SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation

Jesse Sun[1,4], Fatemeh Darbeha[1,2,3], Mark Zaidi[2], Bo Wang[1,2,3]

Peter Munk Cardiac Center[1], University of Toronto[2], Vector Institute[3], University of Waterloo[4]

j294sun@uwaterloo.ca, {fatemeh.darbehani, mark.zaidi}@mail.utoronto.ca

bowang@vectorinstitute.ai

Abstract

Medical image segmentation is a difficult but important task for many clinical operations such as cardiac bi-ventricular volume estimation. More recently, there has been a shift to utilizing deep learning and fully convolutional neural networks (CNNs) to perform image segmentation that has yielded state-of-the-art results in many public benchmark datasets. Despite the progress of deep learning in medical image segmentation, standard CNNs are still not fully adopted in clinical settings as they lack robustness and interpretability. Shapes are generally more meaningful features than solely textures of images, which are features regular CNNs learn, causing a lack of robustness. Likewise, previous works surrounding model interpretability have been focused on post hoc gradient-based saliency methods. However, gradient-based saliency methods typically require additional computations post hoc and have been shown to be unreliable for interpretability. Thus, we present a new architecture called Shape Attentive U-Net (SAUNet) which focuses on model interpretability and robustness. The proposed architecture attempts to address these limitations by the use of a secondary shape stream that captures rich shape-dependent information in parallel with the regular texture stream. Furthermore, we suggest multi-resolution saliency maps can be learned using our dual-attention decoder module which allows for multi-level interpretability and mitigates the need for additional computations post hoc. Our method also achieves state-of-the-art results on the two large public cardiac MRI image segmentation datasets of SUN09 and AC17.

1. Introduction

Cardiovascular magnetic resonance imaging (CMR) is currently used as the gold standard for the non-invasive assessment of various cardiovascular functions [4, 6]. The high spatial resolution and absence of ionizing radiation during CMR confers several advantages over nuclear medicine imaging modalities such as computed tomography (CT), positron emission tomography (PET), and single-photon emission computed tomography (SPECT) [9]. As a result, CMR plays a crucial role in the diagnosis of cardiovascular diseases such as cardiomyopathies, myocarditis, and congenital heart disease [10]. Diagnosis of such diseases requires specific measurements of morphological and intensity-based features in the image. For example, measurements of the left and right ventricular volumes, along with heart rate, can be used to quantify cardiac output. An increase in cardiac output at resting state is associated with the development of cardiovascular disease caused by atherosclerosis, whereas a low cardiac output can be indicative of heart failure and cardiomyopathy [12, 15]. There are numerous diseases that require bi-ventricular volume estimation in disease diagnosis and prognosis [18, 20, 33].

Arguably one of the greatest challenges in bi-ventricular volume estimation is the segmentation of the left and right ventricular endocardium at end-systolic and diastolic time-points. In a clinical setting, there is a high reliance on manual annotations for ventricular volume estimation. CMR accredited operators will typically annotate short-axis slices by drawing a polygon around the left ventricle (LV) on slices that contain it. The inclusion criteria for which slices to annotate the LV on is ambiguous and can range from slices containing at least 50 to 75% of the cavity being surrounded in the myocardium. Furthermore, if a slice contains both ventricle and atrial myocardium, the operator may choose to trace through the junction in a straight line. Since detailed contouring is very time consuming, one method to annotate the LV includes drawing a circle around the endocardium and performing binary thresholding to create a mask of all pixels with an intensity comparable to that of blood. In either case, there is a heavy reliance on a human operator to calculate LV volume. Geometrical models have been developed in an attempt to decrease the time spent performing annotations, however both manual annotations and geometric modeling suffer from poor intra-observer and inter-observer variability [25]. Recent advances in imaging and computing have led to a drastic rise in the use of machine learning for medical imaging [28]. The advent of

deep learning allows for much higher levels of abstraction for feature selection and discretization. Convolutional neural networks (CNNs) have been shown to learn abstractions obtained from multidimensional medical images, learning features hard to define by humans. This is one of the reasons why CNNs excel at object recognition and segmentation [29].

As a result, image segmentation methods in medical imaging have diverted to deep learning solutions. Ronneberger et al. [33] proposed the U-Net architecture which have gained a lot of traction in medical imaging. The appeal of the U-Net architecture is the dual contracting and symmetric expanding path connected via skip connections to capture context and localization, respectively [38]. Unsurprisingly, the versatility of U-Nets has been demonstrated much promise for cell segmentation and tracking. In the ISBI cell tracking challenge in 2015 and the EM segmentation challenge at ISBI 2012, U-Nets came in first place, outperforming DIVE and IDSIA networks [38, 39]. More recently, U-Net inspired architectures have been employed for segmentation of cardiac structures in CMR images [41]. Using the AC17 dataset [4], segmentation of the left and right ventricular cavity (LVC, RVC), and left ventricular myocardium (LVM) were performed at each frame of the cardiac cycle. Dice scores of 0.945 (LVC), 0.908 (RVC), and 0.905 (LVM) were achieved in a training set cross-validation [41, 43]. Furthermore, by permuting the training set with random rotations, mirroring, elastic deformations, and slice misalignments, the resulting model became robust against slice misalignments, cardiac diseases, and ventricular deformations [41].

One disadvantage shared by many neural networks including U-Net is a lack of interpretability. Because these neural networks interface with many convolutional layers simultaneously, it becomes challenging to visualize what features it is learning on. This effectively renders the neural net a black box, which poses a challenge when attempting to find the root cause of a misclassification, and gives an advantage to potential adversarial attacks. Furthermore, CNNs are highly influenced by dense pixel values which are not robust features compared to shapes of objects [8]. Thus, shapes of objects should be learned to allow generalizability and robustness of the model which goes hand-in-hand with transparency. A lack of model transparency and robustness will hinder its translation into a clinical setting.

While CNNs have shown promise in ventricular segmentation of CMR images, a lack of transparency as to what is focused on during the segmentation will limit the translatability of such technology into a clinical setting. There is a dire need to improve the transparency of neural networks and we suggest one way to achieve transparency and robustness is to enforce the model to learn shape considerations. By affording a higher accuracy in segmentation and verify-

ing that an algorithm is not perpetuating biases, a valuable tool can be created to help solve the challenges numerous clinicians face in medical image analysis.

To the best of our knowledge, past attempts in incorporating shape information in medical imaging segmentation involve forming a new loss function [1, 2, 5]. Furthermore, limited works on model interpretability for medical imaging have been published. As such, the contributions of this work are:

- The addition of a secondary stream that processes shape features of the image in parallel with the U-Net. Rather than constructing a new loss function, we suggest learning shape features can be built-in a model. Further, the output of the shape stream is a shape attention map that can be used for interpretability.

- The usage of spatial and channel-wise attention paths in the decoder block for interpretability of features the model learns at every resolution of the U-Net.

We evaluate our model on large public MRI ventricular volume estimation and segmentation datasets SUN09 [31] and AC17 [4] and demonstrate our method yields state-of-the-art results. We then provide an analysis of the spatial and shape attention maps to interpret the features our model learned. Consequently, our method not only yields strong performance, it is also interpretable at multiple resolutions.

2. Related Works

In this section, we briefly review CNN-based segmentation architectures and methods related to this work.

2.1. U-Net

U-Net consists of a contracting path that captures feature information as well as a symmetric expansive path that enables localization. Moreover, a U-Net uses skip connections from encoders to decoders of similar resolutions to pass high-resolution information throughout the network.

In the original proposal of U-Nets, each block in the encoder is comprised of two successive normalized 3x3 convolutional layers. Then, a max-pooling layer with a 2x2 kernel of stride 2 is used to down-sample the image in order to obtain greater contextual and spatial information. In the up-sampling path, the same blocks in the encoder are used, except feature maps of similar resolutions from the down-sampling path are concatenated with the feature maps from the up-sampling path after being cropped and aligned to match dimensions before being filtered by a normalized 3x3 convolutional layer. Unlike the work of [27] which uses element-wise summation of skip connections with up-sampling feature maps, [34] simply concatenates them before applying 3x3 convolutions in the decoder. In both of

these works, the skip connections are used to obtain more fine-grained information that may have been lost during the intermediate stages. Following each decoder block, an up-sampling is applied to the feature maps to increase the resolution by 2. The original proposal uses interpolation but transpose convolutions can be used to learn the up-sampling parameters and have shown better results.

2.2. DenseNet

DenseNets were originally proposed by Huang et al. [16] and lead to significant improvements in state-of-the-art scores over previous models like ResNet [13] and ResNeXt [45] in image classification tasks such as ImageNet. Furthermore, DenseNets achieved state-of-the-art scores while also having fewer parameters than previous models with similar performance (i.e. DenseNet-169 had better performance than ResNet-50 while also having approximately 50% fewer parameters) due to feature reuse from the additional skip connections. A DenseNet is comprised of two main building blocks - a dense block followed by a transition block. A dense block contains many normalized 3x3 convolution layers where the outputs of each layer are concatenated with each of the feature maps entering the following layers to promote feature reuse. With n layers in a dense block, there are $n!$ skip connections in the block. Each layer outputs a feature map with constant depth of k, so $n * k$ channels exit the dense block. The transition block is composed of a normalized 1x1 convolution to reduce the depth of the feature maps followed by a 2x2 average pool with stride 2 to shrink the resolution by half.

2.3. Attention in CNNs

Interpretability in computer vision comes in two flavours - post hoc analysis [36, 37] and trainable attention. Post hoc analysis is employed after the model has been trained and typically uses gradient-based mechanisms to access the network's reasoning. However, these methods generally do not provide a way of interpreting the model's decision making at different stages of forward propagation and have been shown to be an unreliable source of interpretability [24]. Trainable attention are modules in the model that have learnable parameters. The purpose of these modules is to explicitly learn to reassign importance over some dimension. Jetley et al. [20] proposed an attention estimator module that takes in intermediate feature maps at multiple stages of the network. Then, successive convolutions are applied to reduce the depth of these combined feature maps to 1 followed by a sigmoid or softmax to rescale each pixel value of the attention map to $[0, 1]$. The attention map is then multiplied through all channels in the combined feature maps to re-calibrate the relative activations of each map. The authors showed empirically that popular models such as VGG-16 with their proposed attention mechanisms outperform

the same model without it on cross-domain image classification datasets after trained on CIFAR-10 and CIFAR-100. Furthermore, the attention maps can be extracted to produce a visual heat-map that can be used for interpretability. The authors also showed these attention maps can be binarized and then used as maskings for weakly supervised semantic segmentation on image classification datasets. Schempler et al. [35] extended this attention estimator idea further by applying a similar module prior to each decoder block in their U-Net-based architecture tested on 3D CT image datasets.

Squeeze and excitation blocks proposed by Hu et al. [14] can also be thought of as trainable attention mechanisms. These modules are learned to explicitly redistribute the relative importance of each channel in a feature map according to each channel's global information captured by a global average pooling layer. Other authors have proposed U-Net variants that incorporate squeeze and excitation modules for medical image segmentation tasks and have achieved promising results. Guo et al. [11] proposed adding a squeeze and excitation block prior to each decoder block such that it takes in the skip connection and the feature map from the previous lower resolution decoder block. Unfortunately, squeeze and excitation modules do not offer an obvious method of interpretability.

Our work combines the recent advances of trainable channel-wise attention using squeeze and excitation modules proposed by Hu et al. [14] and spatial attention using the attention estimator proposed by Jetley et al. [20]. We show that incorporating attention along all dimensions yields strong results that are also inherently interpretable.

2.4. Active Contour Loss for Learning Shapes

The active contour loss was recently introduced by Chen et al. [5] in an attempt to incorporate shape considerations in the learning process. It was inspired by the work of [22] on Active Contour Models (ACMs) and the authors proposed to incorporate area and size information into the loss function for more precise segmentation. The AC Loss is defined as,

$$Loss_{AC} = Length + \lambda Region \tag{1}$$

where $Length$ is the L_2 norm of the segmentation's contour and $Region$ compares the means of areas inside and outside the object using a pixel-wise L_1 norm loss. λ is a fixed parameter that controls the balance between the above terms. [5] showed that their model performed worse when $\lambda \rightarrow 0$ as only $Length$ contributed to the loss. Otherwise, their model's performance was invariant to the choice of λ.

Chen et al. combined this geometrically-constrained loss function with their U-Net with dense blocks and it outperformed other mainstream loss functions such as cross entropy loss. Although this loss function takes into account

the geometric information of the areas being segmented, it only considers the length and area of inside and outside regions and does not actually consider the shape of the object being segmented. Furthermore, the formulation of the $Region$ term is a L_1 loss which has weak gradients compared to cross entropy. Using this loss function, we experimentally observed the phenomenon of slow learning and poor convergence.

3. Method

We propose a new interpretable image segmentation architecture called Shape Attentive U-Net (SAUNet). A SAUNet is composed of two streams - the texture stream and the gated shape stream. The texture stream is the same structure as an U-Net, however the encoder is replaced with dense blocks from DenseNet-121 similar to the Tiramisu Network proposed by Jégou et al. [21] and the decoder block used is our proposed dual attention decoder block. Dense pixel information and features are learned through the texture stream but not shape features. Takikawa et al. [40] proposed Gated-SCNN, in which the authors first introduced the idea of a gated shape stream to help remove noise and produce finer segmentation on the Cityscape dataset. We propose that on top of producing finer segmentations, the gated shape stream gets the model to learn object shapes (see Figure 3 and *supplementary materials*) and this is consequently interpretable. Hence, we propose the usage of a secondary stream on-top of our U-Net variant that processes shape and boundary information, the gated shape stream, explained in detail in Section 3.1.

3.1. Gated Shape Stream

The gated shape stream processes and refines relevant boundary and shape information using features processed by the encoder from the texture stream. The fusion of shape information from the shape stream flow with texture information is done by the gated convolutional layer.

3.1.1 Gated Convolutional Layer

Let $C_{1\times1}(x)$ denote the normalized 1x1 convolution function applied on feature map x, and let $R(x)$ denote the residual block function applied on feature map x. The residual block used is composed of two normalized 3x3 convolutions with a skip connection. The function $C_{1\times1}(x)$ returns a feature map of the same spatial dimensions as x but shrinks the number of channels down to one. The gated convolutional layer computes an attention map, α_l, of boundaries by using information from the shape stream flow and the texture stream. Formally, denote the shape stream feature maps as S_l and the texture stream feature maps as T_t where l denotes the layer number in our shape stream and t indexes the encoder block the texture stream feature maps

are outputted from. Bilinear interpolation is applied to T_t if needed to match the dimensions of S_l. Since we want precise boundaries of the shape, no pooling layers should be used in the shape stream. We define each residual block as a layer for the shape stream. Then, α_l is computed as,

$$\alpha_l = \sigma(C_{1x1}(S_l||C_{1x1}(T_t))) \tag{2}$$

where σ is the sigmoid function and $||$ denotes the channel-wise concatenation of feature maps. α_l is stacked channel-wise to match the dimensions of S_l. Then, the output of the gated convolutional layer, $\hat{S}_l$, is the inputted shape stream feature map S_l element-wise multiplied with α_l,

$$\hat{S}_l = S_l \otimes \alpha_l \tag{3}$$

where $\otimes$ is the Hadamard product. The feature map of the next layer of the shape stream is computed as,

$$S_{l+1} = R(\hat{S}_l) \tag{4}$$

and the same procedure to refine S_{l+1} is applied (equations 2-3 but now T_{t+1} is used).

3.1.2 Output of Gated Shape Stream

The predicted class boundaries from the shape stream are deeply supervised to produce L_{edge}. L_{edge} is the binary cross entropy loss between the ground truth class boundaries and the predicted class boundaries by the shape stream. Now, an objective of the model is to learn the shapes of the classes correctly. Since the entire gated shape stream is differentiable, the gradients propagate back to even the texture stream encoders. Intuitively, the texture stream encoders will learn some relevant shape information as well. The output of the gated shape stream is the predicted shape feature maps of the classes of interest concatenated channel-wise with the Canny edges from the original image. The output is then concatenated with the texture stream feature maps before the last normalized 3x3 convolution layer of the texture stream.

3.2. Dual Attention Decoder Block

The decoder module fuses feature maps outputted by the encoder from the skip connection along with the feature maps of lower resolution decoder blocks that capture more contextual and spatial information. Naturally, we would like to understand what features the model is detecting in these blocks to make the model less black-box. We propose the dual attention decoder block that is comprised of two new components after the standard normalized 3x3 convolution on the concatenated feature maps. The two new components are the spatial attention path for interpretability and
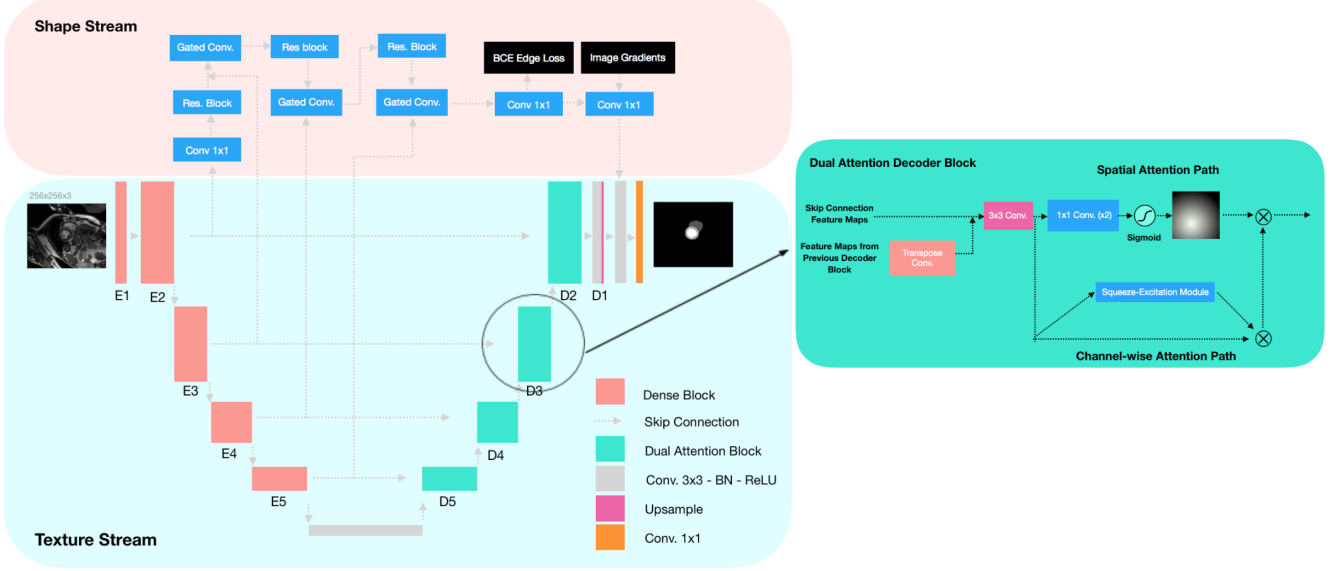
Figure 1. Our proposed Shape Attentive U-Net. The proposed model is composed of two main streams - the shape stream that processes boundary information and the texture stream. The shape stream is composed of gated convolutional layers and residual layers. The gated convolutional layers are used to fuse texture and shape information while the residual layers are used to fine-tune the shape features.

a channel-wise attention path for improved performance as demonstrated by Hu et al. [14].

3.2.1 Spatial Attention Path

Let C denote the number of channels coming into the spatial attention path. Then, the spatial attention path is composed of a normalized 1x1 convolution followed by another 1x1 convolution. The first convolution reduces the number of channels to $\frac{C}{2}$ and the second convolution reduces the number of channels to 1. A sigmoid is applied to map the pixel values in the single channel into the range of $[0, 1]$ to obtain F_s'. F_s' is then stacked channel-wise C times to obtain F_s. This is done to match the dimensions of the spatial attention path output, F_s, with the output from the channel-wise attention path, F_c, in order to perform element-wise multiplication.

3.2.2 Channel Attention Path

The channel-wise attention path is comprised of a squeeze and excitation module that produces a scale coefficient in $[0, 1]$ for each channel from the skip connection. Each channel from the skip connection feature map is then scaled by their respective coefficient to obtain F_c.

3.2.3 Channel and Spatial Attention

The output, F, of our proposed dual-attention decoder block is a fusion of channel and spatial attentions,

$$F = (F_s + 1) \otimes F_c \qquad (5)$$

Operator $\otimes$ denotes the Hadamard product. The $+1$ is included so that the spatial attention originally in the range of $[0, 1]$ can only amplify features rather than zeroing out features that may be valuable in later convolutions.

3.3. Dual-Task Loss Function for Learning Shapes

Our proposed objective function optimizes for precise segmentations and promotes the learning of shapes. We define each term of our objective in the following subsections.

3.3.1 Cross Entropy Loss

Cross entropy loss is a commonly used loss function for image segmentation tasks. It is derived from the concept of entropy in information theory. Formally, we define the cross entropy between two distributions p and q to be the average number of bits required to map an event using an estimated probability distribution q to the ground-truth distribution of p.

For image segmentation, the cross entropy loss is calculated as the average cross entropy over all pixels. Let Ω denote the domain of all pixels with height m, width n, and K classes. Let $y \in M_{m \times n \times K}(\{0, 1\})$ be the ground-truth one-hot matrix encoding the ground truth class of each pixel. Further, let $\hat{y} \in M_{m \times n \times K}([0, 1])$ be a matrix of the predicted probabilities of each individual pixel. The cross entropy loss is defined as,

$$L_{CE}(\hat{y}, y) = \frac{1}{|\Omega|} \sum_i^{\Omega} -y_i log(\hat{y}_i) - (1 - y_i) log(1 - \hat{y}_i) \quad (6)$$

3.3.2 Dice Loss

Dice loss is another common loss function used for image segmentation tasks as it measures the overlap and similarity between two sets. Given two countable sets A and B, the Dice coefficient is formally defined as,

$$Dice(A, B) = \frac{2|A \cap B|}{|A \cap B| + |A \cup B|} \tag{7}$$

Evidently, Dice(A, B) is maximized at 1 when $A = B$ and minimized at 0 when $A \cap B = \varnothing$. The Dice loss function should be inverted and differentiable, so we define the Dice loss as the following,

$$L_{Dice}(\hat{y}, y) = 1 - \frac{2}{K} \sum_{k=0}^{K-1} \frac{\sum_i^{\Omega} y^k \hat{y}_i^k}{\sum_i^{\Omega} y_i^k + \hat{y}_i^k} \tag{8}$$

where K is the total number of classes and y_i^k denotes the ith pixel of the kth indexed class of matrix $y \in M_{m \times n \times K}(\mathbb{R})$.

3.3.3 Dual Task Loss

We propose the loss function used to be composed of the segmentation loss and the shape stream boundary loss. Let L_{CE} and L_{Dice} denote the cross entropy loss and Dice loss of the predicted segmentation respectively. Let L_{Edge} denote the binary cross entropy loss of the predicted shape boundaries. Then, our total loss, L_{total}, is defined as,

$$L_{total} = \lambda_1 L_{CE} + \lambda_2 L_{Dice} + \lambda_3 L_{Edge} \tag{9}$$

where λ_1, λ_2, and λ_3 are hyper-parameters to weigh each measure. In our experiments, we find setting $\lambda_1 = \lambda_2 = \lambda_3 = 1$ works well. Moreover, we found that using L_{CE} without L_{Dice} and vice versa results in the model not learning well and quickly. Hence, both L_{CE} and L_{Dice} are used. Intuitively, the model learns to predict individual pixel values correctly through L_{CE} and also learns to consider overlap through L_{Dice}.

4. Experiments

For this study, we conduct our experiments on the SUN09 and AC17 segmentation datasets. We present our results in the following sections. The experiments were completed on one NVIDIA Quadro RTX 5000 GPU with 16GB of memory. For each experiment, each image slice was z-score normalized. The following data augmentations were performed during runtime: rotations in $[-\pi, \pi]$, horizontal and vertical flips with 50% chance, elastic deformations, and gamma shifts with

a factor sampled from $[0.5, 2.0]$ uniformly distributed. All the code used is publicly available on our GitHub: *https://github.com/sunjesse/shape-attentive-unet*

4.1. SUN09 Left Ventricle Segmentation Dataset

The SUN09 dataset contains separate training datasets for each of the two classes - the endocardium and the epicardium. The dataset for endocardium segmentation consists of 260 2D MRI slices. Likewise, the dataset for epicardium segmentation consists of 135 2D MRI slices. Each slice was center cropped to a resolution of 128px by 128px. Different datasets were given for each of the two classes so two models were trained for SUN09. RAdam [26] optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, along with weight decay value of 1E-4, and initial learning rate of 7E-4 exponentially decayed with parameter 0.99. Each model was trained for 120 epochs with a batch size of 4.

4.2. AC17 Bi-Ventricular Segmentation Dataset

The AC17 dataset contains 200 volumes of varying resolution MRI scans from 100 unique patients from the University Hospital of Dijon [4]. Each volume has a depth of 8-20 slices. To account for varying scales of each image, the pixels of every slice and their respective mask were resampled to 1.25mm by 1.25mm without changing the aspect ratio. The original image used bilinear interpolation while the segmentation mask used nearest-neighbour interpolation to avoid continuous values. The volumes were not rescaled along the z-axis. A center-crop of resolution 256px by 256px was made, and zero-padding was applied if necessary. We noticed that the minimum value of some slices was greater than zero; for consistency with the zero-padding made, the minimum value of every slice was subtracted before being zero-padded so that the minimum pixel intensity of every slice is 0 prior to adding the zero-pad and normalizing. The training dataset was randomly shuffled and split into an 80:20 train-validation split in which we seeded. The model was trained using RAdam for 180 epochs with $\beta_1 = 0.9$, $\beta_2 = 0.999$, along with weight decay value of 1E-4 and initial learning rate of 5E-4 exponentially decayed with parameter 0.99. Two-dimensional batch normalization was used with a batch size of 10. We report and submit the results to the test set of the convergence epoch.

5. Experimental Results

We present our results on the SUN09 left ventricle segmentation dataset and the AC17 left/right ventricle and myocardium segmentation dataset. The Dice Coefficient metric is used for evaluation consistent with other benchmarks and works.

5.1. SUN09

For SUN09, training our proposed model for 120 epochs took 1.5 hours for each class' dataset. No hyper-parameter tuning was done for our model trained with the shape stream. Our scores along with the previous top five scores are reported in Table 1.

Model	Endocardium	Epicardium
Tran [42]	0.92	0.96
Curiale et al. [7]	0.90	0.90
Yang et al. [46]	0.93	0.93
Romaguera et al. [32]	0.90	-
Avendi et al. [3]	0.94	-
SAUNet (Ours)	**0.952**	**0.962**

Table 1. Test set Dice scores for SUN09. Only the previous top five works with the highest reported accuracies (Dice scores) are listed. No hyper-parameter tuning was required for our SAUNet model.

5.2. AC17 Segmentation

For AC17, training our proposed model for 180 epochs took 3.5 hours. The Dice scores for the left ventricle, right ventricle, and myocardium segmentation structures are presented in Table 2 along with the scores from previous published works on AC17. The results presented are for the model trained with RAdam and a learning rate of 5E-4.

	LV	RV	MYO
Wolterink et al. [44]	0.930	0.880	0.870
Patravali et al. [30]	0.925	0.845	0.870
Khened et al. [23]	0.920	0.870	0.860
Ilias et al. [17]	0.905	0.760	0.785
Jang et al. [19]	0.938	0.890	0.879
SAUNet (Ours)	**0.938**	**0.914**	**0.887**

Table 2. AC17 test set results. Our proposed method yields state-of-the-art results for end-to-end models in all classes in terms of Dice score. Notably, our performance in the right ventricle class is significantly better than previous works, and this is perhaps due to the help of the shape stream learning the irregular shape of the right ventricle.

5.3. Ablation Studies

To test the effectiveness of our proposed gated shape stream, we conducted ablation studies on our model using the AC17 dataset. Results reported in Table 3 are from our validation set of our proposed model with and without the shape stream. Both models were trained with learning rate 5E-4 and RAdam. The results reported are the models taken at the convergence epoch. Both models were then evaluated on the test set. Figure 2 shows some cases from the test set where the model with the shape stream visibly outperforms its counterpart.

Model	RV	MYO	LV
SAUNet w/o Shape Stream	91.40	88.31	94.96
SAUNet w/ Shape Stream	**93.11**	**88.64**	**95.71**

Table 3. The Dice scores in percentage of our model with and without the proposed shape stream evaluated on our AC17 validation set split. The accuracy of the right ventricle class increased a significant amount as the irregular shape of the right ventricle was learned using the shape stream.
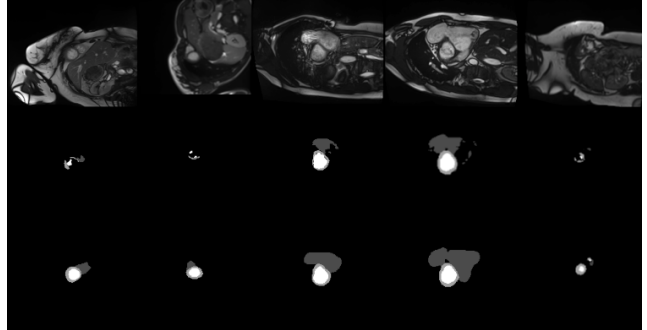

Figure 2. Top: original MRI image. Middle: model without shape stream prediction. Bottom: model with shape stream prediction. Higher quality segmentations with the correct shapes of the classes are made using the shape stream.

5.4. Robustness of Shape Stream

Intuitively, shapes are more robust features than textures. To evaluate the effectiveness of our dual-task objective of concurrently learning shapes through our proposed shape stream, we set up the following experiment: train the model with and without the shape stream on only SUN09 data, and then evaluate each model on our AC17 validation set split. The SUN09 dataset contains only labelled segmentations for the left ventricle, so only the left ventricle class was considered when evaluating on AC17. We trained both models with RAdam, initial learning rate of 5E-4, and a batch size of 4 for 120 epochs. The mIoU scores are reported in Table 4.

Model	SUN09 Train	AC17 Val.	Drop
w/o Shape Stream	90.20	79.33	-10.87
w/ Shape Stream	**90.84**	**81.73**	**-9.11**

Table 4. The mIoU scores in percentage of training on SUN09 and testing on our AC17 validation set split. The percentage drop of using the model with the shape stream is 1.76% less than otherwise suggesting shapes are robust features that generalize well and hence should be learned. The scores under SUN09 Train is the SUN09 test set mIoU score.
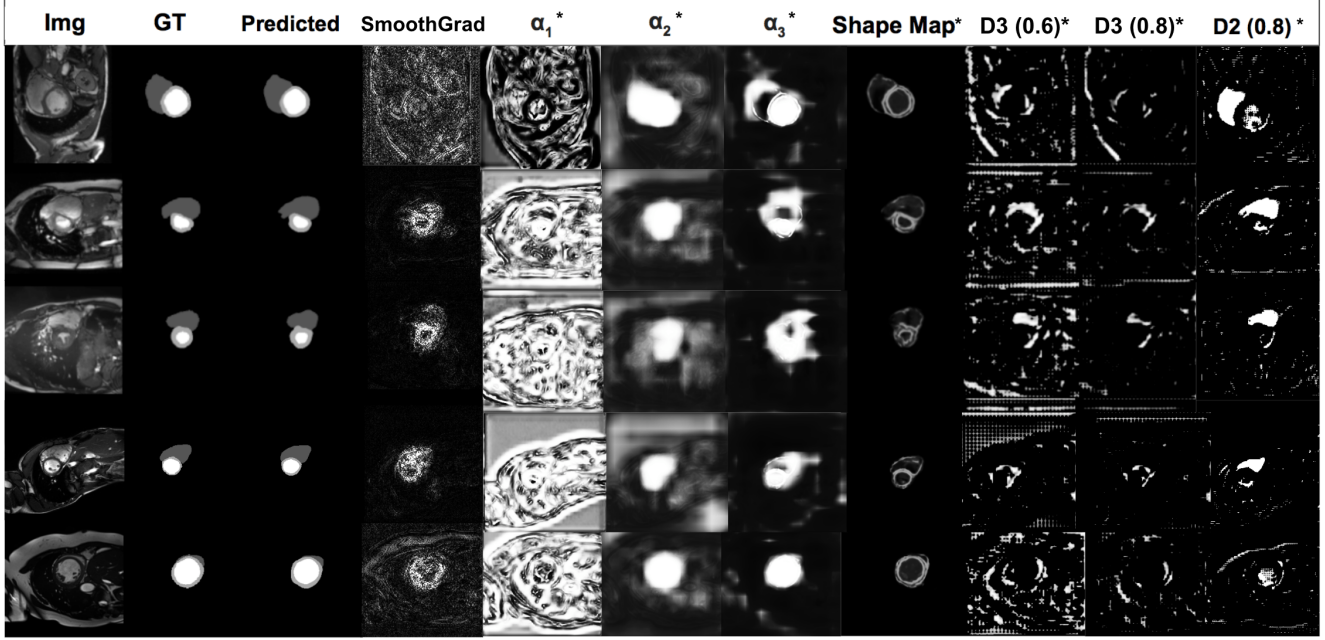
Figure 3. Our models attentions. * denotes proposed method. α_l is l^{th} shape attention map, $DX(Y)$ is X^{th} decoder block threshold of Y.

6. Interpretability

The learned shape and spatial attention maps are extractable from our model. The spatial attention maps can be used to interpret the regions of high activation for each decoder block while the learned shape maps can be used to deduce that the model has learned the correct shape of the classes of interest. Figure 3 contains the spatial attention map of the last two dual-attention decoder blocks, intermediate shape stream attentions α_l, the final shape attention map, and saliency maps generated using SmoothGrad [37].

Higher-resolution decoder blocks learned the feature of the right ventricle primarily according to column 11 (*D2(0.8)*) where the maps are thresholded by 0.8. In the last row, the right ventricle in the image is not visible so high attention was placed on the left ventricle. The right ventricle seems to have the highest attention and priority as its shape is unique and hence serves as a robust feature for localization. The attention maps from decoder block 3 are at a lower resolution level and according to column 10 (*D3(0.8)*) where the spatial attention maps are thresholded by 0.8, fewer areas of high activation are present compared to decoder block 2. Lower resolution blocks process more global information, so similar attention is placed among many regions of the image. To support this claim, column 9 (*D3(0.6)*) presents the same attention maps but thresholded by 0.6. Not only traces of attention around the right ventricle are apparent, many other regions like the brim of the structure are also in focus. Furthermore, the shape stream narrows down on the shape of the structures of interest evident in columns 5-8.

Our method of interpretability has a few advantages over post hoc analysis methods like saliency maps [36, 37]. SmoothGrad [37] is highly regarded currently as it overcomes the pitfalls of previous gradient-based saliency methods through averaging gradients. However, each image's saliency map requires 25-50 forward-backward passes to generate. Using SmoothGrad, it took 24 minutes to generate the saliency maps of all 384 images in our validation set, while it only took 20 seconds using our proposed method. Furthermore, our proposed method offers saliency methods at different levels and resolutions in which SmoothGrad and many other gradient-based methods do not provide. Conversely, GradCam [36] offers a solution to view the activation of every convolutional layer. However, GradCam requires one forward-backward pass post hoc per saliency map while our proposed method generates all of the multi-resolution saliency maps on the initial forward pass.

7. Conclusion

In this work, we present a new inherently interpretable medical image segmentation model called Shape Attentive U-Net. Our proposed method is able to learn robust shape features of objects via the gated shape stream while also being more interpretable than previous works via built-in saliency maps using attention. Furthermore, our method yields state-of-the-art results on the large public bi-ventrciular segmentation datasets of SUN09 and AC17. Through this work, we hope to take a step in making deep learning methods clinically adoptable and hopefully inspire more work on interpretability to be done in the future.

References

[1] S. M. Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh. Shape-aware deep convolutional neural network for vertebrae segmentation. In Ben Glocker, Jianhua Yao, Tomaž Vrtovec, Alejandro Frangi, and Guoyan Zheng, editors, *Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 12–24, Cham, 2018. Springer International Publishing. 2

[2] S. M. Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh. Spnet: Shape prediction using a fully convolutional neural network. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 430–439, Cham, 2018. Springer International Publishing. 2

[3] M. R. Avendi, A. Kheradvar, and H. Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri, 2015. 7

[4] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, and Miguel Angel Gonzalez Ballester. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 1, 2, 6

[5] Xu Chen, Bryan M. Williams, Srinivasa R. Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning active contour models for medical image segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[6] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jrgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851. 1

[7] Ariel H. Curiale, Flavio D. Colavecchia, Pablo Kaluza, Roberto A. Isoardi, and German Mato. Automatic myocardial segmentation by using a deep learning network in cardiac mri, 2017. 7

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018. 2

[9] Maryellen L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018. 1

[10] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016. 1

[11] Y. B. Guo and Bogdan J. Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *VISIGRAPP*, 2019. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[14] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017. 3, 5

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. 1

[16] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2016. 3

[17] Grinias Ilias and Georgios Tziritas. *Fast Fully-Automatic Cardiac Segmentation in MRI Using MRF Model Optimization, Substructures Tracking and B-Spline Smoothing*, pages 91–100. 01 2018. 7

[18] Fabian Isensee, Paul F Jaeger, Peter M Full, Ivo Wolf, Sandy Engelhardt, and Klaus H Maier-Hein. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In *International workshop on statistical atlases and computational models of the heart*, pages 120–129. Springer. 1

[19] Yeonggul Jang, Yoonmi Hong, Seongmin Ha, Sekeun Kim, and Hyuk-Jae Chang. Automatic segmentation of lv and rv in cardiac mri. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 161–169, Cham, 2018. Springer International Publishing. 7

[20] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018. 1, 3

[21] Simon Jgou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, 2016. 4

[22] Michael Kass, Andrew P. Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 3

[23] Mahendra Khened, Varghese Alex, and Ganapathy Krishnamurthi. Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 140–151, Cham, 2018. Springer International Publishing. 7

[24] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schtt, Sven Dhne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2017. 3

[25] A Gerche La, Guido Claessen, A de Bruaene Van, Nele Pattyn, J Cleemput Van, Marc Gewillig, Jan Bogaert, Steven

Dymarkowski, Piet Claus, and Hein Heidbuchel. Cardiac mri: a new gold standard for ventricular volume quantification during high-intensity exercise. *Circulation. Cardiovascular imaging*, 6(2):329–338, 2013. 1

[26] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2019. 6

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. 2

[28] Thomas N Martin, Bjoern A Groenning, Heather M Murray, Tracey Steedman, John E Foster, Alex T Elliot, Henry J Dargie, Ronald H Selvester, Olle Pahlm, and Galen S Wagner. St-segment deviation analysis of the admission 12-lead electrocardiogram as an aid to early diagnosis of acute myocardial infarction with a cardiac magnetic resonance imaging gold standard. *Journal of the American College of Cardiology*, 50(11):1021–1028, 2007. 1

[29] Christopher A Miller, Peter Jordan, Alex Borg, Rachel Argyle, David Clark, Keith Pearce, and Matthias Schmitt. Quantification of left ventricular indices from ssfp cine imaging: Impact of realworld variability in analysis methodology and utility of geometric modeling. *Journal of Magnetic Resonance Imaging*, 37(5):1213–1222, 2013. 2

[30] Jay Patravali, Shubham Jain, and Sasank Chilamkurthy. 2d-3d fully convolutional neural networks for cardiac mr segmentation. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 130–139, Cham, 2018. Springer International Publishing. 7

[31] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. 07 2009. 2

[32] Liset Vzquez Romaguera, Marly Guimares Fernandes Costa, Francisco Perdign Romero, and Cicero Ferreira Fernandes Costa Filho. Left ventricle segmentation in cardiac MRI images using fully convolutional neural networks. In Samuel G. Armato III and Nicholas A. Petrick, editors, *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 760 – 770. International Society for Optics and Photonics, SPIE, 2017. 7

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer. 1, 2

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[35] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images, 2018. 3

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2016. 3, 8

[37] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Vigas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. 3, 8

[38] Rupesh K Srivastava, Klaus Greff, and Jrgen Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385. 2

[39] William Stanford and Brad H Thompson. Coronary atherosclerosis and its effect on cardiac structure and function: evaluation by electron beam computed tomography. *Clinical chemistry*, 44(8):1871–1881, 1998. 2

[40] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation, 2019. 4

[41] Daniel Thomas, Harshali Bal, Jeffrey Arkles, James Horowitz, Luis Araujo, Paul D Acton, and Victor A Ferrari. Noninvasive assessment of myocardial viability in a small animal model: comparison of mri, spect, and pet. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 59(2):252–259, 2008. 2

[42] Phi Vu Tran. A fully convolutional neural network for cardiac segmentation in short-axis mri, 2016. 7

[43] Jean-Louis Vincent. Understanding cardiac output. *Critical care*, 12(4):174, 2008. 2

[44] Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine mr images. In Mihaela Pop, Maxime Sermesant, Pierre-Marc Jodoin, Alain Lalande, Xiahai Zhuang, Guang Yang, Alistair Young, and Olivier Bernard, editors, *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 101–110, Cham, 2018. Springer International Publishing. 7

[45] Saining Xie, Ross Girshick, Piotr Dollr, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2016. 3

[46] X. Yang, Z. Zeng, and S. Yi. Deep convolutional neural networks for automatic segmentation of left ventricle cavity from cardiac magnetic resonance images. *IET Computer Vision*, 11(8):643–649, 2017. 7