures in the three standard viewing planes. Zoom in for details.

Often a requirement for deep networks is the availability of large labeled datasets. However, in medical imaging, data is often not available in large volumes or access is restricted by data governance policies [21]. This poses a problem in medical imaging where current deep-learning based models need large representative datasets to accommodate intra-patient variations, physiological differences due to disease, different acquisition methods and so on. In practice, sufficiently representative data may not be available at the outset but curated over time, or there may be a need to merge datasets from different sites to increase prognosis value. We consider this scenario in the paper.

Humans acquire knowledge and learn new tasks building upon past learning, and this does not lead to a disappearance of past knowledge. In machine learning, *Lifelong Learning* is an approach that uses previously learned information to acquire new knowledge, using the idea that learning the n^{th} task should be easier than learning the $(n\text{-}1)^{th}$ task [1], [2]. This stems from the goal of creating autonomous agents capable of attaining intelligent capabilities over a subsequent range of tasks. Under assumptions of such new knowledge being incrementally added as new classes, the lifelong learning problem is essentially a class incremental learning problem.

How might lifelong learning be relevant to image-based diagnosis? In imaging-based disease diagnosis, available clinical datasets often contain a very small number of samples of disease case versus healthy cases (the baseline). Further, in the dynamic imaging case we consider in the paper, information about a condition may appear over time as a sonographer scans multiple times over an organ. In both cases, lifelong learning needs to address problems of class imbalance over multiple incremental learning stages without prior information on future distributions. Achieving robustness to potential variations in data distributions over new unknown classes in undefined timescales is a non-trivial problem. Addressing such challenges for clinical diagnostic videos requires deep learning based analysis to be amenable to successive learning from small datasets of future classes. This needs to occur without overfitting on new examples or losing the ability to generalize over past information. Availability of past datasets may not be necessarily assumed either. In other words, such models need to be able to accomplish incremental few-shot learning [6].

Contributions. In this paper, we leverage hierarchical feature learning in convolutional networks in an incremental learning algorithm that aggregates a temporally spaced inflow of new information and learns novel classes of data in a self-organizing fashion without significantly altering learned features from previous datasets. We present a proof-of-concept on fetal echocardiography videos. Fetal echocardiography assessment in the 4C (four chamber) and non-4C view requires the identification of a number of anatomical sub-structures [30]. Such a multi-object classification task can be difficult in ultrasound videos and may require visualization over multiple scans. For instance, a sonographer may choose to visualize some sub-structures at higher resolution i.e. zoom-in, in a later part of a full-scan video or may find a better acoustic window in which to image the object at a later point of the full scan. In this case, the desire would be to fuse information from all video clips of the object to build a computational representation of the fetal heart. To our knowledge, this paper is the first to demonstrate incremental learning of different classes on spatio-temporal data and in a medical imaging video application. The proposed method is conceptually scalable to other medical imaging tasks where a hierarchy of features is learned and new classes are appended over time or old classes see an inflow of new data such as diseased instances. The key contributions of the article are that we propose:

1) A modular video classification pipeline for object detection of different anatomical structures evident in different fetal cardiac views, relying on natural hierarchies in ultrasound videos and expanding to account for new data in a self-organized fashion.
2) Incremental learning models that introduce new classes of anatomical structures without significantly affecting prediction accuracy for existing classes.
3) A similarity embedding setup for medical videos in a fetal ultrasound case, for few-shot regimes inspired by typical clinical scenario of obtaining variable length data over time.
4) Class-wise performance on forgetting and retention is studied over incremental additions.
5) Metrics to assess performance on incremental learning tasks for medical imaging, defined in terms of the ability to retain past knowledge and to learn future novel classes.

Unlike several state-of-the-art continual learning paradigms, the presented methods do not require examples of prior classes or generated representations thereof, and can work indefinitely over future incremental stages without past data with minimal loss in overall accuracy. Note that a definition of lifelong learning is not clearly established in machine learning literature nor has this been significantly discussed in the medical image analysis community. Most sources treat incremental learning, continual learning and lifelong learning as interchangeable terms [24]. In this paper, we will use the phrase "incremental learning" and will consider a class incremental condition where new classes are to be adapted for by a trained model. So, all data introduced over future time belongs to previously unseen classes. Further, there has been little consensus on good evaluation metrics for incremental learning algorithms, and therefore we have designed our own metrics suitable for the ultrasound imaging applications considered.

II. Prior work

A. Related Ideas in Computer Vision

Task Adaptation and Forgetting. Acquiring knowledge from examples is a continuous process and needs to evolve dynamically with new data. Most literature has considered the learning process as a function of static data (data known at initial training). There have been relatively few attempts to place learning in a context where data is dynamically available [24]. Attempts to use generic features by transfer learning [3] rely on such features being common across a range of similar tasks [16]. A barrier to incremental learning in connectionist models is the tendency to converge on new tasks with parameters sub-optimal for modelling prior tasks. This occurs

due to the property of connectionist networks that they rely on optimizing an objective function towards a global optimum without a constraint on the parameters used for arriving at such an optimum, implying that there exists no guarantee that parameters responsible for such optima on a given data distribution would be retained during a later optimization step on another distribution. Goodfellow et al. [2] studied this phenomenon, called 'catastrophic forgetting', where features previously learned are lost on re-training for new tasks leading to a diminished performance on old tasks. Techniques like fine-tuning [3] and feature extraction [4] are used to preserve learned features while adding new parameters to adapt to new tasks. Feature extraction is useful for cases where only a few new layers of parameters are initialized over and above learned layers borrowed from other similar tasks and a learning regime affects only these new layers [4]. In fine-tuning or transfer learning, the addition of new parameters is accomplished in a similar fashion but old parameters are selectively or wholly modified while training for new parameters [3]. A shortcoming of such techniques, apart from losing previous training results, is their requirement of similarly sized training datasets with a poor tolerance for class imbalance while re-training for new tasks. Using existing models for similar tasks pertinent to different modalities of datasets, a situation common in multi-modal medical imaging cases, can be tackled using domain adaptation [16]. Ideas like few-shot learning using Siamese networks [5, 20] work with a few samples to improve generalization on trained models for tasks where variations may exist within available data and comparative discrimination in input samples is essential.

Incremental Learning. Approaches mitigating catastrophic forgetting have focused [24] on 1) retraining with regularization to avoid drastic changes to learned weights; 2) multi-headed network expansion, either as modules for subtasks [11] or fixed size modules for each new task [25], and 3) selective retraining with expansion [15]. Learning without Forgetting (LwF) [7] relies on an output level preservation of learned knowledge by using a distillation loss on prior learned logits. An initial dataset was used to train a network and the softmax logits for classes previously trained for were frozen and used to regularize later learning. However, the storage of past prediction probabilities during future task adaptations causes memory overheads. In Elastic Weight Consolidation (EWC) [22], parameter importance is evaluated using a diagonal approximation of the Fisher information matrix, and a quadratic is penalty imposed on the difference in parameters across old and new tasks. Zenke et al. [23] penalized alterations to relevant parameters on past tasks by synaptic relevance obtaining results similar to EWC.

Dynamic Architectures. Xiao et al. [8] presented a node-branch formulation of an error-driven incremental learning system to include classes based on error thresholds. This strategy was used to address the diversity and size of the label space of the ImageNet dataset [14] and to allow for stage-wise classification rather than an end-to-end single classification. Rusu et al. [26] formulated fixed sized networks to be assigned for each new task, thus permanently locking all parameters used in a given task. Following concepts of rehearsal or mingling new class data with samples from past classes, iCaRL [12] was developed to incrementally learn from new classes by including exemplars from previous classes during new class training..

Metric Learning. Exploration of feature similarities in low dimensional spaces has been a mainstay in several areas like information retrieval [11], template matching [16], and increasingly, few-shot learning [6,11,20]. In tasks that are inherently data sparse, achieving a hyperplane separation between clusters of feature similarities becomes challenging and instead, a distance measure in the metric space has been suggested. In incremental learning problems, an implicit step is novelty detection on the presented instance - namely if in prior learning cycles the instance has been encountered or not. This is a stage where one can seek to exploit the few-shot nature of metric-based classification for class incremental learning, as a label different from what is already known by the system can be assigned to a novel exemplar.

Leveraging feature similarities for hierarchical classification of anatomical motion incrementally in modular architectures

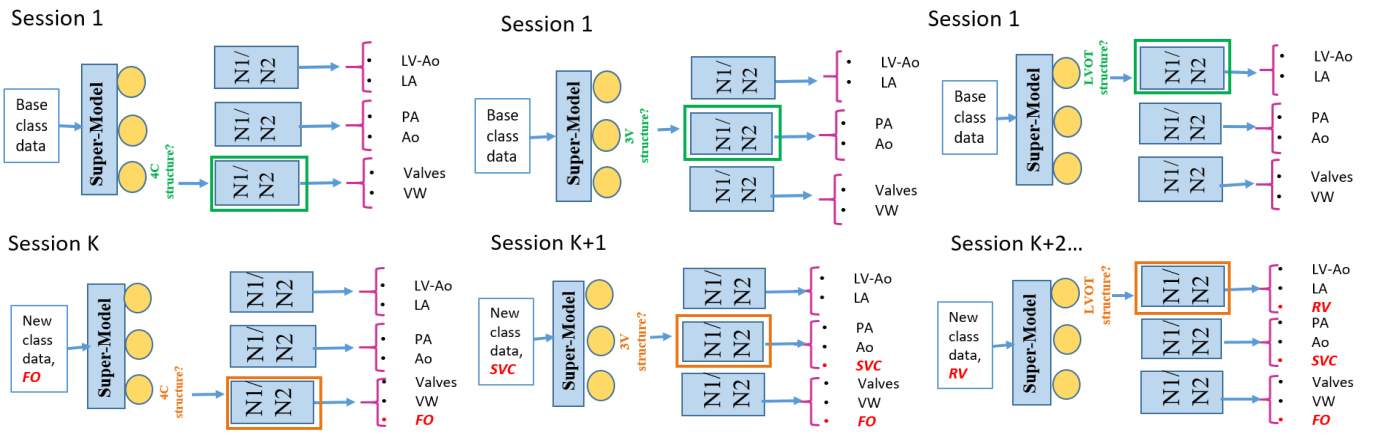

Fig. 2. Overall workflow over multiple learning sessions (K is the number of sessions the model learns over). After initial training on base classes in Session 1 (*top*), the model decides on the top-level class and directs to the corresponding sub-model for fine classification. Session K (and K+1,2..; K=2 in our tasks) is the learning session where the pipeline is shown previously unseen classes – FO, SVC and RV here. Top-level classification decision on previously unseen class data is taken by the super-model and the corresponding sub-model is adapted for the new data arrival. N1 is used as the sub-model configuration for Experiment 1 with balanced new class arrivals. N2 is used as the sub-model for few-shot new class arrivals in Experiment 2. Note the newly added classes are shown in red. Best viewed in Adobe reader.

that are resilient to variable sized datasets remains an open problem. We explore such an incremental task without assuming any presence of past data, a scenario considered in a very limited number of works in the past, notably in Progressive Networks [26], which assigns different networks to different tasks. Such an implementation leads to single networks for each task of a set of multiple 1-class classification tasks, with a redundancy in the parameter space. Thus, primarily transfer learning approaches are the only knowledge extension schemes closest to our task conditions.

Classification of structures can be segregated into models of expert classifiers [11] specializing in different kinds of fine classification, as determined per instance by a coarse category (analogous to referrals from general practitioners to specialists in health practice). Then an incremental data distribution can be adapted to the overall pipeline by restricting fine-tuning for a new addition to the corresponding expert, leaving the integrity of learned representations in other expert models intact. This is different from standard fine-tuning and transfer-learning approaches in that the process of adapting to new data is conditionally segregated using coarse distinctions. This is a trade-off between incrementing fixed network capacity to a new additional task [25] and prior knowledge retention by conditioning access to the parameter space on prior decisions on input instance characteristics than impose a hard bound on model components used to incorporate new task data.

B. *Related Ideas in Medical Imaging*

Medical imaging data is incrementally and continuously acquired in clinical research studies and hospital settings, however incremental learning in medical imaging applications has not received much attention to-date. For instance, Giritharan et al. [10] used incremental learning to diagnose positive and negative cases in endoscopy images by adapting SVM boundaries to new data of diseased instances in subsequent clinical sessions.

Fetal ultrasound image appearance varies due to the variability in pose of the object with respect to the transducer, signal attenuation, shadows and other factors that have made automated fetal ultrasound image analysis technically challenging. Machine learning-based methods are founded on pattern recognition that makes them, in theory, well-suited to ultrasound image analysis. As a result, since the advent of machine learning in medical image there has been growing interest on this topic. Attempts with machine learning approaches includes sequential localization for fetal cardiac anatomies [27], fetal standard plane detection and localization in free-hand ultrasound [28], and gaze-saliency based efficient model compression [29], saliency driven class additions [34], wall enhancement methods for the myocardium [31], visual enhancement of wall motions [32], classification of ventricular wall motions [33], quantification of bulls' eye maps for stress echocardiography [35] and so on. The literature specific to automatic fetal echocardiography is very small with previous work proposing fetal echocardiography view description using Bayesian methods, and CNN-based methods of multi-task view classification and localization [9, 20] without re-training for new tasks. There are no prior publications on incrementally refining models for cardiac ultrasound with previously unseen classes as proposed in this paper.

III. Methodology

A. *Problem Definition*

We define our task as to model fetal echocardiography video dynamics at any instance or session based on currently available

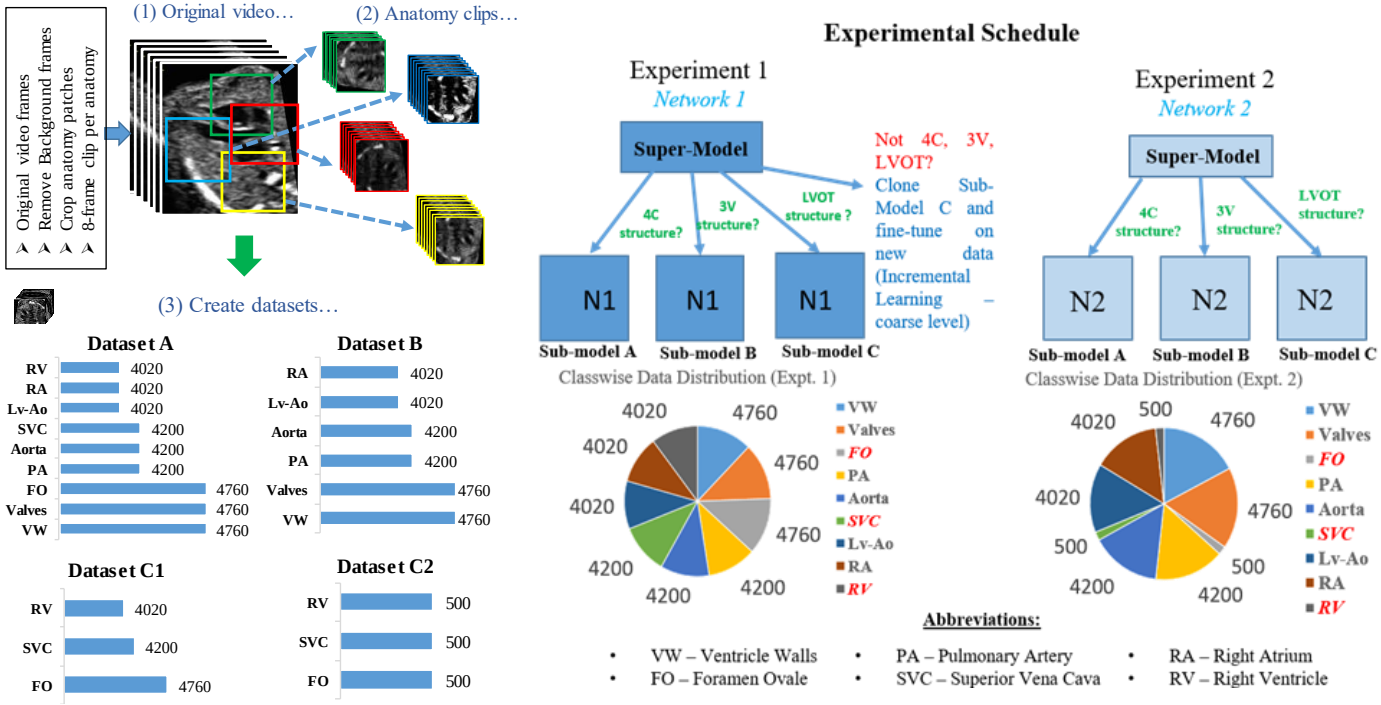

Fig. 3. (*left*) Data preparation from original video to clip generation and dataset aggregation (*right*)Experiments 1 and 2 with their data distribution situations. *Experiment 1* uses *Network 1* with *N1* modules for sub-models. *Experiment 2* uses *Network 2* with *N2* modules for sub-models (*old classes already trained on labeled in grey, new classes to be incrementally trained for labeled in red*).

streams of data. The level of detail in the model is incrementally learned in a manner that does not significantly compromise prior model knowledge. In our fetal echocardiography videos, the fetal heart is visible in three viewing planes –four chamber (4C), three-vessel (3V) and left-ventricular outflow tract (LVOT) [9]. Certain anatomical structures of the fetal heart can be visualized from one or more of these viewing planes. The 4C view includes structures such as the ventricles (their walls can be distinctly observed), the mitral and the tricuspid valves and the foramen ovale (Fig. 1). The 3V view typically includes the pulmonary artery, the aorta, and the superior vena cava. The LVOT includes the left and right ventricles, the aorta (LV and aorta are observed as a continuous cavity and labeled as LV-Ao) and the right atrium. Co-located structures in an anatomical region often have feature similarities suitable for hierarchical grouping and distinction from representations obtained from a different part of the fetus.

Why Hierarchical? With an ability to distinguish between coarse features over top-level divisions, we extend to finer classification modules without a need to train the top-level classifier in response to variable or incremental data distributions for different fine classifications. Anatomical structures seen in each cardiac view have structural and contextual similarities useful for stepwise feature learning. Specifically, structural groups are classified in parallel models. Hence, an incremental addition to the dataset can be facilitated in the same or a different fine-class in one of the coarse groups without compromising models predicting fine-classes in other groups. To initially train a sub-model, we choose the sub-classes to be 'ventricle walls' and 'valves' (left and right ventricle walls grouped into 'ventricle walls'; mitral and tricuspid valves into 'valves') in the 4C case, pulmonary artery and aorta in the 3V case, right atrium and left ventricle-aorta continuum in the LVOT case. Remaining structures are left for incremental learning experiments. The choice of exemplars is motivated by their clinical relevance as clinical biomarkers for diagnostic evaluations of congenital heart disease [21].

B. High-level Network Architecture Design

In this section, we define two network models with different fine classification strategies. These are representative of extreme situations of no class imbalance over new additions and highly imbalanced arrivals of new classes. We also define two experiments utilizing different network topologies conditioned for these two respective distributions handled (Fig. 2 and 3).

Experiment 1. We consider the cases of incremental distributions with dataset sizes of novel classes approximately resembling those of the trained prior classes. Here we use a hierarchy of models with softmax probability-based self-organization throughout. This pipeline is termed ***Network 1***.

Experiment 2. We consider the case of dataset sizes being imbalanced in novel classes as compared to prior learned distributions. We use a second hierarchy of models with the coarse classification stage employing a softmax classification approach to direct instances to sub-models. This adopts a similarity driven few-shot learning regime in a manner similar to metric learning approaches [19]. This pipeline is called ***Network 2***.

Both Network 1 and Network 2 use the same pre-trained model for coarse-classifier stages as this is implemented as a common softmax probability based classifier. While actual distributions could lie in-between these extremes in incremental stages, such a distribution can be attempted in both strategies. Thus in real world use-cases, the choice of architecture can be dictated by specific performances on available distributions.

C. Network Architecture Detail

In this section, we define the various pipeline components and discuss other design aspects. First, we define two model types:

- **Super-model** – A coarse classification stage, with a model implementing convolutional layers leading to a time-marching recurrent stage for taking video sequence input and classifying into pre-defined top-level classes. The softmax probabilities, averaged over a batch of inputs is used to determine the coarse or top-level class. This is different from standard classification pipelines where each instance is classified. While such information is available here, it is of secondary importance since the final outputs are the fine classes and so the flow of control to the appropriate sub-model is prioritized).
- **Sub-model** - Fine classification is implemented by a set of sub-models representing each coarse class the super-model can possibly generate. The models are pre-trained for classification (Network 1, with convolutional – fully connected - softmax architectures as N1 modules) or similar-class distance minimization (Network 2, with a Siamese architecture as N2 modules) on historically available categories.

We implement a conditioned hierarchy of deep learning models, trained separately but working as a multi-stage pipeline. The flow of information is conditioned by top-level decision inputs based on the softmax probabilities or distance metrics depending on the stage of classification and specific implementation. After the convolutional layers is a 512-way fully-connected layer that enables encoding of feature representations for distance computation. In Network 2, each input is separately processed using shared weights leading to two feature representations from each input pair. We encode similarity by a joint embedding function inspired from the Minkowski distance formulations (expressed as $\sum_i (|x_i - y_i|^p)^{1/p}$, $p \in R^+$. The embedding function used here is $\sum_i (|f(x)_i - f(y)_i|^1) + (|f(x)_i - f(y)_i|^2) + (|f(x)_i - f(y)_i|^3)$, where $f(x_i)$ and $f(y_i)$ are feature embeddings from instances compared. This allows the learning routine firstly to learn super-classes (coarse classified stage) based on general sets of features and to use that prior in a second learning step for detailed anatomical structure classification. In the limit of data distributions at the coarse classifier stage belonging to a single class and the original data being available in its entirety, the problem is reduced to a standard transfer learning problem where the addition of the incremental class assumes a new task similar in the same pattern to the old task.

Distance based objective vs Softmax. In *Network 2*, the sub-models learn to discriminate between similar and different images using a Siamese Network [5] with a binary classifier.

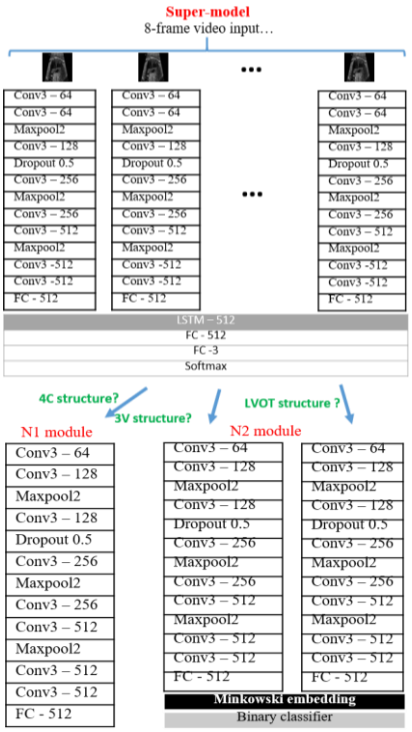

Fig. 4. Super-model for coarse classification(*top*). N1 and N2 modules (*below*).

This is preferred over a direct softmax classification in order to optimize for a distance measure between images of similar and different structures. This is useful if we add to existing datasets a few examples of a novel class and want our classification models to be able to detect instances from this class without a need to retrain with augmentation or generative modelling. Using a training regime with a softmax formulation in this case would suffer from the class imbalance problem of not being able to learn enough discriminative features with only a few examples of a new class, and not being able to predict optimally in the presence of such instances.

Using a distance metric that has been explicitly trained to be minimized for similar images and to be maximized otherwise is useful in this setting. This is because one only needs to compare an input image with representative examples of classes previously trained for and a threshold applied to distances computed used to categorize if a given example belongs to a pre-existing class. In models with a binary cross-entropy loss after distance layers, the same is possible from class output of the softmax function. This enables extension to novel classes of anatomies, as an image of such a structure is likely to structurally differ and encode different feature representations than those of classes already trained on. One can predict if an input instance belongs to a known set of classes or of a different unseen class. Over multiple such examples, it is possible to train a distance minimization for the same, either by fine-tuning the full sub-model or the higher layers alone (feature extraction), without need of a large number of instances in the limit of small learning rates. Incremental learning can be boosted by fine-tuning the relevant sub-model to minimize distances for a novel class using its examples or by creating a new clone of a sub-model (if softmax output from super-model is below a set threshold, in a case of a novel coarse class).

We now provide more architectural detail specific to the fetal echocardiography application model:

Super-model: In both experiments, we first implement a convolutional network, inclusive of LSTM layers for temporal processing, with a 3-way softmax cross-entropy classifier to identify standard fetal heart views – *4C structures*, *3V structures*, and *LVOT structures*. This model essentially performs coarse classification. It consists of multiple convolutional layers, followed by LSTM layers and a final softmax classifier stage. The super-model places input data into one of the coarse classes and provides a confidence score per class based on which subsequent sub-models are called for fine classification.

The super-model is implemented as an architecture with eight convolutional layers with 3x3 filters accompanied by ReLU activation and max-pooling with 2x2 pooling windows and a dropout of 0.5 after the 3rd convolutional layer to avoid overfitting. The reasons for choosing a 3x3 kernel rather than equivalents like 5x5 kernels that map a sequence of convolutional operation is to account for fine features in the ultrasound image space, which may otherwise be not learnt with sufficient information content. This variation has been studied for ultrasound [32] and in the same dataset by multiple earlier works [9, 20]. The convolutional layers are followed by a single fully-connected layer of 512 units feeding into an LSTM layer of 512 recurrent units linking to two 512-way fully-connected layer linked to a *N*-way fully-connected layer feeding to a *N*-way softmax activation, where *N* is the number of super-classes (*N*=3 in our case). The model processes 8-frame video inputs frame by frame and passes the first fully-connected output to the LSTM layer to aggregate the video temporal dynamics. The model layers are initialized in a Xavier-improved scheme [30].

Sub-models: Sub-models are implemented per class of the super-model. Initially, they are trained independently of the super-model. There are two kinds of implementations for the two strategies explored, described in Fig. 4 as *N1* and *N2* modules. An *N1* module is designed using five convolutional layers with alternate max-pooling followed by two fully-connected layers and a softmax classification layer comprises the fine classifier. The *N1* module is replicated thrice into sub-models A, B and C corresponding to the three coarse classes.

We train three *N2* based sub-models (A, B, C) relevant to each coarse class. These are modeled as Siamese networks [5] taking multiple-frame inputs. The strands process each input stream and optimize a distance metric by learning to classify

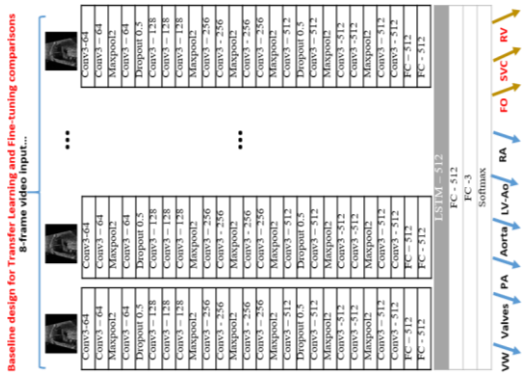

Fig. 5. Baseline architecture for transfer learning based fine-tuning.

'same' and 'different' classes. The model has 8 convolutional layers with 3x3 filters, ReLU activation and alternate max-pooling. The first layer is configured to accept a sequence of eight images to aggregate temporal information directly similar to [9]. While most prior work in metric learning and related areas like few-shot learning has typically used *L1* and *L2* distances, we use a customized embedding function to leverage the numerical magnifications achieved with powers over absolute differences. This helps obtain a more reliable representation of similarity of features in our datasets of anatomical structures from ultrasound videos where structural anatomical definitions are influenced by the presence of speckle, enhancements and noise.

Higher values of p enable a magnified representation of distances between feature representations and discriminative dissimilarities. The value of p is a hyperparameter and we assume $p = 3$, by performing a grid search over integral values between 1 and 8 (Fig. 6), optimizing for overall accuracy after new classes are introduced. The design choice of $p \notin Z^+$ is assumed as non-integral values as number of metric distances added to form the Minkowsky embedding is equal to p. Distance between two representations from fully-connected layers is computed and constrained between 0 and 1 for classification to 'same' and 'different' classes optimized by logistic regression $t.\log(p(x_1 \circ x_2)) + (1-t).\log(1-p(x_1 \circ x_2))$, with $p(x_1 \circ x_2)$ (set equal to the normalized Minkowski embedding value obtained between feature representations of x_1 and x_2) as the probability that *x1* and *x2* are of the same class; $t = 1$ if instances belong to the same class and $t = 0$ otherwise.

In both the variants of the networks, the depths have been so chosen as to keep the overall numbers of parameters involved in a single forward pass in the same order or magnitude as would be in the baselines considered. This is to ensure that the order of parameterization has a minimal role in determining knowledge retention or forgetting. Parameters for temporal extension are delegated to the LSTM stages where feature representations across spatial frames are aggregated in a temporal dimension and final representations can be obtained thereafter. Again, the choice of the number of units is dictated by the requirement to constrain the order of parameterization to be within baseline parameter dimensionality.

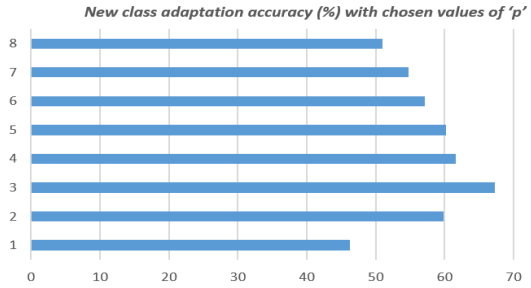

Fig. 6. New class adaptation accuracy (%) with chosen values of '*p*'

IV. Materials and Methods

A. Datasets

The models are trained using a clinical fetal echocardiography videos acquired using a GE Voluson E8 machine. The data was acquired following local ethics guidelines. The dataset consisted of 91 fetal cardiac screening videos from 12 healthy subjects between 20-35 weeks of pregnancy (Fig. 3). The video length ranged between 2-10 s, with a frame rate of 25-76 frames per second with 39556 frames in total. Videos of 10 subjects were used in training and 2 subjects' videos for testing across experiments ('subject-wise split').

B. Pre-processing

The videos were decomposed into frames and data augmentation performed by random top-down and bottom-up flips, besides rotational augmentation in steps of ten degrees. Resizing frames to 400x400 pixels, relevant anatomical patches were identified manually in resolutions of 100x100 pixels. The patches were grouped into 8-frame clips in order of original occurrence in the videos. This resulted in 4760 8-frame clips of 4C view sub-anatomies (ventricle walls, valves, foramen ovale), 4200 clips of the 3V view and its sub-anatomies and 4020 clips of the LVOT view and sub-anatomies. Each video clip was labeled with the fetal heart view it originated from and the structure it highlighted. Video clips of anatomical structures belonging to the 4C view were labeled '*4C structures'*, and

EXPERIMENT SET I

Network 1 used in *Experiment 1* (*non few-shot case*) and comparisons made with a direct transfer learning baseline on performances over incremental classes and retention of accuracy on old tasks (prior 6 classes of VW, Valves, PA, Aorta, LV-Ao, RA). '***Overall drop***' is a proxy for knowledge retention. Network 1 outperforms the baseline (Table 2).

Table 1 : *Network 1* performance on incremental learning tasks (%). A slower decline in performance on old tasks is seen

Stage	4C Structures *(Sub-model A)*			3V Structures *(Sub-model B)*			LVOT Structures *(Sub-model C)*			Average
	VW	Valves	FO	PA	Aorta	SVC	LV-Ao	RA	RV	**(old tasks)**
No Increment	91.25	85.40	--	68.35	60.23	--	74.05	60.50	--	73.29
FO increment	90.22	84.75	67.55	68.35	60.23	--	74.05	60.50	--	73.01
SVC increment	90.22	84.75	67.55	67.10	57.24	64.85	74.05	60.50	--	72.31
RV increment	90.22	84.75	67.55	67.10	57.24	64.85	71.30	59.23	60.20	71.64
All increment	90.22	84.75	67.55	67.10	57.24	64.85	71.30	59.23	60.20	71.64
Overall Drop	1.03	0.65	0.00	1.25	2.99	0.00	2.75	1.27	0.00	**1.10**

Table 2 : Transfer learning (*Baseline*) for Experiment 1 – classes added in stages impact accuracies across coarse groups (%)

Stage	4C Structures			3V Structures			LVOT Structures			Average
	VW	Valves	FO	PA	Aorta	SVC	LV-Ao	RA	RV	**(old tasks)**
Pre-Transfer	*94.15*	*88.24*	--	*71.34*	*70.10*	--	*72.95*	*63.83*	--	*76.77*
FO transfer	*89.60*	*82.10*	*42.35*	*62.15*	*57.15*	--	*68.26*	*59.11*	--	*69.72*
SVC transfer	*88.85*	*81.15*	*41.90*	*61.95*	*56.25*	*41.20*	*67.10*	*55.42*	--	*68.45*
RV transfer	*88.40*	*80.42*	*41.60*	*60.22*	*51.75*	*40.75*	*65.10*	*52.15*	*52.90*	*66.34*
Overall Drop	*5.75*	*4.82*	*0.75*	*11.12*	*18.35*	*0.45*	*7.85*	*11.68*	-	***6.75***

clips from 3V and LVOT structures were labeled equivalently. The set of video clips accumulated with all the labeled sub-anatomies is called Dataset A. Next, we created Dataset B consisting of fine labels. It comprised of only ventricle walls and valves from the 4C view structures set, the pulmonary artery and aorta from the 3V view set and LV-Ao continuum and right atrium from LVOT view set. The foramen ovale clips from the 4C view, superior vena cava from the 3V view and right ventricles from the LVOT view were left for incremental learning. Dataset C had all incrementally added fine structures with their fine labels. C1 had novel fine structures in proportions similar to the old classes. C2 had 500 instances per new class. Note that C1 and C2 were derived from C to simulate experimental conditions of standard incremental versus few-shot incremental learning and as such, dataset C is not the union of C1 and C2 but a parent dataset. Overall, Dataset A is the master dataset (used to derive test instances for assessing knowledge retention and overall drop in accuracies/ degree of forgetting), Dataset B was used for initial trainings, C1 for incremental training in Experiment 1 and C2 for the same in Experiment 2. C1 and C2 were used for incremental fine-tuning in baseline architectures.

C. Baselines

For comparisons with a direct coarse-to-fine classifier using incrementally growing data distributions, we use a baseline architecture composed of convolutional layers for spatial processing feeding into a temporally aggregating long short term memory (LSTM) layer followed by classifier stages (Fig. 5). The design of this model ensures the parameterization is in the same order as that of the super-model and a single sub-model combined. It is trained with all old fine classes (subject-wise, 10 for training and 2 for test, split on Dataset B, batch size of 50 and learning rate of 0.001) initially. Like data distributions used in the models proposed, in additions of new classes in *Experiment 1* and *2* we use datasets of new classes C1 and C2 in a subject-wise split. Class prediction accuracies for old and new classes are used to benchmark *Experiment 1* and *Experiment 2*.

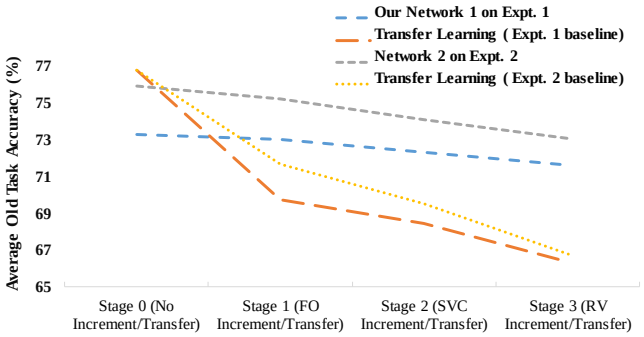

Fig. 7. Diminishing old task performance for transfer learning baselines compared to near constant performance of *Networks 1* and *2*.

V. Experiments

A. Early training

We first train the super-model and the sub-models on the datasets assumed as old classes. Following this, incremental training is performed sequentially on the datasets for the new classes. To begin, we use Dataset A for training and testing the super-model. In this case, the super-model needs to distinguish a structure from 4C, 3V and LVOT views respectively. This step is performed with a mini-batch of 100 and a learning rate of 0.01 using Stochastic Gradient Descent in both cases. The next step is training sub-models for fine classifications into different fine structures of the initial classes. Dataset B is used with a subject wise training-test data split (data originally

EXPERIMENT SET 2

Network 2 is used in *Experiment 2* (*few shot case*) and comparisons made with a direct transfer learning baseline on performances over incremental classes and retention of accuracy on old tasks (6 classes of VW, Valves, PA, Aorta, LV-Ao, RA). 'Overall drop' in accuracy is a proxy for knowledge retention

Table 3: *Network 2* performance on *experiment 2* (%). invariant accuracies in green.

Stage	4C Structures *(Sub-model A)*			3V Structures *(Sub-model B)*			LVOT Structures *(Sub-model C)*			Average
	VW	Valves	FO	PA	Aorta	SVC	LV-Ao	RA	RV	**(old tasks)**
No Increment	93.72	88.30	--	69.55	62.42	--	75.50	66.10	--	75.93
FO increment	91.34	86.54	69.25	69.55	62.42	--	75.50	66.10	--	75.24
SVC increment	91.34	86.54	69.25	68.30	56.90	71.33	75.50	66.10	--	74.11
RV increment	91.34	86.54	69.25	68.30	56.90	71.33	73.63	61.84	61.32	73.09
All increment	91.34	86.54	69.25	68.30	56.90	71.33	73.63	61.84	61.32	73.09
Overall drop	2.38	1.76	0.00	1.25	5.52	0.00	1.87	4.26	0.00	**1.89**

Table 4: Transfer learning (*Baseline*) for Experiment 2 data distribution accuracies (%) catastrophic forgetting compounded by class imbalance – worse new addition accuracies compared to similar stages in table 2

Stage	4C Structures			3V Structures			LVOT Structures			Average
	VW	Valves	FO	PA	Aorta	SVC	LV-Ao	RA	RV	**(old tasks)**
Pre-Transfer	*94.15*	*88.24*	--	*71.34*	*70.10*	--	*72.95*	*63.83*	--	76.77
FO increment	*88.64*	*83.05*	*41.05*	*65.36*	*64.34*	--	*67.82*	*60.91*	--	*71.69*
SVC increment	*86.19*	*82.16*	*40.90*	*63.26*	*61.86*	*40.11*	*66.14*	*57.62*	--	*69.54*
RV increment	*83.85*	*82.11*	*40.16*	*60.10*	*57.16*	*38.26*	*63.89*	*53.65*	*37.90*	*66.79*
Overall drop	*10.30*	*6.13*	*0.89*	*11.24*	*12.94*	*1.85*	*9.06*	*10.18*	--	**6.95**

Table 5: Network 1 used on Experiment 2 Accuracies (%). invariant accuracies in green. poor adaptation of imbalanced new classes in the softmax regime evident despite limited interference on prior learning

Stage	4C Structures *(Sub-model A)*			3V Structures *(Sub-model B)*			LVOT Structures *(Sub-model C)*			Average
	VW	Valves	FO	PA	Aorta	SVC	LV-Ao	RA	RV	**(old tasks)**
Pre-Transfer	91.25	85.40	--	68.35	60.23	--	74.05	60.50	--	73.30
FO Transfer	88.10	81.19	40.10	68.35	60.23	--	74.05	60.50	--	72.07
SVC Transfer	88.10	81.19	40.10	60.25	57.10	43.15	74.05	60.50	--	70.20
RV Transfer	88.10	81.19	40.10	60.25	57.10	43.15	70.15	56.90	40.65	69.45
Post-Transfer	88.10	81.19	40.10	60.25	57.10	43.15	70.15	56.90	40.65	69.45
Overall drop	3.15	4.21	0.00	8.10	3.13	0.00	3.90	3.60	0.00	2.90

sourced from 10 subjects are for training and those from other 2 are for test) and 3 sub-models are trained, one for a coarse class. For *Network 1*, inputs to the sub-models are of the same batch as that of the corresponding super-model as the average softmax probability of a batch is used to select right sub-models for fine classification rather than individual instance predictions. This is implemented as a conditional calling of pre-trained versions of relevant fine classifiers with the same batch for incremental training. As the entirety of a new set is assumed to be of the same class sharing coarse details, a single sub-model is adapted at any incremental stage.

For *Network 2*, inputs are randomly prepared in pairs of same or different classes with labels of 0 or 1, so as to use a binary classification (equal numbers of positive and negative pairs). Encodings are generated using shared weights of a Siamese network, and used to mutually optimize the embedding output normalized between 0 and 1. Here, the training creates models able to find similar or dissimilar images, approximating a 'classification by similarity' task analogous to hard attention based discriminators in [19] with a distinction that a segregated fine-tuning is applied here to establish a footprint of novel data over time. A batch size of 50 is used with a learning rate of 0.0001. A low learning rate helps prevent drastic modifications in the parameter space during fine-tuning.

B. *Incremental training*

We perform incremental learning on sub-classes left out in early training. First, anatomical structures being incrementally learned are passed through the super-model to decide on coarse classification based on softmax probabilities obtained for three coarse classes. The case of incremental learning analyzed here is when the input instance belongs to a coarse class but not an existing fine-class. This is tested with foramen ovale (FO) examples from 4C classes, superior vena cava (SVC) from 3V cases and right ventricles (RV) from LVOT. Exemplar videos for each are fed to the super-model and coarse probabilities computed. Based on the coarse probability, relevant sub-models are adapted with instances from new fine classes in 100 epochs with a learning rate 0.0001. The number of samples is equal to the total dataset size in subject wise train-test split for the incremental learning part of *Experiment 1* and is restricted to 500 samples in *Experiment 2*. We use these 500 samples to generate randomized pairs with same new instances for positive pairs, and negative pairs using pair images from other classes to simulate few-shot increment steps for *Experiment 2*. In both cases, the train:test split, learning rate schedules and batch size are kept the same as in the early training stages.

Note that in the case of *Experiment 2*, the sub-models are implemented as *N2* modules that are essentially networks trained to compare feature representations derived from input video frames. This fundamentally implies that for novel classes, instances can be compared for similarity as long as the classes themselves have an inherent association to the ones used for the first stage training. Thus, novel class comparisons can theoretically be accomplished indefinitely without further incremental training given an oracle that collects and labels the data distribution after its new instances are deemed mutually

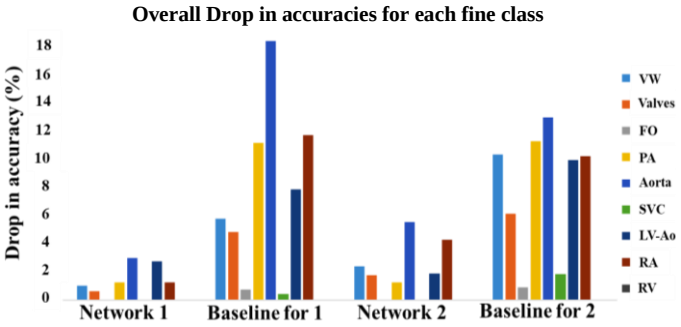

Fig. 8. Catastrophic forgetting quantified by Overall Drop in class accuracies for old classes. Network 1 and 2 outperform baselines across classes.

similar by the early-trained model. A caveat is that the reliance of discriminative power on initial training alone results in diminishing accuracy with diversity of incremental stages. To enable suitable retention of discriminative ability over diverse temporal data, we improve upon standard few-shot learning [19] to fine-tune sub-models on incremental class data. The specific fine-tuning procedure applied for both classes of sub-models in both experiments is different to standard fine-tuning, as there is an enforced segregation on the sub-models actually fine-tuned and the refinement is governed by the coarse classification unlike in standard fine-tuning or transfer learning.

C. *Evaluation protocols*

For evaluation of incremental learning tasks, we consider the retention of old knowledge and ability to generalize to new tasks. The former is defined by the drop in accuracy over validations on old classes during new data addition ('Overall Drop' in Tables and Fig. 8). The second is evident from the validation accuracy on each new class added ('New Task Adaptation' in Fig. 9). At increments, holdout data for each class in Dataset A is used to test the pipeline for accuracies on every class in our experiments. The net decrease is the 'Overall drop', which is the difference in the validation data accuracy before new class addition and that after all new data is trained for. Averaged over fine classes, the capacity of the architecture to mitigate forgetting is found as the Average Overall Drop (this is computed over the total number of classes, old and new, as new class accuracies are affected in subsequent incremental stages in baselines). New Task Adaptation has a reliable proxy in the validation accuracy on new classes after all of them are added. As the trained model is used to make predictions on unseen validation data of new tasks, it shows forward transfer of knowledge without impact on parameters during training.

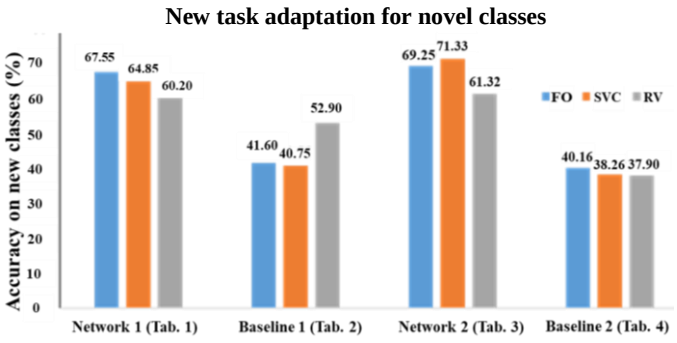

Fig. 9. Ability to adapt to new information. Final accuracy over new classes is a proxy for New Task Adaptation. Network 2 generalizes best overall.

D. *Incremental computational and memory overheads*

The sub-models are initialized for different coarse classes. Thus, the memory requirements are conditional on the number of coarse groups studied. In clinical settings, top-level grouping is often defined and memory budgets can be allocated at the beginning. Arrival of new fine classes likely to share coarse characteristics with a pre-drawn set of top-level classes would thus not lead to increased memory usage during incremental training (though an average softmax probability cut-off at the coarse classification allows creating new sub-models in our pipeline, we do not have a use case relevant to that here). As we start with pre-trained models on old tasks, computational costs on new additions derive only from the fine-tuning on such data. It is a minimal addition on each increment as full training on new classes (or old plus new classes) is not performed.

VI. Results

In both experiments, accuracies are obtained after a 4-fold cross validation using videos originally traceable to separate subjects. This is a significantly different task from a simple viewing plane classification since the available videos are of structures magnified in isolation without the global context of full frame videos. As such, this is a challenging task since several structures are visually not very distinct on ultrasound scans, leading to a certain degree of confusion without the view plane label particularly in LVOT and 3V structures. Distinction between many anatomical structures, often visible as irregularly shaped black patches, is difficult even for human experts. In addition, a fusion of frames of a clip in first layers of sub-models causes an overlap of sequential spatial information, complicating the learning of discriminative spatiotemporal features. In Experiment 1, accuracies with new classes are diminished with direct transfer learning compared to Network 1– Foramen Ovale (67.55% vs 41.60%), Superior Vena Cava (64.85% vs 40.75%), Right Ventricle (60.20% vs 52.90%). In Experiment 2, baseline accuracies for the newly added classes of Foramen Ovale (FO - 41.05%), Superior Vena Cava (SVC - 40.11%) and the Right Ventricle (RV - 37.90%) are much lower than the corresponding values obtained with our hierarchical models in Network 2 (FO - 69.25%, SVC - 71.33%, RV - 61.32%). In comparison to our methods in both experiments, the accuracy for existing classes get modified with additions of new classes to a greater degree in direct transfer learning methods (Tables 2 and 4). 'Overall Drop' quantifies the presence of catastrophic forgetting in both cases - averaged over all fine classes (Fig. 10a). Values for Experiment 1 are 1.10 % vs 6.75 % using the hierarchical pipeline of Network 1 versus using the baseline and values in Experiment 2 are 1.89% vs 6.95% in between the modular Network 2 vs the baselines. The improvement in knowledge retention for several difficult classes is starker – the overall drop in accuracy (Fig. 8) reduces from 11.24% to 1.25% for the PA (Pulmonary artery) additions, from 12.94% to 5.52% for Aorta and from 10.30 to 2.38% for VW (Ventricle Walls) in Experiment 2. Classwise comparisons of the drop in accuracy for the classes seen in initial training are drawn between the methods proposed by Networks 1 and 2, with other incremental learning approaches that have achieved state-of-the-art results in computer vision literature in Fig. 11, with the least accuracy drop for a majority of the old classes. The methods proposed through Networks 1 and 2 are found to exhibit superior incremental learning performance compared to proposed methods in literature across classes, in terms of average overall drop in accuracy (Fig. 12) as they demonstrate the lowest average reduction in accuracy after incremental addition of new classes, compared to other methods considered.

VII. Discussion

Incremental Performance. The progression of learning with the arrival of sparse data from new fine classes is studied for both Networks 1 and 2 and presented in Table 1 and Table 3, with comparisons drawn with a baseline learning applied with and without the novel classes at each incremental stage. |It is evident that in the hierarchical scheme, the effect of such inclusion is localized to the sub-models called with respect to the predicted coarse class and the other fine classification submodels remain unaffected, as they do not participate in the adaptation to such new data. This can be seen as the accuracies stay constant with respect to the 'No Increment' stages unless new data is added within the coarse class (FO increment in 4C structures, Sub-model A; SVC increment in 3V structures, Submodel B; RV increment in LVOT structures, Sub-model C in both modes of hierarchical incremental learning performed in Network 1 and Network 2). Including examples from a novel class and training the relevant sub-model is seen to slightly reduce prediction accuracies for structures the sub-model was previously trained for. This is because the weights adapted for fine classification may not be best suited for predictions on

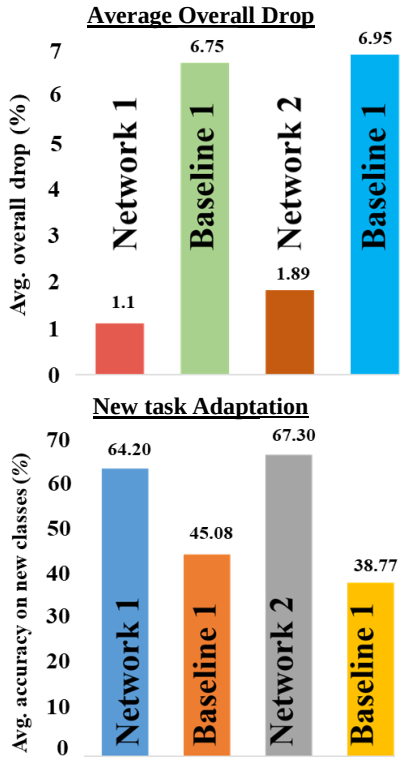

Fig. 10 (a) & (b). Summary of model evaluation - Average overall drop in performance on old classes, and new task accuracies over new classes. *Network 1* shows slightly less forgetting but *Network 2* performs better on new tasks.

previously learned structures but the values predicted are still close to original as compared to the corresponding steps in standard transfer learning baselines. Further, the incorporation of new sub-classes relevant to a single sub-model does not influence the fine-classification performance for the other two sub-models as shown in the prior fine accuracy figures remaining constant (shown in green in the tables) over these stages of class additions compared to other sub-models. The progression of novel data addition in the proposed incremental few-shot scheme is compared to a standard coarse-to-fine model using a standard transfer learning/fine-tuning baseline, which has been a method of incorporating the ability to expand to new tasks in a variety of statistical learning approaches. It is seen that such a transfer learning model adapts poorly to an incorporation of novel classes (Tables 2, 4). This is reflected in the depreciation of accuracies of old fine classes in each coarse bin and a significantly low new task adaptation for novel data.

Forgetting and retention. In Experiment 1, accuracies with new classes are found to be reduced with direct transfer learning as compared to the staged learning approach in Network 1. In Experiment 2, baseline accuracies for the newly added classes of Foramen Ovale, Superior Vena Cava and the Right Ventricle are significantly lower than the corresponding values for the hierarchical models in Network 2. The improvement in knowledge retention for several difficult classes is much more pronounced in this case due to the division of the parameter space utilized for fine grained recognition in the Network 2 modules, for the different sub-anatomies. Such improvements are noted in Experiment 1 too. This is due to drastic weight modifications in baseline architectures, explained using the well-studied limitations of connectionist networks and transfer learning [2], [24] about trainable parameters being successively overwritten in these models. Our models adapt to novel classes better than baselines, quantified by New Task Adaptation term for validation accuracy on new class data after increments, in average final accuracy (Fig. 9) and model averages (Fig. 10b).

Effects of Imbalance. To quantify effects of incremental imbalance, a comparison is also made by using Network 1 on the data of Experiment 2 (Table 5). The initial accuracies before new class additions are at par with values in Table 1 (which is obvious as at this stage the classification tasks are same). The subsequent divergence in accuracies is rapid for new classes (69.25% vs 40.10% for FO, 71.33% vs 43.15% for SVC, 61.32% vs 40.65% for RV), indicating certain misclassification despite modularization. The imbalance during incremental

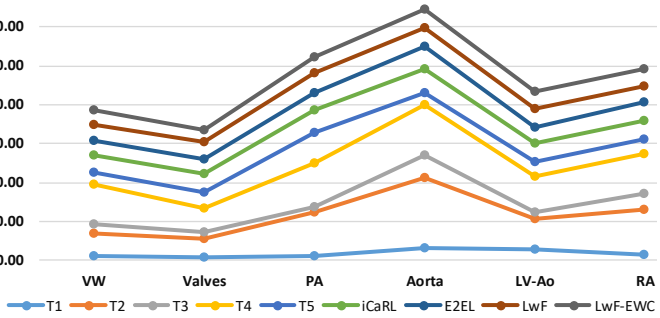

Fig. 11. Drop in accuracy for classes seen in initial training – N1 and N2, with transfer learning (TL) baselines compared to other methods in literature.

finetuning is thus seen to fare poorly in softmax classification settings. Contrarily, similarity methods are explicitly trained on distances among instances as the basis for categorization. This enables a direct extension to small numbers of novel instances, evident in better performance in comparisons on corresponding stages in Table 3 and Table 5. It is to be noted that while methods of over/down sampling and weight loss adaptation are suitable methods for dealing with data imbalance between classes when one has a setting that performs a classification task between multiple such classes in the same learning session. Our setting for incrementally learning unbalanced data distributions is fundamentally different. In the incremental learning session, these classes are observed in different learning sessions – the unbalanced class may be seen in a learning stage when the previous classes are not present. The imbalanced classes that are incrementally added are learnt using a Siamese network with a distance metric based approach, rather than a standard classification task where the feature representations for different classes are segregated in a low-dimensional space, and thus an unbalanced set of classes can be magnified by oversampling them. In our case, while considering the unbalanced classes, the features for exemplars are obtained in the form of embeddings through the existing Siamese N2 network strands, and are compared with existing class embeddings. In this case, an under-sampling of classes previously learnt is of little use as the incremental setting is meant to assume the non-availability of significant amounts of previously available data. On the other hand, such an under-sampling, if carried out during initial stage training, undermines performance and such a possibility for under-sampling during initial training can not be guaranteed during incremental learning scenarios in practice [7], [25].

VIII. Conclusion

We proposed experimental regimes with novel evaluation criteria for designing and assessing deep learning pipelines for continual learning on extremes of incremental data distributions in clinical settings using the task of fetal echocardiographic analysis as an exemplar of approach. Essentially, our hypothesis of observing forgetting was established through classwise results on baselines, and the various strategies for mitigation were explored, with two distribution scenarios representing possible extremities in real-world fetal echocardiographic analysis. Even in cases of arrival of imbalanced datasets over extended time, our hierarchical classification protocols could be adapted at the sub-model level to accomplish optimal incorporation of new information using similarity learning based architectures that retain competitive past performances in a real clinical dataset. The key hypothesis that observations may be missed in initial scans, with later examinations filling in diagnostic gaps, is extendable to multiple clinical imaging tasks. Thus, it is pertinent to attempt our lifelong learning method in application areas where developing abnormalities rely on imaging data. Future work may concern an extension to the inclusion of anomalies in the echocardiography scans being dynamically introduced as novel classes that our pipelines are suitable for adaptation, thereby

leading to avenues of continually learning imaging priors for physiological anomalies such as congenital heart disease and other conditions evident in the fetal anomaly scans.

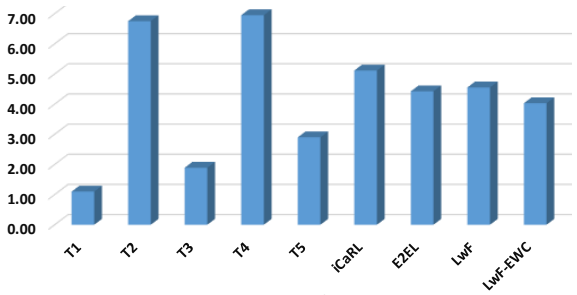

Fig.12. Average overall drop in accuracy for classes in initial training – N1, N2, with the transfer (TL) baselines compared to other methods in literature.

Acknowledgments

The authors acknowledge EPSRC grant EP/M013774/1 (Seebibyte) and ERC Advanced Grant 694581 (PULSE). AP gratefully acknowledges the Rhodes Trust.

REFERENCES

[1] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and Autonomous Systems*, vol. 15, no. 1-2, pp. 25-46, Jul, 1995.

[2] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks," [online] Arxiv.org. Available at: https://arxiv.org/abs/1312.6211, 2015.

[3] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, 2014, pp. 3320-3328.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *IEEE International Conference on Machine Learning*, 2014, pp. 647-655.

[5] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *International Conference on Machine Learning Deep Learning Workshop*. vol. 2, 2015.

[6] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken and C. Sánchez, "A survey on deep learning in medical image analysis", Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[7] Z. Li and D. Hoiem, "Learning without Forgetting", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2017.

[8] T. Xiao, J. Zhang, K. Yang, Y. Peng and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale

[9] image classification," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 177-186, ACM, 2014.

[10] A. Patra, W. Huang, and J. A. Noble, "Learning Spatio-Temporal Aggregation for Fetal Heart Analysis in Ultrasound Video,", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 276-284, Springer, Cham, 2017.

[11] B. Giritharan, X. Yuan and J. Liu, "Incremental classification learning for anomaly detection in medical images," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 7260, pp. 72603W, 2009.

[12] R. Aljundi, P. Chakravarty, P., and T. Tuytelaars, "Expert Gate: Lifelong Learning with a Network of Experts", In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7120-7129.

[13] S. A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5533-5542.

[14] Y. X. Wang, D. Ramanan, and M. Hebert,"Growing a brain: Fine-tuning by increasing model capacity", in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp.2471-2480.

[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.

[16] J. Lee, J. Yun, S. Hwang, and E. Yang, "Lifelong Learning with Dynamically Expandable Networks", in *International Conference on Learning Representations*, 2018.

[17] S. J. Pan, and Y. Qiang, "A survey on transfer learning", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no.10, pp. 1345-1359, 2010.

[18] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C.B. Kendall, M. B. Gotway and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," in IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1299-1312, May 2016..

[19] J. Weston, F. Ratle, H. Mobahi and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural Networks: Tricks of the Trade*, Berlin, Heidelberg: Springer, 2012, pp. 639-655.

[20] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning", in *Advances in Neural Information Processing Systems*, pp. 3630-3638, 2016.

[21] W. Huang, C. P. Bridge, J. A. Noble and A. Zisserman, "Temporal HeartNet: Towards human-level automatic analysis of fetal cardiac screening video." *International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI)*. Springer, Cham, 2017, pp. 341-349.

[22] J. Carvalho, L. Allan, R. Chaoui, J. Copel, G. DeVore, K. Hecher, W. Lee, H. Munoz, D. Paladini, B. Tutschek and S. Yagel, "ISUOG Practice Guidelines: sonographic screening examination of the fetal heart", Ultrasound in Obstetrics & Gynecology, vol. 41, no. 3, pp. 348-359, 2013.

[23] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska and D. Hassabis," Overcoming catastrophic forgetting in neural networks," in *Proc. Nat. Acad. Sci.* ,vol. 14, Mar 2017.

[24] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence", in *IEEE International Conference on Machine Learning*, 2017, pp. 3987–3995.

[25] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan and S. Wermter, "Continual Lifelong Learning with Neural Networks: A Review", Arxiv.org, 2018. [Online]. Available: https://arxiv.org/abs/1802.07569.

[26] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu and R. Hadsell, "Progressive neural networks", Arxiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1606.04671.

[27] A. Patra and J. A. Noble, 'Multi-anatomy localization in fetal echocardiography videos', in *IEEE 16th International Symposium on Biomedical Imaging*, 2019, pp. 1761-1764.

[28] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L.M. Koch and D. Rueckert, 'SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound', IEEE transactions on medical imaging, 36(11), 2204-2215.

[29] A. Patra, Y. Cai, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorghiou and J. A. Noble, 'Efficient Ultrasound Image Analysis Models with Sonographer Gaze Assisted Distillation', In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2019 (pp. 394-402). Springer, Cham.

[30] X. Glorot and Y. Bengio,'Understanding the difficulty of training deep feedforward neural networks', In Proceedings of the thirteenth Int. Conf. on Artificial Intelligence and Statistics, 2010. (pp. 249-256).

[31] H. A. Omar, J. S. Domingos, A. Patra, R. Upton, P. Leeson and J. A. Noble, 'Quantification of cardiac bull's-eye map based on principal strain analysis for myocardial wall motion assessment in stress echocardiography', in *IEEE 15th International Symposium on Biomedical Imaging*, 2018 pp. 1195-1198.

[32] H. A. Omar, J. S. Domingos, A. Patra, R. Upton, P. Leeson and J. A. Noble, 'Improving Visual Detection of Wall Motion Abnormality with Echocardiographic Image Enhancing Methods', in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018 pp. 1128-1131.

[33] H. A. Omar, A. Patra, J. S. Domingos, R. Upton, P. Leeson and J. A. Noble, 'Automated myocardial wall motion classification using handcrafted features vs a deep cnn-based mapping', in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018 pp. 3140-3143.

[34] A. Patra, J.A. Noble, 'Incremental Learning of Fetal Heart Anatomies Using Interpretable Saliency Maps', in *Medical Image Understanding and Analysis, 2019*.

[35] H. A. Omar, A. Patra, J. S. Domingos, R. Upton, P. Leeson and J. A. Noble, 'Myocardial wall motion assessment in stress echocardiography by quantification of principal strain bulls eye maps: P299', European Heart Journal Cardiovascular Imaging, 2017.