Unpaired Multi-Modal Segmentation via Knowledge Distillation

Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker

***Abstract*— Multi-modal learning is typically performed with network architectures containing modality-specific layers and shared layers, utilizing co-registered images of different modalities. We propose a novel learning scheme for unpaired cross-modality image segmentation, with a highly compact architecture achieving superior segmentation accuracy. In our method, we heavily reuse network parameters, by sharing all convolutional kernels across CT and MRI, and only employ modality-specific internal normalization layers which compute respective statistics. To effectively train such a highly compact model, we introduce a novel loss term inspired by knowledge distillation, by explicitly constraining the KL-divergence of our derived prediction distributions between modalities. We have extensively validated our approach on two multi-class segmentation problems: i) cardiac structure segmentation, and ii) abdominal organ segmentation. Different network settings, i.e., 2D dilated network and 3D U-net, are utilized to investigate our method's general efficacy. Experimental results on both tasks demonstrate that our novel multi-modal learning scheme consistently outperforms single-modal training and previous multi-modal approaches.**

***Index Terms*— Unpaired multimodal learning, knowledge distillation, feature normalization, image segmentation.**

I. Introduction

ANATOMICAL structures are imaged with a variety of modalities depending on the clinical indication. For instance, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) show cardiac structures with complementary information important for the assessment of heart diseases [1], [2]. Despite differences between CT and MRI, often a similar analysis is required, such as quantitative assessment via segmentation. Common practice is to develop a segmentation method, e.g., a convolutional neural network (CNN), separately for CT and MRI data. Such separate training leaves potentially valuable cross-modality information unused. By leveraging multi-modal learning, we can exploit shared cross-modality information, possibly making better use of limited datasets and improving overall performance on each modality.

Previous works on multi-modal image segmentation mostly use multi-parametric MRI (e.g., T1, T2, FLAIR). The inputs to a CNN are paired images, i.e., multi-modal data are acquired from the same patient and co-registered across the sequences. To learn representations for multi-modal segmentation, early fusion and late fusion strategies are typically utilized. Specifically, early fusion means concatenating multi-modal images as different channels at the input layer of a network. This strategy has demonstrated effectiveness on segmenting brain tissue [3]–[5] and brain lesions [6]–[8] in multiple sequences of MRI. For late fusion, each modality has modality-specific layers at an early stage of a CNN. The features extracted from different modalities are fused at a certain middle layer of the CNN. The intuition is to initially learn independent features from each modality, and then fuse them at a semantic level. Briefly speaking, late fusion forms a "Y"-shaped architecture, as shown in Fig. 1 (a). It has been widely applied for analyzing brain imaging [9], spinal structures [10], prostate cancer [11], and others. Recently, more complex multi-modal CNNs have been designed, by leveraging dense connections [12], inception modules [13] or multi-scale feature fusion [14]. These more complicated models still follow the idea of combining modality-specific and shared layers.

We identify two main limitations in the current multi-modal segmentation literature. Firstly, input images are typically paired, which requires multiple images from the same patient as well as a registration step at pre-processing. How to leverage unpaired multi-modal images, e.g, data acquired from different cohorts, still remains unclear. Secondly, multi-modality is often limited to different sequences of MRI. An arguably more challenging situation of multi-modal learning combining CT and MRI is less well explored. Due to distinct physical principles of the underlying image acquisition, the very different visual appearance may require new ways to exchange cross-modality information compared to multi-sequence MRI.

To tackle above limitations, studying unpaired multi-modal learning from non-registered CT and MRI has gained

Manuscript received November 12, 2019; revised December 17, 2019; accepted December 24, 2019. Date of publication February 3, 2020; date of current version June 30, 2020. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme, under Grant 757173, in part by the Project MIRA, under Grant ERC-2017-STG, and in part by the Hong Kong Innovation and Technology Commission through ITSP Scheme under Grant ITS/426/17FP and Grant ITS/311/18FP. *(Corresponding author: Qi Dou.)*

Qi Dou and Ben Glocker are with the Biomedical Image Analysis Group, Imperial College London, London SW7 2AZ, U.K. (e-mail: qi.dou@imperial.ac.uk; b.glocker@imperial.ac.uk).

Quande Liu and Pheng Ann Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: qdliu@cse.cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2963882

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

some recent interest [15]–[18]. The very different statistical distributions of CT and MRI makes this a challenging problem in terms of learning shared representations. As the unpaired images have little pixel-to-pixel coherence, cross-modality relationship only exists in a semantic space. Valindria *et al.* [15] is the closest work to this paper, working on CT/MRI multi-organ segmentation by investigating several dual-stream CNNs, demonstrating a benefit of cross-modality learning of CT and MRI. The state-of-the-art performance is achieved by a "X"-shaped network, as shown in Fig. 1 (b). A recent work [19] also demonstrate that such an "X"-shaped model is effective for unsupervised multi-modal learning. The modality-specific encoder and decoder layers are designed to tackle the distribution shift between two modalities, while the shared middle layers fuse multi-modal representations.

Our paper proposes a novel compact model for unpaired CT and MRI multi-modal segmentation, by explicitly addressing distribution shift and distilling cross-modality knowledge. We use modality-specific internal feature normalization parameters (e.g., batch normalization layers), while sharing all the convolutional kernels. Importantly, we further propose to distill semantic knowledge from pre-softmax features. A new loss term is derived by minimizing the KL-divergence of a semantic confusion matrix, to explicitly leverage the shared information across modalities. We extensively evaluate our method on two CT and MRI multi-class segmentation tasks, including cardiac segmentation with a 2D dilated CNN and abdominal multi-organ segmentation with a 3D U-Net. Our method consistently outperforms single-model training and state-of-the-art multi-modal learning schemes on both segmentation tasks. The contributions of this work are summarized as follows:

- We present a novel, flexible, compact multi-modal learning scheme for accurate segmentation of anatomical structures from unpaired CT and MRI.
- We propose a new mechanism to distill semantic knowledge from high-level CNN representations. Based on this, we further derive an effective loss function to guide multi-modal learning.
- We conduct extensive validations on two different multi-class segmentation tasks with 2D and 3D CNN architectures, demonstrating general effectiveness of our method.

Code for our proposed approach is publicly available at https://github.com/carrenD/ummkd.

II. Related Work

Before presenting the proposed approach, we review the literature that inspired the design of our multi-modal learning scheme. The two key aspects are: 1) separating internal feature normalizations for each modality, given the very different statistical distributions of CT and MRI; 2) knowledge distillation from pre-softmax activations, in order to leverage information shared across modalities to guide the multi-modal learning.

A. Independent Normalization of CT and MRI

Representation learning between CT and MRI has attracted increasing research interest in recent years. Zhang *et al.* [16] learn image-to-image translation using unpaired CT and MRI cardiac images. Dou *et al.* [17] present unsupervised domain adaptation of CNNs between CT and MRI for the task of cardiac segmentation using adversarial learning. Huo *et al.* [18] learns a CycleGAN based segmentation model from unpaired CT and MRI, only using segmentation labels from one modality. In terms of supervised multi-modal segmentation, image style transfer techniques may not be necessary because we have precise annotations to fully supervise the learning process. To the best of our knowledge, Valindria *et al.* [15] is the only paper working on supervised unpaired CT and MRI segmentation so far. They extensively investigate four different types of dual-stream architectures, showing that a "X"-shaped architecture obtains the best performance. This indicates that the distribution shift between CT and MRI heavily affects feature-sharing, requiring modality-specific encoders/decoders. We hypothesis that if the features from different modalities are better normalized, learning cross-modality representations may become easier. In literature, independently normalizing features from different domains has demonstrated efficacy for image classification [20] and life-long learning on multi-modal MRI brain segmentation [21].

B. Knowledge Distillation

The concept of knowledge distillation (KD) originates from Hinton *et al.* [22] for model compression, i.e., transferring what has been learned by a large model to a smaller-scale model using soft-label supervision. A key aspect that enables KD is to leverage soft labels instead of hard one-hot labels. Temperature scaling is an essential component to allow this by obtaining softer probability distributions across classes, in order to amplify the inter-class relationships. Previous work has adopted knowledge distillation to address various tasks not limited to original model compression [22], but also a wider scope of scenarios such as domain adaptation [23], life-long learning [24], adversarial attacks [25] and self-supervised learning [26]. In medical imaging, the potential of the knowledge distillation technique is promising yet relatively under-explored as far as we know. Wang *et al.* [27] employ KD for efficient neuronal structure segmentation from 3D optical microscope images with a teacher-student network. Kats *et al.* [28] borrow the concept of KD to perform brain lesion segmentation with soft labels by dilating mask boundaries. Christodoulidis *et al.* [29] utilize KD for multi-source transfer learning on the task of lung pattern analysis. With promising results in prior work, we expect an increase of interest in KD. In this paper, we absorb the spirit of knowledge distillation, and explore how to incorporate effective soft probability alignment with temperature scaling within a novel KD loss derived from processing high-level feature-maps for the task of unpaired multi-modal learning.

III. Methods

An overview of our proposed multi-modal segmentation method is shown in Fig. 1 (d). In this section, we first present a compact model design with modality-specific parameters for feature normalization. Next, we show how to distill knowledge from semantic feature-maps and derive the loss term for multi-modal learning. The training procedures are finally described.

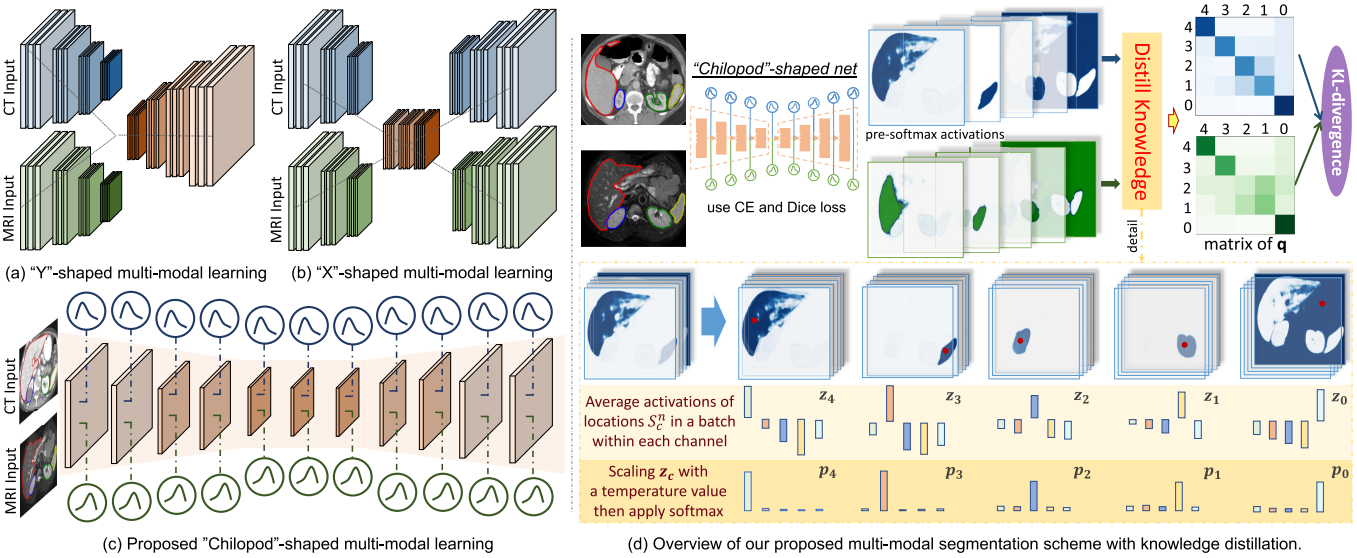

Fig. 1. Overview of proposed multi-modal learning scheme for unpaired image segmentation using knowledge distillation. (a) and (b) are the conventional multi-modal learning architectures using modality-specific encoders or decoders, forming a "X"-shaped or "Y"-shaped model. (c) Our proposed "Chilopod"-shaped multi-modal learning architecture, using modality-specific normalization layers, while all convolution kernels are shared. (d) Our proposed method for distilling semantic knowledge from pre-softmax activations, and deriving a pair of confusion matrix for a KL-divergence based loss term.

A. Separate Internal Feature Normalization

The central architecture of our proposed learning scheme is a separate normalization for internal activations, to mitigate the discrepancy from different data sources. Our design is very different from previous multi-modal learning methods. Rather than using modality-specific encoders/decoders with early/late fusions, we employ the same set of CNN kernels to extract features for both modalities yielding higher parameter efficiency. Using these modality-agnostic kernels gives raise to the hope of extracting universal representations that are more expressive and robust. To achieve this, calibration of the features extracted by the model is important. Normalizing internal activations into Gaussian distribution is common practice for improving convergence speed and generalization of a network. Let $x \in S_g^k$ denote the activation in the k-th layer, S_g^k is the g-th group of activations in the layer for which the mean and variance are computed, the normalization layer is:

$$\hat{x} = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}}, \quad y = \gamma \hat{x} + \beta, \tag{1}$$

where γ and β denote the trainable scale and shift. There are different ways to define the activation set of S_g^k, e.g, Batch Normalization [30], Instance Normalization [31], Layer Normalization [32], and Group Normalization [33].

We employ separate internal normalization for each modality, as the statistics of CT and MRI data are very different and should not be normalized in the same way which otherwise may yield defective features. For instance, if we have $\mu_{\text{CT}} = -\mu_{\text{MRI}}$ in a certain layer, then the mean over both would be zero which is meaningless. Using separate normalization layer for each modality can effectively avoid such problems during multi-modal learning. Our modality-specific normalization gives raise to a compact multi-modal architecture with minimal extra parametrization $\{\gamma, \beta\}$ forming a "Chilopod" shape, as illustrated in Fig. 1 (c). More specifically, the normalization layers for different modalities (i.e., CT and MRI) are implemented under separate variable scopes, while the convolution layers are constructed under shared variable scope. In every training iteration, the samples from each modality are loaded separately with sub-groups, and forwarded into shared convolution layers and independent normalization layers to obtain the logits which are employed to calculate the knowledge distillation loss.

B. Knowledge Distillation Loss

As we share all CNN kernels across modalities, the encoders are expected to extract universal representations, capturing common patterns such as shape, which may be more robust and discriminative across modalities. Training, however, becomes more difficult and we find that ordinary objectives, e.g., cross-entropy or Dice loss, are inadequate. We propose to explicitly enhance the alignment of distilled knowledge from semantic features of both modalities.

The assumption of KD is that the probabilities from softmax contain richer information than one-hot outputs. An additional temperature scaling for the pre-softmax activations gives softer probability distributions over classes, which can further amplify such knowledge. In our approach, we distill semantic knowledge from high-resolution feature maps before softmax, as illustrated in Fig. 1 (d). We average the activations over all locations of each class and compute the soft predictions across all classes. We describe this process for 2D below, while it can be easily extended to 3D.

We denote the activation tensor before softmax by $M \in \mathbb{R}^{N \times W \times H \times C}$, where N is the batch size, W and H denote width and height, C is number of channels which also equals to the number of classes as M is pre-softmax tensor. Let $z_{nwhi} \in M$ denote one neuron activation in M

with index of (n, w, h, i), and $\mathcal{S}_c^n$ denote the set of (w, h) locations in the n-th sample where the pixel's label is class c. Then, for each class c, inside each channel of M, we distill the knowledge over all the locations which belong to class c. There are C channels in total, one for each class. Hence, we get a C-dimensional vector $\mathbf{z}_c \in \mathbb{R}^C$ for class c, with its i-th element $\mathbf{z}_c^i$ as averaging z_{nwhi} over all $\mathcal{S}_c^n$ locations in the i-th class channel. Formally, this procedure is represented as:

$$\mathbf{z}_c^i = \frac{1}{\sum_n |\mathcal{S}_c^n|} \sum_n \sum_{n(w,h)\in S_c^n} z_{nwhi}. \tag{2}$$

Next, we compute scaled $\mathbf{z}_c$ into a probability distribution $\mathbf{p}_c \in \mathbb{R}^C$ using softmax. This distilled knowledge $\mathbf{p}_c$ aggregates how the network's prediction probabilities for the pixels of class c distribute across all other classes. Temperature scaling is employed as:

$$\mathbf{p}_c^i = \frac{\exp(\mathbf{z}_c^i / T)}{\sum_j \exp(\mathbf{z}_c^j / T)}, \tag{3}$$

where $T > 1$ is the temperature scalar [22] for softer output to enhance small values. We set $T = 2$ in our experiments, with $T = 1$ being the ordinary softmax. We empirically observe that the model performance is not very sensitive to this hyper-parameter, within a reasonable range of $T < 10$ considered.

Similarly, we can get an array of distilled semantic knowledge $\mathbf{q} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{C-1}]$. Each vector element conveys how the model's predictions for pixels of a particular class would distribute across all classes. We can distill such knowledge for each modality, denoted by $\mathbf{q}^a$ and $\mathbf{q}^b$ for CT and MRI, respectively. We encourage the network's distilled knowledge from high-level representations of both modalities to be aligned. Intuitively, if one class in CT is often confused by another, this situation may also happen in MRI. For instance, as shown in top-right confusion matrix (the blue and green grid planes) of Fig. 1 (d), the confusions between class *1* and *2* are more obvious, and this is consistent in CT and MRI under our learning scheme. These two classes correspond to the left and right kidney (in human-body view) in abdominal images.

In our scheme, both $\mathbf{q}^a$ and $\mathbf{q}^b$ are updated together during the dynamic training process. We compute their relative entropy between vectors of each class and design a loss term to minimize their Kullback−Leibler (KL) divergence. Our knowledge distillation based loss term (KD-loss) is as follows:

$$\mathcal{L}_{\text{kd}} = \frac{1}{C} \sum_c \left(\mathcal{D}_{\text{KL}}(\mathbf{q}_c^a || \mathbf{q}_c^b) + \mathcal{D}_{\text{KL}}(\mathbf{q}_c^b || \mathbf{q}_c^a) \right),$$
$$\text{where } \mathcal{D}_{\text{KL}}(\mathbf{q}_c^a || \mathbf{q}_c^b) = \sum \mathbf{q}_c^a \log \frac{\mathbf{q}_c^a}{\mathbf{q}_c^b}. \tag{4}$$

We compute a symmetric version of KL-divergence between two modalities, and $\mathcal{D}_{\text{KL}}(\mathbf{q}_c^b || \mathbf{q}_c^a)$ is similar to $\mathcal{D}_{\text{KL}}(\mathbf{q}_c^a || \mathbf{q}_c^b)$. By explicitly enhancing alignment of the distilled knowledge, the shared kernels can extract features capturing cross-modality patterns co-existing in CT and MRI at multi-modal learning.

C. Overall Loss Function and Training Procedure

The compact segmentation network is trained with the loss function as follows:

$$\mathcal{L} = \mathcal{L}_{\text{seg}}^a + \mathcal{L}_{\text{seg}}^b + \frac{\alpha}{2} \mathcal{L}_{\text{kd}} + \eta \left(||\Theta_{\text{kernel}}||_2^2 + ||\Theta_{\text{norm}}^a||_2^2 + ||\Theta_{\text{norm}}^b||_2^2 \right), \tag{5}$$

where $\mathcal{L}_{\text{seg}}^a$ and $\mathcal{L}_{\text{seg}}^b$ are ordinary segmentation loss for each modality, for which we combine Dice loss [35] and pixel-wise weighted cross-entropy loss in our experiments. We multiply a scalar $1/2$ for $\mathcal{L}_{\text{kd}}$ to average over the symmetric KL-divergence. The α is generally set as 0.5, and we also study this hyper-parameter in ablation experiments. The last term is a L2 regularizer for shared kernels and modality-specific normalization parameters. The weight η is fixed as $1e-4$.

The multi-modal images are sampled to similar voxel-spacing, and normalized to zero mean, unit variance for each modality at the input layer. In each training iteration, we input half of the batch as CT and the other half as MRI. The images go through the shared convolutional kernels and respective internal normalization layers. This can be easily implemented in TensorFlow, by defining the scope of involved variables. The KD-loss is computed with activation tensors of the samples from each modality. All trainable parameters $\{\Theta_{\text{kernel}}, \Theta_{\text{norm}}^a, \Theta_{\text{norm}}^b\}$ are updated together with the loss function in Eq. 5 using an Adam optimizer. Note that our multi-modal learning scheme is architecture-independent. It can be easily integrated into various existing 2D and 3D CNN models.

IV. Experiments

We extensively evaluate our multi-modal learning approach on two different multi-class tasks: 1) cardiac structure segmentation; and 2) multi-organ segmentation. We implement 2D and 3D models with different network architectures to demonstrate the flexibility and general efficacy of our method.

A. Datasets and Networks

1) Task-1: We perform multi-class cardiac structure segmentation using the MICCAI 2017 Multi-Modality Whole Heart Segmentation Challenge [36] dataset, which consists of unpaired 20 CT and 20 MRI images from different patients and sites. The multi-class segmentation includes four structures: left ventricle myocardium (LVM), left atrium blood cavity (LAC), left ventricle blood cavity (LVC) and ascending aorta (AA). We crop the heart regions in the images. Both modalities are resampled to a voxel-spacing at around $1.0 \times 1.0 \times 1.0\ mm^3$ with size of 256×256 in the coronal plane. Before inputting to the network, we conduct intensity normalization to zero mean and unit variance separately for each modality. Each modality is randomly divided into 70% for training, 10% for validation and 20% for testing.

2) Network-1: For the cardiac segmentation, we implement a 2D CNN with dilated convolutions, following the architecture used in [17] which employ the same dataset. The inputs are three adjacent slices with the middle slice providing a ground

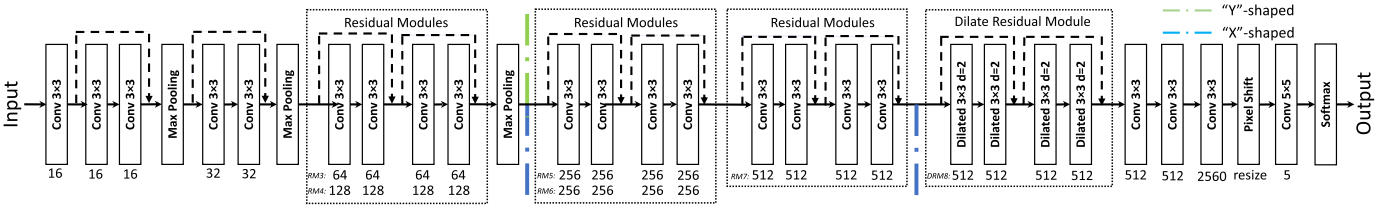

Fig. 2. Architecture of the 2D dilated network for multi-class cardiac structure segmentation, following [17]. The green and blue lines indicate break/merge points for "Y"−/"X"-shaped architectures. Best viewed in the zoomed-in view.

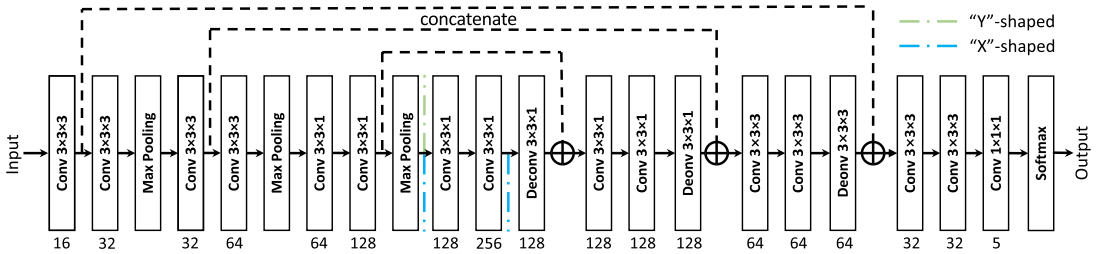

Fig. 3. Architecture of the 3D U-Net [34] for multi-organ segmentation. The black dot lines indicate skip connections. The green and blue lines indicate break/merge points for "Y"-/"X"-shaped architectures.

truth mask. The batch size is set as 8 for each modality. The learning rate is initialized as $1e-4$ and decayed by 5% in every 1000 iterations. The detailed network architecture is presented in Fig. 2, with size of convolution kernels indicated within the each box. The number of channels of feature maps are indicated below each box. The normalization layer follows each convolution operation before applying ReLU non-linearity, except for the last knowledge distillation layer before applying softmax.

The layers at which we separate or merge the data streams in the *"Y"-shaped* and *"X"-shaped* architectures are indicated using the colored lines in the architectures. For the *"Y"-shaped* model, layers before green line are separate, and layers after green line are shared between modalities. For the *"X"-shaped* model, layers before the first blue line and after the second blue line are separate, and layers in-between the blue lines are shared between modalities.

3) Task-2: We perform multi-organ segmentation in abdominal images for liver, spleen, right kidney (R-kdy) and left kidney (L-kdy). We utilize public CT dataset of [37] with 30 patients (but one case was excluded due to low image quality), and our MRI data come from the ISBI 2019 CHAOS Challenge, with 9 cases available at the time we downloaded data. This enables us to observe how our method performs in the situation when one modality has much fewer samples than the other. We crop the original CT and MRI images at the areas of multi-organs by excluding the black margins. Since these two datasets have large variance in voxel-spacing. We resample them into around $1.5 \times 1.5 \times 8.0$ mm^3, with size of 256×256 in transverse plane. The images are normalized to zero mean and unit variance for intensities within each modality before inputting to the network. Again, each modality is randomly divided into 70% for training, 10% for validation and 20% for testing.

4) Network-2: We employ a 3D U-Net [34] wise architecture as shown in Fig. 3, where dot-lines indicate skip connections in the network. The volumetric input has a size of $256 \times 256 \times 8$, considering to contain all organs inside the transverse view and also the constrain from GPU memory. The batch size is set as 4 at training. Noting that at every training iteration, the balance of the number of samples in a mini-batch for each modality still holds, considering training stability. The learning rate is initialized as $1e-4$ and decayed by 5% in every 500 iterations. The indication for *"Y"-shaped* and *"X"-shaped* architectures with green and blue lines are the same as described above for Network-1.

B. Experimental Settings

We generally use BN layer, as it is the most widely-adopted normalization technique in medical image segmentation tasks. Ablation study on other normalization layers is also conducted. For comprehensive analysis and comparison, we design the following seven experimental settings, and implement on our datasets. Network architecture and hyper-parameters are fixed for all the settings, for a fair comparison of different methods.

- **Individual**: independently training a separate model for each individual modality.
- **Joint**: training a single model with all network parameters (convolution kernels and BN) shared for CT and MRI.
- **Joint+KD**: training a joint model for CT and MRI, with adding our proposed KD-loss.
- **"Y"-shaped** [9]: modality-specific encoders and shared decoders, which is a widely-used late fusion scheme for multi-modal learning.
- **"X"-shaped** [15]: modality-specific encoder and decoder layers, with shared bottleneck layers. The [15] demonstrate this two-stream architecture to be the current state-of-the-art for unpaired CT and MRI segmentation.
- **"Chilopod"-shaped**: our proposed architecture, i.e., sharing all CNN kernels while keeping internal feature normalization layers as modality-specific.

TABLE I
MULTI-MODAL LEARNING RESULTS OF CARDIAC SEGMENTATION WITH 2D NETWORK UNDER DIFFERENT SETTINGS

Methods	Param Scale	Cardiac CT					Cardiac MRI					Overall Mean
		LVM	LAC	LVC	AA	Mean	LVM	LAC	LVC	AA	Mean	
		Dice Coefficient (avg.±std., %)										
Payer et al. [38]	-	87.2±3.9	**92.4±3.6**	92.4±3.3	91.1±18.4	90.8	75.2±12.1	81.1±13.8	87.7±7.7	76.6±13.8	80.2	85.5
Individual	78.64M	86.5±2.0	91.7±2.9	90.9±2.0	84.4±13.0	88.4	78.7±4.2	84.4±6.5	92.0±4.0	**84.1±5.4**	84.8	86.6
Joint	39.32M	84.5±2.1	87.6±3.5	90.2±2.5	92.2±4.7	88.6	75.6±3.9	85.3±4.7	93.3±1.8	81.7±5.3	84.0	86.3
Joint+KD	39.32M	84.6±2.6	89.0±3.3	89.8±4.1	92.5±3.7	89.0	76.6±4.2	85.2±6.7	93.2±2.0	83.0±3.0	84.5	86.8
"Y"-shaped [9]	39.99M	86.7±4.8	90.9±5.3	92.2±2.5	87.4±5.1	89.3	78.6±2.6	84.2±9.1	93.0±2.0	81.2±2.1	84.3	86.8
"X"-shaped [15]	65.96M	88.2±4.9	90.6±3.1	92.9±2.6	86.2±5.3	89.5	78.2±3.1	84.9±7.1	93.0±1.3	82.8±4.4	84.7	87.1
"Chilopod"-shaped	39.34M	88.2±3.1	90.3±3.1	91.4±2.8	92.2±5.8	90.5	79.3±5.1	85.8±5.3	93.1±2.3	80.7±4.7	84.7	87.6
Ours	39.34M	**88.5±3.1**	91.5±3.1	**93.1±2.1**	**93.6±4.3**	**91.7**	**80.8±3.0**	**86.5±6.5**	**93.6±1.8**	83.1±5.8	**86.0**	**88.8**
		Hausdorff Distance (avg.±std., mm)										
Payer et al. [38]	-	-	-	-	-	-	-	-	-	-	-	-
Individual	78.64M	3.70±2.86	4.41±2.22	3.80±3.08	3.56±3.43	3.87	1.87±0.99	2.91±3.59	1.66±0.62	2.48±2.47	2.23	3.05
Joint	39.32M	3.97±2.96	5.81±4.31	3.49±2.28	2.40±1.74	3.92	2.07±1.27	2.31±2.12	1.42±0.73	2.04±1.80	**1.96**	2.94
Joint+KD	39.32M	3.05±1.71	4.42±4.94	2.77±2.08	1.88±1.75	3.03	1.97±1.30	**2.29±2.19**	1.36±0.91	2.66±2.49	2.07	2.55
"Y"-shaped [9]	39.99M	2.60±1.64	3.57±2.48	2.51±1.74	2.97±6.03	2.91	2.50±1.89	4.39±3.44	1.75±0.83	1.99±1.21	2.66	2.78
"X"-shaped [15]	65.96M	2.61±1.60	3.46±3.13	2.37±1.63	2.84±7.32	2.82	2.49±1.56	3.86±2.54	2.19±5.10	**1.97±1.40**	2.63	2.72
"Chilopod"-shaped	39.34M	**2.32±1.21**	3.08±2.37	2.32±2.10	2.48±3.32	2.55	2.30±1.34	2.82±3.17	1.38±0.63	3.75±3.93	2.56	2.56
Ours	39.34M	2.38±1.72	**2.43±1.70**	**2.12±2.01**	**1.74±1.91**	**2.17**	**1.85±1.21**	2.59±3.24	**1.36±0.78**	2.30±2.31	2.02	**2.10**

TABLE II
MULTI-MODAL LEARNING RESULTS OF ABDOMINAL MULTI-ORGAN SEGMENTATION WITH 3D NETWORK UNDER DIFFERENT SETTINGS

Methods	Param Scale	Abdominal CT					Abdominal MRI					Overall Mean
		Liver	Spleen	R-kdy	L-kdy	Mean	Liver	Spleen	R-kdy	L-kdy	Mean	
		Dice Coefficient (mean±std, %)										
Individual	3832K	**93.1±1.3**	91.6±2.4	92.4±1.5	85.0±5.6	90.5	87.0±3.3	74.0±3.3	90.9±2.0	83.0±6.5	83.7	87.1
Joint	1916K	89.4±3.4	88.4±1.4	72.3±4.7	74.5±12.5	81.2	81.9±1.8	63.7±6.8	78.0±0.1	82.9±4.0	76.7	79.0
Joint+KD	1916K	84.7±3.1	83.4±3.4	88.4±2.0	82.8±5.1	84.8	83.1±3.3	62.4±8.4	88.0±0.9	74.9±6.2	77.1	81.0
"Y"-shaped [9]	2124K	90.5±2.4	92.6±1.4	92.0±2.4	84.8±8.2	90.0	89.3±3.1	82.4±2.5	89.6±0.6	80.8±8.9	85.5	87.8
"X"-shaped [15]	3389K	92.3±1.2	92.9±0.9	90.9±2.9	84.4±6.7	90.1	88.5±3.0	84.8±1.4	89.2±0.1	86.0±4.1	87.1	88.7
"Chilopod"-shaped	1919K	91.5±1.7	93.0±1.6	92.2±2.1	87.4±4.7	91.0	89.8±1.5	81.7±6.7	**91.1±1.7**	88.2±2.5	87.7	89.3
Ours	1919K	92.7±1.8	**93.7±1.7**	**94.0±0.7**	**89.5±3.9**	**92.4**	**90.3±2.8**	**87.4±1.1**	91.0±1.5	**88.3±1.7**	**89.3**	**90.8**
		Hausdorff Distance (mean±std, mm)										
Individual	3832K	**3.08±4.27**	1.99±2.29	1.31±0.88	2.96±2.98	2.34	4.11±1.87	2.54±1.19	1.49±0.63	4.45±1.21	3.15	2.74
Joint	1916K	4.46±7.34	2.01±1.81	2.93±1.49	2.76±2.61	3.04	4.77±1.82	5.88±6.18	2.84±1.12	5.08±5.56	4.65	3.85
Joint+KD	1916K	4.71±3.23	2.54±2.07	2.08±1.08	2.51±1.71	2.96	5.04±2.43	3.53±1.62	1.97±0.75	5.58±9.82	4.03	3.50
"Y"-shaped [9]	2124K	3.75±4.43	1.59±2.72	1.40±0.86	2.59±2.90	2.45	3.95±4.17	1.84±0.66	1.35±0.51	4.06±1.14	2.80	2.57
"X"-shaped [15]	3389K	3.50±4.84	1.67±1.83	1.51±1.01	2.45±2.18	2.29	4.67±2.96	2.45±2.69	1.56±0.98	**3.85±0.68**	3.14	2.71
"Chilopod"-shaped	1919K	3.68±3.92	**1.53±1.63**	1.32±0.72	1.93±1.53	2.12	3.79±4.68	1.85±0.83	**1.29±0.51**	4.33±4.08	2.81	2.47
Ours	1919K	3.27±3.74	1.59±2.09	**1.20±0.81**	**1.86±1.57**	**1.98**	**3.36±3.15**	**1.82±0.60**	1.38±0.42	4.27±3.40	**2.71**	**2.34**

- **Ours**: our full multi-modal learning scheme, i.e., using "Chilopod"-shaped architecture and KD-loss together.

C. Segmentation Results and Comparison With State-of-the-Arts

We evaluate the segmentation performance with the metrics of volume Dice coefficient (%) and surface Hausdorff distance (mm) by calculating the average and standard deviation of the segmentation results for each class, as listed in Table I and Table II respectively for the two different tasks. The mean Dice coefficient and Hausdorff distance over each modality as well as over two modalities are also presented for a straight-forward comparison. Our implemented *Individual* models are the baselines from single modality training. We compare the performance of different multi-modal learning methods, including two state-of-the-art approaches [9], [15]. We also refer to the available winning performance of the challenge [36], [38] to demonstrate effectiveness of multi-modal learning.

1) Results on Multi-Modal Cardiac Segmentation: In Table I, we see that *Joint* model obtains average segmentation Dice of 88.6% on CT and 84.0% on MRI, which are quite reasonable compared with *Individual* model (88.4% on CT and 84.8% on MRI). This indicates that networks present sufficient capacity to analyze both CT and MRI in a compact model, though the data distributions of these two modalities are very different. On top of *Joint* model, adding our proposed KD-loss can improve the segmentation performance to 89.0% on CT and 84.5% on MRI, thank to the explicit guidance from confusion matrix alignment of the distilled semantic knowledge. Next, we compare the different methods which use modality-specific parameters for multi-modal learning. We see that the models of *"Y"-shaped*, *"X"-shaped* and *"Chilopod"* generally get higher performance over *"Individual"* training. The three models employ the same segmentation loss function (i.e., combining Dice loss and cross-entropy loss), with different ways of designing modality-specific and shared parameters for feature fusion. Both [9] and [19] utilize independent encoders/decoders for each modality, while we just use modality-specific BN layers resulting in a more compact model. By further leveraging KD-loss as sort of cross-modality transductive bias, the segmentation performance is boosted

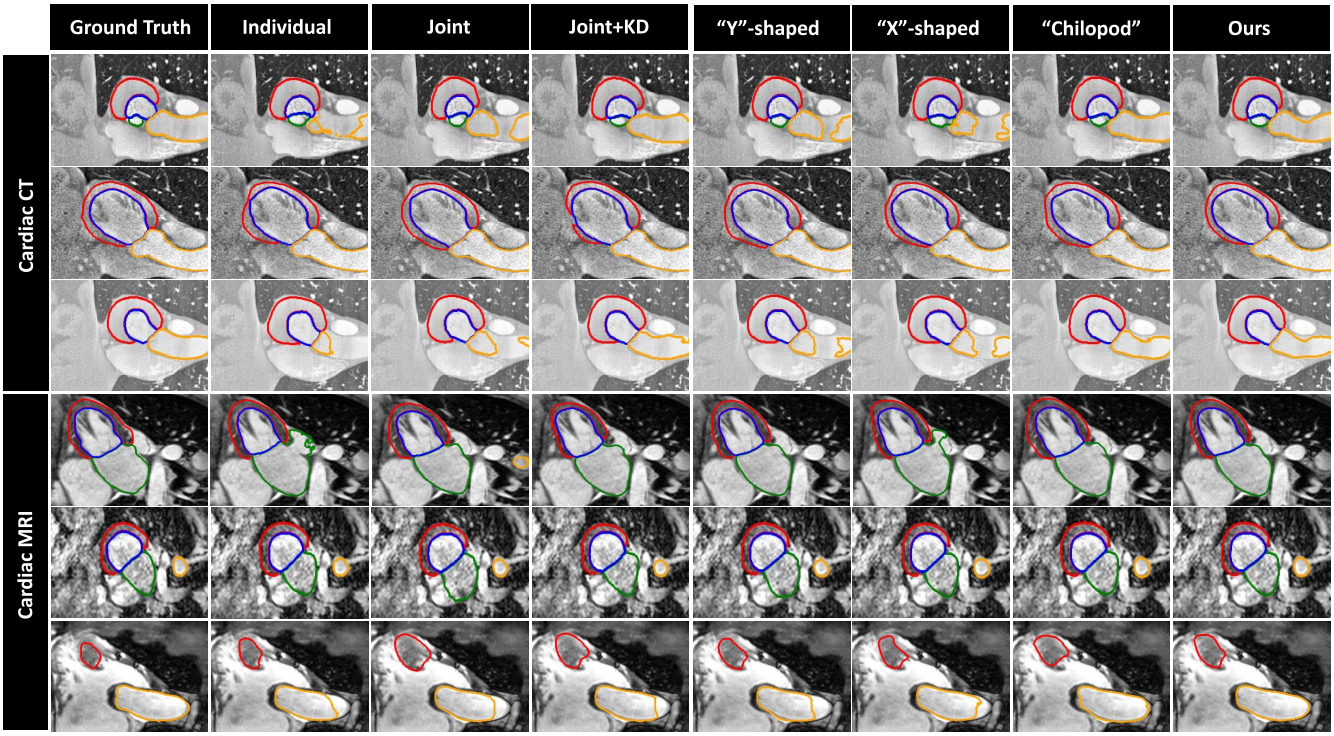

Fig. 4. Qualitative comparison of the segmentation results from the seven different settings on cardiac segmentation. The red, yellow, blue and green colors denote the boundary of LVM, AA, LVC and LAC, respectively.

to overall Dice of 88.8% (specifically, 91.7% on CT and 86.0% on MRI), exceeding our own implemented *"Indivudial"* training as well as the MICCAI-MMWHS challenge winner Payer *et al.* [38] (overall Dice of 85.5%) which used single model learning. Our approach also achieves the lowest overall mean Hausdorff distance (i.e., 2.10mm) among all the compared methods. Fig. 4 presents typical segmentation results of CT and MRI images, for a quantitative comparison of the different methods.

2) Results on Multi-Modal Abdominal Organ Segmentation: Table II lists results of multi-organ segmentation using 3D model with skip connections. When analyzing CT and MRI in a *Joint* model, the performance shows a large decrease compared with *Individual* training, i.e., overall average Dice dropping from 87.1% to 79.0% while mean Hausdorff distance increases from 2.74mm to 3.85mm. This indicates that the multi-modal shift may present more challenges when learning in 3D. Adding our KD-loss to guide the convergence towards extracting universally representative features, the segmentation Dice is improved by 2.0% (from 79.0% to 81.0%). For multi-modal learning methods, We observe that *"X"-shaped* model is superior to *"Y"-shaped* model, which is consistent with the findings in [15]. Our proposed *"Chilopod"-shaped* model, i.e., sharing all the convolution kernels but with modality-specific BN layers, achieves comparable or better performance than the *"X"-shaped* model, with a higher parameter efficiency. Further adding our KD-loss, our full multi-modal learning scheme achieves the best performance on average segmentation Dice of 90.8% and average Hausdorff distance of 2.34mm, outperforming all *Individual* and *"X"-/"Y"-shaped* models by a large margin. The Dice of challenge winners for the two datasets used in this multi-organ segmentation task are currently not available. Last but not least, it is worth noting that, *Ours* significantly improves the segmentation performance on MRI over *Individual* model by 5.6% (from 83.7% to 89.3%). This demonstrates that our method can effectively improve the performance on the modality with fewer training samples, by leveraging multi-modal learning. We notice that such improvement mainly comes from two organs: spleen and left kidney. These are located nearby and have similar appearance in MRI and pose challenges in the context of data scarcity. Multi-modal learning seems beneficial by enhancing high-level inter-class relationship alignment. The Fig. 5 presents typical segmentation results of CT and MRI images, for a quantitative comparison of these different methods.

3) Statistical Analysis on Significance: We have computed p-values using Student's t-tests when comparing our segmentation results with other methods, with numbers given in Table III for 2D cardiac segmentation and Table IV for 3D abdominal organ segmentation. It is observed that we get $p < 0.05$ in all settings, indicating a significant improvement for our approach. The statistical tests are conducted by jointly considering both results of CT and MRI in each setting.

4) Analysis on Parameter Efficiency: A benefit of our multi-modal learning network is its high compactness. With careful internal normalization of features, we can make more sufficient use of the remarkable capacity inherent in neural networks. We compute the parameter scales of all the models for our implemented seven different settings, as listed in the second column of Table I and Table II. A *Joint* model has 39.32M

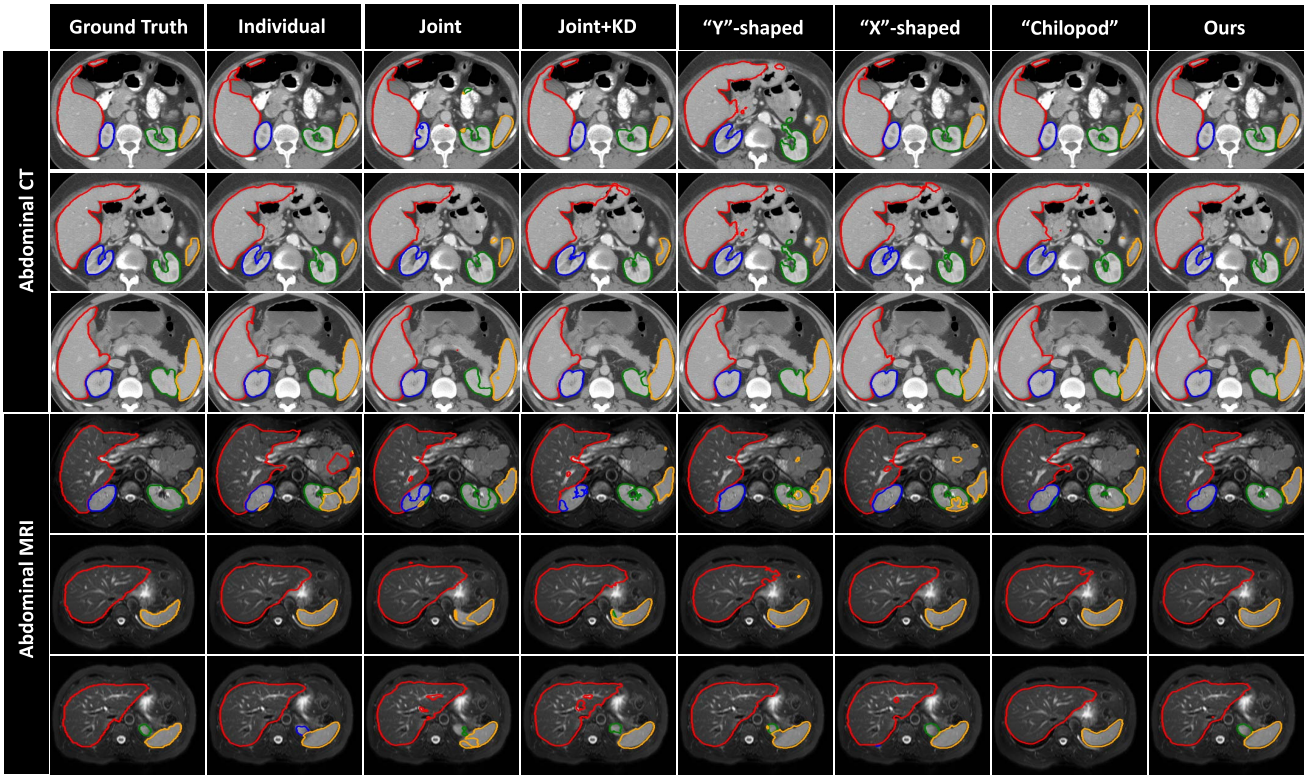

Fig. 5. Qualitative comparison of the segmentation results from the seven different settings on abdominal multi-organ segmentation. The red, yellow, blue and green colors denote the boundary of liver, spleen, R-kdy and L-kdy (in human-body view), respectively.

TABLE III
P-VALUES FOR STATISTICAL ANALYSIS OF ALL SETTINGS FOR OUR PROPOSED METHOD ON 2D CARDIAC SEGMENTATION

Metrics	Individual	Joint	Joint+KD	"Y"-shaped [9]	"X"-shaped [15]	"Chilopod"
Dice coefficient	0.015	6e-4	0.003	0.006	0.002	0.029
Hausdorff distance	0.022	0.036	0.035	0.044	0.035	7e-4

TABLE IV
P-VALUES FOR STATISTICAL ANALYSIS OF ALL SETTINGS FOR OUR PROPOSED METHOD ON 3D ORGAN SEGMENTATION

Metrics	Individual	Joint	Joint+KD	"Y"-shaped [9]	"X"-shaped [15]	"Chilopod"
Dice coefficient	0.009	4e-5	3e-4	0.015	0.005	0.003
Hausdorff distance	0.006	2e-9	5e-7	0.039	0.043	0.036

parameters for 2D model and 1916K parameters for 3D model. Our 2D model has more parameters because it is much deeper and wider than the 3D model. With *individual* training, we need double of the parameter scales (i.e., 78.64M for 2D and 3832K for 3D), since each single modality has its own model separately. Using *Joint+KD* adds no extra parameters, while the distilled knowledge helps stimulate underlying cross-modality information. The conventional *"Y"-shaped* and *"X"-shaped* multi-modal learning schemes consist more parameters due to their modality-specific encoders/decoders. Specifically, the *"X"-shaped* model almost doubles the parameter scale, for the cost of using more modality-specific layers. In comparison, our proposed multi-modal learning schemes (*"Chilopod"* and *Ours*) only need marginally extra parameters for separate internal feature normalization with BN (parameterized by a set of feature channel tied scalars $\{\gamma, \beta\}$. Specifically, our method only adds 0.02M for 2D network and 3K parameters for 3D network, which is quite parameter efficient compared with existing multi-modal schemes.

D. Analytical Ablation Studies

1) Different Internal Normalization Layers: We demonstrate that the effectiveness of separate internal feature normalization is agnostic to different ways of grouping features for S_g^k. Three additional popular feature normalization methods (i.e., Instance Norm [31], Layer Norm [32], and Group Norm [33]) are implemented for five ablation settings of our method, using the cardiac segmentation dataset. The results are listed in Table V. *Ours* consistently presents superior performance over *Individual* learning for all popular feature normalization methods. Comparing the results of *"Joint"* models under these different normalization methods, we notice that Instance Norm has the best average Dice, achieving 87.5%. This indicates that more refined internal feature normalization benefits multi-modal learning under great parameter sharing. This also meets our hypothesis on normalizing multi-modal images separately in our compact model.

2) Different Weights of KD-Loss: We vary the trade-off weight of our KD-loss, to analyze its sensitivity to the hyper-parameter of α in Eq. 5. Specifically, we range $\alpha \in [0.1, 0.9]$ at a step of 0.1, and observe our multi-modal segmentation performance on the cardiac dataset. In Fig. 6, the box-plots present the mean of segmentation Dice across all classes

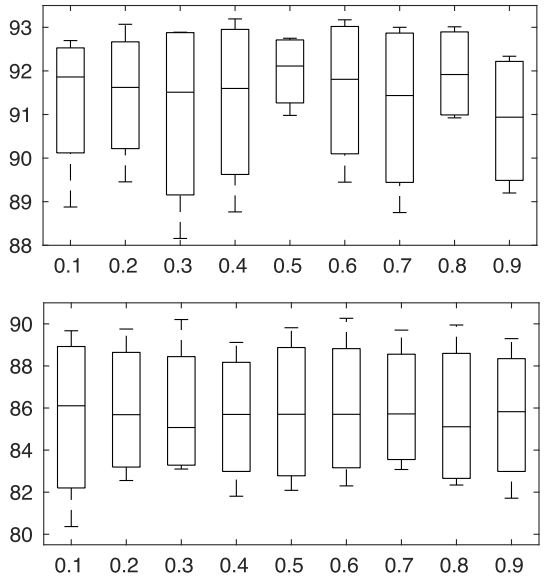

Fig. 6. Box-plots of Dice (average over all structures) on CT (top) and MRI (bottom) for cardiac segmentation, when varying α from 0.1 to 0.9.

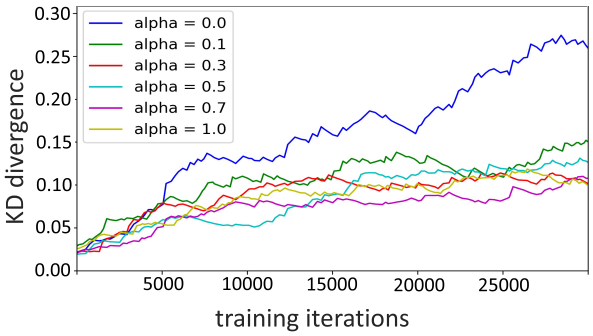

Fig. 7. Comparison of learning curves for KD-loss at different values of α.

on CT (top) and MRI (bottom). Note that the setting of $\alpha = 0$ corresponds to the seventh and twelfth columns in Table I. We observe that our method can generally improve the segmentation performance over baselines (e.g., 88.4% for CT and 84.8% for MRI at *"Individual"* training), while not being very sensitive to the value of α.

3) Learning Curve of KD-Loss: In Fig. 7, we present the learning curve of KD-loss computed on test data at multi-organ segmentation dataset. We observe that without constraints (i.e., $\alpha = 0$), the distilled knowledge from two modalities diverges. By activating KD-loss, the $\mathcal{L}_{\text{kd}}$ is stabilized, reflecting that the probability distributions across classes are better aligned. This observation also explains the performance gain from using the KD-loss as guidance of high-level semantic alignment.

4) Evolution of Prediction Alignment Between Modalities: In Fig. 8, we visualize the evolution of the confusion matrices (i.e., $\mathbf{q}^a$ and $\mathbf{q}^b$) for both modalities, from the beginning of training until model convergence. To more clearly observe alignment of these matrices (i.e., Eq. 4), we compute their absolute difference plane by abstracting one from the other, as illustrated in the bottom row. It is observed that when the model is randomly initialized, the confusion matrices are

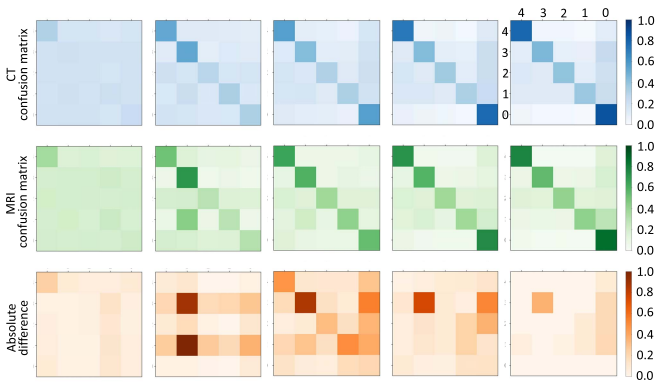

Fig. 8. Visualization of the evolution of the alignment of prediction confusion matrices between CT (first row) and MRI (second row) modalities as training going on. For better illustration of differences, the third row shows their absolute difference planes (with values amplified by a five-times scaling). The classes of $4, 3, 2, 1, 0$ respectively correspond to liver, spleen, right kidney and left kidney in our 3D multi-organ segmentation task.

invariable with no much difference between CT and MRI. As training goes on, the model starts learning and gradually produces meaningful confusion planes with stronger diagonal responses. Notably, with the model converging, the difference plane between CT and MRI returns to clean again, but for the reason of successful alignment of their confusion matrices.

V. Discussion

This work tackles the challenging task of multi-class segmentation on unpaired CT and MRI images. The difficulties mainly arise from the significant distribution shift and absence of registration between the two modalities. This problem has not been well studied so far, compared with segmentation on paired multi-sequence of MRI images. Valindria *et al.* [15] is the current state-of-the-art study on this topic, with empirical results demonstrating that leveraging cross-modality information is superior to single model learning. To more explicitly utilize cross-modality knowledge, we present a novel approach by distilling the semantic representations in a high-level within a compact model. Notably, our multi-modal learning scheme is architecture-independent, therefore can be easily integrated into various existing 2D and 3D CNN models. Our approach is developed for unpaired multi-modal learning, and it does not make use of any information which relies on image alignment. In this regard, we think that matching the distributions of the predictions on *aligned images* would not bring much extra help to our approach as it is. Nevertheless, if such aligned images are available, our method is still directly applicable, and could be extended to leverage the fine-grained pixel-wise alignment.

The remarkable capacity of deep neural networks motivates our design of a highly compact model for multi-modal learning. Bilen and Vedaldi [20] show that sharing all kernels (even including the final classifier) while using domain-specific batch normalization can work reasonably well for different easy tasks, e.g., MNIST and CIFAR-10. In medical imaging, Moeskops *et al.* [39] build a single network for different segmentation tasks in different modalities. Karani *et al.* [21]

TABLE V
MULTI-MODAL CARDIAC SEGMENTATION WITH A 2D NETWORK BY USING DIFFERENT FEATURE NORMALIZATION LAYERS (DICE, %)

Normalization Type	Methods	Cardiac CT					Cardiac MRI					Overall
		LVM	LAC	LVC	AA	Mean	LVM	LAC	LVC	AA	Mean	Mean
Instance Normalization	Individual	87.5±3.2	90.2±3.8	92.1±2.4	87.4±14.7	89.3	80.8±3.3	86.6±5.4	92.7±1.7	82.6±5.6	85.7	87.5
	Joint	86.6±2.4	88.7±2.9	91.3±2.9	92.7±4.6	89.9	82.1±3.1	83.7±8.7	93.9±2.0	80.7±6.6	85.1	87.5
	Joint+KD	86.7±1.8	90.8±3.7	91.7±3.1	92.4±6.6	90.4	80.9±4.3	85.4±7.3	93.9±2.1	82.0±7.4	85.5	88.0
	"Chilopod"	86.5±2.6	91.5±3.0	91.7±3.2	94.2±4.3	91.0	81.2±4.7	85.4±6.8	94.1±1.7	81.0±3.7	85.4	88.2
	Ours	87.3±2.1	90.9±2.8	92.6±3.0	94.0±4.0	91.2	82.4±3.1	85.6±5.6	94.1±2.0	82.6±5.5	86.2	88.7
Layer Normalization	Individual	87.8±3.9	89.9±2.7	92.7±2.2	90.2±9.2	90.2	79.2±4.3	86.2±4.5	91.4±2.9	83.9±3.1	85.2	87.7
	Joint	83.0±4.5	88.2±2.0	90.2±2.9	92.2±1.5	88.4	76.6±3.5	83.3±6.6	92.6±2.1	77.3±6.4	82.5	85.4
	Joint+KD	86.1±3.0	90.7±3.6	92.4±3.6	94.6±4.5	90.9	79.1±3.9	85.8±6.7	92.3±1.9	81.6±5.7	84.7	87.8
	"Chilopod"	86.6±1.9	90.9±3.2	91.5±2.9	94.2±1.5	90.8	78.2±3.4	85.0±7.1	92.7±1.8	80.3±2.7	84.1	87.4
	Ours	85.8±4.0	90.6±3.1	92.3±2.5	94.3±2.7	90.7	79.4±4.2	86.2±6.3	92.5±2.0	82.0±5.7	85.0	87.9
Group Normalization	Individual	88.1±3.7	89.8±3.0	92.4±2.1	84.1±10.5	88.6	79.0±3.8	84.5±5.5	91.5±4.6	82.0±5.4	84.2	86.4
	Joint	86.8±3.0	89.5±2.7	91.6±2.6	91.0±3.8	89.7	79.7±3.2	85.0±5.5	92.7±2.2	79.0±6.2	84.1	86.9
	Joint+KD	85.9±2.6	90.3±3.5	91.6±3.5	93.6±4.2	90.4	80.8±4.1	86.8±6.2	93.4±2.1	82.0±5.2	85.7	88.0
	"Chilopod"	86.1±4.7	91.8±2.0	91.8±4.2	93.6±3.0	90.8	78.5±5.2	85.2±7.6	92.5±1.3	79.1±6.9	83.8	87.3
	Ours	88.4±2.8	90.4±3.1	92.3±2.4	93.7±4.1	91.2	81.8±1.5	85.0±6.8	93.9±1.5	81.9±4.2	85.6	88.4

successfully adapt an MRI brain segmentation model to different scanners/protocols by only fine-tuning the BN layers. Inspired by these findings, we argue that a single shared network has the potential to work well on very different multi-modal data with similar structures (e.g., CT and MRI) requiring only a few modality-specific parameters. Separate internal normalization may be sufficient to unleash the potential model capacity, not necessarily using a separate encoder for each modality. An explicit regularization loss towards cross-modality semantic alignment helps further stimulate the model capacity.

We have conducted case study regarding an outlier case with image artefacts in the abdominal CT dataset. As illustrated in Fig. 9, the outlier case presents clear artefacts with worse image quality compared with other general cases, noting that all the training cases are good-quality images without artefacts. The mean results of the four abdominal organs using seven different settings are listed in Table VI, where we see that all comparison methods obtained a lower performance on this one case (compared with general results in Table II). Our approach achieved a higher Dice score with a smaller Hausdorff distance compared with the other methods, demonstrating our superior robustness at lower image quality.

One limitation of this paper lies in the plain network architectures used for multi-modal segmentation. The employed 2D dilation network for cardiac segmentation and 3D U-Net for abdominal multi-organ segmentation are relatively basic, compared with more complicated network designs for multi-modal learning. This is reasonable as we currently focus on studying separate feature normalization and knowledge distillation between modalities, but this may limit the segmentation accuracy. We plan to integrate our multi-modal learning scheme into more well-designed networks (e.g., consisting multi-scale feature fusion) in our future work, seeking more accurate segmentation of the images with multi-modal learning. This extension is quite natural, thanks to the flexibility of our proposed multi-modal learning scheme.

The characteristics of modality differences may play an important role affecting the performance of multi-modal learning methods. It would be interesting to explore further whether there are limitations on certain functional or statistical

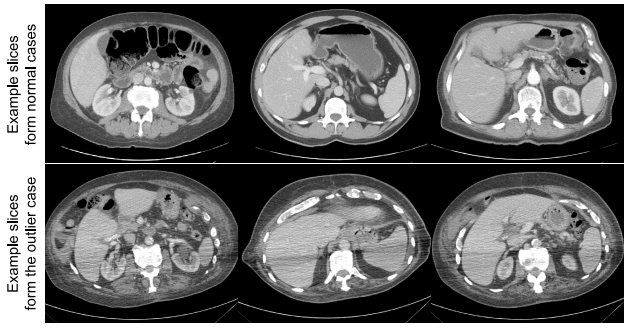

Fig. 9. Illustration of example CT slices from general cases and an outlier case with artefacts, in the task of multi-organ segmentation in abdominal data.

TABLE VI
MEAN RESULTS OF THE FOUR ORGANS IN ABDOMINAL CT ON THE OUTLIER CASE USING DIFFERENT METHODS

Metrics	Individual	Joint	Joint+KD	"Y"-shaped [9]	"X"-shaped [15]	"Chilopod"	**Ours**
Dice coefficient	73.7	74.4	78.0	69.8	76.6	76.4	**81.1**
Hausdorff distance	4.19	4.93	4.23	5.83	5.32	4.89	**3.78**

relationships between intensity distributions that cannot be handled by our proposed approach. The fact that it works for CT and MRI is encouraging that our approach may work for a large family of relationships. In preliminary synthetic experiments, we could also confirm that an anti-correlation between intensity (i.e., one modality is the inverse of the other) does not pose a problem. Regarding differences in image resolution between modalities, we would expect this not to impact the method too much due to the statistical approach of distribution matching. However, this would need to be confirmed for more extreme cases where one of the modalities is of much lower resolution.

In general, our proposed method of separate internal feature normalization and knowledge distillation loss can be applied to many other situations, when we use data with a mixture of distributions or domains. For instance, it can be used for model learning by aggregating images acquired from different clinical sites in real world scenarios. The data from different sites can be separately normalized within a network, so that one can make better use of multiple data sources. The knowledge

distillation loss derived from semantic representations can be applied for domain adaptation problems by leveraging alignment of inter-class relationships. We will explore these extensions in future work.

VI. Conclusion

We present a novel multi-modal learning scheme for unpaired CT and MRI segmentation, with high parameter efficiency and a new KD-loss term. Our method is general for multi-modal segmentation tasks, and we have demonstrated its effectiveness on two different tasks and both 2D and 3D network architectures. Integrating analysis of multi-modal data into a single parameter efficient network helps to ease deployment and improve usability of the model in clinical practice. Moreover, our KD-loss encouraging robust features has a potential to tackle model generalization challenges in medical image segmentation applications.

References

[1] K. Nikolaou, H. Alkadhi, F. Bamberg, S. Leschka, and B. J. Wintersperger, "MRI and CT in the diagnosis of coronary artery disease: Indications and applications," *Insights Imag.*, vol. 2, no. 1, pp. 9–24, Feb. 2011.

[2] R. Karim *et al.*, "Algorithms for left atrial wall segmentation and thickness–Evaluation on an open-source CT and MRI image database," *Med. Image Anal.*, vol. 50, pp. 36–53, Dec. 2018.

[3] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446–455, Apr. 2018.

[4] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, Mar. 2015.

[5] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. N. L. Benders, and I. Isgum, "Automatic Segmentation of MR Brain Images With a Convolutional Neural Network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[6] L. Fidon *et al.*, "Scalable multimodal convolutional networks for brain tumour segmentation," in *Proc. MICCAI*, 2017, pp. 285–293.

[7] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[8] S. Valverde *et al.*, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, Jul. 2017.

[9] D. Nie, L. Wang, Y. Gao, and D. Shen, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1342–1345.

[10] X. Li *et al.*, "3D multi-scale FCN with random modality voxel dropout learning for Intervertebral Disc Localization and Segmentation from Multi-modality MR Images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018.

[11] S. Wang, K. Burtt, B. Turkbey, P. Choyke, and R. M. Summers, "Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research," *BioMed Res. Int.*, vol. 2014, Dec. 2014, Art. no. 789561.

[12] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[13] J. Dolz, C. Desrosiers, and I. B. Ayed, "IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet," in *Proc. Int. Workshop Challenge Comput. Methods Clin. Appl. Spine Imag.*, 2018, pp. 130–143.

[14] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, "MMAN: Multi-modality aggregation network for brain segmentation from MR images," *Neurocomputing*, vol. 358, pp. 10–19, Sep. 2019.

[15] V. V. Valindria *et al.*, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 547–556.

[16] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.

[17] Q. Dou *et al.*, "PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99065–99076, 2019.

[18] Y. Huo *et al.*, "SynSeg-Net: Synthetic segmentation without target modality ground truth," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1016–1025, Apr. 2019.

[19] G. Van Tulder and M. De Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 638–648, Feb. 2019.

[20] H. Bilen and A. Vedaldi, "Universal representations:The missing link between faces, text, planktons, and cat breeds," 2017, *arXiv:1701.07275*. [Online]. Available: https://arxiv.org/abs/1701.07275

[21] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain MR segmentation across scanners and protocols," in *Proc. MICCAI*, 2018, pp. 476–484.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: https://arxiv.org/abs/1503.02531

[23] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. ICCV*, Dec. 2015, pp. 4068–4076.

[24] S. Hou, X. Pan, C. Change Loy, Z. Wang, and D. Lin, "Lifelong learning via progressive distillation and retrospection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 437–452.

[25] N. Papernot, P. Mcdaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[26] S. H. Lee, D. H. Kim, and B. C. Song, "Self-supervised knowledge distillation using singular value decomposition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 339–354.

[27] H. Wang *et al.*, "Segmenting neuronal structure in 3D optical microscope images via knowledge distillation with teacher-student network," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 228–231.

[28] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation," 2019, *arXiv:1901.09263*. [Online]. Available: https://arxiv.org/abs/1901.09263

[29] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 76–84, Jan. 2017.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: https://arxiv.org/abs/1607.08022

[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: https://arxiv.org/abs/1607.06450

[33] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[34] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[35] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[36] X. Zhuang *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge," 2019, *arXiv:1902.07880*. [Online]. Available: https://arxiv.org/abs/1902.07880

[37] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. (2015). *2015 MICCAI Multi-Atlas Labeling Beyond the Cranial Vault Workshop and Challenge*. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/89480

[38] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNs and anatomical label configurations," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, 2017, pp. 190–198.

[39] P. Moeskops *et al.*, "Deep learning for multi-task medical image segmentation in multiple modalities," in *Proc. MICCAI*, 2016, pp. 478–486.