Q-Former 自编码器：一种用于医学异常检测的现代框架

Francesco Dalmonte[1,*] Emirhan Bayar[2,*] Emre Akbas[2,3]

Mariana-Iuliana Georgescu[3] [1]University of Bologna, Italy

[2]Middle East Technical University, Ankara, Türkiye [3]Helmholtz Munich, Germany

Abstract

医学图像中的异常检测是一项重要但具有挑战性的任务，这主要是因为可能出现的异常多样且实际很难收集全面标注的数据集。在这项工作中，我们提出了一种现代化的基于自编码器的框架 Q-Former Autoencoder，以应对无监督的医学异常检测问题，该框架利用了最先进的预训练视觉基础模型，如 DINO、DINOv2 和 Masked Autoencoder。与从头训练编码器不同，我们直接利用冻结的视觉基础模型作为特征提取器，从而无需进行特定领域的微调，即可获得丰富的、多阶段的高级表征。我们建议使用 Q-Former 架构作为瓶颈，这样可以有效地聚合多尺度特征，同时控制重建序列的长度。此外，我们还结合了使用预训练 Masked Autoencoder 特征计算的感知损失，以引导重建朝向语义上有意义的结构。我们的框架在四个不同的医学异常检测基准上进行了评估，在 BraTS2021、RESC 和 RSNA 上取得了最先进的结果。我们的结果突出显示了预训练在自然图像上的视觉基础模型编码器在未经进一步微调的情况下，有效推广到医学图像分析任务的潜力。我们的代码和模型发布在 https://github.com/emirhanbayar/QFAE 。

1. 引言

在医学影像中实现自动化异常检测是一个关键问题，因为它直接影响诊断的准确性、工作流程的效率和患者的结果。然而，手动检查大容量的医学扫描（如磁共振成像（MRI）或计算机断层扫描（CT））天生是耗时的，并且容易受到人为错误的影响，这突显了需要可靠的自动化系统来帮助医生标记潜在异常。然而，自动化的医学异常检测也面临显著挑战。异常表现形式和外观高度多样化，这使得收集所有可能的病理变异的代表性样本变得不可行。结果是，以无监督异常检测方法为恰当，该方法专注于在正常数据上训练模型，以识别偏离情况为异常。

*Equal contribution

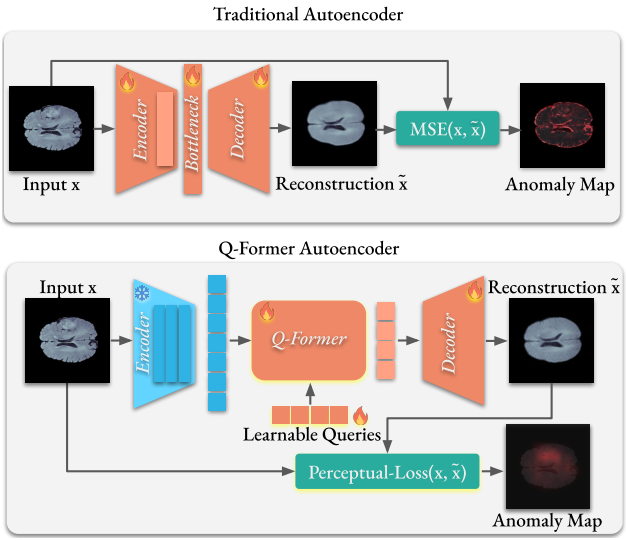

Figure 1. 我们展示了用于异常检测的传统自动编码器（上）与增强了 Q-Former 和感知损失的 Q-Former 自动编码器（下）。传统自动编码器通常使用一个可训练的编码器-解码器对，并依赖均方误差（MSE）进行优化和异常检测。我们的框架包括以下改进（用黄色突出显示）：(i) 一个冻结的编码器（采用强大的预训练视觉基础模型，如 DINO、DINOv2 和 OpenCLIP），(ii) 一个作为动态、可学习瓶颈的 Q-Former，用于高效表示，以及 (iii) 基于 Masked Autoencoder 的感知损失函数的使用。我们的框架能够精确揭示异常区域，从而产生有意义的异常检测（右下，红色）。

早期关于无监督异常检测的研究主要依靠卷积自动编码器，这些编码器经过训练用来重建正常图像。这些传统的自动编码器在表现力上有限，从而限制了它们在异常检测中的有效性。最近在视觉基础模型方面的进展，如 DINO [10]、DINOv2 [43] 和掩码自动编码器（Masked AE） [22]，展示了其在各种任务上的卓越表示迁移能力。尽管这些模型具有潜力，但在医学图像异常检测中却基本被忽视。为数不多的例外之一是 MVFA-AD [24]，它采用了 CLIP 模型 [44] 来执行零样本和小样本的医学异常检测。不幸的是，与任务特定

www.xueshuxiangzi.com

的方法相比，这些方法通常存在性能差距。为弥合这一差距，我们提出了一种新型框架，称为 Q-前置编码器自动编码器，该框架通过集成视觉基础模型和基于 Q-前置的注意机制瓶颈机制，现代化了无监督医学异常检测的自动编码器方法，如图 1 所示。首先，我们利用预训练的视觉基础模型，即 DINO [10] 、DINOv2 [43] 和 OpenCLIP [52] ，作为冻结编码器，在不需要领域特定的重新训练或微调的情况下提取稳健和语义丰富的特征。其次，我们引入一个 Q-前置模型作为灵活瓶颈，它聚合多尺度特征并输出固定长度的潜在表示。这种设计提供了对重建粒度的明确控制，同时提升了模型准确表示正常结构的能力。第三，我们使用预训练遮掩自动编码器提取的特征计算感知损失，鼓励重建过程保持高层次语义而不是低层次像素细节。我们现代化的自动编码器在准确检测和定位异常方面已经显着优于其标准对手，如图 1 所示。为了评估我们的框架，我们在 BMAD [4] 基准测试中的四个数据集上进行了广泛实验：BraTS2021 [2, 3, 41] 、RESC [23] 、RSNA [56] 和 LiverCT [6, 31] 。

我们的框架在所有数据集上都实现了最新的评分，在 BraTS2021 上达到了 AUROC 为 94.3% ，在 RSNA 上达到了 83.8% ，展示了其在不同图像模态上的有效性，包括 MRI、OCT 和 X 射线。总之，我们的贡献有以下三个方面：

- 我们提出了一种现代化和增强的自动编码器方法，该方法结合了冻结的视觉基础模型、Q-Former 瓶颈和感知损失，用于无监督异常检测。
- 我们提出的框架表现优异，在三种医学异常检测基准（即 BraTS2021、RESC、RSNA）上达到最先进的 AUROC 分数，同时不需要特定领域的编码器微调。
- 我们提供了详细的消融实验，展示了主要在自然图像上训练的视觉基础模型如何在结合适当的结构调整时能够有效地推广到医学图像领域。

2. 相关工作

2.1. 异常检测方法的分类

学习策略。基于图像的异常检测（AD）方法通常分为有监督、无监督和零样本方法。有监督的方法，例如基于少样本学习或合成异常生成的方法，需要访问一些标注的异常样本。另一方面，零样本方法旨在识别无分布样本而不访问领域数据，通常依赖于预训练模型。虽然在自然视觉领域有一定的前景 [16] ，但在工业检测或医学成像等需要领域特定知识的专业领域仍然受限。在这种情况下，无监督的 AD 仍然是最相关的设置。这些方法仅在正常样本上训练，并在推理阶段检测偏差。尽管最近的工作已经研究了多类 AD [59] ，但这些方法通常与专门的算法相比表现较差，限制了它们在诸如医学分析等敏感或安全关键领域的适用性。

特征嵌入或重建为基础。一个经典的异常检测方法分类 [36] 将其分为两大类：特征嵌入和基于重建的方法。特征嵌入方法依赖于学习特征空间中的距离或密度估计。相反，基于重建的方法，例如基于自编码器的方法，学习专门重建正常数据，假设模型无法有效重建异常。这些方法展示了强大的性能，即使作为简单基线进行实现，并且架构复杂性极低 [9] ，同时在本质上支持可解释性和异常定位—这在医疗异常检测中特别有价值。在这项工作中，我们提出了一个基于输入重建的框架。

2.2. 用于异常检测的自编码器架构

自编码器学习训练数据的压缩潜在表示，并尝试将其重新投影到输入空间。众所周知，潜在表示的压缩对于自编码器的异常检测（AD）能力至关重要 [8, 49] 。这种方法面临的主要挑战是，在良好重建正常图像的同时，防止模型对异常样本进行泛化。

一个关键方面是重构指标的选择 [39] ：除了 L2 损失之外，还探讨了结构相似性指数（SSIM） [5, 40] ，以及感知损失 [27, 53] 。最近提出的一些最有效的方法是测量特征空间而不是图像空间中的距离，显示出稳健的结果 [20, 21, 40] 。这种方法通常与知识蒸馏技术结合使用，以进一步扩大异常样本的距离 [14, 54] 。其他相关的方法涉及变分自编码器 [38] ，掩码自编码器 [17, 57] 和归一化流机制 [62] 。类似于上述研究 [27, 53] ，我们采用感知损失来训练自编码器。然而，与以往的工作不同的是，我们利用掩码自编码器来引导我们模型的优化。

2.3. 视觉基础模型

最近在大规模模型预训练方面的进展促进了高通用性基础模型在视觉任务中的发展，这主要依赖于 Vision Transformer (ViT) [15] 架构。值得注意的例子包括采用对比学习框架的 CLIP [44] 、使用各种自监督方案训练的 DINOv2 [43] 和 Masked Autoencoder [22] ，以及像 SAM [30] 这样的监督模型。这些模型在大规模数据集上训练，学习到丰富的表示，能够捕获语义和结构化的图像信息，从而在不同的下游任务中实现强大的泛化能力。

对于无监督异常检测（AD）使用高容量视觉基础模型的研究仍然较少。Zhang 等人 [61] 使用冻结的 ViTs 建立了一个多类别 AD 基准。更近期的方法利用了视觉-语言模型：Jeong 等人 [26] 采用组合提示集成和滑动窗口进行分割，并结合记忆库以实现少样本学习。Zhou 等人 [63] 使用与对象无关的模板和提示调优。Huang 等人 [25] 通过专用适应模块解决了领域偏移问题。Gu 等人 [18] 将多模态对话模型重新用于 AD，在工业基准上取得了良好的结果。

虽然这些模型显示了有前途的性能，特别是在零样本和少样本的情况下，但一个可扩展的、统一的方法来充分利用基础模型仍然缺乏，通常导致与任务特定方法相比的性能差距。因此，在这项工作中，我们提出充分利用基础模型，提出一个增强的自编码器框架，配备了 Q-Former 和基于 Mask AE 的感知损失。

www.xueshuxiangzi.com

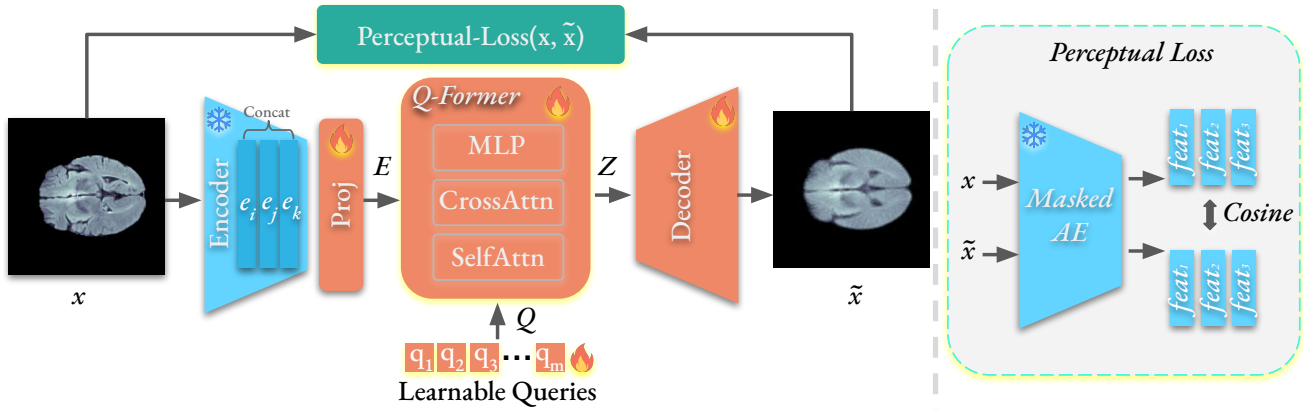

Figure 2. 我们用于医学异常检测的 Q-Former 自编码器的训练。我们的框架使用预训练的基础模型，例如 DINO [10]、DINOv2 [43] 或 OpenCLIP [52]，来提取多尺度特征（E）。这些特征与可学习的查询令牌（Q）一起由 Q-Former 处理，充当动态瓶颈。输出 z 进入解码器以重建 $\tilde{x}$。基于从 Masked AE [22] 提取的多尺度特征的感知损失指导语义重建的训练。

3. Q-Former 自编码器

3.1. 概述

自编码器（AE）模型由于能够学习正常数据的紧凑表达而常用于异常检测。自编码器由编码器、潜在空间（瓶颈）和解码器组成。编码器将输入 x 压缩成潜在表示 z，即 $z = \text{Encoder}(x)$。该潜在表示预期能捕捉数据中最有信息的方面。解码器从 z 重构输入，将其映射回输入空间，即 $\tilde{x} = \text{Decoder}(z)$。AEs 通常通过最小化输入 x 与其重构 $\tilde{x}$ 之间的重构误差来训练。

仅在正常数据上训练自编码器可以实现异常检测，因为模型应该能够准确重建正常数据，而无法重建异常。我们展示了我们 Q-Former 自编码器（QFAE）的训练，强调在图 2 中集成了 Q-Former 和感知损失。

编码器。在视觉 Transformer（ViTs）[15] 引入之前，卷积神经网络（CNNs）是编码器的标准选择。然而，现代基础模型 [10, 22, 43, 44] 主要采用 ViT 架构。ViTs 首先将输入图像分割为不重叠的补丁，并使用浅层神经网络提取补丁表示。随后，对这些补丁表示应用自注意机制，结合其他归一化和前馈层。这些操作重复多次。最终，ViT 编码器输出一个固定长度的补丁嵌入序列。

在这项工作中，我们使用了来自基础模型的预训练编码器，例如 DINO [10]、DINOv2 [43]、CLIP [44]、Masked AE [22]。

受到 BLIP-2 [35] 和 BRAVE [28] 的启发，我们采用 Q-Former 架构作为瓶颈。Q-Former 非常适合，因为它处理可变长度的上下文输入以生成固定长度的潜码，能够结合来自不同层次甚至不同架构的标记。Q-Former 的输入是一组可学习标记，标记的数量控制重建的 patch 的数量。这允许在多个粒度（不同 patch 大小）上重建输出。Q-Former 通过交叉注意力层与编码器特征进行交互。编码器的输出作为键和值，被 Q-Former 的查询交叉注意，如图 2 所示。该设计使 Q-Former 能够有效地汇集编码器潜特征的信息，因为 Q-Former 消除了平方自注意力。对于给定输入 x，我们获得其嵌入为 $[e_i, e_j, \ldots, e_k]$，其中 [.] 是连接操作，e_i, e_j, e_k 是预训练的 ViT 基础模型不同层的特征。这些特征随后通过投影层适应当前任务：$E = \text{Proj}([e_i, e_j, \ldots, e_k])$，如图 2 所示。我们定义可学习的查询 $Q = [q_1, q_2, \ldots, q_m]$，其中 m 是所需输出序列的长度。Q-Former 的一个模块定义为：

根据我们的验证实验，我们在框架中仅使用一个 Q-Former 模块。

解码器。 解码器接收由 Q-Former 产生的潜在表示 Z 作为输入，并重建原始输入图像 x。如前所述，重建序列的长度由 Q-Former 中可学习查询的数量控制。解码器架构是一个只有少数几层的轻量级 Transformer。因此，令牌的重建序列为 $\tilde{x}_{tok} = \text{Decoder}(Z)$。在最后一步中，我们通过重新排列令牌来重建输入图像 $\tilde{x} = \text{unpatchify}(\tilde{x}_{tok})$。

感知损失。 自编码器通常通过最小化输入 x 和其重构 $\tilde{x}$ 之间的平均平方误差或平均绝对误差来进行训练。在实践中，感知损失已被提出以提高重构质量 [27, 53]。我们使用从经过预训练的掩码自编码器（Masked AE）的不同层中提取的特征来计算感知损失。感知损失最小化原始图像和重构图像特征之间的余弦距离：

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

，其中 I 是从中获取特征 feat 的所选层索引集，使用 Masked AE [22]。

我们计算异常分数的方法类似于感知损失，通过比较从预训练的 Masked AE 的多层中提取的特征，这些特征来源于原始图像和重建图像。

www.xueshuxiangzi.com

对于在所选层集合 I 中的每一层 i ，我们分别提取与原始输入及其重构对应的特征图 $\text{feat}_i \in \mathbb{R}^{h_i \times w_i \times c}$ 和 $\tilde{\text{feat}}_i \in \mathbb{R}^{h_i \times w_i \times c}$ 。然后，我们通过计算每个空间位置 (j,k) 上对应的特征向量（patch 嵌入）之间的余弦距离，计算层级异常图 $A_{\text{map},i}$ 。

$$A_{\text{map},i}(j,k) = 1 - \frac{\text{feat}_{i,(j,k)} \cdot \tilde{\text{feat}}_{i,(j,k)}}{||\text{feat}_{i,(j,k)}||_2 ||\tilde{\text{feat}}_{i,(j,k)}||_2}. \quad (2)$$

图像的最终异常评分是一个单标量值，通过从这些逐层的映射中取每个映射的最大值并对这些最大值进行平均计算而得出的：

$$A_{\text{score}} = \frac{1}{|I|} \sum_{i \in I} \max(A_{\text{map},i}). \quad (3)$$

为了便于可视化，通过像素平均所有分层异常图生成一个综合异常图。

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

这个最终的异常图可以实现图像中异常区域的定位，如图 4 所示。

4. 实验

4.1. 数据集

我们报告了三个数据集的结果：BraTS2021 [2, 3, 41] 、RESC [23] 和 RSNA [56] ,具体如下。有关 LiverCT [6, 31] 数据集的更多结果已在补充材料中提供。

BraTS2021。BraTS2021 [2, 3, 41] 数据集，是 BMAD [4] 基准测试的一部分，包含带有各种异常的像素级注释的脑部 MRI 图像。BraTS2021 总共有 11,298 张图像，分为 7,500 个训练样本、83 个验证样本和 3,715 个测试样本。每张切片的分辨率为 240×240 像素。

RESC。RESC [23] ，也属于 BMAD [4] ，包括视网膜 OCT 图像。数据总共包含 6,217 张图像，其中 1,805 张用于测试。所有图像都是高分辨率的，尺寸为 512×1024 像素。

RSNA。RSNA 数据集 [56] ，包含在 BMAD [4] 基准测试中，由具有图像级异常注释的胸部 X 光片组成。它包含 26,684 个分辨率为 1024×1024 的图像，分为 8,000 个训练样本，1,490 个验证样本和 17,194 个测试样本。

我们采用了不同的预训练视觉基础模型作为编码器，包括 DINO [10] 、DINOv2 [43] 、OpenCLIP [52] 和 Masked Autoencoder [22] 。如前所述，编码器在框架的训练过程中保持冻结状态。我们的解码器是一个 Transformer 架构，与 Masked AE 设置一致，具有 6 层、12 个头，以及 768 的隐藏维度。特征从 ViT-L 编码器的第 20 和 22 层以及 ViT-B 架构的第 8 和 10 层提取。Q-Former 的架构仅由一个 Transformer 层组成。

Table 1. 我们在 BraTS2021 [2, 3, 41] 上的医学异常检测框架 QFAE 的消融实验结果。我们逐步展示了如何通过加入组件（如 Q-Former 和感知损失），利用现成的模型，将一个简单的 AE 模型提升为一个强大的医学异常检测器。MAE：平均绝对误差。$\mathcal{L}_{\text{Perceptual}}$ ：基于指定模型的感知损失。

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

Q-Former 中可学习的令牌数量由重建补丁大小决定。对于补丁大小为 8×8 像素和输入分辨率为 224×224 ，可学习的令牌数量为 784 （即：$784 = (224/8)^2$ ）。Q-Former 和解码器均使用感知损失训练 300 个时期。超参数在验证集上进行了调整。更多实现细节在补充材料中提供。

评价指标。与之前的工作 [4, 25] 一致，我们报告用于异常检测的受试者工作特征曲线下面积（AUROC）。由于 AUROC 在严重像素级类别不平衡的情况下往往会产生过于乐观的分数，而这在异常检测中很常见，因此不报告定位的 AUROC。

4.2. 消融研究

建立新架构。我们对 Q-Former Autoencoder 的每个组件进行了消融实验，并在 BraTS2021 [2, 3, 41] 数据集上的结果如表 1 所示。为了创建更新的自动编码器架构，我们从使用预训练的编码器和训练解码器的基本方法开始（行 1 ）。我们选择了 DINOv2 ViT-B/14 作为编码器，因为其在零样本任务上取得了卓越的结果。AE 架构仅包含编码器和解码器，没有引入任何瓶颈，通过最小化解码器输入和输出之间的平均绝对误差进行训练。这种基本版本的 AE 仅达到 66.6 的 AUCROC 分数。加入 Q-Former 模块作为瓶颈（行 2 ）使 AUROC 提升了 12.9 （从 66.6 到 79.5 ），显示出 Q-Former 能够保留正常数据的结构，这使其成为异常检测的良好选择。最后，将优化损失从平均绝对误差改为基于 Masked AE 特征计算的感知损失，使性能提高到 86.8 （行 3 ）。通过应用这些设计选择（Q-Former，感知损失），我们将一个简单的 AE 架构从有序的结果发展为一个强大而准确的框架，实现了强劲的性能。

损失函数的影响。我们评估损失函数对检测医疗异常性能的影响，并在表格 2a 中报告结果。单独使用均绝误差或将其与感知损失相结合会产生较差的结果。仅使用感知损失训练（$\mathcal{L}_{\text{Perceptual}}$ ）达到最佳性能，突出显示了基于深度特征的优化优于像素级重构的优势。

聚合的影响。我们进一步评估了不同聚合策略对异常评分计算的影响，结果如表格 2b 所示。将异常评分定义为最大重建误差可以产生最佳性能，这与异常本质上更难重建的直觉相一致。

感知特征的影响。我们分析了从不同的 Masked AE 模型 [22] 层提取的感知特征的效果，并在表 2c 中报告结果。在我们的初步实验中，我们从第 5 层和第 11 层

www.xueshuxiangzi.com

Table 2. 对 BraTS2021 [2, 3, 41] 数据集的消融实验结果展示了我们架构中不同组件的变更效果。感知损失比简单的平均绝对误差 (MAE) 优化，以及取错误的最大值能实现更高的性能。我们还注意到，使用来自感知编码器的多个隐藏层并结合使用较小的解码器补丁大小效果更好。默认配置在 浅蓝色 中被突出显示。$L_{Perceptual}$：基于遮蔽 AE 的感知损失。

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

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

(b) Aggregation in Eq. 3. Selecting the maximum error within Eq. 3 yields top performance.

Function	AUROC
mean	88.5
max	92.6

(c) Layers from the perceptual model . Using multiple layers from the perceptual model achives 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

Table 3. 在 BraTS2021 [2, 3, 41] 上，改变 Masked AE [22] 的补丁大小时，异常检测结果以 AUROC（%）表示。我们注意到，将输入分割成较大的补丁显著提高了性能。最佳结果用粗体显示。默认配置用 淡蓝色 显示。

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

提取特征以指导模型优化，取得了 92.6 的表现。添加另一层进一步提高了性能至 93.0 ，这表明在 AE 训练期间引入额外信号有助于增强异常检测的鲁棒性。

解码器补丁大小的影响。将 Q-Former 架构作为瓶颈，解除了编码器和解码器输出长度之间的依赖关系。因此，解码器可以在不同的粒度（不同的补丁大小）下重建输入。我们评估了不同的补丁大小，如 8×8 、16×16 和 32×32 ，并在表 2d 中报告了结果。正如预期的那样，较小的补丁大小产生更高的性能，使得解码器能够生成更精确的重建。

感知模型补丁大小的影响。用于计算感知损失的特征提取完全独立于框架的编码器和解码器，这使得可以使用多尺度补丁大小来计算感知特征。有趣的是，表格 3 显示较大的补丁可以提高异常检测性能。然而，94.4 AUROC 在 BraTS2021 上的最佳性能是通过结合两个大补丁大小（32×32 和 56×56 像素）来实现的，有效地创建了一个特征金字塔。这一发现表明，较大的补丁大小更好地捕捉数据的结构，更容易发现差异，从而提高异常检测。

编码器的影响。我们评估了不同编码器的组合，包括 DINO [10] 、DINOv2 [43] 、OpenCLIP [52] 和 Masked AE [22] ，并在 BraTS2021 的异常检测结果中报告这些组合的表现，如表 4 所示。在单一的编码器中，DINOv2 [43] 展示了 94.4 AUROC 的最佳性能，强调了其在零样本任务中的强大能力。当将 DINOv2 [43] 与 DINO [10] 结合时，其性能略有提升，达到 94.5 的 AUROC。然而，我们得出结论，这种提升不足以证明增加额外编码器的计算负担是合理的。因此，DINOv2 [43] 被选择为我们框架的默认单一编码器。值得注意的是，Masked AE [22] 编码器表现较差，即使与 DINO [10] 或 DINOv2 [43] 结合使用，其主要原因是其强大的重构能力妨碍了异常的区分。

Table 4. 在 BraTS2021 [2, 3, 41] 上，不同预训练编码器下异常检测的 AUROC（%）结果。值得注意的是，Masked AE [22] 编码器由于其重建输入的能力而表现不佳。DINO [10] 和 DINOv2 [43] 均取得了强劲表现。默认配置在 浅蓝色 中突出显示。

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

DINO [10] 、DINOv2 [43] 和 OpenCLIP [52] 的结果表明，基础模型对于异常检测是有效的，即使是在

www.xueshuxiangzi.com

Table 5. 对 BraTS2021 [2, 3, 41]、RESC [23] 和 RSNA [56] 的异常检测性能（平均值 + 标准差）。结果是基于五次实验重复得出的，*：表示仅进行了三次重复。最优结果用粗体表示。我们的方法能够超过所有方法，在所有三个数据集上取得最先进的性能。

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

医学领域中也是如此。

我们将我们的框架 QFAE 与几个最新的方法在 BraTS2021 [2, 3, 41]、RESC [23] 和 RSNA [56] 上进行了比较，并在表 5 中展示了结果。我们报告了每个实验从 5 次独立运行中获得的结果的平均值和标准差 (std)。

我们的方法在所有数据集上都达到了最先进的性能。特别是，在 BraTS2021 数据集上，我们的框架实现了 AUROC 为 94.3 ± 0.18，达到与之前表现最佳的方法（由 P-VQ [29] 达到的 94.3 ± 0.23）相当的水平。这个结果表明，通过简单地增强标准 AE 框架，我们在脑成像中的异常检测能力非常强。

此外，我们的方法在所有三个数据集上的表现均优于所有基线。在 RESC 数据集上，我们取得了最高的 AUROC 分数 91.8 ± 0.55，超过了之前的最先进结果 PatchCore [46]（91.6 ± 0.10）。在 RSNA 数据集上，QFAE 取得了 AUROC 分数 83.8 ± 0.46，优于下一个最佳方法 CS-Flow [47]（83.2 ± 0.46）。

这些顶级结果突出了我们框架在多种医学成像模式（MRI、X 光和 OCT）中的稳健性。此外，这项工作进一步强调了基础模型，主要是在自然图像上训练的，也可以成功地应用于不同领域，如医学图像，而无需额外的微调。

我们在图 3 和图 4 中展示了定性结果。我们在图

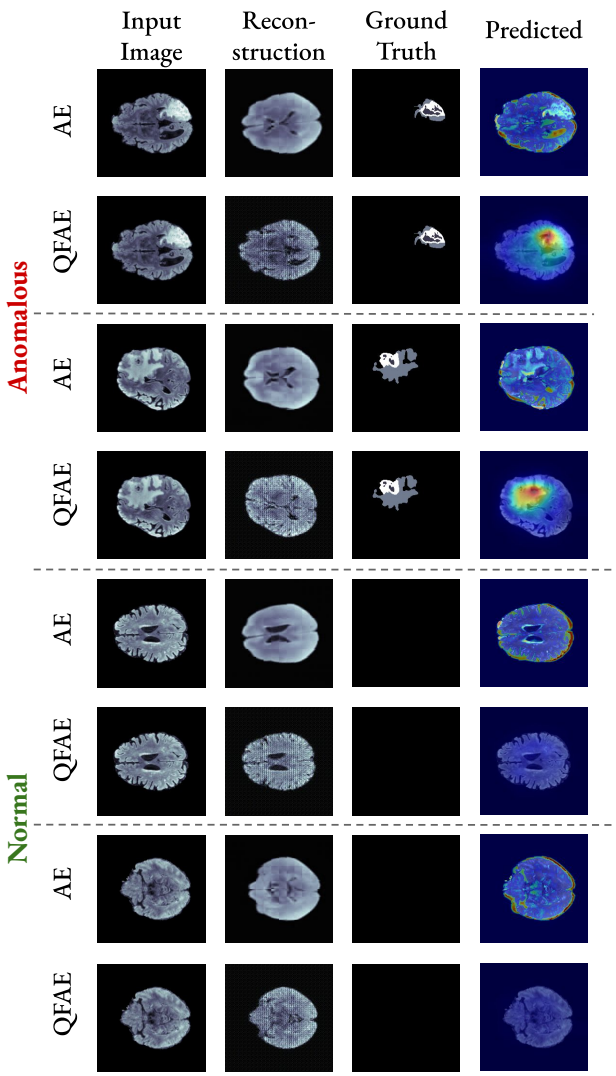

Figure 3. 对几个来自 BraTS2021 [2, 3, 41] 数据集的样本进行异常定位的定性示例。对于每个样本，我们展示了原始输入、重构结果、真实值和预测的异常图。正常和异常样本均被展示。我们的 Q-Former 自编码器（QFAE）与传统自编码器（AE）进行比较。值得注意的是，与基准相比，我们的 QFAE 方法总是能产生更清晰、更准确的异常定位结果，与真实值紧密一致。此外，我们的 QFAE 为正常样本预测了极低的异常分数，能够正确识别为正常样本。

4 中展示了来自 BraTS2021 [2, 3, 41] 和 RESC [23] 数据集的样本，以及输入图像、真实异常掩码、由几种最新技术和我们的 QFAE 框架预测的异常图。在这两个数据集中，我们的框架精准地定位了异常。值得注意的是，在第二个 BraTS2021 样本中，大多数方法都难以定位异常，只有 MVFA-AD [25]（小样本）、DRAEM [60] 和我们改进的 AE（QFAE）在评估的 10 种方法中实现了正确定位。这表明我们的方法能够准确识别微妙且难以检测的异常。

此外，在图 3 中，我们展示了我们的 QFAE 与传统 AE 的异常结果的定性比较。传统 AE 使用了一个

www.xueshuxiangzi.com

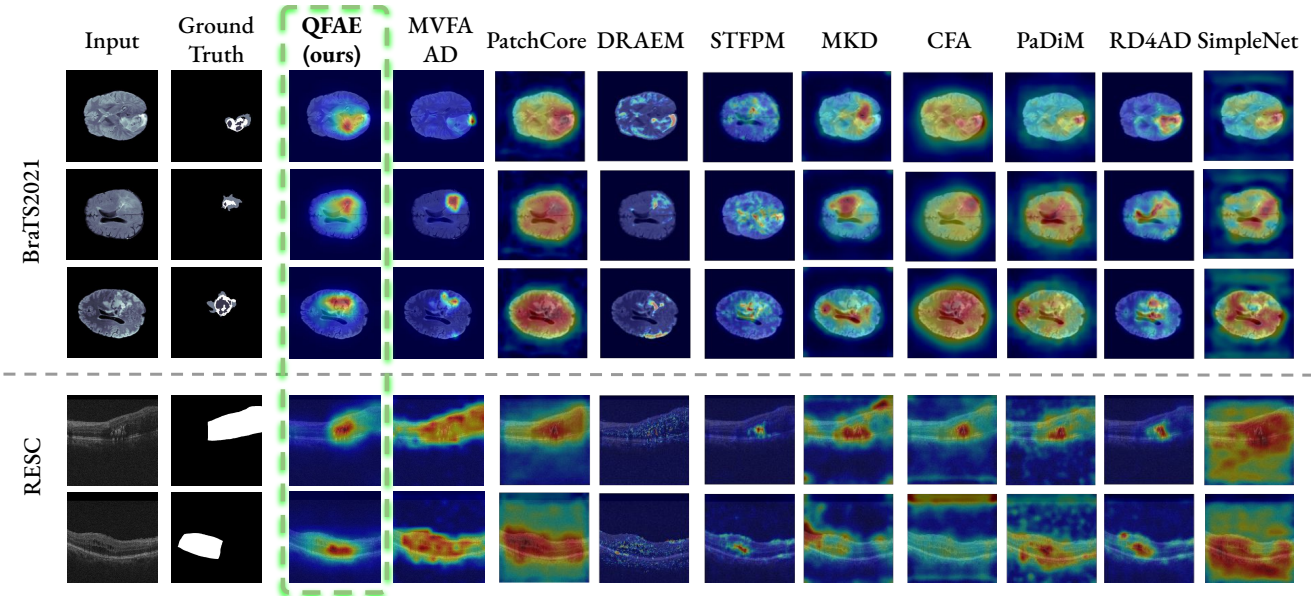

Figure 4. 对来自 BraTS2021 [2, 3, 41] 和 RESC [23] 数据集的几个样本进行异常定位的定性示例。对于每个样本，各列显示原始输入、真实值异常掩码以及由 QFAE（我们的方法）预测的异常图，与各种基线方法并列。我们注意到，MVFA-AD [25] 使用小样本策略，因此不像包括我们的方法在内的其他方法是无监督的。基线方法的预测异常直接从 BMAD [4] 裁剪出来。值得注意的是，与其他方法相比，我们的 QFAE 方法始终产生更锐利和更准确的异常定位，与真实值密切吻合。

预训练的编码器和解码器。我们观察到，我们增强的 AE 预测的异常与真实值有很好的相关性，同时对正常样本的异常分数也非常低。这些结果清楚地表明，使用 Q-Former 作为瓶颈在检测和定位异常方面是有效的。此外，这些发现表明，将 Q-Former 作为瓶颈与利用 Masked AE 的感知损失相结合的方法在增强医学异常检测性能方面非常有效。

在本文中，我们介绍了 Q-Former 自动编码器（QFAE），这是一种现代化的自动编码器框架，利用最先进的预训练视觉基础模型的力量来进行医学异常检测。我们的框架通过集成冻结的预训练编码器（DINO [10] 、DINOv2 [43] 和 OpenCLIP [52] ）来进行稳健的特征提取，使用可训练的 Q-Former 作为动态瓶颈，从变长的上下文输入中生成固定长度的潜在代码，并利用感知损失函数进行语义上有意义的重建，从而解决了传统自动编码器的关键限制。我们在四个不同的医学异常检测基准上对 QFAE 进行了严格评估：BraTS2021、RESC、RSNA 和 LiverCT。我们的结果在这些数据集上持续表现出最先进的性能，实现了优越的 AUROC 分数和精确的异常定位。我们的工作强调了大规模预训练视觉基础模型（最初在自然图像上训练）的成功和稳健应用，用于专业医学成像领域的无监督异常检测，特别是不需要广泛的微调。在未来的工作中，我们计划将 QFAE 应用于多类别医学异常检测。尽管表现优异，我们提出的 QFAE 框架有某些限制。虽然使用如 DINO、DINOv2、Masked AE 等预训练基础模型提高了泛化能力并减少了训练时间，但其固有地限制了模型学习领域特定特征的能力。尽管我们的框架在不同的模式和数据集上取得了一贯良好的结果，但我们不能声称它能推广到所有异常类型或不同复杂程度的输入。

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.

[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.

[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.

[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.

[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 Applica-

www.xueshuxiangzi.com

tions (VISIGRAPP 2019) - Volume 5: VISAPP, pages 372–380. INSTICC, SciTePress, 2019.

[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.

[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.

[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.

[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].

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[17] Mariana-Iuliana Georgescu. Masked Autoencoders for Unsupervised Anomaly Detection in Medical Images. Procedia Computer Science, 225:969–978, 2023.

[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.

[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.

[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.

[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.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, 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.

[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.

[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.

[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.

[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.

[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.

www.xueshuxiangzi.com

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[37] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: 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.

[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.

[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.

[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.

[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.

[42] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021.

[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.

[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.

[45] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022.

[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.

[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.

[48] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep

www.xueshuxiangzi.com

one-class classification. In International conference on machine learning, pages 4393–4402. PMLR, 2018.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[58] Shinji Yamada and Kazuhiro Hotta. Reconstruction student with attention for student-teacher pyramid matching. arXiv preprint arXiv:2111.15376, 2021.

[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.

[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.

[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.

[62] Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. AE-FLOW: Autoencoders with Normalizing Flows for Medical Images Anomaly Detection. 2022.

[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.

www.xueshuxiangzi.com

Q-Former 自编码器：一种用于医学异常检测的现代框架

Supplementary Material

我们在第 5 节提供了额外的实现细节，在第 6 节和第 7 节提供了关于 LiverCT 和 RSNA 的额外实验。

5. 实现细节

本节概述了我们所提出的框架的实现细节，确保我们的结果能够完全重现。所有实验都是在 PyTorch 中进行的。

用于训练和评估的主要超参数分别详见表 6 和表 7 。

Table 6. 实验训练超参数。

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
	Final Projection Out-Features	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	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. 实验评估配置。

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

5.1. 感知损失公式

训练目标是最小化多尺度感知损失。该损失通过三步过程计算：

步骤 1：特征提取。对于输入图像 x 及其重建图像 $\tilde{x}$ ，我们从一组预训练的感知模型中提取特征图。我们使用多个遮掩自编码器 (遮掩 AE) 模型，每个模型通过其输入补丁大小 $p \in P$ 加以区分。对于每个模型，我们从一组变换器块 $i \in I$ 中选择特征。设 $\Phi_{i,p}(x)$ 为从补丁大小为 p 的感知模型的第 i 层提取的形状为 $C_i \times H_i \times W_i$ 的特征图。

步骤 2：异常图计算。对于每个选定的特征图，我们通过计算原始图像和其在每个空间位置处的重建特征之间的余弦距离来计算一个中间异常图 $A_{i,p}$ 。

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

这生成了一组单通道的异常图，每个图对应于层 i 和补丁大小 p 的组合。

步骤 3：分层聚合与最终损失。最终损失是通过两阶段的分层聚合计算得出的。首先，对于每个特征层 $i \in I$ ，我们通过将来自所有不同补丁尺寸模型的相应异常图进行元素级乘法，创建一个鲁棒的、层特定的异常图 $A_{\text{combined},i}$ 。这一步在每个特征层级上强制施加多个尺度的一致性要求。接着，计算总损失 $\mathcal{L}$ 是通过对这些鲁棒的、层特定图的平均值进行平均来实现的。这将每个特征层的误差信号视为对总损失的独立贡献。对于训练，我们使用补丁尺寸 $P = \{32, 56\}$ 并从 Masked AE ViT-Large 编码器的第 16 和第 20 个转换器块中提取特征。

在评估过程中，我们同时生成用于计算 AUROC 的图像级标量评分和像素级异常图。两者都起始于同一组中间异常图 $A_{i,p}$ ，虽然是使用评估配置计算的（表 7 ）。设这个评估集的图为 $\mathcal{A} = \{A_1, A_2, ..., A_N\}$ 。

为了为每张图像得出一个单一的标量得分，我们执行了两步聚合：

步骤 1：空间聚合。对于每个异常图 $A_n \in \mathcal{A}$ ，我们找出最大像素值。这个值，s_n ，代表了由该特征图检测到的最严重的重建误差。

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

步骤 2：跨特征聚合。最终的图像级得分 A_{score} 是这些最大值的平均值，在所有 N 特征图上进行平均。

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

这种方法提供了一个稳健的得分，对强局部异常有敏感性，同时受益于来自不同层次的特征的多样性。

为了生成最终的二维异常图，我们使用一种不同的聚合策略来保留空间信息。在每个空间位置 (j,k) ，我们取所有 N 调整尺寸的异常图的平均值。

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

5.2. 训练和数据增强

使用带有 OneCycleLR 学习率调度器的 Adam 优化器训练模型。为了鼓励模型学习正常数据的鲁棒且可泛化的表示，对训练集应用以下数据增强：

www.xueshuxiangzi.com

- 随机调整大小裁剪：图像被裁剪成随机大小（为原始大小的 90 % 到 100 %）和长宽比（为原始长宽比的 80 % 到 120 %），然后调整为最终的输入尺寸。
- 随机旋转：图像被随机旋转一个在-10 度和 +10 度之间的角度。
- 随机垂直翻转：图像以 50 % 的概率进行垂直翻转。
- 颜色抖动：图像的亮度和对比度被随机调整达 0.1 倍。
- 归一化：图像像素值被归一化为均值为 0.449 和标准差为 0.226。

6. 肝脏 CT 实验

我们在 LiverCT [6, 31] 基准数据上进行了几步预处理。在本节中，我们逐一介绍这些技术，并完成表格 8

6.1. 数据预处理

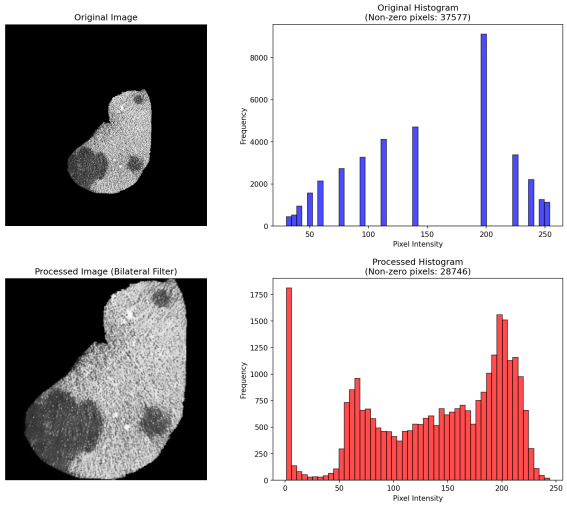

Figure 5. 来自 LiverCT [6, 31] 的原始图像和处理后的图像及其像素直方图。

首先，数据集中图像的尺寸为 512x512 像素，只有一小部分图像包含肝脏区域。将这些图像调整为 224x224 像素，即我们模型的输入尺寸，会导致肝脏部分变小。为了解决在调整大小时不丢失感兴趣区域（即肝脏部分）细节的问题，我们使用了以下算法，将图像调整为 224x224。请注意，此过程是全自动的，并且可以应用于任何分割的肝脏图像。

1. ROI 识别：对于每个 512×512 输入图像，我们首先识别包含肝脏的区域。这是通过计算一个紧密包围所有非零像素的并集边界框来实现的。
2. ROI 裁剪：图像使用计算出的边界框坐标进行裁剪，从空白背景中分离出肝脏部分。
3. 画布准备：创建一个新的、黑色的、目标尺寸为（224×224 ）的画布作为最终模型输入的背景。
4. 条件调整大小和放置：裁剪后的肝脏区域感兴趣区（ROI）采用依赖于大小的策略放置到画布上：
   - 如果 ROI 小于或等于 224×224 ：裁剪的部分直接粘贴到画布的中心，而无需任何调整大小。这保留了肝脏组织的本机分辨率。
   - 如果 ROI 大于 224×224 ：该区域将调整大小以适应 224×224 框架，同时保持其原始的纵横比以防止失真。调整大小后的 ROI 随后在画布上居中。
5. 最终输入：将肝脏段显著居中的 224×224 图像作为模型的输入。

关于这个数据集的另一个问题是，由于计算机断层扫描成像本身的限制，它进行了几种窗口调节和直方图均衡化技术 [4, 7, 34] 。结果，这些图像可能不在我们所用感知损失模型训练的标准数据集（如 ImageNet）的分布内。为了缓解这个问题，我们在将每幅经过处理的 224×224 图像输入网络之前，对其应用双边滤波器 [55] 。图 5 展示了预处理的效果，其中保留了感兴趣区和异常区域，并且图像的直方图看起来更加自然。

使用这个新的预处理算法重新训练和重新评估模型得到了表 8 中第 2 行的结果。该结果是使用 5 个不同种子（42, 7, 13, 65, 91）训练的 5 个不同模型的评估均值和标准差。

Table 8. 在 LiverCT 数据集上的消融研究。

	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] $->$ [8, 16]	65.5 ± 1.96

6.2. 新的评估配置

如图 6 所示，新的数据预处理流程（第二列）在异常图的质量方面比原始配置（第一列）有所改进，并使异常区域更明显。然而，预测的异常图仍未能捕捉到异常区域中的纹理变化。根据视觉感知研究 [42, 45] ，这些研究指出较小的补丁尺寸更偏向于纹理，而较大的补丁尺寸更偏向于形状，我们将感知模型用于异常分数计算的补丁大小从 [16, 32, 56] 改为 [8, 16]。从表格 6 的第三列可以看出，使用此评估配置后，异常图在异常区域上更好地捕捉到纹理变化。这反映在表格 8 的 3 行的 AUROC 分数上。在 LiverCT 上获得最佳结果的评估配置如表格 9 所示，修改的部分以粗体突出显示。训练配置保持不变。

Table 9. 肝脏 CT 的最佳评估配置。

Component	Parameter	Value
General	Batch Size Test Data Augmentation	64 None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model Distance Metric Layers Used for Feature Extraction Multi-Scale Input Patch Sizes	MAE with ViT-Large Encoder Cosine Distance From the 12th, 16th, and 20th transformer blocks 8x8x, 16x16
Image-Level Score Aggregation	Spatial Aggregation per Feature Map Cross-Feature Map Aggregation	Max Mean
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

www.xueshuxiangzi.com

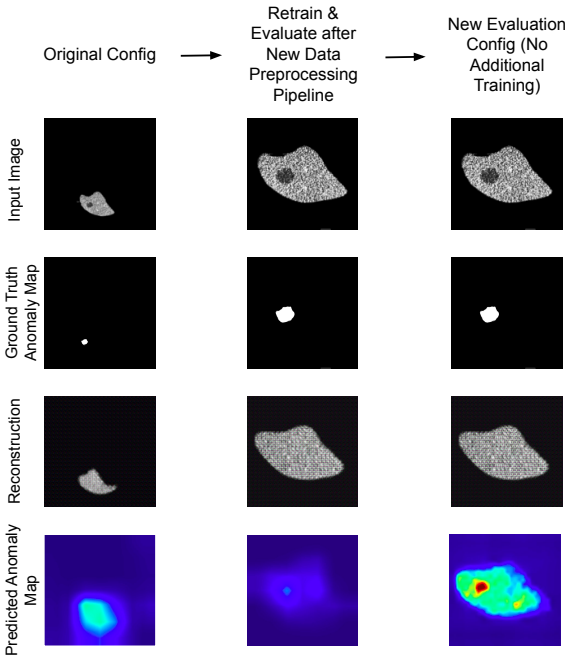

Figure 6. 数据预处理流程和评估配置等修改的效果。我们首先避免在调整大小过程中减小异常区域。然后，根据视觉感知文献的见解，配置感知损失，使其更偏向于文本线索。

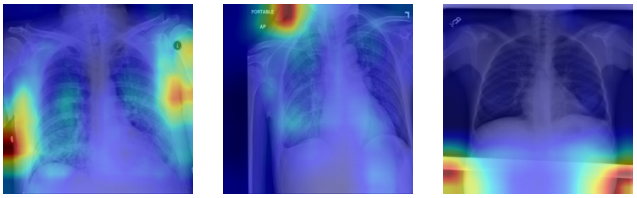

Figure 7. 光学特征和伪影主导了来自异常区域的响应。

7. 胸部 RSNA 的不同聚合

在胸部图像上，通常可以看到不同的光学特征和伪影。它们的位置是变化的。当这些伪影出现时，它们会主导异常信号，使得无法根据主方程 4 中描述的聚合方法正确计算异常分数。由于它们的位置是变化且不可预测的，我们无法设计出一个预处理算法。

为了缓解这个问题，我们决定在 Chest RSNA 数据集的验证集上尝试不同的聚合方法。作为替代方案，我们首先尝试在异常图的每个位置取平均值，然后在不同层之间取最大值。我们观察到 AUROC 从 78.6 % 增加到验证集上的 84.3 %。因此，我们决定保留这种方法，并在测试集上报告了 83.8 % 的 AUROC，如主表 5 所示。在整个 Chest RSNA 上取得最佳结果的评估配置在 10 中展示，修改的部分以粗体显示。训练配置保持不变。

Table 10. 在 Chest RSNA 上的最佳评估配置

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

8. 每个数据集的 SOTA 结果

Table 11. 大脑 MRI 的最佳训练配置。

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

如表 12 所示，我们在 BraTS2021 [2, 3, 41]、RESC [23] 和 RSNA [56] 上达到了最先进的性能，并在 LiverCT [6, 31] 上排名第二。

www.xueshuxiangzi.com

Table 12. 在 BraTS2021、肝脏 CT（BTCV + LiTs）、RESC 和 RSNA 上的异常检测性能（平均值 + 标准差）。结果是对该实验的五次重复进行报告的。* ： 表示仅进行了三次重复。最佳结果用粗体显示。

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	$\mathbf{69.2 \pm 3.86}$	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)	$\mathbf{94.3 \pm 0.18}$	65.5 ± 1.96	$\mathbf{91.8 \pm 0.55}$	$\mathbf{83.8 \pm 0.46}$

www.xueshuxiangzi.com