M^3-UDA: A New Benchmark for Unsupervised Domain Adaptive Fetal Cardiac Structure Detection

Bin Pu[1*] Liwen Wang[2*] Jiewen Yang[1*] Guannan He[3] Xingbo Dong[2] Shengli Li[4]
Ying Tan[4] Ming Chen[5] Zhe Jin[2] Kenli Li[6†] Xiaomeng Li[1†]
[1] The Hong Kong University of Science and Technology.
[2] Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, Anhui University, China. [3] Sichuan Provincial Maternity and Child Health Care Hospital.
[4] Shenzhen Maternity and Child Healthcare Hospital. [5] Harbin Red Cross Central Hospital.[6] Hunan University.
{eebinpu, eexmli}@ust.hk,jyangcu@connect.ust.hk,{jinzhe,xingbo.dong}@ahu.edu.cn,lkl@hnu.edu.cn

Abstract

*The anatomical structure detection of fetal cardiac views is crucial for diagnosing fetal congenital heart disease. In practice, there is a large domain gap between different hospitals' data, such as the variable data quality due to differences in acquisition equipment. In addition, accurate annotation information provided by obstetrician experts is always very costly or even unavailable. This study explores the unsupervised domain adaptive fetal cardiac structure detection issue. Existing unsupervised domain adaptive object detection (UDAOD) approaches mainly focus on detecting objects in natural scenes, such as Foggy Cityscapes, where the structural relationships of natural scenes are uncertain. Unlike all previous UDAOD scenarios, we first collected a **F**etal **C**ardiac **S**tructure dataset from two hospital centers, called **FCS**, and proposed a multi-matching UDA approach (M^3**-UDA**), including **H**istogram **M**atching (HM), **S**ub-structure **M**atching (SM), and **G**lobal-structure **M**atching (GM), to better transfer the topological knowledge of anatomical structure for UDA detection in medical scenarios. HM mitigates the domain gap between the source and target caused by pixel transformation. SM fuses the different angle information of the sub-structure to obtain the local topological knowledge for bridging the domain gap of the internal sub-structure. GM is designed to align the global topological knowledge of the whole organ from the source and target domain. Extensive experiments on our collected FCS and CardiacUDA, and experimental results show that M^3-UDA outperforms existing UDAOD studies significantly. Datasets and source code are available at https://github.com/xmed-lab/M3-UDA*

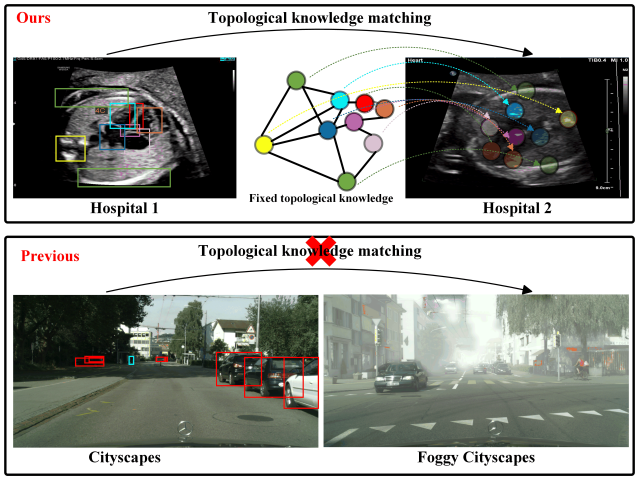


Figure 1. Differences between our FCS detection and the previous typical UDAOD task.

1. Introduction

In practical clinical prenatal screening, the detection of cardiac anatomical structures in fetal ultrasound images is essential for standard view localization [37], quality control [30], and the diagnosis of congenital heart disease (CHD) [2]. For example, ventricular septal defect, as a typical CHD, is due to the absence of ventricular septum structure [6]. In real-world applications, there are always some domain gaps in fetal ultrasound cardiac datasets from different hospital centers, e.g., data collected by different medical devices and scanned by obstetricians in various centers, resulting in a wide variation in the quality. This leads to performance degradation in the detection of anatomical structures testing in different hospitals. For example, several related studies [1, 3, 32] have also explored the structure detection issue but have not addressed the domain gap. Mean-

*Equal Contribution.
†Corresponding authors.

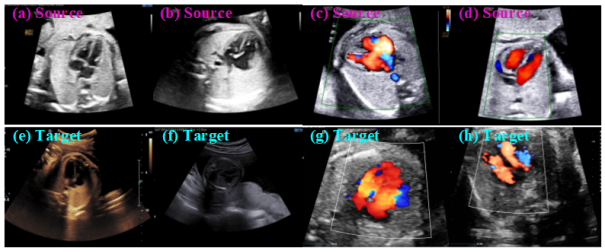


Figure 2. Visual domain shift in medical scenario.

Table 1. The comparison of FCS, CardiacUDA [47], CAMUS [24], and EchoNet [34].

Dataset	**Our FCS dataset**	CardiacUDA	CAMUS	EchoNet
Annotated Images	2,879	4,960	1,000	20,060
Multiple Centers	✓	✓	×	×
Cardiac Views	2	4	1	1
Resolution	640-1920p	720p	480p	120p
Annotated Regions	LV, RV, LA, RA, DAO, VS, SP, CR, RIB, PTDA, SVC, T, AOA	LV, RV, LA, RA	LV, LA	LV

while, accurate annotation information provided by obstetrician experts is costly or even unavailable in practice, especially in low-resource settings with a severe shortage of well-trained obstetricians.

Unsupervised domain adaptation (UDA) techniques aim to maximize the performance of the target domain/hospitals while minimizing expert supervision through invariant feature learning [44], self-training [22, 49], image translation [9, 20], domain randomization [23, 38], etc. Most of the previous UDAOD methods focus on detecting objects, e.g., Cityscapes [11] and Foggy Cityscapes [39], where the relationship between these objects, e.g., boat and bus, is uncertain. In fetal anatomy, there's topological understanding aiding UDA task, while in natural scenes, the relationships between objects are often chaotic and unclear, as shown in Figure 1. For instance, the fetal ventricular septum (VS) is always midway between the left ventricle (LV) and right ventricle (RV). This means that most existing UDAOD methods may not be available in our new medical scenario. To address the challenges for the new UDAOD task, we propose a novel multi-matching method called M^3-UDA by histogram ***M***atching, sub-structure ***M***atching, and global-structure ***M***atching for better aligning and transferring topological knowledge of medical scenarios in the source and target domain. First, there are domain shifts caused by variations in device specifications and calibration, lighting conditions, and screening conditions, an example of which is shown in Figure 2 (a) and (e). We can notice that the gap between different hospital domains is the existence of different brightness and gain-like conditions of ultrasound images. Moreover, color Doppler images are acquired in the screening of cardiac views, which mainly observe the prevalence of cardiac blood flow, as shown in Figure 2 (c) and (g). These situations in the source and target domains are mainly pixel-wise transformations. To this end, we propose a simple, direct, and efficient histogram matching to transform the fetal ultrasound images in the source domain to be more like the target domain by non-linear pixel-wise transformations, which is more suitable for different hospital scenarios.

Second, the relative positions of the sub-anatomical structures are generally fixed, which is reflected in the angles of the sub-structures, e.g., the angular ranges of the four ventricular atria of the cardiac are fixed. To align this detailed substructure topology information, we embed multiple substructure angle information into the matrix representation to match the abundant angle information of the source and target hospital.

Third, fetal cardiac views of anatomical structures that have well-defined global topological information. For example, the standard four-chamber cardiac view has nine anatomical structures and the relative positions are fixed, e.g., the ribs are always on either side of the four-chamber. In other words, the global semantic topological information of the anatomical structures for different domains can be shared. We can utilize this disciplined global semantic information to perform alignment between different domains. To this end, unlike the prototype approach to model the conditional distribution of categories, a global graph based on the relationships of all key structures is proposed to represent the category distribution, which is more conducive to aligning and extracting global topological knowledge in the different domains. To conclude, our contributions include:

- We collected the first ***F***etal ***C***ardiac ***S***tructure dataset, called ***FCS***, for UDAOD in the medical application, containing two views (four-chamber cardiac view (4CC) and three vessels and trachea view (3VT)) of the fetal cardiac with 13 anatomical structures/objects (Tables 1 and 2) from two hospital centers, totaling 2879 fetal images. These fetal cardiac views will be released, facilitating the exploration of UDA tasks, object-based detection methods, and diagnosis.
- We proposed M^3-UDA method to address the challenges for the new release datasets, which served as a new benchmark for better detecting fetal cardiac structures in different hospital centers. For special medical scenarios, we introduce histogram ***M***atching to deal with the contrast, tone, and gain matching issues of ultrasound images. Further, we design sub-structure ***M***atching and global-structure ***M***atching for extracting the strong fixed topological knowledge of the source and target domain.
- Extensive experiments show that the proposed new benchmark outperforms existing state-of-the-art methods. For example, the performance of adapting hospital 2→1 on 4CC is improved by 7.08% mAP compared to the SOTA method, i.e., SIGMA++ [27].

Table 2. Key structure and its abbreviation (Abb) in two fetal cardiac views in our FCS dataset.

4CC View		3VT View	
Structure	Abb	Structure	Abb
Left ventricle	LV	Superior vena cava	SVC
Left atrium	LA	Arch of Aorta	AOA
Left ventricle	LV	Trachea	T
Descending aorta	DAO	Spine	SP
Right atrium	RA	Pulmonary trunk & ductus arteriosus	PTDA
Ventricular septum	VS	Ascending Aorta	DAO
Spine	SP	/	/
Rib	RIB	/	/
Cross	CR	/	/

2. Related Work

2.1. Unsupervised Domain Adaptive Object Detection

To solve the domain offset issue in object detection tasks, many UDAOD methods have been proposed, which can be roughly categorized into invariant feature learning [40, 44, 45, 50], pseudo-label based self-training [22, 49], image translation [9, 10, 20, 21, 48], domain randomization [23, 38] and mean-teacher training [8, 13, 14, 28]. Domain invariant feature learning methods [4, 16] can generate domain invariant features by adversarial training of object detector models with the assistance of a domain discriminator. Pseudo-label-based self-training approaches [49] utilize high-confidence predictions from source-trained detector models to train the target. These popular methods are always combined with semi-supervised DA studies [15]. Then, some researchers proposed to use the unpaired image-to-image translation-based method, like Cycle-GAN [51], to map the target domain to the source domain and vice versa. Another interesting approach is the domain randomization strategy [23], which creates multiple stylized versions of the source domain data to train the detection model, thus allowing the model to minimize source-style bias and generalize better to the target domain. In addition, mean-teacher [13, 41] is an effective approach to improve model generalization by progressively training the detector model within a student-teacher framework using unlabeled data. However, all of these methods ignore the graph topology knowledge, which causes degradation performance, especially in UDAOD of a medical scene.

2.2. Graph Knowledge Reasoning for UDA

Recently, utilizing graph knowledge reasoning to enhance the performance of UDAOD has become a hot topic. Earlier studies [7, 46] utilized inter-object and intra-object relationships present in the detection dataset, and these object relations are modeled through knowledge graphs that can assist detectors in the target domain to perform the same object relations through training. SCAN [25] uses category-specific semantics of different images in the source domain as input by graph-based aggregation and learns the unbiased semantic paradigm for better adaptability. The key idea of SIGMA [26] and SIGMA++ [27] is to solve the problem of the graph matching adaptation between dense feature points. CIGAR [31] mainly addresses better fusion and interaction of semantic-based and vision-based graph information. However, previously designed graph-based methods are suitable for natural scenarios and do not necessarily extend well to UDA detection of medical anatomical structures. In addition, most of the previous graph knowledge is disorganized and uncertain, and the matching is relatively coarse. Unlike previous methods, we propose a multi-matching approach that captures most of the fixed topological knowledge in medical scenarios from both sub-structure and global-structure.

3. Method

Figure 3 shows the overview of our method. Beginning with an annotated source image and an unlabeled target image, our approach involves several key modules. Initially, a histogram matching (HM) module is employed to minimize the domain gap within the spatial domain, specifically tailored for ultrasound images (Section 3.1). Subsequently, a shared feature extractor is utilized to derive features denoted as F^s and F^t, followed by the integration of the FCOS head. Then, we incorporate both a sub-structure matching (SM) module and a global-structure matching (GM) module. These modules play pivotal roles in augmenting the training process of the core object detection network and function as regularization terms, elucidated in Sections 3.2 and 3.3, respectively.

3.1. Histogram Matching

Ultrasound images collected from different health centers reveal differences in brightness and contrast due to diverse medical equipment and practices. To address this domain shift, we leverage histogram matching for image-level feature alignment. Specifically, given images $X^{s/t} \in \mathbb{R}^{H\times W\times 3}$ from source and target domains, respectively, where H and W is the height and width of the image. We first normalize all pixels of images to ensure their intensity values are within the same range $k = [0, 255]$. Then, we calculated the numerical density distribution $P^{s/t} = \{p_i^{s/t}\}_{i=0}^{k}$ for X^s and X^t in both domains. Finally, we conduct the histogram matching via optimal mapping $\sigma(\cdot)$ with the following:

$$X^{t'} = \sigma(X^t),\ \sigma := \arg\min \sum_{i=1}^{k} \mathbf{d}(M(P^s), P^t), \quad (1)$$

where $\mathbf{d}(\cdot)$ is a distance metric between two histograms for each k value, and we measure its corresponding value in-

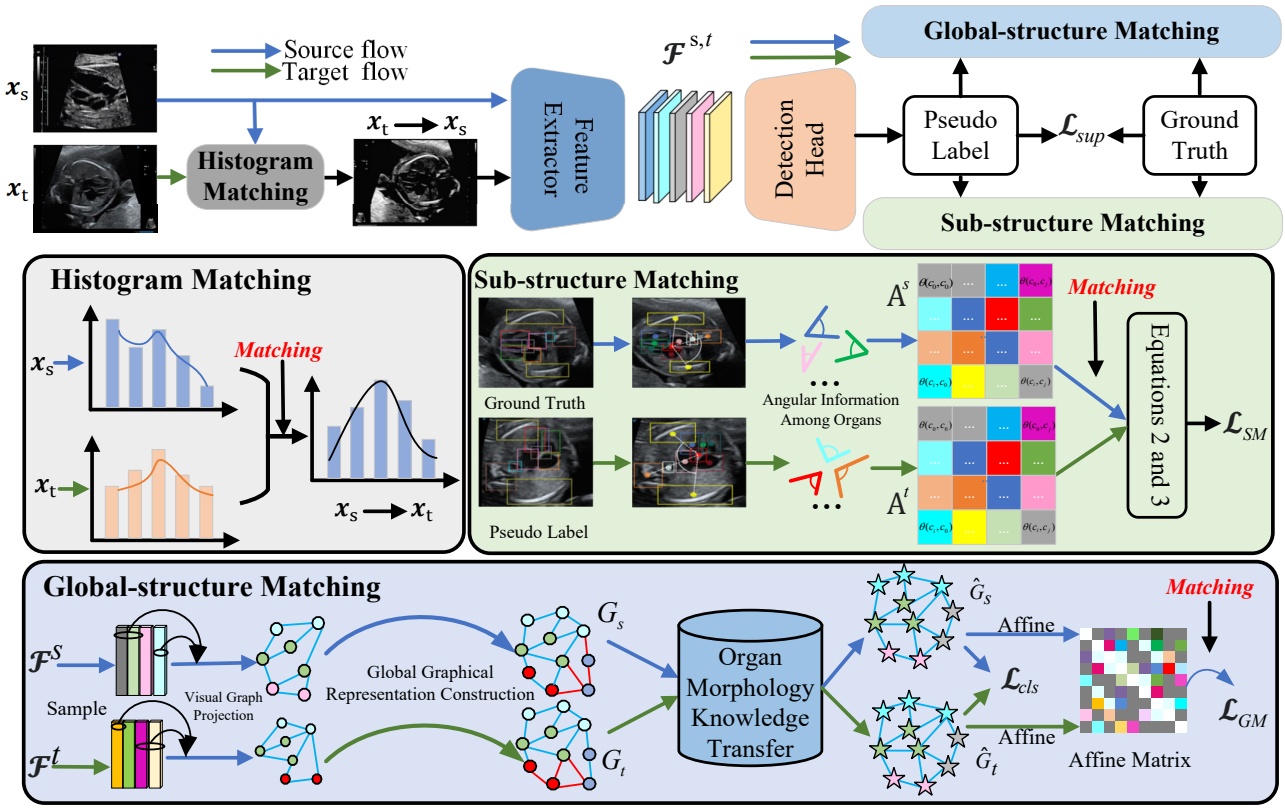


Figure 3. **Overview of M^3-UDA.** First, the source/target images are randomly selected to be fed into the **Histogram Matching Module** to obtain the matched source image for *the first matching*. Second, the target image and the matched image are sent to the shared backbone to generate the feature maps and produce the detection results from FCOS [42] detection head. Third, in **Sub-structure Matching Module**, the pseudo-label of the target image and the ground truth of the matched source image are utilized to generate an adjacency matrix incorporating the detailed angle information of sub-structures. Fourth, we employed the Wasserstein distance method to compute the loss of the sub-structure matching module for *the second matching*. Then, in **Global-structure Matching Module**, feature nodes are produced from the feature maps, and a methodology involving node sampling, visual graph projection, global graphical representation construction, and organ morphology Knowledge transfer is used to compute the loss on the global structure for *the third matching*. Finally, the trained detection head predicts the final result.

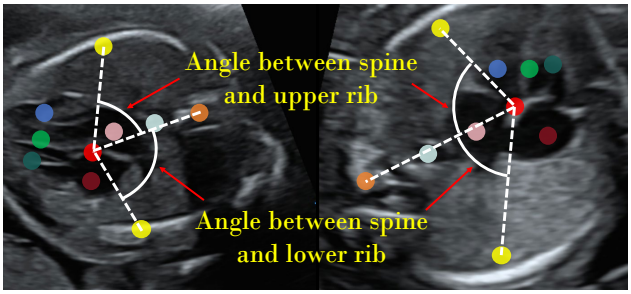


Figure 4. The orientation-based sub-structure remains consistent for the same view.

tensity and use the $\sigma(\cdot)$ to represent the histogram matching across source and target domains.

3.2. Sub-structure Matching

In real scenarios, obstetricians usually scan organs from fixed orientation to acquire fetal ultrasound images clinically. Inspired by this, we consider this principle able to be presented with a consistent representation of sub-structures within the same ultrasound plane across different medical centers. As illustrated in Figure 4, the angles between the spine and two ribs of the interventricular septum intersection remain consistent across domains. To utilize this knowledge for narrowing the domain gap across source and target domains, we present the **S**ub-structure **M**atching (SM) module that innovatively introduces the angular information for constructing the domain relationship. Essentially, the SM module consists of two components: Sub-structure Relation Construction and Structure Matching. We detail them below.

Sub-structure Relation Construction. During sub-structure construction in the training phase, the main goal of this module is to compute the angular 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 $n^{s/t}$ indicate the total number of organs. We first compute the centroid location $c^s \in \mathbb{R}^{n^s \times 2}$ 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. In order to ensure that the number of detected organs and types of all organs remain the same for the target domain, we will complete the missing class by using the annotation of the corresponding class from the source domain and remove the extra class that has not been detected in the source domain. After we complete the missing class and ensure the number of detected targets remains the same, the centroid location of each target thus can be reformulated as $c^{s/t} \in \mathbb{R}^{n \times 2}$, where n is the total classes in the source domain. Subsequently, angular information is obtained by calculating the angles among the centroid location of each organ in $c^{s/t}$. Adjacency matrices can represent the angular relationship of organs within a domain as the following definition:

$$A^{s/t} = \begin{pmatrix} \theta(c_0, c_0) & \cdots & \theta(c_0, c_j) \\ \vdots & \vdots & \vdots \\ \theta(c_i, c_0) & \cdots & \theta(c_i, c_j) \end{pmatrix}^{s/t}, \quad (2)$$

where $\theta(\cdot, \cdot)$ calculates the angle between organs, referred by the centroid location provided from Y^s and $\hat{Y}^t$, i and j are the i-th and j-th element of the c^s and c^t, where $i, j \leq n$.

Sub-structure Matching. In the matching stage, we maximize the similarity of the adjacency matrices. Each element in matrices $A^{s/t}$ from source and target domains is a one-to-one correspondence. Hence, we use the L2 norm to compute the distance of matrices A^s and A^t, and the optimization objective function is formulated as:

$$\mathcal{L}_{\text{SM}} = \sum_{i=0}^{n} \sum_{i=0}^{n} ||A_{i,j}^s - A_{i,j}^t||_2. \quad (3)$$

3.3. Global-structure Matching

In the previous subsection, we solely focus on the orientation information of the organ, which is a unique prior knowledge of anatomical structures in ultrasound images of fetuses. In this section, we further observe the consistent characteristics of each organ in images. For example, as shown in Figure 4, the morphology information of each specific organ within the same scanned plane, such as shape and size, should remain consistent across different hospitals. This is due to the same organs of different fetuses having consistent biological anatomy information. Inspired by the above knowledge, we construct a graph to represent the morphology information of an organ. Since graphs are a more flexible approach to modelling irregular shapes compared with other methods. Also, we propose the **G**lobal-structure **M**atching (GM) module to ensure that the graphical representation of all organs across different domains remains the same.

Organ Feature Extraction and Visual Node Sampling. For the input X^s and $X^{t'}$ from the source and target domains, the share-parameter equipped with Feature Pyramid Networks [29] encodes their feature as $F^{s/t}$, where $F^{s/t}$ consists of feature maps from all layers of the encoder. In the first step, we uniformly sample total M visual nodes from $F^{s/t}$ as $\mathcal{V}^{s/t} \in \mathbb{R}^{M \times d}$, according to the corresponding position of each organ from the Y^s and $\hat{Y}^t$, where d is the channel number of the feature nodes. Similar to the sub-structure matching module, we are not able to ensure that the total number of visual nodes from the source and target domains remain the same. Hence, we apply the node completion approach that refers to SIGMA [26], which builds memory banks for each class to acquire and store the average representation for both domains.

Global Graphical Representation Construction. To facilitate semantic graphs and conduct the graph completion, we first compute the edge $\mathcal{E}^{s/t} \in \mathbb{R}^{M \times M}$ that represents the connectivity between each node of organs in $\mathcal{V}^{s/t}$ by using spectral clustering. Hence, the visual graph can be written as $\{\mathcal{V}, \mathcal{E}\}^{s/t}$. In order to acquire the complete graphical representation, we introduce the graph convolutional neural network as GNN$(\cdot)$ to transform the visual graph as the following formulation:

$$\mathcal{G}^{s/t} = \text{GNN}(\{\mathcal{V}, \mathcal{E}\}^{s/t}) \in \mathbb{R}^{M \times d}. \quad (4)$$

Organ Morphology Knowledge Transfer. Since the morphology of the same organs across domains should remain consistent, we can build the connection across $\mathcal{G}^s$ and $\mathcal{G}^t$ to exchange the information and acquire a more robust graphical representation. Motivated by the above discussion, we introduce the attention mechanism to conduct the inter- and intra-interaction of graphs. For the inter-interaction, the graphs from each domain will compute the attentive graph feature via self-attention. For the intra-interaction between the graphs across domains, we formulate the cross-attention to build the connections between graphs. The formulation can be written as:

$$\begin{aligned} \dot{\mathcal{G}}^{s/t} &= \lambda \cdot \left(W_q \mathcal{G}^{s/t} (W_k \mathcal{G})^{\mathsf{T}} \right) W_v \mathcal{G} \in \mathbb{R}^{M \times d}, \\ \mathcal{G} &= \mathcal{G}^s \oplus \mathcal{G}^t \in \mathbb{R}^{2M \times d}, \end{aligned} \quad (5)$$

where W_q, W_k and $W_v \in \mathbb{R}^{d \times d}$ is the learnable weight of three independent projection layers that project the graph to different latent space as query, key and value vector, λ is the scaling factor set as $1/\sqrt{d}$, and $\oplus$ denotes the concatenation operation.

Graph Matching and Node Classification. After knowledge transfer, we also maximum the similarly between $\dot{\mathcal{G}}^s$ and $\dot{\mathcal{G}}^t$. Due to the graphical representation being disordered, the conventional element-wise distance computation is not able to measure their discrepancy. To tackle this problem, we introduce the Wasserstein distance that measures

the discrepancy of feature distribution between $\dot{\mathcal{G}}^s$ and $\dot{\mathcal{G}}^t$. With the Sinkhorn [12] iteration to acquire the approximate numerical solution. The formulation of our measurement can be written as follows:

$$\mathcal{L}_{\text{GM}} = \sum_{i=0}^{d} \sum_{j=0}^{M} \left\| \dot{\mathcal{G}}^s_{i,j}, \dot{\mathcal{G}}^t_{i,\pi(j)} \right\|_2^2, \tag{6}$$

where $\dot{\mathcal{G}}_{i,j}$ denote the i-th channel in j-th node of $\dot{\mathcal{G}}$, and $\pi(\cdot)$ denotes the mapping across overall nodes. In this optimization, we apply the Sinkhorn that uses the affine matrices as the mapping function for optimizing Eq. (6).

Since we sample node according to the Y^s and $\hat{Y}^t$, each node thus has the corresponding class as the annotation. We also introduce the cross-entropy (denotes as $\text{CE}(\cdot,\cdot)$) for the classification loss of each node across domains.

$$\mathcal{L}_{\text{cls}} = \text{CE}(\dot{\mathcal{G}}^s, Y^s) + \text{CE}(\dot{\mathcal{G}}^t, \hat{Y}^t) \tag{7}$$

Overall Loss of Our Framework. The overall loss (Eq. (8)) can be separated into 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 is the combination of $\mathcal{L}_{\text{SM}}$ from SM module, $\mathcal{L}_{\text{GM}}$ and $\mathcal{L}_{\text{cls}}$ from GM module.

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{SM}} + \lambda_3 \mathcal{L}_{\text{GM}} + \lambda_4 \mathcal{L}_{\text{cls}}, \tag{8}$$

where $\lambda_{1,2,3,4}$ is the weight of different loss terms, which are set as 1.0, 0.5, 0.5 and 1.0 in our experiment setting.

4. Experiments

4.1. Datasets and Evaluation

FCS Dataset. We collected two cardiac views, including 4CC and 3VT, from two hospital centers, and these views are used to regularly screen for CHDs. All dataset collection and experiments were approved by the local hospital ethics committee. Each fetal view contains anatomic structure information, as shown in Table 2. These datasets come from different equipment, such as Samsung, Sonoscape, and Philips. The gestational week of the fetus also varies for all datasets, ranging from 20-34 weeks. All datasets were annotated by ultrasonographers with more than seven years of clinical experience, and each fetal image has view category labels and category and location labeling information for the anatomical structures. A standard 4CC image should contain 9 categories of anatomical structures, but not every view obtained by sonographers of our dataset is standardized, so some views have fewer than 9 structures.

CardiacUDA Dataset [47]. CardiacUDA was used as a video-based cardiac structure segmentation task from two different hospitals (site G and site R), as shown in Table 1. We convert the segmentation mask of LA, LV, RA, and RV to box-level annotation and use it to conduct the UDAOD task for comparison. Since there are so many overlapping frames in the video, we sample one frame per video.

We conducted experiments on adaptation tasks mainly on hospital1 and hospital2. Taking as an example, the hospital1 (source) adapts to the hospital2 (target), denoted as hospital 1→2. During UDA procedure, we use training splits of the source and target domain and use test splits of the target domain for performance evaluation. For comparison, we used the mean average precision (mAP) metric with a threshold of 0.5 for Intersection over Union. The training, testing, and validation datasets for each view are 7:2:1, respectively.

4.2. Implementation Details

We use ResNet-101 [19] as the feature extractor and FCOS [42] as the detector, which is implemented in PyTorch [35]. We trained the M^3-UDA using the stochastic gradient descent (SGD) optimizer [5] with an initial learning rate of 0.01, a fixed epoch of 300, a batch size of 4, and a weight decay of 5×10^{-4}. Due to the variation of acquisition devices in different hospital centers, the image size is uncertain; we uniformly resize it to 600×1000. Because the prediction pseudo label of the model is not trustworthy at the early stage of training, which leads to fewer correct matching at the early stage, the SM module starts to work only after 30 epochs have been reached.

4.3. Comparison with State-of-the-arts

In this experiment, we evaluate M^3-UDA on the collected new benchmark and publicly commonly used CardiacUDA, where the object detector needs to adapt from a hospital center to another hospital.

Hospital 1→2 on 4CC. As shown in Table 3, our M^3-UDA method significantly outperforms all state-of-the-art studies, including semi-supervised UDA object detection methods, by an absolute margin of 8.26% mAP over CMT [8]. This indicates that the previous methods for the UDA object detection task in natural scenarios cannot be well suited to our medical scenarios. Specifically, we have achieved the best detection performance in the detection of eight anatomical structures, e.g., small structure DAO and larger structure RIB. These compelling results clearly demonstrate that M^3-UDA can be adapted to the task of cross-hospital anatomical structure detection through the alignment of multiple topological knowledge.

Hospital 2→1 on 4CC. Similarly, if hospital2 is the source domain and hospital1 is the target domain, we also conduct adaptive detection on hospital 2→1. As listed in Table 3, we can observe that our method obtained a 74.55% mAP, outperforming all existing benchmark methods by a notable margin of 7.08% mAP higher than the previous state-of-the-art SIGMA++ [27]. This demonstrates the robustness of M^3-UDA in different cross-domain scenarios.

Table 3. Quantitative adaptation results on 4CC.

	Hospital 1→2										Hospital 2→1									
Method	RA	RV	LV	VS	SP	LA	CR	DAO	RIB	mAP	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	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 [17] ECCV'22	25.52	21.04	24.64	25.86	26.32	24.64	26.75	25.74	27.24	25.30	37.31	39.86	38.84	40.47	35.13	39.54	39.88	37.64	27.13	37.02
AsyFOD [18] CVPR'23	48.31	48.66	48.66	49.93	48.36	50.00	49.54	47.51	47.43	48.71	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 [33] WACV'23	62.40	65.30	65.20	61.80	59.70	63.20	69.50	67.80	62.70	64.20	58.90	64.00	63.40	61.50	63.40	55.40	64.90	60.00	46.30	59.80
SIGMA [26] CVPR'22	70.64	56.57	64.16	64.58	66.91	61.31	74.2	68.64	69.97	66.33	70.64	56.57	64.16	64.58	66.91	61.31	74.2	68.64	**69.97**	66.33
LRA [36] TNNLS'23	63.23	53.25	58.24	59.56	64.30	**84.32**	66.97	55.98	59.18	62.78	75.43	37.08	49.24	51.84	54.88	48.92	52.38	58.70	58.58	54.11
CMT [8] CVPR'23	79.18	64.87	66.31	64.34	74.84	66.23	71.61	60.93	68.66	68.55	**81.53**	63.30	66.35	68.98	77.30	**76.31**	76.70	67.67	58.42	70.40
SIGMA++ [27] TPAMI'23	56.10	47.01	52.72	51.38	63.65	52.11	60.28	62.99	67.30	57.06	70.29	54.26	55.10	63.12	70.15	62.36	75.23	74.23	55.47	67.47
Ours	**79.94**	**69.77**	**72.84**	**71.73**	**81.02**	77.99	**81.70**	**77.99**	**78.28**	**76.81**	81.35	62.41	68.94	**74.06**	**81.78**	73.87	**82.57**	**83.65**	62.30	**74.55**
Upper Bound	82.30	74.37	78.29	79.25	88.26	83.02	86.82	87.22	85.16	82.74	86.61	82.75	83.48	85.93	90.16	82.92	89.09	89.61	72.51	84.78

Table 4. 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 [17] 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 [18] CVPR'23	49.11	51.32	60.29	44.54	53.11	61.63	53.33	57.43	59.18	**61.74**	**56.92**	**50.65**	**62.71**	**58.10**
Unupervised Domain Adaptation Object Detection Methods														
ConfMix [33] WACV'23	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 [26] 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 [36] 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 [8] 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++ [27] 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	**59.48**	59.65	70.05	51.92	52.35	**68.88**	**60.39**	55.12	**61.97**	46.36	43.63	36.80	52.25	49.36
Upper Bound	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

Hospital 1→2 on 3VT. To further demonstrate the performance of our method, we performed a cross-hospital evaluation on 3VT, another critical view of the fetal cardiac for CHD diagnosis. As shown in Table 4, M^3-UDA achieves 60.39% mAP, which greatly exceeds all other studies with hospital 1→2 on 3VT. This further validates the efficacy of our approach to improve the performance of detection key structures by considering multiple topological knowledge of the matching source and target domain.

Hospital 2→1 on 3VT. Again, as reported in Table 4, for hospital 1→2 adaptive detection task on 3VT, we still achieved the best results over the UDA object detection comparison baseline. This reaffirms the effectiveness of our method. Meanwhile, we can observe that AsyFOD, as a few-shot DAOD method, requires fewer labeled data and outperforms our method by 8.74%. However, getting some experienced physicians at the target hospital to annotate the data is sometimes not available.

site G→R on CardiacUDA. As shown in Table 5, we have also conducted experiments in the public CardiacUDA dataset for adaptive detection in different hospitals. First, it is worth noting that in the adaptive detection on CardiacUDA dataset, the detection performance of most of the baseline methods is very high. For example, SIGMA reaches 92.98% mAP. Second, *Upper Bound* (97.60% mAP) as well as *Without DA* (91.44% mAP) are also close, which suggests that the domain gap is small. The CardiacUDA dataset is labeled only with LA, RA, LV, and RV, which are four anatomical structures with relatively simple relationships and non-overlapping. This indicates a single topological knowledge of the CardiacUDA dataset. In contrast, our dataset contains 9 key structures, many of which have complex positional relationships and are affected by multiple gestational weeks and differences in acquisition devices, resulting in a larger gap and a greater challenge. Still, our method outperforms all UDA object detection baselines by 1.97% mAP over SIGMA [26]. This shows that our method is also effective in alignment of single topological knowledge from different hospitals. Since the domain gap is very small, the few-shot DAOD approach to obtain the labeled data for a portion of the target domain may exhibit good performance, e.g., AcroFOD [17] is 1.26% mAP higher than our method.

Table 5. Quantitative adaptation results on 4CC 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 [17] ECCV'22	**99.51**	**99.03**	98.76	87.57	**96.21**
AsyFOD [18] CVPR'23	94.73	94.73	93.47	94.73	94.41
Unupervised Domain Adaptation Object Detection Methods					
ConfMix [33] WACV'23	53.90	65.80	66.40	59.30	61.40
SIGMA [26] CVPR'22	97.21	84.48	94.96	**95.28**	92.98
CMT [8] CVPR'23	90.89	81.32	87.86	74.64	83.68
SIGMA++ [27] TPAMI'23	90.17	87.66	99.08	94.69	92.90
Ours	99.42	90.39	100	90.00	94.95
Upper Bound	100	90.70	**100**	99.71	97.60

4.4. Further Empirical Analysis

Ablation Studies. We performed ablation studies by adding each component of our method, and the ablation results are listed in Table 6. First, our method is a significant improvement over the baseline method. For Hospital 1→2 and Hospital 2→1 adaptive detection on 4CC, our method improves 19.61% and 10.23% mAP when compared to baseline, respectively. Similarly, on 3VT, our method improved by 23.61% and 7.75%, respectively, compared to the baseline.

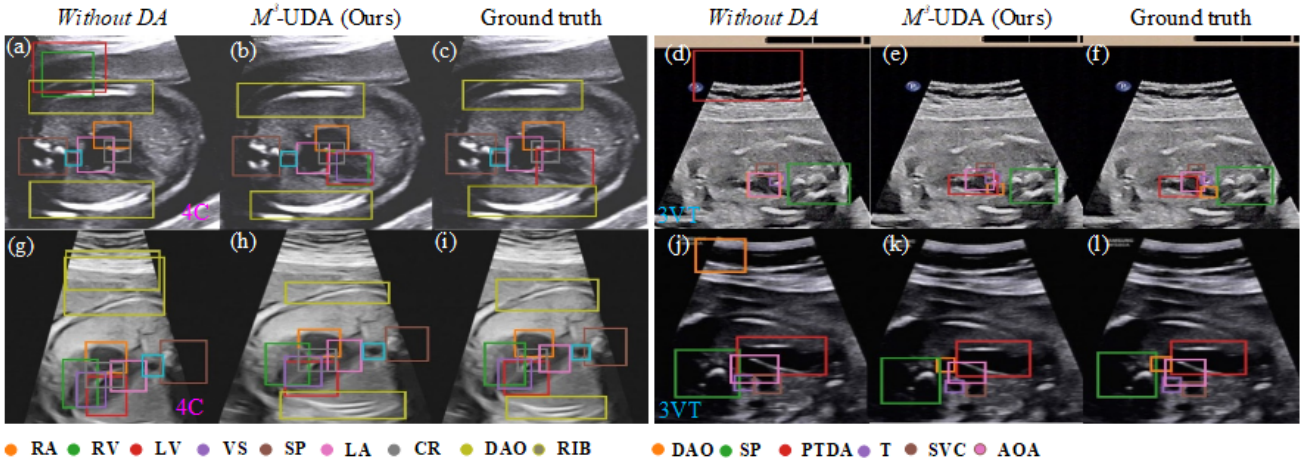
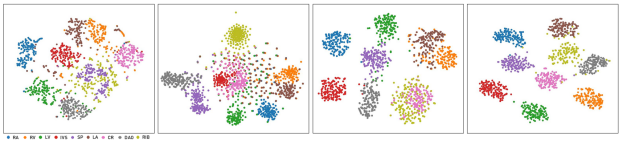


Figure 5. Qualitative result comparison on 4CC and 3VT of target domain.



(a) Baseline (b) SIGMA (c) SIGMA++ (d) M³-UDA(ours)

Figure 6. Visualization comparison of feature representation is performed by t-SNE [43].

Table 6. Ablation experiments on 4CC and 3VT.

	Hospital 1→2 on 4CC				Hospital 1→2 on 3VT			
Method	*HM*	*GM*	*SM*	mAP	*HM*	*GM*	*SM*	mAP
Baseline	✗	✗	✗	57.20	✗	✗	✗	36.78
Ours	✓	✗	✗	69.16	✓	✗	✗	51.49
	✓	✓	✗	73.17	✓	✓	✗	55.32
	✓	✓	✓	**76.81**	✓	✓	✓	**60.39**
	Hospital 2→1 on 4CC				Hospital 2→1 on 3VT			
Baseline	✗	✗	✗	64.32	✗	✗	✗	41.61
Ours	✓	✗	✗	67.41	✓	✗	✗	44.44
	✓	✓	✗	73.23	✓	✓	✗	47.70
	✓	✓	✓	**74.55**	✓	✓	✓	**49.36**

Second, we can observe the effectiveness of each matching component. For example, for Hospital 1→2 adaptive detection on 4CC, HM improves by 11.96%, SM enhances by 4.01%, and after adding GM, the detection mAP reaches 76.81%. Similarly, each matching module has a boost to our method, as reported in Table 6. This shows that our multi-matching approach has significant advantages and provides effective internal and global topology knowledge in the target domain, which is well-suitable for structure detection tasks in different hospital scenarios.

Feature Visualization Comparison. For each structure category of the fetus, we randomly sample the same number of pixels for the target domain on ResNet-101-based features and show t-SNE visualization with some comparison methods and the proposed method in Figure 6. As shown in Figure 6, on the 4CC view, similar categories such as RA, RV, LV, IVS, SP, LA, CR, DAO, and RI can be clearly separated in feature representation by our method. Similarly, on the 3VT view, our method can clearly distinguish DAO, IVS, PTDA, T, SVC, and AOA, which is beneficial for the later detector head in object recognition.

Qualitative Result Comparison. Figure 5 illustrates the qualitative results of our approach and *Without DA* in target hospital adaptation scenarios. We can observe that our method avoids the misdetection of one RV and one LV in (a) and (c) and a misdetection of the RIB in (g) and (h). Similarly, we can also see that our method prevents one false positive error of RV, one miss-detection of RV, and one false positive error of DAO in (d) and (h).

5. Conclusion

This paper presents a new Fetal Cardiac Structure (*FCS*) dataset and a multi-matching approach (M^3-UDA) as a new benchmark by histogram matching, sub-structure matching, and global-matching for the unsupervised adaptation fetus cardiac structure detection. Extensive experiments on our proposed *FCS*, as well as on the publicly available CardiacUDA dataset, the experimental results show that M^3-UDA outperforms the existing state-of-the-art UDA object detection methods significantly. Additional analysis and component ablation experiments also demonstrate the effectiveness of our method. We believe *FCS* and M^3-UDA will provide new insights for object detection-based methods, UDA-based approaches, and disease detection studies in medical scenarios.

Acknowledgements

This work is partially supported by a research grant from NSFC under Grant 62306254, a research grant from the Hong Kong Innovation and Technology Fund under Grant ITS/030/21, a research grant from the Beijing Institute of Collaborative Innovation (BICI) in collaboration with HKUST under Grant HCIC-004, a research grant from the National Key R&D Program of China (No. 2022YFF0606302) and NSFC (Nos. 62227808 and 62306003).

References

[1] Shan An, Haogang Zhu, Yuanshuai Wang, Fangru Zhou, Xiaoxue Zhou, Xu Yang, Yingying Zhang, Xiangyu Liu, Zhicheng Jiao, and Yihua He. A category attention instance segmentation network for four cardiac chambers segmentation in fetal echocardiography. *Computerized Medical Imaging and Graphics*, 93:101983, 2021. 1

[2] Rima Arnaout, Lara Curran, Yili Zhao, Jami C Levine, Erin Chinn, and Anita J Moon-Grady. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat. Med.*, 27(5):882–891, 2021. 1

[3] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017. 1

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79:151–175, 2010. 3

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT*, pages 177–186. Springer, 2010. 6

[6] Margarita Brida, Massimo Chessa, David Celermajer, Wei Li, Tal Geva, Paul Khairy, Massimo Griselli, Helmut Baumgartner, and Michael A Gatzoulis. Atrial septal defect in adulthood: a new paradigm for congenital heart disease. *Eur. Heart J.*, 43(28):2660–2671, 2022. 1

[7] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of CVPR*, pages 11457–11466, 2019. 3

[8] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of CVPR*, pages 23839–23848, 2023. 3, 6, 7

[9] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of CVPR*, pages 8869–8878, 2020. 2, 3

[10] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *Int. J. Comput. Vis.*, 129(7):2223–2243, 2021. 3

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of CVPR*, pages 3213–3223, 2016. 2

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. 6

[13] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of CVPR*, pages 4091–4101, 2021. 3

[14] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of CVPR*, pages 23829–23838, 2023. 3

[15] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1): 1087–1105, 2022. 3

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, 2016. 3

[17] Yipeng Gao, Lingxiao Yang, Yunmu Huang, Song Xie, Shiyong Li, and Wei-Shi Zheng. Acrofod: An adaptive method for cross-domain few-shot object detection. In *Proceedings of ECCV*, pages 673–690. Springer, 2022. 7

[18] Yipeng Gao, Kun-Yu Lin, Junkai Yan, Yaowei Wang, and Wei-Shi Zheng. Asyfod: An asymmetric adaptation paradigm for few-shot domain adaptive object detection. In *Proceedings of CVPR*, pages 3261–3271, 2023. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 6

[20] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of WACV*, pages 749–757, 2020. 2, 3

[21] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of CVPR*, pages 5001–5009, 2018. 3

[22] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of CVPR*, pages 6092–6101, 2019. 2, 3

[23] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of CVPR*, pages 12456–12465, 2019. 2, 3

[24] Sarah Leclerc, Erik Smistad, Pedrosa, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Trans. Med. Imaging*, 38(9):2198–2210, 2019. 2

[25] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *Proceedings of AAAI*, pages 1421–1428, 2022. 3

[26] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of CVPR*, pages 5291–5300, 2022. 3, 5, 7

[27] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 2, 3, 6, 7

[28] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of CVPR*, pages 7581–7590, 2022. 3

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[30] Zehui Lin, Shengli Li, Dong Ni, Yimei Liao, Huaxuan Wen, Jie Du, Siping Chen, Tianfu Wang, and Baiying Lei. Multi-task learning for quality assessment of fetal head ultrasound images. *Med. Image Anal.*, 58:101548, 2019. 1

[31] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of CVPR*, pages 23776–23786, 2023. 3

[32] Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ugne Klimiene, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, Markus Denzinger, David Niederberger, Sven Wellmann, Ece Ozkan, et al. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, 2024. 1

[33] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang. Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In *Proceedings of WACV*, pages 423–433, 2023. 7

[34] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Ebinger, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. 2

[35] Adam Paszke, Sam Gross, Massa, et al. Pytorch: An imperative style, high-performance deep learning library. *Proceedings of NeurIPS*, 32, 2019. 6

[36] Zhengquan Piao, Linbo Tang, and Baojun Zhao. Unsupervised domain-adaptive object detection via localization regression alignment. *IEEE Trans. Neural Netw. Learn. Syst.*, 2023. 7

[37] Bin Pu, Kenli Li, Shengli Li, and Ningbo Zhu. Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. *IEEE Trans. Ind. Inform.*, 17(11):7771–7780, 2021. 1

[38] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019. 2, 3

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.*, 126:973–992, 2018. 2

[40] Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of ECCV*, pages 763–780. Springer, 2020. 3

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 3

[42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of CVPR*, pages 9627–9636, 2019. 4, 6

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11):2579–2605, 2008. 8

[44] Vibashan Vs, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of CVPR*, pages 4516–4526, 2021. 2, 3

[45] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *Proceedings of CVPR*, pages 9603–9612, 2021. 3

[46] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of CVPR*, pages 12355–12364, 2020. 3

[47] Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of CVPR*, pages 11878–11887, 2023. 2, 6

[48] Fuxun Yu, Di Wang, Chen, et al. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In *Proceedings of WACV*, pages 382–391, 2022. 3

[49] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *Proceedings of ECCV*, pages 86–102. Springer, 2020. 2, 3

[50] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of CVPR*, pages 14217–14226, 2022. 3

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of ICCV*, pages 2223–2232, 2017. 3