Mutual Information-based Disentangled Neural Networks for Classifying Unseen Categories in Different Domains: Application to Fetal Ultrasound Imaging

Qingjie Meng, Jacqueline Matthew, Veronika A. Zimmer, Alberto Gomez, David F.A. Lloyd, Daniel Rueckert, *Fellow, IEEE*, and Bernhard Kainz, *Senior member, IEEE*

***Abstract*— Deep neural networks exhibit limited generalizability across images with different entangled domain features and categorical features. Learning generalizable features that can form universal categorical decision boundaries across domains is an interesting and difficult challenge. This problem occurs frequently in medical imaging applications when attempts are made to deploy and improve deep learning models across different image acquisition devices, across acquisition parameters or if some classes are unavailable in new training databases. To address this problem, we propose Mutual Information-based Disentangled Neural Networks (MIDNet), which extract generalizable categorical features to transfer knowledge to unseen categories in a target domain. The proposed MIDNet adopts a semi-supervised learning paradigm to alleviate the dependency on labeled data. This is important for real-world applications where data annotation is time-consuming, costly and requires training and expertise. We extensively evaluate the proposed method on fetal ultrasound datasets for two different image classification tasks where domain features are respectively defined by shadow artifacts and image acquisition devices. Experimental results show that the proposed method outperforms the state-of-the-art on the classification of unseen categories in a target domain with sparsely labeled training data.**

***Index Terms*— Representation disentanglement, domain adaptation, semi-supervised learning, image classification.**

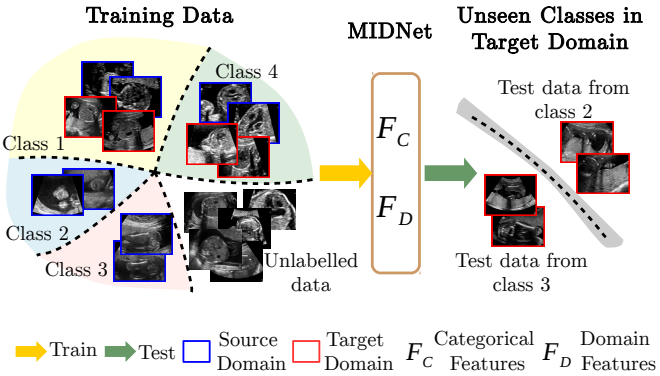


Fig. 1: The proposed method (MIDNet) learns to extract generalized features ($\mathcal{F}_C$, $\mathcal{F}_D$) across domains from sparsely labeled training data. Generalized categorical features ($\mathcal{F}_C$) are able to correctly classify unseen categories in the target domain. This is of particular importance for real-world applications such as improving diagnostic classification in medical imaging, especially when some categories are unavailable in new datasets for training.

I. Introduction

DEPLOYING deep neural networks (DNNs) in real-world clinical scenarios is challenging due to the problem of *domain shift* [1]. Domain shift corresponds to the feature distribution difference between training data and test data, which leads to performance degradation of DNNs from training to testing. This problem is ubiquitous in many clinical applications such as image classification [2]–[4] and image segmentation [5]–[9]. For example, performance degradation can be observed when applying a model that has been trained on images from one particular imaging device to images from another device. Addressing the problem of domain shift can contribute to a wider and efficient utilization of DNNs for image analysis in various clinical applications.

Domain shift can be categorized into (a) covariate shift (different latent feature distributions), (b) prior probability shift (change of labels) and (c) concept shift (different relationship between latent features and the desired label) [10], [11]. Covariate shift is the key reason for the lack of generalizability of DNNs. In medical imaging, covariate shift can be the result of the use of different imaging modalities (*e.g.*, magnetic resonance imaging and ultrasound), different image acquisition devices within the same modality or different combinations of specific image features (*e.g.*, anatomical structures and artifacts).

In contrast to the human visual system, DNNs exhibit weak

This work is supported by the Wellcome Trust IEH [102431], Intel and Nvidia. We thank the ultrasound specialists Emily Skelton, Tom Day and David F.A. Lloyd.

Q. Meng, D. Rueckert and B. Kainz are with Imperial College London, SW7 2AZ, UK, (e-mail: q.meng16|bkainz|drueckert@imperial.ac.uk).

D. Rueckert is also with Faculty of Informatics and Medicine, Technical University Munich, Germany.

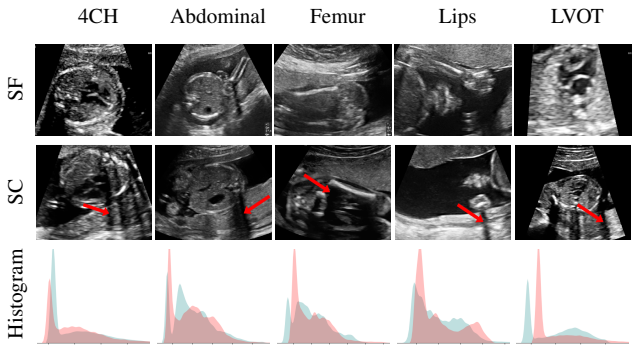
J. Matthew, V. Zimmer, A. Gomez, D. Lloyd are with King's College London, WC2R 2LS, UK. (e-mail: jacqueline.matthew|veronika.zimmer|alberto.gomez|david.lloyd@kcl.ac.uk)

generalizability when confronted with previously unseen entangled image features. This is the problem which we address in this paper. We postulate that DNNs should be able to learn generalizable features to transfer the knowledge from known entangled image features to new entangled image features. As outlined in Fig. 1, we want to improve the performance of DNNs on unseen categories in a target domain where all categories from a source domain and a subset of categories from a target domain are available for training. This task can greatly contribute to diagnostic classification in medical imaging. For example, detecting a certain pathology which rarely occurs in a particular geographic region but might be common in other places.

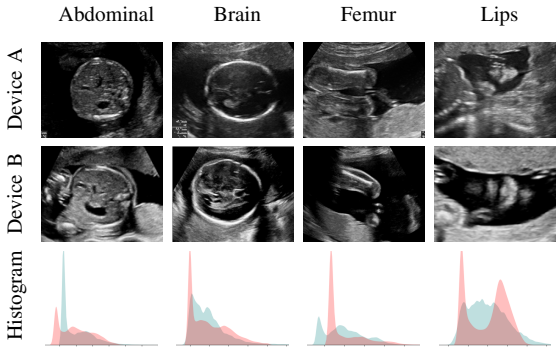
Fine-tuning DNNs on task-specific datasets is a possible solution but often infeasible due to the lack of sufficient annotated data in the target domain. Domain adaptation algorithms have been widely studied to tackle the domain shift problem by extracting domain-invariant features (*e.g.*, [5]–[9] in medical applications), aiming to transfer knowledge from a source domain to a target domain [12]. Previous approaches can be categorized into three main groups: (1) *Discrepancy measurement approaches* aim to align the feature distributions of source and target domain by measuring the discrepancy between representations, such as Maximum Mean Discrepancy (MMD) [13]–[15] or correlation distance [16]; (2) *Adversarial-based approaches* use DNNs to encourage the extracted features to be invariant for domain discrimination, instead of computing the discrepancy metric, and includes non-generative models [5], [17], [18] and generative models [19], [20]; (3) *Reconstruction-based approaches* align the source and the target domain by image reconstruction which uses a cycle-consistency constraint to preserve domain-invariant features [21], [22].

Existing domain adaptation methods can be practically prohibitive in real applications because a large amount of labeled data from source domains is needed. Although adversarial adaptation alternatives can perform well, optimizing adversarial objectives remains difficult and unstable in practice [23]. Most importantly, previous methods make no explicit attempt to disentangle domain-invariant features from domain features, which results in the inability of dealing with previously unseen categories in the target domain.

In this paper, we propose mutual information-based disentangled networks (MIDNet) for representation disentanglement to address the problem outlined in Fig. 1. The proposed approach extracts generalized categorical features by explicitly disentangling categorical features and domain features via mutual information minimization [24]. Note that the *categorical features* in this paper refer to the features relevant to identities of classes or categories. We introduce the supervision from labeled images for an enhanced disentanglement via a feature clustering module, which estimates the similarity of categorical features from both domains. Image reconstruction is used to guarantee that the separated categorical and domain features are not random noise and are representative for the input images. To further explore an improved categorical classification, we structure a categorical feature space by considering inter-class relationships. On top of the proposed



(a) Fetal US dataset



(b) Cross-device dataset

Fig. 2: Examples of fetal US images and intensity histograms. (a) The fetal US dataset, including *shadow-free* (SF) and *shadow-containing* (SC) images of different anatomical structures. Red arrows show examples of acoustic shadow artifacts. (b) The fetal US dataset with images acquired by different devices, device A (GE Voluson E8) and device B (Philips EPIQ V7 G).

MIDNet model, we incorporate distance metric learning to increase inter-class variance. The proposed method is a non-adversarial method which mitigates the difficulty and instability of adversarial model training. Our method is a semi-supervised learning method, which only requires a small number of labeled samples during training while unlabeled data is integrated using a strategy similar to the MixMatch approach [25].

We demonstrate the practical applicability of our method on a challenging medical application, the classification of standardized fetal ultrasound (US) views during prenatal screening. In many countries, US imaging is clinical routine for prenatal health care. The classification of standardized views is important for a consistent, cross-institutional identification of possible abnormalities [26]. Early detection of pathological development can inform following treatment and delivery options [27], [28]. DNNs have shown promising performance to support this task [29]. However, ultrasound images are often hard to interpreted [30]. Anatomical structures have inconsistent appearance [29] and contain different orientations and shapes of anatomical landmarks [31]. Labeled training data is often insufficient as annotating medical images requires significant expertise and is prohibitively expensive in both

time and labor. Manifestation of acoustic shadows [32], [33] as shown in Fig. 2(a) as well as different imaging devices as shown in Fig. 2(b) can lead to a domain shift problem for vanilla DNN classifiers. Exploring domain adaptation in fetal US enables DNN classifiers to be effectively utilized on a wider range, which supports identification of abnormalities from varying data sources. This can benefit prenatal healthcare.

The main contributions of this paper are summarized as follows:

- We investigate a challenging domain adaptation problem for medical image classification: the translation of decision boundaries to a target domain, which lacks training samples for several categories. We propose end-to-end trainable Mutual Information-based Disentangled Networks (MIDNet) for learning generalized categorical features to classify unseen categories in the target domain.
- We develop MIDNet as a non-adversarial learning approach to show a more effective alternative to difficult and unstable adversarial training. Mutual information is utilized to separate categorical features from domain features, which is further supervised by labeled images via a feature clustering module. Image reconstruction is introduced to ensure the separated features are representative and meaningful for the input images.
- The proposed method extends the body of literature about semi-supervised domain adaptation (SSDA), which integrates unlabeled data from both source and target domain to alleviate the demand for annotated data, and thus MIDNet can be considered a new SSDA variant.
- We utilize our method for anatomical classification in fetal US, which, to the best of our knowledge, is the first exploration of transferring knowledge to unseen data in a practical application in medical imaging.

II. Related work

1) Representation disentanglement: Disentangling representations aims at interpreting underlying interacted factors within data [34], [35] and enables the manipulation of relevant representations for specific tasks [36]–[38]. Traditional models include techniques such as Independent Component Analysis (ICA) [39] and bilinear models [40] as well as learning-based models such as InfoGAN [41] and β-VAE [42], [43]. Recent work by Mathieu et al. [44] proposes a conditional generative model to disentangle latent representations into specified and unspecified factors of variation via adversarial training. For the same task, Hadad et al. [38] proposes a simpler two-step adversarial training approach for more efficient learning of various unspecified features. Their method directly utilizes the encoded latent space for unspecified factors instead of assuming the underlying distribution. On top of adversarial training, Peng et al. [45] adds mutual information to disentangle more specific features, including class-irrelevant features, domain-invariant features and domain specific features. To use disentangled representations for identifying images with unseen entangled features in real applications, Meng et al. [46] proposed to disentangle category and domain-specific features using an adversarial regularization in a multi-task learning framework. In contrast to these previous works, we propose a non-adversarial method that evaluates mutual information between latent features to disentangle categorical features and domain features (Sec. III-C). Additionally, our method uses sparsely labeled data during training.

2) Semi-supervised learning (SSL): The goal of SSL is to address the scarcity of labeled data by leveraging unlabeled data. Various approaches have been proposed for SSL [47]–[52]. Recently, Zhang et al. [51] proposed MixUp as learning paradigm to train a model on convex combinations of samples and their corresponding labels. This principle encourages the model to favor linear behavior between samples and alleviates problems arising from mislabelled examples. Extending this work, Berthelot et al. [25] introduced a SSL method, MixMatch, which estimates low-entropy labels for unlabeled samples and then applies MixUp to mixed labeled and unlabeled samples for training the model. In this paper, we utilize MixMatch to integrate unlabeled samples from both source and target domain during training (Sec. III-E).

3) Domain adaptation: Previous domain adaptation approaches consider three domain adaptation settings regarding to the number of domains, including one-to-one domain adaptation, multi-source domain adaptation and multi-target domain adaptation. One-to-one domain adaptation considers a single source domain and a single target domain. Unsupervised domain adaption [11], [17], [53] is a typical one-to-one domain adaptation, which requires plenty of labeled samples from the source domain during training and categories in both domains are the same. Multi-source domain adaptation learns universal knowledge from multiple source domains to a single target domain [54]. Domain generalization [7], [55] is a special case of multi-source domain adaptation, which learns knowledge for an unseen target domain from many labeled samples of multiple source domains. In domain generalization, each category in the target domain has been seen in at least one source domain. Multi-target domain adaptation learns knowledge from a single source domain to multiple target domains. Domain agnostic learning [45] is a multi-target domain adaptation method, which also requires plenty of labeled samples from the source domain. In domain agnostic learning, all categories in target domains have been seen in the source domain.

In this paper, we consider a one-to-one domain adaptation setting. In contrast to other works of one-to-one domain adaptation, in our work, the categories in the target domain are a subset of the categories in the source domain. Our ultimate goal is to learn categorical-discriminative knowledge from available categories in both domains to separate unseen categories in the target domain. Fig. 3 compares the different task settings between our work and other domain adaptation works.

4) Transfer learning: Our problem setting also contains flavors of transfer learning, according to the nomenclature in [12]. Transfer learning is a broader field that tackles domain shift and transfers knowledge between different datasets. Domain adaptation is a subcategory of transfer learning [12]. Problem settings with different tasks between source and target domains are close to inductive transfer learning [56].

Methods	Train				Test	
	Source		Target		Target	
	Domain	Category	Domain	Category	Domain	Category
Unsupervised Domain Adaptation, e.g., [17]	S_1	M categories	T_1	M categories	T_1	M categories
Domain Generalization e.g., [7]	S_1, $S_2 ... S_N$	M categories	--	--	T_1	M categories
Domain Agnostic Learning, e.g., [44]	S_1	M categories	T_1, $T_2 ... T_K$	M categories	T_1, $T_2 ... T_K$	M categories
MIDNet (Ours)	S_1	**M categories**	T_1	$M - C_1$ categories	T_1	C_1 categories

Fully-labeled Semi-labeled Unlabeled

Fig. 3: The differences between our method and other existing domain adaptation methods. We compare two aspects, the problem setup and the training paradigm in this taxonomy.

Problem settings considering domain shift between source and target domains within a single task are similar to domain adaptation or transductive transfer learning [57]. Unsupervised transfer learning is a special case of inductive transfer learning, where only unlabeled data is available in both domains during training. In this paper, we focus on transferring knowledge from a source domain to a target domain for a single task and tackle covariate shift (Sec. I). Therefore, as suggested in [12], we frame our problem setting within the context of domain adaptation.

III. Method

Our goal is to disentangle categorical features from domain features to obtain generalizable features, so that our model can classify the categories in the target domain which have not been seen during training. We formulate our task as follows: let $\mathcal{X}^S = \{\mathbf{x}_i^S\}_{i=1}^{|C^S|}$ be the images from a source domain which contain categories C^S and $\mathcal{X}^T = \{\mathbf{x}_i^T\}_{i=1}^{|C^T|}$ be images from a target domain with categories $C^T, C^T \subset C^S$. In both domains, categorical labels are available for part of the images as $\mathcal{Y}^S, \mathcal{Y}^T$. We want to train a network to maximize the categorical prediction performance of the classifier on images in the target domain from new categories $\{\mathbf{x}^T | \mathbf{x}^T \in C^S - C^T\}$.

To solve this task, we propose MIDNet in combination with semi-supervised learning. The architecture of our model is shown in Fig. 4. Two independent encoders E are utilized to respectively extract categorical features $\mathcal{F}_C$ and domain features $\mathcal{F}_D$ from labeled data $\{\mathcal{X}_L, \mathcal{Y}_L\} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}^S \cup \mathcal{X}^T, y_i | y_i \in \mathcal{Y}^S \cup \mathcal{Y}^T\}$ and unlabeled data $\mathcal{X}_U = \{\mathbf{x}_j | \mathbf{x}_j \in \mathcal{X}^S \cup \mathcal{X}^T\}$. The classifier C is responsible for predicting class distributions from $\mathcal{F}_C$ for both $\mathcal{X}_L$ and $\mathcal{X}_U$ while the decoder D combines $\mathcal{F}_C$ and $\mathcal{F}_D$ for the reconstruction of input images. The mixer M aims to linearly mix labeled and unlabeled samples so that the model is trained to show linear behavior between samples for further leveraging of unlabeled data. For representation disentanglement, mutual information between $\mathcal{F}_C$ and $\mathcal{F}_D$ is minimized to encourage $\mathcal{F}_C$ to become domain-invariant and maximally informative for categorical classification. Feature clustering contains feature alignment and distance metric learning. Feature alignment aims at keeping the feature consistency between labeled images to promote the independence of $\mathcal{F}_C$. Distance metric learning considers inter-class relationships, which clusters similar samples while separating dissimilar samples to optimize $\mathcal{F}_C$ for improving classification performance.

A. Image reconstruction

The first step of MIDNet is to employ an Encoder-Decoder framework for independent extraction of two internal representations from the input data $\mathbf{x}$. Two encoders E_1, E_2 are built to respectively generate latent vectors that aim to represent categorical features $\mathcal{F}_C$ and domain features $\mathcal{F}_D$, where $\mathcal{F}_C = E_1(\mathbf{x}; \phi_1)$ and $\mathcal{F}_D = E_2(\mathbf{x}; \phi_2)$. The decoder D is utilized to guarantee that the combination of these features is capable of recovering original input data, where $\widehat{\mathbf{x}} = D(\mathcal{F}_C, \mathcal{F}_D; \psi)$. Here, $\mathbf{x} \in \mathcal{X}_L \cup \mathcal{X}_U$ and ϕ_1, ϕ_2, ψ are the parameters of E_1, E_2, D, respectively. The cost function of this reconstruction is

$$\mathcal{L}_{rec} = \|\widehat{x} - x\|_F^2. \tag{1}$$

We concatenate layers between E_2 and D to integrate high-frequency features from E_2 into the reconstruction, which helps $\mathcal{F}_D$ to contain valid information instead of noise. This image reconstruction extracts two groups of features from internal representations of original data. The rest of our networks are designed and trained to enable $\mathcal{F}_C$ to only contain categorical information, thus becoming separated from $\mathcal{F}_D$ that only contains domain information.

B. Classification

We use a classifier C to predict $|C^S|$ labels for labeled data, which encourages $\mathcal{F}_C$ to be maximally informative about categorical classification. E_1, E_2 and C are updated by minimizing the cross-entropy loss

$$\mathcal{L}_{cls} = -\mathbb{E}_{\{\mathbf{x},y\}\sim\{\mathcal{X}_L,\mathcal{Y}_L\}} \sum_{t=1}^{|C^S|} \mathbb{1}[y = t] log(C(\mathcal{F}_C; \delta)). \tag{2}$$

Here δ refers to the parameters of C. At the same time, C predicts the class distribution of the unlabeled data $\mathbf{x} \in \mathcal{X}_U$ as $P_C(\widehat{y}|\mathbf{x}; \delta), \widehat{y} \in \widehat{\mathcal{Y}}_U$. The predicted $P_C(\widehat{y}|\mathbf{x}; \delta), \mathbf{x} \in \mathcal{X}_U$ will be utilized in SSL-based regularization (Sec. III-E). The classifier on its own is unlikely to ensure that categorical features $\mathcal{F}_C$ are domain-invariant. This is because the training objective in Eq. 2 only ensures that $\mathcal{F}_C$ contains as much information as possible for the target classification task.

C. Mutual information disentanglement

To address the problem from Sec. III-B, we minimize the mutual information between $\mathcal{F}_C$ and $\mathcal{F}_D$. This minimization forces $\mathcal{F}_C$ to contain less domain information and thus separates categorical features from domain features. Mutual information is defined as

$$I(\mathcal{D}_{\mathcal{F}_C}; \mathcal{D}_{\mathcal{F}_D}) = \int_{\mathcal{X}\times\mathcal{Z}} log \frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \bigotimes \mathbb{P}_Z} d\mathbb{P}_{XZ}, \tag{3}$$

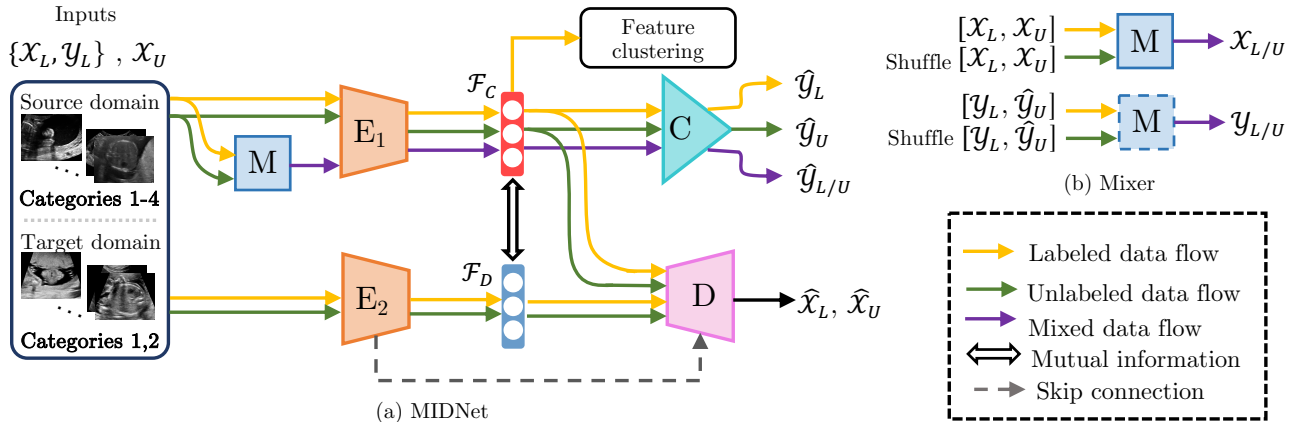


Fig. 4: (a) An overview of MIDNet. We extract disjoint features $(\mathcal{F}_C, \mathcal{F}_D)$ through mutual information minimization in the latent space and apply a feature consistency constraint to extract domain-invariant categorical features $\mathcal{F}_C$ for further disentanglement. E_1, E_2 are separate encoders, C is a classifier and D is a decoder. (b) We integrate unlabeled data by a mixer M for semi-supervised learning.

where $\mathbb{P}_{XZ}$ is the joint probability distribution of $(\mathcal{D}_{\mathcal{F}_C}, \mathcal{D}_{\mathcal{F}_D})$, $\mathbb{P}_X = \int_{\mathcal{Z}} d\mathbb{P}_{XZ}$ and $\mathbb{P}_Z = \int_{\mathcal{X}} d\mathbb{P}_{XZ}$ are respectively marginal distributions of $\mathcal{D}_{\mathcal{F}_C}$ and $\mathcal{D}_{\mathcal{F}_D}$. We utilize Mutual Information Neural Estimation (MINE) [24] to approximate the lower-bound of mutual information on n samples by a neural network T with parameters $\theta \in \Theta$,

$$I(\widehat{\mathcal{D}_{\mathcal{F}_C}; \mathcal{D}_{\mathcal{F}_D}})_n = \sup_{\theta \in \Theta} E_{\mathbb{P}^{(n)}_{XZ}}[T_\theta] - log(E_{\mathbb{P}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Z}[e^{T_\theta}]). \tag{4}$$

Practically, the expectations in Eq. 4 are estimated by Monte-Carlo integration [45] with shuffled samples along the batch axis ($\mathcal{F}'_D$), and thus the cost function of the mutual information disentanglement is

$$\begin{aligned}\mathcal{L}_{MI} &= -I(\widehat{\mathcal{D}_{\mathcal{F}_C}; \mathcal{D}_{\mathcal{F}_D}})_n \\ &= -(\frac{1}{n}\sum_{i=1}^{n} T(\mathcal{F}_C, \mathcal{F}_D, \theta) - log(\frac{1}{n}\sum_{i=1}^{n} e^{T(\mathcal{F}_C, \mathcal{F}'_D, \theta)})).\end{aligned} \tag{5}$$

Here, $(\mathcal{F}_C, \mathcal{F}_D)$ are sampled from joint distributions while $(\mathcal{F}_C, \mathcal{F}'_D)$ are sampled from the product of marginal distributions.

D. Feature clustering

To introduce the supervision from labeled samples for an enhanced disentanglement, we aim at aligning categorical features between source and target domains. We hypothesize that categorical features of a certain category are supposed to be consistent between different domains. Therefore, we further enhance $\mathcal{F}_C$ to be domain-invariant by minimizing the distance of categorical features between source domain and target domain for samples in $\mathcal{X}_L$ with

$$\mathcal{L}_{align} = \frac{1}{|C^T|}\sum_{i=1}^{|C^T|}(\frac{1}{n_{c_i}}\sum_{j=1}^{n_{c_i}} \| f^S_{c_{ij}} - f^T_{c_{ij}} \|^2_F), \tag{6}$$

where n_{c_i} is the number of samples in category c_i, $c_i \in C_T$ and $f^S_{c_{ij}}, f^T_{c_{ij}}$ are the categorical features of the jth sample from category c_i in the source domain and the target domain, respectively. This loss is computed within categorical features $\mathcal{F}_C$ in order to enhance mutual information disentanglement. This is different from feature alignment in other domain adaptation approaches (*e.g.*, [16], [58]), where the feature alignment directly aligns the whole latent features of inputs. In addition, our categorical feature alignment uses the Frobenius norm as the distance metric since labeled samples from both domains are available, whereas many other domain adaptation approaches (*e.g.*, [14], [15]) utilize Maximum Mean Discrepancy (MMD) [13] to diminish the discrepancy between a labeled source domain and an unlabeled target domain.

After feature alignment, We additionally consider inter-class relationships. To further cluster samples from the same category and separate samples from different categories in the latent space, we introduce distance metric learning with triplet loss [59] on $\mathcal{F}_C$ of labeled images from both domains,

$$\mathcal{L}_{trip} = max\{0, d(f_{c_{iq}}, f_{c_{ip}}) - d(f_{c_{iq}}, f_{c_{kn}}) + \xi\}. \tag{7}$$

Here, $d(\cdot,\cdot)$ is the squared Euclidean distance. $f_{c_{iq}}$ is the categorical feature of one query sample from category c_i. $f_{c_{ip}}$ and $f_{c_{kn}}$ are respectively categorical features of one support sample from the same category c_i and one negative sample from a different category c_k, where $c_i, c_k \in C_S \cup C_T$. The selection of the query sample, the support sample and the negative sample is essential for converge. For implementation, we utilize the online semi-hard triplet mining strategy in [59].

Considering both feature alignment and separation, the feature clustering loss is

$$\mathcal{L}_{clus} = \mathcal{L}_{align} + \eta\mathcal{L}_{trip} \tag{8}$$

where η is a hyper-parameter. In practice, we mainly focus on feature alignment and set $\eta = 0$ as the proposed method because feature alignment is the key point to keep consistency between both domains. In our applications, we conduct both experiments, *i.e.*, with and without $\mathcal{L}_{trip}$, to evaluate the importance of inter-class relationships.

E. SSL-based regularization

Conventional fully supervised learning requires a large number of annotated input images with categorical labels and domain labels. However, in practice, labeled data is not easily available at any granularity. Berthelot et al. [25] propose a SSL-based method, MixMatch, integrating unlabeled data during training to reduce the dependency on labeled data. We adopt a simplified MixMatch using a mixer M to leverage unlabeled data. As shown in Fig. 4 (b), M is utilized to linearly combine two random samples $(\mathbf{x}_1, \mathbf{x}_2)$ and their corresponding labels (y_1, y_2) from labeled data and unlabeled data by

$$\begin{aligned} \mathbf{x}_{mix} &= M(\mathbf{x}_1, \mathbf{x}_2; \beta) = \beta \mathbf{x}_1 + (1-\beta)\mathbf{x}_2, \\ y_{mix} &= M(y_1, y_2; \beta) = \beta y_1 + (1-\beta) y_2. \end{aligned} \tag{9}$$

where $\beta = max(\xi, 1-\xi)$, $\xi \sim Beta(\alpha, \alpha)$. Here, $\mathbf{x}_1 \in \mathcal{X}_{cat}$, $\mathbf{x}_2 \in \mathcal{X}'_{cat}$. $\mathcal{X}_{cat} = [\mathcal{X}_L, \mathcal{X}_U]$ is the concatenation of $\mathcal{X}_L$ and $\mathcal{X}_U$. $\mathcal{X}'_{cat}$ is the shuffled $\mathcal{X}_{cat}$ along the batch axis. Similarly, $y_1 \in \mathcal{Y}_{cat}$ with $\mathcal{Y}_{cat} = [\mathcal{Y}_L, \widehat{\mathcal{Y}}_U]$, and $y_2 \in \mathcal{Y}'_{cat}$. Note that $\widehat{\mathcal{Y}}_U$ is the collection of the predicted labels for unlabeled data according to Sec. III-B. We denote that $\mathbf{x}_{mix} \in \mathcal{X}_{L/U}$, $y_{mix} \in \mathcal{Y}_{L/U}$. The goal of this SSL-based regularization is to encourage the linear behavior of the classifier, and thus the objective function is

$$\mathcal{L}_{SSL} = \|y_{mix} - P_C(\widehat{y}_{mix}|\mathbf{x}_{mix}; \delta)\|_F^2, \tag{10}$$

where $P_C(\widehat{y}_{mix}|\mathbf{x}_{mix}; \delta) = C(\mathbf{x}_{mix}; \delta)$ is the predicted label of $\mathbf{x}_{mix}$ via classifier C.

F. Optimization

Our model is an end-to-end trainable framework and the overall objective is a linear combination of all cost functions

$$min\{\lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{MI} + \lambda_4 \mathcal{L}_{clus} + \lambda_5 \mathcal{L}_{SSL}\}, \tag{11}$$

where λ_1 to λ_5 are hyper-parameters chosen experimentally depending on the dataset. We optimize the MINE and the rest of our model in an alternating fashion. Inspired by [24], we use the Adam optimizer (beta $= 0.9$, learning rate $= 10^{-5}$) to train the network parameters θ based on Eq. 5 and use Stochastic Gradient Descent (SGD) with momentum optimizer (momentum $= 0.9$, learning rate $= 10^{-5}$) to update the parameters of encoders, decoders and classifier based on Eq. 11. We apply L2 regularization (scale $= 10^{-5}$) to all weights during training to prevent over-fitting and we apply random image flipping as data augmentation. Classes are kept balanced on labeled data during training. Our model is trained on a Nvidia Titan X GPU with 12 GB of memory.

IV. Experiments

We evaluate the proposed method on two fetal US standard plane classification tasks, where the domain shifts are respectively caused by shadow artifacts (Fig. 2(a)) and different image acquisition devices (Fig. 2(b)). For both tasks, images from source and target domains are unpaired and collected independently. We show the key results in the main paper and detailed implementation, network architectures as well as more results in the supplemental Appendix.

A. Experiment settings

We conduct comprehensive evaluations for each application in the following. (1) We compare the proposed method with the state-of-the-art algorithms for domain adaptation. (2) We explore the effectiveness of different key components in MIDNet via an ablation study. (3) To demonstrate the importance of unlabeled data, we compare the classification performance with and without unlabeled training data on the target domain. (4) By training MIDNet with different percentage of labeled data, we evaluate the performance of our model in a semi-supervised setting. In addition, we discuss the influence of common categories on the classification performance of unseen categories in the target domain.

We utilize three groups of test data for the evaluation: (*i*) test data from the source domain T_{Source}, (*ii*) test data from the target domain whose image features have been observed during training T_{Target}, and (*iii*), most importantly, test data from the target domain whose image features are absent during training T_{Target}^{New}.

We adopt commonly-used statistical metrics, F1-score, recall and precision, to quantitatively evaluate classification performance. Recall $= TP/(TP+FN)$, Precision $= TP/(TP+FP)$ and F1-score is the the harmonic mean of precision and recall. We report the average scores of these metrics for all examined methods. As suggested by [60], we utilize the $\mathcal{A}$-distance as a measure of domain divergence to quantitatively evaluate the separation of categorical and domain features. Similar to [17], [61], we train a SVM as a domain classifier to compute ϵ (the error of classifier) for the $\mathcal{A}$-distance, $\hat{d}_{\mathcal{A}} = 2(1-2\epsilon)$.

B. Comparison methods and ablation study

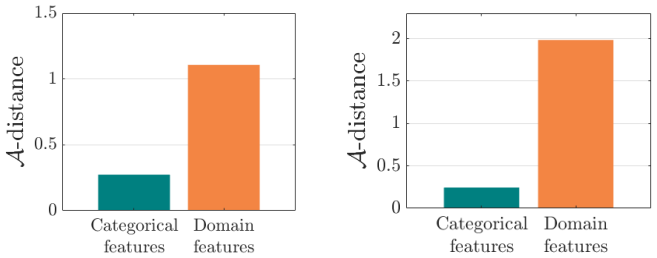
We evaluate a VGG network [62] which is trained on data only from the source domain, namely *Source only*, as a baseline to demonstrate that the domain shift problems affects the generalizability of deep models. To verify that MIDNet is able to extract generalized features across domains, we compare MIDNet with a VGG network [62] and a VGG network with residual unit [63] (Res-VGG). We further compare MIDNet to the state-of-the-art feature disentanglement algorithms for addressing the task in this work, including a two-step disentanglement method[1] [38] and a multi-task learning based disentanglement method[2] [46]. Note that we implement the method in [38] differently from the original paper. Specifically, we train the model simultaneously to enable it to be suitable for our task setup. We denote *Two-step-fair* as [38] with an adversarial network using unspecific features (Z) for category classification and denote *Two-step-Unfair* as [38] with an adversarial network using specific features (S) for domain classification. We keep the original experimental settings for the method in [46] (namely Multi-task). All comparison methods above are fully-supervised. Additionally, we compare the proposed method with the state-of-the-art domain adaptation methods, including domain-adversarial training of

[1]https://github.com/naamahadad/A-Two-Step-Disentanglement-Method
[2]https://github.com/qmeng99/Multi-task-Representation-Disentanglement

TABLE I: Different combination of key components in MIDNet for the ablation study. $\mathcal{L}_{rec}$ is for reconstruction, $\mathcal{L}_{cls}$ is for classification, $\mathcal{L}_{MI}$ is for mutual information disentanglement and $\mathcal{L}_{SSL}$ is for integrating unlabeled data. For $\mathcal{L}_{clus}$, $\eta = 0$ only considers feature alignment and $\eta \neq 0$ considers both feature alignment and inter-class relationships.

Methods	$\mathcal{L}_{rec}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{MI}$	$\mathcal{L}_{clus}$ $\eta = 0$	$\mathcal{L}_{SSL}$	$\mathcal{L}_{clus}$ $\eta \neq 0$
MIDNet-I	√	√				
MIDNet-II	√	√	√			
MIDNet-III	√	√		√		
MIDNet-IV	√	√			√	
MIDNet w/o $\mathcal{L}_{SSL}$	√	√	√	√		
MIDNet w/o $\mathcal{L}_{clus}$	√	√	√		√	
MIDNet w/o $\mathcal{L}_{MI}$	√	√		√	√	
MIDNet	√	√	√	√	√	
MIDNet+$\mathcal{L}_{trip}$	√	√	√		√	√

η is experimentally selected for different tasks when $\eta \neq 0$.



(a) Source: shadow-free; Target: shadow-containing

(b) Source: device A; Target: device B

Fig. 5: Domain discrepancy based on categorical features and domain features in MIDNet. A high $\mathcal{A}$-distance score means a high domain difference.

neural networks (DANN)[3] [17] and semi-supervised domain adaptation via MiniMax Entropy (MME)[4] [61]. These two comparison methods are semi-supervised. For all comparison methods with hyperparameters, we run several sets of parameter values (including the values in corresponding papers). The hyperparameters with the best experimental results are selected for the following evaluation.

For the ablation study, we remove different loss components to obtain different combinations of components in MIDNet. The detailed combinations of ablations are shown in Table. I.

C. Experiments on fetal US with and without shadows

The fetal US dataset consists of $\sim 7k$ 2D fetal US images sampled from 2694 2D US examinations with gestational ages between $18-22$ weeks (iFIND Project [5]). Eight different US systems of identical make and model (GE Voluson E8) were used for the acquisitions to eliminate as many unknown image acquisition parameters as possible. Six different anatomical standard plane locations have been selected by an experienced sonographer, including *Four Chamber View (4CH)*, *Abdominal*, *Femur*, *Lips*, *Left Ventricular Outflow Tract (LVOT)* and *Right Ventricular Outflow Tract (RVOT)*. The images have additionally been classified by an expert observer as shadow-containing or shadow-free. In this experiment, the source domain contains shadow-free images (see Fig. 2 (b) SF) while the target domain has shadow-containing images from less favorable imaging conditions (see Fig. 2 (b) SC). Training data consists of all six standard planes from the source domain as well as Abdominal, LVOT and RVOT from the target domain. We aim to separate anatomical features (categorical features) and shadow artifacts features (domain features) to obtain generalized anatomical features for achieving high performance of standard plane classification on T^{New}_{Target} (4CH, Femur and Lips from target domain). Here, T_{Source} contains 4CH, Abdominal, Femur, Lips, LVOT and RVOT from the source domain and T_{Target} contains Abdominal, LVOT and RVOT from the target domain. Hyper-parameters λ_1 to λ_5 in Eq. 11 are $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10^{-4}, \lambda_4 = 10, \lambda_5 = 10$ for the proposed MIDNet model and $\eta = 0.005$ is additionally for MIDNet+$\mathcal{L}_{trip}$.

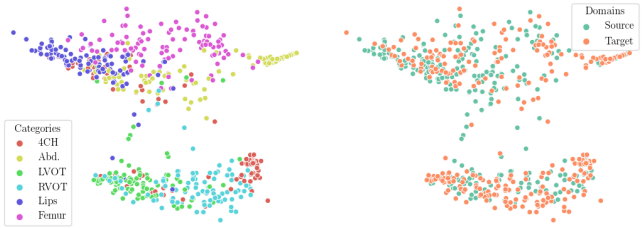
Results: We compare the $\mathcal{A}$-distance of categorical features and domain features. Fig. 5 (a) shows that domain difference is higher in domain features than in categorical features. This indicates that domain features contain more domain information whereas categorical features are more domain-invariant. Fig. 6 shows the t-SNE plot of categorical features in both domains for MIDNet. From Fig. 6 (a), we observe that the categorical features learned by MIDNet enable the anatomical classification. Fig. 6 (b) shows that the learned categorical features are domain-invariant.

The experimental results of the state-of-the-art and the ablation study are shown in Table. II. In this table, we observe that the MIDNet model outperforms all the state-of-the-art methods on the most important test data T^{New}_{Target} for average F1-score, recall and precision. MIDNet+$\mathcal{L}_{trip}$ performs better than MIDNet on T_{Source} and T_{Target}, demonstrating that metric learning is important and efficient for improving classification performance on images whose features have been observed during training. In the ablation study, MIDNet outperforms other variant models, especially *MIDNet w/o* $\mathcal{L}_{SSL}$, *MIDNet w/o* $\mathcal{L}_{clus}$ and *MIDNet w/o* $\mathcal{L}_{MI}$, illustrating the effectiveness of all proposed components in MIDNet. In addition, Fig. 7 (a) shows that the $\mathcal{A}$-distance of *MIDNet w/o* $\mathcal{L}_{MI}$ is higher than that of MIDNet. This demonstrates that mutual information disentanglement ($\mathcal{L}_{MI}$) contributes to learn domain-invariant categorical features.

We further compare the performance of MIDNet with and without unlabeled data on the target domain. Here, the *with unlabeled data* setting utilizes the training data containing 30% labeled data and 70% unlabeled data, while the *without unlabeled data* setting only uses the 30% labeled data. The confusion matrices in Fig. 8(a) show the effectiveness of unlabeled data in the proposed method, for example, the classification accuracy of T^{New}_{Target} in MIDNet (e.g., *4CH* and *Lips*) improves when integrating unlabeled data.

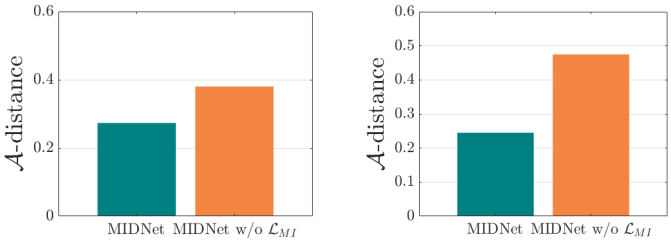
To explore the importance of labeled data, we evaluate the performance of MIDNet and MIDNet+$\mathcal{L}_{trip}$ based on

[3]https://github.com/pumpikano/tf-dann, Jan 2018
[4]https://github.com/VisionLearningGroup/SSDA_MME
[5]http://www.ifindproject.com/



(a) (b)

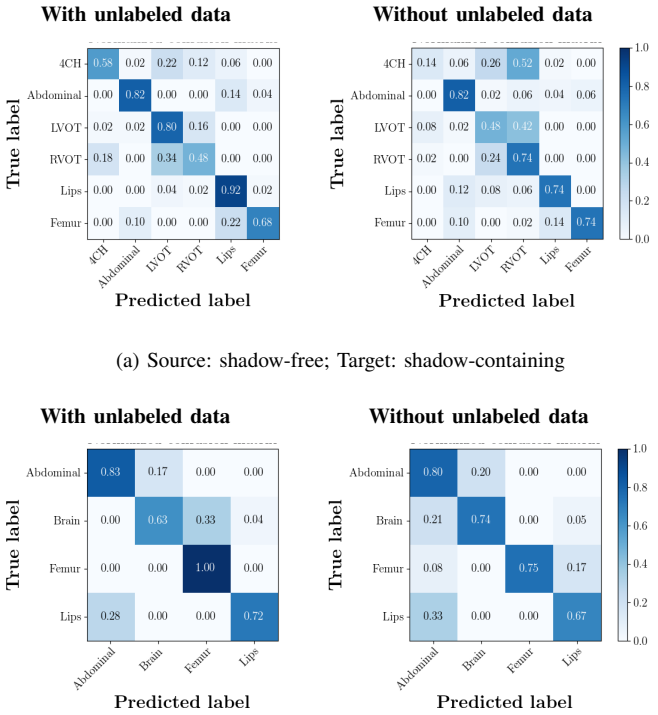
Fig. 6: Feature visualization of MIDNet with t-SNE. We plot categorical features from both domains. (a) The color represents categories. (b) The color represents domains. We observe that the categorical features are domain-invariant and enable the anatomical classification.



(a) Source: shadow-free; Target: shadow-containing

(b) Source: device A; Target: device B

Fig. 7: Domain discrepancy of MIDNet and MIDNet w/o $\mathcal{L}_{MI}$ based on categorical features. A high $\mathcal{A}$-distance score means a high domain difference. Experiment setting is the same with Table II.



(a) Source: shadow-free; Target: shadow-containing

(b) Source: device A; Target: device B

Fig. 8: Confusion matrices of $\mathbf{T_{Target}}$ & $\mathbf{T^{New}_{Target}}$ in MIDNet with and without unlabeled data. For the *with unlabeled data* setting, 30% of training data are labeled and the rest are unlabeled. The *without unlabeled data* setting only uses the 30% labeled data for training, without using unlabeled data. (a) Classification on fetal US with/without shadows. (b) Classification on fetal US from different acquisition devices.

using $15\%, 30\%, 60\%$ and 100% labeled data during training. Fig. 9 (a) shows the average F1-score on three groups of test data. From this figure, we observe that the classification performance improves with increasing labeled data.

Finally, correctly classified and mis-classified examples of T^{New}_{Target} using MIDNet are presented in Fig. 11 (a).

D. Experiments on cross-device fetal US

The previous experiment on fetal US images is supported by data restricted to one US imaging device. Here, we evaluate MIDNet for a standard plane classification task on data from different imaging devices (different device domains). Device A is "GE Voluson E8" which is the same device in Sec. IV-C, which acquired $\sim 6K$ 2D fetal US images. Device B is "Philips EPIQ V7 G" which acquired another $\sim 5K$ images sampled from about 500 2D US examinations with gestational ages between 20-32 weeks (see Fig. 2 (b), iFIND Project [5]). In this experiment, we use four different anatomical standard plane locations with sufficient images in both domains, including *Abdominal*, *Brain*, *Femur* and *Lips*, which are selected by an 10-year-experienced sonographer. In this experiment, the source domain is set as device A while the target domain is device B. Training data consists of all four standard planes from the source domain as well as Abdominal and Brain from the target domain. We aim to separate anatomical features (categorical features) and imaging device features (domain features) to obtain generalized anatomical features for achieving high performance of standard plane classification on T^{New}_{Target} (Femur and Lips). Here, T_{Source} contains Abdominal, Brain, Femur and Lips from the source domain and T_{Target} contains Abdominal and Brain from the target domain. Hyper-parameters λ_1 to λ_5 in Eq. 11 are $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10^{-4}, \lambda_4 = 50, \lambda_5 = 50$ for the proposed MIDNet model and $\eta = 2 \times 10^{-4}$ is additionally for MIDNet+$\mathcal{L}_{trip}$.

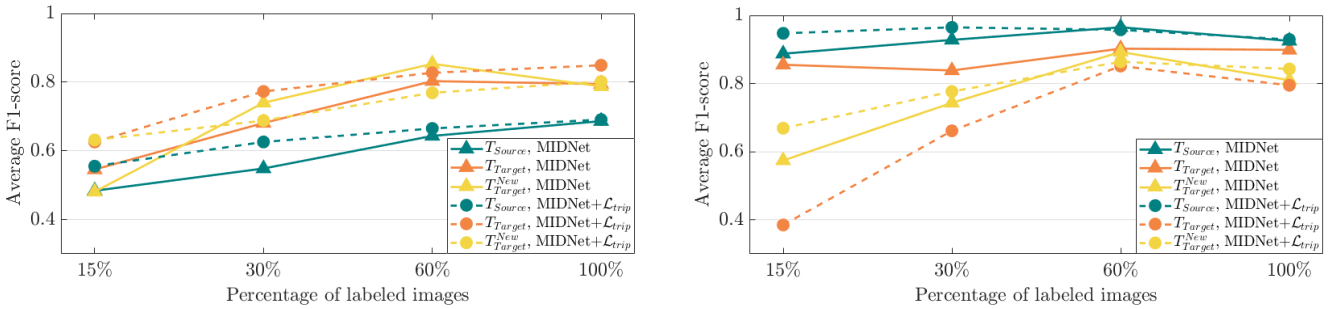
Results: The classification performance of baselines and the proposed model are shown in Table. III. We observe that MIDNet outperforms the state-of-the-art on on the most important test data T^{New}_{Target} for average F1-score and recall. Among all the models in the ablation study, MIDNet+$\mathcal{L}_{trip}$ achieves the best performance on T_{Source} and T_{Target}, demonstrating that metric learning contributes to the separation of seen categories in both domains. MIDNet outperforms *MIDNet w/o* $\mathcal{L}_{SSL}$, *MIDNet w/o* $\mathcal{L}_{clus}$ and *MIDNet w/o* $\mathcal{L}_{MI}$ on T^{New}_{Target}, illustrating the importance of all proposed components in MIDNet for the classification of unseen categories in the target domain. In addition, Fig. 7 (b) shows the effectiveness of mutual information disentanglement ($\mathcal{L}_{MI}$) to learn domain-invariant categorical features.

The confusion matrices in Fig. 8(b) show the effectiveness

TABLE II: Comparison of the state-of-the-art and ablation study for the **fetal US standard plane classification task using data with/without shadow artifacts**. 30% of training data are labeled data and the rest are unlabeled data. Average F1-score, Recall and Precision are measured on three groups of test data. Best results are shown in bold.

Methods	T_{Source}			T_{Target}			T_{Target}^{New}		
	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision
Source only	0.4878	0.4933	0.4914	0.4341	0.4067	0.4693	0.5496	0.5867	0.5170
VGG [62]	0.5347	0.5400	0.5335	0.4492	0.4667	0.4337	0.5495	0.5400	0.5614
Res-VGG [62], [63]	0.4669	0.4800	0.5229	0.5482	0.5867	0.5293	0.5722	0.5467	0.6182
Two-step-fair [38]	0.4531	0.4500	0.4572	0.4400	0.4467	0.4338	0.5008	0.4933	0.5095
Two-step-Unfair [38]	0.4894	0.4933	0.4895	0.4515	0.4733	0.4319	0.4571	0.4400	0.4769
Multi-task [46]	0.4622	0.4667	0.5524	0.5787	0.5667	0.6220	0.6393	0.6533	0.6491
DANN [17]	0.5939	0.6100	0.6598	0.6642	0.7000	0.6400	0.5525	0.5533	0.5700
MME [61]	0.5723	0.5700	**0.7496**	**0.8163**	0.8000	**0.8458**	0.4852	0.5200	0.5006
MIDNet-I	0.4643	0.4767	0.5891	0.5796	0.5933	0.5944	0.6280	0.6133	0.6947
MIDNet-II	0.4760	0.4867	0.5336	0.6185	0.6533	0.6056	0.6559	0.6200	0.7412
MIDNet-III	0.4929	0.5100	0.5498	0.5620	0.5800	0.5512	0.6887	0.6667	0.7267
MIDNet-IV	0.4636	0.4833	0.5403	0.5746	0.5867	0.5705	0.6378	0.6400	0.6732
MIDNet w/o $\mathcal{L}_{SSL}$	0.5379	0.5533	0.6007	0.5976	0.6600	0.5612	0.6603	0.6000	0.8119
MIDNet w/o $\mathcal{L}_{clus}$	0.4195	0.4367	0.5102	0.5657	0.5800	0.5637	0.6025	0.6000	0.6539
MIDNet w/o $\mathcal{L}_{MI}$	0.5339	0.5467	0.5948	0.6654	0.7067	0.6449	0.7091	0.6600	0.8255
MIDNet	0.5484	0.5667	0.6683	0.6809	0.7000	0.6803	**0.7399**	**0.7267**	0.7830
MIDNet+$\mathcal{L}_{trip}$	**0.6257**	**0.6367**	0.7082	0.7728	**0.8533**	0.7146	0.6880	0.6200	**0.8396**

The baselines and ablation study models are introduced in Sec. IV-B. DANN, MME and the proposed models are semi-supervised while other comparison methods are fully-supervised which only use the 30% labeled images.



(a) Source: shadow-free; Target: shadow-containing

(b) Source: device A; Target: device B

Fig. 9: Average F1-score of standard plane classification with different percentage of labeled data (15%, 30%, 60%, 100%) for semi-supervised learning based on MIDNet model and MIDNet+$\mathcal{L}_{trip}$. (a) Classification on fetal US with/without shadows. (b) Classification on fetal US from different image acquisition devices.

of unlabeled data in the proposed method. The classification accuracy of *Femur* and *Lips* in MIDNet improves when using unlabeled data.

From Fig. 9 (b), we observe that classification performance improves with the increase of labeled data in most cases. However, the performance reaches its peak after a certain percentage of labeled data is added. For example, the peak point is 60% in this experiment.

In addition, we utilize t-SNE plots for feature visualization in Fig. 10. Comparing Fig. 10 (a) and Fig. 10 (b), we observe that with mutual information disentanglement, (1) samples from the same category are more tightly clustered (see the top row) and (2) the source domain and the target domain are overlap more (see the bottom row). This indicates that mutual information disentanglement is important for learning categorical-focused and domain-invariant features. Fig. 10 (a), (c)-(d) show that the proposed method outperforms other state-of-the-art methods for learning category-discriminative and domain-invariant features, especially for unseen categories in the target domain (*e.g.*, (a) vs. (d)).

We further present correctly classified and mis-classified examples of T_{Target}^{New} using MIDNet in Fig. 11 (b).

The influence of common categories: We further explore the influence of common categories on this cross-device fetal US classification task. We evaluate the classification performance on the unseen category in the target domain (*i.e.*, Lips) with an increased number of common categories in both domains. In this experiment, 60% of training data are labeled data and the rest are unlabeled data. Table. IV shows that the classification performance of the unseen categories in the

TABLE III: Comparison of *Source only*, the state-of-the-art methods and ablation study for the **fetal US standard plane classification task with data from different acquisition devices (Source domain: Device A, Target domain: Device B)**. 30% of training data are labeled data and the rest are unlabeled data. Best results in bold.

Methods	T_{Source}			T_{Target}			T_{Target}^{New}		
	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision
Source only	0.7665	0.7700	0.7656	0.6971	0.6750	0.7305	0.6742	0.7050	0.6899
VGG [62]	0.7565	0.7600	0.7553	0.6853	0.6750	0.6964	0.7039	0.7250	0.7011
Res-VGG [62], [63]	0.9196	0.9200	0.9229	0.5582	0.6200	0.5092	0.6880	0.6300	0.8728
Two-step-fair [38]	0.7979	0.7975	0.8007	0.3965	0.4050	0.3919	0.7491	0.7400	0.7644
Two-step-Unfair [38]	0.7899	0.7900	0.7903	0.3869	0.4050	0.3831	0.6069	0.6150	0.6013
Multi-task [46]	0.8124	0.8150	0.8383	0.3964	0.4200	0.4186	0.7522	0.7800	0.7955
DANN [17]	0.9572	0.9575	0.9588	0.5507	0.6100	0.5521	0.5611	0.5050	0.7784
MME [61]	0.9526	0.9525	0.9537	0.5400	0.7150	0.4345	0.4293	0.3600	0.9595
MIDNet-I	0.9623	0.9625	0.9642	0.5572	0.5600	0.6256	0.3236	0.5000	0.2393
MIDNet-II	0.9520	0.9525	0.9592	0.6665	0.6600	0.6731	0.3139	0.3950	0.6811
MIDNet-III	0.8948	0.8950	0.8948	0.7125	0.8000	0.7383	0.7400	0.6750	**0.9630**
MIDNet-IV	0.9446	0.9450	0.9486	0.1992	0.1300	0.4261	0.2473	0.5000	0.1471
MIDNet w/o $\mathcal{L}_{SSL}$	0.9100	0.9100	0.9102	0.6099	0.6700	0.5738	0.7961	0.7100	0.9074
MIDNet w/o $\mathcal{L}_{clus}$	0.9421	0.9425	0.9466	0.3406	0.2350	0.6586	0.2264	0.4900	0.1472
MIDNet w/o $\mathcal{L}_{MI}$	0.9425	0.9425	0.9436	0.5015	0.4950	0.5118	0.7913	0.7950	0.7904
MIDNet	0.9281	0.9275	0.9327	0.7434	0.7300	**0.7676**	**0.8383**	**0.8600**	0.8497
MIDNet+$\mathcal{L}_{trip}$	**0.9649**	**0.9650**	**0.9658**	**0.7768**	**0.8500**	0.7177	0.6614	0.6000	0.8009

(a) MIDNet (b) MIDNet w/o $\mathcal{L}_{MI}$ (c) DANN [17] (d) MME [61]

Fig. 10: Feature visualization with t-SNE. We plot categorical features on both domains. The color in the top row represents categories. The color in the bottom row represents domains. Experiment setting is the same with Table III. We observe that the categorical features learned by the proposed method are more domain-invariant and categorical-discriminative than other methods.

target domain improves as increasing the number of common categories between the source and the target domain. This might because more common categories can introduce more information about data distributions in the target domain.

V. Discussion

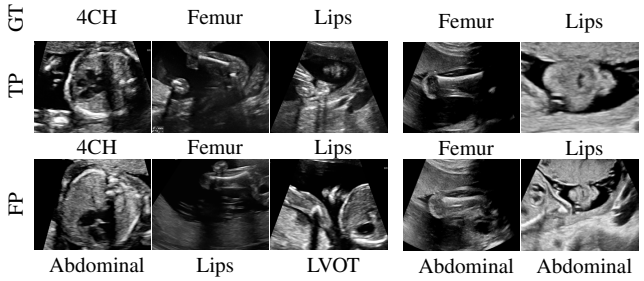
Current disentanglement methods and domain adaptation methods are able to extract domain-invariant categorical features. However, these methods can hardly transfer knowledge to unseen categories in the target domain because (1) they only consider labeled data and extract domain-invariant features with imitated generalizability (*e.g.*, Two-step-Unfair [38] and Multi-task [46]), (2) they have model-specific limitations, for example, MME [61] optimizes categorical clusters only from available categories of both domains and may negatively affect classification of unseen categories, and (3) they neglect intra-class relationships between both domains. Our method extracts domain-invariant categorical features with a semi-supervised paradigm and further explicitly aligns the these features within each category.

The utilization of distance metric learning in feature clustering (Sec. III-D) contributes to further increasing inter-class variance. This results in improved classification performance on T_{Source} and T_{Target} (MIDNet vs. MIDNet+$\mathcal{L}_{trip}$ in Ta-

TABLE IV: Average F1-score of the unseen categories in the target domain (*i.e.*, Lips), when an increasing number of common categories between the source domain ($\mathcal{S}$) and the target domain ($\mathcal{T}$) is available during training. Best results in bold. 60% of training data are labeled and the rest are unlabeled.

Methods	$\mathcal{S}$: Abd., Brain, Femur, Lips		
	$\mathcal{T}$: Abd.	$\mathcal{T}$: Abd., Brain	$\mathcal{T}$: Abd., Brain, Femur
MIDNet	0.7816	0.8047	**0.8187**
MIDNet+$\mathcal{L}_{trip}$	0.6442	0.7179	**0.7727**

Abd. is *Abdominal* and hyperparameters are the same as in Sec. IV-D.



(a) Source: shadow-free; Target: shadow-containing (b) Source: device A; Target: device B

Fig. 11: Examples of classification results on T^{New}_{Target} using MIDNet in Table II and Table III. Top row (GT) contains ground truth labels. Middle row (TP) contains true positive results and bottom row (FP) contains false positive results.

(a) MIDNet (b) MIDNet+$\mathcal{L}_{trip}$

Fig. 12: Comparing categorical feature clusters between MIDNet and MIDNet+$\mathcal{L}_{trip}$. The clusters are visualized by t-SNE plots. The top row is the source domain (device A) and the bottom row is the target domain (device B). We observed that, in MIDNet+$\mathcal{L}_{trip}$, the features of unseen categories are mixed with those of seen categories in the target domain (see the right bottom), although the features of seen categories are more separated.

ble. II and Table. III). However, MIDNet+$\mathcal{L}_{trip}$ models are sometimes outperformed by MIDNet for unseen categories in the target domain T^{New}_{Target}. This may be caused by supervised distance metric learning. The distance metric learning encourages latent feature clusters of different seen categories in both domains to move away from each other, shown in the top row of Fig. 12. Such movement of the seen categories may lead to the inter-class mixture between unseen categories and seen categories (*e.g.*, Fig. 12 bottom row, lips and femur), and results in increased difficulty for identifying unseen categories in the target domain.

Medical images contain complex entangled image features. For example, shadow artifacts in US imaging are caused by anatomies through blocking the propagation of sound waves or destructive interference. Traditional DNN-based classifiers jointly learn shadow features and anatomical features without understanding the underlying semantics. By observing classification performance on source and target domain after separating entangled image features, our model can be potentially used to interpret the effective factors for target tasks. For example, the classification performance of T_{Source} and T_{Target} (Table. II MIDNet and MIDNet+$\mathcal{L}_{trip}$) indicate that shadow features can be more informative for some categories than the actual anatomy in anatomical classification.

The performance of semi-supervised learning usually positively correlates with the percentage of labeled data. In our experimental setting, excessive labeled data may lead to increased class imbalance. This may result in slightly decreased classification performance in Fig. 9 (b) when the percentage of labeled training data increases from 60% (i.e., all classes contain the same number of labeled images: 951 images) to 100% (i.e., *Abdominal*: 1448 labeled images vs. *Lips*: 1760 labeled images). A similar observation is made for a digits classification task in Appendix.

The proposed model is effective for learning knowledge from a source domain with a large training set and a target domain with a relatively small training set. This is demonstrated in the cross-device fetal US classification experiment, where the data from the source domain and the target domain contain respectively 2694 and 500 examinations. In the reversed scenario, the decision boundaries that successfully classify categories in a source domain with a small training set might have difficulty to classify unseen categories in a target domain with many samples. This might be the reason for a low classification performance for T^{New}_{Target} when switching the source and the target domains in the cross-device fetal US classification task (Sec. IV-D). Nevertheless, our model is meaningful for practical applications since it is reasonable to initially learn from a domain containing data from many patients and subsequently integrate knowledge from secondary datasets with smaller patient populations.

In experiments where the domain shift is caused by shadow artifacts, shadow-containing and shadow-free images are separated by an expert observer. To evaluate potential inter-observer variability, another two expert observers have been consulted to annotate a subset of 120 images as shadow-containing and shadow-free (10 shadow-containing/shadow-free images are randomly chosen for each category). The results show that 93.3% and 94.2% of the labels from these two experts are identical with the labels from the initial expert observer. A paired sample t-test between the label distributions

TABLE V: Classification performance on T_{Target}^{New} with MIDNet and DADA [45]. Best results in bold.

Methods	Source: device A; Target: device B		
	F1-score	Recall	Precision
DADA [45]	0.6198	0.6950	0.6067
MIDNet (Ours)	**0.8383**	**0.8600**	**0.8497**

yields p-values of 0.7973 and 0.3674. These high p-values indicate that there is no significant difference between the label distributions, hence the task is relatively straight-forward for skilled sonographers and thus inter-observer variability is negligible.

Hyper-parameters are highly debated in many related works. Our method has five hyper-parameters which are determined by extensive experiments. We fix $\lambda_1 = 1$ in Eq. 11 and adjust the other four hyper-parameters as follows. (1) We initialize these parameters according to their importance, *e.g.*, the parameter for the cross-entropy loss is initially relatively high since classification is our main task. (2) We adjust these parameters to ensure that losses decrease as expect on validation data. This adjustment is performed because losses may have different magnitudes, *e.g.*, in Eq. 11, $\mathcal{L}_{cls}$ is cross-entropy, $\mathcal{L}_{MI}$ is mutual information, $\mathcal{L}_{clus}$ and $\mathcal{L}_{SSL}$ are the Frobenius norm. (3) We run hyper-parameter combinations with grid search according to the results of the previous step, $\lambda_2 \in \{10, 20, 50\}, \lambda_3 \in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}, \lambda_4 \in \{10, 50, 100\}, \lambda_5 \in \{10, 50, 100\}$. (4) We select the hyper-parameters combination with the best result on validation data.

In this work, we compare the proposed method with the state-of-the-art methods that aim at one-to-one domain adaptation, *i.e.*, require a single source/target domain. We compare two domain adaptation methods which represent two groups of methods. DANN [17] is a typical adversarial-based method, which is widely used and MME [61] is a recent semi-supervised method using feature alignment for domain adaptation. Other types of domain adaptation frameworks (shown in Fig. 3) focus on a different domain adaptation scenario from ours and need multiple source/target domains. For example, domain generalization approaches requires multiple source domains and aim at learning universal knowledge from these multiple source domains. Domain agnostic learning approaches require multiple target domains and focus on transferring knowledge to multiple different target domains. Therefore, to ensure fair comparisons, our work excludes such multi-source/multi-target approaches. As closest related surrogate, we further compare the proposed method with a domain agnostic learning method that uses mutual information (DADA[6] [45] with a single target domain). Table V shows that our proposed method is better for improving classification performance on unseen categories in the single target domain.

For the classification task with shadow-free/shadow-containing fetal US images, acoustic shadows cause the domain shift. We explore to reduce this effect by 50% dropout in a VGG [62] and Res-VGG [62], [63]. However, Fig. VI shows that MIDNet outperforms VGG/Res-VGG+dropout, illustrating that dropout alone has limitation to reduce the domain shift caused by shadows in this task.

[6]https://github.com/VisionLearningGroup/DAL

TABLE VI: Comparison between domain-relevant data augmentations (elastic transformation) and MIDNet. Average F1-score is shown in the table. *Augment* includes dropout, contrast shift and brightness shift. *Augment*† includes Gaussian blur, contrast shift and brightness shift. Best results in bold.

Methods	Source: shadow-free; Target: shadow-containing		
	T_{Source}	T_{Target}	T_{Target}^{New}
VGG+dropout	**0.5846**	0.6501	0.6438
VGG+Augment	0.5723	0.6424	0.7080
Res-VGG+dropout	0.5003	0.5988	0.6878
Res-VGG+Augment	0.5067	0.5758	0.6492
MIDNet (Ours)	0.5484	**0.6809**	**0.7399**
Methods	Source: device A; Target: device B		
	T_{Source}	T_{Target}	T_{Target}^{New}
VGG+Augment†	0.7515	0.7016	0.7232
Res-VGG+Augment†	0.9149	0.5725	0.5979
MIDNet (Ours)	**0.9281**	**0.7434**	**0.8383**

VGG is VGG [62] and Res-VGG is Res-VGG [62], [63] in Table II and Table III.

A naive alternative to tackle distribution shift is data augmentation with elastic transformations, including contrast shift, brightness shift and Gaussian blur. Table VI compares the proposed method with data augmentations. For the classification task with shadow-free/shadow-containing fetal US images, contrast shift and brightness shift are used as data augmentation for VGG [62] and Res-VGG [62], [63] with dropout. For the cross-device fetal US classification task, contrast shift, brightness shift and Gaussian blur are adopted for data augmentation. Table VI shows that MIDNet outperforms VGG/Res-VGG+Augment and VGG/Res-VGG+Augment†. This demonstrates that the proposed method is better than naive data augmentation for tackling domain shift, especially for classification of unseen categories in the target domain.

VI. Conclusion

In this paper, we discuss a problem that is rarely evaluated but important in practical scenarios: transferring knowledge from known entangled image features (*e.g.*, categorical features and domain features) to unseen entangled image features (*e.g.*, categories from a target domain that are not available during training). We propose Mutual Information-based Disentangled Neural Networks (MIDNet) to extract generalizable features, which are essential for such scenarios. Our model is developed with a semi-supervised learning paradigm. Experiments on fetal US images demonstrate the efficiency and practical applicability of our method compared with the state-of-the-art. Source code will be publicly available after the review phase.

References

[1] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, ser. Neural Information Processing. MIT Press, 2008.

[2] F. Navarro, S. Conjeti, F. Tombari, and N. Navab, "Webly supervised learning for skin lesion classification," in *MICCAI*, 2018, pp. 398–406.

[3] C. Chen, Q. Dou, H. Chen, J. Qin, and P. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *AAAI*, 2019, pp. 865–872.

[4] Y. Tang, Y. Tang, V. Sandfort, J. Xiao, and R. M. Summers, "Tuna-net: Task-oriented unsupervised adversarial network for disease recognition in cross-domain chest x-rays," in *MICCAI*, 2019, pp. 431–440.

[5] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *IPMI*, 2017, pp. 597–609.

[6] J. Jiang *et al.*, "Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *MICCAI*, 2018, pp. 777–785.

[7] Q. Dou, D. C. Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *NeurIPS*, 2019.

[8] A. Chartsias *et al.*, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, 2019.

[9] Q. Dou, Q. Liu, P. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Transactions on Medical Imaging*, 2020.

[10] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *CVPR*, 2018.

[11] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *CVPR*, 2019.

[12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[13] K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, pp. 49–57, 2006.

[14] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv:1412.3474*, 12 2014.

[15] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015, pp. 97–105.

[16] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016, pp. 2058–2065.

[17] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 2962–2971.

[19] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," in *CVPR*, 2018, pp. 8867–8876.

[20] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017, pp. 95–104.

[21] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017, pp. 1857–1865.

[22] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *ICML*, 2018, pp. 1989–1998.

[23] J. Lezama, "Overcoming the disentanglement vs reconstruction trade-off via jacobian supervision," in *ICLR*, 2019.

[24] M. I. Belghazi *et al.*, "Mutual information neural estimation," in *ICML*, 2018, pp. 531–540.

[25] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*, 2019.

[26] NHS, *Fetal anomaly screening programme: programme handbook 2018*. Public Health England, 2018.

[27] L. J. Salomon *et al.*, "Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan," *Ultrasound Obstet Gynecol*, vol. 37, pp. 116–126, 2011.

[28] B. Holland, J. Myers, and C. Woods Jr, "Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise prior to planned neonatal cardiac surgery: a meta-analysis," *Ultrasound Obstet Gynecol*, vol. 45, no. 6, pp. 631–638, 2015.

[29] C. Baumgartner *et al.*, "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[30] A. Maraci, R. Napolitano, A. Papageorghiou, and J. Noble, "Searching for structures of interest in an ultrasound video sequence," in *International Workshop on Machine Learning in Medical Imaging*, 2014.

[31] M. Ahmed and J. Noble, "Fetal ultrasound image classification using a bag-of-words model trained on sonographers' eye movements," *Procedia Computer Science*, vol. 90, pp. 157–162, 12 2016.

[32] M. K. Feldman, S. Katyal, and M. S. Blackwood, "Us artifacts," *Radio Graphics*, vol. 29, p. 1179–1189, 2009.

[33] Q. Meng *et al.*, "Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging," *IEEE transactions on medical imaging*, April 2019.

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[35] X. Chen *et al.*, "Variational lossy autoencoder," in *ICLR*, 2016.

[36] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *NeurIPS*, 2018.

[37] A. Liu, Y. Liu, Y. Yeh, and Y. F. Wang, "A unified feature disentangler for multi-domain image translation and manipulation," in *NeurIPS*, 2018.

[38] N. Hadad, L. Wolf, and M. Shahar, "A two-step disentanglement method," in *CVPR*, 2018, pp. 772–780.

[39] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.

[40] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.

[41] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NeurIPS*, 2016.

[42] I. Higgins *et al.*, "beta-vae: Learning basic visual concepts with a constrained variational framework," *ICLR*, 2017.

[43] C. P. Burgess *et al.*, "Understanding disentangling in β-vae," *arXiv:1804.03599*, 2018.

[44] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representations using adversarial training," in *NeurIPS*, 2016, pp. 5040–5048.

[45] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *ICML*, 2019.

[46] Q. Meng, N. Pawlowski, D. Rueckert, and B. Kainz, "Representation disentanglement for multi-task learning with application to fetal ultrasound," in *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*, 2019, pp. 47–55.

[47] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.

[48] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML WREPL*, 2013.

[49] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017.

[50] T. Miyato, S. ichi Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1979–1993, 2018.

[51] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[52] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," in *ICLR*, 2019.

[53] Q. Meng, D. Rueckert, and B. Kainz, "Unsupervised cross-domain image classificationby distance metric guided feature alignment," *CoRR*, vol. arXiv:2008.08433, 2020.

[54] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019.

[55] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *ICCV*, 2019.

[56] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *ICML*, 2004.

[57] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *ICML*, 2007.

[58] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *ICML*, 2017, pp. 2208–2217.

[59] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[60] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *NeurIPS*, 2007, pp. 137–144.

[61] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *ICCV*, 2019.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[64] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[65] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

APPENDIX

A. Experiments on handwritten digits data

In this section, we demonstrate the efficiency of our method on an additional handwritten digits classification task (Fig. 13). MNIST is the source domain while MNIST-M is the target domain. Except *Source only*, all the methods are trained on digits 0 to 9 from the source domain and digits 0 to 4 from the target domain. We aim to separate digital features (categorical features) from domain features to obtain generalized digital features, and thus to achieve high digit classification performance on T_{Target}^{New} (digits 5 to 9 from target domain). Here, T_{Source} contains digits 0 to 9 from the source domain and T_{Target} contains digits 0 to 4 from the target domain. Detailed data split of this experiment is shown in Table. VIII. In this experiment, we focus on the effectiveness of our method on the unseen categories in the target domain. According to the two applications in the main paper, the feature clustering component mainly improves the classification performance of categories that are available during training. Therefore, we eliminate the feature clustering component in the proposed model and conduct the same evaluation on the digits classification task. Hyper-parameters λ_1 to λ_5 in Eq. 11 are experimentally chosen as $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10^{-3}, \lambda_4 = 10^2, \lambda_5 = 10^3$.

Fig. 13: Examples of handwritten digits dataset, containing MNIST [64] and MNIST-M [65].

Results: The experimental results of baselines and the ablation study are shown in Table. IX. From this table, we observe that the MIDNet model outperforms other baselines on all test data for average F1-score, recall and precision. For example, MIDNet achieves average F1-score of 0.9906, 0.8204 and 0.7166 for T_{Source}, T_{Target} and T_{Target}^{New}, respectively, while the highest average F1-score of other baselines on the corresponding test data are 0.9802 (Res-VGG [62], [63]), 0.7357 (MME [61]) and 0.5794 (Two-step-fair [38]). Additionally, MIDNet-III performs slightly better than MIDNet on T_{Target} and T_{Target}^{New}, demonstrating that feature consistency is important for digit classification. The results of MIDNet-IV and MIDNet-I (similarly, MIDNet-VI vs. MIDNet-II and MIDNet-VIII vs. MIDNet-V) illustrate the effectiveness of SSL based regularization in the proposed MIDNet.

We further compare the performance of MIDNet in a semi-supervised setting and a fully-supervised setting. Here, the semi-supervised setting utilizes the training data containing 30% labeled data and 70% unlabeled data, while the fully supervised setting only uses the 30% labeled data. The confusion matrix in Fig. 14 shows the effectiveness of unlabeled data in our proposed method, for example, the classification accuracy of T_{Target}^{New} greatly improves when integrating unlabeled data (semi-supervised).

To explore the importance of labeled data, we evaluate the performance of MIDNet based on using $15\%, 30\%, 60\%$ and 100% labeled data during training. Fig. 15 shows the average F1-score of these experiments on three groups of test data. From this figure, we observe that the classification performance only slightly improves with increasing labeled data. This indicates that MIDNet is capable of achieving expected performance with sparsely labeled data. Additionally, excessive labeled data may lead to increased class imbalance. This may result in decreased classification performance as shown in Fig. 15 when the percentage of labeled training data increased from 60% (i.e., all classes contain the same number of labeled images: 4294 images) to 100% (i.e., *Digit 1*: 10768 labeled images vs. *Digit 5*: 4336 labeled images).

B. Network architectures

We use residual units for the Encoders E_1/E_2 and Decoder D of the proposed model. The implementation of the residual units has been integrated from the publicly available DLTK framework [7]. Our implementation is on Tensorflow. The parameter settings (*e.g.*, filters and strides of encoders and decoders, hidden units of the classifier) for these experiments are shown in Table. VII.

TABLE VII: The parameter settings of the MIDNet architecture for different experiments. N is the batch size. Digits refers to the handwritten digits classification task that separates digit features from domain features. Fetal US refers to the fetal US standard plane classification tasks that disentangles (1) anatomies from shadow artifacts and (2) anatomies from image acquisition devices.

	Digits	Fetal US
Input dimension	(N, 28, 28, 3, 1)	(N, 224, 288, 1, 1)
Filters ($E_1/E_2/D$)	(8, 16, 32, 8)	(8, 16, 32, 64, 8)
Strides ($E_1/E_2/D$)	(1, 2, 2, 1)	(1, 2, 2, 2, 1)
Hidden units (C)	(128, 128)	(128, 128)

C. Details of data split

In this section, we provide detailed train/validation/test data split for fetal US classifications on two different datasets. Table. X is the fetal US data split for standard plane classification that **separate anatomical features from artifacts features**. Table. XI is the fetal US data split for standard plane classification that **disjoin anatomical features from imaging devices features**.

D. Additional classification results

We further show randomly selected true positive and false positive inference examples (images) from the three classification tasks in Fig. 16.

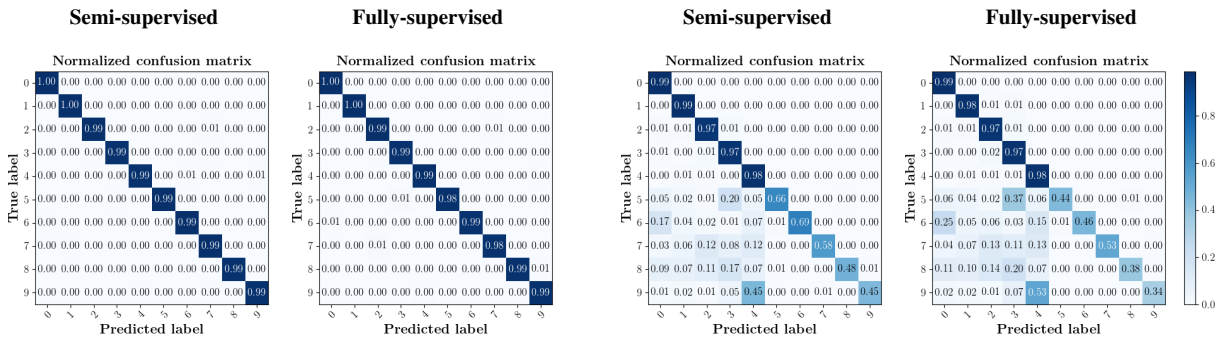
[7]https://dltk.github.io/

TABLE VIII: Data split of digits classification task. Blue test data is T_{Source}, Green test data is T_{Target} and red test data is T_{Target}^{New}. In this case, all labeled data is $\sim 30\%$ of all training data. Train:validation is about 8:2.

	MNIST (Source domain)										
	0	1	2	3	4	5	6	7	8	9	Unlabeled
Train	1087	1087	1087	1087	1087	2174	2174	2174	2174	2174	31690
Validation	1185	1349	1192	1227	1169	1085	1184	1253	1171	1190	–
Test	980	1135	1032	1010	982	892	958	1028	947	1009	–
	MNIST-M (Target domain)										
Train	1087	1087	1087	1087	1087	–	–	–	–	–	19039
Validation	1185	1349	1192	1227	1169	–	–	–	–	–	–
Test	980	1135	1032	1010	982	892	958	1028	947	1009	–

TABLE IX: Comparison of baselines, the state-of-the-art and ablation study (MIDNet-I to MIDNet) for **digit classification task**. 30% of training data are labeled data and the rest are unlabeled data. Average F1-score, Recall and Precision are measured on three groups of test data. Best results in bold.

Methods	T_{Source}			T_{Target}			T_{Target}^{New}		
	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision
Source only	0.9253	0.9254	0.9256	0.5309	0.5293	0.5340	0.5114	0.5118	0.5213
VGG [62]	0.9151	0.9162	0.9146	0.7334	0.8412	0.6517	0.6152	0.5208	0.7552
Res-VGG [62], [63]	0.9802	0.9802	0.9802	0.7236	0.9338	0.5953	0.5228	0.3595	0.9631
Two-step-fair [38]	0.8704	0.8707	0.8704	0.6908	0.7806	0.6203	0.5794	0.5002	0.6911
Two-step-Unfair [38]	0.7465	0.7492	0.7591	0.5839	0.6407	0.5428	0.2983	0.2598	0.3894
Multi-task [46]	0.9318	0.9315	0.9332	0.6203	0.7824	0.5171	0.5053	0.3695	0.8368
DANN [17]	0.9678	0.9679	0.9681	0.6818	0.8901	0.5579	0.4506	0.3023	0.9091
MME [61]	0.9709	0.9704	0.9726	0.7357	0.9426	0.6205	0.4858	0.3287	0.9722
MIDNet-I	0.9836	0.9837	0.9835	0.7039	0.9115	0.5797	0.4956	0.3376	0.9431
MIDNet-II	0.9841	0.9842	0.9842	0.7059	0.9160	0.5809	0.4916	0.3322	0.9501
MIDNet-III	0.9869	0.9869	0.9870	**0.8333**	0.9780	**0.7298**	**0.7511**	**0.6137**	0.9765
MIDNet-IV	0.9858	0.9860	0.9859	0.7439	0.9569	0.6169	0.5207	0.3566	0.9771
MIDNet-V	0.9863	0.9862	0.9864	0.8051	0.9766	0.6903	0.6821	0.5295	**0.9807**
MIDNet-VI	0.9868	0.9869	0.9868	0.7532	0.9602	0.6253	0.5541	0.3900	0.9689
MIDNet-VII	0.9881	0.9881	0.9881	0.8223	0.9779	0.7140	0.7280	0.5820	0.9791
MIDNet	**0.9906**	**0.9905**	**0.9906**	0.8204	**0.9803**	0.7108	0.7166	0.5704	0.9806



(a) $\mathbf{T_{Source}}$ (b) $\mathbf{T_{Target}}$ & $\mathbf{T_{Target}^{New}}$

Fig. 14: Confusion matrices for digit classification: semi-supervised setting versus fully-supervised setting with MIDNet-VIII model as backbone. For the semi-supervised setting, 30% of training data are labeled data and the rest are unlabeled data. The fully-supervised learning in this experiment only uses the 30% labeled data for training, without using unlabeled data.

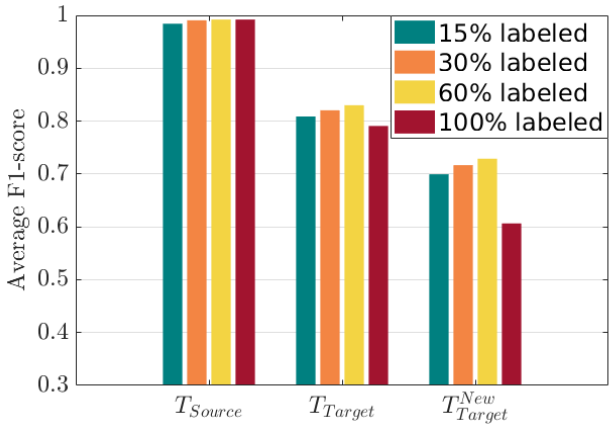


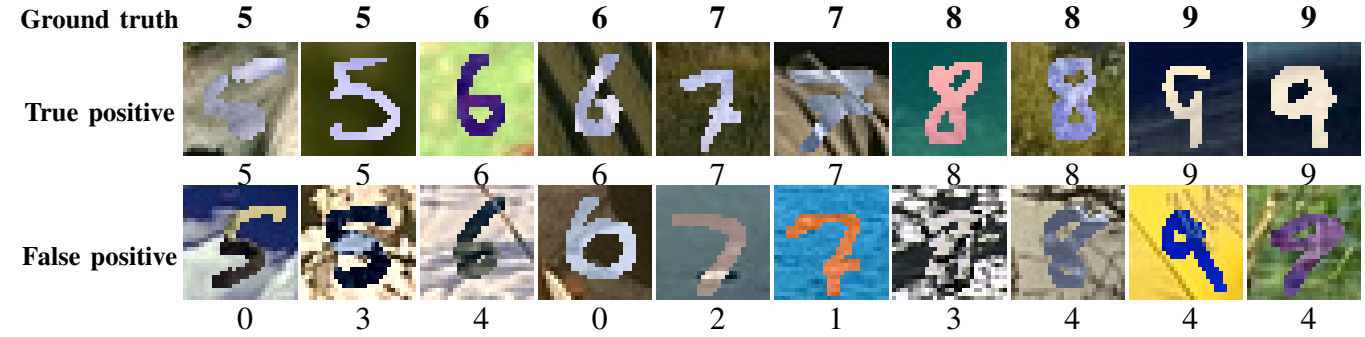
Fig. 15: Average F1-score of digit classification with different percentage of labeled data ($15\%, 30\%, 60\%, 100\%$) for semi-supervised learning based on MIDNet model.

TABLE X: Data split of fetal US standard plane classification (**anatomies vs. artifacts**). Blue test data is T_{Source}, Green test data is T_{Target} and red test data is T_{Target}^{New}. In this case, all labeled data is $\sim 30\%$ of all training data. Train:validation is about 8:2.

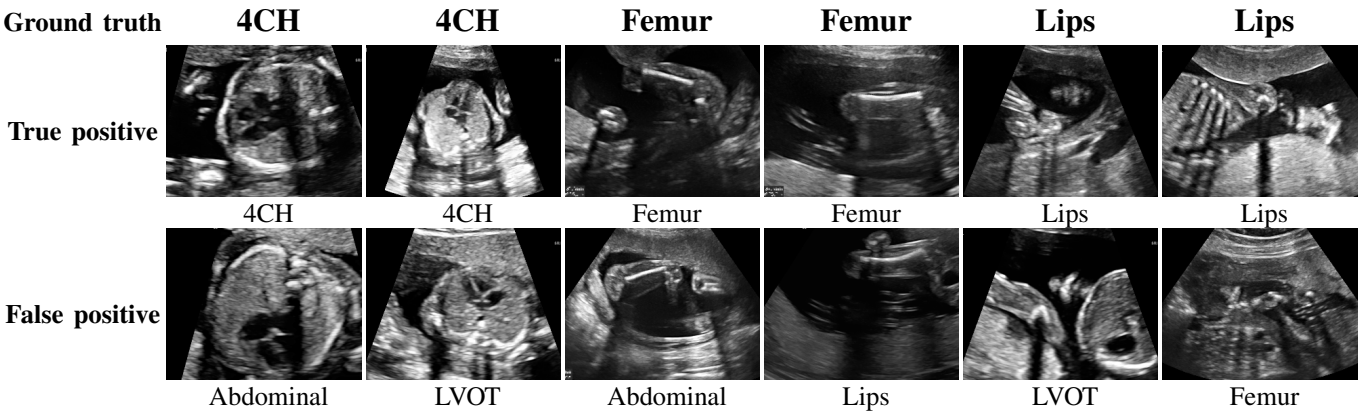
	Shadow-free fetal US (Source domain)						
	4CH	Abdominal	LVOT	RVOT	Lips	Femur	Unlabeled
Train	202	101	101	101	202	202	2125
Validation	139	63	115	106	166	167	–
Test	50	50	50	50	50	50	–
	Shadow-containing fetal US (Target domain)						
Train	–	101	101	101	–	–	710
Validation	–	119	73	60	–	–	–
Test	50	50	50	50	50	50	–

TABLE XI: Data split of fetal US standard plane classification (**anatomies vs. imaging devices**). Blue test data is T_{Source}, Green test data is T_{Target} and red test data is T_{Target}^{New}. In this case, all labeled data is $\sim 30\%$ of all training data. Train:validation is about 8:2.

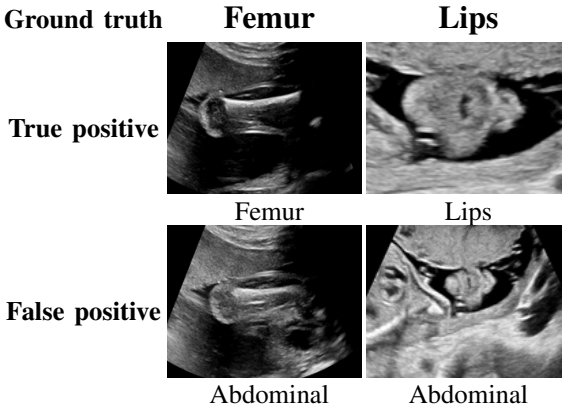
	Fetal US from imaging device A (GE Voluson E8)(Source domain)				
	Abdominal	Brain	Femur	Lips	Unlabeled
Train	237	237	475	475	3456
Validation	180	180	420	440	–
Test	100	100	100	100	–
	Fetal US from imaging device B ((Philips EPIQ V7 G)(Target domain)				
Train	237	237	–	–	992
Validation	182	184	–	–	–
Test	100	100	100	100	–



(a) Handwritten digits $\mathbf{T}^{\mathbf{New}}_{\mathbf{Target}}$ (target domain is MNIST-M)



(b) Fetal US data $\mathbf{T}^{\mathbf{New}}_{\mathbf{Target}}$ (target domain is shadow-containing image)

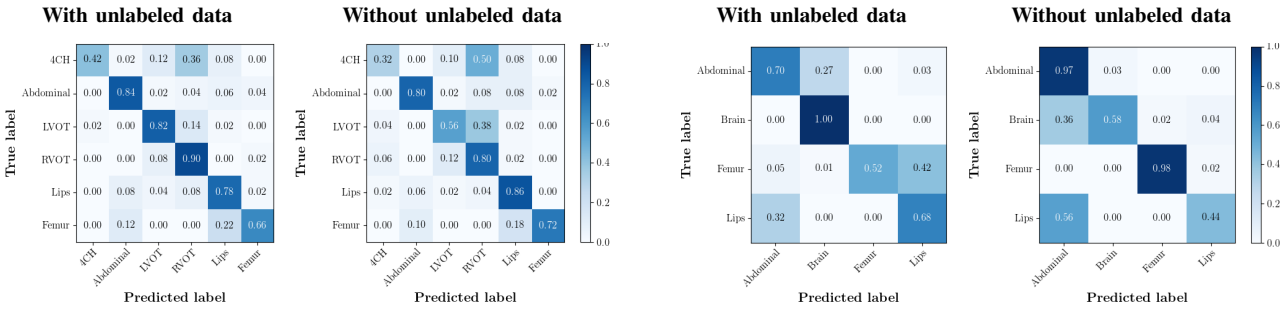


(c) Fetal US data $\mathbf{T}^{\mathbf{New}}_{\mathbf{Target}}$ (target domain is image from device B, Philips EPIQ V7 G)

Fig. 16: Inference examples of three classification tasks on three datasets using MIDNet. We show true positive and false positive samples from T^{New}_{Target}. **Ground truth** contains ground truth labels and labels under images are predicted labels. (a) Handwritten digits classification. (b) Fetal US standard plane classification that separate anatomical features from shadow artifacts features. (c) Fetal US standard plane classification that disjoin anatomical featurs from imaging device features.

E. MIDNet+$\mathcal{L}_{\text{trip}}$ with/without unlabeled data

We further compare the performance of MIDNet+$\mathcal{L}_{trip}$ with and without unlabeled data on the target domain (extended experiments of Fig. 8). Here, the *with unlabeled data* setting utilizes the training data containing 30% labeled data and 70% unlabeled data, while the *without unlabeled data* setting only uses the 30% labeled data. The confusion matrices in Fig. 17 show the effectiveness of unlabeled data. For example, the classification accuracy of T_{Target} in MIDNet+$\mathcal{L}_{trip}$ (e.g., *Abdominal*, *LVOT* and *RVOT* in Fig. 17 (a) and *Brain* in Fig. 17 (b)) improves when integrating unlabeled data.



(a) Source: shadow-free; Target: shadow-containing

(b) Source: device A; Target: device B

Fig. 17: Confusion matrices of $\mathbf{T}_{\mathbf{Target}}$ & $\mathbf{T}^{\mathbf{New}}_{\mathbf{Target}}$ in MIDNet+$\mathcal{L}_{trip}$ with and without unlabeled data. For the *with unlabeled data* setting, 30% of training data are labeled data and the rest are unlabeled data. The *without unlabeled data* setting only uses the 30% labeled data for training, without using unlabeled data. (a) Classification on fetal US with/without shadows. (b) Classification on fetal US from different image acquisition devices.

F. Clinical Relevance

In this section, we discuss in detail the connection between our work and clinical scenarios. The problem of domain shift is probably one of the key reasons hindering deployment of machine learning models in patient care at scale. Specifically, when a machine learning model is trained in one clinic, it rarely performs well in other clinics where different imaging devices and acquisition protocols are used. New medical devices, software and hardware alike, have to show consistent performance to be approved by regulators and to become acceptable for wide adoption.

A naive solution is to train a model on data from potentially thousands of sites but this is infeasible because of regulatory/economical constraints and data sharing limitations. Data often can not be shared because of (1) legality and (2) potential future revenue considerations.

Another possible solution is to train a new model at each clinical site. However, this solution is infeasible because it requires new clinical trials at high costs and introduces more risks for patients every time a new model is established.

Therefore, we need domain adaptation methods to reliably transfer models from one clinical site to another and understand their error margins. Fig. 18 shows the utilization of deep learning models in clinical scenarios with and without domain adaptation.

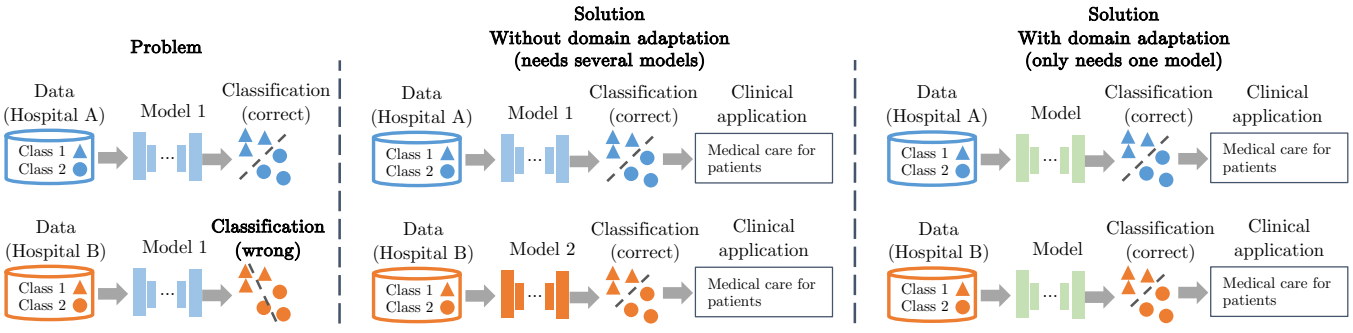


Fig. 18: A diagram of the utilization of deep learning models in clinical scenarios with and without domain adaptation.

Generalizing models across datasets with different feature distributions, *i.e.*, domain adaptation, enables a wider and more effective utilization of deep learning models for clinical applications. This helps clinicians from different clinical sites in a wide range of geographic areas to use the same task-specific model for the analysis of their own data (*e.g.*, obtained by

different image acquisition devices), and subsequently provide diagnostic decisions and treatment suggestions to their patients. In the medical image analysis community, many previous studies (*e.g.*, [5]–[9]) have focused on domain adaptation in different image modalities (*e.g.*, MRI and CT) and organs (*e.g.*, brain and heart) for various clinical applications (*e.g.*, brain lesion and cardiac segmentation). MICCAI (the leading international conference for medical image analysis) has a dedicated session and workshops (*e.g.*, DART) about the topic "domain adaptation". This demonstrates that domain adaptation addresses a common problem in medical imaging and is important for various clinical applications, and thus is relevant to clinical usage and patient care.

Specifically in our work, the proposed method is applied to a real-world clinical application, the classification of standardized fetal ultrasound views during prenatal screening. Standardization of anatomical view planes is key to empower the front-line-of-care during screening, making measurements comparable across patients and to accurately predict outcomes. We focus on domain adaptation between fetal ultrasound datasets with different feature distributions, which are caused by imaging artifacts and different acquisition devices. Our study enables learning-based classifiers to be effectively utilized on a wider range of fetal ultrasound images. This helps early detection of pathological development independent from the used imaging setup. The detection of abnormalities can inform downstream treatment decisions and delivery options [28]. Fig. 19 illustrates and compares the potential impact of our method and other domain adaptation methods on machine learning for patient care in the context of fetal screening.

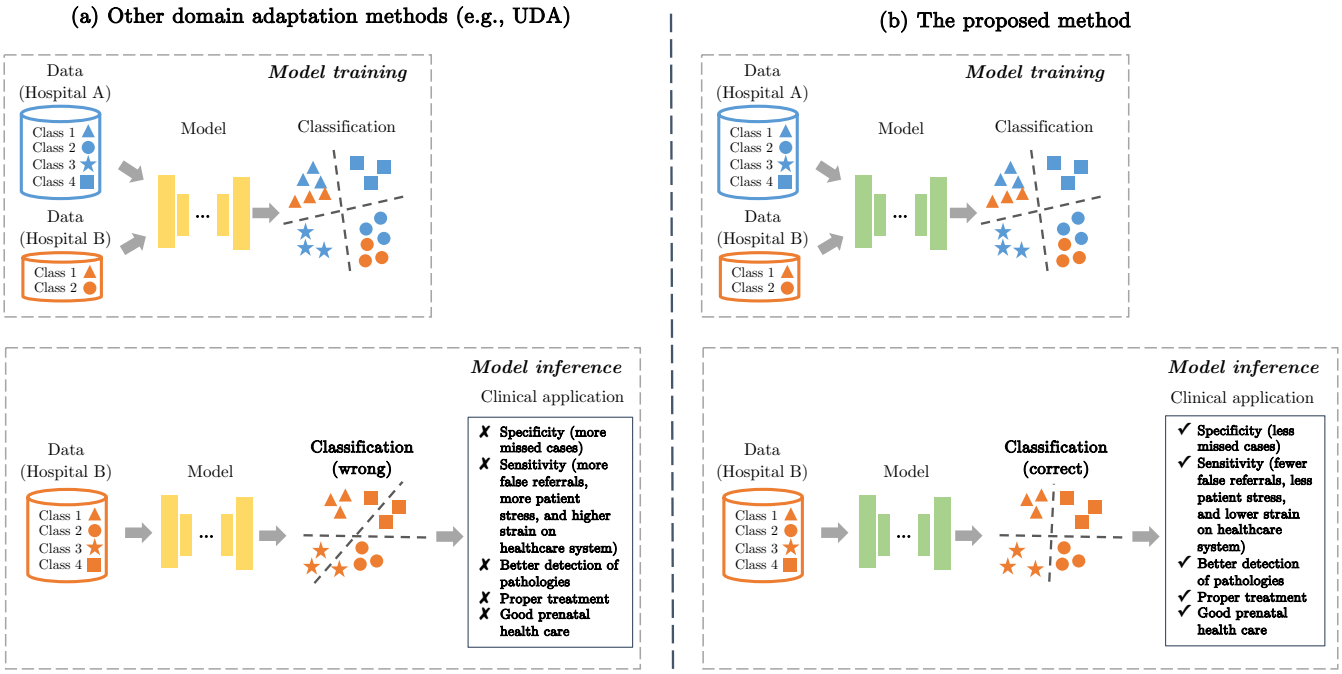


Fig. 19: (a) The use of other domain adaptation methods. (b) The proposed method.