Medical Image Analysis 80 (2022) 102526

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Constrained unsupervised anomaly segmentation

Julio Silva-Rodríguez[a,*], Valery Naranjo[b], Jose Dolz[c]

[a] Institute of Transport and Territory, Universitat Politécnica de Valéncia, Valencia, Spain
[b] Institute of Research and Innovation in Bioengineering, Universitat Politécnica de Valéncia, Valencia, Spain
[c] École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

ARTICLE INFO

Article history:
Received 16 February 2022
Revised 29 May 2022
Accepted 24 June 2022
Available online 25 June 2022

Keywords:
Unsupervised anomaly localization
Constraint segmentation
Brain lesions

ABSTRACT

Current unsupervised anomaly localization approaches rely on generative models to learn the distribution of normal images, which is later used to identify potential anomalous regions derived from errors on the reconstructed images. To address the limitations of residual-based anomaly localization, very recent literature has focused on attention maps, by integrating supervision on them in the form of homogenization constraints. In this work, we propose a novel formulation that addresses the problem in a more principled manner, leveraging well-known knowledge in constrained optimization. In particular, the equality constraint on the attention maps in prior work is replaced by an inequality constraint, which allows more flexibility. In addition, to address the limitations of penalty-based functions we employ an extension of the popular log-barrier methods to handle the constraint. Last, we propose an alternative regularization term that maximizes the Shannon entropy of the attention maps, reducing the amount of hyperparameters of the proposed model. Comprehensive experiments on two publicly available datasets on brain lesion segmentation demonstrate that the proposed approach substantially outperforms relevant literature, establishing new state-of-the-art results for unsupervised lesion segmentation.

1. Introduction

Deep learning models are driving progress in a wide range of visual recognition tasks, particularly when they are trained with large amounts of annotated samples. This learning paradigm, however, carries two important limitations. First, obtaining such curated labeled datasets is a cumbersome process prone to annotator subjectivity, limiting the access to sufficient training data in practice. This problem is further magnified in the context of medical image segmentation, where labeling involves assigning a category to each image pixel or voxel. In addition, even if annotated images are available, there exist some applications, such as brain lesion detection, where large intra-class variations are not captured during training, failing to cover the broad range of abnormalities that might be present in a scan. This results in trained models which are potentially tailored to discover lesions similar to those seen during training. Thus, considering the scarcity and the diversity of target objects in these scenarios, lesion segmentation is typically modeled as an anomaly localization task, which is trained in an unsupervised manner. In this setting, the training dataset contains only *normal* images and *abnormal* images are not ideally accessible during training.

A popular strategy to tackle unsupervised anomaly segmentation is to model the distribution of normal images in the training set. To this end, generative models, such as generative adversarial networks (GANs) (Schlegl et al., 2017; Schlegl et al., 2019; Andermatt et al., 2019; Ravanbakhsh et al., 2019; Baur et al., 2020; Sun et al., 2020) and variational auto-encoders (VAEs) (Chen and Konukoglu, 2018; Nick Pawlowski, 2018; Sabokrou et al., 2019; Chen et al., 2020; Zimmerer et al., 2020) have been widely employed. In particular, these models are trained to reconstruct their input images, which are drawn from a normal, i.e., *healthy*, distribution. At inference, input images are compared to their reconstructed normal counterparts, which are recovered from the learned distribution. Then, the anomalous regions are identified from the reconstruction error.

As an alternative to these methods, a few recent works have integrated class-activation maps (CAMs) during training (Venkataramanan et al. (2020), Liu et al. (2020)). In particular, Venkataramanan et al. (2020) leverage the generated attention maps as an additional supervision cue, enforcing the network to provide attentive regions covering the whole context in normal images. This term was formulated as an equality constraint with the form of a L_1 penalty over each individual pixel. Nevertheless,

* Corresponding author.
E-mail address: jjsilva@upv.es (J. Silva-Rodríguez).

https://doi.org/10.1016/j.media.2022.102526

we found that explicitly forcing the network to produce maximum attention values across each pixel does not achieve satisfactory results in the context of brain lesion segmentation. In addition, recent literature in constrained optimization for deep neural networks suggests that simple penalties –such as the function used in Venkataramanan et al. (2020)– might not be the optimal solution to constraint the output of a CNN (Kervadec et al., 2019c).

Based on these observations, we propose a novel formulation for unsupervised semantic segmentation of brain lesions in medical images. The key contributions of our work can be summarized as follows:

- A novel constrained formulation for unsupervised lesion segmentation, which integrates an auxiliary constrained loss to force the network to generate attention maps that cover the whole context in normal images.
- In particular, we leverage *global* inequality constraints on the generated attention maps to force them to be activated around a certain target value. This contrasts with the previous work in Venkataramanan et al. (2020), where *local* pixel-wise equality constraints on Grad-CAMs (Selvaraju et al. (2020) are employed. In addition, to address the limitations of penalty-based functions, we resort to an extended version of the standard log-barrier.
- Furthermore, we consider an alternative regularization term that maximizes the Shannon entropy of the attention maps, reducing the amount of hyperparameters with respect to the extended log-barrier model, while yielding at par performances.
- We benchmark the proposed model against a relevant body of literature on two public lesion segmentation benchmarks: BraTS and Physionet-ICH datasets. Comprehensive experiments demonstrate the superior performance of our model, establishing a new state-of-the-art for this task.

This journal version provides a substantial extension of the conference work presented in Silva-Rodríguez et al. (2021). First, we extended the literature survey, particularly for unsupervised medical image segmentation. Then, in terms of methodology, the current version introduces several important modifications. In particular, we further investigate the role of the gradients on the attention maps derived from Grad-CAM in the task of unsupervised anomaly detection. Based on our empirical observations, we modify the formulation in Silva-Rodríguez et al. (2021) to constraint directly the activation maps without involving any gradient information. Furthermore, we propose an alternative learning objective for our constrained problem based on the Shannon entropy. More concretely, we replace our log-barrier formulation by a maximizing entropy term on the softmax activation of brain tissue pixels, which reduces the complexity in terms of hyperparameters with respect to the former model. Last, we add comprehensive experiments to empirically validate our method, including an additional dataset and extensive ablation studies on several design choices.

2. Related work

2.1. Unsupervised anomaly segmentation

Unsupervised anomaly segmentation aims at identifying abnormal pixels on test images, containing, for example, lesions on medical images (Baur et al., 2020; Chen and Konukoglu, 2018), defects in industrial images (Bergmann et al., 2019; Liu et al., 2020; Venkataramanan et al., 2020) or abnormal events in videos (Abati et al., 2019; Ravanbakhsh et al., 2019). A main body of the literature has explored unsupervised deep (generative) representation learning to learn the distribution from normal data. The underlying assumption is that a model trained on normal data will not be able to reconstruct anomalous regions, and the reconstructed difference can therefore be used as an anomaly score. Under this learning paradigm, generative adversarial networks (GAN) (Goodfellow et al., 2014) and variational auto-encoders (VAE) (Kingma and Welling, 2014) are typically employed. Nevertheless, even though GAN and VAE model the latent variable, the manner in which they approximate the distribution of a set of samples differs. GAN-based approaches (Schlegl et al., 2017; Schlegl et al., 2019; Andermatt et al., 2019; Ravanbakhsh et al., 2019; Baur et al., 2020; Sun et al., 2020) approximate the distribution by optimizing a generator to map random samples from a prior distribution in the latent space into data points that a trained discriminator cannot distinguish. On the other hand, data distribution is approximated in VAE by using variational inference, where an encoder approximates the posterior distribution in the latent space and a decoder models the likelihood (Sabokrou et al., 2019; Dehaene et al., 2020). Recent literature on unsupervised anomaly segmentation also includes non VAE and GAN based approaches. For instance, Bergmann et al. (2020) exploits the teacher-student learning paradigm, highlighting anomalies on those outputs where the student networks and teacher model predictions differ. Additionally, feature-based methods (Shi et al., 2021; Bergmann et al., 2020), which identify anomalies in the feature space can be also employed.

2.2. Unsupervised anomaly segmentation in medical imaging

In the context of medical images, most current literature resorts to VAEs, proposing several improvements to overcome specific limitations of simple VAEs (Chen and Konukoglu, 2018; Nick Pawlowski, 2018; Chen et al., 2020; Zimmerer et al., 2019). For example, to handle the lack of consistency in the learned latent representation on prior works, Chen and Konukoglu (2018) included a constraint that helps mapping an image containing abnormal anatomy close to its corresponding healthy image in the latent space. Zimmerer et al. (2019) presented a context-encoding VAE that combines reconstruction- with density-based anomaly scoring to capture the high-level structure present in the data. More recently, a probabilistic model that uses a network-based prior as the normative distribution on the latent-variable model was proposed in Chen et al. (2020). In particular, this model penalized large deviations between the reconstructed and original input images, reducing false positives in pixel-wise predictions. Generative models have been also employed to tackle the unsupervised lesion segmentation task (Baur et al., 2020; Nguyen et al., 2021). While SteGANomaly (Baur et al., 2020) integrated a CycleGAN-based style-transfer framework to map samples in the latent space much closer to the training distribution, Nguyen et al. (2021) mask out random regions of the input data before they are fed to the GAN model. Note that a detailed survey on unsupervised anomaly localization in medical imaging can be found in Baur et al. (2021). However, despite the recent popularity of these methods, the results from the Medical Out-of-Distribution Analysis Challenge 2020 (Zimmerer et al., 2022) highlight their suboptimal performance on anomaly segmentation, which might impede their usability in clinical practice, as stressed by Meissen et al. (2022).

More recently, Venkataramanan et al. (2020) integrate attention maps derived from Grad-CAM (Selvaraju et al., 2020) during the training as supervisory signals. In particular, in addition to standard learning objectives, authors introduce an auxiliary loss that tries to maximize the attention maps on normal images by including an equality constraint with the form of a L_1 penalty over each individual pixel.

2.3. Constrained segmentation

Imposing global constraints on the output predictions of deep CNNs has gained attention recently, particularly in weakly supervised segmentation. These constraints can be embedded into the network outputs in the form of direct loss functions, which guide the network training when fully labeled images are not accessible. For example, a popular scenario is to enforce the softmax predictions to satisfy a prior knowledge on the size of the target region. Jia et al. (2017) employed a L_2 penalty to impose equality constraints on the size of the target regions in the context of histopathology image segmentation. In Zhang et al. (2017), authors leverage the target properties by enforcing the label distribution of predicted images to match an inferred label distribution of a given image, which is achieved with a KL-divergence term. Similarly, Zhou et al. (2019) proposed a novel loss objective in the context of partially labeled images, which integrated an auxiliary term, based on a KL-divergence, to enforce that the average output size distributions of different organs approximates their empirical distributions, obtained from fully-labeled images.

While the equality-constrained formulations proposed in these works are very interesting, they assume exact knowledge of the target size prior. In contrast, inequality constraints can relax this assumption, allowing much more flexibility. In Pathak et al. (2015), authors imposed inequality constraints on a latent distribution – which represents a fake ground truth– instead of the network output, to avoid the computational complexity of directly using Lagrangian-dual optimization. Then, the network parameters are optimized to minimize the KL divergence between the network softmax probabilities and the latent distribution. Nevertheless, their formulation is limited to linear constraints. More recently, inequality constraints have been tackled by augmenting the learning objective with a penalty-based function, e.g., L_2 penalty, which can be imposed within a continuous optimization framework (Kervadec et al., 2019c; Kervadec et al., 2019a; Bateson et al., 2021), or in the discrete domain (Peng et al., 2020). Despite these methods have demonstrated remarkable performance in weakly supervised segmentation, they require that prior knowledge, *exact* or *approximate*, is given. This contrasts with the proposed approach, which is trained on data without anomalies, and hence the size of the target is zero.

3. Methodology

An overview of our method is presented in Fig. 1. In what follows, we describe each component of our methodology.

Preliminaries Let us denote the set of unlabeled training images as $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{\Omega_i}$ represents the ith image and Ω_i denotes the spatial image domain. This dataset contains only normal images, e.g., healthy images in the medical context, and has therefore no segmentation mask associated with each image. We now define an encoder, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by $\boldsymbol{\theta}$, which is optimized to project normal data points in $\mathcal{D}$ into a manifold represented by a lower dimensionality d, $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^d$. Furthermore, a decoder $f_{\boldsymbol{\phi}}(\cdot) : \mathcal{Z} \rightarrow \mathcal{X}$ parameterized by $\boldsymbol{\phi}$ aims at reconstructing an input image $\mathbf{x} \in \mathcal{X}$ from $\mathbf{z} \in \mathcal{Z}$, which results in $\hat{\mathbf{x}} = f_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}))$.

3.1. Vanilla VAE

A Variational Autoencoder (VAE) is an encoder-decoder style generative model, which is currently the dominant strategy for unsupervised anomaly location. Training a VAE consists in minimizing a two-term loss function, which is equivalent to maximize the evidence lower-bound (ELBO) (Kingma and Welling, 2014):

$$\mathcal{L}_{VAE} = \mathcal{L}_R(\mathbf{x}, \hat{\mathbf{x}}) + \beta \mathcal{L}_{KL}(q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{1}$$

where $\mathcal{L}_R$ is the reconstruction error term between the input and its reconstructed counterpart. The right-hand term is the Kullback-Leibler (KL) divergence (weighted by β) between the approximate posterior $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$, which acts as a regularizer, penalizing approximations for $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ that differ from the prior.

3.2. Size regularizer via VAE attention

Very recent literature (Liu et al., 2020; Venkataramanan et al., 2020) has explored the use of attention maps for anomaly localization. In particular, attention maps $\mathbf{a} \in \mathbb{R}^{\Omega_i}$ are generated from the latent mean vector $\mathbf{z}_\mu$, by using Grad-CAM (Selvaraju et al., 2020) via backpropagation to an encoder block output $f^s_{\boldsymbol{\theta}}(\mathbf{x})$, at a given network depth s. Thus, for a given input image $\mathbf{x}^n$ its corresponding attention map is computed as follows:

$$\mathbf{a}^n = \sigma\left(\sum_k^K \alpha_k f^s_{\boldsymbol{\theta}}(\mathbf{x}^n)_k\right) \tag{2}$$

where K is the total number of filters of that encoder layer, σ a sigmoid operation, and α_k are the generated gradients such that: $\alpha_k = \frac{1}{|\mathbf{a}^n|}\sum_{t\in\Omega_T} \frac{\partial \mathbf{z}_\mu}{\partial \mathbf{a}^n_{k,t}}$, where Ω_T is the spatial features domain.

In Venkataramanan et al. (2020), authors leveraged the Grad-CAMs based attention maps (Eq. (2)) by enforcing them to cover the whole normal image. To achieve this, their loss function was augmented with an additional term, referred to as expansion loss, which takes the form of: $\mathcal{L}_s = \frac{1}{|\mathbf{a}|}\sum_{l\in\Omega_i}(1 - \mathbf{a}^n_l)$. We can easily observe that this term resembles to multiple equality constraints, one at each pixel, forcing the class activation maps to be maximum at the whole image in a pixel-wise manner (i.e., it penalizes each single pixel individually). Contrary to this work, we integrate supervision on attention maps by enforcing inequality constraints on its global target size. Note that the use of the inequality constraints is motivated by the choice of the barrier function in the constrained problem, which is further detailed in Section 3.3. Hence, we aim at minimizing the following constrained optimization problem:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \mathcal{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\phi}) \qquad \text{s.t.} \quad f_c(\mathbf{a}^n) \leq 0, \quad n = 1, \ldots, N \tag{3}$$

where $f_c(\mathbf{a}^j) = (1 - \frac{1}{|\Omega_i|}\sum_{l\in\Omega_i}\mathbf{a}^n_l)$ is the constraint over the attention map from the jth image, which enforces the generated attention map to cover the whole image. It is well-known in optimization that a penalty does not act as a barrier near the boundary of the feasible set (Boyd et al., 2004). In other words, a constraint that is satisfied results in a null penalty and gradient. Therefore, at a given gradient update, there is nothing that prevents a satisfied constraint from being violated, causing oscillations between competing constraints and ultimately resulting in a potential unstable training. This is further exacerbated in the case of many multiple constraints (i.e., Venkataramanan et al., 2020), motivating the use of a single global constraint to achieve a maximum coverage of class-activation maps over the whole image in our scenario. From Eq. (3) we can derive an approximate unconstrained optimization problem by employing a penalty-based method, which takes the hard constraint and moves it into the loss function as a penalty term ($\mathcal{P}(\cdot)$): $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathcal{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \lambda\mathcal{P}(f_c(\mathbf{a}))$. Thus, each time that the constraint $f_c(\mathbf{a}^n) \leq 0$ is violated, the penalty term $\mathcal{P}(f_c(\mathbf{a}^n))$ increases.

3.3. Extended log-barrier as an alternative to penalty-based functions

Despite having demonstrated a good performance in several applications (Kervadec et al., 2019b; Pathak et al., 2015; He et al., 2017; Jia et al., 2017) penalty-based methods have several drawbacks. First, these unconstrained minimization problems have increasingly unfavorable structure due to ill-conditioning

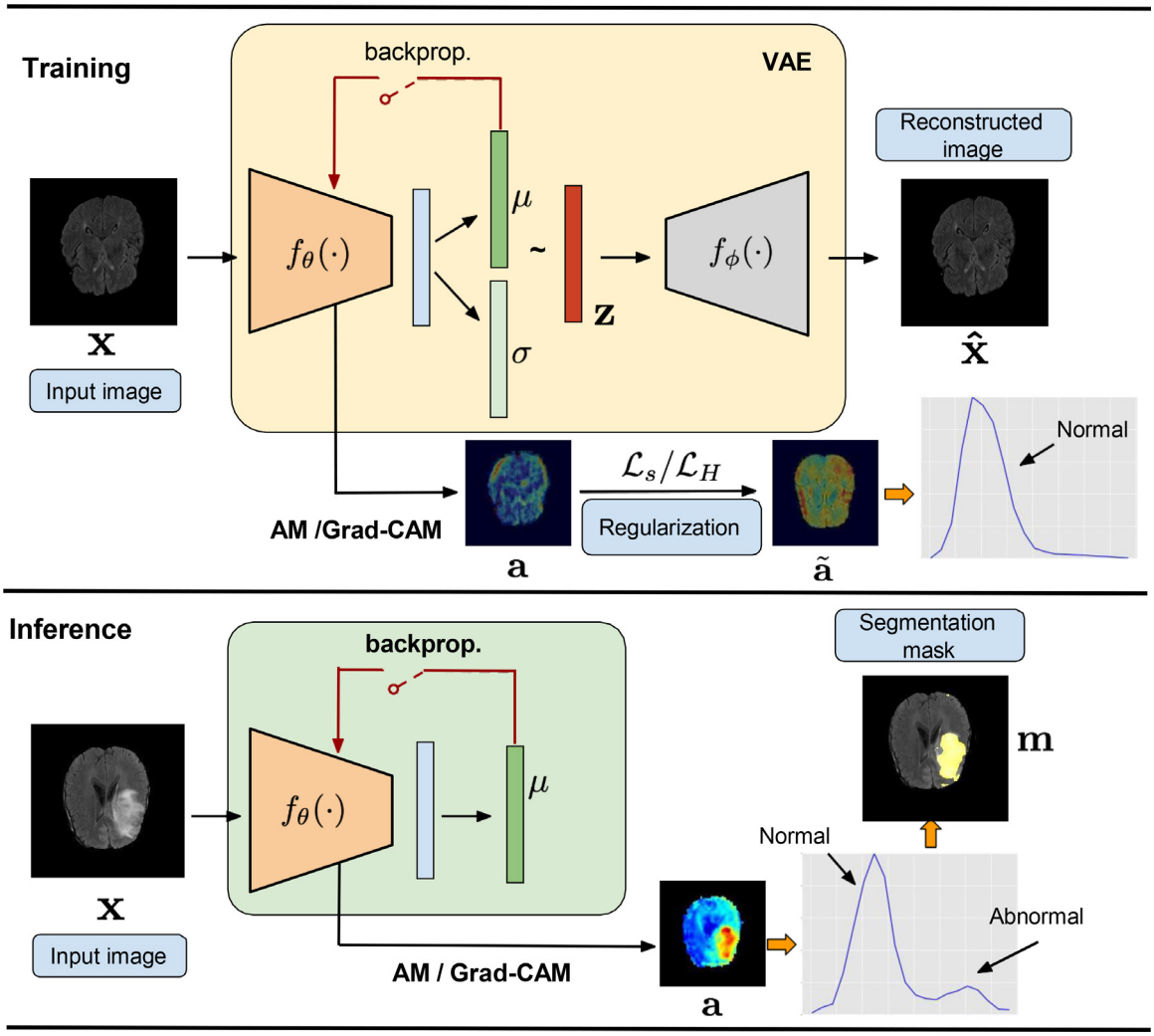

Fig. 1. Method overview. Following the standard literature, the VAE is optimized to maximize the evidence lower bound (ELBO), which satisfies Eq. (1). In addition, we include an attention constraint (in the form of a size-constrained loss $\mathcal{L}_s$ or entropy proxy $\mathcal{L}_H$) on the attention maps **a**, to force the network to search in the whole image. At inference, the attention map is thresholded to obtain the final segmentation mask **m**.

(Fiacco and McCormick, 1990; Luenberger, 1973), which typically results in an exceedingly slow convergence. Second, finding the optimal penalty weight is not trivial. In addition, we advocate for the use of the log-barrier extension versus penalties due to the strictly positive gradient of the latter becomes higher when a satisfied constraint approaches violation during optimization, pushing it back towards the feasible set (See Figure 1 in Kervadec et al., 2019c). As explained in the previous section, this contrasts with penalties, as they deliver null gradients if a given constraint is satisfied. To address these limitations, we replace the penalty-based functions by the approximation of log-barrier[1] presented in Kervadec et al. (2019c). We would like to stress that barrier methods require the interior of the feasible sets to be non-empty and they are used, therefore, in constrained optimization problems with inequality constraints, such as the one defined in Eq. (3) (note that there is no interior for equality constraints). Thus, we can formally define the approximation of log-barrier as:

$$\widetilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(-z) & \text{if } z \leq -\frac{1}{t^2} \\ tz - \frac{1}{t}\log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise,} \end{cases} \tag{4}$$

[1] Note that this function is convex, continuous and twice-differentiable.

where t *controls* the barrier during training, and z is the constraint $f_c(\mathbf{a}^n)$. Thus, by taking into account the approximation in 4, we can solve the following unconstrained problem by using standard Gradient Descent:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \underbrace{\mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi})}_{\text{Standard VAE loss}} + \lambda_s \underbrace{\sum_{n=1}^{N} \widetilde{\psi}_t\left(1 - \frac{1}{|\Omega_i|}\sum_{l\in\Omega_i} \mathbf{a}_l^n\right)}_{\mathcal{L}_s\text{: Size regularizer}} \tag{5}$$

In this scenario, for a given t, the optimizer will try to find a solution with a good compromise between minimizing the loss of the VAE and satisfying the constraint $f_c(\mathbf{a}^n)$. In the following, we refer to this formulation of gradient-CAM constraint as GradCAM-Cons setting.

3.4. On the role of gradients in VAEs

Even though there exist a few initial attempts to integrate attention maps on the task of unsupervised anomaly detection, how gradient-based attention behave on anomalous patterns remains unclear. For instance, Liu et al. (2020) argue that anomalies produce larger gradients in the learned latent representation, which results in higher activated attention maps. On the other hand, Venkataramanan et al. (2020) states that the VAE only focus on

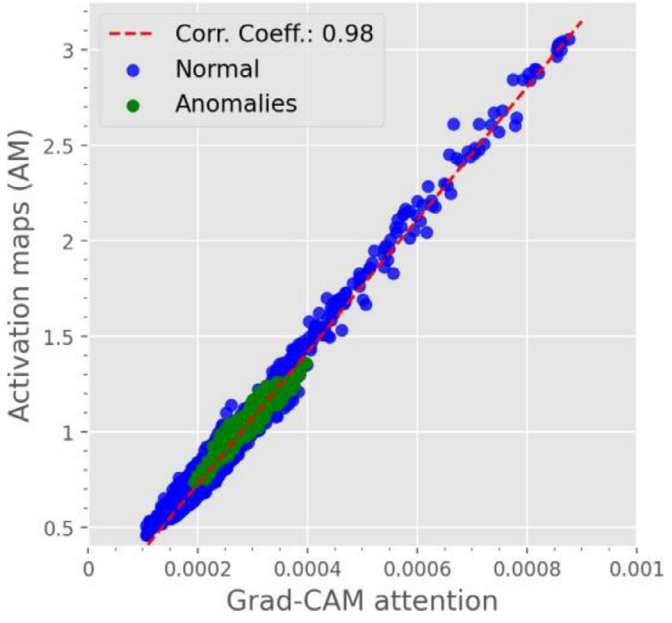

Fig. 2. Relation between the activation values and gradient-weighted attention maps in an unconstrained VAE. These results demonstrate that the values obtained by Grad-CAM based attention are highly correlated (correlation coefficient = 0.98) to those obtained by the attention maps, suggesting that the gradient basically contributes as a scaling factor on the attention maps.

normal patterns (with which it has been trained), thus anomalous regions produce smaller absolute value gradients. These inconsistencies in the literature have motivated us to analyze the underlying role of the gradients in the context of brain images analysis. Thus, we performed several experiments to analyze the behaviour of grad-CAMs in anomaly localization compared to non-weighted activation maps (AMs), which are computed as:

$$\mathbf{a}^n = \frac{1}{K}\sum_{k}^{K} f^s_{\boldsymbol{\theta}}(\mathbf{x}^n)_k \tag{6}$$

In particular, we could not find any benefit on gradients weighting other than serving as a scaling factor for attention maps to fall on non-saturated range of values of typically used activation functions, such as the sigmoid operation in Eq. (2) (see Fig. 2, where we show that the values obtained by both types of attention are highly correlated). Furthermore, we found that the reconstructed images derived from the gradient-based attention contained more errors compared to those reconstructed with attention on the activation maps (Eq. (6)). We refer the reader to Section 1 of Supplemental Material for the detailed results concerning the role of the gradients.

3.5. Entropy maximization as a proxy for the constraint

Based on our previous findings, we advocate that the use of non-weighted activation maps (AMs) should be preferred over their gradient-based counterpart. Nevertheless, this solution has a main limitation that hinders the use of size constraints. As the activation maps are not normalized, the arbitrary activation value to impose the constraint loses the sense of *size* or *proportion*. The activation values produced by neural networks can vary in each application, as well as with the architecture used, which makes it difficult to establish generalizable restrictions on their value. For this reason, we propose to use attention maps derived from normalizing the activation maps over all the pixels of the image, via a softmax activation, similarly to Ilse et al. (2018), such that: $p^n = \tau_{\Omega_B}(\mathbf{a}^n)$[2] Since these attention maps are normalized across pixels and not over classes, the use of global constraints is meaningless, as the sum over all the pixels post-softmax will be equal to 1.0. Nevertheless, we still aim at regularizing the attention distribution p^n to focus on all patterns in the image *homogeneously*. To this end, we propose to minimize the KL distance $D_{KL}(p||q) = H(p,q) - H(p)$ between the attention distribution p, and a constant distribution q, where $H(p,q)$ represents the cross-entropy between both distributions, and $H(p) = H(p,p)$ is the Shannon entropy of the intensity distribution such that $H(p) = -\frac{1}{I}\sum_i p_i \cdot \log(p_i)$. In the scenario where we want p to match a constant distribution, it is straightforward to see that minimizing the KL distance is equivalent to maximizing the entropy $H(p)$:

$$D_{KL}(p||q) = H(p,q) - H(p) =^c -H(p) \tag{7}$$

where $=^c$ indicates equality up to an additive constant.

Thus, the proposed constrained optimization problem integrating an entropy maximization term, referred to as $\mathcal{L}_H$, offers a softer attention constraint compared to the solution in Eq. (5). Furthermore, this formulation allows the VAE to keep the most suitable activation values, while requiring less hyper-parameters to be optimized. Analogously to Eq. (5), we solve the constrained optimization problem with $\mathcal{L}_H$ by using standard Gradient Descent:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \underbrace{\mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi})}_{\text{Standard VAE loss}} - \lambda_H \underbrace{\frac{1}{N}\sum_{n=1}^{N} H(\tau_{\Omega_B}(\mathbf{a}^n))}_{\mathcal{L}_H\text{: Entropy regularizer}} \tag{8}$$

Hereafter, we will refer to this formulation as AMCons.

3.6. Inference

During inference, we use the generated attention as an anomaly saliency map. For the Grad-CAMs based settings we replaced the sigmoid operation by a minimum-maximum normalization in order to avoid saturation caused by large activations. During the experimental stage, we found that anomalies produce larger activation on attention maps than the constrained normal samples, in line to prior literature (Liu et al., 2020). Then, the map is thresholded to create an anomaly mask of the image.

4. Experimental setting

4.1. Datasets

The experiments described in this work are carried out in the context of brain lesions localization. Concretely, we use two relevant neuroimaging challenges: tumour segmentation in MRI volumes and intracranial hemorrhage (ICH) segmentation in CT scans.

Brain tumor segmentation For this task, we used the popular BraTS 2019 dataset (Menze et al., 2015; Bakas et al., 2017; Bakas et al., 2018), which contains 335 multi-institutional multi-modal MR scans with their corresponding Glioma segmentation masks. Following Baur et al. (2019), from every patient, 10 consecutive axial slices of FLAIR modality of resolution 224×224 pixels were extracted around the center to get a pseudo MRI volume. Then, the dataset is split into training, validation and testing groups, with 271, 32 and 32 patients, respectively. Following the standard literature, during training only the slices without lesions are used as normal samples. For validation and testing, scans with less than 0.01% of tumour are discarded, following the standard practices in the literature.

[2] Note that τ is the softmax activation on the brain tissue instances, Ω_B.

Intracranial hemorrhage segmentation We use the Physionet-ICH dataset (Hssayeni, 2020; Hssayeni et al., 2020; Goldberger et al., 2000) to localize intracranial hemorrhage lesions. The dataset is composed of 82 non-contrast CT scans of subjects with traumatic brain injury. From those, 36 cases are diagnosed with intracranial hemorrhage of different types: Intraventricular, Intraparenchymal, Subarachnoid, Epidural and Subdural. ICH Lesions were slice-wise delineated by two expert radiologists. In our work, we join the different ICH types into one single label for binary lesion segmentation. CT scans are skull-stripped, intensity-normalized, and co-registered into a reference scan. Similar to the BraTS dataset, 10 consecutive axial slices of resolution 224×224 pixels around the center were extracted to get CT pseudo volumes. The dataset is divided into training, validation and testing splits. The first one contains only non-ICH cases (n=46), while cases with labeled lesions were used for validation (n=6) and testing (n=30). Although the main core of ablation experiments in this work are described on the BraTS dataset, we use the Physionet-ICH dataset to demonstrate the generalization capabilities of our proposed method on different brain lesions and imaging modalities.

4.2. Evaluation metrics

We resort to standard metrics for unsupervised brain lesion segmentation, as in Baur et al. (2021). Concretely, we compute the dataset-level area under precision-recall curve (AUPRC) at pixel level, as well the area under receptive-operative curve (AUROC). From the former, we obtain the operative point (OP) as threshold to generate the final segmentation masks. Then, we compute the best dataset-level Sørensen-Dice score ($\lceil$DICE$\rceil$) and intersection-over-union ($\lceil$IoU$\rceil$) over these segmentation masks. Finally, we compute the average Sørensen-Dice score (DICE) over single scans. For each experiment, the metrics reported are the average of three consecutive repetitions of the training, to account for the variability of the stochastic factors involved in the process.

4.3. Implementation details

The VAE architecture used in this work is based on the recently proposed framework in Venkataramanan et al. (2020). Concretely, the convolution layers of ResNet-18 (He et al., 2016) are used as the encoder, followed by a dense latent space $\mathbf{z} \in \mathbb{R}^{32}$. For image generation, a residual decoder is used, which is symmetrical to the encoder. It is noteworthy to mention that, even though several methods have resorted to a spatial latent space (Baur et al., 2019; Venkataramanan et al., 2020), we observed that a dense latent space provided better results, which aligns to the recent benchmark in Baur et al. (2021). To train the GradCAMCons formulation in Eq. (5) we first trained the VAE during 50 epochs without any expansion to stabilize the convergence using $\beta = 1$. Then, the proposed regularizer was integrated (Eq. (5)) with $t = 10$ and $\lambda_s = 10^3$ applied to the Grad-CAMs obtained from the first convolutional block of the encoder during 250 epochs. We use a batch size of 8 images, and a learning rate of $1e{-}5$ with ADAM as optimizer. The reconstruction loss, $\mathcal{L}_R$, in Eq. (1) is the binary cross-entropy. Similarly, the AMCons formulation in Eq. (8) was trained by using $\beta = 10$ and $\lambda_H = 0.1$, using a learning rate of $1e{-}4$. Ablation experiments to motivate the choice of values used are presented in Section 5.2 and Section 3 of supplemental materials. The code and trained models are publicly available on (https://github.com/jusiro/constrained_anomaly_segmentation/).

4.4. Baselines

In order to compare our approach to state-of-the-art methods, we implemented prior works and validated them on the dataset used, under the same conditions. First, we use residual-based methods to match the recently benchmark on unsupervised lesion localization in Baur et al. (2021). Then, we implement up-to-date methods based on contrast adjustment on the input image via histogram equalization. We also include recently proposed methods that integrate CAMs to locate anomalies. For both strategies, the AE/VAE architecture was the same as the one used in the proposed method. ***Residual methods***, given an anomalous sample, aim to use the AE/VAE to reconstruct its normal counterpart. Then, they obtain an anomaly localization map using the residual between both images such that $\mathbf{m} = |\mathbf{x} - \hat{\mathbf{x}}|$, where $|\cdot|$ indicates the absolute value. On the AE/VAE scenario, we include methods which propose modifications over vanilla versions, including context data augmentation in Context AE (Zimmerer et al., 2019), Bayesian AEs (Nick Pawlowski, 2018), Restoration VAEs (Chen et al., 2020), an adversarial-based VAEs, AnoVAEGAN (Baur et al., 2019) and a recent GAN-based approach, F-anoGAN (Schlegl et al., 2019). For methods including adversarial learning, DC-GAN (Radford et al., 2016) is used as discriminator. During inference, residual maps are masked using a slight-eroded brain mask, to avoid noisy reconstructions along the brain borderline. ***Equalization-based methods***: very recent methods have highlighted the limits of residual-based approaches to properly discern brain lesions Meissen et al. (2021, 2022). In contrast, they propose to apply an equalization of the histogram of the input image, and to set a threshold on the preprocessed image, considering that brain lesions often show hyperintense patterns in different modalities. Concretely, we include the method proposed in Meissen et al. (2021), which we refer to as HistEq. ***CAMs-based***: we use Grad-CAM VAE (Liu et al., 2020), which obtains regular Grad-CAMs on the encoder from the latent space $\mathbf{z}_\mu$ of a trained vanilla VAE. Concretely, we include a disentanglement variant of CAMs proposed in this work, which computes the combination of individually-calculated CAMs from each dimension in $\mathbf{z}_\mu$, referred to as Grad-CAM$_D$ VAE. We also use the recent method in Venkataramanan et al. (2020) (CAVGA), which applies a L1 penalty on the generated CAM to maximize the attention. In contrast to our model and Liu et al. (2020), the anomaly mask in Venkataramanan et al. (2020) is generated by focusing on the regions not activated on the saliency map such that $\mathbf{a} = 1 - CAM$, hypothesizing that the network has learnt to focus only on normal regions. Then, $\mathbf{a}$ is thresholded with 0.5 to obtain the final anomaly mask $\mathbf{m} \in \mathbb{R}^{\Omega_i}$. For both methods, the network layer to obtain the Grad-CAMs is the same as in our method.

5. Results

5.1. Comparison to the literature.

The quantitative results obtained by the proposed model and baselines on the test cohort are presented in Table 1. Results from residual-based baselines range between $[0.056 - 0.511]$(AUPRC) and $[0.188 - 0.525]$ (DICE), which are in line with previous literature Baur et al. (2021). We can observe that the proposed formulations outperform these approaches by a large margin. Concretely, the AMCons method provides a substantial increase of $\sim$34% and $\sim$26% in terms of AUPRC and DICE, respectively, compared to the best model, i.e., F-anoGAN. Furthermore, the model integrating the $\mathcal{L}_H$ term significantly outperforms our previous method in Silva-Rodríguez et al. (2021). This supports our hypothesis that using non-weighted attention maps with a maximization entropy term as constraint is indeed a better solution for the unsupervised lesion segmentation task. Finally, in comparison with the very recently proposed method of histogram equalization, HistEq, our proposed formulation brings improvements of nearly $\sim$10% in the main figures of merit.

Table 1
Comparison to prior literature on BraTS dataset. Results derived from the proposed methods in gray. Best results in bold. The values in parentheses indicate the standard deviation over the three training repetitions.

Method	AUROC	AUPRC	⌈DICE⌉	⌈IoU⌉	DICE ($\mu \pm \sigma$)
CAVGA (Venkataramanan et al., 2020)	0.726(0.001)	0.056(0.005)	0.188(0.001)	0.104(0.002)	0.182(0.004)±0.096(0.002)
Bayesian VAE (Nick Pawlowski, 2018)	0.922(0.002)	0.193(0.005)	0.342(0.005)	0.206(0.005)	0.329(0.005)±0.115(0.005)
AnoVAEGAN (Baur et al., 2019)	0.925(0.020)	0.232(0.052)	0.359(0.074)	0.221(0.053)	0.349(0.071)±0.115(0.015)
Bayesian AE (Nick Pawlowski, 2018)	0.940(0.002)	0.279(0.009)	0.389(0.012)	0.242(0.009)	0.375(0.010)±0.130(0.011)
AE	0.937(0.002)	0.261(0.011)	0.397(0.011)	0.248(0.008)	0.386(0.010)±0.125(0.004)
Grad-CAM$_D$ VAE (Liu et al., 2020)	0.941(0.003)	0.312(0.010)	0.400(0.009)	0.250(0.012)	0.361(0.014)±0.164(0.005)
Restoration VAE (Chen et al., 2020)	0.934(0.028)	0.352(0.111)	0.403(0.099)	0.252(0.069)	0.345(0.075)±0.186(0.044)
Context VAE (Zimmerer et al., 2019)	0.939(0.004)	0.271(0.017)	0.406(0.020)	0.255(0.016)	0.394(0.017)±0.126(0.007)
Context AE (Zimmerer et al., 2019)	0.940(0.003)	0.278(0.012)	0.411(0.014)	0.259(0.011)	0.399(0.013)±0.126(0.005)
VAE (Baur et al., 2019; Zimmerer et al., 2020)	0.940(0.002)	0.273(0.010)	0.411(0.012)	0.259(0.009)	0.399(0.010)±0.127(0.004)
F-anoGAN (Schlegl et al., 2019)	0.946(0.026)	0.511(0.190)	0.525(0.147)	0.369(0.131)	0.494(0.138)±0.151(0.038)
GradCAMCons w. $\mathcal{L}_S$ (L2 penalty)	0.969(0.015)	0.567(0.138)	0.620(0.085)	0.455(0.086)	0.586(0.079)±0.184(0.028)
HistEq (Meissen et al., 2021)	0.972(0.000)	0.725(0.000)	0.705(0.000)	0.545(0.000)	0.653(0.000)±0.233(0.000)
GradCAMCons w. $\mathcal{L}_S$ (Log Barrier)	0.982(0.001)	0.746(0.034)	0.698(0.034)	0.537(0.041)	0.677(0.021)±0.215(0.019)
AMCons w. $\mathcal{L}_H$	**0.988**(0.000)	**0.850**(0.011)	**0.786**(0.009)	**0.648**(0.013)	**0.741**(0.009)±**0.153**(0.001)

Table 2
Quantitative comparison, in terms of AUPRC, between enforcing the constraint at pixel-level (i.e., Venkataramanan et al., 2020) or at image-level (i.e., proposed approach), and for the impact of the type of regularization.

	L2 (pixel-level)	L2 (image-level)	Log-Barrier (image-level)
AUPRC	0.489(0.098)	0.550(0.160)	0.728(0.034)

5.2. Ablation experiments

The following ablation studies aim at demonstrating, in an empirical way, the motivation of employing the proposed models. First, we provide quantitative evidences about the better performance of using global constraints (model in Eq. (5)) over pixel-level constraints (i.e., Venkataramanan et al., 2020). Second, we show that resorting to the extended log-barrier function is a better alternative than standard L2 penalty functions. Then, we perform an in-depth analysis of the optimal hyperparameters values for the entropy-guided model (Eq. (8)), as well as other important design choices.

Image vs. pixel-level constraint The following experiment demonstrates the benefits of imposing the constraint on the whole image rather than in a pixel-wise manner, such as in Venkataramanan et al. (2020). In particular, we compare the two strategies when the constraint is enforced via a L2-penalty function, whose results are presented in Table 2. In particular, we can easily see that imposing the constraint at image-level consistently outperforms pixel-level constraints. These results support our hypothesis that global constraints, such as the proposed formulation in Eq. (5), should be preferred over multiple pixel-wise constraints, similar to Venkataramanan et al. (2020).

Extended log-barrier vs. penalty-based functions To motivate the choice of employing the extended log-barrier over standard penalty-based functions in the constrained optimization problem in Eq. (3), we compare them in Table 2. It can be observed that imposing the constraint with the extended log-barrier consistently outperforms the L2-penalty, with substantial performance gains.

On the impact of entropy-guided constraints We now perform an in-depth analysis of the effect of integrating the entropy-guided constraint in Eq. (8) for anomaly localization, as well as an extensive validation of the values of the balancing terms β and λ_H. First, we study the impact of $\mathcal{L}_H$ across different β values (i.e. $\beta = \{0.01, 0.1, 1, 10\}$), by fixing its balancing term λ_H to 0.1, a value that empirically showed good stability. These results, which are reported in Fig. 3a, show that the VAE with and without entropy constraint presents different optimal values for β. Nevertheless, the best results are obtained when the contribution of the regularization term is large (i.e. $\beta \geq 1$), and the entropy-based regularization over the activation maps included (i.e., green bars). Furthermore, this configuration is shown to be more stable once a large β weight is set, particularly for the constrained formulation. Then, based on the best configuration ($\beta = 10$), we study how different λ_H weights $\{0.01, 0.1, 1, 10\}$ impact the model performance. These results (Fig. 3b) show that incorporating the entropy regularization always contributes to performance gains, with an optimum weight value of $\lambda_H = 0.1$.

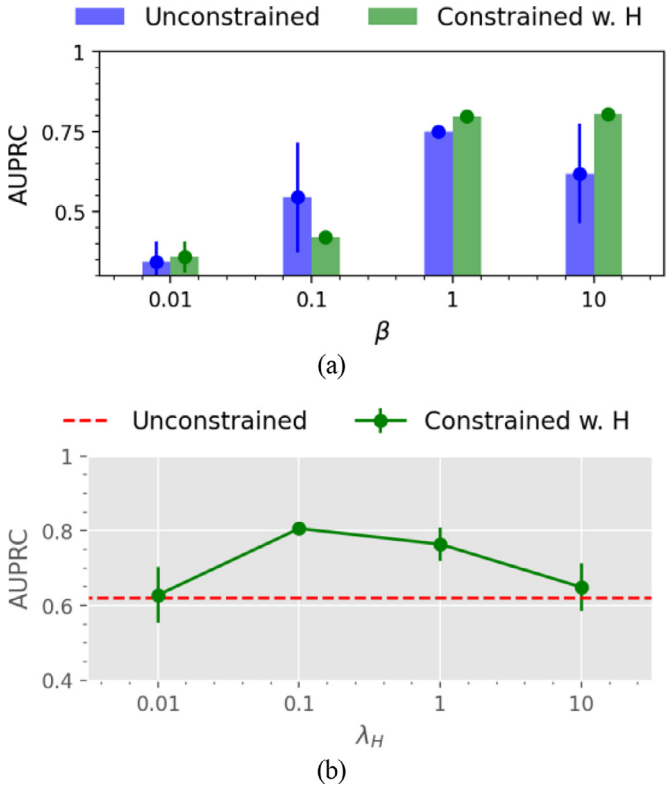

Fig. 3. Ablation study on the AMCons setting. Concretely, the role of the KL regularization (β) in the VAE and the entropy constraint on attention maps (λ_H) from our formulation is studied. (a) Entropy constraint effect and dependency on β. (b) Ablation study on λ_H.

In the next experiment, we show how adding the $\mathcal{L}_H$ term in our formulation impacts the activation maps (AM). Concretely, we first show in Fig. 4 the AM distribution for a normal sample for both the constrained and unconstrained configurations. It

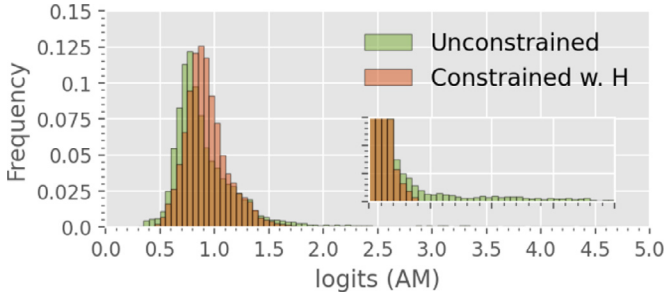

Fig. 4. Influence of the entropy constrained term on the attention maps for AMCons on normal images.

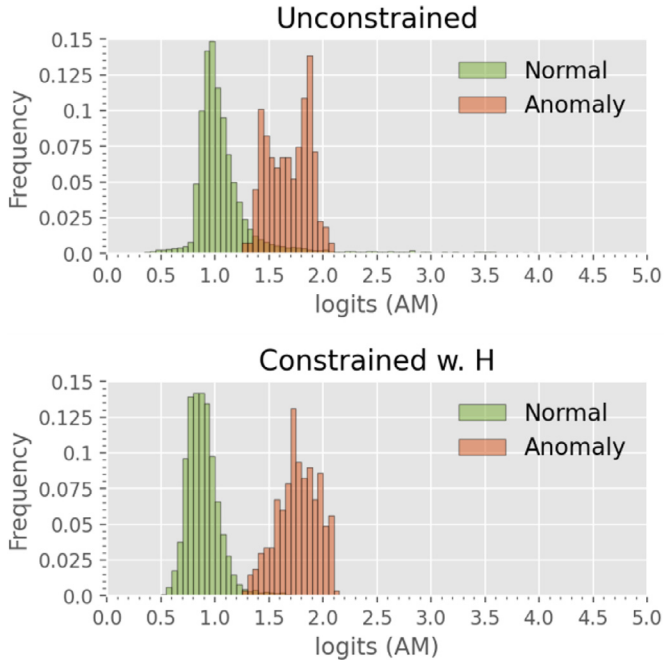

Fig. 5. Influence of the entropy constrained term on the attention maps for AMCons on images with anomalies.

can be observed that, in our constrained formulation, the distribution of activation values is more homogeneous (in orange), unlike the more spread values found in its unconstrained counterpart (in green). Furthermore, we show its impact on unseen, anomalous samples, where the benefits of our model are better highlighted. In particular, we represent the AM distribution for normal and anomalous pixels on the unconstrained formulation (i.e. $\lambda_H = 0$) in Fig. 5 (*top*), and the effect of integrating the L_H term (Fig. 5, *bottom*). Similarly to the normal samples, the distribution of normal pixels produced by the unconstrained setting spreads over a larger range, resulting in a higher overlapping with the distribution of anomalous pixels. Note that, in addition to the overlapping regions, there exist values of normal pixels which overpass anomalous values. In contrast, the more compact distribution provided by the proposed formulation favors a smaller overlap between normal and anomalous pixel intensity distributions. This results in an easier identification of normal *versus* anomalous pixels.

In the following, we explore how the entropy constraint favors the smallest overlap between normal and anomalous distribution on the objective criteria, compared to previous literature. To do so, we depict in Fig. 6 the distribution of both populations for the proposed methods, AMCons and GradCAMCons, and the most promising baselines, F-anoGAN and Histeq. Furthermore, we obtain the overlap between both distributions by dividing the number of samples in the overlapped region of the histograms by the total number of samples. It can be seen how the proposed method based on entropy maximization obtains the smallest overlap (10.2%) and produces a narrower distribution of normal samples in comparison with the GradCAMCons method, based on size constraints.

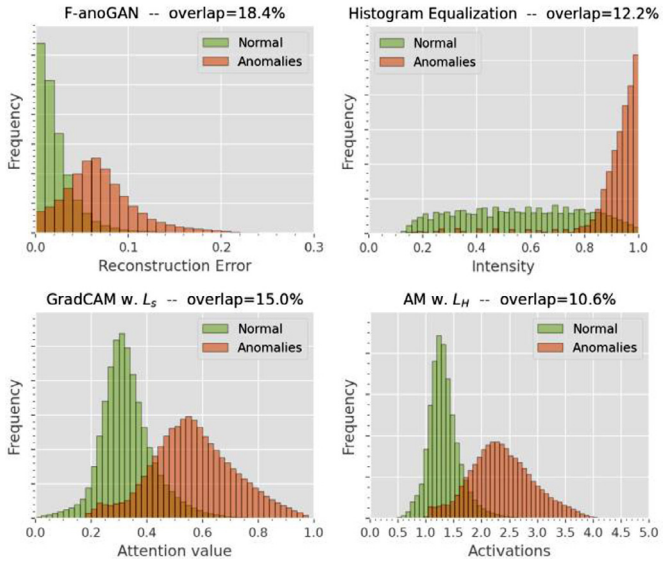

Fig. 6. Histogram analysis on the overlap of normal and anomalous samples for the different proposed methods and baselines (on the whole BRATS dataset).

Table 3
Ablation study on threshold values from normal images. p*X* indicates the average percentile used on the training set (normal images) to compute the segmentation threshold. OP indicates the operative point from area under precision-recall curve, using all validation dataset, which contains anomalous images. The metric presented is the dataset-level DICE.

	OP	th=0.5	p85	p90	p95	p98
F-anoGAN	0.525	-	0.310	0.390	0.505	0.488
HistEq	0.690	-	0.298	0.404	0.624	0.620
GradCAMCons w. $\mathcal{L}_S$	0.693	0.583	0.512	0.611	0.663	0.587
AMCons w. $\mathcal{L}_H$	0.743	-	0.189	0.201	0.265	0.720

Using statistics from normal domain for anomaly localization threshold A common practice on unsupervised anomaly segmentation is to use anomalous images to define the threshold to obtain the final segmentation masks. In particular, these methods look at the AUPRC on the anomalous images, which is then used to compute the optimal threshold value. We refer to this technique in our experiments as OP (Operative Point). To alleviate the need of anomalous samples during the validation stage, several methods (Baur et al., 2019) have discussed the possibility of using a given percentile from the normal images (i.e., no anomalies) distribution to set the threshold. Motivated by this, an ablation study on the percentile value is presented in Table 3 for our proposed formulations and the best performing baselines. First, we can observe that under the OP strategy (i.e., accessing to anomalous images to identify the optimal threshold), both of our models bring substantial improvements over the state-of-the-art on residual-based approaches, ranging from 14% to 22%. If we resort to the percentiles instead, the performance improvements observed are very similar to the OP scenario, with our models outperforming F-anoGAN by a large margin. Nevertheless, we observed that the best results are obtained with different percentile values. While F-anoGAN and AMCons w. $\mathcal{L}_H$ yields the best performance using the 98% percentile, GradCAMCons w. $\mathcal{L}_S$ follows previous observations in Baur et al. (2019), performing better using the 95% percentile.

This suggests that, even though not used directly, anomalous images are still required to find the optimal threshold value. However, the proposed method GradCAMCons shows special properties that suggest that they can achieve large performance gains without having access to anomalous images to define the threshold, unlike prior works. In particular, our GradCAM-based formulation restricts the attention values to $[0, 1]$, which allows to set a typical threshold to 0.5, with still large performance gains (+7%) compared to the baselines. Nevertheless, we can observe that if we resort to the percentile strategy, our method based on maximizing the entropy of the attention maps (i.e., AMCons) is very sensitive to the selected value.

Number of slices to generate the pseudo-volumes In our experiments, we followed the standard literature (Baur et al., 2021) to generate the pseudo-labels for validation and testing. Nevertheless, we concede that this scenario is unrealistic, as the appropriate number of slices used from the MRI scans in unsupervised anomaly detection should be unknown. We now explore the impact of including more slices in these pseudo-volumes, which increase the variability of normal samples. For instance, it is well-known that the target regions in slices farther from the center are incrementally smaller. In this line, we hypothesize that the dimension of the VAE latent space and the importance of the KL regularization may be a determining factors in absorbing this increased variability. Regarding the latent space, the appropriate $\mathbf{z}$ dimension is unclear in the literature. For instance, Baur et al. (2021) uses $\mathbf{z} = 128$, while Baur et al. (2019) uses $\mathbf{z} = 64$, and we obtained better results using $\mathbf{z} = 32$. To validate the proposed experimental setting and latent space dimension, we now present results using increasing number of slices around the axial midline $N = \{10, 20, 40\}$, and two different latent space dimensions $\mathbf{z} = \{32, 128\}$ for both a standard VAE and our proposed models, in Fig. 7a. We can observe that despite the gap between the baselines and the attention based methods is reduced as the number of slides is increased, this difference is still significant, and the relative performance drop is similar for all methods. Finally, we can observe that an increasing on $\mathbf{z}$ dimension (solid *versus* dotted lines in Fig. 7a) does not produce gains in performance in any case. Note that the model hyperparameters used are optimized for $z = 32$, and $N = 10$, which also could produce some underestimation of the proposed model performance when N increases. In the following, we study the performance of the proposed AMCons method using different β values ($\beta = \{1, 10\}$) in the KL term of Eq. 1 across different number of slices, whose results are presented in Fig. 7b. We can observe that, by decreasing the value of β as the number of employed slices increases, we can alleviate the performance degradation observed with a fixed β. Since the KL regularization directly affects the capacity of the VAE for learning different samples, the optimization of its balancing term when increasing the domain of samples used seems necessary. The similar behaviour between the pro-

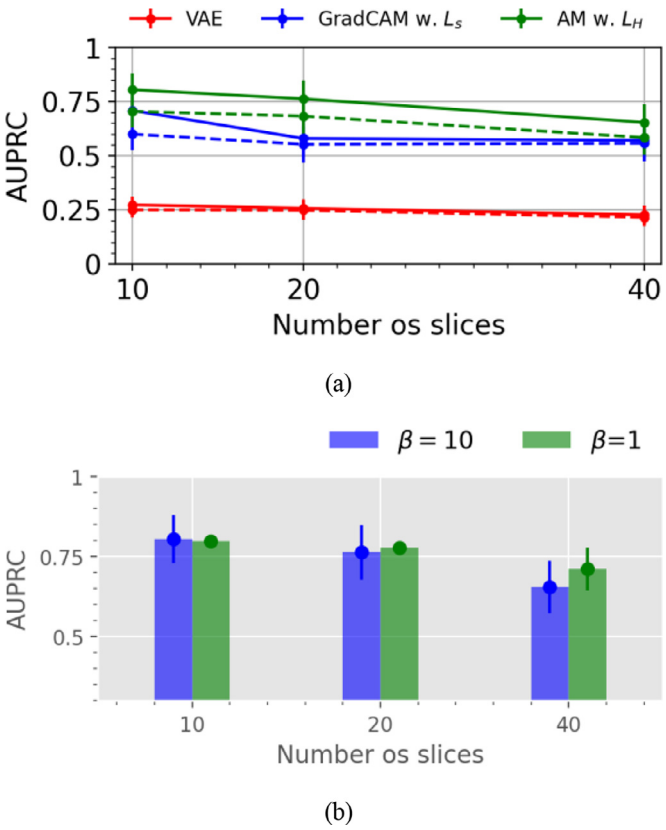

Fig. 7. Ablation study on the effect of increasing the number of axial slices around the center used from MR brain volumes. (a) Study of latent space dimension for the proposed models and an standard VAE. Solid lines indicate $\mathbf{z} = 32$, and dashed lines denote $\mathbf{z} = 128$. (b) Study of the KL component importance (β term) using the proposed AMCons method.

Table 4
Comparison to prior literature on Physionet-ICH dataset, and previous works on ICH segmentation. Results derived from the proposed methods are depicted in gray, and best results are indicated in bold. The values in parentheses indicate the standard deviation over the three training repetitions.

Method	AUROC	AUPRC	⌈DICE⌉	⌈IoU⌉	DICE ($\mu \pm \sigma$)
Other works					
Karkkainen et al. (2021) (Unsupervised)*	-	-	-	-	0.197 ± 0.222
Hssayeni et al. (2020) (Supervised)	-	-	-	-	0.315 ± 0.211
Physionet-ICH dataset					
CAVGA (Venkataramanan et al., 2020)	0.919(0.004)	0.061(0.003)	0.094(0.005)	0.062(0.004)	0.053(0.004)±0.161(0.002)
Grad-CAM$_D$ VAE (Liu et al., 2020)	0.955(0.003)	0.157(0.009)	0.275(0.011)	0.159(0.005)	0.178(0.005)±0.175(0.003)
Bayesian AE (Nick Pawlowski, 2018)	0.961(0.001)	0.188(0.006)	0.309(0.009)	0.183(0.007)	0.242(0.008)±0.181(0.003)
VAE (Baur et al., 2019; Zimmerer et al., 2020)	0.962(0.000)	0.167(0.005)	0.319(0.002)	0.190(0.002)	0.245(0.004)±0.192(0.003)
AnoVAEGAN (Baur et al., 2019)	0.961(0.000)	0.167(0.003)	0.313(0.006)	0.185(0.004)	0.239(0.006)±0.192(0.002)
Bayesian VAE (Nick Pawlowski, 2018)	0.964(0.000)	0.178(0.010)	0.323(0.007)	0.193(0.005)	0.248(0.008)±0.191(0.004)
Context VAE (Zimmerer et al., 2019)	0.963(0.002)	0.170(0.013)	0.321(0.023)	0.191(0.016)	0.243(0.014)±0.191(0.009)
Restoration VAE (Chen et al., 2020)	0.962(0.001)	0.183(0.005)	0.327(0.002)	0.187(0.001)	0.233(0.004)±0.189(0.003)
Context AE (Zimmerer et al., 2019)	0.962(0.001)	0.195(0.005)	0.359(0.010)	0.219(0.007)	0.276(0.004)±0.198(0.004)
F-anoGAN (Schlegl et al., 2019)	0.961(0.000)	0.173(0.007)	0.343(0.007)	0.207(0.005)	0.268(0.007)±0.191(0.005)
AE	0.961(0.001)	0.176(0.006)	0.344(0.007)	0.208(0.006)	0.266(0.002)±0.202(0.005)
GradCAMCons w. $\mathcal{L}_S$ (L2 penalty)	0.967(0.009)	0.261(0.013)	0.361(0.067)	0.231(0.029)	0.276(0.046)±0.243(0.029)
HistEq (Meissen et al., 2021)	0.963(0.000)	0.313(0.000)	0.385(0.000)	0.239(0.000)	**0.348**(0.000)±**0.213**(0.000)
GradCAMCons w. $\mathcal{L}_S$ (Log Barrier)	0.970(0.008)	0.295(0.073)	0.401(0.044)	0.251(0.049)	0.286(0.076)±0.233(0.039)
AMCons w. $\mathcal{L}_H$	**0.971**(0.006)	**0.420**(0.068)	**0.522**(0.046)	**0.354**(0.043)	0.319(0.054)±0.266(0.011)

* Results reported on a different (private) dataset.

posed method and baselines suggest that this could be a limitation of self-training features based on VAEs, which struggle to encode heterogeneous sample information.

5.3. Generalization to other datasets

In order to empirically demonstrate the generalization properties of the proposed methodology, we evaluate its performance on a different dataset for brain lesion detection. Concretely, as previously described, we resort to Physionet-ICH dataset for non-contrast CT on ICH localization. Implementation details are analogous as the ones used on the BraTS dataset, although we decreased the learning rate to $1e-5$, and we set a larger latent dimension, i.e. $\mathbf{z} \in \mathbb{R}^{128}$, along all baselines and methods to favour model convergence. Obtained results for anomaly localization are reported in Table 4. Even though there exist slight differences in the comparison between residual methods in the literature compared to the results obtained on BraTS dataset (i.e. the simple AE outperforms variations approaches), the proposed attention-based anomaly localization methods still achieve remarkable results. Again, the AMCons configuration yields the best performance, and it reaches improvements of nearly ∼25% and ∼18% in terms of AUPRC and DICE, respectively, compared to previous literature. The observed results suggest that the proposed methodology is able to generalize to other unsupervised brain lesion segmentation challenges, even using different imaging modalities. It should be noted, however, that the absolute results in terms of segmentation are lower than those obtained in BraTS. Among other reasons, this may be due to the greater heterogeneity observed in the ICH dataset, the lower degree of standardization and size of the database used, and the small size of ICH lesions, which penalizes metrics such as DICE. Nevertheless, the values obtained are in line with the scarce previous literature on ICH segmentation, as reflected in Table 4. Indeed, the obtained results are at par with previous works using a fully supervised learning approach Hssayeni et al. (2020), which shows the difficulty of the task.

5.4. Qualitative evaluation

Visual results of the proposed and existing methods for both datasets are depicted in Fig. 8. We can observe that our approach identifies as anomalous more complete regions of the lesions, whereas existing methods are prone to produce a significant amount of false positives (*first, third and seventh* rows) and fail to discover many abnormal pixels (*third row*). These visual results are in line with the quantitative validation performed in previous sections. However, there is a known problem about segmenting only hyperintense regions in the state-of-the-art methods of unsupervised anomaly localization of brain lesions (Meissen et al., 2021). Although the proposed method still suffers from this limitation (*fourth row, red arrow*), the positive results regarding true negative segmentation obtained in some normal, hyperintense tissue (*second row, green arrow*) suggest an improvement in relation to this problem.

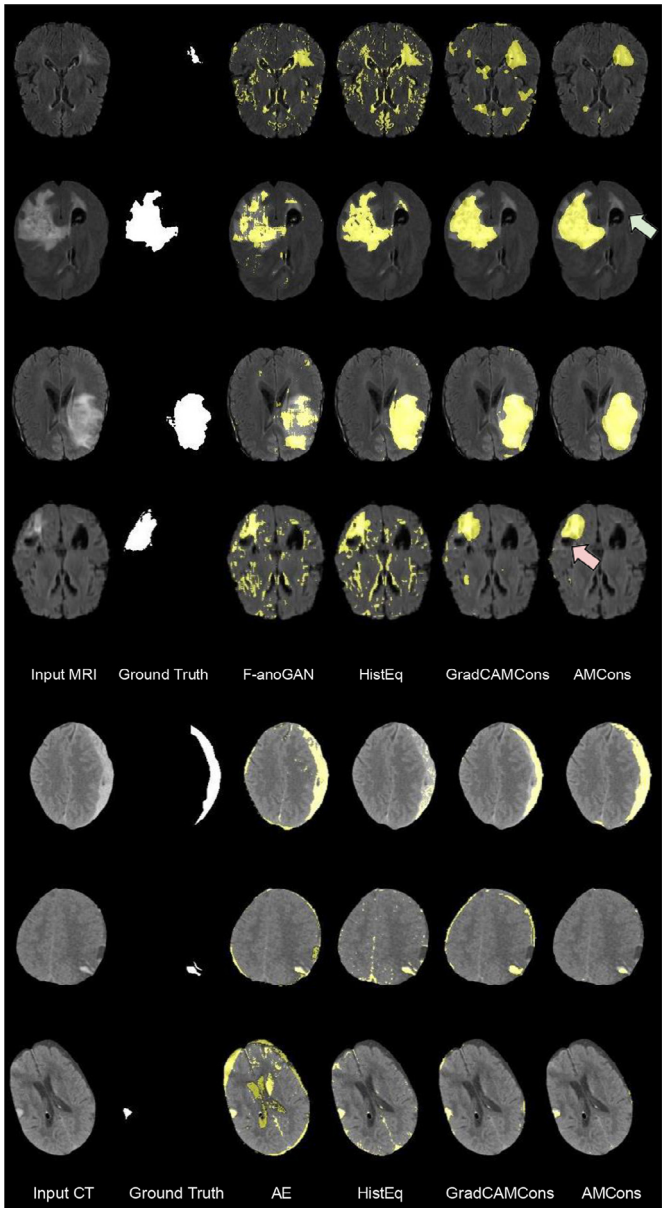

Fig. 8. Qualitative evaluation of the proposed and existing high-performing methods for anomaly localization on BraTS MRI flair volumes (top) and on Physionet-ICH non-contrast CT images (bottom). A failure case is depicted with the red arrow (*fourth column*).

6. Discussion

Despite the recent advances of unsupervised anomaly segmentation in medical problems, existing literature still provides limited performance, with most methods yielding suboptimal results in popular segmentation benchmarks. In this work, we have presented a novel approach that substantially differs from prior literature in several aspects.

First, we resort to generated attention maps to identify anomalous regions, which contrasts with most existing works that rely on the pixel-wise reconstruction error. Second, our formulation integrates a size-constrained loss that enforces the attention maps to cover the whole image in normal images. This differs from very recent works Venkataramanan et al. (2020), as we tackle this problem by imposing inequality constraints on the whole target attention maps. Another important difference lies on the manner the constrained problem is addressed. While Venkataramanan et al. (2020) leverages a L2 penalty function, we resort to an extension of standard log-barrier methods, which overcome the well-known limitations of penalty-based methods. Quantitative results demonstrate that this model significantly outperforms prior literature on unsupervised lesion segmentation.

A drawback of the log-barrier based formulation is that it requires to find the optimal value for several hyperparameters. Motivated by this, we have proposed an alternative model, which integrates a regularization term that maximizes the Shannon entropy on the generated attention maps. This new formulation only

adds the entropy balancing term $\mathcal{L}_H$, which reduces the complexity compared to the constrained problem in eq. 5. Furthermore, as reported in the results, the maximum-entropy model yields better performance than the size regularizer formulation. Note, in addition, that the alternative entropy-based model better separates the intensity distributions between normal and abnormal tissue. This allows us to employ a higher percentile value to obtain the final anomalous regions, with a substantial performance improvement compared to previous methods. Thus, based on the reported empirical validation, the proposed models represent a novel state-of-the-art for unsupervised anomaly segmentation.

We believe that there exist potential research directions to further improve the performance of unsupervised segmentation methods. For example, brain images are typically acquired along multiple modalities. Learning how to combine multiple modalities in the scenario of anomalous regions detection might indeed enhance the learned representation by the VAE, ultimately resulting in better identification of abnormal pixels. In addition, unsupervised segmentation methods have been only evaluated from a discriminative perspective. Nevertheless, assessing their performances in terms of the quality of the uncertainty estimates, i.e., calibration, might give a better overview of the quality of a segmentation model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

J. Silva-Rodríguez work was supported by the Spanish Government under FPI Grant PRE2018-083443. The DGX-A100 used in this work was partially funded by Generalitat Valenciana / European Union through the European Regional Development Fund (ERDF) of the Valencian Community (IDIFEDER/2020/030).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102526.

References

Abati, D., Porrello, A., Calderara, S., Cucchiara, R., 2019. Latent space autoregression for novelty detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Andermatt, S., Horváth, A., Pezold, S., Cattin, P., 2019. Pathology segmentation using distributional differences to images of healthy origin. Medical Image Computing and Computer Assisted Intervention (MICCAI) - Brainlesion Workshop.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4 (September), 1–13.

Bakas, S., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ayed, I.B., 2021. Constrained domain adaptation for image segmentation. IEEE Trans Med Imaging 40 (7), 1875–1887.

Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S., 2021. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Med. Image Anal. 69 (8), 1–16.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N., 2020. SteGANomaly: inhibiting CycleGAN Steganography for unsupervised anomaly detection in brain MRI. Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer International Publishing.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. Medical Image Computing and Computer Assisted Intervention (MICCAI).

Bergmann, P., Fauser, M., Sattlegger, D., Steger, C., 2020. Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4183–4192.

Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C., 2019. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP).

Boyd, S., Boyd, S.P., Vandenberghe, L., 2004. Convex Optimization. Cambridge university press.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. Medical Imaging with Deep Learning (MIDL).

Chen, X., You, S., Tezcan, K.C., Konukoglu, E., 2020. Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. 64.

Dehaene, D., Frigo, O., Combrexelle, S., Eline, P., 2020. Iterative energy-based projection on a normal data manifold for anomaly localization. In: Proceedings of the International Conference on Learning Representations (ICLR).

Fiacco, A.V., McCormick, G.P., 1990. Nonlinear Programming: Sequential Unconstrained Minimization Techniques. SIAM.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-k., Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. Advances in Neural Information Processing Systems (NeurIPS), Vol. 63.

He, F.S., Liu, Y., Schwing, A.G., Peng, J., 2017. Learning to play in a day: faster deep reinforcement learning by optimality tightening. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–13.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).

Hssayeni, M., 2020. Computed tomography images for intracranial hemorrhage detection and segmentation. PhysioNet.

Hssayeni, M.D., Croock, M.S., Salman, A.D., Al-Khafaji, H.F., Yahya, Z.A., Ghoraani, B., 2020. Intracranial hemorrhage segmentation using a deep convolutional model. Data 5 (1), 1–18.

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. In: 35th International Conference on Machine Learning (ICML).

Jia, Z., Huang, X., Chang, E.I.C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging 36 (11), 2376–2388.

Kervadec, H., Dolz, J., Granger, E., Ben Ayed, I., 2019. Curriculum Semi-supervised Segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI).

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ben Ayed, I., 2019. Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. 54, 88–99.

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E., Ayed, I. B., 2019c. Constrained deep networks: lagrangian optimization via log-barrier extensions. arXiv preprint arXiv:1904.04205.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, (ICLR).

Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R.J., Camps, O., 2020. Towards visually explaining variational autoencoders. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Luenberger, D.G., 1973. Introduction to Linear and Nonlinear Programming. Addison-wesley Reading, MA.

Meissen, F., Georgios, K., Rueckert, D., 2021. Challenging current semi-supervised anomaly segmentation methods for brain MRI. MICCAI 2021 BrainLes Workshop.

Meissen, F., Weistler, B., Kaissis, G., Rueckert, D., 2022. On the pitfalls of using the residual error as anomaly score. Medical Image with Deep Learning (MIDL).

Menze, B., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Nguyen, B., Feldman, A., Bethapudi, S., Jennings, A., Willcocks, C.G., 2021. Unsupervised region-based anomaly detection in brain MRI with adversarial image inpainting. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1127–1131.

Nick Pawlowski, M.C.H.L., 2018. Unsupervised lesion detection in brain CT using Bayesian convolutional autoencoders. Medical Imaging with Deep Learning (MIDL).

Pathak, D., Krahenbuhl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1796–1804.

Peng, J., Kervadec, H., Dolz, J., Ben Ayed, I., Pedersoli, M., Desrosiers, C., 2020. Discretely-constrained deep network for weakly supervised segmentation. Neural Netw. 130, 297–308.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (ICLR).

Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N., 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), vol. 2.

Sabokrou, M., Pourreza, M., Fayyaz, M., Entezari, R., Fathy, M., Gall, J., Adeli, E., 2019. AVID: adversarial visual irregularity detection. In: Proceedings of the Asia Conference on Computer Vision (ACCV).

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI).

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad–CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 128 (2), 336–359.

Shi, Y., Yang, J., Qi, Z., 2021. Unsupervised anomaly segmentation via deep feature reconstruction. Neurocomputing 424, 9–22.

Silva-Rodríguez, J., Naranjo, V., Dolz, J., 2021. Looking at the whole picture: constrained unsupervised anomaly segmentation. In: British Machine Vision Conference (BMVC).

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J., 2020. An adversarial learning approach to medical image synthesis for lesion detection. IEEE J. Biomed. Health Inform. 24, 2303–2314.

Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A., 2020. Attention guided anomaly localization in images. In: Proceedings of the European COnference on Computer Vision (ECCV).

Zhang, Y., David, P., Gong, B., 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Zhou, Y., Li, Z., Bai, S., Chen, X., Han, M., Wang, C., Fishman, E., Yuille, A., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Zimmerer, D., Full, P.M., Isensee, F., Jager, P., Adler, T., Petersen, J., Kohler, G., Ross, T., Reinke, A., Kascenas, A., Jensen, B.S., O'Neil, A.Q., Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., Shvetsova, N., Fedulova, I., Dylov, D.V., Yu, B., Zhai, J., Hu, J., Si, R., Zhou, S., Wang, S., Li, X., Chen, X., Zhao, Y., Marimont, S.N., Tarroni, G., Saase, V., Maier-Hein, L., Maier-Hein, K., 2022. MOOD 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE Trans. Med. Imaging.

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K., 2020. Abstract: unsupervised anomaly localization using variational auto-encoders. Informatik Aktuell.

Zimmerer, D., Kohl, S., Petersen, J., Isensee, F., Maier-Hein, K., 2019. Context-encoding variational autoencoder for unsupervised anomaly detection. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track.