IWCAE: IMPORTANCE-WEIGHTED CONDITIONAL AUTOENCODERS FOR DOMAIN-ADAPTIVE GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models characterize full data distributions and support a broad range of downstream tasks, but their performance may deteriorate when the training and testing distributions differ. Existing domain adaptation methods are often developed under covariate shift or label shift assumptions, which may be restrictive under substantial domain shift. We instead consider a more flexible setting in which there exists a representation under which the conditional response distribution is invariant across domains. We propose an importance-weighted conditional variational autoencoder, termed IWCAE, for transferable conditional distribution learning. IWCAE learns the full conditional response distribution through an importance-weighted conditional likelihood objective while using a maximum mean discrepancy penalty to align the marginal representation distributions across the training and testing domains. This design seeks to preserve predictive information while suppressing domain-specific variation. We further develop a sampling-importance-resampling procedure for approximate predictive sampling when direct sampling from the conditional latent prior is difficult. Theoretically, we establish the asymptotic validity of the proposed sampling procedure. Experiments on three medical datasets demonstrate the effectiveness of IWCAE in improving cross-domain predictive performance and quantifying predictive uncertainty under distribution shift.

1 INTRODUCTION

Generative models have become an important tool for analyzing complex data (Kingma & Welling, 2014; Goodfellow et al., 2020; Papamakarios et al., 2021). Unlike traditional prediction methods that estimate only the conditional mean, generative models learn the full conditional distribution of the outcome given the predictors. As a result, they can generate realistic samples and provide a complete description of uncertainty in the target variable (Kohl et al., 2018; Yu et al., 2024). These properties have led to successful applications in many fields, including medical imaging, electronic health records, and computer vision (Pu et al., 2016; Kebaili et al., 2023; Qiu et al., 2025).

Most generative models assume that the training and testing data follow the same distribution, so that a model learned from the training data can be directly applied to new observations. In practice, this assumption is often violated due to differences in data collection, measurement procedures, or population characteristics (Hendrycks & Dietterich, 2019). For example, medical images collected at different hospitals may be acquired using different scanners and clinicians with different levels of experience may record electronic health records differently. These distribution differences raise an important question: can a generative model trained on one dataset accurately generate the target distribution for another?

A substantial body of literature has examined discrepancies between training and test data distributions under the frameworks of covariate shift and label shift. Under covariate shift, the distribution of the predictors changes across domains, while the conditional distribution of the outcome given the predictors remains the same (Shimodaira, 2000; Huang et al., 2006; Sugiyama et al., 2007). Under label shift, the conditional distribution of the predictors given the outcome is assumed to be invariant, whereas the marginal distribution of the outcome may differ across domains (Zhang et al., 2013;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Garg et al., 2020). Building on these assumptions, a variety of domain adaptation methods have been developed, including importance reweighting (Shimodaira, 2000; Sugiyama et al., 2008; Bickel et al., 2009), marginal feature alignment (Pan et al., 2010; Long et al., 2013; Peng et al., 2019), and prediction-conditioned alignment (Long et al., 2018; Liu et al., 2021).

These approaches have several limitations. Importance reweighting can be unstable in high-dimensional settings with limited distributional overlap between the training and testing domains, while marginal feature alignment does not guarantee conditional invariance across domains (Gong et al., 2016; Johansson et al., 2019). Prediction-conditioned methods are mainly developed for classification and can be sensitive to prediction errors. More fundamentally, the invariance assumptions underlying covariate shift and label shift may fail when populations, measurement processes, outcome prevalence, or labeling mechanisms differ across domains (Rojas-Carulla et al., 2018; Tachet des Combes et al., 2020).

These limitations motivate a more flexible setting. We assume that there exists a predictor representation that aligns the training and testing distributions while preserving an invariant conditional relationship with the outcome. Rather than imposing invariance in the original input space, this formulation allows the model to learn a representation that satisfies the invariance assumption while retaining outcome-relevant information.

More specifically, we propose an importance-weighted conditional autoencoding framework (IWCAE) that jointly learns a domain-invariant predictor representation and a conditional latent-variable model for the outcome distribution. By incorporating importance weighting, our approach provides a tighter approximation to the marginal log-likelihood than the standard variational autoencoder objective, thereby improving distributional modeling. Meanwhile, joint optimization of representation alignment and conditional distribution estimation balances predictive performance and robustness to distribution shift.

Our proposal is related to, but fundamentally distinct from, existing approaches to invariant representation learning (Long et al., 2015; Arjovsky et al., 2019; Li et al., 2022; Wu et al., 2025a). Invariant representation learning aims to learn a representation that is both predictive of the outcome and stable across domains (Zhao et al., 2022). However, existing methods are primarily designed for point prediction by optimizing objectives such as the squared error or cross-entropy loss, rather than modeling the full conditional outcome distribution. Consequently, they do not naturally provide uncertainty quantification for predictions. In contrast, our method directly models the conditional outcome distribution, enabling both robust prediction under distribution shift and principled uncertainty quantification. Our main contributions are summarized as follows.

- We introduce IWCAE, a distributional domain-adaptation framework that jointly learns the full conditional response distribution and aligns the represented training and testing marginals.
- We develop an importance-weighted training objective and a sampling-importance-resampling procedure for inference, and establish their asymptotic properties.
- Experiments on three medical datasets with heterogeneous data modalities the effectiveness of IWCAE for conditional generative prediction under cross-domain distribution shifts.

2 Related Work

Domain adaptation and invariant representation learning. Existing domain adaptation methods can be broadly categorized into instance weighting and feature transformation approaches. Instance weighting methods typically rely on the assumption of an invariant conditional distribution and address domain shift by reweighting training-domain samples according to the discrepancy between training and testing input distributions (Sugiyama et al., 2007; 2010; Sun et al., 2011). Instead of directly estimating density ratios, iterative reweighting strategies based on AdaBoost have also been explored (Dai et al., 2007; Yao & Doretto, 2010; Taherkhani et al., 2020). Feature transformation methods instead seek a domain-invariant representation in which the discrepancy between training and testing domains is reduced (Daumé III, 2007; Pan et al., 2010). Various distribution discrepancy measures have been adopted for representation alignment, including maximum mean discrepancy (MMD) (Pan et al., 2010; Long et al., 2013; Li et al., 2018), Kullback-Leibler (KL) divergence (Zhuang et al., 2015; Wang et al., 2021), and Jensen-Shannon divergence (Chen et al., 2010; Dey

et al., 2016). Optimal transport-based approaches further utilize Wasserstein distances to measure and minimize domain discrepancy (Lee et al., 2019; Zhou et al., 2021; He et al., 2024).

Conditional generative modeling. Variational autoencoders (VAE) provide a flexible latent variable framework for learning complex data distributions (Kingma & Welling, 2014; Rezende et al., 2014). Their conditional extensions model the distribution of an outcome given observed covariates and have been widely used for conditional generation and probabilistic prediction (Sohn et al., 2015; Ivanov et al., 2019; Kim et al., 2021; Zhao et al., 2024). By sampling from the latent space, conditional VAE can generate multiple plausible predictions and thereby characterize predictive uncertainty. This capability has been exploited in ambiguous tumor segmentation (Kohl et al., 2018), combined with conformal prediction (Yu et al., 2024), and applied to probabilistic time series forecasting (Wu et al., 2025b). Standard conditional VAE, however, do not explicitly account for distribution shifts between the training and testing domains. Existing studies have addressed this limitation by combining conditional VAE with domain adaptation techniques, including marginal representation alignment (Wu & Zhuang, 2021; Lee et al., 2022; Wang & Breckon, 2023) and domain-adversarial learning (Wang et al., 2022; Hassan Pour Zonoozi et al., 2025). Although these strategies may improve cross-domain generalization, the role of the latent variable structure in preserving predictive information and facilitating domain adaptation remains insufficiently understood (Du et al., 2020).

3 Domain-adaptive conditional generative modeling

In this section, we formulate domain adaptation for conditional generative modeling and develop an invariant-representation framework for learning the conditional response distribution across domains. We first introduce the notation and data structure used throughout the paper. Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^q$ denote the predictor and response, respectively. For the training domain, let P_X^{tr}, $P_{Y|X}^{\mathrm{tr}}$, and P_{XY}^{tr} denote the marginal, conditional, and joint distributions, with corresponding densities $p^{\mathrm{tr}}(\mathbf{x})$, $p^{\mathrm{tr}}(\mathbf{y} \mid \mathbf{x})$, and $p^{\mathrm{tr}}(\mathbf{x}, \mathbf{y})$. We use P_X^{te}, $P_{Y|X}^{\mathrm{te}}$, and P_{XY}^{te} and their corresponding densities analogously for the testing domain. The observed data consist of labeled training samples $\mathcal{D}^{\mathrm{tr}} = \{(\mathbf{x}_i^{\mathrm{tr}}, \mathbf{y}_i^{\mathrm{tr}})\}_{i=1}^n$ and unlabeled testing inputs $\mathcal{D}^{\mathrm{te}} = \{\mathbf{x}_j^{\mathrm{te}}\}_{j=1}^m$, where testing responses are unavailable during training.

3.1 Problem formulation

Let $\mathcal{T}$ denote a class of measurable mappings from $\mathcal{X} \subseteq \mathbb{R}^p$ to $\mathbb{R}^{\tilde{p}}$. Rather than requiring the conditional distribution of Y given the original input X to remain invariant across domains, we assume that there exists a representation $\tau \in \mathcal{T}$ such that the conditional distribution of Y given $\tau(X)$ is shared by the training and testing domains. This formulation allows domain-specific variation to be removed at the representation level, even when the conditional distributions given the original input differ across domains. We formalize this assumption below.

Assumption 1 (Conditional invariant representation) *There exists $\tau \in \mathcal{T}$ such that*

$$P_{Y|\tau(X)}^{\mathrm{tr}} = P_{Y|\tau(X)}^{\mathrm{te}} \text{ and } P_{\tau(X)}^{\mathrm{tr}} = P_{\tau(X)}^{\mathrm{te}}$$

Assumption 1 is related to invariant representation assumptions commonly used in domain adaptation (Pan et al., 2010; Rojas-Carulla et al., 2018; Tachet des Combes et al., 2020) and provides a structural condition under which transfer remains possible even when the classical covariate-shift condition does not hold.

3.2 Learning transferable representations

We are interested in learning a transformation function τ that contains as much information as possible about the target Y, while at the same time filtering out information related to distribution discrepency between testing and training samples. More specifically, Let $\{P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x})) : \tau \in \mathcal{T}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a parametric family of conditional distributions on $\mathcal{Y}$, indexed by the representation value $\tau(\mathbf{x})$. We estimate the conditional model on the training domain by minimizing the logarithmic loss, a population formulation is

$$(\boldsymbol{\theta}^*, \tau^*) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \tau \in \mathcal{T}}{\arg\min}\, \mathcal{L}(\boldsymbol{\theta}, \tau) \quad \text{subject to} \quad D\left(P_{\tau(X)}^{\mathrm{tr}}, P_{\tau(X)}^{\mathrm{te}}\right) = 0, \tag{1}$$

where $\mathcal{L}(\boldsymbol{\theta}, \tau) := \mathrm{E}_{(X,Y)\sim P^{\mathrm{tr}}_{XY}}[-\log p_{\boldsymbol{\theta}}(Y \mid \tau(X))]$ denotes the population negative log-likelihood and $D(\cdot,\cdot)$ is a discrepancy between two distributions.

The optimization of (1) consists of two components: minimizing the logarithmic loss and aligning the marginal distributions of the learned representations. When the conditional likelihood $p_{\boldsymbol{\theta}}(\mathbf{y} \mid \tau(\mathbf{x}))$ has a tractable and differentiable form, it can be optimized directly using gradient-based methods (Fu & Hu, 2012; Martens & Grosse, 2015). However, because $\tau(X)$ is a representation of the predictor X, it generally preserves only part of the information contained in X. We therefore introduce a latent variable Z to capture the residual information in X that is not encoded by $\tau(X)$. Building on this idea, we adopt a conditional latent-variable model (Kingma & Welling, 2014; Sohn et al., 2015) and propose an importance-weighted conditional autoencoder (IWCAE). The importance-weighted objective provides a tighter approximation to the marginal log-likelihood than the conventional variational autoencoder objective, resulting in a more accurate estimate of the conditional outcome distribution. In addition, we choose $D(\cdot,\cdot)$ to be the maximum mean discrepancy (MMD) (Gretton et al., 2006), a kernel-based measure of the discrepancy between probability distributions, to encourage alignment between the learned representation distributions in the training and testing domains.

3.3 Generative model estimation

This section presents the estimation procedure corresponding to the population objective in (1). It combines importance-weighted variational learning for conditional distribution estimation with marginal representation alignment.

We first develop a conditional latent variable model for estimating the predictive distribution $P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}))$ appearing in the optimization problem (1). Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \boldsymbol{\Theta} := \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2$ and introduce a latent variable $\mathbf{z} \in \mathbb{R}^d$. We model the conditional density as

$$p_{\boldsymbol{\theta}}(\mathbf{y} \mid \tau(\mathbf{x})) = \int p_{\boldsymbol{\theta}_1}(\mathbf{y} \mid \mathbf{z}, \tau(\mathbf{x}))\, p_{\boldsymbol{\theta}_2}(\mathbf{z} \mid \tau(\mathbf{x}))\, \mathrm{d}\mathbf{z} \tag{2}$$

where $p_{\boldsymbol{\theta}_2}(\mathbf{z} \mid \tau(\mathbf{x}))$ denotes the latent prior conditioned on the learned representation, $p_{\boldsymbol{\theta}_1}(\mathbf{y} \mid \mathbf{z}, \tau(\mathbf{x}))$ is the conditional decoder. The corresponding posterior density is

$$p_{\boldsymbol{\theta}}(\mathbf{z} \mid \tau(\mathbf{x}), \mathbf{y}) = \frac{p_{\boldsymbol{\theta}_1}(\mathbf{y} \mid \mathbf{z}, \tau(\mathbf{x}))\, p_{\boldsymbol{\theta}_2}(\mathbf{z} \mid \tau(\mathbf{x}))}{p_{\boldsymbol{\theta}}(\mathbf{y} \mid \tau(\mathbf{x}))}.$$

The marginal likelihood in (2) is generally intractable (Diederik & Max, 2019), and hence so is the posterior in $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \tau(\mathbf{x}), \mathbf{y})$. We therefore introduce a variational distribution $Q_{\boldsymbol{\gamma}}(\cdot \mid \tau(\mathbf{x}), \mathbf{y})$ with density $q_{\boldsymbol{\gamma}}(\mathbf{z} \mid \tau(\mathbf{x}), \mathbf{y})$ where $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$. Throughout this paper, we assume that $q_{\boldsymbol{\gamma}}$ belongs to a tractable parametric family.

In this paper, we use the following importance-weighted object to estimate the variational parameter $\boldsymbol{\gamma}$:

$$\epsilon_K(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \tau; \mathbf{x}, \mathbf{y}) = \mathrm{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_K \sim Q_{\boldsymbol{\gamma}}(\cdot \mid \tau(\mathbf{x}), \mathbf{y})}\left[\log\left(\frac{1}{K}\sum_{j=1}^{K} \frac{p_{\boldsymbol{\theta}_1}(\mathbf{y} \mid \mathbf{z}_j, \tau(\mathbf{x}))\, p_{\boldsymbol{\theta}_2}(\mathbf{z}_j \mid \tau(\mathbf{x}))}{q_{\boldsymbol{\gamma}}(\mathbf{z}_j \mid \tau(\mathbf{x}), \mathbf{y})}\right)\right]. \tag{3}$$

The ϵ_K is a lower bound of the conditional log-likelihood. When $k = 1$, the ϵ_K is equivalent to the evidence lower bound (ELBO) used in VAE models. When $K \to \infty$, Burda et al. (2016, Theorem 1) show that this lower bound becomes monotonically tighter as the number of importance samples K increases and converges to the exact log-likelihood for the standard VAE. We establish the analogous result for the conditional VAE setting in Appendix B. Consequently, maximizing ϵ_K is equivalent to maximizing the conditional log-likelihood when $K \to \infty$. Compared with the standard ELBO, the importance-weighted objective therefore provides a tighter approximation to the conditional log-likelihood and may yield a more accurate characterization of the conditional distribution.

We measure the discrepancy between the training and testing representation distributions using the maximum mean discrepancy (MMD). Let $k_{\tau(X)}$ be a bounded kernel on the representation space $\tau(\mathcal{X})$, and let $\mathcal{H}_{k_{\tau(X)}}$ denote its associated reproducing kernel Hilbert space (RKHS). For probability

distributions $P^{\mathrm{tr}}_{\tau(X)}$ and $P^{\mathrm{te}}_{\tau(X)}$, the MMD is defined as

$$\mathrm{MMD}_{k_{\tau(X)}}\left(P^{\mathrm{tr}}_{\tau(X)}, P^{\mathrm{te}}_{\tau(X)}\right) = \sup_{f\in\mathcal{F}}\left|\mathrm{E}_{\tau(X)\sim P^{\mathrm{tr}}_{\tau(X)}}\left[f\left(\tau(X)\right)\right] - \mathrm{E}_{\tau(X)\sim P^{\mathrm{te}}_{\tau(X)}}\left[f\left(\tau(X)\right)\right]\right|,$$

where $\mathcal{F} = \{f : \|f\|_{\mathcal{H}_{k_{\tau(X)}}} \leq 1\}$. MMD is an integral probability metric induced by the unit ball of an RKHS (Sriperumbudur et al., 2012) and is closely related to the energy distance under appropriate choices of kernels (Sejdinovic et al., 2013). Marginal representation alignment based on MMD has been widely adopted in unsupervised domain adaptation (Pan et al., 2010; Long et al., 2013; Li et al., 2018). Then we define $D\left(P^{\mathrm{tr}}_{\tau(X)}, P^{\mathrm{te}}_{\tau(X)}\right)$ in (1) to be $D\left(P^{\mathrm{tr}}_{\tau(X)}, P^{\mathrm{te}}_{\tau(X)}\right) = \mathrm{MMD}^2_{k_{\tau(X)}}\left(P^{\mathrm{tr}}_{\tau(X)}, P^{\mathrm{te}}_{\tau(X)}\right)$. Combining this choice with (3),we define the IWCAE objective function to be defined as

$$\begin{aligned}&\left(\widehat{\boldsymbol{\theta}}_1^{\mathrm{IW}}, \widehat{\boldsymbol{\theta}}_2^{\mathrm{IW}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{IW}}, \widehat{\tau}^{\mathrm{IW}}\right)\\ &= \underset{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_1, \boldsymbol{\theta}_2\in\boldsymbol{\Theta}_2, \boldsymbol{\gamma}\in\boldsymbol{\Gamma}, \tau\in\mathcal{T}}{\arg\min} -\frac{1}{n}\sum_{i=1}^{n}\epsilon_K\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \tau; \mathbf{x}_i, \mathbf{y}_i\right) + \lambda \mathrm{MMD}^2_{k_{\tau(X)}}\left(\widehat{P}^{\mathrm{tr}}_{\tau(X)}, \widehat{P}^{\mathrm{te}}_{\tau(X)}\right), \quad (4)\end{aligned}$$

where λ is a penalty parameter that controls the trade-off between prediction accuracy and domain alignment, $\mathrm{MMD}^2_{k_{\tau(X)}}\left(\widehat{P}^{\mathrm{tr}}_{\tau(X)}, \widehat{P}^{\mathrm{te}}_{\tau(X)}\right)$ is an unbiased estimator of the squared MMD computed by the sample version of marginal distributions $\widehat{P}^{\mathrm{tr}}_{\tau(X)}, \widehat{P}^{\mathrm{te}}_{\tau(X)}$. In the experiments, we adopt the distance-induced kernel associated with the Euclidean metric, whose equivalence to the Euclidean energy distance is established in the Appendix A.1. The Appendix A describes the distributional specifications of the encoders and decoder, the representation classes and the explicit loss function used in our implementation.

3.4 Inference

We approximate the predictive distribution using the sampling-importance-resampling (SIR) procedure (Rubin, 1987), which is widely used in variational latent variable models (Bauer & Mnih, 2019; Aneja et al., 2021).
To facilitate sampling, we introduce a proposal encoder $Q_{\boldsymbol{\varphi}}\left(\cdot \mid \tau(\mathbf{x})\right)$. For an input $\mathbf{x}_i \in \mathcal{X}$, Algorithm 1 generates M samples $\{\mathbf{y}_{ij}\}_{j=1}^{M}$ from L proposal samples. Section 4 shows that, as $L \to \infty$, these samples converge in distribution to draws from $P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}_i))$. The conditional mean is approximated by $\bar{\mathbf{y}}_i = \frac{1}{M}\sum_{j=1}^{M}\mathbf{y}_{ij}$, which converges to $\mathrm{E}_{\boldsymbol{\theta}}[Y \mid \tau(X) = \tau(\mathbf{x}_i)]$ as $L \to \infty$ and subsequently $M \to \infty$.

Algorithm 1 Sampling–importance–resampling

Require: Transformed input $\tau(\mathbf{x}_i)$; proposal $Q_{\boldsymbol{\varphi}}$; conditional prior $P_{\boldsymbol{\theta}_2}$; decoder $P_{\boldsymbol{\theta}_1}$; sample sizes L and M

Draw $\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iL} \sim Q_{\boldsymbol{\varphi}}(\cdot \mid \tau(\mathbf{x}_i))$
for $l = 1, \ldots, L$ **do**
  $r_l \leftarrow \dfrac{p_{\boldsymbol{\theta}_2}(\mathbf{z}_{il} \mid \tau(\mathbf{x}_i))}{q_{\boldsymbol{\varphi}}(\mathbf{z}_{il} \mid \tau(\mathbf{x}_i))}$
end for
$w_l \leftarrow r_l / \sum_{s=1}^{L} r_s$ for $l = 1, \ldots, L$
Draw $I_1, \ldots, I_M \sim \mathrm{Categorical}(w_1, \ldots, w_L)$
for $j = 1, \ldots, M$ **do**
  Draw $\mathbf{y}_{ij} \sim P_{\boldsymbol{\theta}_1}(\cdot \mid \tau(\mathbf{x}_i), \mathbf{z}_{iI_j})$
end for
return $\{\mathbf{y}_{ij}\}_{j=1}^{M}$

Since the distribution of SIR samples $\{\mathbf{y}_{ij}\}_{j=1}^{M}$ converge to $P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}_i))$, the proposed sampling procedure can recover the distribution of Y_i. We utilize the quantiles of the distribution to quantify the prediction uncertainty.

Binary outcome: Suppose that $y \in \{0, 1\}$. For each input $\mathbf{x}_i$, let $p_i := P_{\boldsymbol{\theta}}\left(Y = 1 \mid \tau(X) = \tau(\mathbf{x}_i)\right)$. For each resampled latent variable $\mathbf{z}_{iI_j}$ obtained by Algorithm 1, let $p_{ij} = P_{\boldsymbol{\theta}_1}\left(Y = 1 \mid \tau(\mathbf{x}_i), \mathbf{z}_{iI_j}\right)$. Let $\widehat{q}_{\frac{\alpha}{2}}\left(\tau(\mathbf{x}_i)\right)$ and $\widehat{q}_{1-\frac{\alpha}{2}}\left(\tau(\mathbf{x}_i)\right)$ denote the empirical $\alpha/2$ and $(1-\alpha/2)$ quantiles of $(p_{ij})_{j=1}^{M}$, respectively. We define $\widehat{C}_{1-\alpha}\left(\tau(\mathbf{x}_i)\right) := \left[\widehat{q}_{\alpha/2}\left(\tau(\mathbf{x}_i)\right), \widehat{q}_{1-\alpha/2}\left(\tau(\mathbf{x}_i)\right)\right]$, whose length summarizes the predictive uncertainty in the probability of class $Y = 1$ induced by the latent distribution.

Continuous outcome: Suppose that $y \in \mathbb{R}$. For each input $\mathbf{x}_i$, let $(y_{ij})_{j=1}^{M}$ be the samples generated by Algorithm 1. Let $\widehat{q}_{\frac{\alpha}{2}}\left(\tau(\mathbf{x}_i)\right)$ and $\widehat{q}_{1-\frac{\alpha}{2}}\left(\tau(\mathbf{x}_i)\right)$ denote the empirical $\alpha/2$ and $(1-\alpha/2)$ quantiles of $(y_{ij})_{j=1}^{M}$, respectively. We define the empirical prediction interval as

$\widehat{C}_{1-\alpha}(\tau(\mathbf{x}_i)) := \left[\widehat{q}_{\alpha/2}(\tau(\mathbf{x}_i)), \widehat{q}_{1-\alpha/2}(\tau(\mathbf{x}_i))\right]$, whose length summarizes the predictive uncertainty of the continuous outcome.

For multivariate responses, componentwise prediction intervals may be obtained from the corresponding empirical quantiles of the SIR samples. These intervals provide componentwise measures of predictive uncertainty and are used in the experiments in Section 5.

In practice, the $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau)$ in Algorithm 1 are replaced by the fitted estimates $\left(\widehat{\boldsymbol{\theta}}_1^{\text{IW}}, \widehat{\boldsymbol{\theta}}_2^{\text{IW}}, \widehat{\tau}^{\text{IW}}\right)$ obtained from (4). The proposal parameter $\widehat{\boldsymbol{\varphi}}^{\text{IW}}$ is learned by minimizing the KL divergence between the fitted variational distribution $Q_{\widehat{\boldsymbol{\gamma}}^{\text{IW}}}\left(\cdot \mid \widehat{\tau}^{\text{IW}}(\mathbf{x}), \mathbf{y}\right)$ and the proposal distribution $Q_{\widehat{\boldsymbol{\varphi}}^{\text{IW}}}\left(\cdot \mid \widehat{\tau}^{\text{IW}}(\mathbf{x})\right)$.

4 Theoretical Analysis

In this section, we establish the asymptotic validity of the proposed SIR procedure by showing that the distribution of the generated samples converges to $P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}))$ as the number of samples tends to infinity.

Theorem 1 (Asymptotic validity of SIR predictive sampling) *For any fixed $\mathbf{x}_i \in \mathcal{X}$, suppose that the importance ratio $p_{\boldsymbol{\theta}_2}(\mathbf{z} \mid \tau(\mathbf{x})) / q_{\boldsymbol{\varphi}}(\mathbf{z} \mid \tau(\mathbf{x}))$ is uniformly bounded. Let $(\mathbf{y}_{ij})_{j=1}^{M}$ be the samples generated by Algorithm 1. Then, as $L \to \infty$, $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iM})$ converges in distribution to $(Y_1, \ldots, Y_M)$, where $Y_1, \ldots, Y_M \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}_i))$.*

The following corollary is an immediate consequence of Theorem 1 and establishes asymptotically valid prediction intervals for continous univariate responses.

Corollary 1 (Asymptotic validity of SIR prediction intervals) *Suppose that $\mathcal{Y} \subseteq \mathbb{R}$. For a fixed $\mathbf{x}_i \in \mathcal{X}$, let the predictive cumulative distribution function denoted by $F_{\boldsymbol{\theta}}(y \mid \tau(\mathbf{x}_i)) = P_{\boldsymbol{\theta}}(Y \leq y \mid \tau(X) = \tau(\mathbf{x}_i))$. Let $(y_{ij})_{j=1}^{M}$ be the samples generated by Algorithm 1, and let $\widehat{q}_{\frac{\alpha}{2}}(\tau(\mathbf{x}_i))$ and $\widehat{q}_{1-\frac{\alpha}{2}}(\tau(\mathbf{x}_i))$ denote the empirical $\alpha/2$ and $(1-\alpha/2)$ quantiles of $(y_{ij})_{j=1}^{M}$, respectively. Define the empirical prediction interval $\widehat{C}_{1-\alpha}(\tau(\mathbf{x}_i)) := \left[\widehat{q}_{\alpha/2}(\tau(\mathbf{x}_i)), \widehat{q}_{1-\alpha/2}(\tau(\mathbf{x}_i))\right]$. Assume that: (i) the importance ratio $\frac{p_{\boldsymbol{\theta}_2}(\mathbf{z}|\tau(\mathbf{x}))}{q_{\boldsymbol{\varphi}}(\mathbf{z}|\tau(\mathbf{x}))}$ is uniformly bounded, and (ii) $F_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}_i))$ is continuous and strictly increasing in neighborhoods of the $\alpha/2$ and $(1-\alpha/2)$ quantiles. Then*

$$\lim_{L\to\infty, M\to\infty} P_{\boldsymbol{\theta}}\left(Y \in \widehat{C}_{1-\alpha}(\tau(\mathbf{x}_i)) \mid \tau(X) = \tau(\mathbf{x}_i)\right) = 1 - \alpha.$$

Corollary 1 the validity of uncertainty quantification based on the SIR procedure. This result primarily applies to continuous responses, as it requires $F_{\boldsymbol{\theta}}(\cdot \mid \tau(\mathbf{x}_i))$ to be continuous in neighborhoods of the relevant quantiles. The proofs of the main results are provided in Appendix C.

5 Real data application

In the real-data experiments, we evaluate the domain generalization capability of the proposed framework on three medical applications: (i) dementia classification from clinical consultation notes, (ii) magnetic resonance imaging (MRI) reconstruction from undersampled k-space measurements, and (iii) 4D Flow MRI super-resolution. These tasks span a diverse range of data modalities, progressing from one-dimensional text data to two-dimensional imaging data and three-dimensional velocity field data. In each application, the training and testing domains differ, requiring the learned representation to preserve predictive information while facilitating cross-domain transfer. An overview of the datasets and domain configurations is provided in Figure 1.

5.1 Dementia classification

The UCSF dementia dataset comprises 39381 clinical consultation notes collected across Neurology, Geriatric Medicine, Psychiatry, and Primary Care, with substantial cross-department variation in writing style, terminology, and documentation practices. We formulate dementia identification as a

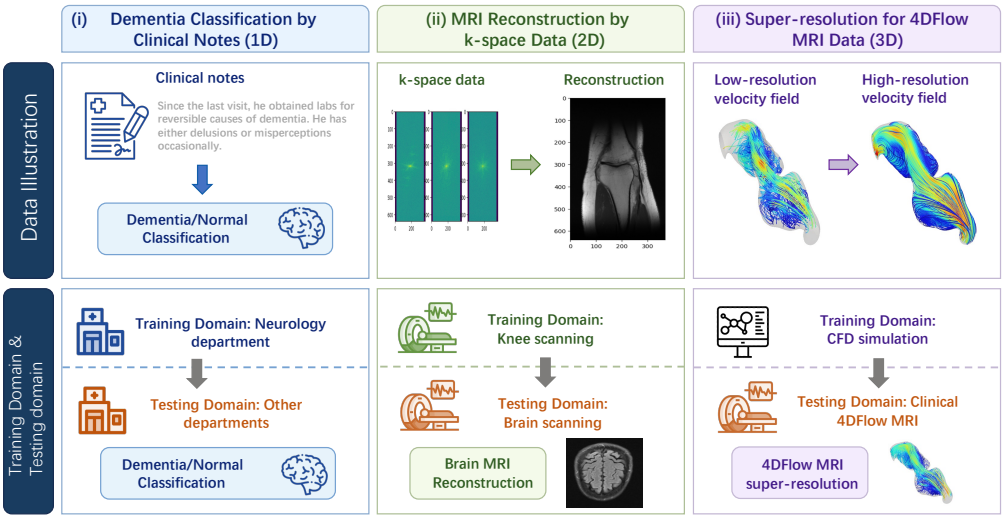


Figure 1: Illustration of the datasets in real data analysis and domain-shift settings.

Table 1: Comparison of metrics for dementia binary classification.

		CNN	DANN	MDD	CNN+Eng	CVAE	IWCAE
Accuracy	In-domain	0.832	0.821	0.832	0.809	0.825	**0.848**
	OOD1	0.731	0.698	0.752	0.724	0.737	**0.790**
	OOD2	0.668	0.621	0.703	0.672	0.682	**0.753**

binary classification task. The model is trained using clinical notes from the Neurology Department, with corresponding labels indicating the presence or absence of dementia. The in-domain data were split into training and testing sets at an 80:20 ratio, comprising 2280 training notes (757 dementia and 1523 non-dementia) and 571 test notes (189 dementia and 382 non-dementia). The out-of-domain (OOD) test set comprised 918 clinical notes describing neurological symptoms, including 280 dementia cases and 638 non-dementia cases. We consider two out-of-domain settings: OOD1 combines Neurology and non-Neurology notes, whereas OOD2 contains only non-Neurology notes.

We generate embeddings from the clinical notes using ClinicalBERT (Huang et al., 2019). These embeddings serve as the input to the deep learning models. We compare IWCAE with CNN, domain adversarial neural network (DANN) (Ganin et al., 2016), margin disparity discrepancy (MDD) (Zhang et al., 2019), Engression with CNN backbone (CNN+Eng) (Shen & Meinshausen, 2025), and Conditional Variational Autoencoder (CVAE) (Sohn et al., 2015); the CNN-based methods use the same backbone to ensure a fair comparison. Table 1 reports the classification accuracy across testing domains. IWCAE achieves the best overall performance in both OOD settings. Furthermore, our method quantifies the prediction uncertainty by the width of the 95% prediction interval obtained from the SIR samples. Figure 2 shows that correctly classified notes consistently exhibit lower uncertainty than misclassified notes across all settings. The difference between the two groups is most prominent in the test sample without domain shift. This suggests that lower uncertainty is generally associated with better prediction performance. In practice, uncertainty estimates can therefore be used to assess the reliability of individual predictions. Additional details on preprocessing, model architectures, hyperparameter settings, classification metrics, and ROC curves are provided in Appendix A.3.1.

5.2 MRI RECONSTRUCTION

We evaluate IWCAE on accelerating MRI reconstruction using the multi-coil FastMRI dataset (Zbontar et al., 2018). The study includes 1000 knee MRI images for training, 500 knee MRI images for in-domain evaluation, and 500 brain images for out-of-domain evaluation. For each image, 10 receiver coils are retained, the complex-valued k-space data are cropped to 600×320, and the real

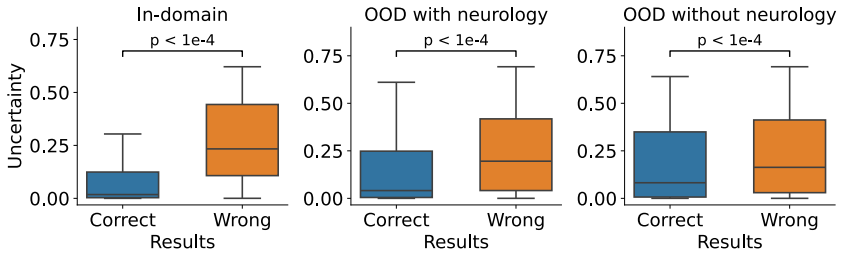


Figure 2: Box plots of predictive uncertainty for dementia classification under the in-domain and two out-of-domain evaluation settings. The $p < 1e^{-4}$ indicates correctly classified samples exhibit significantly lower predictive uncertainty than misclassified samples.

Table 2: Comparison of NMSE and SSIM for MRI reconstruction. Values are reported as mean (standard deviation).

		VN	VN+Eng	CVAE	CVAE+MMD	IWCAE
NMSE	Knee	0.017(0.007)	0.018(0.007)	0.019(0.008)	0.029(0.024)	**0.016**(0.006)
	Brain	0.074(0.063)	0.077(0.067)	0.038(0.026)	0.519(0.130)	**0.034**(0.022)
SSIM	Knee	0.830(0.081)	0.830(0.081)	0.827(0.084)	0.791(0.095)	**0.849**(0.074)
	Brain	0.742(0.132)	0.740(0.135)	0.827(0.130)	0.715(0.111)	**0.848**(0.125)

and imaginary components are represented as separate channels. We apply a four-fold undersampling mask to the k-space data to generate the model input, while the fully sampled image reconstructed using the inverse Fourier transform serves as the target. The goal is to accelerate MRI acquisition by collecting only a fraction of the k-space data, thereby reducing scan time. The trained model then reconstructs a high-quality image that closely matches the fully sampled MRI. We compare IWCAE with the Variational Network (VN) (Hammernik et al., 2018), Engression with VN backbone (VN+Eng) (Shen & Meinshausen, 2025), CVAE, and CVAE+MMD (Lotfollahi et al., 2020) on the task of reconstructing fully sampled MRI images.

Reconstruction performance is evaluated using normalized mean squared error (NMSE) and structural similarity index measure (SSIM) (Wang et al., 2003). NMSE and SSIM are computed for each test image, and the mean and standard deviation across all knee and brain MRI test images are reported in Table 2. IWCAE achieves the best overall performance in both the in-domain (knee) and OOD (brain) settings. Using the SIR procedure in Algorithm 1, IWCAE generates multiple reconstruction samples for each input, from which pixelwise 95% prediction-interval widths and image-level mean relative uncertainty are computed. Figure 3 shows a clear negative association between mean relative uncertainty and SSIM in both knee and brain MRI, indicating that the proposed uncertainty measure is informative of reconstruction fidelity and may help assess prediction quality when ground-truth images are unavailable. Additional details on preprocessing, model architectures, hyperparameter settings, uncertainty calculation, and results comparison are provided in Appendix A.3.2.

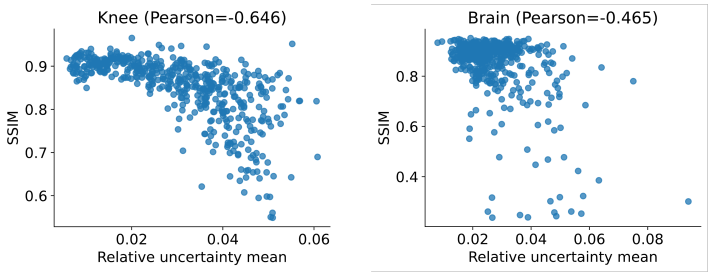


Figure 3: Scatter plots of the mean relative uncertainty versus SSIM for knee and brain MRI. The Pearson correlation coefficient between the two quantities is reported in the title of each panel.

5.3 SUPER-RESOLUTION OF 4DFLOW MRI DATA

In this section, we consider the super-resolution of 4D Flow MRI (4DF) data. 4D Flow MRI provides in vivo measurements of blood flow dynamics while suffers from low spatial resolution and imaging artifacts. A common strategy to improve the resolution of 4DF data is to learn a mapping from low-resolution to high-resolution flow fields using computational fluid dynamics (CFD) simulations. Specifically, high-resolution CFD data and their downsampled counterparts are used for training, while real 4D Flow MRI data are used for testing, creating a substantial domain shift between simulated and acquired flow fields. The dataset contains nine vascular geometries: seven with CFD data only and two with paired CFD and 4DF data, comprising one normal geometry and one aneurysmal geometry. The preprocessing procedure is followed from Wen & Jiang (2026).

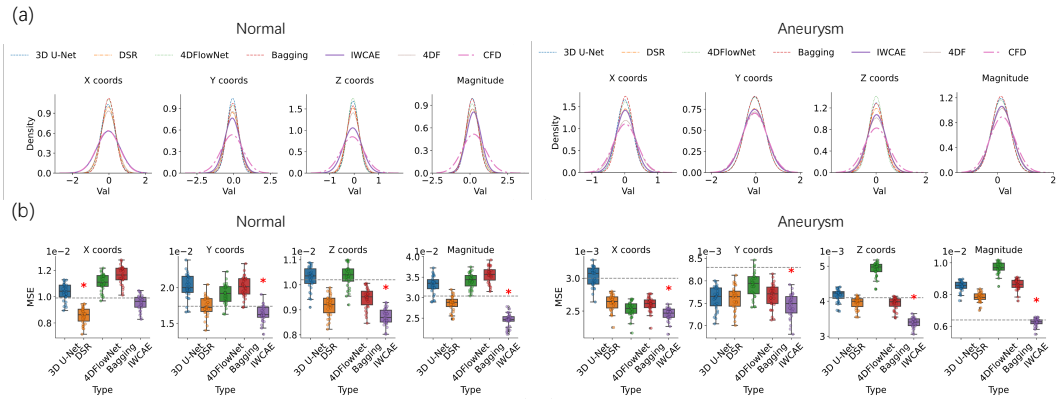


Figure 4: Comparison of 3D U-Net, DSR, 4DFlowNet, Bagging and IWCAE on normal and aneurysmal vascular geometries. (a) Estimated velocity distributions produced by each method, compared with the high-resolution CFD reference and the low-resolution 4D Flow MRI measurements. (b) MSE of the reconstructed velocity fields for the five methods. The gray dashed lines indicate the MSE between the low-resolution 4D Flow MRI data and the high-resolution CFD reference. The red star marks the method achieving the lowest mean MSE over 30 independent trials.

We compare proposed IWCAE model with 3D U-Net (Çiçek et al., 2016), distributional super-resolution (DSR) (Wen & Jiang, 2026), 4DFlowNet (Ferdian et al., 2020), and Bagging (Ericsson et al., 2024). As shown in Figure 4(a), our method provides estimated velocity distributions that closely match those obtained from high-resolution CFD simulations in both normal and aneurysmal settings. In addition, Figure 4(b) shows that IWCAE achieves the lowest MSE across all three velocity components and velocity magnitude for both normal and aneurysmal geometries, demonstrating improved robustness to the domain shift from CFD to 4D Flow MRI. Note that, we do not provide uncertainty analysis results because all the data are patches extracted from two 4DF test sample, the uncertainty measures on two samples do not offer clinicial meaning intepretation. Additional details on preprocessing, model architectures and hyperparameter settings are provided in Appendix A.3.3.

6 CONCLUSION

In this work, we propose an importance-weighted conditional variational autoencoder (IWCAE) that combines distributional learning with marginal representation alignment for unsupervised domain adaptation. Theoretically, we establish the asymptotic validity of the proposed resampling-based inference procedure. Experiments on three medical datasets demonstrate that IWCAE consistently outperforms existing methods under distribution shift, highlighting its effectiveness and broad applicability.

Reproducibility statement

All implementation code and data are publicly available at `https://anonymous.4open.science/r/IWCAE-BEE4/`. Additional implementation details, including hyperparameter settings and technical proofs, are provided in the Appendix.

References

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. In *Advances in Neural Information Processing Systems*, pp. 480–493, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Technical report, 2019. (`https://arxiv.org/abs/1907.02893`).

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *Conference on Artificial Intelligence and Statistics*, pp. 66–75, 2019.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9):2137–2155, 2009.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016.

Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. Location and scatter matching for dataset shift in text mining. In *IEEE International Conference on Data Mining*, pp. 773–778, 2010.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pp. 424–432, 2016.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *International Conference on Machine Learning*, pp. 193–200, 2007.

Hal Daumé III. Frustratingly easy domain adaptation. In *45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, 2007.

Subhadeep Dey, Srikanth Madikeri, and Petr Motlicek. Information theoretic clustering for unsupervised domain-adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5580–5584, 2016.

P Kingma Diederik and Welling Max. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pp. 200–216, 2020.

Leon Ericsson, Adam Hjalmarsson, Muhammad Usman Akbar, Edward Ferdian, Mia Bonini, Brandon Hardy, Jonas Schollenberger, Maria Aristova, Patrick Winter, Nicholas Burris, et al. Generalized super-resolution 4d flow mri-using ensemble learning to extend across the cardiovascular system. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7239–7250, 2024.

Edward Ferdian, Avan Suinesiaputra, David J Dubowitz, Debbie Zhao, Alan Wang, Brett Cowan, and Alistair A Young. 4dflownet: super-resolution 4d flow mri using deep learning and computational fluid dynamics. *Frontiers in Physics*, 8:138, 2020.

Michael C Fu and Jian-Qiang Hu. *Conditional Monte Carlo: Gradient estimation and optimization applications*. Springer Science & Business Media, 2012.

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(59):1–35, 2016.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, pp. 3290–3300, 2020.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pp. 2839–2848, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520, 2006.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

Mahta Hassan Pour Zonoozi, Vahid Seydi, and Mahmood Deypir. An unsupervised adversarial domain adaptation based on variational auto-encoder. *Machine Learning*, 114(5):128, 2025.

Yifei He, Haoxiang Wang, Bo Li, and Han Zhao. Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25(361):1–40, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, pp. 601–608, 2006.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. Technical report, 2019. (`https://arxiv.org/abs/1904.05342`).

O Ivanov, M Figurnov, and D Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019.

Fredrik D. Johansson, David A. Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 527–536, 2019.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: a review. *Journal of Imaging*, 9(4):81, 2023.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pp. 6965–6975, 2018.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

Sang Min Lee, Sang-Youn Park, and Byoung-Ho Choi. Application of domain-adaptive convolutional variational autoencoder for stress-state prediction. *Knowledge-Based Systems*, 248:108827, 2022.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *AAAI Conference on Artificial Intelligence*, pp. 7399–7407, 2022.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI Conference on Artificial Intelligence*, pp. 3579–3587, 2018.

Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In *IEEE International Conference on Computer Vision*, pp. 10367–10376, 2021.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and Philip S Yu. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2013.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1647–1657, 2018.

Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J Theis, and F Alexander Wolf. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36(26):i610–i617, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision*, pp. 1406–1415, 2019.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems*, pp. 2360–2368, 2016.

Jiajun Qiu, Yao Hu, Li Li, Abdullah Mesut Erzurumluoglu, Ingrid Braenne, Charles Whitehurst, Jochen Schmitz, Jatin Arora, Boris Alexander Bartholdy, Shrey Gandhi, et al. Deep representation learning for clustering longitudinal survival data from electronic health records. *Nature Communications*, 16(1):2534, 2025.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286, 2014.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Donald B Rubin. Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):542–543, 1987.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(7): 653–677, 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pp. 3483–3491, 2015.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5):985–1005, 2007.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *Conference on Artificial Intelligence and Statistics*, pp. 781–788, 2010.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 505–513, 2011.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems*, pp. 19276–19289, 2020.

Aboozar Taherkhani, Georgina Cosma, and T Martin McGinnity. Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, 404:351–366, 2020.

Qian Wang and Toby P Breckon. Generalized zero-shot domain adaptation via coupled conditional variational autoencoders. *Neural Networks*, 163:40–52, 2023.

Yixin Wang, Shuang Qiu, Dan Li, Changde Du, Bao-Liang Lu, and Huiguang He. Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1612–1626, 2022.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pp. 1398–1402, 2003.

Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *International Conference on Pattern Recognition*, pp. 9756–9763, 2021.

Xiaoyi Wen and Fei Jiang. Distributional deep learning for super-resolution of 4d flow mri under domain shift. Technical report, 2026. (`https://arxiv.org/abs/2602.15167`).

Fuping Wu and Xiahai Zhuang. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12):3555–3567, 2021.

Keru Wu, Yuansi Chen, Wooseok Ha, and Bin Yu. Prominent roles of conditionally invariant components in domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 26(110):1–92, 2025a.

Xingjian Wu, Xiangfei Qiu, Hongfan Gao, Jilin Hu, Bin Yang, and Chenjuan Guo. k^2 vae: A koopman-kalman enhanced variational autoencoder for probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 67562–67583, 2025b.

Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1855–1862, 2010.

Jialin Yu, Andreas Koukorinis, Nicolò Colombo, Yuchen Zhu, and Ricardo Silva. Structured learning of compositional sequential interventions. In *Advances in Neural Information Processing Systems*, pp. 115409–115439, 2024.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. Technical report, 2018. (`https://arxiv.org/pdf/1811.08839`).

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413, 2019.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. *Journal of Machine Learning Research*, 23(340):1–49, 2022.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *AAAI Conference on Artificial Intelligence*, pp. 19643–19651, 2024.

Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-Draa. Domain generalization via optimal transport with metric similarity learning. *Neurocomputing*, 456:469–480, 2021.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI*, pp. 4119–4125, 2015.