A New Semi-supervised Learning Benchmark for Classifying View and Diagnosing Aortic Stenosis from Echocardiograms

Zhe Huang[1], **M.S.**   ZHE.HUANG@TUFTS.EDU

Gary Long[2], **M.S.**   GARY@CVAI.SOLUTIONS

Benjamin Wessler[3], **M.D.**   BWESSLER@TUFTSMEDICALCENTER.ORG

Michael C. Hughes[1], **Ph.D.**   MICHAEL.HUGHES@TUFTS.EDU

[1] *Dept. of Computer Science, Tufts University, Medford, MA, USA*

[2] *CVAI Solutions, Dorchester, MA*

[3] *Division of Cardiology, Tufts Medical Center, Boston, MA*

Abstract

Semi-supervised image classification has shown substantial progress in learning from limited labeled data, but recent advances remain largely untested for clinical applications. Motivated by the urgent need to improve timely diagnosis of life-threatening heart conditions, especially aortic stenosis, we develop a benchmark dataset to assess semi-supervised approaches to two tasks relevant to cardiac ultrasound (echocardiogram) interpretation: view classification and disease severity classification. We find that a state-of-the-art method called MixMatch achieves promising gains in heldout accuracy on both tasks, learning from a large volume of truly unlabeled images as well as a labeled set collected at great expense to achieve better performance than is possible with the labeled set alone. We further pursue *patient-level* diagnosis prediction, which requires aggregating across hundreds of images of diverse view types, most of which are irrelevant, to make a coherent prediction. The best patient-level performance is achieved by new methods that *prioritize* diagnosis predictions from images that are predicted to be clinically-relevant views and *transfer* knowledge from the view task to the diagnosis task. We hope our released dataset[1] and evaluation framework[2] inspire further improvements in multi-task semi-supervised learning for clinical applications.

1. Introduction

Our motivating task is to improve timely diagnosis and treatment of aortic stenosis (AS), a common cardiac valve condition. If left untreated, severe AS has lower 5-year survival rates than several metastatic cancers (Howlader et al., 2020; Clark et al., 2012). With timely diagnosis and surgical or transcatheter aortic valve replacement, AS becomes a treatable condition with very low mortality (Lancellotti et al., 2018). Unfortunately, in current practice up to 2/3 of symptomatic AS patients may never get referred for care (Tang et al., 2018; Brennan et al., 2019). There is an urgent need in improve timely detection of this life-threatening condition. In this study, we develop and validate machine learning methods for automating the preliminary interpretation of cardiac ultrasound (echocardiogram) images,

---

1. Tufts Medical Echocardiogram Dataset (TMED): https://TMED.cs.tufts.edu
2. Open-source code: https://github.com/tufts-ml/ssl-for-echocardiograms

with the goal of expanding access to rapid and accurate diagnosis of AS while overcoming constraints on the availability of labeled data needed to train methods effectively.

Recent advances in computer vision and machine learning have made it possible to *automate* the way medical images are turned into actionable knowledge for diagnosing and treating disease across cardiology (Chen et al., 2020a), radiology, and other areas of medicine (Shen et al., 2017). However, in order to work well, modern deep learning methods require *large* amounts of *labeled* training examples, where each labeled example consists of an image and its desired class label. While images themselves are routinely collected and easily available in electronic health records, obtaining *labels* for images often requires manual effort from a clinical expert. Thus, a key barrier to deploying deep learning image classifiers for specialty areas of medicine is that it is prohibitively *difficult* and *expensive* to acquire a large labeled dataset whose scale matches the tens of thousands of labeled examples available in common non-medical benchmarks such as CIFAR-10 (Krizhevsky, 2009) or Street View Housing Numbers (SVHN, Netzer et al. (2011)). Privacy and regulatory concerns further inhibit sharing of labeled datasets even if they are collected within a single healthcare system.

A promising technology to overcome the need for abundant labeled data is semi-supervised learning (SSL) (Zhu, 2005; Chapelle et al., 2010; van Engelen and Hoos, 2020). SSL methods train classifiers simultaneously from two data sources: a *small* labeled set and a *large* unlabeled set. Recent semi-supervised deep learning methods have produced remarkable progress (Miyato et al., 2019; Berthelot et al., 2019a,b; Sohn et al., 2020; Xie et al., 2019; Chen et al., 2020b). On the standard SVHN image classification benchmark, a typical WideResNet achieves an error rate of 12.8% when trained using only a *small* labeled set of 1000 total labeled images (100 for each of 10 digit categories). In contrast, a recent SSL method called MixMatch (Berthelot et al., 2019b) can reduce error rate to 3.3% using the small labeled set plus 60,000 unlabeled examples, or even to 2.2% with 600,000 unlabeled examples. These improvements are comparable to training on 50x larger *fully-labeled* datasets, but avoid the time and expense of collecting so many labels (which for our intended applications are reliably obtained only from clinical experts). While SSL methods are promising, the application of modern SSL methods to real medical imaging tasks is largely *untested* and requires development of new methods to address issues such as class imbalance and the need to aggregate many image-specific predictions into coherent decisions for the whole patient.

In parallel to our work, several recent efforts do explore modern SSL methods to analyze medical images. Meng et al. (2020) proposed a SSL domain adaptation method for classifying views of fetal ultrasounds. Calderon-Ramirez et al. (2021) utilize MixMatch to detect COVID-19 based on chest X-ray images. Chen et al. (2021) leverage unlabeled data to improve vein segmentation performance using ultrasound images. Wang et al. (2021) propose an SSL method for classifying breast lesion and ophthalmic diseases. Our study contributes to this growing literature by applying SSL to improve patient-level screening in cardiology.

As a case study to test the promise of SSL methods, we develop SSL classifiers to diagnose aortic stenosis (AS). AS is diagnosed using ultrasound imaging of the heart, known as *echocardiography*. While ultrasound imaging itself is widely available and performed routinely for many patients for a variety of reasons, accurate interpretation of echocardiograms to make complex imaging diagnoses such as AS requires significant expertise that is not widely available. Diagnostic errors may contribute to treatment delays because assessment is challenging and requires integrating information across many hemodynamic parameters (Baumgartner
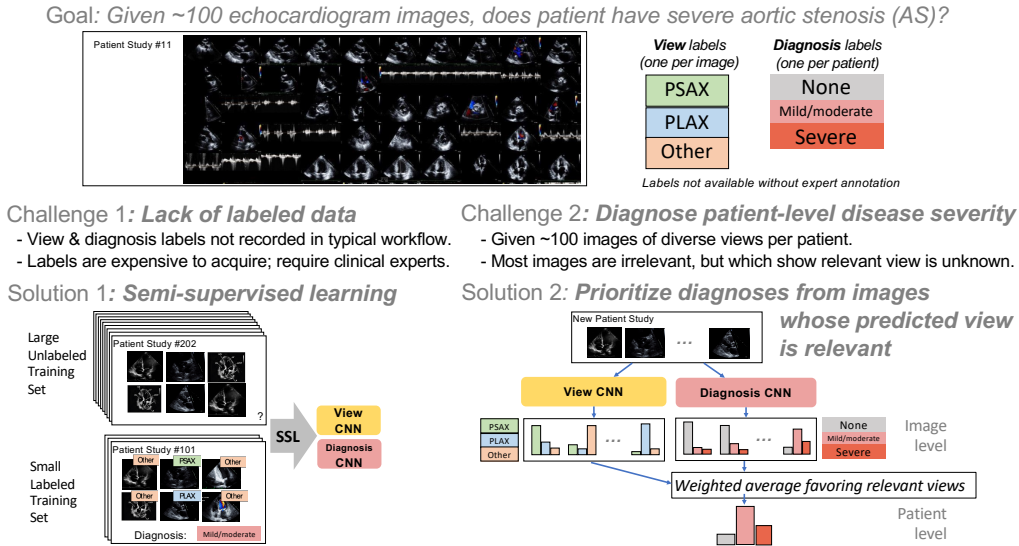
Figure 1: Illustration of our study's goal – automating diagnosis of the severity of *aortic stenosis* (AS) given hundreds of echocardiogram images collected in a typical exam – as well as the key technical challenges and proposed contributions that address these challenges.

et al., 2017), that are often discordant (Minners et al., 2008) and have low inter-reader reliability (Sacchi et al., 2018). Automated grading of AS has the potential to increase the accuracy and reproducibility of disease detection and reduce barriers to access (Batchelor et al., 2019), especially as a first-line screening tool in geographic areas without expert cardiologists. We believe that automated preliminary assessment of AS, with timely follow-up care by an expert clinical team, will improve patient outcomes by better identifying patients with this life-threatening condition that require treatment.

Contributions

In this work, we make the following contributions:

1. **New open-access SSL dataset – the Tufts Medical Echocardiogram Dataset (TMED) – to benchmark view and disease severity classification.** Our dataset is directly inspired by the need for automated preliminary assessment of aortic stenosis (AS). The labeled set of 260 patients contains an AS disease severity label for each patient as well as a view label for all images, all provided by expert clinicians. Furthermore, our dataset is designed to assess the true potential of semi-supervised learning, because in addition to the labeled set it contains a large *unlabeled* set from 2645 patients captured in the course of standard cardiac care. Common SSL benchmarks such as CIFAR-10 (Krizhevsky, 2009) or STL-10 (Coates et al., 2011) do not contain truly unlabeled data but instead artificially "forget" existing labels. This makes the unlabeled data used in these benchmarks unrealistically clean, class-balanced, and relevant to the task. We hope our new dataset will inspire work on effectively using minimally-curated unlabeled data to improve medical image understanding.

2. **Evaluation of SSL methodology to find what works best and why.** We carefully compare standardized implementations of several state-of-the-art SSL methods. We find that

MixMatch (Berthelot et al., 2019b) performs best on both view and diagnosis tasks, reliably beating labeled-set-only methods by over 2% balanced accuracy on view classification (see Table 4) and by over 3% balanced accuracy on patient-level diagnosis (see Table 7). These gains are also consistent for smaller versions of our dataset. Further ablation studies suggest that the surprisingly effective "mix-up" data augmentation strategy underlying MixMatch is a primary reason for success. In contrast, virtual adversarial training (VAT, Miyato et al. (2019)), which by its adversarial design might be expected to perform better in medical imaging domains, only marginally improves over labeled-set-only methods.

3. **New methods for patient-level severity diagnosis without manually preselecting relevant views.** A patient study may contain over 100 echocardiogram images from diverse view types (see Fig. 1). Only some of these images are relevant to diagnosing AS (only some views show the aortic valve that AS impacts). View type labels are not usually available. Previous work has relied on manual preselection of relevant views (Madani et al., 2018b; Ouyang et al., 2020). We develop methods that directly consume *all available images*, as would be needed in a fully automated deployment. Rather than simply average diagnostic predictions across all images, we *prioritize* diagnoses from images that our view classifiers suggest belong to the view types known to be relevant for AS. Using only the labeled set and trying to distinguish between 3 levels of AS severity (none, mild/moderate, and severe), prioritizing relevant views achieves patient-level balanced accuracy of 86.6% compared to 81.6% with simple averaging and 33% for random guessing. Our best SSL methods that prioritize relevant views achieve 87.9%, further improving to 90.1% when we *pretrain* a view classifier and use this to warm start our diagnosis classifier. These results suggest that fully-automated preliminary diagnosis of aortic stenosis may be soon realizable in practice.

Generalizable Insights about Machine Learning in the Context of Healthcare

The key barrier to applying deep learning to address many medical tasks is the lack of abundant labeled data. Recent SSL methods appear to reach competitive performance with only modest labeled datasets, but lack authentic evaluation using truly unlabeled clinical data exhibiting issues such as irrelevance to the task or class imbalance. Thus, existing benchmarks may be too "clean" and over-state the potential of SSL methods. As a remedy, this study develops clinically-motivated semi-supervised learning tasks and releases a dataset useful for measuring the gains possible with SSL methods. Next, we offer insight about *which SSL methods work best and why.* Our careful evaluation and ablation studies can help practitioners identify promising methods like MixMatch. Furthermore, our analysis encourages work to *look beyond simple averaging* when aggregating across multiple images to make a patient-level diagnosis. Overall, we hope our work helps unlock the potential of easily-accessible unlabeled data to improve patient outcomes.

2. Background and Related Work

2.1. Background: Single-task SSL with Neural Networks for Image Classification

We consider the problem of *semi-supervised* image classification. For training, we are given two datasets. First, a (small) *labeled* dataset $\mathcal{D}^L$ containing N^L pairs of images x and corresponding labels y. Second, a (large) *unlabeled* dataset $\mathcal{D}^U$ of N^U examples of images

only, presumed to be sampled from a similar distribution as the labeled set. Each image x is represented as a standard tensor of pixel intensity values (one entry for each pixel and each color channel). Each image-specific label $y \in \mathcal{C}$ indicates one of the possible classes in set $\mathcal{C}$.

Given a deep neural network parameterized by weight vector θ, a standard SSL training procedure tries to find an (approximate) solution to the following loss minimization problem:

$$\theta^* = \arg\min_{\theta} \frac{1}{N^L} \sum_{x,y \in \mathcal{D}^L} \mathcal{L}^L(y, x, \theta) + \lambda \frac{1}{N^U} \sum_{x \in \mathcal{D}^U} \mathcal{L}^U(x, \theta) \tag{1}$$

Here, $\mathcal{L}^L$ indicates a *labeled loss* function such as cross-entropy, and $\mathcal{L}^U$ indicates an *unlabeled loss*. We show how each SSL method we explore instantiates this framework in Sec. 4.1.

For our specific application, we separately train neural networks to perform two tasks. First, for *view* classification we wish to map each image x to a real-valued score for each of the possible view classes in set $\mathcal{C}_V$. We use a network f parameterized by weights θ_V. To obtain probability distributions over the classes, we use the softmax transformation, denoted as $S(\cdot)$, which produces a probability vector with same size as the given input vector. Thus, the probability that an image's view label indicates the c-th element of $\mathcal{C}_V$ is then $S(f_{\theta_V}(x))_c$.

Second, for *diagnosis* classification we wish to map each image x to a real-valued vector containing scores for each of 3 posssible diagnoses (no AS, mild/moderate AS, severe AS) in set $\mathcal{C}_D$. We use network g with weights θ_D. We always use the same architecture as view network f. The probability of assigning the c-th label in $\mathcal{C}_D$ to image x is $S(g_{\theta_D}(x))_c$.

2.2. Semi-supervised learning: Datasets and evaluation

Our work builds upon a previous study by Oliver et al. (2018) on best practices for SSL evaluation. That work emphasizes the need to compare SSL methods that claim advantages from unlabeled data to strong baselines that only use the labeled set. We follow this best practice in our work. Their independent evaluation suggests that, if trained properly, modern deep SSL methods can leverage large unlabeled sets to obtain meaningful improvements.

However, a key limitation of existing SSL research is that popular evaluations, even those used in Oliver et al. (2018), are confined to well-worn non-medical datasets, such as SVHN (Netzer et al., 2011) or CIFAR-10 (Krizhevsky, 2009). These datasets consist of carefully curated *balanced* class distributions, where the so-called "unlabeled" set is obtained by forgetting known labels. Performance numbers are likely too optimistic compared to using truly unlabeled images. As performance saturates on these benchmarks, SSL research needs to move towards datasets like ours featuring truly unlabeled images.

2.3. Related Work: Supervised learning for cardiology

Deep learning research efforts for cardiac imaging applications are plentiful (Chen et al., 2020a). For echocardiogram analysis, the task of *view* classification has been pursued by several efforts (Madani et al., 2018a; Zhang et al., 2018; Long and Wessler, 2018). Accuracies above 90% can be achieved given a *large labeled* dataset of echocardiogram images and corresponding views. Training CNN view classifiers on 200,000 images from 240 patients, Madani et al. (2018a) report accuracy of 91.7% on 15 view types using low-resolution images, exceeding the performance of board-certified cardiographers. Zhang et al. (2018) also report promising performance for CNNs to classify 23 different view types (e.g. 96% accuracy

for the PLAX view). However, we emphasize that these efforts use *proprietary datasets* that other researchers cannot build upon and extend. They also require *large* labeled sets and do not address the key motivation of our work, the downstream diagnosis of *aortic stenosis*.

Recently, researchers at Stanford developed an "EchoNet" deep learning methodology for echocardiography (Ghorbani et al., 2020), predicting measurements related to ejection fractions as well as some binary decisions such as "does the patient have a pace-maker?" or "is there left ventricular hypertrophy?" Ghorbani et al. (2020) produce *patient-level* decisions just by simple averaging, reporting the surprising conclusion that "alternative methods were explored in order to aggregate frame-level predictions into one patient-level prediction and did not yield better results compared to simple averaging." Our later results suggest that smarter aggregation *does* produce notable benefits.

The same team later produced "EchoNet Dynamic" methodology to learn from videos (not static images) (Ouyang et al., 2020). releasing 10,030 videos of echocardiogram imagery for one particular view (apical-4 chamber) and associated measurement and diagnostic labels for thousands of subjects. This is a welcome step forward, but this dataset pursues different goals than ours: there is no focus on semi-supervised learning and this data focuses on only one view type (apical 4 chamber or A4C) out of the dozens of possible views that make up a complete study. While this A4C view is relevant to other measurement and diagnostic tasks in cardiology, it is not helpful for assessing valvular heart diseases generally nor AS in particular.

Another public echocardiogram dataset is the CAMUS dataset (Leclerc et al., 2019). CAMUS' purpose is to evaluate *segmentation* methods. The data contains images of apical 2 chamber (A2C) and apical 4 chamber (A4C) views from 500 patients, together with detailed annotations of anatomical structures (e.g. the myocardium and the left atrium). Like other prior work, these views are not relevant for the diagnosis of aortic valve disease.

Several efforts have looked at diagnosing *aortic stenosis* as the target outcome of a machine learning classifier, though none we are aware of use echocardiograms. Yang et al. (2020) used wearable sensors to build binary AS classifiers using data from 34 total subjects, reporting 96% accuracy with random forests. Kwon et al. (2020) used *electro*cardiogram (ECG) signals to build binary AS classifiers from over 30,000 patients at 2 hospitals in Korea. Their best 12-lead models achieve 0.861 AUROC on external validation. Hata et al. (2020) used ECGs from 700 patients in Japan. Their 12-lead models achieve 84.2% precision and 72.7% recall on a heldout set. In clinical practice, echocardiograms (ultrasound images) are the primary imaging modality used to assess the aortic valve. To the best of our knowledge, there does not seem to be previous work on detecting aortic stenosis (AS) from echocardiograms, likely because echocardiograms contain complex data (video clips and Doppler recordings) that are not routinely annotated as part of clinical care. Acquiring labels for these images is prohibitively time consuming and expensive. Among early efforts to automate AS screening, our study stands out for its focus on ultrasound (the clinical standard for diagnosing AS) and for showing how *semi-supervised learning* can overcome limited labeled data.
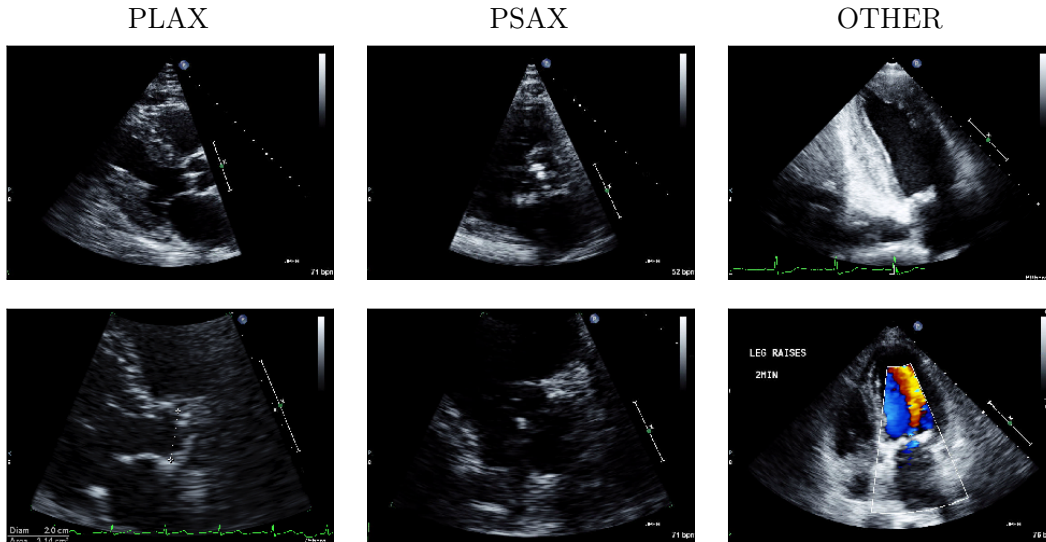
Figure 2: Examples of images with their view label in our dataset. *From left to right:* Two examples of parasternal long axis (PLAX) view, two examples of parasternal short axis (PSAX) view, and two examples of other kinds of view in our "Other" class. (Note: Images shown here are the original high-resolution images. When training the model, we further downsized the images to 64 x 64, see sec 3.2)

2.4. Related Work: Semi-supervised learning for echocardiography

Work on *semi-supervised* learning for echocardiography is still in early stages. Previously, Madani et al. (2018b) pursued semi-supervised classification of echocardiograms using *generative adversarial networks* (GANs) (Goodfellow et al., 2014), including a 15-way view classification task using 267 patients as well as a diagnostic task for left ventricular hypertrophy (LVH). All 2269 images from 455 patients were *manually preselected* for a particular relevant view. Impressively, they report over 92% accuracy at the LVH diagnosis task using GANs; their view classifiers were similarly competitive.

Our work builds upon Madani et al. (2018b) in three key ways. First, we pursue diagnosis *without manually preselecting relevant views.* Our setting matches what is needed in a real deployment where view labels would not be available. Second, we develop methods to aggregate image-level predictions to produce *patient-level* decisions, showing we can do much better than simple averaging. Finally, we offer more streamlined and competitive methodology. Madani et al. (2018b) suggest *different* methods for semi-supervised and fully-supervised settings, using GANs for the former and convolutional neural nets (CNNs) for the later. Instead, modern SSL methods can coherently pursue the same task no matter how much labeled data we have. Our approach builds on well-established CNNs and extensions to wide residual architectures (Zagoruyko and Komodakis, 2017). We do not need the complexities of GANs or their well-known training difficulties (Metz et al., 2017; Arora and Zhang, 2017). Recent evaluations (Miyato et al., 2019, Table 4) suggest the methods we build upon reduce error rates by 50% over GANs on SSL benchmarks like SVHN.

7

3. Dataset

3.1. Dataset extraction and approval

The dataset used in this paper contains a total of 2905 echocardiogram studies[3]. We use transthoracic echocardiogram (TTE) imagery acquired in the course of routine care consistent with American Society of Echocardiography (ASE) guidelines (Mitchell et al., 2019). Each patient study contains multiple cineloop video clips of the heart depicting various anatomical views. To collect this imagery, a sonographer manipulates a handheld transducer over the patient's chest, manually choosing different acquisition angles in order to fully assess the heart's complex anatomy. For this study we focus on still images extracted from all available video clips. While pulsed-wave Doppler (PW), continuous-wave Doppler (CW), and m-mode recordings are also available, we leave these for future work.

The echocardiograms originate from the last 5 years of records at Tufts Medical Center, a high-volume tertiary care center in Boston, MA. Echocardiograms are generally performed to assess for structural heart disease. These studies are done for a variety of reasons, from evaluating symptoms (e.g. chest pain, shortness of breath), to caring for a patient experiencing a cardiac event (e.g. myocardial infarction or acute heart failure), to providing follow up care for a known condition (e.g. aortic stenosis or cardiomyopathy). Studies were sampled from archived image folders that were organized by month of acquisition.

The use of these *deidentified* images for research has been approved by our Institutional Review Board (Tufts IRB #MODCR-03-12678). All images were acquired from The Cardiovascular Imaging & Hemodynamic Laboratory, part of the Tufts CardioVascular Imaging Center. This lab is Intersocietal Accreditation Commission (IAC) accredited and performs roughly 10,000 ultrasound examinations per year using devices from major vendors (Philips®, Toshiba®, Siemens®). By using standardized image formats, the released data and subsequent ML methods are intended to be vendor-independent.

3.2. Image preprocessing.

Each study's raw imaging data contains multiple cineloop videos. Typically there are around 100 - 200 videos per study. From each cineloop file, we take one image to analyze. Clinical collaborators suggested that any single frame could be used; in practice we took the first frame of each video. The resulting data contains both color images and gray scale images with various resolutions. To prepare for neural network training, we convert each image to gray-scale, pad along its shorter axis to achieve a square aspect ratio, and then resize the image to 64x64 pixels. We filtered out Doppler recordings completely using aspect ratio since Doppler recordings have distinct aspect ratios. More data processing details are available in App. C.1.

3.3. Annotating each study with view and diagnostic labels

As part of routine clinical care, there are no annotations applied to individual cineloops or still images for either view or diagnosis when imaging is collected. Instead, images for a

---

3. Our released TMED dataset contains 2773 studies, each guaranteed to be from a distinct patient. The experiments here use a slightly larger dataset including multiple studies from the same patient. We have ensured each patient in the released data has exactly one study for simplicity.

given study are reviewed in aggregate by an echocardiography-board-certified cardiologist to create a summary report that is merged into the electronic medical record. This report contains diagnostic labels, including the presence or absence of AS and also the grade of AS, reliably collected for most patients. All complete study reports produced in routine care have an assigned grade of aortic stenosis (range: no AS to severe AS).

However, while the diagnostic AS grade is available for all studies for which an expert reader has prepared a summary report, as implemented in our institution's current system it requires substantial manual effort to extract this label from the report into a form amenable to machine learning. Furthermore, view labels are not available for any of the imagery we use. Below, we detail how we obtain suitable annotations for a subset of patient-studies, which we call the *labeled* set. For the remaining studies in the *unlabeled* set, we have only images: no view or diagnostic labels are easily available.

Diagnostic labels of AS disease severity. For this investigation, we were able to extract the AS grade from the relevant summary report in the EMR for a subset of all studies. We refer to this subset as the labeled set. Each labeled patient in our study thus has an ordinal class label indicating one of 3 possible levels of severity: "no AS", "mild/moderate AS" and "severe AS". These patient-level diagnostic labels were assigned in standard fashion during routine care, integrating information across all available views for a given patient by a cardiologist with specialty training in echocardiography.

We chose a 3-level granularity[4] for AS severity classification – no AS, mild/moderate AS, and severe AS – as a good balance between simplicity and clinical utility. In Supplement Fig. A.2 and Fig. A.3, we show example images from a patient with severe AS and a patient without AS; distinguishing these two categories is difficult to an untrained eye.

View labels. Because view labels were not available for any imagery, we undertook a significant post-hoc annotation effort. We used a novel view labeling tool that displays a grid of multiple study images and facilitates rapid expert annotation. A board certified cardiologist provided all view label annotations. Our expert annotator provided a view label for each image in our labeled set, selecting one of 3 possible labels: parasternal long axis (PLAX), parasternal short axis (PSAX) and a final category (Other) indicating all other possible view types that are not PLAX or PSAX. We chose to focus on PLAX and PSAX view labels because the aortic valve's structure and function is visible. PLAX and PSAX views are used in the routine clinical assessment of aortic valve disease. Fig. 2 shows examples of each view type; more samples can be found in Supplement Fig. A.1.

Summary of available labeled and unlabeled data. Out of all 2905 studies, 260 studies were assigned both view and diagnosis labels; 174 additional studies were assigned diagnosis but not view labels (while still difficult to automate in our current system, extracting a diagnosis severity label is easier than assigning a view label). The remaining 2471 studies are *truly unlabeled*, with neither diagnosis nor view annotations available.

This data is further processed into two versions for standardized evaluation of SSL classification. The full-size version – TMED-156-52 – is described in Sec. 3.4 and smaller version – TMED-18-18 – in Sec. 3.5.

---

4. The raw data contains 5-level labels (our "mild/moderate AS" category can be further distinguished into "mild", "mild to moderate" and "moderate"). We leave these finer-grained levels to future work.

3.4. Full dataset with authentic unlabeled examples: TMED-156-52.

The full-size dataset used in this investigation consists of a *labeled set* of all 260 fully-labeled patient studies (both view and diagnosis labels), as well as a much larger *unlabeled set*. We review the design of each labeled and unlabeled set below. All methods using the labeled set have access to diagnosis labels for each patient and view labels for each image in that set. No such labels are available in the unlabeled set.

Labeled train/valid/test sets. To evaluate the performance of classifiers on heldout data, we divide the labeled set of 260 patients using a 3:1:1 ratio into a labeled train set of 156 patients and evaluation sets (validation and test) of 52 patients each. We call our full-size dataset the TMED-156-52 dataset, so that the true number of patients used for training and evaluation is apparent. We repeat this partitioning 4 separate times, resulting in 4 independently-chosen train/valid/test sets, with summary counts in Table 2.

Unlabeled set. To build the full-size unlabeled set used to train semi-supervised methods, we combine the 2471 truly unlabeled patient-studies together with the 174 patient-studies that only have diagnosis labels (we discard any labels and treat them as unlabeled). This results in a combined full-size unlabeled set with data from 2645 total patients. We have $\sim 18x$ more unlabeled images than we have labeled training images.

Bonus heldout set for diagnosis. Because our available fully-labeled data is limited, to further assess diagnosis classifiers, we use the 174 studies with diagnosis labels as a *bonus heldout set*. This use lets us evaluate if the rankings of methods on the original labeled test sets (52 patients) are repeatable in the larger 174 patient corpus.

3.5. Smaller dataset: TMED-18-18.

Our full-size dataset described above contains labeled data from hundreds of patients. To simulate the practical scenario where we have access to only a few dozen labeled patient studies (e.g. in early prototyping of a medical imaging ML pipeline), we also perform experiments comparing SSL methods on a *smaller* version of our dataset, where both labeled and unlabeled sets are significantly smaller than the full-size data described above. Because it is easier to train methods on smaller datasets, this smaller version also allows us to evaluate many more methods on a fixed computational budget than our full-size dataset.

Labeled train/valid/test sets for smaller version. We select 54 patients to comprise the smaller labeled set, from the entire full labeled dataset of 260 patients. Within the selected labeled set, we do a 1:1:1 train/validation/test split, favoring larger heldout size ratios here than in the full-size dataset to be sure we can assess real differences between models. Thus, the *labeled training set* contains data from 18 patients and each *labeled heldout set* (validation and test) contains data from 18 patients (each patient's data is exclusively used for either training, validation, or test). We call our smaller dataset the TMED-18-18 dataset (again to signify that 18 patient studies are available for training, and 18 for evaluation). We repeat this partitioning 3 times, resulting in 3 independently-chosen train/valid/test partitions that balance the frequencies of each view and diagnostic label. Summaries of each label's frequency are shown in Table 1.

By Diagnosis	Num. Patients				Num. Images			
	Total	No AS	Mild/Mod.	Severe	Total	No AS	Mild/Mod.	Severe
Train (labeled)	18	6	6	6	1805, 1935	542, 635	634, 700	583, 647
Valid. (labeled)	18	6	6	6	1833, 2018	520, 663	586, 723	652, 704
Test (labeled)	18	6	6	6	1834, 1957	561 609	577, 668	691, 728
By View					Total	PLAX	PSAX	Other
Train (labeled)					1805, 1935	208, 222	58, 72	1520, 1650
Valid. (labeled)					1833, 2018	212, 221	70, 81	1542, 1716
Test (labeled)					1834, 1957	215, 233	66, 77	1549, 1674
Unlabeled Train	380				41183, 41428			

Table 1: **Smaller TMED-18-18 dataset** summary, showing patient (left) and image (right) counts by diagnosis (top) and view (bottom). We use 3 different train/valid./test splits of the labeled set, all with same patient counts by diagnosis. To count images, we report (min., max.) values across splits.

Diagnosis	Num. Patients				Num. Images			
	Total	No AS	Mild/Mod.	Severe	Total	No AS	Mild/Mod.	Severe
Train	156	49	47	60	16463, 16852	4681, 4801	5072, 5102	6589, 7076
Valid.	52	16	16	20	5470, 5617	1458, 1544	1769, 1859	2153, 2330
Test	52	16	16	20	5377, 5855	1542, 1636	1781, 1901	1931, 2443
View					Total	PLAX	PSAX	Other
Train					16463, 16852	1883, 1918	595, 672	13908, 14297
Valid.					5470, 5617	639, 652	184, 221	4644, 4749
Test					5377, 5855	624, 646	172, 212	4559, 5027
Unlabeled	2645				303498			

Table 2: **Full-size TMED-152-52 dataset** summary, showing patient (left) and image (right) counts by diagnosis (top) and view (bottom). We use 4 different train/valid/test splits of the labeled set, all with same patient counts by diagnosis. To count images, we report (min., max.) across splits.

Unlabeled set for smaller version.　To build the unlabeled set for TMED-18-18, we combine the remaining 206 patients from the full-size labeled set with the 174 patients that only have diagnosis labels. Together, these 380 patients form the unlabeled set; even though we technically have labels for these studies, they are not used at all in training or evaluation. In TMED-18-18, our unlabeled set has $\sim 21x$ more images than the labeled train set.

4. Methods

4.1. Methods for SSL Image Classification

Inspired by Oliver et al. (2018), we wish to carefully evaluate semi-supervised learning methods for image classification, focusing on a simple SSL baseline (Pseudo-Label) as well as two recent high-performing methods: virtual adversarial training (VAT, (Miyato et al., 2019)) and MixMatch (Berthelot et al., 2019b). Below we review the key ideas behind how each method learns from labeled and unlabeled data. All descriptions below use the neural network notation defined in Sec. 2.1. Hyperparameters and other settings are found in App. C.2.

Pseudo-Label.　The pseudo-label method (Lee, 2013) is a natural way to use unlabeled data to help train a neural network. At each minibatch of unlabeled data during stochastic

gradient descent, we use the existing classifier to make predictions, obtaining the *pseudo-label* $\hat{y}(x) \in \mathcal{C}$ indicating the most likely predicted class for image x. If the predicted probability of the most likely class is above a user-specified *confidence threshold* τ, we include this example in the loss, treating the pseudo-label as the true label. Thus, for our tasks the Pseudo-Label unlabeled loss $\mathcal{L}^U(x, \theta)$ is either weighted cross-entropy ℓ or zero (if the image is excluded):

$$\mathcal{L}^U(x, \theta) = 1[S(f_\theta(x))_{\hat{y}(x)} > \tau] \cdot \ell(\text{one-hot}(\hat{y}(x)), S(f_\theta(x)), \quad \hat{y}(x) = \arg\max_{c \in \mathcal{C}} f_\theta(x)_c \quad (2)$$

where $1[\cdot]$ is an indicator function that returns either one (if the expression is true) or zero.

Virtual Adversarial Training (VAT). Recently, Miyato et al. (2019) present *virtual adversarial training* as a way to improve robustness for both supervised and semi-supervised classifiers. The key idea is that for each image x we can easily find a nearby *perturbed* version of the image $x' = x + \Delta^*$, where Δ^* is the vector that leads to greatest change in predicted label distribution. To achieve *smooth* and *consistent* predictions, we wish to penalize cases where the predictions for x differ from those for the nearby x' (in KL divergence). Every training image x (both labeled and unlabeled) is assessed for this loss:

$$\mathcal{L}^U(x, \theta) = \text{KL}\left(S(f_\theta(x)), S(f_\theta(x + \Delta^*)))\right), \quad \Delta^* = \arg\max_{\Delta:||\Delta||^2 \leq \epsilon} \text{KL}(S(f_\theta(x)), S(f_\theta(x + \Delta)))$$

The perturbation vector Δ^* is constrained to have magnitude below a given perturbation size $\epsilon > 0$. Its value can be found efficiently using routines from Miyato et al. (2019).

MixMatch. MixMatch (Berthelot et al., 2019b) learns from unlabeled data by combining two key ideas: data augmentation and a variation of pseudo-label's unlabeled loss function.

How MixMatch performs augmentation. MixMatch uses the unlabeled set as a key input to its data augmentation procedure. The core of this procedure is MixUp (Zhang et al., 2017), which *linearly interpolates* between two given images. During training, MixMatch visits each minibatch (which contains labeled and unlabeled data). Each source labeled image (and its label) is transformed via MixUp with a randomly selected other image-label pair in that minibatch. If an unlabeled example is selected for pairing, we create a pseudo-label $q(x)$ from the probabilistic vector output of the classifier: $q(x) = S(f_\theta(x))$. The resulting transformed labeled minibatch is fed into the labeled loss. Thus, unlike other SSL methods described above, unlabeled data can inform training via augmentation alone, even if the unlabeled loss $\mathcal{L}^U$ is omitted.

How MixMatch calculates unlabeled loss. MixMatch also transforms each unlabeled image x in a minibatch via MixUp to obtain x', mixing with either labeled or unlabeled images. These transformed examples are fed into a pseudo-label inspired unlabeled loss:

$$\mathcal{L}^U(x, x', \theta) = \frac{1}{C} \sum_{c=1}^{C} \left(q(x)_c - S(f_\theta(x'))_c\right)^2. \quad (3)$$

Unlike the original pseudo-label method, within MixMatch the pseudo-label $q(x)$ is a probability vector (rather than a one-hot vector). Overall, this loss has *no* threshold-based exclusion, so all unlabeled examples can contribute to the loss, and uses mean squared error (rather than alternatives like KL) as recommended for robustness by Berthelot et al. (2019b).

Ablation: Augmentation-Only MixMatch. Given MixMatch's complexity, a natural question arises: does MixMatch benefit from unlabeled data because it informs the labeled loss via augmentation or because of the unlabeled loss directly? While the original work

did many other ablations (Berthelot et al., 2019b), this question was not directly answered. Our experiments thus assess a variant called **Augment-Only MixMatch**, which omits the unlabeled loss but still uses unlabeled data to inform the labeled loss via augmentation.

4.2. Standardized architecture, training, and hyperparameter selection.

Our main methodological interest is assessing *how much* the SSL paradigm improves task performance on our datasets by incorporating unlabeled data, as well as *which techniques* might work the best, as to our knowledge methods like VAT and MixMatch have not yet been applied to echocardiograms. To put all methods on a fair footing, we use a common architecture (a WideResNet) and develop a standardized protocol for training parameters (ADAM with modest L2 regularization) used by all methods. All hyperparameters are selected via a grid search to maximize balanced accuracy on the validation set. Details about all architectures, hyperparameters, and training procedures are available in App. C.2. Our open-source code provides everything needed for reproducing our implementation[5].

Below, we highlight two implementation choices that give consistent gains to all methods.

Unlabeled loss weight hyperparameter. For every SSL method we study, the unlabeled loss weight hyperparameter $\lambda > 0$ matters. We follow the recommendations of Lee (2013) and use a *deterministic annealing* schedule to slowly increase λ over epochs from an initial value of 0.0 to its maximum value linearly over the course of many training iterations. We select the maximum value of λ for all methods by monitoring the model performance on validation set, and select the λ that gives best validation performance. A well-chosen unlabeled loss schedule is important to achieve good performance for MixMatch and other methods. Supplementary Table B.3 suggests that tuning the schedule can improve balanced accuracy by over 1%.

Ensembling models over one training run to improve generalization. A recent study by Huang et al. (2017) suggests that rather than using the final checkpoint of an SGD training run, or even the best single checkpoint as ranked by validation loss, an ensemble of the checkpoints along the training trajectory can achieve better generalization. We apply this ensemble method to *every method* (SSL and baseline labeled-set only methods). We further incorporate the *weighted average* idea from Caruana et al. (2004) , allowing better-performing checkpoints larger influence. The final performance of each method is determined via an *ensemble* of the last 25 checkpoints (one per epoch). Supplement Table B.1 shows that this ensembling improves performance by a modest but noticeable amount.

4.3. Methods for aggregating many images into patient-level predictions

A key aspect of our study is producing useful diagnosis predictions for a specific patient (indexed by n), based on multiple echocardiogram images collected for that patient (indexed by $i \in \{1, 2, \ldots I_n\}$). Each patient may have a different number of images I_n, with typical I_n values in the 100s. Given all images, we wish to predict the *patient's* diagnosis label y_n (one of no AS, mild/moderate AS, or severe AS). We will use the image-specific view classifier network f_{θ_V} and diagnosis classifier network g_{θ_D} introduced in Sec. 2.1.

Below, we present several strategies for aggregating probabilistic predictions from several images to produce a *patient-level* prediction. Previous studies have often manually prescreened

---

5. https://github.com/tufts-ml/ssl-for-echocardiograms

a subset of images whose view types are known to be relevant to the prediction tasks. Only these prescreened images are used to make patient-level predictions. Instead, we consider the task faced in a real deployment where no manual prescreening is available, and we must compute the probability of the patient's diagnosis from all I_n images: $p(y_n|x_{n,1:I_n})$.

Simple average. One aggregation strategy is to simply average over the diagnosis predictions for each of the I_n images available for patient n, treating each image equally:

$$p(y_n = c|x_{n,1:I_n}) = \frac{1}{I_n} \sum_{i=1}^{I_n} S(g_{\theta_D}(x_{ni}))_c. \tag{4}$$

While simple and used in previous work (Ghorbani et al., 2020), this method will be error-prone if many images do not depict anatomical features relevant to AS diagnosis.

Prioritize diagnoses from relevant views. To diagnose AS, the PLAX and PSAX views show the anatomical structures that are *relevant*, while our catch-all "Other" view contains many diverse view types that are mostly (but not completely) irrelevant. Our view classifier network (with weights θ_V) can predict which images depict a *relevant* view (PLAX or PSAX). Thus, we suggest an aggregation procedure for diagnosis predictions that uses a *weighted* average over images. Each image's weight $w(x_{ni}) \in (0,1)$ is the view classifier's probabilistic confidence that the image shows a *clinically-relevant* view for our task:

$$p(y_n = c|x_{n,1:I_n}) \propto \sum_{i=1}^{I_n} w(x_{ni})S(g_{\theta_D}(x_{ni}))_c, \quad w(x_{ni}) = p(v_{ni} \in \mathcal{R}|x_{ni}, \theta_V). \tag{5}$$

Here, the set of *relevant* view types $\mathcal{R}$ contains PLAX and PSAX but not "Other."

We explored an alternative strategy of prioritization which *thresholds* to identify a subset of relevant images (all treated equally) rather than probabilistic weighting in which all images contribute proportional to their weight. We found this strategy's performance is slightly inferior to our weighting strategy. Details can be found in Appendix B.4.

Learned image-to-patient prediction function. We can further imagine *training a model* that can produce the predicted probabilities needed for a patient, given relevant features from its component images. We explored a few possible approaches based on manually engineered features and logistic regression classifiers, but did not find these delivered benefits worth the extra implementation effort. We leave this idea as a possible future direction.

4.4. Transfer from view to diagnosis by pretraining

A unique property of our problem and dataset is that we have two types of labels for the same input: view labels and diagnosis labels. We further have clinical knowledge that the tasks are closely related (successful diagnosis requires the ability to identify relevant views). We leverage this relation to improve training of our diagnosis classifiers. Specifically, we *pretrain* a single-image view classification network and then use this network's weights as a warm-start for our diagnosis classifier. Note this is different from common transfer learning practice where a network is pretrained on some other dataset. Our method does not require an additional dataset, merely other labels on the same dataset.

4.5. Multi-task learning using both diagnosis and view labels

Another way to improve diagnosis using view labels is via multitask learning (Ruder, 2017; Zhang and Yang, 2021), thought it is sometimes challenging in practice to determine whether auxiliary tasks will be helpful or harmful to the main task (Zhang and Yang, 2021; Ruder, 2017). We investigate whether a simple multi-task approach that trains the same network to jointly recognize view and diagnosis could be effective. Our multi-task labeled loss is:

$$\mathcal{L}^L(\theta) = \ell(y, S(g_{\theta_D}(x))) + \gamma\ell(v, S(f_{\theta_V}(x))), \tag{6}$$

where y represents the image's one-hot diagnosis label, v is the one-hot view label, and ℓ is a weighted cross-entropy loss. Hyperparameter γ controls the strength of the view loss, as we ultimately care most about diagnosing the AS severity. We focus on *labeled-set-only* evaluation of this strategy. Future work could explore semi-supervised multi-task learning.

5. Results

To evaluate whether state-of-the-art SSL can improve real-world echocardiogram classification, we now compare several SSL methods against a baseline network of the same architecture that learns from only the labeled training set. We compare Pseudo-label (Lee, 2013), virtual adversarial training (Miyato et al., 2019), and MixMatch (Berthelot et al., 2019b).

We first investigate image-level view classification in Sec. 5.1, then image-level diagnosis classification in Sec. 5.2. Finally, in Sec. 5.3 we investigate how well our methods perform at patient-level diagnosis, using the image aggregation strategies from Sec. 4.3.

Performance metric. For all view and diagnosis tasks, we use *balanced accuracy* as our primary performance metric. Given a dataset of N true labels $y_{1:N}$ and N predicted labels $\hat{y}_{1:N}$, with each label $y_n \in \{1, \ldots C\}$, we compute balanced accuracy as

$$\text{balanced-accuracy}(y_{1:N}, \hat{y}_{1:N}) = \frac{1}{C}\sum_{c=1}^{C} \frac{\text{TP}_c(y_{1:N}, \hat{y}_{1:N})}{N_c(y_{1:N})}. \tag{7}$$

Let $\text{TP}_c(\cdot)$ count *true positives* for class c (that is, the number of correctly classified examples whose true label is c), and $N_c(\cdot)$ gives the total number of examples with true label c.

We select balanced accuracy because standard accuracy does not adequately reflect performance on tasks with *label imbalance*. In our view classification test set, the "Other" category is far more common, representing over 80% of all images. Trivally guessing "Other" for every image would thus reach over 80% accuracy, but only 33.3% balanced accuracy.

5.1. View classification

We first compare all selected methods on the small TMED-18-18 dataset in Table 3. All SSL methods provide gains over methods that only use the labeled set. The largest gains come from MixMatch, which improves the baseline by over 9% in balanced accuracy.

Next, we study the best performing methods on the larger TMED-156-52 dataset in Table 4. Again, we see visible gains from MixMatch over the labeled-set-only baseline, improving over 2.5% in balanced accuracy. The relative gain of SSL is smaller here because the amount of labeled training data is larger (eventually, with enough labeled data the performance of all methods should saturate). Since Pseudo-Label and VAT perform worse

than MixMatch in our experiments on the smaller TMED-18-18 dataset, we did not assess these methods on the larger dataset to keep computation costs low.

We further observe that our simpler variant of MixMatch that only uses unlabeled data for augmentation, which we called Augment-Only MixMatch, captures most of the gains of MixMatch (around **74**% for both datasets), suggesting that augmentation (rather than a well-designed unlabeled loss) is the primary reason for MixMatch's success.

Number of *Unlabeled*			Split-specific Test Set			
Patients	Images	Method	0	1	2	average
0	0	Basic WRN	81.37	81.84	82.70	81.97
380	∼41,000	Pseudo-Label	81.81	85.07	85.82	84.23
380	∼41,000	VAT	87.95	87.61	86.36	87.31
380	∼41,000	Augment-Only MixMatch	87.77	92.43	86.06	88.75
380	∼41,000	MixMatch	92.11	93.07	88.15	**91.11**

Table 3: View classification on our **smaller TMED-18-18** dataset, showing test-set balanced accuracy across 3 separate train/test splits each with 18 labeled patients for training and 18 for test.

Number of *Unlabeled*			Split-specific Test Set				
Patients	Images	Method	0	1	2	3	average
0	0	Basic WRN	92.37	93.24	93.72	93.87	93.30
2645	303726	Augment-Only MixMatch	95.50	95.62	95.00	94.66	95.20
2645	303726	MixMatch	96.22	95.79	95.77	95.74	**95.88**

Table 4: View classification on our **full-size TMED-156-52** dataset, showing test-set balanced accuracy across 4 train/test splits, each with 156 labeled patients for training and 52 for test. For simplicity, we focus on the best SSL method (MixMatch) identified in the smaller data comparisons.

5.2. Diagnosis classification from a single image

To examine how SSL might improve AS diagnosis when given a single image, we first compare all candidate methods on the smaller dataset in Table 5. Like in the view task, we see that MixMatch beats all other SSL methods and the baseline by a substantial margin. Further experiments on the full-size dataset in Table 6 again show MixMatch improves accuracy.

We also assess the added-value of our proposed pretrain-on-view transfer learning methods (Sec. 4.4). Both Table 5 and Table 6 suggest that this pretraining strategy offers gains over simply initializing the weights of our single-image diagnosis classifier from scratch (across splits we average +3.4% on the smaller dataset and +0.5% on the larger dataset). Notably these gains are *consistent* across splits: we see a gain from pretraining visible in each of the 4 train/test splits.

In addition to pretraining, multitask learning also seems effective. We can see an average balanced accuracy gain of +4.4% on the smaller dataset and +2.2 % on the larger dataset. This demonstrates the added benefit of utilizing both view and diagnosis labels (not just diagnosis alone) to improve generalization performance.

Number of *Unlabeled*				Split-specific Test Set			
Patients	Images	Pretrain	Method	0	1	2	average
0	0		Basic WRN	42.25	51.26	36.73	43.41
0	0		Multitask WRN,	43.46	54.30	45.80	47.85
0	0	view	Basic WRN	41.76	52.21	40.40	44.79
380	∼41,000		Pseudo-Label	41.16	49.93	42.94	44.68
380	∼41,000		VAT	42.05	51.13	42.79	45.32
380	∼41,000		MixMatch	43.89	55.17	44.76	47.94
380	∼41,000	view	MixMatch	50.01	56.46	47.73	**51.43**

Table 5: AS severity diagnosis classification for *individual images* on the **smaller TMED-18-18** dataset, showing balanced accuracy averaged over the test sets from multiple train/test splits of the labeled set. Each split's test set contains all images from 18 patients, and each split's labeled training set contains all images from 18 patients. The labeled-only Basic WRN transfers from corresponding Basic WRN view classifier, while Mixmatch transfers from corresponding MixMatch view classifier

Number of *Unlabeled*				Split-specific Test Set				
Patients	Images	Pretrain	Method	0	1	2	3	average
0	0		Basic WRN	62.95	61.03	56.58	63.84	61.13
0	0		Multitask WRN	66.16	62.41	58.70	66.98	63.31
2645	303726		MixMatch	65.57	62.69	60.87	66.29	63.86
2645	303726	view	MixMatch	67.39	62.79	61.02	67.36	**64.64**

Table 6: AS severity diagnosis classification for *individual images* on the **full-size TMED-156-52** dataset, showing balanced accuracy across multiple splits. Each split's test set contains all images from 52 patients, and each split's labeled training set contains all images from 152 patients.

5.3. Diagnosis performance using all images related to a patient

We finally consider how SSL methods might improve AS diagnosis when making predictions for a patient, aggregating information from many images with diverse view types, using the methods from Sec. 4.3. The results on the full-size dataset in Table 7 suggest that SSL with MixMatch, when combined with our other key innovations (prioritizing relevant views, pretraining on view) offers real value, achieving 90.1% balanced accuracy compared to the baseline's 81.57%. Ablations in Table 7 help further disentangle how each piece (adding semi-supervised learning, adding prioritization, adding pretraining) help. The results suggest that prioritization of relevant views offers the largest and most consistent gains, followed by the semi-supervised learning, and then pretraining on view classification.

To further understand the source of these gains, we examine *confusion matrices* in Fig 3 across 4 independent train/test splits of our full-size TMED-156-52 dataset. This figure compares side-by-side our best classifier (pretrained MixMatch that prioritizes relevant views) and a labeled-set-only baseline using a simple average aggregation strategy. We see that our proposed method consistently makes fewer mistakes across all splits: 3 fewer mis-diagnosed patients on split 1, 3 fewer on split 2, 4 fewer on split 3, and 6 fewer on split 4. For every severity level and split, the proposed method achieves equal or better recall than the baseline.

Our dataset is limited: even the full-size dataset has only 52 patients in the test set to evaluate results. Therefore, to better assess the significance of our claims (that SSL

Pretrain	Method	Aggregation across images	Split-specific Test Set				
			0	1	2	3	average
	Basic WRN	Simple average	87.92	81.67	72.92	83.75	81.57
	Basic WRN	Prioritize relevant view	90.83	80.00	83.33	92.50	86.66
	MixMatch	Simple average	90.00	77.08	80.83	94.17	85.52
	MixMatch	Prioritize relevant view	88.75	83.75	85.00	94.17	87.92
view	MixMatch	Simple average	87.50	77.08	76.67	85.83	81.77
view	MixMatch	Prioritize relevant view	93.75	87.50	82.92	96.25	**90.11**

Table 7: Patient-level AS severity diagnosis classification on the **full-size TMED-156-52 dataset**, showing balanced accuracy on the test set across multiple train/test splits. Each split's labeled training set contains 156 patients, and each test set contains 52 patients. MixMatch methods are trained with *semi-supervised* learning on both labeled and unlabeled data.

learning with MixMatch delivers improved performance, which is further boosted by smart prioritization of relevant views), we revisit the portion of our large dataset that had only diagnosis labels (and no view labels) for 174 patients. For this experiment, we call this the "bonus heldout set". Results comparing all methods on this bonus heldout set are in Supplementary Table B.2. We find our claims are consistent: among methods that use simple averaging, MixMatch improves over the Basic WRN baseline by over 1% balanced accuracy, while when using prioritized views MixMatch improves by over 3.5%. We do note that these "bonus set" images were included in the *unlabeled* training set. However, we emphasize that their labels were *not used* during training and that they make up less than 6% of the total unlabeled set.

5.4. ROC analysis of No AS vs Some AS

Finally, we assess our method's ability as a first-line screening tool by reporting a receiver operating curve in Supplementary Fig. B.5. for the simpler binary task of distinguishing between "no AS" and "some AS" (combining the "mild/moderate AS" and "severe AS" labels). We compared a labeled-set-only model with simple averaging, a labeled-set-only model with prioritized voting, and our SSL-trained MixMatch method with prioritized voting. While all compared methods achieve high area-under-the-ROC scores, we find that prioritized voting consistently shows gains, achieving a remarkable average AUC of **0.98** across the 4 splits. This performance suggests that we may not be far from effective deployment of these models as a first-line screening tool, provided we can replicate this performance in future external validation.

6. Discussion

We have developed and evaluated a semi-supervised learning pipeline that can leverage abundant unlabeled data to deliver competitive patient-level diagnostic predictions for the fully-automated preliminary assessment of aortic stenosis. These methods overcome two challenges. First, echocardiograms are challenging to label so the amount of labeled training data is limited. Second, a patient's record will contain hundreds of images, many of which are not relevant for diagnosing AS. Here, we briefly review the limitations and advantages of our approach.
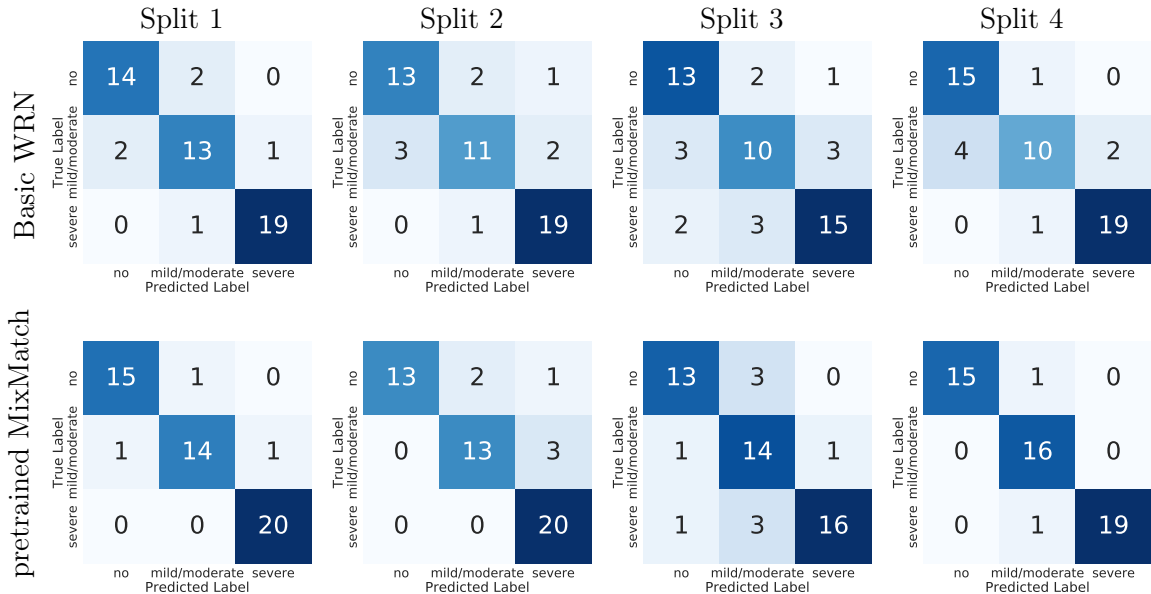
Figure 3: Confusion matrices for the patient-level AS diagnosis classification, across all four train/test splits of the **full-size TMED-156-52** dataset. Each split's test set contains 52 patients. We show the labeled-set only basic WRN (top) as well as the best SSL method (bottom). Our method improves accuracy compared to the baseline, and most remaining mistakes confuse nearby classes (e.g. "no AS" vs. "mild/moderate AS") instead of distant classes (e.g. "no AS" vs. "severe AS").

Limitations. The most important caveat to this work is the need for further independent validation of our methodology. For logistical reasons, all our data come from one institution. A detailed evaluation at another institution would be needed to properly assess our proposed pipeline's utility in a prospective setting when used with different patient populations, imaging devices, and sonographers.

A critical and well-known issue with interpretation of echocardiograms is *inter-rater reliability* (Sacchi et al., 2018). In particular, the distinction between mild and moderate or moderate and severe diagnostic levels can vary across annotators. All labels in our dataset come from less than 5 expert annotators at one institution. Further study is needed to understand if our approach could match the consensus of a broader population of annotators.

Several other opportunities to improve our pipeline exist. Our image processing approach prioritizes simplicity but does not take advantage of recent larger CNN architectures or region-specific attention or segmentation as in some past work on cardiac imaging (Chen et al., 2020a). We could use higher-resolution images. We could include other easily-measured covariates (besides imaging) in our diagnostic model, such as age, demographics, comorbidities, and other cardio-mechanical signals.

A final limitation is that further effort to qualitatively understand what visual signals are driving predictions is needed to build trust. We plan to investigate saliency maps (Simonyan et al., 2014; Selvaraju et al., 2020), though we will be mindful of the limitations of these methods (Adebayo et al., 2018). Qualitative insight is key, because fundamentally, MixMatch works by *interpolating* images. It is surprising to us that MixMatch delivered consistently

improved results (replicated across several train/test splits) in a real medical imaging scenario, because interpolated echocardiogram images have questionable meaning to human experts.

Advantages. For potentially fatal conditions like AS, echocardiograms remain the gold standard source of information to produce a diagnosis. Our approach can already reach performance levels (90% balanced accuracy) that might be useful in a deployed setting (naturally, these must first be reliably replicated in a prospective setting on an external cohort). Automatic diagnostic classification pipelines have the potential to identify individuals who would benefit from further screening who otherwise would not be discovered due to limited access to expert cardiologists.

A key aspect of our approach is demonstrating the value of *semi-supervised* learning for a real medical task with class-imbalance (for our view task over 80% of the images are "Other"). Our dataset also includes truly unlabeled data from over 2400 patients, which represents a more authentic test of SSL than previous benchmark datasets. Overall, our work motivates modern SSL as a promising cost-effective way to improve performance if unlabeled data is abundant, even for real clinical images with substantial diversity. Especially if labeled sets are small, the gains from SSL may be even greater (see Table 3).

A final advantage of our work is the demonstration that patient-level diagnosis benefits from *prioritizing relevant views*. Building on Madani et al. (2018b), who showed promising SSL diagnosis given manually-curated views, our SSL methods can deliver effective diagnoses given an uncurated set of *all* available images, even when most depict irrelevant views.

We hope our study marks a step toward effective early detection of aortic stenosis that can enable helpful interventions. We further hope this study and the accompanying dataset release offer a reproducible template for improving patient outcomes for other diseases where medical imaging is key and labeled data is scarce.

Acknowledgements

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Sanjeev Arora and Yi Zhang. Do GANs actually learn the distribution? An empirical study. *arXiv:1706.08224 [cs]*, 2017. URL http://arxiv.org/abs/1706.08224.

Wayne Batchelor, Saif Anwaruddin, Laura Ross, Oluseun Alli, Michael N. Young, Aaron Horne, Abby Cestoni, Frederick Welt, and Roxana Mehran. Aortic Valve Stenosis Treat-

ment Disparities in the Underserved: JACC Council Perspectives. *Journal of the American College of Cardiology*, 74(18), 2019.

Helmut Baumgartner, Judy Hung, Javier Bermejo, John B. Chambers, Thor Edvardsen, Steven Goldstein, Patrizio Lancellotti, Melissa LeFevre, Fletcher Miller, and Catherine M. Otto. Recommendations on the echocardiographic assessment of aortic valve stenosis: A focused update from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *European Heart Journal Cardiovascular Imaging*, 18(3):254–275, 2017.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019b.

Matthew J. Brennan, Alex Bryant, Isabel Boero, Tracy Wang, Sreekanth Vemulapalli, Vinod Thourani, and Eric Peterson. Provider-level variability in the treatment of patients with severe symptomatic aortic valve stenosis. *Journal of the American College of Cardiology*, 73(9 (Supplement 1)):1949, 2019.

Saul Calderon-Ramirez, Raghvendra Giri, Shengxiang Yang, Armaghan Moemeni, Mario Umana, David Elizondo, Jordina Torrents-Barrena, and Miguel A Molina-Cabello. Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images. In *International Conference on Pattern Recognition (ICPR)*, 2021.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, 2020b. URL http://arxiv.org/abs/2006.10029.

Yu Chen, Yuxuan Wang, Bolin Lai, Zijie Chen, Xu Cao, Nanyang Ye, Zhongyuan Ren, Junbo Zhao, Xiao-Yun Zhou, and Peng Qi. Venibot: Towards autonomous venipuncture with semi-supervised vein segmentation from ultrasound images. *arXiv preprint arXiv:2105.12945*, 2021.

Mary Ann Clark, Suzanne V. Arnold, Francis G. Duhay, Ann K. Thompson, Michelle J. Keyes, Lars G. Svensson, Robert O. Bonow, Benjamin T. Stockwell, and David J. Cohen. Five-year clinical and economic outcomes among patients with medically managed severe aortic stenosis: Results from a Medicare claims analysis. *Circulation. Cardiovascular Quality and Outcomes*, 5(5):697–704, 2012.

Adam Coates, Honglak Lee, and Andrew Y Ng. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *Artificial Intelligence and Statistics*, 2011. URL https://cs.stanford.edu/~acoates/stl10/.

Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1), 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014. URL https://papers.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Erika Hata, Chanjin Seo, Masafumi Nakayama, Kiyotaka Iwasaki, Takaaki Ohkawauchi, and Jun Ohya. Classification of Aortic Stenosis Using ECG by Deep Learning and its Analysis Using Grad-CAM. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2020:1548–1551, 2020.

N Howlader, A. M. Noone, M Krapcho, D. Miller, A. Brest, M. Yu, J Ruhl, Z Tatalovich, A Mariotto, D R Lewis, H S Chen, E J Feuer, and K A Cronin. SEER Cancer Statistics Review, 1975-2017. Technical report, National Cancer Institute, 2020. URL https://seer.cancer.gov/csr/1975_2017.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Joon-Myoung Kwon, Soo Youn Lee, Ki-Hyun Jeon, Yeha Lee, Kyung-Hee Kim, Jinsik Park, Byung-Hee Oh, and Myong-Mook Lee. Deep Learning–Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *Journal of the American Heart Association*, 9 (7), 2020.

Patrizio Lancellotti, Julien Magne, Raluca Dulgheru, Marie-Annick Clavel, Erwan Donal, Mani A. Vannan, John Chambers, Raphael Rosenhek, Gilbert Habib, Guy Lloyd, Stefano Nistri, Madalina Garbi, Stella Marchetta, Khalil Fattouch, Augustin Coisne, David Montaigne, Thomas Modine, Laurent Davin, Olivier Gach, Marc Radermecker, Shizhen Liu,

Linda Gillam, Andrea Rossi, Elena Galli, Federica Ilardi, Lionel Tastet, Romain Capoulade, Robert Zilberszac, E. Mara Vollema, Victoria Delgado, Bernard Cosyns, Stephane Lafitte, Anne Bernard, Luc A. Pierard, Jeroen J. Bax, Philippe Pibarot, and Cécile Oury. Outcomes of Patients With Asymptomatic Aortic Stenosis Followed Up in Heart Valve Clinics. *JAMA cardiology*, 3(11):1060–1068, 2018.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9), 2019.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, 2013.

Gary Long and Benjamin S Wessler. Identification of Echocardiographic Imaging View Using Deep Learning | Circulation: Cardiovascular Quality and Outcomes. In *Circulation: Cardiovascular Quality and Outcomes*, 2018. URL https://www.ahajournals.org/doi/10.1161/circoutcomes.11.suppl_1.276.

Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1), 2018a.

Ali Madani, Jia Rui Ong, Anshul Tibrewal, and Mohammad RK Mofrad. Deep echocardiography: Data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine*, 1(1), 2018b.

Qingjie Meng, Jacqueline Matthew, Veronika A Zimmer, Alberto Gomez, David FA Lloyd, Daniel Rueckert, and Bernhard Kainz. Mutual information-based disentangled neural networks for classifying unseen categories in different domains: application to fetal ultrasound imaging. *IEEE Transactions on Medical Imaging*, 40(2), 2020.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. *arXiv:1611.02163 [cs, stat]*, 2017. URL http://arxiv.org/abs/1611.02163.

Jan Minners, Martin Allgeier, Christa Gohlke-Baerwolf, Rolf-Peter Kienzle, Franz-Josef Neumann, and Nikolaus Jander. Inconsistencies of echocardiographic criteria for the grading of aortic valve stenosis. *European Heart Journal*, 29(8):1043–1048, 2008.

Carol Mitchell, Peter S. Rahko, Lori A. Blauwet, Barry Canaday, Joshua A. Finstuen, Michael C. Foster, Kenneth Horton, Kofo O. Ogunyankin, Richard A. Palma, and Eric J. Velazquez. Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography*, 32(1):1–64, 2019.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 41(8), 2019. URL https://ieeexplore.ieee.org/document/8417973/.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL http://ufldl.stanford.edu/housenumbers.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802), 2020.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Stefania Sacchi, Niti M. Dhutia, Matthew J. Shun-Shin, Massoud Zolgharni, Nilesh Sutaria, Darrel P. Francis, and Graham D. Cole. Doppler assessment of aortic stenosis: A 25-operator study demonstrating why reading the peak velocity is superior to velocity time integral. *European Heart Journal Cardiovascular Imaging*, 19(12):1380–1389, 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *arXiv:1312.6034 [Cs]*, 2014. URL http://arxiv.org/abs/1312.6034.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Liang Tang, Mario Gössl, Aisha Ahmed, Ross Garberich, Steven M. Bradley, Hiroki Niikura, Dawn Witt, Wesley R. Pedersen, Richard Bae, John R. Lesser, Kevin M. Harris, Benjamin Sun, Karol Mudy, and Paul Sorajja. Contemporary Reasons and Clinical Outcomes for Patients With Severe, Symptomatic Aortic Stenosis Not Undergoing Aortic Valve Replacement. *Circulation: Cardiovascular Interventions*, 11(12), 2018.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Xi Wang, Hao Chen, Huiling Xiang, Huangjing Lin, Xi Lin, and Pheng-Ann Heng. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis*, 70, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Chenxi Yang, Banish D. Ojha, Nicole D. Aranoff, Philip Green, and Negar Tavassolian. Classification of aortic stenosis using conventional machine learning and deep learning methods based on multi-dimensional cardio-mechanical signals. *Scientific Reports*, 10(1): 17521, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. *arXiv:1605.07146 [cs]*, 2017. URL http://arxiv.org/abs/1605.07146.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H. Tison, Laura A. Hallock, Lauren Beussink-Nelson, Mats H. Lassen, Eugene Fan, Mandar A. Aras, ChaRandle Jordan, Kirsten E. Fleischmann, Michelle Melisko, Atif Qasim, Sanjiv J. Shah, Ruzena Bajcsy, and Rahul C. Deo. Fully Automated Echocardiogram Interpretation in Clinical Practice. *Circulation*, 138(16):1623–1635, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Technical Report Technical Report 1530, Department of Computer Science, University of Wisconsin Madison., 2005.

Appendix A. Dataset Visualizations
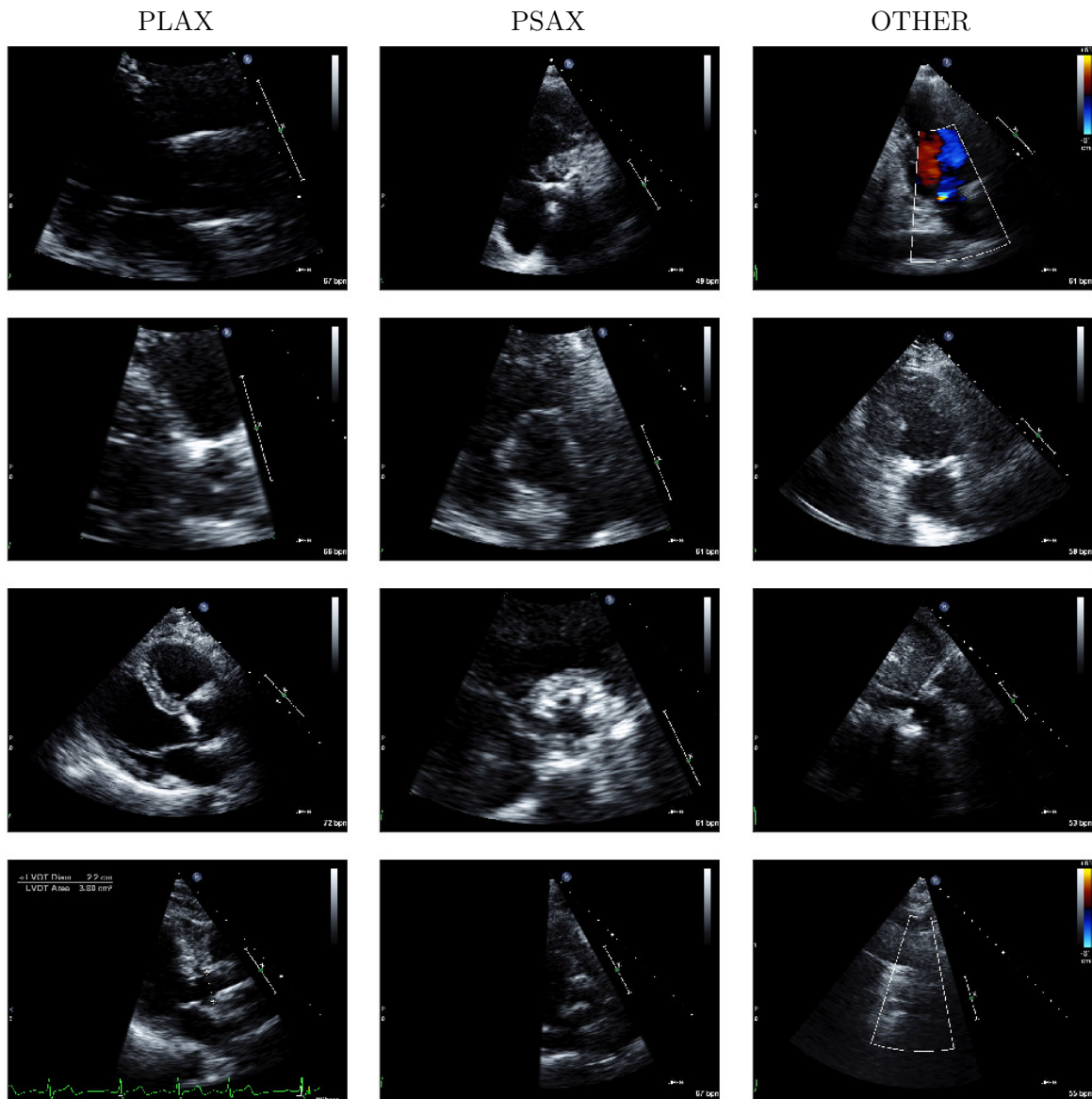
A.1. Examples of each view class

Figure A.1: Examples of images for each possible view label in our dataset. *From left to right:* Four examples of peristernal long axis (PLAX) view, four examples of peristernal short axis (PSAX) view, and four examples of other kinds of view in our "Other" class.

A.2. Examples of each view for a Severe AS patient

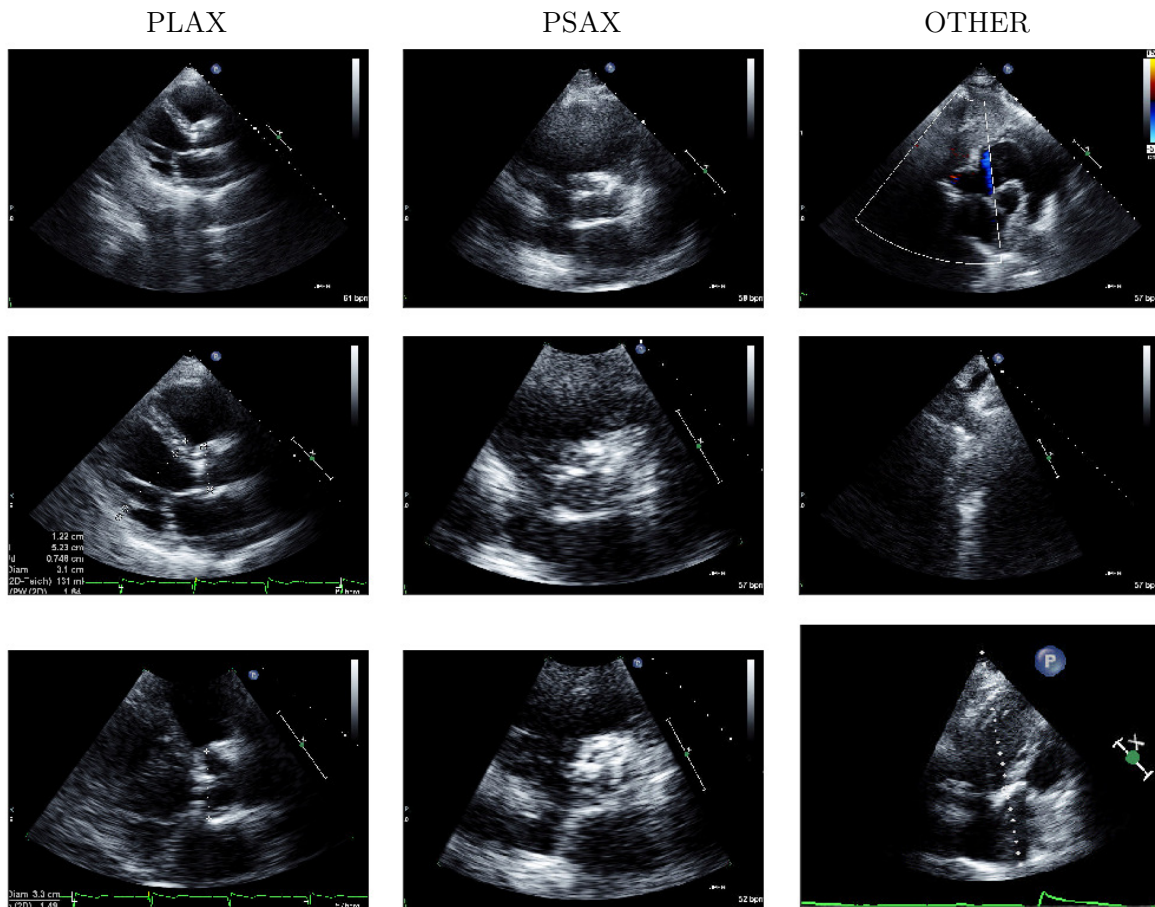PLAX   PSAX   OTHER

Figure A.2: Examples of images from a patient with Severe AS in our dataset. *From left to right:* Three examples of parasternal long axis (PLAX) view, three examples of parasternal short axis (PSAX) view, and three examples of other kinds of view in our "Other" class.

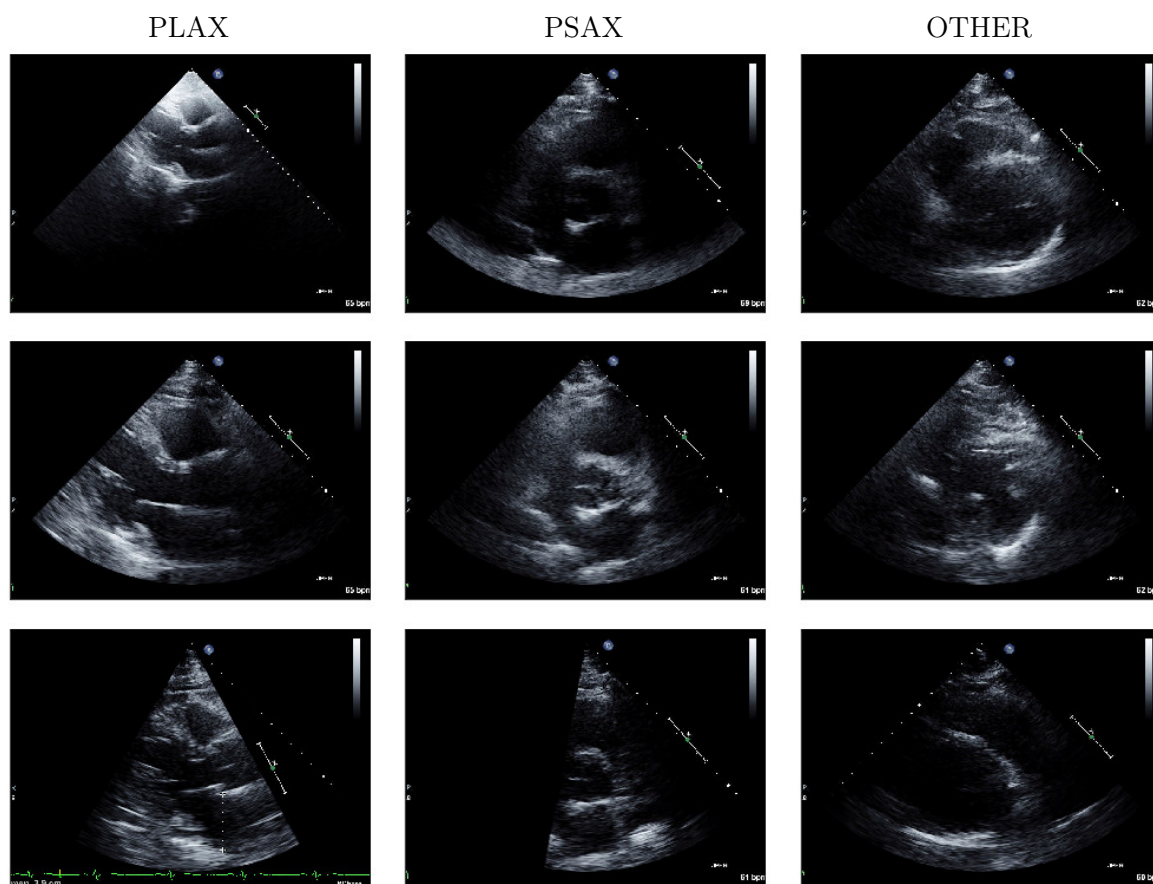### A.3. Examples of each view for a No AS patient

Figure A.3: Examples of images from a patient with No AS in our dataset. *From left to right:* Three examples of parasternal long axis (PLAX) view, three examples of parasternal short axis (PSAX) view, and three examples of other kinds of view in our "Other" class.

Appendix B. Further Results

B.1. Assessment of ensembling

Table B.1 compares using a single checkpoint (one point estimate of neural network weight vector θ) to using an ensemble of parameters aggregated from the last 25 checkpoints (one per epoch).

Diagnosis classification	Split 1	Split 2	Split 3	Split 4	Average
Best single checkpoint	61.81	59.79	56.05	64.21	60.46
Ensemble	62.95	61.03	56.58	63.84	**61.13**
View classification					
Best single checkpoint	93.03	93.24	92.39	93.79	93.11
Ensemble	92.37	93.24	93.72	93.87	**93.30**

Table B.1: Comparing best single checkpoint performance with ensemble performance on **Full-size TMED-156-52**

B.2. Patient-level diagnosis performance on bonus heldout set

Table B.2 examines the performance of the best labeled-set-only methods and MixMatch methods on the 174 patient studies that have diagnosis but no view labels. While the images used here were originally included in the unlabeled training set (which was used to train SSL methods like MixMatch), the diagnosis labels were not provided at all during training time. We thus still believe this is an authentic test of generalization given the scarcity of labeled data available for our task. Of course, additional independent evaluation (especially from another institution) is needed.

Pretrain	Method	Voting	Split 1	Split 2	Split 3	Split 4	average
	Basic WRN	Simple average	76.73	75.25	76.87	81.88	77.68
	Basic WRN	View-prioritized	73.63	83.21	79.70	80.08	79.18
	MixMatch	Simple average	85.32	76.29	74.14	79.95	78.93
view	MixMatch	Simple average	83.36	77.96	75.61	81.37	79.58
	MixMatch	View-prioritized	83.27	83.76	82.34	82.83	**83.05**
view	MixMatch	View-prioritized	82.53	86.15	79.62	83.27	82.89

Table B.2: Patient-level AS Severity Diagnosis Classification on the **bonus heldout set** of 174 patients for whom we have diagnosis labels only (no view labels). We show balanced accuracy on models trained on each of the four folds on four **full-size TMED-156-52** dataset.

B.3. Assessment of MixMatch hyperparameter sensitivity

In Table B.3, we consider four possible strategies for setting the hyperparameters of MixMatch, varying two key settings for the weight on unlabeled loss λ. First, we vary whether the final value of λ is set to its *best* value among a grid of candidates (based on validation set performance), or *fixed* to a constant. Second, we vary whether λ remains fixed over iterations

throughout a training run, or is updated over iterations on a linear ramp schedule from 0 to its final target value.

From this comparison, we see we consistent gains across splits (average gain across splits of over 1.6% balanced accuracy) for using a delayed ramp up schedule with target value selected via grid search.

Final λ value	λ update schedule	Split 1	Split 2	Split 3	Split 4	Average
best on val	Delayed ramp-up	65.57	62.69	60.87	66.29	63.86
best on val	Immediate ramp-up	65.07	61.87	60.82	65.37	63.28
best on val	Constant	65.03	61.52	58.87	65.22	62.66
100 (fixed)	Constant	63.94	61.79	58.87	64.35	62.24

Table B.3: Ablation study of different settings of the unlabeled loss weight λ for MixMatch. AS severity diagnosis classification for individual images on the **full-size TMED-156-52** dataset. showing balanced accuracy averaged over the test sets from multiple folds (each fold's test set contains all images from 52 patients).

B.4. Assessment of alternative view prioritization strategy using thresholding

An anonymous reviewer suggested an alternative strategy for prioritizing images of relevant view. The alternative strategy works as follows: for each image, we compute the predicted probability that the image is a "relevant view" (either PLAX and PSAX) by summing the probabilities of each view type. However, instead of using this raw probability as a weight (as our chosen method does), we use a *cutoff threshold* and simply average the diagnosis predictions of images whose relevant view probability is above the cutoff. For each patient, we use the majority vote prediction of the diagnosis from the images of relevant views. The value of the cutoff threshold is selected using the validation set to maximize balanced accuracy.

Table B.4 shows the performance of this strategy ("threshold-then-average") on the full-size dataset. Using this alternative prioritization strategy together with our suggested methodology for patient-level diagnosis (using MixMatch, pretraining on view), we find the average test set balanced accuracy is around 85.8%, while the weighted average strategy in the main paper achieves over 90% balanced accuracy. We take this as reasonably decisive evidence that a weighted average (rather than a simple cutoff) should be preferred.

Pretrain	Method	Aggregation across images	Split 1	Split 2	Split 3	Split 4	average
	Basic WRN	Threshold-then-Average	85.42	86.25	79.17	92.50	85.84
	MixMatch	Threshold-then-Average	83.33	84.17	77.50	94.58	84.90
view	MixMatch	Threshold-then-Averagen	86.67	80.00	82.50	94.17	85.84

Table B.4: Alternative view-prioritizing strategy for patient-level AS severity diagnosis classification on the **full-size TMED-156-52** dataset, showing balanced accuracy on the test set across multiple folds (each fold's test set contains 52 patients).

B.5. ROC Curve of patient-level diagnosis: no AS vs. mild/moderate/severe AS

Fig. B.5 shows receiver operating curves for several methods for the task of distinguishing no AS vs Some AS (which aggregates both the mild/moderate and severe levels in the 3-level diagnosis task of the main paper).
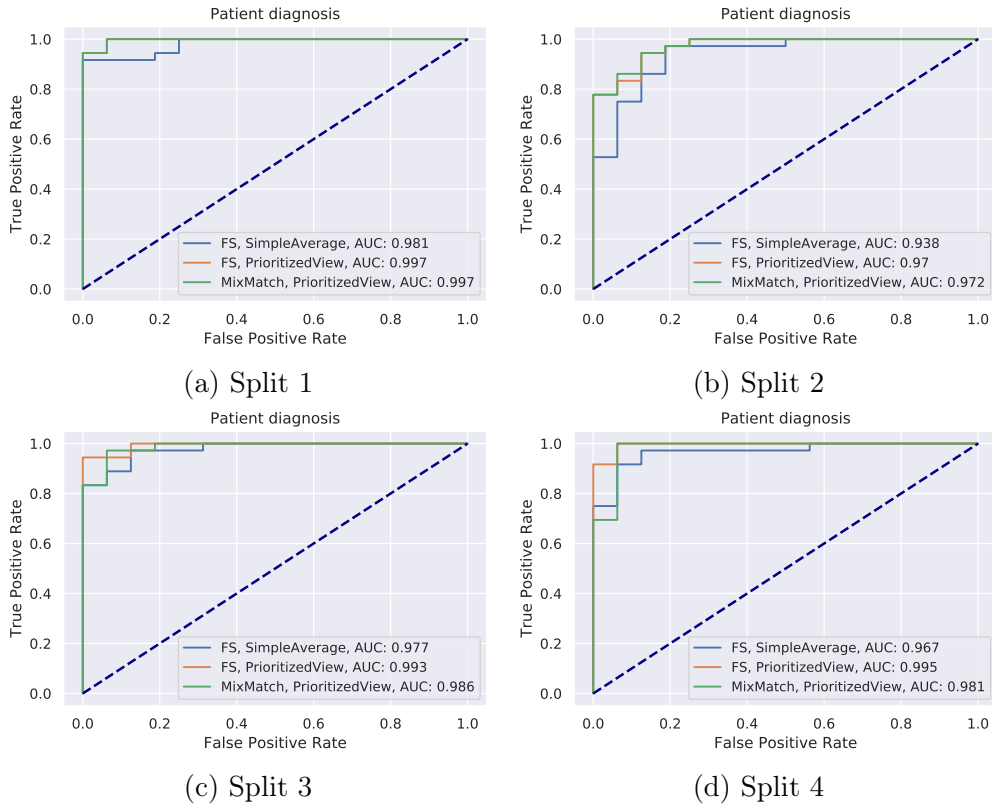

(a) Split 1   (b) Split 2

(c) Split 3   (d) Split 4

Figure B.5: ROC curves for binary diagnosis task (no AS vs "mild/moderate/severe AS") on **full-size TMED-156-52**.

Appendix C. Methodological Details

C.1. Image processing details

Removing doppler images.   In the raw data of all imagery available for an echocardiogram study, we obtained TIFF files that represent both cineloops and Doppler images.

We verified in our labeled set that all Doppler images have one of the following landscape aspect ratio $(831, 323)$, $(901, 384)$, $(901, 390)$, $(704, 305)$, $(831, 421)$, $(901, 469)$ or $(563, 294)$. Only the Dopplers have these aspect ratios. We thus filtered out Doppler completely via these aspect ratios.

Downsizing   The original images are provided as high-resolution TIFF format images (hundreds of pixels per side) of varying aspect ratios. Generally, we can expect that both view and diagnosis classifiers would perform better given higher-resolution input (and holding

other factors the same). The main trade-off of processing higher-resolution images is increased runtime and memory requirements. In our preliminary experiments, we compared downsizing all images to a standard square aspect ratio at 3 possible sizes: 32x32, 64x64 and 128x128. We found that 64x64 achieves a good balance between model performance and computation cost. A prior study by Madani et al. (2018b) provides a more extensive study of optimal resolution size. The interested reader can refer to their work for more details.

C.2. Architecture Settings and Hyperparameters

Weighted cross-entropy for labeled loss   To counteract the effect of class imbalance in the dataset, we use weighted cross-entropy for the labeled loss. For an input image x whose true label y indicates it belongs to class c, the weighted cross-entropy assumes the following form:

$$\mathcal{L}^L(\theta, x) = -w_c \log \hat{p}_c(\theta, x), \tag{8}$$

where $\hat{p}_c$ is the predicted probability of class c. The weight w_c is calculated using the training set statistics as follow:

$$w_c = \frac{\prod_{k \neq c} N_k}{\sum_j \prod_{k \neq j} N_k} \tag{9}$$

where N_k is the number of images of class k in the training set.

Common architecture.   Following Oliver et al. (2018), for all considered methods, we use the *same* backbone neural network architecture: a wide residual network (Zagoruyko and Komodakis, 2017) with 28 layers (WRN-28), which has total of 5,931,683 parameters. This same network architecture is used in the original MixMatch evaluation (Berthelot et al., 2019b) with promising results.

Common training protocol.   All SSL methods we consider follow the loss minimization framework with two primary losses (one for "labeled" data and one for "unlabeled" data) in Eq. (2.1). We allow every method to train for 32 epochs (where each epoch processes 2^{16} images, as in Berthelot et al. (2019b)). Our preliminary experiments suggest that after 30 epochs all methods effectively converge in terms of validation balanced accuracy.

Common regularization.   For all methods, we expect performance will be vulnerable to overfitting, so we impose an L2-norm penalty on the weights θ, also known as weight decay. Each method selects its preferred value of this penalty strength hyperparameter. We searched values in [0.0002, 0.002, 0.02].

Common optimization.   We use ADAM (Kingma and Ba, 2014) to optimize each model. Each method selects the value of the step size (learning rate) as a hyperparameter. We experimented with 0.002 and 0.0007

Hyperparameters for Pseudo-Label.   Beyond the usual hyperparameters for our loss-minimization SSL framework, another important hyperparameter for pseudo-label is the threshold τ. We find that performance is not very sensitive to the chosen τ value as long as it is within a certain range. We set τ to 0.95, as done in past literature that evaluates Pseudo-Label as an SSL method (Oliver et al., 2018; Berthelot et al., 2019b,a; Sohn et al., 2020).

Hyperparameters for VAT. Beyond the usual hyperparameters for our SSL framework, for VAT we need to select a value for ϵ. In Miyato et al. (2019), the authors claimed that they can achieve superior performance by tuning only ϵ and fixing λ to 1. In our experiment, we used the default λ as in Berthelot et al. (2019b) and searched the value of ϵ in [2, 6, 18], together with learning rate and weight decay. We select the best hyperparameters using validation set performance.

Hyperparameters for MixMatch. Beyond the usual hyperparameters for our SSL framework, the key hyperparameters for MixMatch include the number of augmentations K, the temperature $T > 0$ used for sharpening, interpolation hyperparameter α and unlabeled loss coefficient λ. We set $K = 2$, $T = 0.5$, and $\alpha = 0.75$ as done in Berthelot et al. (2019b), and search for λ in the range [10, 30, 75, 100, 130] using validation set.

Hyperparameters for Multitask training. We searched γ, the hyperparameter that control the strength of the auxilliary view loss in Eq. (6), in the range [10, 3, 1, 0.3, 0.1]. The best α is selected together with other hyperparameters on validation set.