This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Complementary Prototype Mapping for Efficient Multimodal Anomaly Detection

Yuan Zhao[1], Xiaoqin Zhang[2], Huchuan Lu[1], Lihe Zhang[1]*
[1]IIAU-Lab, Dalian University of Technology [2]Construction Project Management Branch,
National Petroleum and Natural Gas Pipeline Network Group Co., Ltd.
zhaoyuan@mail.dlut.edu.cn, zhangxq18@pipechina.com.cn, {lhchuan,zhanglihe}@dlut.edu.cn

Abstract

*Multimodal unsupervised anomaly detection has garnered increasing attention for robust defect localization. Recent approaches rely on establishing cross-modal matching relationships under normal conditions without explicit guidance. However, in practice, a single modality may have multiple distinct representations corresponding to another modality, and such unconditional mappings struggle to adaptively capture these variations, resulting in mapping ambiguity and the misclassification of diverse yet normal variations as anomalies. Moreover, existing methods suffer from slow inference speed and high memory overhead, hindering their deployment in real-world production lines. To address these issues, we propose an efficient and effective Complementary Prototype Mapping (**CPMAD**) framework, which dynamically extracts consensus and supplementary prototypes to serve as complementary priors, thereby guiding and disambiguating cross-modal mappings. The framework comprises three key components: (1) Consensus Extraction Module (CEM) learns a dynamic anchor, transforming multimodal features into anomaly-free consensus prototypes to improve cross-modal consistency and suppress latent anomalies; (2) Supplementary Query Module (SQM) employs a Complementary Residual Attention mechanism to capture the discrepancy between the consensus and modality-specific spaces, thereby exploring the most representative cues as supplementary prototypes; and (3) Complementary Mapping Module adaptively integrates both prototypes to perform feature mapping. Extensive experiments across industrial and medical scenarios show that CPMAD outperforms existing methods in both full-data and few-shot settings, while offering faster inference and lower memory usage. Code:https://github.com/yuanzhao-CVLAB/CPMAD.*

1. Introduction

Unsupervised anomaly detection, which relies solely on normal samples without the need for labeled anomalies, has been extensively studied in industrial inspection [3, 35, 38], medical diagnosis [2, 16, 21], and video surveillance [30, 31, 37]. Among them, multimodal anomaly detection (MAD) has recently emerged as a promising solution, significantly enhancing robustness and reliability by leveraging complementary sensory cues. In particular, combining RGB images with 3D measurements enables the joint detection of texture and geometric defects that may remain imperceptible in unimodal anomaly detection.

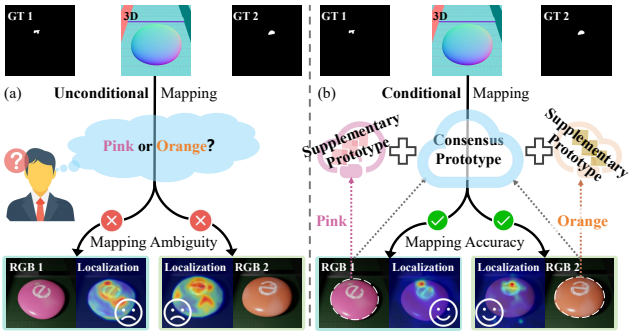


Figure 1. (a) Advanced methods [10, 24] employ unconditional cross-modal mapping and may struggle to map diverse color distributions under identical 3D shapes. (b) The proposed method integrates consensus and supplementary prototypes to provide complementary cues (e.g., either *pink* or *orange*), effectively guiding adaptive and accurate cross-modal mapping.

Early multimodal anomaly detection methods identify defects by measuring reconstructed multimodal feature discrepancies. For example, MMRD [13] adopts parameter-free modulation to blend RGB and depth representations and measures teacher–student distances, while AST [23] concatenates RGB-D features to learn joint representations and estimate density via normalizing flows. However, detecting anomalies directly in the multimodal common space can lead to representation entanglement, where modality-specific information may be lost. Consequently, M3DM [29] uses contrastive learning to align RGB and 3D representations and stores details in multiple memory banks. To mitigate memory overhead, methods based on feature mapping [10, 24], such as CFM [10], learn cross-modal matching relationships to recover modality-specific details and locate defects by mapping differences. In practice, inter-modal re-

*Corresponding authors

lationships are inherently a many-to-many and complicated mapping. Their unconditional mapping may lead to mapping deviations for normal samples with diverse attributes. For instance, identical 3D geometric structures may exhibit diverse colors in the RGB modality (e.g., pink and orange regions, as shown in Fig. 1), leading to erroneous anomaly localization, and vice versa. Furthermore, existing methods suffer from inefficient memory banks and time-consuming decoding processes, hindering their deployment in real-world applications. Hence, the key challenge is *how to effectively and efficiently reconstruct distribution-diverse, modality-specific details while maintaining cross-modal consistency*.

To address this fundamental challenge, we propose the Complementary Prototype Mapping (CPMAD) framework. The core idea is to dynamically capture consensus and supplementary prototypes that represent diverse variations both within and across modalities, enabling adaptive and precise mapping. The consensus prototypes construct a shared, anomaly-free subspace that enhances cross-modal consistency and filters out latent anomalies. Conversely, the supplementary prototypes dynamically capture modality-specific, pivotal cues that are not represented by the consensus prototypes. Together, both prototypes complement and enhance each other, thereby guiding adaptable reconstruction.

Building upon this framework, we design three collaborative components. (1) Consensus Extraction Module (CEM) extracts dynamic anchors via a cross-modal anchor generator to establish a reference center across modalities. It then projects multimodal features into an anomaly-free prototype subspace guided by these anchors, preserving modality-invariant consensus information while suppressing anomalous variations. (2) Supplementary Query Module (SQM) introduces a Complementary Residual Attention mechanism that jointly attends to the consensus and modality-specific spaces and learns a residual attention map between them. This mechanism enables the model to adaptively capture modality-specific variations that are susceptible to cross-modal ambiguity and often overlooked by the consensus prototypes. (3) Complementary Mapping module adaptively integrates the consensus and supplementary prototypes to reconstruct cross-modal features. The proposed framework effectively reduces mapping ambiguity through complementary prototype learning, while eliminating the need for a heavy decoder and maintaining linear computational complexity. Experimental results indicate that our CPMAD and CPMAD-S models deliver strong performance with high efficiency. Notably, the CPMAD-S variant achieves a **5.05× faster inference**, a **2.59× reduction in memory usage**, and **11.86× fewer FLOPs** than the most efficient prior method, while maintaining competitive detection performance, as evidenced by Fig. 2 and Tab. 7. Additionally, the model shows excellent scalability with increasing data and model sizes.

Our main contributions can be summarized as follows:

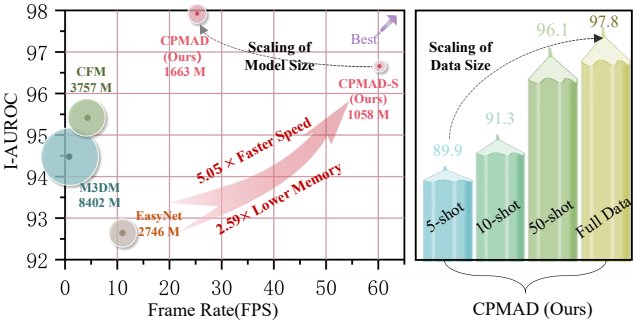


Figure 2. Left: Performance, inference speed, and memory usage of MAD methods on an NVIDIA RTX 2080Ti. Right: I-AUROC of CPMAD under few-shot settings.

- We propose an efficient and effective CPMAD framework that adaptively and accurately reconstructs cross-modal features through complementary prototype learning.
- We design a Consensus Extraction Module with a Cross-modal Anchor Generator and a Supplementary Query Module with Complementary Residual Attention to collaboratively extract complementary prototypes.
- CPMAD achieves state-of-the-art performance across multiple benchmarks, significantly outperforming existing methods in accuracy and efficiency, and demonstrates strong few-shot learning capability.

2. Related Work

2.1. Multimodal Anomaly Detection

Multimodal anomaly detection [10, 34, 36] leverages complementary sensory cues, such as RGB images, depth maps, surface normals, or 3D point clouds, to improve defect localization. Early methods such as AST [23] employ normalizing flows to estimate feature density, while EasyNet [8] utilizes attention-based entropy fusion to mitigate inter-modal interference. Recent memory-based methods, such as M3DM [29] and Shape-guided [9], store normal feature prototypes for anomaly detection. To reduce memory usage, CFM [10] learns cross-modal mappings under normal conditions. Cycle-CFM [24] extends this approach by constructing a bidirectional feature alignment mechanism. However, these methods either require substantial memory to store normal patterns, or suffer from mapping ambiguity due to unconditional feature mapping. In contrast, our CPMAD framework dynamically learns consensus and supplementary prototypes that encode normal and complementary priors, enabling precise cross-modal reconstruction with faster inference and lower memory consumption.

2.2. Prototype Learning

Prototype learning [25] has gained traction in anomaly detection for encoding normality as a compact set of representative features. In unimodal anomaly detection, methods like HVQ-Trans [1] and DSR [33] use learnable or

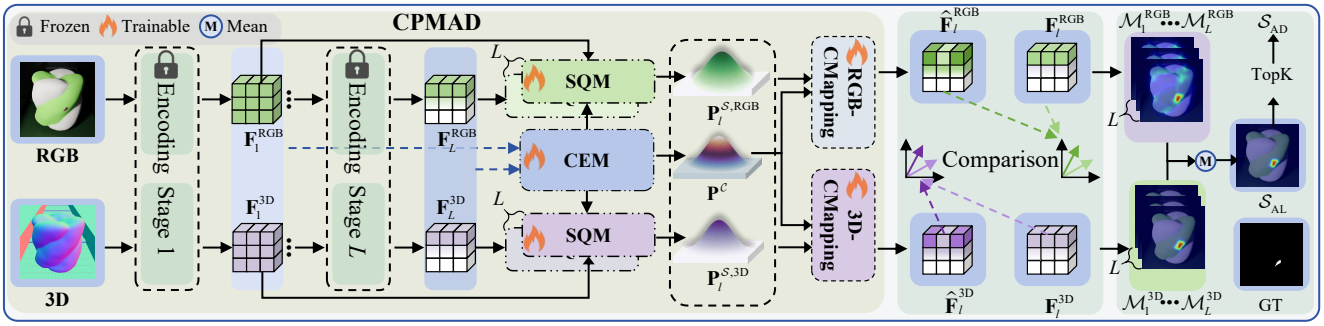


Figure 3. Overview of the proposed CPMAD framework. The architecture consists of paired pre-trained encoders with L encoding stages, a Consensus Extraction Module (CEM), L Supplementary Query Modules (SQMs), and L Complementary Mapping (CMapping) Modules. CEM and SQMs extract complementary prototypes, which guide CMapping modules in reconstructing per-modality features.

vector-quantized prototypes to prevent shortcut solutions and enhance reconstruction fidelity. INP-Former [19] extends this idea by learning intrinsic prototypes to form a compact normal feature space, facilitating anomaly discrimination. However, prototype-based unimodal methods struggle to generalize to multimodal anomaly detection, as learning isolated modality-specific prototypes fails to capture cross-modal consistency. The proposed complementary prototype mapping mechanism jointly learns consensus prototypes to model cross-modal consistency and supplementary prototypes to encode modality-specific characteristics, thereby enhancing the robustness of MAD.

3. Method

3.1. Overview

We propose the Complementary Prototype Mapping (CPMAD) framework, which jointly learns consensus and supplementary prototypes to guide cross-modal mapping. As shown in Fig. 3, the framework consists of a pair of pre-trained ViT-based encoders spanning 1-L encoding stages, a Consensus Extraction Module (CEM), and multiple Supplementary Query Modules (SQMs) and Complementary Mapping (CMapping) modules. Given a pair of 3D and RGB images, the inputs are first processed by the pre-trained encoders to produce multi-layer patch-token representations $\{\mathbf{F}_l^{\text{3D}}, \mathbf{F}_l^{\text{RGB}} | \mathbf{F}_l \in \mathbb{R}^{N\times D}\}_{l=1}^{L}$, where l, N and D denote the layer index, the length of tokens and the embedding dimension, respectively. The patch-token maps are then fed into the CEM, which extracts cross-modal shared semantics to obtain consensus prototypes $\mathbf{P}^{\mathcal{C}}$. Each SQM explores supplementary prototypes $\mathbf{P}^{\mathcal{S}}$, which complement the consensus prototypes by capturing modality-specific, typical components that are not covered in $\mathbf{P}^{\mathcal{C}}$. The consensus and supplementary prototypes are aggregated by the Complementary Mapping module to collaboratively reconstruct per-modality features at each layer, while reconstruction discrepancies are leveraged for anomaly localization.

3.2. Consensus Extraction Module

As shown in Fig. 4, Consensus Extraction Module (CEM) aims to extract anomaly-free consensus prototypes by capturing cross-modal consistent semantics and suppressing anomalous components. To this end, CEM introduces a cross-modal anchor generator to produce dynamic anchors, thereby projecting the multimodal representations to obtain a purified prototype.

Given the per-layer RGB and 3D tokens, CEM first concatenates them to form $\mathbf{F}^{\text{3D}} = [\mathbf{F}_1^{\text{3D}}; \cdots; \mathbf{F}_L^{\text{3D}}]$ and $\mathbf{F}^{\text{RGB}} = [\mathbf{F}_1^{\text{RGB}}; \cdots; \mathbf{F}_L^{\text{RGB}}]$, then computes their mean to obtain a consensus token $\overline{\mathbf{F}} \in \mathbb{R}^{NL\times D}$. This averaging process smooths modality inconsistencies, thereby suppressing modality-independent anomaly signals. However, anomalies that co-occur across modalities may still contaminate the consensus token during averaging. To address this issue, a dynamic anchor is introduced to project the tokens into an anomaly-free prototype space.

Specifically, CEM initializes a set of learnable embeddings $\mathcal{H} \in \mathbb{R}^{I\times D}$ as initial anchors, where I denotes the anchor size. The cosine similarity between each anchor and the per-modality features (RGB and 3D) is independently computed and averaged to produce the affinity scores:

$$\mathbf{W}_{i,n} = \frac{1}{2}\Big(\frac{(\mathbf{Z}_n^{\text{3D}})^\top \mathcal{H}_i^{\text{3D}}}{\|\mathbf{Z}_n^{\text{3D}}\|\,\|\mathcal{H}_i^{\text{3D}}\|} + \frac{(\mathbf{Z}_n^{\text{RGB}})^\top \mathcal{H}_i^{\text{RGB}}}{\|\mathbf{Z}_n^{\text{RGB}}\|\,\|\mathcal{H}_i^{\text{RGB}}\|}\Big), \quad (1)$$

where $\mathbf{Z}^{\text{3D}}, \mathbf{Z}^{\text{RGB}}$ denote the projected features from $\mathbf{F}^{\text{3D}}, \mathbf{F}^{\text{RGB}}$. $\mathcal{H}^{\text{3D}}, \mathcal{H}^{\text{RGB}}$ are the projections of $\mathcal{H}$. Subsequently, CEM concatenates the RGB and 3D tokens along channel dimension and projects them to form $\mathbf{F}^{\text{cat}} \in \mathbb{R}^{NL\times D}$. Using the affinity scores $\mathbf{W}$, the dynamic anchors are updated by aggregating the cross-modal features, enabling anchors to adaptively capture cross-modal semantics:

$$\mathcal{H}_i^{\text{Dyn}} = \sum_{n=1}^{NL} \mathbf{W}_{i,n}\mathbf{F}_n^{\text{cat}}. \quad (2)$$

Finally, cross-attention mechanisms [27] are employed to emphasize tokens near the dynamic anchors, thereby ob-

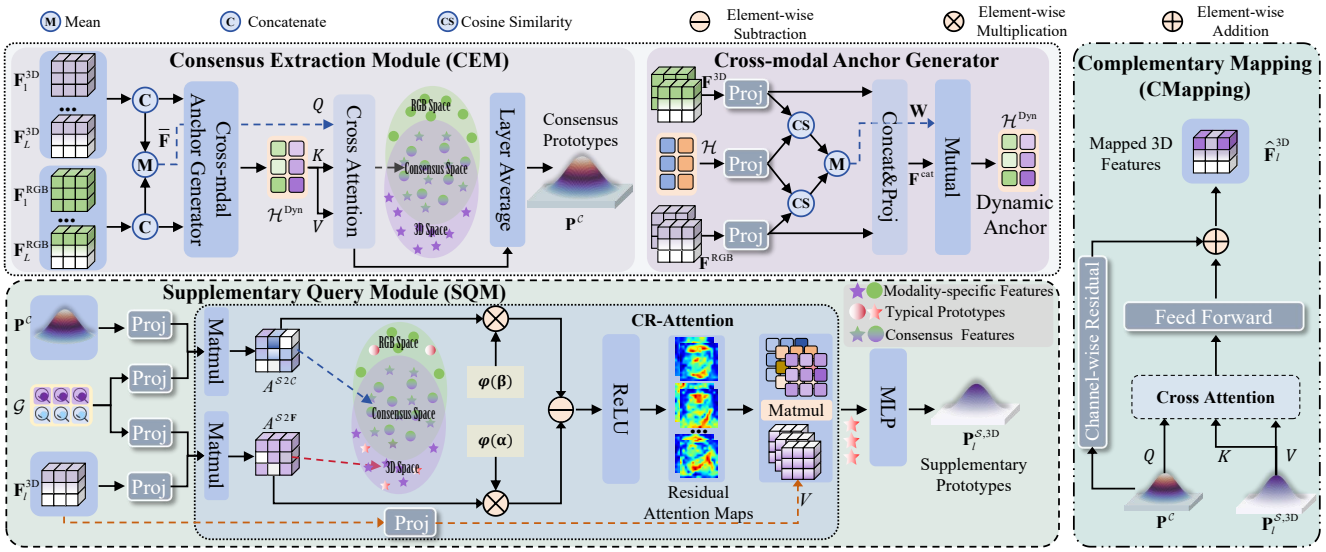


Figure 4. Structure of CEM, SQM, and CMapping. CEM utilizes a Cross-modal Anchor Generator to produce Dynamic Anchors for extracting consensus prototypes. SQM features CR-Attention to query supplementary prototypes , while CMapping integrates both prototypes to reconstruct modality-specific features.

taining consensus prototypes that encode anomaly-free and cross-modal consistent features:

$$\mathbf{P}^{\mathcal{C}} = LAvg(CA(\overline{\mathbf{F}}, \mathcal{H}^{\text{Dyn}})), \tag{3}$$

where CA and $LAvg$ denote cross-attention and layer-wise averaging, respectively.

3.3. Supplementary Query Module

As shown in Fig. 4, Supplementary Query Module (SQM) aims to capture modality-specific, typical variations that are not represented by the consensus prototypes. The key idea is to enable a small set of learnable queries to attend jointly to both the consensus and modality-specific spaces by Complementary Residual Attention (CR-Attention), thereby adaptively emphasizing discriminative cues that resolve cross-modal mapping ambiguities.

Given the modality-specific tokens $\mathbf{F}_l$ and the consensus prototypes $\mathbf{P}^{\mathcal{C}}$, SQM first initializes a set of learnable queries $\mathcal{G} \in \mathbb{R}^{G\times D}$, where G represents the query size. It then employs learnable queries to attend to the consensus prototypes and modality-specific tokens, generating supplementary-to-consensus $A^{\mathcal{S}2\mathcal{C}}$ and supplementary-to-feature $A^{\mathcal{S}2\mathbf{F}}$ attention scores via CR-Attention:

$$\begin{aligned} A^{\mathcal{S}2\mathcal{C}} &= Q^{\mathcal{C}}(K^{\mathcal{C}})^T, Q^{\mathcal{C}} = \mathcal{F}_Q^{\mathcal{C}}(\mathcal{G}), K^{\mathcal{C}} = \mathcal{F}_K^{\mathcal{C}}(\mathbf{P}^{\mathcal{C}}), \\ A^{\mathcal{S}2\mathbf{F}} &= Q^{\mathbf{F}}(K^{\mathbf{F}})^T, Q^{\mathbf{F}} = \mathcal{F}_Q^{\mathbf{F}}(\mathcal{G}), K^{\mathbf{F}} = \mathcal{F}_K^{\mathbf{F}}(\mathbf{F}_l), \end{aligned} \tag{4}$$

where $\mathcal{F}$ denotes the multi-head linear projection. Since the two attention scores represent responses to the consensus and modality-specific spaces, CR-Attention introduces a residual attention mechanism that adaptively emphasizes the supplementary information, as follows:

$$\mathbf{P}_l^{\mathcal{S}} = \text{ReLU}\Big(\frac{\varphi(\alpha)A^{\mathcal{S}2\mathbf{F}} - \varphi(\beta)A^{\mathcal{S}2\mathcal{C}}}{\sqrt{D}}\Big)V, \tag{5}$$

where $V = \mathcal{F}_V^{\mathbf{F}}(\mathbf{F}_l)$ represents the multi-head linear projection of modality-specific features. The scalar parameters α and β are learnable parameters, and $\varphi(\cdot)$ is the Sigmoid function used to normalize parameters to the range of 0 to 1, thereby balancing the two attention paths. By subtracting the supplementary-to-consensus attention scores and leveraging the suppressive characteristic of the ReLU function, CR-Attention dynamically filters out negative score regions that are well explained by the consensus prototypes. In contrast, regions with significantly higher residual score are emphasized to generate supplementary prototypes, thereby covering modality-exclusive cues that are able to mitigate ambiguity. Since supplementary cues generally contain fewer components, and an excessive number of queries increases complexity while potentially reintroducing modality-dependent noise, the supplementary prototypes are constrained to be low-rank by keeping the number of queries $G \ll N$.

3.4. Complementary Mapping

We propose the Complementary Mapping module, which adaptively integrates the shared semantics of consensus prototypes and the modality-specific cues encoded in supplementary prototypes. As shown in Fig. 4, it performs a lightweight cross attention [27] between consensus and supplementary prototype representations, followed by a feed-forward refinement network to enhance feature expressiveness, ultimately obtaining per-layer mapped 3D features $\hat{\mathbf{F}}_l^{\text{3D}}$ and RGB features $\hat{\mathbf{F}}_l^{\text{RGB}}$. Through this mechanism,

each token adaptively attends to the most relevant prototype component, aggregating complementary cues from both the consensus and supplementary subspaces to form semantically coherent reconstructions. Unlike conventional cross attention, which employs token-wise residual and may suffer from the "identical shortcut" problem [32], CPMAD adopts a channel-wise residual strategy to preserve shared contextual information within the consensus prototypes. Specifically, this channel-wise residual is realized via an MLP-based refinement followed by spatial averaging, which maintains cross-modal consistency while suppressing latent anomalies.

3.5. Complexity, Training, and Inference

Computational complexity. Our framework employs complementary prototype learning, eliminating the need for a heavy decoder and utilizing lightweight attention mechanisms, resulting in fast inference and low memory overhead. Specifically, the computational complexity of the CEM module is $\mathcal{O}(NLID)$, where $L = 2$ and $I = 12$, which is significantly smaller than the feature map size (N=784) and is computed only once. Similarly, both the SQM and Complementary Mapping modules exhibit relatively low computational complexities of $\mathcal{O}(NGD)$, where $G \ll N$. The overall complexity scales linearly with the token count, making CPMAD substantially more efficient than full self-attention architectures when processing long token sequences.

Training. We employ mapping loss to align the mapped feature with the corresponding per-modality feature, as follows:

$$\mathcal{L}_{\text{map}} = \frac{1}{LHW} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathcal{M}_l^{\text{3D}}(h,w) + \mathcal{M}_l^{\text{RGB}}(h,w), \tag{6}$$

where H and W represent the height and width of the feature map, respectively. $\mathcal{M}_l^{\text{3D}}(h,w)$ is the anomaly map of 3D modality on the l-th feature map corresponding to coordinates (h,w), and is calculated by cosine similarity as:

$$\mathcal{M}_l^{\text{3D}}(h,w) = 1 - \frac{(\hat{\mathbf{F}}_l^{\text{3D}}(h,w))^T \cdot \mathbf{F}_l^{\text{3D}}(h,w)}{||\hat{\mathbf{F}}_l^{\text{3D}}(h,w))|| \quad ||\mathbf{F}_l^{\text{3D}}(h,w)||}, \tag{7}$$

The $\mathcal{M}_l^{\text{RGB}}$ of RGB modality is acquired in the same way.

Inference. We average anomaly maps across all layers and modalities to obtain an anomaly localization score $\mathcal{S}_{\text{AL}}$. The image-level anomaly detection score $\mathcal{S}_{\text{AD}}$ is calculated by the top-K mean of anomaly localization scores, where K is set to 0.1% of the total pixel count. The pseudocode of the proposed CPMAD is provided in Appendix Algorithm 1.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the proposed method on two widely used industrial multimodal benchmarks: **MVTec-3D** [4] and **Eyecandies** [5], each containing 10 object categories. MVTec-3D is a real-world dataset collected under industrial inspection settings, while Eyecandies is a more challenging dataset featuring diverse color distributions and varying viewpoints. In addition, we construct the BraTS-AD dataset from BraTS 2020 under an unsupervised setting for brain tumor anomaly detection. It comprises four modalities (T1, T2, T1ce, and Flair) with a total of 1,350 images.

Metrics. Following mainstream research [10, 29], we evaluate performance using both image-level and pixel-level metrics. For image-level assessment, we report the Area Under the Receiver Operating Characteristic Curve (I-AUROC). For pixel-level evaluation, we report the Area Under the Receiver Operating Characteristic Curve (P-AUROC) and the Area Under the Per-Region-Overlap (AUPRO).

Implementation Details. All experiments were conducted over 300 epochs with a batch size of 10 on one NVIDIA RTX 4090 GPU. CPMAD utilizes a DINO ViT-B/8 backbone [7, 12, 20] trained on ImageNet [17], the same as that used by mainstream MAD methods [10, 26], to extract two feature groups (L = 2), corresponding to low- and high-level representations [14]. Surface normals are used as the 3D modality alongside RGB images, both with an input resolution of 224×224, corresponding to $N = 784$ tokens (28×28). The anchor size I in CEM is set to 12, and the query size G in SQM is set to 8. Additional details are provided in the Appendix to ensure full reproducibility.

4.2. Quantitative Results

Industrial MAD. Tab. 1 and Tab. 2 summarize the comparisons on MVTec-3D and Eyecandies, respectively. The proposed CPMAD achieves the best overall performance on both benchmarks. On MVTec-3D, CPMAD attains a mean I-AUROC of 97.8% and a mean AUPRO of 97.9%, surpassing the SOTA methods. On Eyecandies, CPMAD achieves a mean AUPRO of 93.4%, outperforming the best prior method by 4.0%. Furthermore, the per-class analysis shows consistent improvements across most categories. Notably, CPMAD yields significant gains on classes where RGB appearance exhibits high intra-class variability for similar geometries (e.g., color-variant *Licorice Sandwich* and *Gummy Bear*), validating the effectiveness of CPMAD for mitigating confusion mapping.

Medical MAD. Tab. 3 presents the results on the multimodal BraTS-AD dataset, a challenging benchmark due to its high inter-modal diversity and diagnostic complexity (Fig. 6). For comparison, we extend established methods (PatchCore [22], RD [11]) with parameter-free fusion in MMRD [13] to handle four modalities. Our approach, however, generalizes seamlessly to four modalities by incorporating additional SQM and Complementary Mapping modules, achieving performance gains of 4.2% in I-AUROC and 5.6% in AUPRO. These results highlight our framework's scalability and gen-

Table 1. Quantitative results on the MVTec-3D dataset, with the best I-AUROC/AUPRO scores highlighted in **bold**.

Method	Publication	Bagel	Cable.	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
PatchCore [22]	CVPR'22	77.8/97.6	89.3/97.1	96.7/97.9	92.1/97.3	88.5/**97.3**	93.1/88.8	84.3/88.8	68.8/97.1	78.2/97.5	79.0/94.9	84.8/95.2
BTF [15]	CVPR'23	91.8/97.6	74.8/96.9	96.7/97.9	88.3/97.3	93.2/93.3	58.2/88.8	89.6/97.5	91.2/98.1	92.1/95.0	88.6/97.1	86.5/95.9
M3DM [29]	CVPR'23	99.4/97.0	90.9/**97.9**	97.2/97.1	97.6/95.0	96.0/95.0	94.2/93.2	97.3/97.7	89.9/**98.9**	97.2/97.5	94.5/93.6	94.5/96.4
AST [23]	WACV'23	98.3/97.0	87.3/94.7	97.6/98.1	97.1/93.9	93.2/91.3	88.5/90.6	97.4/97.9	**98.1**/98.2	100./88.9	79.7/94.0	93.7/94.4
EasyNet [8]	MM'23	99.1/83.9	99.8/86.4	91.8/95.1	96.8/61.8	94.5/82.8	94.5/83.6	90.5/94.2	80.7/88.9	99.4/91.1	79.3/52.8	92.6/82.1
CPMF [6]	PR'24	98.3/95.8	88.9/94.6	98.9/97.9	99.1/86.8	95.8/89.7	80.9/74.6	98.8/98.0	95.9/98.1	97.9/96.1	**96.9**/97.7	95.2/92.9
CFM [10]	CVPR'24	99.4/97.9	88.8/97.2	98.4/98.2	**99.3**/94.5	98.0/95.0	88.8/96.8	94.1/98.0	94.3/98.2	98.0/97.5	95.3/98.1	95.4/97.1
Cycle-CFM [24]	ESWA'25	**99.7**/98.0	91.6/96.0	97.9/97.6	98.6/96.8	**98.3**/97.0	89.4/94.6	98.3/97.9	95.6/97.5	98.4/95.8	96.3/92.8	96.0/96.4
M3DM-NR [28]	TPAMI'25	99.3/97.4	91.1/97.1	97.7/97.8	97.6/94.5	96.0/93.8	92.2/94.7	97.3/97.8	89.9/97.1	95.5/97.2	88.2/97.4	94.5/96.5
3D-ADNAS [18]	AAAI'25	**99.7**/84.1	**100.**/80.4	97.1/90.2	98.6/64.5	96.6/82.7	94.8/78.9	89.7/87.0	87.3/91.9	**100.**/91.7	86.7/85.6	95.1/83.7
G^2SF [26]	ICCV'25	**99.7**/98.2	92.3/97.7	99.3/98.2	96.7/**97.9**	96.6/97.1	**99.1**/97.6	99.4/98.2	**98.8**/98.3	96.6/97.8	92.2/98.1	97.1/**97.9**
CPMAD-S	Ours	97.0/97.6	95.5/97.6	100./98.3	99.3/96.4	96.4/95.4	98.1/98.0	99.6/98.2	94.5/98.3	95.4/98.1	89.7/97.7	96.6/97.6
CPMAD	Ours	99.1/**98.3**	98.6/**97.9**	**99.4/98.4**	97.4/96.9	96.0/96.1	98.1/**98.1**	**100./98.3**	95.8/98.3	99.0/**98.4**	94.8/**98.2**	**97.8/97.9**

Table 2. Quantitative comparison on Eyecandies datasets, with the best I-AUROC/AUPRO scores highlighted in **bold**. "-" indicates results not reported in the original paper.

Method	Publication	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh-mallow	Peppermint Candy	Mean
PatchCore [22]	CVPR'22	44.8/70.9	95.0/93.3	77.9/73.7	92.8/95.2	88.8/90.2	41.6/40.7	91.2/91.9	83.1/86.6	**100.**/96.9	96.3/92.9	81.1/84.0
Eyecandies [5]	ACCV'22	58.7/85.2	84.6/90.3	80.7/74.1	83.3/93.5	83.3/89.9	54.3/53.6	74.4/86.7	87.0/86.4	94.6/94.5	83.5/84.3	78.4/83.9
M3DM [29]	CVPR'23	59.7/88.9	95.4/92.1	93.1/80.8	99.0/**98.2**	88.3/88.9	66.6/67.5	92.3/87.2	88.8/90.1	99.5/96.4	**100.0**/97.3	88.2/88.7
AST [23]	WACV'23	94.5/51.4	89.1/83.5	82.7/71.4	95.8/90.5	85.7/58.7	74.8/59.0	91.1/73.6	90.7/76.9	96.4/91.8	97.2/87.8	89.8/74.4
EasyNet [8]	ACMMM'23	73.7/-	93.4/-	86.6/-	96.6/-	71.7/-	82.2/-	84.7/-	86.3/-	97.7/-	96.0/-	86.9/-
CPMF [6]	PR'24	77.3/94.1	79.5/75.6	75.0/58.4	84.6/79.5	74.0/72.8	58.4/56.1	74.6/67.9	82.0/90.8	71.0/69.6	81.3/78.4	75.8/74.3
CFM [10]	CVPR'24	68.0/94.2	93.1/90.2	95.2/83.1	88.0/96.5	86.5/87.5	78.2/**76.2**	91.7/79.1	84.0/91.3	99.8/93.9	96.2/94.9	88.1/88.7
3D-ADNAS [18]	AAAI'25	**89.6**/94.5	**100.**/89.1	97.0/82.7	**100.**/95.8	82.7/85.7	**88.2**/74.8	93.1/91.1	95.0/90.7	**100.**/96.4	**100.**/97.2	94.6/89.8
G^2SF [26]	ICCV'25	69.6/92.8	96.3/89.7	96.7/84.3	**100./98.**2	90.7/89.0	70.1/68.7	93.9/88.8	85.5/91.3	98.9/97.1	**100./98.1**	90.2/89.8
CPMAD-S	Ours	76.1/94.8	99.6/92.1	98.2/**93.9**	94.7/97.8	97.3/93.9	83.1/69.7	95.7/95.5	88.7/94.2	**100.**/98.2	99.4/97.9	93.3/92.8
CPMAD	Ours	81.3/**96.7**	**100./93.4**	**99.7**/93.5	99.4/98.1	**97.9/93.3**	80.2/72.6	**96.0/96.2**	**95.2/94.3**	**100./98.3**	**100./98.1**	**95.0/93.4**

Table 3. Quantitative comparison on the BraTS-AD dataset.

Methods	I-AUROC	P-AUROC	AUPRO
RD [11]+MMRD [13]	90.9	94.6	77.3
PatchCore [22]+MMRD [13]	91.8	95.7	82.6
DecoupleMAD	**96.4**	**98.0**	**86.3**

eralization ability across complex modality combinations.

Few-shot comparison. To evaluate generalization under low-data regime, we follow the standard few-shot protocol in MAD [10] and report results for 5/10/50-shot settings in Tab. 4. CPMAD demonstrates strong cold-start capability: with only 5 samples per class, it achieves I-AUROC scores of 89.9% on MVTec-3D and 88.0% on Eyecandies, surpassing the previous SOTA by 8.8% and 10.6%, respectively. The performance gap remains evident at 10-shot and 50-shot, and CPMAD exhibits favorable scaling behavior as additional normal samples become available. This few-shot robustness shows that decoupling cross-modal features into consensus and supplementary prototypes effectively reduces data dependency and the difficulty of learning normal patterns, thereby mitigating overfitting in data-scarce scenarios.

4.3. Qualitative Results

Fig. 5 and Fig. 6 compare the proposed CPMAD with existing methods in both industrial and medical domains. In industrial scenarios, memory-bank-based methods such as M3DM [29] and PatchCore [22] exhibit dispersed attention

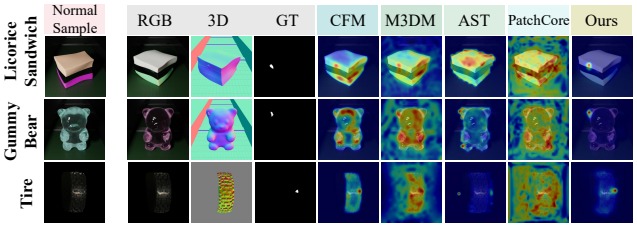


Figure 5. Comparison of localization heatmaps between ours and SOTA methods on the MVTec-3D and Eyecandies.

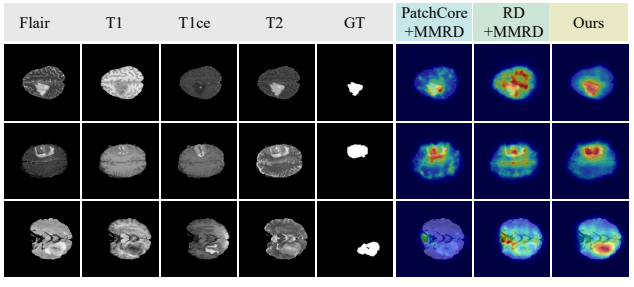


Figure 6. Comparison of localization heatmaps on the BraTS-AD.

responses due to their lack of explicit optimization on normal samples. While CFM [10] achieves superior localization performance on *Tire* with limited color or shape variations, it tends to produce scattered responses in normal regions with diverse color distributions (e.g., *Licorice Sandwich*). In contrast, our method generates sharp and well-localized anomaly maps. In medical scenarios, the proposed CPMAD

Table 4. Quantitative comparison under 5/10/50-shot settings, reporting mean performance (I-AUROC/P-AUROC/AUPRO) over five runs.

Method	Publication	5-shot		10-shot		50-shot		Full data	
		MVTec-3D	Eyecandies	MVTec-3D	Eyecandies	MVTec-3D	Eyecandies	MVTec-3D	Eyecandies
AST [23]	WACV'23	68.0/95.0/90.3	63.3/74.1/69.1	68.9/94.6/83.5	67.1/76.7/62.4	79.4/97.4/92.9	73.9/86.2/71.5	93.7/97.6/94.4	78.0/90.2/74.4
M3DM [29]	CVPR'23	82.2/98.4/93.7	76.4/87.1/80.7	84.5/98.6/94.3	82.4/89.0/81.2	90.7/98.9/95.5	83.6/93.3/84.6	94.5/99.2/96.4	88.2/97.7/88.7
CFM [10]	CVPR'24	81.1/98.6/94.9	79.5/87.9/80.1	84.5/98.7/95.4	83.8/90.3/82.5	90.6/99.1/96.5	85.2/92.6/85.1	95.4/99.3/97.1	88.1/97.4/88.7
3D-ADNAS [18]	AAAI'25	82.6/-/-	77.5/87.5/70.4	84.8/-/-	80.7/86.9/76.7	89.0/-/-	86.8/91.2/89.8	95.1/-/-	94.6/97.0/89.8
CPMAD	Ours	**89.9/99.2/96.6**	**88.0/98.1/90.8**	**91.3/99.3/96.9**	**90.5/98.3/91.4**	**96.1/99.5/97.4**	**93.4/98.6/92.8**	**97.8/99.6/97.9**	**95.0/99.1/93.4**

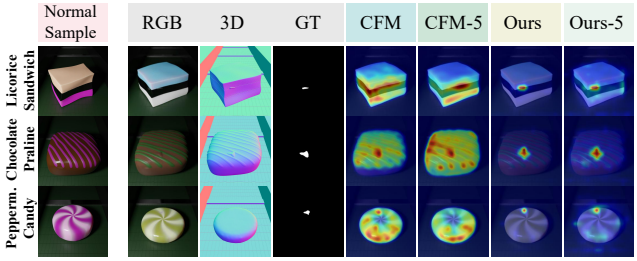


Figure 7. Comparison of localization heatmaps between the proposed method and CFM (CFM-5 denotes CFM trained under the 5-shot setting).

efficiently aggregates cues from the four MRI modalities to accurately localize brain tumors in low-contrast regions, as shown in Fig. 6. Overall, the complementary prototype design effectively enhances cross-modal consistency and captures crucial cues across diverse domains, enabling adaptive and precise defect localization.

Fig. 7 illustrates qualitative comparisons under few-shot settings for a distribution-diverse scenario. Under the 5-shot condition, CFM exhibits more severe misdetections, while our CPMAD localization maps focus more precisely on defects. Specifically, in scenarios involving subtle geometric perturbations co-occurring with complex RGB textures (e.g., *Chocolate Praline*), CPMAD shows a more significant advantage, demonstrating that the complementary prototype design can alleviate mapping ambiguity and improve localization performance even in low-data domains.

4.4. Ablation Studies

4.4.1. Architecture ablation

Quantitative analysis. We conduct ablation studies to evaluate the contribution of each prototype component. Tab. 5 presents results on the MVTec-3D dataset to evaluate the contribution of each prototype component. We design three configurations: (1) **Baseline** employs paired ViT encoders followed by a 3-layer MLP that performs hierarchical cross-modal mapping. (2) **+Consensus**, which incorporates consensus prototypes for feature mapping; and (3) **+All (Consensus+Supplementary)**, which combines both components into the full complementary prototype learning framework. "+Consensus" variant yields an I-AUROC improvement of 6.5% over the baseline. Combining both prototypes further boosts performance, highlighting their synergistic effect.

Qualitative analysis. Fig. 8 visualizes the effects of incorporating supplementary and consensus prototypes. The baseline model, which relies on unconditional mappings, generates ambiguous localization maps in regions with diverse distributions. For instance, in the *Confetto* class, varying product colors cause normal areas to be misclassified as anomalous in the 3D→RGB mapping. Similarly, in the *Cookie* class, the model misinterprets normal dark spots as defective holes in the RGB→3D mapping. By incorporating both complementary prototypes, the model dynamically enhances the adaptive reconstruction of normal regions, resulting in clearer and more accurate defect boundaries.

Distribution analysis. Fig. 9 shows the kernel density estimation of image-level anomaly scores for the three con-

Table 5. Ablation study of Consensus Prototypes (Con) and Supplementary Prototypes (Sup) on the MVTec-3D dataset.

Architecture			Performance		
Baseline	Con	Sup	I-AUROC	P-AUROC	AUPRO
✔	✘	✘	89.5	99.2	96.4
✔	✔	✘	95.9	99.4	96.9
✔	✔	✔	**97.8**	**99.6**	**97.9**

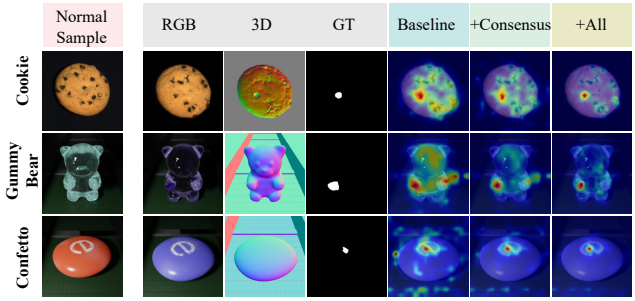


Figure 8. Qualitative ablation of complementary prototypes.

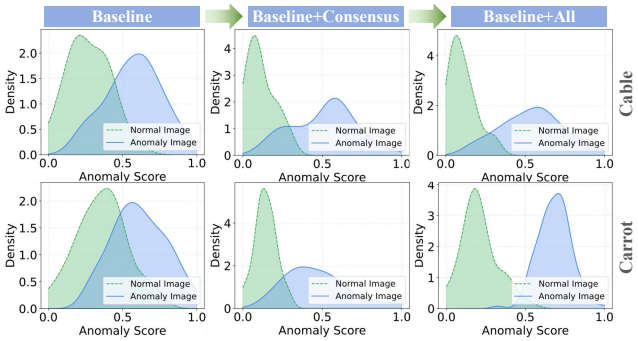


Figure 9. Kernel Density Estimation of anomaly scores for three configuration on the MVTec-3D.

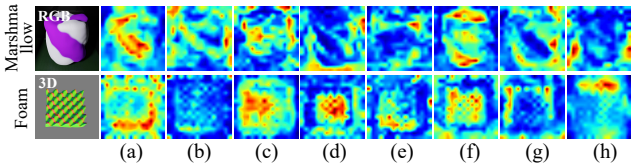


Figure 10. Visualization of supplementary prototypes. (a)-(h) Attention maps of eight different prototypes.

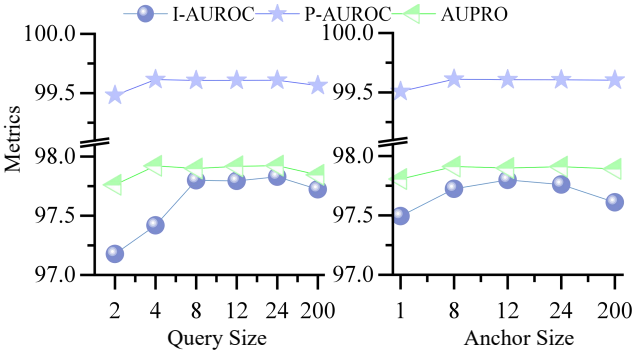


Figure 11. Ablation on SQM Query size and CEM Anchor size.

figurations. The complementary design markedly reduces the distribution overlap between normal and anomalous samples, which can be attributed to its use of both prototypes for restoring normal regions while reconstructing anomalous regions in a way that enhances their distinguishability.

Visualization of prototypes. Fig. 10 illustrates the attention maps of supplementary prototypes for representative RGB and 3D samples. For the *Marshmallow* RGB, the supplementary prototypes concentrate on regions exhibiting large color variations, such as the purple bulges (a, f). In contrast, for the uneven *Foam* 3D surface, they emphasize ambiguous regions, including central depressions (c, d, f) and discontinuous boundary areas (a, e, g, h). These visualizations demonstrate that the supplementary prototypes dynamically cooperate with the consensus prototypes to provide adaptive priors, thereby mitigating mapping ambiguity.

4.4.2. Modules Ablation

Tab. 6 presents detailed ablation on each proposed module. (1) For the SQM, replacing the CR-Attention with standard cross-attention, which uses the queries to extract supplementary prototypes from modality-specific features, results in a noticeable performance degradation. This indicates that when the consensus prototypes are not visible, the supplementary queries may capture redundant shared semantics rather than focusing on complementary modality-specific cues. (2) For the CEM, replacing the non-learnable average with a learnable weighted average also reduces performance, suggesting that learnable aggregation may exploit shortcut cues to reconstruct anomalies. Moreover, replacing the dynamic anchor with a static one (i.e., a learnable embedding) fails to adaptively capture the cross-modal semantics, thereby degrading cross-modal consistency and potentially introducing anomalous patterns into the prototype space. (3) Replacing the Complementary Mapping module with standard cross-attention leads to further degradation, as the traditional token-wise residual introduces more noisy responses from anomalous regions compared to the channel-wise residual.

Table 6. Detailed ablation of modules on the MVTec-3D dataset.

Modules	I-AUROC	P-AUROC	AUPRO
W/o CR-Attention (SQM)	96.8	99.4	97.5
W/o Average (CEM)	97.3	99.5	97.7
W/o Dynamic Anchor (CEM)	97.2	99.5	97.6
W/o Channel-wise Residual	97.6	**99.6**	97.8
CPMAD (Ours)	**97.8**	**99.6**	**97.9**

Table 7. Performance and efficiency comparison of MAD methods and our CPMAD variants with ViT-S/B backbones using an NVIDIA RTX 4090 and RTX 2080Ti, with I-AUROC on MVTec-3D/Eyecandies. "*" are taken from the original publications.

Methods	Metric↑	FPS↑ (4090)	FPS↑ (2080Ti)	FLOPs↓ (GFlops)	Param.↓ (M)	Memory↓ (MB)
G^2SF^* [26]	97.1/90.2	6.04	-	-	-	-
3D-A.* [18]	95.1/94.6	24.70	-	-	-	-
M3DM [29]	94.5/88.2	2.07	1.10	-	-	8402
CFM [10]	95.4/88.1	13.18	5.71	431.14	112.56	3757
EasyNet [8]	92.6/86.9	36.54	11.97	465.65	140.61	2746
CPMAD-S	96.6/93.3	**101.73**	**60.49**	**36.34**	**35.55**	**1058**
CPMAD	**97.8/95.0**	64.66	25.73	145.01	141.88	1663

4.4.3. Hyper-parameter and Efficiency Analysis

Hyper-parameter sensitivity. Fig. 11 shows the sensitivity of CPMAD to query size in SQM and anchor size in CEM. For SQM, query sizes larger than 8 maintain optimal performance, but excessively large sizes may leak anomalous information. Similarly, CEM exhibits a comparable trend. Overall, CPMAD remains robust to hyperparameter variations within a reasonable range.

Inference Efficiency. As shown in Tab. 7, the lightweight CPMAD-S variant achieves rapid inference, approaching millisecond-level latency on an RTX 4090 GPU, while maintaining real-time performance at 64.66 FPS even on a less powerful RTX 2080 Ti. Furthermore, it demonstrates reduced computational cost and memory footprint compared to existing approaches. This efficiency stems from its complementary prototype-based lightweight architecture, which avoids the computational overhead of heavy decoders, making the method well-suited for practical deployment.

5. Conclusion

We propose CPMAD, a lightweight and scalable framework that dynamically learns consensus and supplementary prototypes to mitigate mapping ambiguity. Extensive experiments show that CPMAD achieves superior accuracy, faster inference, and reduced memory usage, while scaling efficiently with model size and data volume. These results indicate that CPMAD provides an efficient and generalizable solution for practical cross-modal inspection systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62431004.

References

[1] Mohammad Adiban, Marco Siniscalchi, Kalin Stefanov, and Giampiero Salvi. Hierarchical residual learning based vector quantized variational autoencorder for image reconstruction and generation. In 33rd British Machine Vision Conference, 2022.

[2] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4042–4053, 2024.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten 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, pages 9592–9600, 2019.

[4] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. arXiv preprint arXiv:2112.09045, 2021.

[5] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In Proceedings of the Asian Conference on Computer Vision, pages 3586–3602, 2022.

[6] Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. Pattern Recognition, 156:110761, 2024.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.

[8] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection. In Proceedings of the 31st ACM International Conference on Multimedia, pages 7038–7046, 2023.

[9] Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3d anomaly detection. In Proceedings of the 40th International Conference on Machine Learning, pages 6185–6194, 2023.

[10] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17234–17243, 2024.

[11] 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.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[13] Zhihao Gu, Jiangning Zhang, Liang Liu, Xu Chen, Jinlong Peng, Zhenye Gan, Guannan Jiang, Annan Shu, Yabiao Wang, and Lizhuang Ma. Rethinking reverse distillation for multi-modal anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 8445–8453, 2024.

[14] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 20405–20415, 2025.

[15] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2968–2977, 2023.

[16] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11375–11385, 2024.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[18] Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 12273–12281, 2025.

[19] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 9974–9983, 2025.

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[21] YeongHyeon Park, Myung Jin Kim, and Hyeong Seok Kim. Contrastive language prompting to ease false positives in medical anomaly detection. In 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), pages 1–4. IEEE, 2025.

[22] 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.

[23] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2592–2602, 2023.

[24] Yikang Shi, Xin Zhan, Yaqian Li, Zhongqiang Wu, Wenming Zhang, and Haibin Li. Cycle-cfm: An unsupervised framework for robust multimodal anomaly detection in industrial settings. Expert Systems with Applications, page 129745, 2025.

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017.

[26] Chengyu Tao, Xuanming Cao, and Juan Du. G2sf: Geometry-guided score fusion for multimodal industrial anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 20551–20560, 2025.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[28] Chengjie Wang, Haokun Zhu, Jinlong Peng, Yue Wang, Ran Yi, Yunsheng Wu, Lizhuang Ma, and Jiangning Zhang. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[29] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8032–8041, 2023.

[30] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 6074–6082, 2024.

[31] Haiyan Yang, Shuning Liu, Mingxuan Wu, Hongbin Chen, and Delu Zeng. Attention-based misaligned spatiotemporal auto-encoder for video anomaly detection. Signal, Image and Video Processing, pages 1–13, 2024.

[32] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems, 35:4571–4584, 2022.

[33] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In European conference on computer vision, pages 539–554. Springer, 2022.

[34] Yuan Zhao, Youwei Pang, Lihe Zhang, Hanqi Liu, Jiaming Zuo, Huchuan Lu, and Xiaoqi Zhao. Unimmad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decompression. arXiv preprint arXiv:2509.25934, 2025.

[35] Yuan Zhao, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Focusclip: Focusing on anomaly regions by visual-text discrepancies. IEEE Transactions on Circuits and Systems for Video Technology, 2024.

[36] Qihang Zhou, Jiangtao Yan, Shibo He, Wenchao Meng, and Jiming Chen. Pointad: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection. Advances in Neural Information Processing Systems, 37:84866–84896, 2024.

[37] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. Advances in Neural Information Processing Systems, 37:89943–89977, 2024.

[38] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In European Conference on Computer Vision, pages 392–408. Springer, 2022.