Looking at the whole picture: constrained unsupervised anomaly segmentation. Supplemental Material

Julio Silva-Rodríguez
jjsilva@upv.es
Institute of Transport and Territory
Universitat Politècnica de València
Valencia, Spain

Valery Naranjo
vnaranjo@dcom.upv.es
Institute of Research and Innovation in Bioengineering
Universitat Politècnica de València
Valencia, Spain

Jose Dolz
jose.dolz@etsmtl.ca
LIVIA Laboratory
École de Technologie Supérieure (ETS)
Montreal, Canada

1 Additional dataset details

A summary of the used dataset, with the corresponding training, validation and testing splits, after the pre-processing detailed in Section 4.1, is presented in Table 1.

Partition	Cases	Training Images
Training	271	268
Validation	26	–
Testing	25	–

Table 1: Dataset, partition and training images used.

2 Additional ablation studies

Model hyperparameters. To better understand the behaviour of the attention constrains in the proposed model, we resort to extensive ablation experiments to determine the optimal values of several model hyperparameters: the log-barrier t term, the size term p, the weights of the attention loss on the training, λ and, finally, the network depth used to compute the CAMs. Firstly, we empirically fix $\lambda = 10$ and use the first convolutional block output to compute CAMs, to evaluate the impact of our model with p values included in $\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$ and t values in $\{10, 15, 20, 25, 50\}$. These results are reported in Table 2. Please note that all the results reported on the ablation studies are obtained on the validation set.

We now validate the level depth from the encoder used to obtain the CAMs (i.e., network depth s in Section 3.2), with the best configuration from the previous ablation in Table 2.

t	Size (proportion) term p						
	0	0.05	0.10	0.15	0.20	0.25	0.30
10	0.614	0.408	0.662	0.504	0.601	0.623	0.500
15	0.575	0.546	0.498	0.614	0.638	0.599	0.641
20	0.682	0.664	0.646	0.641	**0.710**	0.640	0.610
25	0.536	0.606	0.575	0.545	0.679	0.671	0.680
50	0.476	0.606	0.636	0.685	0.539	0.657	0.607

Table 2: Ablation study on the impact of p and t in the proposed formulation, where dataset specific AUPRC results are presented. Bold highlights the best performing configuration.

Results are presented in Table 3, from which we can observe that maximizing the attention in early layers leads to better results than in deeper layers. This could be produced by the better spatial definition of early layers, and the benefits that the proposed constrain produces in its later layers, which receive information from the whole image.

	Conv1	Conv2	Conv3	Conv4
AUPRC	**0.710**	0.621	0.456	0.274
[DICE]	**0.661**	0.454	0.292	0.276

Table 3: Ablation study on network depth to compute CAMs. Dataset specific AUPRC is presented for each possible configuration. Best performance highlighted in bold.

Next, in Table 4 we study the optimal weight to balance the proposed attention loss, by evaluating the performance of our formulation across several λ values. The experiments presented on the main paper are obtained using the best configuration: $t = 20$, $p = 0.20$, $\lambda = 10$, with CAMs being obtained form the first convolutional block.

	λ				
	0.01	0.1	1	10	100
AUPRC	0.150	0.443	0.609	**0.710**	0.587
[DICE]	0.207	0.502	0.609	**0.661**	0.587

Table 4: Ablation study on the influence of attention expansion losses in relation to its relative weight, λ. Dataset specific AUPRC and DICE are presented for each validated value. Bold numbers indicate the best performance.

Number of slices to generate the pseudo-volumes. In our experiments, we followed the standard literature [7] to generate the pseudo-labels for validation and testing. Nevertheless, it is unclear in unsupervised anomaly detection of brain lesions the appropriate number of slices used from the MRI scans. We now explore the impact of including more slices in these pseudo-volumes, which increase the variability of normal samples. In this line, we hypotetize that the dimension of the VAE latent space may be a determining factor in absorbing this increased variability. The appropriate $\mathbf{z}$ dimension is also unclear in the literature. For instance, [9] uses $\mathbf{z} = 128$, while [7] uses $\mathbf{z} = 64$, and we obtained better results using $\mathbf{z} = 32$. To validate the proposed experimental setting and lattent 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

model, in Table 5. We can observe that despite the gap between the two method is reduced as the number of slides is increased, this difference is still significant. Finally, we can observe that an increasing on **z** dimension 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.

Method	**zdim**	**N slices**		
		10	20	40
Proposed	32	0.710	0.581	0.572
		0.661	0.577	0.576
	128	0.601	0.554	0.559
		0.599	0.562	0.556
VAE	32	0.275	0.259	0.229
		0.398	0.373	0.362
	128	0.252	0.250	0.217
		0.368	0.384	0.347

Table 5: Ablation study on number of axial slices around the center used from MR brain volumes, and lattent space dimension, for the proposed model and an standard VAE. We report the AUPRC (*top row*) and the best [DICE] (*bottom row*).

On the impact of the reconstruction losses. We evaluate the effect of including several well-known reconstruction losses in our formulation: SSIM [38] and L_2-norm. Table 6 reports the results from these experiments, where we can observe that, while BCE and SSIM reconstruction losses yield the best performances, integrating the L_2-norm loss in our formulation degrades the performance of the proposed model.

	BCE	**L2 norm**	**SSIM**
AUPRC	**0.710**	0.600	0.679
[DICE]	**0.661**	0.612	0.649

Table 6: Ablation study on the reconstruction losses for the proposed approach. Best results in bold.

Using statistics from normal domain for anomaly localization threshold As mentioned along the manuscript, a main limitation of unsupervised anomaly localization methods is the need of using anomalous images to set a threshold on the obtained heatmaps to locate anomalies. Several methods [7] have discussed the possibility of using a given percentile from the normal images (i.e., no anomalies) distribution to set the threshold. An ablation study on the percentile value is presented in Table 7 for our proposed model and the best performing baseline. Compared to the best baseline method in Table 1 of the main manuscript, i.e., F-anoGAN, our model substantially yields better performance. Nevertheless, we found that the best results are obtained on the percentile 95%, whereas [7] found the operative performance on the percentile 98%. This suggests that, even though not used directly, anomalous images are still required to find the optimal value.

Model parameters. In this section, we compare our formulation to existing approaches in terms of model complexity. Since previous residual-based methods require the generation

	OP	th=0.5	p85	p90	p95	p98
Proposed	0.661	0.579	0.498	0.602	0.657	0.597
F-anoGAN	0.525	–	0.310	0.390	0.505	0.488

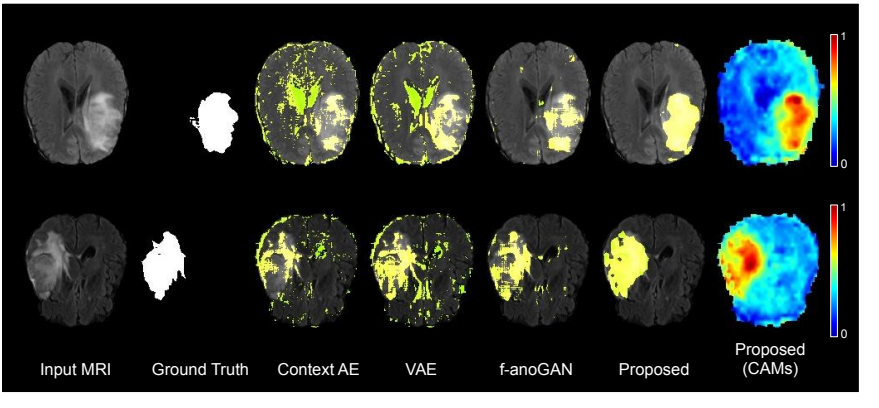
Table 7: Ablation study on threshold values from normal images. pX indicates the average percentile used on the training set (normal images) to compute the segmentation threshold. OP indicates the operative point from area under precision-recall curve, using all validation dataset, which contains anomalous images. The metric presented is the dataset-level DICE.

of normal counterparts from anomalous images, they typically integrate an additional discriminator to create more realistic images, and require to use the trained generative decoder during inference. On the other hand, our proposed formulation only requires the encoder part of the network to localize anomalies, which reduces the number of required parameters, as indicated in Table 8. On the other hand, as highlighted in previous works [20] the cost of adding a single constraint is negligible.

Method	**∼Parameters (millions)**	
	Train	Inference
Context VAE [40]	15.0	15.0
VAE [6, 41]	15.0	15.0
F-anoGAN [34]	17.8	15.0
Proposed	15.0	13.3

Table 8: Parameters of the proposed method and best performing baselines during both, training and inference stages.

Additional qualitative results. In the following Figure 1, we show complementary examples of the proposed method performance.



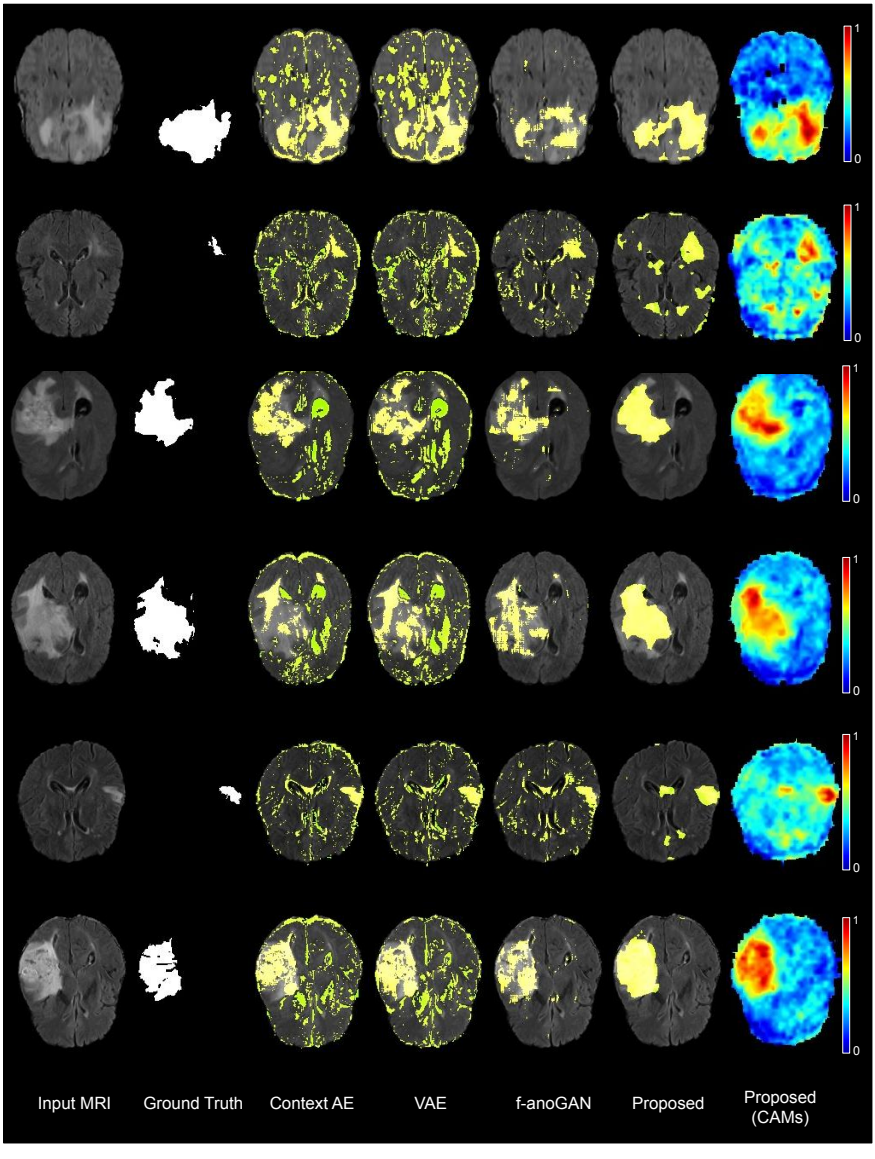


Figure 1: Qualitative evaluation of our method.