

Q-Former Autoencoder: A Modern Framework for Medical Anomaly Detection

Francesco Dalmonte^{1,*}

Emirhan Bayar^{2,*}

Emre Akbas^{2,3}

Mariana-Iuliana Georgescu³

¹University of Bologna, Italy

²Middle East Technical University, Ankara, Türkiye

³Helmholtz Munich, Germany

Abstract

Anomaly detection in medical images is an important yet challenging task due to the diversity of possible anomalies and the practical impossibility of collecting comprehensively annotated data sets. In this work, we tackle unsupervised medical anomaly detection proposing a modernized autoencoder-based framework, **the Q-Former Autoencoder**, that leverages state-of-the-art pretrained vision foundation models, such as DINO, DINOv2 and Masked Autoencoder. Instead of training encoders from scratch, we directly utilize frozen vision foundation models as feature extractors, enabling rich, multi-stage, high-level representations without domain-specific fine-tuning. We propose the usage of the Q-Former architecture as the bottleneck, which enables the control of the length of the reconstruction sequence, while efficiently aggregating multi-scale features. Additionally, we incorporate a perceptual loss computed using features from a pretrained Masked Autoencoder, guiding the reconstruction towards semantically meaningful structures. Our framework is evaluated on four diverse medical anomaly detection benchmarks, achieving state-of-the-art results on BraTS2021, RESC, and RSNA. Our results highlight the potential of vision foundation model encoders, pretrained on natural images, to generalize effectively to medical image analysis tasks without further fine-tuning. We release the code and models at <https://github.com/emirhanbayar/QFAE>.

1. Introduction

Automated anomaly detection in medical imaging is a crucial problem, as it directly impacts diagnostic accuracy, workflow efficiency, and patient outcomes. However, manual inspection of large-volume medical scans, such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT), is inherently time-consuming and susceptible to human error, highlighting the need for reliable auto-

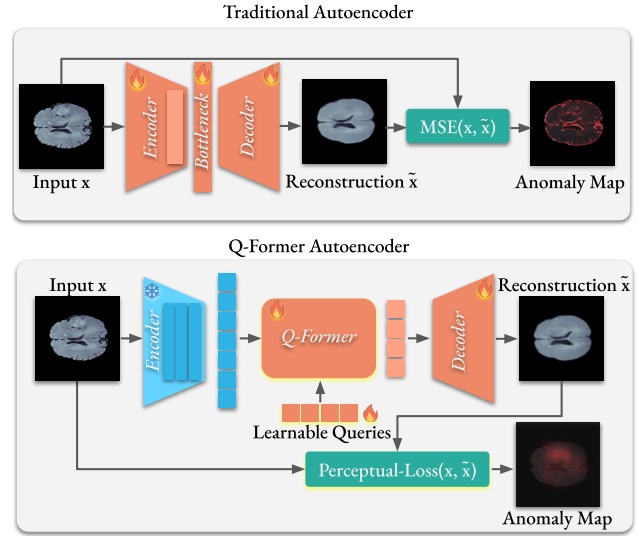


Figure 1. We illustrate the traditional autoencoder for anomaly detection (top) versus our Q-Former Autoencoder enhanced with Q-Former and perceptual loss (bottom). The traditional autoencoder typically uses a trainable encoder-decoder pair and relies on Mean Squared Error (MSE) for optimization and anomaly detection. Our framework includes the following improvements (highlighted in yellow): (i) a **frozen encoder** (employing powerful pretrained vision foundation models, such as DINO, DINOv2 and OpenCLIP), (ii) a **Q-Former** acting as a dynamic, learnable bottleneck for efficient representation, and (iii) the use of a **perceptual loss** function based on Masked Autoencoder. Our framework is able to produce meaningful anomaly detection *precisely highlighting the anomalous regions* (bottom-right, **in red**).

ated systems that can assist physicians in flagging potential anomalies. However, automated medical anomaly detection presents significant challenges too. Anomalies manifest in highly diverse forms and appearances, making it infeasible to collect representative samples of all possible pathological variations. As a result, unsupervised anomaly detection approaches, which train models exclusively on normal data to identify deviations as anomalies, are appropriate for this domain.

*Equal contribution

Early work in unsupervised anomaly detection has predominantly relied on convolutional autoencoders trained to reconstruct normal images. These conventional autoencoders suffered from limited representational power, restricting their effectiveness in anomaly detection. Recent advances in vision foundation models, such as DINO [10], DINOv2 [43], and Masked Autoencoders (Masked AE) [22], have demonstrated remarkable representation transferability to diverse tasks. Despite their potential, these models have been largely overlooked in the detection of medical image anomalies. One of the few exceptions is MVFA-AD [24] which employed the CLIP model [44] to perform zero-shot and few-shot medical anomaly detection. Unfortunately, these methods often suffer a performance gap compared to task-specific methods.

To bridge this gap, we propose a novel framework, **Q-Former Autoencoder**, that modernizes the autoencoder approach for unsupervised medical anomaly detection by integrating vision foundation models and an attention-based bottleneck mechanism based on Q-Former, as illustrated in Figure 1. First, we leverage pretrained vision foundation models, namely DINO [10], DINOv2 [43] and OpenCLIP [52], as frozen encoders, extracting robust and semantically rich features without requiring domain-specific retraining or fine-tuning. Second, we introduce a Q-Former model as a *flexible bottleneck*, which aggregates multi-scale features and outputs a fixed-length latent representation. This design provides *explicit control over the reconstruction granularity* while simultaneously improving the model’s capacity to accurately represent normal structures. Third, we employ a perceptual loss computed using features extracted by a pretrained Masked Autoencoder, which encourages reconstructions that preserve high-level semantics rather than low-level pixel details.

Our *modernized autoencoder* significantly outperforms its standard counterpart in accurately detecting and localizing anomalies, as illustrated in Figure 1. To evaluate our framework, we perform extensive experiments on four data sets from the BMAD [4] benchmark: BraTS2021 [2, 3, 41], RESC [23], RSNA [56], and LiverCT [6, 31].

Our framework achieves state-of-the-art scores on all data sets reaching an AUROC of 94.3% on BraTS2021 and 83.8% on RSNA, showcasing its effectiveness across diverse image modalities, including MRI, OCT and X-rays.

In summary, our contributions are threefold:

- We propose a *modernized and enhanced autoencoder* approach that integrates frozen vision foundation models, a Q-Former bottleneck, and a perceptual loss for unsupervised anomaly detection.
- Our proposed framework achieves strong performance, reaching state-of-the-art AUROC scores on three medical anomaly detection benchmarks (namely BraTS2021, RESC, RSNA) without requiring domain-specific en-

coder finetuning.

- We provide detailed ablation experiments showing how vision foundation models, which are primarily trained on natural images, are able to generalize effectively to the medical image domain when combined with proper architectural adaptations.

2. Related Work

2.1. Taxonomy of Anomaly Detection Approaches

Learning Strategy. Image-based Anomaly Detection (AD) methods are commonly categorized into supervised, unsupervised, and zero-shot approaches. Supervised methods, such as those based on few-shot learning or synthetic anomaly generation, require some access to annotated abnormal samples. Zero-shot methods, on the other hand, aim to identify out-of-distribution samples without any access to domain data, often relying on pretrained models. Although promising in natural visual domains [16], they remain limited in specialized fields such as industrial inspection or medical imaging, where domain-specific knowledge is critical. In such cases, unsupervised AD remains the most relevant setting. These methods train exclusively on normal samples and aim to detect deviations during inference. Although recent work has investigated multiclass AD [59], these approaches typically perform poorly compared to specialized algorithms, limiting their applicability in sensitive or safety-critical contexts such as medical analysis.

Feature-embedding or Reconstruction-based. A classical taxonomy of AD methods [36] divides them into two macro-categories: feature embedding and reconstruction-based approaches. Feature embedding methods rely on distances or density estimation in learned feature spaces. In contrast, reconstruction-based approaches, such as those based on autoencoders, learn to exclusively reconstruct normal data, assuming that anomalies cannot be effectively reconstructed from the model. These approaches have demonstrated strong performance, even when implemented as simple baselines with minimal architectural complexity [9], while inherently supporting explainability and anomaly localization—particularly valuable in medical AD. In this work, we propose a framework based on reconstruction of the input.

2.2. Autoencoder Architectures for AD

Autoencoders learn a compressed latent representation of training data and attempt to re-project it to the input space. It is well-established that the compression of the latent representation is central to the anomaly detection (AD) capabilities of autoencoders [8, 49]. The main challenge that such approaches face is to find a balance between a good reconstruction of normal images, while preventing the model to generalize to the anomalous samples.

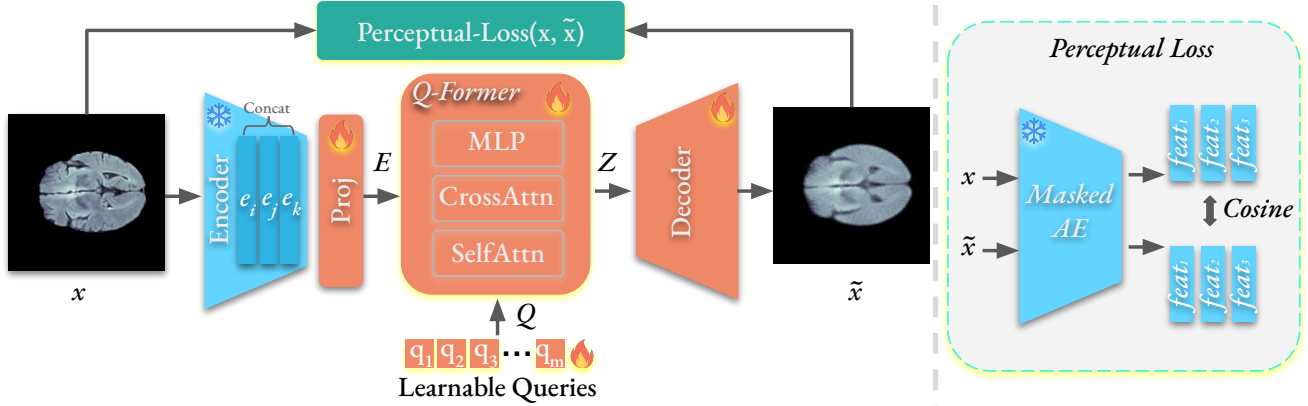


Figure 2. The training of our **Q-Former Autoencoder** for medical anomaly detection. Our framework uses a *pretrained foundation model*, such as DINO [10], DINOv2 [43] or OpenCLIP [52], to extract *multi-scale features* (E). These features, along with learnable query tokens (Q), are processed by Q-Former acting as a *dynamic bottleneck*. The output z goes into the decoder to reconstruct \tilde{x} . The *Perceptual Loss* based on multi-scale features extracted from Masked AE [22] guides the training for semantic reconstruction.

A critical aspect is the choice of the reconstruction metric [39]: in addition to L2 loss, structural similarity index metric (SSIM) has been explored [5, 40], as well as perceptual losses [27, 53]. Some of the most effective approaches recently proposed measure the distance in the feature space rather than the image space, showing robust results [20, 21, 40]. This approach is often used in combination with knowledge-distillation techniques, to further amplify the distance of anomalous samples [14, 54]. Other relevant related methods involved the use of variational autoencoders [38], masked autoencoders [17, 57] and normalizing flow mechanisms [62]. Similarly to the aforementioned works [27, 53], we employ the perceptual loss to train the autoencoder. However, different from the previous work, we utilize the Masked Autoencoder to guide the optimization of our model.

2.3. Vision Foundation Models

Recent advances in large-scale model pre-training have enabled the development of highly versatile foundation models for vision tasks, predominantly leveraging Vision Transformer (ViT) [15] architectures. Notable examples include CLIP [44], which employs a contrastive learning framework; DINOv2 [43] and Masked Autoencoder [22] trained with various self-supervised schemes; and supervised models like SAM [30]. These models, trained on large-scale data sets, learn rich representations that capture semantic and structural image information, enabling strong generalization across diverse downstream tasks.

The use of high-capacity vision foundation models for unsupervised anomaly detection (AD) remains underexplored. Zhang et al. [61] established a multiclass AD baseline using frozen ViTs. More recent approaches leverage vision-language models: Jeong et al. [26] employed com-

positional prompt ensembles and a sliding window for segmentation, and integrates memory banks to enable few-shot learning. Zhou et al. [63] used object-agnostic templates and prompt tuning. Huang et al. [25] addressed domain shift through a dedicated adaptation module. Gu et al. [18] repurposed multimodal conversational models for AD, achieving strong results on industrial benchmarks.

While these models show promising performances, especially in zero- and few-shot regimes—a scalable, unified approach to fully leverage foundation models remains lacking, often resulting in performance gaps compared to task-specific methods. Therefore, in this work, we propose to fully leverage the foundation models, proposing an enhanced autoencoder framework equipped with Q-Former and Masked AE-based perceptual loss.

3. Q-Former Autoencoder

3.1. Overview

Autoencoder (AE) models are commonly employed in anomaly detection due to their ability to learn compact representations of normal data. An autoencoder consists of an encoder, a latent space (bottleneck), and a decoder. The encoder compresses the input x into a latent representation z , i.e., $z = \text{Encoder}(x)$. This latent representation is expected to capture the most informative aspects of the data. The decoder reconstructs the input from z , mapping it back to the input space, i.e., $\tilde{x} = \text{Decoder}(z)$. AEs are typically trained to minimize the reconstruction error between the input x and its reconstruction \tilde{x} .

Training autoencoders only on normal data enables anomaly detection, as the model should reconstruct normal data accurately while failing to reconstruct anomalies.

3.2. Evolving Autoencoders

We present the training of our Q-Former Autoencoder (QFAE), highlighting the integration of Q-Former and Perceptual Loss in Figure 2.

Encoders. Prior to the introduction of Vision Transformers (ViTs) [15], Convolutional Neural Networks (CNNs) were the standard choice for encoders. Modern foundation models [10, 22, 43, 44], however, mainly employ ViT architectures. ViTs begin by splitting the input image into non-overlapping patches and extracting patch representations using a shallow neural network. Subsequently, self-attention operation is applied on these patch representations, together with other normalization and feed-forward layers. These operations are repeated multiple times. As a result, the ViT encoder outputs a fixed-length sequence of patch embeddings.

In this work, we employ pretrained encoders from foundation models, such as DINO [10], DINOv2 [43], CLIP [44], Masked AE [22].

Latent Space. Inspired by BLIP-2 [35] and BRAVE [28], we employ the Q-Former architecture as the bottleneck. Q-Former is well-suited because it processes variable-length contextual input to produce fixed-length latent codes, enabling the combination of tokens from different levels and even different architectures. The input to Q-Former is a set of learnable tokens, where the number of tokens controls the number of reconstructed patches. This enables the reconstruction of the output at multiple granularities (different patch sizes). Q-Former interacts with the encoder features through a cross-attention layer. The encoder output serves as keys and values, being cross-attended by the Q-Former queries, as illustrated in Figure 2. This design enables Q-Former to aggregate information from the latent features of the encoder efficiently, as Q-Former eliminates quadratic self-attention.

For a given input x , we obtain its embedding as $[e_i, e_j, \dots, e_k]$ where $[\cdot]$ is the concatenation operation and e_i, e_j, e_k are features from different layers of a pretrained ViT foundation model. These features are then adapted to the current task via a projection layer: $E = \text{Proj}([e_i, e_j, \dots, e_k])$, shown in Figure 2. We define the learnable queries $Q = [q_1, q_2, \dots, q_m]$, where m is the desired length of the output sequence. A block of Q-Former is defined as:

$$\begin{aligned} Q &= \text{SelfAttn}(Q) \\ Q &= \text{CrossAttn}(Q, E) \\ Z &= \text{MLP}(Q). \end{aligned} \quad (1)$$

Based on our validation experiments, we employ only a single Q-Former block in our framework.

Decoder. The decoder receives as input the latent representation Z produced by Q-Former and reconstructs the original input image x . As noted earlier, the length of the reconstructed sequence is controlled by the number of learnable queries in Q-Former. The decoder architecture is a lightweight Transformer with only a few layers. Therefore, the reconstructed sequence of tokens $\tilde{x}_{tok} = \text{Decoder}(Z)$. In the last step, we reconstruct the input image by rearranging the tokens $\tilde{x} = \text{unpatchify}(\tilde{x}_{tok})$.

Perceptual Loss. Autoencoders are typically trained by minimizing the mean squared or mean absolute error between the input x and its reconstruction \tilde{x} . In practice, perceptual loss has been proposed to improve reconstruction quality [27, 53]. We compute perceptual loss using features extracted from different layers of a pretrained Masked Autoencoder (Masked AE) [22]. The perceptual loss minimizes the cosine distance between features from the original and reconstructed images:

$$\mathcal{L}_{\text{Perceptual}} = \frac{1}{|I|} \sum_{i \in I} \left(1 - \frac{\text{feat}_i \cdot \tilde{\text{feat}}_i}{\|\text{feat}_i\|_2 \|\tilde{\text{feat}}_i\|_2} \right) \quad (2)$$

where I is the set of selected layer indices from which features feat are obtained using the Masked AE [22].

Anomaly Score Computation. We compute the anomaly score similarly to the perceptual loss, by comparing features extracted from multiple layers of the pretrained Masked AE, derived from the original and reconstructed images.

For each layer i in a set of selected layers I , we extract the feature maps $\text{feat}_i \in \mathbb{R}^{h_i \times w_i \times c}$ and $\tilde{\text{feat}}_i \in \mathbb{R}^{h_i \times w_i \times c}$, corresponding to the original input and its reconstruction, respectively. We then compute a layer-wise anomaly map, $A_{\text{map},i}$, by calculating the cosine distance at every spatial location (j, k) between the corresponding feature vectors (patch embeddings).

$$A_{\text{map},i}(j, k) = 1 - \frac{\text{feat}_{i,(j,k)} \cdot \tilde{\text{feat}}_{i,(j,k)}}{\|\text{feat}_{i,(j,k)}\|_2 \|\tilde{\text{feat}}_{i,(j,k)}\|_2}. \quad (3)$$

The final anomaly score for an image, a single scalar value, is calculated from these layer-wise maps by taking the maximum value from each map and averaging these maximums:

$$A_{\text{score}} = \frac{1}{|I|} \sum_{i \in I} \max(A_{\text{map},i}). \quad (4)$$

For visualization purposes, a consolidated anomaly map is generated by pixel-wise averaging all the layer-wise anomaly maps.

$$A_{\text{map, final}} = \frac{1}{|I|} \sum_{i \in I} A_{\text{map}, i}. \quad (5)$$

This final anomaly map enables the localization of anomalous regions within the image, as illustrated in Figure 4.

4. Experiments

4.1. Data sets

We report results on three data sets: BraTS2021 [2, 3, 41], RESC [23] and RSNA [56], as detailed below. Additional results on the LiverCT [6, 31] data set are provided in the supplementary material.

BraTS2021. The BraTS2021 [2, 3, 41] data set, part of the the BMAD [4] benchmark, contains brain MRI images with pixel-level annotations of various anomalies. BraTS2021 has a total of 11,298 images, split into 7,500 training, 83 validation and 3,715 test samples. Each slice has a resolution of 240×240 pixels.

RESC. RESC [23], also part of BMAD [4], contains retinal OCT images. The data includes 6,217 images in total, with 1,805 used for testing. All images are high-resolution with the size of 512×1024 pixels.

RSNA. The RSNA data set [56], included in the BMAD [4] benchmark, consists of chest X-ray images with image-level anomaly annotations. It contains 26,684 images of resolution 1024×1024 , split into 8,000 training, 1,490 validation and 17,194 test samples.

4.2. Implementation Details

We employed different pretrained vision foundation models as encoders, including DINO [10], DINOv2 [43], OpenCLIP [52] and Masked Autoencoder [22]. As previously stated, the encoder remains frozen during the training of the framework. Our decoder is a Transformer architecture, consistent with the Masked AE setting, with 6 layers, 12 heads, and a hidden dimension of 768. Features are extracted from layers 20 and 22 of the ViT-L encoder and layers 8 and 10 of the ViT-B architecture. The architecture of Q-Former consists of only one Transformer layer. The number of learnable tokens in Q-Former is determined by the reconstruction patch size. With a patch size of 8×8 pixels and an input resolution of 224×224 , the number of learnable tokens is 784 (i.e.: $784 = (224/8)^2$). Both the Q-Former and the decoder are trained for 300 epochs using perceptual loss. Hyperparameters were tuned on the validation sets. Further implementation details are presented in the supplementary material.

Evaluation Metrics. Consistent with previous work [4, 25], we report the Area Under the Receiver Operating Characteristic (AUROC) curve for anomaly detection. AUROC for localization is not reported due to its tendency to pro-

Table 1. Ablation results on BraTS2021 [2, 3, 41] of our medical anomaly detection framework, QFAE. We demonstrate step by step how incorporating components, such as Q-Former and perceptual loss, elevates a simple AE model to a strong medical anomaly detector using off-the-shelf models. MAE: Mean Absolute Error. $\mathcal{L}_{\text{Perceptual}}$: Perceptual loss based on the specified model.

	Q-Former	Loss	AUROC (%)
1	✗	MAE	66.6
2	✓	MAE	79.5
3	✓	$\mathcal{L}_{\text{Perceptual}}$ (Masked ViT)	86.8

duce overly optimistic scores in cases of severe pixel-wise class imbalance, which is prevalent in anomaly detection.

4.3. Ablation Study

Establishing the New Architecture. We ablate each component of our Q-Former Autoencoder and present the results on the BraTS2021 [2, 3, 41] data set in Table 1. To create an updated autoencoder architecture, we start with the basics of employing a pretrained encoder and training a decoder (row 1). We selected DINOv2 ViT-B/14 as the encoder due to its exceptional results on zero-shot tasks. The AE architecture contains only the encoder and the decoder without any bottleneck introduced, and it was trained by minimizing the mean absolute error between the input and the output of the decoder. This basic version of AE reaches an AUCROC score of only 66.6. Adding the Q-Former module as the bottleneck (row 2) improves the AUROC by 12.9 (from 66.6 to 79.5) showing that Q-Former is capable of retaining the structure of normal data, which makes it a good choice for anomaly detection. Lastly, changing the optimization loss from the mean absolute error to the perceptual loss computed based on Masked AE features, increases the performance to 86.8 (row 3). By applying these designed choices (Q-Former, perceptual loss), we evolve a simple AE architecture with modest results to a powerful and accurate framework that reaches strong performances.

The Impact of the Loss Function. We evaluate the impact of the loss function on the detection performance for medical anomalies, reporting the results in Table 2a. Using the mean absolute error alone or even combining it with the perceptual loss produces poor results. Training solely with the perceptual loss ($\mathcal{L}_{\text{Perceptual}}$) achieves the best performance, highlighting the superiority of deep feature-based optimization over pixel-level reconstruction.

The Impact of the Aggregation. We further evaluate the influence of different aggregation strategies on the anomaly score computation, with results reported in Table 2b. Defining the anomaly score as the maximum reconstruction error produces the best performance, which aligns with the intuition that anomalies are inherently harder to reconstruct.

The Impact of Perceptual Features. We analyze the effect

Table 2. Ablations results on the BraTS2021 [2, 3, 41] data set changing different components of our architecture. Perceptual loss achieves higher performance than the simple mean absolute error (MAE) optimization, along with taking the maximum of the error. We also notice that using multiple hidden layers from the perceptual encoder is better along with using a smaller patch size for the decoder. The default configuration is highlighted in light blue. $L_{\text{Perceptual}}$: Perceptual loss based on Masked AE.

(a) **Loss function.** Mean Absolute Error (MAE) decreases the performance when combined with the perceptual loss. The top performance is obtained with $L_{\text{Perceptual}}$.

Loss	AUROC
MAE	79.0
MAE, $L_{\text{Perceptual}}$	79.2
$L_{\text{Perceptual}}$	88.5

(b) **Aggregation in Eq. 4.** Selecting the maximum error within Eq. 4 yields top performance.

Function	AUROC
mean	88.5
max	92.6

(c) **Layers from the perceptual model.** Using multiple layers from the perceptual model achieves top performance.

Layers	AUROC
5, 11	92.6
11, 15, 19	93.0

(d) **Decoder patch size.** Reconstructing the input using smaller patch sizes achieves top performance.

Patch size	AUROC
8	93.0
16	92.5
32	91.1

of perceptual features when extracted from different layers of the Masked AE model [22], reporting the results in Table 2c. In our initial experiment, we extracted features from layers 5 and 11 to guide model optimization, achieving a performance of 92.6. Adding another layer further improves the performance to 93.0, demonstrating that incorporating additional signals during AE training is beneficial for robust anomaly detection.

The Impact of the Decoder Patch Size. Employing the Q-Former architecture as the bottleneck decouples the dependency between encoder and decoder output lengths. Therefore, the decoder can reconstruct the input at varying granularities (different patch sizes). We evaluated different patch sizes such as 8×8 , 16×16 and 32×32 , reporting the results in Table 2d. As anticipated, smaller patch sizes produce higher performance, enabling the decoder to generate more precise reconstructions.

Impact of Perceptual Model Patch Size. Feature extraction for computing the perceptual loss is entirely independent of the framework’s encoder and decoder, which enables the use of multi-scale patch sizes to compute the perceptual features. Interestingly, Table 3 reveals that larger patches lead to improved anomaly detection performance. However, the best performance of 94.4 AUROC on BraTS2021 is achieved by combining two large patch sizes (32×32 and 56×56 pixels), effectively creating a pyramid of features. This finding suggests that larger patch sizes better capture the structure of the data, making it easier to spot the differences, thus improving anomaly detection.

Impact of the Encoder. We evaluate different combination of encoders including DINO [10], DINOv2 [43], OpenCLIP [52] and Masked AE [22] reporting the anomaly detection results on BraTS2021 in Table 4. Among the single encoders, DINOv2 [43] demonstrated the best performance of 94.4 AUROC, underscoring its strong capability on zero-shot tasks. When combining DINOv2 [43] with DINO [10], the performance slightly improves reaching an AUROC of 94.5. However, we concluded that this improvement does not justify the computational burden of adding

Table 3. Anomaly detection results on BraTS2021 [2, 3, 41] in terms of AUROC (%) when changing the patch size of the Masked AE [22]. We notice that dividing the input into larger patches significantly improves the performance. Top results are highlighted in bold. The default configuration is highlighted in light blue.

Masked AE Input patch size	AUROC
16	72.7
56	92.8
16, 32, 56	93.0
32, 56	94.4

Table 4. Anomaly detection results on BraTS2021 [2, 3, 41] in terms of AUROC (%) when different pretrained encoders are employed. Notably, Masked AE [22] encoder obtains poor performance due to its ability to reconstruct the input. Both DINO [10] and DINOv2 [43] achieve strong performance. The default configuration is highlighted in light blue.

Encoders	AUROC
DINO ViT-B/8	94.3
OpenCLIP ViT-L/14	94.0
Masked AE ViT-L/16	71.5
DINOv2 ViT-L/14	94.4
DINOv2 ViT-L/14 + DINO ViT-B/8	94.5
DINOv2 ViT-L/14 + OpenCLIP ViT-L/14	93.6
DINOv2 ViT-L/14 + OpenCLIP ViT-B/32	94.3
DINOv2 ViT-L/14 + Masked AE ViT-B/16	76.7
DINOv2 ViT-L/14 + Masked AE ViT-L/16	74.3

an extra encoder. Therefore, DINOv2 [43] was selected as the default single encoder for our framework. Notably, the Masked AE [22] encoder exhibits poor performance, even when combined with DINO [10] or DINOv2 [43], primarily due to its strong reconstructive capacity that hinders anomaly discrimination.

The results from DINO [10], DINOv2 [43], and OpenCLIP [52] demonstrate that foundation models are effective

Table 5. Anomaly detection performance (mean + std) on BraTS2021 [2, 3, 41], RESC [23] and RSNA [56]. The results are reported for five repetitions of the experiment. *: denotes only three repetitions. The top results are reported in bold. Our method is able to outperform all methods obtaining state-of-the-art performance on all three data sets.

Methods	BraTS2021	RESC	RSNA
f-AnoGAN [51]	77.3 \pm 0.18	77.4 \pm 0.85	55.6 \pm 0.09
GANomaly [1]	74.8 \pm 1.93	52.6 \pm 3.95	62.9 \pm 0.65
DRAEM [60]	62.4 \pm 9.03	83.2 \pm 8.21	67.7 \pm 1.72
UTRAD [11]	82.9 \pm 2.32	89.4 \pm 1.92	75.6 \pm 1.24
DeepSVDD [48]	87.0 \pm 0.66	74.2 \pm 1.29	64.5 \pm 3.17
CutPaste [33]	78.8 \pm 0.67	90.2 \pm 0.61	82.6 \pm 1.22
SimpleNet [37]	82.5 \pm 3.34	76.2 \pm 7.46	69.1 \pm 1.27
MKD [50]	81.5 \pm 0.36	89.0 \pm 0.25	82.0 \pm 0.12
RD4AD [13]	89.5 \pm 0.91	87.8 \pm 0.87	67.6 \pm 1.11
STFPM [58]	83.0 \pm 0.67	84.8 \pm 0.50	72.9 \pm 1.96
PaDiM [12]	79.0 \pm 0.38	75.9 \pm 0.54	77.5 \pm 1.87
PatchCore [46]	91.7 \pm 0.36	91.6 \pm 0.10	76.1 \pm 0.67
CFA [32]	84.4 \pm 0.87	69.9 \pm 0.26	66.8 \pm 0.23
CFLOW [19]	74.8 \pm 5.32	75.0 \pm 5.81	71.5 \pm 1.49
CS-Flow [47]	90.9 \pm 0.83	87.3 \pm 0.58	83.2 \pm 0.46
P-VQ* [29]	94.3 \pm 0.23	89.0 \pm 0.48	79.2 \pm 0.04
QFAE (ours)	94.3 \pm 0.18	91.8 \pm 0.55	83.8 \pm 0.46

for anomaly detection, even within the medical domain.

4.4. Comparison with State-of-the-Art

We compare our framework QFAE against several state-of-the-art methods on BraTS2021 [2, 3, 41], RESC [23] and RSNA [56], presenting the results in Table 5. We report the mean and standard deviation (std) of the results obtained from 5 independent runs for each experiment.

Our method achieves state-of-the-art performance on all data sets. In particular, on the BraTS2021 data set, our framework achieves an AUROC of 94.3 ± 0.18 , on par with the previous best performing method (94.3 ± 0.23 achieved by P-VQ [29]). This result demonstrates strong anomaly detection capabilities in brain imaging obtained simply by enhancing the standard AE framework.

Furthermore, our method outperforms all baselines on all three data sets. On RESC, we achieve the highest AUROC score of 91.8 ± 0.55 , surpassing the previous state-of-the-art results of PatchCore [46] (91.6 ± 0.10). On RSNA, QFAE achieves an AUROC score of 83.8 ± 0.46 , outperforming the next best method, CS-Flow [47] (83.2 ± 0.46).

These top results highlight the robustness of our frame-

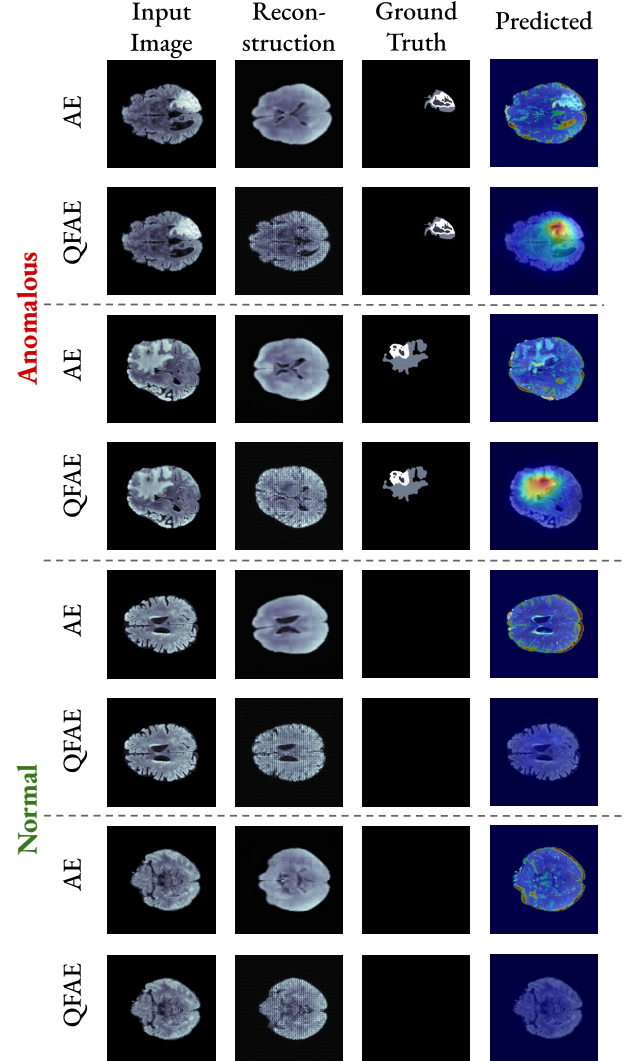


Figure 3. Qualitative examples of anomaly localization on several samples from the BraTS2021 [2, 3, 41] data set. For each sample, we present the original input, the reconstruction, the ground-truth and the predicted anomaly map. Both normal and abnormal samples are presented. Our Q-Former AutoEncoder (QFAE) with a traditional AutoEncoder (AE). Notably, our QFAE method consistently produces sharper and more accurate anomaly localizations compared to the baseline, closely aligning with the ground truth. Moreover, our QFAE predicts very low anomaly scores for normal samples, being able to correctly identify them as normal samples.

work across several medical imaging modalities (MRI, X-rays, and OCT). Additionally, this work further highlights that foundation models, primarily trained on natural images, can be successfully employed in a different domain, such as medical images, without additional finetuning.

4.5. Qualitative Results

We present qualitative results in Figure 3 and Figure 4. We illustrate samples from the BraTS2021 [2, 3, 41] and

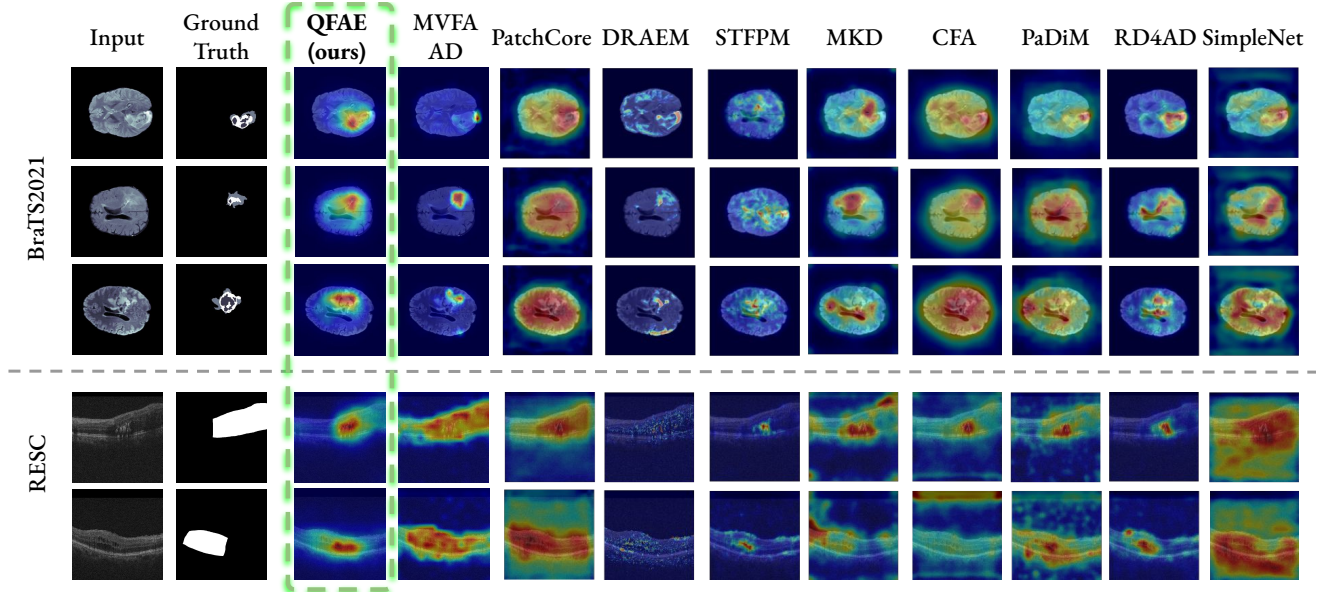


Figure 4. Qualitative examples of anomaly localization on several samples from the BraTS2021 [2, 3, 41] and RESC [23] data sets. For each sample, the columns show the original input, ground truth anomaly masks, and anomaly maps predicted by QFAE (ours) alongside various baseline methods. We note that MVFA-AD [25] uses a few-shot strategy, therefore it is not unsupervised as the rest of the methods including ours. The predicted anomalies for the baselines are cropped directly from BMAD [4]. Notably, our QFAE method consistently produces sharper and more accurate anomaly localizations compared to other approaches, closely aligning with the ground truth.

RESC [23] data sets, along with the input image, the ground-truth anomaly mask, the anomaly map predicted by several state-of-the-art methods and our QFAE framework in Figure 4. On both data sets, our framework precisely localizes the anomalies. Notably, on the second BraTS2021 sample, anomaly localization proved challenging for most methods, with only MVFA-AD [25] (few-shot), DRAEM [60], and our enhanced AE (QFAE) achieving correct localization among the 10 evaluated approaches. This highlights that our approach is capable to accurately identify subtle and difficult-to-detect anomalies.

Additionally, in Figure 3, we illustrate qualitative results comparing the anomaly results of our QFAE with those obtained by a traditional AE. The traditional AE employed a pretrained encoder and a decoder. We observed that our enhanced AE predicts anomalies that correlate well with the ground truth, while also yielding very low anomaly scores for normal samples. These results clearly show that using the Q-Former as a bottleneck is effective in detecting and localizing anomalies. Additionally, these findings indicate that the combined approach of incorporating Q-Former as a bottleneck and leveraging the Masked AE for perceptual loss is highly effective in enhancing medical anomaly detection performance.

5. Conclusions

In this paper, we introduced the **Q-Former AutoEncoder** (QFAE), a modernized autoencoder framework that lever-

ages the power of state-of-the-art pretrained vision foundation models for medical anomaly detection. Our framework addresses key limitations of traditional autoencoders by integrating frozen pretrained encoders (DINO [10], DINOv2 [43] and OpenCLIP [52]) for robust feature extraction, employing a trainable Q-Former as a *dynamic bottleneck* to produce fixed-length latent codes out of variable-length contextual input, and utilizing a perceptual loss function for semantically meaningful reconstruction. We rigorously evaluated QFAE on four diverse medical anomaly detection benchmarks: BraTS2021, RESC, RSNA, and LiverCT. Our results consistently demonstrate *state-of-the-art performance across these data sets*, achieving superior AUROC scores and precise anomaly localization. Our work highlights the successful and robust application of large-scale, pretrained vision foundation models (initially trained on natural images) for unsupervised anomaly detection in specialized medical imaging domains, notably without requiring extensive fine-tuning. In future work, we plan to apply QFAE to multi-class medical anomaly detection. **Limitations.** Despite its strong performance, our proposed QFAE framework has certain limitations. While using pretrained foundation models, such as DINO, DINOv2, Masked AE, etc., enhances the generalization capabilities and reduces training time, it inherently limits the model’s ability to learn domain-specific features. Despite our framework achieving consistently good results across modalities and data sets, we cannot claim that it will generalize to all anomaly types or varying levels of input complexity.

References

- [1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 7, 15
- [2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 2, 5, 6, 7, 8, 15
- [3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017. 2, 5, 6, 7, 8, 15
- [4] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4042–4053, 2024. 2, 5, 8, 14
- [5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019) - Volume 5: VISAPP*, pages 372–380. INSTICC, SciTePress, 2019. 3
- [6] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 2, 5, 13, 15
- [7] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 14
- [8] Yu Cai, Hao Chen, and Kwang-Ting Cheng. Rethinking Autoencoders for Medical Anomaly Detection from A Theoretical Perspective. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part XI*, pages 544–554, Berlin, Heidelberg, 2024. Springer-Verlag. 2
- [9] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A comparative study of anomaly detection in medical images, 2025. *arXiv:2404.04518 [cs]*. 2
- [10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2, 3, 4, 5, 6, 8
- [11] 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. 7, 15
- [12] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 7, 15
- [13] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 7, 15
- [14] Hanqiu Deng and Xingyu Li. Anomaly Detection via Reverse Distillation from One-Class Embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9727–9736, New Orleans, LA, USA, 2022. IEEE. 3
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 3, 4
- [16] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6568–6576, 2022. 2
- [17] Mariana-Iuliana Georgescu. Masked Autoencoders for Unsupervised Anomaly Detection in Medical Images. *Procedia Computer Science*, 225:969–978, 2023. 3
- [18] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI Conference on Artificial Intelligence*, 2023. 3
- [19] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 7, 15
- [20] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. ReContrast: domain-specific anomaly detection via contrastive reconstruction. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 10721–10740, Red Hook, NY, USA, 2023. Curran Associates Inc. 3
- [21] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Encoder-Decoder Contrast for Unsupervised Anomaly Detection in Medical Images. *IEEE Transactions on Medical Imaging*, 43(3):1102–1112, 2024. 3
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 2, 3, 4, 5, 6
- [23] 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. 2, 5, 7, 8, 15
- [24] Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised

- anomaly detection and localization. *IEEE Transactions on Multimedia*, 2022. 2
- [25] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385, 2024. 3, 5, 8
- [26] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616, 2023. 3
- [27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 3, 4
- [28] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVI*, page 113–132, 2024. 4
- [29] 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. 7, 15
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 3
- [31] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 2, 5, 13, 15
- [32] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 7, 15
- [33] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 7, 15
- [34] He Li, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, and Yen-Wei Chen. An accurate unsupervised liver lesion detection method using pseudo-lesions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–223. Springer, 2022. 14
- [35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 4
- [36] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep Industrial Image Anomaly Detection: A Survey. *Machine Intelligence Research*, 21(1):104–135, 2024. 2
- [37] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 7, 15
- [38] Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767, 2020. 3
- [39] Felix Meissen, Benedikt Wiestler, Georgios Kaissis, and Daniel Rueckert. On the pitfalls of using the residual error as anomaly score. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 914–928. PMLR, 2022. 3
- [40] Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert. Unsupervised Anomaly Localization with Structural Feature-Autoencoders. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 14–24, Cham, 2023. Springer Nature Switzerland. 3
- [41] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 2, 5, 6, 7, 8, 15
- [42] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021. 14
- [43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 3, 4, 5, 6, 8
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3, 4
- [45] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022. 14
- [46] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 7, 15

- [47] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097, 2022. [7](#), [15](#)
- [48] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. [7](#), [15](#)
- [49] Mayu Sakurada and Takehisa Yairi. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, New York, NY, USA, 2014. Association for Computing Machinery. [2](#)
- [50] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. [7](#), [15](#)
- [51] 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. [7](#), [15](#)
- [52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [2](#), [3](#), [5](#), [6](#), [8](#)
- [53] Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly Detection in Medical Imaging With Deep Perceptual Autoencoders. *IEEE Access*, 9:118571–118583, 2021. [3](#), [4](#)
- [54] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24511–24520, 2023. [3](#)
- [55] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998. [14](#)
- [56] 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 the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. [2](#), [5](#), [7](#), [15](#)
- [57] Rui Xu, Yunke Wang, and Bo Du. MAEDiff: Masked Autoencoder-enhanced Diffusion Models for Unsupervised Anomaly Detection in Brain Images. *CoRR*, abs/2401.10561, 2024. arXiv: 2401.10561. [3](#)
- [58] Shinji Yamada and Kazuhiro Hotta. Reconstruction student with attention for student-teacher pyramid matching. *arXiv preprint arXiv:2111.15376*, 2021. [7](#), [15](#)
- [59] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, pages 4571–4584. Curran Associates, Inc., 2022. [2](#)
- [60] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. [7](#), [8](#), [15](#)
- [61] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit features for multi-class unsupervised visual anomaly detection. *Comput. Vis. Image Underst.*, 253:104308, 2025. [3](#)
- [62] Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. AE-FLOW: Autoencoders with Normalizing Flows for Medical Images Anomaly Detection. 2022. [3](#)
- [63] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *ArXiv*, abs/2310.18961, 2023. [3](#)

Q-Former Autoencoder: A Modern Framework for Medical Anomaly Detection

Supplementary Material

We provide additional implementation details in Section 6, additional experiments on LiverCT and RSNA in Section 7 and Section 8.

6. Implementation Details

This section provides an overview of the implementation details for our proposed framework, ensuring full reproducibility of our results. All experiments were conducted in PyTorch.

6.1. Hyperparameters

The main hyperparameters used for training and evaluation are detailed in Table 6 and Table 7, respectively.

Table 6. Training hyperparameters for the experiments.

Component	Parameter	Value
General	Seed	42, 7, 13, 65, 91 (mean of 5 runs are reported)
	Image Resolution (Resize)	224x224
	Batch Size	64
	Epochs	300
	Device	CUDA
Encoder	Pre-trained Model	ViT-Large (ViT-L/14) with register tokens
	Pre-training Method	DINOv2
	Frozen During Training	True
	Hidden States Used	Features from the 2nd and 4th to last blocks
	Final Projection In-Features	1024
Q-Former (Junction)	Final Projection Out-Features	768
	Number of Transformer Blocks	1
	Internal Dimension	768
	Output Dimension	768
	Number of Learnable Queries	784 (for 28x28 output patches)
Decoder	Attention Heads	8
	MLP Expansion Ratio	4.0
	Internal Dimension	768
	Depth (Number of Layers)	6
	Attention Heads	12
Optimization	Output Patch Size	8x8
	Number of Output Patches	28x28
	MLP Expansion Ratio	4.0
	Optimizer	Adam
	Learning Rate (Maximum)	8×10^{-5}
Perceptual Loss	Learning Rate Scheduler	OneCycleLR
	Pre-trained Perceptual Model	Masked Autoencoder (MAE) with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 16th and 20th transformer blocks
	Multi-Scale Input Patch Sizes	32x32, 56x56

Table 7. Evaluation configuration for the experiments.

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	16x16, 32x32, 56x56
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	Max
	Cross-Feature Map Aggregation	Mean
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

6.2. Perceptual Loss Formulation

The training objective is to minimize a multi-scale perceptual loss. This loss is calculated in a three-step process:

Step 1: Feature Extraction. For an input image x and its reconstruction \tilde{x} , we extract feature maps from a set of

pretrained perceptual models. We use multiple Masked Autoencoder (Masked AE) models, each distinguished by its input patch size $p \in P$. For each model, we select features from a set of transformer blocks $i \in I$. Let $\Phi_{i,p}(x)$ be the feature map of shape $C_i \times H_i \times W_i$ extracted from the i -th layer of the perceptual model with patch size p .

Step 2: Anomaly Map Calculation. For each selected feature map, we compute an intermediate anomaly map, $A_{i,p}$, by calculating the cosine distance between the features of the original image and its reconstruction at every spatial location (j, k) .

$$A_{i,p}(j, k) = 1 - \frac{\Phi_{i,p}(x)_{j,k} \cdot \Phi_{i,p}(\tilde{x})_{j,k}}{\|\Phi_{i,p}(x)_{j,k}\|_2 \cdot \|\Phi_{i,p}(\tilde{x})_{j,k}\|_2}$$

This produces a set of single-channel anomaly maps, one for each combination of layer i and patch size p .

Step 3: Hierarchical Aggregation and Final Loss. The final loss is computed using a two-stage hierarchical aggregation. First, for each feature layer $i \in I$, we create a robust, layer-specific anomaly map, $A_{\text{combined},i}$, by performing an element-wise multiplication of its corresponding anomaly maps from all different patch-size models $p \in P$. This step enforces a strict consensus across multiple scales for each feature level.

$$A_{\text{combined},i} = \prod_{p \in P} \text{Resize}_{(H,W)}(A_{i,p})$$

Second, the total loss \mathcal{L} is calculated by averaging the mean value of each of these robust, layer-specific maps. This treats the error signal from each feature layer as an independent contribution to the total loss.

$$\mathcal{L}(x, \tilde{x}) = \frac{1}{|I|} \sum_{i \in I} \text{mean}(A_{\text{combined},i})$$

For training, we use patch sizes $P = \{32, 56\}$ and features from the 16th and 20th transformer blocks of the Masked AE ViT-Large encoder.

6.3. Anomaly Score and Map Generation

During evaluation, we generate both an image-level scalar score for AUROC computation and a pixel-level anomaly map. Both start from the same set of intermediate anomaly maps, $A_{i,p}$, though computed using the evaluation configuration (Table 7). Let this evaluation set of maps be denoted by $\mathcal{A} = \{A_1, A_2, \dots, A_N\}$.

Image-Level Anomaly Score Aggregation. To derive a single scalar score for each image, we perform a two-step aggregation:

Step 1: Spatial Aggregation. For each anomaly map $A_n \in \mathcal{A}$, we find the maximum pixel value. This value, s_n , represents the most severe reconstruction error detected by that specific feature map.

$$s_n = \max_{j,k} (A_n(j, k))$$

Step 2: Cross-Feature Aggregation. The final image-level score, A_{score} , is the mean of these maximum values, averaged over all N feature maps.

$$A_{\text{score}} = \frac{1}{N} \sum_{n=1}^N s_n$$

This method gives a robust score that is sensitive to strong local anomalies while benefiting from the diversity of features from different layers.

Pixel-Level Anomaly Map Generation. To generate a final 2D anomaly map, we use a different aggregation strategy that preserves spatial information. At each spatial location (j, k) , we take the mean value across all N resized anomaly maps.

$$A_{\text{pixel-max}}(j, k) = \max_{n \in \{1..N\}} (A_n(j, k))$$

6.4. Training and Data Augmentation

The model is trained using the Adam optimizer with a OneCycleLR learning rate scheduler. To encourage the model to learn robust and generalizable representations of normal data, the following data augmentations are applied to the training set:

- **Random Resized Crop:** Images are cropped to a random size (90% to 100% of the original) and aspect ratio (80% to 120% of the original) before being resized to the final input dimension.
- **Random Rotation:** Images are rotated by a random angle between -10 and +10 degrees.
- **Random Vertical Flip:** Images are flipped vertically with a 50% probability.
- **Color Jitter:** The brightness and contrast of the images are randomly adjusted by a factor of up to 0.1.
- **Normalization:** Image pixel values are normalized to have a mean of 0.449 and a standard deviation of 0.226.

7. Experiments on LiverCT

We performed a few pre-processing steps on the LiverCT [6, 31] benchmark. In this section, we introduce these techniques one by one and complete the Table 8

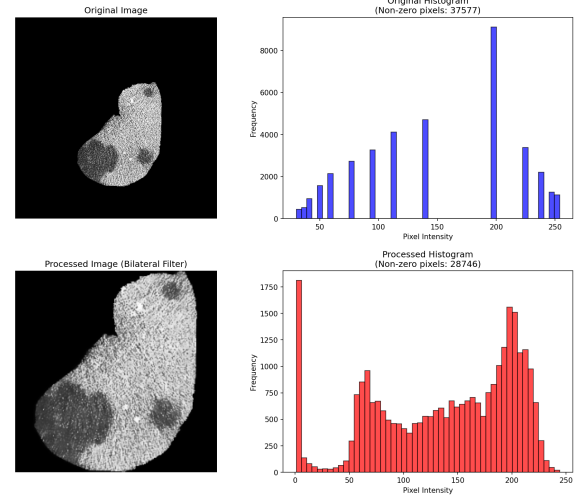


Figure 5. Original and processed images from LiverCT [6, 31] along with their pixel histograms.

7.1. Data Preprocessing

First of all, the dimensions of images in the dataset are 512x512 pixels, and only small portion of the images have the Liver segments. Resizing these images to 224x224 pixels, which is the input size of our model, results in diminished liver sections. To resize without losing details of the region of interest (i.e. liver section) we used the following algorithm to get images resized to 224x224. Note that this process is fully automated and can be applied to any segmented liver images.

1. **ROI Identification:** For each 512×512 input image, we first identify the region containing the liver. This is achieved by computing a union bounding box that tightly encloses all non-zero pixels.
2. **ROI Cropping:** The image is cropped using the coordinates of the calculated bounding box, isolating the liver segment from the empty background.
3. **Canvas Preparation:** A new, black canvas of the target dimensions (224×224) is created to serve as the background for the final model input.
4. **Conditional Resizing and Placement:** The cropped liver ROI is placed onto the canvas using a size-dependent strategy:
 - **If the ROI is smaller than or equal to 224×224 :** The cropped segment is pasted directly onto the center of the canvas without any resizing. This preserves the native resolution of the liver tissue.
 - **If the ROI is larger than 224×224 :** The segment is resized to fit within the 224×224 frame while maintaining its original aspect ratio to prevent distortion. The resized ROI is then centered on the canvas.
5. **Final Input:** The resulting 224×224 image, with the

Table 8. Ablations on LiverCT Dataset.

Version	AUROC
1 Main Config 6	54.1
2 + Train & Eval with New Preprocessing	59.5 ± 1.27
3 + Eval Perceptual Patch Sizes [16, 32, 56] \rightarrow [8, 16]	65.5 ± 1.96

liver segment prominently centered, is used as the input for the model.

Another issue with this dataset is that, due to constraints inherent to Computed Tomography imaging, it underwent several windowing and histogram equalization techniques [4, 7, 34]. As a result, these images can be out of distribution of standard datasets like ImageNet, on which our employed perceptual loss model is trained. To mitigate this, we apply a bilateral filter [55] to each processed 224×224 image prior to feeding it to the network.

The effect of the pre-processing is illustrated in Figure 5, where the ROI and anomalous regions are preserved, and the histogram of the image looks more natural.

Retraining and re-evaluating the model with this new preprocessing algorithm yielded the results in Row 2 of the Table 8. This result is mean and standard deviation of evaluation of 5 different models trained with 5 different seed (42, 7, 13, 65, 91)

7.2. New Evaluation Config

As shown in Figure 6, the new data preprocessing pipeline (column 2) has improved over the original config (column 1) regarding the quality of the anomaly map and made the anomalous region more visible. However, the predicted anomaly map still fails to capture the texture change in the anomalous region. Following the studies on visual perception [42, 45] that state smaller patch sizes are biased towards textures while larger patch sizes are biased towards shape, we changed the patch sizes used by perceptual model for anomaly score calculation from [16, 32, 56] to [8, 16]. As can be seen from the third column of Table 6, with this evaluation config anomaly maps capture texture changes on anomalous regions better. This reflects on the AUROC score of the 3rd row of the Table 8. The evaluation configuration that yields the best result on LiverCT is presented in Table 9, with the modified parts highlighted in bold. The training config is kept the same.

Table 9. Best Evaluation Configuration on LiverCT.

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	8x8s, 16x16
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	Max
	Cross-Feature Map Aggregation	Mean
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

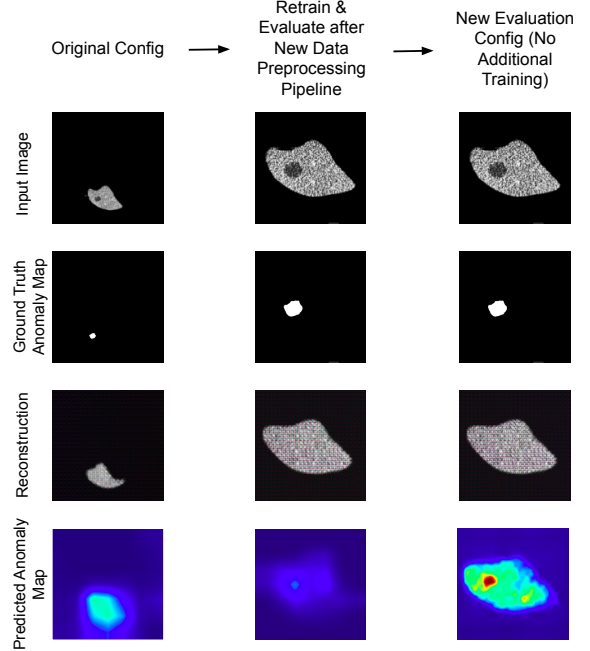


Figure 6. Effect of the modifications such as data preprocessing pipeline and evaluation config. We first avoid diminishing the anomalous region during resizing. Then configured perceptual loss to be more biased towards textual clues following insights from literature on visual perception.

8. Different Aggregation for Chest RSNA

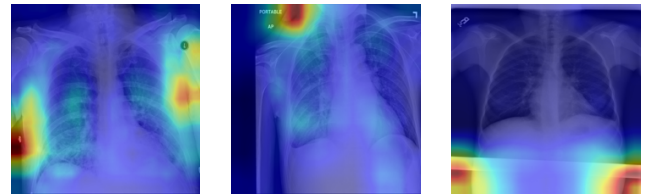


Figure 7. Optical characters and artifacts dominate the response from the anomalous region.

It is usual to see different optical characters and artifacts on Chest images. Their position varies. When these artifacts are present, they dominate the anomaly signals, and the anomaly score cannot be calculated properly with the aggregation method described in Main Eq. 4. Since their positions vary and are unpredictable, we were unable to devise a preprocessing algorithm.

To mitigate this problem, we decided to experiment with different aggregation methods on the validation split of the Chest RSNA dataset. As an alternative, we first tried taking the mean value in the anomaly map from each location, and then taking the maximum across different layers. We

observed an increase in AUROC from 78.6% to 84.3% on the validation split. Therefore, we decided to keep this approach and reported an AUROC of 83.8% on test set as in main Table 5. The evaluation configuration that yields the best result on Chest RSNA is presented in 10, with the modified parts highlighted in bold. The training config is kept the same.

Table 10. Best Evaluation Configuration on Chest RSNA

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	16x16, 32x32, 56x56
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	Mean
	Cross-Feature Map Aggregation	Max
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

9. SOTA Results on Each Dataset

Table 11. Best Training Configuration for Brain MRI.

Component	Parameter	Value
General	Seed	42, 7, 13, 65, 91 (mean of 5 runs are reported)
	Image Resolution (Resize)	224x224
	Batch Size	64
	Epochs	300
	Device	CUDA
Encoder	Pre-trained Model	ViT-L/14 + ViT-B/8
	Pre-training Method	DINOv2 + DINO
	Frozen During Training	True, True
	Hidden States Used	Features from the 2nd and 4th to last blocks
	Final Projection In-Features	1024, 768
	Final Projection Out-Features	768, 768
Q-Former (Junction)	Number of Transformer Blocks	1
	Internal Dimension	768
	Output Dimension	768
	Number of Learnable Queries	784 (for 28x28 output patches)
	Attention Heads	8
	MLP Expansion Ratio	4.0
Decoder	Internal Dimension	768
	Depth (Number of Layers)	6
	Attention Heads	12
	Output Patch Size	8x8
	Number of Output Patches	28x28
	MLP Expansion Ratio	4.0
Optimization	Optimizer	Adam
	Learning Rate (Maximum)	8×10^{-5}
	Learning Rate Scheduler	OneCycleLR
Perceptual Loss	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 16th and 20th transformer blocks
	Multi-Scale Input Patch Sizes	32x32, 56x56

As shown in Table 12, we achieve the state-of-the-art performance in BraTS2021 [2, 3, 41], RESC [23] and RSNA [56] and second on LiverCT [6, 31].

Table 12. Anomaly detection performance (mean + std) on BraTS2021, Liver CT (BTCV + LiTs), RESC and RSNA. The results are reported for five repetitions of the experiment. *: denotes only three repetitions. The top results are reported in bold.

Methods	BraTS2021	Liver CT	RESC	RSNA
f-AnoGAN [51]	77.3 ± 0.18	58.4 ± 0.15	77.4 ± 0.85	55.6 ± 0.09
GANomaly [1]	74.8 ± 1.93	53.9 ± 2.36	52.6 ± 3.95	62.9 ± 0.65
DRAEM [60]	62.4 ± 9.03	69.2 ± 3.86	83.2 ± 8.21	67.7 ± 1.72
UTRAD [11]	82.9 ± 2.32	55.6 ± 5.96	89.4 ± 1.92	75.6 ± 1.24
DeepSVDD [48]	87.0 ± 0.66	53.3 ± 1.24	74.2 ± 1.29	64.5 ± 3.17
CutPaste [33]	78.8 ± 0.67	58.6 ± 4.2	90.2 ± 0.61	82.6 ± 1.22
SimpleNet [37]	82.5 ± 3.34	N/A	76.2 ± 7.46	69.1 ± 1.27
MKD [50]	81.5 ± 0.36	60.4 ± 1.61	89.0 ± 0.25	82.0 ± 0.12
RD4AD [13]	89.5 ± 0.91	60.0 ± 1.4	87.8 ± 0.87	67.6 ± 1.11
STFPM [58]	83.0 ± 0.67	61.6 ± 1.7	84.8 ± 0.50	72.9 ± 1.96
PaDiM [12]	79.0 ± 0.38	50.7 ± 0.5	75.9 ± 0.54	77.5 ± 1.87
PatchCore [46]	91.7 ± 0.36	60.4 ± 0.82	91.6 ± 0.10	76.1 ± 0.67
CFA [32]	84.4 ± 0.87	61.9 ± 1.16	69.9 ± 0.26	66.8 ± 0.23
CFLOW [19]	74.8 ± 5.32	49.9 ± 4.67	75.0 ± 5.81	71.5 ± 1.49
CS-Flow [47]	90.9 ± 0.83	59.4 ± 0.52	87.3 ± 0.58	83.2 ± 0.46
P-VQ* [29]	94.3 ± 0.23	60.6 ± 0.62	89.0 ± 0.48	79.2 ± 0.04
QFAE (ours)	94.3 ± 0.18	65.5 ± 1.96	91.8 ± 0.55	83.8 ± 0.46