A Novel Weakly Supervised Semantic Segmentation Ensemble Framework for Medical Imaging

Erik Ostrowski
Institute of Computer Engineering
Technische Universität Wien (TU Wien)
Vienna, Austria
erik.ostrowski@tuwien.ac.at

Bharath Srinivas Prabakaran
Institute of Computer Engineering
Technische Universität Wien (TU Wien)
Vienna, Austria
bharath.prabakaran@tuwien.ac.at

Muhammad Shafique
eBrain Lab, Division of Engineering
New York University Abu Dhabi (NYUAD)
Abu Dhabi, United Arab Emirates (UAE)
muhammad.shafique@nyu.edu

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650217

***Abstract*—The use of deep learning networks for vision based computer aided diagnostics (CAD) offers a tremendous opportunity for medical practitioners. However, state-of-the-art vision-based CAD systems rely on huge pixel-wise annotated datasets. Such datasets are rarely available, thus severely limiting the applicability of vision-based CAD systems. Hence, semantic segmentation with image labels offers a viable alternative. Semantic segmentation with image labels is well studied in a general context but seldom applied in the medical sector. The major challenge in applying semantic segmentation with image labels in the medical sector is that predicting on medical datasets is more complex than in the general context. Thus, directly applying methods for semantic segmentation with image labels like class activation maps (CAMs) on medical data generates insufficient results. However, state-of-the-art approaches rely on CAMs as a foundation. To address this problem, we propose a framework to extract useful information from particular low-quality segmentation masks. We achieve this by using our observations that the low-quality predictions have very low false negative detections, and multiple low-quality predictions show high variance among each other. We evaluated our framework on the popular multi-modal BRATS and prostate DECATHLON segmentation challenge datasets to demonstrate an improved dice score of up to 8% on BRATS and 6% on DECATHLON datasets compared to the previous state-of-the-art.**

***Index Terms*—Semantic Segmentation, CAMs, Deep Learning, Weakly Supervised Semantic Segmentation, Deep Learning, Machine Learning, Medical Imaging, GradCAM, Deep Neural Networks, DNN**

I. Introduction

Semantic segmentation describes classifying objects in a given image and returning their pixel-wise locations. Semantic segmentation offers, in contrast to simple classification, a particular region of detection within the image, therefore offering a much higher degree of explainability of the network's decision. Hence, their use in medicine would help a doctor or clinician conduct further investigations based on the assistant's decision.

State-of-the-art models for semantic segmentation are achieving new top performances at incredible speed, enabling this approach in the field of computer-aided diagnostics (CAD) and other related to medicine. For example, there are a variety of different applications, like colon crypt segmentation [1], brain tumor detection [2], detection of gastric cancer [3], discovering and tracking medical devices in surgery [4], etc. However, one significant factor limiting the quality of semantic segmentation results in medicine is that state-of-the-art methods are fully supervised. Fully supervised means that to achieve high-quality prediction results, the models require vast training data with pixel-wise annotated ground truth masks. Creating pixel-wise annotations takes over 90 minutes per image in Cityscapes [5]. For training a state-of-the-art fully-supervised semantic segmentation model, we require hundreds, if not thousands of annotations. In the clinical setting, this is, for the most part, not realistic as the personnel required for such work are too expensive and barely available, as such performing semantic segmentation is getting exponentially more difficult. For example, only trained experts can make an accurate diagnosis and pixel-wise annotation of a tumor in a brain MRI, and they are more expensive and less readily available. On the other hand, labeling images requires only context and a duration of less than 20 seconds. Although this time is unrealistic in the clinical setting, it offers a much more manageable environment for the real-world deployment of such systems. Hence, other solutions are needed if we want to leverage the power of semantic segmentation networks in the medical sector on a broader level. One such approach is to create semantic segmentation with only weaker annotations like bounding boxes, point annotations, or image labels. We focus only on semantic segmentation using image label annotations in this work, because image label annotations are the fastest and easiest to acquire, compared to any other weaker form of supervision.

Therefore, enabling the deployment of semantic segmentation with image labels in the medical setting would unlock a new level of applicability, as the availability of pixel-wise annotated open-source datasets is very limited in the medical sector and incredibly expensive to generate. Especially given that the annotation with image labels takes a fraction of the time required for pixel-wise labeling and is therefore also much more cost-efficient. Moreover, compared to simple classification networks semantic segmentation with image labels would provide a much higher degree of explainability since they highlight the area in the image that is responsible for the classification instead of just generating output labels.

Class Activation Maps (CAMs) are the backbone of almost

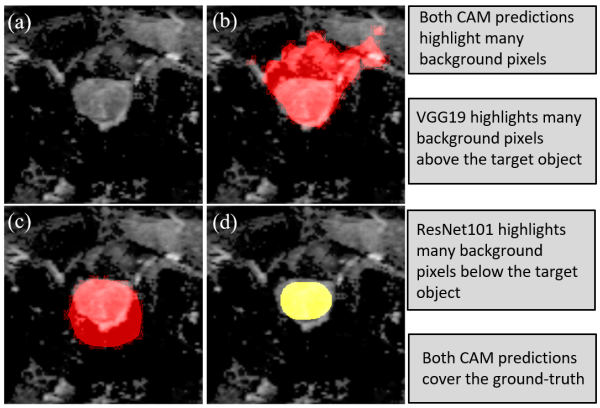

Fig. 1. Example of insufficient base CAM predictions: (a) Input example from Decathlon, (b) prediction of a VGG19 network, (c) prediction of a ResNet101 network, (d) ground truth

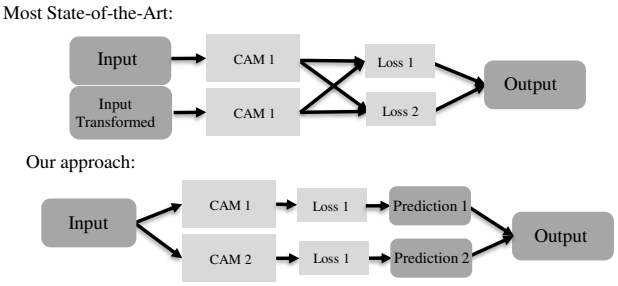

Fig. 2. A comparative overview of the workflow of state-of-the-art approaches and our framework.

all relevant research when it comes to methods for semantic segmentation with image labels. Hence, in this paper, we will focus on CAM-based approaches. Although CAMs have been shown to be reliable for the localization of target objects, their masks lack detail. For example, with a brain tumor, they will highlight the center of the tumor but become more uncertain towards the edges, either including too many background pixels or not covering the whole tumor. Therefore, in the state-of-the-art, most works focus on either enhancing the original CAM approach [6]–[9] or on the refinement of CAMs [10], [11]. However, when it comes to methods for semantic segmentation with image labels in the medical sector, there is very limited literature focusing on applications, and state-of-the-art in general context is not applicable in the medical context without adaptations. The major limitation of semantic segmentation with image labels in the clinical setting is the low level of prediction quality in clinical settings, as shown in Section IV. The classification networks fail to extract relevant information from these datasets due to them being often too small and too complex. Therefore, CAMs, which are built on classifiers, cannot make sufficient base predictions. The state-of-the-art methods do not salvage those shortcomings because they either rely on an already working classifier or on more confident localization, as Fig. 1 highlights.

To address these limitations, we present our approach for the application of semantic segmentation with image labels in the clinical setting. Our framework proposes to improve CAM predictions when the base results are too low quality for conventional methods. Based on our initial experiments, we have made the following observations: First, CAM predictions with a low enough threshold cover the target object with very high confidence. Second, when examining the output of different CAMs, be it different base models or altogether different methods, we noticed that they vary, especially when applying low thresholds. Therefore, we propose a framework that combines the predictions of multiple Grad-CAMs [12] built using different base models by determining the most optimal ensemble of thresholds in the training set. Moreover, we use additional use-case dependent box filters to further improve the prediction quality. Our approach is different from the state of the art in the following ways: Many state-of-the-art methods propose additional regularizations to the classification loss or after-the-fact refinement methods to boost the prediction quality of semantic segmentation with image labels. Fig.2 highlights the differences between our proposed framework and the state-of-the-art. The State-of-the-Art most commonly uses one network architecture, transformations of the input image, and convoluted loss functions, all trained dependent on each other. Whereas we propose to use distinct network architectures trained in our case with one simple loss, all independent of each other, the combination happens in the end. Contrasting that, our approach proposes to combine the predictions of multiple methods, which are on their own insufficient but offer as part of an ensemble valuable information. We conducted an exhaustive analysis of our framework on the popular BRATS 2020 [13]–[15] and DECATHLON [16] datasets and compared it with the relevant state-of-the-art approaches in this context. **Our main contributions can be summarized as follows:**

1) A novel semantic segmentation framework using just medical imaging labels by evaluating ensemble network configurations.
2) Evaluations on the BRATS and DECATHLON datasets have achieved performance improvements of 6% and 8%, respectively, over the current state-of-the-art.
3) Our framework is completely open-source and accessible online at https://github.com/ErikOstrowski/Automated_Ensemble

II. Related Works

In this section, we will highlight a few works that focus on computer vision tasks in the medical sector. Starting with Hirra et al. [17], who proposed a patch-based deep learning model for histopathology images. The unsupervised model is first pre-trained on patches of histopathology slides. Then the extracted features are used as input for the model to make a prediction on the whole slide. Guo et al. [18] use their causal knowledge fusion framework to improve 3D cross-modality cardiac image segmentation. They achieve this by first separating the anatomical factor

from the modality factor and then using a 3D hierarchical attention mechanism to extract the multi-scale information from a 3D cardiac image. In [19], Liu et al. propose a lightweight UNet architecture for Nasopharyngeal carcinoma tumor detection. The network utilizes lightweight modules to form the Compound Scaling Encoder and thus reduces the model's size to 3.55 M parameters. Furthermore, Calisto et al. [20] proposed an ensemble of 2D and 3D for predictions on volumetric data in the medical sector, to optimize both the model's performance and size. Zhang et al. [21] introduced their progressive perception learning framework to address the weaknesses of deep learning in the long-distance semantic relationship capture, the foreground and background interference adaptability, and the boundary detail information preservation when it comes to main coronary segmentation from the X-ray angiography images. Mohammadi et al. [22] proposed a ResU-Net, which uses long and short skip connections to improve the feature extraction for auto-contouring of organs at risk in high-dose-rate brachytherapy. Ke et al. [23] developed a self-constrained 3D DenseNet to detect and segment nasopharyngeal carcinoma automatically in magnetic resource images. In [24] the authors studied the use of deep learning-based autosegmentation for MR-based prostate radiotherapy planning. Messaoudi et al. [25] introduced a network to transfer the efficiency of a 2D classification network trained on natural images to 2D, 3D uni- and multi-modal medical image segmentation applications in a more efficient way. Their proposed network is based on weight transfer by embedding a 2D pre-trained encoder into a higher dimensional U-Net, and dimensional transfer by expanding a 2D segmentation network into a higher dimension one. In [26] CM-SegNet was introduced , a network for the segmentation of lesions in medical images. The model leverages multiscale input and encoding-decoding thoughts and is composed of multilayer perceptron and convolution modules to fully extract global and local image information for the segmentation task. In DCNet [27], the authors proposed a network for Nasopharyngeal Carcinoma segmentation in magnetic resonance imagery. They proposed a densely connected deep convolutional network consisting of an encoder network and a corresponding decoder network, which uses a Skip-connection architecture to propagate spatial information to the decoder network. Zhi et al. [28] introduced a Cross-modality based approach for Vessel contour detection in intravascular images. To overcome the typical label space inconsistency in cross-modality methods, the authors divided the label space into private label space and shared label space. Zhuang et al. [29] proposed to use a YOLOv3 object detector to generate sub-images of left ventricle endocardium images. A Markov random field (MRF) model then performs preliminary identification and binarization of the myocardium on those sub-images. The applications of semantic segmentation with image labels in the field of medicine are still very niche. Patel et al. [30] have tried to apply those methods specifically in a medical context. Their WSS-CMER approach uses a CAM-based approach that extends the ideas proposed by [11]. Furthermore, Guo et al. [31] used image labels to localize organs in 3D images. To adapt the CAM idea to the medical use case, they used a base classifier the U-Net [32] architecture, which proved to be effective in the medical context. Moreover, they train the contracting half of the network as a classifier and use the rescaled output of intermediate layers of the contracting half as pseudo-ground truth for the segmentation loss of the expanding half of the network. Chen et Al. [33] used an adapted CAM approach for segmentation with image labels on images in the medical context as well. Similar to [31], they use a U-Net architecture where the contracting part is trained as a classifier. Additionally, Chen et Al. added modules that try to improve foreground-background differentiation and deal with co-occurrences of objects.

We can summarize, the state-of-the-art approaches mostly tackle the classification loss of the CAM by adding regularizations that aim to guide the network to predict finer and more complete masks. For example, additional regulations for the loss function were used in [30] and [11], like matching the predictions of affine transformations to the original predictions. Alternatively, many approaches refine the CAM prediction after the fact, using methods like pixel-similarity, for example, as in [10]. Nevertheless, those approaches do not work as well in the medical sector. We observed that we were unable to generate reliable classifiers on our target datasets due to their small size and complexity.

Table I summarizes the key contribution of the related work and our proposed approach. We see that our approach is relatively different from the related work and offers the flexibility to incorporate relevant methods into our ensemble.

III. Our Proposed Methods

Fig. 3 presents an overview of our framework. First, we start by training a classifier model on the target dataset. In our case, the ResNet-34 and ResNet-50 [34] models proved to be the most successful, but the framework accommodates the use of any other classifier instances. Second, we use Grad-CAM to create the first masks for the different classifiers. In our case, the standard Grad-CAM worked the best. Next, we test ensemble methods to combine two or more prediction sets. For our final version, we choose the ensemble version that offers the best possible results, followed by the calibration step to determine the best ensemble of thresholds for the highest detection score. Our main contribution lies in our method's approach of combining multiple very low-quality predictions to automatically filter out the wrong predictions and keep a high-quality core prediction. To the best of our knowledge, this approach was not proposed for weakly supervised segmentation works before and offers a different angle for improved prediction quality. Our method is specifically suited for our medical use case as the usual methods for weakly supervised segmentation do not seem to work.

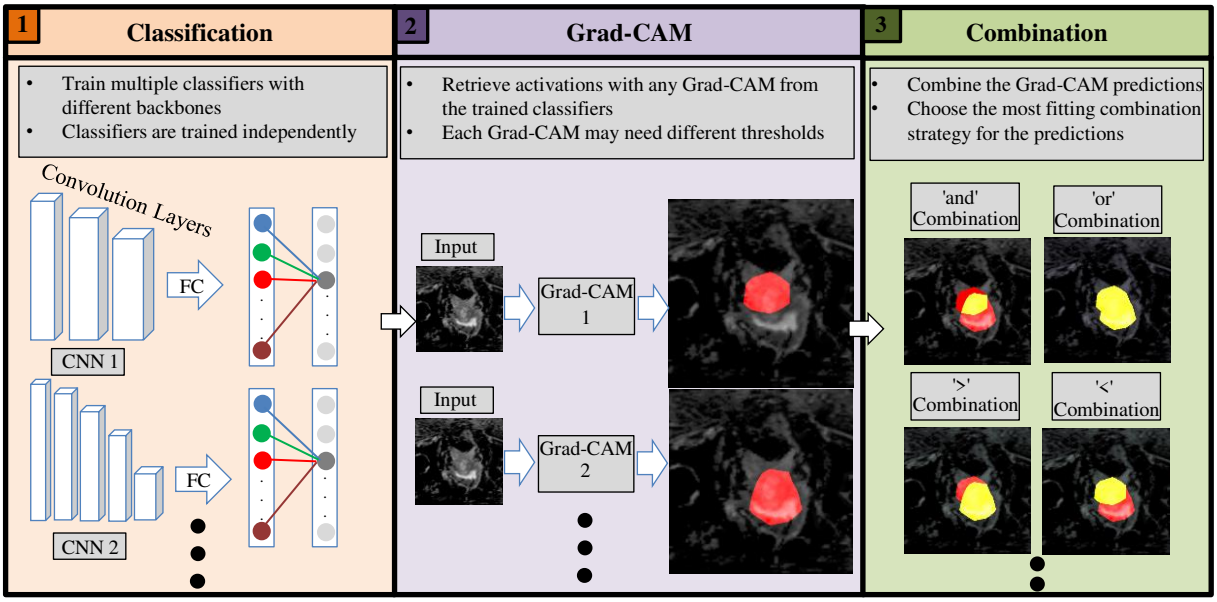

Fig. 3. Overview of our Framework. (1) Stages in the classifier training, which will be used in the Grad-CAM; (2) The Grad-CAM uses trained models and input images to generate CAMs; (3) The Ensemble will test out multiple methods of how to combine the CAMS;

TABLE I
KEY CONTRIBUTIONS OF THE RELATED WORK AND OUR APPROACH

Method	Key contribution
Hirra et al. [17]	Ensemble of patch predictions for global prediction
Guo et al. [18]	3D hierarchical attention for spatial learning
Lie et al. [19]	Lightweight U-Net
Calisto et al. [20]	2D-3D ensemble for efficieny
Zhang et al. [21]	Improved boundary detections
Mohammadi et al. [22]	Dense ResU-Net for auto-contouring of organs
Ke et al. [23]	3D Self-constrained Resnet for nasopharyngeal carcinoma segmentation
Zhong et al. [24]	Deep learning for nasopharyngeal carcinoma segmentation
Messaoudi et al. [25]	Transfer learning from general to medical context
Xing et al. [26]	Multi-Layer perception for the segmentation of lesions
Li et al. [27]	Dense U-Net for nasopharyngeal carcinoma segmentation
Zhi et al. [28]	Cross-modality based approach for Vessel contour detection
Zhuang et al. [29]	Markov Random Field and Yolov3 for segmentation with image labels
Patel et al. [30]	Equivariant Attention for segmentation with image labels on medical images
Guo et al. [31]	U-Net based CAM for segmentation with image labels on medical images
Chen et al. [33]	Causality module for co-occurrence detect. for segmentation with image labels on medical images
Our Approach	CAM ensemble reducing over predictions on small and complex datasets for segmentation with image labels on medical images

A. *Training and Exploration of Classifier Model Instances*

Our framework aims to create an ensemble of different CAM methods because their ensemble evens out their shortcomings and therefore results in more accurate predictions than their singular pieces. Following the observation that the best ensembles are generated from high-quality CAMs, and high-quality CAMs are generated from high-quality classifiers, we have to investigate classifiers for our target datasets. We limited ourselves to multiple ResNet classifiers in our experiments. Note that in our framework, any classifier can be used.

Instead of striving to test more complex networks, we took some inspiration from other methods like the self-supervised Swav [35]. Swav tries to learn to differentiate pictures without guidance instead of using annotations. For this purpose, Swav uses a contrastive loss function that compares pairs of images. The goal of the loss function is to push away images that are different in the feature space while pulling together those from transformations, or views, of the same image in the feature space. In particular, our interest in the Swav approach relies on two reasons: first, using a pre-trained unsupervised model for the medical datasets may improve the quality of the Grad-CAM results. This may be the case since unsupervised models are guided more toward distinguishing between shapes rather than just guessing the correct class. Second, many state-of-the-art approaches add additional regularization to the classification loss. Those regularizations, e.g., the affine transformation in SEAM, are very much in the same spirit as Swav's contrastive learning loss. Therefore, we assume that the activations of an unsupervised trained model may recognize the objects more completely. We have evaluated multiple trained Swav models but observed that their contrastive loss

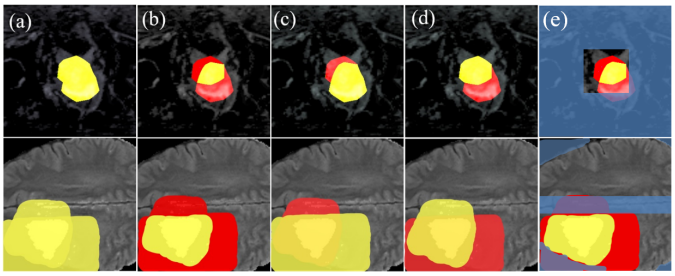

Fig. 4. Visualization of the ensemble methods on a BraTS and Decathlon example: (a) 'or', (b) 'and', (c) '<', (d) '>', (e) 'and' with boxes

approach did not result in higher-quality CAMs compared to the conventionally trained classifiers.

B. Evaluating Grad-CAMs for the trained models

After creating our candidate classifiers, we can focus on generating CAM predictions. For that purpose, we will apply Gradient-weighted Class Activation Maps (Grad-CAMs). (Grad-CAM) takes a network and image as input and returns a rough mask. In detail, Grad-CAM runs an input image through a model and takes the K outputs $A^k \in \mathbb{R}^{u \times v}$ of the final convolutional layer. Next, we calculate the gradient score for each class c, of the logits y^c for the feature map activations A^k, i.e., $\frac{\partial y^c}{\partial A^k}$. Then the gradients are global average pooled across each feature map to give us the importance score α_k^c:

$$\alpha_k^c = \frac{1}{uv} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$$

where k denotes the index of the Z activation maps, and i, j are the feature map coordinates. The importance score weights the relevance of each feature map k for each class c. Then, we multiply each activation map A^k by its importance score α_k^c for each class and take their sum to gain the prediction mask M^c:

$$M^c = ReLU(\sum_k \alpha_k^c A^k)$$

We apply ReLU to the summation to only consider the pixels that positively influence the score of the class of interest. The mask M^c highlights areas of the image that were activated for the respective class. Finally, we resize and rescale M^c so that the mask fits the original image dimensions.

However, it was shown that Grad-CAM fails to properly localize objects in an image if it contains multiple occurrences of the same class. Moreover, it was also found that due to not weighing the average of the partial derivatives, the localization often does not correspond to the entire object but only parts of it. Hence, Grad-CAM++ [36] was introduced, which solves those problems by using a more sophisticated formula for the importance score $\alpha_{i,j}^{c_k}$:

$$\alpha_{i,j}^{c_k} = \frac{\frac{\partial^2 y^c}{(\partial A_{i,j}^k)^2}}{2\frac{\partial^2 y^c}{(\partial A_{i,j}^k)^2} + \sum_a \sum_b A_{ab}^k (\frac{\partial^3 y^c}{(\partial A_{i,j}^k)^3})}$$

As a further optimization, SmoothGrad-CAM++ [37] was introduced. SmoothGrad-CAM++ combined Grad-CAM++ with SMOOTHGRAD [38], which improves the original method by sharpening the Grad-CAM output by taking random samples in a neighborhood of an input x and averaging the resulting outputs.

However, these improvements aim to increase the sharpness of the boundaries of the detected object and alleviate the issues of the original method with multiple objects of the same class within the same image. But, many of those problems do not occur in the observed medical datasets. All images in the BraTS and DECATHLON datasets never contain multiple instances of the target object. Furthermore, those target objects often have a round-ish shape, whose borders are ambiguous even for experts. Nevertheless, we also tested SmoothGrad-CAM++ as it is a more recent iteration of this approach. For CAM generation, we will run our trained candidate models and the images through the candidate Grad-CAM, creating masks for all images.

C. Ensemble methods

Once we have the collected masks of our candidate models and Grad-CAMs, we aim to combine them for higher-quality results.

We considered four simple and intuitive approaches for the ensemble methods, Fig. 4 presents a visualization of those methods. First, we have the 'or' ensemble, which sums up the predictions of candidates. This approach works best when both masks have high true positive rates, generating the biggest possible activation area between the combined masks. Second is the 'and' approach, which multiplies the predictions of candidates. The 'and' approach minimizes the possible detection area of both masks in contrast to the 'or' approach. This approach works best when both models have a high true negative rate. Next, the '<' and '>' approaches only take the masks with the least or most positive classified pixels, respectively. In this way, we address the problem of models tending to predict the size of the target object as too big or too small while the complete prediction of one candidate is still better than the 'and' or 'or' of all candidates. Moreover, we refined the predictions further using dataset-specific boxes, that filter out predictions that are very unlikely to be correct. In particular, for the DECATHLON dataset, we keep only a small box area around the center and any predictions outside of it are removed when using our box filter. We constructed the filter in this way because prostate MRI slides in this dataset are always centered and the target cancer object is almost always located in this region. However, we created a different box filter, for the BraTS dataset. Brain MRIs have the property that the brain does not fill out the complete image and we thus have a region in each slide where we know that there cannot be any cancer. Therefore we filter out any part of the detection that is not on brain pixels. Furthermore, we also made the observation that brain tumors are rarely located along the central horizontal line. Therefore, we removed those pixels from the prediction too. Since the Grad-CAM methods return the predictions in

TABLE II
COMPARISON OF ENSEMBLECAM WITH STATE-OF-THE-ART IMAGE LABEL BASED WSSS TECHNIQUES ON THE BRATS DATASET.

Method	Best AVG DSC	Best AVG mIoU
SEAM	56.1	39.0
WSS-CMER	59.7	42.6
Empty	64.6	47.7
ResNet-34-GradCAM	67.3	50.7
ResNet-50-GradCAM	68.5	52.1
ResNet-34x50 'or'	67.8	51.3
ResNet-34x50 '>'	68.3	51.9
ResNet-34x50 '<'	68.6	52.2
ResNet-34x50 'and'	**70.3**	**54.2**
ResNet-34-GradCAM + Box	70.2	54.1
ResNet-50-GradCAM + Box	71.9	56.1
ResNet-34x50 'or' + Box	68.3	51.8
ResNet-34x50 '>' + Box	68.8	52.4
ResNet-34x50 '<' + Box	71.8	56.0
ResNet-34x50 'and' + Box	**72.4**	**56.8**

logits ranging from 0 to 1, we can determine the threshold of the value at which we consider the pixel a positive or negative classification. This hyperparameter gives us a massive leeway about any given prediction's false positive and false negative rate. Using very high thresholds drastically reduces the area classified as positive, leading to a high false negative rate. And vice-versa, using very low thresholds drastically increases the area classified as positive, leading to a high false positive rate. The most optimal threshold value varies from candidate model to candidate model. Therefore, we decided to test the chosen ensemble approach with all combinations of thresholds from 0 to 1 in steps of 0.1. We conduct these tests on the training set to determine the thresholds we would use on the validation set. Our experiments have shown that the combination of thresholds most optimal for the training set is also one of the most optimal combinations for the validation set.

IV. RESULTS AND DISCUSSION

We compare the results of our framework to relevant prior literature, namely WSS-CMER, and SEAM, which was a baseline used in the WSS-CMER work. To the best of our knowledge, WSS-CMER is the only method in this domain specifically targeted for medical images. Moreover, we include the results of SEAM because it represents a current approach to semantic segmentation with image labels that is not optimized for data from the medical sector. The goal of our framework was to extend current applications of semantic segmentation with image labels to the medical sector. Primarily because we notice a massive demand for dealing with data scarcity in the medical sector, especially with respect to the availability of pixel-wise annotated medical images.

First, we discuss the experimental setup used for the experiments discussed in this section. We ran the experiments on a CentOS 7.9 Operating System executing on an Intel Core i7-8700 CPU with 16GB RAM and 2 Nvidia GeForce GTX 1080 Ti GPUs. We executed our scripts with the following software versions: CUDA 11.5, Pytorch 1.13.0, and torchvision 0.14.0. We tested our framework on the BraTS 2020 and DECATHLON datasets for evaluation and comparison.

To assess the performance of the proposed approach, we employ the common Dice Similarity coefficient (DSC) and mean Intersection over Union (mIoU):

$$DSC(GT, Pred) = \frac{2|GT \cdot Pred|}{|GT| + |Pred|}$$

Where GT and $Pred$ are binary matrices (with values of 1 for elements inside a group and 0 otherwise), GT signifies the ground truth, and $Pred$ signifies the classification result.
The mean Intersection-over-Union (mIoU):

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{p_{i,i}}{\sum_{j=1}^{N} p_{i,j} + \sum_{j=1}^{N} p_{j,i} - p_{i,i}}$$

where N is the total number of classes, $p_{i,i}$ the number of pixels classified as class i when labelled as class i. $p_{i,j}$ and $p_{j,i}$ are the number of pixels classified as class i that were labelled as class j and vice-versa, respectively.

Moreover, we trained and evaluated the datasets in a three-fold cross-validation scheme. For BraTS, we divided the randomly shuffled dataset into three equally sized batches and trained and tested three models, each trained on two of the three samples and evaluated on the remaining sample. For DECATHLON, we downloaded the three-fold cross-validation splits of WSSS-CMER. We reported the average results over the three runs. Note that we could not reproduce the results of WSS-CMER and SEAM. Therefore we used the numbers published in [30], when stating the results from WSS-CMER and SEAM. We ended up using mostly the ensemble of ResNet-34 and ResNet-50 as those two base models achieved the best results on their own. Adding more components or different models to the ensemble resulted in lower prediction quality.

The main challenge that we encountered when trying to apply state-of-the-art image level based semantic segmentation in the clinical setting is that in the medical datasets, the base classifier was unable to obtain an accurate classification. The datasets were too small and/or too complex for a classifier to learn the differences between the different classes. Additionally, in the clinical setting, we are dealing most often with co-occurrences; for example in the BraTS dataset, every image contains a brain, whereas the differences between a brain with and without cancer are too small to be detected by the classifier. Therefore, most state-of-the-art CAMs fail in this setting. Our main goal was to adapt the CAM pipeline in such a way that it could deal with insufficient classifiers.

BraTS: The BraTS 2020 dataset is a multi-modal brain tumor segmentation in Magnetic Resonance images. Three hundred sixty-nine multi-modal scans with their corresponding expert segmentation masks are available for the tumor segmentation task. For comparability with WSS-CMER and SEAM, we will exclude the additional 24 scans of BraTS 2020. These datasets use the same naming convention and direct filename mapping between them. The scans are composed of four modalities, which include T1, T1c, T2, and

T2 Fluid Attenuated Inversion Recovery (FLAIR). We use T1c, T2, and FLAIR for classifier training, and we only use FLAIR for mask generation with Grad-CAM. We achieved the best results in our experiments just using FLAIR. We separated each scan into its set of slides, resulting in 47,232 slides, cropped them to the size 128 x 128, and applied a threshold of 0.8 to the normalized frame, which generated the best results.

For this work, we followed the procedure of WSS-CMER, which only considered a binary segmentation class, i.e., healthy vs. non-healthy targets. Therefore, we merge the different tumor classes into one single 'positive' class. Our framework is not limited to binary scenarios. However, since the used datasets are pretty complex and the work on this domain in the medical sector is still in the early stages, we are constrained to this simplified scenario. Nevertheless, Grad-CAM predictions have been shown to be class-specific in natural images, as widely observed in the computer vision literature. But, in natural images, we do not find the case that certain classes are only present in combination with other classes, which enables the classifier to discriminate between those classes. In both of the used datasets, most classes only appear in combination with the other classes, so our classifier cannot distinguish between them. Since we observed that our classifiers are already struggling to distinguish between images with or without the presence of a tumor, we saw no point in adding different tumor classes. For instance, [39] reported that adding those classes results in worse predictions.

In table II, we compare the results of the BraTS dataset. We added the "Empty" mask, where we just predicted a background mask for every image. Note that [30] do not achieve an improvement over "Empty" masks, which stresses the necessity to improve predictions in this use case. When we tried to reproduce [30]'s results, we experienced that the classifiers had an accuracy of around 50% for both datasets, which is in line with all other classifiers that we tested. This shows that adding a more sophisticated loss to the training process, like in WSS-CMER or SEAM, does not lead to better CAMs if the baseline classifier cannot work with the respective dataset. Our baseline Grad-CAM with ResNet-34 is around 7% better than WSS-CMER. Moreover, Grad-CAM ResNet-50 is 1% better than its ResNet-34 counterpart. Using the bigger ResNet-101 did achieve worse results, as anticipated, due to its larger number of parameters. We did not observe a difference in CAM quality when using SmoothGrad-CAM++ for both datasets. Finally, by combining the two ResNets via the 'and' scheme, we generate our best result, which is 1.5% better than the previous high score. Adding the box filter to the baseline ResNets improved their mIoU score by over 3% for ResNet-34 and 4% for ResNet-50, already besting the results over the ensemble approaches without our box filter. Adding our box filter to the ensemble further improves their results, with the 'and' method achieving the best result with 56.8% mIoU.

DECATHLON: The DECATHLON segmentation challenge is a multi-modal prostate dataset with the task of localizing and detecting the prostate peripheral and transition zones. The dataset consists of 32 volumetric scans containing MRI-T2 and apparent diffusion coefficient (ADC) maps with their corresponding segmentation masks. We only used the ADC maps for classifier training and CAM generation, as they generated the best results. We followed a similar scheme to that of the BraTS dataset, where the prostate peripheral zone and the transition zone are combined into a single 'positive' class. We resized each of the 2D slices to 320 x 320 pixels.

In table III, we compare the results of the DECATHLON datasets. Our baseline Grad-CAM with ResNet-34 is around 7% better than WSS-CMER. However, Grad-CAM ResNet-50 is 2% worse than its ResNet-34 counterpart. The smaller ResNet-18 achieved worse results as it could not learn the relevant information. Finally, by combining the two ResNets via the 'and' scheme, we generate our best result, which is 1.4% better than the previous high score. Adding our box filter to the DECATHLON predictions further improves their mIoU score. ResNet-50 and ResNet-34 improve by almost 10% and 8%. The best result is again generated by the 'and' scheme with 72.8% mIoU, improving the baseline by 1.4%. Moreover, using our box filter has a bigger effect on the 'or' scheme than the '$<$' or '$>$' scheme, increasing its baseline by almost 9%. Note that SEAM and WSS-CMER showed a variation of over 13% between samples, while our Ensemble varied less than 3.5%. The difference between our proposed approach and the previous studies is that our focus relied on adapting the CAM pipeline to work with insufficient base classifiers. The previous methods focus mainly on improving the classifiers or just the CAM method, which is orthogonal to our approach and thereby would improve the benefits offered by our approach. The results presented by Patel [30] et al. illustrate the inability to improve the prediction accuracy beyond the reasonable baseline of 'Empty.' This serves as a motivation that just adding regulations to the CAM loss is not sufficient in the clinical setting and that an approach adapted to deal with such limitations is more desirable. Our approach can function with very low-quality CAM predictions. To our knowledge, our work is the first to address the problem of creating useful semantic segmentation predictions out of completely unreliable base Grad-CAM predictions, as recent works focus on improving or refining the base CAM. But we can see by the comparison of [30] and the 'Empty' prediction in Tables II and III that it is not always straightforward to adapt the CAM to the given use-case. Moreover, refinement is not an option, as the base CAMs have few clues about the target objects.

Fig. 5 presents an overview of our experimental results. The first two rows show examples from the DECATHLON dataset, and the second two are from the BraTS dataset. We notice that column (g) with the 'and' + boxes combination achieves the most significant overlap with the ground truth column (b) in terms of visual and metric-based evaluations, as shown in Section IV. In contrast, the 'or' ensemble used in the (e) column creates the results with the most minor overlap when considering the false positives. This shows our observation that motivated the framework. Both ResNet-34 and ResNet-50

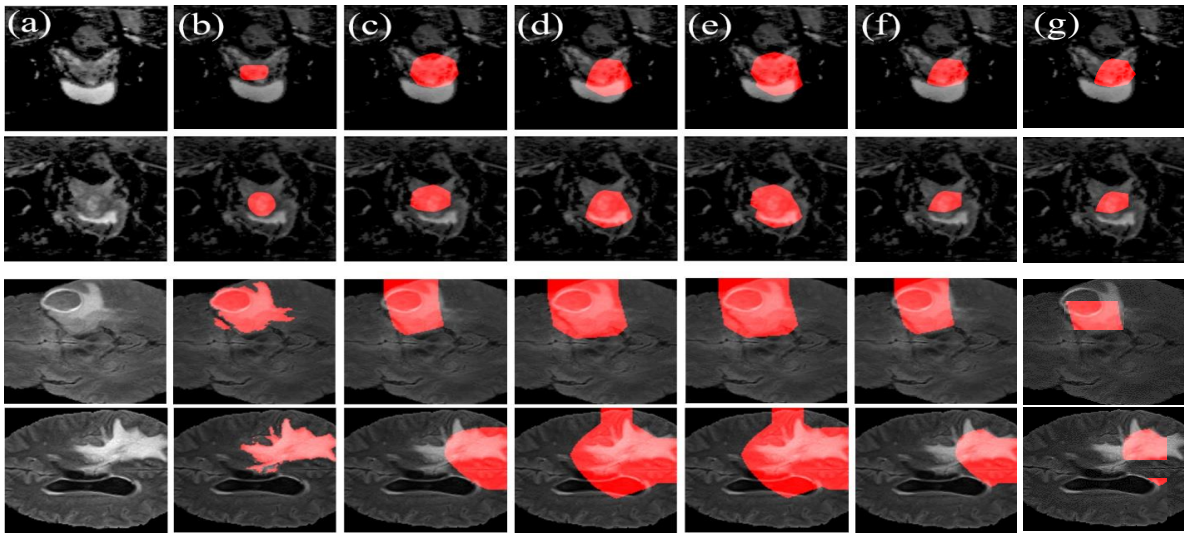

Fig. 5. Example results of our Framework on BraTS and DECATHLON: (a) Source image, (b) ground truth, (c) ResNet34 Grad-CAM, (d) ResNet50 Grad-CAM, (e) 'or' ensemble, (f) 'and', (g) 'and' + boxes ensemble

TABLE III
COMPARISON OF ENSEMBLECAM WITH STATE-OF-THE-ART IMAGE LABEL BASED WSSS TECHNIQUES ON THE DECATHLON DATASET.

Method	Best AVG DSC	Best AVG mIoU
VGG19-GradCAM	63.6	46.6
Empty	65.5	48.7
ResNet101-GradCAM	65.9	49.1
SEAM	65.9	49.1
ResNet-18-GradCAM	67.0	50.3
Swav-GradCAM	70.2	54.1
WSS-CMER	71.3	55.4
ResNet-50-GradCAM	76.1	61.4
ResNet-34-GradCAM	78.0	63.8
Ensemble ResNet-34x50 'or'	77.3	63.0
Ensemble ResNet-34x50 '<'	77.7	63.5
Ensemble ResNet-34x50 '>'	78.4	64.4
Ensemble ResNet-34xSwav 'and'	78.7	64.9
Ensemble ResNet-34x50 'and'	**79.3**	**65.7**
ResNet-50-GradCAM + Box	83.1	71.1
ResNet-34-GradCAM + Box	83.3	71.4
ResNet-34x50 '<' + Box	83.2	71.3
ResNet-34x50 '>' + Box	83.6	71.8
ResNet-34x50 'or' + Box	83.7	71.9
Ensemble ResNet-34xSwav 'and'	83.8	72.1
ResNet-34x50 'and' + Box	**84.3**	**72.8**

cover a lot of ground truth and too many background pixels. However, we notice that the activated background pixels differ between the two models, resulting in the best predictions when using the 'and' + boxes ensemble and the worst when using the 'or' ensemble without boxes.

Fig. 6 presents an overview of the mIoU achieved when different thresholds are used for the two models under consideration. These values are obtained using the training dataset results averaged over all three cross-validation iterations. The left graphic shows the results of BraTS, and the right shows the DECATHLON results. We notice that the left graphic is lighter in color than the right. The reason is that overall, the mIoU results on BraTS are lower than on DECATHLON. Both datasets reach their best results at thresholds 70% for both models. Both graphics show results achieved with the 'and' ensemble; the optimal thresholds appear to vary based on the ensemble method under consideration.

A. Analyzing training and validation loss

Fig. 7 and Fig. 8 show the loss during training on both validation and training sets. The dashed lines show the loss on the training set, and the continuous lines show the loss on the validation set. In Fig. 7, we observe that both models converge to the same loss in the training set, with a slightly lower loss for the ResNet-34 on the validation set, which reflects the marginally better performance of Grad-CAM with ResNet-34 on this dataset. Furthermore, we notice that the validation loss is less volatile than the training loss and saturates already after two epochs. In Fig. 8, we observe that both models converge to a similar loss on the validation set, with the ResNet-34 being faster to saturate. We observe a more significant difference between the models' losses on the training set, but the Grad-CAM performance does not reflect this. As for this dataset, ResNet-50 achieves better results. This can be due to the bigger size and higher complexity of the BraTS dataset compared to DECATHLON. Since ResNet-50 is a deeper model, it can understand more complex data better than ResNet-34.

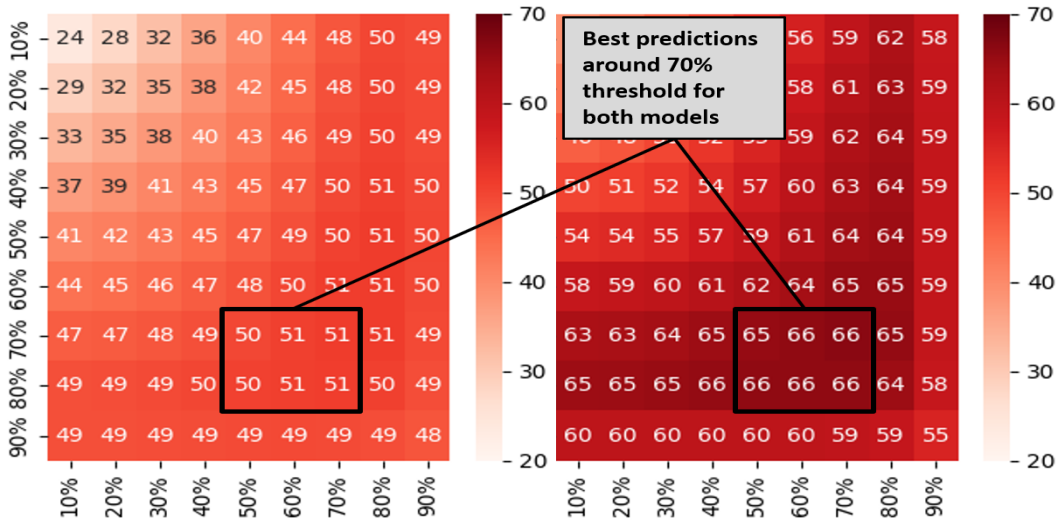

Fig. 6. All threshold combinations of ResNet-34 and ResNet-50 'and' combination on the train set. Y-axis describes the threshold used for ResNet-34, X-axis is the threshold used for ResNet-50. BraTS on the left, DECATHLON on the right.

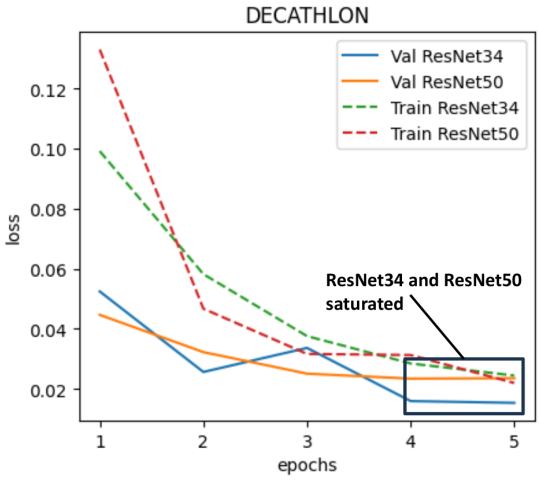

Fig. 7. The training set and validation set loss during training of ResNet-34 and ResNet-50 on DECATHLON.

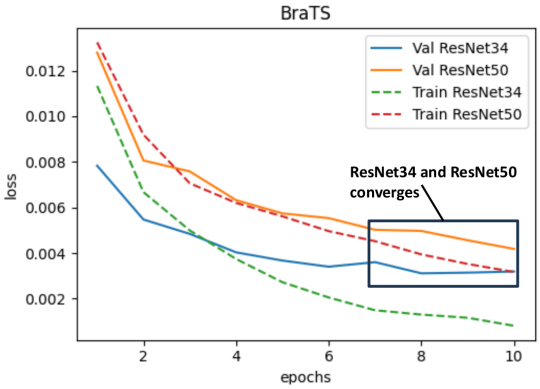

Fig. 8. The training set and validation set loss during training of ResNet-34 and ResNet-50 on BraTS.

V. Conclusion

In this paper, we have proposed our novel framework, which illustrates the approach for finding an ensemble of CAM methods for semantic segmentations with image labels for data from the medical sector to address the lack of research specified for this application. Our framework proposes a scheme to generate useful prediction masks despite the lack of quality prediction of base models due to the complexity and size of the used datasets. Therefore, this framework can also be applied in different contexts where the standard approaches do not generate high-quality results. We showed that the predictions generated by our framework achieve state-of-the-art performance on the BraTS 2020 and DECATHLON datasets, which proves its effectiveness compared to other approaches. Our framework is open-source and accessible at https://github.com/ErikOstrowski/Automated_Ensemble.

Acknowledgments

This work is part of the Moore4Medical project funded by the ECSEL Joint Undertaking under grant number H2020-ECSEL-2019-IA-876190. This work was also supported in parts by the NYUAD's Research Enhancement Fund (REF) Award on "eDLAuto: An Automated Framework for Energy-Efficient Embedded Deep Learning in Autonomous Systems", and by the NYUAD Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award CG010.

References

[1] A. Cohen, E. Rivlin, I. Shimshoni, and E. Sabo, "Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 150–164, 2015.

[2] H. Tek, M. Bergtholdt, D. Comaniciu, and J. Williams, "Segmentation of 3d medical structures using robust ray propagation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 572–579, Springer, 2002.

[3] P. An, D. Yang, J. Wang, L. Wu, J. Zhou, Z. Zeng, X. Huang, Y. Xiao, S. Hu, Y. Chen, *et al.*, "A deep learning method for delineating early gastric cancer resection margin under chromoendoscopy and white light endoscopy," *Gastric Cancer*, vol. 23, no. 5, pp. 884–892, 2020.

[4] G.-Q. Wei, K. Arbter, and G. Hirzinger, "Automatic tracking of laparoscopic instruments by color coding," in *CVRMed-MRCAS'97*, pp. 357–366, Springer, 1997.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[6] J. Xie, X. Hou, K. Ye, and L. Shen, "Clims: Cross language image matching for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4483–4492, 2022.

[7] S. Jo and I.-J. Yu, "Puzzle-cam: Improved localization via matching partial and full features," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 639–643, IEEE, 2021.

[8] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised semantic segmentation via sub-category exploration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8991–9000, 2020.

[9] P.-T. Jiang, Y. Yang, Q. Hou, and Y. Wei, "L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[10] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4981–4990, 2018.

[11] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12275–12284, 2020.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

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

[14] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

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

[16] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 1–13, 2022.

[17] I. Hirra, M. Ahmad, A. Hussain, M. U. Ashraf, I. A. Saeed, S. F. Qadri, A. M. Alghamdi, and A. S. Alfakeeh, "Breast cancer classification from histopathological images using patch-based deep learning modeling," *IEEE Access*, vol. 9, pp. 24273–24287, 2021.

[18] S. Guo, X. Liu, H. Zhang, Q. Lin, L. Xu, C. Shi, Z. Gao, A. Guzzo, and G. Fortino, "Causal knowledge fusion for 3d cross-modality cardiac image segmentation," *Information Fusion*, p. 101864, 2023.

[19] Y. Liu, G. Han, and X. Liu, "Lightweight compound scaling network for nasopharyngeal carcinoma segmentation from mr images," *Sensors*, vol. 22, no. 15, p. 5875, 2022.

[20] M. G. B. Calisto and S. K. Lai-Yuen, "Self-adaptive 2d-3d ensemble of fully convolutional networks for medical image segmentation," in *Medical Imaging 2020: Image Processing*, vol. 11313, pp. 459–469, SPIE, 2020.

[21] H. Zhang, Z. Gao, D. Zhang, W. K. Hau, and H. Zhang, "Progressive perception learning for main coronary segmentation in x-ray angiography," *IEEE Transactions on Medical Imaging*, 2022.

[22] R. Mohammadi, I. Shokatian, M. Salehi, H. Arabi, I. Shiri, and H. Zaidi, "Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer," *Radiotherapy and Oncology*, vol. 159, pp. 231–240, 2021.

[23] L. Ke, Y. Deng, W. Xia, M. Qiang, X. Chen, K. Liu, B. Jing, C. He, C. Xie, X. Guo, *et al.*, "Development of a self-constrained 3d densenet model in automatic detection and segmentation of nasopharyngeal carcinoma using magnetic resonance images," *Oral Oncology*, vol. 110, p. 104862, 2020.

[24] L.-Z. Zhong, X.-L. Fang, D. Dong, H. Peng, M.-J. Fang, C.-L. Huang, B.-X. He, L. Lin, J. Ma, L.-L. Tang, *et al.*, "A deep learning mr-based radiomic nomogram may predict survival for nasopharyngeal carcinoma patients with stage t3n1m0," *Radiotherapy and Oncology*, vol. 151, pp. 1–9, 2020.

[25] H. Messaoudi, A. Belaid, D. B. Salem, and P.-H. Conze, "Cross-dimensional transfer learning in medical image segmentation with deep learning," *Medical Image Analysis*, p. 102868, 2023.

[26] W. Xing, Z. Zhu, D. Hou, Y. Yue, F. Dai, Y. Li, L. Tong, Y. Song, and D. Ta, "Cm-segnet: A deep learning-based automatic segmentation approach for medical images by combining convolution and multilayer perceptron," *Computers in Biology and Medicine*, vol. 147, p. 105797, 2022.

[27] Y. Li, G. Han, and X. Liu, "Dcnet: Densely connected deep convolutional encoder–decoder network for nasopharyngeal carcinoma segmentation," *Sensors*, vol. 21, no. 23, p. 7877, 2021.

[28] Y. Zhi, H. Zhang, Z. Gao, *et al.*, "Vessel contour detection in intracoronary images via bilateral cross-domain adaptation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[29] Z. Zhuang, P. Jin, A. N. Joseph Raj, Y. Yuan, and S. Zhuang, "Automatic segmentation of left ventricle in echocardiography based on yolov3 model to achieve constraint and positioning," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–11, 2021.

[30] G. Patel and J. Dolz, "Weakly supervised segmentation with cross-modality equivariant constraints," *Medical Image Analysis*, vol. 77, p. 102374, 2022.

[31] H. Guo, M. Xu, Y. Chi, L. Zhang, and X.-S. Hua, "Weakly supervised organ localization with attention maps regularized by local area reconstruction," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pp. 243–252, Springer, 2020.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[33] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-cam: Causal cam for weakly supervised semantic segmentation on medical image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11676–11685, 2022.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[35] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[36] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847, IEEE, 2018.

[37] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[38] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[39] K. Wu, B. Du, M. Luo, H. Wen, Y. Shen, and J. Feng, "Weakly supervised brain lesion segmentation via attentional representation learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 211–219, Springer, 2019.