Journal Pre-proof

Towards bridging the distribution gap: Instance to Prototype Earth Mover's Distance for distribution alignment

Qin Zhou, Runze Wang, Guodong Zeng, Heng Fan, Guoyan Zheng

PII: S1361-8415(22)00236-5
DOI: https://doi.org/10.1016/j.media.2022.102607
Reference: MEDIMA 102607

To appear in: *Medical Image Analysis*

Received date : 14 December 2021
Revised date : 28 June 2022
Accepted date : 25 August 2022

Please cite this article as: Q. Zhou, R. Wang, G. Zeng et al., Towards bridging the distribution gap: Instance to Prototype Earth Mover's Distance for distribution alignment. *Medical Image Analysis* (2022), doi: https://doi.org/10.1016/j.media.2022.102607.

Towards Bridging the Distribution Gap: Instance to Prototype Earth Mover's Distance for Distribution Alignment

Qin Zhou[a], Runze Wang[a], Guodong Zeng[b], Heng Fan[c], Guoyan Zheng[a,*]

[a]*Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, No.800 Dongchuan Road, 200240 Shanghai, China*
[b]*sitem Center for Translational Medicine and Biomedical Entrepreneurship, University of Bern, Bern, Switzerland*
[c]*Department of Computer Science and Engineering, University of North Texas, Texas, USA*

Abstract

Despite remarkable success of deep learning, distribution divergence remains a challenge that hinders the performance of many tasks in medical image analysis. Large distribution gap may deteriorate the knowledge transfer across different domains or feature subspaces. To achieve better distribution alignment, we propose a novel module named Instance to Prototype Earth Mover's Distance (I2PEMD), where shared class-specific prototypes are progressively learned to narrow the distribution gap across different domains or feature subspaces, and Earth Mover's Distance (EMD) is calculated to take into consideration the cross-class relationships during embedding alignment. We validate the effectiveness of the proposed I2PEMD on two different tasks: multi-modal medical image segmentation and semi-supervised classification. Specifically, in multi-modal medical image segmentation, I2PEMD is explicitly utilized as a distribution alignment regularization term to supervise the model training process, while in semi-supervised classification, I2PEMD works as an alignment measure to sort and cherry-pick the unlabeled data for more accurate and robust pseudo-labeling. Results from comprehensive experiments demonstrate the efficacy of the present method.

Key words: Distribution Alignment, Instance to Prototype Matching,

*Corresponding author. Email: guoyan.zheng@sjtu.edu.cn (Guoyan Zheng)

Earth Mover's Distance, Unpaired Multi-modal Segmentation,
Semi-supervised Classification

1. Introduction

In medical image analysis, anatomical structures are often imaged with a variety of modalities. Images from different modalities can capture complementary information for disease diagnosis and treatment. Therefore it is important to jointly utilize the cross modality information for better assessment of diseases. However, different imaging mechanisms result in great visual differences, imposing huge feature distribution divergence across different modalities. In some cases, even if the image data is collected from the same or similar distribution, the learned features may be biased towards specific feature subspaces, due to the sampling bias or over-fitting problem (Wang et al., 2019).

To address the above-mentioned issues, distribution alignment across different domains (e.g., cross-modality) or different feature subspaces (e.g., labeled and unlabeled data collected from the same or similar distributions in semi-supervised learning), has drawn growing attention recently. In order to bridge the gap between different modalities, early and late fusion strategies are typically utilized. In early fusion-based methods, inputs from different modalities are concatenated along the color channels before being fed into the network (Pereira et al., 2016; Isensee et al., 2017; Wang et al., 2017; Zhao et al., 2018). As for late-fusion, paired inputs from different modalities are received by separate networks to extract modality-specific features. The extracted features are then fused at the semantic level to generate the final results (Dolz et al., 2018b; Chen et al., 2018; Dolz et al., 2018a). To mitigate the distribution gap across different feature subspaces, various techniques including adversarial training (Li et al., 2020; Dong and Lin, 2019), consistency regularization (Berthelot et al., 2019) and graph-based label propagation (Zhang et al., 2020; Iscen et al., 2019), are proposed.

From a new perspective of instance-to-prototype matching, in this paper, we address the distribution alignment problem by proposing a novel Instance-to-Prototype Earth Mover's Distance (I2PEMD). Specifically, I2PEMD progressively learns shared class-specific prototypes for different modalities (or feature subspaces), and calculates the Earth Mover's Distance (EMD) (Hou et al., 2016) to measure the instance-to-prototype matching degree for loss

minimization or cherry-picking pseudo-labeled samples in downstream tasks. In addition, in our proposed I2PEMD, the important ground distance matrix for measuring cross-class relationships is dynamically updated by the learned prototypes, which can better adapt to the learned feature embedding than a fixed prior.

Unlike previous studies, the core of our proposed I2PEMD lies in shared prototype learning across different modalities (or feature subspaces) and instance-to-prototype EMD estimation. By explicitly learning shared class-specific prototypes, we can pull the high-level features belonging to the same class closer, mitigating the distribution divergence across different modalities. Besides, by carefully considering the cross-class relationships, I2PEMD leads to more robust matching mechanism for distribution alignment.

Our I2PEMD is a flexible module and ready to be plugged in many existing frameworks for handling the distribution alignment problem. To demonstrate its effectiveness, we apply I2PEMD to two different tasks, i.e., unpaired multi-modal image segmentation and semi-supervised classification. Extensive experimental results demonstrate that our I2PEMD matching mechanism is able to effectively alleviate the distribution alignment problem and improve the performance of downstream tasks.

The overall contributions of the proposed I2PEMD are summarized as follows:

- We propose to address the distribution alignment problem from a new perspective of instance-to-prototype matching. This mechanism can be readily plugged into many different frameworks that require distribution alignment during deep feature representation learning.
- We propose to combine shared prototype learning with EMD estimation to take into consideration of both intra-class compactness and cross-class relationships during distribution alignment.
- We conduct comprehensive experiments to evaluate the effectivenss of the proposed I2PEMD on both unpaired cross-modality segmentation and semi-supervised classification tasks, generating superior performance compared with state-of-the-art methods.

2. Related Works

Our work is closely related to the field of distribution alignment as well as methods concerning multi-modal image segmentation and semi-supervised

classification. We will briefly review related literature respectively in the following sections.

2.1. Distribution Alignment

Distribution alignment plays a vital role in many computer vision and image analysis tasks. Although there is no official categorization of distribution alignment methods, we divide them into discrepancy minimization-based ones and subspace learning-based ones. Discrepancy minimization-based distribution alignment methods aim to reduce the distribution divergence between domains by minimizing a specific metric (Long et al., 2015, 2017; Zellinger et al., 2017; Li et al., 2021), while subspace learning-based ones seek to learn domain-invariant feature representations or intermediate subspaces by adversarial training (Ganin and Lempitsky, 2015; Long et al., 2018; Wang et al., 2019) or metric learning (Fernando et al., 2013; Gong et al., 2012; Gopalan et al., 2011; Hu et al., 2015; Zhang et al., 2017). Our proposed distribution alignment shares a similar spirit with the metric learning-based methods. However, instead of learning linear projection mapping functions (Fernando et al., 2013; Zhang et al., 2017), we try to learn shared prototypes for the same semantic class across different domains. Besides, by introducing the EMD for measuring feature representation distance, we further consider the cross-class relationships, leading to more robust distribution alignment.

2.2. Multi-modal Image Segmentation

Multi-modal images (including CT, MRI, PET et al.) can provide complementary information in performing medical image diagnosis (Bhatnagar et al., 2015). Classic methods require paired and registered multi-modal inputs, where images of different modalities belong to the same patient. Representative works include earlier fusion-based (Pereira et al., 2016; Isensee et al., 2017; Wang et al., 2017; Zhao et al., 2018; Myronenko, 2018) and the later fusion-based ones (Dolz et al., 2018b; Chen et al., 2018; Dolz et al., 2018a). The earlier fusion-based methods integrate multi-modality images channel by channel as the multi-channel inputs to learn a fused feature representation (Zhou et al., 2019). In the later fusion-based methods, each modality has modality-specific layers at an early stage of a Convolutional Neural Networks (CNN). The features extracted from different modalities are fused at a certain middle layer of the CNN, forming a Y-shaped architecture (Dou et al., 2020). However, as pointed out in (Chen et al., 2020),

supervised feature learning is often modality dependent. Besides, obtaining paired and spatially well-aligned multi-modal images is itself a costly task and often infeasible. Therefore, it is of demanding importance to design multi-modal image segmentation methods for unpaired inputs (Valindria et al., 2018; Zhang et al., 2018; Huo et al., 2018; Dou et al., 2020; Chen et al., 2020). In (Dou et al., 2020), independent normalization statistics and knowledge distillation from high-level CNN representations are introduced to bridge the cross-modality divergence. Chen et al. (2020) further propose the class-specific affinity matrix to enhance the cross-modality generalization.

2.3. Semi-supervised Classification

Semi-supervised classification (SSL) aims to train a classifier from a small amount of labeled data and a large amount of unlabeled data, such that it outperforms supervised classifier trained only on the small amount of labeled data. The basic assumptions in SSL are the smoothness assumption and the low-density assumption. The smoothness assumption states that if two or more data points are close in the sample space, they should belong to the same class. Similarly, the low-density assumption states that the decision boundary for a classification model should not pass the high-density region of sample space. Based on above assumptions, the semi-supervised learning can be categorized into consistency regularization-based ones (including mean teacher (Tarvainen and Valpola, 2017), temporal ensembling (Samuli and Timo, 2017), unsupervised data augmentation (Xie et al., 2020), et al.) and proxy-label-based ones (including self-training). Self-training favors low-density separation by using model's own predictions as pseudo-labels. Pseudo-labeling (Lee, 2013) picks the most confident predictions as hard (one-hot) pseudo-labels. MixMatch (Berthelot et al., 2019) uses the average of predictions on the image under multiple augmentations as the soft pseudo-label. FixMatch (Sohn et al., 2020) finds an effective combination of image augmentation techniques and pseudo-labeling. We follow the same semi-supervised learning strategy as done in FixMatch (Sohn et al., 2020) which combines the pseudo-labeling-based self-training strategy with the consistency constraints on weak and strong augmented predictions. Different from FixMatch (Sohn et al., 2020), we propose a novel I2PEMD-guided (weak vs. strong) alignment process. By sharing prototype learning and by considering the cross-class relationship priors, our framework can mitigate the distribution shift between labeled and unlabeled data, resulting in more

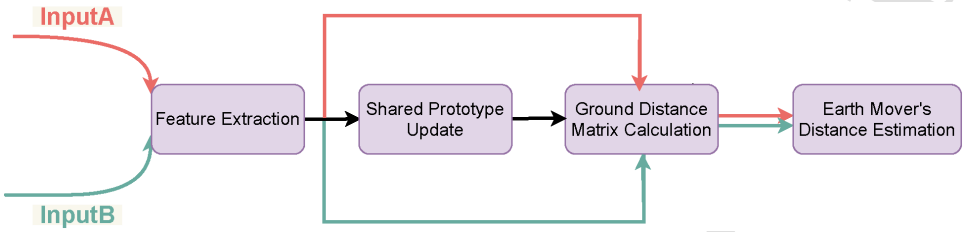


Figure 1: Illustration of the process of our proposed instance-to-prototype Earth Mover's Distance (I2PEMD). Please note that the red and green arrows denote the forward data flows of inputs from different domains, and black arrows indicate common operations.

accurate pseudo-labels for subsequent supervision, thus alleviating the error accumulation problem of self-training methods.

3. Method

In this section, we elaborate on the details of the proposed I2PEMD and its applications to unpaired multi-modal segmentation and to semi-supervised classification tasks as well.

3.1. Instance to Prototype Earth Mover's Distance (I2PEMD)

In this subsection, we will present details about the proposed I2PEMD. Figure 1 illustrates the overall framework of our approach. Specifically, inputs from different domains (or feature subspaces) are fed into the backbone CNN to extract feature embeddings. Then momentum update is introduced to estimate shared prototypes for each class. Finally, EMD is calculated based on the learned prototypes. The important ground distance matrix in EMD estimation is designed to reflect the cross-class relationships among the learned prototypes. Details are presented in the following parts.

3.1.1. Revisiting Earth Mover's Distance (EMD)

The EMD is defined as the minimum cost to transport the mass of one distribution (histogram) to the other (Hou et al., 2016). EMD has the formulation of the transportation problem (Hitchcock, 1941) and the global minimum can be achieved by solving a linear programming problem. Specifically, denote $\mathbf{p} = \{p_i | i = 1, 2, \cdots, m\}$ as a set of sources or suppliers, and $\mathbf{q} = \{q_j | j = 1, 2, \cdots, n\}$ as a set of consumers. The items p_i, q_j in each set

refer to the supply and demand units of the i-th supplier and the j-th consumer, respectively. Denote the per-unit cost for transporting from supplier i to consumer j as g_{ij}, and the units transported from supplier i to consumer j as x_{ij}. The goal of the transportation optimization problem is to find the optimal transportation flow $\mathcal{X} = \{x_{ij} | i = 1, 2, \cdots, m, j = 1, 2, \cdots, n\}$ such that the following objective is minimized,

$$\begin{aligned} & \sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij}x_{ij} \\ s.t. \quad & x_{ij} \geq 0, i = 1, \cdots, m, j = 1, \cdots, n \\ & \sum_{j=1}^{n} x_{ij} \leq p_i, i = 1, \cdots, m \\ & \sum_{i=1}^{m} x_{ij} \leq q_j, j = 1, \cdots, n \end{aligned} \tag{1}$$

In Eq. (1), the non-negative condition constrains that the amount of mass transported must be positive. The second condition guarantees that the amount of mass transported from a supplier must not exceed its total mass p_i. And the last condition ensures that the amount of mass transported to a consumer must not exceed its total demand q_j. The global optimal matching flow $\mathcal{X}$ can be achieved by solving the above linear programming problem. And the EMD between $\mathbf{p}$ and $\mathbf{q}$ can be calculated as the minimum transportation cost in Eq. (1), normalized by the total flow,

$$\mathrm{EMD}(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij}x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \tag{2}$$

3.1.2. Ground Distance Matrix Estimation

The ground distance matrix $\mathbf{G} = \{g_{ij}, i = 1, \cdots, m, j = 1, \cdots, n\}$ indicates the per-unit transportation cost from the i-th class in $\mathbf{p}$ to the j-th class in $\mathbf{q}$. It is an important prior in modeling the cross-class relationships. However, In many medical image analysis tasks (e.g., multi-class classification/segmentation), the prior cross-class relationships can not be accessed. To tackle this problem, we resort to the feature distances among the class-specific prototypes for measuring the cross-class relationship priors. Specifically, class-specific prototypes are learned to represent the center of each class in the feature embedding space, then the feature distances

between different prototypes are calculated to generate the ground distance matrix. In our method, the supplier $\mathbf{p} = \{p_i, i = 1, \cdots, m\}$ and the consumer $\mathbf{q} = \{q_j, j = 1, \cdots, n\}$ refer to the logits for classification or segmentation from different modalities (or feature subspaces), They share the same class sets (i.e., m = n, equals to the number of semantic classes). To perform distribution alignment, we further constrain inputs from different modalities or feature subspaces to share the same prototype for each class.

Denote the set of prototype features as $\mathcal{C} = \{\mathbf{c}_i \in \mathrm{R}^d, i = 1, \cdots, m\}$, where $\mathbf{c}_i$ is the prototype for the i-th class, d is the dimension of the feature embedding. Then the matching cost g_{ij} is calculated as:

$$g_{ij} = ||\mathbf{c}_i - \mathbf{c}_j||_1 \tag{3}$$

where $||\cdot||_1$ denotes the L_1 norm. The ground distance matrix is designed under the assumption that classes with similar visual patterns (small prototype distance) are supposed to have small transportation cost. Please note that we adopt the momentum trick for robust update of the prototypes. And for different downstream tasks, the way to update the class-specific prototypes may be slightly different. Details for progressively updating the prototype features are presented in Sec. 3.2 and Sec. 3.3, respectively.

In practice, before convergence, the learned prototype features may not sufficiently separate different classes. To address this issue, we follow (Hou et al., 2016) to map each row of $\mathbf{G}$ onto uniformly distributed values: each entry is mapped to its percentile value in its row. Denote the transformed matrix as $\overline{\mathbf{G}}$, then the i, j-th element $\overline{g_{ij}}$ is calculated as,

$$\overline{g_{ij}} = \frac{1}{m}\mathrm{R}(g_{ij}, \{g_{i1}, \cdots, g_{im}\}) \tag{4}$$

where $\mathrm{R}(g_{ij}, \{g_{i1}, \cdots, g_{im}\})$ returns the number of elements in the set $\{g_{i1}, \cdots, g_{im}\}$ that is smaller than g_{ij}. In this case, all entries of the transformed matrix $\overline{\mathbf{G}}$ is mapped to range $[0, 1)$. The final ground distance matrix $\mathbf{G}$ is a symmetric matrix obtained as,

$$\mathbf{G} = (\overline{\mathbf{G}} + \overline{\mathbf{G}}^{\mathrm{T}})/2 \tag{5}$$

As the diagonal entries indicate the intra-class matching cost, $G_{ii} = 0$ for all i.

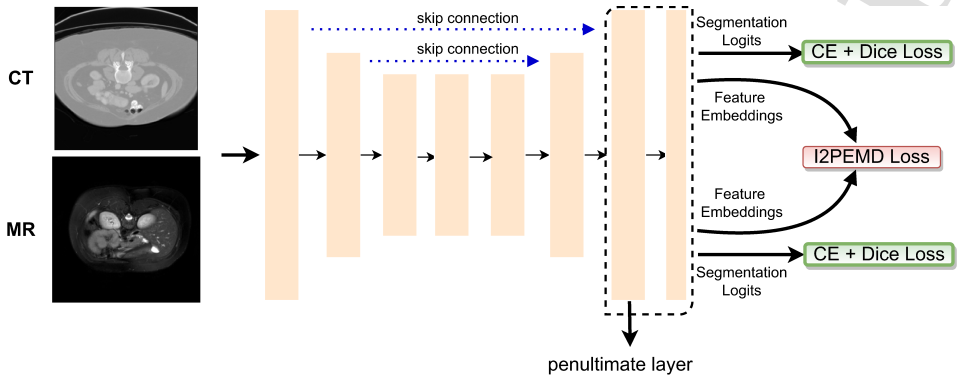


Figure 2: Illustration of the overall framework of the I2PEMD-regularized, unpaired multi-modal segmentation. Layers in the dotted bounding box demonstrate the output layers, where the feature embeddings are extracted from the penultimate layer, and the segmentation logits are generated from the last layer. Please note "CE" refers to the cross entropy loss in the framework.

3.1.3. I2PEMD Estimation

Here we will present how to perform I2PEMD estimation. Specifically, in multi-modal segmentation or semi-supervised classification, the supplier's mass $\mathbf{p}$ refers to the softmax logits of a certain instance (a pixel or an image), while the consumer's mass $\mathbf{q}$ is a binary vector where only the index of the ground truth class k equals to 1: $q_k = 1$. According to the constraints in Eq. (1), all mass in $\mathbf{p}$ must be transported to q_k. Thus the optimal transformation matrix $\mathcal{X}$ satisfies $x_{ij} = 0$ if $j \neq k$, otherwise $x_{ij} = p_i$. According to Eq. (2), the Instance-to-Prototype EMD (I2PEMD) is calculated as,

$$\begin{aligned} \text{I2PEMD}(\mathbf{p}, \mathbf{q}) =& \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} g_{ij}x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \\ =& \frac{\sum_{i=1}^{m} p_i g_{ik}}{\sum_{i=1}^{m} p_i} = \sum_{i=1}^{m} p_i g_{ik} \end{aligned} \tag{6}$$

where g_{ik} represents the i-th item in the k-th column of the ground distance matrix $\mathbf{G}$. Please note here $\sum_{i=1}^{m} p_i = 1$. The I2PEMD can then be utilized directly as a regularization term to supervise the training process of downstream tasks.

3.2. I2PEMD for Unpaired Multi-modal Segmentation

3.2.1. Overall Framework

Unpaired multi-modal segmentation plays a vital role in jointly utilizing multi-source medical images for disease diagnosis. It has the following merits: 1) it does not require the multi-source inputs to be paired and well-aligned, making it flexible to take advantage of images collected from different patients; 2) it can handle multi-source inputs with large visual divergence (e.g., CT and MRI), transferring knowledge across different domains. To address the large distribution gap across different imaging modalities, in this subsection, we will demonstrate how the I2PEMD-based distribution alignment can benefit the task of unpaired multi-modal segmentation.

Figure 2 illustrates the overall diagram of the I2PEMD-regularized, unpaired multi-modal segmentation. Firstly, unpaired multi-modal inputs $\mathbf{I} \in \mathrm{R}^{H\times W\times 3}$ are fed into the backbone network to generate high-level features $\mathbf{F} \in \mathrm{R}^{H\times W\times d}$ and segmentation logits $\mathbf{A} \in \mathrm{R}^{H\times W\times K}$, where d is the dimension of the feature embedding, K is the number of semantic classes (i.e., K equals to m, n in Eq. (1)). Then classic segmentation losses including multi-class cross entropy loss and Dice loss are imposed on the logits. To enhance the distribution alignment from different modalities, we further introduce the I2PEMD regularization term to make multi-modal features benefit from shared prototype learning. The complete objective function for unpaired multi-modal segmentation is thus formulated as,

$$\mathcal{L}_{Seg} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} + \lambda \mathcal{L}_{\mathrm{I2PEMD}} \tag{7}$$

where λ is the tradeoff parameter to balance between segmentation loss and the distribution alignment regularization term. $\mathcal{L}_{\mathrm{I2PEMD}}$ is the average I2PEMD calculated on all the semantic classes as,

$$\mathcal{L}_{\mathrm{I2PEMD}} = \frac{1}{K}\sum_{k=1}^{K} \ell^{k}_{\mathrm{I2PEMD}} \tag{8}$$

where $\ell^{k}_{\mathrm{I2PEMD}}$ is the mean instance-to-prototype distance of the k-th class,

$$\ell^{k}_{\mathrm{I2PEMD}} = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathrm{I2PEMD}(\mathbf{a}^{\mathrm{i}}_{\mathrm{k}}, \mathbf{e}_{\mathrm{k}}) \tag{9}$$

where I2PEMD$(\cdot,\cdot)$ denotes the operation in Eq. (6). $\mathbf{a}_i^k \in \mathbf{A}$ represents the segmentation logit vector corresponding to the i-th pixel belonging to class k, and $\mathbf{e}_k$ is the one-hot vector where the k-th element equals to 1, otherwise 0. For I2PEMD estimation, the class-specific prototypes are progressively updated as described in the following.

3.2.2. Momentum Prototype Update

Prototype features can be viewed as cluster centroids of each semantic class in the feature embedding space. Denote the segmentation mask corresponding to class k as $\mathrm{M}_k \in \mathrm{R}^{H\times W}$, where $\mathrm{M}_k^i = 1$ if pixel i belongs to class k. Then the prototype feature $\mathbf{c}_k$ of class k is updated as,

$$\mathbf{c}_k^t = \alpha * \mathbf{c}_k^{t-1} + (1-\alpha) * \mathbf{c}_k^{\text{update}} \tag{10}$$

where $\mathbf{c}_k^{\text{update}}$ is updated in each iteration as,

$$\mathbf{c}_k^{\text{update}} = \frac{1}{N_k}\sum_i \mathrm{M}_k^i \cdot \mathbf{F}_k^i \tag{11}$$

where N_k represents the number of pixels belonging to class k, $\mathbf{F}_k^i$ is the i-th feature vector of the k-th class and α is the momentum update coefficient, which is empirically set to 0.8 throughout our experiments.

3.3. I2PEMD for Semi-supervised Classification

In medical image analysis, it is usually laborious to get large amount of labeled data due to the requirement of expertise. Therefore, semi-supervised learning has emerged as a more and more important topic recently (Chartsias et al., 2018; Nie et al., 2018; Aviles-Rivero et al., 2019; Dong et al., 2018). In semi-supervised learning, we are equipped with a few labeled images for each class, as well as large amount of unlabeled ones. Pseudo-labeling is frequently utilized to augment the labeled images with confident unlabeled samples. However, the sampling bias problem leads to large distribution gap between labeled and unlabeled samples, undermining the accuracy of the generated pseudo-labels.

To address this issue, we introduce the proposed I2PEMD as a supplement measure for choosing accurate pseudo-labeled samples. In our proposed I2PEMD, both the shared prototype learning and prototype-based EMD estimation help to better align the feature distribution between labeled and selected pseudo-labeled samples, leading to more robust and generalized feature learning.

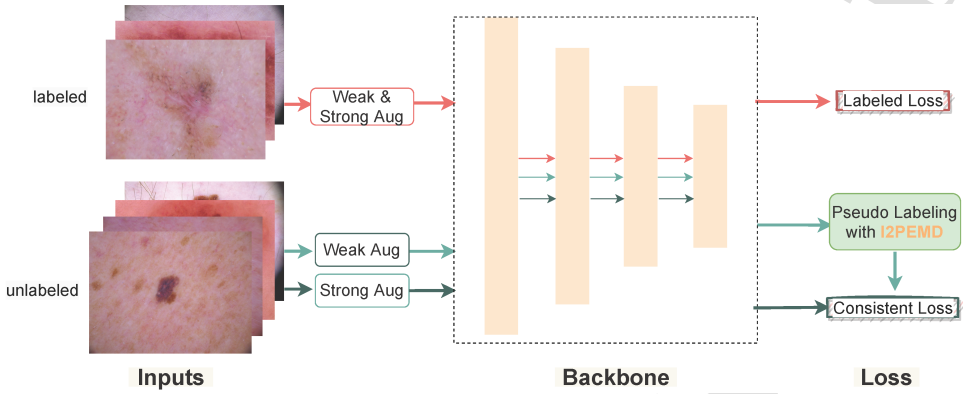


Figure 3: Illustration of the overall framework of our I2PEMD-guided semi-supervised classification, where "Weak Aug" and "Strong Aug" refer to the weak and strong augmentations; Consistent loss denotes the alignment loss between the weak and strong augmented predictions defined in Eq. (13).

3.3.1. Overall Framework

Figure 3 illustrates the overall framework of our I2PEMD-guided semi-supervised classification. Specifically, during each training iteration, both labeled and unlabeled samples are fed into the backbone network to extract features (Similar to the segmentation task, features are extracted from the penultimate layer before the final classifier), then each unlabeled sample is assigned with a pseudo-label. Finally, the selected pseudo-labeled samples are combined together with the labeled samples to update the network. In semi-supervised classification, better pseudo-labeling strategy plays a vital role in improving the overall performance. To validate the effectiveness of our proposed I2PEMD in better sample selection, we adopt the augmentation anchoring (Berthelot et al., 2020; Sohn et al., 2020) strategy in our base model, which is widely utilized in state-of-the-art semi-supervised learning algorithms. For completion, before delving into the details of cherry-picking unlabeled samples with I2PEMD, we first give brief introduction to the base method.

3.3.2. Introduction to the Base Model

In our base model, apart from the supervision from ground truth labeled data, we adopt the augmentation anchoring strategy to align strongly augmented samples to pseudo-labels generated from corresponding weakly aug-

mented samples. Specifically, we adopt random cropping followed by affine transformations and a random horizontal flip as weak augmentations. We use RandAugment (Cubuk et al., 2020) that contains difficult transformations (e.g., color jittering) to generate strong augmentations.

Denote the labeled set as $S_L = \{(x_i, y_i)\}_{i=1}^{N}$ and the unlabeled set as $S_U = \{x_i\}_{i=N+1}^{N+M}$, where x_i is the 2D medical image, y_i is the one-hot ground-truth label. The optimization objective of the whole framework can be formulated as following:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_s(f(x_i|\{\eta, \eta^{'}\}; \theta), y_i) + \lambda \min_{\theta} \sum_{i=N+1}^{N+M} \mathcal{L}_u(f(x_i|\{\eta, \eta^{'}\}; \theta)) \tag{12}$$

where $\mathcal{L}_s$ denotes the supervised loss (i.e., cross-entropy loss) imposed on the labeled data; $\mathcal{L}_u$ represents the consistency loss calculated from the weak/strong augmented unlabeled samples.

In Eq. (12), $f(\cdot)$ refers to the classification network; θ are the parameter weights of the model; η and $\eta^{'}$ represent the weak and strong augmentations applied to the input; λ is a trade-off hyper-parameter to balance between the supervised and unsupervised term.

In our base model, the output logits of strongly augmented samples are aligned to the pseudo-labels generated from the weakly augmented samples. Denote y_i^p as the pseudo one-hot label generated from the weakly augmented version of x_i, then the alignment loss is calculated as,

$$\sum_{i=N+1}^{N+M} \mathcal{L}_u(f(x_i|\eta^{'}; \theta), y_i^p), \tag{13}$$

where $\mathcal{L}_u$ denotes the alignment loss (cross entropy in our method) imposed on the unlabeled data. The overall objective function to be optimized is thus formulated as,

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u, \tag{14}$$

where λ is the trade-off parameter.

3.3.3. Sample Selection with I2PEMD

In the formulation of Eq. (13), the alignment loss is imposed on all the unlabeled samples. However, due to the sampling bias problem, there may

exist large distribution gap between the labeled and unlabeled data. Consequently, the predicted pseudo-labels on weakly augmented unlabeled data are highly likely to be different from the ground truth labels, deteriorating the performance of the learned model. Therefore, it is of key importance to select accurate pseudo-labeled samples in the alignment loss. Since the weak generalization ability to the unlabeled samples is mainly caused by the feature distribution gap between the labeled and unlabeled data, it is important to mitigate the gap, as well as to pick out the truly confident pseudo-labeled data. To this end, we introduce our proposed I2PEMD as a measure to cherry-pick the unlabeled samples. The samples chosen by our I2PEMD have the following strengths:

- The learned prototypes help to narrow down the feature distribution gap between feature subspaces of the labeled and unlabeled data;
- By taking into consideration the cross-class relationships, prototype-based EMD distance allows to cherry-pick truly confident pseudo-labeled samples (which are close to the corresponding class-specific center and faraway from other centers) to augment the supervised training data.

Denote $\mathrm{o}_i \in \mathrm{R}^K, i \in \{N+1, \cdots, N+M\}$ as the logits after softmax of the unlabeled data, then the classification confidence score and the corresponding pseudo-labels can be obtained as,

$$\begin{aligned} s_i &= max(\mathrm{o_i}), \\ r_i &= argmax(\mathrm{o_i}), \end{aligned} \tag{15}$$

In the following, we detail how to exploit our proposed I2PEMD for cherry-picking accurate pseudo-labeled samples. Firstly, we follow the literature (Zhang et al., 2021; Sohn et al., 2020) to filter out less-confident samples with small classification scores. Denote the selected index set as N_s^1, then N_s^1 can be calculated as $N_s^1 = \{i\}, s_i > \tau^{low}$, where τ^{low} is the pre-defined classification score threshold. Denote $\mathrm{l_i}$ as the pseudo one-hot vector of $\mathrm{o_i}$ where only the r_i-th element equals to 1. In our method, we further constrain the selected unlabeled samples to have top T ranked I2PEMD distances compared with the learned prototypes. The index set selected by the I2PEMD distance can be expressed as,

$$N_s^2 = top^T(\mathrm{I2PEMD}(\mathrm{o}_i, \mathrm{l}_i)), i \in N+1, \cdots, N+M, \tag{16}$$

where $top^T(\cdot)$ returns the index set of the top T smallest distances (in our method, T is empirically set to half of the batchsize); I2PEMD$(\cdot,\cdot)$ is calculated according to Eq. (6). Then the final unlabeled set selected by I2PEMD is obtained as,

$$N_s^{low} = N_s^1 \cap N_s^2, \tag{17}$$

At the beginning of the training phase, many pseudo-labeled samples have low classification scores, our I2PEMD-guided instance selection in Eq. (17) can help to find out truly confident pseudo-labeled samples that are well aligned with the feature subspace of the labeled samples. Since we limit the selected number to half of the batchsize in N_s^{low}, this may decrease the diversity of the training samples (especially in the later iterations). To address this issue, we further add the pseudo-labeled samples with higher classification scores in the alignment loss. The final selected unlabeled set is then obtained as,

$$N_s = N_s^{low} \cup N_s^{high} \tag{18}$$

where $N_s^{high} = \{i\}, s_i > \tau^{high}$, and τ^{high} is a larger classification score threshold ($\tau^{high} > \tau^{low}$), $\cup$ represents the union operation. Our overall objective function for semi-supervised classification is then formulated as,

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_s(f(x_i;\theta), y_i) + \lambda \sum_{i \in N_s} \mathcal{L}_u(f(x_i|\eta^{'};\theta), \mathrm{l_i}). \tag{19}$$

3.3.4. Momentum Prototype Update

To better align the labeled and unlabeled feature subspaces, we also include the selected confident unlabeled data during shared prototype learning. Denote the prototype of class k at the t-th iteration as $\mathbf{c}_k^t$, then $\mathbf{c}_k^t$ is progressively updated as,

$$\mathbf{c}_k^t = \alpha * \mathbf{c}_k^{t-1} + (1-\alpha) * \mathbf{c}_k^{\text{update}} \tag{20}$$

where $\mathbf{c}_k^{\text{update}}$ is updated by both the labeled and selected pseudo-labeled data in each iteration,

$$\mathbf{c}_k^{\text{update}} = \frac{1}{|Z_k|}\Big(\sum_{i=1}^{N} f^{'}(x_i|y_i = k) + \sum_{i \in N_s} f^{'}(x_i|r_i = k)\Big) \tag{21}$$

where $f^{'}(\cdot)$ represents the feature extractor of the network (i.e., output of the layer before the linear classifier in our method); $|Z_k|$ is the total number of the samples for updating prototype k.

4. Experiments

In this section, we design and conduct comprehensive experiments to demonstrate effectiveness of the proposed I2PEMD. Specifically, in the task of unpaired multi-modal segmentation, the proposed I2PEMD is utilized to bridge the gap between the CT and MRI domains, mutually benefiting the segmentation performance of both domains. As for the task of semi-supervised classification, I2PEMD acts as a measure to select truly confident samples by taking into consideration the cross-class relationships.

4.1. Unpaired multi-modal image segmentation

In this subsection, firstly, we will introduce the experimental setup (including datasets, network architecture and implementation details). Then we will elaborate on the quantitative and qualitative segmentation results to demonstrate effectiveness of our proposed I2PEMD.

4.1.1. Experimental Setup

Datasets.

We evaluate unpaired multi-modal image segmentation on the task of 3D multi-organ segmentation (including liver, spleen, right kidney (R-kdy) and left kidney (L-kdy)) in abdominal images. Specifically, we perform multi-organ segmentation with 3D CT and MRI volumes of the abdominal images. We utilize public CT dataset of (Landman et al., 2015), with 30 patients (different from (Dou et al., 2020), we adopt the whole dataset without removing the case in low image quality) and all the 20 MRI images from the 2019 International Symposium on Biomedical Imaging (ISBI) Combined Healthy Abdominal Organ Segmentation (CHAOS) Challenge (please note that (Dou et al., 2020) uses only 9 cases). Following the protocol of (Dou et al., 2020), we crop the original CT and MRI images at the areas of multi-organs. The cropped images from both the CT and MRI modality are resampled into a resolution around $1.5 \times 1.5 \times 8.0mm^3$, resulting in a size of $256 \times 256 \times D$, where D is the number of slices after resampling. The 3D images are normalized to zero mean and unit variance for intensities within each modality before being fed into the backbone network. Images of each modality are randomly divided into 70%/30% splits for training and testing, respectively.

Table 1: Comparison results on 3D abdominal multi-organ segmentation with different experimental settings. For fair comparison with the state-of-the-art methods, we reimplemented the method introduced in (Dou et al., 2020) and (Valindria et al., 2018) to train/test on our own dataset split. The best results are marked in bold font.

Methods	Dice Overlap Coefficient (DOC)									
	CT				CT Mean	MRI				MRI Mean	Overall Mean
	Liver	R-kdy	L-kdy	Spleen		Liver	R-kdy	L-kdy	Spleen		
Individual	93.78 ± 0.66	91.34 ± 0.28	88.42 ± 0.36	91.31 ± 0.14	90.8	**91.71** ± 0.65	90.49 ± 0.77	90.97 ± 0.52	86.24 ± 3.1	89.1	90.0
Joint	92.30 ± 0.27	88.65 ± 3.2	88.14 ± 0.50	89.86 ± 1.22	89.7	87.44 ± 0.57	90.28 ± 0.45	91.52 ± 0.47	84.67 ± 0.05	88.5	89.1
Joint+I2PEMD	92.30 ± 0.61	91.94 ± 0.13	89.46 ± 0.11	89.3 ± 1.45	90.8	87.65 ± 1.64	90.42 ± 0.15	90.91 ± 0.28	87.39 ± 0.61	89.1	90.0
Joint+sepBN	93.90 ± 0.05	92.36 ± 0.49	88.48 ± 0.75	91.33 ± 0.37	91.5	91.51 ± 0.58	91.51 ± 0.67	92.58 ± 0.76	83.55 ± 1.29	89.8	90.7
X-shaped (Valindria et al., 2018)	93.11 ± 0.44	90.92 ± 1.29	88.6 ± 0.1	91.16 ± 0.69	91.0	91.35 ± 0.68	91.03 ± 0.22	91.13 ± 0.60	81.17 ± 1.64	88.7	89.9
Y-shaped (Valindria et al., 2018)	94.06 ± 0.34	91.29 ± 0.23	88.36 ± 1.36	91.21 ± 0.32	91.0	91.58 ± 0.05	91.21 ± 0.32	91.06 ± 0.67	84.04 ± 1.21	89.5	90.3
UMMKD (Dou et al., 2020)	93.2 ± 0.23	91.95 ± 0.65	88.93 ± 0.62	91.18 ± 1.27	91.3	90.37 ± 0.12	92.25 ± 0.31	92.70 ± 0.08	86.79 ± 3.32	90.5	90.9
Joint+sepBN+I2PEMD$_{fix}$	94.03 ± 0.32	92.62 ± 0.54	88.44 ± 0.86	90.23 ± 0.66	91.3	91.45 ± 0.46	92.12 ± 0.39	**92.85** ± 0.52	86.79 ± 1.87	90.8	91.1
Joint+sepBN+I2PEMD (Ours)	**94.15** ± 0.24	**92.62** ± 0.27	**89.78** ± 0.17	**92.31** ± 0.17	**92.2**	91.31 ± 0.23	**92.26** ± 0.18	92.77 ± 0.23	**88.66** ± 0.82	**91.3**	**91.8**

Network Architecture.

To evaluate the effectiveness of the proposed I2PEMD, we conduct comprehensive experiments on 3D multi-modal image segmentation. For fair comparison with the existing state-of-the-art methods, we adopt the same backbone as in (Dou et al., 2020). Specifically, we adopt the 3D U-Net (Çiçek et al., 2016) as the backbone model in our framework. The 3D U-Net contains a few 3D Conv-BN-Relu groups followed by Max Pooling. Then 3D DeConv-BN-Relu groups are adopted to recover to the original input size. For better segmentation results, skip connections are introduced in the 3D U-Net architecture. For detailed network architecture, please refer to (Çiçek et al., 2016; Dou et al., 2020).

Implementation Details.

During the training process, the batchsize is set to 6. The learning rate is initialized to $5e-4$ and decayed by 5% per 500 iterations. All the experiments are carried out on a NVIDIA Tesla V100 GPU with 32G memory. For fair comparison, we report the mean and the standard deviation (SD) over three runs for each experimental setting.

4.1.2. Comparison with State-of-the-art (SOTA) Methods

We evaluate the segmentation performance with the metric Dice Overlap Coefficient (DOC), and compare with UMMKD (Dou et al., 2020), X-shaped and Y-shaped architectures (Valindria et al., 2018). We also implement the individual training and joint training settings to demonstrate that by explicitly aligning the feature distributions between CT and MRI domains, I2PEMD can indeed boost the joint training performance. We record the mean DOC over each modality as well as over two modalities for a straightforward comparison.

Table 1 lists results of multi-organ segmentation using 3D models. In the first group, "Individual" denotes the setting where CT and MRI domains do not share network parameters; the "Joint" setting means that CT and MRI domains share the same network parameters, and "Joint+I2PEMD" represents the "Joint" setting supervised by our I2PEMD (apart from the segmentation loss). As shown in Table 1, the "Joint" setting performs poorly than the "Individual" setting, which can be attributed to the reason that large visual difference makes it inferior to directly sharing parameters across the CT and MRI domains. By introducing the proposed I2PEMD, the average performance can be boosted (+0.9% in terms of average DOC). In the second group, the "Joint+sepBN" baseline model indicates the parameters of batch normalization layers are separately learned for different domains, while the remaining parameters are shared across different domains. As shown in Table 1, with the "Joint+sepBN" baseline model, the performance can be further improved (from 89.1% to 90.7% in terms of average DOC). The method introduced in (Dou et al., 2020) exceeds the "Joint+sepBN" baseline by 0.2% in terms of average DOC. Our framework achieves the best results (+1.1% in terms of average DOC when compared to the "Joint+sepBN" baseline). We also report the performance of our method with entries of the ground distance matrix fixed to one (i.e., the "Joint+sepBN+I2PEMD$_{fix}$" setting). As shown, Our method outperforms the fixed ground distance matrix setting (i.e., the "Joint+sepBN+I2PEMD$_{fix}$" setting in Table 1) by 0.7% in terms of average DOC, which demonstrates the merit of dynamically learned ground distance matrix in Eq. (5).

Figure 4 and Figure 5 demonstrate the qualitative visualization results on the 3D abdominal multi-organ segmentation. Specifically, Figure 4 shows the visualized segmentation results on the 3D model, while Figure 5 presents segmentation results on 2D slices. As shown in both figures, compared with other SOTA methods, our proposed segmentation framework can generate more accurate segmentation results with better recalls and less cross-class mis-classification.

4.1.3. Ablation Study

Evaluation of the Influence of Feature Embedding Dimension and learning behavior of the proposed method.

Since I2PEMD is exploited on the feature embeddings, we also study the influence of the shared feature embedding dimension. Detailed results on the 3D abdominal multi-organ segmentation are presented in Figure 6 (a).

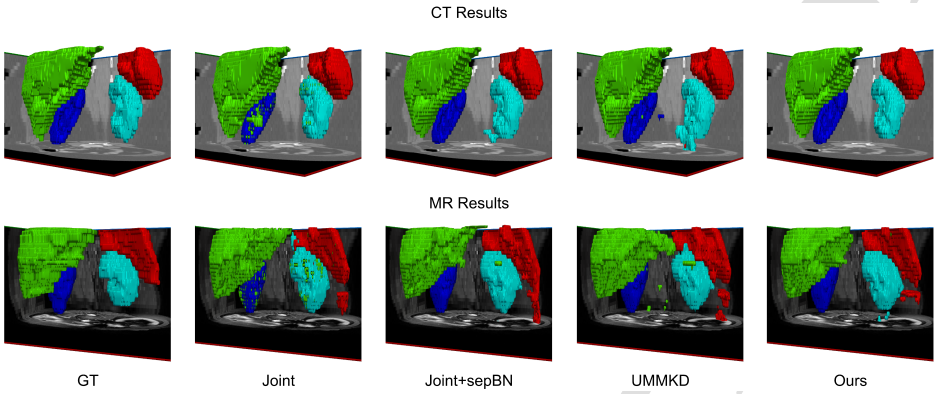


Figure 4: 3D visualization of the results obtained by different methods on 3D Abdominal multi-organ segmentation.

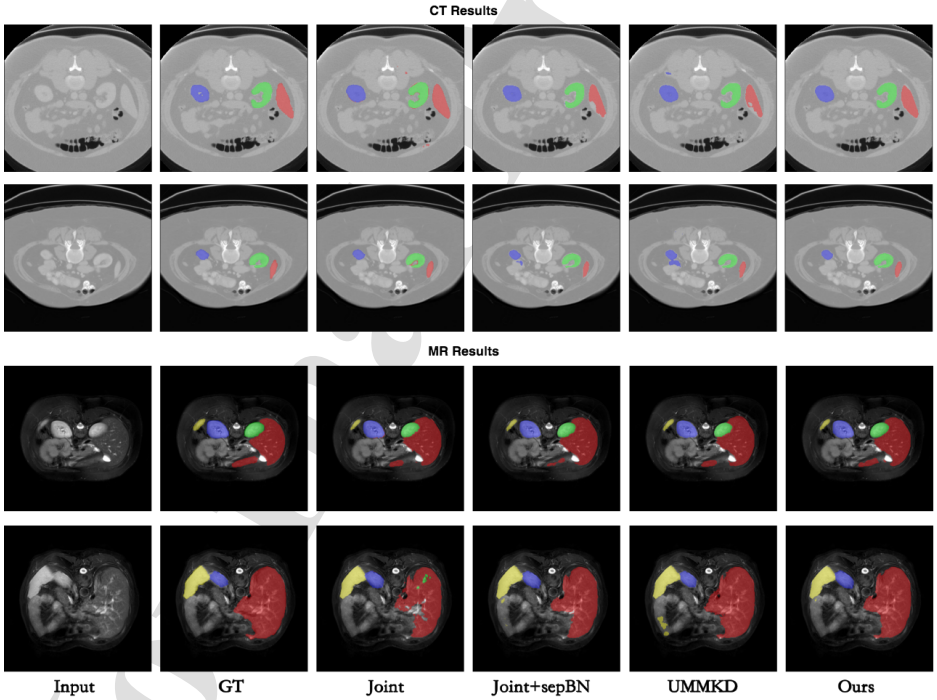


Figure 5: Visualized segmentation results on the 2D slices obtained by different methods on 3D Abdominal multi-organ segmentation.

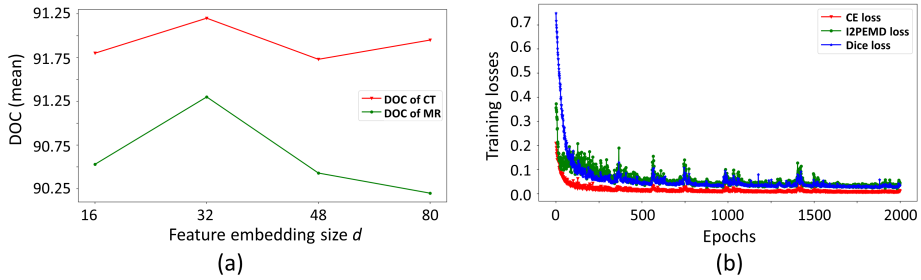


Figure 6: (a) Mean DOC plot for both CT and MR image segmentation when the shared feature embedding dimension varies; (b) loss curve of each item in the loss function of our method.

According to Figure 6 (a), $d = 32$ exceeds all the other listed settings. We can see that the best segmentation results are generated with a moderate feature dimension. This is reasonable, since the model capacity is limited with too small feature dimension. At the same time, it may easily overfit to the training dataset (due to the small size of training set) when the feature dimension is too large. A bonus of a moderate feature dimension is that it can lower the computational load.

In order to analyze the learning behavior of the proposed method, we also plot the loss curve of each item in the loss function in Figure 6 (b). From this figure, one can clearly see that as the training is going on, all three items in the loss function synchronously converge to flat minima. To further analyze the learning behavior, for each epoch during training, we also evaluate the trained model on the testing data. To this end, we record three metrics during training: a) mean I2PEMD when $\mathcal{L}_{\text{I2PEMD}}$ is evaluated on the testing data (referred as mI2PEMD); b) the cross-domain distribution divergence on the testing data measured by Maximum Mean Discrepancy (MMD) with gaussian kernel when evaluated on features extracted from the penultimate layer (referred as MMD) (Long et al., 2015); c) the mean DOC when the trained model is evaluated on the testing data (referred as mDOC). We then calculate Pearson's Correlation Coefficient (PCC) (Zou et al., 2003) between mI2PEMD and mDOC, as well as the PCC between mI2PEMD and MMD. The results are presented in Table 2. As shown in this table, mI2PEMD has a strong negative correlation with mDOC, indicating that the decreasing of I2PEMD is correlated to the increasing of the segmentation performance of our proposed method on testing data. Furthermore, mI2PEMD has a moder-

Table 2: PCC between mI2PEMD and mDOC as well as PCC between mI2PEMD and MMD when the trained model is evaluated on the testing data at each epoch during training.

Settings	PCC	p-value
mI2PEMD vs. mDOC	- 0.79	< 0.001
mI2PEMD vs. MMD	+ 0.56	< 0.001

ate positive correlation with MMD, demonstrating that minimizing I2PEMD could lead to a reduction of distribution gap between CT and MR domains. Thus, combining results presented in Table 1, Figure 6 (b) and Table 2, we conclude that for the task of unpaired multi-modal image segmentation, the proposed I2PEMD regularizer is effective in aligning the feature distributions between the CT and MR domains for a better performance.

Investigation of the effect of individual components in I2PEMD.
The proposed I2PEMD module consists of two components: the instance-to-prototype matching mechanism and the adoption of EMD. In this study, we investigate the effect of each individual component of I2PEMD on the performance of the proposed method. Specifically, we replace EMD with Euclidean distance, forming the Instance to Prototype Euclidean Distance (we refer it as I2PED). We then replace the I2PEMD regularizer with the I2PED regularizer in the unpaired multi-modal segmentation task, and report its performance when the same experimental setup as presented above is used. Detailed results are presented in Table 3. From this table, one can see that the "Joint+sepBN+I2PED" model achieves better performance than all the listed state-of-the-art methods while the proposed method further outperforms the "Joint+sepBN+I2PED" model by considering the cross-class relationship during cross-domain feature distribution alignment. From the results reported in Table 3, we can see that each individual component of the proposed I2PEMD module helps to improve the segmentation results while the instance-to-prototype mechanism plays a vital role in boosting the performance.

4.2. Semi-supervised Classification

In the task of semi-supervised classification, we conduct comprehensive experiments on the 2018 International Skin Imaging Collaboration (ISIC 2018) challenge dataset (Codella et al., 2018) to demonstrate the performance of our I2PEMD-guided semi-supervised learning framework. Similar to Sec. 4.1, we will present the experimental setup and the quantitative and

Table 3: Investigation of the effect of individual components in I2PEMD. Results are reported on the 3D unpaired multi-modal abdominal image segmentation task in terms of mean DOC.

Methods	mean DOC
X-shaped (Valindria et al., 2018)	89.9
Y-shaped (Valindria et al., 2018)	90.3
Joint + sepBN	90.7
UMMKD (Dou et al., 2020)	90.9
Joint + sepBN + I2PED	91.5
Joint + sepBN + I2PEMD (Ours)	91.8

qualitative comparison results to show the merit of our proposed I2PEMD in aligning the labeled and unlabeled feature subspaces via selecting more accurate pseudo-labeled samples.

4.2.1. Experimental Setup

Datasets.

The ISIC 2018 challenge dataset (Codella et al., 2018) contains 10,015 skin lesion dermoscopy images, which are categorized into 7 different types of skin lesions. Before being fed into the network, all images are resized into 224×224. In order to use the pre-trained model on ImageNet (Russakovsky et al., 2015), each image is normalized with statistics calculated for ImageNet. The entire dataset was randomly split to 70% for training, 20% for testing and 10% for validation, since we don't have access to the ground truth labels of the official validation and testing set. We adopt DenseNet121 (Huang et al., 2017) pre-trained on ImageNet as the backbone model.

Implementation Details.

All the experiments are conducted on a NVIDIA Tesla V100 GPU. We adopt the Adam optimizer (Diederik et al., 2015) to train our model. The batchsize is set to 48, with 12 labeled images and 36 unlabeled ones in each mini-batch. The learning rate is originally set to $1e-4$ and decays with 0.9 per 6 epoches. In total, the model is trained for 600 epochs. τ^{low}, τ^{high} are empirically set to 0.7, 0.8 respectively. The trade-off parameter λ is set to 1.0.

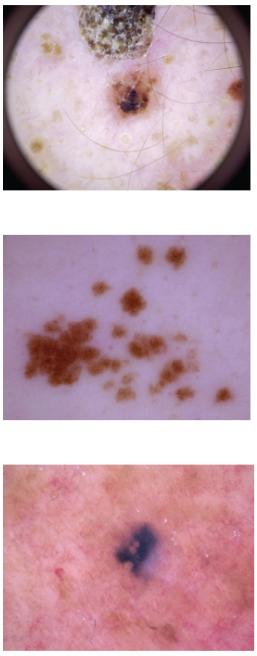
4.2.2. Comparison with SOTA Methods

We compare our method with SOTA methods including mean teacher (MT) (Tarvainen and Valpola, 2017), self-relation-consistency regularized mean teacher (SRC-MT) (Liu et al., 2020), FlexMatch (Zhang et al., 2021),

Table 4: Comparison results on semi-supervised classification with state-of-the-art methods. Best results are marked in bold font. Please note that p-values refer to the paired t-test results between Ours and listed state-of-the-art methods in terms of instance-level accuracy score. ** indicates $p < 0.01$.

Methods	**Percentage**		**Metrics**					
	Labelled	Unlabelled	AUC	Sensitivity	Specificity	Accuracy	F1	p-value
Upper Bound	100%	0	95.84	83.87	96.19	96.11	80.04	–
Baseline	20%	0	90.65	68.68	93.69	93.42	60.98	–
FlexMatch (Zhang et al., 2021)	20%	80%	87.83	64.79	93.18	93.00	51.90	0.000 **
MT (Tarvainen and Valpola, 2017)	20%	80%	88.52	**74.99**	90.98	90.09	50.74	0.000 **
SRC-MT (Liu et al., 2020)	20%	80%	90.84	74.63	91.76	91.11	54.64	0.000 **
augMT (Yu et al., 2019)	20%	80%	92.35	68.98	93.96	93.15	58.65	0.000 **
VAT (Miyato et al., 2018)	20%	80%	94.30	67.82	92.90	93.65	63.43	0.000 **
MixMatch (Berthelot et al., 2019)	20%	80%	94.48	72.12	90.21	93.42	64.31	0.001 **
FixMatch (Sohn et al., 2020)	20%	80%	93.71	68.41	94.04	93.94	63.13	0.000 **
Baseline+I2PEMD$_{fix}$	20%	80%	94.00	70.84	93.93	93.35	63.02	–
Ours	20%	80%	**94.79**	69.73	**94.05**	**94.54**	**66.99**	–

uncertainty-aware mean teacher (augMT) (Yu et al., 2019), MixMatch (Berthelot et al., 2019), Virtual Adversarial Training (VAT) (Miyato et al., 2018) and FixMatch (Sohn et al., 2020) in order to demonstrate effectiveness of our I2PEMD-guided semi-supervised classification framework. Five metrics were adopted to evaluate performance of these methods on the ISIC 2018 challenge dataset including Area Under Curve (AUC), accuracy, sensitivity, specificity and F1. AUC is computed as the area under the Receiver Operating Characteristic (ROC) curve, which describes the true-positive rate (sensitivity) versus the false-positive rate (100% - specificity) at various thresholds. An AUC of 100% represents a perfect test while an AUC of 50% indicates random predictions. For fair comparison, we adopt the same DenseNet121 model pretrained on ImageNet as the backbone and train/test on our own dataset split. Table 4 shows the performance of these methods using 20% labeled data for model training. The upper bound performance was obtained by training the backbone network with 100% labeled data in a fully supervised manner. The baseline was obtained by training the backbone network with 20% labeled data in a fully supervised manner. As shown in Table 4, except for the sensitivity metric, our I2PEMD-guided semi-supervised classification framework consistently outperforms other listed state-of-the art methods. Furthermore, compared with other pseudo-label based methods (including FixMatch (Sohn et al., 2020), FlexMatch (Zhang et al., 2021)), the superior performance of our method demonstrates that the proposed I2PEMD can help to select more accurate pseudo-labeled samples. We also report the performance of our method with entries of the ground distance matrix



Ma	*Mn*	*Bcc*	*Ak*	*Bk*	*Da*	*Vl*	
0.91	0.04	0.01	0.01	0.01	0.02	0	*MT*
1.0	0	0	0	0	0	0	*SRC-MT*
1.0	0	0	0	0	0	0	*FixMatch*
0.04	0.96	0	0	0	0	0	*Ours*
0.28	0.08	0.4	0.06	0.04	0.09	0.05	*MT*
0	0.04	0.95	0.01	0	0	0	*SRC-MT*
0	0	1.0	0	0	0	0	*FixMatch*
0	1.0	0	0	0	0	0	*Ours*
0.03	0.03	0.03	0.02	0.78	0.04	0.07	*MT*
0.01	0.09	0.01	0	0.16	0.06	0.67	*SRC-MT*
0	0	0	0	0	0	1.0	*FixMatch*
0	1.0	0	0	0	0	0	*Ours*

Figure 7: Demonstration of some difficult samples that have severe intra-class visual difference. The digits refer to the classification scores, where each row represents the results of the corresponding method, and each column illustrates the scores of the corresponding class. Yellow area highlight the results on the ground truth class. "Ma", "Mn", "Bcc", "Ak", "Bk", "Da", "Vl" are the short for "Melanoma", "Melanocytic nevus", "Basal cell carcinoma", "Actinic keratosis", "Benign keratosis", "Dermatofibroma", "Vascular lesion", respectively.

fixed to one (denoted as "Baseline+I2PEMD$_{fix}$" in Table 4). As shown, dynamic ground distance matrix based I2PEMD brings more performance gain to semi-supervised classification than its fixed counterpart.

We further conduct paired t-test to compare the results on instance-level accuracy scores between our method and other state-of-the-art methods. The detailed results are presented in the last column of Table 4. As shown, the p-values of each paired t-test are smaller than 0.01. Therefore, we can conclude that the differences achieved by the proposed method and other SOTA methods are statistically significant. Figure 7 illustrates the prediction results on some samples. As shown in this figure, although with severe intra-

Table 5: Results of the ablation study on the effectiveness of different sample selection methods. Best results are marked in bold.

Methods	**Metrics**				
	AUC	Sensitivity	Specificity	Accuracy	F1
Baseline	90.65	68.68	93.69	93.42	60.98
Filter_{all}	88.11	59.2	92.33	92.59	47.0
Filter_{thresh}	93.71	68.41	94.04	93.94	63.13
$\text{Filter}_{\text{I2PEMD}}$ (Ours)	**94.79**	**69.73**	**94.05**	**94.54**	**66.99**

class variations, our method can correctly classify them into the ground truth class, while other methods all predict them into wrong classes.

4.2.3. Ablation Study

Effectiveness of I2PEMD on Sample Selection.

To further demonstrate the effectiveness of our proposed I2PEMD in choosing better pseudo-labeled data for enriching the supervised information, we design following sample selection strategies: 1) all the pseudo-labeled samples are included in N_s to supervise the classification loss in Eq. 19 (denoted as Filter_{all}); 2) only samples with classification confidence score (s_i in Eq. 15) greater than τ^{high} are selected (denoted as Filter_{thresh}); and 3) our I2PEMD-guided sample selection in Eq. 18 (denoted as $\text{Filter}_{\text{I2PEMD}}$). For fair comparison, the tradeoff parameter λ in Eq. (19) is set to 1.0 for all the settings. The comparison results are presented in Table 5. Comparing the results of Filter_{all} with the "Baseline" setting, we can see that due to the error accumulation problem, directly using all the unlabeled samples for weak/strong alignment generally does not bring benefits to the classification performance. On the other hand, Filter_{thresh} and $\text{Filter}_{\text{I2PEMD}}$ both consistently outperform the baseline results, demonstrating the critical impacts of pseudo-labeling quality. Furthermore, comparing $\text{Filter}_{\text{I2PEMD}}$ with Filter_{thresh}, we can see that, by learning shared prototypes and considering the cross-class semantic priors, I2PEMD-guided semi-supervised classification framework can help to alleviate the error accumulation problem via generating more accurate pseudo-labels.

Results of the ablation study on influence of hyper-parameters $(\tau^{low}, \tau^{high})$.

We conduct experiments on the different combinations of hyper-parameters $(\tau^{low}, \tau^{high})$ to shed some light on better parameter configurations. Detailed

Table 6: Ablation study on the influence of different combinations on (τ^{low},τ^{high}). Best results are marked in bold.

Settings	Metrics				
	AUC	Sensitivity	Specificity	Accuracy	F1
(0.6, 0.7)	92.95	70.26	94.01	93.13	60.81
(0.7, 0.8)	**94.79**	69.73	94.05	94.54	66.99
(0.8, 0.9)	94.14	70.15	93.98	94.37	68.18
(0.6, 0.9)	94.56	**71.1**	**94.29**	94.14	65.51
(0.7, 0.9)	94.41	69.38	93.95	**94.61**	**68.69**

comparison results are illustrated in Table 6. As shown, the $(0.6, 0.7)$ setting achieves inferior performance than the other settings, while other settings bring mild performance changes in terms of different metrics. Comparing the results between setting $(0.6, 0.7)$ and $(0.6, 0.9)$, we can see that the performance is more sensitive to τ^{high}, which is reasonable, since τ^{high} controls the number of samples selected with high confidence.

5. Discussion

In this section, we will mainly discuss how our proposed I2PEMD benefits the unpaired multi-modal segmentation task and the semi-supervised classification task through distribution alignment. In the framework of unpaired multi-modal segmentation, I2PEMD functions as a regularization term to directly supervise the training process. Specifically, I2PEMD constrains the model to learn domain-invariant prototypes for CT and MRI inputs. The shared prototypes align the two domains from a global view, such that the shared semantic parts of the CT and MRI images (i.e., pixels with the same semantic class) can generate the same cluster in the learned feature embedding space. On the other hand, the EMD further helps to align the feature of each local pixel with the globally learned prototypes. By considering the cross-class relationships, EMD enables larger penalty when less visually similar classes are mis-classified. Since the joint base model shares parameters across the CT and MRI domains, distribution alignment can improve the segmentation performance by aligning the embedding distribution of different domains, which was confirmed by the results presented in Table 1 and Table 2.

As for the task of semi-supervised classification, I2PEMD works as a

less-confident sample filter to mitigate the "error accumulation" problem in pseudo-labeling. Due to sample bias phenomenon, there exists feature distribution gap between the labeled and unlabeled samples. Consequently, directly utilizing the predictions on unlabeled data may lead to incorrect assignments. Incorrect assignments during training may cause further misclassifications in subsequent iterations, resulting in a feedback loop of self-reinforcing errors that ultimately yields a low-accuracy classifier. In our framework, the proposed I2PEMD can mitigate the incorrect assignment from the following two aspects: 1) in Eq. (17), EMD helps to filter less confident samples that deviate from its corresponding or visually similar semantic classes (prototypes); 2) by learning shared prototypes via momentum update, I2PEMD progressively mitigate the distribution gap between labeled and unlabeled data, improving the generalization ability of the learned model, thus resulting in more accurate predictions on the unlabeled data.

This progressive strategy also shares a similar spirit with curriculum learning (Bengio et al., 2009; Wang et al., 2021). Specifically, I2PEMD helps to define the ranking function in measuring sample difficulty in curriculum learning. At the beginning of the model training, I2PEMD can help to select easier samples that both satisfy the high-prediction-confidence criteria and accord with the cross-class-relationship prior defined in the Earth Mover's Distance. In this way, we can progressively build a more robust and generalized model that can generate more accurate predictions on the unlabeled data. Therefore, we can enlarge the number of training samples by selecting more pseudo-labeled samples according to Eq. (18). This "filter-enlarge" loop shares the similar concept of "training from easier data (tasks) to harder data (tasks)" in curriculum learning (Wang et al., 2021).

6. Conclusion

We propose a novel distribution alignment algorithm, where the alignment is achieved by explicit shared prototype learning and consideration of the cross-class relationships during the instance-to-prototype matching. The proposed distribution alignment module can be flexibly plugged into many frameworks to benefit the tasks which need to bridge gap between different domains or feature subspaces. Comprehensive experiments on the unpaired multi-modal segmentation task and the semi-supervised classification task demonstrate effectiveness of the proposed I2PEMD. Specifically, in multi-modal segmentation, I2PEMD can boost the performance of joint training

of the CT and MRI domains by aligning the cross-domain feature embeddings. Meanwhile, involved in the "filter-enlarge" loop, I2PEMD helps to progressively build a more robust and generalizable model, leading to better performance in semi-supervised classification.

7. Acknowledgments

The work was partially supported by Shanghai Municipality Science and Technology Commission under grant 20511105205, by the Natural Science Foundation of China under grant U20A20199, and by the National Key R&D Program of China under grant 2019YFC0120603.

References

Aviles-Rivero, A. I., Papadakis, N., Li, R., Sellars, P., Fan, Q., Tan, R. T., Schönlieb, C.-B., 2019. Graph$X^{NET}-$ chest x-ray classification under extreme minimal supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 504–512.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2020. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: International Conference on Learning Representations.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C. A., 2019. Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems (NeurIPS), 5050–5060.

Bhatnagar, G., Wu, Q. J., Liu, Z., 2015. A new contrast based multimodal medical image fusion framework. Neurocomputing 157, 143–152.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D., Dharmakumar, R., Tsaftaris, S. A., 2018. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 490–498.

Chen, J., Li, W., Li, H., Zhang, J., 2020. Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 187–196.

Chen, L., Wu, Y., DSouza, A. M., Abidin, A. Z., Wismüller, A., Xu, C., 2018. Mri tumor segmentation with densely connected 3d cnn. In: Medical Imaging: Image Processing. p. 105741F.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 424–432.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, pp. 168–172.

Cubuk, E. D., Zoph, B., Shlens, J., Le, Q. V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703.

Diederik, K., Jimmy, B., et al., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). pp. 273–297.

Dolz, J., Desrosiers, C., Ben Ayed, I., 2018a. Ivd-net: Intervertebral disc localization and segmentation in mri with a multi-modal unet. In: International workshop and challenge on computational methods and clinical applications for spine imaging. Springer, pp. 130–143.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I. B., 2018b. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. IEEE transactions on medical imaging 38 (5), 1116–1126.

Dong, J., Lin, T., 2019. Margingan: Adversarial training in semi-supervised learning. Advances in Neural Information Processing Systems (NeurIPS) 32, 10440–10449.

Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E., 2018. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 544–552.

Dou, Q., Liu, Q., Heng, P. A., Glocker, B., 2020. Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging 39 (7), 2415–2425.

Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T., 2013. Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision. pp. 2960–2967.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. PMLR, pp. 1180–1189.

Gong, B., Shi, Y., Sha, F., Grauman, K., 2012. Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp. 2066–2073.

Gopalan, R., Li, R., Chellappa, R., 2011. Domain adaptation for object recognition: An unsupervised approach. In: 2011 international conference on computer vision. IEEE, pp. 999–1006.

Hitchcock, F. L., 1941. The distribution of a product from several sources to numerous localities. Journal of mathematics and physics 20 (1-4), 224–230.

Hou, L., Yu, C.-P., Samaras, D., 2016. Squared earth mover's distance-based loss for training deep neural networks. CoRR abs/1611.05916, 1466–1475.

Hu, J., Lu, J., Tan, Y.-P., 2015. Deep transfer metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 325–333.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T. K., Savona, M. R., Abramson, R. G., Landman, B. A., 2018. Synseg-net: Synthetic segmentation without target modality ground truth. IEEE transactions on medical imaging 38 (4), 1016–1025.

Iscen, A., Tolias, G., Avrithis, Y., Chum, O., 2019. Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5070–5079.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K. H., 2017. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In: International MICCAI Brainlesion Workshop. pp. 287–297.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. Vol. 5. p. 12.

Lee, D.-H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. p. 896.

Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., Shen, H. T., 2021. Maximum density divergence for domain adaptation. IEEE transactions on pattern analysis and machine intelligence 43 (11), 3918–3930.

Li, W., Wang, Z., Yue, Y., Li, J., Speier, W., Zhou, M., Arnold, C., 2020. Semi-supervised learning using adversarial training with good and bad samples. Machine Vision and Applications 31 (6), 1–11.

Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P. A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging 39 (11), 3429–3440.

Long, M., Cao, Y., Wang, J., Jordan, M., 2015. Learning transferable features with deep adaptation networks. In: International conference on machine learning. PMLR, pp. 97–105.

Long, M., Cao, Z., Wang, J., Jordan, M. I., 2018. Conditional adversarial domain adaptation. Advances in Neural Information Processing Systems (NeurIPS), 1647–1657.

Long, M., Zhu, H., Wang, J., Jordan, M. I., 2017. Deep transfer learning with joint adaptation networks. In: International conference on machine learning. PMLR, pp. 2208–2217.

Miyato, T., Maeda, S.-i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence 41 (8), 1979–1993.

Myronenko, A., 2018. 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. Springer, pp. 311–320.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. Asdnet: attention based semi-supervised deep networks for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 370–378.

Pereira, S., Pinto, A., Alves, V., Silva, C. A., 2016. Brain tumor segmentation using convolutional neural networks in mri images. IEEE transactions on medical imaging 35 (5), 1240–1251.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115 (3), 211–252.

Samuli, L., Timo, A., 2017. Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (ICLR).

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., Li, C.-L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems 33, 596–608.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30, 1195–1204.

Valindria, V. V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E. O., Rockall, A. G., Rueckert, D., Glocker, B., 2018. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp. 547–556.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2017. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: International MICCAI brainlesion workshop. Springer, pp. 178–190.

Wang, Q., Li, W., Gool, L. V., 2019. Semi-supervised learning by augmented distribution alignment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1466–1475.

Wang, X., Chen, Y., Zhu, W., 2021. A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020. Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems 33, 6256–6268.

Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S., 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. In: International Conference on Learning Representation (ICLR).

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T., 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems 34.

Zhang, J., Li, W., Ogunbona, P., 2017. Joint geometrical and statistical alignment for visual domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1859–1867.

Zhang, Y., Deng, B., Jia, K., Zhang, L., 2020. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In: European Conference on Computer Vision. Springer, pp. 781–797.

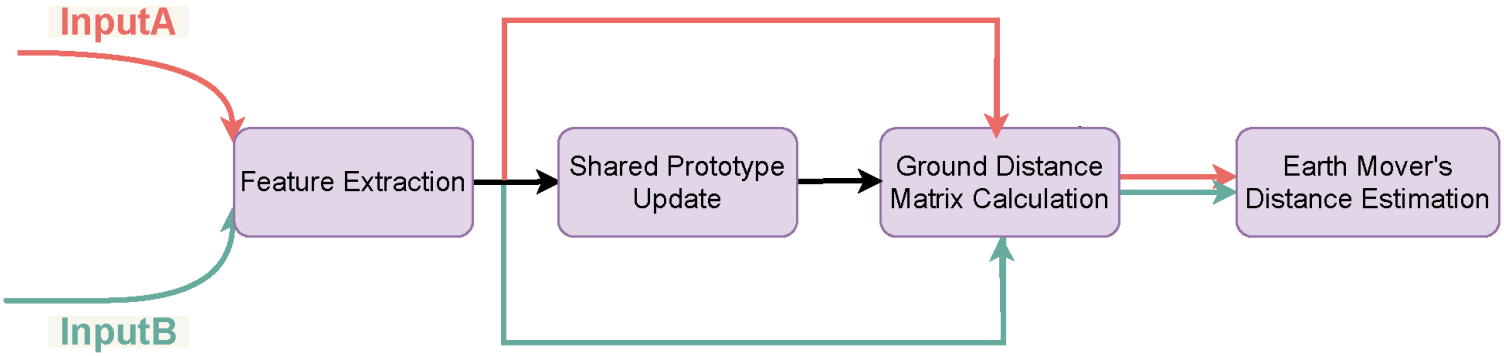
Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 9242–9251.

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y., 2018. A deep learning model integrating fcnns and crfs for brain tumor segmentation. Medical image analysis 43, 98–111.

Zhou, T., Ruan, S., Canu, S., 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3, 100004.

Zou, K. H., Tuncali, K., Silverman, S. G., 2003. Correlation and simple linear regression. Radiology 227 (3), 617–628.

- Propose instance to prototype Earth Mover's distance (I2PEMD) to address the distribution alignment problem.
- The proposed instance to prototype Earth Mover's distance combines shared prototype learning with EMD estimation to take into consideration of both intra-class compactness and cross-class relationships during distribution alignment
- Extensive validation on two typical yet challenging tasks, i.e., unpaired multi-modal image segmentation and semi-supervised image classification.



CRediT authorship contribution statement

Qin Zhou: Methodology, Software, Validation, Figure preparation, Writing – Methodology & Results.

Runze Wang: Software , Validation, Editing & Review

Guodong Zeng: Software, Editing & Review

Heng Fan: Software, Editing & Review

Guoyan Zheng: Conceptualization, Editing & Review, Supervision, Funding Acquisition.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Shanghai, 28 June 2022

Declaration of Conflict of Interest

We wish to confirm to the Editor that there is no known conflict of interest associated with this submission and that there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there is no other person who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all co-authors.

We understand that the Corresponding Author is the sole contact for the Editorial Process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other author about progress, submissions of revisions and final approval of proofs.

Yours Sincerely,

Yours Sincerely,

Prof. Dr. Guoyan Zheng

Institute of Medical Robotics
Shanghai Jiao Tong University,
China
Tel: 0086 13062730961
Email: guoyan.zheng@sjtu.edu.cn