

Unsupervised contrastive analysis for anomaly detection in brain MRIs via conditional diffusion models[☆]

Cristiano Patrício^{a,b,d,*}, Carlo Alberto Barbano^c, Attilio Fiandrotti^c, Riccardo Renzulli^c,
Marco Grangetto^c, Luís F. Teixeira^d, João C. Neves^{a,b}

^a Universidade da Beira Interior, Covilhã, Portugal

^b NOVA LINCS, Portugal

^c University of Turin, Turin, Italy

^d INESC TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

ARTICLE INFO

Keywords:

Contrastive Analysis
Unsupervised Anomaly Detection
Brain MRI
Contrastive learning
Neuroimaging
Diffusion models

ABSTRACT

Contrastive Analysis (CA) detects anomalies by contrasting patterns unique to a target group (e.g., unhealthy subjects) from those in a background group (e.g., healthy subjects). In the context of brain MRIs, existing CA approaches rely on supervised contrastive learning or variational autoencoders (VAEs) using both healthy and unhealthy data, but such reliance on target samples is challenging in clinical settings. Unsupervised Anomaly Detection (UAD) learns a reference representation of healthy anatomy, eliminating the need for target samples. Deviations from this reference distribution can indicate potential anomalies. In this context, diffusion models have been increasingly adopted in UAD due to their superior performance in image generation compared to VAEs. Nonetheless, precisely reconstructing the anatomy of the brain remains a challenge. In this work, we bridge CA and UAD by reformulating contrastive analysis principles for the unsupervised setting. We propose an unsupervised framework to improve the reconstruction quality by training a self-supervised contrastive encoder on healthy images to extract meaningful anatomical features. These features are used to condition a diffusion model to reconstruct the healthy appearance of a given image, enabling interpretable anomaly localization via pixel-wise comparison. We validate our approach through a proof-of-concept on a facial image dataset and further demonstrate its effectiveness on four brain MRI datasets, outperforming baseline methods in anomaly localization on the NOVA benchmark.

1. Introduction

Despite substantial progress in image classification models, they still face two significant challenges: (i) dependence on extensive amounts of labeled data; and (ii) limited interpretability. These challenges are particularly critical in sectors such as healthcare, where explaining the decision process is essential for clinicians to trust the model outcome, and unhealthy samples are scarce and hard to obtain [1]. Contrastive Analysis (CA) represents a promising solution to this problem [2–4] by learning the fundamental generative factors that distinguish a target (TG) dataset from a background (BG) dataset (referred to as anomalous patterns) and are shared across both datasets (referred to as common patterns). However, state-of-the-art CA methodologies have several limitations: (i) they often result in blurred images when utilizing Variational Autoencoders (VAEs) or are susceptible to mode collapse

and unstable training when using Generative Adversarial Networks (GANs) [5,6]; (ii) they depend on the availability of healthy (BG) and unhealthy (TG) samples during training; (iii) they struggle to preserve common patterns of the original image during the reconstruction process. On the other hand, UAD has emerged as an established approach for modeling the distribution of healthy anatomy and identifying deviations from this distribution as potential anomalies, thus removing the dependence on annotated unhealthy (TG) samples. While Denoising Diffusion Probabilistic Models (DDPM) [7] are commonly adopted as generative models in UAD, their training process lacks explicit constraints to preserve common patterns during reconstruction, hindering the fidelity of normal (healthy) reconstructions. Throughout this paper, we use the terms CA and UAD interchangeably to denote the task of detecting anomalous patterns in images. This choice is motivated by their shared objective: identifying deviations from a background

[☆] This article is part of a Special issue entitled: ‘HCBIA’ published in Pattern Recognition Letters.

* Corresponding author at: Universidade da Beira Interior, Covilhã, Portugal.

E-mail addresses: cristiano.patricio@ubi.pt (C. Patrício), carlo.barbano@unito.it (C.A. Barbano), attilio.fiandrotti@unito.it (A. Fiandrotti), riccardo.renzulli@unito.it (R. Renzulli), marco.grangetto@unito.it (M. Grangetto), luisft@fe.up.pt (L.F. Teixeira), jcneves@ubi.pt (J.C. Neves).

<https://doi.org/10.1016/j.patrec.2025.11.014>

Received 30 June 2025; Received in revised form 30 September 2025; Accepted 6 November 2025

Available online 8 November 2025

0167-8655/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

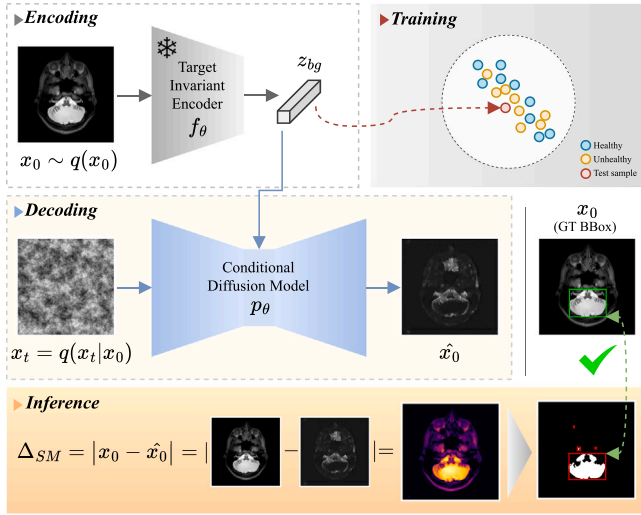


Fig. 1. Overview of the proposed contrastive-guided conditional diffusion model. The target-invariant encoder f_θ encodes the input image x_0 into a latent representation z_{bg} , which captures common (healthy) features. The conditional diffusion model p_θ receives the noisy sample x_t jointly with the common feature z_{bg} to reconstruct the normal (healthy) version of the input image. The saliency map is obtained by the absolute difference between the reconstructed image and the input image ($\Delta_{SM} = |x_0 - \hat{x}_0|$). The high intensity values in the saliency map indicate the potential anomalous regions. Intuitively, the class of the image can be determined based on the magnitude of Δ_{SM} .

(healthy) distribution. The key distinction lies in the data assumptions. While CA requires access to both background and target samples, UAD relies solely on background data. Our work bridges this gap by reformulating contrastive analysis principles for the unsupervised setting, thereby combining the interpretability of CA with the practicality of UAD.

In this work, we propose an unsupervised methodology for the task of anomaly detection in brain MRIs. Specifically, we build upon a contrastive-guided conditional diffusion model (Fig. 1) to reconstruct a healthy version of the input image while preserving its common anatomical patterns. This is achieved by conditioning the denoising diffusion process on a target-invariant latent feature. To this end, we employ a self-supervised contrastive encoder (referred to as target-invariant encoder) trained exclusively on samples from the healthy (BG) dataset. To approximate potential anomalous variation factors, we apply data augmentation techniques such as random cutout and Gaussian noise, which simulate features like tumor presence in brain MRIs. By relying exclusively on healthy (BG) samples during training, our approach reflects the clinical scenarios where annotated pathological brain MRI scans are scarce. During inference, the diffusion model processes unseen samples belonging to the healthy (BG) or unhealthy (TG) distributions by reconstructing the input image such that the anomalous patterns are replaced with learned common (healthy) patterns. By computing pixel-wise differences between the original image and its generated counterpart, high-intensity values can be used to (i) classify the input image as background (normal) or target (anomalous), based on the magnitude of the reconstruction error; and (ii) localize potential anomalies, as illustrated in Fig. 1 (bottom-right), which can aid radiologists in localizing potential tumor regions. Recent studies [8,9] indicate that combining AI with clinical expertise not only enhances diagnostic accuracy but can also reduce interpretation time by up to 23% [8]. These findings underscore the importance of developing AI-driven approaches to support radiologists and alleviate their workload.

Proof-of-concept validation on facial imagery data demonstrates the utility of our method. Additionally, an extensive comparative analysis

against state-of-the-art methods for UAD on four brain MRI benchmarks highlights the effectiveness of our approach, achieving top performance on the recent and challenging evaluation-only NOVA benchmark [10]. Our major contributions are as follows:

- We propose a methodology for learning common features between background (healthy) and target (unhealthy) distributions, enabling the generation of a healthy (normal) counterpart of the input image by encoding only common information;
- We introduce a target-invariant encoder that learns representations which are both class-invariant and instance-aligned, facilitating common features preservation during the generative process;
- We conduct a comprehensive evaluation of our approach on a facial images and four brain MRI datasets, encompassing both healthy (background) and tumor (target) images, to validate the generalization of our approach. Our method outperforms other baselines in anomaly localization on the challenging NOVA benchmark.

2. Related work

Early work in contrastive analysis (CA) relied on the use of Variational Autoencoders (VAEs) [11–13]. A recent method proposed by [13], SepVAE, aims to distinguish common (healthy) from class-specific (unhealthy) patterns in image data. They utilized VAEs with regularization to encourage disentanglement between common and salient representations, along with a classification loss to separate target and background salient factors. However, the resulting reconstructions were often blurry, limiting their interpretability and utility. Alternative approaches [5,6,14] leveraging Generative Adversarial Networks (GANs) have been explored, but they suffer from issues like mode collapse and unstable training. More recently, Diffusion Denoising Probabilistic Models [15–17] (DDPMs) have emerged as a promising alternative for high-quality image generation, addressing the drawbacks of both GANs and VAEs. Our work aligns closely with the setting of CA and UAD. In the context of brain MRI, several works have been proposed [18–20]. For example, [18] introduced a patch-based diffusion model for UAD in brain MRIs. They divide the input image into predefined patches and apply noise to each patch individually in the forward pass. In the backward pass, the partly noised image is utilized to recover the noised patch. One drawback of their work is the extensive duration required for inference. More recently, [20] proposed to conditioning a DDPM to better capture intensity characteristics and domain shifts in brain MRIs during reconstruction, thereby improving segmentation performance. In contrast, our method incorporates contrastive learning during encoder training to explicitly enforce a target-invariant embedding space.

3. Background and notations

3.1. Diffusion models

Denoising diffusion probabilistic models [7] work by corrupting a training image $x_0 \sim q(x_0)$ with a predefined multi-step scheduled noise process to transform it into a sample from a Gaussian distribution. Then, a DNN is trained to revert the process, i.e., starting with a sample from a Gaussian distribution to generate a sample from the data distribution $q(x)$ through a sequence of T sampling steps.

Forward encoder. Given a training image x_0 , the noising process consists of gradually noise-corrupting the image x_0 by adding Gaussian noise according to some variance schedule given by β_t :

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I). \quad (1)$$

As shown in [15], the noisy version of an image x_0 at time t is another Gaussian $q(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$ where $\alpha_t = \prod_{s=1}^t(1-\beta_s)$, which can be written in the form:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \epsilon \sim \mathcal{N}(0, I). \quad (2)$$

Reverse decoder. Since the reverse process $q(x_{t-1}|x_t)$ is intractable, a DNN is used to approximate the distribution $p_\theta(x_{t-1}|x_t)$, where θ represents the weights and biases of the network. The reverse process is then modeled using a Gaussian distribution of the form:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \epsilon_\theta(x_t, t), \beta_t I), \quad (3)$$

where $\epsilon_\theta(x_t, t)$ is a deep neural network governed by a set of parameters θ . From $p_\theta(x_{1:T})$ in Eq. (3), one can generate a sample x_{t-1} from a sample x_t via:

$$x_{t-1} = \sqrt{\alpha_t - 1} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t) + \sigma_t \epsilon_t, \quad (4)$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ is standard Gaussian noise. For training the decoder, [7] reformulated the variational lower bound objective function and considered the objective of predicting the total noise component added to the original image to create the noisy image at a given step. The loss function is then given by the squared difference between the predicted noise $\epsilon_\theta(x_t, t)$ and the actual noise ϵ_t , for a given time step t , using a U-Net [21]:

$$\mathcal{L}_{diff} = \|\epsilon_\theta(x_t, t) - \epsilon_t\|_2^2. \quad (5)$$

By setting $\sigma_t = 0$ in Eq. (4), the coefficient before the random noise ϵ_t becomes zero, resulting in the following deterministic process of generating a new sample x_t . This modification allows to speed up the sampling process without degrading the quality of the generated samples [22].

3.2. Contrastive learning

Contrastive Learning (CL) approaches aim at pulling positive samples' representations (e.g. of the same class) closer together while repelling representations of negative ones (e.g. different classes) apart from each other. Contrasting positive pairs against negative ones is an idea that dates back to previous research [23–25] and has seen various applications in different tasks, such as face recognition [26]. Let $x \in \mathcal{X}$ be an *anchor* sample, x^+ a positive sample (wrt the anchor), and x^- a negative sample. CL methods look for a parametric mapping function $f : \mathcal{X} \rightarrow \mathbb{S}^{d-1}$ that maps semantically similar samples close together in the representation space (i.e. a hypersphere) and dissimilar samples far away from each other. Once pre-trained, f is fixed, and its representation is evaluated on a downstream task, such as classification, through linear evaluation on a test set. Depending on how positive and negative samples are defined, CL can be employed in self-supervised [27] or supervised [28] settings.

4. Method

Our method employs a target-invariant encoder to extract a latent representation capturing common features of the input image. A diffusion-based model is then conditioned on this latent representation to reconstruct a normal (healthy) counterpart of the input. The absolute difference between the reconstructed and original images reveals potential anomalies, where brighter areas indicate abnormalities. An overview of the proposed approach is presented in Fig. 1.

4.1. Target-invariant encoder

The first block of our proposed approach is represented by the target-invariant encoder. This encoder has the goal of learning an input representation which is invariant to the target variable (e.g. presence of eyeglasses in facial images or tumor in brain MRIs) but retains the common information of the input sample. Fig. 1 (top-right) presents a visualization of the embedding space after projecting both healthy and unhealthy samples from the trained encoder. The rationale of this approach is that an invariant representation can allow us to correctly reconstruct a realistic normal version of the input image, as it only encodes its common (healthy) information. To explain the formulation of our encoder, we introduce three definitions specifying the properties that ensure that the encoder preserves the common information of the image.

Definition 4.1 (Instance-aligned Encoder). Given an anchor x , a positive sample x^+ of the same subject, and the set of negative samples x_j^- (all other subjects), we say that an encoder f is instance-aligned if:

$$\|f(x) - f(x_j^-)\|_2^2 - \|f(x) - f(x^+)\|_2^2 \geq \epsilon \quad \forall j \quad (6)$$

where $\epsilon \geq 0$. As the margin ϵ increases, f will provide a better separation between different subjects. In practice, Eq. (6) can be expressed in terms of cosine similarity¹: $\text{sim}(f(x), f(x_j^-)) - \text{sim}(f(x), f(x^+)) \leq -\epsilon \quad \forall j$ which corresponds to the ϵ -InfoNCE loss [29]:

$$\mathcal{L}_{\epsilon\text{-InfoNCE}} = -\log \left(\frac{\exp(s^+)}{\exp(s^+ - \epsilon) + \sum_j \exp(s_j^-)} \right) \quad (7)$$

where s^+ and s_j^- are shorthand notations for $\text{sim}(f(x), f(x^+))$ and $\text{sim}(f(x), f(x_j^-))$ respectively. To obtain a sample x^+ of the same subject of x , if it is not available in the training data, it is possible to employ an augmentation scheme such as in SimCLR [27], i.e. $x^+ = t(x)$ where $t \sim \mathcal{T}$ is an augmentation operator sampled from a family of standard augmentation $\mathcal{T}$ (e.g. random transformations, cropping, etc.).

Definition 4.2 (Class-invariant Encoder). Denoting with $\mathcal{H} \subset \mathcal{X}$ the set of samples which share the same target attribute value (e.g. healthy), and assuming a binary case for simplicity, we say that f is class-invariant, if $x \in \mathcal{H}$ and $x^+ \in \mathcal{X} \setminus \mathcal{H}$. This means that the alignment in the latent space will be performed between samples with a different target attribute, hence achieving invariance. In a SSL setting, to avoid the dependence on anomalous samples, we leverage data augmentation and data manipulation techniques [30] for approximating the distribution of the target attribute. For example, the appearance of tumors in brain MRIs can be approximated by employing random cutout, or Gaussian noise [18] (more details in the supplementary material).

Definition 4.3 (Target-Invariant Encoder). An encoder f preserves common patterns of the image if it is both instance-aligned and class-invariant.

The above definitions provide the theoretical support of the contrastive learning approach used for training the target-invariant encoder. Considering that the learning process of the encoder is based on a contrastive learning strategy that is instance-aligned and class-invariant, our encoder is expected to produce features which preserve common information of the input image, regardless of whether the image belongs to the background or target set.

¹ As representations are normalized, i.e. $\|f(x)\|_2 = 1$, then $\text{sim}(f(x), f(x^+)) = 1 - d(f(x), f(x^+))$ where d is a L2-distance function.

4.2. Conditional diffusion-based decoder

Our conditional diffusion decoder $p_\theta(x_{t-1}|x_t, z_{bg})$ takes as input the noisy sample x_t and the common feature $z_{bg} \in \mathbb{R}^{1 \times d} = f_\theta(x_0)$, a non-spatial vector of dimension d that encodes common patterns observed in the input sample, derived from the properties of the target-invariant encoder f_θ (Section 4.1). Our primary objective is to reconstruct the background version of the input image, preserving its common information. Hence, we deviate from the step-wise sampling process in Eq. (3), which is typically employed to generate new images from noise $p_\theta(x_{T:1}), x_T \sim \mathcal{N}(0, I)$. Instead, we directly estimate the input image $\hat{x}_0 \sim p_\theta(x_0|x_t, z_{bg})$ given the noisy sample x_t . This is achieved by revising Eq. (2), following [15], allowing prediction of the denoised observation, which is an estimation of x_0 given x_t :

$$g_\theta(x_t, t, z_{bg}) := (x_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(x_t, t, z_{bg})) / \sqrt{\alpha_t} \quad (8)$$

The model is then trained by minimizing the following objective function [17], which is a modified version of the MSE objective in Eq. (5):

$$\mathcal{L}_{diff} = \sum_{t=1}^T \mathbb{E}_{x_0, \epsilon_t} \left[\left\| \epsilon_\theta(x_t, t, z_{bg}) - \epsilon_t \right\|_2^2 \right] \quad (9)$$

We condition the U-Net by incorporating timestep and conditioning embeddings into each residual block through scaling and shifting of the normalized activations. Additional details are provided in the Supplementary Material.

5. Experimental setup

5.1. Datasets

Facial images. We first conduct proof-of-concept experiments on CelebA [31] dataset, containing 202,599 facial images with diverse attributes. We create a subset focusing on subjects wearing eyeglasses and those without accessories, resulting in 15,353 images divided into two distinct classes: (1) **Eyeglasses (EG)**: Images with the ‘Eyeglasses’ attribute and no other accessory-related attributes, used solely for evaluation as *target* images; (2) **No Eyeglasses (NEG)**: Images without the ‘Eyeglasses’ attribute or any accessory-related attributes, used for training as *background* images. Our target task with this dataset is to remove glasses from input images while preserving subject-common information. Facial images are chosen due to their ease of visual interpretation, as identity preservation can be readily assessed during reconstruction. More details on dataset partitions and preprocessing are provided in the supplementary material.

Brain MRIs. We evaluate our approach in brain MRIs for the task of tumor detection and localization. For a comprehensive comparison, we utilize the IXI² dataset as a background (healthy) reference for training, as in previous studies [18,20]. Evaluation is conducted in the Multimodal Brain Tumor Segmentation Challenge 2021 (BraTS21) [32–34] and the multiple sclerosis dataset (MSLUB) [35], featuring tumor and Multiple Sclerosis (MS) samples, respectively. Furthermore, we evaluate our model on the recently evaluation-only NOVA [10] dataset, comprising 906 brain MRI scans spanning 281 rare and diagnostically diverse pathologies. Following [18], we utilized T2-weighted images from all datasets across our experiments, as well as performed some pre-processing steps involving brain registration, skull stripping and downsampling. Further details on data pre-processing partitioning are provided in the supplementary material.

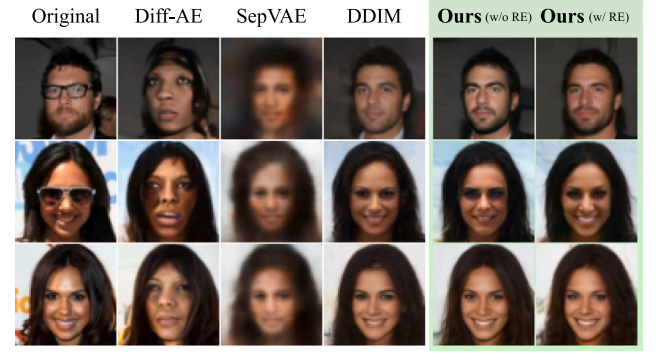


Fig. 2. Reconstruction results of CelebA images using different methods. In contrast to Diff-AE, SepVAE and DDIM, our method produces images with well-preserved common information from the input image.

5.2. Implementation details

As target-invariant encoder, we use a ResNet-50 backbone with a fully connected layer and a 128-dimensional output. For brain MRI experiments, we adapt the Spark2D [36] framework to incorporate the ϵ -InfoNCE loss. Notably, the encoder is pre-trained in the healthy training set. At the inference phase, we fine-tune it along with the denoising network. For the diffusion model, structured simplex noise [37] is applied. During training, timesteps $t \in [1, 1000]$ are sampled uniformly, while inference uses a fixed value of $t = 500$. The denoising network follows a U-Net architecture. Brain MRI volumes are processed on a slice-wise fashion, with slices uniformly sampled with replacement during training and all slices iterated over at test time to reconstruct the full volume. Training runs for up to 1600 epochs on NVIDIA A40 GPUs with Adam optimizer (learning rate of $1e-4$, and a batch size of 32). To assess statistical significance, we conduct hypothesis testing using the `scipy.stats` library. We used a t -test with a significance level of $\alpha = 0.05$ to verify the null hypothesis that two related methods share the same average metrics. Code is publicly available at <https://github.com/CristianoPatrício/unsupervised-contrastive-cond-diff>.

6. Results

We begin by validating the effectiveness of our method on the CelebA dataset. Subsequently, we evaluated its performance in anomaly detection and localization tasks, comparing it with established state-of-the-art UAD methods.

6.1. Proof-of-concept with CelebA dataset

Table 1 presents the reconstruction quality results of our method against baseline models [13,17,38]. These models were trained in the CelebA dataset using only *background* samples (NEG) and evaluated in both *background* (NEG) and *target* (EG) images. Evaluation metrics include SSIM ($\uparrow$), LPIPS ($\downarrow$), and MSE ($\downarrow$). Our method (w/ RE) achieves the highest SSIM (0.9763) and the lowest MSE (0.0003) and LPIPS (0.0059) when reconstructing *background* examples (NEG). Although our method (w/o RE) and DDIM produce similar results, they fail to capture common information and high-level details compared to our method (w/ RE), as visually demonstrated in Fig. 2. Furthermore, the supplementary material includes ablation studies analyzing reconstruction quality for different corruption timesteps $t \in [1, 1000]$ and various values of ϵ in ϵ -InfoNCE.

² <https://brain-development.org/ixi-dataset/>

Table 1

Reconstruction quality of methods trained on CelebA (training set) and tested on CelebA (test set).
NEG - No Eyeglasses, EG - Eyeglasses. $|\Delta| = |NEG - EG|$. RE - Random Erasing.

Model	SSIM $\uparrow$			MSE $\downarrow$			LPIPS $\downarrow$		
	NEG	EG	$ \Delta $	NEG	EG	$ \Delta $	NEG	EG	$ \Delta $
Diff-AE [17]	0.6287	0.6079	0.0208	0.0104	0.0119	0.0015	0.0847	0.1041	0.0194
SepVAE [13]	0.5796	0.4773	0.1023	0.0153	0.0268	0.0115	0.2912	0.3322	0.0410
DDIM [38]	0.9690	0.6568	0.3122	<u>0.0004</u>	0.0095	0.0091	0.0077	0.1385	0.1308
Ours (w/o RE)	<u>0.9756</u>	0.6737	0.3019	0.0003	0.0097	0.0094	<u>0.0060</u>	0.1198	0.1138
Ours (w/ RE)	0.9763	0.6701	<u>0.3062</u>	0.0003	0.0099	0.0096	0.0059	0.1213	<u>0.1154</u>

Table 2

Comparison of evaluated models in brain MRI datasets. Best results are shown in **bold**, and second-best results are underlined. * indicate results obtained by reproducing the method. All metrics are reported as mean $\pm$ standard deviation across 5 folds. Dice score and AUPRC are reported in percentage (%).

Model	Brats21		MSLUB		IXI
	DICE $\uparrow$	AUPRC $\uparrow$	DICE $\uparrow$	AUPRC $\uparrow$	$\ell_1(1e-3) \downarrow$
AE [11]	32.87 $\pm$ 1.25	31.07 $\pm$ 1.75	7.10 $\pm$ 0.68	5.58 $\pm$ 0.26	30.55 $\pm$ 0.27
VAE [11]	31.11 $\pm$ 1.50	28.80 $\pm$ 1.92	6.89 $\pm$ 0.09	5.00 $\pm$ 0.40	31.28 $\pm$ 0.71
SVAE [12]	33.32 $\pm$ 0.14	33.14 $\pm$ 0.20	5.76 $\pm$ 0.44	5.04 $\pm$ 0.13	28.08 $\pm$ 0.02
DAE [16]	37.05 $\pm$ 1.42	44.99 $\pm$ 1.72	3.56 $\pm$ 0.91	5.35 $\pm$ 0.45	10.12 $\pm$ 0.26
f-AnoGAN [5]	24.16 $\pm$ 2.94	22.05 $\pm$ 3.05	4.18 $\pm$ 1.18	4.01 $\pm$ 0.90	45.30 $\pm$ 2.98
DDPM* [37]	43.04 $\pm$ 1.26	47.98 $\pm$ 1.84	4.75 $\pm$ 1.65	6.12 $\pm$ 1.54	15.84 $\pm$ 1.30
pDDPM* [18]	47.14 $\pm$ 1.53	51.36 $\pm$ 1.75	9.69 $\pm$ 1.15	9.93$\pm$0.74	12.96 $\pm$ 0.97
cDDPM* [20]	<u>50.33$\pm$3.24</u>	<u>55.44$\pm$3.70</u>	<u>9.62$\pm$1.35</u>	<u>9.52$\pm$1.67</u>	<u>9.87$\pm$0.28</u>
Ours (w/o target-invariant constraint)	49.38 $\pm$ 2.42	53.88 $\pm$ 2.42	8.64 $\pm$ 1.24	8.47 $\pm$ 1.30	11.56 $\pm$ 0.21
Ours (full)	51.04$\pm$1.62	56.36$\pm$1.92	9.81$\pm$1.81	9.41 $\pm$ 1.44	9.75$\pm$0.44

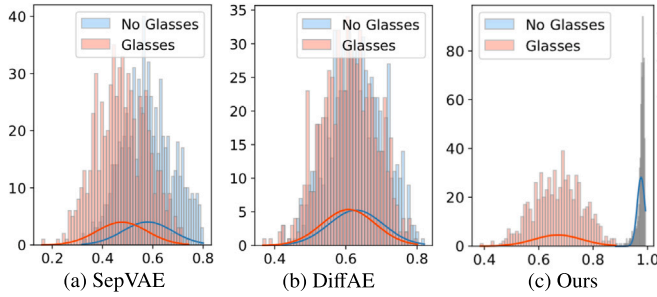


Fig. 3. Histograms of SSIM distributions for CelebA on NEG vs. EG classification. Our method shows well-separated distributions.

EG-NEG classification performance. Tab. 1 provides metrics for distinguishing between the two classes (NEG vs. EG), with the magnitude $|\Delta| = |NEG - EG|$ indicating distinctiveness (histogram plots shown in Fig. 3). Diff-AE and SepVAE exhibit overlapping distributions across metrics, while our method demonstrates well-separated distributions, yielding the highest ROC AUC score (see Fig. C.3 in the supplementary material).

Instance-alignment. Varying the margin ϵ in Eq. (7) can influence identification accuracy by affecting the separation of different classes within the latent space. We conduct an ablation study on ϵ to observe its impact on identification accuracy using a k -NN classifier with $k = 3$. Results depicted in Fig. 4 (right) show that the highest accuracy is achieved when ϵ is chosen from the ranges [0.5, 2.5] with the random erasing (RE) transformation and [0.15, 1.0] without it.

Importance of class-invariance. We conduct a linear probing analysis to assess the encoder’s class-invariance (Definition 4.2). The objective is to evaluate the encoder’s capability to distinguish between EG and NEG samples. Subsequently, a logistic regression classifier was trained on the extracted image latent features. Results in Fig. 4 (left) show lower target accuracy with RE, suggesting stronger common information preservation. However, instance classification performs worse with RE, possibly due to built-in augmentations in contrastive learning.

6.2. Application on brain MRI datasets

Table 2 compares different UAD methods on three brain MRI datasets using evaluation metrics such as DICE score, AUPRC, and ℓ_1 reconstruction error (calculated only on healthy IXI dataset). We reproduce results for DDPM [37], pDDPM [18], and cDDPM [20], while results for other methods are taken from [18].

Reconstruction quality. To assess overall reconstruction quality, we use the held-out test set from the healthy IXI dataset and report the ℓ_1 -error, calculated as the average absolute difference between the reconstructed images and their corresponding inputs. As shown in Table 2, our approach achieves competitive performance with cDDPM and significantly outperforms all other baselines ($p < 0.05$). These results suggest that our model effectively preserves healthy anatomy from the input image during reconstruction. This can be attributed to the influence of the ϵ -InfoNCE regularization (Eq. (7)) within the target-invariant encoder, which improves the representation space with regard to the anatomy. Ablating the target-invariant constraint from our encoder leads to suboptimal reconstruction quality, with the average ℓ_1 error increasing from 9.75 to 11.56 on IXI, as shown in Table 2 (last two rows). Furthermore, Fig. 5 illustrates the anomaly maps obtained with and without the target-invariant constraint.

Tumor segmentation. To evaluate segmentation performance in the context of UAD, we use all unhealthy test sets and report two metrics: (i) Dice score and (ii) Area Under Precision–Recall Curve (AUPRC). As shown in Table 2, our method achieves performance comparable to the top baseline (cDDPM) and significantly outperforms all other baselines ($p < 0.05$), except pDDPM in MSLUB. We achieve the best results with $\epsilon = 0.5$ (Eq. (7)). Additional results with other ϵ values are available in the supplementary material. Analogous to the experimental findings on the CelebA dataset, this model can be used to distinguish between healthy and unhealthy MRIs by analyzing the generated saliency maps, as shown in Fig. 1 (bottom-right). For instance, if the difference between a reconstructed brain MRI and its original counterpart shows high-intensity values across large regions, this may indicate presence of tumor and can be flagged for further review by a radiologist. This strategy can be seen as a form of clinical triage, while also contributing to the model’s explainability by providing a rationale for its decisions. Examples of reconstructions and corresponding tumor segmentation (anomaly maps) are shown in Fig. 6.

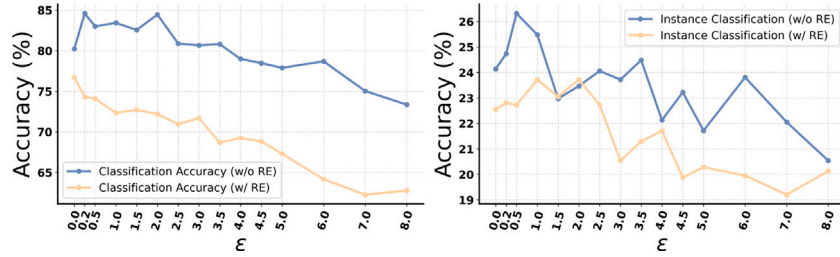


Fig. 4. Left: Classification accuracy (NEG vs. EG) for different values of ϵ . Right: Instance classification (accuracy in %) for different values of ϵ .

Table 3

Localization performance on NOVA benchmark. Reported metrics include standard object detection measures (mean Average Precision at multiple IoU thresholds), detection accuracy (ACC50), number of true positives (TP30), and number of false positives (FP30). Best results are shown in **bold**, and second-best results are underlined. * indicate results obtained by reproducing the method. All metrics are reported as mean $\pm$ standard deviation across 5 folds.

Model	mAP30	mAP50	mAP50-95	ACC50	TP30	FP30 $\downarrow$
DDPM* [37]	17.50 $\pm$ 4.05	8.23 $\pm$ 2.68	2.75 $\pm$ 1.06	7.36 $\pm$ 2.42	63/394	1517
pDDPM* [18]	0.30 $\pm$ 0.19	0.12 $\pm$ 0.15	0.03 $\pm$ 0.04	0.10 $\pm$ 0.12	1/394	1445
cDDPM* [20]	11.51 $\pm$ 4.01	5.19 $\pm$ 2.10	1.51 $\pm$ 0.69	4.72 $\pm$ 2.02	44/394	1423
Ours (all images)	4.50 $\pm$ 0.33	2.34 $\pm$ 0.10	0.79 $\pm$ 0.05	7.68 $\pm$ 0.58	159/1068	20741
Ours (only T2W + all bboxes)	5.42 $\pm$ 0.60	2.85 $\pm$ 0.17	0.96 $\pm$ 0.09	9.29 $\pm$ 1.02	71/394	7958
Ours (only T2W + 1 largest bbox)	17.13 $\pm$ 1.76	8.66 $\pm$ 0.91	3.15 $\pm$ 0.27	7.36 $\pm$ 0.77	57/394	270
Ours (only T2W + 3 largest bboxes)	18.52 $\pm$ 2.06	9.38 $\pm$ 0.98	3.32 $\pm$ 0.28	8.98 $\pm$ 0.92	69/394	892
Ours (only T2W + 5 largest bboxes)	18.57 $\pm$ 2.07	9.41 $\pm$ 1.00	3.33 $\pm$ 0.29	9.29 $\pm$ 1.02	71/394	1492
Ours (only T2W + 8 largest bboxes)	18.56 $\pm$ 2.08	9.41 $\pm$ 1.00	3.33 $\pm$ 0.29	9.29 $\pm$ 1.02	71/394	2348

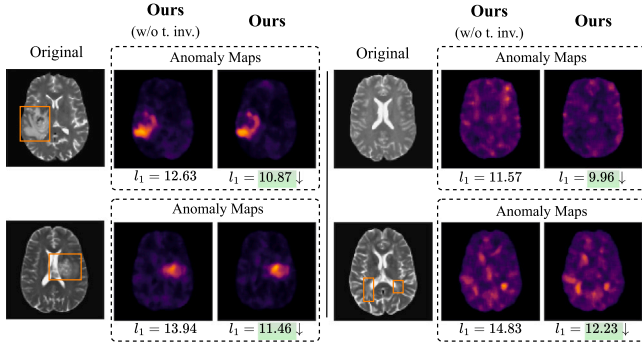


Fig. 5. Anomaly maps generated by our model without the target-invariant constraint (w/o t. inv.) versus our full approach.

Anomaly localization on NOVA benchmark. Anomaly localization involves identifying and detect abnormal regions within brain MRIs. This task holds significant clinical importance, as many medical errors arise from failing to detect a pathology altogether. The objective is to predict one or more bounding boxes per image that correspond to abnormal regions, using annotations provided by radiologists as ground truth. As shown in Table 3, our method outperforms cDDPM and pDDPM with statistical significance ($p < 0.05$), except for DDPM, where the improvements are marginal. Specifically, we achieve these results by considering only T2-weighted images and selecting the top-5 predicted largest bounding-boxes. When considering more than the top-5 bounding-boxes the performance tends to saturate while the number of false positives increases (1492 to 2348). We empirically found that including all modalities available in the NOVA dataset degrades performance, which is expected, as the models were pre-trained only on T2-weighted images from IXI. Compared to pDDPM and cDDPM, our pre-training strategy, which is based on contrastive learning, appears to be more robust than simply using data augmentation. As illustrated in Fig. 7, our model is capable of detecting larger and more accurately localized anomalies, while baseline methods tend to produce smaller and more fragmented predictions. For example, pDDPM underestimates

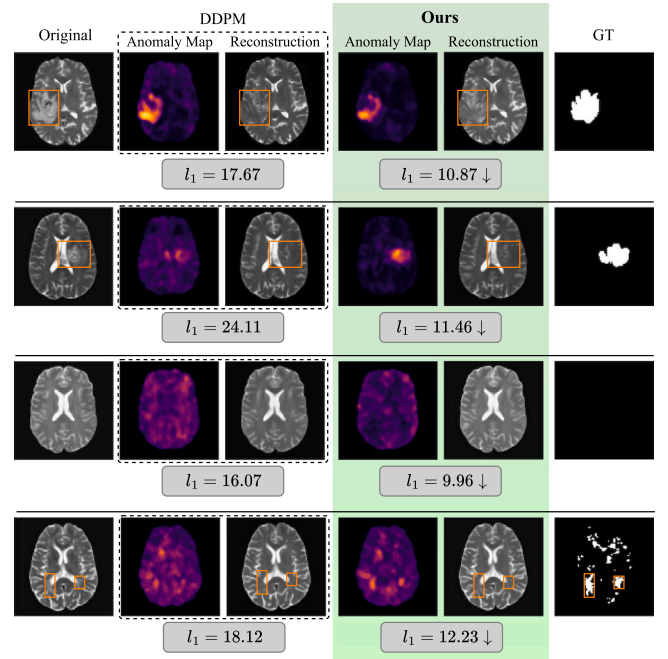


Fig. 6. Examples of brain MRI reconstructions and corresponding anomaly maps on BraTS21 (first two rows), IXI (third row), and MSUB (last row). GT: Ground truth annotation. Our method consistently produces reconstructions that better preserve healthy anatomy, leading to lower reconstruction error (l_1) compared to the baseline.

lesion size, likely due to its patching strategy, which may rely on suboptimal patch sizes. Notably, considering the top-3 largest bounding boxes our method reduces the number of false negatives by nearly 40%, suggesting that future work could explore post-processing strategies to remove isolated or implausible detections, thus improving trust in clinical workflows. These results highlight the effectiveness of our

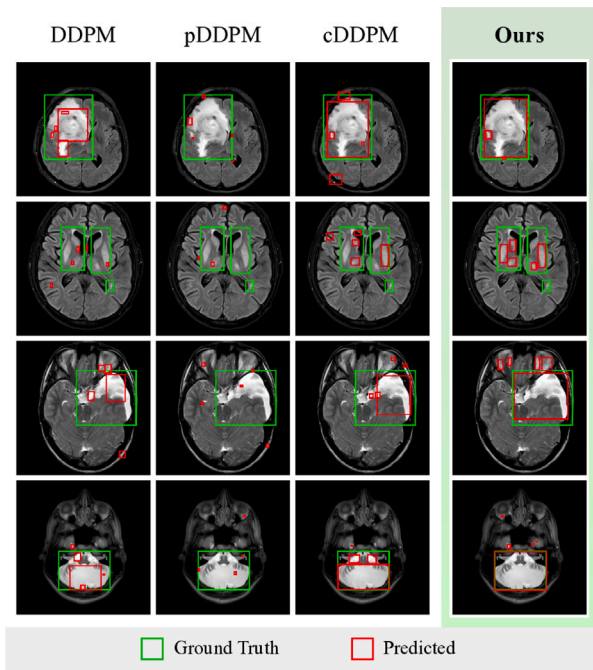


Fig. 7. Examples of model predictions for anomaly localization on NOVA. Compared to baseline models, our method produces more precise and spatially focused bounding boxes that better align with ground-truth annotations (green). Each row corresponds to a distinct pathology.

methodology, particularly given that NOVA is considered a challenging and realistic evaluation-only benchmark dataset.

7. Conclusion and future work

In this work, we proposed an unsupervised framework for anomaly detection in brain MRIs that integrates contrastive learning with conditional diffusion models. Our method reconstructs a healthy version of input images while preserving common patterns, enabling both accurate classification and precise and interpretable localization of anomalies. By relying exclusively on healthy data, our approach mitigates the reliance on scarce or poorly annotated unhealthy samples. Moreover, the use of saliency maps derived from reconstruction errors provides intuitive visual evidence that enhances model interpretability, addressing key challenges in clinical adoption. Providing clinicians with reconstructed healthy images for comparison could help reduce diagnostic errors by enhancing visualization of abnormalities and supporting the differential diagnosis process. Our results demonstrate the effectiveness of the proposed method, which achieves competitive performance compared to existing UAD methods and ranks first in anomaly localization on the challenging NOVA benchmark. Nevertheless, our work has limitations and open challenges that are important to acknowledge. The exclusive use of 2D T2-weighted images during training may limit generalization to other structural modalities. In addition, proper clinical validation is needed to assess the potential of our approach to reduce radiologists' workload and to facilitate early and fast diagnosis. Future work could explore integrating multimodal imaging data and investigate the model's robustness across diverse pathological conditions. In addition, ethical aspects regarding potential misdiagnosis and privacy concerns remain crucial for safe adoption in practice.

CRedit authorship contribution statement

Cristiano Patrício: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization.

Carlo Alberto Barbano: Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Attilio Fian-drotti:** Supervision, Funding acquisition. **Riccardo Renzulli:** Writing – original draft. **Marco Grangetto:** Supervision. **Luís F. Teixeira:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization. **João C. Neves:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

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

This work was funded by the Portuguese Foundation for Science and Technology (FCT) under the PhD grant “2022.11566.BD” (<https://doi.org/10.54499/2022.11566.BD>), and supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINC) with the financial support of FCT/IP.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.patrec.2025.11.014>.

Data availability

The code is publicly available and the link is provided in the manuscript.

References

- [1] C. Patrício, J.C. Neves, L.F. Teixeira, Explainable deep learning methods in medical image classification: A survey, *ACM Comput. Surv.* 56 (4) (2023) 1–41.
- [2] J.Y. Zou, D.J. Hsu, D.C. Parkes, R.P. Adams, Contrastive Learning Using Spectral Methods, *Adv. Neural Inf. Process. Syst. (NeurIPS)* 26 (2013).
- [3] A. Abid, M.J. Zhang, V.K. Bagaria, J. Zou, Exploring patterns enriched in a dataset with contrastive principal component analysis, *Nat. Commun.* 9 (1) (2018) 2134.
- [4] E. Weinberger, N. Beebe-Wang, S.-I. Lee, Moment matching deep contrastive latent variable models, 2022, arXiv preprint [arXiv:2202.10560](https://arxiv.org/abs/2202.10560).
- [5] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks, *Med. Image Anal.* 54 (2019) 30–44.
- [6] F. Carton, R. Louiset, P. Gori, Double InfoGAN for Contrastive Analysis, in: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2024, pp. 172–180.
- [7] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, *Adv. Neural Inf. Process. Syst. (NeurIPS)* 33 (2020) 6840–6851.
- [8] I. Adamchik, S.R. Kantelhardt, H.-J. Wagner, M. Burbelko, Artificial intelligence can help detecting incidental intracranial aneurysm on routine brain MRI using TOF MRA data sets and improve the time required for analysis of these images, *Neuroradiology* (2024) 1–10.
- [9] D.S.W. Lim, A. Makmur, L. Zhu, W. Zhang, A.J. Cheng, D.S.Y. Sia, S.E. Eide, H.Y. Ong, P. Jagmohan, W.C. Tan, et al., Improved productivity using deep learning-assisted reporting for lumbar spine MRI, *Radiology* 305 (1) (2022) 160–166.
- [10] C.I. Bercea, J. Li, P. Raffler, E.O. Riedel, L. Schmitzer, A. Kurz, F. Bitzer, P. Roßmüller, J. Canisius, M.L. Beyrle, et al., NOVA: A benchmark for anomaly localization and clinical reasoning in brain MRI, 2025, arXiv preprint [arXiv:2505.14064](https://arxiv.org/abs/2505.14064).
- [11] C. Baur, S. Denner, B. Wiestler, N. Navab, S. Albarqouni, Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study, *Med. Image Anal.* 69 (2021) 101952.
- [12] F. Behrendt, M. Bengs, D. Bhattacharya, J. Krüger, R. Opfer, A. Schlaefer, Capturing inter-slice dependencies of 3D brain MRI-scans for unsupervised anomaly detection, in: *Proceedings of the Medical Imaging with Deep Learning, MIDL*, 2022.

- [13] R. Louiset, E. Duchesnay, A. Grigis, B. Dufumier, P. Gori, SepVAE: a contrastive VAE to separate pathological patterns from healthy ones, 2023.
- [14] S. Marina, Improving Diagnostic Accuracy of Brain Tumor MRI Classification Using Generative AI and Deep Learning Techniques, *Babylon. J. Artif. Intell.* 2025 (2025) 55–63.
- [15] J. Song, C. Meng, S. Ermon, Denoising Diffusion Implicit Models, in: *Proceedings of the International Conference on Learning Representations, ICLR*, 2021.
- [16] A. Kascenas, N. Pugeault, A.Q. O’Neil, Denoising autoencoders for unsupervised anomaly detection in brain MRI, in: *Proceedings of the International Conference on Medical Imaging with Deep Learning, MIDL*, 2022, pp. 653–664.
- [17] K. Preechakul, N. Chatthee, S. Wizadwongsa, S. Suwajanakorn, Diffusion Autoencoders: Toward a Meaningful and Decodable Representation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022, pp. 10619–10629.
- [18] F. Behrendt, D. Bhattacharya, J. Krüger, R. Opfer, A. Schlaefer, Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI, in: *Proceedings of the International Conference on Medical Imaging with Deep Learning, MIDL, PMLR*, 2024, pp. 1019–1032.
- [19] H. Iqbal, U. Khalid, C. Chen, J. Hua, Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model, in: *Proceedings of the International Workshop on Machine Learning in Medical Imaging, MLMI, Springer*, 2023, pp. 372–381.
- [20] F. Behrendt, D. Bhattacharya, R. Mieling, L. Maack, J. Krüger, R. Opfer, A. Schlaefer, Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris, *Comput. Biol. Med.* 186 (2025) 109660.
- [21] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer*, 2015, pp. 234–241.
- [22] C.M. Bishop, H. Bishop, *Deep Learning: Foundations and Concepts*, Springer, 2024.
- [23] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality Reduction by Learning an Invariant Mapping, *CVPR*, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1735–1742.
- [24] A.v.d. Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, 2018, *arXiv preprint arXiv:1807.03748*.
- [25] Y. Tian, D. Krishnan, P. Isola, Contrastive Multiview Coding, in: *Proceedings of the European Conference on Computer Vision, ECCV*, 2020, pp. 776–794.
- [26] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015, pp. 815–823.
- [27] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *Proceedings of the International Conference on Machine Learning, ICML*, 2020, pp. 1597–1607.
- [28] P. Khosla, et al., Supervised contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), in: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18661–18673.
- [29] C.A. Barbano, B. Dufumier, E. Tartaglione, M. Grangetto, P. Gori, Unbiased supervised contrastive learning, in: *Proceedings of the International Conference on Learning Representations, ICLR*, 2023.
- [30] B. Dufumier, C.A. Barbano, R. Louiset, E. Duchesnay, P. Gori, Integrating prior knowledge in contrastive learning with kernel, in: *Proceedings of the International Conference on Machine Learning, ICML*, 2023.
- [31] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: *Proceedings of International Conference on Computer Vision, ICCV*, 2015, pp. 3730–3738.
- [32] B.H. Menze, et al., The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), *IEEE Trans. Med. Imaging* 34 (10) (2014) 1993–2024.
- [33] S. Bakas, et al., Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, *Sci. Data* 4 (1) (2017) 1–13.
- [34] U. Baid, et al., The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification, 2021, *arXiv preprint arXiv:2107.02314*.
- [35] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, Ž. Špiclin, A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus, *Neuroinformatics* 16 (2018) 51–63.
- [36] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, Z. Yuan, Designing bert for convolutional networks: Sparse and hierarchical masked modeling, 2023, *arXiv preprint arXiv:2301.03580*.
- [37] J. Wyatt, A. Leach, S.M. Schmon, C.G. Willcocks, AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 649–655.
- [38] P. Dhariwal, A. Nichol, Diffusion Models Beat GANs on Image Synthesis, *Adv. Neural Inf. Process. Syst. (NeurIPS)* 34 (2021) 8780–8794.