Feature Modeling for Anomaly Detection in Neuroimaging Data

Denis Kovačević
Student ID: 6707752
MSc Machine Learning
University of Tübingen

Abstract

We propose a 3D anomaly detection pipeline for neuroimaging that combines contrastive representation learning with voxel-wise statistical modeling. A 3D convolutional network is pre-trained on brain MRI volumes using a SimCLR-style contrastive framework, after which multi-scale features from selected layers are spatially aligned and concatenated. For each voxel, a Gaussian distribution is fitted to healthy training data using Welford's algorithm, and anomalies are scored via Mahalanobis distance followed by quantile-based sigmoid normalization. Evaluated on an anomalous-only validation set of registered brain MRIs, the method achieves a Dice score of 0.112, highlighting both the voxel-wise explainability and the current limitations of this approach.

1 Introduction

Anomaly detection in neuroimaging is a crucial step in identifying structural abnormalities that may indicate neurological disorders such as multiple sclerosis, Alzheimer's disease, or brain tumors. Unlike supervised lesion segmentation methods, which rely on large, manually annotated datasets, anomaly detection aims to identify abnormal regions without requiring pathology labels during training. This is especially relevant in clinical research, where labeled datasets are scarce, annotation is time-consuming, and anomalies can vary substantially between patients.

Deep learning methods for anomaly detection have seen rapid progress in the 2D domain, particularly in industrial inspection tasks. In this setting, models learn the distribution of "normal" images and detect deviations in unseen samples [3, 5]. These approaches often combine deep feature extraction with probabilistic modeling to generate anomaly maps. However, extending such methods to 3D brain MRI introduces significant challenges: the higher dimensionality increases computational cost, variability between subjects complicates voxel-wise modeling, and full-volume context is key to accurate detection.

Self-supervised contrastive learning methods such as SimCLR [4] have shown promise in learning image representations without labels. In neuroimaging, such methods can be leveraged to learn features that capture healthy anatomical structures across individuals. Previous works in medical imaging have explored related ideas, combining pretraining with downstream anomaly detection [25], but voxel-wise statistical modeling in feature space remains underexplored for 3D MRI.

In this work, we adapt the principles of feature-based anomaly detection to volumetric brain MRI. Our approach uses a 3D convolutional network with a later discarded projection head pretrained in a SimCLR-like framework. Features from selected layers are spatially aligned and concatenated, and per-voxel Gaussian statistics are fitted using training data, all from healthy samples. At test time, anomalies are scored via the Mahalanobis distance between voxel features and their learned distributions, producing voxel-level heatmaps.

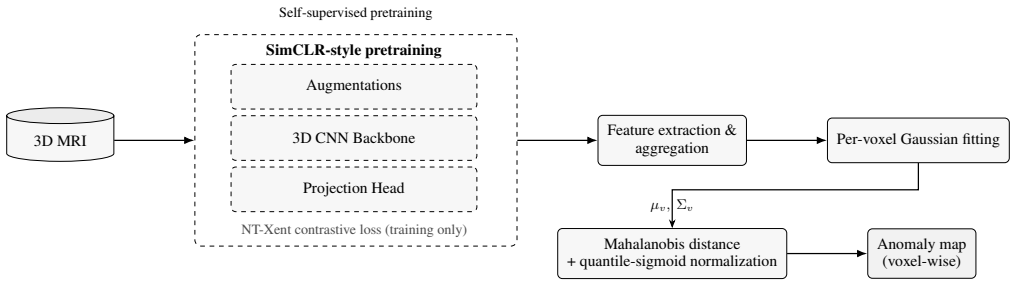

Figure 1: Simplified overview of the 3D anomaly detection pipeline. A 3D CNN is pretrained with SimCLR (left). After freezing the backbone, features are extracted and aggregated, per-voxel Gaussian parameters are learned on healthy data (training only), and inference computes voxel-wise Mahalanobis scores with quantile–sigmoid normalization to yield an anomaly map.

2 Methods

Our anomaly detection framework consists of three main stages: (i) self-supervised pretraining of a 3D convolutional backbone using a SimCLR-like contrastive learning scheme, (ii) feature extraction and per-voxel Gaussian modeling on healthy training data, and (iii) inference via Mahalanobis-based anomaly scoring with quantile–sigmoid normalization. The overall pipeline is illustrated in Figure 1.

2.1 Self-Supervised Pretraining

We employ a 3D convolutional neural network (CNN) as the feature backbone. The network is pretrained on healthy brain MRI volumes using a SimCLR-inspired self-supervised learning framework [4]. Unlike the original formulation we initially adopted this setup but found that applying the full augmentation pipeline to only one view while keeping the other unaltered led to better downstream anomaly detection performance in our setting.

To address the small batch size limitation inherent in 3D medical imaging, we integrate a momentum-updated *memory bank* [8] that stores a large set of L2-normalized feature vectors from past iterations. This mechanism effectively increases the pool of negative samples beyond the current mini-batch, which is crucial for stable contrastive learning in low-batch regimes. The memory bank is updated after each optimization step using a momentum coefficient of 0.99, following the approach of MoCo.

For each input volume, two positive views are generated. Unlike the original formulation, where both views are augmented, we found that applying the full augmentation pipeline to only one view while keeping the other unaltered led to better downstream anomaly detection performance in our setting. The augmentation pipeline, implemented with `TorchIO` [16], consists of (D, H, W being depth, height and width dimensions): For each input volume, two positive views are generated: one unaltered and one augmented. The final augmentation pipeline, implemented with `TorchIO` [16], consists of (D, H, W being depth, height, and width dimensions):

- **Geometric:** `CropOrPad` to $(D, H+16, W+16)$ for variability in random cropping; random flips along sagittal/coronal axes (prob. 0.5); `RandomAffine` with scales $(0.8, 1.3)$, rotation $\pm 15°$, translation ± 12 voxels (prob. 0.9); `RandomElasticDeformation` with 7 control points and max displacement 6 (prob. 0.4); `RandomMotion` with $8°$ rotation and 3-voxel translation (prob. 0.2).
- **Intensity:** `RandomGamma` with $\log \gamma \in (-0.5, 0.5)$ (prob. 0.6); `RandomNoise` with $\sigma \in (0, 0.04)$ (prob. 0.5); `RandomBiasField` with coefficients 0.5 (prob. 0.3).
- **Small corruptions:** `RandomBlur` with $\sigma \in (0.5, 1.0)$ (prob. 0.2); `RandomSwap` with patch size $(\min(28, D), 1, 1)$ for 12 iterations (prob. 0.4).
- **Final preprocessing:** `CropOrPad` back to (D, H, W); `RescaleIntensity` to $(0, 1)$; `Lambda` to `float32`.

The augmented and unaugmented views are processed by the shared 3D CNN backbone, followed by a two-layer projection head mapping features into a low-dimensional contrastive space. The model is

trained with the normalized temperature-scaled cross-entropy (NT-Xent) loss, maximizing agreement between features from the same source volume while pushing apart features from different volumes. After pretraining, the projection head is discarded and the backbone weights are frozen.

2.2 Feature Extraction and Gaussian Modeling

In the final CNN we ended up using, there were 5 layers. We got the best results when extracting features from the layers `"0"`, `"1"`, `"2"`, and `"3"`. The selected source layers are resampled via trilinear interpolation to match the spatial resolution of the target layer `"2"`. The aligned feature maps are concatenated along the channel dimension, resulting in a combined feature tensor of size $(C_{\text{sum}}, D', H', W')$, where C_{sum} is the sum of channel counts from the selected layers.

To model healthy anatomy, we fit a multivariate Gaussian distribution to the features at each voxel location v using the extracted features from the training set. Let $\mathbf{x}_v \in \mathbb{R}^{C_{\text{sum}}}$ denote the concatenated feature vector at voxel v. We estimate the mean μ_v and covariance Σ_v for each voxel using a batch adaptation of Welford's online algorithm [22], which provides numerically stable and memory-efficient updates without storing all feature vectors.

2.3 Inference and Anomaly Scoring

During inference, a test volume is processed by the frozen backbone and the same feature extraction pipeline to obtain $\mathbf{x}_v$ for each voxel v. The anomaly score is computed as the Mahalanobis distance:

$$d_v(\mathbf{x}) = \sqrt{(\mathbf{x}_v - \mu_v)^\top \Sigma_v^{-1} (\mathbf{x}_v - \mu_v)}, \tag{1}$$

yielding a coarse-resolution anomaly map (D', H', W').

First, the map is upsampled to the original MRI resolution. We then apply quantile-based sigmoid normalization to rescale scores into $[0, 1]$, reducing the effect of skewness and extreme values. Let p_{low} and p_{high} denote the lower and upper percentiles (5% and 95%, respectively) computed from a random subsample of scores, and $\text{IQR} = p_{\text{high}} - p_{\text{low}}$. After $\log(1 + x)$ transformation, scores are standardized by subtracting the median and dividing by IQR, and finally mapped to $(0, 1)$ using the sigmoid function. The classification threshold is selected to maximize the Dice score on the validation set.

3 Experiments

We evaluate our proposed pipeline on diverse neuroimaging datasets and benchmark it against recent voxel-wise anomaly detection methods. This section describes the datasets, preprocessing steps, and baseline approaches used for comparison.

3.1 Dataset

For training, we use the Cambridge Centre for Ageing and Neuroscience dataset (CAMCAN)[20], the Human Connectome Project (HCP) Young Adult (S1200) dataset[21], the HCP Development dataset[19] and the IXI dataset[12]. The validation dataset is aggregated from the 2020 version of the Multimodal Brain Tumor Segmentation (BraTS) Challenge[2], the second version of the Anatomical Tracings of Lesions After Stroke (ATLAS)[11], the MSSEG dataset and Ljubljana MS lesion datasets, which are both part of the 2021 Shifts challenge[13], the white matter hyperintensities (WMH) segmentation[9] challenge from MICCAI 2017 and contains 92 volumes in total.

3.2 Preprocessing

For preprocessing the training data and parts of the validation dataset, a pipeline similar to the UK Biobank pipeline was applied[1]. First skull-stripping is applied by using an inverse non-linear registration. Then the skull-stripped scans are rigidly registered to the SRI24 ATLAS / template[17] and interpolated to 1mm isovoxel space.

3.3 Baselines

We compare our method against multiple unsupervised voxel-wise baselines all trained on the slices of the described training dataset:

- **Reconstruction-by-Inpainting Anomaly Detection** (RIAD)[24] : A self-supervised approach trained to predict masked connected regions of varying size in an image. It utilizes multiple reconstruction and averages the reconstruction losses using a novel multi-scale gradient magnitude similarity.
- **Iterative Spatial Mask-Refining** (IterMask)[10] : A self-supervised approach utilizing two models that are trained with multiple masking strategies to predict the original image. Anomalies are detected from voxel-wise reconstruction residuals.
- **Aggregated Normative Diffusion** (ANDi)[7] : A Diffusion Model trained with Gaussian pyramidal noise to effectively reason about low-frequency anomalies during the complete denoising process. Anomalies are detected from voxel-wise reconstruction residuals.
- **Diffusion-Inspired Synthetic Restoration** (Disyre)[15] : A Diffusion Model inspired method that uses synthetic anomalies to generate a corruption process and learns to reverse it. Anomalies are detected from voxel-wise reconstruction residuals.
- **Feature-Autoencoders** (FAE)[14]: A Convolutional Neural Network with an autoencoder structure. It takes embeddings from different layers of a pre-trained ResNet on ImageNet as input and is learned to reconstruct them using the Structural Similarity Index Measure (SSIM) as the loss function.
- **Unified Model for Multi-class Anomaly Detection** (UniAD)[23]: A Transformer network utilizing embeddings from EfficientNet that is trained to reconstruct these embeddings for the healthy data.
- **Reverse Distillation** (RD)[6]: A Knowledge Distillation framework for the task of anomaly detection. A pre-trained encoder is used as the teacher network, a trainable bottleneck embedding module is used to map the embeddings of the teacher to a more compact code, and a decoder is trained to reconstruct the embeddings of the teacher out of the bottleneck embedding.
- **PatchCore** [18]: Utilizes a memory bank of pre-processed embeddings from a pretrained CNN for anomaly detection.

4 Results

4.1 Main Findings

We evaluate voxel-level localization on an anomalous-only validation set (registered brain MRIs with lesion masks). Our method achieves a Dice similarity coefficient (DSC) of **0.111693**. For context only, when evaluated on a combined set that includes healthy and unhealthy volumes (with the same thresholding protocol), the Dice drops to **0.055937**. Because Dice is undefined on healthy volumes without any true positives, including many such volumes effectively penalizes localization even if anomaly scores remain calibrated; we therefore report the anomalous-only result as the primary metric for the rest of the paper. We also show the different distribution of representation of healthy and anomalous voxels as shown in Figure 2.

4.2 Context with Prior Art

To contextualize our results, we plot the Dice scores reported by a set of unsupervised anomaly detection baselines. While these methods are evaluated under different experimental protocols (mostly slice-based), the comparison (Figure 3) provides a rough reference point for our voxel-wise result of **0.112**.

4.3 Training Loss (NT-Xent)

We tracked the SimCLR-style NT-Xent objective throughout pretraining (Figure 4). As is standard for contrastive learning, we use this curve as a *convergence and stability diagnostic* (e.g., to detect

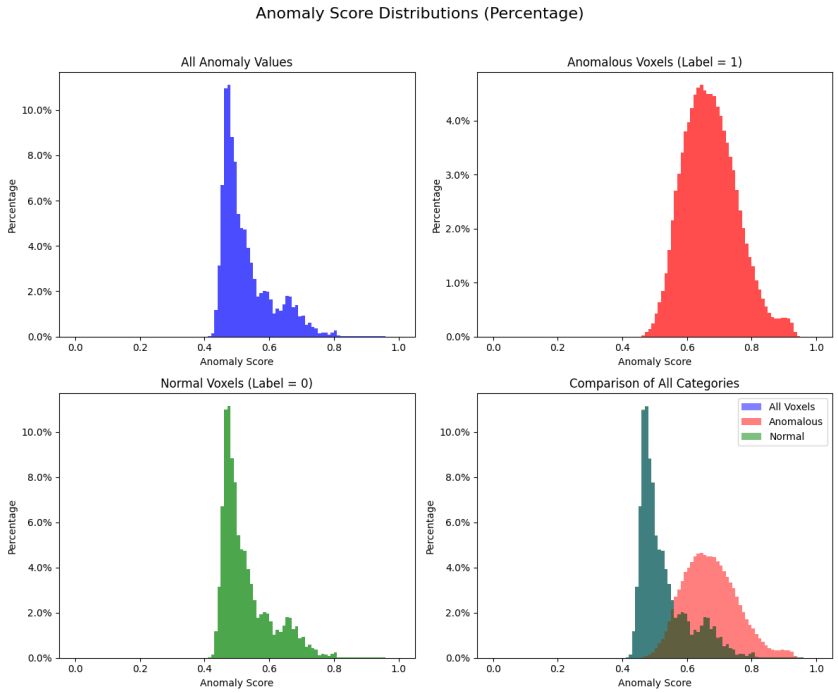

Figure 2: Training NT-Xent loss during self-supervised pretraining (lower is better) on different experiments. The curve is used as a stability and convergence diagnostic; its absolute value is not directly comparable across runs due to differences in hyperparameters.

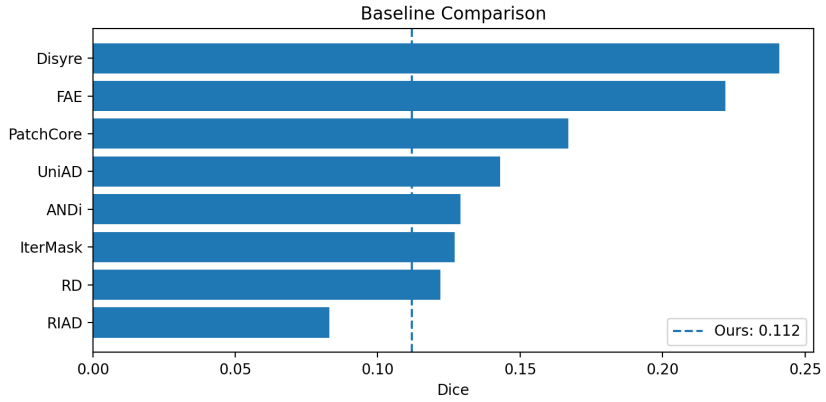

Figure 3: Comparison of Dice scores against unsupervised voxel-wise anomaly detection baselines. Our method (dashed line) achieves a Dice score of 0.112, which, while below the strongest reconstruction-based and diffusion-inspired baselines, remains competitive with several methods.

divergence, optimizer issues, or memory-bank mishandling), not as a comparable performance metric across runs: the absolute loss value depends on temperature, effective negative count (mini-batch size plus memory-bank content), and augmentation strength. We can see that in every run the loss decreases substantially, approaching 0. However, that doesn't give us definite information about the final performance of the pipeline, only an idea.

4.4 Explainability via Mahalanobis Heatmaps

A key advantage of modeling a Gaussian per voxel in feature space is the direct explainability of the anomaly score. For each voxel v, we compute a Mahalanobis distance d_v to the healthy

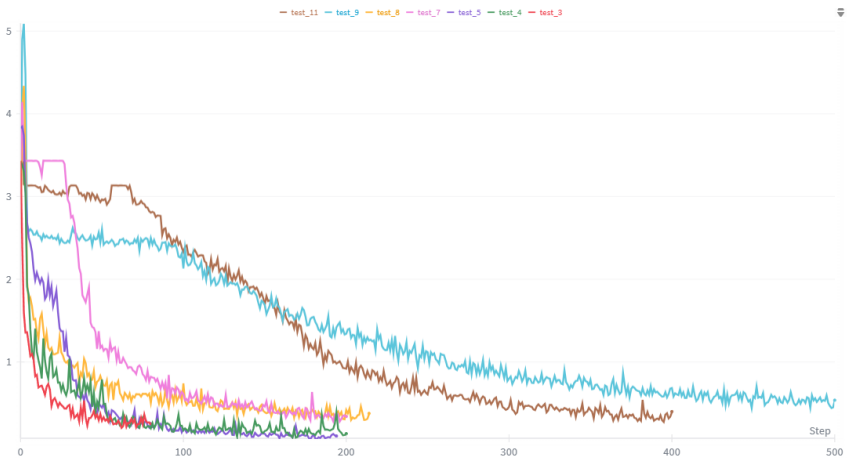

Figure 4: Training NT-Xent loss during self-supervised pretraining (lower is better) on different experiments. The curve is used as a stability and convergence diagnostic; its absolute value is not directly comparable across runs due to differences in hyperparameters.

distribution parameters (μ_v, Σ_v) learned from the training set. The resulting 3D field of distances can be visualized as a heatmap overlaid on the MRI (Figure 5), making it explicit *where* the volume deviates from healthy anatomy and, through the magnitude of d_v, *how strongly* it deviates.

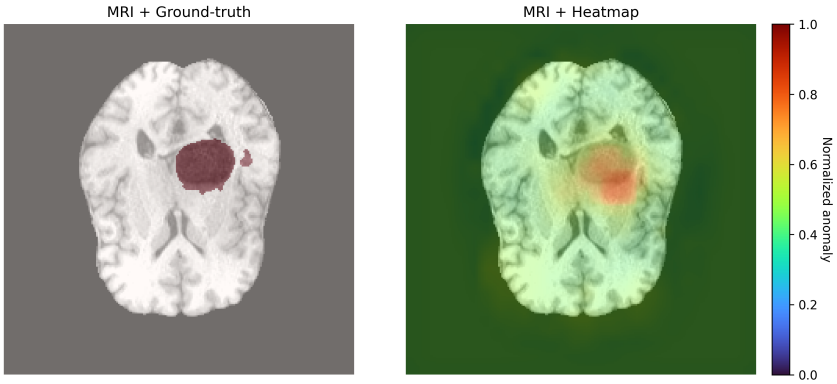

Figure 5: Qualitative examples: axial slices with ground-truth masks (left) and corresponding Mahalanobis heatmaps (right). The voxel-wise distance d_v highlights regions whose features are unlikely under the healthy per-voxel distributions.

5 Discussion

The results in Section 4 highlight both the strengths and current limitations of the proposed pipeline. In this section, we analyze the likely causes of performance gaps, outline concrete ways to improve this specific approach, and give final remarks.

5.1 Failure Analysis

While the proposed pipeline produces voxel-wise heatmaps which produce explanations, the overall Dice score of 0.112 remains low. Several factors likely contribute:

- **Representation limitations:** Small effective batch sizes in 3D contrastive learning reduce the diversity of negative samples, even with a memory bank.

- **Alignment:** Imperfect voxel-wise alignment of the original MRIs across subjects, even after registration, can introduce mismatches that degrade the quality of the per-voxel Gaussian statistics.
- **Resolution:** Down-sampling to the target feature map resolution and subsequent up-sampling of anomaly maps can lead to information loss and spatial blurring, particularly for small lesions.
- **Thresholding:** A single global threshold may fail to adapt to variations in lesion size, intensity, and location across subjects.
- **Augmentation:** In a lot of the samples, the final heatmap looks similar to the ground truth, with a bit of a shift. This might be due to geometrical augmentations, which could make the network invariant to shifts.

5.2 Directions to Improve This Pipeline

Several targeted modifications could strengthen the current approach:

- **Alignment-aware modeling:** Conditioning the Gaussian statistics on anatomical regions or volume patches could reduce variability in healthy feature distributions.
- **Multi-resolution scoring:** Computing anomaly scores at multiple feature resolutions and combining them may better capture both coarse and fine lesion structures.
- **Augmentation tuning:** While our best results came from augmenting only one view, further refinement of the augmentation set might improve representation quality without destroying subtle anatomical cues.
- **Adaptive thresholding:** Learning per-structure or per-subject thresholds from a small calibration set could help address lesion size and intensity variability.

Concluding Remarks. This work shows that self-supervised 3D feature learning combined with voxel-wise Gaussian modeling can produce explanations via anomaly maps without lesion labels. However, the results also indicate that precise alignment and resolution handling are critical for effective localization. Incremental improvements to these aspects, together with augmentation and scoring refinements, may substantially enhance the pipeline's practical performance.

References

[1] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed Students: Student-Teacher Anomaly Detection with Discriminative Latent Embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4182–4191, June 2020. arXiv:1911.02357 [cs].

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, July 2020. arXiv:2002.05709.

[5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization, November 2020. arXiv:2011.08785 [cs].

[6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022.

[7] Alexander Frotscher, Jaivardhan Kapoor, Thomas Wolfers, and Christian F Baumgartner. Unsupervised anomaly detection using aggregated normative diffusion. *arXiv preprint arXiv:2312.01904*, 2023.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, March 2020. arXiv:1911.05722 [cs].

[9] Hugo Kuijf, Matthijs Biesbroek, Jeroen de Bresser, Rutger Heinen, Christopher Chen, Wiesje van der Flier, Barkhof, Max Viergever, and Geert Jan Biessels. Data of the White Matter Hyperintensity (WMH) Segmentation Challenge, 2022.

[10] Ziyun Liang, Xiaoqing Guo, J Alison Noble, and Konstantinos Kamnitsas. Itermask 2: Iterative unsupervised anomaly segmentation via spatial and frequency masking for brain lesions in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–348. Springer, 2024.

[11] Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

[12] Imperial College London. *IXI Dataset*. `https://brain-development.org/ixi-dataset/` [Accessed: 09.04.2025].

[13] Andrey Malinin, Neil Band, Alexander Ganshin, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panos Tigar, and Boris Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

[14] Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert. Unsupervised anomaly localization with structural feature-autoencoders. In *International MICCAI Brainlesion Workshop*, pages 14–24. Springer, 2022.

[15] Sergio Naval Marimont, Vasilis Siomos, Matthew Baugh, Christos Tzelepis, Bernhard Kainz, and Giacomo Tarroni. Ensembled cold-diffusion restorations for unsupervised anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 243–253. Springer, 2024.

[16] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, September 2021.

[17] Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5):798–819, 2010.

[18] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022.

[19] Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468, 2018.

[20] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144:262–269, 2017.

[21] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[22] B. P. Welford. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419–420, August 1962. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1962.10490022.

[23] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

[24] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021.

[25] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised Anomaly Localization using Variational Auto-Encoders, July 2019. arXiv:1907.02796 [cs].

A Experimental Configurations

This appendix lists the different experimental setups explored during development. The goal is to provide a complete overview of what was attempted, regardless of final outcome.

A.1 Pretraining Setups

We explored several variations of the self-supervised pretraining stage:

- **View augmentation:** (i) Both views augmented (original SimCLR setup). (ii) Only one view augmented, the other left unaltered. This latter variant produced the best results.
- **Batch size:** Typically in the range of 8–16 due to GPU memory limits.
- **Memory bank:** MoCo-style memory bank sizes from 128 up to 1024 were tested to increase the pool of negatives. Larger banks stabilized training.
- **Projection head:** One-layer vs. two-layer MLP heads. Two-layer heads yielded more stable contrastive learning.
- **Optimizers:** AdamW was used, with cosine and constant learning rate schedules.
- **Loss:** NT-Xent contrastive loss, temperature in range $[0.05, 0.1]$.

A.2 Backbones and Models

Several different 3D CNN architectures were implemented and tested as feature extractors for anomaly detection. These models varied in depth, channel width, normalization, and convolutional structure. Each network produced intermediate feature maps through registered forward hooks, and a projection head was used for embedding the pooled representations.

- **Simple3DCNN:** A lightweight encoder with three convolutional blocks (16–128 channels), max-pooling for downsampling, and a small projection head. This served as the baseline 3D architecture.
- **Enhanced3DCNN1:** A deeper variant using strided convolutions, Group Normalization, and GELU activations, with a capacity up to 256 channels. Emphasis was on stable initialization and efficient downsampling. This model proved to be the best.
- **Enhanced3DCNN2:** Similar to the above but introducing dilated convolutions in intermediate layers to enlarge receptive fields while maintaining spatial detail.
- **Enhanced3DCNN3:** Based on a modified ResNet-18 (R3D-18) video backbone, adapted for single-channel medical data. The architecture retained residual connections and was combined with gradient checkpointing to reduce memory footprint.
- **Enhanced3DCNN4:** A factorized R(2+1)D variant, separating spatial and temporal convolutions. The network followed a ResNet-like design with residual blocks and channel scaling for adjustable width.

All models used global average pooling to obtain compact feature representations and a projection head (two linear layers with non-linearity) for downstream similarity-based evaluation. Intermediate feature maps at various resolutions were captured for experimentation with different source–target layer combinations.

A.3 Source and Target Layers

We systematically varied the set of feature maps used as source layers and the resolution of the target layer:

- Source layers tested: `"0"`, `"1"`, `"2"`, `"3"`, `"4"`.
- Target layers: mostly `"1"` or `"2"` for balance between resolution and semantics.
- Best performing configuration: `["0","1","2","3"]` as source layers, resampled to target layer `"2"`.

A.4 Augmentation Configurations

The augmentation pipeline included geometric, intensity, and corruption-based transforms. Parameters were varied (scale factors, rotation degrees, noise levels, swap iterations). The final best setup is reported in Section 2, but weaker and stronger variants were tried in early experiments.

A.5 Gaussian Fitting

Voxel-wise Gaussian statistics were estimated after feature extraction. The procedure was implemented with Welford's online algorithm:

1. Extract features for each batch across the chosen source layers.
2. Resample features to the target resolution if needed.
3. Concatenate channel-wise into a unified representation $(B, C_{\text{total}}, D', H', W')$.
4. Reorder to voxel-major form $(D', H', W', B, C_{\text{total}})$.
5. For each voxel (d, h, w) update:
   - Count of samples.
   - Running mean vector.
   - Running covariance matrix via $M2$ (sum of outer products of centered differences).
6. Store per-voxel Gaussian parameters: N, μ, Σ.

Both diagonal and full covariance estimation were attempted. Per-voxel independent Gaussians were found to be tractable; shared covariance across voxels was also tested but performed worse.

A.6 Other Notes

- Experiments also tested different normalization strategies before Gaussian fitting (min–max, z-score, and finally sigmoid).
- Gaussian fitting results were stored to compressed `.npz` files for later evaluation.

A.7 Summary

Across these experiments, we varied backbone layers, memory bank sizes, augmentation strengths, and Gaussian modeling details. The final reported setup (source `["0","1","2","3"]` $\rightarrow$ target `"2"`, one augmented view, two-layer projection head, memory bank of 1024, per-voxel full covariance) represents the configuration that produced the best performance.