Semi-Supervised Deep Regression with Uncertainty Consistency and Variational Model Ensembling via Bayesian Neural Networks

Weihang Dai[1], **Xiaomeng Li**[1, 2*], **Kwang-Ting Cheng**[1, 2]

[1] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
[2] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
wdaiaj@connect.ust.hk, eexmli@ust.hk, timcheng@ust.hk;

Abstract

Deep regression is an important problem with numerous applications. These range from computer vision tasks such as age estimation from photographs, to medical tasks such as ejection fraction estimation from echocardiograms for disease tracking. Semi-supervised approaches for deep regression are notably under-explored compared to classification and segmentation tasks, however. Unlike classification tasks, which rely on thresholding functions for generating class pseudo-labels, regression tasks use real number target predictions directly as pseudo-labels, making them more sensitive to prediction quality. In this work, we propose a novel approach to semi-supervised regression, namely Uncertainty-Consistent Variational Model Ensembling (UCVME), which improves training by generating high-quality pseudo-labels and uncertainty estimates for heteroscedastic regression. Given that aleatoric uncertainty is only dependent on input data by definition and should be equal for the same inputs, we present a novel uncertainty consistency loss for co-trained models. Our consistency loss significantly improves uncertainty estimates and allows higher quality pseudo-labels to be assigned greater importance under heteroscedastic regression. Furthermore, we introduce a novel variational model ensembling approach to reduce prediction noise and generate more robust pseudo-labels. We analytically show our method generates higher quality targets for unlabeled data and further improves training. Experiments show that our method outperforms state-of-the-art alternatives on different tasks and can be competitive with supervised methods that use full labels. Code is available at https://github.com/xmed-lab/UCVME.

Introduction

Deep learning has achieved state-of-the-art results on a variety of tasks such as classification (Dosovitskiy et al. 2020), segmentation (Chen et al. 2021), image generation (Bodla, Hua, and Chellappa 2018), and others. These methods tend to require large amounts of labeled data for training, however, which can be costly to annotate. State-of-the-art image classifiers such as ViT are trained on the JFT-300M dataset, for example, which consists of 300 million images (Dosovitskiy et al. 2020). Labeling can also be prohibitively expensive for medical image analysis, where life-saving tasks such as medical disease diagnoses (Li et al. 2021) and tumor segmentation (Li et al. 2018a) require domain expertise. The ability to train neural networks with reduced labels is therefore highly valuable and an active research area.

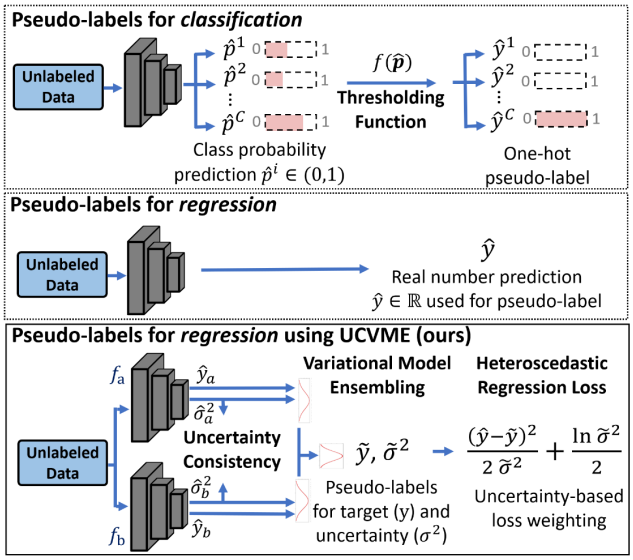


Figure 1: Differences between standard pseudo-labeling approaches and our method (UCVME). Classification tasks typically apply thresholding functions to probability predictions, $\hat{p}^i$, to obtain one-hot pseudo-labels, $\hat{y}^i$. Regression tasks use real number target predictions, $\hat{y}$, directly as pseudo-labels and are therefore more sensitive to prediction quality. Our UCVME improves pseudo-labels for regression by considering pseudo-label uncertainty, σ^2, and robustness. We use a novel uncertainty consistency loss to improve uncertainty-based loss weighting and a variational model ensembling method to improve pseudo-label quality.

Semi-supervised learning uses unlabeled data together with a smaller labeled dataset for model training. These methods reduce reliance on labeled data and sometimes outperform state-of-the-art techniques on fully labeled datasets. Chen *et al.* (Chen et al. 2021) propose CPS, a semi-supervised algorithm for image segmentation, which is time-consuming to label. Li *et al.* (Li et al. 2021) enforce con-

*Corresponding author

sistency between transformed inputs for medical diagnosis, which requires specialist knowledge for annotation. However, comparatively less attention has been paid to deep regression problems, which cover practical applications such as age estimation (Berg, Oskarsson, and O'Connor 2021) and pose estimation (Yang et al. 2019) from images. Deep regression is particularly important in the medical field as it is used to obtain measurements for disease diagnosis and progression tracking, such as bone mineral density estimation for osteoporosis (Hsieh et al. 2021) and ejection fraction estimation for cardiomyopathy (Ouyang et al. 2020).

Regression problems are fundamentally different from classification problems because they generate real number predictions instead of class probabilities. Existing semi-supervised classification techniques *cannot be applied to semi-supervised regression* because they rely on class probabilities and thresholding functions to generate pseudo-labels (Zhang et al. 2021; Sohn et al. 2020) (see Fig. 1). Limited efforts have been devoted to exploring semi-supervised approaches for deep regression. Recent works by Jean *et al.* (Jean, Xie, and Ermon 2018) propose deep kernel learning for semi-supervised regression, but their method is designed for tabular data. Pretrained feature extractors are used for image inputs, which prevents task-specific feature learning and limits performance. Wetzel *et al.* (Wetzel, Melko, and Tamblyn 2021) propose TNNR, which estimates the difference between inputs with deep networks and uses loop consistency for unlabeled data. Loop consistency regulates training, but poor-quality predictions can still reduce the effectiveness of the constraints (see Tables 1 and 4).

Unlike classification tasks, which can smooth predictions using thresholding functions for class pseudo-labeling, regression tasks directly use real number target predictions as pseudo-labels. Therefore, model performance highly depends on the quality of pseudo-labels, *i.e.* predictions. In this paper, we propose a novel Uncertainty-Consistent Variational Model Ensembling method, namely UCVME, that adjusts for the uncertainty of pseudo-labels during training and increases pseudo-label robustness. Our method is based on two key ideas: enforcing uncertainty consistency between co-trained models to improve uncertainty-based loss weighting, and using ensembling techniques to reduce prediction variance for obtaining higher quality pseudo-labels.

We make use of Bayesian neural networks (BNNs), which predict aleatoric uncertainty of observations jointly with the target value. The uncertainty estimates are used for heteroscedastic regression, which assigns sample weightings based on uncertainty to reduce the impact of noisier samples (Kendall and Gal 2017). We observe that aleatoric uncertainty, which by definition is dependant only on input data, *should be equal for the same input,* and propose a novel consistency loss for uncertainty predictions of co-trained models. Our proposed loss notably improves aleatoric uncertainty estimates on unlabeled data, such that higher quality pseudo-labels are given greater importance through heteroscedastic regression (see Fig. 5). This is non-trivial since unreliable uncertainty estimates can lead to adverse loss-weighting and unstable training. Our proposed method is *the first to address uncertainty estimation quality for regression.*

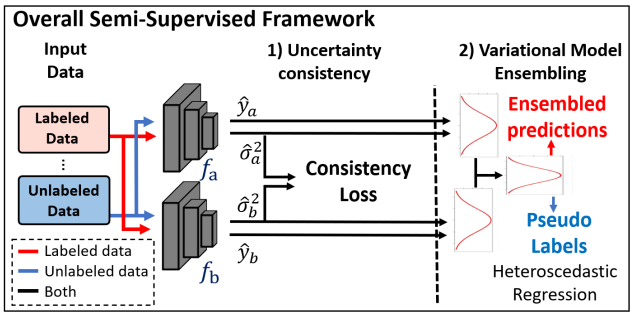


Figure 2: Semi-supervised deep regression framework for our UCVME method. UCVME improves overall pseudo-label quality and assigns greater sample weights to pseudo-labels with low uncertainty.

BNNs also use variational inference during prediction to approximate the underlying distribution of estimates. To improve robustness of pseudo-labels, we introduce variational model ensembling, which uses ensembling methods with variational inference to reduce prediction noise. We analytically show our approach generates higher quality targets for unlabeled data and validate results experimentally (see Table 2). The combined improvements in uncertainty estimation and pseudo-label quality lead to state-of-the-art performance. Fig. 2 illustrates the overall framework.

We demonstrate our method on two regression tasks: age estimation from photographs and ejection fraction estimation from echocardiogram videos. Results show our method outperforms state-of-the-art alternatives and is competitive with supervised approaches using full labels (see Tables 1 and 4). Ablations demonstrate individual contributions from uncertainty consistency and variational model ensembling (Table 2). We summarize our main contributions as follows:

- We propose UCVME, a novel semi-supervised method that improves uncertainty estimates and pseudo-label robustness for deep regression tasks.
- We introduce a novel consistency loss for aleatoric uncertainty predictions of co-trained models, based on the insight that estimates should be equal for the same input.
- We introduce variational model ensembling for generating pseudo-labels on unlabeled data, which we analytically show is more accurate than deterministic methods.
- Results show our method outperforms existing state-of-the-art alternatives on two separate regression tasks.

Related Works

In this section, we review works on learning from unlabeled data, general approaches for semi-supervised learning, state-of-the-art methods for semi-supervised regression, and existing methods for uncertainty estimation.

Unsupervised Representation Learning

One way to learn from unlabeled data is to learn unsupervised feature representations, which can then be fine-tuned for specific tasks using classifiers. Techniques such as

PCA (Bengio, Courville, and Vincent 2013) and data clustering (Huang, Loy, and Tang 2016) learn intermediate features by reducing input dimensionality. With the increasing effectiveness of deep learning, pre-text tasks such as input reconstruction (Kingma and Welling 2013), augmentation prediction (Zhang et al. 2019), and order prediction (Noroozi and Favaro 2016) have been explored for unsupervised training of deep feature extractors. Current state-of-the-art approaches are based on contrastive learning, which has been shown in some cases to outperform supervised learning (Chen, Xie, and He 2021; Chen et al. 2020).

Semi-Supervised Learning

Semi-supervised learning uses both labeled and unlabeled data for training. This reflects realistic settings where raw data is easy to obtain but annotations can be costly. State-of-the-art methods include enforcing consistency on augmented inputs and using pseudo-labels for unlabeled samples. For example, CCT (Ouali, Hudelot, and Tami 2020) applies prediction consistency after perturbing intermediate features. CPS (Chen et al. 2021) enforces consistency of segmentation predictions between co-trained models. Temporal ensembling (Laine and Aila 2016) and mean-teacher (Tarvainen and Valpola 2017) methods use prediction and model-weight ensembling respectively to generate pseudo-labels. FixMatch (Sohn et al. 2020) and FlexMatch (Zhang et al. 2021) use class probability thresholding for pseudo-labeling to achieve state-of-the-art results on semi-supervised classification. Similar techniques have been applied to video action recognition (Xu et al. 2021), image generation (Bodla, Hua, and Chellappa 2018), medical image segmentation (Li et al. 2020, 2018b; You et al. 2022; Lin et al. 2022), and other tasks.

Semi-Supervised Regression

Regression problems are fundamentally different from classification as they involve predicting real numbers in $\mathbb{R}$ instead of class probabilities. Semi-supervised classification methods, which use thresholding functions to select high-probability class pseudo-labels (Sohn et al. 2020; Zhang et al. 2021), cannot be adapted to regression tasks *because there is no equivalent to probability thresholding for real number predictions.* Different formulations must be used instead to quantify prediction uncertainty for regression tasks.

Less attention has been paid to semi-supervised deep regression despite its importance (Jean, Xie, and Ermon 2018; Wetzel, Melko, and Tamblyn 2021; Yin et al. 2022). Semi-supervised regression is especially valuable in medical image analysis, since regression tasks are widely used and annotation costs are high (Ouyang et al. 2020; Hsieh et al. 2021; Dai et al. 2022). COREG (Zhou, Li et al. 2005) is a semi-supervised regression technique originally proposed in 2005 but still commonly used today. Two KNN regressors are co-trained and used to generate pseudo-labels for unlabeled data. Co-training schemes have also been extended to support vector regression by Xu *et al*. (Xu et al. 2011). Graph-based methods proposed in (Timilsina et al. 2021) make use of input proximity for pseudo-labeling. More recent works by Jean *et al*. (Jean, Xie, and Ermon 2018) and Mallick *et al*. (Mallick et al. 2021) make use of deep kernel learning for regression.

One major disadvantage of these methods is that they are primarily designed for structured inputs, where samples consist of one-dimensional tabular data. Feature extractors cannot be trained end-to-end for unstructured inputs such as images and video. Jean *et al*. (Jean, Xie, and Ermon 2018) for example rely on feature extractors pretrained on ImageNet (Deng et al. 2009) to obtain one dimensional embeddings from images. This limits performance as task-specific features cannot be learned (see results in Tables 1 and 4). TNNR (Wetzel, Melko, and Tamblyn 2021) is an alternative method that uses deep networks to predict differences between input pairs. Loop consistency is applied to ensure looped differences sum to zero. Although loop consistency helps regularize training on unlabeled data, inaccurate predictions can limit its effectiveness (see Tables 1 and 4).

Unlike previous works, we address semi-supervised deep regression by improving uncertainty estimation and pseudo-label quality for real number targets. Our UCVME method, which proposes a novel uncertainty consistency loss and variational model ensembling, allows training to be focused on high-quality, robust pseudo-label targets and achieves state-of-the-art results on different regression tasks.

Uncertainty Estimation

Uncertainty estimation is commonly used in semi-supervised learning to adjust for pseudo-label quality of unlabeled samples. UA-MT (Yu et al. 2019) and UMCT (Xia et al. 2020) both use Monte Carlo dropout to estimate pseudo-label uncertainty, which is then used to filter pseudo-labels or weight unlabeled samples for segmentation tasks. Yao *et al*. (Yao, Hu, and Li 2022) and Lin *et al*. (Lin et al. 2022) estimate uncertainty based on prediction differences between co-trained models for segmentation of medical images. Semi-supervised classification methods such as FixMatch (Sohn et al. 2020) and FlexMatch (Zhang et al. 2021) implicitly filter out uncertain pseudo-labels by setting confidence thresholds for predictions.

Uncertainty estimation approaches designed for semi-supervised deep regression have not been explored in existing works however. Although methods such as heteroscedastic regression can be used to estimate uncertainty, it can only be done through joint prediction with the target label (Kaufman 2013). Naïve implementation using pseudo-labels give unreliable estimates, which can leads to inaccurate pseudo-labels being assigned larger weights (see Fig. 5). In this work, we propose a novel uncertainty consistency loss that significantly improves the quality of uncertainty estimates on unlabeled data. This results in more effective uncertainty-based sample weighting and leads to state-of-the-art performance on different semi-supervised deep regression tasks.

Methodology

UCVME is based on two novel ideas: enforcing aleatoric uncertainty consistency to improve uncertainty-based loss weighting, and variational model ensembling for generating high-quality pseudo-labels. We make use of Bayesian

neural networks, which differ from regular neural networks by their usage of aleatoric uncertainty prediction and variational inference (Kendall and Gal 2017). We denote $\mathcal{D} \coloneqq \{(x_i, y_i)\}_{i=1}^{N}$ as the labeled dataset consisting of N samples, where x_i is the input data and y_i is its corresponding label. We denote $\mathcal{D}' \coloneqq \{x'_{i'}\}_{i'=1}^{N'}$ as the unlabeled dataset consisting of input data only. We train two BNNs, f_m where $m \in \{a, b\}$, in a co-training framework and use Monte Carlo dropout for training and inference. We denote $\hat{y}_{i,m}$ as model m's prediction for target label y_i. We denote σ_i^2 as aleatoric uncertainty but predict log-uncertainty $\ln \sigma_i^2$ in practice, which is always done to avoid obtaining negative predictions for variance. We denote predicted log-uncertainty using $\hat{z}_{i,m}$.

Aleatoric Uncertainty Consistency Loss for Improved Heteroscedastic Regression

Aleatoric uncertainty, σ_i^2, refers to uncertainty relating to input data. It is used in BNNs as the variance parameter for heteroscedastic regression loss:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=0}^{N} \frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2} + \frac{\ln \sigma_i^2}{2} \,. \tag{1}$$

where $\hat{y}_i$ is the prediction for target label y_i. Intuitively, the loss function weighs error values dynamically based on aleatoric uncertainty. Samples with high uncertainty are regarded as having lower quality labels with higher noise, and these are given less importance compared to those with greater certainty (Kendall and Gal 2017). Its formal derivation is based on maximum likelihood estimation, assuming observation errors are distributed with different levels of variance (Kaufman 2013). In contrast, standard mean squared error (MSE) loss assumes homoscedastic errors, *i.e.* uncertainty values σ_i^2 have equal variance, which is a more restrictive and unrealistic assumption. We refer interested readers to Sup-1 of the supplementary materials for a review of formal derivations and comparisons.

Heteroscedastic regression can be beneficial for unlabeled data as it allows samples to be weighted based on pseudo-label uncertainty. In practice however, uncertainty prediction is difficult because uncertainty has no ground truth label and must be jointly predicted with the target value. Unstable predictions that do not reflect label quality can adversely affect training by assigning noisier samples with larger weights. *Stable training is even more difficult for unlabeled data* because the target ground truth value is also unavailable, which is why heteroscedastic regression has not been successfully used in existing semi-supervised works. We show this effect in Fig. 5, where we see uncertainty predictions obtained using heteroscedastic regression only can be unreliable.

We observe that aleatoric uncertainty for the same input data *should be equal by definition* and introduce a novel consistency loss to enforce consistent uncertainty predictions between co-trained models. Prediction consistency is known to be an effective regularizer (Chen et al. 2021) and can be applied to both labeled and unlabeled data to improve estimates. By ensuring uncertainty predictions from co-trained models are consistent, we provide an extra training signal in addition to joint estimation with the target label, which helps the model learn more reliable predictions. For labeled inputs, we introduce consistency loss, $\mathcal{L}_{unc}^{lb}$:

$$\mathcal{L}_{unc}^{lb} = \frac{1}{N} \sum_{i=1}^{N} (\hat{z}_{i,a} - \hat{z}_{i,b})^2 \,, \tag{2}$$

which is based on L2 distance. Heteroscedastic regression loss is calculated using the uncertainty predictions:

$$\mathcal{L}_{reg}^{lb} = \frac{1}{N} \sum_{m=a,b} \sum_{i=1}^{N} \left(\frac{(\hat{y}_{i,m} - y_i)^2}{2\exp(\hat{z}_{i,m})} + \frac{\hat{z}_{i,m}}{2} \right) \,. \tag{3}$$

For unlabeled data, ground truth target labels for y are unavailable, which makes joint uncertainty prediction challenging. We instead make use of variational model ensembling to obtain pseudo-labels for log-uncertainty, $\widetilde{z}_i$, which is used as the training target. We describe variational model ensembling for unlabeled samples in the subsection below.

Variational Model Ensembling for Pseudo-label Generation

BNNs use Monte Carlo dropout and variational inference to estimate the distribution of the predictor $\hat{y}$. To reduce prediction noise, we can use ensembling techniques that reduce predictor variance, which can be demonstrated through bias-variance decomposition. The performance of predictor $\hat{y}$ can be evaluated using expected MSE, which we decompose using bias-variance decomposition as follows:

$$E[(\hat{y}_i - y_i)^2] = (E[\hat{y}_i] - y_i)^2 + E[(\hat{y}_i - E[\hat{y}_i])^2] \,, \tag{4}$$

where the first right-hand side term is the bias and the second is the variance. If we take individual sample predictions from variational inference, ${\hat{y}_i}^t$, and obtain an ensemble to form a new predictor $\widetilde{y}_i$, we have:

$$\widetilde{y}_i = \frac{1}{T} \sum_{t=1}^{T} {\hat{y}_i}^t \,, \tag{5}$$

where T is the number of samples used. The expected MSE loss of the predictor $\widetilde{y}_i$ is then:

$$E[(\widetilde{y}_i - y_i)^2] = (E[\widetilde{y}_i] - y_i)^2 + E[(\widetilde{y}_i - E[\widetilde{y}_i])^2] \,. \tag{6}$$

The bias terms of the predictors are equal, but the variance term in equation 6 cannot be greater than in equation 4 because more samples are observed (see Sup-2 of supplementary materials for more detailed derivations). This means predictor $\widetilde{y}_i$ will have expected MSE lower than or equal to $\hat{y}_i$ and will always have higher quality.

Based on this effect, we propose variational model ensembling for generating pseudo-labels on both target value $\widetilde{y}_i$ and log aleatoric uncertainty $\widetilde{z}_i$. Whereas pseudo-labels for co-trained models typically rely on cross-supervision in state-of-the-art approaches (Xu et al. 2021; Chen et al.

2021), we ensemble the average estimate of the co-trained models and apply variational inference:

$$\widetilde{y}_i = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{y}_{i,a}^{\ t} + \hat{y}_{i,b}^{\ t}}{2}\ , \tag{7}$$

$$\widetilde{z}_i = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{z}_{i,a}^{\ t} + \hat{z}_{i,b}^{\ t}}{2}\ . \tag{8}$$

and use this as the pseudo-label for training. Compared to cross-supervision, pseudo-labels calculated using variational model ensembling are *more accurate because of reduced predictive variance* and better reflect the true target and uncertainty values. This is especially important for regression targets because pseudo-labels directly use real number predictions and do not rely on thresholding functions for smoothing. Uncertainty consistency on unlabeled data is then calculated using $\widetilde{z}_i$ as the training target:

$$\mathcal{L}_{unc}^{ulb} = \frac{1}{N'}\sum_{m=a,b}\sum_{i=1}^{N'}(\hat{z}_{i,m} - \widetilde{z}_i)^2\ . \tag{9}$$

Heteroscedastic regression loss for unlabeled data is calculated using $\widetilde{y}_i$ as the target and $\widetilde{z}_i$ as the log-uncertainty:

$$\mathcal{L}_{reg}^{ulb} = \frac{1}{N'}\sum_{m=a,b}\sum_{i=1}^{N'}\left(\frac{(\hat{y}_{i,m} - \widetilde{y}_i)^2}{2\exp(\widetilde{z}_i)} + \frac{\widetilde{z}_i}{2}\right)\ . \tag{10}$$

The improved pseudo-labels lead to more stable heteroscedastic regression, which generates better training signals on unlabeled data.

Overall Semi-Supervised Framework

During training, we calculate heteroscedastic regression loss $\mathcal{L}_{reg}^{lb}$ and aleatoric uncertainty consistency loss $\mathcal{L}_{unc}^{lb}$ using labeled samples. Pseudo-labels for unlabeled data are generated using Eq. 7 and 8 at the start of every training iteration with the most current model weights. The loss values for $\mathcal{L}_{reg}^{ulb}$ and $\mathcal{L}_{unc}^{ulb}$ are calculated for the unlabeled data and jointly optimized with labeled data using the total loss:

$$\mathcal{L} = \mathcal{L}_{reg}^{lb} + \mathcal{L}_{unc}^{lb} + w_{ulb}\,(\,\mathcal{L}_{reg}^{ulb} + \mathcal{L}_{unc}^{ulb}\,)\,, \tag{11}$$

where w_{ulb} is the weighting parameter for unlabeled data. Variational model ensembling is also used for test-time inference to obtain $\widetilde{y}_i$ as the final prediction. Pseudo-code is given in S-Algorithm 1 of the supplementary materials.

Experiments

We demonstrate our method on two semi-supervised deep regression problems: age estimation from photographs and ejection fraction estimation from echocardiogram videos. [1]

[1]Code is available at https://github.com/xmed-lab/UCVME.

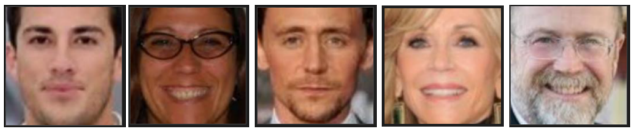


Figure 3: Sample data from UTKFace dataset (Zhang, Song, and Qi 2017) for age estimation. Pre-cropped images are paired with age labels for training.

Age Estimation from Photographs

Age estimation involves predicting a person's age based on their photograph, and is commonly used as a benchmark task for deep regression. Facial images can be easily obtained, but accurate age labels may not always be available given concerns over data privacy. Semi-supervised deep regression methods can provide label-efficient approaches for training.

Dataset We use the UTKFace dataset (Zhang, Song, and Qi 2017) and follow the train-test split in previous works (Cao, Mirjalili, and Raschka 2019). A total of 13,144 images are available for training and 3,287 images for testing. We use a subset of the training dataset for validation. Faces have been pre-cropped and age labels range from 21 to 60 (see Figure 3 for examples). For our semi-supervised setting, we use subsets of the training data as labeled data and the remaining as unlabeled data. Label distributions are shown in S-Fig. 1 of the supplementary materials.

Settings We use ResNet-50 (He et al. 2016) as our encoder and add additional dropout layers after each of the four main residual blocks. The model is trained for 30 epochs using learning rate 10^{-4}, weight decay 10^{-3}, and the Adam optimizer. We use a batch size of 32 for both labeled and unlabeled data. We set dropout probability as 5% and use $T = 5$ for variational inference. We set $w_{ulb} = 10$ which we choose empirically (see S-Table 1 in supplementary materials). Mean absolute error (MAE) and R^2 are used for evaluation on the test set. Experiments are run six times and mean results are reported with standard deviation.

Comparison with state-of-the-art We compare our method with alternative state-of-the-art approaches for semi-supervised regression, specifically COREG (Zhou, Li et al. 2005), SSDPKL (Mallick et al. 2021), and TNNR (Wetzel, Melko, and Tamblyn 2021), and also adapt mean-teacher (Tarvainen and Valpola 2017) and temporal ensembling (Laine and Aila 2016) methods for regression. To highlight the impact of our proposed components, we introduce a baseline method *(Baseline)* that uses two co-trained BNNs with heteroscedastic regression loss, but *without* aleatoric uncertainty consistency loss and variational model ensembling. We perform training under different semi-supervised settings using only 5%, 10%, and 20% of the available training labels. The remaining samples are treated as unlabeled data. For reference, we also show results using the supervised state-of-the-art method by Berg *et al.* (Berg, Oskarsson, and O'Connor 2021) *(RNDB)* for the same settings using reduced labels, as well as on the fully labeled dataset.

MAE Values ↓

Type	Method	Encoder	5% labeled	10% labeled	20% labeled	All labels
Supervised	RNDB (Berg, Oskarsson, and O'Connor 2021)	ResNet50	6.21 ± 0.12	5.69 ± 0.09	5.38 ± 0.10	**4.83 ± 0.06**
Semi-Supervised	Mean-teacher (Tarvainen and Valpola 2017)	ResNet50	6.15 ± 0.08	5.54 ± 0.07	5.29 ± 0.05	-
	Temporal ensembling (Laine and Aila 2016)	ResNet50	6.09 ± 0.07	5.53 ± 0.05	5.25 ± 0.04	-
	SSDPKL (Jean, Xie, and Ermon 2018)	ResNet50	6.08 ± 0.06	5.50 ± 0.01	5.27 ± 0.08	-
	TNNR (Wetzel, Melko, and Tamblyn 2021)	ResNet50	5.94 ± 0.04	5.41 ± 0.11	5.08 ± 0.05	-
	COREG (Zhou, Li et al. 2005)	ResNet50	5.97 ± 0.06	5.39 ± 0.04	4.97 ± 0.03	-
	Baseline	ResNet50	5.92 ± 0.07	5.40 ± 0.03	4.96 ± 0.03	-
	Ours	ResNet50	**5.84 ± 0.06**	**5.26 ± 0.02**	**4.85 ± 0.03**	-

$\mathbf{R}^2$ Values ↑

Type	Method	Encoder	5% labeled	10% labeled	20% labeled	All labels
Supervised	RNDB (Berg, Oskarsson, and O'Connor 2021)	ResNet50	43.8% ± 7.5	51.0% ± 3.1	57.5% ± 2.7	**65.3% ± 0.3**
Semi-Supervised	Mean-teacher (Tarvainen and Valpola 2017)	ResNet50	45.7% ± 1.1	54.3% ± 0.4	58.0% ± 0.5	-
	Temporal ensembling (Laine and Aila 2016)	ResNet50	46.1% ± 1.0	54.2% ± 0.4	58.9% ± 0.3	-
	SSDPKL (Jean, Xie, and Ermon 2018)	ResNet50	46.2% ± 1.3	54.2% ± 0.2	58.1% ± 0.9	-
	TNNR (Wetzel, Melko, and Tamblyn 2021)	ResNet50	48.6% ± 0.3	53.1% ± 1.5	58.6% ± 0.5	-
	COREG (Zhou, Li et al. 2005)	ResNet50	47.4% ± 1.0	56.6% ± 0.4	62.7% ± 0.2	-
	Baseline	ResNet50	47.9% ± 1.1	56.3% ± 0.5	62.5% ± 0.2	-
	Ours	ResNet50	**49.4% ± 0.7**	**57.9% ± 0.3**	**64.3% ± 0.5**	-

Table 1: Comparison with state-of-the-art methods for age estimation from photographs. We use settings where only 5%, 10%, and 20% of training labels are available. "Supervised" methods are only able to use labeled data while "Semi-supervised" methods can use labeled and remaining unlabeled data. "Baseline" method uses two co-trained BNNs without uncertainty consistency loss and variational model ensembling. Bold numbers represent the best result.

The same ResNet-50 encoder (He et al. 2016) is used in all methods for fair comparison. We also modify COREG to use co-trained deep regression models instead of KNN regression and use a pretrained feature encoder for SSDPKL to obtain image features. Additional implementation details are included in Sup-3 of supplementary materials. We show results in Table 1 and visually plot them in Fig. 4.

We can see from Fig. 4 that the supervised approach (blue) under-performs semi-supervised approaches in general. Our method (red) gives the best results and achieves the lowest MAE values for all settings. We also note that our method achieves performance competitive with fully supervised results using only 20% of available training labels (MAE 4.85 *v.s.* 4.83). UCVME therefore effectively reduces reliance on labeled data for deep regression.

Ablation Study We analyze the performance contribution of different components through ablation. We compare results with the baseline model *(Baseline)* after adding uncertainty consistency loss *(Baseline + Con.)*, variational model ensembling *(Baseline + Ens.)*, and both modules *(Ours)* in separate runs to understand the gains from each component. The model is trained with 10% of the training labels and the rest is used as unlabeled data. Results are shown in Table 2.

We can see consistency loss and variational model ensembling have individual contributions and separately reduce MAE by roughly 0.10. Best results are achieved using both.

Impact of consistency loss on uncertainty estimates We analyze the impact of consistency loss on uncertainty estimates by visualizing its relationship with pseudo-label quality. Intuitively, improved uncertainty estimates will show a stronger negative relationship with pseudo-label quality,

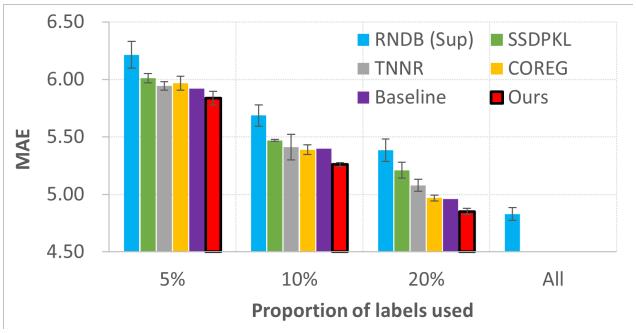


Figure 4: MAE of different state-of-the-art methods for age estimation. Our proposed method (red) consistently achieves the best results under all settings and is competitive with the supervised approach trained with full labels (blue).

Method	Con.	Ens.	MAE↓	$\mathbf{R}^2$ ↑
Baseline			5.40 ± 0.03	56.3% ± 0.5
Baseline + Con.	✓		5.30 ± 0.01	57.6% ± 0.2
Baseline + Ens.		✓	5.31 ± 0.02	57.4% ± 0.2
Ours	✓	✓	**5.26 ± 0.02**	**57.9% ± 0.3**

Table 2: Ablation study with 10% of available labels. Remaining samples are used as unlabeled data. "Baseline" method uses two co-trained BNNs without uncertainty consistency loss and variational model ensembling. "Cons." refers to the use of aleatoric uncertainty consistency loss. "Ens." refers to the use of variational model ensembling.

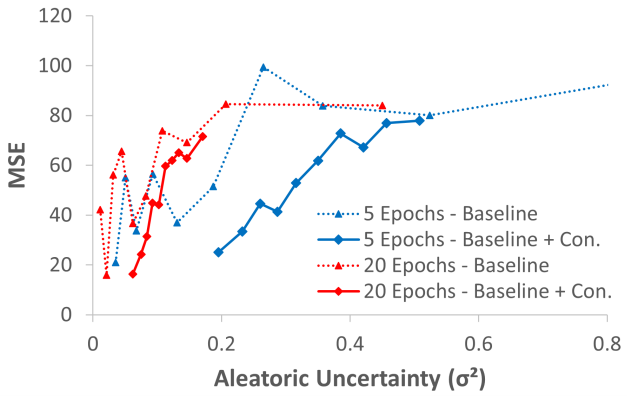


Figure 5: MSE of pseudo-labels plotted against predicted aleatoric uncertainty. Improved uncertainty estimates should display a stronger negative relationship with quality, since higher uncertainty indicates lower quality pseudo-labels with higher MSE. We can see that uncertainty estimates become more reliable after applying uncertainty consistency loss as the relationship between predicted uncertainty and pseudo-label quality is much clearer (solid *v.s.* dashed lines).

since higher uncertainty means more prediction noise and lower quality labels. We obtain pseudo-labels and uncertainty predictions for unlabeled samples using the *Baseline* and *Baseline + Con.* models. Samples are grouped into ten equal bins based on sorted uncertainty predictions. Pseudo-label quality is measured using MSE against the ground truth target value. Average aleatoric uncertainty is calculated for each group. The two values are plotted against each other in Fig. 5 using models trained after five and twenty epochs.

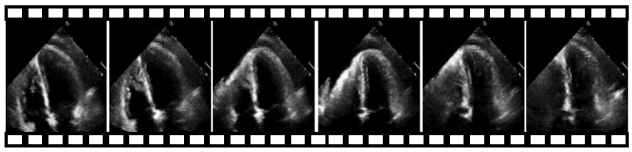
For the *Baseline + Con.* model, we see overall MSE and uncertainty both decrease after training for more epochs (solid red line *v.s.* solid blue line). Pseudo-labels with lower uncertainty have lower MSE and are of higher quality. In contrast, uncertainty predictions from the *Baseline* model are more extreme (dashed lines) and are not significantly reduced with training (dashed red line *v.s.* dashed blue line). The relationship between uncertainty and pseudo-label quality is not strong, which can lead to noisier samples being assigned higher loss weightings. Aleatoric uncertainty consistency loss therefore *improves uncertainty estimates significantly and helps prevent adverse sample weighting*.

Computational Cost We show in Table 3 the computational cost of different semi-supervised deep regression approaches in gigaFLOPS per image (G). We note that the cost of UCVME is dependant on T, the number of iterations used for variational model ensembling, which is set to 5. For reference, we also show the cost from using a single iteration, $T = 1$, which is equivalent to enforcing uncertainty consistency only without using variational model ensembling.

We can see our UCVME method with $T = 1$ incurs the same cost as COREG. Regression performance outperforms COREG however due to the use of uncertainty consistency, which can be seen from results in Tables 1 and 2 (MAE 5.30 *v.s.* 5.39). Using variational model ensembling by setting $T = 5$ leads to more computation but better predictions. Although mean-teacher, temporal ensembling, and SSDPKL require less computation, they do not perform as well.

Method	Computation Cost (G)
Mean-teacher (Tarvainen and Valpola 2017)	4
Temporal ensembling (Laine and Aila 2016)	4
SSDPKL (Jean, Xie, and Ermon 2018)	4
COREG (Zhou, Li et al. 2005)	21
TNNR (Wetzel, Melko, and Tamblyn 2021)	77
Ours w/o variational model ensembling ($T = 1$)	21
Ours ($T = 5$)	49

Table 3: Computation cost of state-of-the-art semi-supervised regression methods in gigaFLOPS per image (G). We also show the cost for our method UCVME without variational model ensembling by setting $T = 1$.



Patient 0X1A58C9DFE12C7953, $LVEF = 44\%$

Figure 6: Sample data from EchoNet-Dynamic (Ouyang et al. 2019). Echocardiogram video sequences are paired with LVEF values.

Ejection Fraction Estimation from Echocardiogram Videos

Left ventricular ejection fraction (LVEF) is the most commonly used medical indicator for diagnosing cardiac disease (Hughes et al. 2021). It is the percentage difference between the maximum and minimum volume of a heart's left ventricle (LV) and measures blood pumping capability. LVEF is manually labeled from echocardiogram video by estimating maximum and minimum volume based on LV segmentations using method of disks (Foley et al. 2012) and finding their percentage difference. An illustration of this process is given in S-Fig. 3 of the supplementary materials for interested readers. State-of-the-art methods for automating LVEF estimation use spatial-temporal models to perform end-to-end regression on raw video. Video regression requires large amounts of labeled data however, and existing methods by Ouyang *et al*. (Ouyang et al. 2020) and Dai *et al*. (Dai et al. 2021) use up to 10,030 samples for training. Like most medical tasks, annotation requires domain expertise and can be costly, motivating the need for label efficient methods.

Dataset We use the EchoNet-Dynamic dataset (Ouyang et al. 2019), which consists of 10,030 echocardiogram videos with LVEF labels (see Fig. 6 for examples). The videos have been rescaled to 112 × 112 pixels and predivided into 7,465 videos for training, 1,288 videos for validation, and 1,277 videos for testing. For our semi-supervised setting, we use subsets of the training labels as our labeled

MAE Values ↓

Type	Method	Encoder	1/16 labeled	1/8 labeled	1/4 labeled	All labels
Supervised	Ouyang *et al*. (Ouyang et al. 2020)	R2+1D	6.04 ± 0.20	5.57 ± 0.21	4.78 ± 0.11	**4.13 ± 3.85**
Semi-Supervised	Mean-teacher (Tarvainen and Valpola 2017)	R2+1D	6.01 ± 0.09	5.51 ± 0.06	4.71 ± 0.07	-
	Temporal ensembling (Laine and Aila 2016)	R2+1D	5.97 ± 0.08	5.52 ± 0.06	4.67 ± 0.06	-
	SSDPKL (Jean, Xie, and Ermon 2018)	R2+1D	6.01 ± 0.04	5.47 ± 0.01	4.68 ± 0.07	-
	TNNR (Wetzel, Melko, and Tamblyn 2021)	R2+1D	5.90 ± 0.11	5.46 ± 0.08	4.79 ± 0.08	-
	COREG (Zhou, Li et al. 2005)	R2+1D	5.94 ± 0.07	5.31 ± 0.02	4.57 ± 0.02	-
	Baseline	R2+1D	5.93 ± 0.10	5.36 ± 0.05	4.58 ± 0.03	-
	Ours	R2+1D	**5.77 ± 0.04**	**5.10 ± 0.05**	**4.37 ± 0.05**	-

$\mathbf{R}^2$ Values ↑

Type	Method	Encoder	1/16 labeled	1/8 labeled	1/4 labeled	All labels
Supervised	Ouyang *et al*. (Ouyang et al. 2020)	R2+1D	55.3% ± 2.6	62.5% ± 2.2	71.6% ± 1.4	**80.4% ± 1.2**
Semi-Supervised	Mean-teacher (Tarvainen and Valpola 2017)	R2+1D	55.1% ± 1.4	62.9% ± 0.7	72.5% ± 0.4	-
	Temporal ensembling (Laine and Aila 2016)	R2+1D	55.2% ± 1.3	62.9% ± 0.7	73.2% ± 0.3	-
	SSDPKL (Jean, Xie, and Ermon 2018)	R2+1D	56.3% ± 1.0	61.2% ± 0.3	74.1% ± 1.0	-
	TNNR (Wetzel, Melko, and Tamblyn 2021)	R2+1D	55.9% ± 1.2	63.4% ± 0.8	73.7% ± 0.6	-
	COREG (Zhou, Li et al. 2005)	R2+1D	55.1% ± 0.7	64.5% ± 0.4	74.1% ± 0.1	-
	Baseline	R2+1D	55.2% ± 1.4	64.9% ± 0.3	74.5% ± 0.1	-
	Ours	R2+1D	**57.8% ± 0.6**	**66.6% ± 0.5**	**76.3% ± 0.6**	-

Table 4: Comparison with state-of-the-art methods for ejection fraction estimation from echocardiogram video. We use settings where only 1/16, 1/8, and 1/4 of training labels are available. "Supervised" methods are only able to use labeled data while "Semi-supervised" methods can use labeled and remaining unlabeled data. "Baseline" method uses two co-trained BNNs without uncertainty consistency loss and variational model ensembling. Bold numbers represent the best result.

data and remaining samples as unlabeled data. Label distributions are given in S-Fig. 2 of the supplementary materials.

Settings We use the R2+1D ResNet encoder (Tran et al. 2018) pretrained on Kinetics 400 (Kay et al. 2017) and add additional dropout layers between the four main residual blocks. We set dropout probability as 5% and use $T = 5$ for variational inference. The model is trained using SGD with 10^{-4} learning rate and 0.9 momentum for 25 epochs. Learning rate is decayed by 0.1 at epoch 15. Clips of 32 frames are sampled from videos at a rate of 1 in every 2 frames for input. Batches of 10 clips are used for labeled and unlabeled videos. We set $w_{ulb} = 10$ which we choose empirically (see S-Table 2 of supplementary materials). We evaluate performance using MAE and R^2. Experiments are run five times and mean results with standard deviation are reported.

Comparison with state-of-the-arts We compare our method with mean-teacher (Tarvainen and Valpola 2017), temporal ensembling (Laine and Aila 2016), COREG (Zhou, Li et al. 2005), SSDPKL (Mallick et al. 2021), TNNR (Wetzel, Melko, and Tamblyn 2021), and our baseline model *(Baseline)*. We perform training under settings where one-sixteenth, one-eighths, and one-quarter of the training labels are used, with the remainder treated as unlabeled data. For reference, we also show results using the supervised method by Ouyang *et al*. (Ouyang et al. 2020) on the reduced labels as well as on the fully labeled dataset. The Kinetics pretrained R2+1D ResNet encoder (Tran et al. 2018) is used in all methods for fair comparison. Additional implementation details are given in Sup-4 of the supplementary materials. Results are shown in Table 4 and plotted in Fig. 7.

Our method consistently achieves the best results for all settings by significant margins. We are also able to achieve an MAE of 4.37 using a quarter of the labels, which is only a relative 5.8% higher than the 4.13 MAE achieved by Ouyang *et al*. on fully labeled data. Our proposed method therefore reduces the number of labels required for training which is highly valuable for medical regression tasks.

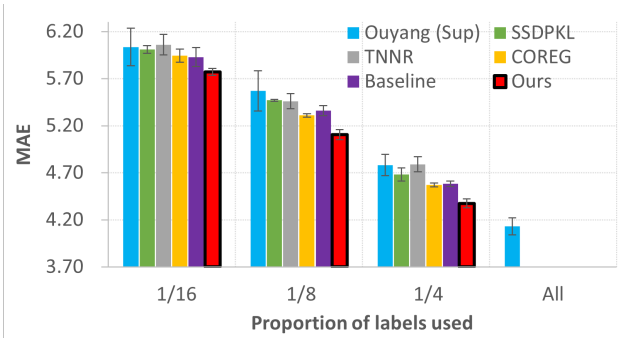


Figure 7: MAE of different state-of-the-art methods for ejection fraction estimation. Our proposed method (red) consistently outperforms alternatives by significant margins.

Conclusion

In this work, we introduce a novel Uncertainty-Consistent Variational Model Ensembling (UCVME) method for semi-supervised deep regression. Our method improves training on unlabeled data by adjusting for pseudo-label quality and improving pseudo-label robustness. We introduce a novel consistency loss on uncertainty estimates, which we demon-

strate significantly improves heteroscedastic loss weighting, especially for unlabeled samples. We also use variational model ensembling to reduce prediction noise and generate better training targets for unlabeled data. Our method has strong theoretical support and can be applied to different tasks and datasets. We demonstrate this using two deep regression tasks based on image and video data and achieve state-of-the-art performance for both. Results are also competitive with supervised methods using full labels. UCVME is therefore a valuable method for reducing the amount of labels required for deep regression tasks.

Acknowledgements

The work described in this paper was supported by a grant from Hong Kong Research Grants Council General Research Fund (GRF) (16203319), a grant from HKUST-BICI Exploratory Fund under HCIC-004, and a grant from Hong Kong Innovation and Technology Commission (Project no. ITS/030/21).

References

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8): 1798–1828.

Berg, A.; Oskarsson, M.; and O'Connor, M. 2021. Deep ordinal regression with label diversity. In *ICPR*, 2740–2747. IEEE.

Bodla, N.; Hua, G.; and Chellappa, R. 2018. Semi-supervised FusedGAN for conditional image generation. In *ECCV*, 669–683.

Cao, W.; Mirjalili, V.; and Raschka, S. 2019. Rank-consistent ordinal regression for neural networks. *arXiv preprint arXiv:1901.07884*, 1(6): 13.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *CVPR*, 2613–2622.

Dai, W.; Li, X.; Chiu, W. H. K.; Kuo, M. D.; and Cheng, K.-T. 2021. Adaptive Contrast for Image Regression in Computer-Aided Disease Assessment. *IEEE Transactions on Medical Imaging*, 41(5): 1255–1268.

Dai, W.; Li, X.; Ding, X.; and Cheng, K.-T. 2022. Cyclical Self-Supervision for Semi-Supervised Ejection Fraction Prediction from Echocardiogram Videos. *arXiv preprint arXiv:2210.11291*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Dosovitskiy, A.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Foley, T. A.; et al. 2012. Measuring left ventricular ejection fraction-techniques and potential pitfalls. *Eur Cardiol*, 8(2): 108–114.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hsieh, C.-I.; et al. 2021. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nature communications*, 12(1): 1–9.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Unsupervised learning of discriminative attributes and visual representations. In *CVPR*, 5175–5184.

Hughes, J. W.; et al. 2021. Deep learning evaluation of biomarkers from echocardiogram videos. *EBioMedicine*, 73: 103613.

Jean, N.; Xie, S. M.; and Ermon, S. 2018. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. *NeurIPS*, 31.

Kaufman, R. L. 2013. *Heteroskedasticity in regression: Detection and correction*. Sage Publications.

Kay, W.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; and Heng, P.-A. 2018a. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE transactions on medical imaging*, 37(12): 2663–2674.

Li, X.; Hu, X.; Qi, X.; Yu, L.; Zhao, W.; Heng, P.-A.; et al. 2021. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 40(9): 2284–2294.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; and Heng, P.-A. 2018b. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *BMVC*.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 523–534.

Lin, Y.; Yao, H.; Li, Z.; Zheng, G.; and Li, X. 2022. Calibrating Label Distribution for Class-Imbalanced Barely-Supervised Knee Segmentation. In *MICCAI*.

Mallick, A.; Dwivedi, C.; Kailkhura, B.; Joshi, G.; and Han, T. Y.-J. 2021. Deep kernels with probabilistic embeddings for small-data learning. In *Uncertainty in Artificial Intelligence*, 918–928. PMLR.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 69–84. Springer.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 12674–12684.

Ouyang, D.; et al. 2019. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS*.

Ouyang, D.; et al. 2020. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802): 252–256.

Sohn, K.; et al. 2020. Fixmatch: Simplifying SSL with consistency and confidence. *NeurIPS*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Timilsina, M.; Figueroa, A.; d'Aquin, M.; and Yang, H. 2021. Semi-supervised regression using diffusion on graphs. *Applied Soft Computing*, 104: 107188.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 6450–6459.

Wetzel, S. J.; Melko, R. G.; and Tamblyn, I. 2021. Twin Neural Network Regression is a Semi-Supervised Regression Algorithm. *arXiv preprint arXiv:2106.06124*.

Xia, Y.; Liu, F.; Yang, D.; Cai, J.; Yu, L.; Zhu, Z.; Xu, D.; Yuille, A.; and Roth, H. 2020. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *CVPR*, 3646–3655.

Xu, S.; An, X.; Qiao, X.; Zhu, L.; and Li, L. 2011. Semi-supervised least-squares support vector regression machines. *Journal of information & computational science*, 8(6): 885–892.

Xu, Y.; et al. 2021. Cross-Model Pseudo-Labeling for Semi-Supervised Action Recognition. *arXiv preprint arXiv:2112.09690*.

Yang, T.-Y.; Chen, Y.-T.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, 1087–1096.

Yao, H.; Hu, X.; and Li, X. 2022. Enhancing Pseudo Label Quality for Semi-Supervised Domain-Generalized Medical Image Segmentation. In *AAAI*.

Yin, Y.; Cai, Y.; Wang, H.; and Chen, B. 2022. FisherMatch: Semi-Supervised Rotation Regression via Entropy-based Filtering. In *CVPR*, 11164–11173.

You, C.; Zhao, R.; Staib, L. H.; and Duncan, J. S. 2022. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In *MICCAI*, 639–652. Springer.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer.

Zhang, B.; et al. 2021. Flexmatch: Boosting SSL with curriculum pseudo labeling. *NeurIPS*.

Zhang, L.; Qi, G.-J.; Wang, L.; and Luo, J. 2019. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *CVPR*, 2547–2555.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 5810–5818.

Zhou, Z.-H.; Li, M.; et al. 2005. Semi-supervised regression with co-training. In *IJCAI*, volume 5, 908–913.