Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation

Xiao Liu[1], Spyridon Thermos[1], Alison O'Neil[1,3] and Sotirios A. Tsaftaris[1,2]

[1] School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK
[2] The Alan Turing Institute, London, UK
[3] Canon Medical Research Europe Ltd., Edinburgh, UK
Xiao.Liu@ed.ac.uk, SThermos@ed.ac.uk, Alison.ONeil@mre.medical.canon,
S.Tsaftaris@ed.ac.uk

Abstract. Generalising deep models to new data from new centres (termed here domains) remains a challenge. This is largely attributed to shifts in data statistics (domain shifts) between source and unseen domains. Recently, gradient-based meta-learning approaches where the training data are split into meta-train and meta-test sets to simulate and handle the domain shifts during training have shown improved generalisation performance. However, the current fully supervised meta-learning approaches are not scalable for medical image segmentation, where large effort is required to create pixel-wise annotations. Meanwhile, in a low data regime, the simulated domain shifts may not approximate the true domain shifts well across source and unseen domains. To address this problem, we propose a novel semi-supervised meta-learning framework with disentanglement. We explicitly model the representations related to domain shifts. Disentangling the representations and combining them to reconstruct the input image allows unlabeled data to be used to better approximate the true domain shifts for meta-learning. Hence, the model can achieve better generalisation performance, especially when there is a limited amount of labeled data. Experiments show that the proposed method is robust on different segmentation tasks and achieves state-of-the-art generalisation performance on two public benchmarks. Code is publicly available at: https://github.com/vios-s/DGNet.

Keywords: Domain generalisation · Disentanglement · Medical image segmentation.

1 Introduction

Despite recent progress in medical image segmentation [3,8,18], inference performance on unseen datasets, acquired from distinct scanners or clinical centres, is known to decrease [5,38]. Such reduction is mainly caused by shifts in data statistics between different clinical centres i.e. *domain shifts* [45], due to variation in patient populations, scanners, and scan acquisition settings [42]. The variation

in population impacts the underlying anatomy and pathology due to factors such as gender, age, ethnicity, which may differ for patients in different locations [28,39,43]. The variation in scanners and scan acquisition settings impacts the characteristics of the acquired image, such as brightness and contrast [45].

The naive approach to handling domain shift is to acquire and label as many and diverse data as possible, the cost implications and difficulties of which are known to this community. Alternatively one can train a model on source domains to generalise for a target domain with some information on the target domain available i.e. domain adaptation [4] such as cross-site MRI harmonisation [37] to enforce the source and target domains to share similar image-specific characteristics [9]. A more strict alternative is to *not use any* information for the target domain, known as *domain generalisation* [22]. Herein, we focus on this more challenging and more widely applicable approach.

In domain generalisation, the overarching goal is to identify suitable representations that encode information about the task at hand whilst being insensitive to domain-specific information. There are several active research directions aiming to address this goal, including: direct augmentation of the source domain data [45], feature space regularisation [6,14,27,35,46], alignment of the source domain features or output distributions [25], and learning domain-invariant features with gradient-based meta-learning [11,24,29]. Of the above, gradient-based meta-learning methods have the advantage of not overfitting to dominant source domains which account for the more populous data in the training dataset [11]. Gradient-based meta-learning [22,26] exploits an episodic training paradigm [23] by splitting the source domains into meta-train and meta-test sets at each iteration. The model is trained to handle domain shift by simulating it during training. By using constraints to implicitly eliminate the information related to the simulated domain shifts, the model can learn to extract domain-invariant features. Previous work introduced different constraints in a fully supervised setting e.g. global class alignment and local sample clustering [11], shape-aware constraints [29] or simply the task objective [20,22], where [20] extends [22] to medical image segmentation.[1] These approaches do not scale in medical image segmentation as pixel-wise annotation is time-consuming, laborious, and requires expert knowledge. Meanwhile, in a low data regime where centres only provide a few labeled data samples, these methods may only learn to extract domain-invariant features from an under-represented data distribution [20,45]. In other words, the simulated domain shifts may not well approximate the true domain shifts between source and unseen domains.

To address this problem, we propose to explicitly disentangle the representations related to domain shifts for meta-learning as we illustrate in Fig. 1. Learning these complete and sufficient representations [1] via reconstruction brings the benefit of unsupervised learning, thus we can better simulate the domain shifts by also using unlabeled data from any of the source domains. We consider two sources of shifts: one due to scanner and scan acquisition setting varia-

---

[1]With the exception of [41] which clusters unlabeled data to generate pseudo labels, but unfortunately is not applicable to segmentation.
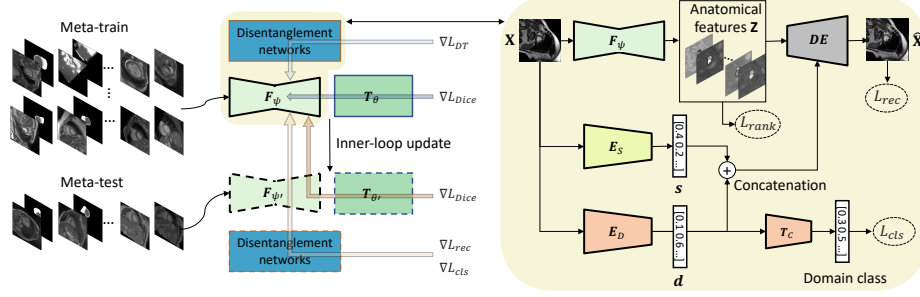
Fig. 1. At each iteration, the training dataset is split into meta-train and meta-test sets including labeled and unlabeled data. A feature network $\boldsymbol{F_\psi}$ extracts features $\mathbf{Z}$ for a task network $\boldsymbol{T_\theta}$ to predict segmentation masks. The model is trained in a semi-supervised setting, where $\mathcal{L}_{DT}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$ do not require pixel-wise annotation. In the inner-loop update, ψ' and θ' are computed for the meta-test step (see Eq. 1). Finally, all the gradients are computed to update $\boldsymbol{F_\psi}$ and $\boldsymbol{T_\theta}$ as in Eq. 2. The disentanglement networks decompose image $\mathbf{X}$ to common $\mathbf{s}$ and specific to the domain $\mathbf{d}$ representations to be disentangled with $\mathbf{Z}$ for meta-train and meta-test sets with the constraints ($\mathcal{L}_{DT}$ and $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$). See Section 2 for loss definitions.

tion, and one due to population variation. Because our task is segmentation, we want to be sensitive to changes in anatomy but insensitive to changes in imaging characteristics be it some common across domains or domain-specific. We use spatial (grid-like) features as a representation of anatomy ($\mathbf{Z}$) and two vectors ($\mathbf{s}$,$\mathbf{d}$) to encode common or domain-specific imaging characteristics. We apply specific design and learning biases to disentangle the above. For example, a spatial $\mathbf{Z}$ is equivariant to segmentation and this has been shown to improve performance [7,16]. We further encourage $\mathbf{Z}$ to be disentangled from $\mathbf{s}$ and $\mathbf{d}$ by exploiting also low-rank regularisation [25]. Gradient-based meta-learning also encourages $\mathbf{Z}$, $\mathbf{s}$, and $\mathbf{d}$ to generalise well to unseen domains whilst at the same time improves (implicitly) their disentanglement. Our **contributions** are summarised as follows:

1. We propose the first, to the best of our knowledge, semi-supervised domain-generalisation framework combining meta-learning and disentanglement.

2. Use of low-rank regularisation as a learning bias to encourage better disentanglement and hence improved generalisation performance.

3. Extensive experiments on cardiac and gray matter datasets show improved performance over several baselines especially for the limited annotated data case.

2   Proposed method

Consider a multi-domain training dataset $\mathcal{D} = \{\mathbf{X}_i^k, \mathbf{Y}_i^k\}_{i=1}^{N_k}, k \in \{1, 2, \cdots, K\}$ that is defined on a joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathbf{X}_i^k$ is the ith training datum from the kth source domain with corresponding ground-truth segmentation mask $\mathbf{Y}_i^k$,

and N_k denotes the number of training samples in the kth source domain. We aim to learn a model containing a feature network $\boldsymbol{F}_\psi : \mathcal{X} \to \mathcal{Z}$ to extract the anatomical features $\mathbf{Z}$ and a task network $\boldsymbol{T}_\theta : \mathcal{Z} \to \mathcal{Y}$ to predict the segmentation masks, where ψ and θ denote the network parameters.

2.1   Gradient-based meta-learning for domain generalisation

In gradient-based meta-learning for domain generalisation, the domain shift is simulated by training the model on a sequence of episodes [23, 24]. Specifically, the meta-train set $\mathcal{D}_{tr}$ and the meta-test set $\mathcal{D}_{te}$ are constructed by randomly splitting the source domains $\mathcal{D}$ for each iteration of training. Each iteration comprises a meta-train step followed by a meta-test step. For the meta-train step, the parameters ψ and θ of $\boldsymbol{F}_\psi$ and $\boldsymbol{T}_\theta$ are calculated by optimising the meta-train loss $\mathcal{L}_{meta-train}$ with data from $\mathcal{D}_{tr}$ (inner-loop update), as defined by:

$$(\psi', \theta') = (\psi, \theta) - \alpha \nabla_{\psi,\theta} \mathcal{L}_{meta-train}(\mathcal{D}_{tr}; \psi, \theta), \qquad (1)$$

where α is the learning rate for the meta-train update step. Typically, $\mathcal{L}_{meta-train}$ is the task objective, e.g. the Dice loss [10] for a segmentation task. This step rewards accurate predictions on the meta-train source domains. For the meta-test step, the meta-test source domains $\mathcal{D}_{te}$ are processed by the updated parameters (ψ', θ') and the model is expected to contain certain properties quantified by the $\mathcal{L}_{meta-test}$ loss. $\mathcal{L}_{meta-test}$ is computed using the updated parameters (ψ', θ'), whilst the gradients are computed towards the original parameters (ψ, θ). The final objective is defined as:

$$\underset{\psi,\theta}{\operatorname{argmin}} \, \mathcal{L}_{meta-train}(\mathcal{D}_{tr}; \psi, \theta) + \mathcal{L}_{meta-test}(\mathcal{D}_{te}; \psi', \theta'). \qquad (2)$$

The intuition behind this scheme is that the model should not only perform well on the source domains, but its future updates should also generalise well to unseen domains. Below, we will describe our meta-train and meta-test objectives but first we present how we disentangle representations related to domain shifts.

2.2   Learning disentangled representations

To model appearance in a single-domain setting, typically a single vector-based variational representation is used [7]. Here, due to our multi-domain setting, inspired by [17, 44], we separately encode domain-specific imaging characteristics as an additional vector-based variational representation. Hence, we aim to learn two independent vector representations, where one ($\mathbf{s}$) captures common imaging characteristics across domains and the other one ($\mathbf{d}$) captures specific imaging characteristics for each domain. In addition, we encode spatial anatomy information in a separate representation $\mathbf{Z}$, which we encourage to be disentangled from $\mathbf{s}$ and $\mathbf{d}$.

In particular, the input image $\mathbf{X}$ is first encoded in a common (appearance) representation $\mathbf{s} = \boldsymbol{E}_S(\mathbf{X})$, and a domain representation $\mathbf{d} = \boldsymbol{E}_D(\mathbf{X})$

that is followed by a shallow domain classifier $\boldsymbol{T}_C(\mathbf{d})$ which predicts the source domain ($\hat{\mathbf{c}}$) label of $\mathbf{X}$. Then, a decoder $\boldsymbol{DE}$ combines the extracted features $\mathbf{Z} = \boldsymbol{F}_\psi(\mathbf{X})$ and the representations $\mathbf{s}$ and $\mathbf{d}$ to reconstruct the input image, i.e. $\hat{\mathbf{X}} = \boldsymbol{DE}(\mathbf{Z}, \mathbf{s}, \mathbf{d})$. Note that $\boldsymbol{DE}$ combines $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$ using adaptive instance normalisation (AdaIN) layers [15]. As shown in [16], AdaIN improves disentanglement and encourages $\mathbf{Z}$ to encode spatially equivariant information, i.e. anatomical information useful for segmentation, and $\mathbf{s}, \mathbf{d}$ to only encode common or domain-specific appearance.

To achieve such "triple" disentanglement we consider several losses: **1)** KL divergences $\mathcal{L}_{KL}(\mathbf{s}, N(0,1)), \mathcal{L}_{KL}(\mathbf{d}, N(0,1))$ to induce a Gaussian $N(0,1)$ prior in $\mathbf{s}$ and $\mathbf{d}$, encouraging the representations to be robust on unseen domains [13]; **2)** Hilbert-Schmidt Independence Criterion (HSIC) loss $\mathcal{L}_{HSIC}(\mathbf{s}, \mathbf{d})$, to force $\mathbf{s}$ and $\mathbf{d}$ to be independent from each other [34]; **3)** a classification loss $\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}})$ such that the domain representation $\mathbf{d}$ is highly correlated with the domain-specific information [17]; and **4)** a reconstruction loss $\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}})$, defined as the ℓ_1 distance between $\mathbf{X}$ and $\hat{\mathbf{X}}$, to learn representations without supervision [7,16].

We further encourage the extracted features $\mathbf{Z}$ to be equivariant across the meta-train source domains and improve disentanglement between $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$ by applying rank regularisation [25]. Specifically, consider a batch $\{\mathbf{X}_{i_1}^1, \mathbf{X}_{i_2}^2, \cdots, \mathbf{X}_{i_{K_{tr}}}^{K_{tr}}\}$ from K_{tr} meta-train source domains, and K_{tr} features $\{\mathbf{Z}_{i_1}^1, \mathbf{Z}_{i_2}^2, \cdots, \mathbf{Z}_{i_{K_{tr}}}^{K_{tr}}\}$ extracted using the feature network $\boldsymbol{F}_\psi$. By flattening and concatenating these features, we end up with a matrix $\mathbb{Z}$ with dimensions $[C, K_{tr} \times H \times W]$, where C, H, W denote the number of channels, height, and width of $\mathbb{Z}$. Then, by forcing the rank of $\mathbb{Z}$ to be m (i.e. the number of the segmentation classes), $\mathbf{Z}$ is encouraged to encode only globally-shared information across K_{tr} source domains in order to predict the segmentation mask as discussed in [25]. We achieve that by minimising the $(m+1)^{th}$ singular value σ_{m+1} of $\mathbb{Z}$, i.e. $\mathcal{L}_{rank} = \sigma_{m+1}$. Overall, $\mathcal{L}_{DT}$ is defined as:

$$
\begin{aligned}
\mathcal{L}_{DT} =& \lambda_{rank}\mathcal{L}_{rank}(\mathbf{Z}) + \lambda_{KL}(\mathcal{L}_{KL}(\mathbf{s}, N(0,1)) + \mathcal{L}_{KL}(\mathbf{d}, N(0,1))) \\
& + \lambda_{rec}\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}}) + \lambda_{HSIC}\mathcal{L}_{HSIC}(\mathbf{s}, \mathbf{d}) + \lambda_{cls}\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}}),
\end{aligned}
\tag{3}
$$

where $\mathbf{c}$ is the domain label. We adopt hyperparameter values according to our extensive early experiments and discussion from [7,25] as $\lambda_{rank} = 0.1$, $\lambda_{KL} = 0.1$, $\lambda_{rec} = 1$ and $\lambda_{cls} = 1$. Note that all the losses do not need ground-truth masks. The domain class label is available, as we know the centre where the data belong.

2.3   Meta-train and meta-test objectives

Our *meta-train* objective contains two components:

$$
\mathcal{L}_{meta-train} = \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \mathcal{L}_{DT},
\tag{4}
$$

where $\lambda_{Dice} = 5$ when labeled data are available.

For the *meta-test* step, the model is expected to: **1)** accurately predict segmentation masks (by applying the task objective), and **2)** disentangle $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$

to the same level as meta-train sets. A naive strategy for the latter is to use $\mathcal{L}_{DT}$ for meta-test sets. However, as analysed in [2, 29], the meta-test step is unstable to train: the gradients from the meta-test loss are second-order statistics of ψ and θ. Our experiments revealed that including the unsupervised losses $\mathcal{L}_{KL}$ and $\mathcal{L}_{HSIC}$ make training even more unstable (even leading to model collapse). In addition, we use one domain for meta-test in experiments, while $\mathcal{L}_{rank}$ requires multiple domains. According to [31, 33], considering fixed learning and design biases, the level of disentanglement can be proxied by the reconstruction quality (with ground-truth image $\mathbf{X}$) and the domain classification accuracy (with ground-truth label $\mathbf{c}$). Hence, we adopt as the meta-test loss:

$$\mathcal{L}_{meta-test} = \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_{rec}\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}}) + \lambda_{cls}\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}}). \quad (5)$$

Note that for unlabeled data, $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$ do not need ground-truth masks.

3   Experiments

3.1   Tasks and datasets

Multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5]: The M&Ms challenge dataset contains 320 subjects. Subjects were scanned at 6 clinical centres in 3 different countries using 4 different magnetic resonance scanner vendors (Siemens, Philips, GE, and Canon) i.e. domains A, B, C and D. For each subject, only the end-systole and end-diastole phases are annotated. Voxel resolutions range from $0.85 \times 0.85 \times 10$ mm to $1.45 \times 1.45 \times 9.9$ mm. Domain A contains 95 subjects. Domain B contains 125 subjects. Both domains C and D contain 50 subjects.
Spinal cord gray matter segmentation (SCGM) dataset [38]: The data from SCGM [38] are collected from 4 different medical centres with different MRI systems (Philips Achieva, Siemens Trio, Siemens Skyra) i.e. domains 1, 2, 3 and 4. The voxel resolutions range from $0.25 \times 0.25 \times 2.5$ mm to $0.5 \times 0.5 \times 5$ mm. Each domain has 10 labeled subjects and 10 unlabelled subjects.

3.2   Baseline models

nnUNet [19]: is a self-adapting framework based on 2D and 3D U-Nets [40] which does not specifically target domain generalisation. Given a labelled training dataset, nnUNet automatically adapts its model design and hyperparameters to obtain optimal performance. In the M&Ms challenge, methods based on nnUNet achieved the top performance [5].
SDNet+Aug. [30]: disentangles the input image to a spatial anatomy and a non-spatial modality factors. Here we use intensity- and resolution- augmented data in a semi-supervised setting. Compared to our method, "SDNet+Aug." only poses disentanglement to the latent features without meta-learning.
LDDG [25]: is the latest state-of-the-art model for domain-generalised medical image analysis. It also uses a rank loss and when applied in a fully supervised

Table 1. Dice (%) results and the standard deviations on M&Ms dataset. For "SD-Net+Aug." and our method, the training data contain all the unlabeled data and 2% or 5% of labeled data from source domains. The other models are trained by 2% or 5% labeled data only. Bold numbers denote the best performance.

	Source	Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
	B,C,D	A	52.87_{19}	54.48_{18}	59.47_{12}	56.31_{13}	$\mathbf{66.01_{12}}$
	A,C,D	B	64.63_{17}	67.81_{14}	56.16_{14}	56.32_{15}	$\mathbf{72.72_{10}}$
2%	A,B,D	C	72.97_{14}	76.46_{12}	68.21_{11}	$75.70_{8.7}$	$\mathbf{77.54_{10}}$
	A,B,C	D	73.27_{11}	74.35_{11}	68.56_{10}	$69.94_{9.8}$	$\mathbf{75.14_{8.4}}$
	Average		$65.94_{8.3}$	$68.28_{8.6}$	$63.16_{5.4}$	$64.57_{8.5}$	$\mathbf{72.85_{4.3}}$
	B,C,D	A	65.30_{17}	71.21_{13}	$66.22_{9.1}$	67.11_{10}	$\mathbf{72.40_{12}}$
	A,C,D	B	79.73_{10}	77.31_{10}	$69.49_{8.3}$	$76.35_{7.9}$	$\mathbf{80.30_{9.1}}$
5%	A,B,D	C	78.06_{11}	$81.40_{8.0}$	$73.40_{9.8}$	$77.43_{8.3}$	$\mathbf{82.51_{6.6}}$
	A,B,C	D	$81.25_{8.3}$	$79.95_{7.8}$	$75.66_{8.5}$	$78.64_{5.8}$	$\mathbf{83.77_{5.1}}$
	Average		$76.09_{6.3}$	$77.47_{3.9}$	$71.29_{3.6}$	$74.88_{4.6}$	$\mathbf{79.75_{4.4}}$

setting, LDDG achieved the best generalisation performance on SCGM.
SAML [29]: is another gradient-based meta-learning approach. SAML proposed to constraint the compactness and smoothness properties of segmentation masks across meta-train and meta-test sets in a fully supervised setting.

3.3   Implementation details

Models are trained using the Adam optimiser [21] with a learning rate of $2e^{-5}$ for 50K iterations using batch size 4. Images are cropped to 224×244 for M&Ms and 144×144 for SCGM. $\boldsymbol{F}_\psi$ is a 2D UNet [40] to extract $\mathbf{Z}$ features with 8 channels of same height and width as input image. We follow the designs of SDNet [7] for $\boldsymbol{E}_S$, $\boldsymbol{T}_\theta$ and $\boldsymbol{DE}$. $\boldsymbol{E}_D$ has the same architecture as $\boldsymbol{E}_S$. Both $\mathbf{s}$ and $\mathbf{d}$ have 8 dimensions. $\boldsymbol{T}_C$ is a single fully-connected layer. All models are implemented in PyTorch [36] and are trained using an NVidia 2080 Ti GPU. In the semi-supervised setting, we use specific percentages of the subjects as labeled data and the rest as unlabeled data. We use Dice (%) and Hausdorff Distance [12] as the evaluation metrics.

3.4   Results and discussion

Tables 1, 2, 3 and 4 show that we consistently achieve the best generalisation performance on cardiac and gray matter segmentation. Particularly in the low data regime we improve Dice by $\approx 5\%$ on M&Ms and $\approx 3\%$ on SCGM compared to the best performing baseline. For 100% annotations in M&Ms (see Appendix), our model still outperforms the baselines. We also show the qualitative results in Appendix, where the improved performance is visually observed.
M&Ms: Compared to "SDNet+Aug." which can also use (due to disentanglement) unlabeled data, our model performs consistently better. The results agree with the conclusion in [32]: without specific designs tuned to the tasks, disentanglement can not provide guaranteed generalisation ability. For LDDG and

Table 2. Hausdorff distance results and the standard deviations on M&Ms dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 2% or 5% of labeled data from source domains. The other models are trained by 2% or 5% labeled data only. Bold numbers denote the best performance.

	Source	Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
2%	B,C,D	A	$26.48_{7.5}$	$24.69_{7.0}$	$25.56_{5.9}$	$25.57_{5.7}$	$\mathbf{23.55_{6.5}}$
	A,C,D	B	$23.11_{6.8}$	$21.84_{6.2}$	$25.44_{5.2}$	$24.91_{5.5}$	$\mathbf{19.95_{6.3}}$
	A,B,D	C	$16.75_{4.6}$	$16.57_{4.2}$	$18.98_{3.9}$	$16.46_{3.5}$	$\mathbf{16.29_{4.0}}$
	A,B,C	D	$17.51_{4.9}$	$17.57_{4.1}$	$18.08_{3.8}$	$17.94_{3.8}$	$\mathbf{17.48_{4.7}}$
	Average		$20.96_{4.0}$	$20.17_{3.3}$	$22.02_{3.5}$	$21.22_{4.1}$	$\mathbf{19.32_{2.8}}$
5%	B,C,D	A	$23.04_{6.7}$	$22.84_{6.3}$	$23.35_{5.7}$	$23.10_{5.9}$	$\mathbf{22.55_{6.6}}$
	A,C,D	B	$\mathbf{18.18_{4.7}}$	$20.26_{5.5}$	$20.56_{4.7}$	$18.97_{4.9}$	$19.37_{6.4}$
	A,B,D	C	$16.44_{4.2}$	$16.22_{3.9}$	$17.14_{3.3}$	$16.29_{3.2}$	$\mathbf{15.77_{3.8}}$
	A,B,C	D	$15.24_{4.2}$	$15.15_{3.3}$	$15.80_{3.2}$	$15.58_{3.2}$	$\mathbf{14.24_{2.8}}$
	Average		$18.22_{3.0}$	$18.62_{3.1}$	$19.21_{3.0}$	$18.49_{2.9}$	$\mathbf{17.98_{3.2}}$

Table 3. Dice (%) results and the standard deviations on SCGM dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 20% or 100% of labeled data from source domains. The other models are trained by 20% or 100% of labeled data only. Bold numbers denote the best performance.

	Source	Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
20%	2,3,4	1	59.07_{21}	83.07_{16}	$77.71_{9.1}$	78.71_{25}	$\mathbf{87.45_{6.3}}$
	1,3,4	2	69.94_{12}	$80.01_{5.2}$	44.08_{12}	75.58_{12}	$\mathbf{81.05_{5.2}}$
	1,2,4	3	$60.25_{7.2}$	58.57_{10}	$48.04_{5.5}$	$54.36_{7.6}$	$\mathbf{61.85_{7.3}}$
	1,2,3	4	$70.13_{4.3}$	$85.27_{2.2}$	$83.42_{2.7}$	$85.36_{2.8}$	$\mathbf{87.96_{2.1}}$
	Average		$64.85_{5.2}$	76.73_{11}	63.31_{17}	73.50_{12}	$\mathbf{79.58_{11}}$
100%	2,3,4	1	$75.27_{8.3}$	$\mathbf{90.25_{4.5}}$	$88.21_{4.9}$	$90.22_{5.6}$	$90.01_{4.9}$
	1,3,4	2	$76.32_{2.9}$	$84.13_{4.2}$	$83.76_{3.1}$	$\mathbf{86.65_{3.5}}$	$85.48_{2.3}$
	1,2,4	3	$62.59_{6.9}$	62.18_{10}	$56.11_{9.3}$	$58.27_{9.4}$	$\mathbf{64.23_{9.7}}$
	1,2,3	4	$71.87_{2.5}$	$88.93_{1.9}$	$89.08_{2.7}$	$88.66_{2.6}$	$\mathbf{89.26_{2.5}}$
	Average		$71.51_{5.4}$	81.37_{11}	79.29_{13}	80.95_{13}	$\mathbf{82.25_{11}}$

SAML, the generalisation performance significantly drops with small amounts of labeled data. Note that nnUNet adapts the model design per each run/training set. However, adapting the model design for different training data limits the scalability of nnUNet. In the Appendix, we also show that nnUNet possibly overfits the source domains in some cases.

SCGM: We obtain consistent improvements also on SCGM, demonstrating application in other organs. Our model benefits from the additional 10 unlabeled subjects of each domain leading to better performance overall.

3.5   Ablation study

Here we conduct ablations on key losses crucial to disentanglement and the extraction of good anatomical features for good generalisation performance. We omit ablations on the KL losses as [13, 17] showcase that variational encoding

Table 4. Hausdorff distance results and the standard deviations on SCGM dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 20% or 100% of labeled data from source domains. The other models are trained by 20% or 100% of labeled data only. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
20%	2,3,4	1	$3.09_{0.25}$	$1.52_{0.33}$	$1.75_{0.26}$	$1.53_{0.38}$	$\mathbf{1.50_{0.30}}$
	1,3,4	2	$3.16_{0.09}$	$1.97_{0.16}$	$2.73_{0.33}$	$2.07_{0.35}$	$\mathbf{1.91_{0.16}}$
	1,2,4	3	$3.38_{0.27}$	$2.45_{0.27}$	$2.67_{0.25}$	$2.52_{0.24}$	$\mathbf{2.23_{0.23}}$
	1,2,3	4	$4.31_{0.14}$	$2.34_{0.21}$	$2.37_{0.14}$	$2.30_{0.18}$	$\mathbf{2.22_{0.13}}$
	Average		$3.49_{0.49}$	$2.07_{0.36}$	$2.38_{0.39}$	$2.11_{0.37}$	$\mathbf{1.97_{0.30}}$
100%	2,3,4	1	$3.26_{0.21}$	$\mathbf{1.37_{0.25}}$	$1.50_{0.23}$	$1.43_{0.36}$	$1.43_{0.29}$
	1,3,4	2	$3.19_{0.09}$	$1.88_{0.16}$	$2.19_{0.19}$	$\mathbf{1.80_{0.19}}$	$1.81_{0.15}$
	1,2,4	3	$3.37_{0.27}$	$2.34_{0.24}$	$2.64_{0.28}$	$2.43_{0.33}$	$\mathbf{2.23_{0.32}}$
	1,2,3	4	$4.30_{0.15}$	$2.13_{0.17}$	$2.12_{0.15}$	$2.15_{0.15}$	$\mathbf{2.11_{0.13}}$
	Average		$3.53_{0.45}$	$1.93_{0.36}$	$2.11_{0.41}$	$1.95_{0.38}$	$\mathbf{1.92_{0.31}}$

helps to learn robust vector representation for better generalisation. To illustrate that $\mathcal{L}_{rank}$ helps to disentangle $\mathbf{Z}$ to $(\mathbf{s}, \mathbf{d})$, and improves performance, we use Distance Correlation (DC) [31] to measure disentanglement (lower DC means a higher level of disentanglement). For M&Ms 5% cases, without $\mathcal{L}_{rank}$, the average DC on the test dataset between $\mathbf{Z}$ and $(\mathbf{s}, \mathbf{d})$ is 0.22 (an increase compared to 0.19 with $\mathcal{L}_{rank}$), and the average Dice is 78.54% (a decrease compared to 79.75% with $\mathcal{L}_{rank}$). We also ablate $\mathcal{L}_{cls}$ and $\mathcal{L}_{HSIC}$. The proposed model on M&Ms 5% cases had an average Dice 79.75% but without $\mathcal{L}_{cls}$, average Dice drops to 77.45% and without $\mathcal{L}_{HSIC}$, average Dice drops to 77.86%.

4   Conclusion

We have presented a novel semi-supervised meta-learning framework for domain generalisation. Using disentanglement our approach models domain shifts, and thanks to our reconstruction approach to disentanglement, our model can be trained also with unlabeled data. By applying the designed constraints (including the low-rank regularisation) to the gradient-based meta-learning approach, the model extracts robust anatomical features useful for predicting segmentation masks in a semi-supervised manner. Extensive quantitative results, especially when insufficient annotated data are available, indicate remarkable improvements compared to previous state-of-the-art approaches. The performance of our method might improve with the use of additional unlabeled data from other domains, which we leave as future work.

5   Acknowledgement

References

1. Achille, A., Soatto, S.: Emergence of invariance and disentanglement in deep representations. JMLR **19**(1), 1947–1980 (2018)
2. Antoniou, A., Edwards, H., Storkey, A.: How to train your maml. In proc. ICLR (2019)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE TMI **37**(11), 2514–2525 (2018)
4. Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., Zheng, Y.: Uncertainty-aware domain alignment for anatomical structure segmentation. MedIA **64**, 101732 (2020)
5. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. IEEE TMI (2021)
6. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalisation by solving jigsaw puzzles. In: Proc. CVPR. pp. 2229–2238 (2019)
7. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. MedIA **58** (2019)
8. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7**(25), 1–33 (2020)
9. Dewey, B.E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E.M., Newsome, S., Oh, J., Calabresi, P.A., Prince, J.L.: A disentangled latent space for cross-site mri harmonization. In: Proc. MICCAI. pp. 720–729. Springer (2020)
10. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
11. Dou, Q., Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalisation via model-agnostic learning of semantic features. Proc. NeurIPS (2019)
12. Dubuisson, M.P., Jain, A.K.: A modified hausdorff distance for object matching. In: In proc. ICPR. vol. 1, pp. 566–568. IEEE (1994)
13. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: Proc. ICLR. ICLR (2016)
14. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In proc. CVPR (2021)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: In proc. ICCV. pp. 1501–1510 (2017)
16. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proc. ECCV. pp. 172–189 (2018)
17. Ilse, M., Tomczak, J.M., Louizos, C., Welling, M.: Diva: Domain invariant variational autoencoders. In: Proc. MIDL. pp. 322–348. PMLR (2020)

18. Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H.: Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In: International workshop on STACOM. pp. 120–129. Springer (2017)

19. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)

20. Khandelwal, P., Yushkevich, P.: Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging. In: Proc. DART workshp, pp. 73–84. Springer (2020)

21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In proc. ICLR (2015)

22. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalise: Meta-learning for domain generalisation. In: Proc. AAAI (2018)

23. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalisation. In: Proc. ICCV. pp. 1446–1455 (2019)

24. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalisation with adversarial feature learning. In: Proc. CVPR. pp. 5400–5409 (2018)

25. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.C.: Domain generalisation for medical imaging classification with linear-dependency regularization. Proc. NeurIPS (2020)

26. Li, X., Yu, L., Jin, Y., Fu, C.W., Xing, L., Heng, P.A.: Difficulty-aware meta-learning for rare disease diagnosis. In: Proc. MICCAI. pp. 357–366. Springer (2020)

27. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: In proc. ECCV. pp. 624–639 (2018)

28. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In: Proc. MICCAI. pp. 614–623. Springer (2020)

29. Liu, Q., Dou, Q., Heng, P.A.: Shape-aware meta-learning for generalising prostate mri segmentation to unseen domains. In: Proc. MICCAI. pp. 475–485. Springer (2020)

30. Liu, X., Thermos, S., Chartsias, A., O'Neil, A., Tsaftaris, S.A.: Disentangled representations for domain-generalised cardiac segmentation. International workshop on STACOM (2020)

31. Liu, X., Thermos, S., Valvano, G., Chartsias, A., O'Neil, A., Tsaftaris, S.A.: Metrics for exposing the biases of content-style disentanglement. arXiv preprint arXiv:2008.12378 (2020)

32. Llera Montero, M., JH Ludwig, C., Ponte Costa, R., Malhotra, G., Bowers, J.: The role of disentanglement in generalisation. In: Proc. ICLR. ICLR (2021)

33. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: Proc. ICML. pp. 4114–4124. PMLR (2019)

34. Ma, W.D.K., Lewis, J., Kleijn, W.B.: The hsic bottleneck: Deep learning without back-propagation. In: Proc. AAAI. pp. 5085–5092 (2020)

35. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalisation via invariant feature representation. In: Proc. ICML. pp. 10–18. PMLR (2013)

36. Paszke, A., Gross, S., Massa, F., Lerer, A., et. al: PyTorch: An imperative style, high-performance deep learning library. In: In proc. NeurIPS. pp. 8026–8037 (2019)

37. Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., et al.: Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage **208**, 116450 (2020)
38. Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M.J., Conrad, B.N., Datta, E., Dávid, G., De Leener, B., et al.: Spinal cord grey matter segmentation challenge. Neuroimage **152**, 312–329 (2017)
39. Puyol-Anton, E., Ruijsink, B., Piechnik k., S., Neubauer, S., Petersen E., S., Razavi, R., King P., A.: Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. arXiv preprint arXiv:2106.12387 (2021)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: In proc. MICCAI. pp. 234–241. Springer (2015)
41. Sharifi-Noghabi, H., Asghari, H., Mehrasa, N., Ester, M.: Domain generalisation via semi-supervised meta learning. arXiv preprint arXiv:2009.12658 (2020)
42. Tao, Q., Yan, W., Wang, Y., Paiman, E.H., Shamonin, et al.: Deep learning–based method for fully automatic quantification of left ventricle function from cine mr images: a multivendor, multicenter study. Radiology **290**(1), 81–88 (2019)
43. Wang, J., Zhou, S., Fang, C., Wang, L., Wang, J.: Meta corrupted pixels mining for medical image segmentation. In: Proc. MICCAI. pp. 335–345. Springer (2020)
44. Yu, X., Chen, Y., Li, T., Liu, S., Li, G.: Multi-mapping image-to-image translation via learning disentanglement. Proc. NeurIPS (2019)
45. Zhang, L., Wang, X., Yang, D., Sanford, T., et al.: generalising deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE TMI **39**(7), 2531–2540 (2020)
46. Zhao, S., Gong, M., Liu, T., Fu, H., Tao, D.: Domain generalization via entropy regularization. In proc. NeurIPS **33** (2020)

6   Appendix

6.1   More quantitative results

In Table A5 and Table 2, we report the Dice (%) and Hausdorff Distance results on the cases of giving 100% labeled data for each model. For M&Ms, the unlabeled data (of labeled subjects) from phases between end-systole and end-diastole phases still gives the proposed model slightly better performance i.e. 0.65% of improvement compared with the best baseline model.

Table A5. Dice (%) results and the standard deviations on M&Ms dataset. For "SD-Net+Aug." and our method, the training data contains all the unlabeled data and all labeled data from source domains. The other models are trained by labeled data only. Bold numbers denote the best performance.

Source	Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
	B,C,D	A	$80.84_{\pm 11}$	$81.50_{\pm 7.7}$	$82.62_{\pm 6.3}$	$81.33_{\pm 7.2}$	$\mathbf{83.21_{\pm 7.4}}$
	A,C,D	B	$\mathbf{86.76_{\pm 5.8}}$	$85.04_{\pm 6.1}$	$85.68_{\pm 5.7}$	$84.15_{\pm 5.9}$	$86.53_{\pm 5.3}$
100%	A,B,D	C	$84.92_{\pm 7.1}$	$85.64_{\pm 6.5}$	$86.49_{\pm 6.3}$	$84.52_{\pm 6.2}$	$\mathbf{87.22_{\pm 6.1}}$
	A,B,C	D	$86.94_{\pm 5.9}$	$84.96_{\pm 5.2}$	$86.73_{\pm 6.1}$	$83.96_{\pm 5.9}$	$\mathbf{87.16_{\pm 4.9}}$
	Average		$84.87_{\pm 2.5}$	$84.29_{\pm 1.6}$	$85.38_{\pm 1.6}$	$83.49_{\pm 1.3}$	$\mathbf{86.03_{\pm 1.7}}$

Table A6. Hausdorff distance results and the standard deviations on M&Ms dataset. For "SDNet+Aug." and our method, the training data contains all the unlabeled data and all labeled data from source domains. The other models are trained by labeled data only. Bold numbers denote the best performance.

Source	Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
	B,C,D	A	$17.86_{\pm 5.5}$	$17.39_{\pm 4.5}$	$17.48_{\pm 4.1}$	$17.70_{\pm 4.2}$	$\mathbf{17.28_{\pm 3.9}}$
	A,C,D	B	$\mathbf{14.82_{\pm 3.4}}$	$15.55_{\pm 3.7}$	$15.42_{\pm 3.4}$	$16.05_{\pm 3.7}$	$14.99_{\pm 3.6}$
100%	A,B,D	C	$13.72_{\pm 3.3}$	$13.67_{\pm 3.0}$	$13.52_{\pm 2.8}$	$14.21_{\pm 3.3}$	$\mathbf{13.11_{\pm 2.8}}$
	A,B,C	D	$12.81_{\pm 3.4}$	$13.64_{\pm 2.9}$	$13.11_{\pm 3.0}$	$14.12_{\pm 2.8}$	$\mathbf{12.72_{\pm 2.6}}$
	Average		$14.80_{\pm 1.9}$	$15.06_{\pm 1.6}$	$14.88_{\pm 1.7}$	$15.52_{\pm 1.5}$	$\mathbf{14.53_{\pm 1.8}}$

6.2   Disentanglement networks

We use AdaIN module in the decoder $\boldsymbol{DE}$ to combine $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$. Each AdaIN layer performs the following operation:

$$\text{AdaIN} = \sigma(\mathbf{s}_i, \mathbf{d}_i)\frac{\mathbf{Z}_i - \mu(\mathbf{Z}_i)}{\sigma(\mathbf{Z}_i)} + \mu(\mathbf{s}_i, \mathbf{d}_i), \tag{6}$$

where each feature map $\mathbf{Z}_i$ is first normalized separately, and then is scaled and shifted based on the style representation and domain representation $\mathbf{s}_i, \mathbf{d}_i$ mean and standard deviation scalars.

**Table A7.** Dice (%) results and the standard deviations on M&Ms dataset.

	Source	Target A	Target B	Target C	Target D
100%	Trained on B	78.82_{11}	$94.58_{4.3}$	$83.60_{8.0}$	$85.81_{6.8}$
	Trained on D	80.03_{10}	$84.35_{7.0}$	$84.01_{8.9}$	$94.74_{5.4}$

6.3   Analysis of M&Ms data

To explore why nnUNet can outperform other models on cases of A,C,D to B and A,B,C to D, we train nnUNet models with 100% labeled data only from domain B or domain D. Then we test the models with data from domain A, B, C and D. As shown in Table A7, the model trained on domain B can achieve 85.81% Dice on domain D (highest Dice compared to A and C). Also, the model trained on domain D can achieve 84.35% Dice on domain B (highest Dice compared to A and C). Hence, nnUNet possibly overfits the source domains e.g. B, to achieve the best performance on domains (e.g. D) similar to B. However, other methods have to generalise to distinct domains i.e. A and C, which causes slightly worse performance on the domains similar to source domains.

6.4   Qualitative results

We show the qualitative results, i.e. predicted segmentation masks, in Fig. A2.
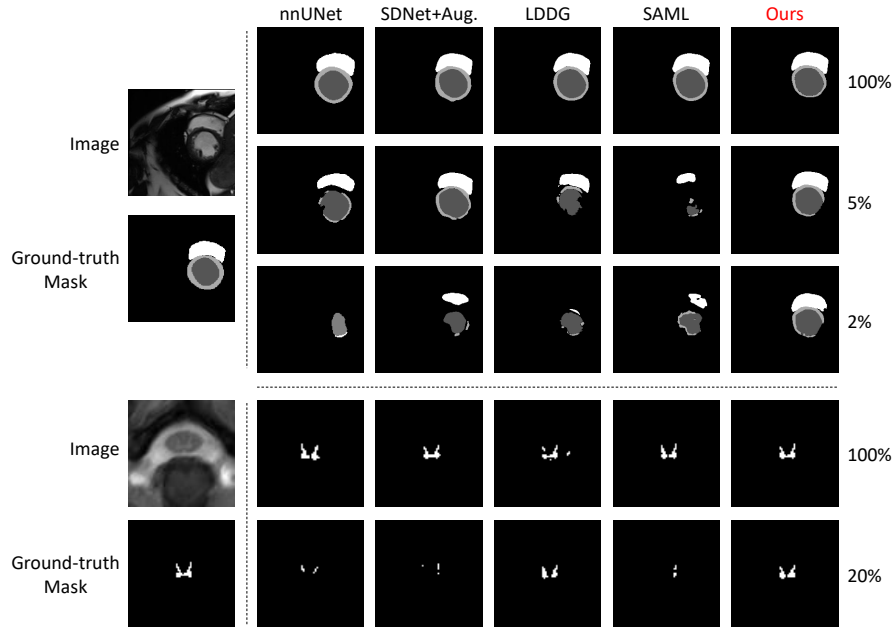
Fig. A2. We show the example images and predicted segmentation masks of each model for different cases. When training the baseline models with less labeled data, the performance drops significantly. In contrast, our model can produce satisfactory masks in every case.