A Simple and Robust Framework for Cross-Modality Medical Image Segmentation applied to Vision Transformers

Matteo Bastico¹, David Ryckelynck¹, Laurent Corté¹, Yannick Tillier², Etienne Decencière³

Mines Paris, Université PSL
¹Centre des Matériaux (MAT), UMR7633 CNRS, 91003 Evry, France
²Centre de Mise en Forme des Matériaux (CEMEF), UMR7635 CNRS, 06904 Sophia Antipolis, France
³Centre de Morphologie Mathématique (CMM), 77300 Fontainebleau, France

{matteo.bastico, etienne.decenciere}@minesparis.psl.eu

Abstract

When it comes to clinical images, automatic segmentation has a wide variety of applications and a considerable diversity of input domains, such as different types of Magnetic Resonance Images and Computerized Tomography scans. This heterogeneity is a challenge for cross-modality algorithms that should equally perform independently of the input image type fed to them. Often, segmentation models are trained using a single modality, preventing generalization to other types of input data without resorting to transfer learning techniques. Furthermore, the multi-modal or cross-modality architectures proposed in the literature frequently require registered images, which are not easy to collect in clinical environments, or need additional processing steps, such as synthetic image generation. In this work, we propose a simple framework to achieve fair image segmentation of multiple modalities using a single conditional model that adapts its normalization layers based on the input type, trained with non-registered interleaved mixed data. We show that our framework outperforms other cross-modality segmentation methods, when applied to the same 3D UNet baseline model, on the Multi-Modality Whole Heart Segmentation Challenge. Furthermore, we define the Conditional Vision Transformer encoder, based on the proposed cross-modality framework, and we show that it brings significant improvements to the resulting segmentation, up to 6.87% of Dice accuracy, with respect to its baseline reference. The code to reproduce our experiments and the trained model weights are publicly available at https://github.com/matteo-bastico/MI-Seg.

1. Introduction

Medical image segmentation is one of the most challenging and studied research fields of deep learning and artificial intelligence in recent years [18, 28, 41, 54, 66], as shown in surveys [45, 48]. The precision of segmentation achieved competitive results thanks to recent efforts in developing convolution [33, 41] and transformer-based neural networks [17, 18]. These algorithms are generally trained on a single medical imaging modality, such as T1- or T2-Weighted Magnetic Resonance Images (MRIs) or Computed Tomography (CT) scans. Therefore, they often suffer from data variability when tested on images different from those seen during training. Data variability is a common issue in the medical field and it is due to several uncontrollable factors that depend on the specific clinic or patient needs [47], e.g., different imaging methods, different scanners, different acquisition settings or patient limitations. In this scenario, the most straightforward approach would be to train a separate model for each domain, requiring a massive amount of different annotated training data and not exploiting inter-domain information.

Recently, to tackle this issue, some multi-modal [58] and cross-modality [28, 54, 63, 66] medical image segmentation techniques have been proposed. The first group of methods aims to produce better segmentation by simultaneously exploiting information from multiple sources. In this case, it is good practice to stack several images of different types to generate a combined input for the segmentation model [65]. Furthermore, multi-modality can also be leveraged to produce synthetic images of missing domains at inference time [8]. Here, different modalities are rather seen as separate inputs for the model that generates a common representation and produces synthetic images. However, in both scenarios, registered medical images [20] are needed. Although registration is relatively simple, the acquisition of several images of the same patient is often limited by local resources and time constraints. The second group of methods, i.e. cross-modality methods, focus on producing segmentation independently of the data type provided in input.
More specifically, an assistant modality, of which several annotated data are available, is often used to improve the segmentation performance on another modality with fewer annotated data, i.e., target modality. Fine-tuning [50] or transfer learning [40] are straightforward ways of adapting deep learning models to target modalities after they have been trained on some specific assistant modality. However, these approaches do not fully exploit shared cross-modality information. One intuitive way to overcome this problem is joint training [54], where a model is trained simultaneously with multiple modalities. Nevertheless, it is relatively hard to directly learn common features when the domain shift is significant, e.g., MRIs and CTs. Therefore, to directly learn from multi-modality data with large appearance discrepancy, the use of several feature extractors, i.e., one per modality, has been first proposed in an X-shaped architecture [54], implying a significant overhead on the segmentation model and requiring adjustment to generalize to other clinical tasks. To address this issue, several techniques based on synthetic image generation prior to segmentation have been introduced [7, 27, 66]. In many cases, the assistant modality is used to synthesize the target, or vice versa, and train the model on a mix of real and synthetic data [28]. The additional pre-processing step required by these methods to make the clinical images segmentation-ready adds computational complexity, which is often critical in real-time applications such as computer-assisted surgery [51].

In this work, we propose a general framework, applicable to any encoder-decoder literature architecture, aiming to produce high-quality cross-modality segmentation without introducing overhead on the segmentation model nor needing registered clinical images for training. With our framework, medical images from different domains can be fed individually directly into a single modality-conditioned model, which generates the desired segmentation by self-adapting its encoder normalization layers. The adaptation is based on Conditional Instance Normalization (CIN), which has been originally proposed for the transfer of arbitrary artistic styles by Dumoulin et al. [11]. Thanks to that, and similarly to joint training, our models can be trained in an end-to-end way. By randomly mixing different data of several modalities, in what we call interleaved mixed training fashion, we get rid of prior synthetic image style transfer. Furthermore, based on the proposed framework, we formally define the Conditional Vision Transformer (C-ViT) encoder architecture to build modality-agnostic ViT-based models [10] for image segmentation or classification, as shown in Figure 1. We extensively evaluate our method on the Multi-Modality Whole Heart Segmentation (MM-WHS) 2017 Challenge [67]. We show that C-ViT improves the baseline model and other multi-modal techniques and that our framework outperforms other cross-modality segmentation methods, when applied to the same baseline model. Our contributions are summarized as follows:

• We present a novel framework to enhance the quality of cross-modality segmentation, while removing expensive input pre-processing steps and the need of registered data for training.

• Following the framework, we introduce the Conditional Vision Transformer encoder, which adapts its normalization layers based on the input modality and can be used for several tasks.

• We show that the proposed framework outperforms other cross-modality learning techniques both on UNet and ViT baselines.

2. Related Works

Multi-Modality Learning in Medical Imaging. In recent years, several deep learning architectures have been proposed for image segmentation, achieving extremely good performance [6, 17, 18, 35, 43]. Among them, UNet [41] is surely the most popular and is used as a baseline to create better performing models [45]. More recently, some variation based on vision transformers (ViT) [10] and swin-transformers [31], such as TransUNET [6], UNETR [18] and Swin-UNETR [17], have been shown to outperform previous versions of UNet. In the clinical imaging

Figure 1: Overview of the proposed cross-modality clinical image segmentation technique applied to the Swin-UNETR segmentation model [17]. The input modality can be arbitrarily switched to obtain the desired segmentation, while keeping the model unchanged. In conditional models, the modality is also used as input to activate the encoder normalizations corresponding to the data type and generate common latent spaces.
field, multi-modality learning is a similar, yet challenging, task. Multiple medical image domains have been exploited for synthetic image generation based on Generative Adversarial Networks (GAN) [8, 9, 15, 49, 61] or multi-modal image segmentation [13, 16, 58, 65]. Interestingly, the two tasks have also been merged by Zhang et al. [62] in a cross-task feedback fusion GAN that first generates synthetic CT and then performs multi-modal segmentation. The latter uses cross-domain information to achieve better performances, but it is restricted to when registered images are available. To relax the need of aligned medical images and exploit inter-domain features during training, cross-modality segmentation approaches have recently gained attraction. Firstly, Zheng et al. [64] investigated the effectiveness of shape priors learned from an assistant modality to improve the segmentation on a target modality by marginal space learning. Valindria et al. [54] developed dual-stream encoder-decoder models having different branches for each modality and implementing weights sharing techniques to extract cross-modality features. Similarly to our work, some researchers tried to play with the normalization layers of the models to improve their generalization capacities [12, 38, 42, 44, 66]. Among others, Pan et al. [38] introduced the IBN-Net which simultaneously exploiting Instance and Batch Normalization to capture both appearance changes and content information. Segu et al. [42] proposed to collect prior domain-dependent statistics by training ad-hoc Batch Normalization (BN) to map the modalities on a shared latent space.

With the aforementioned advances in synthetic images generation, many works have tried to tackle the challenge by assisting segmentation models with prior image translation [7, 22, 27, 28, 63, 66]. Among them, Zhang et al. [63] tried to improve the segmentation for modalities with limited training samples by easing the appearance gap with other modalities using a GAN. Similarly, Li et al. [28] proposed an Image Alignment Module to reduce the appearance gap between assistant and target modality and introduced a Mutual Knowledge Distillation scheme [21] to exploit modality-shared knowledge. More recently, Zhou et al. [66] proposed to simulate the possible appearance changes of a target domains by non-linear transformation to augment source-similar and source-dissimilar images.

Different from these methods, which mainly rely on image translation or prior trainings, our framework tries to shrink the model complexity and, at the same time, improve the segmentation accuracy extracting all the available cross-modality features by conditioning the segmentation models and training them jointly with assistant and target modality. Moreover, some of the previous techniques [28, 64] do not allow for arbitrary switch of the input modality, but they just exploit assistant modality information to improve the target segmentation.

**Normalization Layers.** Our work is also related with normalization layers used in neural networks, which may differ based on the spatial dimension to which they are applied. The four main classes of normalization layers are batch, layer, group, and instance normalization. Batch Normalization was the first technique introduced to deal with the variability of layers input during training by normalizing the features across mini-batches [24]. Layer Normalization (LN) [2], used mainly in Natural Language Processing (NLP) and Recurrent Neural Networks (RNNs), was introduced to relax the dependency on the batch, normalizing only across features. Similarly, Group Normalization [60] computes the mean and standard deviations in groups of channels rather than using the entire feature space. Finally, Instance Normalization (IN) has been introduced, in the imaging context, to prevent instance-specific mean and covariance shift, simplifying the learning process [53]. Huang et al. [23] proposed a further variation of IN, called Adaptive IN to perform arbitrary image style transfer, that is, synthetic image generation, by adjusting the mean and variance of the content input to match those of the desired style input. The latter does not have learnable parameters, but it computes them directly from the input style. In our framework, we propose to use a similar variation of IN recently proposed by Dumoulin et al. [11], called Conditional Instance Normalization (CIN), which learns a different set of parameters $\gamma_m$ and $\beta_m$ for each style $m \in [1, M]$. Given a three-dimensional volume with $C$ channels, $z \in \mathbb{R}^{C \times H \times W \times D}$, of modality $m$, the CIN is calculated as

$$CIN(z,m) = \gamma_m \left( \frac{z - \mu(z)}{\sigma(z)} \right) + \beta_m \tag{1}$$

where $\mu(z)$ and $\sigma(z)$ are computed across spatial dimensions independently for each channel and each sample [23]. It allows the sharing of all convolutional weights of a style transfer network across many image styles by only changing the learnable normalization parameters. Inspired by that, in this work, CIN is exploited to generate a shared latent space for different input modalities, which is used to obtain the segmentation output independently of the medical input image domain.

**3. Method**

The overview of the proposed cross-modality segmentation framework is shown in Figure 2. The segmentation model, which can be chosen among those available in the literature, is conditioned by the input data modality to adapt its normalization layers to the fed clinical image. Therefore, the aim is to produce the desired segmentation independently of the input data domain.

In the following, the general framework for creating conditional models is introduced and it is applied to the transformer encoder, to obtain the C-ViT encoder. Finally, more
The conditional model can then be expressed as

\[ \text{encoder}(\cdot), \]

where \( E(\cdot, m) \) generates \( \mathcal{C} \) common latent spaces, \( \{\mathcal{C}^1, \ldots, \mathcal{C}^S\} \), one for each multi-resolution step of the encoder, adapting itself to the input modality \( m \). The decoder, \( D(\cdot) \), is unique and does not need adaptation.

details about the framework training procedure, including the loss function, are reported.

**Conditional Segmentation Models.** Let any state-of-the-art encoder-decoder segmentation architecture be considered as baseline for the proposed framework. In a straightforward case, for a given data modality \( m \in [1, M] \), let \( E_m(\cdot) : \mathcal{I}_m \rightarrow \mathcal{H}_m \) and \( D_m(\cdot) : \mathcal{H}_m \rightarrow \mathcal{O} \) be the two functions that describe the encoder and decoder modules of the chosen architecture, respectively. The role of \( E_m(\cdot) \) is to transform the input, \( x_m \in \mathcal{I}_m \), into a latent feature space, \( \mathcal{H}_m \), which can better represent the information carried by raw data. On the other hand, the purpose of \( D_m(\cdot) \) is to transform this representation into the final segmentation \( y \in \mathcal{O} \). In general, the space \( \mathcal{H}_m \) is learned ad-hoc by the model, based on the type of data used during training.

To make the model independent of the input modality, we propose to create a general conditional encoding module, \( E(\cdot, m) : \mathcal{I}_m \rightarrow \mathcal{C} \), which transforms any input \( x_m \in \mathcal{I}_m, 1 \leq m \leq M \), into a common shared latent feature space \( \mathcal{C} \). In this way, a single decoder \( D(\cdot) : \mathcal{C} \rightarrow \mathcal{O} \) is needed to generate the segmentation of all the modalities. The relation between any input \( x_m \) and the output \( y \) in a conditional model can then be expressed as

\[ y = D(E(x_m, m)), \quad 1 \leq m \leq M \]  

(2)

More generally, in an architecture with skip connections, e.g. UNet [41], the encoder may be seen as a sequence of \( S \) sub-modules, each corresponding to a different resolution step. In this case, to obtain robust cross-modality segmentation, the framework is applied to each step, generating a set of multi-resolution common latent spaces, \( \mathcal{C}^s, 0 \leq s \leq S \), as shown in Figure 2. Therefore, the decoder will simply gather the data from all the shared latent spaces to generate a modality-agnostic segmentation output.

In practice, we propose to generate the conditional encoder \( E(\cdot, m) \), starting from any simple encoder \( E_m(\cdot) \), by adding the data modality as input (often available from meta-data, without manual intervention) and replacing all its normalization layers with CIN. In such way, the only overhead on the segmentation model is given by the extra normalization learnable parameters, \( \gamma_m \) and \( \beta_m \), \( 1 \leq m \leq M \), of Equation 1, which are negligible with respect to the rest of the architecture. Furthermore, we avoid any pre-processing phase, such as synthetic image generation through GANs, which can heavily impact on the network performances, and we let the model directly learn a common shared latent space for all the input modalities, trying to fully exploit the inter-modality information.

**Conditional Vision Transformer.** The proposed C-ViT encoder is shown in Figure 3. As originally proposed by Vaswani et al. [55], it has two sub-layers, the first is a multi-head self-attention mechanism (MSA), and the second is a simple, position-wise fully connected Multi-Layer Perceptron (MLP). Both sub-layers are preceded by a normalization [10, 56] and a residual connection is employed around each of them. Here, we replace the Layer Normalization (LN) [2], which has historically been used in transformers, with CIN. In other words, a switch mechanism that selects the IN corresponding to the input modality is introduced before each sub-layer.

In a general scenario with \( M \) different input modalities,
Table 1: Quantitative comparison with other methods for cross-modality medical image segmentation on the target modality (CT). All the techniques have the same U-Net baseline [54], are trained using 20 MRIs and 10 CTs and are evaluated on the test set of MM-WHS dataset. The mean DICE score is reported, as well as the ones for all the heart substructures.

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg Dice</th>
<th>Dice of substructures of heart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MYO</td>
<td>LA</td>
</tr>
<tr>
<td>Baseline</td>
<td>87.06</td>
<td>87.02</td>
</tr>
<tr>
<td>Fine-tune</td>
<td>87.69</td>
<td>87.16</td>
</tr>
<tr>
<td>Joint-training</td>
<td>87.43</td>
<td>86.65</td>
</tr>
<tr>
<td>X-shape [54]</td>
<td>87.67</td>
<td>87.19</td>
</tr>
<tr>
<td>Zhang et al. [63]</td>
<td>88.50</td>
<td>87.81</td>
</tr>
<tr>
<td>Li et al. [28]</td>
<td>90.12</td>
<td>89.34</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>90.77</strong></td>
<td><strong>90.06</strong></td>
</tr>
</tbody>
</table>

Let \( L, N \) and \( K \) be the number of stacked transformer blocks that make up a transformer encoder, the number of input patches and the dimensions of the embedding space, respectively. For a given input embedding of the \( i \)-th transformer block, \( z_{i-1} \in \mathbb{R}^{N \times K} \), we have

\[
\hat{z}_i = \text{MSA}(\text{CIN}(z_{i-1}, m)) + z_{i-1}, \quad 1 \leq i \leq L
\]

\[
z_i = \text{MLP}(\text{CIN}(\hat{z}_i, m)) + \hat{z}_i, \quad 1 \leq i \leq L
\]

where \( 1 \leq m \leq M \) is the modality corresponding to the input and \( \text{CIN}(\cdot, \cdot) \) is defined in Equation 1. Note that \( z_0 \) corresponds to the linear projection of the input patches, i.e. non-overlapping sub-patches of dimension \( P \times P \times P \) of the input volume, \( \{p_j \in \mathbb{R}^{P^3} : 1 \leq j \leq N \} \), into a \( K \)-dimensional embedding space using a projection matrix \( E \in \mathbb{R}^{P^3 \times K} \) [18]. Additionally, learnable positional embeddings, \( E_{\text{pos}} \in \mathbb{R}^{N \times K} \), are included to code spatial information, finally obtaining

\[
z_0 = [p_1 E; \ldots; p_J E] + E_{\text{pos}}.
\]

The rationale for substituting LN, which is inherited from the applications of the first transformers in NLP [59], for CIN, or more generally IN, is also motivated on the zoom of Figure 3. LN computes the average and variance for CIN, or more generally IN, is also motivated on the same batch. Therefore, the overall loss will provide an overview of the goodness of the segmentation independently from the input modality, pushing the network to be fair with all different types of data.

4. Experiments

In this section we describe the experiments carried out to evaluate the proposed cross-modality framework with interleaved mixed data training. The experiment settings are described, including the dataset, the implementation details, and the algorithms to tune the models hyper-parameters. Finally, the results are shown and discussed.
Table 2: Performance comparison on the target modality (CT) of our framework applied to the Swin-UNETR model [17]. All the models are trained using 20 MRI and 10 CT volumes and are evaluated on the test set of MM-WHS dataset. The mean DICE score is reported, as well as the ones for all the heart substructures and Whole Heart (WH) segmentation.

<table>
<thead>
<tr>
<th>Model</th>
<th>WH</th>
<th>Avg Dice</th>
<th>Dice of substructures of heart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MYO</td>
</tr>
<tr>
<td>Baseline</td>
<td>83.11</td>
<td>85.32</td>
<td>84.79</td>
</tr>
<tr>
<td>Fine-tune</td>
<td>86.33</td>
<td>86.09</td>
<td>82.57</td>
</tr>
<tr>
<td>Joint-training</td>
<td>89.03</td>
<td>87.99</td>
<td>86.25</td>
</tr>
<tr>
<td>C-ViT [ours]</td>
<td><strong>89.98</strong></td>
<td><strong>89.33</strong></td>
<td><strong>88.13</strong></td>
</tr>
</tbody>
</table>

Table 3: Performance comparison on MRIs, used as assistant modality, of our framework applied to the Swin-UNETR model [17]. The baseline is the model used to do fine-tuning on the target modality.

<table>
<thead>
<tr>
<th>Model</th>
<th>WH</th>
<th>Mean Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>83.32</td>
<td>81.87</td>
</tr>
<tr>
<td>Joint-training</td>
<td>86.65</td>
<td>84.82</td>
</tr>
<tr>
<td>C-ViT [ours]</td>
<td>85.57</td>
<td>84.30</td>
</tr>
</tbody>
</table>

4.1. Experiment settings

**Dataset.** We evaluate the proposed method on the Multimodality Whole Heart Segmentation Challenge 2017 (MM-WHS 2017) dataset [67], which contains non-registered 20 MRI and 20 CT volumes for training and the ground-truth (GT) annotations of 7 cardiac substructures including the Left Ventricle blood cavity (LV), the Right Ventricle blood cavity (RV), the Left Atrium blood cavity (LA), the Right Atrium blood cavity (RA), the MYocardium of the left ventricle (MYO), the Ascending Aorta (AA), and the Pulmonary Artery (PA). Following previous cross-modality segmentation works on this dataset [28], we used MRI as assistant modality and CT as target modality, since MRI has better contrast for soft tissue and ideally provide better information for heart substructures segmentation. We evenly and randomly split the CT data in order to perform a two-fold cross-validation and train the conditional models with 20 MRIs and 10 CTs at a time, to simulate data shortage for the target modality. Furthermore, MRI volumes are pre-processed using the N4 bias field correction algorithm to correct low frequency intensity non-uniformity present in MRI image data known as gain field [52]. Finally, all the volumes are re-sampled to an isotropic space and their values are normalized between 0 and 1.

**Implementation Details.** To provide a fair comparison, we first applied our framework to the same baseline used by other cross-modality previous works [28, 54]. Namely, we implemented the network architecture with encoder-decoder structure proposed by Valindria et al. [54], consisting of a 3D UNet [41] with residual blocks with pre-activation [19]. After that, we evaluated the efficiency of the proposed C-ViT encoder applied to the Swin-UNETR [17] segmentation model, originally consisting of a Swin-transformer encoder [31] and UNet style decoder, as shown in Figure 1. For all the models, we define an input size of $96 \times 96 \times 96$ pixels and sub-patches, in the case of transformer-based architectures, of $16 \times 16 \times 16$ pixels. Therefore, at each epoch, a fixed number of random crops of input size is extracted, for each volume. Note that, in this way, the effective batch size is given by $N_s \times N_c$, where $N_s$ and $N_c$ are the number of samples and crops, respectively. For the transformer-based models, we follow the original implementation of Hatamizadeh et al. [17], with $L = 8$ blocks and hidden size $K = 768$ in the encoder. Furthermore, data augmentation is applied to artificially increase training data and deal with inter-modality variations by randomly shifting and scaling the intensity of the input volumes, along with random flips in all three dimensions. Finally, to validate and test the models, sliding window inference is used with an overlap of 0.5.

The code is implemented in PyTorch, with the help of the MONAI library [5], and trained on a cluster equipped with multiple NVIDIA V100 GPUs. Optimization is performed using Adam with decoupled weight decay [32] and a cosine-decay Learning Rate (LR) scheduler with warm-up, for a maximum of 2500 epochs. All the models are tuned to find the best set of hyper-parameters as described below.

**Hyper-parameters Tuning.** To take full advantage of the segmentation models, we tuned them using the Optuna framework [1], following the guidelines from [14]. Therefore, the batch size is not tuned but is kept unchanged among trials as the maximum that can fit our memory availability, that is, 16 three-dimensional crops ($N_s = 4$ and $N_c = 4$). The optimal LR is found for each model using the cyclical learning rate technique [46] and a logarithmic search space centered on that value is defined for tuning. The number of LR Warm-Up epochs is chosen among the values of 2, 4 and 6% of the maximum training epochs [25] and the optimizer weight decay is also tuned. Finally, the loss function of Equation 6 is used with $\lambda_D = \lambda_F = 1$.

For each trial, the hyper-parameters are sampled form...
the search space using the Tree-structured Parzen Estimator (TPE) algorithm [4, 37]. This technique aims to find the best set of parameters by fitting two Gaussian Mixture Models (GMMs) to the good and bad trials and minimizing their ratio. Furthermore, Asynchronous Successive Halving (ASHA) [29] is used to early-stop bad trials and save computational time and resources.

4.2. Experiment results

We report here the results of our experiments using the proposed framework on the MM-WHS dataset [67], quantitative using the Dice similarity coefficient and qualitative by comparing segmentation images. In the following, we first compare our proposal with other techniques and then we evaluate the accuracy of the proposed C-ViT.

Comparison with Other Methods. The comparative quantitative results for heart substructures segmentation are shown in Table 1. In fine-tuning the model is first trained using the assistant modality only and then, the acquired knowledge is transferred to the target modality with a subsequent training. In joint-training the baseline model is trained with both modalities simultaneously by alternating them on different batches. We also compare the accuracy with the X-shape architecture proposed by Valindria et al. [54] and the online training with synthetic image generation approach of Zhang et al. [63]. Finally, a knowledge distillation cross-modality segmentation technique is also considered for comparison [28]. As highlighted by the latter, fine-tuning, joint-training and the X-shaped architecture bring marginal improvements on the segmentation accuracy since they do not fully exploit cross-modality information. Online GAN-based synthetization with and without mutual knowledge distillation start to significantly improve the mean accuracy of target modality segmentation, up to 3.06%. Nevertheless, in both cases, significant overhead is introduced on the model, limiting its possible real-time applications.

Our method achieves significant improvements in the segmentation accuracy, while reducing the complexity of both training and inference. We observe an enhancement of 0.65% in the mean Dice and up to 2.71% on the single heart substructures, namely in the RV. However, a small decrease in performances is registered for AA and PA, probably due to the fact that the manual segmentation of these structures generally covers beyond the real size of the vessels while the test tool cuts them to a limited length [67].

Evaluation of C-ViT and Ablation Study. We compare the performance of the C-ViT-based segmentation model with its baseline [17] and the result obtained by fine-tuning and joint training the same model (with interleaved mixed fashion). Table 2 summarizes the quantitative results in terms of Dice accuracy of Whole Heart Segmentation (WHS) and single substructures segmentation. Fur-
thermore, Figure 4 shows a qualitative comparison on single slices of the segmentation obtained on CT by using different methods, and Figure 5 directly compares the 3D segmentation of a whole sample. Our framework enhances of 4% the mean Dice of heart substructures, with significant improvements in each of them, and 6.87% the accuracy of WHS, with respect to the baseline model. Qualitative, we can see important improvements when using C-ViT with respect to other methods, that may also result in refinement of the GT (e.g. the segmentation of PA on the second row of Figure 4). In the 3D comparison, we observe the improvement in the segmentation produced by C-ViT which does not contain false positives outside the region of the GT. The proposed framework is robust with respect to the fold used for cross-validation (mean Dice of 91.18% ± 0.68 for the ViT-based and 90.76% ± 0.55 for UNet-based). Moreover, the ViT-based conditional model performs better than the UNet one in validation, but it has slightly worst performances on the test set, probably due to the fact that transformers usually need more data for finer generalization [39].

Differently from some other related works [27, 28] our framework do not just exploit assistant modality to improve the segmentation of the target one, but it keeps the ability of producing high-quality segmentation for the assistant modality by switching the input domain. To confirm that, In Table 3, WHS and Mean Dice of MRI segmentation is reported, where the baseline is considered to be the model trained only with assistant modality and used to fine-tune the target modality. Also the assistant modality segmentation accuracy benefits of our framework, as well as of joint-training, improving with respect to the baseline model. Moreover, Figure 6 also shows some qualitative improvements of the MRI segmentation, especially from baseline to C-ViT. As previously mentioned, transformers are “data-hungry” [57], in the sense that they suffer from significant performance drops on small-size datasets. We have shown that, with our conditional framework applied to ViT, it is possible to relax the need of an huge dataset for training by gathering several small datasets of different modalities.

5. Conclusions

In this work, we proposed a simple framework that aims to segment different types of medical images, while reducing the model overhead and the need of registered data for training, using a single cross-modality conditional model trained with interleaved mixed data. We introduced a general definition of a conditional model, based on CIN, that can be applied to all state-of-the-art medical image segmentation architectures. Furthermore, we developed a new C-ViT encoder that can be used to create conditional ViT-based models. We validated our framework using a public multi-modal dataset for heart substructures segmentation, which collects both MRI and CT data. Our framework achieves new state-of-the-art performance for the cross-modality medical image segmentation with assistant and target modality. We have shown that not only the first one helps the learning on the latter when few annotated data are available, but also the segmentation of the assistant modality benefits of our framework. We believe that our work can be an important contribution for developing robust cross-modality medical image segmentation methods, to tackle cases in which the acquisition of a specific kind of image is limited by environmental or patient-specific restrictions. Furthermore, in the future, other datasets with several modalities [3, 36, 34] may be tested and the proposed framework may be extended in an Unsupervised Domain Adaptation fashion, to make possible the training on a single labeled modality with adaptation to unlabeled domains.

6. Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945304-Cofund AI4theSciences hosted by PSL University.

This work was granted access to the HPC/AI resources of IDRIS under the allocation 2022-AD011013902 made by GENCI.
References


[48] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. A review on deep learning in medical image analysis. In-


