MTL-MAD: MULTI-TASK LEARNERS ARE EFFECTIVE MEDICAL ANOMALY DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly detection in medical images is a challenging task, since anomalies are not typically available during training. Recent methods leverage a single pretext task coupled with a large-scale pre-trained model to reach state-of-the-art performance. Instead, we propose to learn multiple self-supervised and pseudo-labeling tasks from scratch, using a joint model based on Mixture-of-Experts (MoE). By carefully integrating multiple proxy tasks, the joint model effectively learns a robust representation of normal anatomical structures, so that anomaly scores can be derived based on how well the multi-task learner (MTL) solves each task during inference. We perform comprehensive experiments on BMAD, a recent benchmark that comprises a broad range of medical image modalities. The empirical results indicate that our multi-task learner is an effective anomaly detector, outperforming all state-of-the-art competitors on BMAD. Moreover, our model produces interpretable anomaly maps, potentially helping physicians in providing more accurate diagnoses. To reproduce the reported results, we publicly release our code at `https://anonymous.4open.science/r/MTL-MAD`.

1 INTRODUCTION

Medical anomaly detection from images is a difficult problem, largely due to its inherent open-set nature (Wang et al., 2023; Bercea et al., 2023b). In many clinically important settings, the abnormalities are rare, heterogeneous, and expensive to annotate (Baur et al., 2021; Kascenas et al., 2023; Cai et al., 2025), whereas large collections of normal or weakly curated data are substantially easier to obtain (Baur et al., 2021; Bao et al., 2024; Cai et al., 2025). This makes unsupervised anomaly detection (Schlegl et al., 2017; 2019; Baur et al., 2021) especially attractive for medical imaging: instead of requiring exhaustive pathology annotations, a model can learn the regularities of normal anatomy and then flag deviations that warrant downstream review. The difficulty, however, is not only the scarcity of positive (anomalous) examples. Medical anomalies vary widely in scale, texture, semantics, and prevalence (Kascenas et al., 2023; Bao et al., 2024; Cai et al., 2025), while the appearance of normal anatomy is itself strongly dependent on organ, modality, acquisition protocol, and dataset construction. Bao et al. (2024) make the heterogeneity explicit by introducing BMAD, an unsupervised anomaly detection benchmark spanning six datasets over brain MRI, liver CT, two retinal OCT settings, chest X-ray, and histopathology. The authors found that no single pretext task consistently dominates across settings, and that different approaches exhibit different trade-offs between domains.

Nevertheless, existing unsupervised methods usually rely on a single proxy (pretext) objective (Schlegl et al., 2019; Li et al., 2021; Tan et al., 2022; Deng & Li, 2022; Roth et al., 2022), and, by extension, on a single notion of normality. For instance, reconstruction-based approaches learn to restore or translate images, and then score abnormalities through residuals (Schlegl et al., 2017; 2019; Chen & Konukoglu, 2018; Baur et al., 2021). Yet, both benchmark analyses and recent medical anomaly detection studies note a recurring limitation: well-trained generators can partially reconstruct pathology and may fail to reflect semantically meaningful or context-level abnormalities in simple pixel-space errors (Heer et al., 2021; Bercea et al., 2023a; Baur et al., 2021). Self-supervised and feature-based alternatives address different weaknesses, yet each proxy objective emphasizes only one failure mode, such as local corruption, distributional deviation, or one-class compactness (Li et al., 2021; Tan et al., 2022; Roth et al., 2022; Deng & Li, 2022). The mixed results reported on medical benchmarks are consistent with this picture: no single family of objectives transfers uniformly across modalities, organs, and anomaly types (Bao et al., 2024; Cai et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

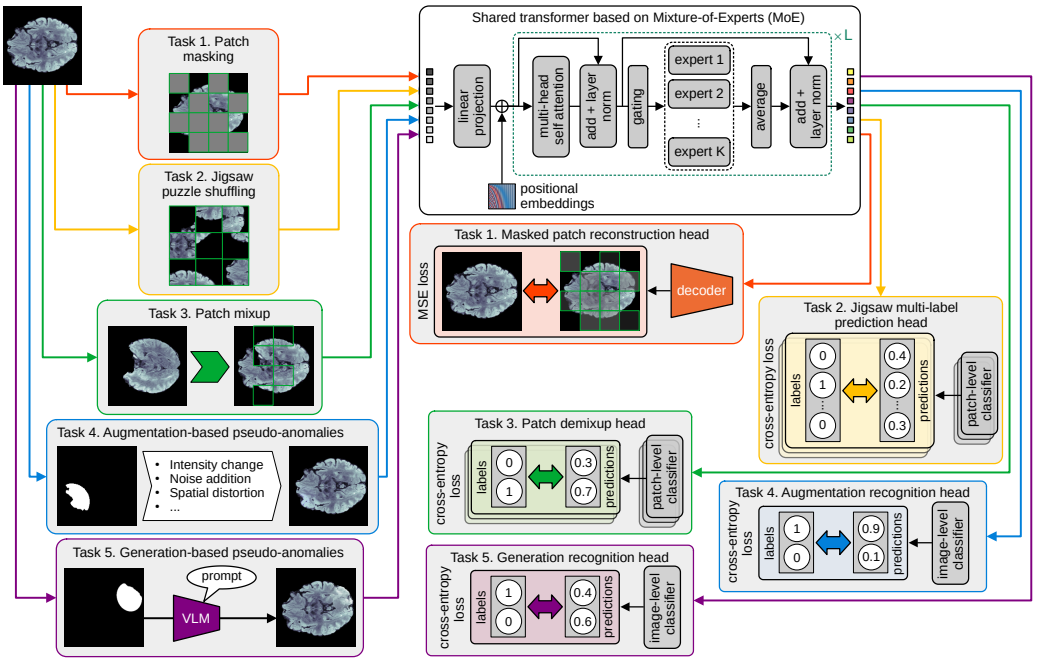


Figure 1: Overview of our multi-task learner for medical anomaly detection (MTL-MAD). By meticulously integrating multiple proxy tasks, the joint model learns a robust representation of normal anatomical structures. During inference, anomaly scores are derived based on how well the multi-task learner solves each task. Positional embeddings are discarded for tokens corresponding to the jigsaw puzzle solving task, to avoid making the task trivial. Best viewed in color.

To address the limitations of casting medical anomaly detection as a single proxy task, we propose MTL-MAD, a multi-task learning (MTL) framework for self-supervised medical anomaly detection (MAD), which combines multi-task learning with task-conditioned Mixture-of-Experts (MoE) (Shazeer et al., 2017) inside the transformer backbone. MTL-MAD shares global visual context across tasks, while allowing task-specific specialization through routed feed-forward experts. This design is motivated by prior work on MoE-based multi-task learning (Ma et al., 2018; Liang et al., 2022), showing that task-conditioned routing can disentangle parameter subspaces and mitigate training conflicts that arise under naïve shared optimization (Sener & Koltun, 2018; Ma et al., 2018; Liang et al., 2022). In our setting, different proxy tasks are assigned to different experts or expert subsets, and each task is paired with its own prediction head. As illustrated in Figure 1, we aggregate a variety of proxy tasks, namely masked image modeling (He et al., 2022; Hondru et al., 2025; Xie et al., 2022), jigsaw puzzle solving (Noroozi & Favaro, 2016; Taleb et al., 2021; Wang et al., 2022), synthetic anomaly prediction (Li et al., 2021; Tan et al., 2022), and our newly proposed DEMIXUP, a patch-level mixing-and-disentanglement objective that mimics local corruptions. These tasks are meticulously integrated to precisely elicit certain abilities during training on normal tissue: masked image modeling induces the ability to generate normal anatomical structures based on neighboring tissue, jigsaw puzzle solving allows the model to understand where certain normal anatomical features are located, DEMIXUP makes the model comprehend when a certain structure does not belong to the surrounding context, and pseudo-anomaly detection allows the model to form a rough boundary separating normal samples from everything else. Together, these proxy tasks induce a comprehensive characterization of normal anatomical structures. Moreover, the proposed self-supervised tasks have direct implications towards accurately detecting anomalies, grounded in the following working hypotheses: reconstructing masked patches is more accurate for normal patches than for anomalous patches, rearranging patches with normal tissue is easier than rearranging patch containing abnormal tissue, detecting tissue that differs from surrounding tissue is more obvious when the respective tissue is anomalous, labeling anomalies via a rough boundary around the normality class is easier than having no clue how tight this boundary should be.

Our key intuition is that the proposed objectives expose complementary abnormality cues that are difficult to capture with any single pretext task. Rather than forcing these signals to collapse into a single score during training, MTL-MAD allows them to be absorbed into the joint backbone, before

converting the outputs of the task-specific heads into anomaly scores that are fused at inference by a weighted aggregation scheme. This produces a detector that can leverage heterogeneous evidence without assuming that all tasks should share the same calibration or dominate the final decision in every modality. To validate our hypothesis, we conduct experiments on BMAD (Bao et al., 2024), a comprehensive benchmark comprising six medical anomaly detection datasets. Our empirical results indicate that MTL-MAD yields state-of-the-art performance on all six benchmarks, outperforming all competing methods. Unlike state-of-the-art competitors, MTL-MAD achieves high performance without any model pre-training, thereby confirming that its power stems solely from the joint learning of multiple and diverse proxy tasks.

In summary, our contributions are as follows:

- We propose MTL-MAD, a task-conditioned MoE framework for unsupervised medical anomaly detection within a single ViT backbone.
- We introduce DEMIXUP, a patch-level self-supervised objective designed to capture local corruption and disentanglement cues complementary to reconstruction and spatial reasoning.
- We show that reconstruction, spatial reasoning, local corruption detection, and synthetic anomaly discrimination provide complementary views of normality, when trained jointly, but with specialized experts.
- We achieve state-of-the-art results on all six BMAD datasets spanning five medical domains.

2 RELATED WORK

Anomaly detection in medical imaging has been explored through several complementary approaches. Reconstruction-based methods (Cai et al., 2024; Walluscheck et al., 2024; Zimmerer et al., 2019; Zhou et al., 2022; Baur et al., 2021) train exclusively on normal anatomy, treating deviations in reconstructed outputs as anomaly signals. Recent work has pushed this paradigm forward through diffusion processes (Pinaya et al., 2022; Behrendt et al., 2023; Baugh et al., 2025; Bi et al., 2025). Behrendt et al. (2024) improved robustness to imperfect reconstructions by sampling multiple pseudo-healthy outputs and measuring pixel-level deviation via the Mahalanobis distance. In contrast, Beizaee et al. (2025) employed a masked latent diffusion approach that focuses exclusively on correcting anomalous regions, thereby preserving structural information that is often degraded in full-image diffusion methods. The Q-Former Autoencoder (Dalmonte et al., 2026) modernized the autoencoder paradigm by coupling frozen encoders based vision foundation models with a learnable bottleneck and perceptual loss, achieving strong unsupervised performance across diverse modalities.

Vision-language models (VLMs), particularly CLIP, have contributed to the increasing prevalence of few-shot and zero-shot anomaly detectors (Zhou et al., 2024; Huang et al., 2024; Zhang et al., 2024). Huang et al. (2024) inserted residual adapters at multiple CLIP encoder stages and combined zero-shot text-based with few-shot memory-bank comparison. Zhang et al. (2024) proposed MediCLIP, a CLIP model fine-tuned on synthetic anomaly generation tasks designed to mimic clinically plausible patterns, achieving substantial gains with less than 1% of the training set. Shiri et al. (2025) introduced MadCLIP, a dual-branch adapter architecture with contrastive training under SigLIP loss, explicitly separating normal and abnormal feature subspaces without relying on memory banks or synthetic anomalies. Ma et al. (2025) proposed AA-CLIP to mitigate the intrinsic anomaly-unawareness of CLIP via a two-stage sequential adaptation of both text and visual encoders. Beyond binary detection, Zhou et al. (2025) designed UltraAD by extending a few-shot CLIP model to fine-grained multi-class ultrasound classification, incorporating image-aware dynamic prompting and a memory-boosted support module for cross-domain robustness. Unified and cross-domain anomaly detection has emerged as a key frontier for VLMs (Jeong et al., 2023; Zhou et al., 2024; Gu et al., 2025). Zhu & Pang (2024) developed InCTRL, a single model trained across industrial, medical, and semantic anomaly benchmarks without any target-domain fine-tuning, using few-shot normal images as in-context prompts and computing multi-scale CLIP residuals. The cross-modal generalization setting (Mao et al., 2025) requires models trained on known modalities to transfer to entirely unseen ones. Their *transferable visual prototypes* address this by learning modality-agnostic normal and abnormal semantics in visual feature space, unified at inference using visual similarity weighting.

Multi-task learning offers a complementary mechanism for building generalizable representations. However, dense parameter sharing across tasks frequently induces inter-task interference, where competing gradients corrupt shared weights and degrade overall performance. In natural language processing (NLP), this was addressed through sparsely activated Mixture-of-Experts architectures (Gupta

et al., 2022) with task-specific routing, directing tokens from different tasks to specialized expert sub-networks, while preserving the computational budget of a dense model. This design provided measurable gains in low-resource transfer and resistance to catastrophic forgetting on multiple NLP tasks. To the best of our knowledge, we are the first to transfer these properties to medical anomaly detection, where a shared model must reconcile heterogeneous supervision signals, across different imaging modalities as diverse as MRI, CT, chest X-ray and histopathology.

Some self-supervised approaches compose multiple pretext tasks, such as jigsaw, rotation, and re-colorization (Jézéquel et al., 2023), or pair contrastive learning with domain-aware synthetic anomalies (Tian et al., 2023). However, these methods typically aggregate their signals through independent score fusion at inference or decoupled sequential pipelines (Bozorgtabar et al., 2023), rather than optimizing all objectives jointly within a single model. Our method adopts a fundamentally different strategy: rather than inheriting representations from pre-trained backbones, adapting foundation models through lightweight modules, or treating tasks as independent components, we propose a joint Mixture-of-Experts model that learns multiple self-supervised and pseudo-labeling objectives simultaneously from scratch, with experts specializing in distinct proxy tasks. Anomaly scores emerge naturally from how poorly each task is solved at inference, additionally leading to interpretable pixel-level anomaly maps that some competing approaches do not provide.

Our framework is connected to self-supervised multi-task learning frameworks that have been successful in video anomaly detection (Barbalau et al., 2023; Georgescu et al., 2021). While such methods perform joint training via multiple proxy tasks, there are two fundamental differences from our approach. First, the proxy tasks for video anomaly detection are designed to capture normal temporal dynamics and object appearance, while our tasks, e.g. DEMIXUP, are designed to grasp a deep understanding of normal anatomical structures. Second, the shared encoders used in video anomaly detection are based on 2D or 3D convolutions, which may experience difficulties in learning too many, potentially conflicting, proxy tasks, as empirically observed by Barbalau et al. (2023). In contrast, we train a Mixture-of-Experts framework, in which tasks are routed to different experts, enabling direct task decoupling to avoid conflicting objectives.

Some works (Jézéquel et al., 2021; Jézéquel et al., 2023; Zhang et al., 2023) have applied multi-task learning in image anomaly detection. However, these methods are not designed for the medical domain. For instance, the discriminative and generative objectives of Jézéquel et al. (2023) live in two separate networks trained with separate gradient steps. Rather than having different encoders and separate branches for each task, our method employs a shared encoder with mixture-of-experts, where each expert specializes on a specific task. Moreover, we integrate Fast Adaptive Multitask Optimization (FAMO) Liu et al. (2023a) to balance objectives and eliminate hyperparameters for loss weighting. Unlike (Jézéquel et al., 2021; Jézéquel et al., 2023; Zhang et al., 2023), we introduce MoE and FAMO to eliminate the negative effects of proxy task misalignment during training. Our configuration has multiple advantages: parameter efficiency, better scalability to additional tasks, and a richer low-level representation, while related information can still be shared across tasks. More importantly, the set of pretext tasks differs. Jézéquel et al. (2023) relies on colorization, which is not applicable to medical modalities such as CT, MRI, OCT, and X-ray, since these images are single-channel rather than color. This limits the applicability of (Jézéquel et al., 2023) to medical image anomaly detection, while our MTL-MAD framework is mainly designed for medical anomaly detection, where images usually have no color. Moreover, the proxy tasks in (Jézéquel et al., 2023) target structure, colorimetry, and texture, but none of them captures contextual incompatibility. Our DeMixUp task addresses exactly this cue. It requires the model to determine, for each patch, whether the tissue originates from the image itself or from a donor image, thereby learning what does not belong in a given anatomical context. Just as CutPaste (Li et al., 2021), our proxy task is inspired by mixup (Zhang et al., 2018) and cutmix (Yun et al., 2019) augmentations. Unlike CutPaste (Li et al., 2021), which discriminates between original and altered images[1], we classify each patch as pertaining to the original image or the donor one. This enables the model to learn a more contextualized understanding of what constitutes an anomaly within the image, rather than a holistic representation.

3 METHOD

We propose a multi-task learning framework for unsupervised medical anomaly detection. Our framework is constructed on a Vision Transformer (ViT) backbone (Dosovitskiy et al., 2021), enhanced with Mixture-of-Experts (MoE) layers (Shazeer et al., 2017), where experts specialize themselves on

[1]An altered image contains patches transferred from another image.

designated proxy tasks. Our model jointly solves three self-supervised pretext tasks: masked image modeling, jigsaw puzzle solving and patch-level DEMIXUP, alongside two explicit pseudo-anomaly classification objectives trained on synthetically generated/augmented anomalies. Each task is routed through a dedicated subset of experts, enabling task-specific feature specialization, while retaining a shared attention-based backbone. At inference time, an anomaly score is obtained for each task, and then aggregated into a unified anomaly score. An overview of our approach is illustrated in Figure 1.

3.1 ARCHITECTURE

Backbone. We adopt a standard ViT architecture consisting of L transformer encoder layers operating on non-overlapping image patches, i.e. the input image $X \in \mathbb{R}^{H\times W\times C}$ is split into $N = (H/P) \times (W/P)$ equally-sized patches of $P \times P$ pixels. Each patch is linearly projected to a d-dimensional embedding (token). A learnable positional embedding is further added to each token. Each transformer layer consists of a multi-head self-attention (MHSA) block, followed by a feed-forward network (FFN), with layer normalization (LN) and a residual connection:

$$X^l = X^{l-1} + \mathrm{MHSA}(\mathrm{LN}(X^{l-1})), \quad X^l = X^l + \mathrm{FFN}(\mathrm{LN}(X^l)), \tag{1}$$

where X^l represents the sequence of tokens (patches) at an arbitrary transformer layer at index l.

Task-controlled Mixture-of-Experts. We replace a subset of the standard FFN layers with a MoE module, where the experts are specifically designated to each task. More precisely, unlike standard MoE formulations, our routing is deterministic, such that the tokens corresponding to a task only go through the specifically allocated experts, and the output is computed by taking the average of the assigned experts. Each expert consists of a two-layer feed-forward network with GeLU activations. During training, we apply expert dropout, randomly dropping experts, but ensuring that at least one expert per task is active. This acts as a regularizer that prevents co-adaption of same-task experts.

3.2 PROXY TASKS

For each proxy task, a custom independent head is attached to the shared encoder. The tasks are carefully chosen and designed such that they empower the model to learn a rich and structured representation of the normal healthy tissue, all of them having complementary capabilities. Combined together, the proposed tasks produce a greater effect: reconstruction-based objectives capture local consistency, permutation-based objectives encode global structure, mixture-based perturbations highlight contextual incompatibilities, and synthetic anomalies provide coarse discrimination. This multi-faceted training paradigm equips the model with both generative and discriminative cues, ultimately enhancing its ability to detect subtle deviations from normal anatomical patterns at inference time.

Task 1: Masked image modeling. Masked image modeling (MIM) (He et al., 2022; Hondru et al., 2025) consists of inferring masked regions from visible context, thus learning to reconstruct plausible normal anatomical structures. Consequently, abnormal regions will violate the learned reconstruction priors. Starting from the assumption that normal anatomical structures exhibit strong local continuity, when the model is forced to reconstruct the missing regions, it encourages the model to internalize the distribution of healthy tissue. Given N non-overlapping patches $\{x_i\}_{i=1}^N$, a random subset $\mathcal{M} \subset \{1, \ldots, N\}$ is masked, and only the visible tokens $\{x_i \mid i \notin \mathcal{M}\}$ are passed through the shared backbone. Using a lightweight transformer decoder with 8 blocks (16 heads with an embedding dimension of 8), it reconstructs the full set of patches $\hat{x}_i$, with the objective of minimizing the reconstruction error for the masked tokens through the mean squared error (MSE):

$$\mathcal{L}_{\mathrm{MIM}}(X) = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \|x_i - \hat{x}_i\|_2^2 . \tag{2}$$

Task 2: Jigsaw puzzle solving. While MIM captures local dependencies, it does not explicitly enforce global spatial awareness. To encourage this, we introduce the jigsaw puzzle proxy task, which requires the model to recover the correct spatial arrangement of shuffled patches. Working with many small patches would make the task intractable. Therefore, the first step is to group the N non-overlapping patches $\{x_i\}_{i=1}^N$ into larger tiles, to form a grid of $T \times T$ tiles, such that $T^2 \ll N$, reducing the complexity of the proxy task. A random permutation is applied on the T^2 tiles before being fed into the model. The shuffled tiles are encoded via the shared backbone, but not before discarding the positional embeddings, which would make the task trivial. A simple two-layer FFN is trained to predict the original tile positions, for each individual tile. Unlike original formulations for solving jigsaw puzzles (Noroozi & Favaro, 2016), we treat the task as a multi-label classification instead of trying to predict the exact permutation of tiles. For each tile, the model predicts the

probability vector $\hat{y}_i \in [0,1]^{T^2}$, not enforcing mutual exclusivity between predictions, e.g. allowing the model to predict the same position for two different tiles. This formulation avoids over-penalizing the model for small mistakes. We employ the binary cross-entropy loss over tile positions, where $y_{i,j} = 1$ if tile i originates from position j, and 0 otherwise:

$$\mathcal{L}_{\text{Jigsaw}}(X) = -\frac{1}{T^2}\sum_{i}^{T^2}\sum_{j}^{T^2} y_{i,j} \cdot \log(\hat{y}_{i,j}). \tag{3}$$

Task 3: Patch DEMIXUP. This task sharpens the sensitivity of the model to contextual mismatches, which are often indicative of anomalies. Inspired by mixup (Zhang et al., 2018) and cutmix (Yun et al., 2019), patch mixup consists of blending patches from distinct images, then predicting whether each patch token originates from the base or donor image. More precisely, given a base image X' and a donor image X'', a composite image X^{mix} is constructed by replacing a subset of patches $\mathcal{C} \in \{x'_i\}$ with patches at corresponding locations in $\{x''_i\}$. Patches from $\{x_i^{\text{mix}}\}$ are encoded and passed through a classification head based on a two-layer FFN. For each patch x_i^{mix}, the classifier predicts a binary label, indicating whether it originates from the base or donor image. The task is trained with binary cross-entropy loss:

$$\mathcal{L}_{\text{DEMIXUP}}(X^{\text{mix}}) = -\frac{1}{N}\sum_{i}^{N} y_i \cdot \log \hat{y}_i + (1-y_i)\cdot \log(1-\hat{y}_i). \tag{4}$$

Task 4: Augmentation-based anomaly classification. We design a pseudo-anomaly classification task, where we employ aggressive image augmentations to produce pseudo-anomalous samples. The aim is to enforce an explicit separation between normal and pseudo-abnormal patterns. When introducing pseudo-anomalies, we provide a weak supervisory signal that helps the model to better approximate a decision boundary around the normal data manifold. One method for generating pseudo-anomalies is applying augmentations on healthy anatomical structures. We first generate anomaly masks by applying unsupervised k-means clustering over pixels, where the number of clusters (superpixels) k is randomly chosen in a given interval. One cluster is randomly chosen for the image manipulation process, which involves operations such as intensity changes, noise addition, spatial distortions, and synthetic artifact generation (opacity and structural defects). Let $X^{\text{aug}} \in \mathbb{R}^{H\times W\times C}$ denote an augmented image corresponding to an original training image X. The classification head is composed of a two-layer FFN that takes the `[CLS]` token from the backbone as input. The training loss function is the binary cross-entropy:

$$\mathcal{L}_{\text{Aug-cls}}(Z) = -(y \cdot \log \hat{y} + (1-y)\cdot \log(1-\hat{y})), \tag{5}$$

where Z is either an original image X, or an augmented image X^{aug}.

Task 5: Generation-based anomaly classification. We introduce an additional pseudo-anomaly classification task, with an analogous classification head and training objective. Different from the augmentation-based pseudo-anomaly generation method, the artificial samples are obtained via generative vision-language models (VLMs), where a healthy sample and a prompt are used to insert an anomaly in the input image, via inpainting. The prompts are specific to the medical domain of the real image that is perturbed. We detail these prompts in the supplementary material. To generate a sample $X^{\text{gen}} \in \mathbb{R}^{H\times W\times C}$, we randomly select a frozen VLM from the following three: Qwen-Image (Wu et al., 2025), Flux (Labs et al., 2025), and Stable Diffusion (Rombach et al., 2022). The classification head is trained via cross-entropy:

$$\mathcal{L}_{\text{Gen-cls}}(Z) = -(y \cdot \log \hat{y} + (1-y)\cdot \log(1-\hat{y})), \tag{6}$$

where Z is either an original image X, or a partially generated image X^{gen}. The inpainting masks are generated using overlapping ellipses with randomly assigned quantities, sizes, and center locations. For images dominated by background space, we constrain the ellipse centers to fall within the organ bounds to ensure that the inpainting process targets actual tissue. We emphasize that the random generation and the overlapping of these ellipses naturally produce final masks with varied and complex shapes.

3.3 MULTI-TASK LEARNING

Curriculum learning / Progressive training. To improve training stability, we take a different approach from the standard practice: rather than starting with all tasks from the beginning, we integrate the tasks gradually, in a specific order, inspired by curriculum learning (Bengio et al., 2009; Soviany et al., 2022). During the first phase, only self-supervised tasks (MIM, Jigsaw and DEMIXUP)

are active, allowing the encoder to learn robust visual representations. In the second phase, the synthetic anomaly detection tasks are activated for joint training, harnessing and further boosting the already learned representations.

Dynamic task weighting. A common issue in multi-task learning is the conflicting gradients problem (Liu et al., 2021). Performance on some tasks might degrade, while attempting to improve performance on others, because their individual task gradients point in opposite directions. To mitigate this problem, we train our model via Fast Adaptive Multitask Optimization (FAMO) (Liu et al., 2023a), aiming to decrease all losses at an equal relative rate. In our scenario, we define $\mathbf{L}^t = \left[\mathcal{L}^t_{\text{MIM}}, \mathcal{L}^t_{\text{Jigsaw}}, \mathcal{L}^t_{\text{DeMixUp}}, \mathcal{L}^t_{\text{Aug-cls}}, \mathcal{L}^t_{\text{Gen-cls}}\right]$ as the vector of loss terms at training iteration t. Let $\mathbf{L}^t_i$ denote the loss function at index i from the vector $\mathbf{L}^t$. The final loss function of our joint training pipeline is defined as a weighted sum with terms from this vector, as follows:

$$\mathcal{L}_t(X, \theta) = c_t \sum_{i=1}^{5} p^t_i \cdot \log \mathbf{L}^t_i, \tag{7}$$

where c_t is a normalization constant, θ are the parameters of the multi-task learner, and $\mathbf{p}^t = \{p^t_i\}^5_{i=1}$ represent the weights automatically associated to each $\mathbf{L}^t_i$. These weights are computed using the softmax function over a set of learnable logits $\mathbf{p}^t = \text{softmax}(\mathbf{w}^t)$. The update rule of the logits at training iteration $t+1$ is based on the relative loss change from the previous step:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \beta \cdot \left(\frac{\partial \mathbf{p}^t}{\partial \mathbf{w}^t}\right)^\top \cdot \left(\log \frac{\mathbf{L}^t}{\mathbf{L}^{t+1}}\right), \tag{8}$$

where β is a learning rate and the logits $\mathbf{w}^t$ are all initialized with zero for the first iteration. With this update rule, FAMO forces the weights to automatically shift toward tasks that are not improving fast enough, balancing the relative decrease rates of the individual losses over time.

3.4 Anomaly Scoring

At test time, we compute an anomaly score based on each proxy task, and combine them into a final prediction. For MIM, the score is the mean reconstruction error over the masked patches, while for jigsaw puzzle solving, the cross-entropy serves as the score, in the end both being normalized via $1 - \exp(-x)$. For DeMixUp, we take the sigmoid output as the anomaly probability of each patch, then use the mean of the top-$k(\%)$ ratio patches, where k is set to 10%. Similarly, for both pseudo-anomaly classification tasks, the sigmoid output is used directly as an anomaly score. To generate anomaly maps, we compute per-patch anomaly scores for the tasks that produce spatially-aligned outputs (MAE, Jigsaw, and DeMixUp), then upsample each score map to the original image resolution via interpolation. Once all maps share the same resolution, we normalize them to a common scale and combine them linearly, using the same per-task weights applied in image-level scoring.

4 Experiments

Datasets. We evaluate our method on the BMAD benchmark (Bao et al., 2024), which comprises six datasets spanning five medical domains. Specifically, the benchmark includes BraTS2021 (Baid et al., 2021) for brain magnetic resonance imaging (MRI) analysis. RESC (Hu et al., 2019) and OCT2017 (Kermany et al., 2018) represent the optical coherence tomography (OCT) domain. The liver computed tomography (CT) domain is represented by a combination of BTCV (Landman et al., 2015) and LiTs (Bilic et al., 2023) datasets. The chest X-ray modality is covered by RSNA (Wang et al., 2017), and cellular-level analysis is captured through the CAMELYON16 (Ehteshami Bejnordi et al., 2017) pathology dataset. We train our pipeline on the official training splits of each of the six datasets, which range in size from 3,000 samples to 26,000 samples. For each training sample, we synthesize two pseudo-anomalies, one via a generative VLM, and the other through classic pixel-level perturbations. We randomly select one of the two pseudo-abnormal images during mini-batch construction.

Hyperparameters. We train the pipeline for 100 epochs with a batch size of 64 samples and a starting learning rate of 10^{-3}. We employ a learning rate scheduler that decays the learning rate down to a minimum of 10^{-6} using a factor of 0.5 when the validation loss does not change significantly. The images are resized to a resolution of 256×256 pixels, and we use a patch size of $P = 16$ in all cases. The hyperparameter configurations of our pipeline can vary between datasets. Specifically, we set the number of transformer encoder blocks L and the number of attention heads within the range of $[4, 12]$, the hidden dimension $d \in [128, 768]$, and the number of experts K between 5 and 10. The

Table 1: Performance comparison of MTL-MAD against leading methods from the literature. The comparison is performed on the entire BMAD benchmark, covering all datasets to demonstrate the consistency of our approach compared to previous state-of-the-art methods. The best score on each dataset is highlighted in **blue bold**, while the second-best score is highlighted in orange underline. We report average image-level AUROC scores and standard deviations over three runs.

Method	BraTS2021	RESC	RSNA	OCT2017	BTCV+LiTs	CAMELYON16
f-AnoGAN (Schlegl et al., 2019)	77.3 ± 0.18	77.4 ± 0.85	55.6 ± 0.09	73.4 ± 1.85	58.5 ± 0.21	69.5 ± 1.98
GANomaly (Akcay et al., 2019)	74.8 ± 1.93	52.6 ± 3.95	62.9 ± 0.65	70.5 ± 9.98	54.6 ± 3.06	54.4 ± 2.57
DRAEM (Zavrtanik et al., 2021)	62.4 ± 9.03	83.2 ± 8.21	67.7 ± 1.72	88.0 ± 8.36	69.9 ± 3.86	52.3 ± 0.77
UTRAD (Chen et al., 2022)	82.9 ± 2.32	89.4 ± 1.92	75.6 ± 1.24	96.8 ± 0.56	55.8 ± 5.66	69.9 ± 4.6
DeepSVDD (Ruff et al., 2018)	87.0 ± 0.66	74.2 ± 1.29	64.5 ± 3.17	76.8 ± 1.37	53.9 ± 1.84	60.9 ± 1.82
CutPaste (Li et al., 2021)	78.8 ± 0.67	90.2 ± 0.61	82.6 ± 1.22	96.8 ± 0.62	59.3 ± 4.86	75.2 ± 0.41
SimpleNet (Liu et al., 2023b)	82.5 ± 3.34	76.2 ± 7.46	69.1 ± 1.27	94.7 ± 2.17	72.3 ± 2.68	62.4 ± 3.71
MKD (Salehi et al., 2021)	81.5 ± 0.36	89.0 ± 0.25	82.0 ± 0.12	96.7 ± 0.26	60.7 ± 1.19	77.5 ± 0.27
RD4AD (Deng & Li, 2022)	89.5 ± 0.91	87.8 ± 0.87	67.6 ± 1.11	97.3 ± 0.79	60.4 ± 1.17	66.8 ± 0.71
STFPM (Yamada & Hotta, 2022)	83.0 ± 0.67	84.8 ± 0.50	72.9 ± 1.96	96.8 ± 0.23	61.8 ± 1.58	66.4 ± 1.01
PaDiM (Defard et al., 2021)	79.0 ± 0.38	75.9 ± 0.54	77.5 ± 1.87	91.8 ± 0.96	50.8 ± 0.61	67.3 ± 0.32
PatchCore (Roth et al., 2022)	91.7 ± 0.36	91.6 ± 0.10	76.1 ± 0.67	98.6 ± 0.03	60.3 ± 0.76	69.3 ± 0.21
CFA (Lee et al., 2022)	84.4 ± 0.87	69.9 ± 0.26	66.8 ± 0.23	79.5 ± 0.56	62.0 ± 1.08	65.6 ± 0.10
CFLOW (Gudovskiy et al., 2022)	74.8 ± 5.32	75.0 ± 5.81	71.5 ± 1.49	85.4 ± 2.11	50.8 ± 4.47	55.7 ± 1.97
CS-Flow (Rudolph et al., 2022)	90.9 ± 0.83	87.3 ± 0.58	83.2 ± 0.46	98.5 ± 0.28	59.4 ± 0.54	68.4 ± 0.42
P-VQ (Kim et al., 2024)	94.3 ± 0.23	89.0 ± 0.48	79.2 ± 0.04	-	-	-
QFAE (Dalmonte et al., 2026)	94.3 ± 0.18	91.8 ± 0.55	83.8 ± 0.46	-	65.5 ± 1.96	-
MTL-MAD (Ours)	$\mathbf{95.5 \pm 0.19}$	$\mathbf{92.8 \pm 1.28}$	$\mathbf{86.0 \pm 0.01}$	$\mathbf{98.9 \pm 0.22}$	$\mathbf{82.7 \pm 2.34}$	$\mathbf{78.1 \pm 1.51}$

Table 2: Ablation study on BraTS2021, highlighting the impact of each task on the final performance. The performance is measured in terms of image-level AUROC.

MIM	Jigsaw	DEMIXUP	Augmented AD	Generated AD	AUROC
✓	✗	✗	✗	✗	65.54
✗	✓	✗	✗	✗	84.38
✗	✗	✓	✗	✗	87.77
✗	✗	✗	✓	✗	88.36
✗	✗	✗	✗	✓	90.72
✓	✓	✗	✗	✗	84.78
✓	✓	✓	✗	✗	93.40
✓	✓	✗	✓	✗	90.50
✓	✓	✗	✗	✓	91.80
✓	✓	✓	✓	✗	93.87
✓	✓	✓	✓	✓	95.49

exact values are established via grid search on the official validation splits. All baselines received equivalent effort in hyperparameter tuning on the same validation splits.

Results. In Table 1, we present our main results, comparing MTL-MAD and several leading models from the literature using the complete BMAD benchmark. Our analysis reveals that MTL-MAD outperforms all existing methods on all six datasets, with no exception. MTL-MAD surpasses its competitors on both challenging datasets, such as BTCV+LiTs and CAMELYON16, as well as on nearly saturated datasets, namely OCT2017. We observe a significant improvement on the BTCV+LiTs dataset, where our model achieves absolute gains of over 10% in terms AUROC over the strongest competitor. This result is particularly important, since BTCV+LiTs is the most challenging dataset in BMAD. Due to its difficulty, some recent methods struggle (Dalmonte et al., 2026) or do not even attempt (Kim et al., 2024) to perform anomaly detection on BTCV+LiTs.

Ablation studies. In Table 2, we present an ablation study on BraTS2021, which highlights the contribution of each proxy task to the final performance. The results demonstrate the complementarity between the proposed proxy tasks. Moreover, our novel DEMIXUP task yields one of the strongest individual performance levels, and contributes with the largest enhancement to the overall pipeline. The impact is clearly reflected in the evaluation measure, where the AUROC increases from 84.78% to 93.40% upon including DEMIXUP. We include additional ablation studies in Section B from our supplementary material.

Qualitative results. In Figure 2, we present our anomaly maps for the examples illustrated in the BMAD paper (Bao et al., 2024), which allows us to compare our visualizations with many other methods. Our visual interpretation is carried out by an experienced physician, with a background in oncology and radiotherapy. Across all three evaluated benchmarks (BTCV+LiTs, BraTS2021 and RESC), the anomaly maps generated by MTL-MAD show a substantially more accurate visual correspondence with the pathological regions indicated by the ground-truth masks. Unlike the other models, MTL-MAD accurately delineates the tumor region, while significantly reducing false activations within healthy tissue. On liver CT, the proposed method identifies lesions with compact

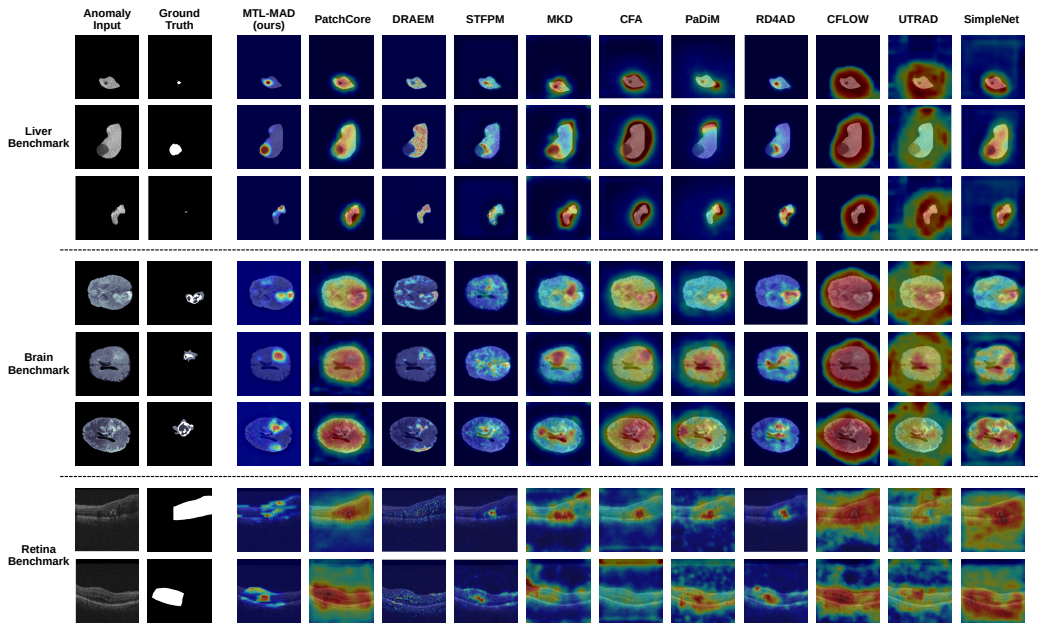


Figure 2: Qualitative visualizations of anomaly maps for the three benchmarks that have ground-truth segmentation maps: BTCV+LiTs, BraTS2021 and RESC. The examples are exactly those selected by Bao et al. (2024) for the BMAD introductory paper. Best viewed in color.

and well-defined localization, without the pathological over-extension observed in methods such as CFA, CFLOW, or UTRAD. On brain MRI, MTL-MAD preserves the normal anatomical structure, while selectively highlighting only the affected regions, demonstrating high sensitivity and specificity. In contrast, the other approaches tend either to underestimate lesion size or to generate diffuse anomaly maps that are difficult to interpret clinically. Similarly, on the retina benchmark, MTL-MAD provides the clearest separation between physiological and pathological areas, maintaining focal localization of anomalies, while minimizing background noise. These findings indicate that MTL-MAD maintains stable performance across different types of medical imaging modalities, including liver CT, brain MRI, and retinal OCT, demonstrating an increased capacity to adapt to multiple clinical and pathological contexts. Moreover, our method exhibits improved accuracy in detecting small lesions that are difficult to identify, as illustrated in both liver CT and brain MRI domains. The maps generated by MTL-MAD more closely reflect the true distribution of disease, which facilitates medical interpretation and future clinical integration. Another important advantage is the significant reduction in over-detection observed in competing methods, where entire anatomical structures are frequently marked as abnormal. By contrast, MTL-MAD maintains an optimal balance between sensitivity and localization precision. This performance demonstrates that the proposed method can extract clinically relevant pathological features without compromising the integrity of normal anatomical regions. Overall, MTL-MAD is not only capable of detecting the presence of disease, but also of providing a more faithful and clinically interpretable localization compared with the other methods.

5 Conclusion

In this study, we introduced a multi-task learner for medical anomaly detection. By meticulously integrating complementary proxy tasks, MTL-MAD outperformed state-of-the-art methods on six datasets from the recently introduced BMAD benchmark. More importantly, we demonstrated exceptional performance levels without any model pre-training, further confirming that the power of MTL-MAD stems solely from the pretext tasks. By routing task-related tokens to specific experts of the MoE-based transformer, we obtained a versatile framework, able to bypass potentially conflicting loss terms via effective task decoupling embedded in the proposed architectural design. Finally, a subjective interpretation from a trained physician clearly indicated the precise anomaly localization abilities of MTL-MAD, which streamlines the subsequent manual diagnosis provided by physicians.

In future work, we aim to integrate our anomaly detection framework into downstream diagnosis, to assist AI-based diagnosis tools in taking better informed decisions.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets and implement methods. We have not used generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting dataset, qualitative and thematic data analysis or results interpretation. The rest of the required disclosure tasks (formulating mathematical claims, obtaining critical ingredients for proving mathematical claims, assisting in the writing of proofs) are not applicable to this work. Additionally, we used generative AI tools to edit software code ans suggest experimental parameters.

We have reviewed all AI-assisted work. To improve the performance of anomaly detection in medical images, our method relies on generating synthetic anomalies to augment the training data, a process detailed in Section 3 and Supplementary B. Because this synthetic data pipeline is built upon the anonymized BMAD dataset, the resulting generated samples are decoupled from any sensitive patient information, thereby avoiding potential privacy risks and ethical concerns. In terms of code development, we utilized generative AI tools to assist in implementing specific methods and classes. However, to guarantee accuracy, four authors verified and tested all AI-generated code. Additionally, while the AI models suggested default experimental parameters during code generation, we overrode most of them with custom values optimized for our approach. Therefore, we take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility reasons, we share our code at the link provided in the abstract. The link contains both the source code and the parameters configurations required to run experiments for each of BMAD subsets. We underline that the link does not provide access to the generated anomalies, but we include generation details in Section 3 and Supplementary B. Moreover, although we did not share the synthetic anomalies because of anonymization concerns, we plan to release them after the manuscript acceptance.

ETHICAL STATEMENT

This work has a high potential of generating a positive impact in healthcare by improving anomaly detection in medical images, which may support early diagnosis and reduce clinical workload. However, deployment in clinical settings may also introduce several risks. Model errors (false positives or false negatives) could lead to missed diagnoses or unnecessary interventions, particularly if clinicians over-rely on automated outputs. Additionally, performance may vary across demographic groups or imaging settings, raising fairness concerns if training data is not representative. Privacy is another key consideration, as medical imaging data is highly sensitive. To mitigate these risks, systems based on our work should pass thorough clinical validation, to ensure diverse training data and secure data handling practices.

Since the use of multi-task learning and Mixture-of-Experts increases computational cost during training, our approach may contribute to a higher energy consumption and a negative environmental impact. For limited training resources, we thus recommend checking if methods based on a single proxy task provide good-enough results, before considering multi-task learning.

REFERENCES

Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In *Proceedings of ACCV*, pp. 622–637, 2019.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo,

Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv preprint arXiv:2107.02314*, 2021.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: Benchmarks for Medical Anomaly Detection. In *Proceedings of CVPRW*, pp. 4042–4053, 2024.

Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023.

Matthew Baugh, Hadrien Reynaud, Sergio Naval Marimont, Sarah Cechnicka, Johanna P. Müller, Giacomo Tarroni, and Bernhard Kainz. Image-Conditioned Diffusion Models for Medical Anomaly Detection. In *Proceedings of UNSURE*, pp. 117–127, 2025.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. *Medical Image Analysis*, 69:101952, 2021.

Finn Behrendt, Debayan Bhattacharya, Julia Kruger, Roland Opfer, and A. Schlaefer. Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI. In *Proceedings of MIDL*, pp. 1019–1032, 2023.

Finn Behrendt, Debayan Bhattacharya, Robin Mieling, Lennart Maack, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Leveraging the Mahalanobis Distance to enhance Unsupervised Brain MRI Anomaly Detection. In *Proceedings of MICCAI*, pp. 394 – 404, 2024.

Farzad Beizaee, Gregory Lodygensky, Christian Desrosiers, and Jose Dolz. MAD-AD: Masked diffusion for unsupervised brain anomaly detection. In *Proceedings of IPMI*, pp. 139–153, 2025.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of ICML*, pp. 41–48, 2009.

Cosmin I. Bercea, Daniel Rueckert, and Julia A. Schnabel. What Do AEs Learn? Challenging Common Assumptions in Unsupervised Anomaly Detection. In *Proceedings of MICCAI 2023*, pp. 304–314, 2023a.

Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Generalizing Unsupervised Anomaly Detection: Towards Unbiased Pathology Screening. In *Proceedings of MIDL*, pp. 39–52, 2023b.

Yuan Bi, Lucie Huang, Ricarda Clarenbach, Reza Ghotbi, Angelos Karlas, Nassir Navab, and Zhongliang Jiang. Synomaly noise and multi-stage diffusion: A novel approach for unsupervised anomaly detection in medical images. *Medical Image Analysis*, 105:103737, 2025.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin

Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró i Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The Liver Tumor Segmentation Benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023.

Behzad Bozorgtabar, Dwarikanath Mahapatra, and Jean-Philippe Thiran. AMAE: Adaptation of Pre-trained Masked Autoencoder for Dual-Distribution Anomaly Detection in Chest X-Rays. In *Proceedings of MICCAI*, pp. 195–205, 2023.

Yu Cai, Hao Chen, and Kwang-Ting Cheng. Rethinking Autoencoders for Medical Anomaly Detection from A Theoretical Perspective. In *Proceedings of MICCAI*, pp. 544–554, 2024.

Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A Comparative Study of Anomaly Detection in Medical Images. *Medical Image Analysis*, 102:103500, 2025.

Liyang Chen, Zhiyuan You, Nian Zhang, Juntong Xi, and Xinyi Le. UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, 147:53–62, 2022.

Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. In *Proceedings of MIDL*, 2018.

Francesco Dalmonte, Emirhan Bayar, Emre Akbas, and Mariana-Iuliana Georgescu. Q-Former Autoencoder: A modern framework for medical anomaly detection. In *Proceedings of WACV*, pp. 7985–7995, 2026.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *Proceedings of ICPR*, pp. 475–489, 2021.

Hanqiu Deng and Xingyu Li. Anomaly Detection via Reverse Distillation From One-Class Embedding. In *Proceedings of CVPR*, pp. 9737–9746, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*, 2021.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A.W.M. van der Laak, the CAMELYON16 Consortium, Meyke Hermsen, Quirine F. Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory Crf van Dijk, Peter Bult, Francisco Beca, Andrew H. Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA: The Journal of the American Medical Association*, 318(22): 2199–2210, 2017.

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of CVPR*, pp. 12742–12752, 2021.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A Training-Free Unified Model for Few-Shot Visual Anomaly Detection. In *Proceedings of CVPR*, pp. 15194–15203, 2025.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. In *Proceedings of WACV*, pp. 1819–1828, 2022.

Shashank Gupta, Subhabrata Mukherjee, Krishan Subudhi, Eduardo Gonzalez, Damien Jose, Ahmed H. Awadallah, and Jianfeng Gao. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arXiv:2204.07689*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of CVPR*, pp. 16000–16009, 2022.

Matthäus Heer, Janis Postels, Xiaoran Chen, Ender Konukoglu, and Shadi Albarqouni. The OOD Blind Spot of Unsupervised Anomaly Detection. In *Proceedings of MIDL*, pp. 286–300, 2021.

Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *International Journal of Computer Vision*, 133(10):7154–7200, 2025.

Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in OCT using deep neural networks. *Medical Image Analysis*, 55:216–227, 2019.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of CVPR*, pp. 11375–11385, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In *Proceedings of CVPR*, pp. 19606–19616, 2023.

Loïc Jézéquel, Ngoc-Son Vu, Jean Beaudet, and Aymeric Histace. Efficient Anomaly Detection Using Self-Supervised Multi-Cue Tasks. *IEEE Transactions on Image Processing*, 32:807–821, 2023.

Loïc Jézéquel, Ngoc-Son Vu, Jean Beaudet, and Aymeric Histace. Fine-grained anomaly detection via multi-task self-supervision. In *Proceedings of AVSS*, pp. 1–8, 2021.

Antanas Kascenas, Pedro Sanchez, Patrick Schrempf, Chaoyang Wang, William Clackett, Shadia S. Mikhael, Jeremy P. Voisey, Keith Goatman, Alexander Weir, Nicolas Pugeault, Sotirios A. Tsaftaris, and Alison Q. O'Neil. The Role of Noise in Denoising Models for Anomaly Detection in Medical Images. *Medical Image Analysis*, 90:102963, 2023.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131, 2018.

Taejune Kim, Yun-Gyoo Lee, Inho Jeong, Soo-Youn Ham, and Simon S. Woo. Patch-wise vector quantization for unsupervised medical anomaly detection. *Pattern Recognition Letters*, 184: 205–211, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*, 2025.

Bennett Landman, Zhoubing Xu, J. Igelsias, Martin Styner, T. Langerak, and Arno Klein. Multi-Atlas Labeling Beyond the Cranial Vault–Workshop and Challenge. In *Proceedings of BCV*, pp. 12, 2015.

Sungwook Lee, Seunghyun Lee, and Byung Song. CFA: Coupled-Hypersphere-Based Feature Adaptation for Target-Oriented Anomaly Localization. *IEEE Access*, 10:78446–78454, 2022.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In *Proceedings of CVPR*, pp. 9664–9674, 2021.

Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M^3ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *Proceedings of NeurIPS*, pp. 28441–28457, 2022.

Bo Liu, Xingchao Liu, Xiaojie Jin, et al. Conflict-averse gradient descent for multi-task learning. In *Proceedings of NeurIPS*, pp. 18878–18890, 2021.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Proceedings of NeurIPS*, pp. 57226–57243, 2023a.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A Simple Network for Image Anomaly Detection and Localization. In *Proceedings of CVPR*, pp. 20402–20411, 2023b.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of KDD*, pp. 1930–1939, 2018.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of CVPR*, pp. 4744–4754, 2025.

Kai Mao, Ping Wei, Yiyang Lian, Yangyang Wang, and Nanning Zheng. Beyond single-modal boundary: Cross-modal anomaly detection through visual prototype and harmonization. In *Proceedings of CVPR*, pp. 9964–9973, 2025.

Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Proceedings of ECCV*, pp. 69–84, 2016.

Walter H. L. Pinaya, Mark S. Graham, Robert Gray, Pedro F. da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models. In *Proceedings of MICCAI*, pp. 705–714, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of CVPR*, pp. 10674–10685, 2022.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of CVPR*, pp. 14318–14328, 2022.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully Convolutional Cross-Scale-Flows for Image-based Defect Detection. In *Proceedings of WACV*, pp. 1829–1838, 2022.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep One-Class Classification. In *Proceedings of ICML*, pp. 4393–4402, 2018.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution Knowledge Distillation for Anomaly Detection. In *Proceedings of CVPR*, pp. 14897–14907, 2021.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In *Proceedings of IPMI*, pp. 146–157, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks. *Medical Image Analysis*, 54:30–44, 2019.

Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Proceedings of NeurIPS*, pp. 525–536, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proceedings of ICLR*, 2017.

Mahshid Shiri, Cigdem Beyan, and Vittorio Murino. MadCLIP: Few-shot Medical Anomaly Detection with CLIP. In *Proceedings of MICCAI*, pp. 416–426, 2025.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal Self-Supervised Learning for Medical Image Analysis. In *Proceedings of IPMI*, pp. 661–673, 2021.

Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting Outliers with Foreign Patch Interpolation. *Machine Learning for Biomedical Imaging*, 1:1–27, 2022.

Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W. Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical Image Analysis*, 90:102930, 2023.

Sina Walluscheck, Annika Gerken, Ivana Galinovic, Kersten Villringer, Jochen B. Fiebach, Jan Klein, and Stefan Heldmann. Generative adversarial network-based reconstruction of healthy anatomy for anomaly detection in brain CT scans. *Journal of Medical Imaging*, 11(4):044508, 2024.

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Proceedings of ECCV*, pp. 494–511, 2022.

Meng Wang, Tian Lin, Lianyu Wang, et al. Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications*, 14:6757, 2023.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of CVPR*, pp. 3462–3471, 2017.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of CVPR*, pp. 9653–9663, 2022.

Shinji Yamada and Kazuhiro Hotta. Reconstruction Student with Attention for Student-Teacher Pyramid Matching. *arXiv preprint arXiv:2111.15376*, 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of ICCV*, pp. 6023–6032, 2019.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DRAEM - A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. In *Proceedings of ICCV*, pp. 8330–8339, October 2021.

Chenkai Zhang, Yueming Wang, and Wenming Tan. Mthm: Self-supervised multitask anomaly detection with hard example mining. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of ICLR*, 2018.

Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. MediCLIP: Adapting CLIP for Few-shot Medical Image Anomaly Detection. In *Proceedings of MICCAI*, pp. 458–468, 2024.

Kang Zhou, Jing Li, Weixin Luo, Zhengxin Li, Jianlong Yang, Huazhu Fu, Jun Cheng, Jiang Liu, and Shenghua Gao. Proxy-Bridged Image Reconstruction Network for Anomaly Detection in Medical Images. *IEEE Transactions on Medical Imaging*, 41(3):582–594, 2022.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. In *Proceedings of ICLR*, pp. 49705–49737, 2024.

Yue Zhou, Yuan Bi, Wenjuan Tong, Wei Wang, Nassir Navab, and Zhongliang Jiang. UltraAD: Fine-Grained Ultrasound Anomaly Classification via Few-Shot CLIP Adaptation. In *Proceedings of MICCAI*, pp. 625–635, 2025.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of CVPR*, pp. 17826–17836, 2024.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised Anomaly Localization Using Variational Auto-Encoders. In *Proceedings of MICCAI*, pp. 289–297, 2019.

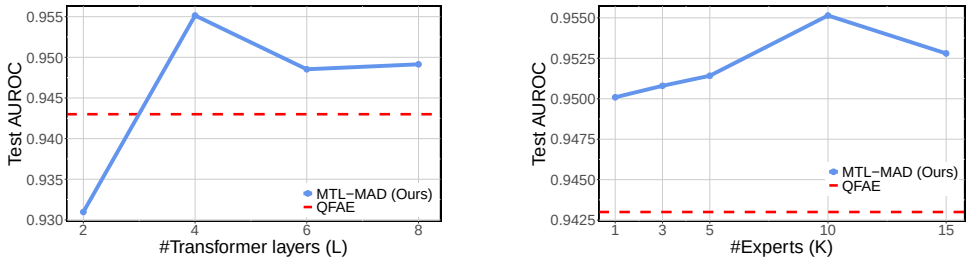
A LIMITATIONS

While the proposed multi-task self-supervised framework demonstrates strong performance in medical anomaly detection, several limitations should be noted. First, the approach relies on proxy self-supervised and pseudo-labeling tasks. Taken alone, each task is only indirectly related to pathological abnormalities. As a result, there is an inherent risk of encountering mismatched anomalies and proxy tasks, where learned representations may not fully capture clinically relevant deviations from normal anatomy. This can potentially lead to reduced sensitivity for certain subtle or rare anomaly types, as well as false responses to irrelevant image variations. We mitigate such potential mismatches by aggregating multiple proxy tasks, so that mismatches caused by certain proxy tasks can still be recovered by other (well-aligned) proxy tasks. Hence, our design is based on a reasonable assumption, namely that proxy tasks are less likely to fail together, at exactly the same time.

Second, multi-task learning introduces the possibility of task interference. Competing objectives may lead to gradient conflicts or uneven optimization across tasks, where some tasks dominate the shared representation, while others contribute marginally. This can result in suboptimal feature learning

Table 3: Chosen hyperparameter configurations for MTL-MAD.

Dataset	Image resolution	L	d	K
BraTS2021	256×256	4	128	10
RESC	256×256	4	128	5
RSNA	256×256	4	128	5
OCT2017	256×256	4	128	5
BTCV+LiTs	256×256	4	512	10
CAMELYON16	256×256	12	768	10



(a) Ablation for the depth L of the shared encoder. (b) Ablation for the number of experts K inside MoE.

Figure 3: Ablation study on BraTS2021 that showcases the importance of two hyperparameters, namely the number of transformer layers (L) and the number of experts (K). The performance is measured in terms of image-level AUROC. Best viewed in color.

compared to more carefully isolated single-task models. Our attempt to mitigate this limitation is (i) to employ Mixture-of-Experts (MoE) and decouple the tasks by routing task-specific tokens to different experts, as well as (ii) Fast Adaptive Multitask Optimization (FAMO) (Liu et al., 2023a) to automatically balance objective during training.

Third, the proposed framework increases computational complexity during training. The inclusion of multiple tasks naturally leads to increased training cost. We try to counter this effect by using a relatively lightweight shared encoder, with at most $L = 12$ transformer blocks. As reference, both ViT-Small and ViT-Base contain $L = 12$ transformer blocks (Dosovitskiy et al., 2021).

Finally, the performance of MTL-MAD may vary depending on the number of experts and depth of the shared encoder. To obtain optimal results in other medical imaging domains, hyperparameters should be established on a small BMAD-style validation set that can easily be collected in practice.

B Hyperparameter Configuration

In Table 3, we report the chosen hyperparameter configuration for each of the six datasets for the results reported in Table 1. We employ larger backbones for the more complex BTCV+LiTs ($L = 4$, $d = 512$) and CAMELYON16 ($L = 12$, $d = 768$) datasets, while keeping the backbone fairly lightweight ($L = 4$, $d = 128$) for the other datasets.

C More Ablations Studies

In Figure 3, we present the impact of two important hyperparameters, namely the number of transformer layers (L) and the number of experts (K) within each layer, respectively. These ablation studies are conducted on BraTS2021. In both cases, most of the considered values yield better performance than the previous state-of-the-art method (Dalmonte et al., 2026). Moreover, in Figure 3b, we study scenarios where $K < 5$, and in all these cases, the experts are shared among tasks. The results indicate that there is some benefit of having distinct specialized experts per task. Since MTL-MAD comprises five tasks, we always utilize $K \geq 5$ to reach optimal performance in all our experiments.

We extend the results analysis in Table 4 with an anomaly-size stratification on BraTS2021, comparing MTL-MAD against QFAE (Dalmonte et al., 2026) (the second-best method) across small, medium, and large lesions defined by the available ground-truth masks. Specifically, we partition the anomalous

Table 4: Image AUROC of MTL-MAD and QFAE on large, medium and small anomalies from BraTS2021.

Method	AUROC on Large	AUROC on Medium	AUROC on Small
MTL-MAD	98.92	98.08	87.94
QFAE	98.60	97.84	85.95

Table 5: Results with various hyperparameter combinations for MTL-MAD on BTCV+LiTs.

L	d	K	AUROC
4	128	5	81.0
6	128	5	77.9
8	128	5	82.6
10	128	5	79.8
12	128	5	83.1
4	256	5	83.5
4	512	5	84.2
4	756	5	82.9
4	128	10	79.5
4	256	10	83.0
4	512	10	82.7
4	128	15	82.4

Table 6: Results with various top-k ratios of patches selected for anomaly score computation from the patch DEMIXUP task on BraTS2021.

Top-k ratio (%)	AUROC
5	95.44
10	95.51
15	94.64
20	94.16
25	93.70

samples into three equally-sized subsets according to the ground-truth anomalous area and report the Image AUROC for each. The results confirm that MTL-MAD exhibits clear gains for smaller anomalies, that are typically harder to detect, while performing on par with QFAE on medium and large anomalies. This provides a quantitative counterpart to the physician's observation regarding small/subtle abnormalities.

We include additional results with different hyperparameter configurations on BTCV+LiTS in Table 5. While certain configurations also lead to performance degradation, all variants remain well above all competing methods from Table 1. Importantly, the overall variation across configurations remains limited, indicating that our method remains robust to the choice of these architectural hyperparameters.

In Table 6, we present an ablation study on the BraTS2021 dataset, analyzing the effect of the top-k ratio of patches selected to compute the anomaly score in the patch DEMIXUP task. As expected, performance degrades at higher ratios because an increasing number of normal patches are included in the calculation. Conversely, selecting an insufficient number of patches also yields worse results, as the anomaly score estimation becomes overly sensitive to noise. Across our experiments, the optimal value is $k = 10\%$.

We supplement Figure 3 from the paper with Table 7, where we have already provided ablations for the number of experts (K) and the depth of the model (L). As before, 12 of 13 configurations outperform QFAE (Dalmonte et al., 2026), and 4 of 13 configurations surpass our result reported in Table 1, suggesting that a more extensive hyperparameter search can further improve the performance of our pipeline. This further confirms that we fairly tuned our model hyperparameters.

We conduct an additional ablation for MoE and FAMO on BraTS2021 in Table 8. The reported numbers confirm that both MoE and FAMO lead to an overall performance gain of 2.5, so they are contributing factors for achieving SOTA performance on BraTS.

Table 7: Ablation study on BRaTS2021 for the number of experts (K), the hidden dimension (d) and the depth of the model (L).

L	d	K	AUROC
4	128	5	95.14
6	128	5	95.37
2	128	10	95.14
4	128	10	95.51
6	128	10	94.85
8	128	10	94.91
4	256	10	95.53
6	256	10	96.05
8	256	10	93.61
4	128	15	95.28
6	128	15	95.56
6	256	15	95.85
8	256	15	96.39

Table 8: Ablation study for MoE and FAMO on BraTS 2021.

Dataset	Base MTL	+MoE	+FAMO
BraTS2021	93.0	93.9	95.5

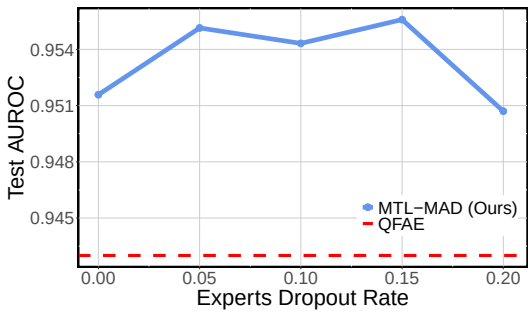


Figure 4: Impact of varying expert dropout rates on performance. The study is conducted on BraTS2021.

In Figure 4, we present the impact of varying the expert dropout rate on the final performance. Notably, dropout rates between 5% and 15% yield a performance improvement over the no-dropout baseline.

D Discussion on improvement margins

In Table 1, the $+10.4\%$ AUROC improvement on BTCV+LiTs is particularly substantial. In addition, we would like to clarify that the smaller margins on the other datasets are not measured against one single consistently strong competing method. Instead, the second-best method changes across datasets: P-VQ/QFAE on BraTS2021, QFAE on RESC and RSNA, PatchCore on OCT2017, SimpleNet on BTCV+LiTs, and MKD on CAMELYON16. Nevertheless, MTL-MAD ranks first on all six datasets, with gains of $+1.2\%$, $+1.0\%$, $+2.2\%$, $+0.3\%$, $+10.4\%$, and $+0.6\%$ AUROC, respectively. We also note that OCT2017 is nearly saturated: the previous best is already 98.6% AUROC, leaving little room for large absolute gains.

When comparing against individual methods evaluated on the complete benchmark, the advantage is considerably larger. For example QFAE (Dalmonte et al., 2026) is the second best method on three datasets, but our MTL-MAD surpasses it significantly on BTCV+LiTs (65.5% for QFAE vs. 82.7% for MTL-MAD). This supports our central motivation: existing methods perform strongly in particular domains but inconsistently across modalities, whereas MTL-MAD remains strong throughout the whole benchmark.

Table 9: List of prompts employed for synthetic anomaly generation. For each medical domain, we randomly sample one prompt from the list to generate diverse pseudo-anomalies.

Medical domain	Prompts
Brain MRI	> Brain lesion > White area, high intensity white blob consistent with the surrounding area, white matter lesion > Hyperintense spot > Brain tumor > Bright white spot
Liver CT	> Generate a solid, hypervascular nodule in the masked region. > Show bright arterial phase hyperenhancement and portal venous phase washout. > Inpaint a hypodense lesion that shows peripheral, nodular enhancement in the arterial phase and gradually fills in centripetally (from outside-in) on the portal venous phase. > Generate a hypodense, ring-enhancing lesion in the masked area, most visible in the portal venous phase. > A simple hepatic cyst. The lesion must be sharply defined, non-enhancing, and hypodense (near-water density, 0-20 HU). > Inpaint the masked liver parenchyma to be hypodense (darker) than the spleen, consistent with hepatic steatosis. > Add an anomaly to the liver image. > Inpaint black blobs.
Retinal OCT	> Add a tightly packed cluster of bright, waxy dots with very sharp, distinct edges. > Draw several irregular, dark crimson splatters and deep dots that look like wet ink stains on the surface. > Create a few soft, fluffy, pale white patches with blurred and faded edges, resembling thin clouds. > Generate a chaotic, tangled web of extremely thin, bright, thread-like lines that branch out irregularly. > Create a large, sharply outlined area where the background looks faded, revealing pale and exposed thick white vessels underneath.
Chest X-ray	> Create multiple small, calcified nodules consistent with old granulomatous disease. > Airspace consolidation with air bronchograms, consistent with pneumonia, in the masked region. > Opacities, suggestive of atypical pneumonia. > Pneumothorax in the masked apex, showing a visible visceral pleural line and an absence of lung markings. > Create a pleural effusion in the masked area, characterized by blunting of the costophrenic angle and a meniscus sign. > Add an anomaly to the chest image.
Histopathology	> Cells exhibit nuclei of vastly different sizes and irregular, angular shapes. Some nuclei are hyperchromatic (deep purple) with thickened nuclear membranes. > Large, distinct, reddish-purple dots are visible inside the enlarged, pale nuclei. > Histological image showing multiple atypical mitotic figures. Visible tripolar and quadripolar spindles (star-shaped or multi-pronged dark chromatin clumps) indicating rapid, disorganized cell division. > Cells are haphazardly arranged and 'piling up' rather than forming organized layers. > Cells are highly primitive, bizarre, and unrecognizable compared to normal tissue.

E TEXT PROMPTS AND GENERATED IMAGES

In Table 9, we show the text prompts used to generate synthetic anomalies via frozen VLMs. To generate a pseudo-anomaly for a certain medical domain, we uniformly sample one prompt from the corresponding list of prompts. In Figure 5, we showcase several examples of pseudo-anomalies obtained via generative VLMs or data augmentations, respectively. While pseudo-anomalies are sometimes too mild (hard to spot) or unrealistic (e.g. colorful spots in grayscale images), the corresponding pseudo-anomaly classification tasks remain a critical addition to the success of MTL-MAD.

F COMPUTE ENVIRONMENT

The experiments have been conducted on four local nodes with different configurations. The best-equipped node has a single NVIDIA A100 80GB PCIe GPU, 64 GB of memory and a CPU featuring 56 physical cores. The least capable configuration is equipped with a single NVIDIA RTX 3090 24 GB, 64 GB of system memory and 10 physical cores.

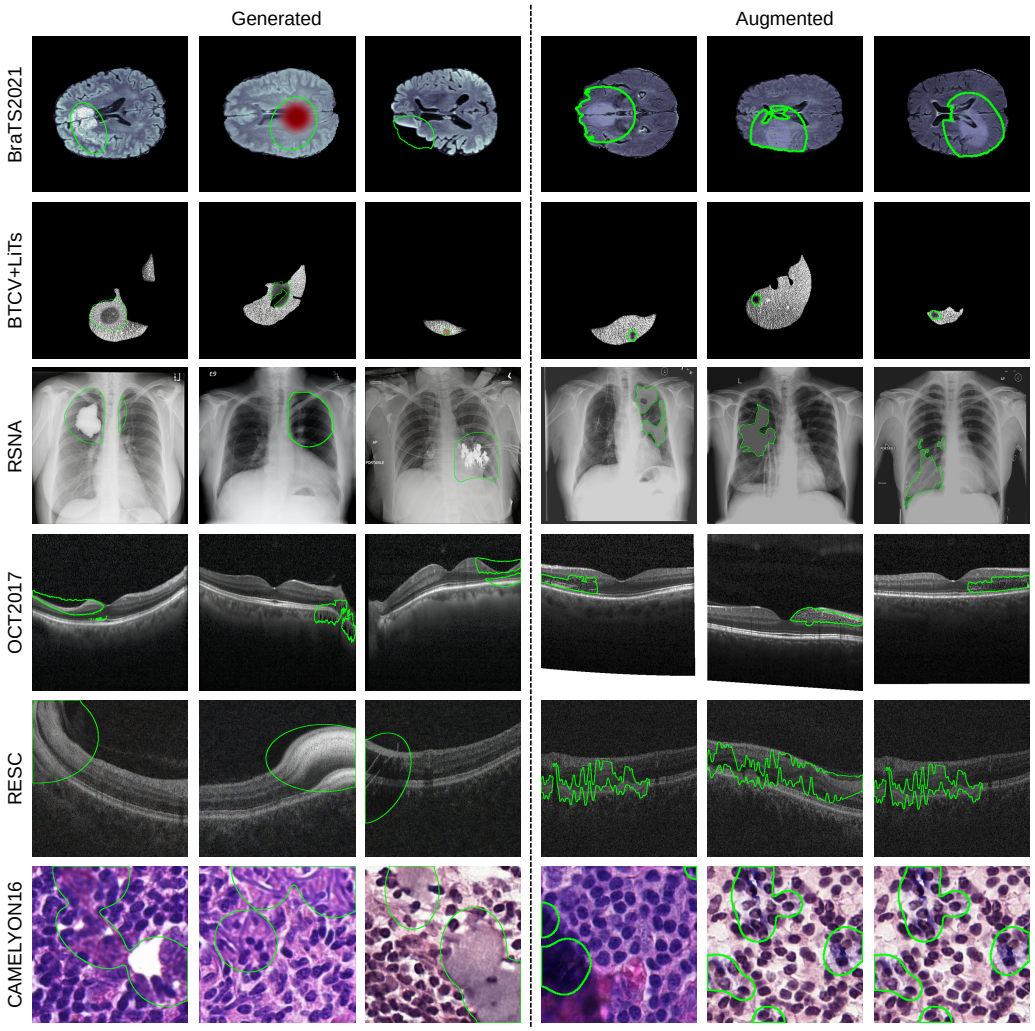


Figure 5: Examples of pseudo-anomalies generated via VLMs (left-hand side) or data augmentation (right-hand side), which are used by our pipeline in the pseudo-anomaly classification tasks (Task 4 and Task 5). The full sets of generated and augmented anomalies are released along with the code. Best viewed in color.