Multi-scale self-guided attention for medical image segmentation

Ashish Sinha and Jose Dolz

***Abstract*—Even though convolutional neural networks (CNNs) are driving progress in medical image segmentation, standard models still have some drawbacks. First, the use of multi-scale approaches, i.e., encoder-decoder architectures, leads to a redundant use of information, where similar low-level features are extracted multiple times at multiple scales. Second, long-range feature dependencies are not efficiently modeled, resulting in non-optimal discriminative feature representations associated with each semantic class. In this paper we attempt to overcome these limitations with the proposed architecture, by capturing richer contextual dependencies based on the use of guided self-attention mechanisms. This approach is able to integrate local features with their corresponding global dependencies, as well as highlight interdependent channel maps in an adaptive manner. Further, the additional loss between different modules guides the attention mechanisms to neglect irrelevant information and focus on more discriminant regions of the image by emphasizing relevant feature associations. We evaluate the proposed model in the context of semantic segmentation on three different datasets: abdominal organs, cardiovascular structures and brain tumors. A series of ablation experiments support the importance of these attention modules in the proposed architecture. In addition, compared to other state-of-the-art segmentation networks our model yields better segmentation performance, increasing the accuracy of the predictions while reducing the standard deviation. This demonstrates the efficiency of our approach to generate precise and reliable automatic segmentations of medical images. Our code is made publicly available at: https://github.com/sinAshish/Multi-Scale-Attention**

***Index Terms*—Convolutional neural networks, Deep learning, Medical image segmentation, Deep attention, Self-attention**

I. Introduction

Semantic segmentation of medical images is a crucial step in diagnosis, treatment and follow-up of many diseases. Despite the automation of this task has been widely studied in the past, manual annotations are still typically used in clinical practice, which is a time-consuming and prone to inter and intra-observer variability process. Thus, there is a high demand on accurate and reliable automatic segmentation methods that allow to improve the work flow efficiency in clinical scenarios, alleviating the workload of radiologists and other medical experts.

Recently, convolutional neural networks (CNNs) have achieved state-of-the-art performance in a breadth of visual recognition tasks, becoming very popular due to their powerful, nonlinear feature extraction capabilities. These deep models dominate the literature in medical image segmentation [1] and have achieved outstanding performance in a broad span of applications, including brain [2] or cardiac [3] imaging, for example, becoming the *de facto* solution for these problems. In this scenario, fully convolutional neural networks [4] or encoder-decoder architectures [5], [6] are typically the standard choice. These architectures are commonly composed of a contracting path, which collapses an input image into a set of high-level features, and an expanding path, where high-level features are used to reconstruct a pixel-wise segmentation mask at a single [4] or multiple upsampling steps [5], [6]. Nevertheless, despite their strong representation power, these multi-scale approaches lead to a redundant use of information flow, e.g., similar low-level features are extracted multiple times at different levels within the network. Furthermore, the discriminative power of the learned feature representations for pixel-wise recognition may be insufficient for some challenging tasks, such as medical image segmentation.

Recent works to improve the discriminative ability of feature representations include the use of multi-scale context fusion [7], [8], [9], [10]. Zhao et al. [8] proposed a pyramid network that exploited global information at different scales by aggregating feature maps generated by multiple dilated convolutional blocks. Aggregation of contextual multi-scale information can also be achieved through pooling operations [11]. Even though these strategies may help to capture objects at different scales, contextual dependencies for all image regions are homogeneous and non-adaptive, ignoring the difference between local representation and contextual dependencies for different categories. Further, these multi-context representations are manually designed, lacking flexibility to model the multi-context representations. This makes that long-range object relationships in the whole image cannot be fully leveraged in these approaches, which is of pivotal importance in many medical imaging segmentation problems.

Alternatively, attention mechanisms have been widely studied in deep CNNs for many computer vision tasks in order to efficiently integrate local and global features, including human pose estimation [12], emotion recognition [13], text detection [14], object detection [15] and classification [16]. Unlike standard multi-scale features fusion approaches, which compress an entire image into a static representation, attention allows the network to focus on the most relevant features without additional supervision, avoiding the use of multiple similar feature maps and highlighting salient features that are useful for a given task. Semantic segmentation networks have also benefited from attention modules, which has resulted in enhanced models for pixel-wise recognition tasks [17], [18], [19], [20], [21], [22]. For example, Chen et *al.* [17] proposed an attention mechanism to weight multi-scale features extracted

A. Sinha is with the Indian Institute of Technology Roorkee, India. e-mail: asinha@mt.iitr.ac.in.

J. Dolz is with the École de technologie Superieure, Montreal, Canada. email:jose.dolz@etsmtl.ca.

Manuscript received XXX; revised XXX.

at different scales in the context of natural scene segmentation. This method improved the segmentation performance over classical average and max-pooling techniques to merge multi-scale features predictions.

Despite the growing interest on integrating attention mechanisms in image segmentation networks for natural scenes, their adoption in medical images remains scarce [23], [24], [25], [26], [27], being limited to simple attention models. Thus, in this work, we explore more complex attention mechanisms that can boost the performance of standard deep networks for the task of medical image segmentation. Specifically, we propose a multi-scale guided attention network for medical image segmentation. First, the multi-scale approach generates stacks at different resolutions containing different semantics. While lower-level stacks focus on local appearance, higher-level stacks will encode global representations. This multi-scale strategy encourages that attention maps generated at different resolutions encode different semantic information. Then, at each scale, a stack of attention modules will gradually remove noisy areas and emphasize those regions that are more relevant to the semantic descriptions of the targets. Each attention module contains two independent self-attention mechanisms, which focus on modelling position and channel feature dependencies, respectively. This duple allows to model wider and richer contextual representations and improve dependencies between channel maps, resulting in enhanced feature representations. We validate our method in three different segmentation tasks: abdominal organ, cardiovascular structures and brain tumor. Results show that the proposed architecture improves the segmentation performance by successfully modeling rich contextual dependencies over local features.

II. Related work

A. Medical image segmentation

Even though segmentation of medical images has been widely studied in the past [28], [29] it is undeniable that CNNs are driving progress in this field, leading to outstanding performances in many applications. Most available medical image segmentation architectures are inspired from the well-known fully convolutional neural network (FCN) [4] or UNet [5]. In FCN the fully connected layers of standard classification CNNS are replaced by convolutional layers to achieve dense pixel prediction at one forward step. To recover the original resolution of the input image, the prediction is upsampled in a single step. Further, to improve the prediction capabilities, skip connections are included in the network by employing the intermediate feature maps. On the other hand, UNet contains contractive and expansive paths created using the combination of convolutional layers with pooling and upsampling layers. Skip connections are used to concatenate the features from contractive and expansive path layers. Many extensions of these networks have been proposed to solve pixel-wise segmentation problems in a wide range of applications [30], [31], [32], [33], [34], [35], [36], [37], [38], [39].

B. Deep attention

Attention mechanisms aim at emphasizing important local regions captured in local features and filtering irrelevant information transferred by global features, improving the modeling of long-range dependencies. These modules have therefore become an essential part of models that need to capture global dependencies. The integration of these attention modules has been proved very successful in many vision problems, such as image captioning [40], image question-answering [41] or classification [42]. Self-attention [43], [44], [45] has recently attracted the attention of researchers, as it exhibits a good ability to model long-range dependencies while maintaining computational and statistical efficiency. In these modules, the response at each position is calculated by attending to all positions and taking their weighted average in an embedding space. For image vision problems, [18], [19] integrated self-attention to model the relation of local features with their corresponding global dependencies. For instance, the point-wise spatial attention network (PSANet) proposed in [18] allows a flexible and dynamic aggregation of long-range contextual information by connecting each position in the feature map with all the others through self-adaptive attention maps.

Recent works have indicated that attention features generated in a single step may still contain noise introduced from regions that are irrelevant for a given class, leading to sub-optimal results [41], [46]. To overcome this issue, some works have investigated the use of progressive multiple attention layers in the context of visual question answering [41] or zero shot learning [46]. This strategy gradually filters undesired noise and emphasizes the regions highly relevant for the class semantic representations. To the best of our knowledge, the application of stacked attention modules remains unexplored in semantic segmentation.

C. Medical image segmentation with deep attention

Even though attention mechanisms are becoming popular on many vision problems, the literature on medical image segmentation with attention remains scarce, with simple attention modules [23], [24], [25], [26], [27]. Wang et *al.* [23] employed attention modules at multiple resolutions to combine local deep attention features (DAF) with global context for prostate segmentation on Ultrasound images. To model long-range dependencies local and global features were combined in a simple attention module, which contains three convolutional layers followed by a softmax function to create the attention map. A similar attention module, composed of two convolutional layers followed by a softmax, was integrated in a hierarchical aggregation framework integrated in UNet for left atrial segmentation [24]. More recently, additive attention gate modules were integrated in the skip connections of the decoding path of UNet with the goal of better model complimentary information from the encoder [25].

III. Methods

A. Overview

Target structures on medical imaging typically present intra and inter-class diversity on size, shape and texture, particularly if images are processed in 2D. Traditional CNNs for segmentation have a local receptive field, which results in the generation of local feature representations. As long-range

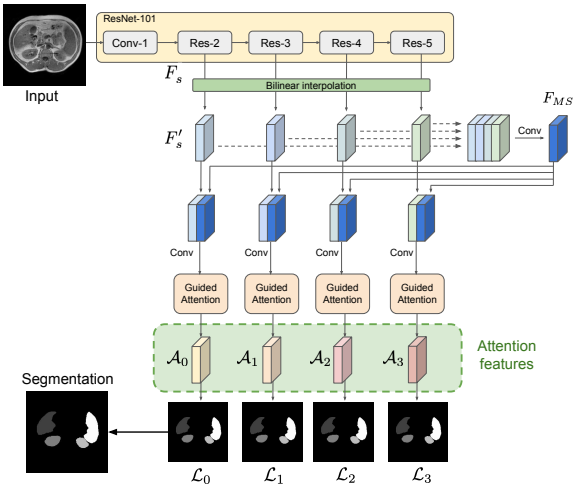


Fig. 1: Overview of the proposed multi-scale guided attention network. We resort to ResNet-101 to extract dense local features. Four feature maps with different sizes –acquired from the outputs of [Res-2, Res-3, Res-4, Res-5]– are employed. The guided attention modules will generate attentive features at multiple scales, removing noisy areas and helping the network to emphasize the regions that are more relevant to the semantic classes.

contextual information is not properly encoded, local features representations may lead to potential differences between features corresponding to the pixels with the same label [19]. This may introduce intra-class inconsistency that can ultimately impact on the recognition performance [47]. To tackle with this problem, we investigate attention mechanisms to build associations between features. First, global context is captured by employing a multi-scale strategy. Then, learned features at multiple scales are fed into the guided attention modules, which are composed by a stack of spatial and channel self-attention modules. While the spatial and channel self-attention modules will help to adaptively integrate local features with their global dependencies, the stack of attention modules will help to gradually filter noise out emphasizing on relevant information. The overview of the proposed framework is depicted in Figure 1.

B. Multi-scale attention maps

Multi-scale features are known to be useful in computer vision problems even before the deep learning era [48]. In the context of deep segmentation networks, the integration of multi-scale features has demonstrated astonishing performance [17], [49], [50]. Inspired by these works we make use of learned features at multiple scales, which help to encode both global and local context. Specifically we follow the multi-scale strategy recently proposed in [23]. In this setting, features at multiple scales are denoted as F_s, where s indicates the level in the architecture (Fig. 1). Since features come at different resolutions for each level s, they are upsampled to a common resolution by employing bilinear interpolation, leading to enlarged feature maps F'_s. Then, F'_s from all the scales are concatenated forming a tensor that is convolved to create a common multi-scale feature map, $F_{MS} = conv([F'_0, F'_1, F'_2, F'_3])$. Thus, F_{MS} encodes low-level detail information from shallow layers as well as high-level semantics learned in deeper layers. Then, this new multi-scale feature map is combined with each of the feature maps at different scales s and fed into the guided attention modules to generate the attention features $\mathcal{A}_s$:

$$\mathcal{A}_s = AttMod_s(conv([F'_s, F_{MS}])) \tag{1}$$

where $AttMod$ represents each guided attention module (Section III-D). As the multi-scale feature maps F_{MS} are combined at each individual layer, complementary low-level information and high-level semantics from F_{MS} are encoded jointly, resulting in a more powerful representation. In the following sections we detail how the attentive features A_s are obtained.

C. Spatial and Channel self-attention modules

As introduced earlier, receptive fields in traditional segmentation deep models are reduced to a local vicinity. This limits the capabilities of modeling wider and richer contextual representations. On the other hand, channel maps can be considered as class-specific responses, where different semantic responses are associated with each other. Thus, another strategy to enhance the feature representation of specific semantics is to improve the dependencies between channel maps [51]. To address these limitations of standard CNNs we employ the position and channel attention modules recently proposed in [19], which are depicted in Figure 2.

Position attention module (PAM): Let denote $F \in \mathbb{R}^{C\times W\times H}$ an input feature map to the attention module, where C, W, H represent the channel, width and height dimensions, respectively. In the upper branch F is passed through a convolutional block, resulting in a feature map $F_0^p \in \mathbb{R}^{C'\times W\times H}$, where C' is equal to $C/8$[1]. Then, F_0^p is reshaped to a feature map of shape $(W \times H) \times C'$. In the second branch, the input feature map F follows the same operations and then is transposed, resulting in $F_1^p \in \mathbb{R}^{C'\times(W\times H)}$. Both maps are multiplied and softmax is applied on the resulted matrix to generate the spatial attention map $S^p \in \mathbb{R}^{(W\times H)\times(W\times H)}$:

$$s^p_{i,j} = \frac{\exp\left(F^p_{0,i} \cdot F^p_{1,j}\right)}{\sum_{i=1}^{W\times H} \exp\left(F^p_{0,i} \cdot F^p_{1,j}\right)} \tag{2}$$

where $s^p_{i,j}$ evaluates the impact of the i^{th} position on the j^{th} position. The input F is fed into a different convolutional block in the third branch, resulting in $F_2^p \in \mathbb{R}^{C\times(W\times H)}$, which has the same shape as F. As in the other branches, F_2^p is reshaped becoming $F_2^p \in \mathbb{R}^{C\times(W\times H)}$. Then it is multiplied by a permuted version of the spatial attention map S, whose output is reshaped to a $\mathbb{R}^{C\times(W\times H)}$. The attention feature map corresponding to the position attention module, i.e., F_{PAM}, can be therefore formulated as follows:

[1]We use the superscript p to indicate that the feature map belongs to the position attention module. Similarly, we will employ the superscript c for the channel attention module features.

$$F_{PAM,j} = \lambda_p \sum_{i=1}^{W \times H} s_{ij}^p F_{2,j}^p + F_j \tag{3}$$

As in [19], the value of λ_p is initialized to 0 and it is gradually learned to give more importance to the spatial attention map. Thus, the position attention module selectively aggregates global context to the learned features, guided by the spatial attention map.

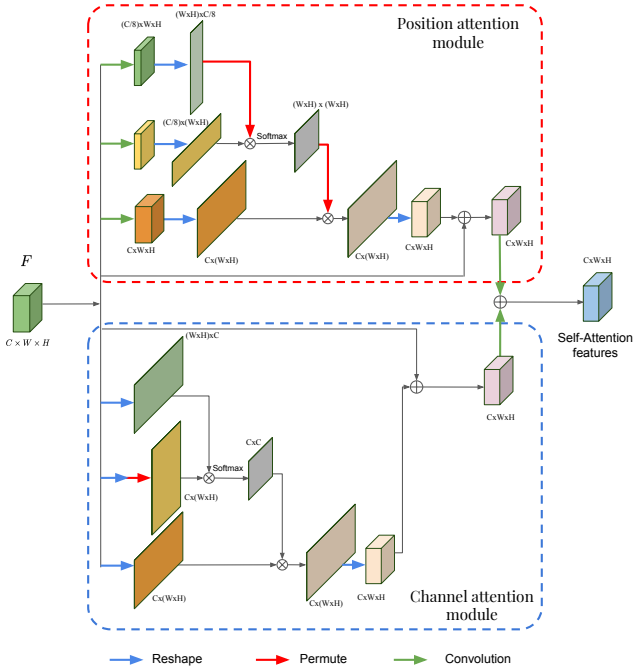


Fig. 2: Details of the position and channel attention modules inspired by [19].

Channel attention module (CAM): The pipeline of the channel attention module is depicted at the bottom of Figure 2. The input $F \in \mathbb{R}^{C \times W \times H}$ is reshaped in the first two branches of the CAM, and permuted in the second branch, leading to $F_0^c \in \mathbb{R}^{(W \times H) \times C}$ and $F_1^c \in \mathbb{R}^{C \times (W \times H)}$, respectively. Then, we perform a matrix multiplication between F_0^c and F_1^c, and obtain the channel attention map $S^c \in \mathbb{R}^{C \times C}$ as:

$$s_{i,j}^c = \frac{\exp\left(F_{0,i}^c \cdot F_{1,j}^c\right)}{\sum_{i=1}^{C} \exp\left(F_{0,i}^c \cdot F_{1,j}^c\right)} \tag{4}$$

where the impact of the i^{th} channel on the j^{th} is given by $s_{i,j}^c$. This is then multiplied by a transposed version of the input F, i.e., F_2^c, whose result is reshaped to $\mathbb{R}^{C \times (W \times H)}$. Then, the final channel attention map is obtained as:

$$F_{CAM,j} = \lambda_c \sum_{i=1}^{C} s_{ij}^c F_{2,j}^c + F_j \tag{5}$$

where λ_c controls the importance of the channel attention map over the input feature map F. Similarly to λ_p, λ_c is initially set to 0 and gradually learned. This formulation aggregates weighted versions of the features of all the channels into the original features, highlighting class-dependent feature maps and increasing feature discriminability between classes. At the end of both attention modules, the new generated features are fed into a convolutional layer before performing an element-wise sum operation to generate the position-channel attention features.

D. Guiding attention

Inspired by recent work to stack attention modules in the context of image classification [46], we propose to add progressive refinement of attentive features through sequential refinement modules. The intuition is that this sequential refinement will progressively weight the importance of different local regions, while masking out irrelevant noise. Particularly, given the feature map F at the input of the guided attention module at scale s–generated by concatenating F_{MS} and F_s'–, it generates attention features via a multi-step refinement (Fig. 3). In the first step, F is used by the position and channel attention modules to generate self-attention features. In parallel, we integrate an encoder-decoder network that compresses the input features F into a compacted representation in the latent space [46]. The objective is that the class information can be embedded into the subsequent guided attention modules by forcing the latent representation of encoder-decoders to be close, which is formulated as:

$$\mathcal{L}_G = \sum_{i}^{M-1} \|\mathbb{E}_i(F_A^{i-1}) - \mathbb{E}_{i+1}(F_A^i)\|_2^2 \tag{6}$$

where $\mathbb{E}_i(\cdot)$ is the encoded representation of the i-th encoder-decoder network, F_A^i denotes the attention features generated after the i-th dual attention module and M the number of iterations. Note that F_A^{i-1} are the features at the input of the semantic guided attention module, F. Specifically, the feature maps reconstructed in the first encoder-decoder ($n = 0$) are combined with the self-attention features generated by the first attention module through a matrix-multiplication to generate F_{SA}. In addition, to ensure that the reconstructed features correspond to the features at the input of the position-channel attention modules, the output of the encoders are forced to be close to their input:

$$\mathcal{L}_{Rec} = \sum_{i}^{M} \|F_i - \hat{F}_i\|_2^2 \tag{7}$$

where $\hat{F}_i$ are the reconstructed feature maps, i.e., $\mathbb{D}_i(\mathbb{E}_i(F))$ of the i-th encoder-decoder networks.

As the guided attention module is applied at multiple scales, the combined guided loss for all the modules will be:

$$\mathcal{L}_{G_{Total}} = \sum_{s=0}^{S} \mathcal{L}_G^s \tag{8}$$

Similarly, the total reconstruction loss becomes:

$$\mathcal{L}_{Rec_{Total}} = \sum_{s=0}^{S} \mathcal{L}_{Rec}^s \tag{9}$$

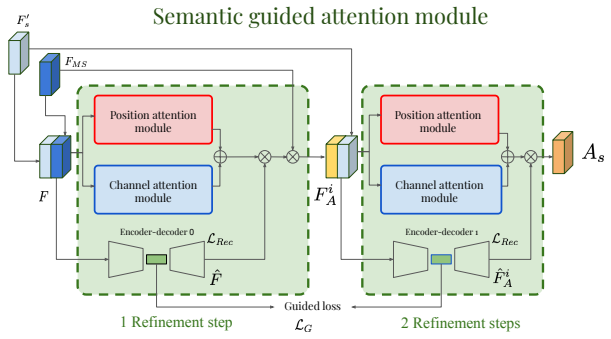


Fig. 3: An illustration of the semantic guided attention module with 2 refinement steps. For each scale s this module provides a set of attentive features, i.e., A_s.

where $\mathcal{L}_{Rec_1}$ and $\mathcal{L}_{Rec_2}$ are the reconstruction losses for the encoder-decoder architectures in the first and second block of the guided attention module.

E. Deep supervision

While the attention modules do not require auxiliary objective functions, we found that the use of extra supervision at each scale [52] improved the segmentation performance of the proposed model, which is in line with similar works in the literature [17], [23], [25].

$$\mathcal{L}_{Seg_{Total}} = \sum_{s=0}^{S} \mathcal{L}^{s}_{Seg_{F'}} + \sum_{s=0}^{S} \mathcal{L}^{s}_{Seg_{A}} \tag{10}$$

where the first term refers to the segmentation results at the raw features F'_s and the second term evaluates the segmentation result provided by the attention features. In all the cases, the multi-class cross-entropy between the network prediction and the ground truth labels is employed as segmentation loss. The final objective function to optimize becomes:

$$\mathcal{L}_{Total} = \alpha \mathcal{L}_{Seg_{Total}} + \beta \mathcal{L}_{G_{Total}} + \gamma \mathcal{L}_{Rec_{Total}} \tag{11}$$

where α, β and γ control the importance of each term in the main loss function.

IV. Experiments

A. Experimental setting

Datasets: We employ three public segmentation benchmarks to evaluate our method. First, the abdominal MRI dataset from the Combined Healthy Abdominal Organ Segmentation (CHAOS) Challenge [53], [54], [55]. Particularly, we focus on the segmentation of abdominal organs (spleen, liver and kidneys) on MRI (T1-DUAL in phase), which includes scans from 20 subjects for training with their corresponding ground truth annotations, and 20 for testing without annotations. Scans have a resolution of 256×256 pixels per slice, and between 26 and 50 slices. Since testing labels are not provided within the dataset, we employed the training dataset for our experiments, splitting it into subsets of 13, 2 and 5 subjects that were used for training, validation and testing. We repeated the process 3 times selecting different subjects and report the average results over the three folds. Then, we evaluated our approach on the task of whole-heart and great vessel segmentation from 3D Cardiovascular MRI in congenital heart disease, provided in the HVSMR 2016 Challenge [56]. Particularly, the myocardium and the blood pool are targeted in this scenario. The training set consists on 10 MRI Axial scans with their corresponding manual segmentations. Image dimensions varied across subjects, with an average of 390 × 390 × 165 voxel volumes. We report results on the training data, by employing a 5-fold cross-validation strategy, where each fold contains 6 scans for training, 2 for validation and 2 for testing. To increase the variability of the data, we rotate, flipped and mirrored the images randomly, but without augmenting the dataset size. For the third task, we employed the brain segmentation dataset provided in the Medical Segmentation Decathlon Challenge[2]. Particularly, this dataset contains multimodal multisite MRI data (FLAIR, T1w, T1gd,T2w) from the BRATS'16 and BRATS'17 Challenges [57], [58], [59]. The focus of this dataset is on the segmentation of necrotic (TC) and active areas (ET), as well as oedema (ED) in brain gliomas. We employed 484 scans that were split into training (388 scans), validation (48 scans) and testing (48 scans). Similarly to previous tasks, we rotate, flipped and mirrored the images randomly, but without augmenting the dataset size.

Network architectures: The multi-scale strategy in the proposed network is based on the recently work in [23], and is considered as the lower baseline in our experiments. First, we perform an ablation study on the different proposed modules to evaluate the impact of each choice in the segmentation performance. The first two networks –i.e., *Proposed (PAM)* and *Proposed (CAM)*– extend the baseline by replacing the attention module by either the spatial or the channel self-attention module (Fig. 2), respectively. Then, both modules are combined simultaneously, leading to the *Proposed (DANet)* model. In the next model –i.e., *Proposed (MS-Dual)*– the attention features generated by the dual attention module are refined in a multi-step process, where a second dual attention module is included. Last, the proposed architecture, referred to as *Proposed (MS-Dual-Guided)* extends the *Proposed (MS-Dual)* model by incorporating the semantic guidance (Fig. 3). We also evaluated the impact of different elements, other than the attention modules, on the proposed multi-scale architectures. First, we remove the deep supervision (first term in eq. 10) on our model. Second, instead of using an encoder-decoder structure to reconstruct the input features at each dual attention module, we remove this and replace the eq. 7 by the mean error square loss between the input and the output of each attention module. This models is referred to as *w/out encoder-decoder (dist)*. And last, we also investigated the effect of not having an encoder-decoder, i.e., no guidance, in the refinements steps, which is referred to as *w/out encoder-decoder)*. In addition, we evaluated the impact of having multiple refinements steps n, with $n = 1, 2, 3$ and 5.

Furthermore we compared the performance of the pro-

[2]http://medicaldecathlon.com

posed network to other state-of-the-art architectures integrating attention: Attention UNet [25], DANet [19] and Pyramidal Attention Network (PAN) [20].

Training and implementation details: We train all the networks using Adam optimizer with mini-batch of size 8, and with β_1 and β_2 set to 0.9 and 0.99, respectively. While most of the networks converged during the first 250 epochs, we found that PAN [20] and DANet [19] needed around 400 epochs to achieve the best results. The learning rate is initially set to 0.001 and multiplied by 0.5 after 50 epochs without improvement on the validation set. As a segmentation objective function, we employ the cross-entropy error at each pixel over all the categories for all the networks. Furthermore, as introduced in Section III, we use the objective function in eq. (11) in the proposed architecture, with α, β and γ set empirically to 1, 0.25 and 0.1, respectively. As input of the networks we employed 2D axial images of size 256 $\times$ 256. Experiments were performed in a server equipped with a Titan V. The code of our model is made publicly available at https://github.com/sinAshish/Multi-Scale-Attention.

Evaluation: Similarity between ground truth and CNN segmentations is assessed by employing several comparison metrics. First, we resort to the widely used Dice similarity coefficient (DSC) to compare volumes based on their overlap. Further, we also assess the segmentation performance based on the volume similarity (VS). Additionally, to measure the sensitivity to segmentation outline, we considered the use of the mean surface distance (MSD). The formulation of these metrics is detailed in the Supplemental materials. Since inter-slice distances and x-y spacing for each individual scan are not provided, we report these results on voxels.

B. Results

Ablation study on the proposed attention modules: To validate the individual contribution of different components to the segmentation performance, we perform an ablation experiment under different settings. Table I reports the results of the different attention modules. Compared to the baseline, we observe that by integrating either a spatial (PAM) or an attention module (CAM) at each scale in the baseline architecture the performance improves between 2-3% in terms of overlapping and volume similarity, and between 12-18% in terms of surface distances, as average. On the other hand, having both modules in parallel –i.e., *Proposed (DANet)*– brings slightly better results in terms of DSC, but achieves lower performance when employing the surface distance metric. However, despite the lower average performance on the MSD, the proposed DANet model still achieves better results in 3 out of 4 structures compared to the channel attention model. This trend is repeated on the DSC metric, where DANet surpasses the proposed CAM architecture in the same 3 structures: liver and both left and right kidneys. This suggests that, even though both spatial and channel attention bring an improvement on the performance, the channel attention module contributes more than the spatial attention when they are combined. If features generated by the proposed DANet model are refined in a second step –network referred to as *Proposed(MS-Dual)*– the average results are further improved by nearly 0.7% and 10% in volume and distance-based metrics, respectively. Last, the introduction of the semantic-guided loss –*Proposed (MS-Dual-Guided)*– results in an additional boost on performance, yielding to the best values in the three metrics: 86.75%(DSC), 93.85%(VS) and 0.66 voxels (MSD). These results represent an improvement of 4.5%, 4% and 26% in DSC, VS and MSD, respectively, compared to the baseline in [23], showing the efficiency of the proposed attention network compared to individual attention components.

Method	DSC (%)	VS (%)	MSD (voxels)
Baseline (DAF [23])	82.48 (±6.06)	89.68 (±4.48)	0.92 (±0.33)
Proposed (PAM)	84.46 (±6.68)	91.84 (±4.77)	0.80 (±0.43)
Proposed (CAM)	85.08 (±5.62)	92.18 (±5.07)	0.74 (±0.32)
Proposed (DANet)	85.52 (±5.86)	92.07 (±5.23)	0.77 (±0.41)
Proposed (MS-Dual)	**86.17 (±5.78)**	**92.74 (±4.76)**	**0.67 (±0.30)**
Proposed (MS-Dual-Guided)	**86.75 (±5.05)**	**93.85 (±3.50)**	**0.66 (±0.27)**

TABLE I: Ablation study on different attention modules on the Chaos dataset. The values show the average result of the experiments averaged over the 3 folds. Best and second best results are represented in red and blue bold, respectively.

Proposed (MS-Dual and MS-Dual-Guided)			
Model	DSC (%)	VS (%)	MSD (voxels)
1 Refinement step			
MS-Dual (No guidance)	85.75 (±5.08)	92.72 (±3.65)	0.71 (±0.28)
MS-Dual-Guided	86.34 (±5.17)	93.47 (±3.78)	0.68 (±0.29)
w/out deep supervision	84.71 (±4.86)	91.39 (±3.55)	0.75 (±0.17)
w/out encoder-decoder (dist)	85.92 (±5.17)	92.94 (±4.04)	0.76 (±0.34)
2 Refinement steps			
MS-Dual (No guidance)	86.17 (±5.78)	92.74 (±4.76)	**0.67 (±0.30)**
MS-Dual-Guided	**86.75 (±5.05)**	**93.85 (±3.50)**	**0.66 (±0.27)**
w/out deep supervision	83.51 (±5.52)	91.80 (±3.66)	0.75 (±0.16)
w/out encoder-decoder (dist)	**86.67 (±4.98)**	93.67 (±3.38)	0.77 (±0.31)
3 Refinement steps			
MS-Dual (No guidance)	86.26 (±5.71)	93.62 (±4.72)	0.71 (±0.34)
MS-Dual-Guided	86.14 (±5.89)	93.50 (±3.98)	**0.67 (±0.36)**
w/out deep supervision	83.22 (±5.72)	90.95 (±4.31)	0.80 (±0.17)
w/out encoder-decoder (dist)	85.88 (±4.78)	93.23 (±3.71)	0.79 (±0.39)
5 Refinement steps			
MS-Dual (No guidance)	86.33 (±4.98)	**93.74 (±3.91)**	0.71 (±0.31)
MS-Dual-Guided	86.30 (±5.05)	93.16 (±4.11)	0.68 (±0.22)
w/out deep supervision	83.88 (±5.78)	91.03 (±3.66)	0.87 (±0.34)
w/out encoder-decoder (dist)	86.16 (±4.23)	92.98 (±2.93)	0.80 (±0.31)

TABLE II: Ablation study on different elements on the MS-Dual and MS-Dual-Guided architectures evaluated on the Chaos dataset. The values show the average result of the experiments on the 3 folds. Best results are represented in red bold, while blue is used to highlight the second best performance.

The impact of the refinement steps, as well as of the several elements on both MS-Dual and MS-Dual-Guided models is reported in Table II. First, we can observe that increasing the number of refinement steps does not typically improve the performance of the methods. Indeed, best results are often obtained with only two attention guided modules. We argue that progressively refining feature maps may produce an excessive focus to the attentive regions, leading to strongly mined attentive features. This has an adverse effect, as the

attentive features may concentrate in the most discriminative areas, not covering the whole extent of the object. Further, we observe that the proposed model including guided-attention outperforms all the variants, particularly in the distance-based metric. In addition, we provide a comparison in terms of complexity, whose results are depicted in Table VIII, in Supplemental Materials.

Comparison to state-of-the-art: The experimental results obtained by several state-of-the-art segmentation networks are reported in Table III. In the first dataset (*top*), compared to other networks that were proposed in the context of medical image segmentation –i.e.,UNet [5], Attention UNet [25] and DAF [23]– our network achieves a mean improvement of 5.6%, 4.3% and 2.0% (in terms of DSC), 4.9%, 4.2% and 2.1% (on VS) and 25%, 26% and 6% (on MSD), respectively. This difference in performance could be explained by the fact that the attention modules integrated in [23] and [25] are much simpler than those proposed in our architecture. On the other hand, attention modules on general computer vision tasks have attracted more attention, resulting in more elaborated strategies which typically achieve better segmentation results. Among these architectures, the PAN model [20] with ResNet101 as backbone –the same as ours– achieved the best results for segmentation networks proposed for natural scenes. Despite these competitive results, the proposed model still outperforms the PAN architecture by 2.4%, 1.9% and 12% in DSC, VS and MSD. As PAN [20] also employed a multi-scale architecture, these differences suggest that the use of dual self-attention and a guided refinement module can actually improve the performance of segmentation networks. Similarly, the proposed model outperforms other networks in the second and third datasets (*middle and bottom*), indicating that it can be broadly applied to segmentation of medical images in general. Individual per-class scores for both datasets are given in Tables V, VI and VII in Supplemental Material. In addition to these values, we also depict the distribution of DSC, VS and MSD values on the 15 subjects used for evaluation in CHAOS for all the models (Fig. 7 in Supplemental Material).

Qualitative evaluation: To visualize the impact of the different attention modules, Fig. 4 displays the segmentation results on three CHAOS subjects. Despite the similar results reported on Table III for several architectures, the qualitative results depict interesting findings. First, we can observe that UNet typically under-segments certain organs and gets confused easily. For example, in the second row it confused the small bowels with the spleen, while the spleen is not even present in that slice. Integrating attention can overcome some of these limitations and improve the segmentation performance by focusing the attention to relevant areas. This can be observed in the results obtained by the other networks, which, up to some extent, reduce the amount of false positives. Nevertheless, it produces smoother segmentations, resulting in a loss of fine grained details. An interesting result is the segmentation in the last row, where all the models except the proposed network get confused to segment the left kidney. While DANet and PAN models confuse the left kidney with the right one, DAF is not able to detect any relevant region related to the kidneys in that area. In addition, both UNet and UNet with attention models generate segmentations of the left kidney that contain three organs, i.e., left and right kidneys and spleen, which is anatomically not plausible. Unlike all these models, the proposed architecture does not get distracted by ambiguous regions and some misclassified structures are now correctly classified.

CHAOS			
Model	**DSC**	**VS**	**MSD**
UNet [5]	81.14 (±7.88)	89.01 (±4.82)	0.91 (±0.49)
DANet [19]	83.89 (±9.54)	91.42 (±4.52)	0.78 (±0.23)
PAN (ResNet34) [20]	82.70 (±6.51)	90.32 (±5.27)	0.86 (±0.29)
PAN (ResNet101)[20]	84.34 (±6.17)	**91.93 (±4.71)**	0.78 (±0.31)
DAF [23]	82.48 (±6.06)	89.68 (±4.48)	0.92 (±0.33)
UNet Attention [25]	**84.77 (±5.27)**	91.79 (±3.53)	**0.72 (±0.24)**
Proposed (MS-Dual-Guided)	**86.75 (±5.05)**	**93.85 (±3.50)**	**0.66 (±0.27)**
HSVM			
	DSC	**VS**	**MSD**
UNet [5]	79.80 (±6.72)	93.41 (±6.44)	1.68 (±1.28)
DANet [19]	**82.55 (±5.91)**	**94.65 (±4.45)**	1.27 (±0.46)
PAN (ResNet34) [20]	80.97 (±7.76)	93.76 (±5.85)	1.62 (±1.19)
PAN (ResNet101)[20]	82.26 (±5.08)	94.33 (±3.69)	1.24 (±0.38)
DAF [23]	81.78 (±5.71)	94.31 (±3.21)	1.48 (±0.50)
UNet Attention [25]	81.58 (±6.84)	**94.61 (±4.17)**	**1.25 (±0.42)**
Proposed (MS-Dual-Guided)	**83.20 (±4.93)**	94.45 (±2.39)	**1.19 (±0.37)**
BRATS'18			
	DSC	**VS**	**MSD**
UNet [5]	73.65 (±12.39)	87.72 (±8.70)	1.65 (±0.57)
DANet [19]	**79.09 (±10.89)**	**93.32 (±6.99)**	**0.95 (±0.33)**
PAN (ResNet34) [20]	74.12 (±12.76)	89.85 (±9.93)	1.42 (±0.52)
PAN (ResNet101)[20]	76.89 (±11.53)	91.76 (±8.11)	1.17 (±0.47)
DAF [23]	76.78 (±11.77)	90.58 (±9.03)	1.21 (±0.46)
UNet Attention [25]	78.61 (±10.58)	92.66 (±6.86)	1.02 (±0.40)
Proposed (MS-Dual-Guided)	**80.37 (±10.74)**	**93.08 (±7.20)**	**0.90 (±0.36)**

TABLE III: Comparison to other state-of-the-art architectures on the four analyzed datasets. Best and second best results are represented in red and blue bold, respectively.

Similar results are observed on the segmentations obtained in the BRATS'18 images (Fig. 5). Particularly, we can see that the proposed network obtains finer details than the other architectures. For example, small ramifications on the oedema (in green) are better captured by the proposed model (*second row*). Likewise, segmentation of necrotic areas (in red) achieved by our method is closer to the ground truth, specially when the region has a complex shape (*first row*). These visual results indicate that our approach can successfully recover finer segmentation details, while avoiding getting distracted in ambiguous regions. The selective integration of spatial information and among channel maps followed by a guided attention module helps to capture context information. This demonstrates that the proposed multi-scale guided attention model can efficiently encode complimentary information to accurately segment medical images.

Visual inspection of feature maps: Showing the performance difference through ablation studies and quantitative evaluations alone may not be enough to fully understand the benefits and behaviour of novel models. Although the proposed modules contribute to the performance improvement, as shown in the results, it is interesting to investigate whether different modules work as expected. To this end, we analyze some attended feature maps from both the spatial and channel

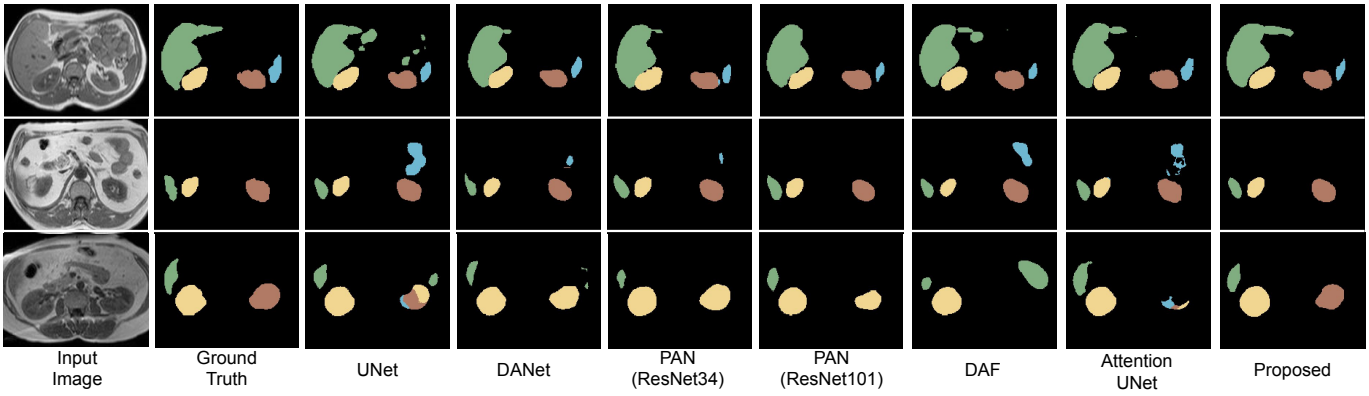


Fig. 4: Results on three subjects on the CHAOS Challenge dataset. The proposed multi-scale guided attention network achieves qualitatively better results than other state-of-the-art networks that also integrate attention modules.

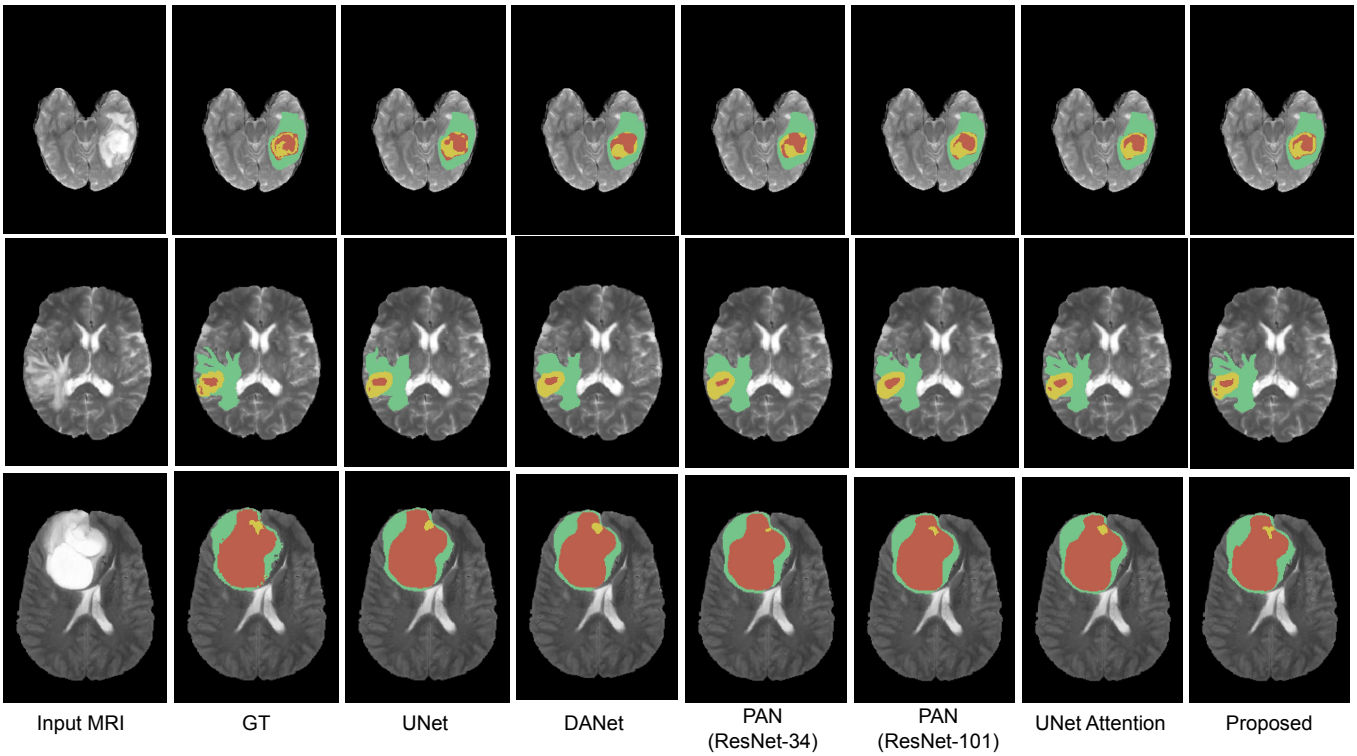


Fig. 5: Results on three subjects on the BRATS Challenge dataset. In these figures, the following tumor structures are depicted: oedema (green), enhancing core (yellow) and necrotic or tumor core (red).

attention modules (Fig. 6). We find that the response of specific semantic classes is more noticeable after the second guided attention modules, i.e., PAM 2 and CAM 2 attentive features. While spatial and channel attention can highlight specific class semantics in the first step of the guided module (second and third column), some non-targeted regions are still highlighted on the semantic maps. Furthermore, highest values are also more spread over the entire image. Contrary, the proposed guided attention module generates features (fourth and fifth columns) that better focus on the specific regions of the structures of interest. Particularly, it can be observed that there exist feature maps whose highlighted areas concentrate on a single organ, avoiding ambiguous regions that might result on misclassification of some regions.

V. Conclusion

In this work, we introduced a novel attention architecture for the task of medical image segmentation. This model incorporates a multi-scale strategy to combine semantic information at different levels and self-attention modules to progressively aggregate relevant contextual features. Last, a guided refinement module filters noisy regions and help the network to focus on relevant class-specific regions in the image. To validate our approach we conducted experiments on three different segmentation tasks: abdominal organ, cardiovascular structures and brain tumor. We provided extensive experiments to evaluate the impact of the individual components of the proposed architecture. Besides, we compared our model to existing approaches that integrate attention, which have been

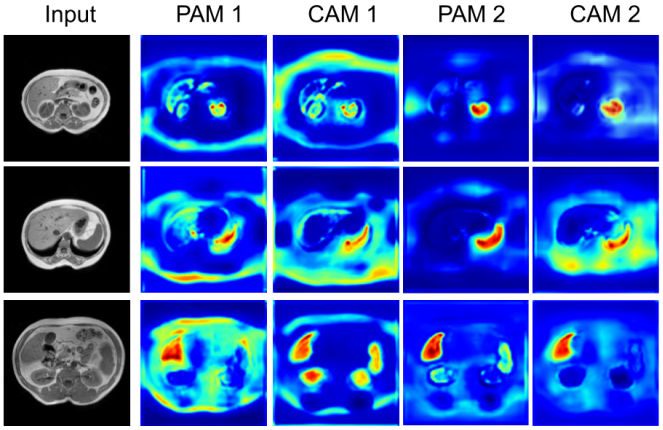


Fig. 6: Visualization results of the channel maps. For each row, we show an input image, and the corresponding channel maps from the outputs of spatial (PAM) and channel (CAM) attention module at guided module of the Fig. 3.

recently proposed for natural scene [19], [20] and medical image [5], [23], [25] segmentation. Experiment results showed that the proposed model outperformed all previous approaches both quantitative and qualitatively, which may be explained by the enhanced ability to model rich contextual dependencies over local features. This demonstrates the efficiency of our approach to provide precise and reliable automatic segmentations of medical images.

ACKNOWLEDGMENTS

We wish to thank NVIDIA for its kind donation of the Titan V GPU used in this work.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.
[2] J. Dolz *et al.*, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE transactions on medical imaging*, 2018.
[3] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.
[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[6] G. Lin *et al.*, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.
[7] L.-C. Chen *et al.*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
[9] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.
[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.
[11] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.
[12] X. Chu *et al.*, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.
[13] A. Gupta *et al.*, "An attention model for group-level emotion recognition," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 611–615.
[14] Z. Huang *et al.*, "Mask R-CNN with pyramid attention network for scene text detection," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 764–772.
[15] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–250.
[16] K. Li *et al.*, "Tell me where to look: Guided attention inference network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9215–9223.
[17] L.-C. Chen *et al.*, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.
[18] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.
[19] J. Fu *et al.*, "Dual attention network for scene segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[20] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *BMVC*, 2018.
[21] C. Yu *et al.*, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.
[22] P. Zhang *et al.*, "Deep gated attention networks for large-scale street-level scene segmentation," *Pattern Recognition*, vol. 88, pp. 702–714, 2019.
[23] Y. Wang *et al.*, "Deep attentional features for prostate segmentation in ultrasound," in *MICCAI*, 2018.
[24] C. Li *et al.*, "Attention based hierarchical aggregation network for 3D left atrial segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 255–264.
[25] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.
[26] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 370–378.
[27] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitationin fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.
[28] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.
[29] J. Dolz, L. Massoptier, and M. Vermandel, "Segmentation algorithms of subcortical brain structures on MRI for radiotherapy and radiosurgery: a survey," *IRBM*, vol. 36, no. 4, pp. 200–212, 2015.
[30] T. Fechter *et al.*, "Esophagus segmentation in CT via 3D fully convolutional neural network and random walk," *Medical physics*, vol. 44, no. 12, pp. 6341–6352, 2017.
[31] X. Li *et al.*, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
[32] Y. Man *et al.*, "Deep Q learning driven CT pancreas segmentation with geometry-aware U-Net," *IEEE transactions on medical imaging*, 2019.
[33] J. Dolz, C. Desrosiers, and I. Ben Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, vol. 170, pp. 456–470, 2018.
[34] A. Carass *et al.*, "Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images," *NeuroImage*, 2018.
[35] C. Zotti, Z. Luo, A. Lalande, and P.-M. Jodoin, "Convolutional neural network with shape prior applied to cardiac mri segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1119–1128, 2018.
[36] J. Dolz *et al.*, "Multiregion segmentation of bladder cancer structures in MRI with progressive dilated convolutional networks," *Medical physics*, vol. 45, no. 12, pp. 5482–5493, 2018.

[37] D. Jin, D. Guo, T.-Y. Ho, A. P. Harrison, J. Xiao, C.-k. Tseng, and L. Lu, "Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 182–191.

[38] M. P. Heinrich, O. Oktay, and N. Bouteldja, "OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions," *Medical image analysis*, vol. 54, pp. 1–9, 2019.

[39] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical image analysis*, vol. 55, pp. 88–102, 2019.

[40] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1242–1250.

[41] Z. Yang *et al.*, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[42] F. Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[43] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *In EMNLP*, 2016.

[44] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[45] X. Wang *et al.*, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[46] Z. Ji *et al.*, "Stacked semantics-guided attention model for fine-grained zero-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5995–6004.

[47] C. Peng *et al.*, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.

[48] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[49] B. Hariharan *et al.*, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 447–456.

[50] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3376–3385.

[51] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[52] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[53] M. A. Selver, "Exploring brushlet based 3D textures in transfer function specification for direct volume rendering of abdominal organs," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 2, pp. 174–187, 2014.

[54] E. Selvi *et al.*, "Segmentation of abdominal organs from MR images using multi-level hierarchical classification," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 30, no. 3, pp. 533–546, 2015.

[55] M. A. Selver, "Segmentation of abdominal organs from CT using a multi-level, hierarchical neural network strategy," *Computer methods and programs in biomedicine*, vol. 113, no. 3, pp. 830–852, 2014.

[56] D. F. Pace, A. V. Dalca, T. Geva, A. J. Powell, M. H. Moghari, and P. Golland, "Interactive whole-heart segmentation in congenital heart disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 80–88.

[57] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[58] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170117, 2017.

[59] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

Supplemental Materials

Evaluation metrics: formulation

In this section, we give the formal definition of the metrics employed to evaluate the proposed architecture.

*a) **Dice Similarity Coefficient (DSC)**:* Given two volumes A and B, their DSC can be defined as:

$$\text{DSC} = \frac{2\,|A \cap B|}{|A| + |B|} \tag{12}$$

In this metric, values close to 1 indicate high degree of overlapping, whereas near 0 represent not overlapping at all.

*b) **Volume Similarity (VS)**:* Further, we also assess the segmentation performance based on the volume similarity, which is formulated as:

$$\text{VS} = 1 - abs(A - B)/(A + B) \tag{13}$$

*c) **Mean Surface Distance (MSD)**:* The MSD between contours A and B is defined as follows:

$$\text{MSD} = \frac{1}{|A| + |B|}\left(\sum_{a \in A} \min_{b \in B} d(a,b) + \sum_{b \in B} \min_{a \in A} d(b,a)\right) \tag{14}$$

$$\text{MSD} = \frac{1}{|A| + |B|}\left(\sum_{a \in A} d(a,b) + \sum_{b \in B} d(b,a)\right) \tag{15}$$

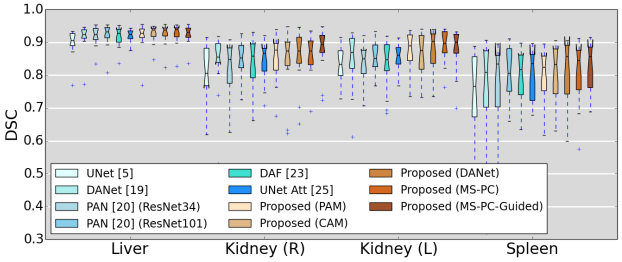
where $d(a,b)$ is the distance between a point a on the surface A and the surface B, which is given by the minimum of the Euclidean norm:

$$d(a,B) = \min_{b \in B} \|a - b\|_2^2 \tag{16}$$

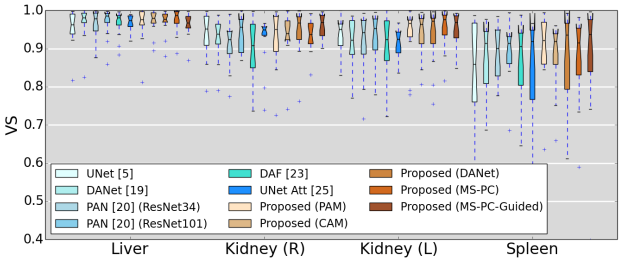
Additional results

Tables IV, V and VI report the extended version of the experimental results on the ablation study and comparison to other state-of-the-art networks. In these tables, individual results on single organs are also included to provide the reader a wider view of the performance of the different methods. We can observe that the proposed architecture is consistently outperforming other models, ranking either first or second in almost all the organs for all the evaluation metrics. The only exception is the result obtained for liver segmentation in terms of volume similarity, where all the models obtain almost identical results.

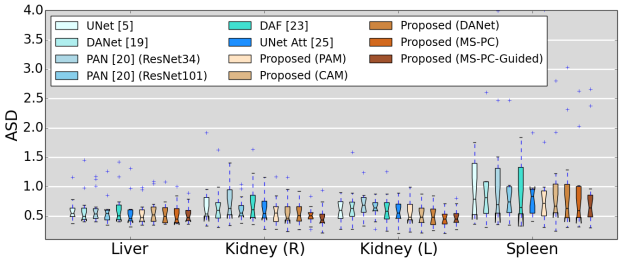
In addition to the values reported on Tables IV and V in the Supplemental Material, we also depict the distribution of DSC, VS and MSD values on the 15 subjects used for evaluation for all the models (Fig. 7). In these plots, we can first observe the impact of the different attention modules in the segmentation performance of the proposed model. As we progressively include the proposed attention modules in the baseline network, the segmentation performance improves, which is reflected in a better distribution of segmentation accuracy values with a smaller variance. This difference on results distribution is more prominent when comparing the proposed network with other state-of-the-art networks, which are represented in bluish box plots. We can also observe that this pattern is constant across organs and metrics, suggesting that the proposed attention network achieves better and more robust segmentation results than current state-of-the-art architectures.



(a) Dice Similarity coefficient (%)



(b) Volume similarity (%)



(c) Average surface distance (voxels)

Fig. 7: These plots depict the distributions of the different evaluation metrics for the four organs segmented. Bluish colors represent the results obtained by other state-of-the-art networks, whereas the results obtained by our proposed models are displayed in with the brownish boxplots.

*1) **Convergence**:* We have also compared the different architectures in terms of convergence, whose results are depicted in Fig. 8. Particularly, the mean DSC value over the four structures on one of the validation folds is shown for each of the networks. It can be observed that, even though most of the networks achieve results which may be considered 'similar' –up to some extent– the convergence behaviour is totally different. While there are three networks with similar convergence curves –i.e., UNet, DANet and DAF–, PAN needs more iterations to convergence, ultimately performing better than these networks after nearly 400 epochs. On the other hand, we found that attention UNet and the proposed network presented the fastest convergence, achieving their best results at epoch 48 and 73, respectively.

Method	Liver	Kidney R	Kidney L	Spleen	Mean
			DSC (%)		
Baseline (DAF [23])	91.66 (±2.99)	79.28 (±18.68)	83.63 (±7.56)	75.35 (±20.41)	82.48 (±6.06)
Proposed (PAM)	91.89 (±4.29)	85.47 (±7.04)	86.84 (±6.53)	73.65 (±22.62)	84.46 (±6.68)
Proposed (CAM)	92.58 (±2.65)	84.52 (±9.34)	86.38 (±6.27)	76.84 (±20.56)	85.08 (±5.62)
Proposed (DANet)	**92.60 (±3.20)**	85.29 (±7.96)	87.74 (±6.37)	76.44 (±22.17)	85.52 (±5.86)
Proposed (MS-Dual)	**92.62 (±3.08)**	**86.29 (±5.98)**	**88.82 (±4.84)**	**76.96 (±19.87)**	**86.17 (±5.78)**
Proposed (MS-Dual-Guided)	92.46 (±2.82)	**87.96 (±6.46)**	**88.01 (±6.16)**	**78.61 (±18.69)**	**86.75 (±5.05)**
			Volume similarity (VS) (%)		
	Liver	**Kidney R**	**Kidney L**	**Spleen**	**Mean**
Proposed(DAF [23])	96.69 (±3.21)	86.75 (±16.41)	90.29 (±8.39)	**84.98 (±14.42)**	89.68 (±4.48)
Proposed (PAM)	96.62 (±4.62)	92.83 (±7.43)	93.96 (±6.46)	83.93 (±20.54)	91.84 (±4.77)
Proposed (CAM)	**97.25 (±2.95)**	93.78 (±6.04)	93.98 (±5.48)	83.72 (±20.97)	92.18 (±5.07)
Proposed (DANet)	97.04 (±3.03)	**94.50 (±5.96)**	93.43 (±7.03)	83.30 (±22.53)	92.07 (±5.23)
Proposed (MS-Dual)	**97.47 (±3.07)**	93.30 (±4.11)	**95.27 (±4.89)**	84.90 (±16.86)	**92.74 (±4.76)**
Proposed (MS-Dual-Guided)	96.44 (±3.15)	**96.14 (±3.15)**	**94.95 (±4.48)**	**87.87 (±15.23)**	**93.85 (±3.50)**
			Average Surface Distance (MSD) (voxels)		
	Liver	**Kidney R**	**Kidney L**	**Spleen**	**Mean**
Baseline(DAF [23])	0.64 (±0.29)	0.97 (±1.08)	0.63 (±0.25)	1.45 (±2.04)	0.92 (±0.33)
Proposed (PAM)	0.55 (±0.19)	0.56 (±0.23)	0.55 (±0.21)	1.54 (±2.40)	0.80 (±0.43)
Proposed (CAM)	0.58 (±0.22)	0.57 (±0.24)	0.52 (±0.20)	1.29 (±1.64)	0.74 (±0.32)
Proposed (DANet)	**0.54 (±0.19)**	0.56 (±0.19)	0.50 (±0.18)	1.49 (±2.29)	0.77 (±0.41)
Proposed (MS-Dual)	**0.53 (±0.18)**	**0.51 (±0.14)**	**0.46 (±0.14)**	**1.19 (±1.42)**	**0.67 (±0.30)**
Proposed (MS-Dual-Guided)	**0.54 (±0.16)**	**0.48 (±0.18)**	**0.48 (±0.14)**	**1.13 (±1.24)**	**0.66 (±0.27)**

TABLE IV: Ablation study on different proposed attention modules on the Chaos dataset (multi-organ segmentation on MRI task). The values show the average result of the experiments averaged over the 3 folds. Best results are represented in red bold, while blue is used to highlight the second best performance.

Method	Liver	Kidney R	Kidney L	Spleen	Mean
			DSC (%)		
UNet [5]	90.94 (±4.01)	79.14 (±15.23)	82.51 (±7.48)	71.95 (±21.61)	81.14 (±7.88)
DANet [19]	91.69 (±4.07)	83.85 (±9.40)	84.49 (±8.60)	75.54 (±16.08)	83.89 (±9.54)
PAN (ResNet34) [20]	91.99 (±2.98)	81.51 (±9.03)	83.62 (±6.21)	73.70 (±19.97)	82.70 (±6.51)
PAN (ResNet101)[20]	**92.13 (±3.51)**	**85.02 (±5.16)**	85.36 (±4.87)	74.84 (±21.23)	84.34 (±6.17)
DAF [23]	91.66 (±2.99)	79.28 (±18.68)	83.63 (±7.56)	75.35 (±20.41)	82.48 (±6.06)
UNet Attention [25]	92.02 (±1.93)	84.33 (±5.91)	**85.57 (±4.09)**	**77.18 (±15.95)**	**84.77 (±5.27)**
Proposed (MS-Dual-Guided)	**92.46 (±2.82)**	**87.96 (±6.46)**	**88.01 (±6.16)**	**78.61 (±18.69)**	**86.75 (±5.05)**
			Volume similarity (VS) (%)		
	Liver	**Kidney R**	**Kidney L**	**Spleen**	**Mean**
UNet [5]	95.54 (±4.43)	87.68 (±5.77)	89.55 (±4.68)	83.28 (±14.78)	89.01 (±4.82)
DANet [19]	96.90 (±4.18)	92.88 (±5.12)	91.52 (±6.73)	84.37 (±16.15)	91.42 (±4.52)
PAN (ResNet34) [20]	96.56 (±3.55)	90.89 (±5.64)	91.83 (±7.75)	81.98 (±20.67)	90.32 (±5.27)
PAN (ResNet101) [20]	**96.99 (±3.64)**	**93.77 (±4.63)**	**92.69 (±6.88)**	84.24 (±17.37)	**91.93 (±4.71)**
DAF [23]	96.69 (±3.21)	86.75 (±16.41)	90.29 (±8.39)	84.98 (±14.42)	89.68 (±4.48)
UNet Attention [25]	**96.95 (±1.89)**	92.29 (±6.41)	91.79 (±3.53)	**85.94 (±11.88)**	91.74 (±3.91)
Proposed (MS-Dual-Guided)	96.44 (±3.15)	**96.14 (±3.15)**	**94.95 (±4.48)**	**87.87 (±15.23)**	**93.85 (±3.50)**
			Average Surface Distance (MSD) (voxels)		
	Liver	**Kidney R**	**Kidney L**	**Spleen**	**Mean**
UNet [5]	0.59 (±0.18)	0.69 (±0.38)	0.61 (±0.19)	1.76 (±2.57)	0.91 (±0.49)
DANet [19]	0.61 (±0.27)	0.65 (±0.31)	0.67 (±0.30)	1.17 (±0.94)	0.78 (±0.23)
PAN (ResNet34)[20]	0.62 (±0.25)	0.75 (±0.31)	0.69 (±0.21)	1.37 (±1.43)	0.86 (±0.29)
PAN (ResNet101) [20]	**0.57 (±0.22)**	**0.61 (±0.19)**	0.64 (±0.15)	1.30 (±1.47)	0.78 (±0.31)
DAF [23]	0.64 (±0.29)	0.97 (±1.08)	0.63 (±0.25)	1.45 (±2.04)	0.92 (±0.33)
UNet Attention [25]	**0.57 (±0.25)**	**0.61 (±0.23)**	**0.56 (±0.18)**	**1.15 (±1.01)**	**0.72 (±0.24)**
Proposed (MS-Dual-Guided)	**0.54 (±0.16)**	**0.48 (±0.18)**	**0.48 (±0.14)**	**1.13 (±1.24)**	**0.66 (±0.27)**

TABLE V: Comparison of the proposed network to other state-of-the-art architectures on the CHAOS dataset (multi-organ segmentation on MRI task). The values show the average result of the experiments averaged over the 3 folds. Best results are represented in red bold, while blue is used to highlight the second best performance.

	DSC		
Method	**Myocardium**	**Blood Pool**	**Mean**
UNet [5]	71.77 (±9.36)	87.84 (±4.35)	79.80 (±6.72)
DANet [19]	**75.85 (±9.10)**	89.24 (±3.56)	**82.55 (±5.91)**
PAN (ResNet34) [20]	72.90 (±11.93)	89.04 (±3.69)	80.97 (±7.76)
PAN (ResNet101)[20]	74.98 (±7.68)	**89.53 (±2.97)**	82.26 (±5.08)
DAF [23]	74.08 (±8.55)	**89.48 (±3.39)**	81.78 (±5.71)
UNet Attention [25]	74.50 (±10.13)	88.66 (±4.25)	81.58 (±6.84)
Proposed	**77.10 (±6.94)**	89.30 (±3.50)	**83.20 (±4.93)**
	Volume similarity (VS)		
	Myocardium	**Blood Pool**	**Mean**
UNet [5]	91.05 (±9.75)	95.78 (±4.04)	93.41 (±6.44)
DANet [19]	91.80 (±8.95)	**97.50 (±3.01)**	**94.65 (±4.45)**
PAN (ResNet34) [20]	90.58 (±10.89)	96.93 (±3.66)	93.76 (±5.85)
PAN (ResNet101) [20]	91.42 (±7.59)	**97.23 (±2.36)**	94.33 (±3.69)
DAF [23]	91.73 (±6.30)	96.89 (±2.33)	94.31 (±3.21)
UNet Attention [25]	**92.52 (±7.66)**	96.69 (±2.20)	**94.61 (±4.17)**
Proposed	**92.08 (±4.39)**	96.82 (±2.76)	94.45 (±2.39)
	Average Surface Distance (MSD)		
	Myocardium	**Blood pool**	**Mean**
UNet [5]	1.82 (±1.48)	1.55 (±1.08)	1.68 (±1.28)
DANet [19]	**1.23 (±0.51)**	1.32 (±0.46)	1.27 (±0.46)
PAN (ResNet34)[20]	1.97 (±1.84)	1.26 (±0.48)	1.62 (±1.19)
PAN (ResNet101) [20]	1.33 (±0.53)	**1.15 (±0.30)**	1.24 (±0.38)
DAF [23]	1.41 (±0.45)	1.44 (±0.46)	1.48 (±0.50)
UNet Attention [25]	1.24 (±0.42)	1.25 (±0.39)	**1.25 (±0.42)**
Proposed	**1.15 (±0.33)**	**1.24 (±0.43)**	**1.19 (±0.37)**

TABLE VI: Comparison of the proposed network to other state-of-the-art architectures on the HVSMR 2016 dataset. The values show the average result of the experiments on the 5 folds.

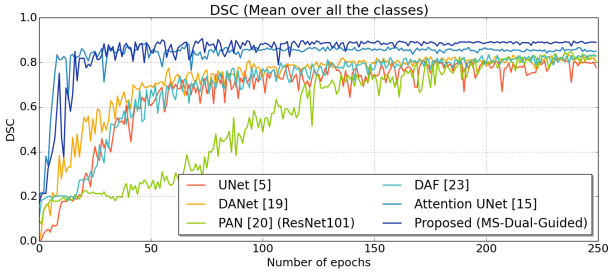


Fig. 8: Evolution of the mean validation DSC over time.

Method	ED	ET	TC	–	Mean
			DSC (%)		
UNet [5]	84.87 (±6.82)	56.38 (±27.55)	79.71 (±11.70)	–	73.65 (±12.39)
DANet [19]	**88.24 (±5.39)**	63.69 (±22.25)	**85.33 (±6.92)**	–	**79.09 (±10.89)**
PAN (ResNet34) [20]	85.25 (±6.64)	55.89 (±27.76)	81.23 (±8.22)	–	74.12 (±12.76)
PAN (ResNet101)[20]	87.07 (±6.67)	60.77 (±24.74)	82.82 (±8.76)	–	76.89 (±11.53)
DAF [23]	86.87 (±5.94)	60.28 (±24.74)	83.18 (±8.39)	–	76.78 (±11.77)
UNet Attention [25]	87.50 (±5.66)	**63.74 (±22.65)**	84.59 (±7.43)	–	78.61 (±10.58)
Proposed (MS-Dual-Guided)	**89.11 (±4.94)**	**65.25 (±.2285)**	**86.76 (±6.49)**	–	**80.37 (±10.74)**
			Volume similarity (VS) (%)		
	ED	**ET**	**TC**	**–**	**Mean**
UNet [5]	96.36 (±4.08)	75.81 (±27.23)	90.99 (±11.63)	–	87.72 (±8.70)
DANet [19]	**99.04 (±1.21)**	**83.47 (±20.11)**	**97.45 (±2.95)**	–	**93.32 (±6.99)**
PAN (ResNet34) [20]	98.05 (±1.98)	75.87 (±28.17)	95.63 (±4.41)	–	89.85 (±9.93)
PAN (ResNet101) [20]	**98.68 (±2.21)**	80.38 (±24.83)	96.22 (±5.89)	–	91.76 (±8.11)
DAF [23]	97.99 (±2.10)	77.86 (±24.92)	95.88 (±5.26)	–	90.58 (±9.03)
UNet Attention [25]	98.14 (±1.88)	**82.99 (±21.09)**	96.84 (±2.87)	–	92.66 (±6.86)
Proposed (MS-Dual-Guided)	98.54 (±1.76)	82.91 (±20.17)	**97.78 (±2.56)**	–	**93.08 (±7.20)**
			Average Surface Distance (MSD) (voxels)		
	ED	**ET**	**TC**	**-**	**Mean**
UNet [5]	0.99 (±0.33)	2.37 (±1.74)	1.56 (±1.34)	–	1.65 (±0.57)
DANet [19]	**0.67 (±0.16)**	**1.43 (±0.95)**	**0.78 (±0.25)**	–	**0.95 (±0.33)**
PAN (ResNet34)[20]	0.86 (±0.20)	2.29 (±1.87)	1.10 (±0.47)	–	1.42 (±0.52)
PAN (ResNet101) [20]	0.74 (±0.19)	1.79 (±1.35)	0.96 (±0.48)	–	1.17 (±0.47)
DAF [23]	0.76 (±0.17)	1.84 (±1.33)	1.02 (±0.66)	–	1.21 (±0.46)
UNet Attention [25]	0.69 (±0.18)	1.58 (±1.12)	0.79 (±0.29)	–	1.02 (±0.40)
Proposed (MS-Dual-Guided)	**0.58 (±0.14)**	**1.40 (±1.02)**	**0.71 (±0.31)**	–	**0.90 (±0.36)**

TABLE VII: Comparison of the proposed network to other state-of-the-art architectures on the BRATS 2018 dataset (multi-organ segmentation on MRI task). The values show the average result of the experiments averaged over the 3 folds. Best results are represented in red bold, while blue is used to highlight the second best performance.

Model	# Params				
		1 Iter	2 Iter	3 Iter	5 Iter
			Model complexity		
UNet	31,030,853	-	-	-	-
PAN (ResNet34)	21,323,991	-	-	-	-
PAN (ResNet101)	42,675,415	-	-	-	-
UNet Attention	34,877,681	-	-	-	-
DANet (ResNet101)	68,475,961	-	-	-	-
Proposed(DAF)	43,482,179	-	-	-	-
Proposed(PAM)	43,486,343	-	-	-	-
Proposed(CAM)	43,485,543	-	-	-	-
Proposed(DANet)	43,980,179	-	-	-	-
MS-Dual (No guidance)	-	43,485,831	44,411,103	45,337,675	47,190,819
MS-Dual-Guided	-	50,531,399	58,499,679	66,470,539	82,412,259
MS-Dual-Guided (No Deep Sup)	-	50,530,099	58,498,379	66,467,939	82,407,059
MS-Dual-Guided (Dist)	-	43,485,831	44,411,103	45,337,675	47,190,819

TABLE VIII: Model complexity, measured in number of parameters, for the evaluated models.