Invariant Shape Representation Learning For Image Classification

Tonmoy Hossain[1*], Jing Ma[1†], Jundong Li[1,2*], Miaomiao Zhang[1,2*]
Computer Science[1], Electrical and Computer Engineering[2]
University of Virginia[*], Case Western Reserve University[†]
{tonmoy,jl6qk,mz8rr}@virginia.edu[*], jing.ma5@case.edu[†]

Abstract

Geometric shape features have been widely used as strong predictors for image classification. Nevertheless, most existing classifiers such as deep neural networks (DNNs) directly leverage the statistical correlations between these shape features and target variables. However, these correlations can often be spurious and unstable across different environments (e.g., in different age groups, certain types of brain changes have unstable relations with neurodegenerative disease); hence leading to biased or inaccurate predictions. In this paper, we introduce a novel framework that for the first time develops invariant shape representation learning (ISRL) to further strengthen the robustness of image classifiers. In contrast to existing approaches that mainly derive features in the image space, our model ISRL is designed to jointly capture invariant features in latent shape spaces parameterized by deformable transformations. To achieve this goal, we develop a new learning paradigm based on invariant risk minimization (IRM) to learn invariant representations of image and shape features across multiple training distributions/environments. By embedding the features that are invariant with regard to target variables in different environments, our model consistently offers more accurate predictions. We validate our method by performing classification tasks on both simulated 2D images, real 3D brain and cine cardiovascular magnetic resonance images (MRIs). Our code is publicly available at https://github.com/tonmoy-hossain/ISRL.

1. Introduction

Deformable shape has a long history of aiding image analysis, as it plays an important role in the processing of human visual information [20, 24, 42, 67]. Recent works on shape-based deep networks have demonstrated the robustness of shape to variations in image intensity and texture (e.g., noisy or corrupted data) for image analysis tasks, such as classification [16, 21, 55]. Despite achieving improved performance in classification accuracy, the aforementioned models were designed to leverage any statistical correlations between the geometric shape features and target image labels. However, these features can contain both "invariant" features that have stable relationships with the labels under any circumstances, and "spurious" features that have varying relationships with labels across different contexts or environments (e.g., a confounding factor can often lead to spurious correlations).

Existing shape-based image classification networks are often incapable of distinguishing the invariant shape features from the spurious ones [16, 55]; hence are easily fooled by illusory patterns and perform unstably in different environments. For example, anatomical brain changes derived from images have been used as predictive features to distinguish healthy subjects from neurodegenerative diseases [55, 58]. However, in different environments such as age groups, if not properly handled (e.g., when the average age of the disease group is significantly older than the healthy controls), a predictor (classifying healthy/disease) can be potentially confounded [64, 68], rather than capturing the invariant features caused by Diseases. Teasing apart spurious features related to age factor is critical for DNNs to provide robust predictions of brain diseases.

A few research groups have initial attempts to capture more reliable information instead of spurious correlations to facilitate the robustness of image classification under different scenarios, mainly including confounding biases [63, 68] and weak/noisy supervision [8, 59, 66]. Invariant learning is one of the subsets of this arena that recently attracted significant attention [1, 2, 10]. The goal of invariant learning is to identify and capture the underlying factors that have a stable relationship with the label across different environments, while disregarding factors with unstable spurious correlations to the label. More intuitively, it aims to capture the real "cause" of the label to enable more robust and accurate predictions, improving outcomes even for unseen environments, which results in achieving promising performance of extracting invariant features in the image space [1, 2, 33, 35, 36, 56]. Another line of work investigated domain generalization approaches to deal with

possible changes across known input distributions by exploiting unsupervised adaptation [46, 47], domain calibration [17, 54], data augmentation [22, 29, 61], or adversarial learning [15, 60]. However, none of the existing studies have investigated invariant learning methods in an integrated shape space. This limits their ability to fully utilize geometric features that have been proven to be important and robust in image classification tasks.

In this paper, we introduce a novel method that for the first time develops a joint learning of *invariant shape representations* for improved performance of image classifiers. In contrast to previous approaches that are limited to learning invariant features solely in the image spaces [49, 59, 68], our algorithm has the main advantages of

(i) Developing a new learning paradigm based on invariant risk minimization (IRM) [1] to learn integrated image and shape representations that are invariant across multiple training distributions/environments.

(ii) Improving the efficiency and adaptability of image classifiers when tested with unseen data/environments, by leveraging learned invariant features with maximally eliminated spurious correlations.

(iii) Opening new avenues for causal representation learning by leveraging invariant image and shape features associated with target labels.

We validate the effectiveness of ISRL on 2D simulated data [25], 3D real brain MRIs [23], and 3D cardiac MRI videos [57]. Experimental results show that our method outperforms the state-of-the-arts by significantly improved robustness to shifted environments with consistently higher classification accuracy.

2. Background: Deformation-based Shape Representations

The literature has studied various representations of geometric shapes, including landmarks [13], binary segmentations [7], and medial axes [38]. These aforementioned techniques often ignore objects' interior structures; hence do not capture the intricacies of complex objects in images. In contrast, deformation-based shape representations (based on elastic deformations or fluid flows) focus on highly detailed shape information from images [6, 32, 50]. This paper will feature deformation-based shape representations that offer more flexibility in describing shape changes and variability of complex structures. However, our developed methodology can be easily adapted to other representations, including those characterized by landmarks, binary segmentations, curves, and surfaces. With the underlying assumption that objects in many generic classes can be described as deformed versions of an ideal template, descriptors in this class arise naturally by transforming/deforming the template to an input image [32]. The resulting transformation is then considered as a shape that reflects geometric changes.

Given a number of N images, $\{I_1, \cdots, I_N\}$, the problem of template-based image registration is to estimate the deformation fields, $\{\phi_1, \cdots, \phi_N\}$, between a template image I and each individual image via minimizing the energy

$$E(I, \phi_n(t)) = \sum_{n=1}^{N} \frac{1}{\sigma^2} \text{Dist}[I \circ \phi_n^{-1}(v_n(t)), I_n] + \int_0^1 (Lv_n(t), v_n(t))\, dt, \quad (1)$$

subject to $d\phi_n(t)/dt = v_n(t) \circ \phi_n(t)$. Here, σ^2 is a noise variance and $\circ$ denotes an interpolation operator that deforms image I with an estimated transformation ϕ_n. The $\text{Dist}[\cdot, \cdot]$ is a distance function that quantifies the dissimilarity between images, i.e., sum-of-squared differences [5]. Here, $L : V \to V^*$ is a symmetric, positive-definite differential operator that maps a tangent vector $v_n(t) \in V$ into its dual space as a momentum vector $m_n(t) \in V^*$. We write $m_n(t) = Lv_n(t)$, or $v_n(t) = Km_n(t)$, with K being an inverse operator of L. The notation $(\cdot, \cdot)$ denotes the pairing of a momentum vector with a tangent vector, which is similar to an inner product. In this paper, we use a metric of the form $L = -\alpha\Delta + \mathbb{I}$, in which Δ is the discrete Laplacian operator, α is a positive regularity parameter, and $\mathbb{I}$ denotes an identity matrix.

The *geodesic shooting* algorithm [52] states that the minimum of Eq. (1) is uniquely determined by solving an Euler-Poincaré differential equation (EPDiff) [4, 34] with a given initial condition of velocity field $v_n(0)$,

$$\frac{\partial v_n(t)}{\partial t} = -K[(Dv_n(t))^T \cdot m_n(t) + Dm_n(t) \cdot v_n(t) + m_n(t) \cdot \text{div}\, v_n(t)], \quad (2)$$

where D denotes a Jacobian matrix, div is the divergence, and $\cdot$ represents element-wise matrix multiplication.

We are now able to equivalently minimize the energy function in Eq. (1) as

$$E(I, v_n(0)) = \sum_{n=1}^{N} \frac{1}{\sigma^2} \text{Dist}[I \circ \phi_n^{-1}(v_n(t)), I_n] + (Lv_n(0), v_n(0)), \text{ s.t. Eq. (2).} \quad (3)$$

For notation simplicity, we will drop the time index in the following sections.

3. Our Model

This section presents a novel algorithm, ISRL, that for the first time develops a joint learning of invariant shape representations for improved image classifiers.

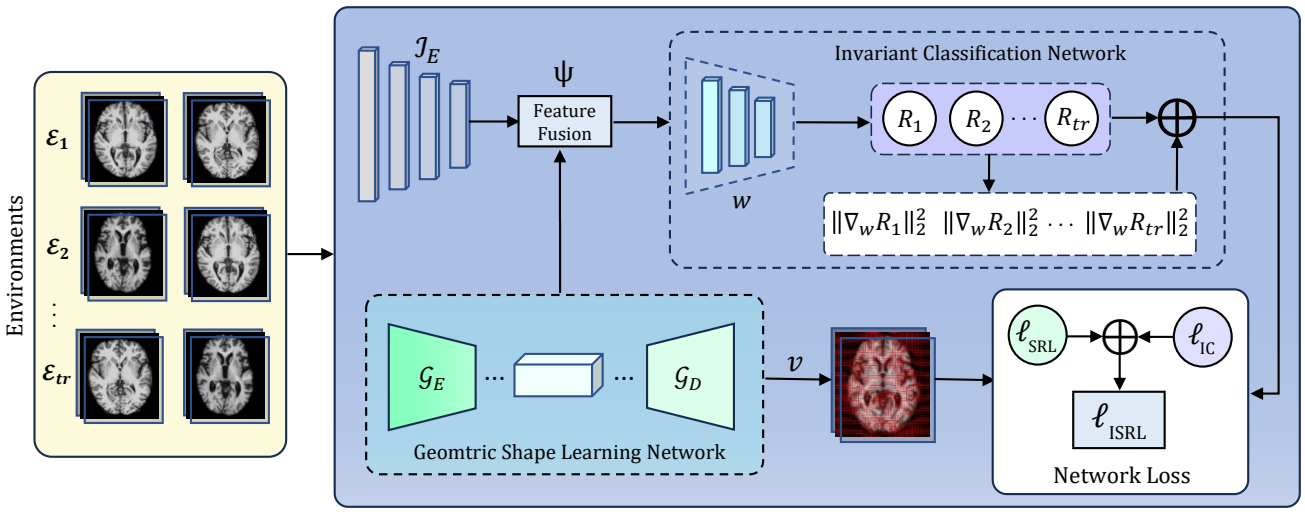

Figure 1. An overview of our proposed network architecture of ISRL. The geometric shape learning ($\mathcal{G}_E, \mathcal{G}_D$) and image network ($\mathcal{I}_E$) is taking images from different environments $\{\mathcal{E}_1, \cdots, \mathcal{E}_{tr}\}$. ISRL combined features from latent spaces passing it to the classifier w. To learn invariant features, we combine geometric shape learning loss with environment-wise risk (R_{tr}) along with their gradient ($||\nabla_w R_{tr}||_2^2$).

Problem setup. Given a number of J image classes, there exists a number of $N_j, j \in \{1, ..., J\}$ images in each class. With a group of training images $\{I_{nj}\}_{n=1,j=1}^{N_j,J}$ and their associated class labels $\{y_{nj}\}$ collected from multiple training environments $\{e\} \in \mathcal{E}_{tr}$, we define the training data as $X = \{(I_{nj}^e, y_{nj}^e)\}_{e=1}^{\mathcal{E}_{tr}}$. Note that each image I_{nj} is considered as a deformable variant of a template T, where $T \in N_j$. The proposed framework of ISRL consists of two modules: (i) an unsupervised learning of geometric deformations via a template-based registration network [55], and (ii) an invariant classification network that takes fused features from latent spaces of image intensities and geometric shapes. An overview of our network architecture is shown in Fig. 1.

3.1. Geometric Shape Representation Learning Network

Let $\Theta = (\mathcal{G}_E, \mathcal{G}_D)$ be the parameters of an encoder-decoder in our geometric shape learning network. The shape representations of images within each class (a.k.a., initial velocity fields $v_{nj}(\Theta)$) will be learned by minimizing the network loss function as

$$\ell_{\text{SRL}}(\Theta) = \sum_{n=1}^{N_j}\sum_{j=1}^{J}[\frac{1}{\sigma^2}\|T \circ \phi_{nj}^{-1}(v_{nj}(\Theta)) - I_{nj}\|_2^2 + \text{reg}(\Theta) + (\mathcal{L}_j v_{nj}(\Theta), v_{nj}(\Theta))], \text{ s.t. Eq. (2)}, \quad (4)$$

where $\text{reg}(\cdot)$ denotes a regularization term on network parameters. Note that this shape representation learning network takes into account images within each class across all environments, i.e., $I_{nj} \triangleq \{I_{nj}^e\}_{e=1}^{\mathcal{E}_{tr}}$. For notation simplicity, we omit the environment variable e.

The template T can be either treated as network parameters to estimate [19, 55], or pre-selected references [26, 62]. In this experiment, we use a pre-selected image as a reference template. As for the shape learning network backbone, we adopt a commonly used UNet architecture [44] in this paper. However, other networks such as UNet++ [69] and TransUNet [12] can be easily applied.

3.2. Invariant Classification Network

Our newly designed classification network is trained to learn the unobserved *invariant features from an integrated image and deformation shape spaces*, aiming to achieve an optimal performance. Inspired by recent works [1, 2], we develop a mechanism of invariant representation learning that captures both geometric shape and image features robust to data distribution shifts across multiple environments. Such features may serve as indicators or proxies for the underlying invariant relationships to determine the labels of image classes. Meanwhile, the biases introduced by spurious factors, will be effectively mitigated.

Let $\mathcal{I}_E$ be the parameters of an encoder that extracts features from image spaces. We first employ a feature fusion module that concatenates the representation of geometric shape, $\mathbf{v}$, and image, $\mathbf{I}$, in a latent space $\mathcal{H}$ [55]. Note that $\mathcal{I}_E$ can be a wide variety of feature extractors, including but not limited to ResNet [18], VGGNet [45], ViT [14], or other state-of-the-art network architectures for image classification tasks. An invariant classifier that maps the latent features in $\mathcal{H}$ to predicted labels, $\mathbf{y}$, is defined as follows.

Definition 3.1. Consider an integrated shape and image representation $\psi(\mathcal{G}_E, \mathcal{I}_E) : \mathbf{v} \times \mathbf{I} \to \mathcal{H}$. A classifier $w(\psi)$ is invariant across environments $\{e\}$, when there exists a

model $w : \mathcal{H} \rightarrow \mathbf{y}$ simultaneously optimal for all environments. That is to say, we have $w \in argmin \sum_{e\in\mathcal{E}} R_e(\psi)$, where $R_e(\psi) := \mathbb{E}_{\mathbf{I}_e,\mathbf{v}_e,\mathbf{y}_e}[l(\psi(\mathbf{I}_e, \mathbf{v}_e), \mathbf{y}_e)]$ is the empirical risk under environment e.

Following a practical implementation of IRM [2], we are now ready to define a loss function of our invariant classification network as a combination of the sum of empirical risk minimization with invariant constraints of the predictor w. With $\mathcal{E}_{tr}$ denoting a set of all training environments, we formulate the loss of our invariant classification network as

$$\ell_{\text{IC}}(\psi) = \min_{\psi} \sum_{e\in\mathcal{E}_{tr}} R_e(\psi) + \lambda \left\| \bigtriangledown_{w|w=1.0} R_e(w \cdot \psi) \right\|_2^2, \tag{5}$$

where $w = 1.0$ is now a scalar (or a fixed dummy classifier) and ψ becomes the entire invariant predictor. The parameter λ is a penalty weight to control the invariance of the predictor ψ across different training environments. Minimizing the loss above encourages the learned feature representations to discard spurious features but embed invariant factors for a more robust classification performance.

In this paper, we employ a cross-entropy loss as the classification loss R_e in each environment, i.e.,

$$R_e(\psi(\mathcal{G}_E, \mathcal{I}_E)) = \tau \sum_{n=1}^{N_j} \sum_{j=1}^{J} -y_{nj} \cdot \log \hat{y}_{nj}(\psi(\mathcal{G}_E, \mathcal{I}_E)) + \text{reg}(\psi(\mathcal{G}_E, \mathcal{I}_E)), \tag{6}$$

where τ is a weighting parameter and $\hat{y}_{nj}$ denotes a predicted label of the n-th subject in class j.

3.3. Joint Learning Network Loss and Optimization.

The loss function of our joint learning framework, ISRL, includes the loss from both geometric learning (Eq. (4)) and invariant classification networks (Eq. (5)). Defining β as the weighting parameter, we are now ready to write the joint network loss as $\ell_{\text{ISRL}} = \ell_{\text{SRL}}(\Theta(\mathcal{G}_E, \mathcal{G}_D)) + \beta\ell_{\text{IC}}(\psi(\mathcal{G}_E, \mathcal{I}_E))$. We employ an alternative optimization scheme [37] to minimize the total loss. More specifically, we jointly optimize all network parameters by alternating between the training of the geometric shape learning and invariant classification network, making it an end-to-end learning [9, 11].

The training inference of ISRL is summarized in Alg. 1.

Algorithm 1: Joint optimization of ISRL model.

Input : A group of input images $\mathbf{I}$ with their associated class labels $\mathbf{y}$, number of environments tr, and convergence threshold ϵ.

Output: Predicted initial velocity fields $\mathbf{v}$ and class labels $\hat{\mathbf{y}}$.

repeat
  for $e \in \forall\mathcal{E}_{tr}$ **do**
    `/* Train our geometric shape learning network */`
    Optimize the geometric shape learning loss in Eq. (4);
    Output the initial velocity fields $\mathbf{v}$ and within group template image $\{T_j\}$ (if treated as network parameters rather than pre-selected images) for all classes.
    `/* Train our invariant classifier */`
    Output the predicted class labels $\hat{\mathbf{y}}$.
    Optimize the invariant classifier loss in Eq. (5) with integrated features from shape and image spaces;
  end
until convergence, $|\Delta\ell_{\text{ISRL}}| < \epsilon$;

4. Experimental Evaluation

We demonstrate the effectiveness of our proposed model on a diverse set of deformable image datasets, including 2D simulated data, 3D real brain MRIs, and 3D video sequences of cardiac MRIs. Examples of all datasets in multiple different training environments (color vs. age vs. patient history of congestive heart failure) are shown in Fig. 2.

2D simulated data. We first randomly choose 3000 2D images with three different classes of circles, squares, and triangles (1000 images per class) from the Google Quickdraw dataset [25]. All images underwent affine transformation and intensity normalization with the size of 224×224.

Analogous to [2], we first assign a random color (red, green, or blue) to each image, introducing a spurious correlation between the color information and class labels. Such a correlation is artificially established to make color more predictive of the label than the actual drawings. As a result, current algorithms focused solely on minimizing training errors, such as ERM (an image classifier extracting features from image space by minimizing the empirical classification training loss) [51], will tend to exploit the color. In this scenario, we expect our proposed algorithm, ISRL, is able to (i) observe the strength of the correlation between color and label varies across multiple training environments; and (ii) leverage the geometric shape information derived from images to eliminate color as a predictive feature; leading to enhanced generalization performance.

Following [2], we then synthesize a colored Google Quickdraw dataset with three environments, including two training and one testing by performing the following steps: i) assign a preliminary label $\tilde{y}$: $\tilde{y} = 0$ for circle, $\tilde{y} = 1$ for square, and $\tilde{y} = 2$ for triangle; ii) obtain the final label y by

flipping $\tilde{y}$ with probability p^l (i.e., 0.25, 0.4, and 0.5 in our experimental settings); iii) sample the color id z by flipping y with different probabilities p^e in each environment. In this paper, we set $p^e = 0.2, 0.1, 0.9$ for the two training and one testing environment, respectively. We then color the image red if $z = 0$, green if $z = 1$, and blue if $z = 2$. We split such a synthesized dataset into 70% for training, 15% for validation, and 15% for testing.

3D brain MRI. For Alzheimer's Disease (AD) classification, we include 690 public T1-weighted brain MRIs from the AD Neuroimaging Initiative (ADNI) [23]. All subjects ranged in age from 50 to 100, with 345 images from patients affected by AD and the others from cognitively normal (CN). All MRIs were preprocessed to be the size of $104 \times 128 \times 120$, $1mm^3$ isotropic voxels, and underwent skull-stripping, intensity min-max normalization, bias-field correction, and affine registration [43].

This experiment focuses on treating age as an environment, which is motivated by the well-established understanding that the brain changes caused by different ages are often confused with the brain changes related to AD [40, 41, 68]. That is, at different ages, certain types of brain changes have varying correlations with AD. In this setting, we expect the model to learn invariant features of brain shape changes that are actual diagnostic criteria for AD. The spurious features are brain shape changes affected by age. In this context, we set the training environments, $\mathcal{E}_{tr}$, of our model into three categories representing individuals in their 50s to 60s, 70s, and 80s to 90s. Each age group or environment is characterized by a uniform distribution and includes a total of 150 images. We split the data into 65%, 25%, and 10% as training, testing, and validation.

3D cardiac MRI videos. We include 510 cine cardiac MRI video sequences with manually delineated left ventricular myocardium segmentation maps collected from 125 subjects [57]. Approximately 40% of these subjects exhibit heart motion abnormalities caused by myocardial scars. All videos were preprocessed to be the size of $224 \times 224 \times 24$, $1mm^3$ isotropic voxels, where 24 represents the number of time frames that cover a full cardiac cycle.

In this experiment, we perform patient scar classification (scar vs. non-scar) from the 3D video sequence of myocardium contour segmentation. We treat congestive heart failure (CHF) as an environment, given evidence indicating that individuals with CHF have an increased risk of heart motion abnormalities [27, 39]. We expect to learn invariant shape features of myocardial motion changes that are actual indicators for the scar, while the spurious features are motions led by CHF. In this context, the two training environments $(\mathcal{E}_1, \mathcal{E}_2)$, are subjects labeled as having CHF or not. Out of the 510 videos (approximately equal number of scar/non-scar), we found 146 videos have CHF ($\mathcal{E}_1$, 48/98 as scar/non-scar) and 364 not ($\mathcal{E}_2$, 210/154 as scar/non-scar). We split the data into 70%, 15%, and 15% for training, validation, and testing, respectively.

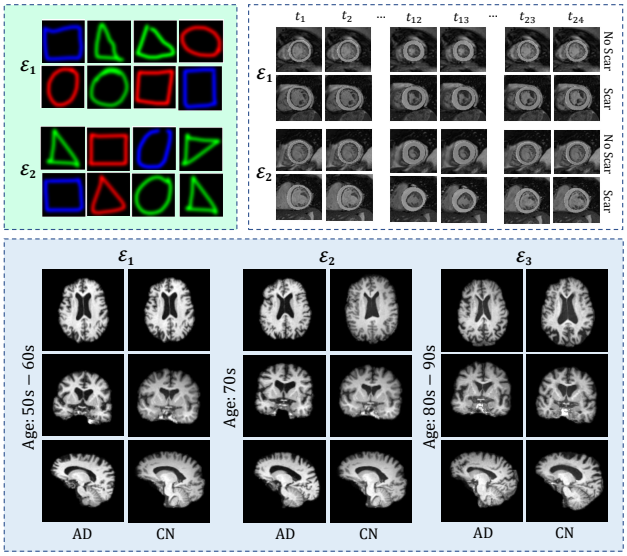

Figure 2. Top to bottom: examples of 2D simulated data vs. 3D brain MRIs vs. 3D video sequence of cardiac MRIs. The training environments of each dataset (color vs. age vs. patient history of congestive heart failure).

4.1. Experiment Settings

We evaluate ISRL from two perspectives: (i) the benefits of incorporating the learning of geometric shape features for classification and (ii) the effectiveness of our developed invariant representation learning. We compare our model's performance with two categories of baseline algorithms: a typical image classifier extracting features solely from image space based on ERM [51] and a recent work that learns invariant classification features solely from the image space based on IRM [2].

We also compare ISRL with four state-of-the-art domain generalization models, including CORAL [47], DANN [15], IWDAN [48], and CLOvE [54] on 3D Brain and Cardiac MRI experiments. Please note that none of these algorithms explicitly model shape feature learning.

Classification evaluation. We compare all methods with four classification network backbones on both simulated 2D and real 3D brain MRI datasets, including a custom four-block convolutional neural network (CNN), AlexNet [28], ResNet18 [18], and VGGNet11 [45]. For the custom CNN, each block comprises a convolutional layer with a kernel size of 5×5, a max-pooling layer with a size of 3×3, and batch normalization. We examine the micro-averaged accuracy (Acc.), precision (Prec.), and F1-score (F1-sc.) for the 2D simulated and 3D brain experiments. For all experiments on the 3D cardiac MRI videos, we use vision transformer (ViT) [14] and Video-ViT (ViViT) [3] as the network

backbone. We evaluate the accuracy and F1-score for classification performance.

To investigate the robustness of ISRL on 2D simulated dataset, we manipulate the images by increasing the number of flipped labels p^e (i.e., from 10% to 50%) [2] and compare the accuracies of the baselines on all network backbones. Besides, we cover an ablation study of ISRL measuring the effectiveness of shape information and invariant learning scheme on the network's prediction.

Evaluation on joint learning. To evaluate the effectiveness of jointly learning the invariant shape representations along with the classification task, we conduct an analysis by comparing ISRL with a disjoint two-step training approach. In the latter, our geometric shape learning network is employed as an initial preprocessing stage to extract geometric shape features. These extracted features are then integrated with image features, which are the inputs to the invariant classification network. We run validation on both 2D simulated data and 3D brain MRIs.

Ablation on invariant shape learning module. We perform an ablation study comparing our invariant shape learning model ISRL against non-invariant ERM (with and without shape features) and invariant IRM (omitting shape features) on all datasets.

Parameter setting. We set the number of time steps for Euler integration in EPDiff (Eq. (2)) as 10, and the noise variance $\sigma = 0.02$. Besides, we set the parameter $\alpha = 3$ for the operator $\mathcal{L}$, and the batch size as 32 and 8 for 2D and 3D experiments, respectively. For network training, we utilize the cosine annealing learning rate scheduler that starts with a learning rate of $\eta =$ 1e-3. We perform a cross-validation with increased penalty weights on all datasets to determine an optimal parameter λ that governs the strength of the invariance penalty in the network loss (in Eq. (6)) of our ISRL model, and the baseline algorithm IRM. All models are trained by the Adam optimizer with the best validation performance until convergence.

4.2. Results

Tab. 1 reports the micro-averaged classification performance on 2D colored Google Quickdraw (with $p^l = 0.25$, meaning 25% flipped labels) for all backbones. Our proposed model consistently outperforms the baselines across all backbones. More specifically, classifiers that incorporate shape features, (such as our ISRL), demonstrate notably enhanced accuracy compared to classifiers that rely solely on image intensities (IRM / ERM). Furthermore, as expected, classifiers utilizing invariant learning (our ISRL or IRM) are more robust to spurious correlations introduced by colors, resulting in much higher classification accuracy when compared to classifiers without invariant learning (ERM).

Fig. 3 illustrates the classification accuracy of all methods with an increased percentage of flipped labels (where $p^e = \{0, 0.25, 0.4, 0.5\}$) on four network backbones. All models demonstrate similar accuracy under ideal conditions of 0% label flipping and no spurious correlations between the color information and image labels. However, as the percentage of label flipping increases, we observe a substantial performance drop. The baseline models experience a decrease of 30-55% in accuracy vs. our proposed model shows a much lower drop of approximately 13-20%. Overall, our model consistently archives higher accuracy ($> 10\%$) than the baselines. This indicates that ISRL is more robust when confronted with unseen data from a different distribution in the testing environment.

Tab. 2 reports a comparison of binary classification (AD vs. CN) results on 3D real brain MRIs for all backbones, as well as related domain generalization methods that fo-

Table 1. A comparison of classification performance on simulated 2D data over all models with different network backbones.

Model	*Network*	*Acc*	*Prec*	*F1-sc*
ERM		48.67	48.52	48.17
IRM	CNN	68.67	68.86	68.61
ISRL		**78.89**	**79.64**	**78.89**
ERM		48.44	48.44	48.43
IRM	AlexNet	70.67	71.16	70.03
ISRL		**81.11**	**83.04**	**80.67**
ERM		48.67	48.95	47.79
IRM	ResNet	71.11	73.84	70.38
ISRL		**87.56**	**87.77**	**84.59**
ERM		47.56	47.35	47.27
IRM	VGGNet	68.89	72.83	68.56
ISRL		**79.79**	**80.96**	**79.77**

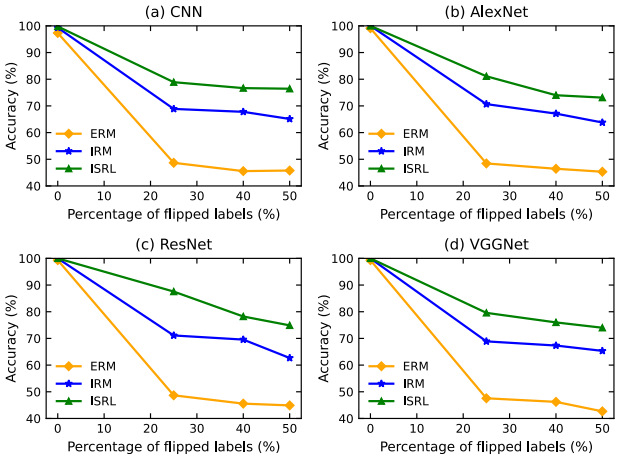

Figure 3. A comparison of the baselines and ISRL on four different backbones over increased probability of label flipping. ISRL exhibits superior generalization across backbones under increasing label noise, suggesting enhanced invariance in its learned representations.

Table 2. A comparison of classification accuracy on real 3D brain MRIs over all models with different network backbones.

	CNN			*AlexNet*			*ResNet*			*VGGNet*		
Model	Acc.	Prec.	F1-sc.	Acc.	Prec.	F1-sc.	Acc.	Prec.	F1-sc.	Acc.	Prec.	F1-sc.
ERM	66.00	69.87	64.27	62.00	62.10	61.92	70.67	73.49	69.76	62.67	72.80	58.00
IRM	69.33	69.56	69.25	68.00	72.00	66.47	72.67	76.65	71.60	65.33	65.89	65.03
CORAL	71.11	70.00	73.16	70.00	72.73	71.54	75.56	72.50	79.31	76.67	82.73	81.54
DANN	73.33	70.34	77.96	71.11	70.00	73.16	75.19	71.67	80.48	77.78	72.12	73.87
IWDANN	73.33	74.29	72.07	72.22	69.32	67.31	75.93	72.50	79.31	77.78	72.90	72.90
CLoVE	74.00	74.04	73.98	75.56	72.50	**77.80**	80.37	**81.43**	78.97	**82.22**	80.83	**83.12**
ISRL	**76.67**	**76.67**	**76.66**	**76.00**	**77.23**	75.72	**84.44**	80.37	**86.77**	**82.22**	**82.73**	81.91

cus on image intensity features (CORAL [47], DANN [15], IWDAN [48], and CLOvE [54]). Our ISRL model shows superior performance with consistently improved classification accuracy. Note that such a classification task on human brain images with large age variations is challenging, given that age has been demonstrated to induce anatomical brain changes [41, 68]. A typical classification network based on ERM achieves $\leq 70\%$ accuracy on our dataset.

Similarly, Tab. 3 reports the classification performance on the real 3D cardiac MRI videos (classifying subjects with scar vs. non-scar). Our model ISRL consistently outperforms all methods, including IRM and domain generalization-based models.

Table 3. Accuracy comparison on real 3D cardiac MRI videos over all models under ViT and ViViT network backbones.

	ViViT		*ViT*	
Model	Acc.	F1-sc.	Acc.	F1-sc.
ERM	79.67	79.77	78.65	77.93
IRM	86.51	86.43	82.02	81.96
CORAL	83.15	83.87	79.78	76.32
DANN	80.90	80.00	79.78	81.63
IWDANN	83.15	82.76	80.90	80.46
CLoVE	84.24	83.67	85.39	85.06
ISRL	**88.76**	**88.76**	**87.64**	**87.63**

Fig. 4 reports the performance comparison of our joint learning model with a disjoint (two-step) training approach on both 2D simulated and 3D brain datasets (based on all network backbones). Our joint learning model consistently outperforms the two-step approach by approximately 2-3%.

Tab. 4 displays the classification performance of varying penalty weights (λ) for both IRM and our model ISRL on different backbones. It shows that a small value of λ weakens the regularizer's effect and reduces the classifier's efforts to search for invariant representations, while a larger value of λ can result in overly emphasized commonalities at the cost of domain-specific performance. As λ increases, both models' performance declines beyond a threshold.

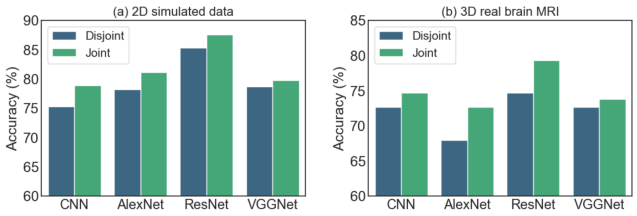

Figure 4. Comparison of disjoint (two-step) learning vs. our joint approach on both 2D simulated data and 3D brain MRIs on different network backbones. Joint optimization of shape learning and invariant classification yields superior performance.

Table 4. Classification accuracy (%) comparison across various penalty weights (λ) under ResNet backbone.

Dataset	*2D simulated data*		*3D brain MRI*	
Penalty	IRM	ISRL	IRM	ISRL
10000	62.67	74.89	65.33	74.67
25000	65.11	76.44	66.67	79.33
50000	69.56	85.33	72.67	**84.44**
75000	71.11	**87.56**	68.00	76.00
100000	68.67	78.72	68.67	78.00

Tab. 5 presents an ablation study evaluating the effectiveness of geometric shape features, $\mathcal{G}_E$, and invariant learning mechanisms. We first compare the classification model in an empirical learning setting without invariant learning, with or without geometric features. We then compare with the introduced invariant learning mechanism. While training with geometric shape features improves classifier performance over using only image features, $\mathcal{I}_E$, our ISRL (integrating invariant learning of image and shape features) achieves the best performance, surpassing the IRM that operates solely in the image space.

Fig. 5 visualizes activation maps [65] of the last convolution layers for various baseline models under ResNet backbones. It shows that ERM, when incorporating shape features alongside image features, captures more relevant object-focused representations than when using image

Table 5. An ablation study on the effectiveness capturing invariant geometric shape features, $\mathcal{G}_E$, with image features, $\mathcal{I}_E$, on both simulated 2D and real 3D datasets.

Datasets	Non-Invariant		Invariant	
	$\mathcal{I}_E$	$\mathcal{I}_E + \mathcal{G}_E$	$\mathcal{I}_E$	*ISRL*
2D Quick-draw	48.67	52.22	71.11	**87.56**
3D Brain	70.67	74.00	72.67	**84.44**
3D Cardiac	79.67	85.39	86.51	**88.76**

features alone. However, our ISRL model, utilizing an invariant feature learning mechanism in integrated shape and image spaces, further improves the network attention to object shapes across all classes, demonstrating its enhanced capacity to capture relevant geometric structures effectively.

Limitation & Discussion. Intuitively, IRM captures factors that have invariant relations with the label to make robust predictions in different environments. The connection between IRM and causality originates from the invariance of causal mechanism in causal inference theory, where the relation between each variable and its "parent" variables is invariant. Such a connection was discussed in the original IRM paper and follow-up works [2, 10, 33, 56]. Existing studies have shown that IRM effectively removes spurious correlations under different causal models with certain assumptions [30, 31, 64]. However, if the real-world data does not meet such assumptions (e.g., when no causal variables are available for the label, or the spurious correlation is stronger than the invariant relations in all environments), IRM may not outperform other methods. Additionally, IRMs can result in suboptimal predictors when training distributions fail to sufficiently cover the full range of potential test scenarios, or when unobserved confounders affect both input features and target variables across environments; leading to limited generalization performance. As a result, the model may struggle to adapt to new data distributions, thereby weakening its robustness and reliability in real-world applications. Addressing the limitation of IRM is out of the scope of this paper, our focus is harnessing its effectiveness in suitable scenarios.

While our current experimental design on the real-world data is in the context of a specified environment, it is noteworthy that our proposed ISRL can be naturally general to multi-environment settings. Future work will explore other potential variables, such as sex and genetics [41] or batch effects due to imaging sites [53, 58].

5. Conclusion

This paper introduces a novel model, named ISRL, that for the first time develops invariant shape representation learning for image classification. In contrast to previous approaches that derive features in image space, our model jointly learns the invariant features to image labels in an integrated space of images and geometric deformations. To achieve this, we develop a new learning paradigm that captures the invariant shape and image features across multiple training environments by minimizing the risk of spurious correlations that might be present in the shifted data distributions. Experimental results show that our method significantly improves the classification performance on simulated 2D data, real 3D brain images, and cardiac MRI videos.

Our work on ISRL is an initial attempt to explore the power of learning causal shape representations for image analysis tasks. Future directions include but not limited to i) investigating the interpretability of ISRL in shape-based deep networks, ii) generalizing the model's ability to learn causally invariant image and shape features from multi-modal image data, and iii) developing a robust classification model to deal with unknown hidden spurious factors. Besides, investigating the problem under more issues in practice (such as insufficient labels, noises, selection biases, and privacy concerns) would also be an interesting direction.

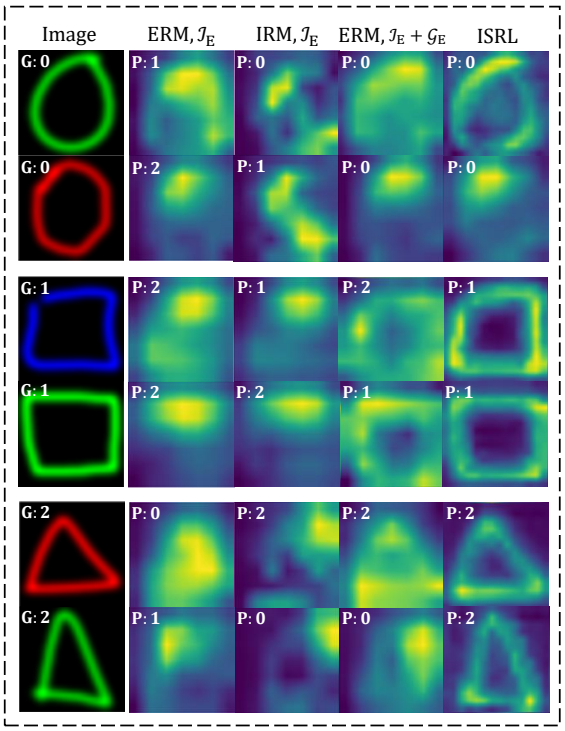

Figure 5. A visual comparison of activation maps generated by ERMs and invariant models (IRMs, including ISRL). G: Ground-Truth, P: Prediction.

Acknowledgements

This work was supported by NSF CAREER Grant 2239977 and NIH 1R21EB032597.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020. 1, 2, 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 3, 4, 5, 6, 8

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 5

[4] Vladimir Arnold. Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In *Annales de l'institut Fourier*, volume 16, pages 319–361, 1966. 2

[5] MIRZA Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005. 2

[6] Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9271–9280, 2018. 2

[7] Ch Brechbühler, Guido Gerig, and Olaf Kübler. Parametrization of closed surfaces for 3-d shape description. *Computer vision and image understanding*, 61(2):154–170, 1995. 2

[8] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020. 1

[9] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Ertunc Erdil, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised task-driven data augmentation for medical image segmentation. *Medical Image Analysis*, 68:101934, 2021. 4

[10] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020. 1, 8

[11] Chen Chen, Chen Qin, Cheng Ouyang, Zeju Li, Shuo Wang, Huaqi Qiu, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert. Enhancing mr image segmentation with realistic adversarial data augmentation. *Medical Image Analysis*, 82:102597, 2022. 4

[12] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3

[13] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[15] 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. 2, 5, 7

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[17] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8967, 2021. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5

[19] Jacob Hinkle, David Womble, and Hong-Jun Yoon. Diffeomorphic autoencoders for lddmm atlas building. 2018. 3

[20] Yi Hong, Polina Golland, and Miaomiao Zhang. Fast geodesic regression for population-based image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2017. 1

[21] Tonmoy Hossain, Fairuz Shadmani Shishir, Mohsena Ashraf, MD Abdullah Al Nasim, and Faisal Muhammad Shah. Brain tumor detection using convolutional neural network. In *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*, pages 1–6. IEEE, 2019. 1

[22] Tonmoy Hossain and Miaomiao Zhang. Mgaug: Multimodal geometric augmentation in latent spaces of image deformations. *arXiv preprint arXiv:2312.13440*, 2023. 2

[23] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008. 2, 5

[24] Nivetha Jayakumar, Tonmoy Hossain, and Miaomiao Zhang. Sadir: shape-aware diffusion models for 3d image reconstruction. In *International Workshop on Shape in Medical Imaging*, pages 287–300. Springer, 2023. 1

[25] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The quick, draw!-ai experiment. *Mount View, CA, accessed Feb*, 17(2018):4, 2016. 2, 4

[26] Minjeong Kim, Guorong Wu, Qian Wang, Seong-Whan Lee, and Dinggang Shen. Improved image registration by sparse patch-based deformation estimation. *Neuroimage*, 105:257–268, 2015. 3

[27] John Kjekshus. Arrhythmias and mortality in congestive heart failure. *The American journal of cardiology*, 65(19):42–48, 1990. 5

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[29] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 2

[30] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022. 8

[31] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 325–335. Springer, 2021. 8

[32] Peter Lorenzen, Brad C Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8*, pages 411–418. Springer, 2005. 2

[33] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. 1, 8

[34] Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2):209–228, 2006. 2

[35] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020. 1

[36] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1

[37] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. 4

[38] Stephen M Pizer, Daniel S Fritsch, Paul A Yushkevich, Valen E Johnson, and Edward L Chaney. Segmentation, registration, and measurement of shape variation via image object shape. *IEEE transactions on medical imaging*, 18(10):851–865, 1999. 2

[39] Philip J Podrid, Richard I Fogel, and Therese Tordjman Fuchs. Ventricular arrhythmia in congestive heart failure. *The American journal of cardiology*, 69(18):82–96, 1992. 5

[40] Cyrius A Raji, OL Lopez, LH Kuller, OT Carmichael, and JT Becker. Age, alzheimer disease, and brain structure. *Neurology*, 73(22):1899–1905, 2009. 5

[41] Rajat Rasal, Daniel C Castro, Nick Pawlowski, and Ben Glocker. Deep structural causal shape models. In *European Conference on Computer Vision*, pages 400–432. Springer, 2022. 5, 7, 8

[42] Sureerat Reaungamornrat. *Deformable Image Registration for Surgical Guidance using Intraoperative Cone-Beam CT*. PhD thesis, Johns Hopkins University, 2017. 1

[43] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012. 5

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2

[47] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 2, 5, 7

[48] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020. 5, 7

[49] Vagan Terziyan and Oleksandra Vitko. Causality-aware convolutional neural networks for advanced image classification and generation. *Procedia Computer Science*, 217:495–506, 2023. 2

[50] Marc Vaillant, Michael I Miller, Laurent Younes, and Alain Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:S161–S169, 2004. 2

[51] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 4, 5

[52] François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241, 2012. 2

[53] Christian Wachinger, Anna Rieckmann, Sebastian Pölsterl, Alzheimer's Disease Neuroimaging Initiative, et al. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, 2021. 8

[54] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227, 2021. 2, 5, 7

[55] Jian Wang and Miaomiao Zhang. Geo-sic: Learning deformable geometric shapes in deep image classifiers. *The Conference on Neural Information Processing Systems*, 2022. 1, 3

[56] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022. 1, 8

[57] Shuo Wang, Hena Patel, Tamari Miller, Keith Ameyaw, Akhil Narang, Daksh Chauhan, Simran Anand, Emeka Anyanwu, Stephanie A Besser, Keigo Kawaji, et al. Ai based cmr assessment of biventricular function: clinical significance of intervendor variability and measurement errors. *Cardiovascular Imaging*, 15(3):413–427, 2022. 2, 5

[58] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020. 1, 8

[59] Chao-Han Huck Yang, I-Te Hung, Yi-Chieh Liu, and Pin-Yu Chen. Treatment learning causal transformer for noisy image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6139–6150, 2023. 1, 2

[60] Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. *Advances in Neural Information Processing Systems*, 34:19448–19460, 2021. 2

[61] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 2

[62] Chenfei Ye, Ting Ma, Dan Wu, Can Ceritoglu, Michael I Miller, and Susumu Mori. Atlas pre-selection strategies to enhance the efficiency and accuracy of multi-atlas brain segmentation tools. *PloS one*, 13(7):e0200294, 2018. 3

[63] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020. 1

[64] Samira Zare and Hien Van Nguyen. Removal of confounders via invariant risk minimization for medical diagnosis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 578–587. Springer, 2022. 1, 8

[65] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 7

[66] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020. 1

[67] Miaomiao Zhang and Polina Golland. Statistical shape analysis: From landmarks to diffeomorphisms, 2016. 1

[68] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. *Nature communications*, 11(1):6010, 2020. 1, 2, 5, 7

[69] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 3