Domain Generalization for Medical Image Analysis: A Review

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk, *Senior Member, IEEE*

Abstract— **Medical image analysis (MedIA) has become an essential tool in medicine and healthcare, aiding in disease diagnosis, prognosis, and treatment planning, and recent successes in deep learning (DL) have made significant contributions to its advances. However, deploying DL models for MedIA in real-world situations remains challenging due to their failure to generalize across the distributional gap between training and testing samples — a problem known as *domain shift*. Researchers have dedicated their efforts to developing various DL methods to adapt and perform robustly on unknown and out-of-distribution data distributions. This paper comprehensively reviews domain generalization studies specifically tailored for MedIA. We provide a holistic view of how domain generalization techniques interact within the broader MedIA system, going beyond methodologies to consider the operational implications on the entire MedIA workflow. Specifically, we categorize domain generalization methods into data-level, feature-level, model-level, and analysis-level methods. We show how those methods can be used in various stages of the MedIA workflow with DL equipped from data acquisition to model prediction and analysis. Furthermore, we critically analyze the strengths and weaknesses of various methods, unveiling future research opportunities.**

Index Terms—**Domain generalization, medical image analysis, out-of-distribution, deep learning**

I. Introduction

Medical image analysis (MedIA) plays a critical role in modern healthcare, enabling accurate diagnosis and treatment planning for various diseases. Over the past few decades, deep learning has demonstrated great success in automating various MedIA tasks such as disease diagnosis [1], prognosis [2], and treatment planning [3]. These achievements have become feasible by the capability of deep learning algorithms to learn from vast amounts of data, identify patterns, and generate predictive models that aid in MedIA tasks. Moreover, the availability of powerful computational resources has greatly expedited the process of training deeper, wider, and more complex models. These have led to impressive performance in relatively well-controlled settings. However, many challenges in real-world scenarios remain.

With homogeneous data distributions, well-designed models perform on par with and often surpass their human counterparts in many applications. However, their reliability and robustness can be compromised when presented with previously unseen, out-of-distribution, or heterogeneous data. This highlights a common challenge in the field of MedIA: the limited capacity of models to generalize to unfamiliar data distributions. Changes in data distribution can result from variations in imaging equipment, protocols, or patient populations. Domain generalization aims to overcome these challenges by developing models that can adapt to new, unseen domains without compromising performance.

A. Domain Generalization for Medical Image Analysis

Domain generalization has emerged as a crucial field in deep learning, particularly in applications where the ability to generalize across diverse domains is of importance. Its significance is particularly high in the context of MedIA, where data is very heterogeneous. To better understand the unique challenges of domain generalization for MedIA, it is important to consider the following factors:

- **Image appearance variability**: Variability in medical imaging refers to differences and inconsistencies typically manifest during the data acquisition process [4]. This variability may arise externally from using different modalities, protocols, scanner types, and patient populations across multiple healthcare facilities, while internal variability may also occur within a controlled setting (*e.g.*, same scanner or healthcare facility) due to factors such as hardware aging, software parameter variations, and human error (*e.g.*, human motion).
- **Complex and high-dimensional data**: Medical images are often high-dimensional and may contain multiple channels or sequences. Many of such datasets span from thousands of pixels to gigapixel [5] and from 2-dimensions to 5-dimensions [6]. This complexity makes it difficult to identify and extract domain-invariant features that can generalize well across different domains.
- **Challenging data acquisition, organization, and labeling**: Large-scale, diverse, and labeled datasets are difficult to obtain due to the cost of data acquisition, privacy concerns, data sharing restrictions, and the labor-intensive nature of manual annotation by medical experts. Furthermore, quality assurance is challenging as the medical image is prone to noise and artifacts, such as patient motion, scanner imperfections, and imaging artifacts from hardware or software limitations.
- **Model interpretability, safety, and privacy**: In Me-

J. Yoon is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: wltjr1007@korea.ac.kr).

K. Oh is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: ksohh@korea.ac.kr).

Y. Shin is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: usxxng@korea.ac.kr).

M. Mazurowski is with the Departments of Radiology, Biostatistics and Bioinformatics, Electrical and Computer Engineering, and Computer Science, Duke University, Durham, NC 27705, USA (e-mail: maciej.mazurowski@duke.edu)

H.-I. Suk is with the Department of Artificial Intelligence and the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea, and the corresponding author (e-mail: hisuk@korea.ac.kr).

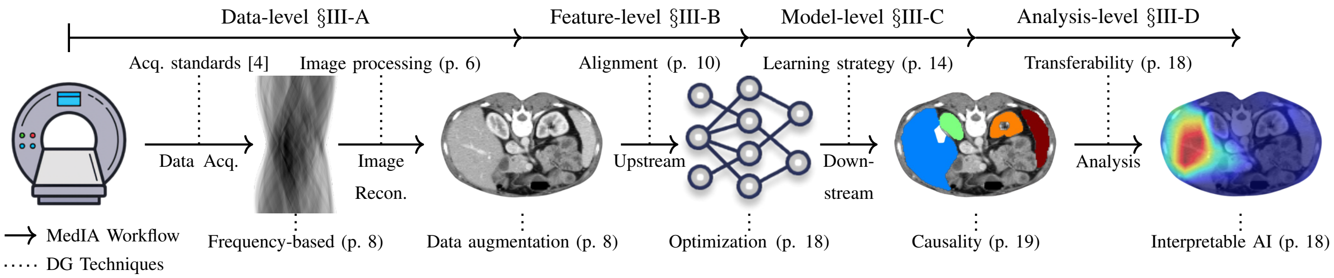


Fig. 1. Overview of the medical imaging analysis (MedIA) pipeline illustrating various stages and their associated domain generalization (DG) techniques. Stages include data acquisition, image reconstruction, upstream feature extraction, downstream task, and analysis. Each stage is associated with references to specific sections (§), pages (p.), or external citations where the techniques are detailed.

dIA, ensuring model interpretability, safety, and compliance with regulatory and ethical standards is crucial. Robustness against adversarial examples and to out-of-distribution samples is essential to prevent adverse effects on patient care. Additionally, privacy-preserving data sharing and collaboration in multi-center contexts add complexity to implementing domain generalization techniques.

B. *Our Contributions*

With these challenging factors in mind, this review provides a comprehensive review of domain generalization techniques specifically tailored to MedIA. There already exist few review papers on domain generalization with a specific focus on MedIA, but these are limited to specific data domain and task, *i.e.*, mammography-based mass detection [7], electroencephalography (EEG)-based emotion assessment [8], and computational pathology [9]. Also, there are several review papers on related topics for MedIA, such as domain adaptation [10], [11] and harmonization [4]. However, domain generalization presents unique challenges compared to these tasks.

Multiple survey papers have been published that offer a comprehensive understanding of domain generalization for general data domains and tasks, presenting broader perspectives [12], [13], [14], [15], [16], [17], [18] as well as focused approaches such as causal models [19], graph models [20], and federated learning [21]. While these surveys serve as a detailed reference for specific algorithms, techniques, and model architecture, they lack an in-depth exploration of the system-level implications of domain generalization on the overall workflow of MedIA.

Our review aims to provide a holistic view of how domain generalization techniques interact within the broader structure of a MedIA system. We go beyond the methodological hierarchy presented in previous surveys and delve into the operational consequences of domain generalization on the entire MedIA workflow (see Fig. 1). Our focus is on understanding how domain generalization can be seamlessly integrated into every step of the decision-making process, including but not limited to data acquisition, pre-processing, model prediction, and analysis. To this end, we categorize domain generalization techniques into each step of the MedIA workflow, *i.e.*, from data preparation to analysis.

C. *Scope of Review*

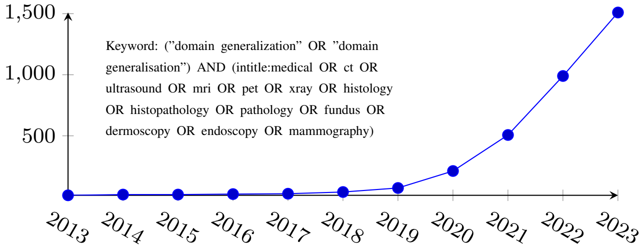


Fig. 2. Number of publications per year on the Google Scholar database.

Literature search and selection were conducted by researchers experienced in machine learning and medical image analysis. We used the Google Scholar search engine with three different search strategies, resulting in a database of 1,621 papers. First, we searched for all papers that cited the existing domain generalization review [12], [13], [14], [19], [20], [11], [8]. Second, we searched Google Scholar using the exact phrase “domain generaliz(s)ation” and medical-related keywords (“Medical, CT, ultrasound, MRI, PET, X-ray, histology, histopathology, pathology, fundus, dermoscopy, endoscopy, mammography”) and selected 1,000 papers sorted by relevancy. Lastly, we searched for the top 1,000 papers with the terms “unseen”, “domain”, and medical-related keywords. The eligibility criteria for papers to be included in this review are that they have conducted at least one experiment involving the use of medical images within the domain generalization problem settings (see §II-A), regardless of their use of the term “domain generalization” in their paper (small number of papers instead use “unseen”, “out-of-distribution”, or their own terms). Peer-reviewed published papers were prioritized, but non-peer-reviewed archive papers (*e.g.*, arXiv, bioRxiv) were also included if they had been deemed particularly suitable for selection (*e.g.*, highly relevant, highly significant, highly cited).

II. Background

A. *Problem Definition*

In this section, we formalize the problem of domain generalization (DG) by following the mathematical notations and formulations used in previous surveys [12], [21] (see Table I for the definition of mathematical notations). Let $\mathcal{X}$ denote a

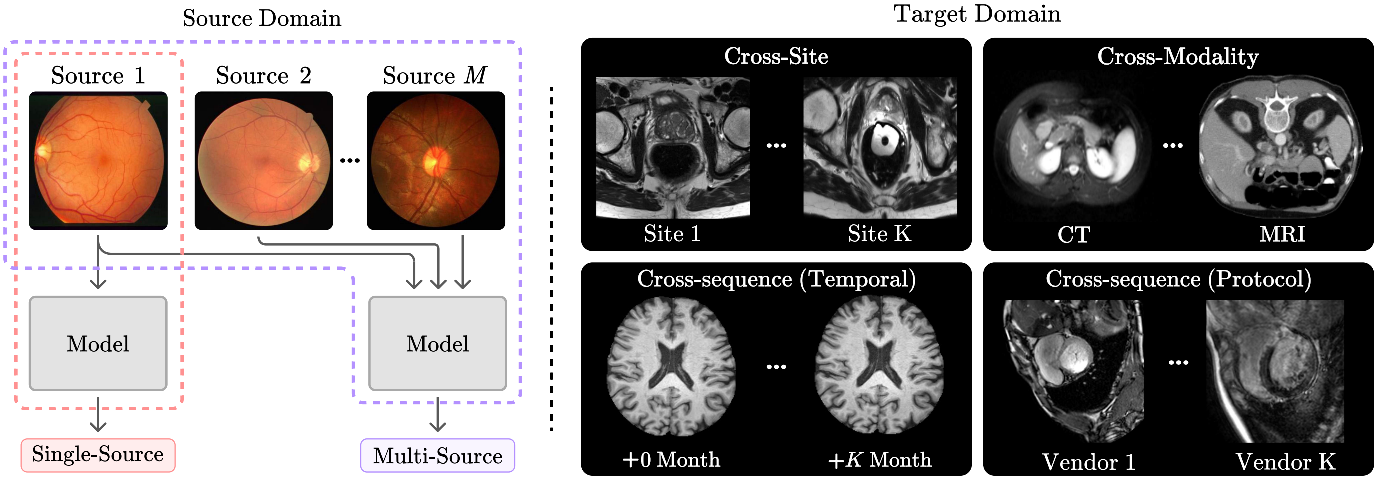


Fig. 3. Settings of source and target domain for domain generalization. Source domain consists of M domains, where $M = 1$ refers to single-source and $M > 1$ refers to multi-source settings. The target domain consists of K domains. Cross-site, cross-sequence, and cross-modality define unique settings of the target domain for MedIA. Cross-site refers to generalization to different sites, *e.g.*, different devices, and healthcare facilities. Cross-sequence refers to generalization to different sequences, *i.e.*, different acquisition times or imaging protocols. Cross-modality refers to generalization to different modalities, *e.g.*, CT to MRI.

TABLE I
DEFINITION OF MATHEMATICAL NOTATIONS, FOLLOWING [12].

Notation	Definition	Notation	Definition
$\mathcal{X}, \mathcal{Z}, \mathcal{Y}$	Input, feature, output space	$\mathbf{x}, \mathbf{z}, y$	Input, feature, output variables
P_{XY}	Probability distribution	$\mathcal{S}$	Domain
n_i	i-th domain data count	M, K	Number of source, target domains
$\mathcal{L}(\cdot,\cdot)$	Loss function	$h(\cdot)$	Predictive function
$\mathcal{M}(\cdot)$	Manipulation function	$f(\cdot)$	Feature mapping function
$\mathcal{D}(\cdot,\cdot)$	Dissimilarity function		

nonempty input space and $\mathcal{Y}$ an output space (*e.g.*, labels). A domain is composed of data that are sampled from a joint distribution of the input sample and output label P_{XY}. We denote a domain as $\mathcal{S} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n} \sim P_{XY}$, where $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, and n is the number of data pairs.

In DG, we are given M training (source) domains $\mathcal{S}_{source} = \{\mathcal{S}^i \mid i = 1, \cdots, M\}$, where $\mathcal{S}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i-th domain with n_i data pairs. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq M$. The goal of DG is to learn a robust and generalizable predictive function $h : \mathcal{X} \rightarrow \mathcal{Y}$ from the M training domains to achieve a minimum prediction error on an K *unseen* test (target) domain $\mathcal{S}_{target} = \{\mathcal{S}^i \mid i = M+1, \cdots, M+K+1\}$ (*i.e.*, $\mathcal{S}_{target}$ cannot be accessed in training). In other words, the goal of DG is to minimize the generalization error:

$$\mathbb{E}_{(\mathbf{x},y)\in\mathcal{S}_{target}}[\mathcal{L}(h(\mathbf{x}), y)]. \tag{1}$$

B. Settings of Domain Generalization

This subsection elucidates different settings for DG in the MedIA workflow, detailing the various configurations and challenges present in both the source and target domains during the implementation process (see Fig. 3). A critical aspect of DG is overcoming the *domain shift* — differences between the source and target domains that hinder the model's ability to generalize. These gaps typically arise due to variations such as:

- Intensity Variations: Differences in image brightness or contrast, *e.g.*, varying exposure levels in chest X-rays or staining intensities in histology images.
- Resolution Differences: Variations in image detail, such as differences between high and low-resolution ultrasound devices or varying pixel densities in fundus photographs.
- Noise Characteristics: Varying levels of image noise, *e.g.*, speckle noise in ultrasound or grain in low-dose CT scans.
- Artifact Patterns: Modality-specific artifacts, such as motion blur in MRI, beam hardening in CT, or light reflections in endoscopy images.
- Anatomical Variations: Differences in patient populations, leading to variations in organ sizes or shapes across different datasets.
- Label Distributions: Varying disease prevalence or severity between domains, affecting class balance in tasks like skin lesion classification from dermatology images.
- Acquisition Protocols: Differences in imaging techniques, such as varying MRI sequences, CT reconstruction kernels, or staining protocols in histopathology.

As it is often challenging to attribute domain shifts to a single cause, we have systematically categorized these variations with into cross-site, cross-modality, cross-temporal, and cross-protocol domain shifts (see §II-B2). These shifts can be further classified into covariate and concept shifts (see §II-B3), providing a comprehensive framework for understanding and addressing domain shifts in MedIA.

1) Settings for Source Domain

DG typically focuses on two settings regarding the number of source domains: *multi-source DG* and *single-source DG* [13]. The multi-source setting assumes multiple distinct but relevant domains are available (*i.e.*, $M > 1$). By leveraging the data from these domains, representations invariant to disparate marginal distributions are learned. This is usually accomplished by minimizing the domain discrepancy among the source domains during the training process. The single-source setting assumes training data is homogeneous (*i.e.*, $M = 1$). Therefore, this setting does not require domain labels during training. Single-source DG tends to be more challenging than multi-source DG as it may not capture the full diversity of data distributions that exist across different domains. Refer to Section IV for some extreme settings for domain generalization (such as, open-set DG, source-free DG, and unsupervised DG).

2) Settings for Target Domain

There are three settings for target domains that are unique to DG for MedIA as follows.

- **Cross-site DG**: As the most prevalent form of DG for MedIA, the goal of cross-site DG is to develop models that can generalize well across different medical imaging datasets collected from multiple healthcare institutions. Cross-site DG helps in creating more robust models that can be deployed across different healthcare institutions without the need for extensive site-specific fine-tuning. Cross-site DG research often define a domain or 'site' as an individual healthcare facility, but some define it as an individual vendor, scanner, or even an individual patient when inter-patient variability is extremely large, which is a common setting in many EEG studies.
- **Cross-sequence DG**: Medical imaging data often consists of multiple types of sequences or series, each capturing different aspects of the underlying anatomy or pathology. The most commonly used sequences are the *cross-temporal* sequences and *cross-protocol* sequences. Temporal sequences are images taken at different time points (*e.g.*, before, during, and after treatment), while protocol sequences are images of the same modality with different acquisition protocols. For example, in magnetic resonance imaging (MRI), protocol sequences like T1-weighted, T2-weighted, and fluid attenuated inversion recovery (FLAIR) images provide different tissue contrasts and diagnostic information.
- **Cross-modal DG**: Medical imaging encompasses a wide range of modalities, such as MRI, computed tomography (CT), and X-ray. Each modality provides different types of information and is suited for specific clinical applications. This type of DG can involve training a model on data from one modality and testing its performance on data from a different, previously unseen modality.

3) Settings for Domain Shift

In the context of DG, domain shift can be categorized into *covariate shfit* and *concept shift*. **Covariate shift** happens when the data distribution between the source and target domains is different, but the functional relationship between the input and output (the "concept") remains the same. Given the source domain and a target domain, we have covariate shift when $P_X^{source} \neq P_X^{target}$ but $P_{Y|X}^{source} = P_{Y|X}^{target}$. Here, P_X and $P_{Y|X}$ denote the marginal distribution of the input features and the conditional distribution of the output given the input, respectively. **Concept shift** occurs when the functional relationship between the input and output changes, *i.e.*, $P_{Y|X}^{source} \neq P_{Y|X}^{target}$.

To illustrate, consider two clinics that perform brain MRI scans on their patients. A covariate shift might be caused by differences in the MRI scanners, patient population, or other factors that affect the appearance of the brain scans. On the other hand, a concept shift might occur when the diagnostic criteria or the diseases of interest vary between the clinics. For example, one clinic might focus on diagnosing Alzheimer's disease, whereas another might concentrate on detecting brain tumors, in which case Alzheimer's disease might not be deemed significant. Additionally, the concept shift can manifest in the differences in diagnoses made by various medical professionals. This type of shift is closely associated with alterations in the assigned labels (label shift) or the interpretation of these labels (semantic shift).

In the context of different DG settings, cross-site DG and cross-temporal DG could lead to covariate shift as the same concept (*e.g.*, the presence or absence of a disease) may be associated with different input features (*e.g.*, different patient populations) across different sites or at different times. In contrast, cross-protocol and cross-modal DG could potentially involve concept shifts. For example, a concept shift could occur when a model trained on MRI images, which highlights detailed information about soft tissues, struggles to correctly interpret CT scans that provide more detailed depictions of bone structures, essentially changing the underlying relationship between image features and the corresponding disease labels.

C. Related Machine Learning Tasks

TABLE II
RELATED MACHINE LEARNING TASKS CATEGORIZED BY COVARIATE AND CONCEPT SHIFT, AND ACCESS TO $\mathcal{S}_{target}$.

Task	Covariate	Concept	$\mathcal{S}_{target}$
Multi-Task Learning			●
Transfer Learning	●	●	●
Harmonization	●		●
Domain Adaptation (DA)	●		●
Unsupervised/Zero-shot DA	●		◐
Zero-shot Learning		●	◐
Test-time Adaptation	●		◐
Out-of-distribution		●	
Domain Generalization	●	●	

●: *Full access,* ◐: *Partial access (e.g., auxiliary information, mini-batch).*

In this subsection, we discuss the relationship between DG and its related machine-learning tasks and clarify their differences. The main takeaway is that DG restricts its access to the target domain data, while other tasks have full or partial access to the target domain distribution. An overview of related tasks is in Table II.

- **Multi-task Learning (MTL)** aims to learn a single model that performs well on multiple related tasks. In the context of DG, MTL can be viewed as learning a predictive function h that minimizes the combined risk

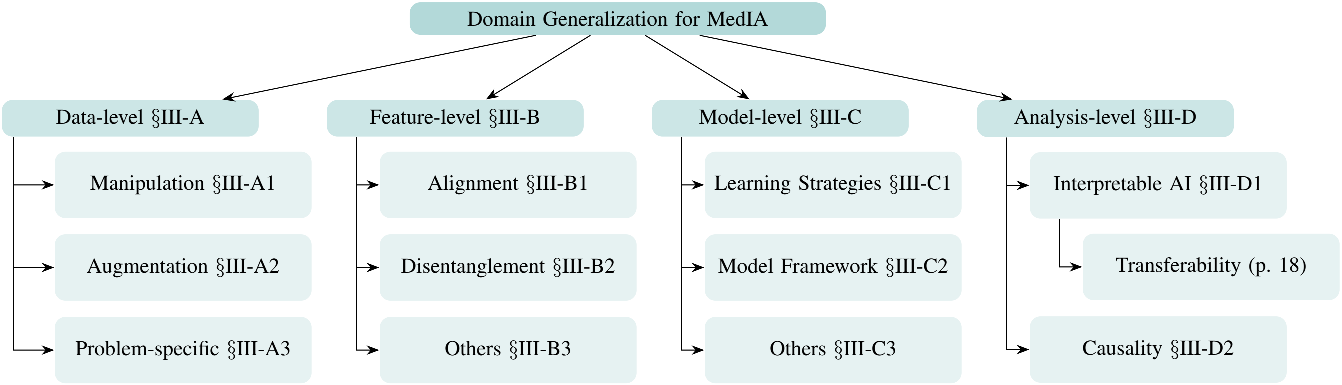


Fig. 4. Hierarchical structure of the different aspects of domain generalization (DG) for medical image analysis (MedIA). This taxonomy divides the DG strategies into four primary levels: data-level, feature-level, model-level, and analysis-level, each encompassing distinct sub-strategies.

over M related tasks. The main difference between MTL and DG is that MTL aims to perform well on the same set of tasks that the model was trained on, while DG aims to generalize to unseen data distributions.

- **Transfer Learning (TL)** aims to transfer the knowledge learned from one or more source domains to a different but related target domain. Both TL and DG deal with situations where the target distribution is different from the source distribution. However, in TL, the target domain is used during training (usually during fine-tuning), whereas in DG we assume no access to the target domain.
- **Harmonization** aims to reduce non-biological heterogeneity caused by cohort bias (*e.g.*, different scanner type or acquisition protocol). However, harmonization primarily focuses on cross-site datasets and does not necessarily impose restrictions on access to the target domain distribution. Most harmonization techniques are performed prior to model training mainly as a preprocessing technique.
- **Domain Adaptation (DA)** aims to tackle the domain shift problem encountered in new test environments. DA assumes the availability of labeled or unlabeled target data (*i.e.*, **unsupervised DA, UDA**) for model adaptation. **Source-free DA (SFDA)** assumes source data is unavailable after pretraining a model (*e.g.*, due to privacy reasons). **Zero-shot DA (ZDA)** limit its access to target domain data, but leverages auxiliary information related to the target domain. The primary distinction between UDA/SFDA/ZDA and DG lies in the (partial) access to target domain data during training.
- **Zero-shot Learning (ZSL)** is closely related to out-of-distribution (OOD) generalization in that it aims to classify test samples with concept shift, but ZSL generally leverages auxiliary information, such as attribute descriptions, related to the target domain.
- **Test-time Adaptation (TTA)** deals with the domain shift problem as well. TTA differs from DA in that only a single or mini-batch of test data is used for model tuning, which is often done in an online manner. TTA and DG both share the constraint of not having access to the target domain during training. However, TTA requires an additional step of fine-tuning at test time, requiring a mini-batch of target data.
- **OOD Generalization** aims to detect the concept shift between in-distribution (ID) and OOD data. While OOD and DG both assume no access to the target domain, the main difference between OOD and DG lies in that they focus on different domain shifts. Specifically, OOD mainly focuses on concept shift whereas DG considers both covariate and concept shift in their problem settings.

III. Methods

In this section, we review and explain a series of DG methods for medical imaging. We employ a bottom-up approach and categorize the methods into data-level, feature-level, model-level, and analysis-level DG methods (see Fig. 4). Then, we explore some DG methods under extreme constraints (see §IV).

- **Data-level generalization** methods focus on manipulating and generating input data to facilitate learning generalizable representations.
- **Feature-level generalization** methods focus on extracting domain-invariant features from input images to improve the generalization performance of models. These methods often involve learning a shared feature representation across multiple domains by extracting domain-invariant features.
- **Model-level generalization** aims to improve DG in medical imaging by refining the learning process, model structure, or optimization techniques.
- **Analysis-level generalization** methods help users understand, explain, and interpret the decision-making process of machine learning models.

This categorization is inspired by the various stages of the MedIA pipeline, *i.e.*, stages starting from data acquisition, image reconstruction, feature extraction, downstream task to analysis (see Fig. 1). Our aim is to provide researchers and practitioners a holistic view of how different approaches can be combined and applied throughout the MedIA pipeline.

It is important to note that some DG techniques can be applied at multiple levels of the MedIA pipeline (see Table III). Augmentation at the data-level (§III-A2) involves directly

TABLE III
DG TECHNIQUES APPLICABLE TO MULTIPLE LEVELS OF THE MEDIA PIPELINE.

	Data-level	Feature-level	Model-level
Augmentation	§III-A2	§III-B3a	
Adversarial	§III-A2b		§III-C1c
Contrastive		§III-B2a	§III-C1b

manipulating input images to increase data diversity, while feature-level augmentation (§III-B3a) operates on extracted feature representations to expand the feature space. Adversarial techniques at the data-level (§III-A2b) focus on generating challenging examples to improve robustness, whereas at the model-level (§III-C1c) they aim to learn domain-invariant features through adversarial training. Contrastive learning at the feature-level (§III-B2a) encourages similar samples to have close representations while pushing dissimilar ones apart, while at the model-level (§III-C1b) it is used as a self-supervised pretraining strategy to learn generalizable representations. These multi-level applications of similar concepts highlight how DG techniques can be integrated throughout the pipeline, each tailored to address domain shift at different stages of processing.

A. Data-level Generalization

The success of machine learning models often hinges on the training data's quality, quantity, and diversity. As the qualitative acquisition of medical images is challenging and costly, data-level generalization methods present an efficient and straightforward approach to enhance a model's generalization capability. These methods focus on manipulating and augmenting input data to increase the diversity and quantity of available samples, ultimately improving the model's adaptability to different domains. Data-level generalization can be divided into two primary techniques: **Data manipulation**, which transforms existing data to expose the model to a broader range of samples, and **data augmentation**, which creates new samples to further expand the model's exposure to various data variations. As these techniques are at the early stages of the MedIA workflow, *e.g.*, data acquisition and image reconstruction, some of them are **problem-specific methods** that require specialized model architectures or algorithms for the task at hand. The theoretical understanding of how these techniques enhance a model's generalization ability has been shown by Wang *et al.* [12], and empirical results [52] also show promising improvements in model performance on both out-of-distribution and in-distribution samples.

The general learning objective of data-level DG can be expressed as:

$$\min_{h} \lambda_1 \mathbb{E}_{\mathbf{x},y}[\mathcal{L}(h(\mathbf{x}),y)] + \lambda_2 \mathbb{E}_{\mathbf{x}',y'}[\mathcal{L}(h(\mathbf{x}'),y')], \quad (2)$$

where $\mathcal{S} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n} \sim P_{XY}$ refers to the source domain, $\mathcal{S}' = \{(\mathbf{x}_j', y_j')\}_{j=1}^{n} \sim P_{XY}$ refers to manipulated domain derived from the distribution of the source domain P_{XY} using data-level DG methods, and λ is a constant hyperparameter. The parameter λ_1 determines the extent to which original data contributes to the learning process, and λ_2 quantifies the influence of manipulated data on the process. Specifically, when $\lambda_1 > 0$ and $\lambda_2 > 0$, data augmentation is employed alongside original data, while if $\lambda_1 = 0$, the learning objective function relies exclusively on manipulated data. Hence, existing data-level DG can further be refined by choosing the manipulated domain $\mathcal{S}'$, resulting in the methods in the following sections.

1) Data Manipulation

In data manipulation methods, $\mathcal{S}'$ can be defined as a transformed version of the original dataset $\mathcal{S}$, where each sample has been modified using a specific manipulation function. This manipulation function, $\mathcal{M}(\cdot)$, can be a closed-form or learnable function that alters the characteristics of the data, making the manipulated data different from the source data. The specific form of the function $\mathcal{M}(\cdot)$ often depends on the data and the task. To this end, a manipulated domain $\mathcal{S}'$ that encapsulates data manipulation methods could be defined as:

$$\mathcal{S}' = \{(\mathcal{M}(\mathbf{x}_j), y_j)\}_{j=1}^{n} \sim P_{XY}. \quad (3)$$

III-A1a Image Processing Methods

Image processing techniques involve closed-form or learnable transformation functions to increase the diversity and quantity of training data. Examples of some traditional image processing methods include registration, resampling, and filtering, which are specifically designed for the distinct characteristics of the medical image data in question. Although many traditional image processing methods have empirically shown to improve the model's generalizability [53], they are predominantly employed as pre-processing tools for downstream tasks, rather than as standalone solutions for DG. Also, with the advancement of deep learning, there has been a gradual shift towards incorporating these techniques directly into deep learning architectures, enabling a more seamless integration of end-to-end learning of domain-invariant features (see Section III-B). Readers are referred to [54] for a comprehensive review of the traditional image processing methods. In the following paragraphs, we explore several deep learning-based image processing methods specifically designed for DG for MedIA tasks.

Intensity normalization methods aim to normalize the raw intensity values or their statistics to reduce the impact of variations in image intensity across different domains. Several deep learning-based works [55] have been proposed for intensity normalization technique, typically utilizing an autoencoder-based approach. For example, inspired by z-score normalization, Yu *et al.* [22] proposed a U-Net-based [56] self-adaptive normalization network (SAN-Net) for the stroke lesion segmentation task. The U-Net encoder of SAN-Net minimizes the inter-site discrepancy by learning the site-invariant representation with a site classifier and a gradient reversal layer, and the decoder outputs an intensity-normalized image that removes any site-related distribution shifts. Karani *et al.* [23] proposed an intensity denoising method for medical image segmentation. The DAE is trained on intensity-perturbed images to produce denoised outputs, which are then used to train a segmentation CNN.

Other image processing techniques often involve applying a linear or non-linear transformation to the image intensi-

TABLE IV
DATA-LEVEL DOMAIN GENERALIZATION METHODS. METHODS CATEGORIZED BY DIFFERENT SETTINGS FOR SOURCE AND TARGET DOMAINS (SEE §II-B), TASK, ORGAN, AND MODALITY USED IN EXPERIMENT.

Method	Specific	Ref.	Source	Target	Task	Organ	Modality
Image Processing	Intensity Normalization	[22]	Multiple	Site	Segmentation	Brain	MRI
		[23]	Single	Site, Sequence	Segmentation	Brain, Prostate, Cardiac	MRI
	Histogram Matching	[24]	Multiple	Site	Segmentation	Cardiac	MRI
		[25]	Multiple	Site	Segmentation	Atrial	MRI
		[26]	Multiple	Site	Segmentation	Retinal	Fundus
	Color Normalization	[27]	Multiple	Site	Detection	Tissue	Histology
		[28]	Multiple	Site	Classification	Retinal	Fundus
		[29]	Multiple	Site	Classification	Skin	Dermatology
		[30]	Single, Multi	Site	Segmentation, Classification	Retinal, Chest	Fundus, X-ray
Surrogate	Frequency-based DG	[31]	Multiple	Site	Segmentation	Brain	MRI
		[32]	Single, Multiple	Site	Segmentation	Retinal	Fundus
		[33]	Multiple	Site	Restoration	Retinal	Fundus
		[34]	Multiple	Site	Segmentation	Retinal	Fundus, OCT
		[35]	Single	Site	Segmentation	Retinal	Fundus
	Using Raw Signals	[36]	Multiple	Site	Segmentation	Brain	MRI
		[37]	Single	Sequence	Segmentation	Brain	MRI
	Dictionary Learning	[38]	Multiple	Sequence	Reconstruction	Brain	MRI
		[39]	Single	Site	Segmentation	Prostate, Retinal	MRI, Fundus
Augmentation	Randomization-based	[40]	Single	Sequence	Segmentation	Cardiac	MRI
		[33]	Multiple	Site	Restoration	Retinal	Fundus
		[41]	Single	Site	Segmentation	Retinal	Fundus
		[42]	Single	Site	Segmentation	Retinal	Fundus
	Adversarial-based	[43]	Multiple	Site	Segmentation	Spinal, Prostate, Colon	MRI, Histology
	Generative Model	[44]	Multiple	Site	Classification	Colon, Breast	Histology
		[45]	Single	Site	Classification	Colon	Histology
		[46]	Single	Sequence, Site	Detection	Liver	PET, CT
Problem-specific	Cross-modal Generative Model	[47]	Multiple	Modality, Sequence	Segmentation	Brain, Prostate, Abdominal	MRI, CT
		[48]	Single	Modality, Sequence	Segmantation	Prostate, Abdominal, Cardiac	MRI, CT
		[49]	Single	Modality	Segmentation	Cardiac, Abdominal	MRI, CT
	Stain Normalization	[50]	Multiple	Site	Detection	Breast	Histology
		[51]	Multiple	Site	Classification, Segmentation	Tissue, Breast	Histology

ties, such as histogram matching and color normalization. **Histogram matching** is a contrast adjustment method that scales pixel values to fit the range of a specified histogram. Ma [24] showed that augmenting the source domain with histogram-matched images improves generalization performance for the cardiac image segmentation task. A subsequent benchmark by Li *et al*. [25] also revealed that histogram matching had the highest performance compared to some commonly used DG methods for atrial segmentation. Gunasinghe *et al*. [26] proposed a randomized histogram matching method for glaucoma detection that sequentially matches a target image's histogram to multiple randomly selected reference images from the source domain. This process iteratively adjusts the target image's intensity distribution, promoting a better representation of the source domain.

Global **color normalization** [57] transfers color statistics by globally altering the image histogram, while local color normalization transfers color statistics of specific regions, preserving intensity information within regions of interest. These color normalization methods are commonly used in histopathology images, and these methods have improved the generalizability of a neural network [58]. Kondo *et al*. [27] employed a color normalization method [59] in their architecture for mitosis detection in histopathology images. This color normalization method decomposes the input image into stain density maps and combines them with the stain color basis of a target image. Xiong *et al.* [28] introduced the Enhanced Domain Transformation, a color transformation method to align the color space distributions of seen and unseen data for diabetic retinopathy classification. Pakzad *et al.* [29] introduced a color transformer utilizing StarGAN [60] to diversify clinical skin images by altering skin types while retaining original visual characteristics, enhancing dataset diversity and reducing skin type biases in skin disease classification.

It is worth noting that some **harmonization** techniques, while primarily aimed at reducing non-biological variability across imaging sites or protocols, can be applied in a DG setting. A notable example is the ComBat harmonization [61], which uses an empirical Bayes framework to adjust for site effects while preserving biological variability. More recently, deep learning-based harmonization techniques have emerged. Those include autoencoders with heavy regularization or normalization layers [62], [63], and generative adversarial networks (GANs) capable of scanner-to-scanner translation [64], [65]. Some of those methods, such as adversarial learning with domain classifiers [66], [67] and conditional variational autoencoders [68], showed the capability of deriving scanner-/cohort-invariant features for reconstructing harmonized samples across unseen scanners.

III-A1b Surrogate Methods

Surrogate methods involve using a surrogate representation, such as summary statistics or closed-form mathematical representations, as a substitute for the original input data to improve the generalization performance.

One traditional example is the **frequency-based DG**, which employs Fourier transformation to separate an image into its amplitude and phase components, typically representing style and content, respectively [32]. This is motivated by a well-known property of Fourier transformation that amplitude contains low-level statistics while phase contains high-level semantics [69]. The goal of frequency-based DG is to manipulate the low-level statistics of the amplitude component without significantly varying the high-level semantics of the phase component. These methods are usually well-suited for tasks where high contrast is advantageous, such as fundus imaging [32] or image segmentation tasks [31]. For the white matter hyperintensity segmentation task, Zhao *et al.* [31] creates amplitude prototypes from source domains and learns a calibrating function that reduces the divergence between source and target amplitudes during inference time. Inspired by Mix-Up [70], Xu *et al.* [32] introduces perturbation to the amplitude by interpolating the amplitudes of images from different domains for the fundus image segmentation task. Lie *et al.* [33] proposed an alternative frequency-based DG for fundus image restoration, which uses a Gaussian filter to decompose low-frequency and high-frequency components from an image. Hu *et al.* [34] uses Hessian matrices of an image for retinal vessel segmentation, as vector fields better capture the morphological features and suffer less from covariate shift.

Distribution shifts in medical imaging often arise from image reconstruction processes, which transform raw device data into interpretable images. An alternative is to train **using raw signals**, such as k-space data in MRI and sinogram-space data in CT, to circumvent domain-specific variations introduced by reconstruction algorithms and scanner parameters. Lee *et al.* [71] found that a sinogram-space CNN was about 3% more accurate than an image-space CNN in body part recognition tasks, demonstrating the advantage of using sinogram-space data over CT images. Their findings, along with the potential for radiomics signature analysis on raw data [72], underscore the benefits of leveraging raw image data to bypass reconstruction biases. For example, Zakazov *et al.* [36] proposed a DG method that operates on k-space data for brain segmentation tasks. The proposed method transfers the contrast and structure-related features by swapping the low-frequency areas (*i.e.*, center) of the target k-space data with that of the source k-space data. Zhang *et al.* [37] tackled motion correction in brain MRI by training their model on synthesized motion-corrupted images, which created by introducing motion artifacts into the k-space data.

Dictionary learning [73], or sparse representation learning, can be considered as a type of surrogate method that seeks to find a sparse representation of input data (*i.e.*, the surrogate) as a linear combination of basic elements, capturing common structures while reducing domain-specific variations [74]. Song *et al.* [38] applied this to multi-contrast MRI reconstruction by learning dictionaries that highlight structural similarities. Similarly, Liu *et al.* [39] used dictionary learning for prostate MRI and fundus image segmentation, constructing a shape dictionary with templates to represent diverse segmentation masks efficiently.

2) Data Augmentation

Data augmentation is one of the most prevalent and important forms of DG in MedIA. It refers to techniques that artificially expand and diversify the training dataset by applying various transformations to existing data. The primary goal is to improve the model's ability to generalize across different domains by exposing it to a wider range of data variations during training. Unlike feature-level augmentation (§III-B3a), which modifies the learned feature representations, data-level augmentation directly alters the input data-space.

The widespread adoption and significance of data augmentation in DG for MedIA stem from its effectiveness, relative simplicity, and broad applicability across different tasks and modalities. It serves several crucial purposes:

- Simulating domain shift: By applying transformations that mimic potential variations across different domains (e.g., changes in image intensity, contrast, or noise levels), models can learn to be more robust to these shifts.
- Addressing data scarcity: In medical imaging, where large, diverse datasets are often challenging to obtain, augmentation can help mitigate the limitations of small sample sizes.
- Enhancing model generalizability: By exposing the model to a broader range of data variations, it can learn more robust and generalizable features.

For a comprehensive review of general (non-DG) data augmentation methods for MedIA, readers are referred to the survey by Chlap *et al.* [54]. In the following subsections, we focus on augmentation techniques specifically designed or adapted for domain generalization in medical image analysis.

III-A2a Randomization-based Augmentation

The idea of random augmentation is to generate novel input data by applying random transformations to the original data. Some conventional techniques include randomly applying flipping, rotation, scaling, cropping, adding noise, *etc.*, which are used extensively to improve a model's generalization performance by reducing overfitting [75]. Li *et al.* [40] developed a novel style transfer network that augments a domain by modifying cardiac images with randomly sampled shape and spatial (*i.e.*, slice index) priors to alleviate the modality-level difference for cardiac segmentation. Liu *et al.* [33] proposed a random amplitude mixup method that randomly mixes the amplitudes of different images for DG for fundus image restoration. Billot *et al.* [76] introduced SynthSeg, a novel approach that leverages domain randomization to train a segmentation network on synthetic brain MRI scans with randomized contrasts and resolutions. This method enables the network to generalize to real scans of varying imaging characteristics without retraining. Similarly, Shen *et al.* [77] proposed RandStainNA, which unifies stain normalization and augmentation techniques by randomly generating virtual stain templates.

III-A2b Adversarial-based Augmentation

Adversarial-based data augmentation methods operate on the principle of creating adversarial examples that aim to maximize the model's uncertainty, thereby improving its robustness and generalizability. In this section, we concentrate on data-level adversarial augmentation, while a discussion on model-level adversarial training can be found in Section III-C1c. Tomar *et al.* [43] developed a method that combines knowledge distillation with adversarial-based data augmentation for cross-site medical image segmentation tasks. The process involves the creation of augmented data that is adversarial to the current model, to push the model's feature representations toward the decision boundary. This is achieved by optimizing and sampling data augmentations that simulate data in the uncertain region of the feature space, thereby improving the model's ability to generalize from the training data to unseen test data.

III-A2c Generative Models

Generative models have been widely used for data augmentation in DG tasks. These models learn to generate new data that mirrors the training data distribution, thus providing additional examples for the model to learn from. Scalbert *et al.* [44] designed a new augmentation strategy based on multi-domain image-to-image translation to enhance robustness in unseen target protocols. By adapting the style encoding method [78] based on generative models, they derive a considerable boost of performances for DG at test time. Yamashita *et al.* [45] proposed a style transfer-based augmentation (STRAP) method for a tumor classification task, which applies the style of non-medical images to histopathology images while preserving their semantic content. The authors argue that the style of these images is specific to their domain and irrelevant to their classification, making STRAP effective in learning domain-agnostic representations.

3) Problem-specific Data-level Methods

Problem-specific manipulation methods are tailored to address unique challenges posed by particular types of medical imaging data.

III-A3a Cross-modal Generative Models

Cross-modal generative models represent a pioneering paradigm for achieving DG, wherein models are trained to gain knowledge of the data distribution across diverse modalities (*e.g.*, CT, MRI, X-ray, and PET). These models, often based on GANs, generate synthetic data [40] or suitable latent representations [79], which bridge the distributional gap among cross-modalities. This strategy allows us to provide a model especially capable in medical imaging where data could vary greatly due to patient cohorts, hospital practices, or different imaging modalities. As obvious advantages of such a model, it can be highly valuable when one modality is unavailable for a particular patient or when the model is required to generalize to an unseen domain where a different imaging modality is used. Readers are referred to Xie *et al.* [80] for a comprehensive review on cross-modal neuroimage synthesis.

Taleb *et al.* [47] introduced a self-supervised learning strategy using multimodal jigsaw puzzles for synthesizing cross-modal medical images, where patches from different imaging modalities are assembled to enhance feature extraction across modalities. They further augmented multimodal data volume by generating synthetic images between modalities through a CycleGAN-based translation model. Xu *et al.* [48] proposed an adversarial domain synthesizer for single-source cross-modality image segmentation, employing adversarial training coupled with a mutual information regularizer to maintain semantic consistency between original and synthetic domains. Su *et al.* [49] introduced the Saliency-balancing Location-scale Augmentation (SLAug) for enhancing cross-modal and cross-sequence medical image segmentation. SLAug modifies image distribution with class-specific adjustments and dynamically tunes location-scale weights via model gradients, effectively mitigating domain shifts in medical imaging.

III-A3b Stain normalization

Stain normalization and stain separation techniques are primarily used in histopathology, where different tissue components (*e.g.*, nuclei, cytoplasm, extracellular matrix) are separated based on their staining patterns. This process helps remove staining artifacts and enhances the precision of MedIA tasks, such as cell counting and segmentation. Xu *et al.* [50] proposed a stain normalization method for cell detection in histopathology images. Specifically, the authors address the limitations of stain transformation performed during network training, which may not perfectly represent the stain color of test images. Thus, their approach involves mixing stain colors of target and source domain images and generating multiple transformed test images for better stain representation during testing. Chang *et al.* [51] proposed Stain Mix-Up for the cancer detection task. By decomposing histopathology images into stain color matrices and density maps, the stain mix-up method allows for combining stain colors from different domains. This approach enhances the color diversity in the training data, improving cancer detection performance. The stain mix-up technique can effectively address stain color variations and staining artifacts, providing more accurate and reliable results for histopathology image analysis.

B. Feature-level Generalization

TABLE V
SUMMARY OF FEATURE-LEVEL DOMAIN GENERALIZATION METHODS.

Method	Formulation
Normalization	$\hat{\mathbf{z}} = \frac{f(\mathbf{x})-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\epsilon}}$
Dissimilarity-based	$\min_f D(f(\mathbf{x}^i), f(\mathbf{x}^j)), \quad \forall 1 \leq i \neq j \leq M$
Information theoretic	$\min_f I(\mathbf{z}_{\text{task}}; \mathbf{z}_{\text{domain}})$
Contrastive	$\mathcal{L}(\mathbf{x}_i, \mathbf{x}_j) = -\log \frac{\exp\left(\text{sim}\left(f(\mathbf{x}_i), f(\mathbf{x}_j)\right)/\tau\right)}{\sum_{k \in a(i)} \exp\left(\text{sim}\left(f(\mathbf{x}_i), f(\mathbf{x}_k)\right)/\tau\right)}$
Variational	$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}q\phi(\mathbf{z}\vert\mathbf{x})[\log p_\theta(\mathbf{x}\vert\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}\vert\mathbf{x})\vert\vert p(\mathbf{z}))$
Explicit	$\min_{h,f} \mathbb{E}_{(\mathbf{x},c)\in\mathcal{S}}[\ell(h(f(\mathbf{x})), c)] + \lambda\, \ell_{reg}$

Feature-level generalization methods aim to utilize the domain-invariant features from the input images to improve the generalization performance of a model. These methods often involve learning a feature representation shared across multiple domains, either by training a domain-invariant feature extractor or adapting the feature extractor on the fly during inference. We denote f as a feature mapping function that maps input data to a feature space. The objective function of domain generalization from Eq. 1 can be modified to include a feature extractor $f : \mathcal{X} \to \mathcal{Z}$ and the redefined predictive function $h : \mathcal{Z} \to \mathcal{Y}$:

$$\min_{h,f} \mathbb{E}_{(\mathbf{x},y)}[\mathcal{L}(h(f(\mathbf{x})), y)]. \tag{4}$$

Refer to Table V for a summary of feature-level methods. In the following paragraphs, we explore feature-level domain generalization techniques.

1) Feature Alignment

Feature alignment aims to align or standardize the feature distributions across different domains. These strategies aim to produce domain-invariant features through statistical and structural adjustments, enhancing generalization across varied domains by minimizing distributional discrepancies and aligning feature distributions to a common representation.

III-B1a Feature Normalization

Feature normalization methods aim to statistically center, scale, decorrelate, standardize, or whiten feature distributions across domains and enhance the model's ability to generalize [107]. By transforming all features to the same statistical distribution, normalization prevents features with larger numerical values from dominating those with smaller ones during training, ensuring a more balanced and accurate model. These methods generally stem from the traditional scaling methods, such as z-score and unit vector normalization, as well as some traditional machine learning methods, such as batch and instance normalization. These methods can be formulated as the following generalized equation for feature normalization:

$$\min_{h,f} \mathbb{E}_{(\mathbf{x},y)}\left[\mathcal{L}\left(h\left(\frac{\mathbf{z}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\epsilon}}\right), y\right)\right], \tag{5}$$

where $\mathbf{z} = f(\mathbf{x})$ is the feature embedding, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the statistics of the feature embedding $\mathbf{z}$ (usually the mean and variance), and ϵ is a constant for numerical stability.

Zhou *et al.* [81] proposed a per-domain batch normalization method for medical image segmentation. When testing the model on the target domain, the model compares the distribution information of the target domain with the stored distribution information (mean and variance) from each domain. Then, the model selects the most suitable domain distribution statistics to normalize the activated features from the target domain. Liu *et al.* [82] introduced spectral-spatial normalization (SS-Norm) for retinal vessel segmentation, merging frequency and spatial normalization to isolate domain-invariant features. The approach uses discrete Fourier transformation for frequency normalization and a convolutional network for spatial normalization, improving the representation of spatial details in activation maps.

III-B1b Dissimilarity-based Alignment

Dissimilarity-based alignment methods attempt to reduce the difference between the feature distributions of different domains by minimizing a dissimilarity measure. This aligns the distributions to a common representation, which helps mitigate the domain shift problem. The goal of dissimilarity-based alignment is to find f to minimize the distribution shift among domains in the feature space. For instance, given the i-th and j-th source domains with input samples $\mathbf{x}^i$ and $\mathbf{x}^j$, we may want to minimize the difference between the distributions of their mapped features: $\mathcal{D}(f(\mathbf{x}^i), f(\mathbf{x}^j))$, where $\mathcal{D}(\cdot,\cdot)$ measures the dissimilarity between two distributions, *i.e.*,

$$\min_f \mathcal{D}(f(\mathbf{x}^i), f(\mathbf{x}^j)), \quad 1 \leq i \neq j \leq M. \tag{6}$$

Numerous statistical metrics exist to measure the dissimilarity between distributions, including ℓ_2 distance, f-divergences, and the Wasserstein distance.

Stacke *et al.* [83] empirically evaluated different dissimilarity metrics for tumor classification in cross-site histopathology images. Among various metrics, Wasserstein-based metrics have been shown to better capture the domain shift in cross-site histopathology images. Lyu *et al.* [84] applied a Wasserstein-based metric, specifically the Sinkhorn distance, to measure divergence between augmented domains created through varied image transformations for retinal image segmentation. This approach facilitated the evaluation of domain shift through the divergence of novel distributions induced by different augmentation sub-policies. Similarly, Li *et al.* [85] developed Linear-Dependency Domain Generalization (LDDG) to improve generalization for lesion classification and spinal cord segmentation by aligning latent feature distributions across multiple source domains using Kullback-Leibler (KL) divergence and linear dependency modeling. This approach seeks to reduce empirical risk on unseen target domains, aiming for a theoretical performance upper bound.

TABLE VI
FEATURE-LEVEL DOMAIN GENERALIZATION METHODS. METHODS CATEGORIZED BY DIFFERENT SETTINGS FOR SOURCE AND TARGET DOMAINS (SEE §II-B), TASK, ORGAN, AND MODALITY USED IN EXPERIMENT.

Method	Specific	Ref.	Source	Target	Task	Organ	Modality
Feature Alignment	Feature Normalization	[81]	Single	Sequence, Modality	Segmentation	Brain, Cardiac, Abdominal	MRI, CT
		[82]	Single	Site	Segmentation	Retinal	Fundus
	Dissimilarity-based	[83]	Multiple	Site	Classification	Colon, Breast	Histology
		[84]	Multiple	Modality	Segmentation	Retinal	Fundus, OCT
		[85]	Multiple	Sequence, Site	Classification, Segmentation	Skin, Spinal	Dermatology, MRI
Implicit Disentanglement	Mutual Information	[86]	Multiple	Site	Classification	Abdominal, Brain, Femur, Lips	Fetal Ultrasound
		[87]	Multiple	Site	Classification	Abdominal, Brain, Femur, Lips	Fetal Ultrasound
		[88]	Single	Site	Segmentation	Carotid	Ultrasound
		[89]	Single, Multiple	Sequence	Classification	Blood Cell	Histology
		[48]	Single	Modality, Sequence	Segmantation	Prostate, Abdominal, Cardiac	MRI, CT
	Contrastive	[90]	Multiple	Site	Detection	Breast	X-ray
		[91]	Multiple	Site	Segmentation	Retinal	Fundus
	Variational	[92]	Multiple	Site	Classification	Blood Cell	Histology
		[93]	Multiple	Site	Classification	Breast	Histology
		[94]	Multiple	Site	Classification	Breast	X-ray
		[95]	Multiple	Site	Classification	Breast	X-ray
Explicit Disentanglement	Conditional Representation Learning	[96]	Multiple	Site	Classification	Retinal	Fundus
		[97]	Multiple	Site	Classification	Retinal	Fundus
		[98]	Multiple	Site	Segmentation	Retinal	Fundus
	Feature Reguarlization	[99]	Multiple	Site	Segmentation	Cardiac	MRI
		[100]	Multiple	Site	Compression, classification	Brain, Skin	MRI, Histology
		[101]	Single	Site	Segmentation	Spinal	MRI
Others	Feature Augmentation	[102]	Single, Multiple	Site	Segmentation	Cardiac, Prostate	MRI
		[103]	Multiple	Site	Segmentation	Prostate	MRI
		[104]	Single	Site		Liver	CT
	Kernel-based	[105]	Single	Site	Classification	Brain	MRI
		[106]	Multiple	Site	Detection	Brain	EEG

2) *Disentanglement Methods*

Disentanglement methods aim to decompose an input sample into a feature vector that reveals various factors of variation where each dimension or subset of dimensions carries information linked to a specific factor. The primary goal of these methods is to create a clear boundary between domain-specific and task-specific features. This distinction is crucial in capturing the universal patterns related to the task. Given this goal, the disentanglement process seeks to isolate task-relevant features from those features intrinsic to the domain, *i.e.*, $\mathbf{z} = [\mathbf{z}_{\text{task}}, \mathbf{z}_{\text{domain}}]$, respectively. The goal is to create a model that emphasizes $\mathbf{z}_{\text{task}}$ while effectively ignoring $\mathbf{z}_{\text{domain}}$, thus ensuring that the model's focus is primarily on the features that contribute to the task at hand and less on those that are domain-specific features. To this end, we further refine disentanglement methods into implicit and explicit methods.

III-B2a Implicit Feature Disentanglement

Implicit feature disentanglement strategies learn to decompose factors of variations by utilizing, for example, the statistical properties of the data and indirect incentives to encourage disentanglement. Such approaches provide scalable and flexible techniques for learning disentangled representations. Typical examples of these methods include information-theoretic methods, contrastive learning, and variational inference.

Information theoretic disentanglement methods often focus on using mutual information to separate and understand the different factors of variations in data. Mutual information, denoted by $I(X;Y)$, measures the information obtained from a random variable X by observing another variable Y. The goal of information-theoretic disentanglement is to minimize the mutual information between the task and domain representations, *i.e.*,

$$\min_{f} I(\mathbf{z}_{\text{task}}; \mathbf{z}_{\text{domain}}), \tag{7}$$

where $f(\mathbf{x}) = [\mathbf{z}_{\text{task}}, \mathbf{z}_{\text{domain}}]$ is a feature mapping function that disentangles the input image into $\mathbf{z}_{\text{task}}$ and $\mathbf{z}_{\text{domain}}$. This minimization process plays a vital role in ensuring that the task-related and domain-specific feature sets are *independently* informative. This disentanglement approach seeks to construct a learning model capable of robustly interpreting and classifying data across a spectrum of domains, making it adaptable to a wide range of task-specific challenges in diverse applications.

Specifically, Meng *et al.* [87] proposed MIDNet, an MI-based model specifically designed for fetal ultrasound classification tasks. MIDNet's primary objective is to distinguish domain-invariant features from domain-specific ones by minimizing the mutual information between these feature sets. To achieve this, they employ the Mutual Information Neural Estimation (MINE) [108] approach to approximate the lower bound of the mutual information. This facilitates the extraction of generalizable features and enables knowledge transfer across unseen categorical features in target domains. Similarly, Bi *et al.* [88] proposed MI-SegNet for ultrasound image segmentation. MI-SegNet employs two encoders that separately extract anatomical and domain features from images, and MINE approximation is used to minimize the mutual information between these features. Rather than minimizing the mutual information between domains, Chen *et al.* [89] and Xu *et al.* [48] proposed to maximize the mutual information for maintaining the consistency between the source domain and augmented samples.

Contrastive Disentanglement aims to make representations of similar instances more alike (low contrast) and those of different instances more dissimilar (high contrast). In the context of domain generalization, this approach can be used to learn domain-invariant features by encouraging similarity between samples that share the same task-relevant characteristics, regardless of their domain, while separating samples with different characteristics. A typical contrastive learning loss function [109] is defined as:

$$\mathcal{L}(\mathbf{x}_i, \mathbf{x}_j) = -\log \frac{\exp\left(\text{sim}\left(f(\mathbf{x}_i), f(\mathbf{x}_j)\right)/\tau\right)}{\sum_{k\in a(i)} \exp\left(\text{sim}\left(f(\mathbf{x}_i), f(\mathbf{x}_k)\right)/\tau\right)}, \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ is a function for cosine similarity, $\mathbf{x}_i$ and $\mathbf{x}_j$ are positive pairs, $k \in a(i)$ are the indexes of selected negative samples, and $\mathbf{x}_i$ and $\mathbf{x}_k$ are negative pairs. In the context of domain generalization for medical image analysis, positive pairs could be defined as images showing the same pathology (*e.g.*, two images of malignant tumors) regardless of whether they come from different hospitals or were acquired using different imaging protocols. Negative pairs would be images showing different pathologies (*e.g.*, an image of a malignant tumor paired with an image of healthy tissue), again regardless of their source domain. This pairing strategy encourages the model to learn features that distinguish between pathologies while being invariant to domain-specific characteristics like image acquisition settings or hospital-specific protocols.

Li *et al.* [90] proposed a novel approach that couples multi-style and multi-view contrastive learning to enhance the generalization capability for mammography lesion detection. Specifically, positive pairs for multi-style contrastive learning were synthesized using a GAN, and different views of the breast (*i.e.*, craniocaudal and mediolateral oblique) were used as multi-view contrastive learning. In a similar approach, Gu *et al.* [110] proposed Contrastive Domain Disentanglement and Style Augmentation (CDDSA) for image segmentation in the fundus and MR images. The unique feature of CDDSA is its implementation of a style contrastive loss function, which ensures that style representations from the same domain bear similarity while those from different domains diverge significantly.

Variational disentanglement is a method that utilizes variational autoencoders (VAEs) to learn a disentangled representation. The typical approach for this method involves encoding input data $\mathbf{x}$ into a latent variable $\mathbf{z}$ using an encoding function $q_\phi(\mathbf{z}|\mathbf{x})$. The decoder, $p_\theta(\mathbf{x}|\mathbf{z})$, then reconstructs the original data from the latent representation $\mathbf{z}$. The objective function of VAEs, or the evidence lower bound (ELBO), can be expressed as:

$$\begin{aligned}\mathcal{L}(\theta, \phi; \mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\\ &= \mathcal{L}_{rec} + \lambda\mathcal{L}_{reg},\end{aligned} \tag{9}$$

where $\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ is the K divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$, which is often chosen to be a normal distribution. ELBO can also be interpreted as minimizing the reconstruction error $\mathcal{L}_{rec}$, *i.e.*, the posterior $p_\theta(\mathbf{x}|\mathbf{z})$, and regularizing the approximate posterior $\mathcal{L}_{reg}$, *i.e.*, the KL term. The key idea behind variational disentanglement involves structuring a latent space so that distinct dimensions capture domain-specific and domain-invariant factors. This is typically achieved by introducing tailored constraints or regularization mechanisms during training [111]. For example, regularization or constraints can be incorporated into the ELBO to specifically encourage the separation of domain-specific and domain-invariant factors in the latent space.

Ilse *et al.* [92] proposed the Domain Invariant Variational Autoencoder (DIVA) for malaria cell image classification [112]. DIVA is an extension to the VAE framework that can partition a latent space into three independent latent subspaces for domain label $\mathbf{z}_d$, class label $\mathbf{z}_y$, and residual variations $\mathbf{z}_x$, which captures any residual variations left in data x. This partitioning aims to encourage the model to disentangle these sources of variation. Specifically, DIVA employs three separate encoders that serve as variational posteriors over the three latent variables. In addition to the ELBO term, DIVA formulates classifier-based auxiliary objectives to further encourage the separation of domain-specific and class-specific information into their respective latent variables:

$$\begin{aligned}\mathcal{L}(\theta, \phi; \mathbf{x}) = \ &\mathbb{E}_{q_{\phi_d}(\mathbf{z}_d|\mathbf{x})q_{\phi_x}(\mathbf{z}_x|\mathbf{x}), q_{\phi_y}(\mathbf{z}_y|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_d, \mathbf{z}_x, \mathbf{z}_y)\right]\\ &- \beta KL\left(q_{\phi_d}(\mathbf{z}_d|\mathbf{x})||p_{\theta_d}(\mathbf{z}_d|d)\right) - \beta KL\left(q_{\phi_x}(\mathbf{z}_x|\mathbf{x})||p(\mathbf{z}_x)\right)\\ &- \beta KL\left(q_{\phi_y}(\mathbf{z}_y|\mathbf{x})||p_{\theta_y}(\mathbf{z}_y|y)\right).\end{aligned} \tag{10}$$

Wang *et al.* [93] introduced the Variational Disentanglement Network (VDN) for breast cancer metastasis classification, which separates domain-invariant and domain-specific features by maximizing information gain and posterior probability. Through adversarial training between a task-specific encoder and a feature discriminator, VDN aligns latent features with a predefined prior and employs a generator network for high-quality reconstruction and effective feature disentanglement, enhancing domain generalization. Wang *et al.* [94], [95] propose a variational causal model for the breast cancer classification task. Specifically, they propose a structural causal model that can decompose the latent factors of medical images into domain-agnostic causal features and domain-aware features. These features are factored into a reformulated ELBO term of VAE, and optimizing the modified ELBO provably disentangles the domain-agnostic causal features from domain-aware features.

III-B2b Explicit Feature Disentanglement

There is an explicit mechanism separating task-relevant features from domain-specific features in disentanglement. These methods often involve supervision or hard constraints in the model. Supervision could take the form of domain labels or auxiliary attributes indicating the values of factor of variations for each data instance. Some methods use constraints or regularization terms in the objective function to encourage the model to separate specific factors of variation in the representations. The loss for these types of methods can be in the form of:

$$\min_{h,f} \mathbb{E}_{(\mathbf{x},c)\in\mathcal{S}}[\mathcal{L}(h(f(\mathbf{x})),c)] + \lambda\,\mathcal{L}_{reg}, \tag{11}$$

where c is an auxiliary attribute or a domain label, $\mathcal{L}_{reg}$ is a regularization term that encourages separation between the task-relevant and domain-specific features, and λ is a hyperparameter controlling the strength of this regularization. The first term in this loss refers to model supervision with an auxiliary attribute or a domain label, while the second term encourages the model to keep the task-relevant and domain-specific features separate.

Conditional representation learning refers to learning a representation of the input data influenced by a certain conditioning variable. This variable can be any additional information, such as domain labels or induced priors. Conditional representation learning aims to create representations that are sensitive to the specific aspects of the data relevant to the condition, and invariant or insensitive to other aspects. This can improve performance on tasks where certain aspects of the data are more relevant than others, or where the relevance of different aspects varies under different conditions.

Liu *et al.* [96], [97] proposed the Recursively Conditional Gaussian (RCG) prior for diabetic retinopathy and congenital heart disease diagnosis task. Their proposed method utilizes the ordinal structure of the class labels to construct an appropriate RCG before the class-related latent space. This RCG prior enforces a poset constraint that aligns the extracted latent vectors with the ordinal class labels. By conditioning the latent space on the ordinal labels, the RCG prior aims to learn a representation sensitive to the relevant aspects of the data for the specific diagnosis task, while invariant to other aspects. Wang *et al.* [98] proposed Domain-oriented Feature Embedding (DoFE) for fundus image segmentation, which incorporates a domain knowledge pool to learn the domain prior information extracted from the multi-source domains. This domain prior knowledge is then dynamically enriched with the image features to make the semantic features more discriminative.

Feature regularization methods focus on incorporating regularization terms into the learning objective to guide the model toward extracting meaningful and generalizable features. These methods often utilize penalties that discourage the model from relying too heavily on individual features or encourage the model to maintain certain structures or properties in the learned representations. Additionally, regularization can be used to encourage the model to learn representations invariant to certain transformations of the data, such as translations or rotations. These kinds of regularization can make the learned features more robust to data variations that are irrelevant to the task at hand. For example, this might be done by promoting sparse representations (*e.g.*, dropout [113], ℓ_1, ℓ_2 regularization), where the model is encouraged to use as few features as possible to achieve its task, or by promoting orthogonality, where the model is encouraged to learn features that are independent of each other.

Islam and Glocker [99] proposed Frequency Dropout (FD) for cardiac image segmentation task. FD uses a random feature map filtering approach that works as a form of feature-level regularization during training. In this method, random filters (*e.g.*, Gaussian smoothing, Laplacian of Gaussian, and Gabor filtering) are applied to the feature maps to prevent the neural network from learning frequency-specific image features. Nguyen *et al.* [100] introduced the Adversarially-Regularized Mixed Effects Deep learning (ARMED) for Alzheimer's disease diagnosis and cell image classification tasks. ARMED incorporates a regularization mechanism that enforces the model to learn features invariant to specific clusters in the data. This is achieved by introducing an adversarial classifier that attempts to predict the cluster membership based on the learned features, while the main model is penalized for enabling this prediction. Wang *et al.* [101] proposed Knowledge Distillation for Domain Generalization (KDDG) for MRI gray matter segmentation task. KDDG applies a form of feature-level regularization that encourages the student model's predictions to align with the teacher's predictions, thus improving the student model's robustness and generalization capability.

3) Other Representation Learning Methods

III-B3a Feature Augmentation

Feature augmentation is a technique used to improve machine learning models' generalization capability by transforming the feature space, rather than the input space. Unlike traditional data augmentation, which directly manipulates raw data, feature augmentation operates on the derived features extracted from the raw data. While data augmentation creates a more comprehensive and diverse source domain by introducing variations at the data level, it is limited by the extent and variety of feasible and meaningful transformations on the raw data. On the other hand, by working directly in the feature space, feature augmentation allows for a richer set of transformations. Feature augmentation can also incorporate domain knowledge more effectively, as transformations can be designed to specifically target and vary important features.

Chen *et al.* [102] proposed a novel feature augmentation framework, MaxStyle, for cardiac MRI segmentation. MaxStyle introduces adversarial noise into the feature styles and conducts a worst-case style composition search through adversarial training. This approach broadens the range of augmented styles and makes the model more robust by exposing it to harder cases. Zhou and Konukoglu [103] proposed a Federated Feature Augmentation (FedFA) for cross-site prostate MRI segmentation. FedFA augments the features by estimating a vicinity distribution at each layer of the neural network during training, thus enhancing the data representation at each client. It manipulates the channel-wise statistics of the features, such as the mean and standard deviation, which often carry significant domain-specific information.

III-B3b Kernel-based Learning

Kernel-based methods are a classic and effective approach within feature-level domain generalization. These methods improve generalization by mapping the original input features into a higher dimensional space. This mapping offers several advantages for domain generalization. In the higher-dimensional space, kernel methods can potentially reveal domain-invariant structures that are not apparent in the original feature space. Kernel methods can also model non-linear relationships in the data, which is particularly useful for capturing intricate patterns in medical images that may be consistent across domains. There are various kernel-based methods for feature-level domain generalization, including Support Vector Machine (SVM) variants, Maximum Mean Discrepancy (MMD), and Transfer Component Analysis (TCA). Kernel trick enables these methods to operate in high-dimensional spaces without explicitly calculating the coordinates of the data in that space, but by simply computing the dot products between the images of all data pairs in the feature space. This makes the calculations more tractable and efficient. These kernel-based methods can benefit medical image analysis as they can handle high-dimensional data and discover complex patterns. They also offer an excellent way to incorporate domain knowledge, such as spatial relationships in images, by defining appropriate kernels.

Wang *et al.* [105] proposed a kernel-based binary classifier for cross-site brain disease diagnosis tasks. In the kernel setting, we can reformulate the regularization term as:

$$\mathcal{L}_{reg} = \frac{||f(\mathbf{x})||^2}{2} = \frac{1}{2}\sum_i k(\mathbf{x}^i, \mathbf{x}), \tag{12}$$

where the norm is the Reproducing kernel Hilbert space (RKHS) norm, and $k(\cdot,\cdot)$ is the kernel function that measures the similarity between two variables. The RKHS norm captures the classifier's complexity or "smoothness" within the chosen kernel space. The authors use this kernel-based classifier to measure the disharmony and utilize it to improve the generalizability of the given model. Ayodele *et al.* [106] proposed a multi-TCA approach for epileptic seizure detection using an EEG dataset. In contrast to utilizing the disharmony [105], the authors use the RKHS norm to measure the shared subspace between source domains. Then, they utilize various dimension reduction techniques to extract a generalized feature vector for a recurrent neural network.

C. Model-level Generalization

Model-level generalization focuses on enhancing the intrinsic ability of machine learning models to generalize across domains by modifying core aspects of the model itself, including the learning process, model architecture, and optimization techniques. Specifically, such strategies encompass several categories of methods: a) *Learning strategy*, which focuses on adequately reflecting the target-suitable knowledge or leveraging distinct representations gained from a variety of subtasks; b) *Model framework*, which exploits modifications to the network architecture or the incorporation of adaptive auxiliary components to more efficiently address the domain shift; and lastly c) *Other model-based DG*, which involve various optimization and adaptation techniques.

1) Learning Strategy

Methods in this category concentrate on harnessing the general learning strategy to enhance the model's generalizability, which mainly involves various techniques as a) *Meta-learning*, wherein the model learns how to rapidly adapt to new tasks, thereby improving its flexibility and generalization capacity; b) *Self-supervised learning*, which is an unsupervised manner that can leverage large amounts of unlabeled data by creating pretext tasks; c) *Adversarial learning*, which strives to minimize the divergence between different domains to enhance the model's transferability.

III-C1a Meta-learning

Meta-learning techniques are closely relevant in medical imaging due to the prevalent scarcity of annotated data coupled with the need to rapidly adapt to unseen data domains. Specifically, a model employing such strategies aims to learn an optimal initialization or update rule that can be quickly fine-tuned to perform well in unseen data domains. By virtue of these advantages, it is possible to improve the model's flexibility and the efficiency of its generalization capabilities. To simulate domain shift, meta-learning methods divide the source domains into meta-training and meta-test sets. Meta-learning can be formulated as follows:

$$\begin{aligned}\phi^* &= \arg\min_{\phi} \mathcal{L}_{meta}(\phi; \mathcal{S}_{mtrain}),\\ \theta^* &= \arg\min_{\theta} \mathcal{L}_{task}(\theta; \mathcal{S}_{mtest}, \phi),\end{aligned} \tag{13}$$

where ϕ^* denotes the meta-learned parameters optimized on the meta-training set $\mathcal{S}_{mtrain}$, which are then used to initialize the task-specific parameters θ^* optimized on the meta-test set $\mathcal{S}_{mtest}$. $\mathcal{L}_{meta}$ and $\mathcal{L}_{task}$ are the meta-learning and task-specific loss functions, respectively. This formulation is inspired by the Model-Agnostic Meta-Learning (MAML) algorithm [143], where the meta-objective is to find an initialization that allows for quick adaptation to new tasks.

TABLE VII
Model-level domain generalization methods. Methods categorized by different settings for source and target domains, task, organ, and modality.

Method	Specific	Ref.	Source	Target	Task	Organ	Modality
Learning Strategy	Meta-learning	[114]	Multi	Site	Segmentation	Spinal	CT
		[115]	Multiple	Site	Segmentation	Brain	MRI
		[116]	Multiple	Site	Segmentation	Prostate	MRI
		[117]	Multiple	Site, Sequence	Segmentation	Cardiac, Spinal	MRI
		[118]	Multi	Site	Segmentation	Retinal	Fundus, OCT, FC
		[119]	Single, Multiple	Site	Classification	Brain	Functional MRI
	Self-supervised Learning	[110]	Multiple	Site	Segmentation	Retinal	Fundus
		[91]	Multiple	Site	Segmentation	Retinal	Fundus
		[120]	Multiple	Modality	Segmentation	Abdominal, Cardiac	CT, MRI
		[121]	Multiple	Site	Classification	Skin, Retinal, Chest, Breast	Histology, Fundus, X-ray, Mammography
		[122]	Single	Site	Segmentation	Retinal	Fundus
	Adversarial Learning	[123]	Single, Multiple	Sequence	Detection	Skin, Hip	MRI
		[124]	Multiple	Site	Segmentation	Cardiac	MRI
		[125]	Multiple	Site	Segmentation	Retinal, Prostate	Fundus, MRI
		[126]	Single	Site	Classification	Breast	Histology
Model Framework	Ensemble Learning	[127]	Multiple	Sequence, Site	Segmentation	Brain	MRI
		[128]	Single	Sequence	Localization	Surgical Scene	Video Frames
		[44]	Single, Multiple	Site	Classification	Colon, Breast	Histology
	Model Distillation	[101]	Single	Site	Segmentation	Spinal	MRI
		[129]	Multi	Site	Detection	Breast	Histology
		[130]	Single, Multi	Site	Classification	Retinal	Fundus
		[131]	Single	Site	Segmentation	Prostate	MRI
		[132]	Single	Site	Segmentation	Prostate	MRI
		[133]	Multi	Site	Segmentation	Abdominal, Prostate, Surgical Scene	CT, MRI, Video
	Distributed Learning	[134]	Multiple	Site	Segmentation	Retinal, Prostate	Fundus, MRI
		[135]	Multiple	Site	Classification	Breast	Histology
		[136]	Multiple	Site	Segmentation	Retinal	Fundus
		[137]	Multiple	Modality	Segmentation	Tumor	MRI
Other	Geometric Learning	[138]	Multiple	Modal, Sequence, Site	Classification, segmentation, detection	Multiple organs, modalities from 55 datasets	
		[139]	Multiple	Site	Segmentation	Surgical Scene	Video Frames
		[140]	Single	Site	Segmentation	Abdominal, Cardiac, Prostate	MRI, CT
	Distributionally Robust Optimization	[141]	Multiple	Site	Classification	Skin	Dermatology
		[142]	Multiple	Site	Classification	Skin	Dermatology

Khandelwal and Yushkevich *et al.* [114] extended the Meta-learning for Domain Generalization [144] for the CT vertebrae segmentation task (MLDG-Seg). The key idea behind MLDG-Seg is to simulate the domain shift during the training process by artificially creating a meta-test set $\mathcal{S}_{mtest}$ from multiple source domains and then training the model in a way that optimizes its performance across these varied domains or tasks. Dou *et al.* [115] proposed Model-agnostic learning of Semantic Features (MASF) for the cross-site brain MRI segmentation task. MASF employs a meta-learning algorithm that enhances generalization to unseen domains by globally aligning class relationships and locally clustering class-specific features, optimizing semantic feature representations. This approach updates model parameters for improved accuracy in source domains during meta-training, and enforces semantically relevant learning through global and local mechanisms during meta-testing. Liu *et al.* [116] proposed a Shape-aware Meta-learning (SAML) approach for the prostate MRI segmentation task. SAML introduces two loss functions specifically designed to improve the compactness and smoothness of segmentation in the presence of domain shift. The compactness loss function encourages segmentations to preserve the complete shape of the prostate, while the smoothness loss function enhances boundary delineation by promoting intra-class cohesion and inter-class separation between contour-relevant and background-relevant embeddings across different domains. Lie *et al.* [117] proposed a semi-supervised meta-learning approach for domain generalization in medical image segmentation tasks. Specifically, they split their training dataset into meta-train and meta-test sets, including labeled and unlabeled data, enabling their model to generalize to unseen domains. Hu *et al.* [118] proposed Meta-Learning on Anatomy-Consistent Pseudo-Modalities (MAP) for the retinal vessel segmentation tasks. MAP employs a mixup technique with episodic training on synthesized pseudo-modalities to emphasize structural vessel features, achieving improved generalization across different imaging domains.

III-C1b Self-supervised Learning

Self-supervised learning (SSL) is a novel learning paradigm where the model is trained to figure out a *pretext* task that learns general but useful feature representations from unlabeled large-scale data. Specifically, the principal idea behind SSL is to design a proxy where the answers can be deduced by a portion of the input data, enabling the model to learn representations under its own supervision. Thanks to such an advantage, creating the pretext task can alleviate the chronic issues induced by a scarcity of annotated data, especially in medical imaging. Further fine-tuning the downstream task via these universally useful features improves the generalization capability, allowing the model to adequately escape overfitting for domain-specific biases. A typical example of SSL is the contrastive learning paradigm introduced in Eq. 8.

Gu *et al.* [110] proposed a contrastive domain disentanglement and style augmentation for domain generalization. In particular, domain-style contrastive learning is to properly decompose an image into domain-invariant representation and domain-specific modality representation (*i.e.*, style code), whereas a style augmentation strategy enhances generalizability by combining the randomly generated style codes with given anatomical representation to reconstruct new styles' images. Meanwhile, Ouyang *et al.* [120] devised a superpixel-based SSL with details in pseudo-label generation for few-shot semantic segmentation. By further designing the adaptive local prototype module, they prevent the local information of each class such that it achieves outstanding segmentation performance while improving generalizability. Azizi *et al.* [121] combines large-scale supervised transfer learning on natural images and intermediate contrastive learning on medical images for specific downstream medical-imaging ML tasks, thereby enhancing the data-efficient generalization performance.

An emerging and powerful strategy in addressing domain shifts is the combination of pretraining and self-supervised adaptation. This approach leverages the benefits of both large-scale pretraining and task-specific fine-tuning to enhance model generalization. A notable example is FINE (Feature-level Instance Normalization and Exchange) method proposed by Zhang *et al.* [145], which incorporates the physical model of data generation into the adaptation process. FINE updates the weights of a pretrained network by minimizing a data fidelity loss for each test case, allowing it to better capture features specific to the target domain while maintaining physical consistency. Similarly, Zhao *et al.* [146] introduced Synthetic Multi-Orientation Resolution Enhancement (SMORE). SMORE is a self-supervised technique for super-resolution and anti-aliasing of MRI images that does not require external training data. It works by training a network on high-resolution in-plane slices and applying it to low-resolution through-plane slices to enhance image quality.

III-C1c Adversarial Learning

Adversarial learning is widely used for learning domain invariant features in machine learning. The key idea of adversarial learning is to introduce adversarial examples during training to make the model more robust to potential attacks or unexpected inputs. These adversarial examples are usually generated by applying minute perturbations to the original input data to deceive the model into making incorrect predictions. By incorporating such adversarial examples, the model can better handle real-world scenarios where it may encounter unseen domains, enhancing its ability to make accurate and reliable diagnoses.

Bekkouch *et al.* [123] proposed the adversarial reconstruction loss to force an encoder to forget style information while extracting useful classification features for hip MRI landmark detection. Chen *et al.* [124] introduces a realistic adversarial intensity transformation model for data augmentation in MRI that simulates intensity inhomogeneities, common artifacts in MR imaging. This method is a simple yet effective framework based on adversarial training to learn adversarial transformations and to regularize the network for segmentation robustness, which can be used as a plug-in module in general segmentation networks. Zhang *et al.* [125] proposed an adversarial intensity attack method for medical image segmentation, which exploits an adversarial attack strategy to adjust the intensity distribution in images without altering their content.

2) Model Framework

Model framework dives into the architectural and structural strategies deployed to tackle the pervasive challenge of domain shift. Within this framework, three pivotal approaches are discussed: Ensemble learning, model distillation, and distributed learning. Together, these strategies represent a comprehensive framework aimed at improving the generalizability of models through innovative architectural solutions and privacy-preserving techniques, ultimately aiming to bridge the gap between diverse medical imaging domains while safeguarding patient privacy.

III-C2a Ensemble Learning

Ensemble learning methods are a fundamental approach in machine learning that can significantly enhance model generalization. The key idea behind ensemble models is to build a predictive model by combining the predictions of several base models trained on different subsets of data or using different network architectures. The diverse models can capture varying aspects of unique patterns and feature representation, so their combination could lead to more robust predictions. In particular, ensemble learning empowers medical imaging systems to achieve robustness and generalization in medical imaging, ultimately contributing to enhanced clinical decision-making and patient care.

Kamraoui *et al.* [127] proposed the Mixture of Calibrated Networks (MCN) for brain tumor segmentation. The proposed MCN utilizes the complementarity of different base models and takes advantage of their strengths, thus improving the overall system performance. Specifically, MCN combines the predictions from multiple base models, each with unique calibration characteristics, to deliver more precise tumor boundary definitions and more accurate segmentation results. Philipp *et al.* [128] proposed a dynamic CNN for surgical instrument localization, which fuses image and optical flow modalities so that the most reliable information contributes to the prediction. Scalbert *et al.* [44] introduces an ensemble strategy based on multi-domain image-to-image translation for various classification tasks using histology images. Specifically, the proposed method performs image-to-image translation by projecting the target image to the source domains and then ensembles the model prediction of these projected images.

III-C2b Model Distillation

Model distillation involves transferring the knowledge from a large, sophisticated *teacher* model to a more compact and efficient *student* model. This process not only preserves the intricate insights and performance capabilities of the teacher model but also ensures that the student model remains lightweight and practical for deployment in environments with stringent computational or memory constraints.

Wang *et al.* [101] proposed Knowledge Distillation for Domain Generalization (KDDG) for the spinal cord gray matter segmentation task. The authors propose a training strategy that utilizes a gradient filter as a novel regularization term, aiming to simplify the learning task and thereby improve the generalization performance of the model. The paper articulates that the ”richer dark knowledge“ [147] derived from the teacher network, along with the proposed gradient filter, can significantly mitigate the learning challenge, leading to better generalization in various tasks. Fernandez-Martín *et al.* [129] proposed Uninformed Teacher-Student (UTS) for the mitosis localization task, employing a method that distills “hard” samples by training a teacher model to identify and retain only the clean, closely matched predictions to annotated mitoses, thus creating a purified training subset. This subset is used to train a student model, incorporating strong image transformations to challenge and refine the model’s focus, enhancing its ability to generalize by learning from a distilled dataset that minimizes noise and irrelevant variability.

III-C2c Distributed learning

Distributed learning techniques, such as *federated learning* and *privacy preservation*, are essential in medical domain generalization due to patient information’s sensitivity in exploiting the data from various institutions [148]. When we have used or shared the data from the decentralized device or different data server, privacy concerns may arise in model updates that could leak patients’ information [136]. Accordingly, it may violate social ethics that are accompanied by potential risks. To alleviate this fatal issue, advanced techniques such as differential privacy [149] and federated learning [150] allow models to learn a wide range of data from different institutions or hospitals without directly accessing it while preserving data privacy.

To secure sensitive patient information, Liu *et al.* [134] proposed a privacy-preserving solution with a boundary-oriented episodic learning scheme, which allows us to aggregate model updates from multiple clients without revealing any individual client’s data or compromising their privacy. Chen *et al.* [135] designed cross-client style transfer using style vectors to improve performance in domain generalization while preserving privacy in federated learning. Meanwhile, Xu *et al.* [136] proposed Federated Adversarial Domain Hallucination (FADH), which encodes the information of multiple domains through weight aggregation, as a surrogate for the domain classifier. Using differential privacy, Li *et al.* [137] brings sparse vector technique to the patient data owners and only shares intermediate model training updates among them, thus preserving patient data privacy. By doing so, these approaches ensure the surveillance and security of sensitive patient information while enabling the incorporation of diverse datasets into the learning process, thereby promoting model generalization.

3) Other Model-based DG

III-C3a Geometric learning

Geometric learning [151] is an approach that leverages the intrinsic geometric structure of data, often residing in non-Euclidean spaces. Non-Euclidean spaces refer to geometric environments that do not adhere to Euclidean geometry, *e.g.*, graphs, topologies, and manifolds, often encountered in MedIA. Here, geometric learning harnesses these intrinsic data geometries, exploiting the geometric information to better generalize across different domains. By modeling the complex correlations in high-dimensional data, geometric learning can better handle irregularities inherent in medical imaging. Readers are referred to [20] for a comprehensive review of the OOD generalization on graphs. The following paragraph explores the geometric learning techniques specifically proposed for MedIA tasks.

Nguyen *et al.* [138] developed a graph-matching algorithm employing a k-nearest neighbors approach to construct graphs from medical images, where nodes represent distinctive image regions and edges define spatial relationships. Their Learned Vertex Matching (LVM) method analyzes structural similarities and differences across images, enhancing abnormality detection and segmentation tasks. In a similar approach, Seenivasan *et al.* [139] proposed a graph network for surgical scene understanding. They utilized graph learning to understand interactions in the surgical scene by embedding the visual and semantic features of instruments and tissues into graph nodes. Santhirasekaram *et al.* [140] proposed a hierarchical topology preservation method for medical image segmentation tasks. Their method constrains the latent space of a deep learning model to a dictionary of base components, which are chosen to capture the limited structural variability found across patients' medical images. This dictionary is learned through vector quantization, and a topological prior is incorporated into the sampling process using persistent homology, which ensures topologically accurate segmentation maps.

III-C3b Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) [161] is a model-level domain generalization method aiming to optimize model performance over the worst-case distribution within a specified uncertainty set. In other words, instead of optimizing the model's performance based on a single training data distribution, DRO tries to ensure good performance across a range of possible data distributions.

Bissoto *et al.* [141] utilized the Group Distributionally Robust Optimization (GDRO) [162] in their skin lesion classification model. GDRO extends the DRO framework by considering groups, or "environments", in the data distribution. They partitioned the training data into different environments based on the presence of various artifacts, such as hair, ruler marks, and dark corners. These environments were then used to train the model under the GDRO framework. Goel *et al.* [142] enhanced Generalized Distributionally Robust Optimization (GDRO) by introducing class-conditional Subgroup DRO (SGDRO) for skin lesion classification, which refines risk minimization by considering both broad data groups and more granular subgroups defined by class-specific traits. SGDRO optimizes for the worst-case scenario within each subgroup across different environments, resulting in a model that is better equipped to handle complex data distributions and more resistant to distributional shifts.

D. Analysis-level Generalization

Analysis-level DG refers to techniques that focus on understanding and interpreting the behavior of domain-generalized models. These methods aim to: (1) provide insights into how models make decisions across different domains, (2) evaluate the extent of a model's generalization, (3) identify potential biases or failure modes in generalized models, and (4) enable trust and adoption of DG models in clinical settings. Unlike data-level, feature-level, or model-level approaches that primarily aim to improve generalization performance, analysis-level methods are concerned with the post-hoc examination and interpretation of already generalized models. This is particularly crucial in MedIA, where understanding model decisions is essential for clinical validation and trust.

Analysis-level DG techniques face unique challenges, as they must provide interpretations that are consistent and meaningful across multiple domains, often with varying characteristics. These methods must balance the trade-off between model performance and interpretability, especially for complex, highly generalizable models.

1) Interpretable AI

Interpretable AI aims to develop techniques that help evaluate and debug a model's decisions making process. Interpreting domain generalization models is more challenging as these models have special architectures and learning paradigms to accommodate the novel DG settings (*e.g.*, cross-modality). Hence, interpretable AI for DG proposes new techniques to visualize the model's output given heterogeneous data, such as multi-modal [153] and temporal [154] data. In MedIA, AI's ability to adapt to new, unseen data from various hospitals or demographic backgrounds is crucial for diagnosing and determining treatment paths accurately. Interpretable AI, especially under DG setting, is essential as it allows healthcare professionals to understand and trust AI's decisions, thereby enhancing patient care and safety through transparency and clinical evidence validation. For a general overview of interpretable AI for MedIA, readers are referred to the survey by Singh *et al.* [163] and Van *et al.* [164]. In the following paragraphs, we present domain generalization techniques for interpretable AI specifically designed for MedIA.

Dong *et al.* [152] proposed a saliency map-based method for lung lesion classification that uses a contrastive learning scheme incorporating synthetic causal interventions. This technique utilizes weighted backpropagation to generate a saliency map that visualizes and highlights causally relevant areas in the data, thereby improving our understanding of the model's decision-making process. Karim *et al.* [153] proposed DeepKneeExplainer, a CAM-based interpretable AI method for multimodal knee osteoarthritis diagnosis. The DeepKneeExplainer uses an explainable neural ensemble method to improve performance by implicitly reducing the generalization error and using CAM to visualize the model's decision. Similarly, Wang *et al.* [154] proposed a novel focal domain generalization loss and used Grad-CAM++ [165] to visualize the pathological activity from stereo-electroencephalogram (sEEG).

III-D1a Transferability

Methods that provide interpretability in one domain may not necessarily transfer well to other domains. Since different medical imaging data and tasks may require different interpretability approaches, ensuring that interpretability methods can be effectively applied across diverse domains is challenging. Gao *et al.* [156] proposed BayeSeg for interpretable medical image segmentation. One of the key advantages of BayeSeg is its ability to control the performance and interpretability

TABLE VIII
ANALYSIS-LEVEL DOMAIN GENERALIZATION METHODS. METHODS CATEGORIZED BY DIFFERENT SETTINGS FOR SOURCE AND TARGET DOMAINS (SEE §II-B), TASK, ORGAN, AND MODALITY USED IN EXPERIMENT.

Method	Ref.	Source	Target	Task	Organ	Modality
Interpretable AI	[152]	Single	Site	Classification, Segmentation	Retinal, Cardiac	OCT, CT
	[153]	Single	Site	Classification, Segmentation	Knee	MRI
	[154]	Multiple	Site	Classification	Brain	EEG
	[155]	Multiple	Sequence	Detection	Gastrointestinal	Endoscopy
Transferability	[156]	Single	Sequence, Site	Segmentation	Cardiac, Prostate	MRI, CT
	[157]	Single, Multiple	Site	Classification	Breast	Histology
Causality	[158]	Single, Multiple	Sequence	Classification	Chest	X-ray
	[159]	Multiple	Site	Classification	Brain	MRI
	[160]	Single	Modality, Sequence, Site	Segmentation	Prostate, Abdominal, Cardiac	MRI, CT

tradeoff. By approximating the posterior distributions of the shape, appearance, and segmentation, BayeSeg captures the statistical relationships between these variables. This statistical modeling allows users to adjust the weights of the variational loss terms in BayeSeg to prioritize different aspects of the segmentation process, allowing them to control the tradeoff between interpretability and performance. Yuan *et al.* [157] proposed a method for augmenting histopathology images using text prompts (*e.g.*, “synthesize image of a lymph node in the style of S*”). To tackle the challenge of transferability, authors proposed to leverage text-to-image (T2I) generators as a means of enabling interpretable interventions for robust representations. The authors argue that T2I generators offer unprecedented capability and flexibility in approximating image interventions conditioned on natural language prompts. By using T2I generators, the proposed method can provide a more interpretable and domain-agnostic approach that can be effectively applied across diverse domains.

2) *Causalility*

Causality refers to the relationship between variables in a causal system, where one variable (the cause) directly affects or influences another variable (the effect). In domain generalization, causality focuses on understanding the underlying causal mechanisms that lead to the differences between source and target domains. It aims to identify the causal factors invariant across different domains and responsible for the targeted MedIA task. By understanding and leveraging causality, domain generalization methods can effectively generalize the learned knowledge from a source domain to target domains with different distributions. Readers are referred to a survey by Seth *et al.* [19] for a deeper insight into the causal perspective of domain generalization for general tasks. In the following paragraph, we explore several approaches of causal learning specifically designed for domain generalization for MedIA tasks.

Mahajan *et al.* [158] proposed a causality-aware domain generalization method for pneumonia detection using chest X-ray images. They used a causal Bayesian network to model the relationships among the domain, the image features, and the class label. By explicitly modeling the causal relationships, they were able to identify the common causal features that are invariant across domains and are important for predicting the presence of pneumonia. Wang *et al.* [159] used a causal graph-based approach for Alzheimer’s disease diagnosis using MRI. They modeled the causal relationships among imaging sites, gender, age, and imaging features using a Structural Causal Model (SCM). By performing counterfactual inference on the model, they could generate harmonized data that simulate the imaging data as if it came from the same site. This approach effectively removed the site-specific confounding factors and improved the generalization of the trained model across different sites. Similarly, Ouyang *et al.* [160] proposed a causal learning framework for single-source domain generalization in CT image segmentation. They introduced a SCM to represent the causal relationships between the input data, the domain shift variables, and the task-specific output. The SCM allows for the identification of invariant causal factors shared across different domains, which can be used to improve the generalization of the models.

IV. DG UNDER LIMITED SOURCE

TABLE IX
COMPARISON OF SOURCE-LIMTED DG

DG Paradigm	Access to Source Domain
Multi-source DG	Full access (*i.e.*, $\{\mathcal{S}^i\}_{i=1}^M$)
Single-source DG	Single source domain (*i.e.*, $\mathcal{S}^1$)
Unsupervised DG	Unlabeled source domain
Open-set DG	With concept (label) shift
Source-free DG	Pretrained source model only
Zero-shot DG	Auxiliary information only

In this section, we examine the challenge of domain generalization under severe restrictions in the source domain, as detailed in Table IX. These particular DG scenarios have received relatively less attention in the field of MedIA due to their extreme conditions. Our review aims to shed light on these under-explored areas and their implications for MedIA.

A. Single-source Domain Generalization

Single-source domain generalization (SSDG) assumes that there is only one source domain to learn from. Due to the lack of diversity in the training data, most SSDG propose augmentation-based solutions, both in data and feature space, to simulate a broader range of domain variability. For example, augmentation-based SSDG inlcude color normalization [30] (§III-A1a), frequency-based [35] (§III-A1b), randomization [41], [42] (§III-A2a), generative models [46] (§III-A2c), feature augmentation [104] (§III-B3a), contrastive learning [122] (§III-C1b), adversarial learning [48], [126] (§III-C1c). Recent *bleeding-edge* SSDG include transferring knowledge from large-scale pretrained models, *i.e.* variations of model distillation [130], [131], [132] (§III-C2b).

B. Open-set Domain Generalization

TABLE X
OPEN-SET DOMAIN GENERALIZATION METHODS. METHODS CATEGORIZED BY DIFFERENT SETTINGS FOR SOURCE AND TARGET DOMAINS (SEE §II-B), TASK, ORGAN, AND MODALITY USED IN EXPERIMENT.

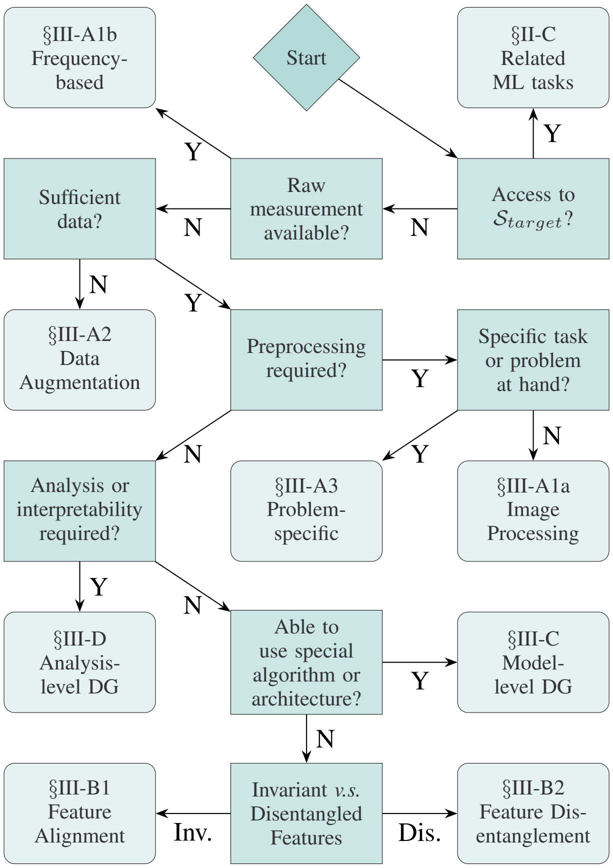
Method	Ref.	Source	Target	Task	Organ	Modality
Open-set DG	[166]	Single	Site	Classification	Chest	X-ray
	[167]	Multiple	Site	Classification	Chest	X-ray
	[168]	Single	Site	Classification	Chest	X-ray
	[169]	Single	Site	Classification	Chest	X-ray
	[170]	Single	Site	Classification	Chest	X-ray
	[171]	Multiple	Site	Classification	Breast	Histology

Open-set domain generalization (OSDG) refers to DG techniques that specialize in capturing and correcting the concept shift in addition to the covariate shift (§II-B3). Yang *et al*. [166] proposed a simple feature-based semantics score function to consider both detecting label shift and being tolerant to covariate shift as in-distribution. Mahajan *et al*. [167] investigated the theoretical relationship of whether better OOD generalization leads to better privacy for ML models in practice and showed that capturing stable features from models represents superior open-set generalization with robustness. Zheng *et al*. [168] proposed Open-Set Single-Domain Generalization for Multiple Cross-Matching (MCM) for the open-set lung cancer diagnosis. This work delves into an open-set single-source DG problem where the source domain only contains data with unique class names, while the target domain contains multiple unseen class names. Puli *et al*. [169], [170] and Gao *et al*. [171] deal with the spurious correlation or variations underlying several confounding variables in terms of causal perspective to circumvent the open-set problem.

C. Other under-explored DGs

In this section, we introduce some under-explored DGs where only preliminary research has been done in the field of MedIA.

Fig. 5. Problem-specific suggestion for strategies for integrating domain generalization into MedIA workflow. Diamond box indicates the start terminator, angled boxes indicate the process, and round boxes indicate the decision.


§III-A1b Frequency-based
Start
§II-C Related ML tasks
Y
Y
Sufficient data?
N
Raw measurement available?
N
Access to $\mathcal{S}_{target}$?
N
Y
§III-A2 Data Augmentation
Preprocessing required?
Y
Specific task or problem at hand?
N
Y
N
Analysis or interpretability required?
§III-A3 Problem-specific
§III-A1a Image Processing
Y
N
§III-D Analysis-level DG
Able to use special algorithm or architecture?
Y
§III-C Model-level DG
N
§III-B1 Feature Alignment
Inv.
Invariant *v.s.* Disentangled Features
Dis.
§III-B2 Feature Dis-entanglement

Unsupervised domain generalization refers to methods that enable a model to learn useful, domain-invariant features from unlabeled source data such that it can perform well on unseen domains. **Source-free domain generalization** places an extreme constraint on privacy-preserving models (§III-C2c) where it assumes source data is inaccessible, but only the pretrained source model is available. **Zero-shot domain generalization** is another extreme case of DG where only the auxiliary information (*e.g*., meta-data) from the source domain is available. A possible solution [130], [131], [132], [133] (§III-C2b) to these extreme cases of DGs involves a pretrained foundation model combined with zero-shot prompting techniques, though these topics fall beyond the scope of this survey due to the lack of literature in this topic.

V. FUTURE DIRECTIONS

Domain generalization for MedIA is a rapidly evolving field with several promising directions for future research. In this section, we outline some important areas that warrant further exploration.

A. Source-limited Domain Generalization

Unlike the common focus on scenarios involving multiple sources, there is a significant gap in research for methods tailored to domain generalization under limited source (§IV). For example, single-source domain generalization requires the

TABLE XI
CRITICAL ANALYSIS OF DOMAIN GENERALIZATION METHODS

Level	Method	Specific	Strengths	Limitations
Data	Image Process.	Intensity Normalization	Uniformity across images; improved ML performance	May reduce contrast; relies on similarity assumptions
		Histogram Matching	Adapts to reference histograms improving consistency	Depends on reference choice; less suitable for multi-domain
		Color Normalization	Standardizes color for consistency and feature recognition	Could alter diagnostic features; mostly for stain images
	Surrogate	Frequency-based	Separates amplitude/phase for style/content manipulation	Sensitive to noise; limited by domain applicability
		Raw Signals	Manages early-stage data manipulation; captures latent info	Complexity/accessibility issues; limited applicability
		Dictionary Learning	Captures sparse representation and structural similarities	Computationally intense; may not handle unique features
	Augment	Randomization	Increases data diversity and simplicity	Limited augmentation scope; may create unrealistic images
		Adversarial	Improves robustness through targeted examples	Computationally intense; risks catastrophic forgetting
		Generative	Generates diverse and realistic data	Complex to train; prone to modal-collapse
	Problem-specific	Cross-modal	Enhances unpaired data utilization	Complex and relies on synthetic data quality
		Stain Normalization	Enhances histopathological analysis	Limited applicability; risks of over-normalization
Feature	Feature Align.	Feature Normalization	Standardizes statistical distributions efficiently	Depends on choice of domain-specific statistics
		Dissimilarity-based	Mitigates domain shift through direct measures	Computationally complex and challenges in metric selection
	Implicit Disent.	Information Theoretic	Enhances interpretability and adapts to complex structures	Difficult to estimate MI; depends on estimation quality
		Contrastive	Improves sample efficiency; robustness to domain shifts	Depends on pair quality; risks representational collapse
		Variational	Models uncertainty; direct latent space regularization	Balance between fidelity and disentanglement
	Explicit Disent.	Conditional Representation	Enhances contextual adaptation; targeted feature learning	Complexity leading to potential overfitting on condition
		Feature Regularization	Robustness to variations; prevents over-reliance on features	Sensitive to choice of hyperparameters
	Others	Feature Augmentation	Incorporates domain knowledge within latent space	Complex and can overfit to augmented features
		Kernel-based Learning	Incorporates domain knowledge efficiently	Selection of kernel is critical; scalability issues
Model	Learning Strategy	Meta-learning	Enables rapid adaptation and efficient learning	Faces complex optimization; overfitting to meta-tasks
		Self-supervised Learning	Utilizes unlabeled data for enhanced features	Dependent on pretext task design
		Adversarial Learning	Learns domain-invariant features and enhances robustness	Vulnerable to adversarial attacks; computationally complex
	Model Frame-work	Ensemble Learning	Increases robustness and leverages model diversity	High computational cost and implementation complexity
		Distillation	Highly efficient; preserves ID performance	Dependent on choice of teacher; complex training process
		Distributed Learning	Preserves privacy and improves scalability	Faces communication overhead and heterogeneity issues
	Others	Geometric	Handles non-Euclidean data well	High complexity; limited applicability
		DRO	Optimizes across worst-case scenarios for robustness	Complex in defining uncertainty sets
Analysis		Interpretable AI	Enhances trust; model debugging	Interpretability-complexity tradeoff
		Causality	Focuses on invariant features for robust generalization	Complex model development and limited scalability

model to generalize well to unseen domains even when only a single source domain is available for training. This scenario often arises in medical imaging, where data collection can be resource-intensive, and privacy concerns may limit access to multiple sources. Future research should explore novel methods that effectively address the challenges of source-limited domain generalization, such as robustness to concept shift in the presence of covariate shift.

B. Medical Foundation Model

Foundation models [172] refer to hyper-scale models that has been trained on massive and diverse datasets. Arguably, this new class of model is a step towards next-level artificial intelligence that has state-of-the-art zero-shot generalizability performances. However, development of foundation model for MedIA is challenging as medical datasets are heterogeneous and hard to collect in large scale. Despite these challenges, research on medical foundation models is very active and shows great potential [173]. Recently, Segment Anything Model (SAM) [132], [131] has shown promising results for various source-limited DG in segmentation tasks for MedIA.

C. Benchmark Datasets

Many DG methods for MedIA rely on custom datasets that mix private and public datasets, treating each dataset as a distinct domain. The reproducibility of these custom datasets is extremely low because the processes for data splitting, preprocessing, and annotation vary widely and are hard to duplicate. Therefore, there is a need to establish standardized benchmark datasets that reflect the diversity and challenges encountered in real-world medical imaging scenarios. These benchmark datasets should cover various imaging modalities, patient populations, and imaging protocols to facilitate the fair and rigorous evaluation of domain generalization methods. Additionally, benchmark datasets for different settings of domain generalization, *i.e.*, multi-source, single-source, cross-site, cross-sequence, cross-modality, covariate shift, and concept shift, should be developed to enable comprehensive evaluation and comparison of different techniques. Readers are referred to the Supplementary and the interactive website (https://milab.korea.ac.kr/dg-dataset) for a review of existing DG benchmark datasets and a summary of public datasets.

D. Suggestions for Domain Generalization for MedIA

There exist a number of empirical evaluations of commonly used domain generalization techniques suggesting appropriate methods for specific tasks at hand. For example, Korevaar *et al*. [174] evaluated methods on benchmark datasets, whereas Zhang *et al*. [175] and Galappaththige *et al*. [130] evaluated them on some custom cross-site datasets consisting of publicly available datasets. While these benchmarks shed light on the capabilities of certain techniques in domain generalization, they do not offer a comprehensive guide for problem-specific and task-specific strategies throughout the MedIA workflow. To this end, we critically analyze different DG methods (Table XI) and suggest strategies to incorporate domain generalization into the model (Fig. 5).

VI. Conclusion

Domain generalization is a crucial capability for modern medical image analysis, aimed at creating machine learning models capable of handling a wide variety of data distributions arising from variations in, for example, imaging protocols, patient demographics, and equipment. This paper provides a comprehensive review of domain generalization techniques, extending beyond the methodological hierarchy of previous surveys and considering the implications of domain generalization on the entire MedIA workflow. Our focus includes every stage of the decision-making process, from data acquisition to pre-processing, prediction, and analysis.

We have also highlighted and discussed the current benchmark datasets, emphasizing the necessity to expand the spectrum of these resources. Moreover, we shed light on potential directions for future research in this field. While domain generalization for MedIA is still a rapidly evolving field, it is clear that it holds significant promise for improving patient care by enhancing the robustness and reliability of MedIA models. As we continue to address the challenges and capitalize on the opportunities, we anticipate seeing substantial improvements in the efficiency and accuracy of MedIA workflows, leading to more personalized and effective patient treatments. This, in turn, will help the healthcare sector move towards the broader goal of precision medicine, providing each patient with care that is uniquely tailored to their health profile.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) RS-2022-II220959 ((Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making) and RS-2019-II190079 (Department of Artificial Intelligence (Korea University)).

References

[1] C. Lian *et al.*, “Hierarchical fully convolutional network for joint atrophy localization and alzheimer’s disease diagnosis using structural MRI,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 880–893, 2020.

[2] P. Mukherjee *et al.*, “A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets,” *Nature Machine Intelligence*, vol. 2, no. 5, pp. 274–282, 2020.

[3] J. Duan *et al.*, “Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process,” *Medical Physics*, vol. 49, no. 4, pp. 2570–2581, 2022.

[4] Y. Nan *et al.*, “Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions,” *Information Fusion*, vol. 82, pp. 99–122, 2022.

[5] R. J. Chen *et al.*, “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” in *Proceedings of the CVPR*. IEEE, 2022, pp. 16 123–16 134.

[6] L. Feng *et al.*, “5D whole-heart sparse MRI,” *Magnetic Resonance in Medicine*, vol. 79, no. 2, pp. 826–838, 2018.

[7] D. Ghosh and D. Ekta Ghosh, “A large-scale multi-centre research on domain generalisation in deep learning-based mass detection in mammography: A review,” *Acta Biology Forum*, vol. 1, pp. 05–09, 2022.

[8] A. Apicella *et al.*, “Machine learning strategies to improve generalization in eeg-based emotion assessment: a systematic review,” 2022.

[9] M. Jahanifar *et al.*, “Domain generalization in computational pathology: Survey and guidelines,” 2023.

[10] H. Guan and M. Liu, “Domain adaptation for medical image analysis: A survey,” *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[11] G. Sarafraz, A. Behnamnia, M. Hosseinzadeh, A. Balapour, A. Meghrazi, and H. R. Rabiee, “Domain adaptation and generalization on functional medical images: A systematic survey,” 2022.

[12] J. Wang *et al.*, “Generalizing to unseen domains: A survey on domain generalization,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2023.

[13] K. Zhou *et al.*, “Domain generalization: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023.

[14] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, “Towards out-of-distribution generalization: A survey,” 2023.

[15] G. Csurka, R. Volpi, and B. Chidlovskii, “Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey,” 2021.

[16] J. Yang, K. Zhou, Y. Li, and Z. Liu, “Generalized out-of-distribution detection: A survey,” 2024.

[17] P. Cui and J. Wang, “Out-of-distribution (OOD) detection based on deep learning: A review,” *Electronics*, vol. 11, no. 21, p. 3500, 2022.

[18] N. Ghassemi and E. Fazl-Ersi, “A comprehensive review of trends, applications and challenges in out-of-distribution detection,” 2022.

[19] P. Sheth, R. Moraffah, K. S. Candan, A. Raglin, and H. Liu, “Domain generalization – a causal perspective,” 2022.

[20] H. Li, X. Wang, Z. Zhang, and W. Zhu, “Out-of-distribution generalization on graphs: A survey,” 2022.

[21] Y. Li, X. Wang, R. Zeng, P. K. Donta, I. Murturi, M. Huang, and S. Dustdar, “Federated domain generalization: A survey,” 2024.

[22] W. Yu *et al.*, “SAN-Net: Learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization,” *Computers in Biology and Medicine*, vol. 156, p. 106717, 2023.

[23] N. Karani *et al.*, “Test-time adaptable neural networks for robust medical image segmentation,” *Medical Image Analysis*, vol. 68, p. 101907, 2021.

[24] J. Ma, "Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation," in *Proceedings of the STACOM*. Cham: Springer, 2021, pp. 177–186.

[25] L. Li *et al.*, "Atrialgeneral: Domain generalization for left atrial segmentation of multi-center lge mris," in *Proceedings of the MICCAI*. Cham: Springer, 2021, pp. 557–566.

[26] H. Gunasinghe *et al.*, "Domain generalisation for glaucoma detection in retinal images from unseen fundus cameras," in *Proceedings of the ACIIDS*. Cham: Springer, 2022, pp. 421–433.

[27] S. Kondo, S. Kasai, and K. Hirasawa, "Tackling mitosis domain generalization in histopathology images with color normalization," in *Proceedings of the MICCAI Workshop*. Cham: Springer, 2023, pp. 217–220.

[28] J. Xiong *et al.*, "Improve unseen domain generalization via enhanced local color transformation," in *Proceedings of the MICCAI*. Cham: Springer, 2020, pp. 433–443.

[29] A. Pakzad, K. Abhishek, and G. Hamarneh, "Circle: Color invariant representation learning for unbiased classification of skin lesions," in *Proceedings of the ECCV Workshops*. Cham: Springer, 2023, pp. 203–219.

[30] R. Zhang *et al.*, "Semi-supervised domain generalization for medical image analysis," in *Proceedings of the ISBI*, 2022, pp. 1–5.

[31] X. Zhao *et al.*, "Test-time fourier style calibration for domain generalization," in *Proceedings of the IJCAI*, 7 2022, pp. 1721–1727.

[32] Q. Xu *et al.*, "Fourier-based augmentation with applications to domain generalization," *Pattern Recognition*, vol. 139, p. 109474, 2023.

[33] H. Liu, H. Li, M. Ou, Y. Zhao, H. Qi, Y. Hu, and J. Liu, "Domain generalization in restoration of cataract fundus images via high-frequency components," in *Proceedings of the ISBI*, 2022, pp. 1–5.

[34] D. Hu *et al.*, "Domain generalization for retinal vessel segmentation with vector field transformer," in *Proceedings of the MIDL*, ser. Proceedings of Machine Learning Research, vol. 172. PMLR, 06–08 Jul 2022, pp. 552–564.

[35] H. Li *et al.*, "Frequency-mixed single-source domain generalization for medical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2023, pp. 127–136.

[36] I. Zakazov *et al.*, "Feather-light fourier domain adaptation in magnetic resonance imaging," in *Proceedings of the MICCAI Workshop*. Cham: Springer, 2022, pp. 88–97.

[37] L. Zhang *et al.*, "Motion correction in mri using deep learning and a novel hybrid loss function," 2022.

[38] P. Song *et al.*, "Coupled dictionary learning for multi-contrast MRI reconstruction," *IEEE Transactions on Medical Imaging*, vol. 39, no. 3, pp. 621–633, 2019.

[39] Q. Liu, C. Chen, Q. Dou, and P. Heng, "Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary," in *Proceedings of the AAAI*. AAAI Press, 2022, pp. 1756–1764.

[40] L. Li *et al.*, "Random style transfer based domain generalization networks integrating shape and spatial information," in *Proceedings of the STACOM*. Cham: Springer, 2020, pp. 208–218.

[41] Z. Zhang, Y. Li, and B.-S. Shin, "Robust color medical image segmentation on unseen domain by randomized illumination enhancement," *Computers in Biology and Medicine*, vol. 145, p. 105427, 2022.

[42] R. Gomathi and S. Selvankumaran, "A novel medical image segmentation model with domain generalization approach," *International Journal of Electrical and Electronics Research*, vol. 10, no. 2, pp. 312–319, 2022.

[43] D. Tomar, G. Vray, B. Bozorgtabar, and J.-P. Thiran, "TeSLA: Test-time self-learning with automatic adversarial augmentation," in *Proceedings of the CVPR*, June 2023, pp. 20 341–20 350.

[44] M. Scalbert *et al.*, "Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology," in *Proceedings of the MICCAI*. Cham: Springer, 2022, pp. 120–129.

[45] R. Yamashita *et al.*, "Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3945–3954, 2021.

[46] X. Yang *et al.*, "Learning with synthesized data for generalizable lesion detection in real pet images," in *Proceedings of the MICCAI*. Cham: Springer, 2023, pp. 116–126.

[47] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Proceedings of the IPMI*. Cham: Springer, 2021, pp. 661–673.

[48] Y. Xu *et al.*, "Adversarial consistency for single domain generalization in medical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2022, pp. 671–681.

[49] Z. Su *et al.*, "Rethinking data augmentation for single-source domain generalization in medical image segmentation," in *Proceedings of the AAAI*, vol. 37, 2023, pp. 2366–2374.

[50] C. Xu *et al.*, "Improved domain generalization for cell detection in histopathology images via test-time stain augmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2022, pp. 150–159.

[51] J.-R. Chang *et al.*, "Stain mix-up: Unsupervised domain generalization for histopathology images," in *Proceedings of the MICCAI*. Cham: Springer, 2021, pp. 117–126.

[52] D. Adila and D. Kang, "Understanding out-of-distribution: A perspective of data dynamics," in *Proceedings of the NeurIPS Workshops*, ser. Proceedings of Machine Learning Research, vol. 163. PMLR, 13 Dec 2022, pp. 1–8.

[53] M. Shah *et al.*, "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis," *Medical Image Analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[54] P. Chlap *et al.*, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[55] J. Panic *et al.*, "Normalization strategies in multi-center radiomics abdominal mri: Systematic review and meta-analyses," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 4, pp. 67–76, 2023.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2015, pp. 234–241.

[57] R. T. Shinohara *et al.*, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[58] J. T. Pontalba *et al.*, "Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 300, 2019.

[59] A. Vahadane *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[60] Y. Choi *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the CVPR*. IEEE Computer Society, 2018, pp. 8789–8797.

[61] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[62] V. Golkov *et al.*, "Q-space deep learning: twelve-fold shorter and model-free diffusion mri scans," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1344–1351, 2016.

[63] S. Koppers *et al.*, "Spherical harmonic residual network for diffusion signal harmonization," in *Proceedings of the MICCAI Workshop*. Springer, 2019, pp. 173–182.

[64] F. Zhao *et al.*, "Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks," in *Proceedings of the MICCAI*. Springer, 2019, pp. 475–483.

[65] M. Ren, N. Dey, J. Fishbaugh, and G. Gerig, "Segmentation-renormalized deep feature modulation for unpaired image harmonization," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1519–1530, 2021.

[66] H. Guan *et al.*, "Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification," *Medical Image Analysis*, vol. 71, p. 102076, 2021.

[67] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal," *NeuroImage*, vol. 228, p. 117689, 2021.

[68] D. Moyer *et al.*, "Scanner invariant representations for diffusion mri harmonization," *Magnetic Resonance in Medicine*, vol. 84, no. 4, pp. 2174–2189, 2020.

[69] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *Proceedings of the CVPR*, 2021, pp. 14 383–14 392.

[70] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proceedings of the ICLR*. OpenReview.net, 2018.

[71] H. Lee *et al.*, "Machine friendly machine learning: Interpretation of computed tomography without image reconstruction," *Scientific Reports*, vol. 9, no. 1, p. 15540, 2019.

[72] L. Gallardo-Estrella *et al.*, "Normalizing computed tomography data reconstructed with different filter kernels: Effect on emphysema quantification," *European Radiology*, vol. 26, pp. 478–486, 2016.

[73] R. Zhao, H. Li, and X. Liu, "A survey of dictionary learning in medical image analysis and its application for glaucoma diagnosis," *Archives of Computational Methods in Engineering*, vol. 28, pp. 463–471, 2021.

[74] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[75] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[76] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523000506

[77] Y. Shen, Y. Luo, D. Shen, and J. Ke, "Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 212–221.

[78] Y. Choi *et al.*, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the CVPR*. IEEE, 2020, pp. 8185–8194.

[79] Y. Xu *et al.*, "Generative adversarial domain generalization via cross-task feature attention learning for prostate segmentation," in *Procedings of the ICONIP*. Cham: Springer, 2021, pp. 273–284.

[80] G. Xie *et al.*, "Cross-modality neuroimage synthesis: A survey," *ACM Computing Surveys*, vol. 56, no. 3, 2023.

[81] Z. Zhou *et al.*, "Generalizable cross-modality medical image segmentation via style augmentation and dual normalization," in *Proceedings of the CVPR*, 2022, pp. 20 824–20 833.

[82] Y.-P. Liu *et al.*, "Ss-norm: Spectral-spatial normalization for single-domain generalization with application to retinal vessel segmentation," *IET Image Processing*, vol. 17, no. 7, pp. 2168–2181, 2023.

[83] K. Stacke *et al.*, "Measuring domain shift for deep learning in histopathology," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 325–336, 2020.

[84] J. Lyu *et al.*, "AADG: Automatic augmentation for domain generalization on retinal image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3699–3711, 2022.

[85] H. Li *et al.*, "Domain generalization for medical imaging classification with linear-dependency regularization," in *Proceedings of the NeurIPS*, vol. 33. Curran Associates, Inc., 2020, pp. 3118–3129.

[86] Q. Meng, D. Rueckert, and B. Kainz, "Learning cross-domain generalizable features by representation disentanglement," 2020.

[87] Q. Meng *et al.*, "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 722–734, 2020.

[88] Y. Bi *et al.*, "Mi-segnet: Mutual information-based us segmentation for unseen domain generalization," in *Proceedings of the MICCAI*. Cham: Springer, 2023, pp. 130–140.

[89] Y. Chen *et al.*, "D-BIN: A generalized disentangling batch instance normalization for domain adaptation," *IEEE Transactions on Cybernetics*, vol. 53, no. 4, pp. 2151–2163, 2023.

[90] Z. Li *et al.*, "Domain generalization for mammography detection via multi-style and multi-view contrastive learning," in *Proceedings of the MICCAI*. Cham: Springer, 2021, pp. 98–108.

[91] R. Gu *et al.*, "CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation," *Medical Image Analysis*, vol. 89, p. 102904, 2023.

[92] M. Ilse *et al.*, "DIVA: Domain invariant variational autoencoders," in *Proceedings of the MIDL*, ser. Proceedings of Machine Learning Research, vol. 121. PMLR, 06–08 Jul 2020, pp. 322–348.

[93] Y. Wang *et al.*, "Variational disentanglement for domain generalization," *Transactions on Machine Learning Research*, 2022.

[94] C. Wang, J. Li, X. Sun, F. Zhang, Y. Yu, and Y. Wang, "Domain invariant model with graph convolutional network for mammogram classification," 2022.

[95] ——, "Learning domain-agnostic representation for disease diagnosis," in *Proceedings of the ICLR*, 2023.

[96] X. Liu *et al.*, "Recursively conditional gaussian for ordinal unsupervised domain adaptation," in *Proceedings of the ICCV*. IEEE, 2021, pp. 744–753.

[97] ——, "Ordinal unsupervised domain adaptation with recursively conditional gaussian imposed variational disentanglement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022.

[98] S. Wang *et al.*, "DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4237–4248, 2020.

[99] M. Islam and B. Glocker, "Frequency dropout: Feature-level regularization via randomized filtering," in *Proceedings of the ECCV Workshops*. Cham: Springer, 2023, pp. 281–295.

[100] K. P. Nguyen, A. H. Treacher, and A. A. Montillo, "Adversarially-regularized mixed effects deep learning (armed) models improve interpretability, performance, and generalization on clustered (non-iid) data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8081–8093, 2023.

[101] Y. Wang *et al.*, "Embracing the dark knowledge: Domain generalization using regularized knowledge distillation," in *Proceedings of the ACM MM*, New York, NY, USA, 2021, p. 2595–2604.

[102] C. Chen *et al.*, "Maxstyle: Adversarial style composition for robust medical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2022, pp. 151–161.

[103] T. Zhou and E. Konukoglu, "FedFA: Federated feature augmentation," in *Proceedings of the ICLR*, 2023.

[104] R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu, "From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation," in *Proceedings of the WACV*, January 2024, pp. 464–474.

[105] R. Wang, P. Chaudhari, and C. Davatzikos, "Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation," *Medical Image Analysis*, vol. 76, p. 102309, 2022.

[106] K. P. Ayodele *et al.*, "Supervised domain generalization for integration of disparate scalp EEG datasets for automatic epileptic seizure detection," *Computers in Biology and Medicine*, vol. 120, p. 103757, 2020.

[107] L. Huang *et al.*, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 173–10 196, 2023.

[108] M. I. Belghazi *et al.*, "Mutual information neural estimation," in *Proceedings of the ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 530–539.

[109] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[110] R. Gu *et al.*, "Contrastive domain disentanglement for generalizable medical image segmentation," 2022.

[111] J. S. Yoon *et al.*, "A plug-in method for representation factorization in connectionist models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3792–3803, 2022.

[112] P. Bándi *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.

[113] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[114] P. Khandelwal and P. Yushkevich, "Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging," in *Proceedings of the MICCAI Workshop*. Cham: Springer, 2020, pp. 73–84.

[115] Q. Dou *et al.*, "Domain generalization via model-agnostic learning of semantic features," in *Proceedings of the NeurIPS*, 2019, pp. 6447–6458.

[116] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *Proceedings of the MICCAI*. Cham: Springer, 2020, pp. 475–485.

[117] X. Liu *et al.*, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2021, pp. 307–317.

[118] D. Hu *et al.*, "Map: Domain generalization via meta-learning on anatomy-consistent pseudo-modalities," in *Proceedings of the MICCAI*. Cham: Springer, 2023, pp. 182–192.

[119] J. Lee *et al.*, "Site-invariant meta-modulation learning for multisite autism spectrum disorders diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[120] C. Ouyang *et al.*, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," in *Proceedings of the ECCV*. Cham: Springer, 2020, pp. 762–780.

[121] S. Azizi *et al.*, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 756–779, 2023.

[122] S. Hu, Z. Liao, and Y. Xia, "Devil is in channels: Contrastive single domain generalization for medical image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2023, pp. 14–23.

[123] I. E. I. Bekkouch *et al.*, "Adversarial reconstruction loss for domain generalization," *IEEE Access*, vol. 9, pp. 42 424–42 437, 2021.

[124] C. Chen *et al.*, "Realistic adversarial data augmentation for mr image segmentation," in *Proceedings of the MICCAI*. Cham: Springer, 2020, pp. 667–677.

[125] Z. Zhang *et al.*, "Domain generalization with adversarial intensity attack for medical image segmentation," 2023.

[126] S. Cheng, T. Gokhale, and Y. Yang, "Adversarial bayesian augmentation for single-source domain generalization," in *Proceedings of the ICCV*, October 2023, pp. 11 400–11 410.

[127] R. A. Kamraoui *et al.*, "DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation," *Medical Image Analysis*, vol. 76, p. 102312, 2022.

[128] M. Philipp *et al.*, "Dynamic cnns using uncertainty to overcome domain generalization for surgical instrument localization," in *Proceedings of the WACV*, January 2022, pp. 3612–3621.

[129] C. Fernandez-Martín *et al.*, "Uninformed teacher-student for hard-samples distillation in weakly supervised mitosis localization," *Computerized Medical Imaging and Graphics*, vol. 112, p. 102328, 2024.

[130] C. J. Galappaththige, G. Kuruppu, and M. H. Khan, "Generalizing to unseen domains in diabetic retinopathy classification," in *Proceedings of the WACV*, January 2024, pp. 7685–7695.

[131] H. Wang, H. Ye, Y. Xia, and X. Zhang, "Leveraging sam for single-source domain generalization in medical image segmentation," 2024.

[132] Y. Gao *et al.*, "Desam: Decoupling segment anything model for generalizable medical image segmentation," 2023.

[133] C. Chen *et al.*, "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation," 2023.

[134] Q. Liu *et al.*, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the CVPR*, June 2021, pp. 1013–1023.

[135] J. Chen *et al.*, "Federated domain generalization for image recognition via cross-client style transfer," in *Proceedings of the WACV*, January 2023, pp. 361–370.

[136] Q. Xu *et al.*, "Federated adversarial domain hallucination for privacy-preserving domain generalization," *IEEE Transactions on Multimedia*, vol. 26, pp. 1–14, 2024.

[137] W. Li *et al.*, "Privacy-preserving federated brain tumour segmentation," in *Proceedings of the MLMI*. Cham: Springer, 2019, pp. 133–141.

[138] D. M. H. Nguyen *et al.*, "Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching," in *Proceedings of the NeurIPS*, vol. 36. Curran Associates, Inc., 2023, pp. 27 922–27 950.

[139] L. Seenivasan *et al.*, "Biomimetic incremental domain generalization with a graph network for surgical scene understanding," *Biomimetics*, vol. 7, no. 2, p. 68, 2022.

[140] A. Santhirasekaram, M. Winkler, A. Rockall, and B. Glocker, "Topology preserving compositionality for robust medical image segmentation," in *Proceedings of the CVPR Workshop*, June 2023, pp. 543–552.

[141] A. Bissoto *et al.*, "Artifact-based domain generalization of skin lesion models," in *Proceedings of the ECCV Workshop*. Cham: Springer, 2023, pp. 133–149.

[142] K. Goel *et al.*, "Model patching: Closing the subgroup performance gap with data augmentation," in *Proceedings of the ICLR*. OpenReview.net, 2021.

[143] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1126–1135.

[144] D. Li *et al.*, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI*. AAAI Press, 2018, pp. 3490–3497.

[145] J. Zhang, Z. Liu, S. Zhang, H. Zhang, P. Spincemaille, T. D. Nguyen, M. R. Sabuncu, and Y. Wang, "Fidelity imposed network edit (fine) for solving ill-posed image reconstruction," *NeuroImage*, vol. 211, p. 116579, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920300665

[146] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince, "Smore: A self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 805–817, 2021.

[147] G. Hinton, O. Vinyals, and J. Dean, "Dark knowledge," *Presented as the keynote in BayLearn*, vol. 2, no. 2, 2014.

[148] G. A. Kaissis *et al.*, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[149] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proceedings of the ACM CCS*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 308–318.

[150] Y. Chen *et al.*, "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020.

[151] M. M. Bronstein *et al.*, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021.

[152] D. Wang *et al.*, "Proactive pseudo-intervention: Causally informed contrastive learning for interpretable vision models," 2021.

[153] M. R. Karim *et al.*, "DeepKneeExplainer: Explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging," *IEEE Access*, vol. 9, pp. 39 757–39 780, 2021.

[154] Y. Wang *et al.*, "SEEG-Net: An explainable and deep learning-based cross-subject pathological activity detection method for drug-resistant epilepsy," *Computers in Biology and Medicine*, vol. 148, p. 105703, 2022.

[155] W. Fan *et al.*, "Invnorm: Domain generalization for object detection in gastrointestinal endoscopy," 2022.

[156] S. Gao *et al.*, "Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability," *Medical Image Analysis*, vol. 89, p. 102889, 2023.

[157] J. Yuan, F. Pinto, A. Davies, and P. Torr, "Not just pretty pictures: Toward interventional data augmentation using text-to-image generators," 2023.

[158] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *Proceedings of the ICML*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 7313–7324.

[159] R. Wang, P. Chaudhari, and C. Davatzikos, "Harmonization with flow-based causal inference," in *Proceedings of the MICCAI*. Cham: Springer, 2021, pp. 181–190.

[160] C. Ouyang *et al.*, "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1095–1106, 2023.

[161] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[162] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *Proceedings of the ICLR*. OpenReview.net, 2020.

[163] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[164] B. H. van der Velden *et al.*, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.

[165] A. Chattopadhay *et al.*, "Grad-Cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of the WACV*. IEEE, 2018, pp. 839–847.

[166] J. Yang, K. Zhou, and Z. Liu, "Full-spectrum out-of-distribution detection," *International Journal of Computer Vision*, vol. 131, no. 10, pp. 2607–2622, Oct 2023.

[167] D. Mahajan, S. Tople, and A. Sharma, "The connection between out-of-distribution generalization and privacy of ml models," 2021.

[168] K. Zheng *et al.*, "From single to multiple: Generalized detection of covid-19 under limited classes samples," *Computers in Biology and Medicine*, vol. 164, p. 107298, 2023.

[169] A. M. Puli, L. H. Zhang, E. K. Oermann, and R. Ranganath, "Out-of-distribution generalization in the presence of nuisance-induced spurious correlations," in *Proceedings of the ICLR*. OpenReview.net, 2022.

[170] A. Puli, N. Joshi, H. He, and R. Ranganath, "Nuisances via negativa: Adjusting for spurious correlations via data augmentation," 2023.

[171] I. Gao *et al.*, "Out-of-distribution robustness via targeted augmentations," in *Proceedings of the NeurIPS Workshop*, 2022.

[172] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," 2022.

[173] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[174] S. Korevaar, R. Tennakoon, and A. Bab-Hadiashar, "Failure to achieve domain invariance with domain generalization algorithms: An analysis in medical imaging," *IEEE Access*, vol. 11, pp. 39 351–39 372, 2023.

[175] H. Zhang *et al.*, "An empirical framework for domain generalization in clinical settings," in *Proceedings of the CHIL*. New York, NY, USA: Association for Computing Machinery, 2021, p. 279–290.

[176] V. M. Campello *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[177] M. Aubreville *et al.*, "Mitosis domain generalization in histopathology images—the midog challenge," *Medical Image Analysis*, vol. 84, p. 102699, 2023.

[178] A. F. Kazerooni *et al.*, "The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds)," 2024.

[179] M. Adewole *et al.*, "The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa)," 2023.

[180] H. Bogunović *et al.*, "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1858–1874, 2019.

[181] N. M. Müller, J. Jacobs, J. Williams, and K. Böttinger, "Localized shortcut removal," in *Proceedings of the CVPR Workshop*, June 2023, pp. 3721–3725.

Jee Seok Yoon received a BS degree in Computer Science and Engineering from Korea University, Seoul, South Korea, in 2018. He is pursuing a PhD with the Department of Brain and Cognitive Engineering at Korea University, Seoul, South Korea.

His research interests include explainable AI, computer vision, and representation learning.

Kwanseok Oh received the BS degree in Electronic Control and Engineering from Hanbat National University, Daejeon, South Korea, in 2020. He is currently pursuing a Ph.D. degree with the Department of Artificial Intelligence at Korea University, Seoul, South Korea.

His current research interests include explainable AI, computer vision, and representation learning.

Yooseung Shin received the BS degree in Computer Science and Engineering from Hallym University, Chuncheon, South Korea, in 2022, and MS degree in Artificial Intelligence from Korea University, Seoul, South Korea, in 2024.

His research interests include transfer learning, computer vision, and machine/deep learning.

Maciej A. Mazurowski is currently an associate professor of radiology and electrical and computer engineering at Duke University. He is also a faculty with the Duke Medical Physics Program and a member of Duke Cancer Institute. His main research interest includes applications of machine learning including deep learning and statistical modeling, as well as computer vision algorithms to medicine. The particular focus in terms of applications is medical imaging in the context of cancer treatment, as well as understanding image interpretation process and error-making in radiology.

Heung-Il Suk is currently a Professor at the Department of Artificial Intelligence and an Adjunct Professor at the Department of Brain and Cognitive Engineering at Korea University. He was a Visiting Professor at the Department of Radiology at Duke University between 2022 and 2023.

He was awarded a Kakao Faculty Fellowship from Kakao and a Young Researcher Award from the Korean Society for Human Brain Mapping (KHBM) in 2018 and 2019, respectively. His research interests include causal machine/deep learning, explainable AI, biomedical data analysis, and brain-computer interface.

Dr. Suk serves as an Editorial Board Member for Clinical and Molecular Hepatology (Artificial Intelligence Sector), Electronics, Frontiers in Neuroscience, Frontiers in Radiology (Artificial Intelligence in Radiology), International Journal of Imaging Systems and Technology (IJIST), and a Program Committee or a Reviewer for NeurIPS, ICML, ICLR, AAAI, IJCAI, CVPR, MICCAI, AISTATS, *etc*.

Supplementary Material

A. Benchmark Datasets

The field of domain generalization for MedIA relies primarily on customized datasets which are created by combining several private and publicly available datasets. Although these datasets have greatly contributed to the field, there is a strong need to expand the spectrum of benchmark datasets. In this section, we introduce three benchmark datasets for the domain generalization task in MedIA: Camelyon17-WILDS [112], the M&Ms challenge dataset [176], and the MIDOG challenge dataset [177]. To the best of our knowledge, these three are the only publicly open benchmark dataset for DG in MedIA. For a comprehensive review of benchmark datasets for domain *adaptation*, refer to [10].

1) Existing benchmark datasets

The **Camelyon17-WILDS** [112] dataset is derived from the Camelyon17 Challenge, which focuses on detecting metastasis in histopathological images of lymph nodes. This dataset, however, has been adapted specifically for domain generalization. It includes pathology images from two separate institutions and contains a total of 100 whole-slide images. The main challenge in this dataset comes from the inter-institutional variations, including differences in staining procedures and scanners used, which can significantly affect the performance of models. The **M&Ms challenge dataset** [176] is a multi-center, multi-vendor, and multi-disease cardiac magnetic resonance (CMR) dataset. The M&Ms challenge focuses on automatically segmenting the left ventricle, right ventricle, and myocardium in cardiac MR images, which are critical for diagnosing and managing various cardiovascular diseases. The dataset contains images from five different sites and five scanners, with patients suffering from five distinct pathologies. This heterogeneity poses a challenge for domain generalization, as models need to overcome variations in imaging protocols, equipment, and patient populations. The **MIDOG challenge dataset** [177] is focused on detecting mitotic figures in histopathological images, which is an important task for cancer diagnosis and grading. This dataset comprises images from five different hospitals, and the images have been digitized using different scanners and have undergone various staining procedures. These inter-hospital variations make it a challenging dataset for domain generalization.

2) Emerging benchmark dataset

One promising dataset for this purpose is the Brain Tumor Segmentation (BRATS) Challenge dataset. The BRATS dataset has been a valuable resource for neuroimaging researchers since its inception. Since the 2022 challenge, the dataset has included additional cohorts from pediatric [178] and African [179] populations. These additions significantly increase the diversity of the dataset, making it an ideal resource for domain generalization research. Including pediatric and African cohorts helps address two key areas of need in the field. First, there is a growing acknowledgment of the need to ensure that machine learning models in healthcare are trained on diverse data representing various age groups. The pediatric cohort in the BRATS dataset provides an opportunity to test and improve the performance of models in analyzing medical images from younger patients. Second, the African cohort provides much-needed diversity in terms of ethnicity, helping to mitigate model biases and improve the generalizability of machine learning models across different ethnic groups. Another emerging benchmark dataset for domain generalization is the Retinal OCT Fluid Challenge (RETOUCH) [180], a cross-site dataset of 70 OCT volumes with 3 sites.

B. Publicly Open Datasets

TABLE XII: List of datasets used by reviewed literature.

Task	Organ	Modality	Abbreviation	Ref.
Classification	Brain	MRI	ADNI[1]	[100], [159]
			iSTAGING[2]	[105], [159]
		EEG	MAYO[3]	[154]
			FNUSA[4]	[154]
	Skin	Dermatology	ISIC[5]	[85], [142], [141], [138]
			HAM10000[6]	[85], [123]
			Fitzpatrick17K[7]	[29]
			Dermofit[8]	[85]

Continued on next page

[1] Alzheimer's Disease Neuroimaging Initiative, https://adni.loni.usc.edu/

[2] Imaging-based coordinate SysTem for AGing and NeurodeGenerative diseases consortium, https://doi.org/10.1093/braincomms/fcac117

[3] Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals, https://doi.org/10.6084/m9.figshare.c.4681208

[4] Multicenter sEEG dataset, https://www.kaggle.com/datasets/nejedlypetr/multicenter-intracranial-eeg-dataset

[5] International Skin Imaging Collaboration, https://api.isic-archive.com/collections/

[6] The Human Against Machine with 10000 training images, https://doi.org/10.7910/DVN/DBW86T

[7] Fitzpatrick17K, https://github.com/mattgroh/fitzpatrick17k

[8] DERMOFIT dataset, https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library

Task	Organ	Modality	Abbreviation	Ref.
Classification	Skin	Dermatology	PH2[9]	[85]
			Derm7pt[10]	[85]
		Histology	openLCH[11]	[100]
	Colon	Histology	Kather16[12]	[43], [44]
			Kather18[13]	[43], [44], [45]
			Kather19[14]	[45]
			CRC-TP[15]	[44]
			AIDA-LNCO[16]	[83]
	Blood Cell	Histology	BCISC[17]	[89]
			LISC[18]	[89]
			BCCD[19]	[89]
	Chest	X-ray	MIMIC-CXR[20]	[121], [169], [170], [30]
			CXR8[21]	[158], [30]
			ChexPert[22]	[121], [167], [169], [170], [158], [30]
			RSNA-PD[23]	[158]
			ChestX-ray14[24]	[121]
			PadChest[25]	[30]
	Breast	Histology	CAMELYON17[26]	[44], [51], [83], [93], [121], [135], [171], [157], [126]
		X-ray	DDSM[27]	[94], [95]
	Knee	MRI	MOST[28]	[153]
	Lung	CT	LIDC-IDRI[29]	[152]
	Retinal	OCT	A2ASDOCT[30]	[152]
Segmentation	Brain	MRI	WMH[31]	[31]
			ATLAS[32]	[22]
			ABIDE[33]	[23], [119]
			ISBI-MS[34]	[127]
			BraTS[35]	[47], [81], [138], [137]
			CC359[36]	[36]

Continued on next page

[9] Ph2 Database, https://www.fc.up.pt/addi/ph2%20database.html

[10] 7-point criteria evaluation Database, https://derm.cs.sfu.ca/

[11] Live Cell Histology, https://doi.org/10.17867/10000161

[12] Kather16, https://doi.org/10.5281/zenodo.53169

[13] Kather18, https://doi.org/10.5281/zenodo.1214456

[14] Kather19, https://doi.org/10.5281/zenodo.2530835

[15] ColoRectal Cancer for Tissue Phenotyping, https://warwick.ac.uk/fac/cross_fac/tia/data/

[16] Regional lymph node metastasis in colon adenocarcinoma dataset, https://doi.org/10.23698/aida/lnco

[17] Blood Cell Images for Segmentation and Classification dataset, https://github.com/fpklipic/BCISC

[18] Leukocyte Images for Segmentation and Classification, https://users.cecs.anu.edu.au/~{}hrezatofighi/Data/Leukocyte%20Data.htm

[19] Blood Cell Count and Detection, https://www.kaggle.com/datasets/paultimothymooney/blood-cells

[20] Medical Information Mart for Intensive Care-Chest X-Ray, https://doi.org/10.13026/s5dg-6s42

[21] ChestX-ray8, https://nihcc.app.box.com/v/ChestXray-NIHCC

[22] Chest eXpert (A Large Chest Radiograph Dataset), https://stanfordmlgroup.github.io/competitions/chexpert/

[23] RSNA Pneumonia Detection Challenge, https://www.rsna.org/rsnai/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018

[24] NIH Chest X-rays, https://www.kaggle.com/datasets/nih-chest-xrays/data

[25] PadChest: A Large Chest X-ray Image Dataset with Multi-Label Annotated Reports, https://bimcv.cipf.es/bimcv-projects/padchest/

[26] CAMELYON, https://camelyon17.grand-challenge.org/

[27] CBIS-DDSM: Breast Cancer Image Dataset, https://wiki.cancerimagingarchive.net/x/lZNXAQ

[28] Multicenter Osteoarthritis Study, https://agingresearchbiobank.nia.nih.gov/studies/most/

[29] The Lung Image Database Consortium and Image Database Resource Initiative, https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

[30] AREDS 2 Ancillary Spectral Domain Optical Coherence Tomography Study, https://clinicaltrials.gov/study/NCT00734487

[31] MICCAI White Matter Hyperintensity Challenge, https://doi.org/10.34894/AECRSD

[32] Anatomical Tracings of Lesions After Stroke, https://doi.org/10.3886/ICPSR36684

[33] Autism Brain Imaging Data Exchange, https://fcon_1000.projects.nitrc.org/indi/abide/

[34] ISBI Longitudinal Multiple Sclerosis Lesion, https://iacl.ece.jhu.edu/index.php/MSChallenge

[35] Brain Tumor Segmentation Challenge, https://www.med.upenn.edu/cbica/brats/

[36] Calgary-Campinas 359, https://www.ccdataset.com/download

Task	Organ	Modality	Abbreviation	Ref.
Segmentation	Brain	MRI	MSSEG[37]	[127]
		fMRI, MRI	HCP[38]	[23]
	Retinal	Fundus	Drishti-GS[39]	[39], [84], [98], [125], [91], [30], [41], [110]
			STARE[40]	[82], [84], [98], [118]
			IDRiD[41]	[84], [96], [97]
			KDR[42]	[84], [96], [97], [35], [130]
			RIM-ONE-r3[43]	[26], [39], [84], [98], [125], [134], [91], [30], [41], [110]
			DRHAGIS[44]	[82]
			REFUGE[45]	[26], [32], [39], [84], [98], [125], [136], [134], [91], [30], [41], [110]
			CHASE[46]	[82], [84], [98]
			E-Ophtha[47]	[84]
			ARIA[48]	[82], [118]
			IOSTAR[49]	[82], [35]
			HRF[50]	[84], [98]
			LES-AV[51]	[35]
			PRIME-FP20[52]	[118]
			RIGA+[53]	[122]
			APTOS[54]	[130]
			MESSIDOR[55]	[130]
		OCT	OCTA-500[56]	[34], [84], [118]
			ROSE[57]	[34], [84], [118]
		FA	RECOVERY-FA19[58]	[118]
	Prostate	MRI	NCI-ISBI[59]	[23], [39], [43], [48], [102], [103], [116], [125], [140], [134], [156], [160], [132]
			I2CVB[60]	[39], [43], [48], [102], [103], [116], [125], [156], [160], [132]
			PROMISE[61]	[23], [39], [48], [102], [103], [116], [125], [134], [156], [160], [132]
	Prostate, Spinal	MRI	SCGM[62]	[43], [85], [101], [117]
	Spinal	CT	CSI[63]	[114]

Continued on next page

[37] MICCAI2016 MS Challenge Dataset, https://portal.fli-iam.irisa.fr/msseg-challenge/

[38] Human Connectome Project, https://www.humanconnectome.org/

[39] Retinal image dataset for optic disc and cup segmentation, http://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php

[40] Structured Analysis of the Retina dataset, https://cecas.clemson.edu/∼{}ahoover/stare/

[41] Indian Diabetic Retinopathy Image Dataset, https://dx.doi.org/10.21227/H25W98

[42] Kaggle Diabetic Retinopathy dataset, https://kaggle.com/competitions/diabetic-retinopathy-detection

[43] RIM-ONE Release 3, https://medimrg.webs.ull.es/

[44] DR HAGIS database, https://personalpages.manchester.ac.uk/staff/niall.p.mcloughlin/

[45] Retinal Fundus Glaucoma Challenge, https://refuge.grand-challenge.org/

[46] Child Heart and Health Study in England dataset, https://researchdata.kingston.ac.uk/id/eprint/96

[47] e-ophtha database, https://www.adcis.net/en/third-party/e-ophtha/

[48] Automated Retinal Image Analysis dataset, https://www.researchgate.net/post/How_can_I_find_the_ARIA_Automatic_Retinal_Image_Analysis_Dataset/5964f84aed99e15c3140b3e6/citation/download

[49] IOSTAR vessel segmentation dataset, https://www.retinacheck.org/download-iostar-retinal-vessel-segmentation-dataset

[50] High-Resolution Fundus Image Database, https://www5.cs.fau.de/research/data/fundus-images/

[51] LES-AV dataset, https://doi.org/10.6084/m9.figshare.11857698

[52] PRIME-FP20: Ultra-widefield Fundus Photography Vessel Segmentation Dataset, https://doi.org/10.21227/ctgj-1367

[53] RIGA+ Dataset for Unsupervised Domain Adaptation in Medical Image Segmentation, https://doi.org/10.5281/zenodo.6325549

[54] APTOS 2019 Blindness Detection, https://www.kaggle.com/c/aptos2019-blindness-detection

[55] Feedback on a Publicly Distributed Image Database: The Messidor Database, https://www.adcis.net/en/third-party/messidor/

[56] A Retinal Dataset for Optical Coherence Tomography Angiography, http://ieee-dataport.org/1951

[57] Retinal OCTA SEgmentation dataset, https://imed.nimte.ac.cn/dataofrose.html

[58] RECOVERY-FA19: Ultra-widefield Fluorescein Angiography Vessel Detection Dataset, https://doi.org/10.21227/m9yw-xs04

[59] NCI-ISBI 2013 Challenge, http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vlOPv

[60] Initiative for Collaborative Computer Vision Benchmarking, http://i2cvb.github.io/

[61] Prostate MR Image Segmentation 2012, https://promise12.grand-challenge.org/

[62] Spinal Cord Grey Matter Segmentation Challenge, http://cmictig.cs.ucl.ac.uk/niftyweb/challenge/

[63] A multi-center milestone study of clinical vertebral CT segmentation, http://spineweb.digitalimaginggroup.ca/

Task	Organ	Modality	Abbreviation	Ref.
Segmentation	Spinal	CT	xVertSeg[64]	[114]
			VerSe[65]	[114]
	Cardiac	MRI	EMIDEC[66]	[156]
			RVSC[67]	[23]
			ACDC[68]	[23], [102], [124], [156]
			M&Ms[69]	[24], [40], [99], [102], [117], [140]
			MS-CMRSeg[70]	[48], [49], [102], [120], [156], [160]
		CT, MRI	MM-WHS[71]	[81], [138], [156]
	Carotid	Ultrasound	SPLab[72]	[88]
	Atrial	MRI	cDEMRIS[73]	[25]
			ASC[74]	[25]
	Abdominal	CT	SABSCT[75]	[48], [49], [81], [120], [140], [160], [133]
			LiTS[76]	[104]
		CT, MRI	CHAOS[77]	[47], [48], [49], [81], [120], [160]
			MSD[78]	[133]
	Breast	Ultrasound	BUID[79]	[138]
	Lung	X-ray	JSRT[80]	[138]
	Surgical Scene	Video Frames	EndoVis-Robot[81]	[139], [133]
	Gastrointestinal	Endoscopy	KvaSir[82]	[138]
Detection	Brain	EEG	CHB-MIT[83]	[106]
			TUSZ[84]	[106]
	Chest	X-ray	COVID-QU-ex[85]	[181]
			BIMCV[86]	[166], [168]
			Hannover-CV[87]	[166]
			COVID-CT[88]	[166]
			ActualMed[89]	[166]
			RSNA-BA[90]	[166]
			VinDr[91]	[138]

Continued on next page

[64] , https://doi.org/10.17605/OSF.IO/NQJYW
[65] VerSe: Large Scale Vertebrae Segmentation Challenge, https://github.com/anjany/verse
[66] Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI, https://emidec.com/
[67] Right Ventricle Segmentation Challenge, https://rvsc.projets.litislab.fr/
[68] Automated Cardiac Diagnosis Challenge, https://www.creatis.insa-lyon.fr/Challenge/acdc/
[69] Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge 2020, https://www.ub.edu/mnms/
[70] Multi-sequence Cardiac MR Segmentation Challenge, https://zmiclab.github.io/zxh/0/mscmrseg19/
[71] Multi-Modality Whole Heart Segmentation, https://zmiclab.github.io/zxh/0/mmwhs/
[72] Signal processing laboratory, Brno University of Technology, http://splab.cz/en/research/zpracovani-medicinskych-signalu/databaze/artery
[73] Cardiac Delayed Enhancement Segmentation Challenge, https://figshare.com/articles/dataset/4214532
[74] Atrial Segmentation Challenge, https://www.cardiacatlas.org/atriaseg2018-challenge/
[75] MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, https://doi.org/10.7303/syn3193805
[76] LiTS - Liver Tumor Segmentation Challenge, http://www.lits-challenge.com/
[77] Combined (CT-MR) Healthy Abdominal Organ Segmentation challenge, https://chaos.grand-challenge.org/
[78] The Medical Segmentation Decathlon, http://medicaldecathlon.com/
[79] Breast Ultrasound Images Dataset, https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset
[80] Japanese Society of Radiological Technology, http://db.jsrt.or.jp/eng.php
[81] Robotic Scene Segmentation Sub-Challenge, https://endovissub2018-roboticscenesegmentation.grand-challenge.org/
[82] A Multi-Class Image-Dataset for Computer Aided Gastrointestinal Disease Detection, https://datasets.simula.no/kvasir/
[83] Children's Hospital Boston (CHB)-MIT dataset, https://doi.org/10.13026/C2K01R
[84] Temple University Hospital (TUH) EEG Seizure Corpus dataset, https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml
[85] COVID-QU-ex, https://doi.org/10.34740/kaggle/dsv/3122958
[86] A large annotated dataset of RX and CT images of COVID19 patients, https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/
[87] COVID-19 Image Repository, https://doi.org/10.25835/0090041
[88] A CT Scan Dataset about COVID-19, https://github.com/UCSD-AI4H/COVID-CT
[89] Actualmed COVID-19 Chest X-ray Dataset Initiative, https://github.com/agchung/Actualmed-COVID-chestxray-dataset
[90] RSNA Pediatric Bone Age Challenge (2017), https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017
[91] VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, https://vindr.ai/datasets/cxr

Task	Organ	Modality	Abbreviation	Ref.
Detection	Breast	Histology	TUPAC[92]	[50], [129]
	Breast	X-ray	INbreast[93]	[90]
	Tissue	Histology	MIDOG[94]	[27], [129]
			MITOS-ATYPIA-14[95]	[129]
			CCMCT[96]	[129]
Restoration	Retinal	Fundus	DRIVE[97]	[33], [34], [82], [84], [98], [138], [35], [118]
			Kag-Cat[98]	[33]
Localization	Surgical Scene	Video Frames	SurgicalActions160[99]	[128]
			Cataract-101[100]	[128]
Reconstruction	Blood	Histology	MalariaScreener[101]	[92]
	Brain	MRI	BrainWeb[102]	[38]

End of table

[92] Tumor Proliferation Assessment Challenge, https://tupac.grand-challenge.org

[93] INbreast: toward a full-field digital mammographic database, https://www.kaggle.com/datasets/ramanathansp20/inbreast-dataset

[94] The Mitosis Domain Generalization Challenge, https://imig.science/midog/

[95] MITOS-ATYPIA 2014 challenge of ICPR, https://mitos-atypia-14.grand-challenge.org/Dataset/

[96] A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor, https://github.com/DeepMicroscopy/MITOS_WSI_CCMCT

[97] Digital Retinal Images for Vessel Extraction, https://drive.grand-challenge.org/

[98] Kaggle cataract dataset, https://www.kaggle.com/datasets/jr2ngb/cataractdataset

[99] The ITEC SurgicalActions160 Dataset, https://ftp.itec.aau.at/datasets/SurgicalActions160/

[100] Cataract-101 Video Dataset, https://ftp.itec.aau.at/datasets/ovid/cat-101/

[101] Malaria Screener, https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html#malaria-datasets

[102] BrainWeb, https://brainweb.bic.mni.mcgill.ca/brainweb/