Semi-supervised Deep Transfer for Regression without Domain Alignment

Mainak Biswas[1,2], Ambedkar Dukkipati[3], Devarajan Sridharan[2,3]
[1]Brain, Computation and Data Sciences, [2]Centre for Neuroscience, [3]Computer Science and Automation
Indian Institute of Science, Bangalore - 560012
mainakbiswas@iisc.ac.in, ambedkar@iisc.ac.in, sridhar@iisc.ac.in

Abstract

Deep learning models deployed in real-world applications (e.g., medicine) face challenges because source models do not generalize well to "domain-shifted" target data. Many successful domain adaptation approaches require full access to source data or reliably labeled target data. Yet, such requirements are unrealistic in scenarios where source data cannot be shared either because of privacy concerns or because it is too large, and incur prohibitive storage or computational costs. Moreover, resource constraints may limit the availability of labeled targets. We illustrate this challenge in a neuroscience setting where source data are unavailable, labeled target data are meager, and predictions involve continuous-valued outputs. We build upon Contradistinguisher (CUDA), an efficient framework that learns a shared model across the labeled source and unlabeled target samples, without intermediate representation alignment. Yet, CUDA was designed for unsupervised DA, with full access to source data, and for classification tasks. We develop CRAFT – a Contradistinguisher-based Regularization Approach for Flexible Training – for source-free (SF), semi-supervised transfer of pretrained models in regression tasks. We showcase the efficacy of CRAFT in two neuroscience settings: gaze prediction with electroencephalography (EEG) data and "brain age" prediction with structural MRI data. For both datasets, CRAFT yielded up to 9% improvement in root-mean-squared error (RMSE) over fine-tuned models when labeled training examples were scarce. CRAFT leveraged unlabeled target data and outperformed four competing state-of-the-art source-free domain adaptation models by more than 3%. Lastly, we demonstrate the efficacy of CRAFT on two other real-world, regression benchmarks. We propose CRAFT as an efficient approach for source-free, semi-supervised deep transfer for regression that is ubiquitous in biology and medicine.

Keywords: deep transfer, source-free domain adaptation, semi-supervised learning, continuous label prediction, saccades, brain age

1. Introduction

For the successful application of deep learning models in the real-world, they must be robust to "domain shift" [50]. For example, in biology and medicine, supervised deep learning models are often trained on large, high-quality source datasets collected in resource-rich settings [11, 39]. However, naive transfer or even simple finetuning, fails to produce high accuracies when tested with smaller target datasets of mixed quality collected [2, 45]. As a result, state-of-the-art domain adaptation (DA) approaches, which enable the source models to generalize successfully to target datasets, must be employed. Many of these approaches involve explicit alignment of source and target, while learning a shared predictive model for both domains [17, 19, 40].

A particular challenge occurs with DA when the source dataset is unavailable. This can happen either because of privacy or proprietary concerns, such as with medical images or patents [43]. Alternatively, source datasets may be too large, incurring prohibitive storage and computation costs in low-resource settings [39]. Source-free (SF) domain adaptation methods that address these challenges are getting increasingly popular [18, 32]. Such methods generalize the source model to unlabeled or partially target data (unsupervised or semi-supervised DA) with one of two approaches: data-based or model-based. While data-based methods create surrogate source data, model-based methods adapt the source model to the target domain (see Section 2).

We address model-based SF-DA in a semi-supervised setting. Specifically, we focus on neuroscience applications where source data are inaccessible, and target labels are reliably available only for a few samples. Moreover, we tackle regression problems due to their relevance in neuroscience – like predicting brain age (Fig. 1C) [39], saccades [29], neural activity [20], and stimulus orientation [8]. We build upon Contradistinguisher [3, 4], a recently-proposed tool for effective unsupervised DA (CUDA). Unlike popular representation alignment DA methods [19, 40], CUDA directly learns a common model for both domains. Yet, CUDA requires full access to source data and addresses classification tasks. Therefore, it is not directly applicable to our problem.

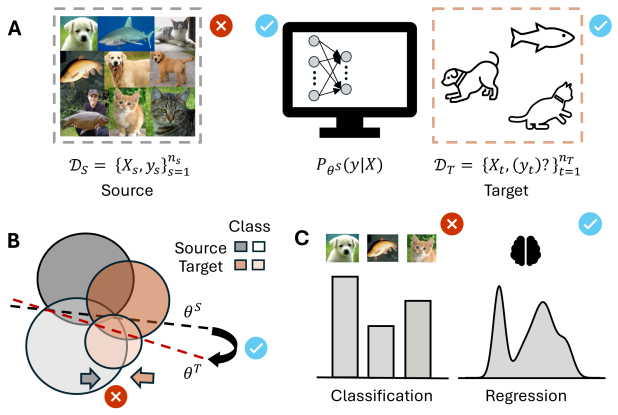

Figure 1. Schematic of Semisupervised Deep Transfer without Domain Alignment. (**A**) The approach seeks to transfer a pretrained model to a target domain with limited labeled data when the source dataset is unavailable. (**B**) Because the source dataset is unavailable, we tune the decision boundaries without explicitly aligning the representation of the domains, combining a supervised loss and the CUDA loss (regularizer). (**C**) The algorithm targets deep transfer in regression tasks, like brain age prediction.

Here, we leverage CUDA to develop a new algorithm – Contradistinguisher-based Regularization Approach for Flexible Training (CRAFT) which performs deep transfer, without domain alignment, even when the source dataset is unavailable (Fig. 1A-B). We also extend CUDA to a regression setting. Previous studies approach regression tasks naively by binning continuous labels into discrete bins [40], sometimes employing a rank-ordinal objective [7]. Regardless, the efficacy of such approaches may depend critically on bin sizes. Our CRAFT model uses a principled approach to extend CUDA to predictions involving continuous-valued outputs. We employ binning as a practical strategy to generate and optimize pseudo-labels, without constraining the model to produce discrete outputs. Our study makes the following main contributions.

- First, we extend CUDA to source-free DA for regression tasks involving continuous-valued outputs.
- Second, we formulate a theoretically-motivated semi-supervised objective, by combining a supervised and unsupervised losses with a tunable hyperparameter (α) (Section 3). We also derive the unsupervised loss as a maximum entropy prior to model parameters (Appendix A.2).
- Third, we show the effectiveness of CRAFT on two neuroscience challenges – gaze prediction from electroencephalographic (EEG) [29] data and brain age prediction from magnetic resonance imaging (MRI) data [45].
- Fourth, we also apply CRAFT to two real-world regression benchmarks involving people counting [44, 48] and tumor size prediction [6, 16].
- Fifth, we compare CRAFT against state-of-the-art SF-DA methods [15, 26, 35, 37] and show that the best improvements occur as unlabeled target samples increases, demonstrating its efficacy for semi-supervised SF-DA.
- Finally, we perform a careful analysis of computational complexity and overheads and demonstrate numbers competitive with sota for our model.

2. Related Work

Domain alignment methods. Popular Unsupervised Domain Adaptation (UDA) approaches rely on learning common intermediate representations across source and target domains. Domain Adversarial training of Neural Networks (DANN) [19] uses a discriminator to learn common representations across both domains adversarially. In Importance Weighting [40] (IW), higher importance is afforded to losses corresponding to source samples more likely to belong to the target domain. Central Moment Discrepancy (CMD) [17] minimizes the central moments of the latent representations between the domains, while others optimize their Maximum Mean Discrepancy (MMD) [22, 34]. Despite being effective, they cannot be used when the source is unavailable and will not be discussed further.

Source-free UDA. To address DA scenarios where source data and its target labels are unavailable, several source-free unsupervised DA approaches have been recently proposed. Broadly, they are categorized as: Data-based and Model-based [32]. Data-based methods generate a source-like distribution [14, 30]. However, they often depend on the availability of generative models trained on the source, which could be unavailable. For example, SF-UDA through domain-specific perturbation (AUGFree) [51] assumes the existence of domain-invariant features for prediction. Since the source is unavailable, AUGFree perturbs the target data and then learns a transformation to the invariant features by adversarially aligning the target and the perturbed target.

On the other hand, model-based methods finetune the source model on the target using strategies like contrastive learning [27], pseudo-labeling [26], or entropy minimization [36]. They are often prone to overfitting to pseudo-labels. For example, Class-Balanced Multicentric Dynamic Prototype Strategy for SF-UDA (BMD) [42] follows a pseudo-labeling strategy using class prototypes based on the source model predictions, and in turn uses these labels for training. Source Data-Absent UDA Through Hypothesis Transfer (SHOT++) [33] combines three objectives – a) maximizing mutual information between the predictions and inputs, b) self-supervised loss based on prototype-based clusters, and c) a Mixup-based (See SF-SSDA) loss [37]. Attracting and Dispersing (AaD) [52] defines a neighboring cluster (K nearest neighbors) and a randomly selected background cluster for each sample; and maximizes the log-likelihood of the ratio of their distributions.

Target-agnostic SF UDA for Regression Tasks (TASFAR) [26] works on the principle that prediction uncertainty is correlated with prediction error. It uses Monte-Carlo dropout to estimate label uncertainty and divides the target dataset into confident and uncertain samples. A discrete density map is estimated using the confident data, which is then used as weights to re-calibrate the pseudo-labels of uncertain data. SF-UDA to measurement Shift via Bottom-Up Feature Restoration (BUFR/DataFree) [15] is designed on the principle of intermediate feature alignment of the source and the target distributions. During adaptation, the model minimizes the symmetric KL divergence between feature distributions of source (fixed) and target data.

Here, we compare our model's performance against two of the above methods - TASFAR and DataFree. TASFAR was selected as it is a pseudo-labeling-based method suitable for regression. On the other hand, DataFree is a pseudo-label-free feature alignment method, not constrained to classification tasks. By adding a supervised loss (Section 4.2), we evaluate these as semi-supervised models.

Source-free SSDA. Source-free semi-supervised domain adaptation (SF-SSDA) models best address the scenario of SF-DA when few target labels are available. For example, Generalized SF-SSDA [1] minimizes source loss alongside an unsupervised target loss which maximizes the cosine similarity between predictions of k-nearest neighbors. The model is then transferred by adversarial training based on source features generated using a Conditional VAE. Similarly, [23] proposes generating surrogate source data from target samples using gradient-based optimization. Learning from Different Samples for SF-SSDA [28] minimizes the contrastive loss on the predicted probabilities of the target domain. Furthermore, it minimizes supervised and contrastive loss on a set comprising labeled, and unlabeled data with low uncertainty. Likewise, Mutual Enhancement training for SS Hypothesis transfer (MESH) [36] minimizes the supervised loss on the labeled samples and entropy on pseudo-labeled samples. Finally, the target manifold is smoothened by making predictions robust to perturbations.

Progressive Mixup [37] is trained on augmentations by convex combinations of data on which the source model is confident and ones on which it is uncertain (hybrid mixup). Similarly, another set is obtained by convex combinations of samples in the (pseudo) labeled dataset (self-mixup). In contrast, Bilateral-Branch Consistency Network (BBCN) [35] computes prototypes as the weighted mean of feature representations of the target dataset by the source model. Pseudo-labels are selected based on the proximity of unlabeled data to these prototypes. A separate target branch is trained using the pseudo-labels, and a consistency loss is added between the representations of both branches. The need for class-level prototype generation makes it challenging to adapt this model to regression tasks.

We compare the performance of CRAFT with two SF-SSDA methods – SF-SSDA via progressive Mixup [37], and BBCN [35], as they have little dependence on the source distribution and its derivatives. Moreover, these methods are based on complementary ideas, and the latter [35] has been applied specifically in medical imaging, for tuberculosis recognition.

3. The CRAFT Model

3.1. Model formulation

Background. Our objective is to develop an approach for source-free semi-supervised domain adaptation, for regression tasks. For this, we build upon the *Contradistinguisher-based Unsupervised Domain Adaptation (CUDA)* model [4]. CUDA obviates the need for learning a common intermediate representation across the source and the target datasets. It optimizes a joint distribution over the target features and category labels to adapt the source model. Because CUDA is intended for unsupervised domain adaptation (UDA), target labels are unavailable. Hence, the algorithm uses a two-step joint optimization over the target labels and model parameters. First, it generates pseudo-labels on the target dataset from the source model [21]. Then, it updates the model parameters by fixing the pseudo-labels.

In this study, we extend CUDA to a source-free semi-supervised setting (SF-SSDA). Briefly, to the supervised label loss, we add a 'regularized", unsupervised loss as a prior on the model parameters. As we will show in the next section, the model can be flexibly applied, either in SF-UDA case or in the SF-SSDA case, depending on the weightage given to the supervised loss. We, moreover, extend the model to continuous label prediction. We call our new formulation CRAFT - **C**ontradistinguisher-based **R**egularization **A**pproach for **F**lexible **T**raining (CRAFT).

Objective. We consider a task in which a source model θ^s trained on a source dataset – involving continuous label prediction (regression) – is already available to us. The target dataset comprises i.i.d samples from an unknown distribution $p^t(\mathbf{x}, y)$: $\mathcal{D}^t = \{\mathbf{x}_i^t, y_i^t\}_{i=1}^{N_l}, \{\mathbf{x}_i^t\}_{i=1}^{N_{ul}}$ where $\mathbf{x}_i^t \in \mathbb{R}^d$ is a feature-vector, and $y_i^t \in \mathbb{R}$ is its corresponding label. Target training data are partially and sparingly labeled; we define the proportion of unlabeled samples (N_{ul}) to the total number of samples (N), as $N_{ul} : (N_l + N_{ul})$.

We formulate, first, the supervised objective. For continuous label prediction problems, the conditional density of the label given the features is assumed to be normally distributed. The mean is modeled using a neural network $f(.;\theta)$, parameterized by θ, as: $p(y^t|\mathbf{x}^t, \theta) = \mathcal{N}(y^t; f(\mathbf{x}^t;\theta), c)$, and the variance, c is assumed to be constant as it simplifies the optimization. In the final formulation, the variance occurs as the weight of the unsupervised loss, α (equation 3); therefore, only one of the two

hyperparameters (c or α) needs to be tuned. Hence, the log-likelihood can be written as:

$$\log p(\mathcal{D}^t|\theta) \propto -\sum_{i=1}^{N_l}(y_i^t - f(\mathbf{x}_i^t;\theta))^2 \quad (1)$$

The unsupervised objective is essentially identical to the CUDA objective and is defined as a joint log density over the labels and model parameters as follows:

$$\log q(\mathbf{x}^t, y^t|\theta) = \log \left[\frac{p(y^t|\mathbf{x}^t,\theta)}{\sum_{i=1}^{N} p(y^t|\mathbf{x}_i^t,\theta)} p(y^t)\right] \quad (2)$$

This joint distribution seeks to reduce the model's predictive bias when faced with a distribution shift in the target domain (see Fig. 1B). The denominator, which reflects the total probability of predicting a particular class, normalizes for biases originating from the source data. In other words, model predictions are normalized as $p(y^t|\mathbf{x}^t,\theta)/\sum_{i=1}^{N} p(y^t|\mathbf{x}_i^t,\theta)$, such that no class has a biased probability of being predicted. Finally, the model predictions align with the target label distribution, using the supplied marginal $p(y^t)$.

Note, that this objective does not depend on whether the samples are labeled or not. Optimizing the joint distribution: 1) normalizes inherent bias towards one label over the other by calibrating the prediction for a label with its total weight in the dataset, 2) and explicitly trains q to match the marginals – i.e., $\sum_{i=1}^{N} q(\mathbf{x}_i^t, y^t|\theta) = p(y^t)$, thereby matching high-level statistics like per-class proportions in the training distribution. Importantly, with an informative prior $p(y^t)$ that specifies the putative distribution of the target labels, pseudo-label prediction can be biased toward labels that are more likely (see section 3.2).

Combining the supervised and the unsupervised objective, the semi-supervised CRAFT objective is defined as:

$$\mathcal{L}(\mathcal{D}^t,\theta) = \sum_{i=1}^{N_l} \log p(y_i^t|\mathbf{x}_i^t,\theta) + \alpha \sum_{i=1}^{N} \log q(\mathbf{x}_i^t, y_i^t|\theta) \quad (3)$$

In the above equation, α is the weight associated with the unsupervised objective and is tuned as a hyperparameter on a held-out cross-validation set during training. For deep-transfer, model parameters are initialized to θ^s derived from the source model, and then gradient descent is performed to obtain θ^t.

Theoretical motivation. Artificial Neural Networks (ANNs) with hypotheses space $\mathbf{\Theta}$ are trained to learn a parametric approximation $p(\mathbf{x}, y|\theta)$, $\theta \in \mathbf{\Theta}$, such that the log-likelihood of the training data given the parameters is maximized. However, training ANNs often does not account for prior distributions over model parameters. In Appendix A.1, we demonstrate that the CRAFT semi-supervised objective can be theoretically derived as MAP estimation of the parameters ($\log p(\theta|\mathcal{D})$), which we write as a sum of the supervised objective, $\log p(\mathcal{D}|\theta)$, and a prior on the model parameters, $p(\theta)$. Specifically, we demonstrate, using the principle of entropy maximization [21], that the unsupervised objective is related to the prior on the model parameters.

3.2. Model optimization

We optimize the supervised loss of the CRAFT objective with conventional gradient descent, jointly optimizing both components in the loss (equation 3). The unsupervised component of the loss is dependent on the labels y^t, which are unavailable for the vast majority of target data. To tackle this, we employ a 2-step optimization process.

Pseudo-label selection. As the first step, we fix the model parameters θ and select the label that maximizes the joint distribution for each datapoint:

$$\tilde{y}_i^t = \underset{y^t \in \mathcal{Y}}{\text{argmax}}\ q(\mathbf{x}_i^t, y^t|\theta) \quad (4)$$

where $\tilde{y}_i$ the pseudo-label for feature vector $\mathbf{x}_i$. For classification tasks, the optimization in equation 4 can be achieved by selecting the best label from a finite set of class labels that maximizes q [4]. But, for regression problems such as ours, class labels map to real numbers spanning the range of the model outputs $f(.;\theta) : \mathbb{R}^d \mapsto [a, b]$. To address this, we rewrite the above equation under the assumption that the continuous-valued labels are distributed normally, with mean $f(\mathbf{x}^t;\theta)$, and variance c. Hence, the above equation can be written as:

$$\tilde{y}_i^t = \underset{y^t \in \mathcal{Y}}{\text{argmax}}\ \frac{\mathcal{N}(y^t; f(\mathbf{x}_i^t;\theta), c)p(y^t)}{\sum_{l=1}^{N} \mathcal{N}(y^t; f(\mathbf{x}_l^t;\theta), c)} \quad (5)$$

Equation 5 can be optimized with gradient ascent with respect to y^t for the objective $\log q(\mathbf{x_i}^t, y^t|\theta)$. But, the pseudo-labels must be estimated before gradient descent can be performed on the model parameters (see next step, equation 6). This requires nested gradient descent, which increases computational complexity. Moreover, to employ an informative prior over target labels, we fit a class of mixture models (described in Appendix A.3) to estimate $p(y)$ from the data; this makes backpropagation more cumbersome. To overcome these challenges, we divide the entire range of labels into small, discrete bins whose midpoints are the candidate pseudo-labels. Hence, this step in the optimization can be achieved faster pseudo-label selection from a discrete set. Note that this approach does not transform the model outputs to be discrete; it merely represents a convenient optimization strategy to update the parameters. In the experimental section, we provide a recipe for choosing bin sizes and study the impact of varying bin sizes.

Maximization. In the second step, we optimize the model parameters by using the labeled data $\{\mathbf{x}_i^t, y_i^t\}_{i=1}^{N_l}$ and fixing the pseudo-labels $\{\mathbf{x}_i^t, \tilde{y}_i^t\}_{i=1}^{N}$, as:

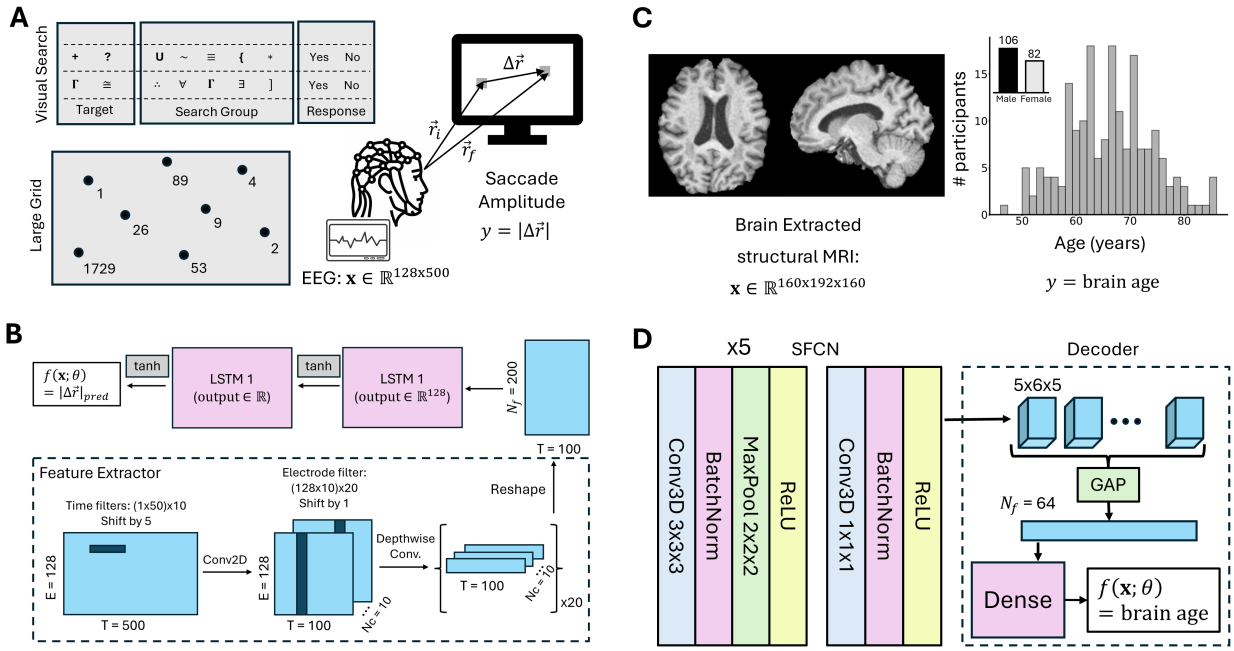

Figure 2. (**A**) Schematic of gaze prediction using EEG signals. (Left) The task schematic of the Visual Search and Large Grid datasets from the EEGEyeNet benchmark. (Right) EEG and eye tracking data were recorded [29]. (**B**) Architecture of the EEGNet-LSTM model, used to decode from EEG data. The parsimonious architecture can be used to decode saccades from EEG data. (**C**) Schematic of the "brain age" prediction task using structural MRI scans. (Left) A sample preprocessed structural MRI scan by SFCN. (right) The age and gender (inset) distribution of the TLSA dataset. (**D**) Model architecture of the SFCN-based decoder. The SFCN backbone was used to extract features from the MRI scan, which were used by a decoder block to predict brain age.

$$\theta^* = \underset{\theta \in \mathbf{\Theta}}{\operatorname{argmax}} \sum_{i=1}^{N_l} \log p(y_i^t | \mathbf{x}_i^t, \theta) + \alpha \sum_{i=1}^{N} \log q(\mathbf{x}_i^t, \tilde{y}_i^t | \theta)$$

After incorporating distributional assumptions in the above equation and subsuming constants with respect to model parameters (e.g., priors on the labels $p(y^t)$):

$$\theta^* = \underset{\theta \in \mathbf{\Theta}}{\operatorname{argmax}} -\sum_{i=1}^{N_l} (y_i^t - f(\mathbf{x}_i^t;\theta))^2 - \alpha \Big(\sum_{i=1}^{N} \big(\tilde{y}_i^t - f(\mathbf{x}_i^t;\theta)\big)^2 - \sum_{i=1}^{N} \log \sum_{l=1}^{N} \exp\left(-(\tilde{y}_i^t - f(\mathbf{x}_i^t;\theta))^2\right)\Big) \quad (6)$$

In practice, each update is performed with a batch of labeled and unlabeled data. First, pseudo-labels are computed for the unlabeled data. Then, keeping the pseudo-labels fixed, the second and third terms of this objective are optimized, while the labeled data are used to optimize the first term, concurrently. Note that the optimization is performed with the model parameters initialized to θ^s, derived from the source model.

Intuitively, the supervised loss in Equation 6 encourages model predictions to align with their true labels; because model parameters were initialized with the source model, this acts as an implicit regularizer discouraging the adapted model from deviating too far from the source model. The second term forces dissimilar predictions for data points with different pseudo-labels. In other words, providing samples with dissimilar label values during training produces disparate representations for data points with distant labels, which in turn helps to learn a better regression line. We use the log-sum-exponent trick to avoid numerical instability during optimizing of the second term.

4. Datasets and Competing Methods

We employ CRAFT to address challenges relating to data scarcity, missing labels, and continuous label prediction in neuroscience.

4.1. Datasets and models

4.1.1. Gaze prediction from EEG signals

Saccades are rapid ballistic eye movements, putatively linked to higher-order cognitive functions like attention and working memory [5, 13]. Decoding eye movements from neural data could help neuroscientists develop better models of these complex processes. With this objective, the EEGEyenet benchmark dataset [29] was collected from 356

participants. It contains $\sim$ 47.5 hours of 128-channel electroencephalogram (EEG) recordings with eye-tracking data collected concurrently.

Here, we predict saccade amplitudes from these EEG data. We consider two datasets – the "Large Grid" and the "Visual Search" datasets – from the EEGEyeNet benchmark [29] (Fig. 2A). In the Large Grid paradigm, participants made saccades to a target location cued by a dot on the screen. In the Visual Search paradigm, participants searched for a specific symbol in the display and, thus, made saccades during the search.

Preprocessing. Each dataset was preprocessed using the pipeline described in the benchmark (details in Appendix B.1). We decode saccade magnitudes ($y \in \mathbb{R}^+$) using the 128-channel EEG timeseries, sampled at 500Hz ($\mathbf{x} \in \mathbb{R}^{128\text{x}500}$), for corresponding trials. Table 5 (see Appendix B.1) lists the details of the train, cross-validation and test distributions for these datasets. The Visual Search paradigm has more training data and was used to train the source model; this was then transferred to predict saccade amplitudes in the Large Grid dataset (target).

Models. The EEGEyeNet benchmark showed that Pyramidal CNNs [47] were most effective at predicting saccade amplitude. As an alternative to 2D convolutional networks, we also evaluate the performance of a novel end-to-end Long-Short-Term-Memory (LSTM) model. This model is trained on features extracted by separable convolutions along time and electrode dimensions, following EEGNet [31] (Fig. 2B, more details in the Appendix B.1).

4.1.2. "Brain age" prediction from MRI scans

"Brain age" is the biological age of the brain and is an important biomarker for cognitive health. Studies have shown that accelerated brain aging is associated with serious cognitive impairments like Alzheimer's disease [9]. Hence, estimating brain age is a critical problem in neuro-medicine.

As a first step, attempts have been made to decode the chronological age of healthy participants using T1-weighted structural MRI scans [41]. A 3D-CNN model (see next, Models) was trained using the large UK-Biobank dataset (N=14,503, 44-80 years, mean age=52.7 years) [39]. We seek to transfer this pretrained model to decode age in a small Indian cohort – the TATA Longitudinal Study of Aging (TLSA) (N=188, 46-88 years, mean age = 66 years) [45], without access to the UKBiobank data.

Preprocessing. We followed a preprocessing pipeline identical to [41]. The T1w-MRI scan was brain-extracted, bias-corrected [46], and registered to standard MNI-152 space [38], and cropped to get preprocessed scans $\mathbf{x} \in \mathbb{R}^{160\text{x}192\text{x}160}$, which was used to predict brain age.

Models. Features from the pretrained Simple Fully Convolutional Network (SFCN, Fig. 2D) were used to make the predictions with a regression model; more details are provided in Appendix B.2) [41]. The entire model was then adapted to the target dataset (TLSA).

We also apply CRAFT to people counting and tumor prediction tasks (see Section B.3, B.4 for dataset details).

4.2. Competing DA Approaches

Baseline model. We define a naive baseline model that always predicts the mean label of the training set, ignoring the input features of the test samples.

Transfer Learning (TL). We naively adapt the source model by minimizing the supervised loss on the labeled samples in the target dataset.

TASFAR. This is an SF-UDA model (See Section 2) for regression problems. It depends on two statistics from the source dataset – the threshold for confident samples and the mapping between uncertainty to prediction errors. Because the source is unavailable, we compute these using the TL model on the target. For a fair comparison, we add the supervised loss on the labeled data to the unsupervised loss, weighted by the same α as in our model.

DataFree. This is another SF-UDA model (See Section 2). The original model depends on the distribution of intermediate latents on the source distribution. Again, because the source data is unavailable, we computed latent distribution using the TL model and the labeled target data. As before, the objective was modified to accommodate a supervised loss (same α as in our model). Bin sizes for both TASFAR and DataFree were selected to be the same as in CRAFT.

Progressive Mixup. This is a semi-supervised, source-free domain adaptation tool (See Section 2). We adapt the model to the regression task by replacing cross-entropy losses with mean-squared error loss. Predictive entropy was replaced by variance to estimate uncertainty.

BBCN. Bilateral Cycle Consistency is a prototype-based clustering SF-SSDA paradigm (See Section 2). It depends on class-level prototypes for each model. To extend it to regression, we discretize the output space into bins of the same width as the other methods.

Metrics. We quantify the prediction performance of continuous valued outputs using two metrics – Root Mean Squared Error (RMSE) and Percentage-Bend Correlation [49] coefficient (R). A *lower* RMSE reflects a closer correspondence between the predicted and actual labels, a *higher* R-value quantifies better predictions along the best-fit line between the actual and predicted labels.

In subsequent sections (5.1 and 5.2), we predict saccade amplitude from EEG and brain age from structural MRI scans, respectively.

5. Experiments: Results

5.1. Saccade Amplitude Prediction

Firstly, we trained an EEGNet-LSTM model on the Visual Search source dataset (RMSE=67.03, R=0.91). This novel

architecture outperformed the benchmark model (Pyramidal CNNs) by $\sim$ 16% (benchmark: RMSE=79.59 pix, R=0.87). Next, this pretrained model was transferred to the Large Grid dataset. The upper 3 rows of Table 1 shows the performance when all the data was used for funetuning. In this case also EEGNet-LSTM outperformed the benchmark (Pyramidal CNN [29]) by a large margin ($\sim$ 19%). Hence, subsequent SSDA experiments were performed using EEGNet-LSTMs.

All models were trained with mini-batches of 128 samples, and only the best-performing checkpoint on the held-out validation set was used for evaluating the test set. Adam optimizer with an initial learning rate of 10^{-4} was used. α in equation 3 was selected with a grid saerch over $\{0.01, 0.1, 1.0\}$; empirically, $\alpha = 0.1$ was always favored.

Table 1. Source Free Semi-supervised Deep Transfer of the model trained on the Visual Search task to the Large-Grid Task. The upper part of the table (rows 1-3) shows the performance ceiling when all labeled data in the target is used for training. The lower part (rows 4-9) shows SF-SSDA results when only 1% of the target samples were labeled (3 seeds).

Method	**R** ↑	**RMSE** (in pix) ↓
Naive Baseline	-	149.12 ± 0.02
Benchmark (TL, 100%)	0.91 ± 0.01	63.84 ± 0.75
EEGNet-LSTM (TL, 100%)	**0.93 ± 0.01**	**51.47 ± 0.63**
SF-SSDA (1% labels)		
EEGNet-LSTM + TL	0.77 ± 0.01	92.26 ± 1.66
EEGNet-LSTM + Mixup	0.48 ± 0.02	135.70 ± 1.25
EEGNet-LSTM + BBCN	0.76 ± 0.01	99.8 ± 3.35
EEGNet-LSTM + TASFAR	0.76 ± 0.01	*86.41 ± 1.05*
EEGNet-LSTM + DataFree	*0.80 ± 0.01*	87.64 ± 3.08
EEGNet-LSTM + CRAFT	**0.81 ± 0.02**	**84.17 ± 3.95**

To demonstrate the efficacy of CRAFT for SF-SSDA, target labels from the training data of the Large Grid dataset were randomly dropped in a stratified manner. Figure 5 (A) shows how the performance of these models is affected by reducing the fraction of labeled data (n_{ul}/N is the fraction of unlabeled samples). A lower proportion of labeled data worsens the performance of all the models (RMSE increases), but CRAFT produces the lowest RMSE for all unlabeled data proportions. Figure 5 (B) shows the prediction of saccade magnitudes ($|\Delta\overline{r}|$) for the best CRAFT-based EEGNet-LSTM model. Table 1 (lower 6 rows) summarizes the performance of CRAFT when 99% of the data were unlabeled; CRAFT outperforms supervised training (TL) by 9% and other SOTA SF-SSDA models by $> 4\%$ in terms of RMSE. The closest competitive methods were the SF-UDA methods of TASFAR and DataFree that incorporated our supervised objective. The correlation coefficient (R) also showed similar trends (Table 1; Figure 6A).

5.2. Brain Age Prediction

Next, we transferred the SFCN model trained on the UK Biobank MRI scans to the TLSA dataset. The TLSA dataset has only 188 MRI scans, so we avoid hyperparameter searches. All models were trained for 25 epochs (batch size = 4) using the Adam optimizer (initial learning rate = 10^{-4}), and only the final model was used for inference. Here, we used $\alpha = 0.1$; we explore the effect of varying α for ours and competing models in Appendix D. Each metric was calculated for the prediction on the entire dataset, leaving out one fold at a time (4-folds). The top of Table 2 (row 2) quantifies the performance ceiling when all the training data are used (3 seeds).

Table 2. Semi-supervised transfer of the SFCN model (source: UKB) to the TLSA dataset. Same as in Table 1, except that the lower part of the table (3-8) shows SF-SSDA results when only 20% of the target samples were labeled. (3 seeds)

Method	**R** ↑	**RMSE** (in years) ↓
Naive Baseline	-	7.91 ± 0.05
SFCN + TL (100%)	0.66 ± 0.01	6.14 ± 0.03
SF-SSDA (20% labels)		
SFCN + TL	0.41 ± 0.07	7.41 ± 0.21
SFCN + Mixup	0.34 ± 0.04	7.71 ± 0.14
SFCN + BBCN	0.28 ± 0.04	8.00 ± 0.15
SFCN + TASFAR	0.42 ± 0.07	7.47 ± 0.15
SFCN + DataFree	*0.50 ± 0.03*	*7.36 ± 0.14*
SFCN + CRAFT	**0.51 ± 0.03**	**7.14 ± 0.11**

For SF-SSDA, we performed experiments closely similar to the saccade task (Section 5.1). Figure 5 (C) shows how these models' performance is affected when the proportion of unlabeled samples increases. Reducing the fraction of labeled data worsens model performance, but CRAFT continues to remain the most competitive of all models. Figure 5 (D) shows the prediction scatter of the best CRAFT-based model. Table 2 summarizes the performance when 80% of the data was unlabeled. CRAFT outperforms supervised training (TL) by $\sim$ 4% and other SOTA SF-SSDA models by $> 3\%$ in terms of RMSE. Similar results were obtained with the R values also (Fig. 6B). Again, TASFAR and DataFree emerge as the most competitive methods.

Finally, we evaluated the effectiveness of CRAFT at two other real-world regression benchmarks: People Counting [44, 48] and Tumor Size prediction [6, 16]. CRAFT outperformed other SF-SSDA methods by $> 5\%$ and $> 2\%$ respectively for each dataset; the details are presented in Appendices C.1 and C.2.

Mitigating Sampling Biases. Next, we show how the unsupervised CRAFT objective could mitigate sampling bias effects in the target training set. In all of the previous simu-

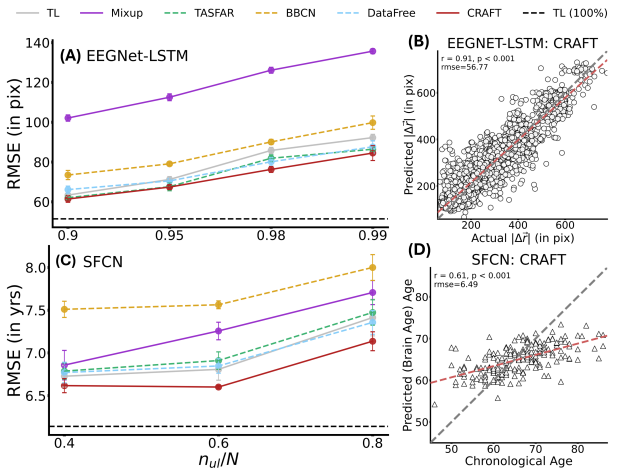

Figure 3. (**A, C**) RMSE (↓) for source-free semi-supervised domain adaptation of saccade (**A**) and brain age (**C**) prediction tasks, trained with varying proportions of unlabeled target data n_{ul}/N. Dashed black line: Performance ceiling. Circles: individual trials. (**B, D**) Prediction for the best CRAFT model for saccade amplitude (**B**) and brain age (**D**), with 90% and 40% unlabeled data, respectively. Triangles: individual participants. Results with a much higher proportion of unlabeled data are shown for the saccade dataset because of the far greater size of this (N ∼12k) compared to the brain age (N ∼180) dataset.

lations, the label distributions of the target train and test data were matched. Here, we deliberately distorted the label distribution of the target training data: 80% of the data above the mean age in the training set were removed, creating a heavy bias toward lower ages. Following this, labels were retained in only 40% of the training data ($n_{ul}/N = 0.6$). No such biases were introduced into the test data. The CRAFT model was then trained by imposing the unbiased (true) marginal label distribution on the unlabeled training set (equation 5). In this scenario, again, we find that CRAFT mitigates the effect of sampling biases in the training data (Table 3). It outperforms transfer learning by $\sim 5\%$, and other sota SF-SSDA models by $> 2.5\%$ (RMSE).

5.3. Analysis of computational complexity

Table 4 shows the computational complexity of CRAFT and the competing models. We compute both the algorithms' theoretical complexity and running time on the saccade and brain age prediction tasks. The theoretical training complexity is defined as the number of gradient updates per epoch. For inference, the complexity is the same for all models as the same architecture is trained using different adaptation paradigms. Training time and complexity are lowest for TASFAR, amongst all other SF-DA tools. The complexity of CRAFT was comparable to DataFree, the next best method.

Table 3. Analyses with datasets with sampling error. 80% of the data above the mean age is removed. Furthermore, only 40% of the labels in this mislabeled data are used for semi-supervised transfer of the SFCN model (source: UKB). (3 seeds)

Method	**R** ↑	**RMSE** (in years) ↓
Naive Baseline	-	8.38 ± 0.22
SFCN + TL	0.41 ± 0.05	8.40 ± 0.27
SFCN + Mixup	0.32 ± 0.02	*8.19 ± 0.15*
SFCN + BBCN	0.23 ± 0.02	8.53 ± 0.16
SFCN + TASFAR	0.39 ± 0.04	8.57 ± 0.22
SFCN + DataFree	*0.41 ± 0.03*	8.58 ± 0.23
SFCN + CRAFT	**0.45 ± 0.03**	**7.98 ± 0.25**

Table 4. **Computational complexity**. Time per training epoch and for all test data (in **min**). Parentheses: P- no. of model parameters; N- samples; D- latent dimensionality; β- batch size, B- no. of bins, m- no. of inferences for MC Dropout uncertainty estimation.

	EEGNet-LSTM ($P\sim 190K$)		SFCN ($P\sim 2.95M$)	
(B) **Method**	**Train /Comp.** ↓ ($N\sim 12k$)	**Test /Comp.** ↓ ($N\sim 2.5k$)	**Train/Comp.** ↓ ($N\sim 150$)	**Test/Comp.** ↓ ($N\sim 50$)
TL	0.15 / $O(N)$	0.02 / $O(N)$	0.12 / $O(N)$	0.04 / $O(N)$
Mixup	0.74 / $O(Nm)$	0.02 / $O(N)$	0.74 / $O(Nm)$	0.04 / $O(N)$
BBCN	2.71 / $O(NBD)$	0.02 / $O(N)$	1.29 / $O(NBD)$	0.04 / $O(N)$
TASFAR	0.30 / $O(N)$	0.02 / $O(N)$	0.19 / $O(N)$	0.04 / $O(N)$
DataFree	0.55 / $O(ND^2)$	0.02 / $O(N)$	0.32 / $O(ND^2)$	0.04 / $O(N)$
CRAFT	0.45 / $O(N(\beta+B))$	0.02 / $O(N)$	0.36 / $O(N(\beta+B))$	0.04 / $O(N)$

6. Concluding Remarks

Data scarcity poses a serious challenge for applying deep-learning models in real-world settings. Training over-parameterized models on a few hundred data points leads to overfitting and poor generalization. Conventional domain adaptation algorithms, to transfer pretrained networks, typically need access to source domain data, but these may not always be available. In this study, we propose a source-free semi-supervised domain adaptation framework that recalibrates the decision boundary without aligning the intermediate representation of the domains. In fields like medicine, resource constraints associated with labeling data are significant; therefore, leveraging the large amounts of unlabeled data becomes critical. We propose CRAFT as an SF-SSDA approach for such source-free, label-sparse settings that are common in real-world applications.

7. Acknowledgements

This work was supported by the following funding sources: a Prime Minister's Research Fellowship (to M.B.), a DST SwarnaJayanti fellowship, a Pratiksha Trust Intramural grant, a Tata Trusts grant and a Google Research grant (to D.S.). We also thank the Tata Longitudinal Study on Aging at the Indian Institute of Science for access to MRI scans from their cohort.

References

[1] Jiayu An, Changming Zhao, and Dongrui Wu. Semi-supervised generalized source-free domain adaptation (ssg-sfda). In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. 3

[2] APTOS. Aptos 2019 blindness detection dataset, 2019. 1

[3] Sourabh Balgi and Ambedkar Dukkipati. CUDA: Contradistinguisher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Data Mining*, 2019. 1

[4] Sourabh Balgi and Ambedkar Dukkipati. Contradistinguisher: A Vapnik's imperative to unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4730–4747, 2022. 1, 3, 4

[5] Scott L. Brincat, Jacob A. Donoghue, Meredith K. Mahnke, Simon Kornblith, Mikael Lundqvist, and Earl K. Miller. Interhemispheric transfer of working memories. *Neuron*, 109: 1055–1066.e4, 2021. 5

[6] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. 2, 7, 3

[7] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140: 325–331, 2020. 2

[8] Thomas A. Carlson. Orientation decoding in human visual cortex: New insights from an unbiased perspective. *Journal of Neuroscience*, 34(24):8373–8383, 2014. 1

[9] J H Cole, S J Ritchie, M E Bastin, M C Valdés Hernández, S Muñoz Maniega, N Royle, J Corley, A Pattie, S E Harris, Q Zhang, N R Wray, P Redmond, R E Marioni, J M Starr, S R Cox, J M Wardlaw, D J Sharp, and I J Deary. Brain age predicts mortality. *Molecular Psychiatry*, 23:1385–1392, 2018. 6

[10] Sébastien M. Crouzet. Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, 10:1–17, 2010. 2

[11] Jorge Cuadros and George Bresnick. Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009. PMID: 20144289. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[13] Heiner Deubel and Werner X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36:1827–1837, 1996. 5

[14] Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *Neural Networks*, 167:92–103, 2023. 2

[15] Cian Eastwood, Ian Mason, Christopher K I Williams, and Bernhard Schölkopf. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022. Tenth International Conference on Learning Representations 2022, ICLR 2022 ; Conference date: 25-04-2022 Through 29-04-2022. 2, 3

[16] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. 2, 7, 3

[17] Saminger-Platz et al. Central moment discrepancy (cmd) for domain-invariant representation learning. *ICLR*, 2017. 1, 2

[18] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. *Neural Networks*, 174:106230, 2024. 1

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016. 1, 2

[20] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies, 2025. 1

[21] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*. MIT Press, 2004. 3, 4

[22] Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009. 2

[23] Yan Hao, Yuhong Guo, and Chunsheng Yang. Source-free unsupervised domain adaptation with surrogate data generation. In *British Machine Vision Conference*, 2021. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing. 3

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[26] Tianlang He, Zhiqiu Xia, Jierun Chen, Haoliang Li, and Shueng-Han Gary Chan. Target-agnostic source-free domain adaptation for regression tasks. *2024 IEEE 40th In-*

ternational Conference on Data Engineering (ICDE), pages 1464–1477, 2023. 2, 3

[27] Xinyang Huang, Chuang Zhu, Bowen Zhang, and Shanghang Zhang. Learning from different samples: A source-free framework for semi-supervised domain adaptation, 2024. 2

[28] Xinyang Huang, Chuang Zhu, Bowen Zhang, and Shanghang Zhang. Learning from different samples: A source-free framework for semi-supervised domain adaptation, 2024. 3

[29] Ard Kastrati, Martyna Plomecka, Damian Pascual Ortiz, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, and Nicolas Langer. Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1, 2, 5, 6, 7

[30] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method, 2021. 2

[31] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15:056013, 2018. 6, 2

[32] Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A Comprehensive Survey on Source-Free Domain Adaptation . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(08):5743–5762, 2024. 1, 2

[33] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2022. 2

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 97–105, Lille, France, 2015. PMLR. 2

[35] Jie Ma, Zhijun Xu, Xinyu Xiong, MaYiDiLi NiJiaTi, Dongyu Zhang, Weijun Sun, Feng Gao, and Guanbin Li. Source-free semi-supervised domain adaptation for tuberculosis recognition. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024. 2, 3

[36] Ning Ma, Jiajun Bu, Lixian Lu, Jun Wen, Zhen Zhang, Sheng Zhou, and Xifeng Yan. Semi-supervised hypothesis transfer for source-free domain adaptation, 2021. 2, 3

[37] Ning Ma, Haishuai Wang, Zhen Zhang, Sheng Zhou, Hongyang Chen, and Jiajun Bu. Source-free semi-supervised domain adaptation via progressive mixup. *Knowledge-Based Systems*, 262:110208, 2023. 2, 3

[38] Pravat K. Mandal, Rashima Mahajan, and Ivo D. Dinov. Structural brain atlases: Design, rationale, and applications in normal and pathological cohorts. *Journal of Alzheimer's Disease*, 31:S169–S188, 2012. 6

[39] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper L R Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature Neuroscience*, 19:1523–1536, 2016. 1, 6

[40] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation, 2020. 1, 2

[41] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, 2021. 6, 2

[42] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *Computer Vision – ECCV 2022*, pages 165–182, Cham, 2022. Springer Nature Switzerland. 2

[43] Xiaofeng Qu, Li Liu, Lei Zhu, Liqiang Nie, and Huaxiang Zhang. Source-free style-diversity adversarial domain adaptation with privacy-preservation for person re-identification. *Knowledge-Based Systems*, 283:111150, 2024. 1

[44] Vishwanath Sindagi, Rajeev Yasarla, and Vishal Patel. Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method . In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1221–1231, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2, 7

[45] Jonas S Sundarakumar, Albert Stezin, Abhishek L Menesgere, and Vijayalakshmi Ravindranath. Rural-urban and gender differences in metabolic syndrome in the aging population from southern india: Two parallel, prospective cohort studies. *eClinicalMedicine*, 47:101395, 2022. 1, 2, 6

[46] J-Donald Tournier, Robert Smith, David Raffelt, Rami Tabbara, Thijs Dhollander, Maximilian Pietsch, Daan Christiaens, Ben Jeurissen, Chun-Hung Yeh, and Alan Connelly. Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116137, 2019. 6

[47] Ihsan Ullah and Alfredo Petrosino. About pyramid structure in convolutional neural networks, 2016. 6, 2

[48] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 7

[49] Rand R. Wilcox. The percentage bend correlation coefficient. *Psychometrika*, 59:601–616, 1994. 6

[50] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), 2020. 1

[51] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems*, 36(8):3746–3766, 2021. 2

[52] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation, 2022. 2

Semi-supervised Deep Transfer for Regression without Domain Alignment

Supplementary Material

A. Derivations

A.1. CRAFT: A MAP Estimation View

In this section, we show that the proposed method can be viewed as regularized training of neural networks. In particular, we show that optimization of the CRAFT objective gives us the maximum a posteriori (MAP) estimate of model parameters. We know the MAP estimate for the parameters θ (given dataset $\mathcal{D}$) is obtained by maximizing the posterior $\log p(\theta|\mathcal{D}) \propto \log p(\mathcal{D}|\theta) + \log p(\theta)$, where, $\log p(\mathcal{D}|\theta)$ is the log-likelihood, and $p(\theta)$ is the prior distribution (regularizer) over the model parameters. The subsections below describe the choices for these components.

A.1.1. Likelihood

We define the log-likelihood for a labeled data as: $\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \theta)$. As this paper deals with unlabeled data, as a design choice, we assume the conditional distribution $p(y|\mathbf{x}, \theta)$ to be constant for unlabeled training data, optimizing the log-likelihood over labeled data only.

Although we only address continuous label prediction (regression – $p(y|\mathbf{x}, \theta) = \mathcal{N}(y; f(\mathbf{x}; \theta), c)$) tasks in this study, the framework can be converted to k-way discrete label prediction (classification) tasks by appropriate assumptions on the conditional distribution – e.g., $p(y|\mathbf{x}, \theta) = \text{Mult}_k(y; f_1(\mathbf{x}; \theta), ..., f_k(\mathbf{x}; \theta))$.

A.1.2. The Prior Distribution

In this subsection, we derive a principled prior distribution that can leverage unlabeled data and/or deal with data scarcity during transfer learning by biasing the selection of parameters that best suit the domain. We achieve this by extending the CUDA [4] framework, originally used for domain adaptation.

We note a few desirable properties for supervised transfer learning algorithms. Firstly, these models are not explicitly trained to match the marginal label distribution, i.e., $p(y) \neq \int_{\mathbf{x}} p(y|\mathbf{x}, \theta) p(\mathbf{x}) d\mathbf{x}$. Hence, the model's performance on unseen data can be compromised if the training and the test label distributions differ, such as due to sampling biases. Hence, $p(y) = p(y|\theta)$ is a property we desire. Secondly, ensuring this property can be particularly useful if we know that the training dataset is small, where sampling biases are likely. Hence, we define a joint distribution using the trained model $p(y|\mathbf{x}, \theta)$ to tackle these challenges:

$$\mathrm{q}(\mathbf{x}, y|\theta) = \left[\frac{p(y|\mathbf{x}, \theta)}{\sum_{i=1}^{N} p(y|\mathbf{x}_i, \theta)} p(y) \right] \tag{7}$$

Now, we incorporate $q(\mathbf{x}, y|\theta)$ as a prior over the model parameters. We use the maximum entropy principle to achieve this. In particular, we wish to maximize the entropy of the model's parameter distribution such that the negative log-probability of the joint distribution q, on average is as small as a constant G, i.e., $\mathbb{E}_{\theta \in \mathbf{\Theta}}[-\log q(\mathbf{x}, y|\theta)] = G$. Solving the Euler-Lagrange equations we get (see Appendix A.2):

$$\log p(\theta) \propto \alpha \log q(\mathbf{x}, y|\theta) \tag{8}$$

where α is the Lagrange multiplier corresponding to G.

A.2. A Maximum Entropy Prior

The optimization objective described above can be expressed as:

$$\theta^* = \underset{\theta \in \Theta}{\text{argmax}} - \int_{\theta} p(\theta) \log p(\theta) \mathrm{d}\theta \tag{9}$$

$$\text{s.t.} \quad \mathbb{E}_{\theta \in \mathbf{\Theta}}[-\log q(\mathbf{x}, y|\theta)] = G, \int_{\theta} p(\theta) \mathrm{d}\theta = 1$$

We ignore the constraint on the density integral, and normalize it at the end. Hence, the Lagrangian can be written as:

$$\mathcal{L}(\theta) = \int_{\theta} \underbrace{p(\theta) \left(-\log p(\theta) + \alpha(\log q(\mathbf{x}, y|\theta) + G)\right)}_{g(\theta,\, p(\theta))} \mathrm{d}\theta$$

Using the Euler-Lagrangian equation, and because g does not depend on the first derivate of p:

$$\frac{\partial g}{\partial p} - \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{\partial g}{\partial p'} = -1 - \log p(\theta) + \alpha(\log q(\mathbf{x}, y|\theta) + G) = 0$$

$$\implies \log p(\theta) = \alpha \log q(\mathbf{x}, y|\theta) + (\alpha G - 1)$$

Dropping the constant terms that are independent of θ:

$$p(\theta) \propto \alpha \log q(\mathbf{x}, y|\theta).$$

A.3. Mixtures of Gaussians and Exponentials

For learning the marginal distribution of labels for CRAFT we fit a mixture of exponential and Gaussian distributions. Given samples $\{\mathbf{x}_i\}_{i=1}^{N}$ from an unknown distribution p_x, s.t., $x_i \in \mathbb{R}_+$. If negative components exist, we can add a constant offset to make the data non-negative. We model

the underlying distribution by k_1 Gaussians and k_2 exponentials. The density of a mixture model p_θ, can be written as;

$$p_\theta(x) = \sum_{i=1}^{k_1+k_2} p_\theta(z)p_\theta(x|z) \tag{10}$$

, where z is the latent variable, s.t., $p_\theta(z) = \text{Mult}_k(z; \beta_1, \beta_2, ..., \beta_{k_1+k_2})$. The conditionals are defined as:

$$p_\theta(x|z=i) = \begin{cases} \mathcal{N}(x; \mu_i, \sigma_i^2) & i \in \{1, 2, ..., k_1\} \\ \text{Exp}(x; \lambda_i) & i \in \{k_1+1, ..., k_1+k_2\} \end{cases}$$

We estimate the parameters $\theta = \{\{\beta_i\}_{i=1}^{k_1+k_2}, \{\mu_i, \sigma_i^2\}_{i=1}^{k_1}, \{\lambda_i\}_{i=k_1+1}^{k_2}\}$, using the expectation-maximization algorithm (code available in Section E.2).

B. Datasets and Methods

B.1. Gaze prediction from brain signals

Preprocessing. All eye-tracking datasets – Visual Search and Large Grid – were preprocessed similarly. Briefly, blinks were removed using Gaussian filtering, and the temporal locations of saccades were found by detecting very high velocity and acceleration. Then, the eye-tracking data was partitioned into 1s windows so that only one saccade event occurred in that window. The displacement vector of the eye was computed, and its magnitude (saccade amplitude) was designated the target variable $y \in \mathbb{R}^+$. Time-matched to each saccade, a 500Hz EEG signal was extracted and used as the feature vector $\mathbf{x} \in \mathbb{R}^{128\text{x}500}$ (128=number of electrodes or channels). Table 5 lists the details of the train, cross-validation and test distributions for these datasets. We followed the leave-participant-out evaluation strategy with an identical train-validation-test split for the Large grid dataset for direct comparison with the benchmark. The Visual Search paradigm has more training data and was used to train the source model, which was transferred to the target datasets - Large Grid (from the benchmark). Before training the model, the target variable (saccade amplitude) was linearly scaled between $[-1, 1]$, and the EEG data were z-scored, by computing a common mean and standard deviation, across time and channels.

Models. The EEG-EyeNet benchmark showed that Pyramidal CNNs [47] performed best at predicting saccade amplitude. However, naive 2D convolution-based neural networks treat the timechannel EEG data as an image, whereas such data typically lack the correlation structure associated with natural images. Moreover, CNN-based models are not ideal for learning temporal autocorrelation structure present in the data.

Table 5. Training, validation and test sets for each dataset. A 60-20-20% leave-participant-out split was used, as in the original benchmark.

Dataset	Train	Validation	Test
Visual Search (Source)	21,191	5,018	5,354
Large Grid (Target)	12,275	2,836	2,719

We address these shortcomings with a novel end-to-end Long-Short-Term-Memory (LSTM) model trained on features extracted by an EEGNet-like [31] architecture. Figure 2B shows the feature extractor, which employs separable convolutions along time and electrode dimensions. The extracted temporal features are then fed to a series of two LSTMs, which predict the saccade amplitude.

In the feature extractor, the EEG data is first convolved along the temporal dimension using 100 ms trainable filters independently for each electrode, as it is the average duration to execute a saccade [10]. The filters are shifted by 10 ms to capture neural dynamics up to $\sim$100 Hz. Finally, depthwise-separable convolution is performed time-point-wise along the electrode dimension to obtain temporal features for the LSTM-based decoder.

B.2. Brain age prediction from structural MRI scans

Models. The Simple Fully Convolutional Network (SFCN, Fig. 2D [41] feature Extractor contains five blocks of 3D Convolution (3x3x3 filters), Batch-Normalizaton, and 3D-MaxPooling (2x2x2 filters), and ReLU activation, followed by a single block of 3D Convolution with 1x1x1 filters, Batch-Normalizaton, and ReLU activation. The outputs of the feature extractor are then the Global Average Pooled (GAP) and fed as input to a classifier. In the original paper, the classifier predicted the probability of each MRI scan belonging to one of the 40 bins (range: 42-82 years). The expected value of the outputs was reported as the predicted age. In our version of SFCN, we avoid binning by replacing the classifier with a regressor.

B.3. People counting from natural scene images

We addressed the challenge of counting people from photographs of crowds – the "people counting" challenge. **Preprocessing**. High-resolution scene images from two datasets – NWPU [48] and JHU-crowd [44] – were resized to 1152x768. Following this, each image was divided into six non-overlapping 384x384 patches. These patches were fed to a ResNet-based feature extractor, to ultimately predict the number of people in the scene. The target dataset (JHU-crowd) contains images with 0-10,000 people, and thus, we restricted the training samples of the source dataset (NWPU) to those that have less than 10k humans in each

image. Table 6 shows the train-test-validation split of the datasets.

Table 6. Training, validation and test sets for each dataset. The same train-test-validation split was used, as in the original benchmark.

Dataset	Train	Validation	Test
NWPU (Source)	3,100	499	1,500
JHU-crowd (Target)	2,269	500	1,600

Models. Figure 4A shows the model used for people counting. The patches from each image were passed through a ResNet50 [25] model, pretrained on ImageNet [12], to get six 2048-dimensional encoding vectors. These vectors were then combined using an "attention" head to obtain a 64-dimensional embedding for the image. This was then fed to a dense layer to predict the number of people in the scene.

B.4. Tumor size estimation from histopathology images

We also addressed the challenge of estimating tumor sizes from cancer histopathology images.

Preprocessing. We employed the Camelyon-16 [16] and Camelyon-17 [6] high-resolution breast-cancer datasets. The patch level annotations available in these datasets render them particularly suitable for our analysis. We used 256x256 high-resolution patches from the whole-slide images (WSIs) for our analysis. The annotations in these patches were used to compute the fraction of the patch that has tumerous tissue. This fraction was predicted as the "tumor size" at the patch level. Table 7 shows the train-validation-test split for the source and target datasets. To render the transfer more challenging, we used only patches containing 20-80% tumor in the image for training the source model, while the target patches had tumor coverage ranging between 0.1-99.9%.

Table 7. Training, validation and test sets for each dataset. A 64-16-20% stratified split was used for both datasets.

Dataset	Train	Validation	Test
Camelyon-16 (Source)	65,337	16,335	20,418
Camelyon-17 (Target)	7,926	1,982	2,477

Models. The 256x256 patch was passed through a Resnet152v2 [24] model initialized to the ImageNet [12] weights (Fig. 4B). Next, the extracted features were passed through two dense layers to predict the fraction of tissue covered by the tumor.

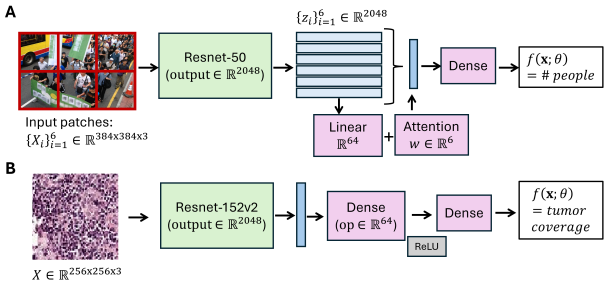

Figure 4. Model Architectures. (**A**) Predicting the number of people in a scene. The Resnet50 model is used to extract features from 6 patches of an image, along with an "attention" head that combines them, followed by a dense layer to make predictions; (**B**) Predicting the fraction of tumor cells in a histopathology image. The Resnet152v2 model is used to extract features, followed by predictions using dense layers.

C. Additional Benchmarks

C.1. People counting

We attempted transfer with models trained on a people counting task. First, we trained the Attn-Resnet model to count people with the NWPU dataset. Source predictions were reasonably effective, with RMSE=459.01 and R=0.74. Next, this pretrained model was transferred to the JHU-crowd dataset (see Appendix B.3). The upper 2 rows of Table 8 show the performance when all the data was used for finetuning. All models were trained with mini-batches of 16 samples, and only the best-performing checkpoint on the held-out validation set was used for evaluating the test set. Adam optimizer with an initial learning rate of 10^{-4} was used. Based on observations in the previous experiments (Section 5) $\alpha = 0.1$ was used as the weight for the unsupervised loss relative to the supervised loss.

To demonstrate the efficacy of CRAFT for SF-SSDA, target labels from the training data of the JHU-crowd dataset were removed randomly, but in a stratified manner. Figure 5A shows how the performance of these models is affected by reducing the fraction of labeled data (n_{ul}/N is the fraction of unlabeled samples). A lower proportion of labeled data worsens the performance of all the models (RMSE increases), but CRAFT outperforms other DA algorithms. Surprisingly, naive transfer learning performed better than SOTA algorithms for this problem. On further investigation, we observed that the model trained on the NWPU dataset readily generalized to the JHU-crowd dataset (RMSE=451.62, R=0.75), already performing close to the ceiling.

Figure 5B shows the prediction of people count in a scene for the best CRAFT-based Attn-Resnet model. Table 8 (lower 6 rows) summarizes the performance of CRAFT when 98% of the data were unlabeled; CRAFT

outperforms SOTA SF-SSDA models by $> 5\%$ in terms of RMSE. But supervised finetuning (TL) beat CRAFT by 17%. The correlation coefficient (R) also showed similar trends (Fig. 6C). Several SF-SSDA methods rely on predicting pseudo-labels from unlabeled target data, and it is possible that subpar pseudo-label prediction exacerbated the poor performance of these algorithms.

Table 8. Source Free Semi-supervised Deep Transfer of the model trained on the NWPU crowd dataset to the JHU dataset. The upper part of the table (rows 1-3) shows the performance ceiling when all labeled data in the target is used for training. The lower part (rows 4-9) shows SF-SSDA results when only 2% of the target samples were labeled (3 seeds).

Method	**R** ↑	**RMSE** (in #people) ↓
Naive Baseline	-	724.17 ± 1.11
Attn-Resnet (TL, 100%)	0.83 ± 0.02	429.25 ± 5.63
SF-SSDA (2% labels)		
Attn-Resnet + TL	$\mathbf{0.73 \pm 0.01}$	$\mathbf{459.05 \pm 4.41}$
Attn-Resnet + Mixup	0.34 ± 0.04	702.05 ± 8.26
Attn-Resnet + BBCN	0.32 ± 0.06	1221.36 ± 237.66
Attn-Resnet + TASFAR	0.27 ± 0.03	620.55 ± 37.28
Attn-Resnet + DataFree	0.67 ± 0.03	582.53 ± 6.63
Attn-Resnet + CRAFT	$\mathit{0.69 \pm 0.02}$	$\mathit{550.77 \pm 9.86}$

C.2. Tumor size estimation

Finally, we attempted transfer with models trained to estimate the fraction of tumor tissue in a high-resolution histopathological slide. The source model (Resnet152v2) was trained to predict what fraction of the patched Camelyon-16 dataset contained a tumor. Source predictions were fairly accurate, with RMSE=0.19 and R=0.46. Next, the model was transferred to make predictions on the Camelyon-17 dataset (see Appendix B.4). All models were trained with mini-batches of 32 samples, and only the best-performing checkpoint on the held-out validation set was used for evaluating the test set. Adam optimizer with an initial learning rate of 10^{-4} was used. As before, $\alpha = 0.01$ was used as the unsupervised loss weightage relative to the supervised loss.

The top of Table 9 (row 2) quantifies the performance ceiling when all the training data are used (3 seeds). In rows 3-8, we compare the performance of all the source-free models when only 1% of the target dataset was labeled. We observe that CRAFT outperformed all SF-SSDA methods by $> 6\%$ on the correlation coefficient (R). Only DataFree performs comparably in terms of RMSE. Figure 5C shows that CRAFT performs better than all competing methods when the fraction of unlabeled data is varied from 90-99%. Similar trends were observed for the correlation coefficient R (Fig. 6D). Figure 5D shows the prediction scatter of the best-performing CRAFT model when 90% of the data was labeled.

Table 9. Source Free Semi-supervised Deep Transfer of the model trained on Camelyon-16 dataset to the Camelyon-17 dataset. The upper part of the table (rows 1-3) shows the performance ceiling when all labeled data in the target is used for training. The lower part (rows 4-9) shows SF-SSDA results when only 1% of the target samples were labeled (3 seeds).

Method	**R** ↑	**RMSE** ↓
Naive Baseline	-	0.313 ± 0.003
Resnet (TL, 100%)	0.773 ± 0.003	0.203 ± 0.003
SF-SSDA (1% labels)		
Resnet + TL	0.577 ± 0.026	0.260 ± 0.005
Resnet + Mixup	0.490 ± 0.009	0.283 ± 0.003
Resnet + BBCN	0.457 ± 0.084	0.307 ± 0.017
Resnet + TASFAR	0.617 ± 0.020	0.290 ± 0.012
Resnet + DataFree	$\mathit{0.630 \pm 0.012}$	$\mathit{0.247 \pm 0.003}$
Resnet + CRAFT	$\mathbf{0.670 \pm 0.009}$	$\mathbf{0.243 \pm 0.003}$

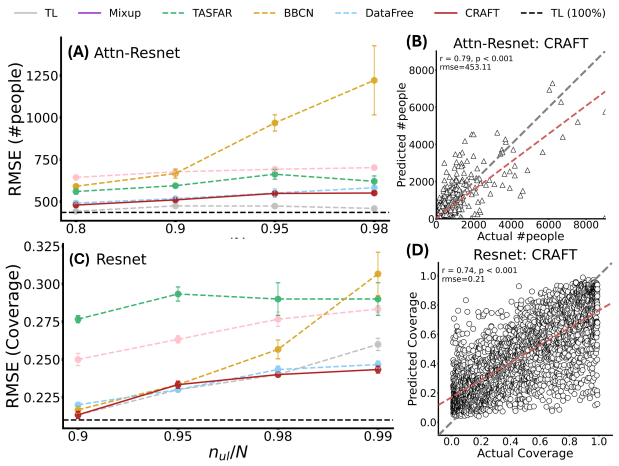

Figure 5. (**A, C**) RMSE (↓) for source-free semi-supervised domain adaptation of people counting (**A**) and tumor size (**C**) prediction tasks, trained with varying proportions of unlabeled target data n_{ul}/N. Dashed black line: Performance ceiling. Circles: individual trials. (**B, D**) Predictions for the best CRAFT model for people counting (**B**) and tumor size prediction (**D**), with 80% and 90% unlabeled data, respectively.

D. Analysis of hyperparameters: bin size, α

Bin Size. We varied the number of bins in the pseudo-label selection step of CRAFT to quantify the effect of bin sizes. We found the algorithm to be fairly robust to the choice of bin sizes for both the neuroscience datasets (Table 10).

Unsupervised loss weight (α). For the Saccade prediction task, a cross-validation set was available and hence, only the best-performing model, based on the corresponding α, is reported. However, a validation set was unavail-

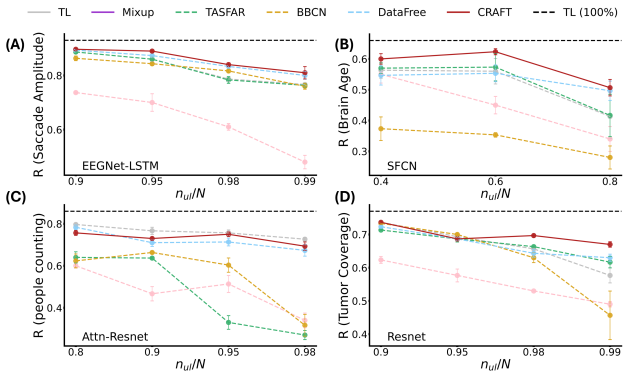

Figure 6. Correlation Coefficient R (↑) for all the competing SF-SSDA models at predicting (**A**) Saccade Amplitude; (**B**) Brain Age; (**C**) Number of People; and (**D**) Tumor Coverage.

Table 10. Effect of bin size on the performance of CRAFT for SF-SSDA for the saccade prediction (5% labeled) and brain-age prediction (60% labeled) tasks.

	EEGNet-LSTM ($P\sim 190K$)		SFCN ($P\sim 2.95M$)	
Bin Size % Range(y)	**R** ↑	**RMSE** ↓	**R** ↑	**RMSE** ↓
0.50	0.88 ± 0.01	67.25 ± 0.46	0.65 ± 0.03	6.53 ± 0.15
0.33	0.89 ± 0.01	65.84 ± 0.98	0.55 ± 0.02	6.91 ± 0.09
0.25	0.89 ± 0.01	67.39 ± 1.39	0.62 ± 0.01	6.60 ± 0.02
0.20	0.89 ± 0.01	65.21 ± 0.89	0.64 ± 0.02	6.36 ± 0.09

able for brain age prediction. Hence, we report the performance across three α values $\{0.01, 0.1, 1.0\}$ in Table 11. While CRAFT outperforms competing models for low values of alpha, large weightages to the unsupervised loss (e.g., $\alpha = 1$) result in comparatively poor performance.

Table 11. Effect of α (unsupervised loss weight) on the performance of CRAFT and other competing methods for SF-SSDA for the brain-age prediction (20% labeled) tasks.

	TASFAR		DataFree		**CRAFT**	
α	**R** ↑	**RMSE** ↓	**R** ↑	**RMSE** ↓	**R** ↑	**RMSE** ↓
0.01	0.40 ± 0.07	7.53 ± 0.18	0.31 ± 0.04	7.82 ± 0.20	**0.51 ± 0.03**	**7.31 ± 0.17**
0.10	0.42 ± 0.07	7.47 ± 0.15	0.50 ± 0.05	7.35 ± 0.15	**0.51 ± 0.03**	**7.14 ± 0.11**
1.00	**0.40 ± 0.06**	**7.47 ± 0.21**	0.07 ± 0.02	8.70 ± 0.06	0.37 ± 0.03	8.64 ± 0.15

E. Further Details

E.1. Compute Resources and Software

All the algorithms were implemented using TensorFlow 2.x, and experiments were run on 24 GB RTX-4090Ti, 32 GB V100, and 40 GB A100 GPUs.

E.2. Code Availability

The implementation of all the methods described in the study is publicly available in the following repository – https://github.com/mainak-biswas1999/CRAFT_ICCV2025.git.