Information Fusion 133 (2026) 104341

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Towards reliable medical unsupervised anomaly detection: A benchmark and dataset for PET/CT with cross-modality knowledge distillation network

Muhao Xu [a,1], Lili Qu [b,1], Zihan Nie [a], Feng Li [c], Zhuangzhuang Chen [d], Qi Liao [a], Yi Wan [a], Sijie Niu [e], Runmin Cong [f], Xin Li [b,*], Weiye Song [a,*]

[a] Department of Mechanical Engineering, Key Laboratory of High Efficiency and Clean Mechanical Manufacture of Ministry of Education, Shandong University, Jinan, 250061, China
[b] Department of Nuclear Medicine, Qilu Hospital of Shandong University, Jinan, 250012, China
[c] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[d] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[e] School of Information Science and Engineering, University of Jinan, Jinan, China
[f] School of Control Science and Engineering,Shandong University, Jinan, China

ARTICLE INFO

Keywords:
Unsupervised anomaly detection
PET/CT
Knowledge distillation
Cross-modality

ABSTRACT

Unsupervised anomaly detection (UAD) in medical imaging has emerged as a promising approach to identify pathological deviations without the need for pixel-level annotations. While most existing UAD methods focus on single-modality data, clinical diagnosis increasingly relies on multi-modality imaging, particularly hybrid PET/CT scans, which combine metabolic and anatomical information for enhanced lesion detection. However, no benchmark dataset or dedicated method exists for UAD using PET/CT image pairs. To bridge this gap, we construct the first PET/CT dataset for unsupervised thoracic anomaly detection, leveraging the complementary strengths of PET in capturing functional metabolic activity and CT in providing high-resolution anatomical structure. Furthermore, we propose a Cross-Modality Anomaly Detection network that extends distillation-based unsupervised anomaly detection to paired PET/CT imaging. The framework follows a reverse knowledge distillation paradigm, where stable modality-aware representations are learned from normal data and reconstructed across modalities to reveal abnormal deviations. Instead of explicit patch-wise cross-attention, the proposed model introduces a token-mediated multimodal interaction mechanism that encourages global cross-modality consistency. A dedicated global refinement strategy is further employed during decoding to enhance contextual feature aggregation and improve robustness to complex anomaly patterns. Extensive experiments on the proposed PET/CT dataset and multiple public benchmarks demonstrate consistent performance gains over strong single-modality baselines, highlighting the effectiveness of modeling normality-driven cross-modality consistency for multi-modal medical anomaly detection. Our code is available at: https://github.com/Xmh-L/TRMUAD

1. Introduction

Unsupervised anomaly detection (UAD) [1–3] has attracted increasing attention in medical imaging as a data-efficient framework for characterizing deviations from normal imaging patterns, particularly in scenarios where exhaustive expert annotations are impractical or unavailable. By learning the distribution of normal anatomical structures during training, UAD methods enable the automatic discovery of abnormalities during inference, thereby addressing the inherent challenges of annotation scarcity and inter-patient heterogeneity. Current UAD approaches can be broadly categorised into three groups: reconstruction-based [4–7], representation-based [8,9], and knowledge distillation-based strategies [10–12].

While reconstruction-based methods typically rely on reconstruction errors between inputs and outputs, they often suffer from identity mapping, which weakens anomaly sensitivity. Representation-based

* Corresponding authors.
E-mail addresses: 202420710@mail.sdu.edu.cn (M. Xu), lili.qu@sdu.edu.cn (L. Qu), 202414371@mail.sdu.edu.cn (Z. Nie), fengli@hfut.edu.cn (F. Li), eezzchen@ust.hk (Z. Chen), 202434508@mail.sdu.edu.cn (Q. Liao), wanyi@sdu.edu.cn (Y. Wan), sjniu@hotmail.com (S. Niu), rmcong@sdu.edu.cn (R. Cong), lixin16@sdu.edu.cn (X. Li), songweiye@sdu.edu.cn (W. Song).
[1] Co-first author

https://doi.org/10.1016/j.inffus.2026.104341
Received 18 September 2025; Received in revised form 2 March 2026; Accepted 30 March 2026
Available online 1 April 2026

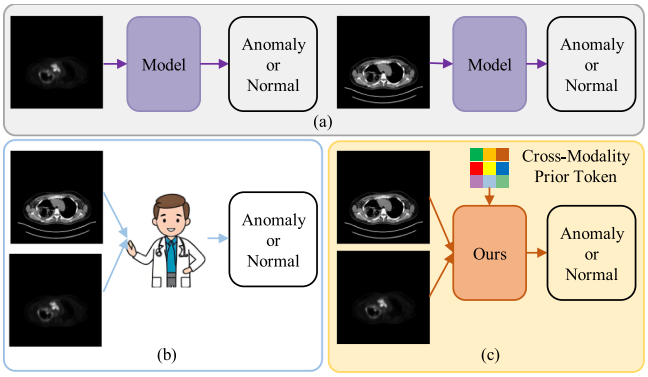

Fig. 1. Comparison of anomaly detection paradigms. (a) Existing unsupervised methods typically rely on single-modality inputs (either PET or CT), which limits diagnostic completeness. (b) In clinical practice, physicians integrate PET and CT information simultaneously to form a holistic judgment of anatomical and functional abnormalities. (c) Our proposed method mimics this clinical reasoning process by jointly modeling co-registered PET/CT pairs, achieving anomaly detection that is more consistent with expert diagnostic practice.

approaches extract compact embeddings of normal anatomy, yet they frequently struggle to capture both local and global dependencies. More recently, knowledge distillation frameworks [13,14] have shown notable improvements by detecting discrepancies between teacher and student networks trained solely on normal data. The efficacy of these techniques has been demonstrated in a range of single-modality datasets, including magnetic resonance imaging (MRI), chest radiography (X-ray), and computed tomography (CT).

However, existing research on unsupervised anomaly detection in medical imaging has remained almost exclusively limited to single-modality datasets, such as CT, MRI, or other modalities in isolation. This constraint restricts algorithmic advances to within-modality modeling and overlooks the complementary nature of functional and anatomical information [15,16]. In clinical practice, it is rare for a single imaging modality to be used for the purpose of disease diagnosis. Conversely, multi-modality imaging has become increasingly indispensable, particularly the combination of Positron Emission Tomography (PET) [17] and Computed Tomography (CT) [18]. Positron emission tomography provides functional and metabolic insights by capturing the spatial distribution of radiotracers, whereas computed tomography offers high-resolution structural information with precise anatomical localisation. The integration of these modalities in hybrid PET/CT imaging enables simultaneous assessment of both metabolism and anatomy, thereby offering complementary and synergistic diagnostic information. From a clinical perspective, its capacity to co-register metabolic activity with anatomical structures provides diagnostic insights that are unattainable with either modality alone, as illustrated in Fig. 1. To the best of our knowledge, no benchmark or method has been specifically developed for unsupervised anomaly detection with paired PET/CT data. The establishment of a dedicated PET/CT UAD dataset would enable the development of more robust and clinically relevant multi-modal UAD algorithms, enhancing both sensitivity and specificity by more accurately reflecting the complex, information-rich diagnostic scenarios encountered in real-world chest disease evaluation.

To establish a realistic benchmark for multi-modal unsupervised anomaly detection in medical imaging, we construct the first paired PET/CT dataset specifically tailored for thoracic anomaly detection. Building upon this dataset, we introduce a multimodal anomaly detection framework that extends distillation-based feature reconstruction methods to the PET/CT setting by explicitly modeling cross-modality consistency under normal conditions. At its core, the network leverages a reverse knowledge distillation strategy, where a frozen encoder ensures stable modality-aware feature extraction while paired decoders collaboratively reconstruct cross-modality representations. To effectively fuse heterogeneous modalities, we propose a Cross-Modality Hybrid (CMH) module that enables token-mediated multimodal interaction between PET and CT through a learnable Cross-Modality Prior Token (CMPT). Rather than performing explicit transformer-style patch-wise cross-attention, the CMPT is learned exclusively from normal data and acts as a spatially invariant modality-bridging representation, encouraging global cross-modality consistency instead of explicit spatial or semantic alignment. Beyond local semantic fusion, we further introduce a Global Fusion Refinement (GFR) module, which facilitates dense multi-level interactions and captures long-range dependencies via the Selective Scan 2D (SS2D) mechanism. This global refinement process enhances contextual aggregation during reconstruction, improving robustness to complex anomaly patterns without relying on explicit cross-modal alignment. Together, these innovations form a unified architecture that explicitly addresses both local semantic alignment and global contextual reasoning, two critical factors for accurate abnormality detection in multi-modal settings. Extensive experiments confirm that the proposed framework consistently outperforms strong single-modality baselines, demonstrating the practical value of modeling normality-driven cross-modality consistency for PET/CT-based anomaly detection.

Our contributions are summarized as follows:

- We construct the first paired PET/CT dataset specifically designed for unsupervised anomaly detection, providing a clinically realistic benchmark for multimodal modeling of thoracic abnormalities.
- We propose a Cross-Modality Anomaly Detection framework that extends distillation-based feature reconstruction methods to the PET/CT setting, employing a reverse knowledge distillation strategy to model normality-driven cross-modality consistency.
- We introduce a Cross-Modality Hybrid module, which uses a learnable CMPT to mediate token-level multimodal interaction under a global consistency constraint and a Global Fusion Refinement module, which enhances multi-level feature integration and long-range dependency modelling to improve robustness to complex anomaly patterns.

The remainder of this paper is organised as follows. Section 2 reviews related work on unsupervised anomaly detection and existing medical imaging benchmarks. Section 3 describes the construction of the proposed PET/CT dataset. Section 4 details the architecture of the proposed method. Section 5 presents extensive experimental results, including quantitative comparisons and qualitative analyses. Section 6 provides ablation studies and further discussion. Finally, Section 7 concludes the paper.

2. Related work and dataset

Unsupervised anomaly detection is a one-class classification framework where models are trained solely on normal examples and identify anomalies in test samples by measuring deviations from this normative distribution. This approach is particularly advantageous in medical imaging, where accurately annotated pathological data is scarce, costly, or unavailable. Existing UAD methods can be broadly categorized into reconstruction-based, representation-based, and knowledge distillation-based approaches.

2.1. The reconstruction-based methods

Reconstruction-based methods [4,5] posit that models trained exclusively on healthy data will struggle to accurately reconstruct regions of abnormality. These methods usually involve training autoencoders (AEs) [24], variational autoencoders (VAEs) [25], generative adversarial networks (GANs) [26], or masked autoencoders [27] in order to learn the distribution of normal anatomy. At inference time, anomalous regions are identified based on elevated reconstruction error,

Table 1
Comparison of PET/CT datasets and public datasets for unsupervised anomaly detection in medical imaging.

Benchmarks	Originations	Total	Train	Test	Sample size	Annotation Level	modality
Retinal OCT	Oct-17 [19]	27,283 images	26,315	968	512*496	Image label	Single
Pathology	HIS [20]	7085 patches	5088	1997	256*256	Image label	Single
Color Fundus Photography	APTOS 2019 [21]	3662 image	1000	2662	256*256	Image label	Single
Retinal OCT	RESC [22]	6102 image	4297	1805	512*1204	Segmentation mask	Single
Brain MRI	BraTS2021 [23]	11,215 slices	7500	3715	240*240	Segmentation mask	Single
Chest PET/CT	Our	10,810 pairs	4,456*2	6,354*2	512*512	Image label	Multiple

under the premise that abnormalities are reconstructed poorly. For example, Schlegl et al. introduced f-AnoGAN [6], a GAN-based architecture that detects lesions by comparing pixel-wise reconstructions. However, it suffers from the identity mapping issue, whereby anomalies may be reconstructed as accurately as normal tissue. It is also computationally inefficient due to its reliance on iterative latent-space optimisation.

2.2. The representation-based methods

Representation-based approaches in unsupervised anomaly detection focus on learning discriminative feature embeddings from normal data distributions [9,28,29]. At inference time, anomalies are identified by quantifying the distance between test-time features and the distribution of embeddings observed during training. This paradigm includes methods based on self-supervised learning, contrastive learning strategies, and memory bank retrieval mechanisms. PatchCore, introduced by Roth et al. [8], exemplifies this category by leveraging patch-level descriptors extracted via pre-trained convolutional backbones. These descriptors are compared to a memory of normal features using nearest-neighbour search, demonstrating state-of-the-art performance on industrial benchmarks such as MVTec-AD [30]. Despite its success, PatchCore and related methods often face limitations in handling medical data, where fine-grained local cues and long-range dependencies are critical, such as in optical coherence tomography or volumetric scans. CFlow-AD [31] constructs hierarchical feature pyramids and extends normalising flows to estimate the likelihood of individual feature vectors for pixel-wise anomaly localisation. However, its per-vector independence assumption can lead to spatial incoherence in localisation maps. Another thread of work explores perturbation and synthesis strategies in the latent space. SimpleNet [32] introduces adversarial noise directly into the feature space to reinforce compactness in the normal distribution boundary.

2.3. Distillation-based methods

Knowledge distillation approaches have become increasingly popular in UAD due to their robustness and reliable performance [33]. In this setting, a teacher network pretrained exclusively on normal images guides a student network to replicate its feature-level outputs. During inference, discrepancies between the student's and teacher's responses serve as indicators of anomalous regions. Bergmann et al. proposed the uninformed student-teacher architecture [10,12], where multiple student networks are trained to regress the feature outputs of a descriptive teacher pretrained on natural image patches. Anomalies are flagged where the students' regression fails, with predictive uncertainty further aiding localisation accuracy. While effective, this approach may falter when anomalies are subtle, as minor feature deviations can be challenging to distinguish. In order to enhance sensitivity to varied anomaly scales, Salehi et al. introduced a multi-resolution distillation technique [12]. This method leverages intermediate activations at multiple levels to enrich the feature discrepancy signal between teacher and student networks. Such multi-layer supervision improves detection performance over methods relying solely on final-layer features. More advanced architectures, such as Dual-Student Knowledge Distillation [34] and Heterogeneous Auto-Encoder [35] implement paired student models with complementary structures, designed to amplify distinction in anomalous regions while maintaining consistency on normal data. Additionally, scale-aware contrastive reverse distillation [13] strategies integrate contrastive learning into the distillation framework and adjust feature modelling across different resolutions, leading to superior anomaly discriminability and localisation precision.

2.4. Unsupervised anomaly detection in PET/CT imaging

While UAD methods have demonstrated strong performance on single-modality datasets, such as brain MRI [36] and retinal OCT [19], their extension to multi-modal medical imaging remains underexplored. In particular, despite the widespread clinical adoption of PET/CT for thoracic and chest disease assessment, there exists no publicly available dataset or dedicated method for unsupervised anomaly detection on fused PET/CT images. Existing multimodal or cross-modal anomaly detection frameworks primarily focus on supervised or weakly supervised settings and commonly rely on explicit spatial alignment or patch-wise cross-attention for modality fusion [37,38]. This is a significant gap, as PET provides functional information while CT offers anatomical detail; jointly leveraging both is crucial for identifying subtle or ambiguous abnormalities that may go undetected in isolation. Therefore, developing UAD methods specifically tailored to PET/CT image pairs and constructing benchmark datasets in this domain is not only timely but essential. Such efforts would enable more accurate, robust, and clinically aligned anomaly detection in real-world chest disease diagnosis and management.

2.5. Related datasets

Unsupervised anomaly detection has been extensively studied on a variety of medical imaging datasets spanning different modalities and clinical applications. Most existing benchmarks are single-modality in nature and are primarily designed to evaluate anomaly detection or segmentation performance within a specific diagnostic domain.

Several representative datasets that have been widely used in prior UAD studies are summarized in Table 1. These include OCT-17, which consists of 27,283 retinal OCT images for anomaly detection; HIS, a histopathology dataset comprising 7085 image patches; APTOS 2019, which contains 3662 fundus photographs for diabetic retinopathy assessment; RESC, a retinal OCT dataset with 6102 images annotated with pixel-level segmentation masks; and BraTS2021, a large-scale brain MRI benchmark including 11,215 slices with lesion segmentation labels. These datasets differ in imaging modality, annotation granularity, and pathological characteristics; notably, although OCT-17 and RESC are both retinal OCT datasets, they are acquired using different imaging devices and target distinct disease categories with different annotation levels, together forming complementary evaluation benchmarks for medical UAD methods. In this work, the above five publicly available single-modality datasets are used exclusively for generalisation experiments.

3. Datasets

In this section, we introduce the proposed dataset resources in detail. For completeness, we also report statistics of widely used

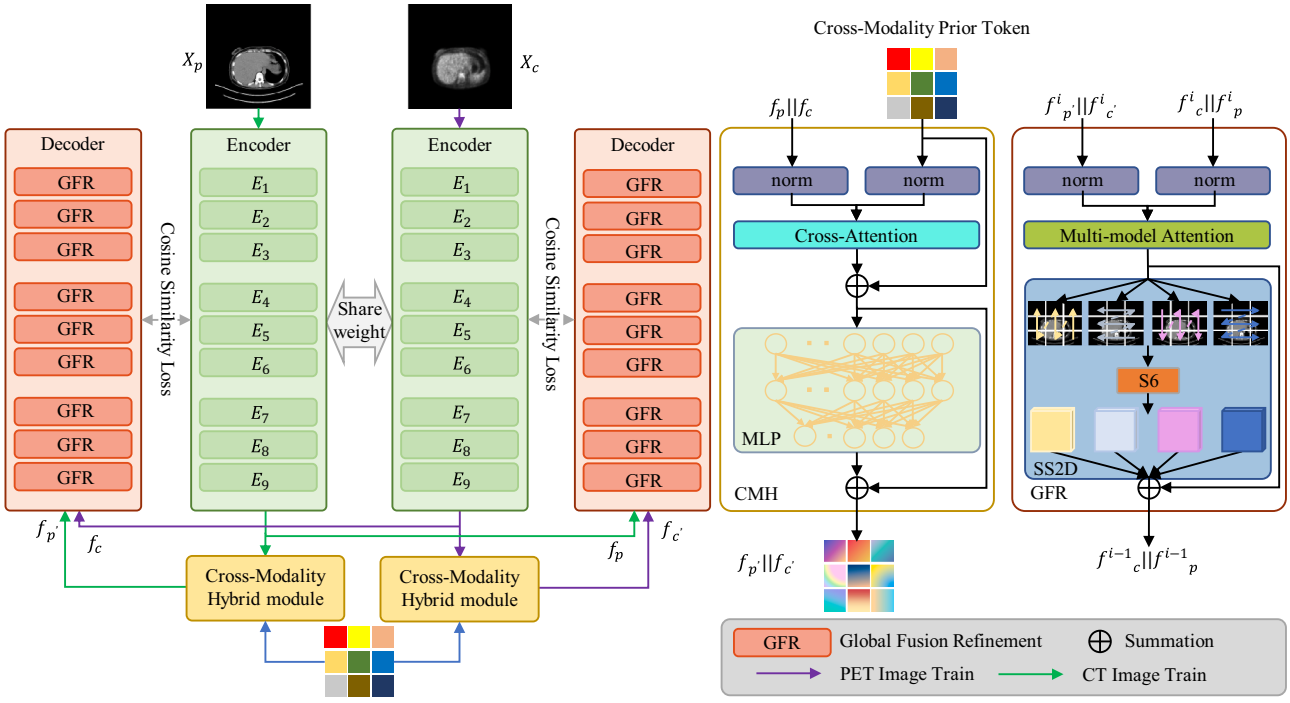

Fig. 2. Overview of the proposed cross-modality anomaly detection (cmad) network. Paired PET and CT images are encoded by a shared frozen encoder and reconstructed by modality-specific decoders under a reverse knowledge distillation framework. cross-modality interaction is enabled by the cross-modality hybrid module with a learnable cross-modality prior token, while global contextual information is refined through the global fusion refinement module.

single-modality benchmarks to facilitate cross-dataset evaluation and demonstrate the generalisability of our method.

3.1. Dataset construction

To address the absence of benchmark resources for multi-modal unsupervised anomaly detection (UAD), we construct a new dataset, termed Chest PET/CT, which represents the first publicly available dataset of co-registered positron emission tomography (PET) and computed tomography (CT) image pairs with image-level annotations. Unlike existing datasets that are predominantly single-modality, Chest PET/CT provides aligned dual-modality scans, enabling the joint modelling of anatomical structure and functional metabolism. This characteristic makes it particularly suitable for advancing anomaly detection methods in clinically realistic thoracic imaging scenarios.

The dataset was collected from 110 patients diagnosed with a spectrum of thoracic abnormalities, including small cell lung carcinoma, lung adenocarcinoma, fibrous hyperplasia, and granulomatous lesions. Imaging was performed on two widely used clinical scanners-GE Discovery STe 16 and Siemens Biograph Vision 600. In total, the dataset contains 10,810 PET/CT pairs. To support the UAD setting, the dataset is split at the slice level within each patient: normal slices are included in the training set, while abnormal slices from the same patients are reserved exclusively for evaluation. This design ensures consistency with UAD protocols while simultaneously reflecting the diagnostic challenges posed by real-world thoracic imaging.

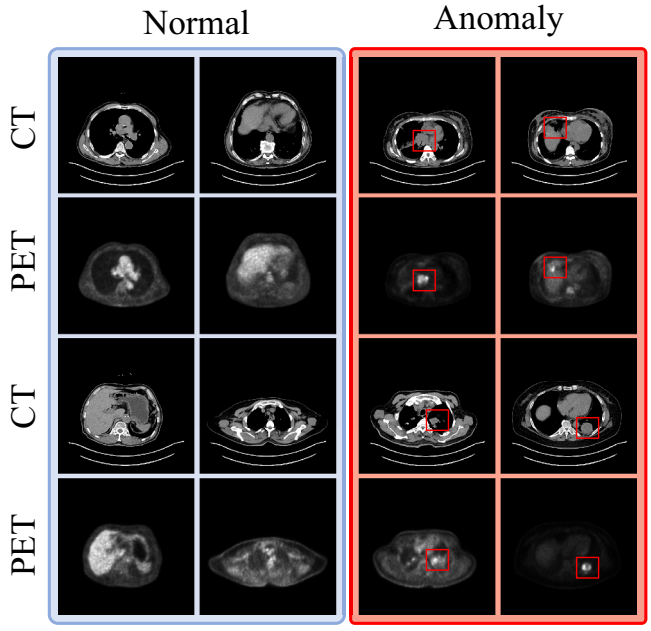

Fig. 3. Sample image of PET/CT dataset, the location marked with the red box is abnormal.

3.2. Dataset statistics

Table 1 summarizes the proposed Chest PET/CT dataset. The dataset consists of paired PET and CT slices acquired from thoracic imaging studies, and is designed specifically for unsupervised anomaly detection in a clinically realistic multimodal setting. Specifically, Chest PET/CT comprises a total of 10,810 paired PET/CT slices, including $4,456 \times 2$ slices for training and $6,354 \times 2$ slices for testing. All samples are annotated at the image level, indicating whether a given PET/CT pair is normal or abnormal. Compared with existing benchmarks, the proposed dataset exhibits three distinctive characteristics: It is, to the best of our knowledge, the first dataset to provide dual-modality PET/CT scans for unsupervised anomaly detection. It covers a broad spectrum of thoracic pathologies, thereby offering high clinical relevance. It provides high-resolution, spatially aligned PET/CT slice pairs, enabling fine-grained investigation of multimodal consistency and anomaly patterns. Fig. 3 illustrates representative examples from the Chest PET/CT dataset, where abnormal regions are highlighted by red bounding boxes. Overall, the

proposed Chest PET/CT dataset substantially extends the scope of medical UAD research by providing a clinically grounded benchmark for multimodal anomaly detection.

In the proposed Chest PET/CT dataset, anomalies are defined at the image level, distinguishing normal scans from those exhibiting abnormal imaging patterns without specifying lesion type, location, or pathological category. This design reflects a realistic, weakly supervised setting, where fine-grained annotations are often unavailable or prohibitively expensive to obtain in large-scale clinical imaging studies. Accordingly, the dataset is not intended to support disease-specific diagnosis or lesion segmentation. Instead, it is designed for evaluating unsupervised anomaly detection methods that aim to identify deviations from normal imaging patterns in a holistic manner, serving as a preliminary screening or quality-control mechanism in clinical workflows.

4. Benchmarks

4.1. Overview of the framework

In this section, we present the proposed Cross-Modality Anomaly Detection Network (CMAD), a dual-branch architecture for unsupervised anomaly detection using paired PET/CT images. An overview of the complete processing pipeline and architectural components is provided in Fig. 2. The network follows a reverse knowledge distillation paradigm with a frozen encoder and two learnable decoders. It further integrates a Cross-Modality Prior Token, a Cross-Modality Hybrid module, and a Global Fusion Refinement module to incorporate normality-driven semantic priors and long-range contextual information.

4.2. Feature extraction

Given paired normal PET and CT images, denoted as X_p and X_c, we first divide each image into non-overlapping patches and project them into a latent space via a patch embedding layer:

$$f_p^1 = \mathrm{Ep}(X_p), \quad f_c^1 = \mathrm{Ep}(X_c), \tag{1}$$

where $f_p^1, f_c^1 \in \mathbb{R}^{H\times W\times d}$ represent the initial patch tokens of PET and CT images, respectively.

Subsequently, we employ a shared Vision Transformer encoder $E(\cdot)$, instantiated with a self-supervised model pre-trained on ImageNet, to extract hierarchical modality-aware features. The encoder is frozen during training and acts as a teacher network to ensure stable and semantically aligned representations across modalities. The multi-level features are computed iteratively as:

$$f_p^{i+1} = E_i(f_p^i), \quad f_c^{i+1} = E_i(f_c^i), \quad i = 1, \ldots, 9, \tag{2}$$

where $E_i(\cdot)$ denotes the i-th layer of the encoder, and $f_p^i, f_c^i \in \mathbb{R}^{H\times W\times d}$ are the intermediate PET and CT features at resolution level i.The final feature representations used for downstream cross-modality decoding is taken from a 9th encoder layer, denoted as f_p, f_c.

4.3. Cross-modality hybrid (CMH) module

To enable fine-grained semantic fusion between modalities, we design a Cross-Modality Hybrid module, which facilitates bidirectional attention-based feature refinement between PET and CT representations. The module leverages a learnable Cross-Modality Prior Token, denoted as f_{CFP}, which serves as a global reference to enhance semantic alignment across modalities.

Instead of relying on naive fusion operations, which simply combine PET and CT features without guaranteeing meaningful correspondence, there is a fundamental need to introduce a mechanism that explicitly promotes semantic alignment across modalities. Given the modality-specific features f_p and f_c, extracted from the frozen encoder, the CMH module applies a symmetric cross-attention mechanism to update each modality by referencing its counterpart.

Formally, the updated PET feature f_p' is computed as:

$$\begin{aligned} \tilde{f}_p &= f_p + \mathrm{CA}(\mathrm{Norm}(f_p), \mathrm{Norm}(f_{CFP})), \\ f_p' &= \tilde{f}_p + \mathrm{MLP}(\mathrm{Norm}(\tilde{f}_p)), \end{aligned} \tag{3}$$

where the Cross-Attention (CA) module allows the PET features to attend to the CT features. The cross-attention operation is defined as:

$$\begin{aligned} &q = f_p W_{qc}, k = f_{CFP} W_{kc}, v = f_{CFP} W_{vc}, \\ &\mathrm{CA}(\mathrm{Norm}(f_p), \mathrm{Norm}(f_{CFP})) = \mathrm{Softmax}\left(\frac{qk^\top}{\sqrt{d}}\right)v, \end{aligned} \tag{4}$$

where $W_{qc}, W_{kc}, W_{vc} \in \mathbb{R}^{C\times C}$ are learnable projection matrices, and $d = C/h$ denotes the dimension per attention head when using h heads.

Similarly, an identical structure is used to update the CT feature f_c, attending to the PET branch. Note that the CMPT is implicitly integrated within the cross-attention as a shared reference during training, providing modality-invariant prior guidance to both PET and CT branches. The CMPT serves as a global semantic prior, constraining attention to align functionally and structurally relevant regions, and the bidirectional update promotes complementary reasoning, enabling subtle anomalies to be captured even when they appear in only one modality. From this perspective, the CMH module can be viewed as introducing an inductive bias towards modality complementarity, bridging the gap between functional and anatomical representations.

4.4. Global fusion refinement (GFR) module

To enhance reconstruction stability and contextual coherence during decoding, we incorporate a Global Fusion Refinement module at each decoding stage. The GFR module performs joint feature processing as an intermediate regularization mechanism, where PET and CT features are temporarily fused to promote spatial continuity and contextual consistency before being projected back to modality-specific reconstruction targets. Through prototype-guided interaction followed by directional state-space modeling, GFR captures long-range spatial dependencies and enforces coherent feature organization during reconstruction. This design improves the robustness of modality-specific feature recovery under complex anatomical variations, thereby stabilizing reconstruction errors and indirectly enhancing anomaly detection performance, without explicitly enforcing lasting cross-modal alignment.

The GFR module consists of two key components: a prototype-guided multi-modal attention block and a global context encoder based on Selective Scan 2D (SS2D). The module is instantiated symmetrically in both PET and CT branches with shared architecture but modality-specific inputs. For clarity, we describe the operations of the CT branch.

Given PET features $f_p^{i'}$ and CT features f_c^i at the i-th decoding layer, the two inputs are first normalized:

$$\tilde{f}_p^{i'} = \mathrm{Norm}(f_p^{i'}), \quad \tilde{f}_c^i = \mathrm{Norm}(f_c^i). \tag{5}$$

To enable dynamic cross-modal feature refinement, we adopt a prototype-guided multi-head attention mechanism. Given normalized CT and PET features $\tilde{f}_c^i \in \mathbb{R}^{B\times N\times C}$ and $\tilde{f}_p^{i'} \in \mathbb{R}^{B\times P\times C}$, the fused representation z^i is computed as:

$$\begin{aligned} &q = \mathrm{Norm}(\tilde{f}_c^i)W_q, \quad [k, v] = \mathrm{Norm}(\tilde{f}_p^{i'})[W_k, W_v], \\ &q = \mathrm{reshape}(q) \in \mathbb{R}^{B\times h\times N\times d}, \quad k, v \in \mathbb{R}^{B\times h\times P\times d}, \\ &\alpha = \mathrm{ReLU}\big((qk^\top) \odot s\big), \\ &z^i = \mathrm{reshape}(\alpha v)W_o \in \mathbb{R}^{B\times N\times C}, \end{aligned} \tag{6}$$

where W_q, W_k, W_v, W_o are learnable projection matrices, h is the number of heads, $d = C/h$, and $s \in \mathbb{R}^{h\times 1\times 1}$ is a learnable head-wise scaling factor. ReLU activation ensures non-negativity of attention weights, facilitating stable learning and sparse selection.

The same process is mirrored in the PET branch, using CT features as key-value inputs. To further refine the fused features and model long-range dependencies, we employ the Selective Scan 2D module, which

Table 2
Quantitative results on PET/CT dataset for anomaly detection, as measured on AUROC/AP/F1[%].

Method	PET			CT		
	AUROC	AP	F1	AUROC	AP	F1
PaDiM [29]	64.22	60.76	67.92	54.84	52.14	67.86
PatchCore [8]	85.33	83.14	79.47	87.74	84.14	82.52
MHKD [34]	83.21	82.35	76.33	87.24	86.86	79.73
RD4AD [13]	80.44	75.98	74.34	86.81	84.10	81.89
INPFormer [39]	78.30	72.03	74.91	82.94	78.92	79.63
Efficient [40]	81.96	77.39	73.16	84.46	81.53	80.97
Ours	85.58	83.47	79.54	87.91	84.97	81.45

	PET/CT		
	AUROC	AP	FI
Ours	91.91	91.36	84.86

integrates spatial-aware convolution with state-space modeling in an efficient and structured manner. Given the fused representation z^i, SS2D first applies a linear projection to produce two latent representations:

$$[x, z] = \text{InProj}(z^i), \quad x, z \in \mathbb{R}^{B\times H\times W\times D}, \tag{7}$$

where x is used for dynamic modeling, and z serves as a modulation gate.

The input tensor x is reshaped and passed through a depthwise convolution with non-linear activation, followed by a learnable state-space model (SSM) for directional global context modeling:

$$y = \text{SSM}(\text{SiLU}(\text{Conv}(x))). \tag{8}$$

The output is modulated by the gating vector z through SiLU activation and projected back to the embedding space, after which a residual connection yields the refined CT representation:

$$f_{dc}^{i-1} = z^i + \text{OutProj}(y \odot \text{SiLU}(z)). \tag{9}$$

An identical process is applied in the PET branch to obtain f_{dp}^{i-1}, where PET features $f_c^{i'}$ and f_p^i serve as inputs to the GFR module. This ensures that both modalities benefit equally from prototype-guided attention and global context modeling, yielding modality-specific representations that are contextually aligned and semantically enriched.

4.5. Loss function

To enable effective unsupervised anomaly detection, our training objective enforces consistency between the reconstructed and original feature representations at multiple semantic levels. Specifically, we compute the reconstruction loss at each decoding stage by comparing the decoder output features to the corresponding encoder features extracted from the frozen teacher network. We define the cosine similarity-based reconstruction loss as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{9} 1 - \frac{\langle f_{dp}^i, f_p^i \rangle}{\|f_{dp}^i\|_2 \cdot \|f_p^i\|_2} + 1 - \frac{\langle f_{dc}^i, f_c^i \rangle}{\|f_{dc}^i\|_2 \cdot \|f_c^i\|_2}, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\| \cdot \|_2$ is the ℓ_2 norm. This hierarchical loss supervision facilitates multi-scale learning of normal anatomical patterns, ensuring better reconstruction fidelity and stronger anomaly sensitivity.

5. Experiments

5.1. Training details.

Our method is implemented in the PyTorch framework and trained from scratch on NVIDIA A100 GPUs. All images are resized to 256×256, and we follow the one-model-per-category protocol as in prior studies. For dataset partitioning, we adopt the official training, validation, and testing splits provided by each benchmark to ensure fair comparison. Importantly, all training sets contain only normal images. For our PET/CT dataset, paired PET-CT slices are used as inputs, whereas for single-modality public datasets, the same image is duplicated across both branches to maintain architectural consistency. Each student model is trained for 80,000 epochs with a batch size of 1. Optimisation is performed using the StableAdamW optimizer with a learning rate of 10^{-3}, betas of $(0.9, 0.999)$, weight decay of 10^{-4}, amsgrad enabled, and $\epsilon = 10^{-10}$. A Warm Cosine scheduler is applied to decay the learning rate smoothly from 10^{-3} to 10^{-4} across the total number of iterations.

5.2. Anomaly detection and localization

Results on PET/CT Dataset To comprehensively assess the effectiveness of our proposed CMAD-Net, we compare it with several state-of-the-art unsupervised anomaly detection (UAD) methods, including PaDiM, PatchCore, MHKD, Reverse Distillation (RD4AD), INPFormer, and EfficientAD. All methods are trained solely on normal samples and evaluated on abnormal cases, following standard UAD protocols. Quantitative results on the Chest PET/CT dataset are presented in Table 2.

We consider three evaluation settings: PET-only input, CT-only input, and dual-modality PET/CT input. Across all scenarios, CMAD-Net consistently achieves superior performance. Specifically, it obtains AUROC scores of 85.58% and 87.91% on PET and CT modalities, respectively, outperforming the best single-modality baselines (e.g., MHKD: 83.21% on PET; PatchCore: 87.74% on CT). This validates the robustness of our framework even under unimodal inputs. More importantly, when both modalities are jointly utilized, CMAD-Net achieves a substantial performance gain, reaching an AUROC of 91.91%, AP of 91.36%, and F1-score of 84.86%. These improvements highlight the critical role of multi-modal synergy in detecting complex thoracic abnormalities. The superior performance of CMAD-Net in the PET/CT setting supports our central hypothesis: functional and structural modalities are inherently complementary, and their joint modeling enhances both sensitivity and specificity in abnormality detection. To systematically exploit this complementarity, CMAD-Net incorporates dedicated architectural components to regulate cross-modality interaction and contextual aggregation. Specifically, the Cross-Modality Hybrid module facilitates token-mediated bidirectional feature interaction between modalities through a learnable prior token, thereby encouraging cross-modality consistency and effective integration of complementary cues. In addition, the Global Fusion Refinement module, built upon the SS2D mechanism, captures long-range contextual dependencies and enhances representation coherence with linear computational complexity. Unlike prior multi-modal anomaly detection methods that depend on supervision or strict spatial alignment, our fully unsupervised PET/CT framework models normality-driven cross-modality consistency through reverse knowledge distillation, enabling more robust and effective multimodal anomaly detection.

In addition to quantitative evaluations, qualitative results are illustrated in Fig. 4. The anomaly maps are normalized to the range of 0 to 1 using image-level min-max scaling. A fixed jet colormap is adopted for visualization to ensure consistent interpretation across different samples. Representative examples from the Chest PET/CT dataset are shown, where the first and second columns correspond to input CT and PET slices, respectively, and the third column presents the anomaly maps generated by CMAD-Net. It can be observed that CMAD-Net accurately localises pathological regions across different patients, effectively integrating structural cues from CT and functional information from PET. The anomaly maps highlight abnormal regions with high fidelity while suppressing the normal background. These visualisations provide intuitive evidence for the advantage of our multi-modal framework in enhancing abnormality localisation and interpretation.

Overall, CMAD-Net consistently outperforms existing UAD approaches, with particularly pronounced gains under dual-modality

Table 3
Quantitative results on Public dataset for anomaly detection/localization, as measured on image-AUROC (AC)/pixel-AUROC (AS)[%].

Method	HIS	OCT17	APTOS	BrainMRI		RESC	
	AC	AC	AC	AC	AS	AC	AS
CFLowAD [41]	54.54	85.43	94.21	73.97	93.52	74.43	93.75
RD4AD [13]	66.59	97.24	92.43	89.38	96.54	87.53	96.17
Patchcore [8]	69.34	98.56	90.45	91.55	96.97	91.50	96.39
URD [42]	67.30	98.04	93.91	77.30	96.73	92.45	96.74
SimpleNet [32]	56.26	98.31	92.80	84.35	93.22	86.74	91.65
DiAD [43]	67.26	98.54	73.29	82.36	96.08	87.77	94.01
Msflow [44]	55.08	96.81	94.90	88.73	95.91	91.56	94.86
INPFormer [39]	69.70	98.91	94.47	89.67	97.46	92.09	95.36
Efficient [40]	62.17	98.65	90.84	82.48	93.63	91.77	95.58
OUR	69.95	99.21	96.36	92.18	97.46	93.36	95.76

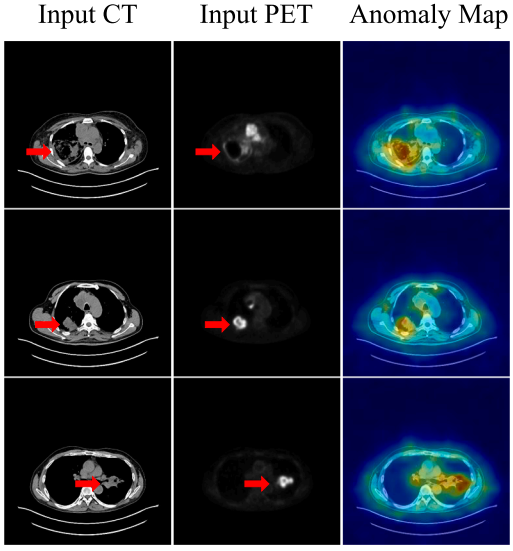

Fig. 4. Qualitative results on the chest PET/CT dataset. each row shows an input CT slice (left), the corresponding PET slice (middle), and the anomaly map generated by cmad-net (right). red arrows mark annotated lesions.

PET/CT inputs, highlighting the effectiveness of its cross-modality consistency modeling strategy.

Results on Public Dataset To validate the generalizability of CMAD-Net beyond PET/CT data, we evaluate its performance on a diverse set of public anomaly detection datasets covering multiple imaging modalities and diagnostic scenarios. The benchmark includes histopathological images (HIS), retinal OCT (OCT17 and RESC), color fundus images (APTOS), and brain MRI (BraTS2021). The area under the Receiver Operating Characteristic curve metric (AUROC) is used to quantify the performance. This metric is a standard in AD evaluation, with separate considerations for image-level AUROC in AC and pixel-level AUROC in AS. Table 3 summarizes the comparative results against a range of recent state-of-the-art UAD methods, including CFLowAD, RD4AD, PatchCore, MsFlow, URD, SimpleNet, DiAD, INPFormer, and EfficientAD.

Our method achieves the best or competitive results across all datasets. In particular, it achieves the highest classification accuracy on OCT17 (99.21%), APTOS (96.36%), and BrainMRI (92.18%), outperforming strong baselines such as INPFormer (98.91%, 94.47%, 89.67%) and PatchCore (98.56%, 90.45%, 91.55%) on the same benchmarks. Similarly, on RESC and BrainMRI segmentation tasks, CMAD-Net maintains superior or equivalent segmentation accuracy (95.76%, 97.46%), demonstrating its capacity for both coarse-level classification and fine-grained localization. In addition to quantitative results, qualitative examples on several public datasets are presented in Fig. 5. Each row corresponds to a distinct dataset, including histopathology (HIS), retinal OCT (OCT17 and RESC), fundus photography (APTOS), and brain MRI (BraTS2021). It can be observed that CMAD-Net accurately captures diverse pathological patterns across modalities, including cellular irregularities in histopathology, fluid accumulation in retinal OCT, vascular lesions in fundus images, and tumours in brain MRI. Importantly, the predicted anomaly maps exhibit high sensitivity to pathological regions while effectively suppressing irrelevant background responses, demonstrating strong localisation capability. These qualitative results indicate that CMAD-Net maintains robust anomaly detection and localisation performance across single-modality datasets, which is consistent with the quantitative improvements reported in Table 3 and highlights the general applicability of the proposed framework across heterogeneous imaging domains.

For single-modality public datasets, the same input image is duplicated across both branches to maintain architectural consistency with the multimodal setting. Under this configuration, the dual-branch structure does not introduce additional modality information but instead preserves the inductive biases of the proposed architecture. In particular, the reverse knowledge distillation scheme and global contextual modelling remain effective in enhancing reconstruction coherence and feature robustness. The consistent performance gains observed across diverse imaging modalities suggest that the core design principles of CMAD-Net generalise beyond the multimodal scenario, without relying on explicit cross-modality cues.

Moreover, CMAD-Net achieves state-of-the-art performance across multiple public benchmarks while maintaining a unified architecture for both single- and multi-modality settings, demonstrating strong generalization capability and practical applicability in medical anomaly detection.

6. Ablation study and discussion

6.1. Ablation study

To evaluate the contributions of the key architectural components within the proposed CMAD-Net, we conducted a comprehensive ablation study on the paired PET/CT dataset. The results, summarized in Table 4, analyze the impact of the Cross-Modality Hybrid (CMH) module and the Global Fusion Refinement (GFR) module against both single-modality and multi-modality baselines. All results are reported as mean ± standard deviation over five independent runs conducted under identical training and evaluation protocols with different random seeds.

Baseline and Single-Modality Performance. To establish a rigorous benchmark for performance comparison, we first evaluated the architecture under single-modality settings. In these configurations, the dual-branch structure is reduced to a single encoder-decoder stream. The empirical results presented in the first and third rows of Table 4 indicate that the CT-only baseline achieves an AUROC of 82.45%, which slightly surpasses the PET-only baseline AUROC of 81.02%. This performance disparity is likely attributable to the superior anatomical

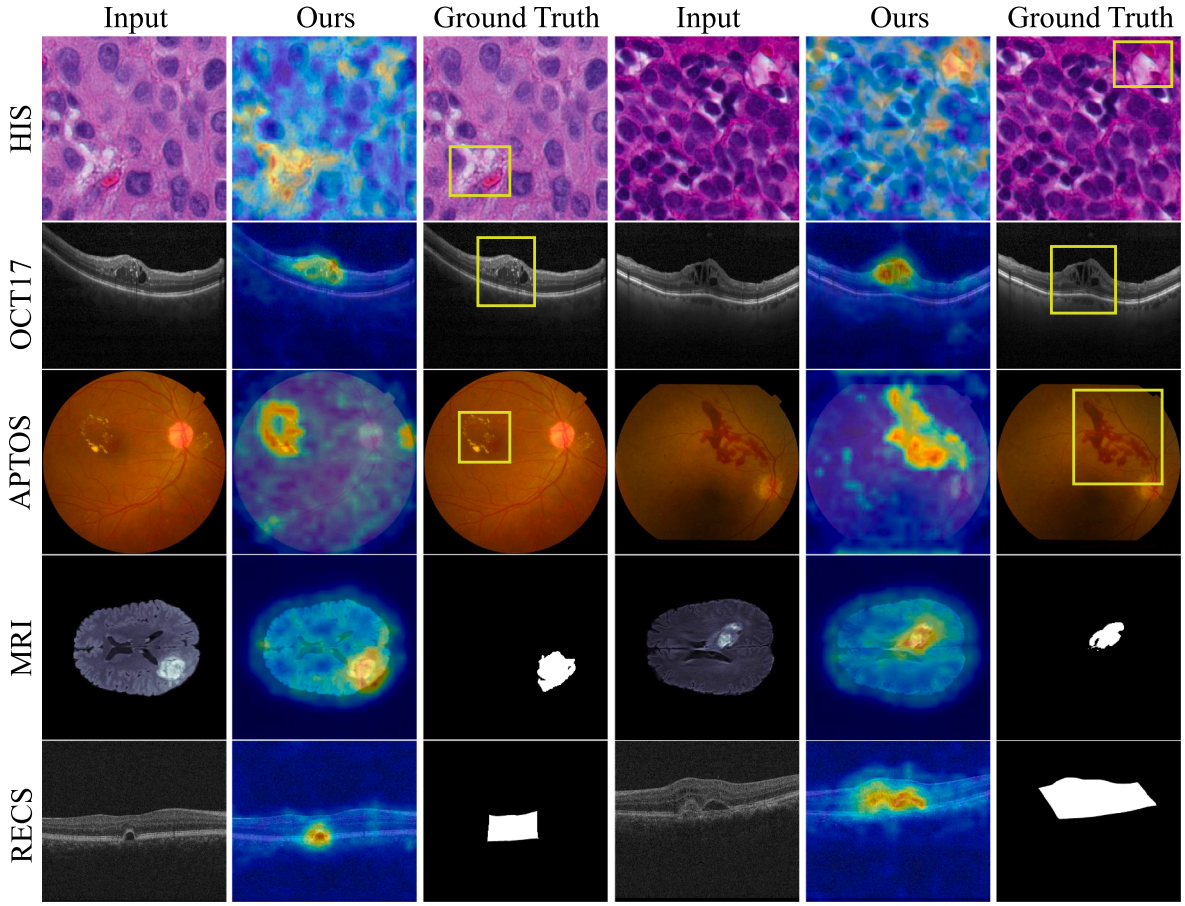

Fig. 5. Qualitative results of CMAD-Net on multiple public datasets. From top to bottom: histopathology (HIS), retinal OCT (OCT17), fundus images (APTOS), brain MRI (BraTS2021), and retinal OCT with segmentation masks (RESC). Each triplet shows the input image (left), the anomaly map predicted by CMAD-Net (middle), and the ground truth annotation (right).

Table 4
Ablation experiments on the PET/CT dataset. Results are reported as mean ± standard deviation over multiple runs. For single-modality PET and CT evaluations, the dual-branch architecture is reduced to a single encoder–decoder branch and all cross-modality modules are disabled. Model complexity is reported in terms of FLOPs.

Modality	Baseline	CMH	GFR	AUROC [%]	AP [%]	F1 [%]	FLOPs (G)
PET	✓			81.02 ± 0.04	75.97 ± 0.05	74.72 ± 0.10	132.03
PET	✓		✓	82.56 ± 0.03	76.81 ± 0.06	75.67 ± 0.03	157.17
CT	✓			82.45 ± 0.05	77.71 ± 0.04	76.10 ± 0.08	132.03
CT	✓		✓	83.89 ± 0.04	78.24 ± 0.05	76.49 ± 0.04	157.17
PET/CT	✓			83.99 ± 0.03	80.67 ± 0.06	77.53 ± 0.07	140.40
PET/CT	✓	✓		85.75 ± 0.04	82.35 ± 0.03	79.73 ± 0.04	142.28
PET/CT	✓		✓	88.54 ± 0.02	87.08 ± 0.03	81.64 ± 0.03	190.67
PET/CT	✓	✓	✓	**91.91 ± 0.02**	**91.36 ± 0.02**	**84.86 ± 0.03**	192.55

structural detail provided by Computerized Tomography compared to Positron Emission Tomography. Additionally, we have introduced GFR modules on this basis. In the single-modality setting, the GFR module degenerates into a modality-specific contextual refinement block, where no cross-modality interaction is involved. Its contribution arises solely from long-range spatial dependency modeling rather than multimodal fusion. The GFR module leverages the Selective Scan 2D mechanism to facilitate dense, long-range dependency modeling across the fused representation. Unlike standard convolutions that are limited to local receptive fields, the GFR module ensures that the reconstruction process benefits from a holistic understanding of the anatomical structure, thereby enhancing robustness against complex physiological variations and detection accuracy.

The impact of the CMH and GFR modules on cross-modality. Subsequently, we evaluated the naive multi-modal baseline where PET and CT features are processed via the shared encoder and paired decoders without the proposed interaction modules. As evidenced by the fifth row of Table 4, this naive fusion strategy yields an AUROC of 83.99%. While this surpasses the single-modality baselines, the marginal gain suggests that implicit feature concatenation is insufficient for exploiting the rich complementary information inherent in multi-modal data. To address this limitation, we introduced the CMH and GFR modules to explicitly model cross-modality consistency and global context. The integration of the CMH module alone improves the AUROC to 85.75% and the AP to 82.35%. This performance increment substantiates the value of the learnable Cross-Modality Prior Token, which

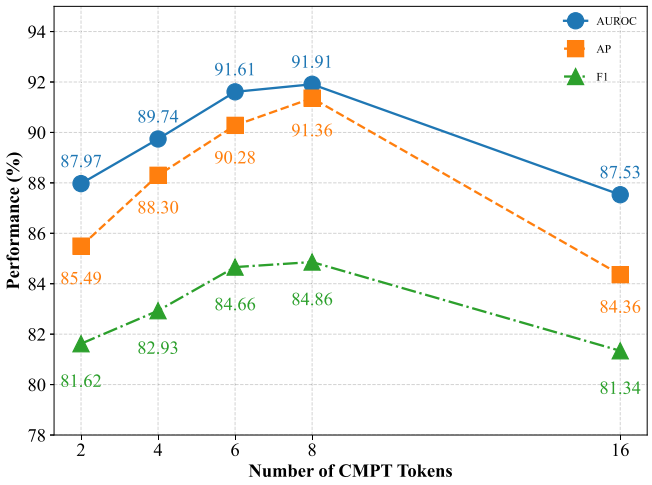

Fig. 6. Ablation on the number of cross-modality prior tokens on the PET/CT dataset.

establishes a spatially invariant bridge between PET and CT features. By encouraging token-mediated semantic interaction rather than rigid spatial alignment, the CMH module effectively reduces false positives arising from anatomical misregistration. Further substantial gains are observed upon incorporating the GFR module, which elevates the AUROC to 88.54% and the AP to 87.08% when applied independently. This underscores the critical role of global contextual reasoning in medical anomaly detection.

In addition to detection performance, we further analyse the computational complexity of the proposed framework. As reported in Table 4, model efficiency is evaluated in terms of floating-point operations (FLOPs). The results show that the introduction of the Cross-Modality Hybrid module incurs only marginal computational overhead, owing to its lightweight token-level interaction design, while the Global Fusion Refinement module accounts for the main increase in FLOPs due to its global contextual modeling via the SS2D mechanism. Overall, the proposed CMAD-Net achieves consistent performance gains with a moderate and well-controlled increase in computational cost.

The synergistic integration of both modules in the complete CMAD-Net architecture yields the most significant performance, achieving state-of-the-art results with an AUROC of 91.91%, an AP of 91.36%, and an F1-score of 84.86%. These results confirm that the proposed modules are highly complementary: the CMH module provides a robust foundation of modality-aware semantic alignment, while the GFR module refines these features through global structural modeling. Together, they constitute a unified framework that significantly outperforms strong single-modality and naive multi-modal baselines.

6.2. Discussion

We further investigate the effect of the number of Cross-Modality Prior Tokens (CMPT) within the Cross-Modality Hybrid (CMH) module. As shown in Fig. 6, increasing the token number from 2 to 8 consistently improves anomaly detection performance across all evaluation metrics. Specifically, AUROC increases from 87.97% to 91.91%, AP from 85.49% to 91.36%, and F1 from 81.62% to 84.86%, with the optimal performance observed at 8 tokens. However, when the token number is further enlarged to 16, performance decreases notably (AUROC 87.53%, AP 84.36%, F1 81.34%), indicating that excessive tokens may introduce redundancy and hinder discriminative representation learning under the unsupervised setting. These results demonstrate that CMPT provides an effective inductive bias for cross-modality alignment, but its capacity must be carefully calibrated to balance representational richness and generalization.

From a theoretical perspective, CMPT can be viewed as modality-agnostic latent anchors that project PET and CT features into a shared subspace, thereby enhancing semantic alignment through cross-attention. The number of tokens effectively controls the dimensionality of this shared basis: too few tokens under-parameterize the latent space, limiting its ability to capture diverse anatomical and metabolic correlations, whereas too many tokens over-parameterize the representation, leading to redundancy and attention diffusion. In the context of reverse knowledge distillation, such redundancy diminishes the discrepancy margin between normal and anomalous features, which is critical for robust anomaly detection. Furthermore, the non-monotonic trend reflects the interaction between CMPT and the SS2D-based Global Fusion Refinement (GFR) module: insufficient tokens restrict the routing of global contextual cues, while excessive tokens overwhelm the global aggregation with noisy priors. A moderate number of tokens, such as 6–8, strikes the best balance, providing a compact yet expressive inductive bias that complements global context modeling while maintaining computational efficiency. We emphasize that the detected anomalies should be interpreted as indicators of abnormal imaging patterns rather than definitive clinical diagnoses. The subtle or highly localized abnormalities that do not significantly alter global feature distributions may be missed. These failure modes are inherent to weakly supervised UAD paradigms and motivate future work on incorporating finer-grained supervision or human-in-the-loop refinement.

Despite the promising results, several practical limitations remain, including the use of image-level annotations, reliance on well-aligned PET/CT pairs, 2D slice-based modeling, and the uniform treatment of all abnormal categories; future work will extend the framework toward lesion-level and disease-specific analysis, improve robustness to cross-institutional variation, explore 3D volumetric modeling, and investigate modality-dependent and cross-disease distribution patterns further to enhance generalization, interpretability, and clinical relevance.

7. Conclusion

In this work, we addressed the unexplored problem of unsupervised anomaly detection from hybrid PET/CT imaging. We curate a paired PET/CT dataset for thoracic anomaly detection and investigate this problem under an unsupervised learning framework. The proposed approach extends reverse knowledge distillation to the PET/CT setting by explicitly modeling cross-modality consistency under normal conditions. To enable effective multimodal interaction without relying on explicit spatial alignment, token-mediated cross-modality fusion and global contextual refinement are incorporated to enhance reconstruction coherence and robustness. Extensive experimental results demonstrate that the proposed method consistently outperforms strong single-modality baselines, substantiating the effectiveness of normality-driven cross-modality consistency modeling for PET/CT-based unsupervised anomaly detection. Our dataset and method provide a solid foundation for advancing multi-modal UAD and developing more reliable computer-aided diagnosis systems.

CRediT authorship contribution statement

Muhao Xu: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization; **Lili Qu:** Data curation; **Zihan Nie:** Data curation; **Feng Li:** Writing – review & editing; **Zhuangzhuang Chen:** Writing – original draft; **Qi Liao:** Data curation; **Yi Wan:** Writing – review & editing; **Sijie Niu:** Writing – review & editing; **Runmin Cong:** Writing – review & editing; **Xin Li:** Writing – review & editing; **Weiye Song:** Writing – review & editing, Funding acquisition, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China (ZR2023QC262); the Key R&D Programme of Shandong Province, China (2024CXGC010106); the Shandong Provincial Technology Innovation Guidance Programme, Shandong (YDZX2024114); the National Key R&D Programme for Young Scientists, China (2025YFC2426400).

References

[1] E. Jove, J.-L. Casteleiro-Roca, H. Quintian, J.-A. Mendez-Perez, J.L. Calvo-Rolle, A new method for anomaly detection based on non-convex boundaries with random two-dimensional projections, Inf. Fusion 65 (2021) 50–57.
[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. (CSUR) 41 (3) (2009) 1–58.
[3] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, C. Adak, Unsupervised anomaly detection for surface defects with dual-siamese network, IEEE Trans. Ind. Inf. 18 (11) (2022) 7707–7717.
[4] J. Guo, S. Lu, L. Jia, W. Zhang, H. Li, Encoder-decoder contrast for unsupervised anomaly detection in medical images, IEEE Trans. Med. Imaging 43 (3) (2023) 1102–1112.
[5] S. Akcay, A. Atapour-Abarghouei, T.P. Breckon, Ganomaly: semi-supervised anomaly detection via adversarial training, in: Proceedings of the Asian Conference on Computer Vision (ACCV), Springer, 2019, pp. 622–637.
[6] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, F-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44.
[7] T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 146–157.
[8] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14318–14328.
[9] T. Reiss, N. Cohen, L. Bergman, Y. Hoshen, Panda: adapting pretrained features for anomaly detection and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2806–2814.
[10] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, Uninformed students: student-teacher anomaly detection with discriminative latent embeddings, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4183–4192.
[11] H. Guo, L. Ren, J. Fu, Y. Wang, Z. Zhang, C. Lan, H. Wang, X. Hou, Template-guided hierarchical feature restoration for anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6447–6458.
[12] M. Salehi, N. Sadjadi, S. Baselizadeh, M.H. Rohban, H.R. Rabiee, Multiresolution knowledge distillation for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14902–14912.
[13] H. Deng, X. Li, Anomaly detection via reverse distillation from one-class embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9737–9746.
[14] M. Xu, Z. Nie, B. Fu, Z. Chen, Z. Li, H. Wei, Y. Wan, W. Song, Beyond feature mapping: dual-Heterogeneous knowledge distillation with mamba for industrial anomaly detection, Expert Syst. Appl. (2026) 131146.
[15] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista, J. Del Ser, Visual tracking in complex scenes: a location fusion mechanism based on the combination of multiple visual cognition flows, Inf. Fusion 96 (2023) 281–296.
[16] S. Liu, S. Wang, X. Liu, A.H. Gandomi, M. Daneshmand, K. Muhammad, V.H.C. De Albuquerque, Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring, IEEE Trans. Multimed. 23 (2021) 2188–2198.
[17] J.B.I. Sap, C.F. Meyer, J. Ford, N.J.W. Straathof, A.B. Dürr, M.J. Lelos, S.J. Paisey, T.A. Mollner, S.M. Hell, A.A. Trabanco, et al., [18F] Difluorocarbene for positron emission tomography, Nature 606 (7912) (2022) 102–108.
[18] P.J. Withers, C. Bouman, S. Carmignato, V. Cnudde, D. Grimaldi, C.K. Hagen, E. Maire, M. Manley, A. Du Plessis, S.R. Stock, X-Ray computed tomography, Nature Rev. Methods Primers 1 (1) (2021) 18.
[19] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.
[20] B. Ehteshami Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (22) (2017) 2199–2210.
[21] Karthik, Maggie, S. Dane, APTOS 2019 Blindness Detection, 2019, (https://kaggle.com/competitions/aptos2019-blindness-detection). Kaggle.
[22] J. Hu, Y. Chen, et al., Automated segmentation of macular edema in OCT using deep neural networks, Med. Image Anal. 55 (2019) 216–227.
[23] B.H. Menze, A. Jakab, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024.
[24] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Special lecture on IE 2 (1) (2015) 1–18.
[25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, C. Steger, Improving unsupervised defect segmentation by applying structural similarity to autoencoders, arXiv preprint arXiv:1807.02011. (2018).
[26] C. Baur, B. Wiestler, S. Albarqouni, N. Navab, Deep autoencoding models for unsupervised anomaly segmentation in brain MR images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4Th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4, Springer, 2019, pp. 161–169.
[27] P. Tang, X. Yan, X. Hu, K. Wu, T. Lasser, K. Shi, Anomaly detection in medical images using encoder-Attention-2Decoders reconstruction, IEEE Trans. Med. Imaging (2025).
[28] M. Rudolph, B. Wandt, B. Rosenhahn, Same same but differnet: semi-supervised defect detection with normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1907–1916.
[29] T. Defard, A. Setkov, A. Loesch, R. Audigier, Padim: a patch distribution modeling framework for anomaly detection and localization, in: International Conference on Pattern Recognition (ICPR), Springer, 2021, pp. 475–489.
[30] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, MVTec AD–A Comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9592–9600.
[31] D. Gudovskiy, S. Ishizaka, K. 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, 2022, pp. 98–107.
[32] Z. Liu, Y. Zhou, Y. Xu, Z. Wang, Simplenet: a simple network for image anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20402–20411.
[33] Y. Zheng, X. Wang, R. Deng, T. Bao, R. Zhao, L. Wu, Focus your distribution: coarse-to-Fine non-Contrastive learning for anomaly detection and localization, arXiv preprint arXiv:2110.04538. (2021).
[34] M. Xu, C. Zhu, G. Feng, S. Niu, Multitask hybrid knowledge distillation for unsupervised anomaly detection, IEEE Trans. Ind. Inf. (2025).
[35] S. Lu, W. Zhang, H. Zhao, H. Liu, N. Wang, H. Li, Anomaly detection for medical images using heterogeneous auto-Encoder, IEEE Trans. Image Process. (2024).
[36] M. Nickparvar, 2021, (https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset).
[37] S. Liu, Z. Luo, W. Fu, Fcdnet: fuzzy cognition-based dynamic fusion network for multimodal sentiment analysis, IEEE Trans. Fuzzy Syst. 33 (1) (2024) 3–14.
[38] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, Sustainability 13 (3) (2021) 1224.
[39] W. Luo, Y. Cao, H. Yao, X. Zhang, J. Lou, Y. Cheng, W. Shen, W. Yu, Exploring intrinsic normal prototypes within a single image for universal anomaly detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 9974–9983.
[40] K. Batzner, L. Heckler, R. König, Efficientad: accurate visual anomaly detection at millisecond-level latencies, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 128–138.
[41] D. Gudovskiy, S. Ishizaka, K. 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 (WACV), 2022, pp. 98–107.
[42] X. Liu, J. Wang, B. Leng, S. Zhang, Unlocking the potential of reverse distillation for anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 5640–5648.
[43] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, L. Xie, A diffusion-based framework for multi-class anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 8472–8480.
[44] Y. Zhou, X. Xu, J. Song, F. Shen, H.T. Shen, Msflow: multiscale flow-based framework for unsupervised anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. (2024).