Learning Frequency and Structure in UDA for Medical Object Detection

Liwen Wang[1*], Xiaoyan Zhang[1*], Guannan He[3], Ying Tan[4], Shengli Li[4], Bin Pu[2], Zhe Jin[1], Wen Sha[1], Xingbo Dong[1(✉)]

[1] Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, School of Artificial Intelligence, Anhui University,Hefei,China
[2] The Hong Kong University of Science and Technology,HKSAR,China
[3] Sichuan Provincial Maternity and Child Health Care Hospital,Chengdu,China
[4] Shenzhen Maternity and Child Healthcare Hospital,Shenzhen,China
xingbod@gmail.com

Abstract. In medical imaging applications, particularly in cardiac and skeletal analysis, the anatomical structure detection is crucial for diagnosing cardiac disease and other disease. However, the domain gap between images acquired from different sources or modalities poses a significant challenge and impedes model generalization across diverse patient populations and imaging conditions. Bridging this gap is particularly essential in image-based diagnosis, where subtle variations in anatomical structures and imaging characteristics can profoundly impact diagnostic performance. Take fetal cardiac ultrasound images as an example, this paper proposes a novel method for unsupervised domain adaptive fetal cardiac structure detection. The method integrates both the frequency-based distributional properties and anatomical structural information inherent in medical images. Specifically, we introduce a Frequency Distribution Alignment (FDA) module and an Organ Structure Alignment (OSA) module to mitigate detection misalignment across different hospital settings. We demonstrates the effectiveness of these modules through extensive experiments. Our method significantly improves the performance of fetal cardiac structure detection tasks, enabling adaptation to diverse hospital scenarios and showcasing its potential in addressing domain gaps in medical imaging.

Keywords: Unsupervised domain adaption · Medical image analysis · Object detection.

1 Introduction

In medical imaging, especially in the analysis of the heart and skeleton, deep learning (DL) has the potential to significantly enhance diagnostic accuracy and clinical decision-making [1]. Nevertheless, the presence of a domain gap between

[*] Equal Contribution.
[✉] Corresponding authors.

medical images acquired from different hospitals or imaging devices poses a substantial challenge [10,39]. Addressing this disparity is crucial in medical imaging applications, as diagnostic performance can be profoundly affected by subtle variations in anatomical structures and imaging characteristics. [33].

Unsupervised domain adaptation (UDA) techniques aim to optimize the performance of the target domain/hospital while minimizing the need for expert supervision through invariant feature learning [35], self-training [41,16], image translation [5,12], domain randomization [17,31], etc. However, directly applying DL-based models to anatomical structure detection in ultrasound data often yields relatively poorer results, particularly when dealing with data sourced from multiple healthcare centers [10]. This is primarily due to domain gaps present in real-world datasets [26,18], resulting from variations in data collection devices and scanning techniques among obstetricians across different hospital centers.

While object relationships may appear chaotic in natural images, medical images adhere to anatomical principles. For instance, in fetal cardiac imaging, paired ribs consistently flank the heart. Moreover, organs in images from different centers exhibit morphologically consistency. For example, in Fig 1 and 3, the SP in the source and target domains are visually similar. In clinical practice, sonographers rely on topology and morphology to diagnose the disease, shedding light on Unsupervised Domain Adaptation (UDA). As shown in Fig 2, anatomical structures of the same view remain consistent in topology and morphology. Here, topological information refers to the angle and distance information of anatomical structures, while morphological information refers to the shape and the morphological features in the frequency domain.

The unique characteristics of medical images indicate that previous methods for UDA object detection (UDAOD) in natural scenarios are not suitable or applicable to our task. Meanwhile, existing domain adaptation techniques often fall short in effectively leveraging the intricate anatomical structures and morphological features specific to medical objects. Motivated by these observations, we propose a novel UDA method named FS (**F**requency and **S**tructure)-UDA for medical object detection. This method includes two modules: **F**requency **D**istribution **A**lignment (FDA) and **O**rgan-**S**tructure **A**lignment (OSA).

In FDA, based on our observations, identical organs tend to exhibit high visual similarity in the ground truth. From this, we propose that the frequency-based morphological information of the correct pesudo-labels in the target domain should resemble those of the labeled annotations in the source domain images. In OSA, the consistent spatial relationships among anatomical structures are characterized by similar distance ratios and specific angular configurations. For example, in Fig 2, the angular and ratio of distance ranges defined by the positions of the two ribs and spine remain consistent across samples. To accurately align this intricate anatomical topology, we incorporate angle and distance ratios into the matrix representation and calculate distances using L2 norm. By integrating both FDA and OSA, our proposed methodology aims to bridge the domain gap. To conclude, our contributions include:

- We propose a novel method called FS-UDA to tackle above challenges, which reduces the domain gap and better detects fetal heart structures across different hospital centres.
- For medical scenarios, we introduce frequency distribution alignment to address detection misalignment by aligning frequency-based morphological knowledge. Additionally, we design organ-structure alignment to synchronize detection in source and target domains by aligning topological representation.
- Extensive experiments demonstrate that the proposed method outperforms existing state-of-the-art methods.

2 Related Work

2.1 Unsupervised Domain Adaptive Object Detection

Recently, the UDAOD task has garnered significant attention for its potential to enhance model generalization across diverse domains [32,36,43,41]. Studies in this field can be broadly categorized into adversarial learning [44,35], self-training [16,38,14], image-to-image translation [17,13], and others [7,21,31,29]. Within adversarial learning, Li et al. proposed Sigma, leveraging semantic complete graph matching to facilitate domain adaptation effectively [19]. In the realm of self-training, Kim et al. introduced a paradigm termed Diversify and Match, emphasizing the importance of domain adaptive representation learning for object detection [17]. Additionally, Gao et al. presented Asyfod, an asymmetric adaptation paradigm tailored for few-shot domain adaptive object detection [8].Hsu et al. introduced a technique called Progressive Domain Adaptation for Object Detection, addressing domain adaptation challenges in object detection by presenting a progressive domain adaptation approach [12]. Mattolin et al. introduced Confmix, focusing on unsupervised domain adaptation via confidence-based mixing to enhance detection accuracy [24]. Cao et al. adopted a contrastive mean teacher approach for training domain adaptive object detectors in an unsupervised setting [4]. Furthermore, Gao et al. proposed Acrofod, an adaptive method specifically designed for cross-domain few-shot object detection [9].

However, most of the above methods are tailored for natural images and do not fully account for the structural information present in medical images, rendering them potentially unsuitable for medical applications.

2.2 Application of Structure Information in Deep Learning

The integration of anatomical structure information into deep learning frameworks has shown promise in various applications. Chen et al. emphasized the importance of harmonizing transferability and discriminability to effectively adapt object detectors, leveraging anatomical structure information [5]. Additionally, Pu et al. proposed ToMo-UDA, which incorporates Topology and Morphology information alignment for effective anatomical structure detection [29]. Ni et al.

introduced the misalignment-robust frequency distribution loss, addressing issues related to structural misalignment during image transformation [25]. These studies demonstrate innovative approaches to domain adaptation and object detection. In the medical imaging domain, Dalca et al. introduced a technique that incorporates anatomical priors into convolutional neural networks for unsupervised biomedical segmentation, resulting in enhanced accuracy in delineating anatomical structures in medical images [6]. Similarly, Sindagi et al. emphasised the significance of anatomical priors in improving the robustness of object detection models [32]. Zhao et al. designing the Adaptive Relation Graph Reasoning (ARGR) module, anatomical structures are treated as nodes, with two kinds of relationships between nodes modeled as edges.[42]. Yu et al. leverages deep learning to emphasize fetal cardiac anatomy structure in ultrasound image segmentation [22]. Additionally, Yang et al. introduced Graphecho, utilizing graph-driven unsupervised domain adaptation techniques for echocardiogram video segmentation [37].

As previously mentioned, the incorporation of anatomical structures and domain-specific features into UDAOD frameworks can address some unique challenges in medical image analysis tasks, ultimately improving model performance and generalization. However, previous studies have not fully explored topology and morphology knowledge in both source and target domains.

3 Proposed Method

As depicted in Fig 1, our approach begins with an annotated source image and an unlabeled target image, and it comprises several essential modules. Initially, we employ histogram matching to alleviate the domain gap within the spatial domain, ensuring that a pair of images is unbiased to lighting conditions. Subsequently, a shared feature extractor is used to derive features denoted as $\mathcal{F}^s$ and $\mathcal{F}^t$, followed by the integration of FCOS (Fully Convolutional One-Stage) heads. Then, we incorporate **Frequency Distribution Alignment (FDA) Module** and **Organ Structure Alignment (OSA) Module**. For more details, see sections 3.1 and 3.2, respectively.

3.1 Frequency Distribution Alignment

The Wasserstein Distance (WD) is commonly used for optimizing neural networks by quantifying the dissimilarity between probability distributions. Unlike focusing on spatial alignment, as emphasized by [40], WD instead estimates discrepancies between the underlying distributions of signals. However, the neglect of spatial information may compromise the structural accuracy of predicted results. As argued by [25], utilizing global information can address this issue more effectively than relying solely on local information. Therefore, computing WD in the frequency domain is expected to better preserve the structural accuracy of predictions due to the richer global information it encompasses, as observed in [3,15,45].

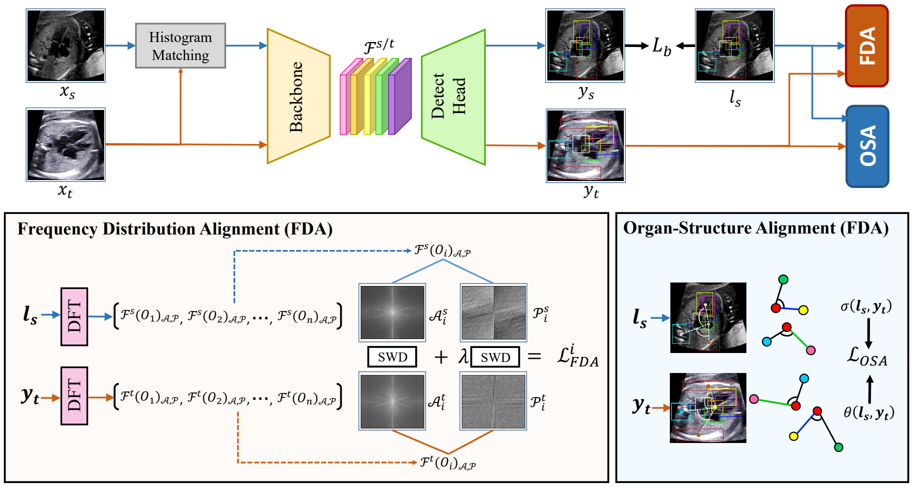

x_s
x_t
Histogram Matching
Backbone
$\mathcal{F}^{s/t}$
Detect Head
y_s
y_t
L_b
l_s
FDA
OSA
Frequency Distribution Alignment (FDA)
l_s
y_t
DFT
DFT
$\{\mathcal{F}^s(O_1)_{\mathcal{A},\mathcal{P}}, \mathcal{F}^s(O_2)_{\mathcal{A},\mathcal{P}}, \cdots, \mathcal{F}^s(O_n)_{\mathcal{A},\mathcal{P}}\}$
$\{\mathcal{F}^t(O_1)_{\mathcal{A},\mathcal{P}}, \mathcal{F}^t(O_2)_{\mathcal{A},\mathcal{P}}, \cdots, \mathcal{F}^t(O_n)_{\mathcal{A},\mathcal{P}}\}$
$\mathcal{F}^s(O_i)_{\mathcal{A},\mathcal{P}}$
$\mathcal{F}^t(O_i)_{\mathcal{A},\mathcal{P}}$
$\mathcal{A}_i^s$
$\mathcal{P}_i^s$
$\mathcal{A}_i^t$
$\mathcal{P}_i^t$
SWD + λ SWD = $\mathcal{L}_{FDA}^i$
Organ-Structure Alignment (FDA)
l_s
y_t
$\sigma(l_s, y_t)$
$\mathcal{L}_{OSA}$
$\theta(l_s, y_t)$

Fig. 1. The overview pipeline of the FS-UDA.

As such, calculating the WD in the frequency domain enhances the structural accuracy of predictions in image transformation. This approach allows for effectively measuring the structural (morphological) similarity between two images by computing the WD in the frequency domain. Given the similarity in appearance of organs in ultrasound images, and the fact that ultrasonographers typically assess organs based on their appearance, we can evaluate the accuracy of detection results by calculating the WD between the sonographers' annotations and the detected pseudo-bounding boxes.

The main objective of this module is to compute the frequency distance of the same organ from both domains. Before that, it's important to address the impact of lighting conditions on frequency information. Since images from the two domains are collected from different medical centers, their lighting distributions may differ, contributing to the occurrence of a domain gap. Therefore, prior to transmitting images from both domains into the network, a histogram matching is performed on the source domain image with target domain image, so that its lighting distribution is closer to that of the target domain image, thereby eliminating part of the domain gap and avoiding the interference of lighting information on the loss of frequency distribution. Given the ground truth bounding boxes $Y^s \in \mathbb{R}^{n^s \times 4}$ from the source domain and pseudo bounding boxes $\hat{Y}^t \in \mathbb{R}^{n^t \times 4}$ from the target domain, where $n^{s/t}$ indicates the total number of organs. Initially, we extract the corresponding regions from the original images based on the provided bounding boxes. Subsequently, a shared encoder based on VGG is utilized to derive features of the cropped regions, denoted as $\hat{\mathcal{F}}^s$ and $\hat{\mathcal{F}}^t$, respectively. To compute WD in the frequency domain, we initially apply the Discrete Fourier Transform (DFT) to convert feature signals into the frequency domain. This transformation yields frequency components, including both amplitude and phase, which encapsulate all the information in the frequency domain. So we obtain the amplitude and phase of these two features through DFT respectively and mix the amplitude of $\hat{\mathcal{F}}^s$ and the phase of $\hat{\mathcal{F}}^t$.

Then, we propose the **F**requency **D**istribution **A**lignment (FDA) between the ground truth and pseudo labels, formulated as:

$$\mathcal{L}_{\text{FDA}}(\hat{\mathcal{F}}^s, \hat{\mathcal{F}}^t) = \sum_{i=0}^{n} \text{SW}\left(\mathcal{A}_{(\hat{\mathcal{F}}_i^s)}, \mathcal{A}_{(\hat{\mathcal{F}}_i^t)}\right) + \lambda \cdot \text{SW}\left(\mathcal{P}_{(\hat{\mathcal{F}}_i^s)}, \mathcal{P}_{(\hat{\mathcal{F}}_i^t)}\right), \tag{1}$$

where $SW(\cdot,\cdot)$ denotes the Sliced Wasserstein Distance (SWD) between the distributions of two signals. $\mathcal{P}_{(\hat{F}_i^t)}$ and $\mathcal{A}_{(\hat{F}_i^t)}$ denotes the phase and amplitude of $\hat{F}_i^t$, respectively. Others are similar. The hyperparameters λ are 0.1 in our experiments. The SWD serves as an approximation to the WD, which lacks a closed-form solution in high-dimensional spaces. In practice, SWD is approximated by employing a simple Monte Carlo scheme defined below::

$$SW_p\left(\mathcal{A}_{(\hat{\mathcal{F}}_i^s)}, \mathcal{A}_{(\hat{\mathcal{F}}_i^t)}\right) \approx \left(\frac{1}{L}\sum_{l=1}^{L} W_p^p\left(\mathcal{A}_{(\hat{\mathcal{F}}_i^s)_l}, \mathcal{A}_{(\hat{\mathcal{F}}_i^t)_l}\right)\right)^{1/p}, \tag{2}$$

where L denotes the length of sliced $\mathcal{A}_{(\hat{F}_i^t)}$ and W_p refers to p^{th} Wasserstein distance. In our methods, we use $p = 2$ for the computation of this module.

3.2 Organ-Structure Alignment Module

In clinical practice, obstetricians typically scan organs from a fixed orientation to obtain fetal ultrasound images. We propose that this principle can be applied to ensure a consistent representation of organ structures within the same ultrasound plane across various medical centers. As shown in Figure 2, the angles between the spine and two ribs at the intersection of the interventricular septum remain consistent across different domains. Additionally, the ratio between the distance of different organs to the center of mass also remains consistent across domains.

To utilize this knowledge for narrowing the domain gap between source and target domains, we introduce the Organ Structure Alignment (OSA) module, which innovatively incorporates angular and distance information to construct the domain relationship.

Organ-Structure Relation Construction. During the organ-structure construction in the training phase, the main objective of this module is to compute the angular and distance information between different organs from both domains. Given the ground truth bounding boxes $Y^s \in \mathbb{R}^{n^s \times 4}$ for the source domain and pseudo bounding boxes $\hat{Y}^t \in \mathbb{R}^{n^t \times 4}$ from the target domain, where $n^{s/t}$ indicates the total number of organs. We first compute the centroid locations $c^s \in \mathbb{R}^{n^s \times 2}$ and $c^t \in \mathbb{R}^{n^t \times 2}$ of each organ according to Y_i^s and $\hat{Y}_i^t$ from both domains. To ensure consistency in the number and types of detected organs within the target domain, we supplement the missing classes by utilizing annotations from corresponding classes in the source domain and remove any extra classes that have not been detected in the source domain. With this operation,

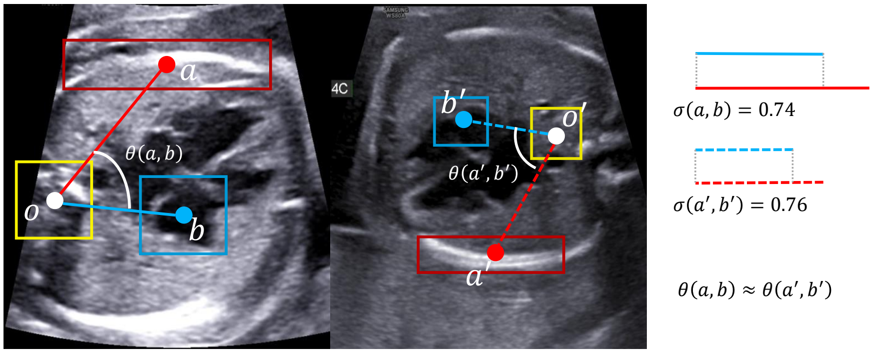


Fig. 2. Consistency of orientation-based organ structures for the same view.

the centroid locations can be reformulated as $c^{s/t} \in \mathbb{R}^{n\times 2}$, where n is the total number of classes in the source domain. Subsequently, angular and distance information is obtained by calculating the angles among the centroid locations of each organ in $c^{s/t}$. Adjacency matrices represent the angular relationship of organs within a domain, as defined below:

$$A^{s/t} = \begin{pmatrix} \theta(c_0, c_0) & \cdots & \theta(c_0, c_j) \\ \vdots & \vdots & \vdots \\ \theta(c_i, c_0) & \cdots & \theta(c_i, c_j) \end{pmatrix}^{s/t}, D^{s/t} = \begin{pmatrix} \sigma(c_0, c_0) & \cdots & \sigma(c_0, c_j) \\ \vdots & \vdots & \vdots \\ \sigma(c_i, c_0) & \cdots & \sigma(c_i, c_j) \end{pmatrix}^{s/t}, \quad (3)$$

where $\theta(\cdot,\cdot)$ calculates the angle between organs, referred by the centroid locations provided from Y^s and $\hat{Y}^t$. Similarly, distance information is represented by matrices $D^{s/t}$, computed using $\sigma(\cdot,\cdot)$. Given two centroids $\mathbf{a}, \mathbf{b}$ and center of mass $\mathbf{o}$, $\theta(\mathbf{a}, \mathbf{b})$ and $\sigma(\mathbf{a}, \mathbf{b})$ as define below:

$$\theta(\mathbf{a}, \mathbf{b}) = \arctan\left(\measuredangle(a - o, b - o)\right), \sigma(\mathbf{a}, \mathbf{b}) = \frac{\|\mathbf{b} - \mathbf{o}\|_2}{\|\mathbf{a} - \mathbf{o}\|_2}. \quad (4)$$

Organ-Structure Alignment. During the alignment stage, we aim to maximize the similarity of the adjacency matrices. Each element in matrices $A^{s/t}$ and $D^{s/t}$ from the source and target domains corresponds one-to-one. Therefore, we compute the L2 norm distance between matrices A^s and A^t, D^s and D^t, and the optimization objective function is formulated as:

$$\mathcal{L}_{\text{OSA}} = \sum\nolimits_{i=0}^{n} \sum\nolimits_{i=0}^{n} ||A^s_{i,j} - A^t_{i,j}||_2 + ||D^s_{i,j} - D^t_{i,j}||_2. \quad (5)$$

3.3 Overall Loss Function

The overall loss (Eq. (6)) comprises supervised loss and domain adaptation loss. The supervised loss $\mathcal{L}_{\text{sup}}$ is the objection detection loss in the source domain, while the domain adaptation loss combines $\mathcal{L}_{\mathrm{FDA}}$ from the FDA module and $\mathcal{L}_{\mathrm{OSA}}$ from the OSA module.

$$\mathcal{L}_{\mathrm{all}} = \lambda_1 \mathcal{L}_{\mathrm{sup}} + \lambda_2 \mathcal{L}_{\mathrm{FDA}} + \lambda_3 \mathcal{L}_{\mathrm{OSA}}. \tag{6}$$

where $\lambda_{1,2,3}$ are the weights of different loss terms, set as 1.0, 1.0, and 0.5, respectively, in our experiment setting.

4 Experiments

4.1 Datasets and Evaluation

In the experiments, we employed two datasets of cardiac standard views as examples. The datasets are described as follows:

Fetal Cardiac Structure (FCS)[30]: The FCS dataset comprises ultrasound data obtained from two medical centers, capturing two cardiac views: the three-vessel and trachea view (3VT) and the four-chamber cardiac view (4C). It consists of four datasets acquired using different medical devices, including Samsung, Sonoscape, and Philips, covering a gestational week range of 20-34 weeks. In this article, we focus on 4C and 3VT views. The **4C** dataset contains nine anatomical structures, including LV, LA, RV, DAO, RA, VS, SP, RIB, and CRO, while the **3VT** dataset includes SVC, AOA, T, SP, PTDA, and DAO.

CardiacUDA Dataset [37]: The CardiacUDA dataset was utilized for video-based cardiac structure segmentation tasks from two hospitals, sites G and R. We converted the segmentation masks of LA, LV, RA, and RV into box-level annotations to facilitate the UDAOD task. Due to numerous overlapping frames in the video, we sampleed one frame per video.

We conducted adaptation experiments primarily between hospitals 1 and 2, where adaptation occurred from hospital 1 (source) to hospital 2 (target), denoted as hospital1→2. Throughout the unsupervised domain adaptation (UDA) process, we utilized training data from both the source and target domains, evaluating performance on the target domain's test data. Performance was measured using the mean average precision (mAP) metric with an Intersection over Union (IoU) threshold of 0.5. The dataset was split into training, testing, and validation sets in a 7:2:1 ratio for each view.

4.2 Implementation Details

We employed ResNet-101 [11] as the feature extractor and FCOS [34] as the detector, implemented in PyTorch [27]. We trained the FS-UDA using the stochastic gradient descent (SGD) optimizer [2] with an initial learning rate of 0.01, fixed for 300 epochs, with a batch size of 4, and weight decay of 5×10^{-4}. Due to device variation across hospitals, images were uniformly resized to 800×1000. As the model's prediction pseudo-labels are unreliable in the early training stages, resulting in fewer correct matches, the FDA and OSA modules were activated only after reaching 5 epochs.

Table 1. Quantitative adaptation results on 4C.

Hospital 1→2										
Method	RA	RV	LV	VS	SP	LA	CR	DAO	RIB	mAP
Without DA	59.59	45.88	53.04	52.30	64.54	57.89	62.84	57.57	61.43	57.20
Few-shot Domain Adaptation Object Detection Methods										
AcroFOD [9] ECCV'22	25.52	21.04	24.64	25.86	26.32	24.64	26.75	25.74	27.24	25.30
AsyFOD [8] CVPR'23	48.31	48.66	48.66	49.93	48.36	50.00	49.54	47.51	47.43	48.71
Unsupervised Domain Adaptation Object Detection Methods										
ConfMix [24] WACV'21	62.40	65.30	65.20	61.80	59.70	63.20	69.50	67.80	62.70	64.20
SIGMA [19] CVPR'22	70.64	56.57	64.16	64.58	66.91	61.31	74.2	68.64	69.97	66.33
LRA [28] TNNLS'23	63.23	53.25	58.24	59.56	64.30	**84.32**	66.97	55.98	59.18	62.78
CMT [4] CVPR'23	79.18	64.87	66.31	64.34	74.84	66.23	71.61	60.93	68.66	68.55
SIGMA++ [20] TPAMI'23	56.10	47.01	52.72	51.38	63.65	52.11	60.28	62.99	67.30	57.06
Ours(FS-UDA)	**80.42**	**81.69**	**79.95**	**78.55**	**77.90**	81.33	**82.22**	**76.60**	**78.86**	**79.72**
Target Only	82.30	74.37	78.29	79.25	88.26	83.02	86.82	87.22	85.16	82.74

Hospital 2→1										
Method	RA	RV	LV	VS	SP	LA	CR	DAO	RIB	mAP
Without DA	70.56	49.82	54.26	61.88	73.04	64.48	74.25	77.32	53.28	64.32
Few-shot Domain Adaptation Object Detection Methods										
AcroFOD [9] ECCV'22	37.31	39.86	38.84	40.47	35.13	39.54	39.88	37.64	27.13	37.02
AsyFOD [8] CVPR'23	71.94	**69.09**	**70.52**	71.61	71.05	70.61	71.71	69.28	55.97	69.08
Unsupervised Domain Adaptation Object Detection Methods										
ConfMix [24] WACV'21	58.90	64.00	63.40	61.50	63.40	55.40	64.90	60.00	46.30	59.80
SIGMA [19] CVPR'22	70.64	56.57	64.16	64.58	66.91	61.31	74.2	68.64	69.97	66.33
LRA [28] TNNLS'23	75.43	37.08	49.24	51.84	54.88	48.92	52.38	58.70	58.58	54.11
CMT [4] CVPR'23	81.53	63.30	66.35	68.98	77.30	76.31	76.70	67.67	58.42	70.40
SIGMA++ [20] TPAMI'23	70.29	54.26	55.10	63.12	70.15	62.36	75.23	74.23	55.47	67.47
Ours(FS-UDA)	**90.11**	**79.77**	**84.54**	**90.17**	**85.78**	**87.27**	**90.01**	**87.56**	**73.09**	**85.37**
Target Only	86.61	82.75	83.48	85.93	90.16	82.92	89.09	89.61	72.51	84.78

Table 2. Quantitative adaptation results on 3VT.

	Hospital 1→2							**Hospital 2→1**						
Method	DAO	SP	PTDA	T	SVC	AOA	mAP	DAO	SP	PTDA	T	SVC	AOA	mAP
Without DA	24.91	31.66	47.89	25.59	38.07	52.54	36.78	38.64	48.80	37.73	34.64	41.39	48.45	41.61
Few-shot Domain Adaptation Object Detection Methods														
AcroFOD [9] ECCV'22	50.19	57.90	64.99	52.04	56.19	60.12	56.90	58.82	58.83	60.45	56.07	48.99	61.81	57.49
AsyFOD [8] CVPR'23	49.11	51.32	60.29	44.54	53.11	61.63	53.33	49.10	49.92	49.25	46.15	41.41	50.47	47.71
Unupervised Domain Adaptation Object Detection Methods														
ConfMix [24] WACV'21	41.80	67.10	54.20	**70.40**	**63.50**	59.20	59.40	60.38	60.37	43.09	21.67	27.17	48.30	43.50
SIGMA [19] CVPR'22	42.92	42.83	59.41	39.63	41.68	59.97	47.74	36.34	42.52	38.62	39.67	35.20	48.64	40.17
LRA [28] TNNLS'23	**53.50**	62.80	37.31	47.91	44.47	75.76	56.32	14.74	17.23	3.36	16.01	4.65	24.09	13.34
CMT [4] CVPR'23	45.43	60.11	**81.46**	27.63	45.64	63.68	53.99	16.53	27.09	27.41	27.63	20.50	40.44	22.74
SIGMA++ [20] TPAMI'23	42.29	37.39	45.36	28.95	31.98	42.87	38.14	33.71	42.77	31.56	34.91	32.07	44.38	36.57
Ours(FS-UDA)	53.44	**71.20**	79.72	45.11	56.37	**71.69**	**62.92**	**66.83**	**62.40**	**64.89**	**59.37**	**58.30**	**72.79**	**64.10**
Target Only	65.14	71.11	71.81	64.77	53.37	71.94	66.36	82.49	81.36	85.20	76.28	73.85	90.34	81.59

4.3 Comparison with State-of-the-Arts

In this experiment, we evaluated the performance of FS-UDA using both a Fetal Cardiac Structure benchmark and the widely used CardiacUDA dataset, focusing on scenarios where object detectors must adapt between hospital centers.

Hospital 1→2 on 4C: Results in Table 1 demonstrate that FS-UDA outperforms all existing state-of-the-art studies, including few-shot DA object detection methods such as CMT [4], by a substantial margin of 11.27% mAP. This

Table 3. Quantitative adaptation results on 4C of CardiacUDA.

site G→site R					
Method	LA	RA	LV	RV	mAP
Without DA	97.33	87.48	90.91	90.03	91.44
Few-shot Domain Adaptation Object Detection Methods					
AcroFOD [9] ECCV'22	99.51	**99.03**	98.76	87.57	96.21
AsyFOD [8] CVPR'23	94.73	94.73	93.47	94.73	94.41
Unupervised Domain Adaptation Methods					
ConfMix [24] WACV'21	53.90	65.80	66.40	59.30	61.40
SIGMA [19] CVPR'22	97.21	84.48	94.96	**95.28**	92.98
CMT [4] CVPR'23	90.89	81.32	87.86	74.64	83.68
SIGMA++ [20] TPAMI'23	90.17	87.66	99.08	94.69	92.90
Ours(FS-UDA)	**99.55**	95.29	**99.42**	94.31	**97.14**
Target Only	100	90.70	**100**	99.71	97.60

highlights the inadequacy of previous UDA object detection methods in natural medical scenarios. Notably, FS-UDA achieves superior detection performance across nine anatomical structures, ranging from small structures like DAO to larger ones like RIB, showcasing its adaptability for cross-hospital anatomical structure detection via topological knowledge integration and frequency distribution alignments.

Hospital 2→1 on 4C: Similarly, with hospital 2 as the source domain and hospital 1 as the target domain, FS-UDA demonstrates robust performance in adaptive detection, yielding an mAP of 85.37%. This surpasses all existing benchmark methods by a notable margin of 19.04% mAP higher than the previous state-of-the-art method SIGMA++ [20], indicating FS-UDA's versatility across different cross-domain scenarios.

Hospital 1→2 on 3VT: To further validate our method, cross-hospital evaluation on 3VT, crucial for congenital heart disease (CHD) diagnosis, demonstrates significant performance improvement. FS-UDA achieving a mAP of 62.92%, surpassing all other studies in the hospital 1→2 scenario on 3VT. This underscores the effectiveness of our approach in enhancing the detection of key structures by leveraging multiple topological knowledge alignments.

Hospital 2→1 on 3VT: Similarly, in the hospital 2→1 adaptive detection task on 3VT, FS-UDA continues to outperform UDA object detection comparison baselines, reaffirming its effectiveness.

Site G→R on CardiacUDA: Experiments on the CardiacUDA dataset reveal notable observations. Despite high detection performance among baseline methods, FS-UDA consistently outperforms all object detection baselines by 0.93% mAP over AcroFOD [9], indicating its efficacy in aligning single topological knowledge from diverse hospital settings. Additionally, the negligible domain gap in the CardiacUDA dataset results in little performance difference between various domain adaptation object detection methods.

Table 4. Ablation experiments on 4C and 3VT.

	Hospital 1→2 on 4CC			Hospital 1→2 on 3VT		
Method	*FDA*	*OSA*	mAP	*FDA*	*OSA*	mAP
Baseline	✗	✗	57.20	✗	✗	36.78
Ours	✓	✗	70.39	✗	✓	53.19
	✗	✓	66.57	✓	✗	51.31
	✓	✓	**79.72**	✓	✓	**62.92**
	Hospital 2→1 on 4CC			Hospital 2→1 on 3VT		
Baseline	✗	✗	64.32	✗	✗	41.61
Ours	✓	✗	71.15	✗	✓	52.93
	✗	✓	72.55	✓	✗	50.36
	✓	✓	**85.37**	✓	✓	**64.10**

4.4 Further Empirical Analysis

Ablation Studies: Ablation studies adding each component of our method, are listed in Table 4. Our method significantly outperforms the baseline. For Hospital 1→2 and Hospital 2→1 adaptive detection on 4C, our method improves mAP by 22.52% and 21.05% compared to the baseline, respectively. Likewise, on 3VT, our method improves by 26.14% and 20.49%, respectively, compared to the baseline. Additionally, the effectiveness of each matching component is evident. For Hospital 1→2 adaptive detection on 4C, FDA improves by 13.19%, OSA enhances by 9.37%, and with together, the detection mAP reaches 79.72%. Each matching module contributes to the enhancement of our method, as reported in Table 4, underscoring the significant advantages of our multi-matching approach. This approach furnishes effective internal and global topology knowledge in the target domain, rendering it suitable for structure detection tasks across various hospital scenarios.

Qualitative Result Comparison: Figure 3 illustrates qualitative results comparing our approach with *Source only* in target hospital adaptation scenarios. Our method avoids false positive error of one IVS in (h). Similarly, our method prevents two miss-detection of PTDA, and false positive error of one SVC and one DAO in (e) and (k). Meanwhile, LV is missing in (i), and our method accurately fills this defect.

Feature Visualization Comparison: Randomly sampling the same number of pixels for each fetal structure category in the target domain, we perform t-SNE visualization of ResNet-101-based features, comparing with other methods in Figure 4. Clear separation of similar categories, such as RA, RV, LV, IVS, SP, LA, CR, DAO, and RIB, can be observed in the 4C view. Similarly, in the 3VT view, our method can clearly distinguish DAO, IVS, PTDA, T, SVC, and AOA, beneficial for subsequent detection.

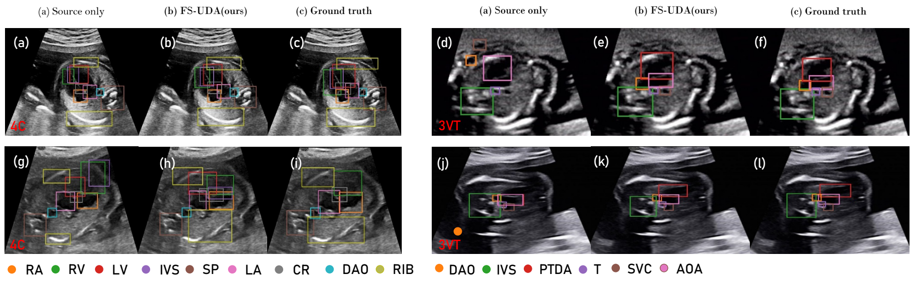
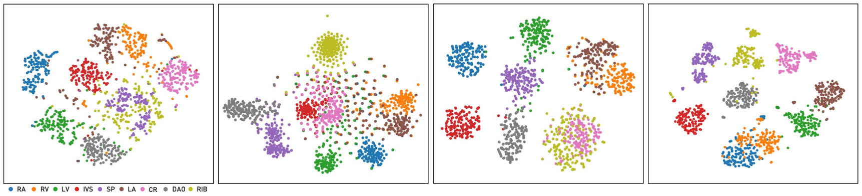


Fig. 3. Qualitative results comparison on 4C and 3VT of target domain



(a) Baseline (b) SIGMA (c) SIGMA++ (d) FS-UDA(ours)

Fig. 4. Visualization comparison of feature representation on 4C between the baseline models, two state of the art models and our FS-UDA is performed by t-SNE [23].

5 Conclusion

This study introduces a novel approach, termed FS-UDA, aimed at tackling the challenge of unsupervised domain adaptive fetal cardiac structure detection in medical scenarios. By aligning frequency distribution and organ-structure information between the source and target domains, FS-UDA effectively mitigates the domain gap inherent in medical imaging applications. Comprehensive experimentation conducted on both proprietary and publicly available datasets validates the efficacy of FS-UDA in unsupervised domain adaptive object detection tasks. Ablation experiments and visualizations provide a deeper understanding of the mechanisms underlying FS-UDA's performance, shedding light on its effectiveness in overcoming domain adaptation challenges in medical imaging.

References

1. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A.: Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine **4**(1), 65 (2021)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT. pp. 177–186. Springer (2010)
3. Cai, M., Zhang, H., Huang, H., Geng, Q., Li, Y., Huang, G.: Frequency domain image translation: More photo-realistic, better identity-preserving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13930–13940 (2021)

4. Cao, S., Joshi, D., Gui, L.Y., Wang, Y.X.: Contrastive mean teacher for domain adaptive object detectors. In: Proceedings of CVPR. pp. 23839–23848 (2023)
5. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of CVPR. pp. 8869–8878 (2020)
6. Dalca, A.V., Guttag, J., Sabuncu, M.R.: Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (Jun 2018). https://doi.org/10.1109/cvpr.2018.00968, http://dx.doi.org/10.1109/CVPR.2018.00968
7. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of CVPR. pp. 4091–4101 (2021)
8. Gao, Y., Lin, K.Y., Yan, J., Wang, Y., Zheng, W.S.: Asyfod: An asymmetric adaptation paradigm for few-shot domain adaptive object detection. In: Proceedings of CVPR. pp. 3261–3271 (2023)
9. Gao, Y., Yang, L., Huang, Y., Xie, S., Li, S., Zheng, W.S.: Acrofod: An adaptive method for cross-domain few-shot object detection. In: Proceedings of ECCV. pp. 673–690. Springer (2022)
10. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR. pp. 770–778 (2016)
12. Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: Proceedings of WACV. pp. 749–757 (2020)
13. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6891–6902 (2021)
14. Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. Advances in Neural Information Processing Systems **34**, 3635–3649 (2021)
15. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021)
16. Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of CVPR. pp. 6092–6101 (2019)
17. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of CVPR. pp. 12456–12465 (2019)
18. Li, M., Zhang, H., Li, J., Zhao, Z., Zhang, W., Zhang, S., Pu, S., Zhuang, Y., Wu, F.: Unsupervised domain adaptation for video object grounding with cascaded debiasing learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3807–3816 (2023)
19. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: Proceedings of CVPR. pp. 5291–5300 (2022)
20. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. IEEE Trans. Pattern Anal. Mach. Intell. (2023)

21. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of CVPR. pp. 7581–7590 (2022)
22. Lu, Y., Li, K., Pu, B., Tan, Y., Zhu, N.: A yolox-based deep instance segmentation neural network for cardiac anatomical structures in fetal ultrasound images. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022)
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
24. Mattolin, G., Zanella, L., Ricci, E., Wang, Y.: Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 423–433 (2023)
25. Ni, Z., Wu, J., Wang, Z., Yang, W., Wang, H., Ma, L.: Misalignment-robust frequency distribution loss for image transformation. arXiv preprint arXiv:2402.18192 (2024)
26. Oza, P., Sindagi, V.A., Sharmini, V.V., Patel, V.M.: Unsupervised domain adaptation of object detectors: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
27. Paszke, A., Gross, S., Massa, et al.: Pytorch: An imperative style, high-performance deep learning library. Proceedings of NeurIPS **32** (2019)
28. Piao, Z., Tang, L., Zhao, B.: Unsupervised domain-adaptive object detection via localization regression alignment. IEEE Trans. Neural Netw. Learn. Syst. (2023)
29. Pu, B., Lv, X., Yang, J., Guannan, H., Dong, X., Lin, Y., Shengli, L., Ying, T., Fei, L., Chen, M., et al.: Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In: Forty-first International Conference on Machine Learning
30. Pu, B., Wang, L., Yang, J., He, G., Dong, X., Li, S., Tan, Y., Chen, M., Jin, Z., Li, K., et al.: M3-uda: A new benchmark for unsupervised domain adaptive fetal cardiac structure detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11630 (2024)
31. Rodriguez, A.L., Mikolajczyk, K.: Domain adaptation for object detection via style consistency. arXiv preprint arXiv:1911.10033 (2019)
32. Sindagi, V.A., Oza, P., Yasarla, R., Patel, V.M.: Prior-based domain adaptive object detection for hazy and rainy conditions. In: Proceedings of ECCV. pp. 763–780. Springer (2020)
33. Stan, S., Rostami, M.: Domain adaptation for the segmentation of confidential medical images. arXiv preprint arXiv:2101.00522 (2021)
34. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of CVPR. pp. 9627–9636 (2019)
35. Vs, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In: Proceedings of CVPR. pp. 4516–4526 (2021)
36. Wang, Y., Zhang, R., Zhang, S., Li, M., Xia, Y., Zhang, X., Liu, S.: Domain-specific suppression for adaptive object detection. In: Proceedings of CVPR. pp. 9603–9612 (2021)
37. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: Proceedings of CVPR. pp. 11878–11887 (2023)
38. Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., Chen, X.: Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. arXiv preprint arXiv:1911.07158 (2019)

39. Zhang, P., Li, J., Wang, Y., Pan, J.: Domain adaptation for medical image segmentation: a meta-learning method. Journal of Imaging **7**(2), 31 (2021)
40. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3770 (2019)
41. Zhao, G., Li, G., Xu, R., Lin, L.: Collaborative training between region proposal localization and classification for domain adaptive object detection. In: Proceedings of ECCV. pp. 86–102. Springer (2020)
42. Zhao, L., Tan, G., Wu, Q., Pu, B., Ren, H., Li, S., Li, K.: Farn: Fetal anatomy reasoning network for detection with global context semantic and local topology relationship. IEEE Journal of Biomedical and Health Informatics (2024)
43. Zhao, L., Wang, L.: Task-specific inconsistency alignment for domain adaptive object detection. In: Proceedings of CVPR. pp. 14217–14226 (2022)
44. Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13766–13775 (2020)
45. Zhou, M., Huang, J., Yan, K., Yu, H., Fu, X., Liu, A., Wei, X., Zhao, F.: Spatial-frequency domain information integration for pan-sharpening. In: European Conference on Computer Vision. pp. 274–291. Springer (2022)