Domain Generalized Medical Landmark Detection via Robust Boundary-Aware Pre-Training

Haifan Gong[1,2], Yu Lu[3,4], Xiang Wan[1], Haofeng Li[1]*
[1]Shenzhen Research Institute of Big Data, Shenzhen, China
[2]The Chinese University of Hong Kong, Shenzhen, China
[3]University of California, Merced, CA, USA
[4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
haifangong@link.cuhk.edu.cn, ylu54@ucmerced.edu, {wanxiang, lhaof}@sribd.cn

Abstract

In recent years, deep learning has revenue in automated medical landmark detection. Nonetheless, prevailing research in this field predominantly addresses single-center scenarios or domain adaptation settings. In practical environments, the acquisition of multi-center data faces privacy concerns, coupled with the time-intensive and costly nature of data collection and annotation. These challenges substantially impede the broader application of deep learning-based medical landmark detection. To mitigate these issues, we propose a novel domain-generalized medical landmark detection framework that relies solely on single-center data for training. Considering the availability of numerous public medical segmentation datasets, we design a simple yet effective method that utilizes single-center segmentation to enhance the domain generalization capabilities of the landmark detection task. Specifically, we introduce a novel boundary-aware pre-training approach to focus the model on regions pertinent to landmarks. To further enhance the robustness and generalization capabilities during pre-training, we have derived a mixing loss term and proved its effectiveness in theory and practice. Extensive experiments conducted on our new domain generalization benchmark for medical landmark detection demonstrate the superiority of our approach.

Code — https://github.com/lhaof/DGMLD

Introduction

Medical landmark detection is pivotal in disease diagnosis (Payette et al. 2021; Avisdris et al. 2022) and surgical planning (Li et al. 2023). For instance, the detection of fetal cerebellar landmarks is crucial for monitoring fetal neurodevelopment (Garel and Garel 2004; Zhou et al. 2021; Vahedifard et al. 2023), and accurate identification of these landmarks is essential for assessing the fetal growth trajectory (Garel and Garel 2004; Hughes et al. 2023; Han et al. 2024; Chen et al. 2024). Given that manual labeling is laborious and subjective, numerous studies have employed deep learning for automate medical landmark detection (Avisdris et al. 2021; Li et al. 2023; Jin, Che, and Chen 2023; Gong et al. 2025). Despite the successes of deep-learning-based algorithms for medical landmark detection, their widespread application is often impeded by the domain gap due to variations in medical imaging equipment. Previous works (Jin, Che, and Chen 2023) have addressed this issue under the unsupervised domain adaptation (UDA) setting, which involves training neural networks using available source domain data-label pairs and target domain data. However, due to privacy concerns, target domain data may not be publicly available in real-world scenarios. Furthermore, data collection can also be labor-intensive (Viswanathan, Parmar, and Madabhushi 2024). In response, this work introduces the domain generalization (DG) setting for medical landmark detection, where only source domain data and corresponding labels are utilized (Zhou et al. 2022; Li et al. 2024b). To the best of our knowledge, there is no publicly available benchmark for the domain-generalized medical landmark detection task. Therefore, we contribute a benchmark encompassing data from three centers for medical landmark DG.

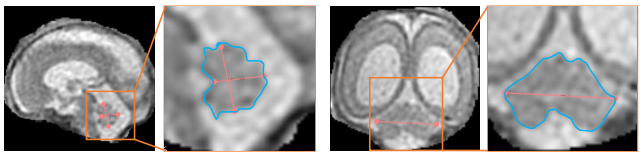

Figure 1: Illustration of medical landmark detection in 3D MRI scans. The fetal cerebellum, highlighted by the blue line, is the anatomical structure of interest, with landmarks indicated by red points. This figure shows that the landmarks are positioned on the boundary of the anatomical structure.

Existing DG approaches have predominantly focused on tasks involving natural images (Zhao et al. 2020; Zhou et al. 2022), addressing challenges in segmentation (Long et al. 2024; Gu et al. 2024) and classification (Li et al. 2020; Qu et al. 2023), often employing augmentation (Peng, Zheng, and Chen 2024) or alignment technologies (Long et al. 2024). However, these approaches face several limitations when applied to DG medical landmark detection: 1. Medical images typically exhibit low contrast, high noise levels, and uncertain boundaries compared to natural images (Xu et al. 2023c), making it challenging to transfer methodologies developed for natural images directly. 2. The

*Corresponding author.

landmark detection task, generally formulated as Gaussian heatmap regression (Li et al. 2023), differs from classification or segmentation tasks. Recognizing that most medical landmarks are located on the boundaries of anatomical structures (Jimenez-del Toro et al. 2016; Avisdris et al. 2022; Li et al. 2023) (as illustrated in Fig. 1), we leverage prior knowledge of anatomical structures to pre-train the neural network, thereby enhancing its ability to generalize to unseen domains. Given the abundance of available medical segmentation datasets (Gholipour et al. 2017; Wu et al. 2021; Fidon et al. 2022), we establish a pre-training pipeline focused on boundary detection using an auxiliary dataset with voxel-level segmentation masks of landmark-related structures. While directly using segmentation masks for pre-training is straightforward, several disparities exist between segmentation and landmark detection tasks: 1. Segmentation focuses on regions (Li et al. 2023), whereas landmark detection targets specific points along boundaries where foreground and background are more imbalanced. 2. Segmentation predicts discrete values, whereas landmark detection typically formulates as heatmap regression with continuous values derived from a Gaussian distribution (Li et al. 2023; Jin, Che, and Chen 2023). To bridge these gaps, we propose a simple yet effective pre-training pipeline by extracting boundary distance maps from the segmentation masks and using these maps to pre-train the neural network. This approach inherently emphasizes corners on a structure's boundary, where landmarks are often situated. Furthermore, to enhance the robustness of the pre-trained backbone, we employ a mixed term of log-cosh loss and mean square error loss for pre-training. Our comprehensive experiments across various network backbones and dataset distributions confirm the effectiveness of our pre-training method in significantly improving the accuracy of medical landmark detection. The contributions of our work are as follows:

1. We investigate the DG medical landmark detection task for the first time, exploring the feasibility of using existing segmentation masks to boost the DG medical landmark detection task.
2. We tailor-design a DG framework for medical landmark detection and introduce a novel boundary-aware pre-training method designed for this context. We also construct a robust mixing loss function for pre-training and provide both theoretical and experimental analyses on the effectiveness of the loss function.
3. We contribute the first benchmark for the DG medical landmark detection task, which includes data from three different centers. Extensive experiments conducted with various network backbones on our benchmark confirm the effectiveness of our approach.

Related Work

Medical Landmark Detection

With the advent of deep learning, a paradigm shift has occurred for landmark detection in medical imaging (Liu et al. 2020; Juneja et al. 2021). Xu et al. (2019) introduced a 3D UNet architecture to accurately identify landmarks in fetal joints. Liu et al. (2020) designed an anatomically constrained neural network to facilitate preoperative measurements for total hip arthroplasty using 2D X-ray images. Xu et al. (2020a) employed a conditional generative adversarial network to deduce fetal poses from 3D imaging. Avisdris et al. (2021) utilized convolutional neural networks to detect fetal landmarks within the scope of 2D brain segmentation tasks. Shankar et al. (2022) devised a method that incorporates clinically pertinent biometric constraints to measure fetal brain biometrics from 2D ultrasound images. A dynamic methodology for determining the orientation of 2D fetal landmarks using ultrasound was developed by (Avisdris et al. 2022). Zhou et al. (2023) innovated a technique to regularize the predicted heatmap using the Swoosh Activation Function. Li et al. (2023) proposed a transformer-based multi-task learning framework for addressing the landmark detection task. Li et al. (2024b) built a feature decouple and gated recalibration network for medical landmark detection. Li et al. (2024a) designed a spatio-temporal graph convolutional network for ultrasound landmark detection. Feng et al. (2024) devised a Bayesian network for simultaneous keyframe and landmark detection in ultrasonic.

Nevertheless, previous studies have primarily focused on single-domain settings, largely overlooking the DG issues inherent in medical landmark detection. While some work (Avisdris et al. 2022) has employed data augmentation to enhance testing performance, these approaches have not proven effective in DG settings. To the best of our knowledge, this is the first study to address domain-generalized medical landmark detection by incorporating additional boundary priors.

Domain Generalization

DG focuses on learning domain-invariant representations and can be categorized into three primary types (Gong et al. 2022b,a; Xu et al. 2023b; Long et al. 2024; Peng, Zheng, and Chen 2024): domain alignment, meta-learning, and augmentation strategies. In domain alignment, Zhao et al. (2020) enhances conditional invariance by incorporating an entropy regularization term to improve classifier generalization, while Matsuura and Harada (2020) iteratively segregates samples into latent domains via clustering. Long et al. (2024) propose a new discriminate learning approach for landmark detection. Regarding meta-learning, Li et al. (2018) introduces a model-agnostic training procedure that simulates domain shifts during training, and Qiao, Zhao, and Peng (2020) tailors meta-learning to single-domain generalization. For augmentation strategies, Zhao et al. (2020) innovates a regularization term derived from the information bottleneck principle for adversarial data augmentation, and Zhao et al. (2022) develops a style hallucination module to generate diversified samples essential for generalization. Peng, Zheng, and Chen (2024) propose a strong-weak augmentation method in conjunction with adversarial learning for domain generalization.

Unlike the above works, which mainly focus on natural images with less noise and high contrast, we take the first step in detecting DG medical landmarks. The work most closely related to ours is proposed by (Jin, Che, and Chen

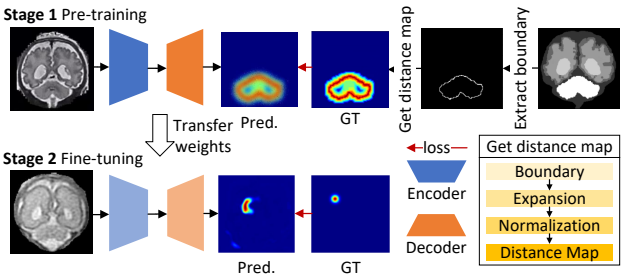

Figure 2: Overview of the proposed boundary-distance map-based pre-training pipeline. Stage 1 involves extracting the boundary distance map from segmentation data to pre-train the model for subsequent tasks. In Stage 2, the model is fine-tuned using the pre-trained weights specifically for the landmark detection task.

2023) and (Gu et al. 2024). The former (Jin, Che, and Chen 2023) focused on the medical landmark detection under the unsupervised domain adaptation setting while we focused on DG. Pre-training has been a popular technology in the medical domain (Gong et al. 2021; Kang et al. 2023). Gu et al. (2024) aims to address DG in medical image segmentation by extracting edge canny features for pre-training, while our work focuses on using the available high-level segmentation masks to boost the DG capability for the medical landmark detection task. Moreover, we first derive and construct a mixing term of loss functions that aims to obtain better generalization ability through the pre-training process.

Methodology

We introduce the Boundary-aware Pre-training (BaP) framework for boosting DG medical landmark detection, as illustrated in Fig. 2. The key idea of BaP is the infusion of structural knowledge into the neural network, thereby encouraging the predicted landmarks to align precisely with the organ's boundary. This is achieved by pre-training the network to be aware of the boundary's position, effectively priming the network for more accurate landmark localization in out-of-domain data.

Pre-training with Boundary Distance Map

We utilize the publicly available Atlas dataset collected by (Xu et al. 2023c) for the BaP, which provides the landmark annotation and organ segmentation masks. The process begins with extracting organ boundaries relative to the dataset's landmarks, then generating boundary heatmaps for regression purposes. We deploy specialized convolution operations to accurately extract the landmark boundary from the segmentation mask M of shape (H, W, D). These are defined through the expansion function F_{exp} and the erosion function F_{ero}. Here, the foreground voxels in M are set to 1, and the background voxels to 0. We then define the boundary β as:

$$\beta = F_{\text{ero}}(F_{\text{exp}}(M)) - F_{\text{ero}}(\mathbf{1} - M), \tag{1}$$

where $\mathbf{1}$ represents a matrix of ones of the same shape as M. Following this, we construct the boundary distance map $\Phi(x)$ from β, assigning weights to voxels that diminish with increasing distance from the boundary. The distance map Φ_o is formed by iteratively applying F_{exp} to the boundary:

$$\Phi_o = \frac{1}{N}\sum_{i=1}^{N}\left(F_{\text{exp}}^{i}(\beta)\right), \tag{2}$$

where $F_{\text{exp}}^{i}(\beta)$ represents the i-th iterative expansion of the boundary β, and N denotes the number of iterations, reflecting the boundary distance range at the voxel level, which is set to 5 according to the kernel size to generate downstream heatmap for landmark detection (Li et al. 2023).

For the BaP loss function, we introduce a robust training pipeline named *generalization all the time*. This is based on the straightforward observation that loss values are typically larger at the beginning of training and decrease as training progresses. To leverage this behavior, we employ two distinct loss functions: one (log-cosh loss) that is more robust at the beginning of training and another (Mean Squared Error, MSE) that is more effective towards the end of training. We have identified that the log-cosh loss function exhibits greater robustness than the MSE loss when dealing with larger loss values, which typically occur at the start of the training phase. Conversely, MSE becomes more effective as the loss values diminish. Let e be the current epoch, and E be the total epochs; we have developed the following loss function L to measure the discrepancy between the network's predicted heatmap and the target boundary distance map in the pre-training stage:

$$L = \frac{e}{E}\text{MSE}(f(\mathbf{x};\theta), \Phi_o) + \frac{E-e}{E}\text{logcosh}\left(f(\mathbf{x};\theta) - \Phi_o\right), \tag{3}$$

where $\frac{e}{E}$ is a linearly increasing weight function, facilitating the transition from **logcosh** loss, which provides robustness in the early training stages. The MSE loss is used for robust training in later stages. This structured approach underscores the importance of adaptive loss functions in achieving more robust initialized weight for the downstream landmark detection task. The full theoretical analysis is shown below.

Generalization Ability Between logcosh and MSE

In the pre-training phase, we construct a mixed term combining the commonly used MSE loss and the **logcosh** loss function. The primary motivation for this choice is the demonstrated superior generalization capability of **logcosh** compared to MSE, particularly under conditions where the predictions exhibit a larger gap to the ground truth—a scenario typically encountered at the beginning of training. Unlike the previous works (Xu et al. 2020b; Saleh and Saleh 2022; Jeendgar et al. 2022; Gong et al. 2023; Chan et al. 2024), we first investigate the **logcosh** on the generalization of medical landmark detection task. The comparative analysis of these loss functions encompasses three key aspects, which are outlined below:

A. Robustness Analysis We start with the loss function itself. For the **logcosh**, we have

$$\mathbf{logcosh}(\delta) \approx \begin{cases} \log(1+\frac{\delta^2}{2}) \approx \frac{\delta^2}{2} & \text{if } \delta \to 0 \\ \log(\frac{\exp(\delta)}{2}) = \delta - \log(2) & \text{if } \delta \to \infty, \end{cases} \tag{4}$$

where we define $\delta = |f(\mathbf{x};\theta) - \Phi_o|$. Hence, it is easy to see that the **logcosh** function grows linearly for large errors. It grows quadratically but at half the rate of MSE for small errors. Next, the sensitivity to changes can be verified by examining the gradients of both functions:

$$\frac{\partial \mathbf{logcosh}(\delta)}{\partial \delta} = tanh(\delta) \approx \begin{cases} \delta & \text{if } \delta \to 0 \\ 1 & \text{if } \delta \to \infty, \end{cases} \tag{5}$$

and

$$\frac{\partial \mathbf{MSE}(\delta)}{\partial \delta} = 2\delta. \tag{6}$$

This demonstrates that **logcosh** prevents large errors from disproportionately influencing the learning process. Finally, we can also examine the Hessian (curvature) of both functions:

$$\frac{\partial^2 \mathbf{logcosh}(\delta)}{\partial \delta^2} = \frac{\partial tanh(\delta)}{\partial \delta} = 1 - tanh^2(\delta), \tag{7}$$

and

$$\frac{\partial^2 \mathbf{MSE}(\delta)}{\partial \delta^2} = 2. \tag{8}$$

Since $\tanh^2(\delta) \in [0,1]$, **logcosh** has less curvature than MSE, implying it penalizes large errors less severely and is more robust to outliers.

B. Generalization Bound Analysis With the gradient and Hessian from the previous section, we now focus on the generalization bound (Sammut and Webb 2011; Akbari et al. 2021), which is defined as

$$R(f_\theta) \leq \hat{R}(f_\theta) + \text{Complexity Term} + \text{Slack Term}, \tag{9}$$

where $R(f_\theta)$is the expected loss of the model f_θ on unseen data and $\hat{R}(f_\theta)$ is the average loss on the training data. These are defined as follows:

$$R(f_\theta) = \mathbb{E}_{(\mathbf{x},\Phi_o)\sim P}[\ell(f(\mathbf{x},\theta),\Phi_o)] \tag{10}$$

and

$$\hat{R}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x},\theta),\Phi_o)). \tag{11}$$

The **Complexity Term** is related to the model's capacity and is often represented by the Rademacher complexity (Bartlett and Mendelson 2002; Yin, Kannan, and Bartlett 2019), which is defined as

$$\hat{C}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{\ell \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(\mathbf{x},\theta),\Phi_o) \right], \tag{12}$$

where σ are i.i.d. Rademacher variables taking values of ± 1 and $\mathcal{H}$ is the model class. The **Slack Term** is influenced by the range of the loss function (VAPNIK 1998) and is derived from Hoeffding's Inequality (Mohri, Rostamizadeh, and Talwalkar 2018):

$$\Pr\left(|\hat{R}(f_\theta) - R(f_\theta)| \geq s\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{13}$$

where n is the number of data points in your training set and $f_\theta \in [a,b]$. Now, we assume the probability of the deviation exceeding t is at most probability p. That is,

$$2\exp\left(-\frac{2ns^2}{(b-a)^2}\right) = p \tag{14}$$

The equivalent expression for slack term s is

$$s = \sqrt{\frac{(b-a)^2}{2n}\log\left(\frac{2}{p}\right)}, \tag{15}$$

and we can define $B = b - a$.

Theorem 1. *Considering the same model with two different loss functions, namely **logcosh** and MSE, assume that the range of reconstruction using **logcosh** is smaller than using MSE, then the upper generalization bound for **logcosh** is smaller than that for MSE.*

Proof. **First**, we show that $\hat{R}_{\mathbf{logcosh}}(f_\theta) \leq \hat{R}_{\mathbf{MSE}}(f_\theta)$. Consider

$$\hat{R}_{\mathbf{logcosh}}(f_\theta) - \hat{R}_{\mathbf{MSE}}(f_\theta) \tag{16}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\log(cosh(\delta)) - \delta^2), \tag{17}$$

where $\delta_i = f_\theta(\mathbf{x_i}) - \Phi_o$. Define function $g(\delta) := (\log(cosh(\delta)) - \delta^2)$. Taking the derivative of $g(\delta)$, we get

$$g'(\delta) = tanh(\delta) - 2\delta. \tag{18}$$

This derivative indicates that $g(\delta)$ increases for $\delta < 0$ and decreases for $\delta > 0$. Thus,

$$\max_\delta \; g(\delta) = g(0) = 0, \tag{19}$$

which shows $\hat{R}_{\mathbf{logcosh}}(f_\theta) \leq \hat{R}_{\mathbf{MSE}}(f_\theta)$ for any model reconstruction f_θ.

Second, we show $\hat{C}_n(\mathcal{H}_{\mathbf{logcosh}}) < \hat{C}_n(\mathcal{H}_{\mathbf{MSE}})$ when the model architectures are the same. Consider

$$\hat{C}_n(\mathcal{H}_{\mathbf{logcosh}}) - \hat{C}_n(\mathcal{H}_{\mathbf{MSE}}) \tag{20}$$

$$= \mathbb{E}_\sigma \left[\sup_{\ell \in \mathcal{H}_{\mathbf{logcosh}}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(\mathbf{x},\theta),\Phi_o) \right] \tag{21}$$

$$- \mathbb{E}_\sigma \left[\sup_{\ell \in \mathcal{H}_{\mathbf{MSE}}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(\mathbf{x},\theta),\Phi_o) \right] \tag{22}$$

$$= \mathbb{E}_\sigma \left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{logcosh}(f(\mathbf{x},\theta) - \Phi_o) \right] \tag{23}$$

$$- \mathbb{E}_\sigma \left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{MSE}(f(\mathbf{x},\theta) - \Phi_o) \right] \tag{24}$$

$$= \mathbb{E}_\sigma \left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\mathbf{logcosh}(\delta) - \mathbf{MSE}(\delta)\right) \right] \tag{25}$$

$$\leq 0, \tag{26}$$

the second equal holds because we use different loss functions but the same model architecture. Therefore, $\hat{C}_n(\mathcal{H}_{\textbf{logcosh}}) \leq \hat{C}_n(\mathcal{H}_{\textbf{MSE}})$ when the model architectures are the same.

Third, we show slack term $s_{\textbf{logcosh}} < s_{\textbf{MSE}}$. Let $B = b - a$ be the range of reconstruction, we have the following

$$s_{\textbf{logcosh}} - s_{\textbf{MSE}} \tag{27}$$

$$= \sqrt{\frac{B^2_{\textbf{logcosh}}}{2n} \log\left(\frac{2}{p}\right)} - \sqrt{\frac{B^2_{\textbf{MSE}}}{2n} \log\left(\frac{2}{p}\right)} \tag{28}$$

$$= \sqrt{\frac{1}{2n} \log(\frac{2}{p})}\ (B_{\textbf{logcosh}} - B_{\textbf{MSE}}) \tag{29}$$

$$< 0. \tag{30}$$

Since all three terms for **logcosh** are less than or equal to those for MSE, it is natural to conclude that MSE will have a higher upper generalization bound compared to **logcosh** in general cases. Thus, we can enhance the model's generalization ability by embedding the **logcosh** function into the design of the loss function. □

C. Tighter Bounds Analysis Although the previous two sections have directly demonstrated the superior generalization of **logcosh**, it is still worth measuring performance differences using Lipschitz continuity (Hager 1979).

Theorem 2. *Consider a model with two loss functions:* ***Logcosh*** *and MSE. The* ***Logcosh*** *loss function is 1-Lipschitz continuous, while the Lipschitz continuity of the MSE depends on the specific data and model.*

Proof. The proof is straightforward. The Lipschitz constant for MSE is not fixed but depends on δ:

$$L_{\text{MSE}} = \sup_{\delta \in [\delta_{\min}, \delta_{\max}]} \left| \frac{\partial \text{MSE}(\delta)}{\partial \delta} \right| = 2\delta_{\max}, \tag{31}$$

where δ represents the difference between the reconstruction and the true data, which can be a very large number in the presence of outliers. In contrast, the Lipschitz constant for **logcosh** is

$$L_{\text{logcosh}} = \sup_{\delta \in \mathbb{R}} \left| \frac{\partial \text{logcosh}(\delta)}{\partial \delta} \right| = \sup_{\delta \in \mathbb{R}} |\tanh(\delta)| = 1. \tag{32}$$

This indicates better stability and robustness of **logcosh**. Consequently, **logcosh** is generally better at generalizing to unseen data under the condition that $\delta > 1/2$. □

Overall, as the loss function for the pre-training model, **logcosh** is more generalized than MSE loss when the training is at the beginning, and thus we embed the **logcosh** to MSE loss in the pre-training stage as shown in Equation.3, making the weight more robust for the downstream landmark detection task.

Transfer the Weights for Landmark Detection Task

After the pre-training phase, we use a weight transfer strategy to adapt the network for the downstream landmark detection task. This adaptation involves the reloading of the pre-trained weights across the network. We construct the Gaussian heatmaps y_{gt} for ground truth landmarks during the fine-tuning stage by following Li et al. (2023). The fine-tuned network for landmark detection, denoted by $\mathcal{F}(\mathbf{x}, \theta)$, is trained with a loss function $L_{landmark}$ defined by the MSE loss:

$$L_{landmark} = \text{MSE}(\mathcal{F}(\mathbf{x}, \theta), y_{gt}), \tag{33}$$

where $\mathcal{F}(\mathbf{x}, \theta)$ yields the predicted heatmaps for input $\mathbf{x}$ with the network parameters θ, and y_{gt} is the vector of ground truth heatmaps for the landmarks. This fine-tuning process harnesses the discriminatory features learned during pre-training to increase the out-domain generalization ability for landmark detection tasks.

	Atlas	LFC		FeTA
Split	Training	Validation	Testing	Testing
Number	40	60	120	55
Segmask	40	-	-	-

Table 1: Detail of our DGMLD benchmark.

Experiment

Dataset and Implementation

In this work, we build the Domain Generalized Medical Landmarks Detection (DGMLD) benchmark, which aims to advance the field of medical imaging by focusing on the detection of anatomical landmarks across various datasets, as detailed in Table 1. The benchmark incorporates multiple datasets tailored for specific medical imaging aspects. The Atlas dataset (Gholipour et al. 2017; Wu et al. 2021; Fidon et al. 2022), used for training, includes 40 cases with segmentation masks, facilitating model development under varying imaging conditions and anatomical variations. The LFC dataset, which we construct to fill the gap in 3D MR benchmarks for fetal cerebellum landmark detection, comprises 180 annotated MR images, split into 60 validation and 120 testing cases, without segmentation masks, annotated at a high-precision resolution of $0.6 \times 0.6 \times 0.6$ mm^3 using the NeSVOR technique (Xu et al. 2022a, 2023a). Additionally, the FeTA benchmark (Payette et al. 2021, 2023), used for external out-of-domain testing, includes 55 testing cases without segmentation masks. We adhere strictly to ethical standards by the Declaration of Helsinki (Goodyear, Krleza-Jeric, and Lemmens 2007). The data splits are also shown in Table 1. We use the following biometric annotations (Garel and Garel 2004) related to the fetal cerebellum: Transverse Cerebellar Diameter (TCD), Height Diameter of the Vermis (HDV), and Anteroposterior Diameter of the Vermis (ADV). In total, the dataset includes six landmarks, with two landmarks for each biometric measure. Initially, we invite the expert to remove the cases with unrecognized landmarks. After that, a clinician labels the data, followed by a review and necessary modifications by an additional senior expert to ensure accuracy and consensus.

Our framework was developed using PyTorch version 2.1.2 with Python 3.9.16 and trained on an NVIDIA A100

Data	Type	Method	TCD1	TCD2	HDV1	HDV2	ADV1	ADV2	Average
LFC	UNet	Baseline	$3.43_{\pm 0.39}$	$3.94_{\pm 0.33}$	$2.26_{\pm 0.09}$	$2.44_{\pm 0.07}$	$2.68_{\pm 0.19}$	$2.52_{\pm 0.34}$	$2.88_{\pm 0.14}$
		AugDG	$3.64_{\pm 0.36}$	$4.00_{\pm 0.21}$	$2.45_{\pm 0.33}$	$2.43_{\pm 0.06}$	$2.54_{\pm 0.21}$	$2.15_{\pm 0.38}$	$2.87_{\pm 0.23}$
		EdgeDG	$3.99_{\pm 0.85}$	$4.30_{\pm 0.57}$	$2.22_{\pm 0.23}$	$\mathbf{2.31_{\pm 0.06}}$	$2.89_{\pm 0.43}$	$2.20_{\pm 0.38}$	$2.99_{\pm 0.13}$
		SegDG	$3.32_{\pm 0.24}$	$\mathbf{3.81_{\pm 0.14}}$	$2.74_{\pm 0.41}$	$2.53_{\pm 0.26}$	$2.84_{\pm 0.22}$	$2.08_{\pm 0.13}$	$2.88_{\pm 0.11}$
		BaPDG	$\mathbf{3.09_{\pm 0.02}}$	$3.99_{\pm 0.25}$	$\mathbf{2.19_{\pm 0.12}}$	$2.32_{\pm 0.15}$	$\mathbf{2.50_{\pm 0.38}}$	$\mathbf{1.96_{\pm 0.08}}$	$\mathbf{2.67_{\pm 0.09}}$
	ViTPose	Baseline	$4.52_{\pm 0.60}$	$5.87_{\pm 1.31}$	$3.61_{\pm 0.42}$	$2.79_{\pm 0.15}$	$3.50_{\pm 0.68}$	$2.41_{\pm 0.11}$	$3.78_{\pm 0.46}$
		AugDG	$5.02_{\pm 0.84}$	$5.45_{\pm 0.31}$	$3.40_{\pm 0.30}$	$2.76_{\pm 0.02}$	$3.09_{\pm 0.71}$	$2.63_{\pm 0.19}$	$3.72_{\pm 0.34}$
		EdgeDG	$4.97_{\pm 1.12}$	$4.99_{\pm 0.08}$	$3.46_{\pm 0.13}$	$2.66_{\pm 0.14}$	$3.33_{\pm 0.48}$	$2.29_{\pm 0.20}$	$3.62_{\pm 0.31}$
		SegDG	$4.97_{\pm 0.19}$	$6.01_{\pm 0.97}$	$3.50_{\pm 0.26}$	$2.80_{\pm 0.16}$	$3.34_{\pm 0.82}$	$2.31_{\pm 0.08}$	$3.82_{\pm 0.26}$
		BaPDG	$\mathbf{4.05_{\pm 0.43}}$	$\mathbf{4.59_{\pm 0.33}}$	$\mathbf{2.91_{\pm 0.14}}$	$\mathbf{2.61_{\pm 0.06}}$	$\mathbf{2.88_{\pm 0.09}}$	$\mathbf{2.21_{\pm 0.08}}$	$\mathbf{3.21_{\pm 0.12}}$
FeTA	UNet	Baseline	$5.07_{\pm 2.81}$	$4.48_{\pm 1.22}$	$6.47_{\pm 2.14}$	$5.13_{\pm 1.86}$	$7.28_{\pm 3.46}$	$9.19_{\pm 3.01}$	$6.27_{\pm 1.49}$
		AugDG	$3.40_{\pm 0.44}$	$\mathbf{3.27_{\pm 0.33}}$	$3.58_{\pm 1.00}$	$2.48_{\pm 0.45}$	$3.60_{\pm 0.29}$	$3.32_{\pm 1.46}$	$3.28_{\pm 0.49}$
		EdgeDG	$11.36_{\pm 4.45}$	$7.35_{\pm 1.07}$	$3.79_{\pm 2.16}$	$3.61_{\pm 2.51}$	$6.63_{\pm 3.73}$	$3.75_{\pm 1.37}$	$6.08_{\pm 2.41}$
		SegDG	$4.54_{\pm 1.92}$	$3.48_{\pm 1.11}$	$2.84_{\pm 0.38}$	$2.21_{\pm 0.34}$	$3.53_{\pm 0.40}$	$2.28_{\pm 0.41}$	$3.15_{\pm 0.38}$
		BaPDG	$\mathbf{2.25_{\pm 0.33}}$	$4.60_{\pm 1.97}$	$\mathbf{2.45_{\pm 0.09}}$	$\mathbf{1.78_{\pm 0.12}}$	$\mathbf{3.28_{\pm 0.68}}$	$\mathbf{2.15_{\pm 0.54}}$	$\mathbf{2.75_{\pm 0.48}}$
	ViTPose	Baseline	$15.54_{\pm 10.81}$	$14.61_{\pm 8.41}$	$6.56_{\pm 3.56}$	$2.45_{\pm 0.34}$	$6.31_{\pm 2.22}$	$3.49_{\pm 0.89}$	$8.16_{\pm 4.29}$
		AugDG	$8.29_{\pm 3.20}$	$10.29_{\pm 1.99}$	$6.28_{\pm 0.60}$	$2.53_{\pm 0.28}$	$6.32_{\pm 1.19}$	$3.94_{\pm 0.52}$	$6.28_{\pm 1.00}$
		EdgeDG	$19.38_{\pm 3.65}$	$17.19_{\pm 5.74}$	$6.09_{\pm 0.87}$	$5.68_{\pm 5.76}$	$5.55_{\pm 0.88}$	$3.73_{\pm 0.97}$	$9.60_{\pm 2.15}$
		SegDG	$12.54_{\pm 5.30}$	$16.58_{\pm 0.61}$	$5.67_{\pm 1.32}$	$\mathbf{2.43_{\pm 0.25}}$	$5.25_{\pm 1.13}$	$3.63_{\pm 0.48}$	$7.68_{\pm 0.84}$
		BaPDG	$\mathbf{6.81_{\pm 1.84}}$	$\mathbf{11.39_{\pm 1.65}}$	$\mathbf{4.71_{\pm 0.65}}$	$2.53_{\pm 0.13}$	$\mathbf{3.77_{\pm 0.26}}$	$\mathbf{3.08_{\pm 0.02}}$	$\mathbf{5.38_{\pm 0.75}}$

Table 2: Comparison of state-of-the-art methods on DGMLD testset. The "UNet" model (Ronneberger, Fischer, and Brox 2015) employs a ResNet backbone (He et al. 2016), and "ViTPose" is adapted from (Xu et al. 2022b). "AugDG" (Avisdris et al. 2022), utilizes a specific data augmentation strategy. "EdgeDG" (Gu et al. 2024) incorporates edge features extracted via the Canny operator for pre-training. "SegDG" uses segmentation mask for pre-training. Results are aggregated from three seeds, with values and subscripts indicating means and standard deviations. The best results and our method are highlighted in **bold**.

GPU with 40 GB memory, driver version 525.85.12, and CUDA 12.0. The CPU is an AMD EPYC 7742. We utilized the Adam optimizer for pre-training and fine-tuning phases, with a batch size of 4, a learning rate of 0.001, and 50 training epochs. For evaluation, we adopted the Mean Radial Error (MRE) and Successful Detection Rate (SDR) as the metrics, as outlined by (Li et al. 2023). All the results are obtained by averaging the results of three different seeds.

Comparison with State-of-the-Art Methods

Table 2 illustrates the comparison of our proposed method with state-of-the-art techniques. Across different backbone networks and various out-of-distribution test sets, our method consistently demonstrated the strongest average MRE. In Figure 3, we present the SDR curves for different methods. It is evident from the curves that our proposed method's curve is closest to the upper left corner, further validating the effectiveness of our approach. Simultaneously, our method exhibits more stable results (indicated by the smaller light-shaded areas), further substantiating our DG strategy's efficacy. Unlike direct supervision with segmentation masks, our proposed method based on distance maps better focuses the model on the critical boundaries where landmarks are located, thereby achieving superior performance. We present our visualization results in Figure 4.

We conducted additional experiments on 2D cephalometric landmark detection using X-ray images in Table 3. We used the ISBI-15 (Wang et al. 2016) training data (150 cases) as the source domain, testing data 2 (100 cases) for validation, testing data 1 (150 cases) as the in-domain test set, and the PKU (Zeng et al. 2021) (102 cases) as the out-of-domain test set. Since segmentation masks were not available in these datasets, we applied the MuGE (Zhou et al. 2024) to extract boundaries and generate distance maps.

	ISBI-15 (In domain)			PKU (Out of domain)		
	MRE	SDR 3mm	SDR 4mm	MRE	SDR 3mm	SDR 4mm
Baseline	2.43	82.52	88.59	3.02	74.09	81.26
EdgeDG	1.98	84.77	90.94	2.73	72.91	83.07
Ours	1.95	85.19	91.26	2.66	74.25	83.33

Table 3: Results on 2D cephalometric landmark detection.

LFC Validation	SegDG	MSE	MSE+logcosh
MRE	$2.14_{\pm 0.05}$	$2.05_{\pm 0.07}$	$1.97_{\pm 0.14}$

Table 4: Ablation study. "SegDG" directly uses the segmentation mask for pre-training, "MSE" first extracts the boundary distance map, then uses MSE loss to regress. "MSE+logcosh" uses the method in Eq.3 for pre-training.

Ablation Study

Table 4 compares MRE across different models on the validation set, formatted to emphasize precision and variability. Methods based on edge heatmap regression yield superior

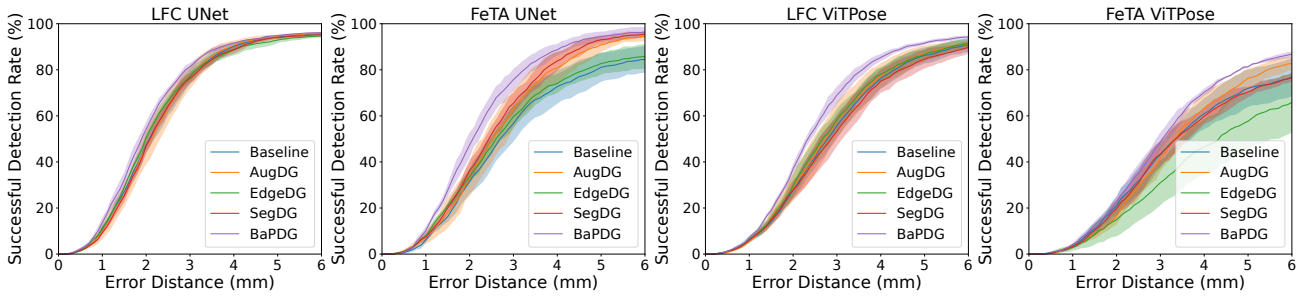

Figure 3: Visualization of the SDR curves. A distinct colored line and the surrounded shaded areas depicts each method and their standard deviation, respectively. Curves that approach the upper left corner demonstrate superior performance.

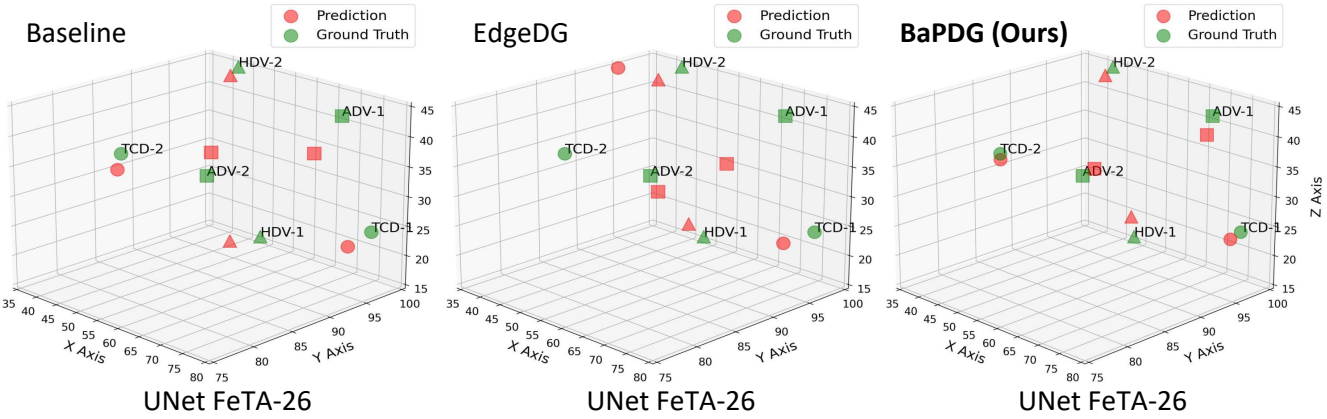

Figure 4: Qualitative analysis of our method and other state-of-the-art methods. Closer proximity between red and green symbols indicates superior accuracy. Our proposed BaPDG method achieves the best prediction accuracy on average.

results. Building on this, the further utilization of the mixed loss function derived in this paper significantly enhances the robustness of the model during the pre-training phase. This mixed loss function integrates multiple error metrics to better cater to the intricate variations in medical imaging, thereby improving the model's ability to generalize from training to real-world scenarios. In addition, this approach mitigates overfitting, contributing to more consistent performance across diverse datasets and conditions.

Conclusion

This work introduces a novel framework explicitly designed for DG medical landmark detection that leverages training solely on single-center data. Our approach significantly improves domain generalization capabilities for landmark detection tasks by using publicly available medical segmentation datasets. This enhancement is primarily achieved through a boundary-aware pre-training strategy that meticulously directs the model's focus to critical regions essential for accurate landmark detection, coupled with a mixed loss function that substantially improves model robustness and generalization across diverse datasets. Extensive experiments conducted on our newly developed DG medical landmark detection benchmark demonstrate the unequivocal superiority of our method. The results clearly illustrate its potential to facilitate more generalized and robust solutions in the field of medical landmark detection, thereby addressing substantial gaps in existing research methodologies.

Acknowledgements

This work is supported in part by the Guangdong Basic and Applied Basic Research Foundation (2023A1515011464), in part by the Shenzhen Science and Technology Program (JCYJ20220818103001002), in part by Project No. (20232ABC03A25), in part by the Longgang District Special Funds for Science and Technology Innovation (LGKCSDPT2023002), in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.

References

Akbari, A.; Awais, M.; Bashar, M.; and Kittler, J. 2021. How does loss function affect generalization performance of deep learning? Application to human age estimation. In *ICML*, 141–151. PMLR.

Avisdris, N.; et al. 2021. Automatic linear measurements of the fetal brain on MRI with deep neural networks. *IJCARS*, 16(9): 1481–1492.

Avisdris, N.; et al. 2022. BiometryNet: Landmark-based Fetal Biometry Estimation from Standard Ultrasound Planes. In *MICCAI*, 279–289. Springer.

Bartlett, P. L.; and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov): 463–482.

Chan, Y. Y.; Li, X.-P.; Mai, J.; Leung, C.-S.; and So, H. C. 2024. Sparse Unmixing in the Presence of Mixed Noise Using l0-norm Constraint and Log-Cosh Loss. *TGRS*.

Chen, C.; Yang, X.; Huang, Y.; Shi, W.; Cao, Y.; Luo, M.; Hu, X.; Zhu, L.; Yu, L.; Yue, K.; et al. 2024. FetusMapV2: Enhanced fetal pose estimation in 3D ultrasound. *MIA*, 91: 103013.

Feng, Y.; Yang, J.; Li, M.; Tang, L.; Sun, S.; and Wang, Y. 2024. A Bayesian network for simultaneous keyframe and landmark detection in ultrasonic cine. *MIA*, 103228.

Fidon, L.; Viola, E.; Mufti, N.; David, A. L.; Melbourne, A.; Demaerel, P.; Ourselin, S.; Vercauteren, T.; Deprest, J.; and Aertsen, M. 2022. A spatio-temporal atlas of the developing fetal brain with spina bifida aperta. *Open Research Europe*, 1: 123.

Garel, C.; and Garel, C. 2004. *MRI of the Fetal Brain*. Springer.

Gholipour, A.; Rollins, C. K.; Velasco-Annis, C.; Ouaalam, A.; Akhondi-Asl, A.; Afacan, O.; Ortinau, C. M.; Clancy, S.; Limperopoulos, C.; Yang, E.; et al. 2017. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific reports*, 7(1): 476.

Gong, H.; Chen, G.; Liu, S.; Yu, Y.; and Li, G. 2021. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *ICMR*, 456–460.

Gong, H.; Chen, G.; Mao, M.; Li, Z.; and Li, G. 2022a. Vqamix: Conditional triplet mixup for medical visual question answering. *TMI*, 41(11): 3332–3343.

Gong, H.; Chen, J.; Chen, G.; Li, H.; Li, G.; and Chen, F. 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *CBM*, 155: 106389.

Gong, H.; Cheng, H.; Xie, Y.; Tan, S.; Chen, G.; Chen, F.; and Li, G. 2022b. Less is More: Adaptive Curriculum Learning for Thyroid Nodule Diagnosis. In *MICCAI*, 248–257. Springer.

Gong, H.; Kang, L.; Wang, Y.; Wang, Y.; Wan, X.; Wu, X.; and Li, H. 2025. nnmamba: 3D biomedical image segmentation, classification and landmark detection with state space model. In *ISBI*.

Goodyear, M. D.; Krleza-Jeric, K.; and Lemmens, T. 2007. The declaration of Helsinki.

Gu, S.; et al. 2024. Train Once, Deploy Anywhere: Edge-Guided Single-source Domain Generalization for Medical Image Segmentation. In *MIDL*.

Hager, W. W. 1979. Lipschitz continuity for constrained processes. *SIAM Journal on Control and Optimization*, 17(3): 321–338.

Han, X.; Yu, J.; Yang, X.; Chen, C.; Zhou, H.; Qiu, C.; Cao, Y.; Zhang, T.; Peng, M.; Zhu, G.; et al. 2024. Artificial intelligence assistance for fetal development: evaluation of an automated software for biometry measurements in the mid-trimester. *BMC Pregnancy and Childbirth*, 24(1): 1–11.

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

Hughes, D. E.; et al. 2023. Genetic patterning for child psychopathology is distinct from that for adults and implicates fetal cerebellar development. *Nature Neuroscience*, 1–11.

Jeendgar, A.; Devale, T.; Dhavala, S. S.; and Saha, S. 2022. LogGENE: A smooth alternative to check loss for Deep Healthcare Inference Tasks. *arXiv preprint arXiv:2206.09333*.

Jimenez-del Toro, O.; Müller, H.; Krenn, M.; et al. 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE TMI*, 35(11): 2459–2475.

Jin, H.; Che, H.; and Chen, H. 2023. Unsupervised domain adaptation for anatomical landmark detection. In *MICCAI*, 695–705. Springer.

Juneja, M.; Garg, P.; Kaur, R.; Manocha, P.; Batra, S.; Singh, P.; Singh, S.; Jindal, P.; et al. 2021. A review on cephalometric landmark detection techniques. *BSPC*, 66: 102486.

Kang, L.; Gong, H.; Wan, X.; and Li, H. 2023. Visual-attribute prompt learning for progressive mild cognitive impairment prediction. In *MICCAI*, 547–557. Springer.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32.

Li, H.; Wang, Y.; Wan, R.; Wang, S.; Li, T.-Q.; and Kot, A. 2020. Domain generalization for medical imaging classification with linear-dependency regularization. In *NeurIPS*, volume 33, 3118–3129.

Li, H.; Yang, J.; Xuan, Z.; Qu, M.; Wang, Y.; and Feng, C. 2024a. A spatio-temporal graph convolutional network for ultrasound echocardiographic landmark detection. *MIA*, 103272.

Li, X.; Lv, S.; Zhang, J.; Li, M.; Rodriguez-Andina, J. J.; Qin, Y.; Yin, S.; and Luo, H. 2024b. FDGR-Net: Feature Decouple and Gated Recalibration Network for medical image landmark detection. *Expert Systems with Applications*, 238: 121746.

Li, X.; et al. 2023. SDMT: Spatial Dependence Multi-Task Transformer Network for 3D Knee MRI Segmentation and Landmark Localization. *IEEE TMI*.

Liu, W.; Wang, Y.; Jiang, T.; Chi, Y.; Zhang, L.; and Hua, X.-S. 2020. Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In *MICCAI*, 670–679. Springer.

Long, S.; Zhou, Q.; Ying, C.; Ma, L.; and Luo, Y. 2024. Rethinking domain generalization: Discriminability and generalizability. *IEEE TCSVT*.

Matsuura, T.; and Harada, T. 2020. Domain generalization using a mixture of multiple latent domains. In *AAAI*, volume 34, 11749–11756.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Payette, K.; de Dumast, P.; Kebiri, H.; Ezhov, I.; Paetzold, J. C.; Shit, S.; Iqbal, A.; Khan, R.; Kottke, R.; Grehten, P.; et al. 2021. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific data*, 8(1): 167.

Payette, K.; Li, H. B.; de Dumast, P.; Licandro, R.; Ji, H.; Siddiquee, M. M. R.; Xu, D.; Myronenko, A.; Liu, H.; Pei, Y.; et al. 2023. Fetal brain tissue annotation and segmentation challenge results. *MIA*, 88: 102833.

Peng, Q.; Zheng, C.; and Chen, C. 2024. A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation. In *CVPR*, 2240–2249.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *CVPR*, 12556–12565.

Qu, S.; Pan, Y.; Chen, G.; Yao, T.; Jiang, C.; and Mei, T. 2023. Modality-agnostic debiasing for single domain generalization. In *CVPR*, 24142–24151.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Saleh, R. A.; and Saleh, A. 2022. Statistical properties of the log-cosh loss function used in machine learning. *arXiv preprint arXiv:2208.04564*.

Sammut, C.; and Webb, G. I. 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

Shankar, H.; et al. 2022. Leveraging Clinically Relevant Biometric Constraints to Supervise a Deep Learning Model for the Accurate Caliper Placement to Obtain Sonographic Measurements of the Fetal Brain. In *ISBI*, 1–5. IEEE.

Vahedifard, F.; Adepoju, J. O.; Supanich, M.; Ai, H. A.; Liu, X.; Kocak, M.; Marathu, K. K.; and Byrd, S. E. 2023. Review of deep learning and artificial intelligence models in fetal brain magnetic resonance imaging. *World Journal of Clinical Cases*, 11(16): 3725–3735.

VAPNIK, V. 1998. Statistical Learning Theory. *Adaptive and Learning Systems for Signal Processing Communications and control*.

Viswanathan, V. S.; Parmar, V.; and Madabhushi, A. 2024. Towards equitable AI in oncology. *Nature Reviews Clinical Oncology*, 1–10.

Wang, C.-W.; Huang, C.-T.; Lee, J.-H.; Li, C.-H.; Chang, S.-W.; Siao, M.-J.; Lai, T.-M.; Ibragimov, B.; Vrtovec, T.; Ronneberger, O.; et al. 2016. A benchmark for comparison of dental radiography analysis algorithms. *MIA*, 31: 63–76.

Wu, J.; Sun, T.; Yu, B.; Li, Z.; Wu, Q.; Wang, Y.; Qian, Z.; Zhang, Y.; Jiang, L.; and Wei, H. 2021. Age-specific structural fetal brain atlases construction and cortical development quantification for Chinese population. *Neuroimage*, 241: 118412.

Xu, J.; Moyer, D.; Grant, P. E.; Golland, P.; Iglesias, J. E.; and Adalsteinsson, E. 2022a. SVoRT: Iterative transformer for slice-to-volume registration in fetal brain MRI. In *MICCAI*, 3–13. Springer.

Xu, J.; Zhang, M.; Turk, E. A.; Grant, P. E.; Golland, P.; and Adalsteinsson, E. 2020a. 3D fetal pose estimation with adaptive variance and conditional generative adversarial network. In *PIPPI, MICCAI Workshop*, 201–210. Springer.

Xu, J.; et al. 2019. Fetal pose estimation in volumetric MRI using a 3D convolution neural network. In *MICCAI*, 403–410. Springer.

Xu, J.; et al. 2023a. NeSVoR: Implicit Neural Representation for Slice-to-Volume Reconstruction in MRI. *IEEE TMI*.

Xu, L.; Gong, H.; Zhong, Y.; Wang, F.; Wang, S.; Lu, L.; Ding, J.; Zhao, C.; Tang, W.; and Xu, J. 2023b. Real-time monitoring of manual acupuncture stimulation parameters based on domain adaptive 3D hand pose estimation. *BSPC*, 83: 104681.

Xu, X.; Li, J.; Yang, Y.; and Shen, F. 2020b. Toward effective intrusion detection using log-cosh conditional variational autoencoder. *IEEE IoTJ*, 8(8): 6187–6196.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022b. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 35: 38571–38584.

Xu, Z.; Gong, H.; Wan, X.; and Li, H. 2023c. ASC: Appearance and Structure Consistency for Unsupervised Domain Adaptation in Fetal Brain MRI Segmentation. In *MICCAI*. Springer.

Yin, D.; Kannan, R.; and Bartlett, P. 2019. Rademacher complexity for adversarially robust generalization. In *ICML*, 7085–7094. PMLR.

Zeng, M.; Yan, Z.; Liu, S.; Zhou, Y.; and Qiu, L. 2021. Cascaded convolutional networks for automatic cephalometric landmark detection. *MIA*, 68: 101904.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NeurIPS*, 14435–14447.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*, 535–552. Springer.

Zhou, C.; Huang, Y.; Pu, M.; Guan, Q.; Deng, R.; and Ling, H. 2024. MuGE: Multiple Granularity Edge Detection. In *CVPR*, 25952–25962.

Zhou, G.-Q.; Miao, J.; Yang, X.; Li, R.; Huo, E.-Z.; Shi, W.; Huang, Y.; Qian, J.; Chen, C.; and Ni, D. 2021. Learn fine-grained adaptive loss for multiple anatomical landmark detection in medical images. *IEEE JBHI*, 25(10): 3854–3864.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE TPAMI*, 45(4): 4396–4415.

Zhou, S.; Ahn, E.; Wang, H.; Quinton, A.; Kennedy, N.; Sridar, P.; Nanan, R.; and Kim, J. 2023. Improving Automatic Fetal Biometry Measurement with Swoosh Activation Function. In *MICCAI*, 283–292. Springer.