Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images

Jo Schlemper[a,1], Ozan Oktay[a,b,1], Michiel Schaap[b], Mattias Heinrich[c], Bernhard Kainz[a], Ben Glocker[a], Daniel Rueckert[a]

[a] *BioMedIA, Imperial College London, SW7 2AZ, London, UK*
[b] *HeartFlow, Redwood City, CA 94063, USA*
[c] *Medical Informatics, University of Luebeck, DE*

Abstract

We propose a novel attention gate (AG) model for medical image analysis that automatically learns to focus on target structures of varying shapes and sizes. Models trained with AGs implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task. This enables us to eliminate the necessity of using explicit external tissue/organ localisation modules when using convolutional neural networks (CNNs). AGs can be easily integrated into standard CNN models such as VGG or U-Net architectures with minimal computational overhead while increasing the model sensitivity and prediction accuracy. The proposed AG models are evaluated on a variety of tasks, including medical image classification and segmentation. For classification, we demonstrate the use case of AGs in scan plane detection for fetal ultrasound screening. We show that the proposed attention mechanism can provide efficient object localisation while improving the overall prediction performance by reducing false positives. For segmentation, the proposed architecture is evaluated on two large 3D CT abdominal datasets with manual annotations for multiple organs. Experimental results show that AG models consistently improve the prediction performance of the base architectures across different datasets and training sizes while preserving computational efficiency. Moreover, AGs guide the model activations to be focused around salient regions, which provides better insights

[1] The corresponding authors contributed equally.

into how model predictions are made. The source code for the proposed AG models is publicly available.

Keywords: Fully Convolutional Networks, Image Classification, Localisation, Segmentation, Soft Attention, Attention Gates

---

1. Introduction

Automated medical image analysis has been extensively studied in the medical imaging community due to the fact that manual labelling of large amounts of medical images is a tedious and error-prone task. Accurate and reliable solutions are required to increase clinical work flow efficiency and support decision making through fast and automatic extraction of quantitative measurements.

With the advent of convolutional neural networks (CNNs), near-radiologist level performance can be achieved in automated medical image analysis tasks including classification of Alzheimer's disease (Sarraf et al., 2017), skin lesions (Esteva et al., 2017; Kawahara and Hamarneh, 2016) and echo-cardiogram views (Madani et al., 2018), lung nodule detection in CT/X-ray (Liao et al., 2017; Zhu et al., 2018) and cardiac MR segmentation (Bai et al., 2017). An extensive list of applications can be found in (Litjens et al., 2017; Zaharchuk et al., 2018). High representation power, fast inference, and weight sharing properties have made CNNs the de facto standard for image classification and segmentation.

Methods for existing applications rely heavily on multi stage cascaded CNNs when the target organs show large inter-patient variation in terms of shape and size. Cascaded frameworks extract a region of interest (ROI) and make dense predictions on that particular ROI. The application areas include cardiac MRI (Khened et al., 2018), cardiac CT (Payer et al., 2017), abdominal CT (Roth et al., 2017, 2018) segmentation, and lung CT nodule detection (Liao et al., 2017). However, this approach leads to excessive and redundant use of computational resources and model parameters; for instance, similar low-level features are repeatedly extracted by all models within the cascade.

To address this general problem, we propose a simple and yet effective

solution, named *attention gates* (AGs). CNN models with AGs can be trained from scratch in a standard way similar to the training of fully convolutional network (FCN) models. Moreover, AGs automatically learn to focus on target structures without additional supervision. At test time, these gates generate soft region proposals implicitly on-the-fly and highlight salient features useful for a specific task. In addition, they do not introduce significant computational overhead and do not require a large number of model parameters as in the case of multi-model frameworks. In return, the proposed AGs improve model sensitivity and accuracy for global and dense label predictions by suppressing feature activations in irrelevant regions. In this way, the necessity of using an external organ localisation module can be eliminated while maintaining the high prediction accuracy. Similar attention mechanisms have been proposed for natural image classification (Jetley et al., 2018) and captioning (Anderson et al., 2017) to perform adaptive feature pooling, where model predictions are conditioned only on a subset of selected image regions. In this paper, we generalise this design and propose image-grid based gating that allows attention coefficients to be specific to local regions.

We demonstrate the performance of AG in real-time fetal ultrasound scan plane detection and CT pancreas segmentation. The first task is challenging due to low interpretability of the images and localising the object of interest is key to successful classification of the plane. To this end, we incorporate AGs into a variant of a VGG network, termed AG-Sononet, to demonstrate that attention mechanism can automatically localise the object of interest and improve the overall classification performance. The second task of pancreas segmentation is challenging due to low tissue contrast and large variability in organ shape and size. Moreover, we extend a standard U-Net architecture (*Attention U-Net*). We choose to evaluate our implementation on two commonly used benchmarks: TCIA Pancreas CT-82 (Roth et al., 2016) and multi-class abdominal CT-150. The results show that AGs consistently improve prediction accuracy across different datasets and training sizes while achieving state-of-the-art performance without requiring multiple CNN models.

1.1. Related Work

Attention Gates: AGs are commonly used in classification tasks such as in the analysis of citation graphs (Veličković et al., 2017) and natural images (Jetley et al., 2018; Wang et al., 2017a). Similarly in the context of natural language processing (NLP), such as image captioning (Anderson et al., 2017) and machine translation (Bahdanau et al., 2014; Luong et al., 2015; Shen et al., 2017; Vaswani et al., 2017), there have been several use cases of soft-attention models to efficiently use the given context information. In particular, given a sequence of text and a current word, a task is to extract a next word in a sentence generation or translation. The idea of attention mechanisms is to generate a *context* vector which assigns weights on the input sequence. Thus, the signal highlights the salient feature of the sequence conditioned on the current word while suppressing the irrelevant counter-parts, making the prediction more contextualised.

Initial work on attention modelling has explored salient image regions by interpreting gradient of output class scores with respect to the input image. Trainable attention, on the other hand, is enforced by design and categorised as hard- and soft-attention. Hard attention (Mnih et al., 2014), e.g. iterative region proposal and cropping, is often non-differentiable and relies on reinforcement learning for parameter updates, which makes model training more difficult. Ypsilantis and Montana (2017) used recursive hard-attention to detect anomalies in chest X-ray scans. Contrarily, soft attention is probabilistic, end-to-end differentiable, and utilises standard back-propagation without need for posterior sampling. For instance, additive soft attention is used in sentence-to-sentence translation (Bahdanau et al., 2014; Shen et al., 2017) and more recently applied to image classification (Jetley et al., 2018; Wang et al., 2017a).

In computer vision, attention mechanisms are applied to a variety of problems, including image classification (Jetley et al., 2018; Wang et al., 2017a; Zhao et al., 2017), segmentation (Ren and Zemel, 2016), action recognition (Liu et al., 2017; Pei et al., 2016; Wang et al., 2017b), image captioning (Lu et al., 2016; Xu et al., 2015), and visual question answering (Nam et al., 2016; Yang et al., 2015). Hu et al. (2017) used channel-wise attention to highlight important

feature dimensions, which was the top-performer in the ILSVRC 2017 image classification challenge. Similarly, non-local self attention was used by Wang et al. (2017b) to capture long range dependencies.

In the context of medical image analysis, attention models have been exploited for medical report generation (Zhang et al., 2017a,b) as well as joint image and text classification (Wang et al., 2018). However, for standard medical image classification, despite often the information to be classified are extremely localised, only a handful of works use attention mechanisms (Guan et al., 2018; Pesce et al., 2017). In these methods, either bounding box labels are available to guide the attention, or local context is extracted by a hard-attention model (i.e. region proposal followed by hard-cropping).

2D Ultrasound Scan Plane Detection: Fetal ultrasound screening is an important diagnostic protocol to detect abnormal fetal development. During screening examination, multiple anatomically standardised (NHS Screening Programmes, 2015) scan planes are used to obtain biometric measurements as well as identifying abnormalities such as lesions. Ultrasound suffers from low signal-to-noise ratio and image artefacts. As such, diagnostic accuracy and reproducibility is limited and requires a high level of expert knowledge and training. In the past, several approaches were proposed (Chen et al., 2015; Yaqub et al., 2015), however, they are computationally expensive and cannot be deployed for the real-time application. More recently, Baumgartner et al. (2016) proposed a CNN architecture called *Sononet*. It achieves very good performance in real-time plane detection, retrospective frame retrieval (retrieving the most relevant frame) and weakly supervised object localisation. However, it suffers from low recall value in differentiating different planar views of the cardiac chambers, which requires the method to be able to exploit the subtle differences in the local structure and it makes the problem challenging.

Pancreas Segmentation in 3D-CT Images: Early work on pancreas segmentation from abdominal CT used statistical shape models (Cerrolaza et al., 2016; Saito et al., 2016) or multi-atlas techniques (Oda et al., 2017; Wolz et al., 2013). In particular, atlas approaches benefit from implicit shape constraints

enforced by propagation of manual annotations. However, in public benchmarks such as the TCIA dataset (Roth et al., 2016), Dice similarity coefficients (DSC) for atlas-based frameworks are relatively low, ranging from 69.6% to 73.9% (Oda et al., 2017; Wolz et al., 2013). A classification based framework was proposed by Zografos et al. (2015) to remove the dependency of atlas to image registration. Recently, cascaded multi-stage CNN models (Roth et al., 2017, 2018; Zhou et al., 2017) have been proposed to address the problem. Here, an initial coarse-level model (e.g. U-Net or Regression Forest) is used to obtain a ROI and then a cropped ROI is used for segmentation refinement by a second model. Similarly, combinations of 2D-FCN and recurrent neural network (RNN) models are utilised by Cai et al. (2017) to exploit dependencies between adjacent axial slices. These approaches achieve state-of-the-art performance in the TCIA benchmark ($81.2\% - 82.4\%$ DSC). Without using a cascaded framework, the performance drops between 2.0% and 4.4%. Recently, Yu et al. (2017) proposed an iterative two-stage model that recursively updates local and global predictions, and both models are trained end-to-end. Besides standard FCNs, dense connections (Gibson et al., 2017) and sparse convolutions (Heinrich et al., 2018; Heinrich and Oktay, 2017) have been applied to the CT pancreas segmentation problem. Dense connections and sparse kernels reduce computational complexity by requiring less number of non-zero parameters.

1.2. Contributions

In this paper, we propose a novel soft-attention gating module that can be utilised in CNN based standard image analysis models for dense label predictions. Additionally, we explore the benefit of AGs to medical image analysis, in particular, in the context of image classification and segmentation. The contributions of this work can be summarised as follows:

- We take the attention approach proposed by Jetley et al. (2018) a step further by proposing grid-based gating that allows attention gates to be more specific to local regions. This improves performance compared to

gating based on a global feature vector. Moreover, our approach is not only limited to adaptive pooling (Jetley et al., 2018) but can be also used for dense predictions as in segmentation networks.

- We propose one of the first use cases of soft-attention in a feed-forward CNN model applied to a medical imaging task that is end-to-end trainable. The proposed attention gates can replace hard-attention approaches used in image classification (Ypsilantis and Montana, 2017) and external organ localisation models in image segmentation frameworks (Khened et al., 2018; Oda et al., 2017; Roth et al., 2017, 2018). This also eliminates the need for any bounding box labels and backpropagation-based saliency map generation used by Baumgartner et al. (2016).

- For classification, we apply the proposed model to real-time fetal ultrasound scan plane detection and show its superior classification performance over the baseline approach. We show that attention maps can used for fast (weakly-supervised) object localisation, demonstrating that the attended features indeed correlate with the anatomy of interest.

- For segmentation, an extension to the standard U-Net model is proposed that provides increased sensitivity without the need of complicated heuristics, while not sacrificing specificity. We demonstrate that accuracy improvements when using U-Net are consistent across different imaging datasets and training sizes.

- We demonstrate that the proposed attention mechanism provides fine-scale attention maps that can be visualised, with minimal computational overhead, which helps with interpretability of predictions.

2. Methodology

2.1. Convolutional Neural Network

CNNs are now the state-of-the-art method for many tasks including classification , localisation and segmentation (Bai et al., 2017; Kamnitsas et al., 2017,

2018; Lee et al., 2015; Litjens et al., 2017; Long et al., 2015; Ronneberger et al., 2015; Roth et al., 2017, 2018; Xie and Tu, 2015; Zaharchuk et al., 2018). CNNs outperform traditional approaches in medical image analysis while being an order of magnitude faster than, e.g., graph-cut and multi-atlas segmentation techniques (Wolz et al., 2013). The success of CNNs is attributed to the fact that (I) domain specific image features are learnt using stochastic gradient descent (SGD) optimisation, (II) learnt kernels are shared across all pixels, and (III) image convolution operations exploit the structural information in medical images in an optimal fashion. However, it remains difficult to reduce false-positive predictions for small objects that show large shape variability. In such cases, in order to improve the accuracy, current frameworks (Guan et al., 2018; Khened et al., 2018; Roth et al., 2017, 2018) rely on additional preceding object localisation models to simplify the task into separate localisation and subsequent classification/segmentation steps, or guide the localisation using weak labels (Pesce et al., 2017). Here, we demonstrate that the same objective can be achieved by integrating attention gates (AGs) in a standard CNN model. This does not require the training of multiple models and a large number of extra model parameters. In contrast to the localisation model in multi-stage CNNs, AGs progressively suppress feature responses in irrelevant background regions without the requirement to crop a ROI between networks.

2.2. Attention Gate Module

We now introduce *Attention Gate* (AG), which is a mechanism which can be incorporated in any existing CNN architecture. Let $\mathbf{x}^l = \{\mathbf{x}_i^l\}_{i=1}^n$ be the activation map of a chosen layer $l \in \{1, \dots, L\}$, where each $\mathbf{x}_i^l$ represents the pixel-wise feature vector of length F_l (i.e. the number of channels). For each $\mathbf{x}_i^l$, AG computes coefficients $\alpha^l = \{\alpha_i^l\}_{=1}^n$, where $\alpha_i^l \in [0, 1]$, in order to identify salient image regions and prune feature responses to preserve only the activations relevant to the specific task as shown in Figure 1. The output of AG is $\hat{\mathbf{x}}^l = \{\alpha_i^l \mathbf{x}_i^l\}_{i=1}^n$, where each feature vector is scaled by the corresponding attention coefficient.

The attention coefficients α_i^l are computed as follows: In standard CNN

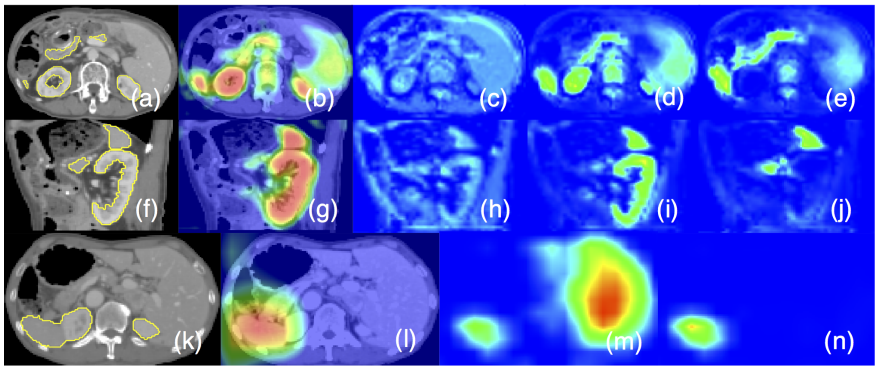

Figure 1: Axial (a) and sagittal (f) views of a 3D abdominal CT scan, (b,g) attention coefficients, feature activations of a skip connection before in (c,h) and after gating (d,e,i,j). Similarly, (k-n) visualise the gating on a coarse scale skip connection. The filtered feature activations (d,e,i,j) are collected from multiple AGs, where a subset of organs is selected by each gate and activations consistently correspond to specific structures across different scans.

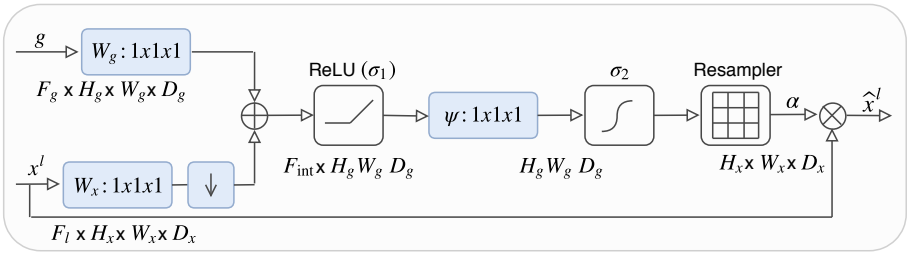

Figure 2: Schematic of the proposed additive attention gate (AG). Input features (x^l) are scaled with attention coefficients (α) computed in AG. Spatial regions are selected by analysing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale. Grid resampling of attention coefficients is performed using trilinear interpolation.

architectures, to capture a sufficiently large receptive field and thus, semantic contextual information, the feature-map is gradually downsampled. The features on the coarse spatial grid level identify location of the target objects and model their relationship at global scale. Let $\mathbf{g} \in \mathbb{R}^{F_g}$ be such global feature vector. The idea is to consider each $\mathbf{x}_i^l$ and $\mathbf{g}$ jointly to attend the features at each scale l that are most relevant to the objective being minimised.

We use additive attention (Bahdanau et al., 2014) to obtain the gating coefficient as can be seen in Figure 2, which is formulated as follows:

$$q_{att,i}^{l} = \boldsymbol{\psi}^{T} \left(\sigma_1 \left(\boldsymbol{W}_x^{T} \boldsymbol{x}_i^{l} + \boldsymbol{W}_g^{T} \boldsymbol{g} + \boldsymbol{b}_{xg} \right) \right) + b_{\psi} \tag{1}$$

$$\alpha^{l} = \sigma_2(q_{att}^{l}(\boldsymbol{x}^{l}, \boldsymbol{g}; \boldsymbol{\Theta}_{att})), \tag{2}$$

where $\sigma_1(x)$ is an element-wise nonlinearity (e.g. rectified linear-unit) and $\sigma_2(x)$ is a normalisation function. For example, one can apply sigmoid to restrict the range to $[0, 1]$, or one can apply softmax operation $\alpha_i^l = e^{q_{att,i}^l} / \sum_i e^{q_{att,i}^l}$ such that the attention map sums to 1. AG is therefore characterised by a set of parameters $\boldsymbol{\Theta}_{att}$ containing: linear transformations $\boldsymbol{W}_x \in \mathbb{R}^{F_l \times F_{int}}$, $\boldsymbol{W}_g \in \mathbb{R}^{F_g \times F_{int}}$, $\boldsymbol{\psi} \in \mathbb{R}^{F_{int} \times 1}$ and bias terms $b_{\psi} \in \mathbb{R}$, $\boldsymbol{b}_{xg} \in \mathbb{R}^{F_{int}}$. The linear transformations are computed using channel-wise $1 \times 1 \times 1$ convolutions for the input tensors.

We note that AG parameters can be trained with the standard back-propagation updates without a need for sampling based update methods used in hard-attention (Mnih et al., 2014). While AG do not require auxiliary loss function to optimise, we found that using deep-supervision (Lee et al., 2015) encourages the intermediate feature-maps to be semantically discriminative at each image scale. This ensures that attention units, at different scales, have an ability to influence the responses to a large range of image foreground content. We therefore prevent dense predictions from being reconstructed from small subsets of skip connections.

2.2.1. Multi-dimensional Attention

In case of where multiple semantic classes are present in the image, one can learn multi-dimensional attention coefficients. This is inspired by the approach of Shen et al. (2017), where multi-dimensional attention coefficients are used to learn sentence embeddings. Thus, each AG learns to focus on a subset of target structures. In case of multi-dimensional AGs, each α_k^l corresponds to a vector and produce $\hat{\boldsymbol{x}}^l = [\alpha_{(1)}^l \odot \boldsymbol{x}^l, \ldots, \alpha_{(m)}^l \odot \boldsymbol{x}^l]$ where $\alpha_{(k)}^l$ is k-th sub AG. In each sub-AG, complementary information is extracted and fused to define the output

of skip connection.

2.2.2. Gating Signal and Grid Attention

As the gating signal $\mathbf{g}$ must encode global information from large spatial context, it is usually obtained from the coarsest scale activation map. For example in classification, one could use the activation map just before the final softmax layer. In the context of medical imaging, however, since most objects of interest are highly localised, flattening may have the disadvantage of losing important spatial context. In fact, in many cases a few max-pooling operations are sufficient to infer the global context without explicitly using the global pooling. Therefore, we propose a *grid attention* mechanism. The idea is to use the coarse scale feature map before any flattening is done. For example, given an input tensor size of $F_l \times H_x \times W_x$, after r max pooling operations, the tensor size is reduced to $F_g \times H_g \times W_g = F_g \times H_x/(2^r) \times W_y/(2^r)$. To generate the attention map, we can either downsample or upsample the coarse grid to match the spatial resolution of $\boldsymbol{x}^l$. In this way, the attention mechanism has more flexibility in terms of what to focus on a regional basis. For upsampling, we chose to use bilinear upsampling. Note that the upsampling can be replaced by a learnable weight, however, we did not opt for this for the sake of simplicity. For segmentation, one can directly use the coarsest activation map as the gating signal.

2.2.3. Backward Pass through Attention Gates

Information extracted from coarse scale is used in gating to disambiguate irrelevant and noisy responses in skip connections. This is performed right before the concatenation operation to merge only relevant activations. Additionally, AGs filter the neuron activations during the forward pass as well as during the backward pass. Gradients originating from background regions are down weighted during the backward pass. This allows model parameters in shallower layers to be updated mostly based on spatial regions that are relevant to a given task. The update rule for convolution parameters in layer $l-1$ can be formulated

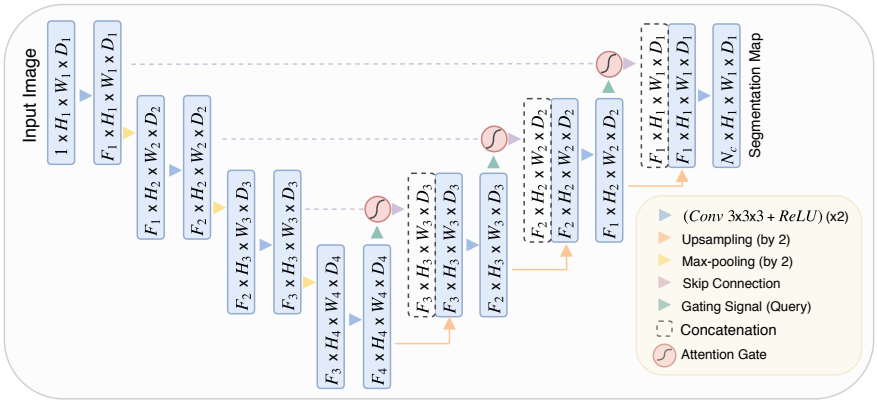

Figure 3: A block diagram of the proposed Attention U-Net segmentation model. Input image is progressively filtered and downsampled by factor of 2 at each scale in the encoding part of the network (e.g. $H_4 = H_1/8$). N_c denotes the number of classes. Attention gates (AGs) filter the features propagated through the skip connections. Schematic of the AGs is shown in Figure 2. Feature selectivity in AGs is achieved by use of contextual information (gating) extracted in coarser scales.

as follows:

$$\frac{\partial(\hat{x}_i^l)}{\partial\left(\Phi^{l-1}\right)} = \frac{\partial\left(\alpha_i^l\, f(x_i^{l-1};\Phi^{l-1})\right)}{\partial\left(\Phi^{l-1}\right)} = \alpha_i^l \frac{\partial(f(x_i^{l-1};\Phi^{l-1}))}{\partial\left(\Phi^{l-1}\right)} + \frac{\partial(\alpha_i^l)}{\partial\left(\Phi^{l-1}\right)} x_i^l \qquad (3)$$

where the first gradient term on the right-hand side is scaled with α_i^l.

2.3. Attention Gates for Segmentation

In this paper, we build our attention-gated segmentation model on top of a standard 3D U-Net architecture. U-Nets are commonly used for image segmentation tasks because of their good performance and efficient use of GPU memory. The latter advantage is mainly linked to extraction of image features at multiple image scales. Coarse feature-maps capture contextual information and highlight the category and location of foreground objects. Feature-maps extracted at multiple scales are later merged through skip connections to combine coarse- and fine-level dense predictions as shown in Figure 3. The proposed AGs are incorporated into the standard U-Net architecture to highlight salient features that are passed through the skip connections. For AGs, we chose sigmoid

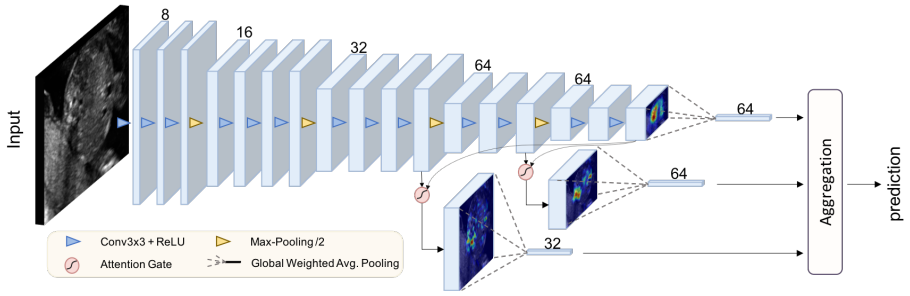

Figure 4: The schematics of the proposed classification network, termed *Attention-Gated Sononet*. The proposed attention units are incorporated in layer 11 and layer 14. The activation map after the attention gate

activation function for normalisation: $\sigma_2(x) = \frac{1}{1+exp(-x)}$. While in image captioning (Anderson et al., 2017) and classification (Jetley et al., 2018) tasks, the softmax activation function is used to normalise the attention coefficients σ_2, however, sequential use of softmax yields sparser activations at the output. For dense prediction task, we empirically observed that sigmoid resulted in better training convergence for the AG parameters.

2.4. Attention Gates for Classification

For attention-gated classifier, we chose our base architecture to be *Sononet* (Baumgartner et al., 2016), which is a variant of VGG network (Simonyan and Zisserman, 2014). The difference is that Sononet can be decoupled into feature extraction module and adaptation module. In the adaptation module, the number of channels are first reduced to the number of target classes C. Subsequently, the spatial information is flattened via channel-wise global average pooling. Finally, a softmax operation is applied to the resulting vector and the entry with maximum activation is selected as the prediction. As the network is constrained to classify based on the reduced vector, the network is forced to extract the most salient features for each class.

The proposed attention mechanism is incorporated in the Sononet architecture to better exploit local information. In the modified architecture, termed Attention-Gated Sononet (*AG-Sononet*), we remove the adaptation module. The

final layer of the feature extraction module is used as gridded global feature map $\mathbf{g}$. We apply the proposed attention mechanism to layer 11 and 14 just before pooling, as shown in Figure 4. After the attention coefficients $\{\alpha_i^l\}_{i=1}^n$ are obtained, the weighted average over the spatial axes is computed, yielding a vector of length F_l at scale l: $\tilde{\boldsymbol{x}}^l = \sum_{i=1}^n \alpha_i^l x_i^l$. In addition, we also perform the global average pooling on the coarsest scale representation. The prediction is given by fitting a fully connected layer on the concatenated feature vector $\{\tilde{\boldsymbol{x}}^{l_1}, \tilde{\boldsymbol{x}}^{l_2}, \tilde{\boldsymbol{x}}^{l_3}\}$ (e.g. $l_1 = 11, l_2 = 14, l_3 = 17$). We note that for AG-sononet, we normalised the attention coefficients as $\alpha_i^l = (\alpha_i^l - \alpha_{min}^l / \sum_j (\alpha_j^l - \alpha_{min}^l))$, where $\alpha_{min}^l = \min_j \alpha_j^l$, as we realised that softmax output was often too sparse, making the prediction more challenging.

2.4.1. Aggregation Strategy

Given the attended feature vectors at different scales, we highlight that the aggregation strategy is flexible and that it can be adjusted depending on the target problem. The simplest is to just fit a fully connected layer on the concatenated vector as mentioned above. However, we noticed that sometimes the network abandons the fine-scale attention mechanisms as it is non-trivial to train. An alternative approach is to fit a separate fully connected (FC) layer at each scale and make separate predictions. The final prediction is then given by either weighted mean or max operations. One can also use deep-supervision (Lee et al., 2015) to force each scale to learn a useful prediction as well as when combined. We empirically observed that first training the network at each scale, then fine-tuning using a new FC layer fitted on the concatenated vector worked the best. In the experimentation, we considered the following variations:

- *AG-Sononet*: simple averaging of the predictions at each scale,
- *AG-Sononet-DS*: model making one final prediction from concatenated vector, which also uses deep supervision at each scale, and
- *AG-Sononet-FT*: AG-Sononet that is subsequently finetuned to predict from the concatenated vector.

3. Experiments and Results

The proposed AG model is modular and independent of application type; as such it can be easily adapted for pixel and image level classification tasks. To demonstrate its applicability to image classification and segmentation, we evaluate the proposed attention based FCN models on challenging abdominal CT multi-label segmentation and 2D ultrasound image plane classification problems. In particular, pancreas boundary delineation is a difficult task due to shape-variability and poor tissue contrast, similarly image quality and subject variability introduce challenges in 2D-US image classification. Our models are compared against the standard 3D U-Net and Sononet in terms of model prediction performance, model capacity, computation time, and memory requirements.

3.1. Evaluation Datasets

In this section, we present the image datasets used in classification and segmentation experiments.

3.1.1. 3D-CT Abdominal Image Datasets

For the experiments, two different CT abdominal datasets are used: (I) 150 abdominal 3D CT scans acquired from patients diagnosed with gastric cancer (CT-150). In all images, the pancreas, liver, and spleen boundaries were semi-automatically delineated by three trained researchers and manually verified by a clinician. The same dataset is used by Roth et al. (2017) to benchmark the U-Net model in pancreas segmentation. (II) The second dataset[2] (CT-82) consists of 82 contrast enhanced 3D CT scans with pancreas manual annotations performed slice-by-slice. This dataset (NIH-TCIA) (Roth et al., 2016) is publicly available and commonly used to benchmark CT pancreas segmentation frameworks. The images from both datasets are downsampled to isotropic 2.00 mm resolution due to the large image size and hardware memory limitations.

[2] `https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT`

3.1.2. 2D Fetal Ultrasound Image Dataset

Our dataset consisted of 2694 2D ultrasound examinations of volunteers with gestational ages between 18 and 22 weeks. The dataset contains 13 types of standard scan planes and background, complying the standard specified in the UK National Health Service (NHS) fetal anomaly screening programme (FASP) handbook (NHS Screening Programmes, 2015). The standard scan planes are: Brain (Cb.), Brain (Tv.), Profile, Lips, Abdominal, Kidneys, Femur, Spine (Cor.), Spine (Sag.), 4CH, 3VV, RVOT, LVOT. The dataset further includes large portions of frames which contains anatomies that are not part of the scan plane, labelled as "background". The details of the image acquisition protocol as well as how scan plane labels are obtained can be found in (Baumgartner et al., 2016). The data was cropped to central 208×272 to prevent the network from learning the surrounding annotations shown in the ultrasound scan screen.

3.2. Model Training and Implementation Details

The datasets used in this manuscript contains large class imbalance issue that needs to be addressed. For ultrasound dataset, due to the nature of screening process, the background label dominates the dataset. To address this, we used a weighted sampling strategy: the sampling "probability" of an image from class c is given by $1/n_c$, where n_c is the number of images in class c. For the background label, we used $13/n_{\text{background}}$, where 13 is the number of the standard scan planes. In this way, we expect to see one background image for every standard scan plane. For the segmentation models, the class imbalance problem is tackled using the Sorensen-Dice loss (Drozdzal et al., 2016; Milletari et al., 2016) defined over all semantic classes. Dice loss is experimentally shown to be less sensitive to class imbalance in segmentation tasks.

For both tasks, batch-normalisation, deep-supervision (Lee et al., 2015), and standard data-augmentation techniques (affine transformations, axial flips, random crops) are used in training attention and baseline networks. Intensity values are linearly scaled to obtain a normal distribution $N(0, 1)$. For classification models, we empirically found that optimising with Stochastic Gradient Descent

Table 1: Multi-class CT abdominal segmentation results obtained on the *CT*-150 dataset: The results are reported in terms of Dice score (DSC) and mesh surface to surface distances (S2S). These distances are reported only for the pancreas segmentations. The proposed Attention U-Net model is benchmarked against the standard U-Net model for different training and testing splits. Inference time (forward pass) of the models are computed for input tensor of size $160 \times 160 \times 96$. Statistically significant results are highlighted in bold font.

Method	U-Net	Att U-Net	U-Net	Att U-Net
Train/Test Split	120/30	120/30	30/120	30/120
Pancreas DSC	0.814±0.116	**0.840±0.087**	0.741±0.137	**0.767±0.132**
Pancreas Precision	0.848±0.110	0.849±0.098	0.789±0.176	**0.794±0.150**
Pancreas Recall	0.806±0.126	**0.841±0.092**	0.743±0.179	**0.762±0.145**
Pancreas S2S Dist (mm)	2.358±1.464	**1.920±1.284**	3.765±3.452	3.507±3.814
Spleen DSC	0.962±0.013	0.965±0.013	0.935±0.095	**0.943±0.092**
Kidney DSC	0.963±0.013	0.964±0.016	0.951±0.019	0.954±0.021
Number of Params	5.88 M	6.40 M	5.88 M	6.40 M
Inference Time	0.167 s	0.179 s	0.167 s	0.179 s

with Nesterov momentum ($\rho = 0.9$) worked the best. The initial learning rate was set to 0.1, which was subsequently reduced by a factor of 0.1 for every 100 epoch. We also used a warm-start learning rate of 0.01 for the first 5 epochs. For segmentation models, we used Adam with $\alpha = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$. The batch size for the Sononet models was set to 64. However, for the 3D-CT segmentation models, gradient updates are computed using small batch sizes of 2 to 4 samples. For larger segmentation networks, gradient averaging is used over multiple forward and backward passes. This is mainly because we propose a 3D-model to capture sufficient semantic context in contrast to the state-of-the-art CNN segmentation frameworks (Cai et al., 2017; Roth et al., 2018). Gating parameters are initialised so that attention gates pass through feature vectors at all spatial locations. Moreover, we do not require multiple training stages as in hard-attention based approaches therefore simplifying the training procedure.

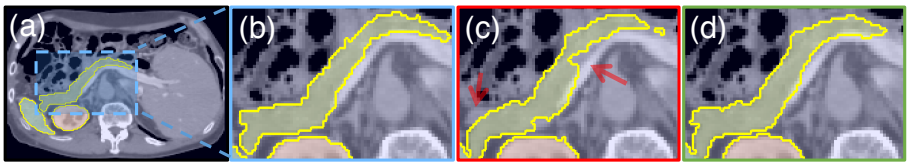

Figure 5: The ground-truth pancreas segmentation (a) is highlighted in blue (b). Similarly, U-Net (c) and Attention U-Net model (d) predictions are shown. The missed dense predictions by U-Net are highlighted with red arrows.

3.2.1. Implementation Details of Sononet:

For the classification network, we modified the baseline Sononet architecture slightly: instead of using 2 convolution layers for the first 2 feature scales and 3 convolution layers for the last 3 feature scales, we used 3 layers for the first 3 and 2 layers for the last 2 feature scales. The architecture for AG-sononet is shown in Fig. 4. The parameters for AG-Sononet was initialised using a partially trained Sononet. We compare our models with different capacities, with initial number of features 8, 16 and 32. For U-net and Attention U-net, the initial number of features is set to $F_1 = 8$, which is doubled after every max-pooling operation. Our implementation using PyTorch (Paszke et al., 2017) is publicly available[3].

3.3. 3D-CT Abdominal Image Segmentation Results

The proposed Attention U-Net model is benchmarked against the standard U-Net (Ronneberger et al., 2015) on multi-class abdominal CT segmentation. We use CT-150 dataset for both training (120) and testing (30). The corresponding Dice scores (DSC) and surface distances (S2S) are given in Table 1. The results on pancreas predictions demonstrate that attention gates (AGs) increase recall values ($p = .005$) by improving the model's expression power as it relies on AGs to localise foreground pixels. The difference between predictions obtained with these two models are qualitatively compared in Figure 5. In the second experiment, the same models are trained with fewer training images (30) to show that the performance improvement is consistent and significant for different sizes

[3] `https://github.com/ozan-oktay/Attention-Gated-Networks`

Table 2: Segmentation experiments on CT-150 dataset are repeated with higher capacity U-Net models to demonstrate the efficiency of the attention models with similar or less network capacity. The additional filters in the U-Net model are distributed uniformly across all the layers. Segmentation results for the pancreas are reported in terms of dice score, precision, recall, surface distances. The models are trained with the same train/test data splits (120/30).

Method	# of Pars	DSC	Precision	Recall	S2S Dist (mm)	Run Time
U-Net	6.44 M	.821±.119	.849±.111	.814±.125	2.383±1.918	.191 s
U-Net	10.40 M	.825±.104	.861±.082	.807±.121	2.202±1.144	.222 s

Table 3: Pancreas segmentation results obtained on the TCIA Pancreas-CT Dataset (Roth et al., 2016). The dataset contains in total 82 scans which are split into training (61) and testing (21) sets. The corresponding results are obtained before (BFT) and after fine tuning (AFT) and also training the models from scratch (SCR). Statistically significant results are highlighted in bold font.

	Method	Dice Score	Precision	Recall	S2S Dist (mm)
BFT	U-Net	0.690±0.132	0.680±0.109	0.733±0.190	6.389±3.900
	Attention U-Net	**0.712±0.110**	0.693±0.115	**0.751±0.149**	**5.251±2.551**
AFT	U-Net	0.820±0.043	0.824±0.070	0.828±0.064	2.464±0.529
	Attention U-Net	**0.831±0.038**	0.825±0.073	**0.840±0.053**	**2.305±0.568**
SCR	U-Net	0.815±0.068	0.815±0.105	0.826±0.062	2.576±1.180
	Attention U-Net	0.821±0.057	0.815±0.093	**0.835±0.057**	**2.333±0.856**

of training data ($p = .01$). For both approaches, we observe a performance drop on spleen DSC as the training size is reduced. The drop is less significant with the proposed framework. For kidney segmentation, the models achieve similar accuracy since the tissue contrast is higher.

In Table 1, we also report the number of trainable parameters for both models. We observe that by adding 8% extra capacity to the standard U-Net, the performance can be improved by 2-3% in terms of DSC. For a fair comparison, we also train higher capacity U-Net models and compare against the proposed model with smaller network size. The results shown in Table 2 demonstrate that the addition of AGs contributes more than simply increasing model capacity

Table 4: State-of-the-art CT pancreas segmentation methods that are based on single and multiple CNN models. The listed segmentation frameworks are evaluated on the same public benchmark (CT-82) using different number of training and testing images. Similarly, the FCN approach proposed in (Roth et al., 2017) is benchmarked on CT-150 although it is trained on an external dataset (Ext).

Method	Dataset	Pancreas DSC	Train/Test	# Folds
Hierarchical 3D FCN (Roth et al., 2017)	CT-150	82.2 ± 10.2	Ext/150	-
Dense-Dilated FCN (Gibson et al., 2017)	CT-82 & Synapse[4]	66.0 ± 10.0	63/9	5-CV
2D U-Net (Heinrich et al., 2018)	CT-82	75.7 ± 9.0	66/16	5-CV
HN 2D FCN Stage-1(Roth et al., 2018)	CT-82	76.8 ± 11.1	62/20	4-CV
HN 2D FCN Stage-2(Roth et al., 2018)	CT-82	81.2 ± 7.3	62/20	4-CV
2D FCN (Cai et al., 2017)	CT-82	80.3 ± 9.0	62/20	4-CV
2D FCN + RNN (Cai et al., 2017)	CT-82	82.3 ± 6.7	62/20	4-CV
Single Model 2D FCN (Zhou et al., 2017)	CT-82	75.7 ± 10.5	62/20	4-CV
Multi-Model 2D FCN (Zhou et al., 2017)	CT-82	82.2 ± 5.7	62/20	4-CV

(uniformly) across all layers of the network ($p = .007$). Therefore, additional capacity should be used for AGs to localise tissues, in cases when AGs are used to reduce the redundancy of training multiple, individual models.

3.3.1. Comparison to State-of-the-Art CT Abdominal Segmentation Frameworks

The proposed architecture is evaluated on the public TCIA CT Pancreas benchmark to compare its performance with state-of-the-art methods. Initially, the models trained on CT-150 dataset are directly applied to CT-82 dataset to observe the applicability of the two models on different datasets. The corresponding results (BFT) are given in Table 3. U-Net model outperforms traditional atlas techniques (Wolz et al., 2013) although it was trained on a disjoint dataset. Moreover, the attention model performs consistently better in pancreas segmentation across different datasets. These models are later fine-tuned (AFT) on a subset of TCIA dataset (61 train, 21 test). The output nodes corresponding to spleen and kidney are excluded from the output softmax computation, and the gradient updates are computed only for the background and pancreas labels. The results in Table 3 and 4 show improved performance compared to concatenated multi-model CNN approaches (Cai et al., 2017; Roth et al., 2018; Zhou

et al., 2017) due to additional training data and richer semantic information (e.g. spleen labels). Additionally, we trained the two models from scratch (SCR) with 61 training images randomly selected from the CT-82 dataset. Similar to the results on CT-150 dataset, AGs improve the segmentation accuracy and lower the surface distances ($p = .03$) due to increased recall rate of pancreas pixels ($p = .09$).

Results from state-of-the-art CT pancreas segmentation models are summarised in Table 4 for comparison purposes. Since the models are trained on the same training dataset, this comparison gives an insight on how the attention model compares to the relevant literature. It is important to note that, post-processing (e.g. using conditional random field) is not utilised in our framework as the experiments mainly focus on quantification of performance improvement brought by AGs in an isolated setting. Similarly, residual and dense connections can be used as in (Gibson et al., 2017) in conjunction with AGs to improve the segmentation results. In that regard, our 3D Attention U-Net model performs similar to the state-of-the-art, despite the input images are downsampled to lower resolution. More importantly, our approach significantly improves the results compared to single-model based segmentation frameworks (see Table 4). We do not require multiple CNN models to localise and segment object boundaries. Lastly, we performed 5-fold cross-validation on the CT-82 dataset using the Attention U-Net for a better comparison, which achieved 81.48 ± 6.23 DSC for pancreas labels.

3.4. 2D Fetal Ultrasound Image Classification Results

The dataset was split to training $(122,233)$, validation $(30,553)$ and testing $(38,243)$ frames on subject basis. For evaluation, we used accuracy, precision, recall, F1, the number of parameters and execution speed. Note that due to aforementioned class imbalance, it is important to take the macro-averaging for precision, recall and F1: e.g $\text{recall}_{\text{macro}} = (\text{recall}_{c_1} + \cdots + \text{recall}_{c_n})/n$. Furthermore, we also qualitatively study the attention map generated to highlight that the network indeed attends salient local regions.

Table 5: Test results for standard scan plane detection. Number of initial filters is denoted by the postfix "-n". Time taken for forward (Fwd) and backward (Bwd) passes were recorded in milliseconds.

Method	Accuracy	F1	Precision	Recall	Fwd/Bwd (ms)	#Param
Sononet-8	0.969	0.899	0.878	0.922	1.36/2.60	0.16M
AG-Sononet-8	0.976	0.921	0.911	**0.933**	1.86/3.46	0.18M
AG-Sononet-DS-8	0.975	0.918	0.907	0.929	1.92/3.51	0.18M
AG-Sononet-FT-8	**0.977**	**0.922**	**0.916**	0.929	1.92/3.47	0.18M
Sononet-16	0.977	0.923	0.916	0.931	1.45/3.92	0.65M
AG-Sononet-16	0.976	0.925	0.917	0.932	1.88/5.13	0.70M
AG-Sononet-DS-16	**0.978**	0.924	0.919	0.929	1.90/5.19	0.71M
AG-Sononet-FT-16	**0.978**	**0.929**	**0.924**	**0.934**	1.94/5.13	0.70M
Sononet-32	0.979	0.931	0.924	**0.938**	2.40/6.72	2.58M
AG-Sononet-32	**0.980**	0.932	0.928	0.937	3.01/8.74	2.79M
AG-Sononet-DS-32	0.978	0.929	0.921	0.937	2.98/8.81	2.80M
AG-Sononet-FT-32	**0.980**	**0.933**	**0.931**	0.935	2.92/8.68	2.79M

Table 6: Class-wise performance for AG-Sononet-FT-8. In bracket shows the improvement over Sononet-8. Bold highlights the improvement more than 0.02.

	Precision	Recall	F1
Brain (Cb.)	0.988 (-0.002)	0.982 (-0.002)	0.985 (-0.002)
Brain (Tv.)	0.980 (0.003)	0.990 (0.002)	0.985 (0.003)
Profile	0.953 **(0.055)**	0.962 (0.009)	0.958 **(0.033)**
Lips	0.976 **(0.029)**	0.956 (-0.003)	0.966 (0.013)
Abdominal	0.963 (0.011)	0.961 (0.007)	0.962 (0.009)
Kidneys	0.863 **(0.054)**	0.902 (0.003)	0.882 **(0.030)**
Femur	0.975 (0.019)	0.976 (-0.005)	0.975 (0.007)
Spine (Cor.)	0.935 **(0.049)**	0.979 (0.000)	0.957 **(0.026)**
Spine (Sag.)	0.936 **(0.055)**	0.979 (-0.012)	0.957 **(0.024)**
4CH	0.943 **(0.035)**	0.970 (0.007)	0.956 **(0.022)**
3VV	0.694 **(0.050)**	0.722 (-0.014)	0.708 **(0.021)**
RVOT	0.691 **(0.029)**	0.705 **(0.044)**	0.698 **(0.036)**
LVOT	0.925 **(0.022)**	0.933 **(0.027)**	0.929 **(0.024)**
Background	0.995 (-0.001)	0.992 (0.007)	0.993 (0.003)

Table 5 summarises the performance of the models. In general, AG-Sononets improve the results over Sononet at all capacity levels. In particular, AG-Sononets achieve higher precision. AG-Sononets reduces false positive examples because the gating mechanism suppresses background noise and forces the network to make the prediction based on class-specific features. As the capacity of Sononet is increased, the gap between the methods are tightened, but we note that the performance of AG-Sononets is also close to the one of Sononet with double the capacity. In addition, the advantage of AG-Sononets is that it can provide attention maps for no extra computational cost (see Appendix). Therefore, attention-mechanism allows the network to allocate all resources on the most salient aspect of the problem, and can achieve higher performance with minimal number of parameters. In Table 6, we show the class-wise F1, precision and recall values for AG-Sononet-FT-8. The improvement over Sononet is indicated in brackets. Baumgartner et al. (2016) noted that the model often confuses between cardiac views as they appear anatomically similar. The situation is notably improved, with statistically significant improvement for 4CH and 3VV ($p < 0.05$) due to fine-scale aggregating differences. However, these views remained challenging. We see that the precision increased by around 5% for kidney, profie and spines, as well as on average 3% for cardiac views. In some cases, we see minor reduction in recall rates. We believe that this is because the network may have become slightly more conservative when predicting the class labels.

3.5. Attention Map Analysis

The attention coefficients of the proposed U-Net model, which are obtained from 3D-CT test images, are visualised with respect to training epochs (see Figure 6). We commonly observe that AGs initially have a uniform distribution and pass features at all spatial locations. This is gradually updated and localised towards the targeted organ boundaries. Additionally, at coarser scales AGs provide a rough outline of organs which are gradually refined at finer resolutions. Moreover, by training multiple AGs at each image scale, we observe that each

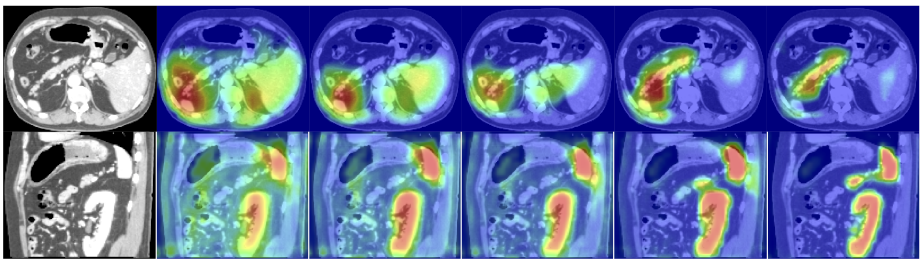

Figure 6: The figure shows the attention coefficients ($\alpha^{l_{s2}}$, $\alpha^{l_{s3}}$) across different training epochs (3, 6, 10, 60, 150). The images are extracted from sagittal and axial planes of a 3D abdominal CT scan from the testing dataset. The model gradually learns to focus on the pancreas, kidney, and spleen.

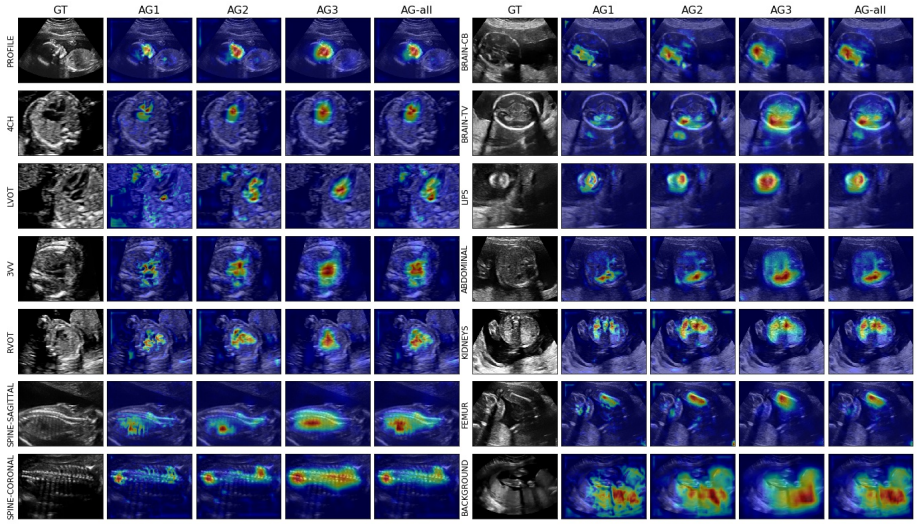

Figure 7: Examples of obtained attention map from AG-sononet. AG1 and AG2 are from layer 11 and 14 respectively. AG3 is obtained using CAM (Zhou et al., 2016). AG-all is obtained by normalising the maximum attented value across all AG's and taking mean over them.

AG learns to focus on a particular subset of organs.

3.5.1. Object Localisation using Attention Maps

In Figure 7, we show the attention map of AG-Sononet. AG-1 and AG-2 are the attention map applied at layer 11 and 14 respectively. AG-3 is the activation map of the final layer (the coarsest), with $F_g \in \{64, 128, 256\}$ channels depending

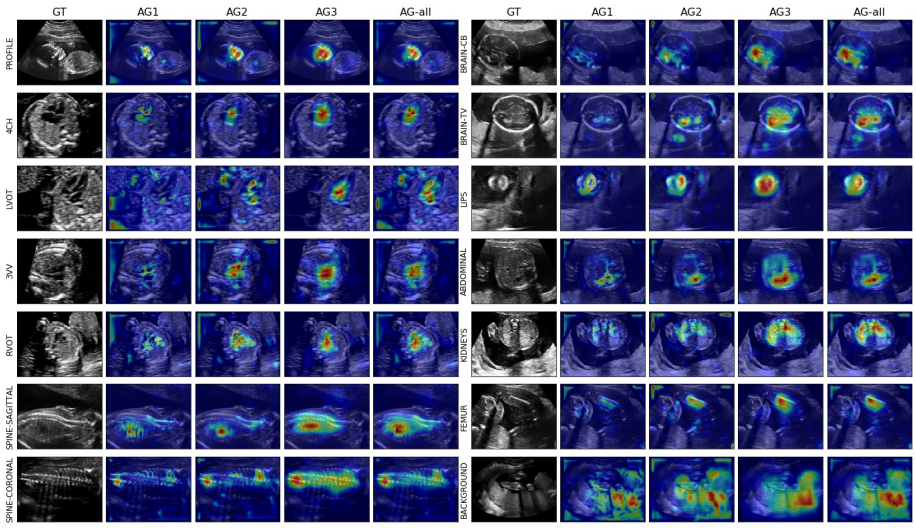

Figure 8: Examples of obtained attention map from AG-sononet-FT. AG1 and AG2 are from layer 11 and 14 respectively. AG3 is obtained using CAM (Zhou et al., 2016). AG-all is obtained by normalising the maximum attented value across all AG's and taking mean over them.

on the capacity. We employed Class Activation Mapping (CAM) (Zhou et al., 2016) to visualise class-specific attention. AG-all is obtained by taking the mean of the attention maps which are all normalised to have the maximum value 1. Recall that AG-Sononet simply obtains mean of the predictions at each image scale. As such, the attention maps pinpoint the class-specific information at all scales. In Figure 8, we show the attention map of AG-Sononet-FT. In this case, the aggregation layer relearns how to optimally combine the features at different scales. In some cases, fine-scale features seem to not learn anything if the prediction can be done by coarser scales.

Finally, in Figure 9, we show the attention maps of AG-Sononet-FT across different subjects, together the bounding box annotation generated using the attention maps (see Appendix for the heuristics). We see that the network consistently focuses on the object of interest, which indicates that the network indeed learnt the most important feature for each class. We note, however, attention map outlines the discriminant region; in particular, it does not nec-

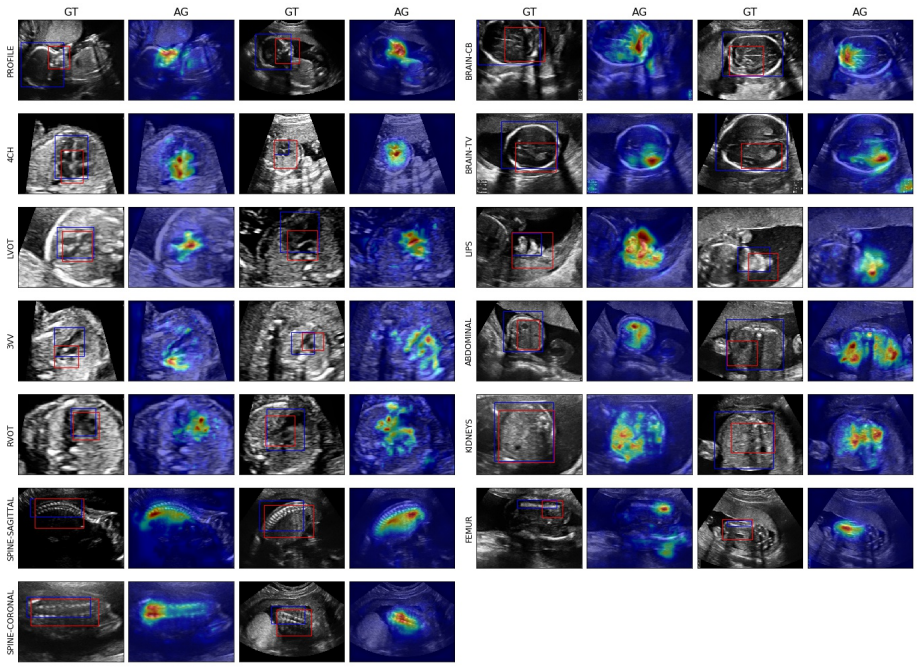

Figure 9: Examples of the obtained attention map and genereated bounding boxes (red) from AG-Sononet-FT across different subjects. The ground truth annotation is shown in blue. The detected region highly agrees with the object of interest.

essarily coincide with the entire object. This behaviour makes sense because some part of object will appear in background label (i.e. when the ideal plane is not reached). Qualitatively, however, the bounding boxes well agree with the annotated ground truth. Most crucially, the attention map is obtained for almost no additional computational cost; In comparison, (Baumgartner et al., 2016) requires guided backpropagation for localisation, which limits the localisation speed. This highlights the advantage of attention model for the real-time applications.

4. Discussion

In this work, we considered soft-attention mechanism and discussed how to incorporate this idea into segmentation and scan plane detection frameworks

to better exploit local structures in CT abdominal and fetal ultrasound images. In particular, we highlighted several aspects: gridded attention mechanisms, a normalisation strategy for the attention map, and aggregation strategies. We empirically observed and reported that using soft-max as the activation function tends to generate a map that sparsely activated and is overly sensitive to local intensity changes. The latter is problematic as in ultrasound imaging, image quality is often low. In the classification setting, We found that dividing the activations by the sum of the activations helped generate attention map with larger contextual support. As demonstrated in the segmentation framework, Sigmoid function is a good alternative as it only normalises the range and allows more information to flow. However, we found that training is non-trivial due to the gradient saturation problem.

We noted that training the attention-mechanism was slightly more complex than the standard network architecture. In particular, we observed that the strategy employed to aggregate the attention maps at different scales affects both the learning of the attention mechanism itself and hence the performance. Having a loss term defined at each scale ensures that the network learns to attend at each scale. We observed that first training the network at each scale separately, followed by fine-tuning was the most stable approach to get the optimal performance.

The proposed AG-Sononet architecture resembles the one of deep-supervision in the sense that we add modules before the final layer which helps back-propagating the gradient at the early layers of the network. However, we argue that without a proper gating mechanism, we will not see any improvement. In fact, we saw that the model trained with deep supervision did not necessarily obtain the best result. Therefore, while the network certainly benefits from backpropagating through additional pathways, the improvement in performance only came in conjunction with the attention mechanism.

There is a vast body of literature in machine learning exploring different gating architectures. For example, highway networks (Greff et al., 2016) make use of residual connections around the gate block to allow better gradient back-

propagation and slightly softer attention mechanisms. Although our segmentation experiments with residual connections have not provided any significant performance improvement, future work will focus on this aspect to obtain a better training behaviour.

Lastly, we note that the presented quantitative comparisons between the Attention 3D-Unet and state-of-the-art 2D cascaded models might not be sufficient enough to draw a final conclusion, as the proposed approach takes advantage of rich contextual information in all spatial dimensions. On the other hand, the 2D models utilise the high resolution information present in axial CT planes without any downsampling. We think that with the advent of improved GPU computation power and memory, larger capacity 3D-CT segmentation models can be trained with larger image grids without the need for image downsampling. In this regard, future research will focus more and more on deploying 3D models, and the performance of Attention U-Net can be further enhanced by utilising fine resolution input batches without any additional heuristics.

5. Conclusion

In this work we proposed a novel and modular attention gate model that can be easily incorporated into existing segmentation and classification architectures. Our approach can eliminate the necessity of applying an external object localisation model by implicitly learning to highlight salient regions in input images. Moreover, in a classification setting, AGs leverage the salient information to perform task adaptive feature pooling operation.

We applied the proposed attention model to standard scan plane detection during fetal ultrasound screening and showed that it improves overall results, especially precision, with much less parameters. This was done by generating the gating signal to pinpoint local as well as global information that is useful for the classification. Similarly, experimental results on CT segmentation task demonstrate that the proposed AGs are highly beneficial for tissue/organ identification and localisation. This is particularly true for variable small size organs

such as the pancreas, and similar behaviour is observed in image classification tasks.

Additionally, AGs allow one to generate fine-grained attention map that can be exploited for object localisation. We envisage that the proposed soft-attention module will also have great impact for explainable deep learning, which is a vital research area for medical imaging analysis.

References

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:170707998 2017;.

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473 2014;.

Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, et al. Human-level cmr image analysis with deep fully convolutional networks. arXiv preprint arXiv:171009289 2017;.

Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, Kainz B, Rueckert D. Real-time detection and localisation of fetal standard scan planes in 2d freehand ultrasound. arXiv preprint arXiv:161205601 2016;.

Cai J, Lu L, Xie Y, Xing F, Yang L. Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function. In: MICCAI. 2017. .

Cerrolaza JJ, Summers RM, Linguraru MG. Soft multi-organ shape models via generalized PCA: A general framework. In: MICCAI. Springer; 2016. p. 219–28.

Chen H, Dou Q, Ni D, Cheng JZ, Qin J, Li S, Heng PA. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent

neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015. p. 507–14.

Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications. Springer; 2016. p. 179–87.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115.

Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson BR, Pereira SP, Clarkson MJ, Barratt DC. Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal ct with dense dilated networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017. p. 728–36.

Greff K, Srivastava RK, Schmidhuber J. Highway and residual networks learn unrolled iterative estimation. arXiv preprint arXiv:161207771 2016;.

Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:180109927 2018;.

Heinrich MP, Blendowski M, Oktay O. TernaryNet: Faster deep model inference without GPUs for medical 3D segmentation using sparse and binary convolutions. arXiv preprint arXiv:180109449 2018;.

Heinrich MP, Oktay O. BRIEFnet: Deep pancreas segmentation using binary sparse convolutions. In: MICCAI. Springer; 2017. p. 329–37.

Hu J, Shen L, Sun G. Squeeze-and-excitation networks. arXiv preprint arXiv:170901507 2017;.

Jetley S, Lord NA, Lee N, Torr P. Learn to pay attention. In: International Conference on Learning Representations. 2018. URL: `https://openreview.net/forum?id=HyzbhfWRW`.

Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, Rajchl M, Lee M, Kainz B, Rueckert D, Glocker B. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham; 2018. p. 450–62.

Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis 2017;36:61–78.

Kawahara J, Hamarneh G. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2016. p. 164–71.

Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. arXiv preprint arXiv:180105173 2018;.

Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: Artificial Intelligence and Statistics. 2015. p. 562–70.

Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network. arXiv preprint arXiv:171108324 2017;.

Litjens GJS, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. CoRR 2017;abs/1702.05747. URL: `http://arxiv.org/abs/1702.05747`. `arXiv:1702.05747`.

Liu J, Wang G, Hu P, Duan LY, Kot AC. Global context-aware attention lstm networks for 3d action recognition. In: CVPR. 2017. .

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431–40.

Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. CoRR 2016;abs/1612.01887. URL: http://arxiv.org/abs/1612.01887. arXiv:1612.01887.

Luong M, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. CoRR 2015;abs/1508.04025. URL: http://arxiv.org/abs/1508.04025. arXiv:1508.04025.

Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. npj Digital Medicine 2018;1(1):6.

Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. IEEE; 2016. p. 565–71.

Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: Advances in neural information processing systems. 2014. p. 2204–12.

Nam H, Ha J, Kim J. Dual attention networks for multimodal reasoning and matching. CoRR 2016;abs/1611.00471. URL: http://arxiv.org/abs/1611.00471. arXiv:1611.00471.

NHS Screening Programmes . Fetal Anomaly Screen Programme Handbook. NHS, 2015.

Oda M, Shimizu N, Roth HR, Karasawa K, Kitasaka T, Misawa K, Fujiwara M, Rueckert D, Mori K. 3D FCN feature driven regression forest-based pancreas localization and segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2017. p. 222–30.

Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch 2017;.

Payer C, Štern D, Bischof H, Urschler M. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: STACOM. Springer; 2017. p. 190–8.

Pei W, Baltrusaitis T, Tax DMJ, Morency L. Temporal attention-gated model for robust sequence classification. CoRR 2016;abs/1612.00385. URL: http://arxiv.org/abs/1612.00385. arXiv:1612.00385.

Pesce E, Ypsilantis PP, Withey S, Bakewell R, Goh V, Montana G. Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv preprint arXiv:171200996 2017;.

Ren M, Zemel RS. End-to-end instance segmentation and counting with recurrent attention. CoRR 2016;abs/1605.09410. URL: http://arxiv.org/abs/1605.09410. arXiv:1605.09410.

Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.

Roth H, Farag A, Turkbey EB, Lu L, Liu J, Summers RM. Data from pancreas-ct. the cancer imaging archive. 2016. URL: http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU.

Roth HR, Lu L, Lay N, Harrison AP, Farag A, Sohn A, Summers RM. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical Image Analysis 2018;45:94 – 107.

Roth HR, Oda H, Hayashi Y, Oda M, Shimizu N, Fujiwara M, Misawa K, Mori K. Hierarchical 3d fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:170406382 2017;.

Saito A, Nawano S, Shimizu A. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. Medical image analysis 2016;28:46–65.

Sarraf S, DeSouza DD, Anderson J, Tofighi G. Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri. bioRxiv 2017;URL: https://www.biorxiv.org/content/early/2017/01/14/070441. doi:10.1101/070441. arXiv:https://www.biorxiv.org/content/early/2017/01/14/070441.full.pdf.

Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C. Disan: Directional self-attention network for rnn/cnn-free language understanding. arXiv preprint arXiv:170904696 2017;.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014;.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. p. 6000–10.

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903 2017;.

Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X. Residual attention network for image classification. arXiv preprint arXiv:170406904 2017a;.

Wang X, Girshick R, Gupta A, He K. Non-local neural networks. arXiv preprint arXiv:171107971 2017b;.

Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. CoRR 2018;abs/1801.04334. URL: http://arxiv.org/abs/1801.04334. arXiv:1801.04334.

Wolz R, Chu C, Misawa K, Fujiwara M, Mori K, Rueckert D. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE transactions on medical imaging 2013;32(9):1723–30.

Xie S, Tu Z. Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1395–403.

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. 2015. p. 2048–57.

Yang Z, He X, Gao J, Deng L, Smola AJ. Stacked attention networks for image question answering. CoRR 2015;abs/1511.02274. URL: `http://arxiv.org/abs/1511.02274`. `arXiv:1511.02274`.

Yaqub M, Kelly B, Papageorghiou AT, Noble JA. Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015. p. 687–94.

Ypsilantis PP, Montana G. Learning what to look in chest X-rays with a recurrent visual attention model. arXiv preprint arXiv:170106452 2017;.

Yu Q, Xie L, Wang Y, Zhou Y, Fishman EK, Yuille AL. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. arXiv preprint arXiv:170904518 2017;.

Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz C. Deep learning in neuroradiology. American Journal of Neuroradiology 2018;.

Zhang Z, Chen P, Sapkota M, Yang L. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017a. p. 320–8.

Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: A semantically and visually interpretable medical image diagnosis network. CoRR 2017b;abs/1707.02485. URL: http://arxiv.org/abs/1707.02485. arXiv:1707.02485.

Zhao B, Feng J, Wu X, Yan S. A survey on deep learning-based fine-grained object classification and semantic segmentation. International Journal of Automation and Computing 2017;14(2):119–35.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE; 2016. p. 2921–9.

Zhou Y, Xie L, Shen W, Wang Y, Fishman EK, Yuille AL. A fixed-point model for pancreas segmentation in abdominal ct scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017. p. 693–701.

Zhu W, Liu C, Fan W, Xie X. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. CoRR 2018;abs/1801.09555. URL: http://arxiv.org/abs/1801.09555. arXiv:1801.09555.

Zografos V, Valentinitsch A, Rempfler M, Tombari F, Menze B. Hierarchical multi-organ segmentation without registration in 3D abdominal CT images. In: International MICCAI Workshop on Medical Computer Vision. Springer; 2015. p. 37–46.

Appendix - Weakly Supervised Object Localisation (WSL)

In (Baumgartner et al., 2016), WSL was performed by exploiting the pixel-level saliency map obtained by guilded-backpropagation, followed by ad-hoc procedure to extract bounding boxes. The same heuristics can be applied for the given network, however, owing to the attention map, we can device a much efficient way of performing object localisation. In particular, we generate object location by simply: (1) blur the attention maps, (2) threshold the low activations, (3) perform connected-component analysis, (4) select a component that overlaps at each scale and (5) apply bounding box around the selected components. In this heuristics, backpropagation is not required so it can be executed efficiently. We note, however, attention map outlines salient region used by the network to perform classification; in particular, it does not necessarily agree with the object of interest. This behaviour makes sense because some part of object will appear both in the class as well as background frame until the ideal plane is reached. Therefore, the quantitative result is shown in 7, however, the result is biased. We however define new metric called *Relative Correctness*, which is defined as 50% of maximum achievable IOU (due to bias). We see that in this metric, the method achieves very high results, indicating that it can detect relevant features of the object of interest in its proximity.

Acknowledgements

We thank the volunteers, radiographers and experts for providing manually annotated datasets, Wellcome Trust IEH Award [102431], NVIDIA for their GPU donations, and Intel.

Table 7: WSL performance for the proposed strategy with AG-Sononet-FT-16. Correctness (Cor.) is defined as $IOU > 0.5$. Relative Correctness (Rel.) is defined as $IOU > 0.5 \times \max(IOU_{class})$.

	IOU Mean (Std)	Cor. (%)	Rel. (%)
Brain (Cb.)	0.69 (0.11)	0.96	0.96
Brain (Tv.)	0.68 (0.12)	0.96	0.96
Profile	0.31 (0.08)	0.00	0.80
Lips	0.42 (0.18)	0.36	0.60
Abdominal	0.71 (0.10)	0.96	0.96
Kidneys	0.73 (0.13)	0.92	0.98
Femur	0.31 (0.11)	0.02	0.58
Spine (Cor.)	0.53 (0.13)	0.56	0.76
Spine (Sag.)	0.53 (0.11)	0.54	0.94
4CH	0.61 (0.14)	0.76	0.86
3VV	0.42 (0.14)	0.34	0.62
RVOT	0.56 (0.15)	0.70	0.76
LVOT	0.54 (0.15)	0.62	0.80