Cross-grained Contrastive Representation for Unsupervised Lesion Segmentation in Medical Images

Anonymous ICCV submission

Paper ID 12

Abstract

Automatic segmentation of lesions in medical images plays a crucial role in the quantitative assessment of disease progression. While supervised deep learning-based methods have been successful in numerous segmentation tasks, they rely on a large number of labelled images for training, which can be expensive and time-consuming to acquire. Although unsupervised learning shows potentials in addressing this challenge, the performance of current unsupervised algorithms is mostly unsatisfactory. To overcome this issue, we propose a new unsupervised framework for medical lesion segmentation using a novel cross-granularity contrastive (CGC) module. Our module contains coarse-grained and fine-grained discrimination paths that enable the network to capture the distinctions between lesions and normal tissues at different levels of context. We evaluate our method on two large public datasets of CT/MRI scans and demonstrate that our approach improves a Gaussian mixture model-based segmentation by up to 9%, which surpasses all other unsupervised segmentation methods by a large margin. Additionally, our module can also be integrated with other existing unsupervised segmentation methods to further enhance their performance. Therefore, our framework shows great potential for use in medical image applications with limited labelled data availability.

1. Introduction

Accurate segmentation of brain lesions plays an essential role in the quantitative assessment of disease progression, as well as pre-and post-operative treatment planning. As reading and annotating MRI/CT scans is a tedious and time-consuming process, there have been significant efforts in recent years to develop deep learning-based algorithms to mitigate this dilemma in clinical practice. However, most of these algorithms focus on supervised training, which requires an enormous amount of annotated datasets. Acquiring such data is highly challenging and even impractical for several reasons. First, labelling 3D MRI/CT volumes is time-consuming and requires specialized medical knowledge. Second, lesion regions are characterized by significant heterogeneity in texture, size, location, and pathological appearances. These issues make it challenging to obtain a comprehensive training dataset that covers all possible cases, potentially compromising the performance and generalization ability of the learned-based network.

Given these constraints, there has been a growing interest in exploring unsupervised strategies, such as unsupervised anomaly segmentation. Unlike supervised methods, unsupervised strategies require no labelled data for training or little labels for fine-tuning. Zimmerer et al [1] proposed Context VAE, an expansion of VAE [2]. By reassembling an input image with clipped patches, it compels the VAE encoder to embed more information. Chen et al [3] proposed Constrained VAE, which employs the encoder to map the recovered pixels to the same location in the latent space as the original. While several previous studies have been introduced to this research area, most of them are reconstruction-based approaches aiming to model healthy brain anatomy distribution. Although such generative models have promising potential for reconstructing medical images, they are not inherently suitable for semantic segmentation tasks as they do not receive explicit constraints and guides during training. To address this issue, a possible solution is leveraging the hidden patterns in images through contrastive learning [4]. However, these existing contrastive learning approaches often rely on pixel-level labels to fine-tune the model, which is difficult to obtain in practice.

In contrast to previous approaches, we propose a completely unsupervised framework to address the aforementioned concerns. Our approach is based on the observation that lesion regions exhibit anomalous contrast compared to healthy tissue, resulting in a semantic content of the foreground object that is inherently distinct from its background. We leverage this information by utilizing different granularity contrastive representations, enabling effective segmentation of lesions. The main contributions of this paper are summarized as follows:

- We propose a novel cross-granularity contrastive (CGC) module that contains coarse-grained and fine-grained discrimination paths. We demonstrate in the experiments that this module can efficiently incorporate context from different levels, thus enhancing representation learning.
- To initiate our segmentation, we employ a Gaussian mixture model (GMM) to generate a foreground probability map, thus keeping our entire segmentation framework unsupervised. However, we also show our CGC module is not restricted to the GMM method and can be used in combination with other unsupervised segmentation methods and further refine their results.
- To mitigate the noise issue during training, we investigate the correlation between performance and the temperature-calibrated logit map, which has rarely been studied in medical image segmentation.
- Our proposed unsupervised framework shows superior performance in lesion segmentation on both MRI and CT images compared to state-of-the-art algorithms, indicating its potential for application to other medical modalities.

2. Related Work

2.1. Unsupervised Anomaly Segmentation

Unsupervised anomaly segmentation aims to identify abnormal voxels in images from test sets. This approach is particularly appealing due to its potential to alleviate challenges in real clinical scenarios where vast labeled data is difficult to obtain or when encountering infrequent diseases. Currently, the majority of prior unsupervised techniques fall into the category of reconstruction-based methods. These techniques rely on trained models that can generate healthy counterparts to input data and then use pixel-wise residuals between the model's generation and the input to detect anomalies and lesions. For example, Variational Auto-encoders (VAE) [1, 5, 6, 7, 8], vector quantized variational autoencoders (VQ-VAE) [9, 10, 11], and Generative Adversarial Networks (GAN) [12, 13] constitute the most common methods in this community. Most these methods attempt to model normal distribution of healthy in a low-dimensional latent space and are constrained by the input through reconstruction loss. They presume that the reconstruction of unseen anomalous regions should be inaccurate, hence yielding large residuals that can be used to localize or segment anomalies in images. Silva-Rodríguez el al [14] used inequality constraints and an alternative regularizer to force the attention map to be activated and maximize its Shannon entropy for enhancing the performance of VAEs segmentation pipeline. Pinaya et al [9] proposed an ensemble of autoregressive transformers combined with a VQ-VAE.

Different from detection and localization tasks, semantic segmentation usually requires that the models understand the high-level information existing in the image context. Unfortunately, the above-mentioned reconstruct-based methods miss the understanding of anomalous texture and lack the ability to discriminate differences between abnormal tissue and normal tissue. Moreover, these methods may drop dramatically in performance when they meet imperfect reconstructions [15], raising concerns about their robustness.

2.2. Contrastive learning

Contrastive learning has emerged as a powerful self-supervised learning technique in the domain of computer vision. The main principle behind contrastive learning is to encourage representations of similar samples to be closer in the latent space while pushing representations of dissimilar samples further apart. By doing so, the model can learn meaningful and discriminative features without the need for explicit labels. In two representative works, SimCLR [4] and MoCo [16] conduct two training strategies that produce SOTA results. SimCLR investigates the usage of in-batch samples for negative sampling, whereas MoCo uses a dictionary as a queue to store negative samples for training. There are many efforts to deploy contrastive learning in medical image analysis, including MRI, CT, and PET [17, 18, 19, 20]. Note that most of these contrastive-based methods generate positive samples using diverse augmentations, making it challenging to explicitly capture the differences between lesions and normal tissues. In contrast to previous work, we construct positive and negative samples mainly by considering the pathological features of lesions directly.

As a consequence, these limitations prompt the need for more robust and effective approaches in the domain of unsupervised anomaly segmentation. To address these concerns in existing methods, we introduce the CGC module to address such issues, which is further detailed in Sec. 3.3.

3. Methods

3.1. Architecture

An overview of our proposed unsupervised lesion segmentation model is shown in Fig. 1. Based on the observation that an intensity distribution discrepancy inherently exists between lesion and normal tissue, we aim to exploit this relevance in MRI/CT images themselves in an unsupervised manner. Concretely, let $X = \{X_1, X_2, \cdots, X_N\}$ be a set of N images from one batch, where $X_i \in \mathbb{R}^{C \times H \times W \times L}$. The encoder $E^c(\cdot)$ codes X into high-dimension feature maps z in the latent content space C at the bottleneck of

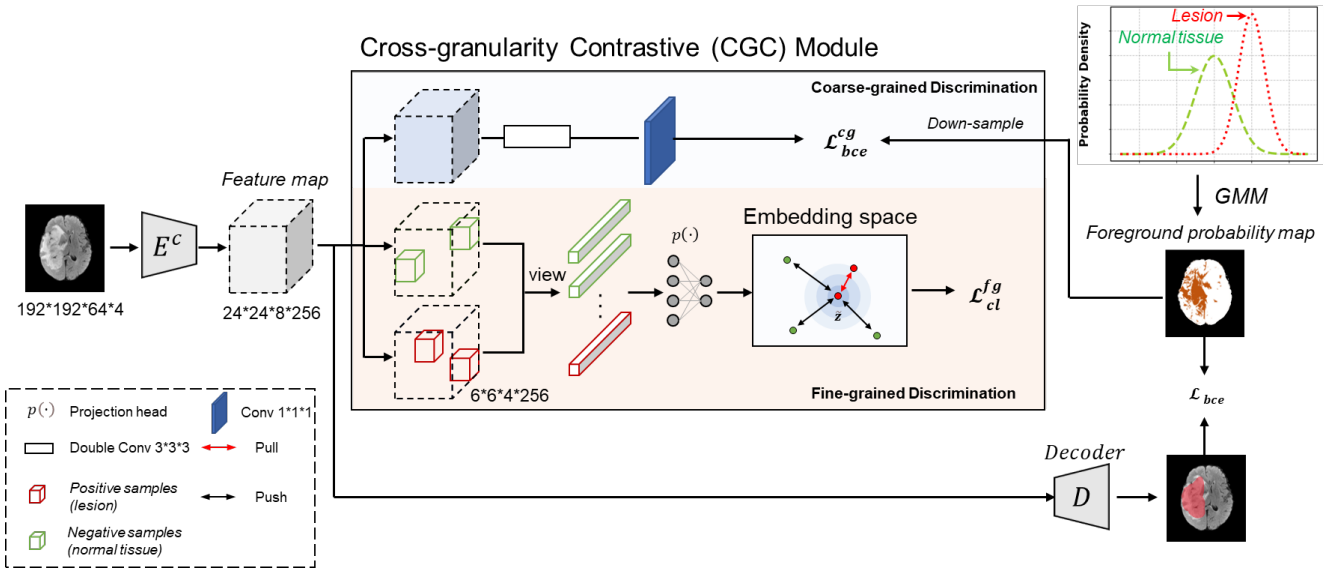

Figure 1. Proposed cross-granularity contrastive module for our segmentation framework that contains coarse-grained (light blue box) and fine-grained discrimination (orange box) paths. See text for a detailed explanation for all components.

the network, in which $z_i \in \mathbb{R}^{c\times h\times w\times l}$. Using the extracted feature maps as a basis, the CGC module leverages the spatial content in different granularity explicitly to promote the capacity of network's representations. The coarse-grained path utilizes global-level aggregation features to enable the network to identify the location of the lesion, while the fine-grained path addresses context divergence among various mini-patches. Then, the enhanced feature maps are upsampled via decoder $D(\cdot)$ to original resolution and supervised with pseudo-label $\hat{Y}$ generated by Gaussian Mixture Model (GMM), which can be formulated as follows:

$$\mathcal{L}_{bce} = -\sum_{k=1}^{K}(\hat{y}_i)_j^k \log(D(E^c(x_i)))_j^k \tag{1}$$

where $\hat{y}_i$ is the foreground probability of pixel x_i, $D(\cdot)_j^k$ denote the probability prediction of voxel j for class k.

3.2. Foreground-background Determination

To make the network aware of the foreground (lesion) vs background (normal tissue), we utilise a GMM probability model based on the idea of heterogeneity among foreground-background contrasts, which locates teach class region hypothesized by Multidimensional Gaussian distribution and can be formulated as follows:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)} \tag{2}$$

$$p_M(x) = \Sigma_{j=1}^{c}\alpha_j \cdot (x|\mu_j, \Sigma_j), \Sigma_j^c \alpha_j = 1 \tag{3}$$

where formula (2) is the multi-dimensional Gaussian distribution, Σ is the covariance matrix, μ is the mean vector. The Gaussian mixture distribution is represented by formula (3), where α_j stands for the mixture coefficient and the J_{th} Gaussian distribution probability. c is the number of Gaussian components. The variables in equation (4) are resolved using the maximum likelihood approach:

$$\{\alpha_j, \mu_j, \Sigma_j\} = argmax_{\alpha,\mu,\Sigma} \ln(\Pi_i^n p_M(x_i)) \tag{4}$$

In order to optimize parameters, the Expectation-Maximization (EM) technique is typically utilized because equation (4) contains hidden variables. Following the discovery of the Gaussian distribution, the elements are divided based on the posterior probability corresponding to the prototype, that is:

$$\lambda_i = \arg\max_{j\in\{1,2,\ldots,k\}} \frac{\alpha_j \cdot (x_i|\mu_j, \Sigma_j)}{\Sigma_{l=1}^{k}\alpha_l \cdot (x_i|\mu_l, \Sigma_l)} \tag{5}$$

where λ_i represents the posterior probability that the voxel x_i belongs to the j-th Gaussian component, where k is the total number of components in the Gaussian mixture.

By utilizing this Gaussian Mixture Model, we can leverage the contrast diversity among lesion and normal tissue, thereby providing the explicit guidances for our proposed two contrastive learning strategies.

3.3. Lesion-Normal Tissue Discrimination

3.3.1 Coarse-grained Discrimination.

We deem that embedding feature $z = E^c(\cdot)$ followed by a shallow multilayer perceptron (MLP) projection head ought to learn different semantic information and thus have different representations in the latent feature space. To

this end, we supervise this via cross entropy loss, named as $\mathcal{L}_{bce}^{cg}$, with down-sampled foreground probability map $p(x_i)$. Specifically, we use the residual block to enrich global content ulteriorly and followed by $f^{1\times1\times1}$ convolution. The formulation of $\mathcal{L}_{bce}^{cg}$ is similar to the formula 1, so we omit it here to avoid redundancy.

Suppose the network is able to differentiate the foreground and background thus assigning corresponding probability p_i of each pixel in the map and finally nurture disentangle the feature map z_i into lesion regions $r_i^{lesion} = p_i \otimes z_i$ and normal tissue regions $r_i^{normal} = (1-p_i)\otimes z_i$ under global texture during the training process, respectively.

3.3.2 Fine-grained Discrimination.

To make the semantic feature disentanglement more precisely, we further split z into internal mini-patch $\widetilde{z} \in \Omega$ for fine-grained discrimination, which focuses more on local texture. Concretely, we deem positive pair $\Omega^+ = \{\widetilde{z}_i(\widetilde{x}_i) | \forall C(\widetilde{x}_i) \in C(r_i^{lesion})\}$, whereas the negative ones $\Omega^- = \Omega \backslash \Omega^+$, thus exploit representations in contrastive manner. Mathematically, we have:

$$\mathcal{L}_{cl}^{fg} = \sum_{i=1}^{N} \frac{-1}{|\Omega^+|} \sum_{x^+ \in \Omega^+} \log \frac{\exp\left(CL^+/\tau\right)}{\sum_{x_i \in \Omega'} \exp\left(CL/\tau\right)}. \quad (6)$$

where $CL^+ = sim(\widetilde{z}_{x_i}, \widetilde{z}_{x^+})$, $CL = sim(\widetilde{z}_{x_i}, \widetilde{z}_x)$, $x^+ \in \Omega^+$ and $\Omega^{'} = \Omega/\{x_i\}$. τ is the temperature scaling parameter. $sim(\cdot,\cdot)$ is a pairwise similarity function that uses cosine distance to determine how similar two vectors are in the latent space:

$$sim(x, y) = \frac{xy^{\mathrm{T}}}{\|x\| + \|y\|} \quad (7)$$

Besides, considering the fact that a foreground probability map P_i cannot guarantee that lesions are completely included, we select a relatively high temperature τ to make logits-softmax values more smooth. The overall objective loss function can be written as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{bce} + \lambda_{cg}\mathcal{L}_{bce}^{cg} + \lambda_{fg}\mathcal{L}_{bce}^{fg} \quad (8)$$

4. Experimental setting

4.1. Dataset

We validate our method on two public datasets: the 2018 Multimodal Brain Tumor Segmentation Challenge (BraTS) dataset [25] and CQ500 dataset [26]. The BraTS dataset consists of 285 annotated MRI subjects with gliomas. Each subject has four aligned modalities, T1, T1ce, T2, and FLAIR. The CQ500 [26] comprises 491 CT scans with clinical radiology reports. 61 scans with intracranial hemorrhages (ICH) diagnosed by three senior radiologists were used in our study as CQ500 dataset. The ground truths of lesion regions are annotated by two senior radiologists.

4.2. Evaluation Metrics

We evaluate the performance of unsupervised brain lesion segmentation at the level of individual voxels, where the consideration of class imbalance becomes crucial since anomalous voxels are typically less common than normal voxels. Hence, we introduce several metrics widely used here in medical imaging analysis, including sensitivity (SEN) and Dice, which can be formulated as follows:

$$\text{SEN} = \frac{TP}{TP + FN}; \quad (9)$$

$$\text{Dice} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (10)$$

where TP, FP, and FN stand for true positive, false positive, and false negative, respectively.

4.3. Experiment Details

For both datasets, we conduct 5-fold cross-validation and randomly split 80% scans as training sets and 20% scans as testing sets at subject-level. In the preprocessing process, we resampled scans from CQ500 to $1 \times 1 \times 1mm$ resolution. We removed the skull by clustering and morphology methods. The scans from the BraTS dataset have been skull-stripped already with the shape $240 \times 240 \times 155$. We randomly cropped patches to $X \in \mathbb{R}^{192\times192\times64}$ from two datasets for training models, respectively.

The foreground probability map P was calculated by GMM, and then the models were trained using images without real ground truth. It should be noted that the network weights training processes were detached from the calculation of P to prevent them from decelerating training speed. The images in one batch were sent to content encoder E^c. For dataset with multi-modalities, we concatenate all modalities into one channel, following the standard literature [27]. After residual blocks and 3 times downsampling, we got embedding features $z \in \mathbb{R}^{24\times24\times8}$ at the bottleneck of the network.

For all training procedures, we used the Adam optimizer with a learning rate of 0.0001, and set the size of the batch to 2. The proposed framework was deployed in the Pytorch library and trained on two NVIDIA A6000 GPUs with 400 epochs. Empirically, we set $\lambda_{cg}, \lambda_{fg} = 1$, respectively. The code will be released to encourage more efforts toward developing unsupervised lesion segmentation algorithms.

4.4. Probability Map Generation

For the BraTS dataset, we used T2-FLAIR scans to generate foreground probability maps, denoted as P, through a modified clustering pipeline called AUCseg [28]. The clusters with higher cluster center values in T2-FLAIR modality

Table 1. Comparison of brain tumor segmentation performance on BraTS and CQ500 dataset.

Methods	BraTS dataset		CQ500 dataset	
	Dice	Sensitivity	Dice	Sensitivity
Full supervision	0.9134± 0.1026	0.9136±0.1150	0.7962±0.1523	0.7840±0.2138
AE	0.3543±0.2462	0.4542±0.2401	0.3381±0.2369	0.5142±0.2616
Context VAE [1]	0.4261±0.1874	0.4371±0.2596	0.3969±0.2446	0.5348±0.2212
GMVAE [21]	0.4418±0.1726	0.5374±0.2127	0.4147±0.2184	0.5362±0.2449
f-AnoGAN [22]	0.4835±0.1675	0.5332±0.2446	0.4024±0.2172	0.5528±0.1882
Bayesian VAE [2]	0.5348±0.1618	0.5575±0.2375	0.4391±0.2474	0.5446±0.2076
AnoVAEGAN [23]	0.5184±0.1560	0.5737±0.2098	0.4467±0.2286	0.5649±0.1844
AMCons [14]	0.7362±0.1642	0.7684±0.2084	0.4741±0.2310	0.4588±0.2469
Mumford-Shah [24]	0.7156±0.1881	0.7063±0.2157	0.5206±0.1937	0.5087±0.2336
Ours (P from [24])	0.7743±0.1365	0.7576±0.1974	0.5569±0.1861	0.5348±0.1875
GMM w/ threshold	0.7585±0.2091	0.7965±0.1892	0.6490±0.1975	0.5535±0.1485
Ours w/o CGC	0.7929±0.1877	0.8013±0.1815	0.6625±0.1945	0.6427±0.2013
Ours (P from GMM)	**0.8405±0.1323**	**0.8178±0.1756**	**0.6993±0.1755**	**0.6768±0.1825**

Table 2. Ablation study for each module and mini-path size on the BraTS dataset.

Methods	BraTS dataset	
	Dice	ASSD [mm]
Full supervision (backbone)	0.9134± 0.1028	1.1581±1.2894
Full supervision (nnUNet)	0.9161± 0.0873	1.0257±1.1316
P w/ threshold	0.7585±0.2091	2.1472±1.9426
Backbone + P	0.7929±0.1877	1.2330±1.1821
Backbone + P + Global granularity	0.8164±0.1631	1.3392±1.6271
Backbone + P + Local granularity	0.8248±0.1548	1.2513±1.2484
Ours ($\widetilde{z} : \mathbb{R}^{12\times12\times4}$)	0.8392±0.1389	1.2367±1.4205
Ours ($\widetilde{z} : \mathbb{R}^{6\times6\times4}$)	**0.8405±0.1323**	**1.1756±1.1439**
Ours ($\widetilde{z} : \mathbb{R}^{3\times3\times4}$)	0.8388±0.1410	1.1943±1.4263

are signed as the foreground in GMM due to the pathological features of lesions, which typically exhibit high intensities/signals in T2-FLAIR modality. In the GMM clustering, clusters with higher cluster center values in the T2-FLAIR modality were identified and considered as the foreground. This choice was guided by the pathological features of lesions, which tend to exhibit high intensities or signals in the T2-FLAIR modality. As such, these higher-intensity clusters were assumed to correspond to the regions of interest, i.e., the lesions, and were therefore labeled as the foreground in the GMM-based probability map. Similarly, we can also acquire P from the CQ500 dataset.

Although the primary ablation experiments in this work are conducted on the BraTS dataset, we employ the CQ500 dataset to showcase the generalization capabilities of our proposed method across diverse brain lesions and imaging modalities.

5. Results and Discussion

5.1. Evaluation on BraTS and CQ500 Dataset

We first compare our method with existing state-of-the-art (SOTA) deep learning based methods for unsupervised anomaly segmentation following the successful practice of [29, 30]. The quantitative evaluation results are shown in Table 1. Our proposed method achieves the Dice scores of 84.05% and 69.93% on BraTS and CQ500 datasets, respectively, which surpassed all other methods by a large margin, demonstrating the effectiveness of our method. Compared to the baseline foreground probability map with threshold of 0.5 reaching an average Dice of 75.85% and 64.90% on the BraTS and CQ500 dataset, respectively, our method enhances Dice by about 8.20%-5.03%, which could be attributed to the learned contrastive representation and noise calibration effect by our module.

It is also worth noting that our method is not bound to GMM-initiated segmentation, but can also be used as a plug-and-play module to improve the results of other un-

Table 3. Ablation study of temperature on the BraTS dataset.

Temperature (τ)	0.07	0.1	0.2	0.3	0.7	0.9
Dice	0.8254	0.8375	**0.8405**	0.8392	0.8248	0.8211
ASSD [mm]	1.4744	1.1809	1.1756	**1.1550**	1.2665	1.1871

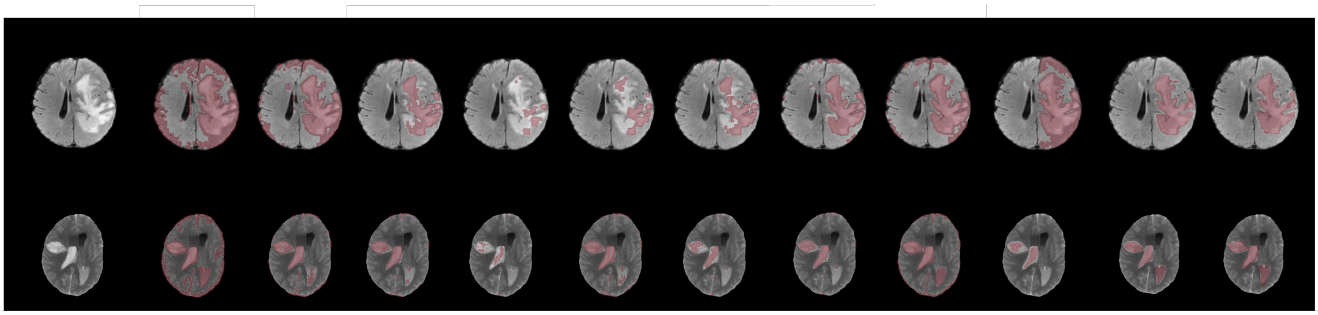

Figure 2. Visualization of exemplary segmentation results on BraTS and CQ500 dataset. From left to right: images to be segmented(1st column), segmentation results of different comparative methods (2rd-9th column), results of our method based on GMM (10th column), and the ground truth segmentation (the last column).

supervised methods, e.g., it also improves Mumford-Shah segmentation by almost 6% on the BraTS dataset.

Fig. 2 illustrates that most anomalous voxels are precisely covered by the prediction of our method. On the contrary, We can also observe that reconstruct-based methods exhibit varying degrees of mis-segmentation, e.g., Context VAE [1] and Bayesian VAE [2], resulting in unsatisfactory performance in both datasets. This occurrence might be attributed to the inherent limitations of solely relying on pixel-wise residuals for anomaly detection. The inability to fully capture the complex and diverse characteristics of anomalies, especially in challenging scenarios, can significantly impact the accuracy of the segmentation results.

In addition, we also conducted full supervision experiments to demonstrate the upper-bound of our framework. Concretely, these experiments were conducted using real labels and the backbone of our method, while maintaining consistent experiment settings, such as dataset splits, with the unsupervised methods.

Taking into account the results presented above, these experiments demonstrate that our method is able to locate and segment lesion regions regardless of the modalities or imaging protocols. These findings reinforce the potential of our method as a versatile and valuable tool for lesion detection and localization in various clinical settings.

5.2. Ablation Studies

5.2.1 Effectiveness of Each Module.

We execute the ablation study on BraTS dataset and report quantitative performances in Table 2. We also perform full supervision experiment using nnUNet [27] as backbone to present whether variations from implementation details could dominate results. As shown in Table 3, when only using foreground probability P with a threshold of 0.5, 75.85% of Dice is reached. Training the backbone only using P achieves 79.29% of Dice. By utilizing either coarse-grained or fine-grained discrimination techniques, we achieved Dice scores of 81.64% and 82.48%, respectively. Combining the two approaches yields the highest Dice score of (84.05%).

We also conduct experiments about the size of mini-patch $\widetilde{z} \in \{\mathbb{R}^{12\times12\times4}, \mathbb{R}^{6\times6\times4}, \mathbb{R}^{3\times3\times2}\}$. We found that the patch size can have a minor impact on performance. Using smaller patches may result in fragmented semantic content, while larger patches may contain both lesion and normal tissues, leading to ambiguous content.

5.2.2 Effectiveness of Temperature-calibrated Logits.

Contrastive training procedure usually uses low temperature (e.g. $\tau = 0.07$) to excavate hard samples [31]. However, in the presence of label noise, employing smoother logits may help improve the model's performance [32]. Hence, we explore different temperature settings as a calibration strategy. Table 3 demonstrates that our model achieves its best performance when $\tau = 0.2$, indicating that our temperature scaling approach strikes a balance between hardness for sample discrimination and softness for label noise.

6. Conclusion

To summarize, we have presented a new unsupervised framework for medical image segmentation using a novel

cross-granularity contrastive module. Our module contains coarse-grained and fine-grained discrimination paths, enabling the network to capture the distinctions between lesions and normal tissues at different levels of context. We evaluate our method on two large public datasets of CT/MRI scans and demonstrate that our approach improves a Gaussian mixture model-based segmentation by up to 9%, which surpasses all other unsupervised segmentation methods by a large margin. Additionally, our module can also be combined with other existing unsupervised segmentation methods to further enhance their performance. Therefore, our framework shows great potential for use in medical image applications with limited labelled data availability.

References

[1] David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018. 1, 2, 5, 6

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 5, 6

[3] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[5] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018. 2

[6] Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. 2018. 2

[7] Suhang You, Kerem C Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pages 540–556. PMLR, 2019. 2

[8] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 161–169. Springer, 2019. 2

[9] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 2

[10] Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767. IEEE, 2021. 2

[11] Lu Wang, Dongkai Zhang, Jiahao Guo, and Yuexing Han. Image anomaly detection using normal data only by latent space resampling. *Applied Sciences*, 10(23):8660, 2020. 2

[12] Christoph Baur, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–727. Springer, 2020. 2

[13] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1127–1131. IEEE, 2021. 2

[14] Julio Silva-Rodríguez, Valery Naranjo, and Jose Dolz. Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80:102526, 2022. 2, 5

[15] Felix Meissen, Benedikt Wiestler, Georgios Kaissis, and Daniel Rueckert. On the pitfalls of using the residual error as anomaly score. In *International Conference on Medical Imaging with Deep Learning*, pages 914–928. PMLR, 2022. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[17] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33:12546–12558, 2020. 2

[18] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2

[19] Ziqi Yu, Xiaoyang Han, Shengjie Zhang, Jianfeng Feng, Tingying Peng, and Xiao-Yong Zhang. Mousegan++: Unsupervised disentanglement and contrastive representation for multiple mri modalities synthesis and structural segmentation of mouse brain. *IEEE Transactions on Medical Imaging*, 42(4):1197–1209, 2022. 2

[20] Yuchen Fei, Chen Zu, Zhengyang Jiao, Xi Wu, Jiliu Zhou, Dinggang Shen, and Yan Wang. Classification-aided high-quality pet image synthesis via bidirectional contrastive gan

with shared information maximization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 527–537. Springer, 2022. 2

[21] Suhang You, Kerem C Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pages 540–556. PMLR, 2019. 5

[22] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 5

[23] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157. Springer, 2017. 5

[24] Boah Kim and Jong Chul Ye. Mumford–shah loss functional for image segmentation with deep learning. *IEEE Transactions on Image Processing*, 29:1856–1866, 2019. 5

[25] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 4

[26] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Development and validation of deep learning algorithms for detection of critical findings in head ct scans. *arXiv preprint arXiv:1803.05854*, 2018. 4

[27] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 4, 6

[28] Botao Zhao, Yan Ren, Ziqi Yu, Jinhua Yu, Tingying Peng, and Xiao-Yong Zhang. Aucseg: An automatically unsupervised clustering toolbox for 3d-segmentation of high-grade gliomas in multi-parametric mr images. *Frontiers in Oncology*, 11:679952, 2021. 4

[29] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021. 5

[30] Julio Silva-Rodríguez, Valery Naranjo, and Jose Dolz. Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80:102526, 2022. 5

[31] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. 6

[32] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020. 6