Tri-VAE: Triplet Variational Autoencoder for Unsupervised Anomaly Detection in Brain Tumor MRI

Hansen Wijanarko, Evelyne Calista, Li-Fen Chen, Yong-Sheng Chen
National Yang Ming Chiao Tung University
Taiwan
chocolemon.cs06@nctu.edu.tw, {evelynecalista.bt09,lfchen, yschen}@nycu.edu.tw

Abstract

The intricate manifestations of pathological brain lesions in imaging data pose challenges for supervised detection methods due to the scarcity of annotated samples. To overcome this difficulty, our focus shifts to unsupervised anomaly detection. In this work, we exclusively train the proposed model using healthy data to identify unseen anomalies during testing. This study entails investigating the triplet-based variational autoencoder to simultaneously learn the distribution of healthy brain data and denoising capabilities. Importantly, we rectify a misconception inherent in prior projection-based approaches which relies on the presumption that healthy regions within images would persist unaltered in the reconstructed output. This inadvertently implied a substantial likeness in latent space representations between lesion and lesion-free images. However, this assumption might not hold true, particularly due to the potential significant impact of lesion area intensities on the projection process notably for autoencoders with single information bottleneck. To overcome this limitation, we disentangled metric learning from latent sampling. This approach ensures that both lesion and lesion-free input images are projected into the same distribution, specifically the lesion-free projection. Moreover, we introduce a semantic-guided gated cross skip module to enhance spatial detail retrieval while suppressing anomalies, leveraging robust healthy brain representation semantics exist in the deeper levels of the decoder. We also discovered that incorporating structure similarity index measure as an extra training objective bolsters the capability of anomaly detection for the proposed model.

1. Introduction

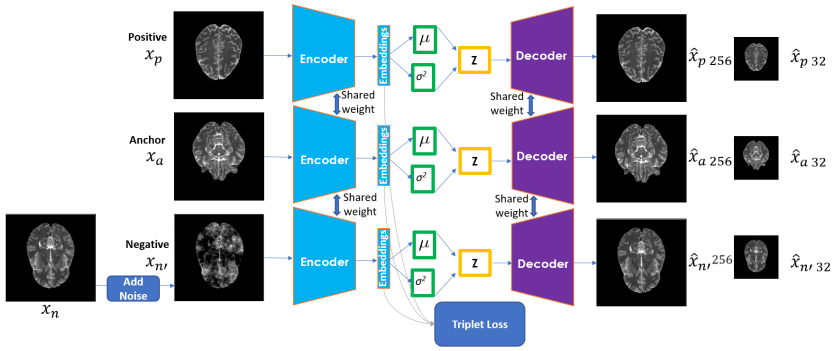
Magnetic Resonance Imaging (MRI) serves as a critical tool for early-stage disease detection, precise disease staging, and meticulous treatment planning. It provides healthcare professionals with detailed insights into anatomical structures, enabling accurate clinical decision-making. However, the analysis and interpretation of MR scans pose a complex challenge, with studies indicating that significant pathologies may go unnoticed in 5 - 10% of cases [4, 5]. Given the intricate nature of pathological brain lesions depicted in MRI images, there is an urgent need for the development of computational approaches to aid radiologists in addressing these complexities.

In light of the challenges posed by the intricate nature of MRI scans, developing a supervised deep learning model requires annotated label data, which is both laborious and reliant on human observers. Consequently, unsupervised learning methods offer a promising solution to mitigate these challenges. Aside from its practical benefits, the adaptability of unsupervised learning to diverse and previously unknown anomalies has enormous potential. This adaptability is especially appealing in the medical domain, where the ability to streamline pre-screening processes and increase radiological assessment efficiency is critical.

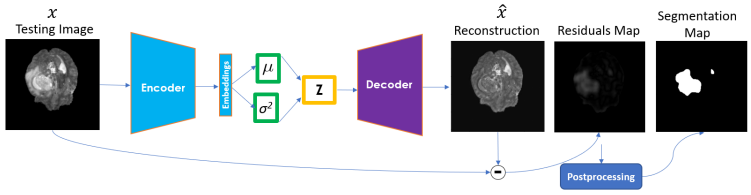
Previous anomaly detection methods [4] commonly employed a technique involving the projection of images containing brain lesions into the latent space of the model. However, this method relied on the assumption that healthy regions in the images would remain unchanged in the reconstructed image. In other words, it assumed that the latent space representations of an image with a lesion and the same image without a lesion would be very similar. Unfortunately, this assumption could be problematic because the presence of a lesion might significantly affect the projection step, leading to large deviations between the latent space representations. In this paper, we introduce *Tri-VAE*, a triplet variational autoencoder, to address the misconception in previous projection-based methods. Our approach integrating a triplet loss and a semantic-guided gated cross skip connection module to enhance spatial detail retrieval while suppressing anomalies.

Our main contributions are summarized as follow:

- We propose a novel unsupervised training scheme with



(a) Training phase architecture.



(b) Testing phase.

Figure 1. Training strategy and inference procedure of the proposed model. (a) The training process involves feeding the network with three randomly selected 2D brain slices from distinct healthy subjects, designated as anchor (x_a), positive (x_p), and negative (x_n) samples. (b) During the testing phase, we refrain from introducing any noise, and the process of anomaly detection is primarily accomplished through the assessment of reconstruction error.

metric learning that surpasses the performance of the previous anomaly detection methods.

- We rectify a misconception in previous projection-based methods, which assumed that healthy regions would remain unchanged in the reconstructed output. This implied a significant similarity in latent space representations between lesion and lesion-free images.
- We propose a novel semantic-guided gated cross skip connection module to enhance the retrieval of low-level non-semantic information.
- We explore the benefits of incorporating the structural similarity index as an supplementary training loss function.

2. Related Work

Baur et al. [3] were the pioneers in the unsupervised anomaly detection (UAD) for brain lesion. They used the spatial variant autoencoders models to better preserve spatial details by incorporating a spatial bottlenecks design and rely solely on a set of normal data. In this approach, the identification and demarcation of pathologies are derived through an assessment of pixel-wise reconstruction errors. Their work rests upon the concept that, due to the constrained anatomical variability of healthy brains, an alternative strategy involves modeling the distribution of healthy brains and identifying deviations from this baseline to detect and delineate anomalies. However, the pursuit of simultaneously achieving high-fidelity reconstruction for normal regions and intentionally introducing errors for anomalous regions may pose a conflicting objective.

In more intricate strategies, *Schlegl et al.* [17] proposed f-AnoGaN, an unsupervised anomaly detection using adversarial networks. They re-purposed the generator and discriminator components of a GAN to employ a rapid mapping technique for new data into the latent space of the GAN. Moreover, *Raunak and Yi* [8] employed the adversarial strategy with ASC-Net that circumvents the necessity for conventional image reconstruction, instead prioritizing direct anomaly detection. ASC-Net innovatively employs a discriminator to assimilate insights from the reference image distribution, effectively partitioning the input image into segments classified as within and outside the distribution. *Baur et al.* [4] executed a comparative assessment of unsupervised anomaly detection methods in brain MRI. Their study scrutinized 17 models encompassing autoencoders, latent variable models, and varied generative adversarial network configurations, in both spatial and dense bottleneck designs. The comparison demonstrated that restoration-VAE [20] and F-AnoGAN [17] achieved superior performance, and concluded that VAE [13, 15] was the most easily optimized among them.

The misconception in prior projection-based method was identified in [20]. Those methods assumed that healthy regions in images would remain unchanged in the reconstructed output, implying a striking similarity in latent space representations between lesion and lesion-free im-

ages. However, this assumption might not be valid, given the potential substantial impact of lesion area intensities on the projection process. Hence, they redefined anomaly detection as a maximum a posteriori (MAP) image restoration challenge. The restoration-VAE method employed gradient ascent and iterative optimization to achieve restored images during testing, using the normative image prior estimated by a deep learning-based model VAE. This approach substituted the reconstruction error with restoration error, traditionally employed for estimating anomaly scores. In addition to the prevalent use of VAEs as the primary models for anomaly detection, several structural modifications have been proposed to enhance their effectiveness.

None of the aforementioned studies in unsupervised anomaly detection have thus far integrated the application of denoising techniques, a strategy increasingly influential in distinct domains. *Kascenas et al.* [12] employed denoising autoencoders with coarse noise for tumor dataset anomaly detection. Moreover, *Wyatt et al.* [19] utilized the diffusion probabilistic model (DDPM) [9], employing simplex noise degradation instead of Gaussian noise. The noise modification improved the performance by enabling the reconstruction of larger anomalous regions as healthy areas, thanks to the coherent and structured randomness generated by noise.

In this study, we integrate the essence of two distinct concepts: the acquisition of the representation distribution of healthy brain data through latent variable modeling and the integration of denoising strategies within a unified framework. This approach not only address the misconception prevalent in the prior projection reconstruction-based method for unsupervised anomaly detection identified in [20], but also demonstrates the efficacy of these concepts.

3. Methodology

3.1. Anomaly detection method

Training phase. In the training phase of our proposed method, we exclusively utilize the healthy brain dataset, a common practice in unsupervised anomaly detection. Our network adopts a triplet of VAE architecture, as illustrated in Fig. 1a. This architecture consists of three instances of the same feed-forward VAE network, with shared weights. Additionally, we introduce noise to x_n to simulate the presence of anomalies in a healthy brain slice, denoted as $x_{n'}$.

Our training strategy combines two advantageous aspects: learning the distribution of healthy brain slices and harnessing denoising capabilities for noise removal and the reconstruction of healthy brain slices. This closely mirrors the conditions encountered during testing, where the network may also encounter brain slices with lesions. Consequently, the network is encouraged to perform as it was explicitly trained: capturing the unique information of each brain slice, projecting its latent representation onto a normal version, and reconstructing its healthy brain counterpart. This comprehensive training scheme and network architecture address the misconception discussed earlier in Section 2. Instead of redefining it as another problem, as done by *You et al.* [20], our approach offers a holistic solution, effectively mitigating the issue.

Noise generation during training. In our experimental setup, we investigate the introduction of two distinct types of noise, namely coarse noise and simplex noise, when transforming x_n into $x_{n'}$. The selection of these noise types is based on their demonstrated effectiveness in enhancing the model's denoising capabilities, as supported by previous studies [12, 19].

The first noise type we consider is coarse noise, as employed in the *DAE* [12]. They highlighted that the use of lower-resolution noise is associated with improved anomaly detection. In line with these recommendations, we adopt a coarse noise specification featuring a resolution of 16×16 pixels and a standard deviation of 0.2 for the Gaussian noise. This noise is upsampled to match our input image size of 256×256 pixels.

We also utilize simplex noise within our experimental framework. Following the strategy used in AnoDDPM [19], our approach incorporates multiple octaves of noise. We set a starting frequency of 2^{-6} and the number of octaves to 6, establish a persistence (or decay rate) of 0.8, and lastly, fix the lacunarity at 2. Subsequently, we introduce perturbations to the input images via a forward diffusion process consisting of 87 steps. This process yields noise patterns that exhibit spectral density visually akin to our coarse noise model.

Testing phase. As illustrated in Fig. 1b, the anomaly detection is performed by computing the absolute difference between the input data x and the corresponding reconstruction $\hat{x}$, denoted as $|x - \hat{x}|$. This evaluation procedure yields a measure known as the *residuals map*, which directly signifies the presence of anomalies through its calculated scores.

Post-processing procedure. To refine the anomaly detection results, firstly, we mask the residuals map with a foreground mask to ensure that only anomalies detected within the brain region are considered. Next, we apply a small thresholding operation to further refine the results. This thresholding step compensates for very subtle reconstruction errors that may not necessarily indicate anomalies. In this study, we choose a threshold value of 0.1, which we consider to be reasonable as smaller values would be less effective. Then, we aggregate the residual images from all slices in the inference mini-batches to create a corresponding 3D residuals volume. To enhance the quality of this volume, we apply a 3D median filtering technique. This process helps to remove small outliers and to create a smoother and more continuous signal while preserving im-

portant edges. In line with prior research [4], we employ a 3D median filter with a kernel size of $5 \times 5 \times 5$. Lastly, to further clean the volume and eliminate small noise or artifacts, we apply a 3D connected component filter. This filter effectively removes any volume components smaller than 8 voxels.

3.2. Overall VAE architecture

The overall structure of our VAE network, follows the encoder-decoder framework inspired by the U-Net model [16], is visually represented in Fig. 2 and Fig. 3. A distinctive feature of our architecture is the inclusion of skip connections (concatenations) in the first three layers while excluding them from deeper layers. This strategic choice enables our network to effectively capture both low-level and high-level features from the input data. By integrating skip connections at the outset, we ensure that the network can leverage fine-grained details captured by the lower layers. Meanwhile, by excluding skip connections from the deeper layers, we encourage the network to concentrate on learning intricate semantic representations that help distinguish healthy brain slices from noisy ones.

3.3. Enhancing brain image representation with semantic-enriched metric learning and sampling

Our approach focuses on restoring input brain images to a healthy state, whether the input comprises healthy brain slices or noisy ones. As the result, we untangle the triplet learning mechanism from the latent space resampling. By doing so, we aim to achieve similar projections for images containing noise and their noise-free counterparts onto the latent space representing normal healthy brain states.

The triplet loss, which is employed for metric learning, is calculated on the bottleneck embeddings. These embeddings capture the high-level semantic distinctions between healthy brain slices and those with artificial noise, which generalized to brain slices with actual lesions on test time. These embeddings are then utilized to predict both the mean and log variance of the data distribution for the input brain slices. Subsequently, we sample the latent embeddings of healthy versions of brain slices. These sampled healthy brain embeddings are concatenated with the metric learning-enriched embeddings. This combination allows us to retain the unique features of the brain slices while preserving their healthiness, thereby enhancing our network's ability to differentiate between normal and anomalous brain slices.

3.4. Enhancing spatial details retrieval with semantic-guided gated cross skip connection

By introducing a coarse-scale loss, we ensure that the network's deeper layers capture essential knowledge and semantic details prior to the introduction of gated cross skip (GCS) connection. This is achieved by evaluating the reconstruction loss on a layer that precedes the skip connection. Following this, to facilitate the transformation of high-level semantics into spatial representations, we incorporate an additional convolution block just before the adjustment of the output dimension using a 1×1 2D convolutional layer for the 32 coarse scale output. The inclusion of an intermediate reconstruction loss encourages the model to preserve detailed information across layers, particularly before the introduction of the GCS connection. This is crucial because the performance of the retrieval process relies on the features sampled and learned in the preceding decoder layers.

The GCS connection aims to address the loss of spatial information in healthy regions due to dimensional reduction and the latent resampling process. We first pass both the encoder and decoder features through a linear layer to reduce the channel dimension to $C//r$, where r is the reduction ratio set to 4 in our case, and to reduce the number of parameters to be computed. Inspired by cross-attention mechanisms [7], this approach is employed to recover spatial details that may have been lost. By calculating cosine similarity from both encoder and decoder features, an attention map is generated. This map facilitates the retrieval of spatial information, assigning higher similarity scores to healthy regions, thus preserving their details. Anomalous regions, conversely, exhibit lower similarity scores, resulting in the suppression of their features in the decoder.

Furthermore, our choice of using the hyperbolic tangent activation function, rather than the sigmoid function, allows values to be bounded between -1 and 1. This flexibility avoids making predefined assumptions about whether anomaly regions should appear brighter or darker, which can vary across different image modalities. To modulate the GCS features and decoder features achieving minimal loss, we introduce the learnable parameters, α and β. Concatenating these features before the convolution block ensures the effective integration of contextual information into our model.

3.5. Loss functions

Reconstruction loss function. In this work, we choose the L_1 loss as the reconstruction loss function:

$$L_1(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i| \,, \tag{1}$$

where N is the number of pixels in image. We employ the L_1 reconstruction loss for positive, anchor, and negative samples with the spatial size 32×32 ensuring the incorporation of strong semantic features. Additionally, we purposefully leave out the L_1 reconstruction loss for healthy brain slices within the anchor and positive samples of full

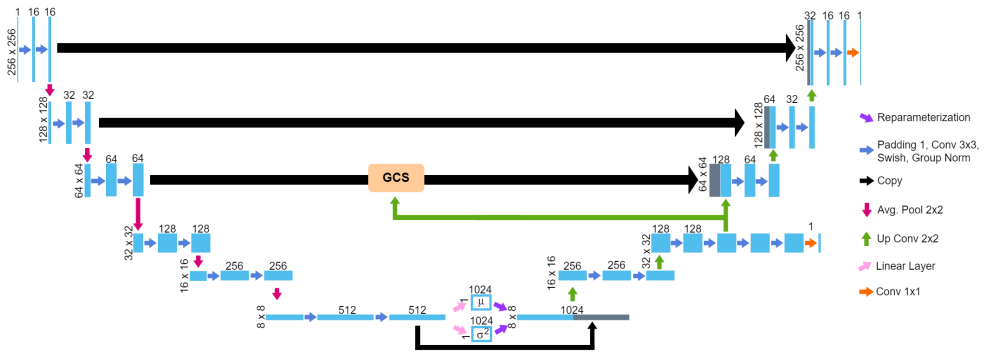


Figure 2. Network VAE architecture.

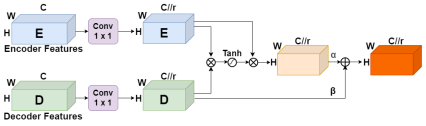


Figure 3. Gated Cross Skip (GCS) connection.

scale 256×256 output. This deliberate omission serves the purpose of preventing the network from relying on the simplistic approach of copying information directly from the input data via skip connections. For both coarse scale and full scale output, denoting anchor samples as anc, positive samples as pos, and negative samples as neg, the equation is as follows:

$$L_{1\,\mathrm{total}} = L_{1\,\mathrm{anc},32\times32} + L_{1\,\mathrm{pos},32\times32} + L_{1\,\mathrm{neg},32\times32} + L_{1\,\mathrm{neg},256\times256}\,. \tag{2}$$

KL divergence loss function. Within our framework of VAEs, the Kullback-Leibler Divergence (KLD) loss plays a crucial role in regulating the behavior of the latent space. The KLD equation is formulated as:

$$\mathrm{KLD}(p\,||\,q) = \frac{1}{2}\sum_{i=1}^{N}\left(\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2) - 1\right)\,, \tag{3}$$

where N is the latent dimension. In our experimental setup, our primary goal is to minimize the KLD loss for the distribution of healthy brain slices within the anchor and positive samples. This strategic choice aligns with the core objective of our network, which focusing on the reconstruction of healthy brain slices regardless of input conditions. Hence, enabling the network to generate latent representations corresponding to the characteristics of healthy brain slices for negative samples. This approach involves incorporating anchor and positive samples in the calculation of KLD:

$$\mathrm{KLD}_{\mathrm{total}} = \mathrm{KLD}_{\mathrm{anc}} + \mathrm{KLD}_{\mathrm{pos}}\,. \tag{4}$$

Triplet loss function. In this work, our specific context is that all healthy normal brain slices exhibit similarity with one another while being distinctly dissimilar from any noisy brain slices. Our approach draws parallels with the TVAE model [11]. Here, our network is tasked with the dual objective of minimizing the upper-bound on the expected negative log-likelihood of data while simultaneously incorporating the triplet loss.

Let's denote the three inputs as x_{a}, x_{p}, and x_{n} for the anchor, positive, and negative input samples, respectively. Also, let the embedded representation of the network be represented as $\mathrm{Enc}(x)$, the distance between two embedding points x and y as $d(x, y)$, and margin as α (in our experiment we set it equals 1). The triplet loss equation is as follows:

$$\mathcal{L}_{\mathrm{triplet}} = \sum_{i=1}^{N}\max\left(d(\mathrm{Enc}(x_{\mathrm{a}_i}), \mathrm{Enc}(x_{\mathrm{p}_i})) - d(\mathrm{Enc}(x_{\mathrm{a}_i}), \mathrm{Enc}(x_{\mathrm{n}_i})) + \alpha, 0\right)\,, \tag{5}$$

with squared L_2 norm for the distance function $d(x, y)$:

$$d(x, y) = \sum_{i=1}^{N}(x_i - y_i)^2\,. \tag{6}$$

SSIM loss function. The structural similarity index (SSIM) [18] serves as a valuable metric for quantifying the structural resemblance between a pair of images. SSIM provides a comprehensive assessment of image similarity, encompassing luminance, contrast, and structural aspects. The equation is as the following:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}\,. \tag{7}$$

We incorporate SSIM into the loss function, which calculates the similarity between the reconstruction output from

the negative samples ($\hat{x}_n$) and the original negative samples before the addition of noise (x_n). This inclusion aims to improve the quality of the reconstructed images. This is achieved by not only examining pixel values in isolation but also by evaluating their structural relationships, taking into account neighboring pixels. Notably, for the boundary regions, the L_1 loss is still employed, ensuring a balanced approach to optimizing image reconstruction quality. By doing so, the network gains a more holistic understanding of image structures, leading to improved reconstruction outcomes.

Total loss function. By amalgamating the VAE loss functions, the metric learning loss function, and the supplementary structural similarity loss function, we derive our comprehensive total loss function, denoted as L_{total}. This total loss function is formally represented as follows:

$$L_{\text{total}} = L_{1\,\text{total}} + \text{KLD}_{\text{total}} + \mathcal{L}_{\text{triplet}} + \text{SSIM}_{\text{neg}}\,. \quad (8)$$

4. Experiments and Results

4.1. Dataset

Data description. We assess the performance of *Tri-VAE* on two publicly available datasets. The first dataset, serve as training dataset, consists of meticulously curated healthy brain MRI scans obtained from the IXI Brain Development Dataset [6]. We carefully select 250 T2 scans to create a representative subset, each featuring dimensions of 256×256 pixels. On average, approximately 120 slices are attributed to each subject, varying based on individual brain size. The T2 scans maintain a spatial resolution of $0.9375 \times 0.9375\ mm^2$, with a slice thickness of 0.125 mm.

Our anomaly dataset, essential for the testing phase, comprises MRI scans of brains afflicted with lesions sourced from the BraTS-2020 Dataset [1, 2, 14]. All data underwent a series of preprocessing steps, including co-registration to a standardized anatomical template, interpolation to a consistent resolution of 1 mm^3, and meticulous skull-stripping. For our study, we randomly select 30 T2 scans from the BraTS training dataset, with each T2 slice maintaining dimensions of 240×240 pixels. On average, each subject's scan comprises approximately 155 slices, with the exact count varying according to the subject's unique brain size.

Data preprocessing. Unlike the tumor dataset, the dataset containing healthy brain scans initially lack of skull-stripping. Therefore, we employ the HD-BET brain MRI extraction tool [10] for the brain extraction. For the anomalous data from the BraTS2020 dataset, we harmonize the labels representing tumor sub-regions, such as the GD-enhancing tumor, peritumoral edema, and necrotic and non-enhancing tumor, into a single "anomaly" label. Subsequently, all slices are resized to a uniform resolution of 256×256 after being normalized.

Evaluation metrics. We utilize standard evaluation metrics to quantitatively assess our results, including the DICE similarity coefficient, AUROC, and AUPRC. AUROC offers insights into a model's ability to correctly rank examples but lacks details about specific tradeoffs between true positives and false positives. In contrast, AUPRC provides a more nuanced perspective on false positives, which is particularly valuable for scenarios involving imbalanced datasets, such as medical diagnostics.

4.2. Results

We compare *Tri-VAE* with existing brain MRI anomaly detection methods: *AnoDDPM* [19], *VAE (restoration)* [20], *F-Anogan* [17], *ASC-Net* [8], and *DAE* [12].

Quantitative comparison. The quantitative evaluation results are presented in Table 1. *Tri-VAE*, along with the latest denoising technique *DAE* [12], which employs coarse noise, demonstrates superior performance in the evaluation. Compared to *DAE* [12], *Tri-VAE* is outperform on both of the noise type. While paired with the simplex noise, *DAE* [12] shows similar performance with ASC-Net [8].

Visual comparison. Despite the denoising method's superiority in anomaly detection, *AnoDDPM* [19] exhibits a tendency to predict a relatively high number of false positives, as depicted in Fig. 4. However, *AnoDDPM* [19] still achieves a better AUPRC compared to other older methods like *VAE* (restoration) [20] and *F-Anogan* [17], indicates improved precision. Meanwhile, *ASC-Net* demonstrates a well-balanced enhancement across all three evaluation metrics compared to its predecessors, with a reduction in false positives while retaining the ability to identify anomalous regions. In the rightmost part of the comparison, both *DAE* [12] and our *Tri-VAE* exhibit minimal false positives. Nevertheless, our *Tri-VAE* successfully detects more anomalous regions compared to *DAE* [12].

Model	DICE	AUROC	AUPRC
AnoDDPM [19] (2022)	0.2363	0.9188	0.3557
VAE (restoration) [20] (2018-2021)	0.3490	0.9446	0.2995
F-Anogan [17] (2018-2021)	0.2971	0.9461	0.2796
ASC-Net [8] (2021)	0.3818	0.9650	0.3644
DAE [12] (2022) (w/ simplex noise)	0.3738	0.9465	0.3293
Tri-VAE (ours) (w/ simplex noise)	**0.4047**	**0.9590**	**0.3965**
DAE [12] (2022) (w/ coarse noise)	0.5687	0.9571	0.4323
Tri-VAE (ours) (w/ coarse noise)	**0.6058**	**0.9682**	**0.4615**

Table 1. Performance comparison among anomaly detection methods. The proposed method, *Tri-VAE*, outperforms others by a significant margin.

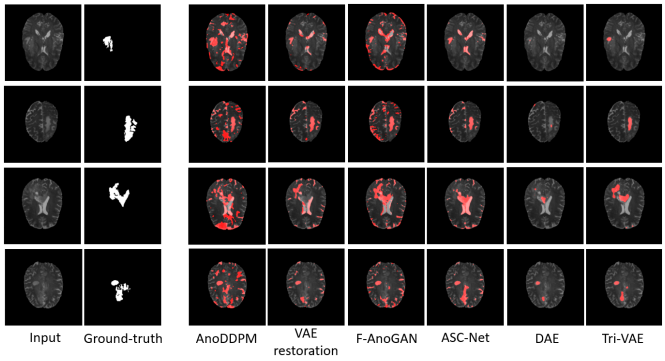


Figure 4. The segmentation results of proposed method (*Tri-VAE*) demonstrate a reduction in false positives while exhibiting a robust ability to capture the majority proportion of anomalous regions.

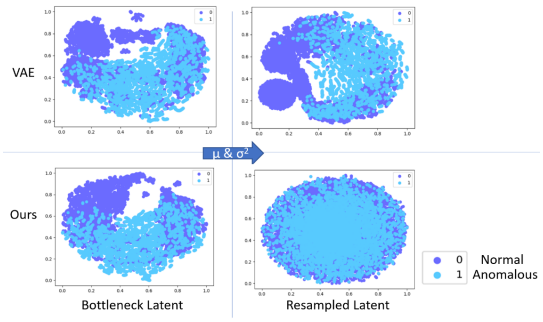


Figure 5. We employ t-SNE with random embedding initialization and a typical perplexity of 30 to project the embeddings of VAE and our *Tri-VAE* onto the same 2D space. The left side is the latent embeddings from the encoder, which initially contain unique high-level semantic information about the slices. The right side is the resampled latent projections, which should contain only the definition of healthy brain slices.

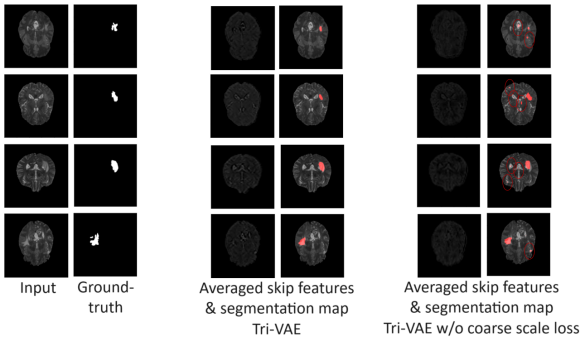


Figure 6. This comparison highlights differences between the feature maps and segmentation results generated by *Tri-VAE* and an alternative model that omits the coarse scale reconstruction loss. Small false positives are marked within the red circles

4.3. Projection of bottleneck and resampled healthy latent

Tri-VAE is designed to reconstruct the normal brain slice version regardless of the input, thus it should encourage the resampling of normal brain slice latents only. This results in the latent representations that originate from the same cluster distribution for both normal and anomalous inputs. In our experiment, we compare the sampled latent projections of the test dataset using our *Tri-VAE* and a VAE as the baseline. As depicted in Fig. 5, *Tri-VAE*'s projected bottleneck latents are denser between brain slices of similar categories (normal and anomalous) due to the metric learning from the triplet loss. This indicates that *Tri-VAE* has more confidence and captures more discriminative high-level semantics in distinguishing healthy and anomalous brain slices. On the other hand, the resampled latent projections of healthy and anomalous brain slices from our model are closely positioned, following the concept explained earlier, while the resampled latents of the VAE still produce points that do not lie in the same distribution, deviating from this concept. This illustrates the effectiveness of our approach in capturing meaningful representations.

4.4. Effect of coarse scale loss on spatial details retrieval

To demonstrate the beneficial impact of the coarse scale reconstruction loss that precedes the GCS connection in enhancing the capacity of deeper network layers to capture robust high-level semantics, we present a visual comparison in Fig. 6. The feature maps are showcased by averaging the feature maps along the channel's axis into a singular channel image. Notably, the feature maps produced by our current model exhibit heightened contrast, indicative of greater confidence in correctly identifying cerebrospinal fluid (CSF) regions, characterized by high pixel intensities. Conversely, the model without the coarse scale reconstruction loss displays reduced confidence in identifying these regions. This observation highlights the effectiveness of the GCS connection in its dual role: facilitating the preserva-

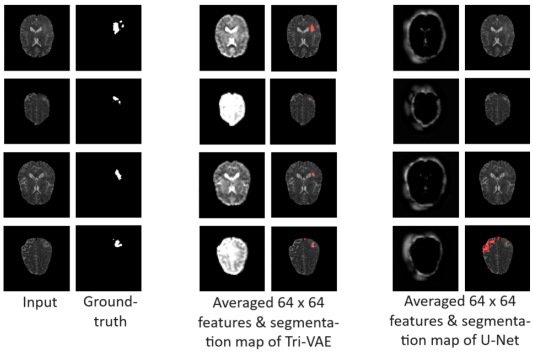


Figure 7. The averaged 64×64 spatial-sized feature maps generated in the midsection of the decoder, just prior to the subsequent upsampling.

tion of low-level details in healthy regions while suppressing spurious information in anomalous regions.

Furthermore, *Tri-VAE* exhibits a notable advantage in terms of producing more precise segmentation maps compared to the model without the coarse scale reconstruction loss. The latter, in contrast, tends to yield more false positives. This outcome reaffirms the significance of the coarse scale reconstruction loss in refining segmentation accuracy and reducing false positive predictions.

4.5. Comparative visualization of network features

As shown in Fig. 7, the feature maps produced by *Tri-VAE* reveals a greater level of details captured before even reaching the end of the network. This stands in sharp contrast to the feature maps produced by the U-Net architecture, which primarily only depict boundaries and offer only limited insight into the CSF regions. Furthermore, in terms of segmentation results, our network demonstrates an exceptional ability to detect subtle anomalous regions that the U-Net architecture struggles to identify. This comparative visualization highlights our network's proficiency in capturing richer and more detailed information at an earlier stage in the encoding-decoding process, showcasing its superior feature extraction capabilities.

4.6. Ablation study

We investigate the contributions of each component within the *Tri-VAE* relative to the VAE baseline, as shown in Table 2. Throughout this ablation study, we use coarse noise into the negative samples due to its superior performance. Incorporating metric learning into the model leads to an approximately 2% enhancement across all metrics. This highlights the model's ability to distinguish between normal brain slices and their noisy counterparts, which extends effectively to anomaly detection during testing. Additionally, the effectiveness of the GCS connection becomes evident when it is supported by the coarse-scale loss from the preceding layer. This demonstrates that retrieving low-level spatial features is less effective without a strong semantic understanding, resulting in minimal or negligible improvements. Moreover, integrating the SSIM loss, which facilitates the learning of structural relationships among image pixels, enhances anomaly detection performance. These findings collectively demonstrate the effectiveness of our proposed method's components and their respective roles in improving anomaly detection.

Triplet + Coarse Noise	Skip Connections (concatenate)	GCS	Coarse Loss	SSIM Loss	DICE	AUROC	AUPRC
baseline - VAE					0.3113	0.9274	0.2793
✓					0.3340	0.9437	0.2986
✓	✓				0.5478	0.9687	0.4240
✓	✓	✓			0.5404	0.9590	0.4311
✓	✓	✓	✓		0.5828	0.9621	0.4495
✓	✓	✓	✓	✓	**0.6058**	**0.9682**	**0.4615**

Table 2. Ablation results on our proposed method (*Tri-VAE*) by employing coarse noise to the model.

5. Discussion & Conclusions

In this paper, we introduce a novel Triplet Variational Autoencoder with metric learning for unsupervised brain anomaly detection in MRI images. Our model design aims to address a persistent misconception in the initial concept of Unsupervised Anomaly Detection (UAD). Additionally, we introduce a novel semantic-guided gated cross skip connection module to enhance spatial detail retrieval while suppressing anomalies. The performance of our proposed model surpasses existing methods for brain MRI anomaly detection, offering potential for more accurate and reliable anomaly detection in real-world applications.

While our current model has demonstrated promising results, there are still areas for further investigation and improvement in the future. As observed from the performance results, the choice of noise type and its characteristics significantly influences the model's ability to detect anomalies. Therefore, we must delve deeper into methods aimed at enhancing the model's robustness to noise variability. We believe adopting an iterative noise exploration strategy to bolster model generalization, particularly in real-world lesion scenarios during testing. This strategy involves training an auxiliary network to dynamically adapt noise patterns used in negative samples during training, thereby challenging the main model's ability to construct a healthy representation. Through active exploration of a range of noise parameters, this dynamic process exposes the main model to progressively complex scenarios, thereby enhancing its robustness. Furthermore, implementing this strategy necessitates rigorous validation across diverse lesion datasets. An open-source benchmark dataset for brain MRI anomaly detection would facilitate comprehensive evaluations, fostering transparency and benchmarking within the research community.

References

[1] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 2017. 6

[2] Spyridon Bakas, Mauricio Reyes, András Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, Jana Lipková, John B. Freymann, Justin S. Kirby, Michel Bilello, Hassan M. Fathallah-Shaykh, Roland Wiest, Jan Stefan Kirschke, Benedikt Wiestler, Rivka R. Colen, Aikaterini Kotrotsou, Pamela J. LaMontagne, Daniel S. Marcus, Mikhail Milchenko, Arash Nazeri, Marc-André Weber, Abhishek Mahajan, Ujjwal Baid, Dongjin Kwon, Manu Agarwal, Mahbubul Alam, Alberto Albiol, Antonio Albiol, Alex Varghese, Tran Anh Tuan, Tal Arbel, Aaron Avery, B. Pranjal, Subhashis Banerjee, Thomas Batchelder, Nematollah K. Batmanghelich, Enzo Battistella, Martin Bendszus, Eze Benson, José Bernal, George Biros, Mariano Cabezas, Siddhartha Chandra, Yi-Ju Chang, and et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *ArXiv*, abs/1811.02629, 2018. 6

[3] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *BrainLes@MICCAI*, 2018. 2

[4] Christoph Baur, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical image analysis*, 69:101952, 2020. 1, 2, 4

[5] Michael Bruno, Eric Walker, and Hani Abujudeh. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35:1668–1676, 2015. 1

[6] IXI dataset. Ixi brain development dataset, Year (if available). Data set of nearly 600 MR images from normal, healthy subjects, along with demographic characteristics, collected as part of the Information eXtraction from Images (IXI) project available for download. Tar files containing T1, T2, PD, MRA and DTI (15 directions) scans from these subjects are available. The data has been collected at three different hospitals in London: * Hammersmith Hospital using a Philips 3T system * Guy"s Hospital using a Philips 1.5T system * Institute of Psychiatry using a GE 1.5T system. 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 4

[8] Raunak Dey and Yi Hong. Asc-net: Adversarial-based selective network for unsupervised anomaly segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 236–247, Cham, 2021. Springer International Publishing. 2, 6

[9] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 3

[10] Fabian Isensee, Marianne Schell, Irada Tursunova, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, H. P. Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence mri using artificial neural networks. *Human Brain Mapping*, 40:4952 – 4964, 2019. 6

[11] Haque Ishfaq, Assaf Hoogi, and Daniel L. Rubin. Tvae: Triplet-based variational autoencoder using metric learning. *ArXiv*, abs/1802.04403, 2018. 5

[12] Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning*, 2022. 3, 6

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 2

[14] Bjoern H Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin S. Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth R. Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andaç Hamamci, Khan M. Iftekharuddin, Rajesh Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José Antonio Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen John Price, Tammy Riklin-Raviv, Syed M. S. Reza, Michael T. Ryan, Duygu Sarikaya, Lawrence H. Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Alberto Silva, Nuno J. Sousa, Nagesh K. Subbanna, Gábor Székely, Thomas J. Taylor, Owen M. Thomas, N. Tustison, Gözde B. Ünal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koenraad Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024, 2015. 6

[15] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014. 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 4

[17] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Margarethe Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54: 30–44, 2019. 2, 6

[18] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 5

[19] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 649–655, 2022. 3, 6

[20] Suhang You, Kerem Can Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64:101713, 2018. 2, 3, 6