OCSVM-Guided Representation Learning for Unsupervised Anomaly Detection

Nicolas Pinon, Carole Lartizien

***Abstract*—Unsupervised anomaly detection (UAD) aims to detect anomalies without labeled data, a necessity in many machine learning applications where anomalous samples are rare or not available. Most state-of-the-art methods fall into two categories: reconstruction-based approaches, which often reconstruct anomalies too well, and decoupled representation learning with density estimators, which can suffer from suboptimal feature spaces. While some recent methods attempt to couple feature learning and anomaly detection, they often rely on surrogate objectives, restrict kernel choices, or introduce approximations that limit their expressiveness and robustness. To address this challenge, we propose a novel method that tightly couples representation learning with an analytically solvable One-Class SVM (OCSVM), through a custom loss formulation that directly aligns latent features with the OCSVM decision boundary. The model is evaluated on two tasks: a new benchmark based on MNIST-C, and a challenging brain MRI subtle lesion detection task. Unlike most methods that focus on large, hyperintense lesions at the image level, our approach succeeds to target small, non-hyperintense lesions, while we evaluate voxel-wise metrics, addressing a more clinically relevant scenario. Both experiments evaluate a form of robustness to domain shifts, including corruption types in MNIST-C and scanner/age variations in MRI. Results demonstrate performance and robustness of our proposed model, highlighting its potential for general UAD and real-world medical imaging applications. The source code is available at https://github.com/Nicolas-Pinon/uad_ocsvm_guided_repr_learning.**

***Index Terms*—Unsupervised anomaly detection, Representation learning, One-class SVM, Autoencoders, Joint optimization, MNIST-C, Medical imaging, Brain MRI**

I. Introduction

UNSUPERVISED anomaly detection (UAD) aims to identify patterns in data that deviate significantly from an underlying distribution learned from unlabeled normal samples. It is a critical problem in domains where anomalies are rare, variable, and costly to label, such as fraud detection or medical imaging. In neuroimaging, for instance, detecting subtle or small lesions in MRI scans without annotated anomalies remains an open challenge [1]. Models must not only detect rare and diverse outliers but also generalize reliably to new data distributions, such as those resulting from data acquired on different scanners, or populations with different demographics.

Existing methods fall into two main categories: reconstruction-based approaches and representation learning combined with support or density estimation methods. Autoencoders and their variants are frequently used in reconstruction-based strategies, under the assumption that anomalies will yield higher reconstruction errors. However, these models typically lack structured latent representations, which can lead to high quality reconstruction of never seen anomalies. To overcome this, other methods decouple representation learning from the anomaly scoring process, for instance by training a feature extractor independently from a classifier such as a one-class support vector machine (OCSVM [2]). However, this separation can yield to representations not optimized for the decision function computation, leading to suboptimal performance and limited generalization. Several recent approaches attempt to couple representation learning and anomaly detection more tightly, including methods inspired by Deep SVDD [3]. Yet these often rely on approximations, suffer from hypersphere collapse, or impose strong inductive biases (e.g., linear kernel methods) that limit flexibility and robustness.

To address these limitations, we propose a novel method for UAD that tightly couples an autoencoder-based representation learning with a one-class SVM. Our core contribution lies in a new loss formulation that guides the encoder to produce latent representations optimized for the OCSVM's boundary. At each training step, the model splits latent samples into two subsets: one to fit the OCSVM boundary and another to enforce that new samples remain within it. This design reduces overfitting to non-relevant features by directly aligning the encoder's output with the SVM's discriminative objective. Crucially, it enables the use of an exact, analytically solved SVM objective, requiring no approximations or kernel restrictions, thereby preserving the full expressivity of the OCSVM.

To evaluate the proposed method, we conduct two experiments. First, we introduce a new benchmark task based on MNIST-C [4], a corrupted version of the MNIST dataset designed to simulate real-world anomalies. This task allows us to rigorously assess the model's performance in a controlled setting and compare it against state-of-the-art UAD methods. Importantly, this experiment evaluates the model's ability to perform anomaly detection under domain shift, as it must generalize across diverse corruption types. Second, we apply the model to a challenging medical imaging task: detecting subtle brain lesions in MRI scans. In medical imaging, many UAD methods have traditionally focused on detecting large, hyperintense lesions, which are often more visible and easier to identify, especially through reconstruction-based methods. Our work, in comparison, tackles the problem of detecting lesions that can be small and not necessarily hyperintense, representing a more subtle and clinically significant problem. Additionally, while a significant portion of UAD studies in

 (Corresponding author: Nicolas Pinon)

Nicolas Pinon and Carole Lartizien are with Univ. Lyon, CNRS UMR 5220, Inserm U1294, INSA Lyon, UCBL, CREATIS, France (e-mails: nicolas.e.pinon@laposte.net; carole.lartizien@creatis.insa-lyon.fr)

medical imaging measure performances at the image level, we also assess voxel-wise anomaly detection, thus evaluating precise localization of anomalies within the image. Furthermore, this experiment evaluates the model's robustness to domain shifts arising from variations in MRI scanners and patient demographics, such as age.

The contributions of this work are threefold:

- **A novel OCSVM-guided representation learning method for general UAD is introduced, which introduces a loss term aiming at optimizing the representation learner to produce more suitable representations when used in conjunction with OCSVM**
- **A new task based on MNIST-C is introduced to evaluate the proposed and state-of-the-art methods under domain shift, providing a standardized framework for future research.**
- **We demonstrate the method's applicability to real-world medical imaging, showing improved sensitivity to subtle and non-hyperintense lesions in public brain MRI datasets.**

The remainder of this paper is organized as follows: section II reviews related work on anomaly detection, and then specifically methods used in medical imaging. Section III describes our proposed method, detailing the OCSVM-guided representation learning strategy. Sections IV and V present our experimental studies: digit distinction under corruptions using the MNIST-C dataset and subtle lesion detection in brain MRI, respectively. Section VI provides a general discussion, including an analysis of the loss components and concludes the paper while outling potential future research directions.

II. Related works

Unsupervised Anomaly Detection (UAD) methods can be broadly categorized into three main families: *reconstruction-based methods*, *density estimation-based methods*, and *support estimation-based methods*, as outlined in the review by Ruff et al. [5].

All methods share a common objective: modeling the distribution of normal (i.e., non-anomalous) data, often referred to as the normative distribution. Once this distribution is learned, anomalies can be detected as samples that significantly deviate from it. Reconstruction-based methods learn this normative distribution implicitly, by trying to learn a mapping that can accurately reconstruct inputs from a compressed representation. Typically, an autoencoder is trained to encode and decode normal data, minimizing reconstruction error. At test time, if the model fails to reconstruct a sample accurately, the resulting high reconstruction error is interpreted as a sign of abnormality. This assumes that the model, having only seen normal data, cannot generalize well to outliers, and thus reconstructs them poorly. Density and support estimation methods, on the other hand, attempt to explicitly characterize the distribution of normal data either by modeling its density or by learning a decision boundary that encloses the normal data. In both cases, the anomaly score corresponds to how far a test sample lies from the estimated normative distribution.

Representation learning lies at the core of most unsupervised anomaly detection approaches, whether they rely on reconstruction, density estimation, or support estimation. In practice, both density- and support-based methods typically do not operate directly on raw data, but instead leverage intermediate feature representations, often learned through neural networks, to better capture the structure of normal data.

In this work, we specifically focus on autoencoders due to their simplicity and widespread use as a foundational method for unsupervised feature learning, whether used at the end for reconstruction or density/support estimation. Our study serves as a case study to illustrate how feature extractors can be driven and enriched by the downstream anomaly detection task, and how this compares to reconstruction methods. While other feature extractors, such as transformer-based models, could also be employed in a similar framework, exploring all possible alternatives is beyond the scope of this work.

In this bibliographic review, we place a slight emphasis on support estimation methods. While density estimation methods solve a more general problem by modeling the entire data distribution, support estimation directly focuses on distinguishing normal from anomalous data. Given the vast range of possible approaches, we choose to primarily focus on support estimation to maintain a more targeted study, while still acknowledging the relevance of density estimation methods in certain contexts.

Section II-A covers reconstruction-based methods where the anomaly score is directly derived from the reconstruction error. Section II-B focuses on support and density estimation methods that use learned representations. We distinguish between decoupled methods, where the representation learning and the anomaly scoring are optimized separately, and coupled methods, which jointly optimize both components, like the method proposed in this work. Finally, section II-C provides an overview of anomaly detection methods specifically applied to medical imaging.

A. Reconstruction-based methods

A widely used approach in UAD is to leverage an autoencoder's (AE) ability to reconstruct normal data while failing to accurately reconstruct anomalies. As detailed in the review by Ruff et al. [5], reconstruction-based methods assume that, after training on normal samples, an autoencoder will learn a compressed representation that captures essential features of the normal data distribution. When presented with an anomalous input, the reconstruction error is expected to be significantly higher due to the model's inability to generalize to unseen, out-of-distribution patterns.

Early approaches relied on simple autoencoders trained with standard mean squared error or cross-entropy loss, where anomalies were detected based on high reconstruction error [6]. This paradigm has been widely applied to image anomaly detection [7], [8], [9] and extended to various domains, such as industrial defect detection [10] or medical images [11]. Variational Autoencoders (VAE) introduced a probabilistic constraint on the latent space, which helps regularize representations, but they often struggle to clearly separate normal from anomalous reconstructions due to their tendency to generate blurry outputs [12].

Hybrid methods, known as restoration methods have emerged, which combine the reconstruction error with an estimation of the density of the distribution of normal samples in the autoencoder's latent space. These methods aim to "heal" the image by restoring it to the normal distribution (thus erasing the anomaly) and then comparing it to the original image through the reconstruction error. One example is the work by Wang et al. [13], which applies this approach to industrial images by using a quantized autoencoder (VQ-VAE) in conjunction with an autoregressive model (PixelSnail [14]) for density estimation in the latent space. Another type of restoration methods has gained recent popularity for anomaly detection in images: diffusion models, where the image is first partially noised, and then denoised with a UNet-like model, effectively providing a restored image [15].

Another alternative direction involves synthetic anomaly detection (also called self-supervised learning strategies [16], [17]), where synthetic anomalies are added to the data during training of a supervised method. This approach, also proved effective in medical imaging [18], [19], suffers from a severe drawback : the synthetic anomalies distribution must match the (unknown) true anomaly distribution, therefore imposing a strong prior on anomalies that can be detected.

Despite their effectiveness, Ruff et al. [5] highlight several limitations of reconstruction-based methods. Autoencoders may generalize too well, inadvertently reconstructing anomalies with low error, which weakens their discriminative power [20]. Also, reconstruction error alone does not explicitly define a geometrically-coherent decision boundary between normal and anomalous data, making it hard to calibrate anomaly scores. These challenges motivate alternative approaches where autoencoders serve as representation learners rather than direct anomaly detectors, as discussed in section II-B.

B. Support/density estimation methods

As previously mentioned, support and density estimation methods typically rely on representation learning to effectively model the structure of normal data. Autoencoders can fill this purpose [21], where the learned representations can then be used for support or density estimation. These representations can be coupled with classical methods like One-Class SVM (OCSVM [2]), Support Vector Data Description (SVDD [22]) and their variants, Gaussian Mixture models, and so on.

In this section, we distinguish between decoupled methods, where the representation learner is trained separately (II-B1) before applying a support or density estimation method, and coupled methods (II-B2), where the representation learning process is influenced by the anomaly detection objective.

1) Decoupled methods: A common approach is to first train an autoencoder to reconstruct its input, thus providing an encoder capable of producing a compressed representation of the input and then apply a separate anomaly detection method on the learned latent representations; the encoder's weights are thus frozen.

One such method is PaDiM [23], which employs a pre-trained convolutional autoencoder to extract patch-level features, followed by a multivariate Gaussian density estimation to detect anomalies. Similarly, Perera and Patel [24] propose an autoencoder-based feature extraction stage, followed by a clustering approach to identify anomalous samples.

Beggel et al. [8] address the challenge of UAD when the training set is contaminated with outliers by using a discriminator in the latent space of an autoencoder. During training this enhances the separation between the normal training distribution and a predefined anomalous distribution supposed to contain the outliers. At inference the discriminator is used to reject anomalies, along with the reconstruction error.

Another example is the use of autoencoder-based representations with OCSVM, where the extracted features are used to learn a decision boundary enclosing the normal data. This approach is applied to industrial images [9] and synthetic aperture radar images [25]. In both cases, a convolutional autoencoder is trained on normal samples, and the encoder's latent features are fed to an OCSVM for anomaly detection. In [25], the features are further reduced via PCA, and as in [8] a discriminator is used.

Decoupled methods often suffer from a sub-optimal alignment between the learned representations and the anomaly detection objective. Since the representation learner is trained independently from the downstream detection task, the extracted features may not be maximally informative for distinguishing normal from anomalous samples. This mismatch can lead to degraded performance, particularly in complex or high-dimensional settings where anomaly structures are subtle.

2) Coupled methods: Coupled methods aim to address this limitation by integrating the representation learning and support/density estimation steps into a unified framework, thereby encouraging the latent space to be more directly optimized for the detection task. A foundational example is Deep SVDD (DSVDD [3]), which replaces the implicit dual space mapping of traditional SVDD by an explicit modeling (thus approximated) with a neural network. The normal data points are projected in a dual space where they must fit into an hypersphere of learned radius (soft-variant) or just compacted around a predefined center (hard-variant). Anomalies are then identified by measuring the distance to the center (hard) or to the hypersphere (soft). The method is evaluated on several standard image datasets, including MNIST and CIFAR-10, where it demonstrates better performance than kernel-based baselines such as OCSVM.

Nguyen et al. [26] propose an autoencoder-based OCSVM, which combines a deep autoencoder for dimensionality reduction with a OCSVM for anomaly detection. The key innovation is the end-to-end training of both components, where the autoencoder learns a latent representation that directly supports the OCSVM in separating anomalies from normal data. The OCSVM uses Random Fourier Features (RFF) to approximate the Radial Basis Function (RBF) kernel, making the method scalable for large datasets. The method is evaluated on both synthetic and real-world datasets, including MNIST or KDDCup99 and compared to several classical baselines such as OCSVM, Isolation Forest [27], and decoupled deep learning methods. They demonstrate improved performance over the compared methods, while not directly compared against coupled methods.

Deep Structure Preservation SVDD (DSPSVDD) [28] enhances Deep SVDD by first pre-training an autoencoder and then adding the deep SVDD term in the loss fur further fine-tuning. The major difference is that the reconstruction loss term is still present in the fine-tuning. This approach is shown to be more competitive than deep SVDD, isolation forest and reconstruction error from autoencoder on datasets such as MNIST, Fashion-MNIST [29] and MVTecAD [10].

In a similar fashion, VAE-based Deep SVDD (DVAESVDD [30]) combines a VAE with Deep SVDD. This method jointly optimizes the VAE's reconstruction loss and the SVDD's hypersphere loss. Similarly to DSPSVDD, the integration of VAE attempts avoiding the "hypersphere collapse" problem, where all data points are mapped to a single point in the latent space, a limitation of the original Deep SVDD. Experiments on MNIST and CIFAR-10 show the superiority of DVAESVDD over OCSVM and AE reconstruction error.

DASVDD [31] is also an example of combination of autoencoder and deep SVDD, where the main difference is that the center of the hypersphere is updated with a customized procedure at each batch instead of fixed at the beginning of the training. This approach shows increased performances over AE and VAE (when used with reconstruction error), OCSVM and deep SVDD on MNIST, fashion-MNIST and CIFAR-10.

In a similar vein, Contrastive Deep SVDD (CDSVDD [32]) leverages contrastive learning to improve the discriminative power of the learned representations. By minimizing both the contrastive loss and the SVDD loss, CDSVDD ensures that the representations of normal data are tightly clustered around the hypersphere center, while anomalies are pushed further away. This approach also addresses the hypersphere collapse issue and achieves state-of-the-art performance on benchmark datasets. This approach shows increase performances compared to deep SVDD and DSPSVDD, notably on CIFAR-10 and Fashion-MNIST.

Beyond SVDD-based formulations, Zong et al. [33] introduce the Deep Autoencoding Gaussian Mixture Model (DAGMM), which combines a compression network with a Gaussian Mixture Model applied in the latent space. The loss function integrates the reconstruction error, the GMM log-likelihood, and a regularization term. DAGMM was originally tested on tabular datasets (KDDCup99, Thyroid, Arrhythmia).

Other coupled methods include Patch SVDD [34] that extends Deep SVDD by incorporating spatial patch-based features, making it particularly effective for texture-based anomaly detection tasks or one-class GAN (OCGAN) [35], which uses adversarial training to enforce that every normal samples are distributed as a uniform distribution and that every interpolated sample from this distribution output a normal-looking image. The method is evaluated on MNIST and CIFAR-10 and compared against Deep SVDD, VAE and OCSVM. Patch SVDD [34] shows improved performance on classification and anomaly localization on MVTecAD, over deep SVDD and AE and VAE reconstruction error.

Overall, coupled methods seem to benefit from end-to-end optimization, where the representation learning and anomaly detection objectives are jointly optimized. This could ensures that the learned features are directly tailored for anomaly discrimination, leading to superior performance compared to decoupled methods. While coupled approaches seem to surpass their decoupled counterpart in the cited studies, the diversity of evaluation protocols and datasets makes generalization of conclusions difficult. To the best of our knowledge, no comprehensive study has been conducted to systematically assess the benefits of coupling representation learning with anomaly detection, compared to decoupled approaches. Also, to this day, no method makes use of the full flexibility offered by the kernel-representation of the OCSVM/SVDD: all methods use approximations or limitations regarding the type of kernel used for dual space mapping.

Moreover, most existing studies focus on standard, low-complexity datasets such as MNIST, Fashion-MNIST, or CIFAR-10, which do not reflect the challenges of real-world applications. In particular, the medical imaging domain, despite its complexity and practical importance, remains largely unexplored in this context. This highlight the need for a dedicated review of UAD methods in medical imaging, which we present in section II-C.

C. Unsupervised anomaly detection for medical images

In this section, we focus on Unsupervised Anomaly Detection (UAD) methods specifically applied to medical imaging. While the broader field of medical anomaly detection encompasses a wide range of modalities and anatomical regions, we restrict our discussion to studies that align with our focus on brain MRI.

Reconstruction-based methods, as discussed in section II-A, have been widely applied to medical imaging. For instance, Baur et al. [11] conducted a comprehensive comparative benchmark of various autoencoder architectures, including classical autoencoders, variational autoencoders, and adversarial autoencoders, for detecting hyperintense lesions in brain MRI datasets such as MSLUB [36] and MSSEG [37]. Their findings highlight the effectiveness of reconstruction-based approaches for identifying small, hyperintense lesions, which are common in conditions like multiple sclerosis. The same authors proposed a hybrid architecture combining autoencoders with UNet [38], leveraging the reconstruction error as the primary anomaly score. Their method was evaluated on the WMH challenge dataset [39], demonstrating strong performance for detecting small hyperintense lesions.

Cai et al. [17] have performed a wide benchmark on image-level anomaly detection on medical imaging datasets, and image-level and voxel-level anomaly detection specifically for brain MRI on the BraTS dataset [40]. While they do not evaluate support/density estimation methods on the voxel-level anomaly detection task, they evaluate a wide variety of methods based on reconstruction error or synthetic anomaly generation and find that reconstruction-based methods outperform other methods for voxel-level anomaly detection. They also state that on certain datasets a basic autoencoder used with reconstruction error outperform every state-of-the-art methods.

Pinaya et al. [41] introduced a restoration-based approach using a Vector Quantized Variational Autoencoder (VQ-VAE) coupled with a transformer model for density estimation in

the latent space. This method was evaluated on multiple neuroimaging datasets, including MSLUB, BraTS, and WMH, further highlighting the utility of reconstruction-based techniques for hyperintense lesion detection. Additionally, Ramirez et al. [42] used VAEs to detect anomalies in Parkinson's patients' brain MRI, showing that more anomalies were detected in patients than in controls, while Zimmerer et al. [43] and Zhao et al. [44] employed VAEs for brain tumor segmentation, leveraging reconstruction errors as anomaly scores.

In addition to reconstruction and synthetic methods, support/density estimation approaches, as discussed in section II-B, have also been applied to medical imaging. For example, we proposed to employ autoencoders as feature extractors, followed by OCSVM for anomaly detection [45], [46], [47]. In [45], we utilized a localized SVM approach to detect challenging epileptogenic lesions in a private dataset, while in [46] and [47], we proposed a patient-specific OCSVM framework evaluated on the WMH dataset in the former and a Parkinson VS control task in the later. Furthermore, Azami et al. [48] and Bowles et al. [49] used OCSVM for brain MRI anomaly detection, the latter applying it to unsupervised brain lesion segmentation by modeling white and gray matter voxels.

Also, a critical issue in the evaluation of medical anomaly detection methods in brain MRI is the predominance of hyperintense lesions in benchmark datasets. As noted by Meissen et al. [50], many state-of-the-art methods are evaluated on anomalies that are significantly brighter than the surrounding tissue in the MRI image (e.g. FLAIR), such as those in the BraTS and WMH datasets. This raises concerns about the generalizability of these methods to more challenging anomalies, such as those with subtle intensity differences or complex morphological characteristics. In fact, Meissen et al. [51] demonstrated that simply thresholding these MRI images could achieve competitive performance on hyperintense lesion detection, highlighting the need for more rigorous evaluation protocols and diverse datasets.

Despite encouraging results on hyperintense lesions, the performance of unsupervised anomaly detection methods on more challenging, publicly available medical imaging datasets remains largely unevaluated. Autoencoder-based reconstruction methods continue to serve as strong baselines, in contrast, support and density estimation approaches (decoupled II-B1) remain underexplored in this context, often evaluated only on private datasets or omitted altogether from comparative benchmarks. Also, to the best of our knowledge, no coupled (II-B2) method that jointly optimizes feature representation and anomaly detection has been applied to medical imaging.

III. Method: OCSVM-guided representation learning

The method we propose is presented in figure 1. An autoencoder is used for representation learning, while a OCSVM is used to estimate the normal data distribution support. The main goal of the term that we add to the loss function of the autoencoder is to use normal samples that are misclassified during training (projected outside the support) to modify the representation space such that these misclassified samples will be included in the estimated support at the next iteration. Section III-A details the main idea of the method without coupling, by describing the representation learning step III-A1, followed by the anomaly detection step III-A2. Section III-B describes our contribution: coupling of the two steps through the OCSVM-guidance of the representation learning.

A. Decoupled representation learning and anomaly detection

As we have seen in section II-B, autoencoders can be used to perform representation learning, to obtain a more compact representation of their input, in their latent space. Our proposed method can be used with any representation learner (e.g. transformers) but we will focus on a description with an autoencoder. Once the input $\mathbf{x}$ is compressed into a latent representation $\mathbf{z}$, we will use the OCSVM algorithm to estimate the support of the normal class in the latent space (i.e. the support of the $\mathbf{z}_i$).

1) Representation learning with autoencoder: To learn efficient and compressed representations, we train the autoencoder to reconstruct as accurately as possible the input batch[1] $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, while reducing its dimension through its latent space bottleneck. In UAD, the autoencoder is only trained on normal data, and thus learn to represent the normal data manifold in its latent space. We train the autoencoder with the classical MSE loss :

$$L_{AE}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \tag{1}$$

Where $\hat{\mathbf{x}}_i$ is the reconstruction of $\mathbf{x}_i$. After training, the decoder is discarded and the encoder is used, frozen, to perform dimensionality reduction of samples $\mathbf{x}$ into their latent representation $\mathbf{z}$.

2) Anomaly detection with one-class SVM: To perform the detection of anomalies, we estimate the support of the normal data (the boundaries of the normative distribution) with a One-Class SVM (OCSVM [2]). This is done by constructing a decision function f, positive on the estimated support of the distribution of normal samples $\mathbf{z}_i$, negative elsewhere and null on the frontier. The normal samples are first mapped to a high dimensional space by a feature map $\mathbf{\Phi}(\cdot)$ associated with a kernel k such that $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{\Phi}(\mathbf{z}_i) \cdot \mathbf{\Phi}(\mathbf{z}_j)$. As the problem is linear in this re-description space, the parameters $\mathbf{w}$ and ρ of the hyperplane $\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{z}) - \rho = 0$ are obtained by solving a convex optimization problem, presented in equation 2, aiming at maximizing the distance of the hyperplane from the origin.

$$\begin{aligned} \min_{\mathbf{w},\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}_i)\rangle \geq \rho - \xi_i \qquad i \in [1, n] \\ & \xi_i \geq 0 \qquad\qquad\qquad\quad i \in [1, n] \end{aligned} \tag{2}$$

The decision function can then be expressed as $f(\mathbf{z}) = \mathbf{w}^* \cdot \mathbf{\Phi}(\mathbf{z}) - \rho^*$, with $\mathbf{w}^*$ and ρ^* the solutions of the optimization problem.

[1]Input batches in experiment 1 will be batches of whole images and in experiment 2 batches of image patches, but this method can be used with any type of data (even non-image, if the autoencoder is adapted).

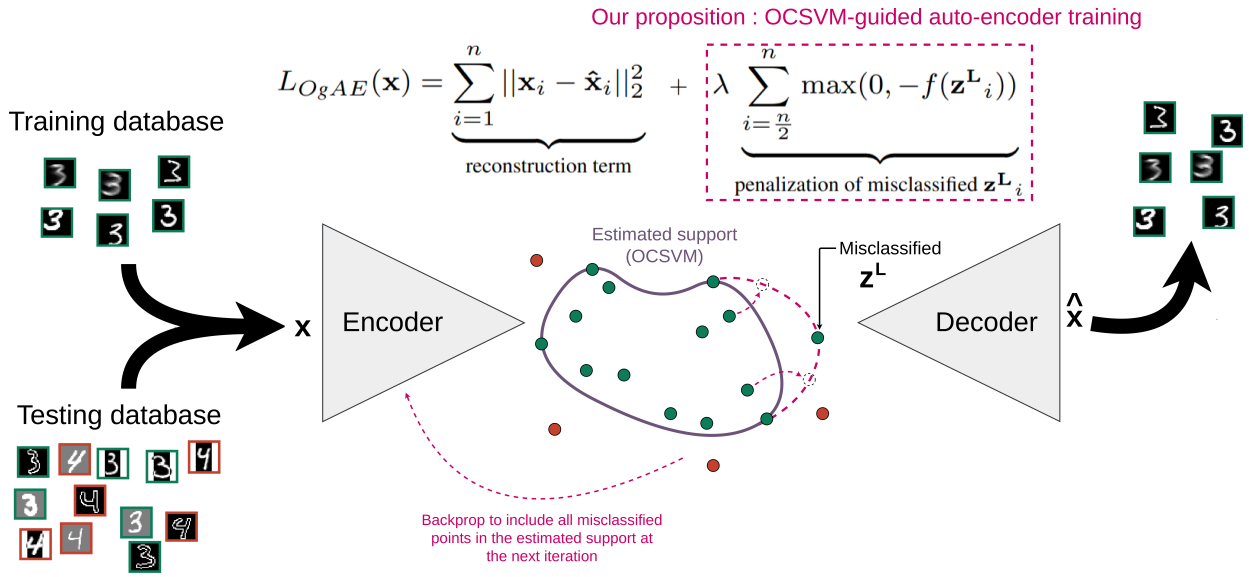

Fig. 1: Graphical abstract of the proposed method. During training, the autoencoder must both minimize the reconstruction error between input and output and a new loss term (section III-B) that guides the encoder towards representations that are more fitted for support estimation with OCSVM.

Through a process known as the kernel trick, the problem is actually solved in its dual form :

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{z}_i,\mathbf{z}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu n} \qquad i \in [1,n] \\ & \sum_{i=1}^{n}\alpha_i = 1 \end{aligned} \tag{3}$$

The decision function is thus expressed as :

$$f(\mathbf{z}) = \sum_{j=1}^{n}\alpha_j^* k(\mathbf{z}_j,\mathbf{z}) - \rho^* \tag{4}$$

which corresponds to a weighted mean of the kernel distance to each normal samples, where many coefficients α_j^* are actually 0. ρ^* is derived using the α_j^*

At inference, to obtain the anomaly score of a new sample $\mathbf{x}$, it must first go through the encoder to obtain its latent representation $\mathbf{z}$, and then through the decision function f. Note that this score will be positive if the sample is within the distribution and negative if outside. The more negative the score, the further the sample is from the normal distribution and thus the more suspicious it will be considered.

B. Coupling: OCSVM-guidance of the representation learning

We describe in this section our contribution: a novel OCSVM-guidance (Og) loss term. The goal of this loss is to align as best as possible the representation of the encoder with the downstream task of estimating the support of the normal distribution with the OCSVM. This is performed by splitting each training batch into two, one used for support estimation ($\mathbf{z^{SVM}}$) and one for loss computation ($\mathbf{z^L}$).

The OCSVM-guidance is divided into two terms: the expander and the compactor, as represented in figure 2. The compactor term makes the estimated support more compact by *moving misclassified normal training samples* inside the

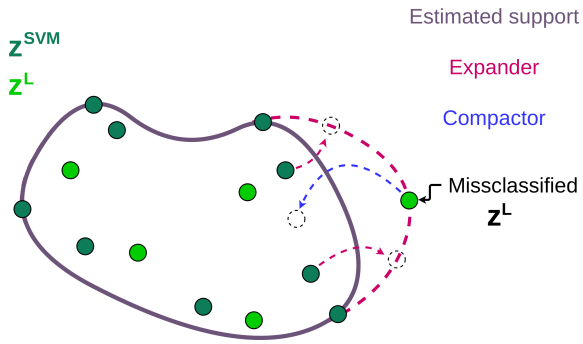

Fig. 2: Visualization of the two terms present in the proposed loss : both terms are based on the idea to use the misclassified $\mathbf{z^L}$ to steer the representations towards SVM-compatible features. While the expander term focus on moving the $\mathbf{z^{SVM}}$ to expand the estimated support, the compactor term focus on moving the $\mathbf{z^L}$ inside the estimated support.

estimated support: this ensures the support stays compact and allow anomalies to fall outside the support. To prevent collapsing of the support, as can happen in deep SVDD, the expander term *moves the boundary* such that misclassified normal training samples fall inside the estimated support. By training the encoder to align with the estimated support (whether by expanding it or compacting it), we implicitly encourage deviations from this manifold to correspond to anomalous behavior, thus learning OCSVM-compatible features while avoiding irrelevant overfitting.

As stated, one batch of samples, after encoding, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is split into two: the part used to solve the OCSVM problem ($\mathbf{z^{SVM}}$) and the other for the loss computation ($\mathbf{z^L}$):

$$\mathbf{z}_i = \begin{cases} \mathbf{z}_i^{SVM} & \text{for } 1 \leq i \leq \frac{n}{2}, \\ \mathbf{z}_i^{L} & \text{for } \frac{n}{2} < i \leq n. \end{cases}$$

At each batch, we solve the OCSVM problem for the $\mathbf{z^{SVM}}_i$, which will give the optimal $\boldsymbol{\alpha}$ and ρ: $\boldsymbol{\alpha}^*$ and ρ^*.

The proposed L_{OgAE} loss is composed of a standard reconstruction error term and the OCSVM-guidance (Og) term, which penalizes the misclassified $\mathbf{z^L}_i$ (that are not used to compute the SVM problem):

$$L_{OgAE}(\mathbf{x}) = \underbrace{\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2}_{\text{reconstruction term}} + \lambda \underbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -f(\mathbf{z^L}_i))}_{\text{penalization of misclassified } \mathbf{z^L}_i} \tag{5}$$

The second term, weighted by λ, indeed penalizes only the misclassified $\mathbf{z^L}_i$, as the decision function outputs positive values for correctly classified $\mathbf{z^L}_i$, and thus $\max(0, -f(\mathbf{z^L}_i))$ is 0. Misclassified $\mathbf{z^L}_i$ are penalized proportionally to their euclidean distance to the estimated hyperplane.

The interest of separating the latent representation vectors into two parts $\mathbf{z^{SVM}}$ and $\mathbf{z^L}$ appears here: as the SVM frontier is estimated on the $\mathbf{z^{SVM}}_i$, most of them are correctly classified. This justifies the use of another set of latent vectors $\mathbf{z^L}$. Penalizing samples not used for the support estimation could also be viewed as a way to penalize bad generalization to unseen samples. We can develop L_{OgAE} with the expression of f from equation 4:

$$\begin{aligned} L_{OgAE}(\mathbf{x}) = &\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \\ &+ \lambda \sum_{i=\frac{n}{2}}^{n} \max\left(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z^{SVM}}_j, \mathbf{z^L}_i) - \rho^*\right) \end{aligned} \tag{6}$$

Recall that $\boldsymbol{\alpha}^*$ and ρ^* are functions of the $\mathbf{z^{SVM}}_i$. If we separate the second term into what depends on the $\mathbf{z^{SVM}}_i$ and what depends on the $\mathbf{z^L}_i$, using the stopgradient operator sg[.] and $\beta_1 + \beta_2 = 1$, we can write L_{OgAE} as:

$$\begin{aligned} &L_{OgAE}(\mathbf{x}) = \textstyle\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \\ &\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad}^{\text{Gradient flow only through the } \mathbf{z^{SVM}}_i} \\ &+\lambda\beta_1 \sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k(\mathbf{z^{SVM}}_j, \text{sg}[\mathbf{z^L}_i]) - \rho^*) \\ &+\lambda\beta_2 \underbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \text{sg}[\alpha_j^*] k(\text{sg}[\mathbf{z^{SVM}}_j], \mathbf{z^L}_i) - \text{sg}[\rho^*])}_{\text{Gradient flow only through the } \mathbf{z^L}_i} \end{aligned} \tag{7}$$

This formulation of L_{OgAE} allows separating the influence of the $\mathbf{z^{SVM}}_i$ and the $\mathbf{z^L}_i$. We argue that the term weighted by β_1, which gradient flows through the $\mathbf{z^{SVM}}_i$, influences the frontier of the SVM, as it moves samples in directions such that it includes the misclassified $\mathbf{z^L}_i$ in the frontier: we call this term the *expander*. The term weighted by β_2, which gradient flows through the $\mathbf{z^L}_i$, will influence the misclassified $\mathbf{z^L}_i$, as it moves the samples in directions such that they enter the boundary drawn by the $\mathbf{z^{SVM}}_i$: we call this term the *compactor*.

C. Algorithm and implementation details

The whole training procedure is divided into two parts. First part is training of the auto-encoder with guidance from the OCSVM loss term, on the normal data, divided per batches. Second part is a final OCSVM-training on the encoded normal data, undivided. The weights of the OCSVM could be computed and averaged along the batched training of the autoencoder such that the whole procedure would be in one step, but we believe a final training on the whole data is quick and increases stability. The whole procedure is summarized in the algorithm presented in the supplementary material S-A and technical details given in S-B.

IV. Experiment 1: Digit distinction under corruptions

We propose in this first experiment a use-case to evaluate the performance of the proposed model in a controlled setting against state-of-the-art. The proposed task will be to evaluate if the models can correctly classify handwritten digits of the normal class versus digits of other classes when presented with a wide variety of corruption noises.

A. Experimental setup and dataset

1) Corrupted MNIST database: MNIST-C [4] is a corrupted variant of the MNIST dataset, designed to evaluate model robustness under distribution shifts. It applies 15 different types of corruptions, such as noise, blur, and geometric transformations, to the original MNIST digits. Examples of such corruptions are shown in the supplementary material figure 5.

2) Compared methods: To evaluate our proposed method, we benchmark it against a set of commonly used approaches in UAD that align with the two main paradigms discussed in Section II: reconstruction-based methods (Section II-A) and support estimation-based methods (Section II-B), both using autoencoders for representation learning.

First, we include standard Autoencoder (AE), Variational Autoencoder (VAE) and Siamese Autoencoder (SAE) models, assessing their anomaly detection performance based on reconstruction error. These models are widely used as baseline approaches in anomaly detection, as discussed in [5] and [11]. Additionally, we evaluate their combination (non-coupled) with a OCSVM trained on the learned latent space, following prior works [25], [52], [45].

Since our proposed method explicitly integrates support estimation within the representation learning process, we also compare against coupled methods (Section II-B2). Specifically, we benchmark against Deep SVDD (both *hard* and *soft* versions) [3], Deep Structure Preservation SVDD (DSPSVDD) [28], and Deep VAE-SVDD (DVAESVDD) [30]. These methods share the objective of refining representations through direct integration with an anomaly detection criterion, making them particularly relevant for comparison with our approach.

This selection of methods allows us to contrast different paradigms of autoencoder-based anomaly detection, from pure reconstruction-based approaches to support estimation-based strategies, both in their decoupled and coupled forms.

For each method, we benchmark a set of their corresponding hyperparameters, and choose the best performing hyperparameter on a separate validation set. Performances are then

reported for the testing set. The benchmarked hyperparameter and the training/validation/testing split are both detailed in supplementary materials S-C and S-D. The training, validation and testing sets roughly contains 17 000, 2000 and 36 000 images. The same autoencoder and training procedure are used for every method, ensuring fair comparison.

3) Proposed task: We propose here to evaluate the capability of the different models to perform anomaly detection under distribution shift, meaning that we train the networks on a "normal" digit, here *3*, under specific corruptions, here *identity*, *motion blur* and *translate*, and then evaluate the networks ability to distinguish "normal" from anomalous digit (here *8*), under another distribution of corruptions, here *stripe*, *canny edges* and *brightness*. This allows evaluating anomaly detection performance in a difficult setting where there is a domain shift between the training and the output, as illustrated on Figure 5. We evaluate another outlier digit in the supplementary material S-E.

The setting where the training and testing corruptions would be the same was found too easy to discriminate the different UAD methods in this analysis. The setting where the method must distinguish between uncorrupted and corrupted digits is also fairly easy, with basic methods such as autoencoder reconstruction error reaching near perfect accuracy [5]. We propose the outlier digit *8*, because it can be very similar to a *3*, and thus is supposed to offer a more challenging setup. Also, some corruptions were found to naturally project to the same latent space locations, thereby making the density/support estimation trivial and the reconstructions naturally erase the corruptions. To provide a difficult setup for both kind of methods, the corruptions used in the experiments have been selected such that when training a basic autoencoder, they would each be separated in its latent space, which we verified using UMAP.

4) Metrics and statistical testing: In our experiments, we evaluate anomaly detection performance using the *AUROC*, *AUROC30*, and Area Under the Precision-Recall Curve (*AUPR*) metrics. *AUROC* measures the model's ability to distinguish between normal and anomalous samples across all decision thresholds but may overestimate performance when anomalies are rare. *AUROC30* focuses on the low false-positive rate regime ($\leq$30%), better reflecting practical scenarios with strict anomaly detection constraints. *AUPR* is more sensitive to class imbalance, making it particularly relevant for highly imbalanced datasets, a setup that is very common in anomaly detection, but this will not be the case in the following experiment where the number of normal and abnormal digits are roughly the same.

We perform statistical testing among the compared models, by generating 1000 bootstrap samples by resampling the testing set with replacement, compute the evaluation metrics (*AUROC*, *AUPR*, and *AUROC30*) for each model on each bootstrap sample, and identify the best-performing model based on mean metric values. We then perform a paired bootstrap test, computing p-values as the fraction of bootstrap samples where a competing model matches or outperforms the best model. To account for multiple comparisons, we apply Bonferroni correction, adjusting the significance threshold accordingly.

TABLE I: PERFORMANCE OF STUDIED MODELS ON DISCRIMINATING *3* VS *8* UNDER CORRUPTION. BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE < 0.01 AFTER PAIRED BOOTSTRAP TEST WITH BONFERRONI CORRECTION) ARE UNDERLINED.

3 vs 8	*AUROC*	*AUPR*	*AUROC30*
AE *recons*	0.56	0.66	0.66
AE *ocsvm*	0.54	0.65	0.67
VAE *recons*	0.54	0.65	0.66
VAE *ocsvm*	0.52	0.63	0.65
SAE *recons*	0.55	0.65	0.67
SAE *ocsvm*	0.53	0.64	0.66
O*g*AE *recons* [ours]	0.53	0.64	0.66
O*g*AE *ocsvm* [ours]	<u>0.59</u>	**0.70**	**0.71**
h-DSVDD [3]	0.51	0.62	0.65
s-DSVDD [3]	0.52	0.63	0.66
DSPSVDD [28]	0.51	0.62	0.65
DVAESVDD [30]	**0.59**	0.67	0.65

B. Results and discussion

Table I presents the performance metrics obtained by all benchmarked models when distinguishing *3* from *8* under corruption. On a side note, we find from the start that the 3 studied metrics show good correlation for all results, indicating that only the study of one of them could suffice.

We observe that representation models coupled with OCSVM (AE *ocsvm*, VAE *ocsvm*, SAE *ocsvm* and O*g*AE *ocsvm*) seem to be on par with their reconstruction-based counterparts (AE *recons*, VAE *recons*, SAE *recons* and O*g*AE *recons*) with the exception of O*g*AE *ocsvm*, significantly outperforming its *recons* counterpart. This finding is slightly counter-weighted by the additional experiments (table IV) which suggest that *ocsvm* models perform slightly better than *recons*. Overall, the basic methods (AE-based) remain competitive, consistently performing within 5 points of the best model for every metric. Our proposed model, tailored for coupling with OCSVM (O*g*AE *ocsvm*) achieves better performances than any other model on all metrics (except when being on par with DVAESVDD for *AUROC*).

For deep SVDD, the results consistently show that the hard-margin variant of Deep SVDD (h-SVDD) outperforms or at least matches the performance of the soft-margin version (s-SVDD). This seems to align with the original paper results [3] and the literature, as the version that has been widely adapted is the hard one [28], [30], [34], [31], which suggests that the added complexity of the soft-margin approach does not translate into a performance gain.

When comparing all coupled models (O*g*AE *ocsvm*, h-DSVDD, s-DSVDD, DSPSVDD and DVAESVDD), we find that O*g*AE and DVAESVDD outperform their competitors. We also find that on this non-trivial task, some coupled models are outperformed by basic baselines (AE *recons*), aligning with previous findings [17].

Additionally, we find that DVAESVDD consistently outperforms DSPSVDD. This could highlight the advantage of using the VAE for more compact latent space or adapting the center of the hyper-sphere at each batch. Both methods consistently outperform traditional Deep SVDD approaches, aligning with the findings in their original papers [30], [28].

A global analysis of the results suggests several global

patterns. Models that leverage representation learning combined with explicit support estimation generally outperform or are on par with reconstruction-based methods. Coupled approaches, where representation learning and anomaly detection are jointly optimized, tend to yield better results than decoupled methods. Recent methods that build upon Deep SVDD frameworks demonstrate improved performance over earlier variants. Finally, our proposed method achieves superior results on this benchmark, surpassing existing state-of-the-art models.

It is worth noting that the dataset corruptions introduced in our experiments can be interpreted as a form of domain shift (i.e. corruptions in the test set are not the same as those in the training set), further emphasizing the adaptability of the evaluated models in real-world scenarios. Also, all comparative claims regarding model performance are supported by rigorous statistical testing, ensuring the robustness of our experimental findings.

V. Experiment 2: Subtle lesion detection in brain MRI

In this experiment, we evaluate the models capabilities to detect subtle lesions in brain MRI scans. This setup is more challenging than the one in experiment 1, as brain MRI images typically contain more complex structures and noise, making anomaly detection more difficult. Example illustrative images are shown in the first column of Figure 3, where a transverse slice of an MRI T1 image is shown on top, and the lesion mask overlaid on this MR image is shown on the bottom. We propose two tasks: classification at the image-level (3D) and a localization (segmentation) task at the voxel-level. Both evaluations are derived from a single anomaly score map output by each model. The goal is twofold: to determine whether the model can differentiate between controls and patients (classification) and to assess its ability to accurately localize anomalies in patient images (localization).

A. Experimental setup and dataset

For this experiment, we consider three 3D MRI T1 image databases described below, two databases of normal control subjects (V-A2), one for training and one for testing and a pathological database (V-A1) comprising exams of patients with brain lesions, referred to as the patient database. The three databases undergo the same preprocessing procedure, described in the supplementary material S-F, to obtain 3D volumes of size $186\times218\times135$ with 1mm^3 voxel size.

1) Pathological database: The White Matter Hyperintensities (WMH) dataset originates from the WMH Segmentation Challenge [39]. More recently, it has also been employed for unsupervised anomaly detection [38], [41], [53], [51], [46], using separate normative datasets for training and leveraging WMH data exclusively for evaluation. Examples of transverse 2D slice extracted from the 3D volume are presented in left columns of Figures 3 and 8. The dataset consists of MRI scans from 60 patients acquired from three different hospitals (20 per hospital), along with expert-annotated segmentation masks of the different pathologies. The patient cohort has a mean age of 70.1 ± 9.3 years, significantly older than the general population, introducing a domain shift when used as a test set for anomaly detection models trained on younger subjects, notably as the process of normal-brain aging results in a slight brain shrinkage. Additionally, because WMH lesions are mostly found in the white matter, models that inherently score white matter as more anomalous (regardless of lesions) may perform artificially better. The dataset also exhibits a wide range of lesion volumes (0.78 cm^3 to 195.15 cm^3), making it particularly challenging due to inter-subject variability and scanner differences.

2) Control databases: The CERMEP control dataset [54] is used for training and validation of the UAD models. This semi-public dataset, available upon request, comprises 75 healthy controls. The subjects in this control group have an average age of 38 ± 11.5 years, which is relatively younger compared to the WMH patient cohort. The control dataset used for testing is a subset of the openly available IXI dataset, which comprises nearly 600 MRI scans from healthy subjects. For this study, we selected 60 IXI controls for testing, age-matched to the WMH dataset (70.1 ± 9.3 years), to mitigate potential age-related bias in the model's classification performance. Without this age-matching correction, the model could have learned to distinguish datasets based on age rather than pathological features. As a result, the control (age-matched IXI) and patient (WMH) datasets used during inference contain the same number of subjects and exhibit identical mean and standard deviation in age, ensuring a fair evaluation of the model's ability to detect pathology rather than demographic differences.

3) Compared Methods: Due to the size of the MR images, we will use our proposed method on small 2D patches (15×15), as we have already done in previous work [45], [46] and such that it will approximately match the size of the images on experiment 1. The 2D anomaly map is obtained by moving this patch in increments of 1 in all directions across the entire 2D image and calculating the score of the central pixel for each position. The 3D score map is obtained by concatenating the 2D anomaly maps (see first row of figure 3 for examples of anomaly score maps superposed with MRIs).

We compare our proposed method with state-of-the-art Unsupervised Anomaly Detection (UAD) approaches that have been evaluated on the WMH dataset. SAE + *localized* OCSVM [45] and SAE + *patient specific* OCSVM [46] are two methods that also work by patch, and we will thus use the same autoencoder for our proposed method and these two: it follows a structure similar to the one used in experiment 1, both detailed in the supplementary material S-H. We also include the methods proposed by Baur et al. [38] and Pinaya et al. [41], which both process full 2D slices, and then by concatenation obtain the 3D anomaly score map.

For a fair comparison, we implement each method using the hyperparameters provided in their respective publications. Hyperparameters for our proposed method are taken to be the best performing for experiment 1. The training of the models is done on 80% of the CERMEP control dataset, while the remaining 20% is used for early stopping during training. Testing is performed on both the control IXI-age-matched subset and on the pathological WMH database for

the classification task and on the WMH database only for the localization task.

4) Proposed task: ***Classification****:* Once a model is trained on the CERMEP control database, it can produce at inference 3D anomaly score maps on the IXI test control database and on the WMH patient database. The first task we propose is a classification one, more precisely we evaluate how we can distinguish healthy controls (IXI) from pathological patients (WMH) from the information contained in each score map. To obtain a single anomaly score per patient from their anomaly score map, we tested different aggregation methods (2% percentile, mean, median, with or without ventricle removal) and found that it had little impact on the overall results. In the end, we used the 2nd percentile threshold of the anomaly scores (meaning 2% of scores fall below this value) while excluding the ventricles. This exclusion was motivated because ventricles tend to exhibit high anomaly scores due to age-related differences between the control and patient databases.

5) Proposed task: ***Localization****:* For the second task, we directly use the anomaly score maps of the WMH patients and compare it voxel-wise to the ground-truth maps of the lesions, to obtain localization metrics. IXI controls are not used here because they contain no lesions.

6) Metrics and statistical testing: We use the same evaluation metrics as in experiment 1 (*AUROC*, *AUROC30*, and *AUPR*) as detailed in section IV-A4, both for the classification task (distinguishing controls from patients) and the localization task (identifying lesions within patient images). Unlike the first experiment with balanced classes, this setup introduces imbalance in the localization task, where lesion voxels are rare. AUPR is thus critical, as it better reflects performance under imbalanced training.

For the **classification** task, we perform statistical testing among the different compared models by generating 1000 bootstrap samples by resampling the subjects with replacement, compute the evaluation metrics for each model on each subject (control or patient), and identify the best model by mean performance. Then, as for experiment 1, we perform a paired bootstrap test with Bonferroni correction.

For the **localization** task, we compute one *AUROC*, *AUPR*, and *AUROC30* per patient, thus introducing natural variability across samples (patients). We employ a Kruskal-Wallis test to detect overall differences among models, followed by Dunn's test for pairwise comparisons (with Bonferroni correction).

The main difference between the two tasks is that in the classification tasks, we only get one score per sample and thus, for example, one *AUROC* for the whole task. We thus have to use bootstrapping to produce multiple *AUROC* and simulate variability, whereas in the localization task, we have one *AUROC* per patient (multiples localizations and lesions) and thus we have a natural inter-patient variability.

B. Results and discussion

Results of the classification experiment are presented in table III, while results of the localization experiment are presented in Table IV. Figures 3 and 4 (plus 8 and 9 in the supplementary material) present visualization of the obtained score maps. For the AE/UNet method the dynamic range of the image had to be enhanced to [5%, 95%] quantile for enhanced visibility.

On the classification task, Table III shows that most models achieve very high accuracy, with the exception of the SAE+*p.s.*OCSVM. The anomaly maps (figure 3 and 4) for this model suggest that it is highly overfitted to detecting anomalies in the ventricles and cortex, which could be due to registration errors rather than actual pathological features (see registration pipeline in supplementary material S-F). For classification, the three evaluation metrics exhibit strong correlations. Although SAE+*loc.*OCSVM emerges as the best-performing method, its advantage over other approaches is not statistically significant (except when compared to SAE+*p.s.*OCSVM).

It is important to recall that the test databases are age-matched, meaning that models should not be able to distinguish images based solely on age-related degenerative changes. This ensures that any detected anomalies are not confounded by age effects.

On the localization task, results reported in Table IV show that AE/UNet, VQ-VAE+Transformer and SAE+*p.s.*OCSVM) are not capable of localizing correctly the lesions, as their performance are at chance level or below. For SAE+*p.s.*OCSVM, this result is expected as it did not succeed to classify patients from controls. For AE/UNet and VQ-VAE+Transformer, however, this result is surprising, as these models are capable of distinguishing between control and patient, but not by directly identifying the lesions' localizations. This could suggest that these models may have found other discriminant anomalies than those annotated by the clinicians, or other confounding features enabling discriminating the IXI from the WMH subjects.

In contrast, quantitative performances of both SAE+*loc.*OCSVM and O*g*AE+*loc.*OCSVM show that they both successfully localize lesions, with O*g*AE+*loc.*OCSVM demonstrating superior performance in detecting small lesions, as reflected by the *AU PR*. Note that for this localization task, the baseline *AU PR* (random classifier) is 0.007.

Visualization of the score maps in Figure 3 and 4 indicates that most methods are sensitive to registration errors (particularly at the outer brain regions) and brain shrinkage, which is expected due to the difference in age (which is a form of domain shift, adding difficulty to the task) between the training and test datasets. We see, for instance, on Figure 4 that most models flag the lower right ventricle of this example patient as anomalous, as it is quite enlarged compared to a younger control (see supplementary material figure 7).

In this study, we used the T1 MRI modality, where lesions are challenging to detect, unlike all previous studies performed on this database which also included MR FLAIR images where WMH lesions appear as hyperintense [53], [41], [38], [46], [51]. A broader trend emerges where models initially designed to detect hyperintense lesions struggle with this task. SAE+*loc.*OCSVM, originally developed for epileptogenic lesions detection (which even experts struggle to see [55]), performs better in this context. Overall, our proposed method outperforms state-of-the-art methods on this difficult task, particularly for identifying small lesions.

Methods
AE/UNet [38]
VQ-VAE + Transformer [41]
SAE + *localized* OCSVM [45]
SAE + *patient specific* OCSVM [46]
O*g*AE + *localized* OCSVM [ours]

TABLE II: BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE < 0.01 AFTER PAIRED BOOTSTRAP TEST WITH BONFERRONI CORRECTION) ARE UNDERLINED.

Classification IXI vs WMH		
AU ROC	*AU ROC 30*	*AU PR* (0.5)
0.98	0.97	0.99
0.96	0.93	0.96
0.99	**0.98**	**0.99**
0.09	0.41	0.32
0.90	0.89	0.93

TABLE III: BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE < 0.01 AFTER KRUSKAL-WALLIS AND DUNN WITH BONFERRONI CORRECTION) ARE UNDERLINED.

Localization WMH		
AU ROC	*AU ROC 30*	*AU PR* (0.007)
0.38	0.42	0.005
0.52	0.51	0.008
0.62	0.59	0.017
0.32	0.41	0.004
0.61	**0.72**	**0.066**

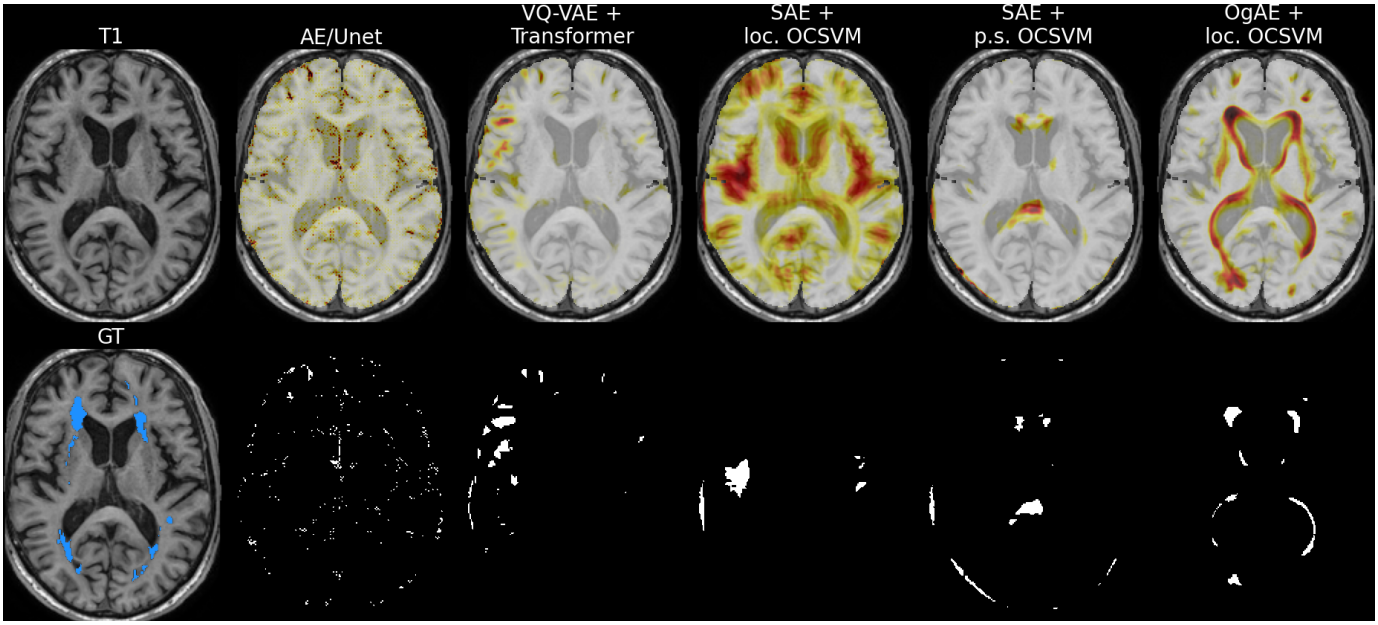

Fig. 3: Visualization of a central slice from the T1-weighted brain MRI of a WMH patient (AM126). The ground truth (GT) is overlaid, with light blue indicating pathological lesions. Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.

VI. GENERAL DISCUSSION AND CONCLUSION

In this work, we introduced a novel method for UAD that addresses limitations of existing approaches: most state-of-the-art methods rely either on reconstruction-based models, which tend to reconstruct anomalies too well and fail to produce discriminative representations, or on decoupled architectures where feature learning and anomaly scoring are optimized separately resulting in misaligned feature spaces. Recent attempts to couple these processes often rely on surrogate objectives, linear kernel formulations, or approximations that compromise flexibility and robustness. To overcome these challenges, we proposed a coupled framework in which the representation learning process is explicitly guided by an analytically solvable OCSVM loss that steers the encoder toward producing latent features aligned with the OCSVM decision boundary, thereby directly optimizing the feature space for anomaly detection. By enforcing this alignment during training, the encoder is encouraged to focus on features that are genuinely relevant for modeling the normative distribution, reducing overfitting to irrelevant patterns.

We evaluated our approach on two tasks: digit distinction under corruption, and subtle lesion detection in brain MRI. In the first task, our proposed O*g*AE paired with OCSVM outperformed both classical and state-of-the-art UAD methods, and additionally demonstrating robustness to domain shifts across diverse corruptions. In the medical imaging task, O*g*AE effectively distinguished pathological from control subjects in brain MRI, despite the challenge of detecting small, non-hyperintense lesions. It showed superior localization capabilities, particularly for small lesions (by improved *AUPR*).

A key contribution of our work is the OCSVM-guided representation learning, which adresses the limitations of existing coupled approaches: it avoids the pitfalls of traditional deep SVDD approaches, which often suffer from hypersphere collapse, by ensuring that the learned representations maintain sufficient variance while still being well-clustered within the normal class. In deep SVDD, soft-margin methods explicitly model the dual-space projection through a neural network, reducing expressivity, also, the widely used hard-margin variant focuses on compacting points around a predefined center without a notion of radius. Our approach, in contrast, does not rely on a neural network projection, preserving the full expressivity of the original OCSVM formulation. Furthermore, unlike methods that arbitrarily steer all points toward a center, our model allows them to remain in place if they lie within the estimated boundary, ensuring a sufficient level of variance in the learned representation. Additionally, we think computing

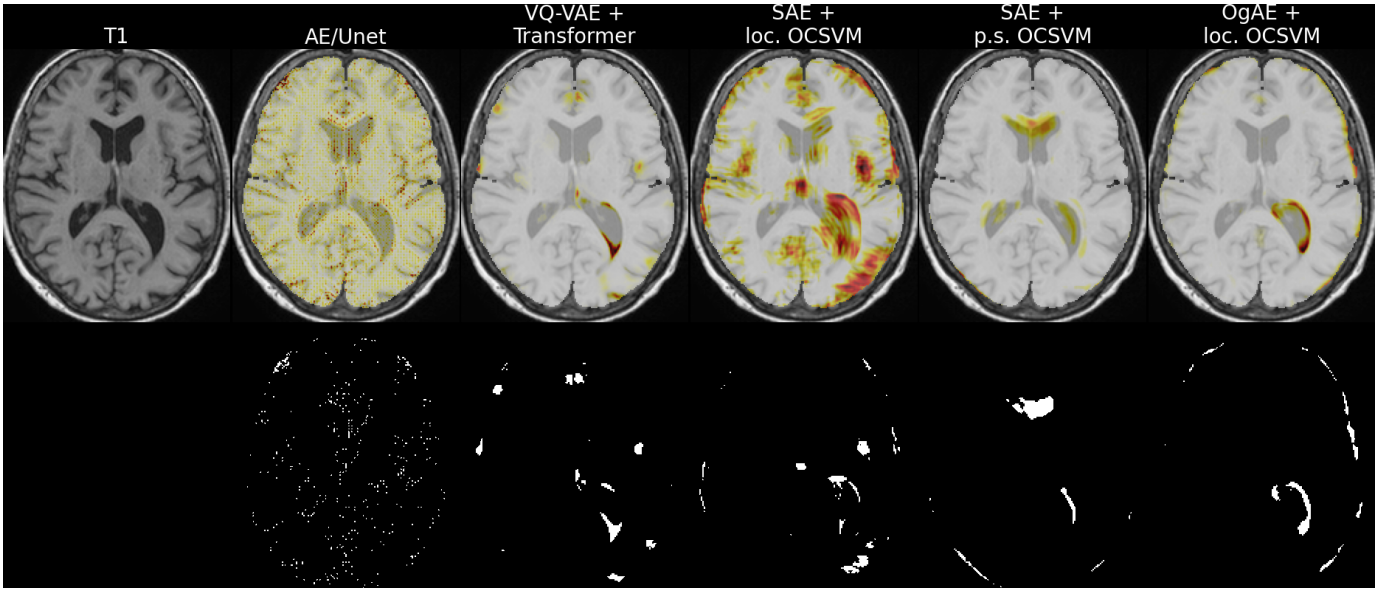

Fig. 4: Visualization of a central slice from the T1-weighted brain MRI of a IXI control (IXI158-Guys-0783). Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.

the loss on a holdout portion of each batch can enhance generalization.

We show, on Figure 6 in supplementary material, an example of training with the expander term (equation 7 $\beta_1 = 1, \beta_2 = 0$) for the first 5 epochs followed by the compactor term ($\beta_1 = 0$, $\beta_2 = 1$) for 5 other epochs. We study the average pairwise MSE between the latent representations, which is an indicator of their spread. We clearly see that during the expanding phase the spread of the latent representation is growing and that in the compaction phase it is decreasing. The best performing strategy (evaluated on experiment 1) was found to be expander term first followed by expander + compactor with the same weight, aligning with the intuition that increasing the representation's variety at first benefits learning, but ultimately, the boundary size must be controlled and fixed. We believe further research is needed to explore optimal training strategies.

For the medical image experiment, we did not employ any post-processing for our approach, unlike other works ([38], [41]), suggesting that further refinement could improve performances, particularly in the localization task. Additionally, transitioning to 3D representations for medical images could enhance the model's spatial awareness. Previous research [46] suggests that patch size has minimal impact on performance, reinforcing the generalizability of our approach. Given that the autoencoder+*localized* OCSVM method [45] was effective for epilepsy detection, we should evaluate the potential of our proposed O*g*AE on epilepsy datasets as well, e.g. [56].

Several avenues for future research remain open. While our study focused on autoencoders, the OCSVM-guided framework could be applied to other feature extraction methods (e.g. transformers). Additionally, since we have focused our study to support estimation models (see section II-B), exploring density estimation techniques, which have proven competitive in anomaly detection, could provide further insights. Our method was designed for UAD (training only on normal samples), but in a semi-supervised setting, it could be extended by incorporating anomalous samples to refine the decision boundary: instead of only enforcing that normal samples remain inside the estimated boundary, anomalous samples could be explicitly pushed outside (or the frontier compacted such the sample remain outside). Also, an SVDD-guided variant could be implemented and evaluated, despite being similar when using the RBF kernel.

REFERENCES

[1] I. Lagogiannis, F. Meissen, G. Kaissis, and D. Rueckert, "Unsupervised pathology detection: a deep dive into the state of the art," *IEEE transactions on medical imaging*, vol. 43, no. 1, pp. 241–252, 2023.
[2] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
[3] L. Ruff *et al.*, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.
[4] N. Mu and J. Gilmer, "Mnist-c: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019.
[5] L. Ruff *et al.*, "A Unifying Review of Deep and Shallow Anomaly Detection," *Proceedings of the IEEE*, vol. 109, pp. 756–795, May 2021.
[6] M. A. Kramer, "Autoassociative neural networks," *Computers & chemical engineering*, vol. 16, no. 4, pp. 313–328, 1992.
[7] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA, Machine Learning for Sensory Data Analysis*. ACM, 2014, pp. 4–11.
[8] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust anomaly detection in images using adversarial autoencoders," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany*. Springer, 2020, pp. 206–222.
[9] N. Pinon, "Unsupervised anomaly detection in neuroimaging: Contributions to representation learning and density support estimation in the latent space," Ph.D. dissertation, INSA Lyon, 2024, chapter III, Sec 1.
[10] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.
[11] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," *Medical Image Analysis*, vol. 69, p. 101952, 2021.

[12] D. Zimmerer, J. Petersen, and K. Maier-Hein, "High-and low-level image component decomposition using vaes for improved reconstruction and anomaly detection," *arXiv preprint arXiv:1911.12161*, 2019.

[13] L. Wang, D. Zhang, J. Guo, and Y. Han, "Image Anomaly Detection Using Normal Data Only by Latent Space Resampling," *Applied Sciences*, vol. 10, no. 23, p. 8660, Dec. 2020.

[14] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International conference on machine learning*. PMLR, 2018, pp. 864–872.

[15] H. He *et al.*, "A diffusion-based framework for multi-class anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 8472–8480.

[16] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems*, 2018, pp. 9758–9769.

[17] Y. Cai, W. Zhang, H. Chen, and K.-T. Cheng, "Medianomaly: A comparative study of anomaly detection in medical images," *Medical Image Analysis*, vol. 102, p. 103500, 2025.

[18] A. Kascenas, R. Young, B. S. Jensen, N. Pugeault, and A. Q. O'Neil, "Anomaly detection via context and local feature matching," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[19] J. Tan, B. Hou, J. Batten, H. Qiu, B. Kainz *et al.*, "Detecting outliers with foreign patch interpolation," *Machine Learning for Biomedical Imaging*, vol. 1, no. April 2022 issue, pp. 1–27, 2022.

[20] Y. Zhang *et al.*, "Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features," *Proceedings of the AAAI*, vol. 35, no. 9, pp. 10 777–10 785, 2021.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, pp. 504–507, 2006.

[22] D. M. Tax and R. P. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, Jan. 2004.

[23] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021.

[24] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5450–5463, 2019.

[25] S. Mabu, S. Hirata, and T. Kuremoto, "Anomaly detection using convolutional adversarial autoencoder and one-class svm for landslide area detection from synthetic aperture radar images." *J. Robotics Netw. Artif. Life*, vol. 8, no. 2, pp. 139–144, 2021.

[26] M.-N. Nguyen and N. A. Vien, "Scalable and interpretable one-class svms with deep learning and random fourier features," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin*. Springer, 2019, pp. 157–172.

[27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[28] Z. Zhang and X. Deng, "Anomaly detection using improved deep SVDD model with data structure preservation," *Pattern Recognition Letters*, vol. 148, pp. 1–6, Aug. 2021.

[29] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[30] Y. Zhou, X. Liang, W. Zhang, L. Zhang, and X. Song, "Vae-based deep svdd for anomaly detection," *Neurocomputing*, vol. 453, 2021.

[31] H. Hojjati and N. Armanfard, "Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 8, pp. 3739–3750, 2023.

[32] H.-J. Xing and P.-P. Zhang, "Contrastive deep support vector data description," *Pattern Recognition*, vol. 143, p. 109820, 2023.

[33] B. Zong *et al.*, "Deep autoencoding gaussian mixture model fore unsupervised anomaly detection," in *International Conference on Learning Representations (ICLR)*, 2018, p. 19.

[34] J. Yi and S. Yoon, "Patch svdd: Patch-level svdd for anomaly detection and segmentation," in *Computer Vision-ACCV 2020: 15th Asian Conference on Computer Vision, Kyoton*. Springer, 2021, pp. 375–390.

[35] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF*, 2019, pp. 2898–2906.

[36] Ž. Lesjak *et al.*, "A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus," *Neuroinformatics*, vol. 16, pp. 51–63, 2018.

[37] O. Commowick *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific reports*, vol. 8, no. 1, p. 13650, 2018.

[38] C. Baur, B. Wiestler, M. Muehlau, C. Zimmer, N. Navab, and S. Albarqouni, "Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly Detection in Brain MRI," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e190169, May 2021.

[39] H. J. Kuijf *et al.*, "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 11, pp. 2556–2568, Nov. 2019.

[40] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[41] W. H. Pinaya *et al.*, "Unsupervised brain imaging 3D anomaly detection and segmentation with transformers," *Medical Image Analysis*, vol. 79, p. 102475, Jul. 2022.

[42] V. M. Muñoz-Ramírez, V. Kmetzsch, F. Forbes, and M. Dojat, "Deep Learning Models to Study the Early Stages of Parkinson's Disease," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA, USA: IEEE, Apr. 2020, pp. 1534–1537.

[43] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019 Shenzhen*. Springer, 2019, pp. 289–297.

[44] Y. Zhao, Q. Ding, and X. Zhang, "Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection," in *The Eleventh International Conference on Learning Representations*, 2022.

[45] Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien, "Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening," *Medical Image Analysis*, vol. 60, p. 101618, 2020.

[46] N. Pinon, R. Trombetta, and C. Lartizien, "One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities," in *MIDL 2023, International Conference on Medical Imaging with Deep Learning*. PMLR, 2023.

[47] N. Pinon, G. Oudoumanessah, R. Trombetta, M. Dojat, F. Forbes, and C. Lartizien, "Brain subtle anomaly detection based on auto-encoders latent space analysis : application to de novo parkinson patients," in *ISBI 2023 - IEEE 20th International Symposium on Biomedical Imaging*. Cartegena de Indias, Colombia: IEEE, Feb. 2023.

[48] M. El Azami, A. Hammers, J. Jung, N. Costes, R. Bouet, and C. Lartizien, "Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem," *PloS one*, vol. 11, 2016.

[49] C. Bowles *et al.*, "Brain lesion segmentation through image synthesis and outlier detection," *NeuroImage: Clinical*, vol. 16, pp. 643–658, 2017.

[50] F. Meissen, B. Wiestler, G. Kaissis, and D. Rueckert, "On the pitfalls of using the residual as anomaly score," in *Medical Imaging with Deep Learning*, 2021.

[51] F. Meissen, G. Kaissis, and D. Rueckert, "Challenging current semi-supervised anomaly segmentation methods for brain mri," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 63–74.

[52] N. Pinon, R. Trombetta, and C. Lartizien, "Détection d'anomalies dans l'espace image ou l'espace latent d'auto-encodeurs par patch pour l'analyse d'images industrielles," in *GRETSI 2023, XXIXème Colloque Francophone de Traitement du Signal et des Images*, 2023.

[53] W. H. Pinaya *et al.*, "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 705–714.

[54] I. Mérida *et al.*, "CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F]FDG PET, T1 and FLAIR MRI, and CT images available for research," *EJNMMI Research*, vol. 11, p. 91, Dec. 2021.

[55] T. Wehner *et al.*, "Factors influencing the detection of treatable epileptogenic lesions on mri. a randomized prospective study," *Neurological research and practice*, vol. 3, pp. 1–11, 2021.

[56] F. Schuch *et al.*, "An open presurgery mri dataset of people with epilepsy and focal cortical dysplasia type ii," *Scientific Data*, vol. 10, 2023.

[57] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter, "Differentiable convex optimization layers," in *Advances in Neural Information Processing Systems*, 2019.

[58] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[59] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "Osqp: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

SUPPLEMENTARY MATERIAL

A. Algorithm

Autoencoder training with OCSVM-guidance

Input: Normal samples $(\mathbf{x}_i)_{1\leq i\leq N}$
Output: Trained encoder E ▷ Decoder D is discarded

for each epoch **do**
  for every batch of samples $(\mathbf{x}_i)_{1\leq i\leq b}$ **do** ▷ Batch size b
    Compute latent representations of samples :
    $(\mathbf{z}_i)_{1\leq i\leq b} = E[(\mathbf{x}_i)_{1\leq i\leq b}]$
    Split in two the $\mathbf{z}_i$ to obtain ${\mathbf{z}^{\text{SVM}}}_i$ and ${\mathbf{z}^{\text{L}}}_i$
    Solve the OCSVM problem (3) for the ${\mathbf{z}^{\text{SVM}}}_i$ to obtain:
    $(\alpha^*_j)_{1\leq j\leq \frac{b}{2}}$ and ρ^*
    Compute the reconstructions of latent representations:
    $(\hat{\mathbf{x}}_i)_{1\leq i\leq b} = D[(\mathbf{z}_i)_{1\leq i\leq b}]$
    Compute the loss (7) and apply a gradient step to E and D
  end for
end for

OCSVM final training

Input: Normal samples $(\mathbf{x}_i)_{1\leq i\leq N}$ and trained encoder E
Output: Decision function f of OCSVM

Compute latent representations of samples:
$(\mathbf{z}_i)_{1\leq N} = \mathbf{D}[(\mathbf{x}_i^h))_{1\leq N}]$
Solve the OCSVM problem (3) for the $(\mathbf{z}_i)_{1\leq N}$ to obtain the parameters of the final decision function

B. Technical details for the OCSVM-guidance model

This section outlines the technical implementation of the OCSVM-guidance model, particularly the gradient computation through the dual solution, the numerical stabilization techniques, and the kernel matrix reformulation.

When computing the expander term in equation 7, we have to differentiate through $\boldsymbol{\alpha}^*$ and ρ^*, thus through a convex optimization problem (problem 3): to do this we use [57]. We also study configurations in which the gradient flow only trough $\mathbf{z}^{\text{SVM}}$ for the expander term and thus apply the $\text{sg}[.]$ operator to $\boldsymbol{\alpha}$ and ρ. For solver-related manner, the problem 3 has to be written in a way that it is linear in parameters, not quadratic. We thus utilize the fact that $\mathbf{K}$ is positive semi-definite (because it is a gram matrix), to express it as: $\mathbf{K} = \mathbf{K}^{\frac{1}{2}^T}\mathbf{K}^{\frac{1}{2}}$. Where $K_{ij} = k(\mathbf{z}_i, \mathbf{z}_j)$. As recommended in [58], because $\frac{1}{\nu_j n}$ can get very small as n increase, this only leaves a tight bound for the constraint $0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n}$. Thus, for numerical stability reasons, we solve a scaled problem of variable $\tilde{\alpha}_{ji} = n\nu_j\alpha_{ji}$. As also recommended in [58] for numerical stability, to compute ρ, we average the ρ obtained for every support vector. Finally, also for numerical stability, we computed $\mathbf{K}$ as $\mathbf{K} + 1e^{-8}\mathbf{I}$. We used the OSQP solver [59].

C. Benchmarked hyperparameters for experiment 1

These hyperparameters are:

- weight coefficients for KL divergence (VAE)
- cosine similarity (SAE)
- λ (OgAE)
- γ (DSPSVDD, see article [28])
- α (DVAESVDD, see article [30])
- expander/compactor strategy (OgAE), i.e. $\beta_1, \beta_2 = \{0, 1\}$
- ν, γ_{RBF} and scaling of the latent variables (every method using OCSVM)

The same autoencoder and training procedure are used for every method, to ensure fair comparison. The architecture is the one used for the MNIST experiment in DVAESVDD [30], it is detailed in the supplementary material S-H1. Note that the auto-encoder inputs here are batches of full-sized images. For the models using reconstruction error, the mean of the reconstruction error map ($||\mathbf{x} - \hat{\mathbf{x}}||_2^2$) is used as the anomaly score. For the models using a OCSVM, the OCSVM is trained on the training set and the anomaly score is computed using the decision function (equation 4), RBF kernel is used. For deep SVDD (soft and hard), DSPSSVDD and DVAESVDD, the anomaly scoring function described in the original articles is used.

D. Training/validation/testing split for experiment 1

The training set is composed of the 6131 handwritten *3* images from the training set of MNIST, corrupted with the training corruptions (*identity*, *motion blur* and *translate*), for a total of 18393 images. 90% are used for model training and 10% are used for early stopping. The validation set is composed of both 974 handwritten *8* and 1010 *3* images from the testing set of MNIST, corrupted with the testing corruptions *s1* (*stripe*, *canny edges* and *brightness*), for a total of 5952 images. The testing set is composed of both 5851 handwritten *8* and 6131 *3* images from the training set of MNIST, corrupted with the testing corruptions (*stripe*, *canny edges* and *brightness*), for a total of 35946 images. Note that the use of the testing set for validation and the training set for testing is done to give the testing set the most samples and thus the most statistical power for drawing reliable conclusions. Also note that while the performances could be a little bit over-estimated because the hyperparameter optimization is done on the same corruptions as the testing set, the validation set and the testing set have no samples in common. The training set and the testing set, obviously, do not have the same corruptions. For the 3 vs 4 experiment, presented in the supplementary material S-E in table IV, the validation set is composed of both 982 handwritten *4* and 1010 *3* images from the testing set of MNIST, corrupted with the testing corruptions, for a total of 5976 images. The testing set is composed of both 5842 handwritten *4* and 6131 *3* images from the training set of MNIST, corrupted with the testing corruptions for a total of 35919 images.

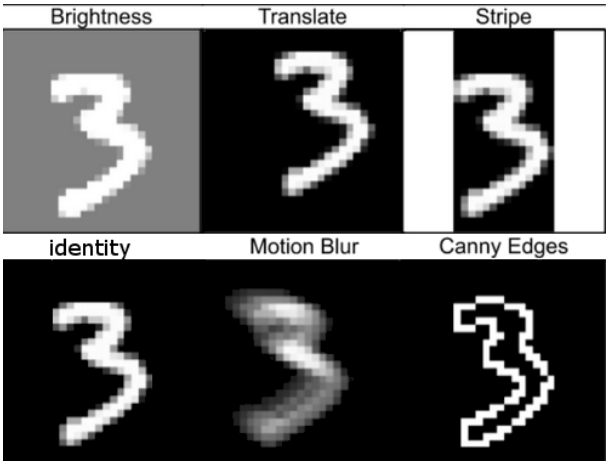

Fig. 5: Corruptions of the MNIST dataset (MNIST-C [4]) used throughout this article, on the digit 3.

TABLE IV: PERFORMANCE OF STUDIED MODELS ON DISCRIMINATING *3* VS *4* UNDER CORRUPTION (WHICH DIFFER FROM *3* VS *8* IN THE MAIN BODY). BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE < 0.01 AFTER PAIRED BOOTSTRAP TEST WITH BONFERRONI CORRECTION) ARE UNDERLINED.

3 vs 4 *s1*	*AUROC*	*AUPR*	*AUROC30*
AE *recons*	0.51	0.64	0.66
AE *ocsvm*	0.52	0.64	0.67
VAE *recons*	0.51	0.63	0.65
VAE *ocsvm*	0.57	0.67	0.68
SAE *recons*	0.47	0.60	0.64
SAE *ocsvm*	0.58	0.68	0.69
OgAE *recons* [ours]	0.55	0.66	0.67
OgAE *ocsvm* [ours]	**0.74**	**0.76**	**0.73**
h-DSVDD [3]	0.54	0.64	0.65
s-DSVDD [3]	0.46	0.59	0.63
DSPSVDD [28]	0.56	0.66	0.67
DVAESVDD [30]	0.59	0.69	0.69

E. Additional results for experiment 1

We propose to extend the evaluation carried out in experiment 1, by evaluating another, easier outlier digit : *4*. Table IV presents the performances on the different evaluated models (main body section IV-A2) for this outlier digit.

Slightly differently to the experiments of the main body (table I) we observe that representation models coupled with OCSVM (AE *ocsvm*, VAE *ocsvm*, SAE *ocsvm* and OgAE *ocsvm*) seem to outperform their reconstruction-based counterparts (AE *recons*, VAE *recons*, SAE *recons* and OgAE *recons*) for the 3 vs 4 task. The only exception is the AE, which performs on par with its *ocsvm* counterpart. Overall the OgAE, paired with *ocsvm*, seem to be the best-performing model on this extended experiment, confirming the trend presented in the main body.

As in the main body, the extended results show that the hard-margin variant of Deep SVDD (h-SVDD) outperforms the performance of the soft-margin version (s-SVDD). Also, as in the main body, we find that DVAESVDD outperforms DSPSVDD.

When examining the highest-performing models, we observe a decline in performance when transitioning from the 3 vs 4 task to the 3 vs 8 task, which as we suggested could be caused by the *8* being more similar to a *3* than the *4*. Note however that this trend is not systematic across all models.

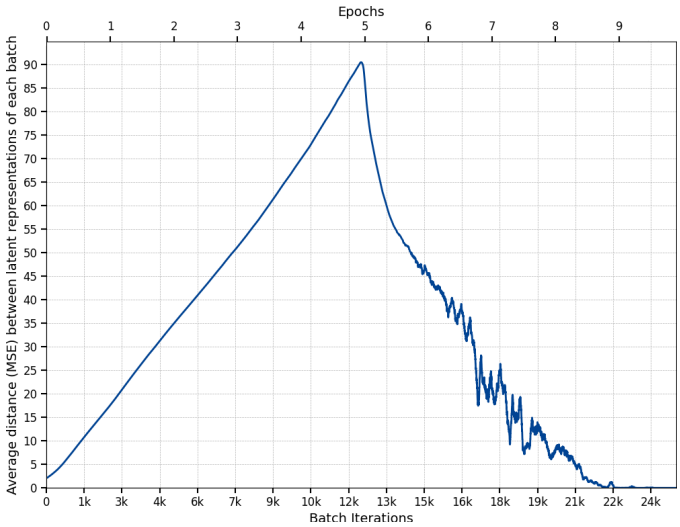

Fig. 6: Average MSE between latent representations during the training of the OgAE model for experiment 2. For the first 5 epochs the expander term is used, followed by the compactor term.

As in the main body, we find that on average the basic methods (AE-based, with *recons* or *ocsvm*) remain strong competitors, especially when compared with state-of-the-art propositions [30], [28], [3]. We also find that our proposed method achieves superior results on this benchmark, surpassing existing state-of-the-art models.

We also benchmarked a different set of corruptions from those used in the main experiments and observed similar results

F. Brain MRI registration and preprocessing pipeline for experiment 2

The brain MRI T1 preprocessing applied in this paper is based on a pipeline implemented in SPM12 and fully described in [45]. This pipeline includes a critical registration step that enables precise voxel-wise comparisons across subjects by aligning all images to a standardized anatomical space. Spatial normalization was performed using the unified segmentation algorithm (UniSeg) which includes segmentation of grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), correction for magnetic field inhomogeneities and spatial normalization to the standard brain template of the Montreal Neurological Institute (MNI). In this work, we used the default parameters for normalization and a voxel size of 1 × 1 × 1 mm. The cerebellum and brain stem were excluded from the spatially normalized images. The masking image in the reference MNI space was derived from the Hammersmith maximum probability atlas. On top of that, each image was intensity-normalized with: $X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$.

G. Additional results and visualizations for experiment 2

Figure 7 shows a control of the training database: notice the ventricles, considerably smaller than the ventricles of the other older control subjects in Figure 4 and Figure 9. Figure 8 and Figure 9 show two additional examples of both WMH

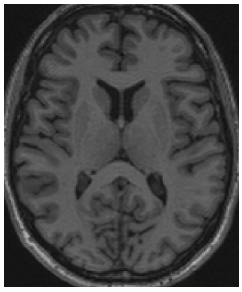

Fig. 7: Visualization of a central slice from the T1-weighted brain MRI of a control used for training (database [54], with younger mean age then the test databases).

patient and IXI controls, with their associated anomaly score maps for all the benchmarked methods.

As mentioned in the main body, figure 6 shows an example of training, in experiment 2, with the expander term (equation 7 $\beta_1 = 1, \beta_2 = 0$) for the first 5 epochs followed by the compactor term ($\beta_1 = 0$, $\beta_2 = 1$) for 5 other epochs. We study the average pairwise MSE between the latent representations, which is an indicator of their spread.

H. Autoencoder architectures

1) Experiment 1: The autoencoder architecture for all models of experiment 1 is the one used for the MNIST experiment in DVAESVDD [30]. It consists of a convolutional encoder and a symmetric decoder. The encoder comprises two convolutional layers (5×5 kernels, 4 and 8 filters), each followed by batch normalization, LeakyReLU activation, and 2×2 max pooling. The latent representation is obtained via a fully connected layer of dimension 32 (meaning reduction factor of 24.5). The decoder mirrors the encoder, employing a dense layer to reshape the latent space, followed by two transposed convolutional layers (5×5 kernels, 8 and 4 filters) interleaved with batch normalization, LeakyReLU activation, and 2×2 upsampling. A final transposed convolution (5×5, 1 filter) with a sigmoid activation reconstructs the input. The model is trained with mean squared error as the reconstruction loss, optimized with Adam (learning rate: 1e-3), with a batch size of 100, for 20 epochs.

2) Experiment 2: We present in this section the autoencoder architecture used in experiment 2. The encoder consists of four convolutional layers: a 5×5 layer with 3 filters, followed by three successive 3×3 layers with 4, 12, and 16 filters, respectively. Each convolutional layer is paired with batch normalization and GELU activation. The decoder mirrors this structure precisely. It begins with three 3×3 transposed convolutional layers with 12, 4, and 3 filters, each followed by batch normalization and GELU activation, and concludes with a 5×5 transposed convolution and a sigmoid activation. For training, we optimize the model using mean squared error (MSE) with the Adam optimizer (learning rate: 1e-3), trained for 10 epochs with a batch size of 100.

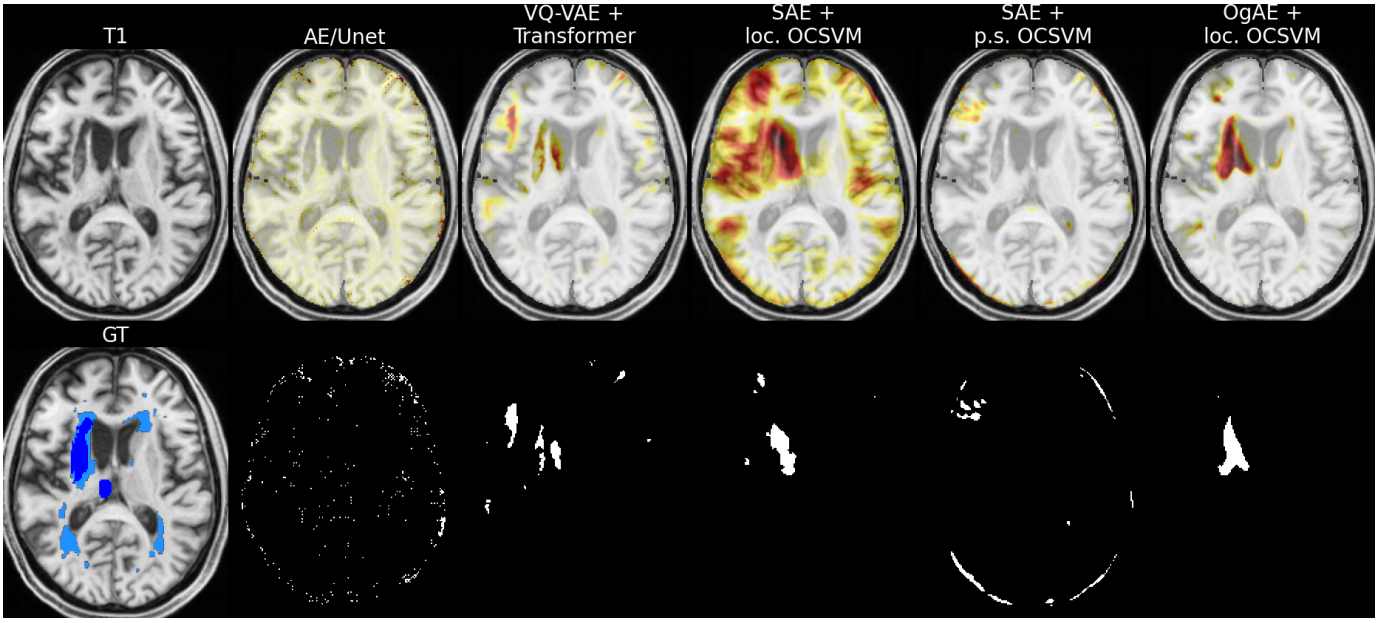

Fig. 8: Visualization of a central slice from the T1-weighted brain MRI of a WMH patient (SIN67). The ground truth (GT) is overlaid, with light blue indicating white matter lesions ("hyperintensities" on FLAIR MRI but not on T1) and blue representing other pathologies. Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.

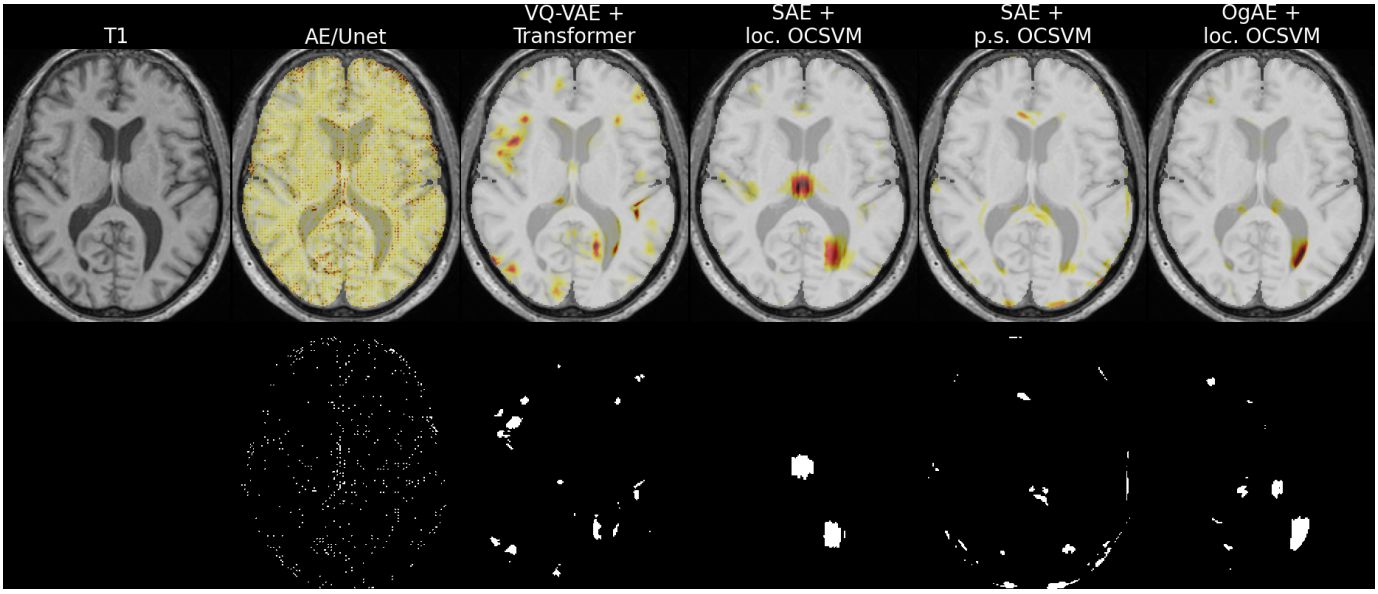

Fig. 9: Visualization of a central slice from the T1-weighted brain MRI of a IXI control (IXI072-HH-2324). Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.