Unsupervised Domain Adaptation for Anatomical Structure Detection in Ultrasound Images

Bin Pu [*1] **Xingguo Lv** [*2] **Jiewen Yang** [*1] **Guannan He** [3] **Xingbo Dong** [2] **Yiqun Lin** [1] **Shengli Li** [4] **Ying Tan** [4] **Fei Liu** [2] **Ming Chen** [5] **Zhe Jin** [2] **Kenli Li** [6] **Xiaomeng Li** [1]

Abstract

Models trained on ultrasound images from one institution typically experience a decline in effectiveness when transferred directly to other institutions. Moreover, unlike natural images, dense and overlapped structures exist in fetus ultrasound images, making the detection of structures more challenging. Thus, to tackle this problem, we propose a new Unsupervised Domain Adaptation (UDA) method integrated with the Topology Knowledge Transfer (TKT) and the Morphology Knowledge Transfer (MKT) module for fetus structure detection, named **ToMo-UDA**. TKT leverages prior knowledge of the medical anatomy of fetal as topological information, reconstructing and aligning anatomy features across source and target domains. Then, MKT formulates a more consistent and independent morphological representation for each substructure of an organ. To evaluate the proposed ToMo-UDA for ultrasound fetal anatomical structure detection, we introduce **FUSH**^{2}, a new **F**etal **U**ltra**S**ound benchmark, comprises **H**eart and **H**ead images collected from **Two** health centers, with 16 annotated regions. Our experiments show that utilizing topological and morphological anatomy information in ToMo-UDA greatly improves organ structure detection. This expands the potential for structure detection tasks in medical image analysis.

[*]Equal contribution [1]The Hong Kong University of Science and Technology [2]Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, Anhui University [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. Correspondence to: Kenli Li <lkl@hnu.edu.cn>, Xiaomeng Li <eexmli@ust.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

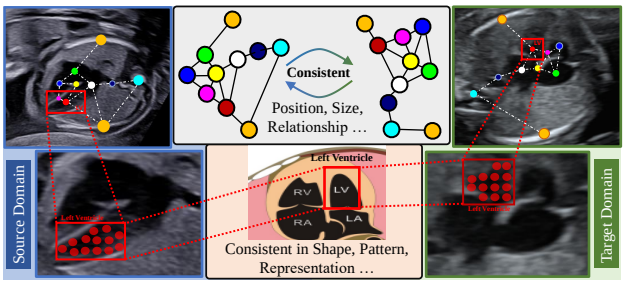

Figure 1: Ultrasound images contain prior knowledge across domains. All substructures of the same view remain consistent in topology (position, size, relationship) and morphology (shape, patterns, representation), which provides us with new insights into UDA, enhancing our capacity to bridge the domain gaps effectively (Best view in color).

1. Introduction

In clinical practice, the observation of anatomical structures allows for the direct diagnosis of many diseases (Xue et al., 2021; Lin et al., 2019; Arnaout et al., 2021; Zheng et al., 2023; Dai et al., 2022; Li et al., 2018). For example, the absence of cavum septi pellucidi structure in fetal head view is diagnosed as a severe disease called holoprosencephaly (Monteagudo, 2020). Therefore, anatomical structure detection serves as an essential foundation for disease diagnosis. Recently, deep learning (DL)-based methods as a powerful tool have already achieved significant progress in fetal anatomical structure detection, such as standard view quality control (Pu et al., 2021; Chen et al., 2017; Zhao et al., 2022; Wu et al., 2017), and disease diagnosis (Gong et al., 2019; Xu et al., 2022).

Nonetheless, applying DL-based models directly to anatomical structure detection in ultrasound data often yields suboptimal results, especially for data from multiple health centers (Guan & Liu, 2021). This is because real-world datasets have domain gaps (Oza et al., 2023; Li et al., 2023a) due to variations in data collection devices and obstetricians' scanning techniques across different hospital centers. Fine-tuning the DL models on the target data may solve the problem, but obtaining accurate annotations from obstetrician experts is either costly or unavailable. The diversity of machines equipped with various transducers further challenges annotations, presenting a significant hurdle for DL

approaches.

Unsupervised domain adaptation (UDA) has been proposed to solve the aforementioned challenges by mitigating the domain gaps. UDA aims to maximize the performance of the target domain while minimizing expert supervision through invariant feature learning (Vs et al., 2021), self-training (Zhao et al., 2020; Kim et al., 2019a), image translation (Chen et al., 2020; Hsu et al., 2020), domain randomization (Kim et al., 2019b; Rodriguez & Mikolajczyk, 2019), etc. In natural images, for example, the relationship between the objects is always chaotic and lacks specific patterns. In contrast, in ultrasound images, the relationship of anatomical structures, e.g., left ventricle and right ventricle, conforms to the theory of human anatomy and knowledge of topology regardless of health centers (domains). For example, as illustrated in Figure 2, the thalamus (T) structure in the fetal head view always appears in symmetrical pairs. Similarly, the two ribs (R) flanking the heart are another common example. In addition, sonographers diagnose mainly based on topological and morphological features (Chen et al., 2023), which provides us with new insights into UDA. As shown in Figure 1, substructures of the same view remain consistent in topology and morphology. In medical images, topological information focuses on the relationship between anatomical composition and positional relationship, while morphological information refers to the textural, shape, and morphology features of the interior of the anatomical structures.

The unique characteristics of ultrasound images indicate that previous methods for UDA object detection (UDAOD) in natural scenarios are not suitable or available for our task. UDAOD methods for natural scenarios do not consider a priori knowledge of medical images, yet this is one of the most significant properties in medical scenarios. For example, previous medical UDAOD methods have not considered topology knowledge and morphology information characteristic consistency for different domains. Motivated by the above discussion, we propose a novel UDA method named **ToMo-UDA** for fetus anatomical structure detection. The method includes two modules - Topology Knowledge Transfer (TKT) and Morphology Knowledge Transfer (MKT). TKT aligns features by reconstructing anatomy features, while MKT formulates consistent and independent representations for each substructure of an organ.

Collecting datasets from different health centers is challenging, and annotating multi-structure for these datasets is especially difficult, as it requires the participation of numerous experienced obstetricians. Therefore, multicenter ultrasound datasets with multiple structures of detailed box-level annotations are currently unavailable and scarce. To address the above discussion, the proposed **FUSH**^{2} that serves as **F**etal **U**ltra**S**ound benchmark with 1,978 **H**eart and 1,391 **H**ead views, is collected from **Two** health centers.

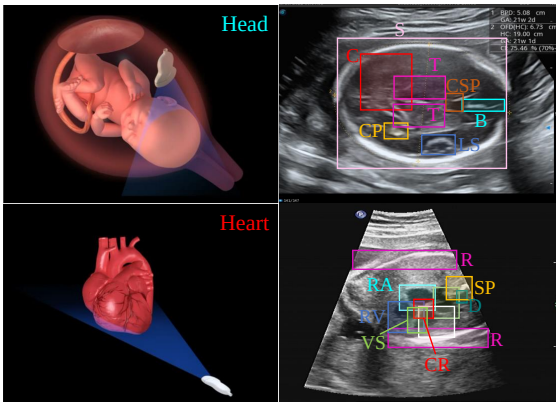

Figure 2: The fetal head and heart view with the key anatomical structures. Fetal ultrasound examination, which is required to screen for disease during pregnancy, relies on the presence of anatomical structures that are more challenging to examine than in adults.

Moreover, ultrasound images of **FUSH**^{2} are collected from different equipment, including Samsung, Sonoscape, and Philips. The gestational age of the fetus ranges from 20 to 34 weeks. All data were annotated with 16 anatomy regions and 2 view labels by ultrasonographers who have more than seven years of clinical experience. **In summarize, our contributions include:**

1. A comprehensive real-world fetal ultrasound dataset from two health centers with 1,978 heart and 1,391 head views, namely **FUSH**^{2}, is released. **FUSH**^{2} is labeled with 16 anatomy regions by experienced sonographers and includes various equipment and gestational weeks ranging from 20 to 34 weeks.

2. A new UDA method, namely ToMo-UDA, has been proposed. ToMo-UDA consists of two modules, i.e., TKT and MKT. TKT and MKT align ultrasound features, focusing on anatomical structures and morphological features for accurate diagnosis, respectively.

3. Extensive experiments show that the proposed ToMo-UDA outperforms all UDAOD baseline and state-of-the-art (SOTA) structure detection techniques with a clear margin. Our work opens up new possibilities for accurate and reliable object detection in medical image analysis. Datasets and source code are available at https://github.com/xmed-lab/ToMo-UDA.

2. Related Work

2.1. UDA Object Detection in Natural Scenarios

Recently, the UDAOD task has become a hot topic (Sindagi et al., 2020; Wang et al., 2021; Zhao & Wang, 2022; Zhao et al., 2020; Yu et al., 2022; Chen et al., 2021), and there have been many studies that can be broadly grouped into adversarial learning (Ganin & Lempitsky, 2015; Zheng et al., 2020; Vs et al., 2021), self-training (Kim et al., 2019a; Yu et al., 2019; Huang et al., 2021b), image-to-image transla-

tion (Kim et al., 2019b; Arruda et al., 2019; Huang et al., 2021a), and others (Deng et al., 2021; Li et al., 2022b; Rodriguez & Mikolajczyk, 2019). (He & Zhang, 2019) proposed to find the invariance feature from the source and target domain through multi-adversarial training. The self-training technique (Yang et al., 2022) utilizes unlabeled target data by training with target pseudo-labels, and a typical study (Liu et al., 2021) explores cycle self-training, a principled self-training algorithm that explicitly enforces cross-domain generalization of pseudo-labels. Recently, some remarkably new UDAOD algorithms have also emerged, e.g., mean-teacher training (Chen et al., 2022; Cao et al., 2023; Deng et al., 2023) and graph-based reasoning methods (Li et al., 2022a; 2023b; Liu et al., 2023b). Medical UDAOD differs from natural scenarios, leading to suboptimal performance with general object detection methods.

2.2. UDA in Medical Scenarios

Few works in the literature focus on UDAOD studies in medical scenarios. One pioneering work (Jin et al., 2023) uses adaptive adversarial training to learn domain-invariant features to minimize domain shifts. The more relevant topics are source-free (Liang et al., 2020) domain adaptive medical object detection (Liu et al., 2023a; Liu & Yuan, 2022; Xing et al., 2023) and UDA for segmentation (Shin et al., 2023; Yang et al., 2023; Huai et al., 2023; Liu et al., 2020). (Liu et al., 2023a) systematically analyzed the bias in source-free domain adaptation medical object detection by constructing a structural causal model and proposed an unbiased source-free domain adaptation framework based on the decoupled unbiased teacher. In another popular work, SMPT (Liu & Yuan, 2022) transfers the domain-invariant knowledge stored in the pre-trained source model to the target model via source knowledge distillation. Recently, in an unsupervised segmentation task, (Yang et al., 2023) proposed a mining prior knowledge of echocardiogram videos by aligning global and local features from source and target domains. In a nutshell, few studies have been conducted on UDAOD in medical scenarios due to the unavailability of datasets with detailed box-level annotations from multiple centers. The release of our dataset will benefit UDAOD in medical scenarios. In addition, previous studies have yet to fully explore topology and morphology knowledge in both source and target domains.

3. Method

Taking heart as an example, Figure 3 shows the overall pipeline of our ToMo-UDA, which consists of a source domain flow and a target domain flow. First, for both domains, a shared encoder $E(\cdot)$ based on feature pyramid network (Lin et al., 2017) is leveraged to extract features $\{f_k\}_{k=1}^{K}, f_k \in \mathbb{R}^{h_k \times w_k \times d}$ from input images, where d and K denote the total number of channels and feature map layers, respectively. Subsequently, the feature maps are passed to an object detection head (e.g., FCOS head), thus substructure centroid $\{c_i\}_{i=1}^{N}, c_i \in \mathbb{R}^{1\times d}$, bounding boxes $y^b \in \mathbb{R}^{N\times 4}$ and organ class $y^c \in \mathbb{R}^{N}$ are predicted from the detection head, here N represent the total number of organs. For the source domain, the ground truth annotation and prediction results are formulated as

$$\mathcal{L}_{supervised} = \mathcal{L}_{class}(y^c) + \mathcal{L}_{reg}(y^b), \tag{1}$$

for the supervision loss in object detection, where $\mathcal{L}_{class}$ is cross-entropy loss and $\mathcal{L}_{reg}$ is the $L1$ loss. On top of the above common object detection pipeline, we propose two modules in ToMo-UDA, named **T**opology **K**nowledge **T**ransfer (TKT) and **M**orphology **K**nowledge **T**ransfer (MKT), to bridge the domain gap from different hospitals. TKT allows for transferring the heart topology knowledge from the source to the target domain (see Section 3.1 and Appendix Section A2). MKT builds the complete inter-graph knowledge for different substructures, improving morphological representation consistency of the same substructure from different domains by minimizing their feature discrepancy (see Section 3.2 and Appendix Section A2).

3.1. Topology Knowledge Transfer

Unlike datasets like Cityscapes (Cordts et al., 2016) and COCO (Lin et al., 2014) for object detection in nature images, large domain gaps in ultrasound images actually depend on equipment manufacturers and physician experience across different medical centers. Despite the significant differences between the domains, we/sonographers have observed that substructures of the fetal heart consistently maintain their relative location. For instance, as shown in Figure 1, the substructure locations, such as locations of the left ventricle and left atrium of the heart in ultrasound images, remain consistent. Therefore, this consistent information can be utilized as the robust prior topology knowledge for our domain adaptation problem. Motivated by the above discussion, we concluded that fully annotated location labels of structures from the source domain can serve as complete structural information, making them suitable to be used as a standard reference for aligning heart knowledge in the target domain.

The TKT module is designed to align topology knowledge across domains to tackle the above problems. In the TKT module, we first take the centroid feature of each substructure $\{c_i\}_{i=1}^{N}$ generated by $E(\cdot)$ from both the source and target domain. Subsequently, we construct topology graphs $(\mathcal{V}, \mathcal{E})$ for each domain to represent the heart topology, where $\mathcal{V}$ denote the representation of N (e.g., $N = 9$ for heart) substructures, and $\mathcal{E}$ denote the set of edges connecting each substructure, respectively. To construct the representation of substructures, we introduce memory banks to maintain substructure features from large-scale data samples. Then, acquire centroid representation $\{\theta_i\}_{i=1}^{N}, \theta_i \in \mathbb{R}^{1\times d}$

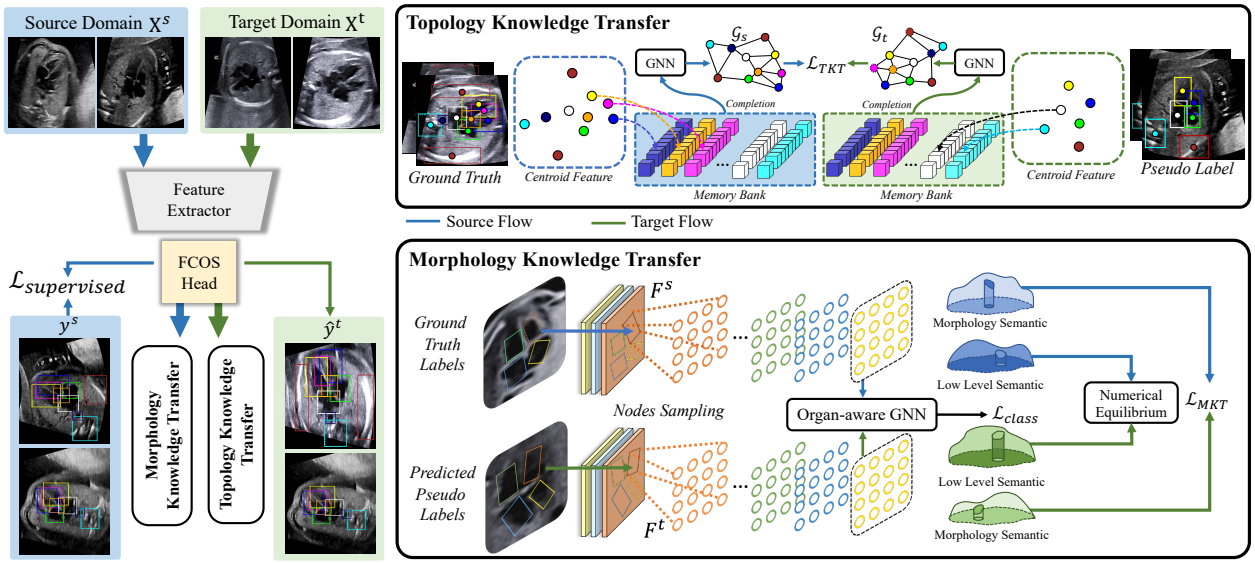

Figure 3: **The overview of the proposed ToMo-UDA.** A shared-parameter backbone and a detection head generate detection results from inputs of source and target domains. In the Topology Knowledge Transfer module, we obtain centroid representation from ground truth labels and pseudo-labels for the source and target domain, respectively, for constructing overall topological representations. In Morphology Knowledge Transfer, we sample nodes from feature maps from bounding boxes as morphological representations for each organ and align low-level and morphological semantics through the numerical equilibrium approach. In our testing stage, the trained extractor and detector are used for inference.

via average features of the bank of i-th substructure for $\mathcal{V}$. To ensure that banks can be updated synchronously, we also use the centroid θ_i to fill the empty nodes where the network misses detection.

When averaging the centroid feature of the i-th substructure, only intra-substructure discrepancy is considered, not inter-substructure discrepancy. The representation of each substructure should be distinguished with clear margins, and the centroid c_i in different samples should be close to its corresponding clustering centers θ_i. Thus, to complete the topology graph $(\mathcal{V}, \mathcal{E})$ for a sample, the edge $\mathcal{E}$ is computed by the pairwise distance between the centroid c_i of current sample and clustering center θ_i in the i-th substructure, formulated as $\mathcal{E} = \{c_i \cdot \theta_i^{\mathsf{T}}\}_{i=1}^{N}$, and $\mathcal{V} = \{c_i\}_{i=1}^{N}$.

To obtain the topological representation via graph, we apply the graph neural network (GNN) (Kipf & Welling, 2016) to acquire a more cohesive representation $\mathcal{G}$ from $(\mathcal{V}, \mathcal{E})$ through $\mathcal{G} = GNN(\mathcal{V}, \mathcal{E}), \mathcal{G} \in \mathbb{R}^{N \times d}$. In the GNN module, the embedding size is set to d to stay in line with the input $(\mathcal{V}, \mathcal{E})$. Then, to narrow the discrepancy across source and target domains, we optimize their transport distance between the graph $\mathcal{G}_s$ and $\mathcal{G}_t$ as:

$$\mathcal{L}_{dis}(s,t) = \inf_{\gamma \in \Gamma(s,t)} \left(\mathbb{E}_{(\mathcal{G}_s, \mathcal{G}_t) \sim \gamma}[\mathcal{G}_s, \mathcal{G}_t]^p\right)^{\frac{1}{p}}, \quad (2)$$

where $\gamma \in \Gamma(s,t)$ is the set of all couplings of training samples from source and target domains, γ and Γ denotes a joint probability measure and all joint probability distribution of $\gamma(\mathcal{G}_s, \mathcal{G}_t)$, respectively. In subsequent content, we use subscript letters s and t to represent the source and target domains, respectively.

Directly optimizing Equation 2 is challenging. Thus, we store features of heart substructures from data samples in memory banks and use centroid clustering to approximate the overall representation. This allows us to reformulate Equation 2 as a discrete form:

$$\mathcal{L}_{dis}(s,t) = \inf_{\pi} \left(\sum\nolimits_{i=1}^{N} ||\mathcal{G}_{t,i} - \mathcal{G}_{s,\pi(i)}||^p\right)^{\frac{1}{p}}, \quad (3)$$

where the i in $\mathcal{G}_{s/t,i}$ denotes the graph nodes of i-th substructure in graph $\mathcal{G}$ from source/target domain, and the infimum (inf) is over all permutations π of N heart substructures, computed by using the Sinkhorn (Cuturi, 2013) iteration. We use the $\mathcal{L}_{TKT} = \mathcal{L}_{dis}(\mathcal{G}_t, \mathcal{G}_\theta)$ to represent the overall loss of module TKT.

3.2. Morphology Knowledge Transfer

As shown in Figure 1, topology refers to the spatial relationships that remain constant globally, considering the overall representation of a specific view of the heart. While morphology deals with the form and shape of each substructure itself. To maintain the consistent representation of substructures across domains, we propose a technique called MKT to align the morphology representation of substructures across different domains.

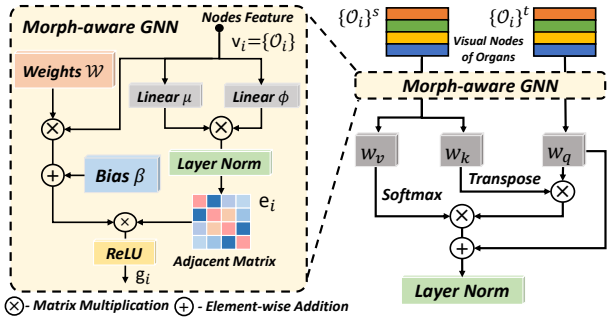

Figure 4: **The overview pipeline of the proposed Morph-aware GNN (MAGNN) and cross-domain graphical attention.** The weights and bias are the learnable parameters, where weights $\mathcal{W} \in \mathbb{R}^{d\times d}$ and $\beta \in \mathbb{R}^{1\times d}$.

Substructure Morphology Representation Formulation. To construct heart substructures representation, we sparsely sample features from each layer of feature map F according to the bounding boxes y_t^b and y_s^b. As the MTK module shown in Figure 3, we equidistantly sample M feature nodes $\{o_{i,j}\}_{i,j=1}^{N,M}$, $o_{i,j} \in \mathbb{R}^d$ for each substructure and concatenate sample nodes from deep to shallow layers of F (see Appendix A3: line 4-9 of Algorithm 1), where the sampled nodes can be represented by $\{\mathcal{O}_i\}_{i=1}^{N} = \{o_{i,j}\}_{i,j=1}^{N,M}$. For each substructure, to transfer their morphological information across domains, we introduce the auxiliary network named morph-aware GNN (MAGNN) to complete the morphological representation $\mathbf{g}_i = \text{MAGNN}(\{\mathcal{O}_i\})$, $\mathbf{g}_i \in \mathbb{R}^{M\times d}$. Refer to Figure 4, the input nodes is $\mathbf{v}_i = \{o_{i,j}\}_{i,j=1}^{N,M}$, and adjacent matrix $\mathbf{e}_i$ is computed by $\mathbf{e}_i = \{\text{LN}[\mu(o_{i,j}) \cdot \phi(o_{i,j})^\mathsf{T}]\}_{j=1}^{M}$, where LN denotes the layer normalization, μ, ϕ are Linear layers. Finally, the cross-domain graphical attention is introduced for cross-domain interaction via graph $\mathbf{g}_i^s$ and $\mathbf{g}_i^t$, formulated as $\text{LN}\left[\left(\mathcal{W}_q(\mathbf{g}_i^t) \cdot \mathcal{W}_k(\mathbf{g}_i^t)^\mathsf{T}\right)\mathcal{W}_v(\mathbf{g}_i^s)\right]$, the $\mathcal{W}_q, \mathcal{W}_k, \mathcal{W}_v \in \mathbb{R}^{d\times d}$ are the linear projection layers.

Essentially, all nodes in the source and target domains must be classified correctly based on the substructure label of each graph node of $\mathbf{g}_i$ from the i-th heart substructure, cross-entropy loss for node classification is defined to train the model:

$$\mathcal{L}_{class} = -\left(\sum_i^N \sum_j^M y_i^c log\left(\mathbf{g}_{i,\pi(j)}\right)\right)_{s/t}, \quad (4)$$

where $\pi(j)$ is the over all permutations of graph node in $\mathbf{g}_i$. The label y_i^c of the source and target domains comes from annotation and the prediction result, respectively.

As discussed, same substructure across different fetal hearts should remain a consistent morphological representation, i.e., intra-class similarity. Similar to the Equation 2 in Section 3.1, with the help of MAGNN network that constructs $\mathbf{g}_i$. We can perform the optimal transport to narrow the distribution discrepancy of i-th substructure across domains, formulated as:

$$W_p(\mathbf{g}_s, \mathbf{g}_t) = \left(\int_{M\times M} ||\mathbf{g}_s - \mathbf{g}_t||^p d\sigma\left(\mathbf{g}_s, \mathbf{g}_t\right)\right)^{\frac{1}{p}}, \quad (5)$$

where $\sigma(\mathbf{g}_s, \mathbf{g}_t)$ indicates how much "mass" must be transported between source and target domains in order to transform their graphical representation. $||\mathbf{g}_s - \mathbf{g}_t||^p$ represent the cost of transporting a unit mass between $\mathbf{g}_s$ and $\mathbf{g}_t$, measure on $M \times M$ feature nodes of the homologous pair of heart substructures from source and target domain with p-norm distance.

Feature Numerical Equilibrium. Equation 5 is designed to transfer the knowledge of morphology and reduce the domain gap, while the low-level feature alignment is not considered. For low-level feature alignment, we noticed a significant numerical distribution discrepancy between the source and target domains, which leads graphical representations $\mathbf{g}_s$ and $\mathbf{g}_t$ to the unequal sum of masses. Thus, the cost of transporting the unit mass may be dominated by $||\mathbf{g}_s - \mathbf{g}_t||$ instead of $\gamma\left(\mathbf{g}_s, \mathbf{g}_t\right)$, which hinders the optimization during training. To tackle this problem, we introduce the feature numerical equilibrium approach.

The numerical equilibrium approach first normalizes the feature $\{\mathcal{O}_i\}_{i=1}^N$ to $0 \sim 1$. Specifically, for each substructure, compute their cumulative distribution $\mathcal{P}(\{\mathcal{O}_i\})$ by Probability Density Function (PDF) $p(k)$ for both domains as the Equation 6:

$$\mathcal{P}(\{\mathcal{O}_i\}) = \int_{k=0}^{1} p(\{o_{i,j}\})dk. \quad (6)$$

Given a mapping $\mathcal{M}$, for numerical equilibrium between source and target domains, we optimize the mapping function through Equation 7:

$$\mathcal{M}^* = argmin_{\mathcal{M}} \sum_{j=0}^{N} \mathcal{D}(\mathcal{M}[\mathcal{P}_t(\{\mathcal{O}_{t,i}\})], \mathcal{P}_s(\{\mathcal{O}_{s,i}\})), \quad (7)$$

where $\mathcal{D}(\cdot,\cdot)$ is the p-norm distance metric, $\{\mathcal{O}_{s/t,i}\}$ denotes that sampled nodes from source/target domain. According to Figure 6 in ablation study Section 4.4, the numerical distribution can be shifted from shallow to deep layer of feature maps, which denotes that this process can be first implemented in the input image X and formulated as:

$$\mathcal{M}^* = argmin_{\mathcal{M}} \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=0}^{K} \mathcal{D}(\mathcal{M}[p(o_{t,i,j})], p(o_{s,i,j})), \quad (8)$$

where K denotes the grayscale value, $o_{s/t,i,j}$ denotes the j-th node of i-th substructure from source/target domain.

According to Equation 3 and Equation 8, we then formulate

Equation 5 to the discrete form as following:

$$\mathcal{L}_{dis}\left(\mathbf{g}_s, \mathbf{g}_t\right) = \sum_{i=1}^{N} \inf_{\pi} \left(\sum_{j=1}^{M} ||\mathbf{g}_{t,i,j} - \mathcal{M}(\mathbf{g}_{s,i,\pi(j)})||^p \right)^{\frac{1}{p}}, \tag{9}$$

where the infimum is over all permutations π of M graph nodes, $\mathbf{g}_{s/t,i,j}$ denotes that the j-th graph node of i-th sub-structure from source/target domain, and mapping $\mathcal{M}$ is formulated by Equation 8. The loss $\mathcal{L}_{MKT}$ of the MKT module in Section 3.2 is written as $\mathcal{L}_{MKT} = \mathcal{L}_{class} + \mathcal{L}_{dis}\left(\mathbf{g}_s, \mathbf{g}_t\right)$.

Finally, the overall loss of our proposed ToMo-UDA can be summarized as:

$$\mathcal{L}_{all} = \alpha\mathcal{L}_{TKT} + \beta\mathcal{L}_{MKT} + \mathcal{L}_{supervised}, \tag{10}$$

where α and β is the decay ratio of the loss $\mathcal{L}_{TKT}$ and $\mathcal{L}_{MKT}$. Our experiment found that when both α and β are set to 0.1, it can achieve the best domain adaptation performance.

4. Experiment

4.1. Datasets and Evaluation

The proposed FUSH^2 dataset was collected from two medical centers and includes fetal head and heart images. The dataset collection and experiment are approved by the local ethics committee with the approval number LLYJ2022-014-005. A senior and experienced sonographer annotated the bounding box of the organ structures and its class name. These ultrasound images are obtained by local sonographers from various ultrasound devices such as Samsung and SonoScape. There are a total of 1,391 fetal head images and 1,978 heart images in the FUSH^2 dataset.

When compared to other counterparts like **CAMUS** (Leclerc et al., 2019) and **EchoNet** (Ouyang et al., 2020), FUSH^2 collects data from multiple health centers with a wide range of resolutions. In contrast, CAMUS and EchoNet use data from a single health center. Additionally, when compared to **CardiacUDA** (Yang et al., 2023), FUSH^2 has wide-ranged resolutions and remarkably 16 annotated regions (9 for heart and 7 for head), whereas CardiacUDA only has 4. The benefits of our dataset compared to existing datasets are shown in Table 1. The main anatomical structure abbreviations are shown in Table 2.

MMWHS (Zhuang et al., 2019) consists of 20 unpaired MRI and 20 CT volumes with corresponding pixel-level segmentation ground truth. We use the pre-processing methods of PnP-AdaNet (Dou et al., 2019) and convert the segmentation masks into bounding boxes for four regions present in both MRI and CT modalities: the ascending aorta (AA), the left atrial blood cavity (LA-blood), the left ventricular blood cavity (LV-blood) and the left ventricular myocardium (LV-MYO). The dataset split ratios are also consistent with PnP-AdaNet.

Table 1: The comparison of our **FUSH^2**, CardiacUDA (Yang et al., 2023), CAMUS (Leclerc et al., 2019), and EchoNet (Ouyang et al., 2020).

Dataset	**Our FUSH^2 dataset**	CardiacUDA	CAMUS	EchoNet
Annotated Images	3,369	4,960	1,000	20,060
Multiple Centers	✓	✓	×	×
Views	2	4	1	1
Resolution	480-1080p	720p	480p	120p
Annotated Regions	LV, RV, LA, RA, DAO, VS, SP, CR, R, LS CSP, BM, T, S, C, CP	LV, RV, LA, RA	LV, LA	LV

Table 2: Professional terms and abbreviations of FUSH^2.

Heart		Head	
Structure	Abb	Structure	Abb
Left Atrium	LA	Thalamus	T
Right Atrium	RA	Lateral Sulcus	LS
Left Ventricle	LV	Choroid Plexus	CP
Right Ventricle	RV	Cavum Septi Pellucidi	CSP
Cross	CR	Brain Midline	BM
Rib	R	Skull	S
Ventricular Septum	VS	Cerebellum	C
Spine	SP	/	/
Descending Aorta	DAO	/	/

4.2. Implementation Details

We use ResNet101 (He et al., 2016) as our feature extractor. For the detection head, we choose one-stage (Tian et al., 2019) (two-stage (Ren et al., 2015) is shown in the Table A1) detection strategies. During training, we apply the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001, a batch size of 6, a momentum of 0.9, and a total of 100 training epochs. We apply proportional scaling, random flipping, and random erasing as preprocessing operations. For each dataset, we split it into training, validation, and test sets with a ratio of 7:1:2, respectively.

4.3. Comparison with SOTA

We performed extensive domain adaptive detection experiments on fetal heart and head from different medical centers. For example, the adaptive detection from center1 to center2 is denoted as center 1→2. The detection results were evaluated by mean Average Precision (mAP) with Intersection over Union (IoU) (Lin et al., 2014) thresholds over 0.5. In our table, the *Source Only* refers to evaluating weights trained only on the source domain directly to the target, without using any DA methods, while the *Target Only* refers to training on the target domain in a fully supervised manner.

UDA on heart. As reported in Table 3, On both center 1→2 and center 2→1, ToMo-UDA outperforms all the latest UDA SOTA methods, achieving the significant performance of 71.33% mAP and 88.71% mAP respectively. As shown in Figure 2, organs such as LA, RA, LV, RV, VS, and CR

Table 3: Domain adaptation results on the heart.

Methods	Center 1→2										Center 2→1									
	LA	RA	LV	RV	CR	R	VS	SP	DAO	mAP (%)	LA	RA	LV	RV	CR	R	VS	SP	DAO	mAP (%)
Source Only	35.06	38.77	34.46	36.67	52.86	52.51	42.19	54.66	48.01	43.90	70.58	80.29	76.08	65.34	79.98	61.64	75.96	88.07	79.21	75.23
Few-shot Methods																				
AcroFOD (Gao et al., 2022)	22.51	28.43	22.94	21.92	26.02	26.04	29.02	24.51	26.60	25.33	39.54	38.81	41.59	41.17	39.48	27.25	41.46	35.32	38.43	38.11
AsyFOD (Gao et al., 2023)	55.57	59.79	57.04	54.90	58.24	54.91	53.71	66.02	59.00	57.78	56.64	61.72	58.52	56.93	57.34	58.04	59.28	62.63	60.45	59.03
Unsupervised Methods																				
ConfMix (Mattolin et al., 2023)	43.90	59.53	61.37	53.55	46.12	51.67	50.83	60.33	61.94	54.36	59.97	66.83	67.22	62.30	56.76	45.77	63.32	66.11	53.97	60.25
SIGMA (Li et al., 2022a)	50.11	62.10	49.52	51.30	58.94	55.61	46.68	54.00	47.85	52.04	83.86	86.42	83.60	78.08	82.47	68.87	78.06	90.46	85.64	81.94
LRA (Piao et al., 2023)	41.68	24.37	34.43	32.18	41.05	70.21	46.14	27.88	25.32	38.14	88.43	71.82	79.30	84.66	84.83	64.46	83.76	82.33	80.53	80.01
CMT (Cao et al., 2023)	56.39	57.30	61.42	55.51	58.58	**75.80**	59.23	60.94	59.08	60.47	79.89	**91.51**	85.90	83.79	87.87	65.53	85.74	85.18	75.37	82.31
SIGMA++ (Li et al., 2023b)	57.12	56.97	60.21	58.32	56.05	58.87	55.46	59.08	60.11	57.83	86.88	88.19	82.83	79.55	87.17	66.66	82.20	82.62	86.48	82.50
Ours	**64.23**	**75.62**	**70.40**	**64.32**	**66.69**	75.03	**75.48**	**77.24**	**72.99**	**71.33**	**93.91**	90.25	**94.72**	**88.20**	**89.92**	**69.86**	**90.66**	**90.91**	**90.06**	**88.71**
Target Only	65.95	75.63	72.39	73.88	72.86	76.60	76.70	77.41	71.55	73.66	82.91	86.61	83.47	82.75	89.09	72.50	85.93	90.15	89.61	84.78

Table 4: Domain adaptation results on the head.

Methods	Center 1→2								Center 2→1							
	T	LS	CP	CSP	BM	S	C	mAP (%)	T	LS	CP	CSP	BM	S	C	mAP (%)
Source Only	63.88	64.02	53.86	53.40	90.91	59.78	45.25	61.58	42.33	77.52	79.89	50.73	43.43	82.89	30.83	58.23
Few-shot Methods																
AcroFOD (Gao et al., 2022)	42.98	36.11	40.25	41.61	47.52	41.47	38.52	41.20	61.28	60.42	61.34	60.95	59.38	61.87	36.16	57.34
AsyFOD (Gao et al., 2023)	39.96	35.07	39.79	39.20	47.01	39.17	34.03	39.17	60.92	60.06	60.95	60.61	58.98	61.51	36.16	57.03
Unsupervised Methods																
ConfMix (Mattolin et al., 2023)	62.38	60.88	53.79	71.44	48.97	95.23	50.34	63.29	**69.31**	73.40	52.45	**66.53**	36.82	96.11	40.51	62.16
SIGMA (Li et al., 2022a)	72.62	64.94	62.15	67.96	64.72	95.63	58.58	69.51	48.40	71.97	79.89	54.42	49.13	77.68	28.58	58.58
LRA (Piao et al., 2023)	59.00	61.74	66.93	71.29	64.62	**99.43**	67.92	70.13	54.04	30.98	32.68	27.60	46.77	93.25	62.11	49.63
CMT (Cao et al., 2023)	79.53	**80.84**	75.56	77.57	52.33	98.00	64.73	75.94	33.84	66.16	**86.17**	42.42	47.92	**99.97**	29.17	57.95
SIGMA++ (Li et al., 2023b)	80.72	65.70	69.63	77.27	67.63	98.22	61.02	74.31	53.81	**75.04**	85.17	53.86	51.83	79.08	55.99	64.96
Ours	**86.57**	76.86	**77.71**	**84.31**	**74.34**	90.91	**78.39**	**81.30**	64.00	72.69	72.04	64.19	**62.36**	88.68	**67.96**	**70.27**
Target Only	79.76	88.17	77.62	82.12	71.10	96.51	70.02	80.75	97.00	99.05	90.54	87.81	87.06	100.00	73.83	90.75

Table 5: Cross-modal adaptation results on MMWHS.

Methods	CT ⟶ MRI					MRI ⟶ CT				
	LV-MYO	LA-blood	LV-blood	AA	mAP (%)	LV-MYO	LA-blood	LV-blood	AA	mAP (%)
Source Only	27.02	9.09	59.06	22.14	29.32	34.51	56.80	32.36	32.22	38.97
ConfMix (Mattolin et al., 2023)	58.43	32.00	70.29	24.01	46.18	45.77	53.20	59.88	39.03	49.46
SIGMA (Li et al., 2022a)	62.12	24.52	80.09	16.30	45.76	51.71	65.50	53.33	48.14	54.67
LRA (Piao et al., 2023)	75.18	8.85	74.97	10.74	42.43	71.34	55.79	79.32	61.00	66.86
CMT (Cao et al., 2023)	**84.41**	17.37	82.14	39.98	56.97	67.75	72.59	66.49	74.72	70.39
SIGMA++ (Li et al., 2023b)	60.91	44.59	79.52	23.63	52.16	67.10	61.25	74.24	60.89	65.87
AT (Li et al., 2022b)	81.87	17.87	79.49	23.73	50.74	65.52	74.25	66.20	**78.27**	71.06
Ours	77.10	**51.16**	**82.60**	**40.35**	**62.80**	**82.43**	**80.07**	**81.00**	70.02	**78.38**
Target Only	85.38	74.49	86.45	77.00	80.83	83.68	86.43	81.78	80.72	83.15

Table 6: Domain adaptation results on CardiacUDA dataset.

Methods	Site R → Site G				
	LV	RV	LA	RA	mAP (%)
Source Only	72.76	76.68	75.49	66.52	72.86
ConfMix (Mattolin et al., 2023)	66.38	71.76	64.62	51.29	63.51
SIGMA (Li et al., 2022a)	78.37	81.20	75.83	69.12	76.13
CMT (Cao et al., 2023)	**87.13**	80.46	74.85	57.11	74.89
SIGMA++ (Li et al., 2023b)	84.71	**85.76**	75.44	66.08	77.99
Ours	85.12	83.83	**85.38**	**86.53**	**85.21**
Target Only	83.73	81.92	81.58	82.17	82.35
Methods	Site G → Site R				
	LV	RV	LA	RA	mAP (%)
Source Only	97.33	87.48	90.91	90.03	91.44
ConfMix (Mattolin et al., 2023)	53.90	65.80	66.40	59.30	61.40
SIGMA (Li et al., 2022a)	**97.21**	84.48	94.96	95.28	92.98
CMT (Cao et al., 2023)	90.89	81.32	87.86	74.64	83.68
SIGMA++ (Li et al., 2023b)	90.17	87.66	**99.08**	94.69	92.90
Ours	90.12	**90.34**	98.83	**99.03**	**94.58**
Target Only	96.33	90.79	99.07	99.71	96.48

Table 7: Ablation results on heart and head datasets.

Methods	Center 1→2 on heart				Center 1→2 on head			
	TKT	*MKT*	*NE*	mAP (%)	*TKT*	*MKT*	*NE*	mAP (%)
Baseline	-	-	-	43.90	-	-	-	61.58
Ours	✓	✗	✗	58.60	✓	✗	✗	77.16
	✗	✓	✗	62.24	✗	✓	✗	76.89
	✓	✓	✗	68.97	✓	✓	✗	78.96
	✓	✓	✓	**71.33**	✓	✓	✓	**81.30**
Methods	Center 2→1 on heart				Center 2→1 on head			
	TKT	*MKT*	*NE*	mAP (%)	*TKT*	*MKT*	*NE*	mAP (%)
Baseline	-	-	-	75.23	-	-	-	58.23
Ours	✓	✗	✗	82.97	✓	✗	✗	59.96
	✗	✓	✗	84.72	✗	✓	✗	62.72
	✓	✓	✗	86.66	✓	✓	✗	67.56
	✓	✓	✓	**88.71**	✓	✓	✓	**70.27**

exhibit dense overlapping in these views. Traditional detection methods are prone to false positives and negatives under these conditions. In contrast, our approach, which synergizes topological and morphological knowledge, effectively overcomes these challenges. As shown in Table 3, our detection capability on center 1→2 is approaching the target only, and remarkably, it surpasses the target only on center 2→1. This superior performance can be attributed largely to the diverse styles captured in the center2 dataset, which endows the trained models with enhanced generalization capabilities and robustness. One observation is that individual few-shot UDA methods that are lower than ***Source Only***, which may be caused by the samples selected not being the representation of the overall target domain, may lead to domain bias.

UDA on Head. We conducted experiments focused on the adaptive detection of key structures in the fetal head. As depicted in Figure 2, although the overlap of brain structures is less compared to cardiac sections, ultrasound detection of the brain remains a significant challenge. This is due to

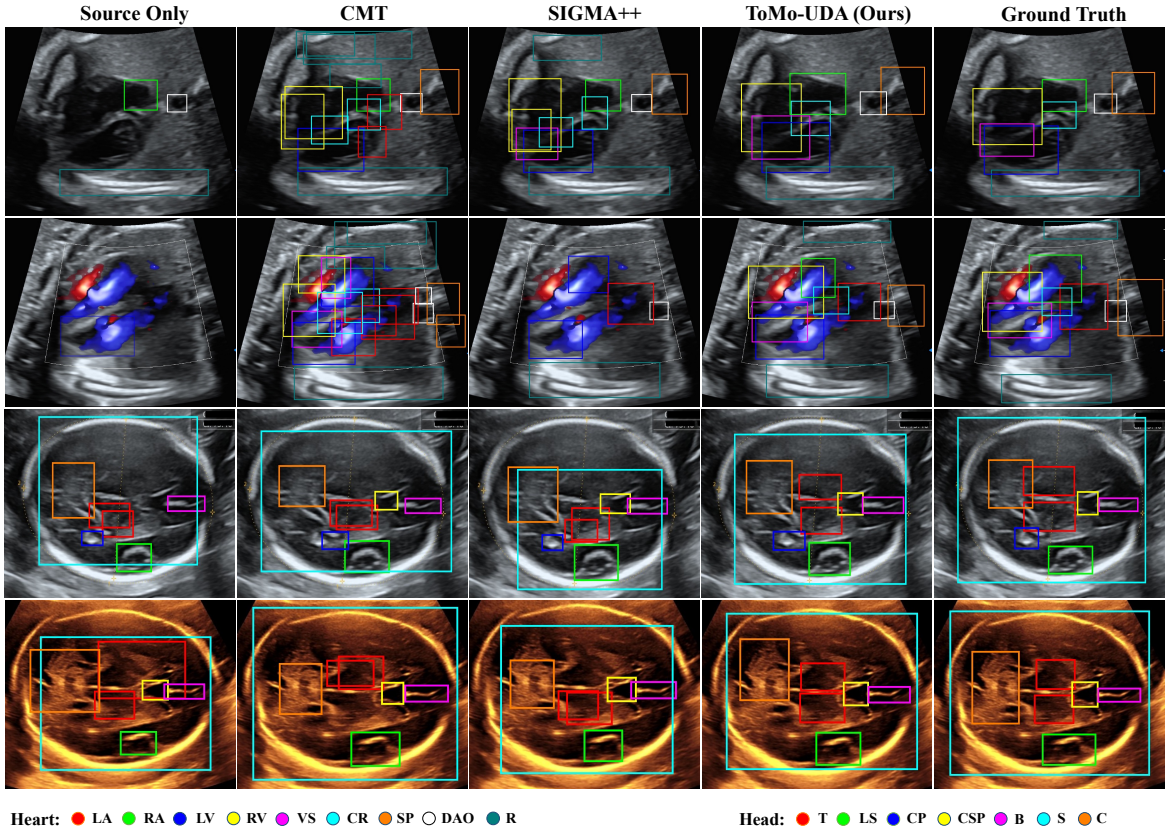

Figure 5: Qualitative result comparison of ***Source Only***, CMT (Cao et al., 2023), SIGMA++ (Li et al., 2023b), ToMo-UDA and Ground Truth. The first and second rows show the results from center 1→2 and center 2→1 on the heart of the FUSH^2 dataset. The third and last rows show the results from center 1→2 and center 2→1 on the head of the FUSH^2 dataset.

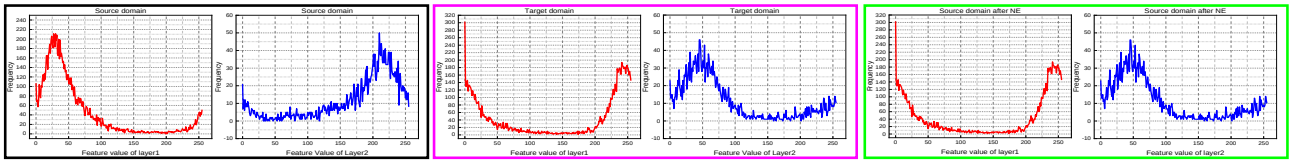

Figure 6: The distribution of the numerical value from the feature maps. Black box: source domain. Magenta box: target domain. Green box: source domain after the proposed Numerical Equilibrium (NE).

the considerable individual variability in fetal development, coupled with the complexity of brain anatomy. Despite these challenges, our method demonstrated excellent detection performance. As shown in Table 4, we achieved the best detection results between the two centers. On center 1→2, our method improved by 5.36%, and on center 2→1, we observed an improvement of 5.31%. It's notable that on center 2→1, despite achieving the best detection metrics, there remains a considerable gap from the target only. We attribute this to the lower resolution of images in center2 compared to center1, with resolutions typically around 648×480 in center2, whereas center1 images are generally around 1280×872. This resolution disparity makes it challenging for models trained on center2 to extract detailed organ structural information, leading to less effective domain adaptation.

UDA on CardiacUDA Dataset. CardiacUDA was originally proposed in (Yang et al., 2023) to explore unsupervised domain adaption for echocardiogram video segmentation. We test the proposed ToMo-UDA using the heart view from the CardiacUDA dataset for comparison with existing methods. Specifically, for CardiacUDA, each video was labelled with 5 frames, and the segmentation mask annotations were transformed into the bounding box, including LA, RA, LV, and RV structures. The performance of ToMo-UDA on the CardiacUDA is summarized in Table 6. The results suggest that our method still outperforms all UDAOD methods. From Site R → Site G, our method slightly underperforms other methods by 2.01% and 1.93% on the LV and RV structures. However, for the LA and RA, which are relatively more challenging to detect, our method surpasses the second-best method by 9.55% and 17.41%,

respectively. Existing UDAOD methods often experience performance degradation when detecting smaller objects. By incorporating topological and morphological priors from medical images, our approach achieves balanced detection results for each organ structure.

UDA on MMWHS Dataset. Table 5 shows the comparisons of our ToMo-UDA with other SOTA methods for cross-modal adaptation. Our method outperforms the second-best by 5.83% mAP on CT → MRI and by 7.32% mAP on MRI → CT. This demonstrates the effectiveness and scalability of our ToMo-UDA.

4.4. Analysis and Ablation Study.

Ablation on Source Only. As shown in Table 7, compared to the Source Only baseline, ToMo-UDA substantially improves the cross-domain fetal structure detection task. For example, ToMo-UDA outperforms Source Only by 27.43% mAP and 19.72% mAP for the adaptation center 1→2 detection on fetal heart and head, respectively, which demonstrates the effectiveness of the proposed method.

Ablation on TKT. Compared to Source Only, TKT increases mAP by 14.70% and 7.74% for adaptive center 1→2 and center 2→1 heart detection, respectively, as shown in Table 7. The same pattern can be found in head detection. For example, TKT improves mAP by 15.58% compared to Source Only in adaptive center 1→2 head detection. This indicates that TKT can transfer topological knowledge between source and target domains to facilitate better adaptive detection through invariant topological knowledge.

Ablation on MKT. Similarly, compared to the Source Only again, MKT enhances mAP by 15.31% and 4.49% in center 1→2 and center 2→1 head detection, respectively. In heart structure detection, MKT significantly improves detection performance. These experimental results demonstrate the effectiveness of aligning morphological knowledge within the structure.

Ablation on Numerical Equilibrium. As noted in Section 3.2, inter-domain discrepancies in low-level numerical distributions can impact the performance of our model. Given the challenges of computing continuous distributions, we have adopted a domain-mapping approach based on the numerical distribution of the input image. Figure 6 visualizes the numerical distributions of feature maps from the first two layers, demonstrating the impact of our numerical equilibrium. It highlights significant differences in numerical distributions of the feature map layers between the source and target domains (magenta box). However, after applying the numerical equilibrium operation (green box), the low-level feature distribution of the source domain has been adjusted closely to the target domain, and this equilibrium effect is equally applicable to deeper layers. The performance improvement is shown in Table 7. These results show the effectiveness of the proposed numerical equilibrium.

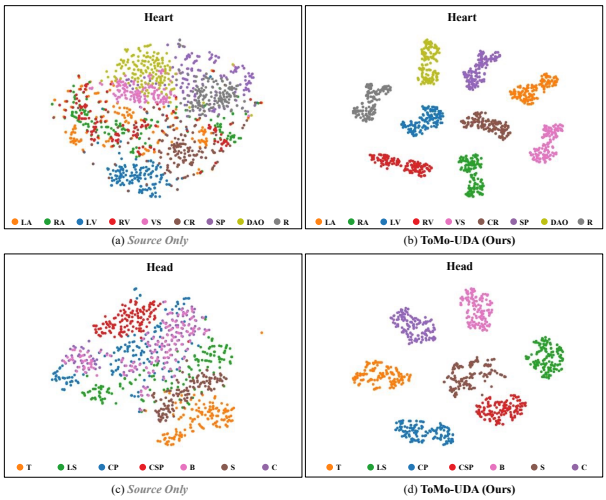

Figure 7: Feature visualization by t-SNE technique is performed by ***Source Only*** and our ToMo-UDA on the fetal heart and head of the target domain.

Qualitative Result Comparison. A comparison of the quantitative detection results is presented in Figure 5. We selected Source Only and the latest UDA methods CMT (Cao et al., 2023) and SIGMA++ (Li et al., 2023b) for visual comparison. On complex views such as the heart, Source Only struggles to detect intricate and overlapping structures, especially in color ultrasound. Both CMT and SIGMA++ have varying degrees of false detections, missed detections or duplicate detections. In contrast, the results of our ToMo-UDA closely match the ground truth annotations. The same is true for the head view.

Figure 7 shows the feature distribution visualization by our method and ***Source Only***. In Figure 7(a) and (b), we can clearly observe that our method can distinguish the various anatomical structures. However, the entangled distribution of categories shows that ***Source Only*** is difficult to separate the key structures. Similarly, in Figure 7(c) and (d), we find the same advantage of our ToMo-UDA.

5. Conclusion

This work proposes the ToMo-UDA for the issue of adaptive detection of fetal key structures in medical scenarios by aligning the morphological knowledge and topology knowledge of the source and target domains. Extensive experiments verify the effectiveness of ToMo-UDA in UDAOD on the collected and public datasets. We intuitively understand how the proposed ToMo-UDA works in the UDAOD task through ablation experiments and visualizations. In addition, we will release a new valuable dataset (FUSH^2) for fetal structure detection across domains, and we believe that FUSH^2 and ToMo-UDA can further inspire the community to address object detection and domain adaptive problems. Please see *Appendix Section A3* for Limitations.

Acknowledgements

This work is partially supported by a research grant from NSFC under Grant 62306254, a research grant from the RGC HKSAR (Project Reference Number: T45-401/22-N), 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).

Impact Statement

This study released the first multicenter ultrasound dataset with multi-structure box-level annotations that will greatly benefit the medical image analysis community. In addition, this work is dedicated to enhancing the accuracy of detecting key anatomical structures in the fetal heart and brain through ultrasound imaging technology, aiming to support early diagnosis and intervention, thereby positively impacting pregnancy management and fetal health. Ethically, we strictly adhere to the ethical standards of medical research, and ensure that all image data collection and experiments are approved by the local ethics committee.

Looking to the future, the development and application of this technology are expected to have profound impacts on society. It has the potential not only to improve health outcomes for pregnant women and families, reducing the risk of neonatal mortality and developmental disabilities, but also to alleviate the burden on the healthcare system by reducing long-term healthcare costs through early intervention.

References

Arnaout, R., Curran, L., Zhao, Y., Levine, J. C., Chinn, E., and Moon-Grady, A. J. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nature Medicine*, 27(5):882–891, 2021.

Arruda, V. F., Paixao, T. M., Berriel, R. F., De Souza, A. F., Badue, C., Sebe, N., and Oliveira-Santos, T. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Cao, S., Joshi, D., Gui, L.-Y., and Wang, Y.-X. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23839–23848, 2023.

Chen, C., Zheng, Z., Ding, X., Huang, Y., and Dou, Q. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878, 2020.

Chen, D., Xia, Y., Zhao, Y., Qian, X., Yin, J., Li, D., Zhang, J., and Zhang, J. Topology-aware brain vessel segmentation in ultrafast doppler imaging. In *2023 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4. IEEE, 2023.

Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.-Z., Ni, D., and Heng, P.-A. Ultrasound standard plane detection using a composite neural network framework. *IEEE Transactions on Cybernetics*, 47(6):1576–1586, 2017.

Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al. Learning domain adaptive object detection with probabilistic teacher. In *International Conference on Machine Learning*, pp. 3040–3055. PMLR, 2022.

Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Dai, W., Li, X., Ding, X., and Cheng, K.-T. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. *IEEE Transactions on Medical Imaging*, 2022.

Deng, J., Li, W., Chen, Y., and Duan, L. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4091–4101, 2021.

Deng, J., Xu, D., Li, W., and Duan, L. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23829–23838, 2023.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., and Heng, P.-A. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189. PMLR, 2015.

Gao, Y., Yang, L., Huang, Y., Xie, S., Li, S., and Zheng, W.-S. Acrofod: An adaptive method for cross-domain few-shot object detection. In *European Conference on Computer Vision*, pp. 673–690. Springer, 2022.

Gao, Y., Lin, K.-Y., Yan, J., Wang, Y., and Zheng, W.-S. Asyfod: An asymmetric adaptation paradigm for few-shot domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3261–3271, 2023.

Gong, Y., Zhang, Y., Zhu, H., Lv, J., Cheng, Q., Zhang, H., He, Y., and Wang, S. Fetal congenital heart disease echocardiogram screening based on dgacnn: adversarial one-class classification combined with video transfer learning. *IEEE Transactions on Medical Imaging*, 39(4): 1206–1222, 2019.

Guan, H. and Liu, M. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, Z. and Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6668–6677, 2019.

Hsu, H.-K., Yao, C.-H., Tsai, Y.-H., Hung, W.-C., Tseng, H.-Y., Singh, M., and Yang, M.-H. Progressive domain adaptation for object detection. In *Proceedings of Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 749–757, 2020.

Huai, Z., Ding, X., Li, Y., and Li, X. Context-aware pseudo-label refinement for source-free domain adaptive fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 618–628. Springer, 2023.

Huang, J., Guan, D., Xiao, A., and 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, 2021a.

Huang, J., Guan, D., Xiao, A., and Lu, S. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021b.

Jin, H., Che, H., and Chen, H. Unsupervised domain adaptation for anatomical landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 695–705. Springer, 2023.

Kim, S., Choi, J., Kim, T., and Kim, C. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6092–6101, 2019a.

Kim, T., Jeong, M., Kim, S., Choi, S., and Kim, C. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12456–12465, 2019b.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E. A. R., Jodoin, P.-M., Grenier, T., et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9): 2198–2210, 2019.

Li, M., Zhang, H., Li, J., Zhao, Z., Zhang, W., Zhang, S., Pu, S., Zhuang, Y., and 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, 2023a.

Li, W., Liu, X., and Yuan, Y. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5291–5300, 2022a.

Li, W., Liu, X., and Yuan, Y. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

Li, Y.-J., Dai, X., Ma, C.-Y., Liu, Y.-C., Chen, K., Wu, B., He, Z., Kitani, K., and Vajda, P. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022b.

Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., and Lei, B. Multi-task learning for quality assessment of fetal head ultrasound images. *Medical Image Analysis*, 58:101548, 2019.

Liu, D., Zhang, D., Song, Y., Zhang, F., O'Donnell, L., Huang, H., Chen, M., and Cai, W. Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. *IEEE Transactions on Medical Imaging*, 40(1): 154–165, 2020.

Liu, H., Wang, J., and Long, M. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021.

Liu, X. and Yuan, Y. A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Transactions on Medical Imaging*, 41(7):1897–1908, 2022.

Liu, X., Li, W., and Yuan, Y. Decoupled unbiased teacher for source-free domain adaptive medical object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Liu, Y., Wang, J., Huang, C., Wang, Y., and Xu, Y. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23776–23786, 2023b.

Mattolin, G., Zanella, L., Ricci, E., and 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.

Monteagudo, A. Holoprosencephaly. *American Journal of Obstetrics & Gynecology*, 223(6):B13–B16, 2020.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580 (7802):252–256, 2020.

Oza, P., Sindagi, V. A., Sharmini, V. V., and Patel, V. M. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Piao, Z., Tang, L., and Zhao, B. Unsupervised domain-adaptive object detection via localization regression alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Pu, B., Li, K., Li, S., and Zhu, N. Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. *IEEE Transactions on Industrial Informatics*, 17(11): 7771–7780, 2021.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Rodriguez, A. L. and Mikolajczyk, K. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019.

Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., and Hwang, D. Sdc-uda: Volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7412–7421, 2023.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sindagi, V. A., Oza, P., Yasarla, R., and Patel, V. M. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, pp. 763–780. Springer, 2020.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

Vs, V., Gupta, V., Oza, P., Sindagi, V. A., and Patel, V. M. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

Wang, Y., Zhang, R., Zhang, S., Li, M., Xia, Y., Zhang, X., and Liu, S. Domain-specific suppression for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9603–9612, 2021.

Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., and Ni, D. Fuiqa: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Transactions on Cybernetics*, 47(5):1336–1349, 2017.

Xing, F., Yang, X., Cornish, T. C., and Ghosh, D. Learning with limited target data to detect cells in cross-modality images. *Medical Image Analysis*, 90:102969, 2023.

Xu, M., Zhang, T., and Zhang, D. Medrdf: A robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack. *IEEE Transactions on Medical Imaging*, 41(8):2130–2143, 2022.

Xue, C., Zhu, L., Fu, H., Hu, X., Li, X., Zhang, H., and Heng, P.-A. Global guidance network for breast lesion segmentation in ultrasound images. *Medical Image Analysis*, 70:101989, 2021.

Yang, J., Shi, S., Wang, Z., Li, H., and Qi, X. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6354–6371, 2022.

Yang, J., Ding, X., Zheng, Z., Xu, X., and Li, X. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11878–11887, 2023.

Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., and Chen, X. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*, 2019.

Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., and Chen, X. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 382–391, 2022.

Zhao, G., Li, G., Xu, R., and Lin, L. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pp. 86–102. Springer, 2020.

Zhao, L. and Wang, L. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14217–14226, 2022.

Zhao, L., Li, K., Pu, B., Chen, J., Li, S., and Liao, X. An ultrasound standard plane detection model of fetal head based on multi-task learning and hybrid knowledge graph. *Future Generation Computer Systems*, 135:234–243, 2022.

Zheng, Y., Huang, D., Liu, S., and 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.

Zheng, Z., Yang, J., Ding, X., Xu, X., and Li, X. Gl-fusion: Global-local fusion network for multi-view echocardiogram video segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 78–88. Springer, 2023.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M. P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical Image Analysis*, 58:101537, 2019.

A1. More Ablation Studies

The results of different backbones and detection heads for the adaptive detection task are shown in Table A1

Table A1: Quantitative adaptation results on the heart.

Center 1→2											
Detection head	Feature extractor	RA	RV	LV	VS	SP	LA	CR	DAO	R	mAP (%)
Faster-RCNN (Ren et al., 2015)	VGG-16 (Simonyan & Zisserman, 2014)	61.21	62.31	64.17	60.18	69.14	61.70	70.47	65.61	66.30	64.57
	ResNet-50 (He et al., 2016)	65.27	77.12	64.40	73.69	68.53	60.63	72.04	69.99	77.20	69.88
	ResNet-101 (He et al., 2016)	65.47	75.36	71.96	67.61	75.21	75.36	60.57	78.17	71.96	71.29
	ResNet-152 (He et al., 2016)	66.19	73.91	78.82	72.88	77.40	70.17	64.14	78.22	64.46	71.80
FCOS (Tian et al., 2019)	VGG-16 (Simonyan & Zisserman, 2014)	60.78	62.69	63.12	60.89	61.98	71.63	61.40	69.31	75.08	65.21
	ResNet-50 (He et al., 2016)	61.04	64.47	60.49	70.49	65.39	77.18	72.19	76.99	64.73	68.11
	ResNet-101 (He et al., 2016)	64.23	75.62	70.40	64.32	66.69	75.03	75.48	77.24	72.99	71.33
	ResNet-152 (He et al., 2016)	60.94	71.24	76.39	72.14	69.67	76.40	72.93	67.83	71.01	70.95

We have tabulated the number of trainable parameters, FLOPs, and the time cost per step (forward and backward propagation) for various settings. The results are presented in the following Table A2. In the baseline, we used ResNet101 with an FCOS detection head. Upon adding the TKT module, the number of trainable parameter increased by 0.92M, FLOPs increased by 28.54G, and the time increased by 0.01s. Continuing with the addition of the MKT module, the parameters, FLOPs, and time increased by 2.21M, 113.52G, and 0.06s respectively. Upon incorporating NE, there was no significant change in parameters and FLOPs, but the time increased by 0.48 seconds. This is due to NE's inability for parallel computation, involving interactions between CPU and GPU, which affects network speed.

The ablation experiments of sensitivity for α and β in Equation 10. We trained on heart center 1→2 for 50 epochs and selected the weights from the last epoch for testing on the test set. We set α and β at four different levels each. The final results are shown in the Table A3.

Table A2: Additional implementation details.

Module	*TKT*	*MKT*	*NE*	Params. (M)	GFLOPs	Time (s)
Baseline	✗	✗	✗	55.24	742.57	0.63
Ours	✓	✗	✗	56.16	771.11	0.64
	✓	✓	✗	58.37	884.63	0.70
	✓	✓	✓	58.37	884.81	1.18

Table A3: The mAP(%) under different α and β.

α \ β	0.01	0.1	0.5	1
0.01	67.15	69.23	69.80	68.03
0.1	68.34	**71.08**	70.03	70.24
0.5	68.57	70.44	69.11	67.29
1	67.90	69.98	68.27	66.47

A2. Algorithm Pipeline.

Algorithms 1 and 2 outline the basic procedures of our proposed Topology Knowledge Transfer (TKT) and Morphology Knowledge Transfer (MKT), respectively.

A3. Limitations

In medical images, different structures of the cardiac or head actually do not overlap in the real scenario. However, the bounding box in object detection may include some irrelevant information, such as background or even information from other structures (see Figures 2 and 5). Hence, when we construct the topology and morphology information for the ultrasound image, that irrelevant information will also be introduced for training, which may degrade the performance of our method.

Our method focuses more on the data that have a fixed structure, which is suitable for most of the human body, such as the cardiac and brain. However, in some cases, such as the fundus photographs and optical coherence tomography that are not able to define the fixed structural information, where our method may not obtain efficient performance in the UDA task in such a situation. Also, in some cases, due to the ultrasound image is in a very low-quality format, which may limit its performance. For example, the cardiac view scanned under Doppler mode contains blood (see Figure 5 row 2), which may obscure part of the structures, leading to the failure of detection.

Algorithm 1 Topology Knowledge Transfer (TKT)

Output: $\mathcal{L}_{TKT}$: The overall loss of TKT;

Input: $f_{s,t}$: Feature maps from source and target domains;
y_s^b: The ground truth bounding boxes of the source domain;
y_t^b: The pseudo bounding boxes of the target domain;
N: The total classes number of detection organ;

(1). For Source Domain

for each $i \in [1, N]$ **do**
  $c_i \leftarrow$ The organ centroid feature obtained from f_s and y_s^b.
  Build memory banks: $\theta_i \leftarrow c_i$.
end for
Build visual graph: $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{c_i\}_{i=1}^N$, and $\mathcal{E} = \{c_i \cdot \theta_i^{\mathsf{T}}\}_{i=1}^N$.
Build topological representation graph: $\mathcal{G}_s \leftarrow \text{GNN}(\mathcal{V}, \mathcal{E})$, GNN is the graph neural network.

(2). For Target Domain

for each $i \in [1, N]$ **do**
  $c_i \leftarrow$ The organ centroid feature obtained from f_t and y_t^b.
  $c_i \leftarrow \theta_i$ IF c_i is NULL.
end for
$\mathcal{G}_t \leftarrow \text{GNN}(\mathcal{V}, \mathcal{E})$.
$\mathcal{L}_{dis}\left(\mathcal{G}_s, \mathcal{G}_t\right) = \inf_{\pi} \left(\sum_{i=1}^N ||\mathcal{G}_{t,i} - \mathcal{G}_{s,\pi(i)}||^p\right)^{\frac{1}{p}}$, where π are all permutations of N organs.

Overall Loss of TKT

$\mathcal{L}_{TKT} = \mathcal{L}_{dis}\left(\mathcal{G}_s, \mathcal{G}_t\right)$.

Algorithm 2 Morphology Knowledge Transfer (MKT)

Output: $\mathcal{L}_{MKT}$: The overall loss of MKT;
$\mathcal{L}_{supervised}$: The supervised loss for training the detection head;
$\mathcal{L}_{cls}$: Cross-entropy loss of the classification;
$\mathcal{L}_{bbox}$: Regression loss of bounding box;
$\mathcal{L}_{class}$: The classification loss of the MKT;
$\mathcal{L}_{dis}$: The discrete form of distribution discrepancy;
$f_{s,t}$: Feature maps of source and target domains;

Input: $X_{s,t}$: Input images from source and target domain;
$y_s^{b,c}$: The bounding boxes and organ class ground truth annotations from the source domain;
E: Feature extractor;
D: Detector head;
N: The total classes number of detection organ;

Get the feature maps: $f_{s,t} \leftarrow E(X_{s,t})$.
Get the predict detection result: $y_{s,t}^b, y_{s,t}^c \leftarrow D(f_{s,t})$.
$\mathcal{L}_{supervised} \leftarrow \mathcal{L}_{cls}(y_s^c) + \mathcal{L}_{bbox}(y_s^b)$.

(1). Nodes Sampling

for each $i \in [1, N]$ **do**
  $f_{s,t}^i \leftarrow$ Get different substructure feature maps $\{f_{s/t,k}^i\}_{k=1}^K$ from feature maps $f_{s,t}$ according to $y_{s,t}^b$, $y_{s,t}^c$.
  Collect all feature nodes of i-th substructure form $f\{f_{s/t,k}^i\}_{k=1}^K$.
  Concatenate feature nodes of i-substructure from feature maps $\{f_{s/t,k}^i\}_{k=1}^K$, from layer k to 1.
  $\{o_{i,j}\}_{j=1}^M \leftarrow$ Sample M feature nodes $\{o_{i,j}\}_{i,j=1}^{N,M}$ for i-th substructure with the sample rate equal to the number of all collected feature nodes divided by M, and prioritize sample nodes close to k-th layer.
end for

(2). Morphological Representation

Get the morphological representation: $\mathbf{g}_i \leftarrow$ MAGNN$(\{\mathcal{O}_i\})$, where $\{\mathcal{O}_i\} \leftarrow$ Concatenate$(\{o_{i,j}\}_{i,j=1}^{N,M})$.
Get the cross-domain interaction of morphological representation: $\mathbf{g}_{i,\pi(j)} \leftarrow GA(\mathbf{g}_i)$, GA is graphical attention.

(3). Node Classification

$\mathcal{L}_{class} = -\left(\sum_i^N \sum_j^M y_i^c log\left(\mathbf{g}_{i,\pi(j)}\right)\right)_{s/t}$.

(4). Numerical Equilibrium

$\mathcal{M}^* \leftarrow argmin_{\mathcal{M}} \sum_{i=1}^N \sum_{j=1}^M \sum_{k=0}^K \mathcal{D}(\mathcal{M}[p(o_{t,i,j})], p(o_{s,i,j}))$, where $p(k)$ is the probability density function.
$\mathcal{L}_{dis}\left(\mathbf{g}_s, \mathbf{g}_t\right) = \sum_{i=1}^N \inf_{\pi} \left(\sum_{j=1}^M ||\mathbf{g}_{t,i,j} - \mathcal{M}^*(\mathbf{g}_{s,i,\pi(j)})||^p\right)^{\frac{1}{p}}$.

Overall Loss of MTK

$\mathcal{L}_{MKT} = \mathcal{L}_{class} + \mathcal{L}_{dis}\left(\mathbf{g}_s, \mathbf{g}_t\right)$.