Beyond Single-Modal Boundary: Cross-Modal Anomaly Detection through Visual Prototype and Harmonization

Kai Mao, Ping Wei*, Yiyang Lian, Yangyang Wang, Nanning Zheng
National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
kaimao@stu.xjtu.edu.cn, pingwei@xjtu.edu.cn

Abstract

Anomaly detection is a significant task for its application and research value. While existing methods have made impressive progress within the same modality, cross-modal anomaly detection remains an open and challenging problem. In this paper, we propose a cross-modal anomaly detection model that is trained using data from a variety of existing modalities and can be generalized well to unseen modalities. The model consists of three major components: 1) the Transferable Visual Prototype directly learns normal/abnormal semantics in visual space; 2) the Prototype Harmonization strategy adaptively utilizes the Transferable Visual Prototypes from various modalities for inference on the unknown modality; 3) the Visual Discrepancy Inference under the few-shot setting enhances performance. In the zero-shot setting, the proposed method achieves AUROC improvements of 4.1%, 6.1%, 7.6%, and 6.8% over the best competing methods in the RGB, 3D, MRI/CT, and Thermal modalities, respectively. In the few-shot setting, our model also achieves the highest AUROC/AP on ten datasets in four modalities, substantially outperforming existing methods. Codes are available at https://github.com/Kerio99/CMAD.

1. Introduction

Identifying various unknown anomalies only given normal samples is an essential and challenging problem. It has attracted increasing attention for its practical significance in numerous scenarios such as industrial quality inspection and medical diagnosis[2, 9, 12, 13, 22, 25, 27, 29, 30, 33].

For the recent success of vision-language models [15, 26, 39, 41], some methods [17, 20, 40] focus on detecting unknown anomalies in zero-shot setting. A strong correlation between the auxiliary datasets and test images is necessary in this setting, indicating that they should belong to the same modality. However, in practical settings, the requirement to perform anomaly detection on cross-modal data frequently arises across various scenarios. For example, sensors with varying wavelengths in industry are used to capture different modalities of product images, such as RGB, 3D, and thermal. Instruments across different spectral ranges in medicine are utilized to produce MRI and CT scan images. The ability of detecting anomalies on cross-modal data is becoming a necessity. We define the cross-modal anomaly detection task as: the anomaly detection model is trained with data of known modalities and tested on data of unseen modalities.

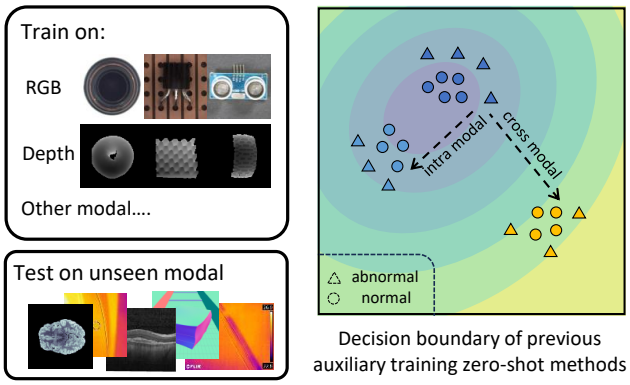

Figure 1. Illustration of cross-modal anomaly detection. Data from different modalities exhibit a more dispersed distribution in the feature space, making the decision boundaries developed by previous zero-shot methods on the auxiliary dataset inapplicable in cross-modal scenarios.

Cross-modal anomaly detection can be regarded as a zero-shot problem. Thus existing zero-shot methods can be tentatively employed for this task. These methods [10, 11, 17, 40] utilize prior knowledge about normal/abnormal semantics in the text space to guide anomaly detection in the visual space using vision-language models like CLIP [26]. However, although the text and visual space have been aligned to some extent in these vision-language mod-

*Corresponding author.

els, a modality gap still exists [21, 31]. Consequently, learning normal/abnormal semantics in the text space may suffer from modality bias, which does not accurately capture the normal/abnormal information in the visual space. Besides, these zero-shot methods require a strong correlation between the auxiliary datasets and test images. When the auxiliary datasets and test images belong to different modalities, as shown in Figure 1, the performance of these methods significantly declines due to the salient discrepancies between data from different modalities.

Constructing generalized universal representations about anomalies using data from known modalities can help mitigate the above discrepancies. The integration of these representations can further address the discrepancies between different modalities. Inspired by the observations, we propose a cross-modal anomaly detection method consisting of the Transferable Visual Prototype and a Prototype Harmonization strategy. The Transferable Visual Prototype (TVP) aims to learn transferable anomaly prototypes from known modalities while avoiding modality bias by directly learning normal/abnormal semantics in the visual space. It addresses diverse anomaly manifestations by leveraging both global representations and fine-grained insights to distinguish normal conditions from anomalies. TVPs trained on datasets from various modalities depict the concept of anomaly from different perspectives and exhibit significant complementarity. The Prototype Harmonization strategy seeks to leverage this complementarity to construct a more generalizable concept of anomaly by employing TVPs developed across different modalities to guide anomaly detection on unknown modalities. It adaptively utilizes TVPs from various modalities for inference, assigning different weights based on the visual characteristics of the images and their similarity to the prototypes.

Besides, in many scenarios, a small number of normal samples from unknown modalities can be obtained during testing. These few normal samples can provide prior information about normality. Our model can also estimate anomalies by measuring the discrepancy between test images and these few normal samples through the proposed Visual Discrepancy Inference. It can seamlessly integrates with prototype harmonization, further improving performance in few-sample scenarios.

Extensive experiments on ten datasets from four modalities demonstrate the effectiveness of our method. In the zero-shot setting, our method achieve the best performance on all modalities on average, while in the few-shot setting, it outperforms other methods across all ten datasets.

Our key contributions are summarized below:

- We propose the task of cross-modal anomaly detection to evaluate the generalization ability of various methods in identifying anomalies from unknown modalities, highlighting its challenge and application value.
- We propose a method consisting of Transferable Visual Prototype and Prototype Harmonization strategy for cross-modal anomaly detection. And we propose the Visual Discrepancy Inference process, further enhancing the performance when few samples are available at test phase.
- Extensive experiments are conducted on ten datasets from four different modalities. The quantitative results demonstrate the effectiveness of our method.

2. Related Work

2.1. Zero-shot Anomaly Detection

Existing anomaly detection methods can be categorized into several major branches, such as embedding-based methods [12, 13, 29], reconstruction-based methods [34–36, 38], and knowledge distillation-based methods [3, 5, 14, 37], most of which are conducted under the premise of having a large number of normal samples. Recently, due to the success of pre-trained vision-language models such as CLIP [17], their robust language-driven zero-shot capabilities have been applied in the field of anomaly detection. The prior knowledge about normal/abnormal semantics in the text space is utilized to guide anomaly detection in the visual space. Some methods [11, 17] set fixed textual descriptions of normal/abnormal semantics, which are encoded by the text encoder of the multi-modal model to generate corresponding features in the text space. Other methods [20, 40] incorporate learnable word embeddings into the textual descriptions, aiming to learn more flexible and general information about normal/abnormal semantics by training on the auxiliary datasets. These methods mark a significant contribution towards more general anomaly detection.

WinCLIP [17] is one of the pioneers using vision-language models for anomaly detection. It describes normal and abnormal semantics using numerous hand-crafted text prompts. During inference on test images, anomalies are detected by comparing the similarity between the images' visual features and the text features. WinCLIP has achieved remarkable zero-shot performance due to alignment between vision and language. Building on WinCLIP, VAND [11] introduces learnable projection techniques to further align local visual and textual features. AnomalyCLIP [40] designs an object-agnostic prompt template to learn generic abnormal and normal prompts through the use of auxiliary data. This approach significantly simplifies the process of prompt design by eliminating the need for task-specific templates. AdaCLIP [10] employs both static and dynamic parts of learnable prompts for anomaly learning and designs a hybrid-semantic fusion module to extract region-level context about anomaly regions. However, these methods all describe or learn about anomalies in the textual embedding space. Given the modality gap [21, 31] present in multi-modal models, their learning about anomalies may

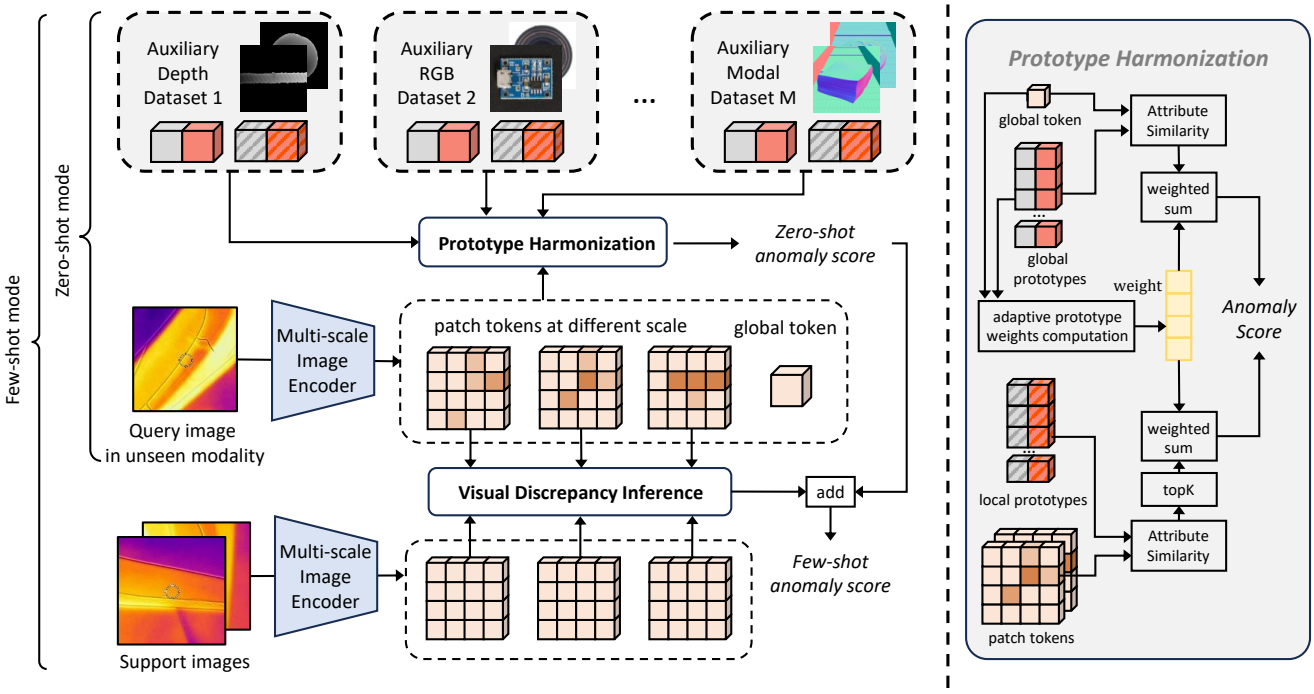

Figure 2. Overview of our framework for inference in zero and few-shot settings. In the zero-shot setting, the query image are input to a multi-scale image encoder to extract global and local features in the visual space. Subsequently, Transferable Visual Prototypes from different datasets are integrated using our designed Prototype Harmonization strategy for anomaly detection. In the few-shot setting, local features of support images are firstly extracted through the multi-scale image encoder. The visual discrepancy is then inferred by comparing the differences in local features between the query image and the support images. This discrepancy information is combined with the prediction results from the zero-shot branch to serve as the few-shot results.

be biased. Our method learns directly in the visual space, allowing for a more accurate and comprehensive representation of anomalies.

2.2. Few-shot Anomaly Detection

The goal of few-shot anomaly detection is to detect anomalies with only a small number of normal images provided. Traditional methods based on embedding distance, such as SPADE [12], PaDiM [13], and Patchcore [29], calculate the anomaly score of test images based on the distance from the features of the few available samples, requiring no training but offering limited performance. Inspired by the human ability to detect anomalies by comparing an query image with normal images, RegAD [16] enhances the performance under few-shot setting by training a feature registration module using an additional dataset to align features. However, the relevance between test images and images in the additional dataset is required. PromptAD [20] explores the feasibility of prompt learning under few-shot conditions. It transposes the semantics of normal prompts by concatenating anomaly suffixes and use the explicit anomaly margin to control the feature distance. InCTRL [42] employs in-context residual learning to distinguish anomalies by detecting residuals between test images and in-context few-shot normal sample prompts. These methods have significantly advanced the development of few-shot anomaly detection. However, they rely on relatively strong correlations between auxiliary training data and test set, which limits their performance in cross-modal scenarios.

3. Method

3.1. Overview

For the cross-modal anomaly detection task, consider a training set containing M datasets from several different modalities, denoted as $\mathcal{X}_{train} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_M\}$. Let $\mathcal{D}_m = \{X_m, Y_m\}$ denote the m-th dataset from $\mathcal{X}_{train}$, where $X_m = \{x_i\}, i = 1, ..., N_m$ contains N_m normal and anomalous images, and $Y_m = \{y_i\}, i = 1, ..., N_m$ represents the corresponding labels. $y_i = 0$ indicates a normal image, and $y_i = 1$ indicates an anomalous image. The dataset from an unseen modality, $\mathcal{D}_U = \{X_{test}, Y_{test}\}$, is used as the test set. The goal of the cross-modal anomaly detection task is to train a generalized anomaly scoring model on the training set that can effectively generalize to an unseen test set from a novel modality. In zero-shot setting, no extra data is available other than the test query image. In the few-shot setting, a support set $\mathcal{I} =$

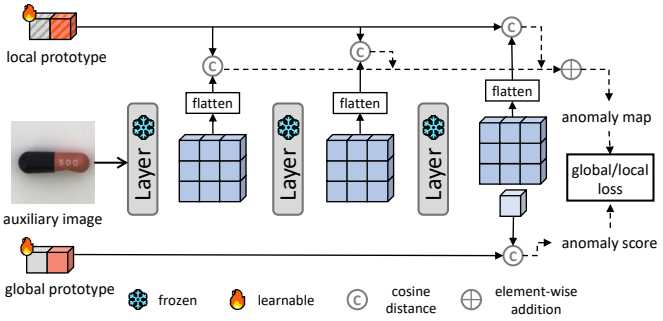

Figure 3. Overview of the training process of our Transferable Visual Prototype. The normal and abnormal semantics are learned by maximizing the cosine similarity between the Transferable Visual Prototype and the visual embedding of the auxiliary image for both global and local parts.

$\{I_1, I_2, ..., I_K\}$ consisting of several normal images from the target modality is provided. However, $\mathcal{I}$ remains unknown during model training.

Our approach comprises two principal phases: training and inference. Figure 2 demonstrates our approach for zero-shot and few-shot inference. In the zero-shot setting, Transferable Visual Prototypes trained on each dataset from different modalities are integrated using our designed Prototype Harmonization strategy to predict anomalies. Under the few-shot setting, embeddings in the visual space of query and support images are analyzed through Visual Discrepancy Inference to deduce anomaly scores, further enhancing the detection capabilities. Figure 3 demonstrates the training phase of our method. For each dataset $\mathcal{D}_m$ within $\mathcal{X}_{train}$, we independently train a Transferable Visual Prototype to delineate the concepts of normality and anomaly within that particular dataset.

3.2. Transferable Visual Prototype

The existing zero-shot methods heavily rely on the text space of vision-language models. WinCLIP [17] designs fixed text prompts, utilizing features from its text space to guide the anomaly detection in the visual space. PromptAD [20] and AnomalyCLIP [40] incorporate learnable word embeddings into the text prompts, facilitating the detection by learning on these embeddings. Although the vision and language modality have been aligned to some extent in the joint embedding space in vision-language models like CLIP [26], ImageBind [15] and LanguageBind [41], a modality gap still exists [21, 31]. Consequently, learning normal/abnormal semantics in the text space may suffer from modality bias, which does not accurately capture the normal/abnormal information in the visual space.

Inspired by the observations above, we design the Transferable Visual Prototype to capture and articulate prototypical expressions of normality and anomaly, ensuring effective transferability across different modalities. It comprises two parts: $\{\boldsymbol{p}_g, \boldsymbol{p}_l\}$, where $\boldsymbol{p}_g$ and $\boldsymbol{p}_l$ represent the global and loacal part, respectively. The purpose of the global part is to learn normal/abnormal characteristics on the global visual embeddings of images. The local part is designed to capture normal/abnormal features in fine-grained, local abnormal regions. For both global and local components, we adopt two prototypes: one prototype describes features of normality, while the other describes features of abnormalities, as shown in Eq. (1):

$$\boldsymbol{p}_g = \{p_g^n, p_g^a\}, \boldsymbol{p}_l = \{p_l^n, p_l^a\}. \tag{1}$$

p_g^n and p_g^a represent the global part describing normal and abnormal features, respectively. p_l^n and p_l^a are the local part describing normal and abnormal features, respectively.

We optimize the Transferable Visual Prototype through the following process. Given a dataset $\mathcal{D}_m$ from the training set $\mathcal{X}_{train}$, let $f_g^i \in \mathbb{R}^C$ denote its global visual embedding and $f_l^i \in \mathbb{R}^{h \times w \times C}$ denote its local visual embedding for x_i in $\mathcal{D}_m$. Both f_g^i and f_l^i are generated by the visual encoder of CLIP [26]. We define an anomaly score function to measure the similarity between a feature and the visual prototype, with f_g^i and $\boldsymbol{p}_g$ as an example:

$$S(f_g^i, \boldsymbol{p}_g) = \frac{exp(\langle p_g^a, f_g^i \rangle / \tau)}{exp(\langle p_g^n, f_g^i \rangle / \tau) + exp(\langle p_g^a, f_g^i \rangle / \tau)}, \tag{2}$$

where τ is a temperature hyperparameter. $\langle \cdot, \cdot \rangle$ represents the cosine similarity operation. The anomaly score based on the global part of TVP is denoted as $s_g = S(f_g^i, \boldsymbol{p}_g)$. We compute the cross-entropy loss between s_g and the label y_i as the loss of the global part:

$$L_{global} = cross\_entropy(s_g, y_i). \tag{3}$$

Let $s_l = \{s_l^{u,v} | u \in [1, h], v \in [1, w]\}$ denote the fine-grained prediction result for x_i. $s_l^{u,v}$ denotes the score at location (u, v), where h and w are the feature dimensions. $s_l^{u,v}$ is calculated as $s_l^{u,v} = S(f_l^{i,u,v}, \boldsymbol{p}_l)$, where $f_l^{i,u,v}$ represents the local visual embedding of f_l^i at the location (u, v). Let $G \in \mathbb{R}^{H \times W}$ be the ground-truth mask. We upsample s_l to the original image size, denoted as $Up(s_l) \in \mathbb{R}^{H \times W}$. We use the following loss to optimize the local part:

$$\begin{aligned} L_{local} = & focal(Up([\mathbb{1} - s_l, s_l]), G) + \\ & dice(Up(\mathbb{1} - s_l), \mathbb{1} - G) + \\ & dice(Up(s_l), G), \end{aligned} \tag{4}$$

where $focal(\cdot, \cdot)$ and $dice(\cdot, \cdot)$ represent the focal loss [28] and dice loss [19], respectively. The operator $[\cdot, \cdot]$ denotes the concatenation along with the channel, and $\mathbb{1}$ represents the full-one matrix. To utilize visual features from different

levels of abstraction, we conduct learning on local features across multiple layers. Let $\mathcal{T}$ be the set of layers we used, we minimize the following loss to learn the Transferable Visual Prototype:

$$L_{total} = L_{global} + \beta \sum_{t \in \mathcal{T}} L^t_{local}, \tag{5}$$

where L^t_{local} denotes the loss calculated using the local feature from layer t.

3.3. Inference for Zero-Shot

Considering that TVPs trained on datasets from different modalities portray the concepts of normality and abnormality from diverse perspectives, they may demonstrate strong complementarity. Synthesizing these TVPs from various modalities, a more comprehensive depiction of the normal/abnormal concepts can be achieved.

Based on the above observations, we design a prototype harmonization strategy for inference. Let $\boldsymbol{p}_{g,m}$ and $\boldsymbol{p}_{l,m}$ denote the global and local parts of Transferable Visual Prototype learned from $\mathcal{D}_m$, respectively. $\mathcal{P}_g$ and $\mathcal{P}_l$ contain all the global and local Transferable Visual Prototypes, respectively:

$$\begin{aligned} \mathcal{P}_g &= \{\boldsymbol{p}_{g,m}\}, m = 1, ..., M \\ \mathcal{P}_l &= \{\boldsymbol{p}_{l,m}\}, m = 1, ..., M. \end{aligned} \tag{6}$$

For each image $\hat{x}_i$ in the test set $\mathcal{D}_U$, let $\hat{f}^i_g \in \mathbb{R}^C$ denote the global visual embedding of $\hat{x}_i$, $\hat{f}^i_l \in \mathbb{R}^{h \times w \times C}$ denote the local visual embedding of $\hat{x}_i$ and $\hat{f}^{i,u,v}_l$ indicates the local visual embedding of $\hat{f}^i_l$ at location (u, v). We calculate the cosine similarity between $\hat{f}^i_g$ and each element in $\mathcal{P}_g$:

$$d_m = Max(\left\langle p^n_{g,m}, \hat{f}^i_g \right\rangle, \left\langle p^a_{g,m}, \hat{f}^i_g \right\rangle), \tag{7}$$

where $p^n_{g,m}, p^a_{g,m}$ denote the normal and abnormal part of global prototype $\boldsymbol{p}_{g,m}$, respectively. We use d_m as the relevance between $\hat{x}_i$ and the Transferable Visual Prototype on the m-th dataset in certain modality, with higher values of d_m indicating larger relevance. The softmax operation is then used to get the weight related to each global Transferable Visual Prototype, as illustrated in Eq. (8).

$$w_m = \frac{exp(d_m/\tau')}{\sum_{m=1}^{M} exp(d_m/\tau')} \tag{8}$$

Then, the prediction of anomaly score is calculated:

$$\hat{s}_g = \textstyle\sum_{m=1}^{M} w_m * S(\hat{f}^i_g, \boldsymbol{p}_{g,m}). \tag{9}$$

For fine-grained prediction, we employ the same computational manner, with weights consistent with w_m:

$$\hat{s}^{u,v}_l = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (\textstyle\sum_{m=1}^{M} w_m * S(\hat{f}^{i,u,v}_{l,t}, \boldsymbol{p}_{l,m})), \tag{10}$$

where $\hat{f}^{i,u,v}_{l,t}$ represents the local visual embedding from layer t.

The final prediction score for zero-shot setting integrates the prediction results based on both global and local perspectives.

$$s_z = \lambda * \hat{s}_g + (1 - \lambda) * Topk(\hat{s}_l) \tag{11}$$

where $Topk(\cdot)$ is the operation of selecting the top k largest predicted pixel values from the fine-grained prediction results and computing their average.

3.4. Inference with Few-Normal-Shots

Across many scenarios, securing a limited quantity of normal images during the testing period is achievable. Hence, we propose the Visual Discrepancy Inference to augment the detection performance within the aforementioned zero-shot setting.

Given the few normal images $\mathcal{I} = \{I_1, I_2, ..., I_K\}$ belong the same modality as the test image $\hat{x}_i$, we first extract the local visual features of the images in $\mathcal{I}$ and put features from all locations into $\mathcal{M}$. Subsequently, we infer potential anomalies by calculating the visual discrepancy between $\hat{f}^i_l$ and $\mathcal{M}$, as illustrated in Eq. (12).

$$q^{u,v} = \frac{1}{2}(1 - \left\langle \hat{f}^{i,u,v}_l, h(\hat{f}^{i,u,v}_l | \mathcal{M}) \right\rangle) \tag{12}$$

$q^{u,v}$ represents the prediction result at the spatial location (u, v) based on the visual discrepancy, and $h(\hat{f}^{i,u,v}_l | \mathcal{M})$ returns the feature in $\mathcal{M}$ that has the highest similarity to $\hat{f}^{i,u,v}_l$. We aggregate the predictions from multiple layers to formulate the few-shot prediction, as follows:

$$q^{u,v} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} q^{u,v}_t, \tag{13}$$

where $q^{u,v}_t$ represents the prediction result using the local feature from layer t. Let $q = \{q^{u,v} | u \in [1, h], v \in [1, w]\}$ denote the prediction result. The prediction from the zero-shot setting is integrated as part of the few-shot prediction results.

$$s_f = \lambda' * s_z + (1 - \lambda') * Topk(q) \tag{14}$$

4. Experiment

4.1. Experimental Details

Datasets and Cross-Modal Setting. We utilize ten datasets from four different modalities to validate the effectiveness of our method for cross-modal anomaly detection task. These datasets include the RGB modality (MvTec AD[4], VisA [43], BTAD [24], MPDD [18]), 3D modality (MvTec 3D-AD [6], Eyecandies [8]), MRI/CT modality (Brain MRI [1], OCT2017 [1]), and thermal modality

Metric	Methods	Publication	RGB				3D		MRI/CT		Thermal	
			MvTec AD	VisA	BTAD	MPDD	MvTec 3D-AD	Eyecandies	Brain MRI	OCT2017	WTB	Leaves
AUROC	VAND[11]	*CVPRW 23*	86.1	77.5	73.5	**76.8**	65.5	70.3	54.7	46.7	69.9	80.6
	WinCLIP[17]	*CVPR 23*	**91.8**	**78.1**	68.2	63.6	70.5	73.9	71.3	49.1	86.3	66.7
	AnomalyCLIP[40]	*ICLR 24*	86.1	74.8	82.6	**74.9**	63.8	74.5	64.5	**85.2**	86.6	**85.7**
	AdaCLIP[10]	*ECCV 24*	90.4	76.5	**95.3**	70.6	**74.1**	**77.0**	**73.1**	68.8	**87.8**	72.5
	Ours	*CVPR 25*	**92.3**	**86.8**	**97.5**	72.8	**79.6**	**83.7**	**73.7**	**91.2**	**94.8**	**91.1**
AP	VAND[11]	*CVPRW 23*	93.5	80.9	69.7	**82.5**	87.3	71.2	86.2	75.8	75.3	96.6
	WinCLIP[17]	*CVPR 23*	**96.5**	**81.2**	70.9	69.9	89.0	74.2	**91.9**	77.8	89.0	92.9
	AnomalyCLIP[40]	*ICLR 24*	93.8	78.6	81.6	**81.9**	87.5	77.1	87.9	**94.9**	88.2	**97.6**
	AdaCLIP[10]	*ECCV 24*	95.9	78.9	**97.5**	73.1	**90.6**	**79.7**	90.8	87.0	88.8	95.3
	Ours	*CVPR 25*	**96.3**	**89.3**	**98.9**	79.7	**93.1**	**86.0**	**91.8**	**97.1**	**95.1**	**98.6**

Table 1. AUROC% and AP% results of different methods under the zero-shot setting. The best and second-best performance results are highlighted in red and blue, respectively.

Setup	Methods	RGB				3D		MRI/CT		Thermal	
		MvTec AD	VisA	BTAD	MPDD	MvTec 3D-AD	Eyecandies	Brain MRI	OCT2017	WTB	Leaves
1 shot	PatchCore[29]	(83.8, 92.6)	(77.5, 80.8)	(85.4, 86.8)	(61.6, 65.4)	(59.3, 84.7)	(65.5, 67.3)	(**73.1**, 91.9)	(75.0, 89.3)	(56.3, 57.0)	(88.5, 97.7)
	RegAD[16]	(73.3, 87.7)	(62.3, 66.7)	(84.2, 90.0)	(53.1, 62.5)	(60.5, 84.6)	(61.9, 64.7)	(47.3, 81.9)	(45.4, 74.8)	(55.9, 56.7)	(78.0, 95.1)
	WinCLIP[17]	(91.1, 95.9)	(80.1, 81.6)	(84.9, 82.4)	(64.6, 70.3)	(73.3, 90.2)	(75.6, 76.1)	(72.5, **92.2**)	(52.8, 79.5)	(**89.9**, **91.4**)	(70.9, 93.9)
	PromptAD[20]	(**93.1**, **96.7**)	(**86.7**, **88.5**)	(**91.4**, **94.3**)	(73.0, 76.7)	(75.4, 91.1)	(**79.7**, **82.3**)	(57.6, 85.5)	(74.6, 88.8)	(85.3, 86.9)	(**92.9**, **98.6**)
	InCTRL[42]	(92.1, 94.9)	(85.4, 86.4)	(73.3, 82.9)	(**74.5**, **76.8**)	(**78.2**, **92.3**)	(76.9, 81.5)	(71.2, 91.2)	(**86.5**, **94.7**)	(88.5, 90.6)	(87.9, 97.4)
	Ours	(**96.0**, **97.5**)	(**91.4**, **92.7**)	(**97.5**, **98.9**)	(**79.6**, **84.3**)	(**82.3**, **94.2**)	(**82.2**, **85.8**)	(**77.2**, **93.1**)	(**90.9**, **97.0**)	(**94.0**, **94.8**)	(**94.5**, **99.0**)
2 shot	PatchCore[29]	(87.1, 93.8)	(81.5, 83.4)	(**90.7**, 88.5)	(60.1, 67.2)	(64.6, 87.0)	(69.0, 72.1)	(74.2, 91.5)	(88.6, 95.4)	(56.4, 56.8)	(91.7, 98.4)
	RegAD[16]	(75.6, 88.4)	(65.5, 69.6)	(89.1, 91.9)	(53.2, 61.6)	(65.0, 86.3)	(66.6, 68.3)	(51.4, 82.3)	(43.4, 72.8)	(55.9, 55.4)	(78.0, 95.2)
	WinCLIP[17]	(91.7, 96.1)	(81.9, 82.8)	(87.3, 88.1)	(64.9, 70.4)	(73.7, 90.2)	(76.7, 77.1)	(72.4, 92.2)	(55.4, 80.8)	(**90.2**, **91.6**)	(72.7, 94.2)
	PromptAD[20]	(**94.8**, **97.5**)	(**87.5**, **89.5**)	(90.6, **94.9**)	(**76.4**, **78.4**)	(**79.2**, **92.6**)	(**79.1**, **80.6**)	(72.9, 90.9)	(89.4, **95.9**)	(86.5, 87.6)	(**93.6**, **98.7**)
	InCTRL[42]	(91.1, 94.7)	(86.9, 88.2)	(86.6, 87.8)	(75.3, 76.8)	(78.8, 92.5)	(76.8, 79.5)	(**76.4**, **93.3**)	(**89.7**, 95.2)	(86.5, 87.8)	(91.5, 98.5)
	Ours	(**96.5**, **97.8**)	(**92.5**, **93.6**)	(**96.8**, **98.8**)	(**83.2**, **87.2**)	(**83.8**, **94.6**)	(**83.2**, **86.4**)	(**78.0**, **93.5**)	(**93.7**, **97.9**)	(**94.2**, **95.0**)	(**93.8**, **98.9**)
4 shot	PatchCore[29]	(87.8, 94.1)	(85.0, 86.4)	(92.3, 93.7)	(68.3, 71.5)	(70.4, 88.9)	(70.3, 72.7)	(75.3, 91.8)	(90.3, 96.1)	(63.6, 63.8)	(**93.8**, **98.7**)
	RegAD[16]	(80.0, 89.8)	(70.7, 73.0)	(88.8, 94.7)	(57.7, 64.7)	(64.7, 86.5)	(69.4, 71.2)	(57.4, 84.4)	(50.4, 76.5)	(61.0, 61.5)	(83.1, 96.1)
	WinCLIP[17]	(93.3, 96.8)	(82.6, 83.4)	(90.2, 94.6)	(65.3, 70.6)	(73.8, 90.1)	(77.1, 77.8)	(72.5, 92.2)	(56.8, 81.5)	(**90.3**, **91.8**)	(73.3, 94.5)
	PromptAD[20]	(**95.1**, **97.6**)	(**88.7**, **90.6**)	(**90.7**, **95.6**)	(80.6, **83.2**)	(**79.9**, 92.6)	(**80.8**, **82.5**)	(76.1, 92.2)	(**93.1**, **97.6**)	(87.2, 87.9)	(93.2, **98.7**)
	InCTRL[42]	(91.8, 95.4)	(84.1, 83.6)	(88.4, 73.4)	(**80.7**, **83.2**)	(79.7, **93.0**)	(78.7, 81.5)	(**77.5**, **93.2**)	(92.0, 96.2)	(88.9, 89.9)	(91.7, 98.4)
	Ours	(**97.0**, **98.3**)	(**93.7**, **94.9**)	(**97.5**, **99.0**)	(**85.4**, **88.9**)	(**84.2**, **94.7**)	(**84.0**, **87.5**)	(**80.5**, **94.5**)	(**95.9**, **98.6**)	(**94.2**, **95.0**)	(**96.6**, **99.4**)
8 shot	PatchCore[29]	(92.7, 96.3)	(88.7, 90.1)	(**92.5**, 95.3)	(73.3, 75.1)	(73.3, 90.1)	(74.2, 75.5)	(77.6, 92.0)	(95,2, 98.2)	(65.6, 66.3)	(**94.1**, **98.7**)
	RegAD[16]	(85.0, 93.0)	(77.7, 78.4)	(90.7, 95.9)	(61.1, 67.3)	(65.7, 86.7)	(72.0, 72.9)	(63.8, 86.3)	(56.3, 77.7)	(67.7, 68.0)	(80.9, 95.8)
	WinCLIP[17]	(94.1, 97.0)	(83.8, 85.4)	(90.6, 95.6)	(65.8, 71.0)	(74.2, 90.2)	(78.0, 78.8)	(72.2, 92.1)	(57.1, 81.6)	(90.4, 91.8)	(73.1, 94.4)
	PromptAD[20]	(**95.8**, **98.0**)	(89.0, **91.7**)	(91.3, **96.9**)	(**84.2**, **88.0**)	(**82.8**, **93.8**)	(**80.5**, **83.1**)	(**81.2**, 93.8)	(**95.8**, **98.4**)	(87.2, 88.0)	(93.4, **98.7**)
	InCTRL[42]	(93.9, 96.4)	(**89.2**, 90.0)	(79.3, 86.7)	(82.7, 86.5)	(81.6, 93.6)	(79.2, 83.1)	(81.1, **94.2**)	(95.2, 98.0)	(**91.9**, **92.5**)	(91.5, 98.3)
	Ours	(**97.4**, **98.5**)	(**94.7**, **95.9**)	(**98.0**, **99.1**)	(**87.2**, **90.1**)	(**84.7**, **94.9**)	(**84.8**, **88.4**)	(**82.3**, **95.2**)	(**96.6**, **98.9**)	(**95.7**, **96.2**)	(**96.7**, **99.4**)

Table 2. Comparison of different methods under various few-shot settings. The results are presented in the form of (AUROC%, AP%). The best and second-best results are highlighted in red and blue, respectively.

(WTB [23], Leaves [7]). We follow the cross-modal setup described in Section 3.1. Specifically, when testing datasets of the RGB modality, we use datasets from the 3D modality as the auxiliary datasets; for testing datasets of the 3D modality, RGB datasets are used as the auxiliary datasets; for the MRI/CT and thermal modality, datasets from both RGB and 3D modalities are used as auxiliary datasets. For the few-shot setting, we evaluate the cases of $K = 1, 2, 4, 8$.

Competing Methods and Evaluation Metric. For the zero-shot setting, we compare our method with VAND [11], WinCLIP [17], AnomalyCLIP [40] and AdaCLIP [10]. For the few-shot setting, we compare our method with PatchCore [29], RegAD [16], WinCLIP [17], PromptAD [20] and InCTRL [42]. All the methods follow the same cross-modal settings described above. Two popular metrics AUROC (Area Under the Receiver Operating Characteristic) and AP (Average Precision) are utilized to evaluate the anomaly detection performance.

Implementation details. We employ the image encoder of the publicly available CLIP [26] as our backbone. Local features from intermediate layers 5, 15, and 25 are used for both zero-shot and few-shot settings. Images are resized to 224×224 for both training and testing. For training the Transferable Visual Prototype, the Adam optimizer is utilized with a learning rate of 1×10^{-3}. The batch size is set to 16. Each group of Transferable Visual Prototype is trained on the specific modality dataset using a single NVIDIA GeForce RTX 3090 GPU for 1500 iterations.

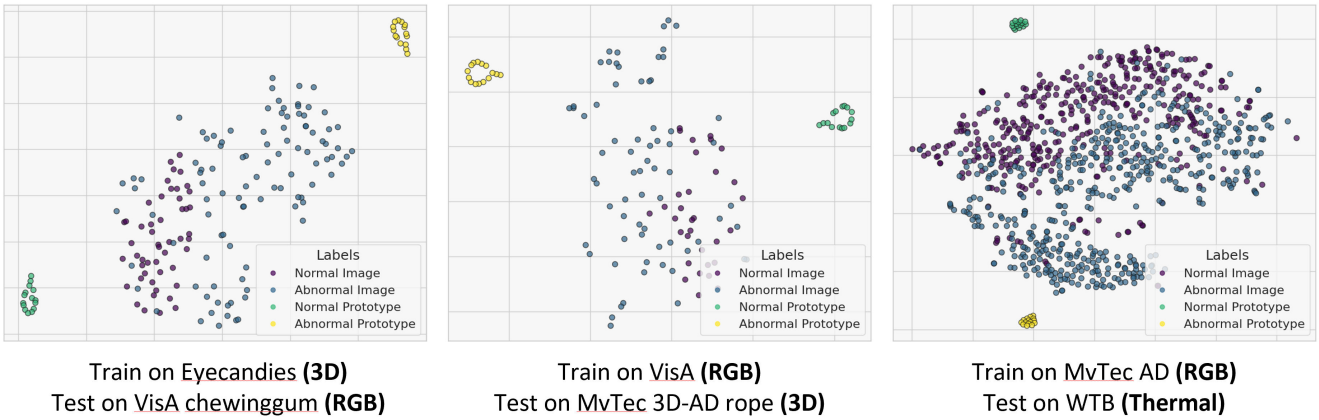

Figure 4. The t-SNE visualization of our Transferable Visual Prototypes trained on datasets from different modalities. "Train on" refers to training the TVP on a dataset under a specific modality. "Test on" indicates testing on a different modality's dataset within a specific category. During the training of TVP, we preserve the global part of TVP from the intermediate stages of training, corresponding to the Normal Prototype and Abnormal Prototype. Normal Image and Abnormal Image represent the global visual embeddings corresponding to normal and abnormal images in the test category, respectively.

4.2. Results for zero-shot setting

Table 1 presents the results of different methods in zero-shot setting. Our method achieve the best AUROC scores across nine datasets and outperform all other methods on seven datasets in terms of AP. For each modality, we calculate the mean AUROC across all datasets within that modality. Our method achieves AUROC improvements of 4.1%, 6.1%, 7.6%, and 6.8% and AP improvement of 4.7%, 4.4%, 3.0%, 4.0% over the best competing methods in the RGB, 3D, MRI/CT, and thermal modalities, respectively. It is worth mentioning that while other methods demonstrated strong detection capabilities on certain modalities, such as WinCLIP's performance on MvTec AD with an AUROC of 91.8%, it only managed an AUROC of 49.1% on OCT2017. Similarly, AdaCLIP achieved an AUROC of 95.3% on BTAD but only 72.5% on Leaves. This indicates that although they can achieve competitive detection performance under certain modalities, their performance may be poor in others. By training on known modalities to obtain the Transferable Visual Prototypes, our method learns directly in the visual space to build generalized representations of normal/abnormal patterns. Moreover, when predicting data of unknown modalities, we use the proposed prototype harmonization strategy, allowing to comprehensively utilize abnormal representations across different modalities, thereby achieving stronger cross-modal detection performance.

Furthermore, we visualize the Transferable Visual Prototypes trained on different datasets using t-SNE [32], as shown in Figure 4. It can be seen that the Transferable Visual Prototypes trained on known modalities demonstrate good discriminability for data with unseen modalities,. This suggests that our Transferable Visual Prototypes possess excellent cross-modal capabilities, enabling them to learn more universal abnormal representations across different modalities, which in turn guide anomaly detection in unknown modalities.

4.3. Results for few-shot setting

In the few-shot setting, we demonstrate the anomaly detection performance of different methods in Table 2. Across all settings of $K = 1, 2, 4, 8$, our method achieves the highest detection performance for both AUROC and AP. Notably, our method significantly surpasses the second-best results among all methods. This advantage stems from our use of Visual Discrepancy Inference, where we compare the visual space feature discrepancies between the test image and the known few-shot images, effectively enhancing the model's detection capabilities.

4.4. Ablation Studies

Study on global and local prototype. To analyze the significance of global and local components in our Transferable Visual Prototype, we conduct the following ablation study in zero-shot setting. During inference, we use only the global part $\boldsymbol{p}_g$ and the local part $\boldsymbol{p}_l$, respectively. Their results, along with those using both global and local components, are displayed in Table 3. The results show that relying solely on the global component for inference delivers highly competitive performance across most datasets. Although exclusive reliance on the local component typically results in poorer performance across the majority of datasets, it significantly outperforms the global-only approach on a few datasets (such as OCT2017). This highlights that global and local components characterize anoma-

Metric	Global	Local	MvTec AD	VisA	MvTec 3D-AD	OCT2017	Leaves
AUROC%	✓	-	**91.6**	**86.0**	**78.6**	74.2	**92.0**
	-	✓	84.6	73.9	74.4	**96.6**	76.1
	✓	✓	**92.3**	**86.8**	**79.6**	**91.2**	**91.1**
AP%	✓	-	**96.0**	**88.8**	**92.7**	90.7	**98.7**
	-	✓	91.5	77.9	90.9	**98.7**	94.9
	✓	✓	**96.3**	**89.3**	**93.1**	**97.1**	**98.6**

Table 3. Ablation study of the global and local parts for our Transferable Visual Prototype in zero-shot setting.

Metric	Manner	MvTec AD	VisA	MvTec 3D-AD	OCT2017	Leaves
AUROC%	Best	**90.9**	**82.5**	73.4	**93.3**	**85.8**
	Max	89.9	75.4	71.7	**93.3**	84.1
	Mean	90.8	81.2	**75.4**	82.2	83.3
	PTH	**92.3**	**86.8**	**79.6**	**91.2**	**91.1**
AP%	Best	**95.5**	**86.1**	90.9	**97.7**	**97.7**
	Max	95.4	79.2	89.6	**97.7**	97.4
	Mean	**95.5**	85.0	**91.7**	94.0	97.1
	PTH	**96.3**	**89.3**	**93.1**	**97.1**	**98.6**

Table 4. Ablation study of different prototypes aggregation methods in zero-shot setting.

Metric	Unit	MvTec AD	VisA	MvTec 3D-AD	OCT2017	Leaves
AUROC%	Modal	90.8	80.5	74.2	80.9	**82.8**
	Dataset	**92.3**	**86.8**	**79.6**	**91.2**	**91.1**
	SubDataset	**92.0**	**85.8**	**78.7**	**95.3**	81.4
AP%	Modal	95.5	84.5	91.2	93.6	**97.1**
	Dataset	**96.3**	**89.3**	**93.1**	**97.1**	**98.6**
	SubDataset	**96.0**	**87.5**	**92.8**	**98.3**	97.0

Table 5. Ablation study of different training set units for our Transferable Visual Prototype in zero-shot setting.

Metric	TVP	VDI	MvTec AD	VisA	MvTec 3D-AD	OCT2017	Leaves
AUROC%	✓	-	92.3	86.8	79.6	**91.2**	**91.1**
	-	✓	**95.7**	**89.7**	**80.7**	89.7	90.3
	✓	✓	**96.5**	**92.5**	**83.8**	**93.7**	**93.8**
AP%	✓	-	96.3	89.3	93.1	**97.1**	**98.6**
	-	✓	**97.1**	**90.5**	**93.3**	96.1	98.0
	✓	✓	**97.8**	**93.6**	**94.6**	**97.9**	**98.9**

Table 6. Ablation study of the Visual Discrepancy Inference in two-shot setting.

lies from different perspectives and exhibit strong complementarity. Therefore, using both global and local components together achieves better results.

Study on prototype aggregation strategy. We explore the different ways of using multiple Transferable Visual Prototypes trained on different known datasets in zero-shot setting, as shown in Table 4. Firstly, for data from unknown modalities, we use each set of Transferable Visual Prototypes separately for inference and report the highest detection performance (***Best***). Secondly, we report results using the Transferable Visual Prototype that has the highest similarity d_m with the test image (***Max***). Besides, we assign equal weight to each Transferable Visual Prototype during inference (***Mean***). ***PTH*** denotes the results based on our prototype harmonization strategy. It is evident that, even when compared with the ***Best*** approach, our method achieves optimal performance in the vast majority of datasets, benefiting from our ability to comprehensively utilize different groups of Transferable Visual Prototypes for inferring on the test images.

Study on different training set units. In our prototype harmonization strategy, complementarity is a critical factor for enhancing performance. We wonder how further splitting or merging the dataset would affect complementarity. We initially train single TVP jointly across all datasets within each modality, using modality as the smallest unit (***Modal***). Subsequently, we train TVP separately for each dataset, considering the dataset as the minimal unit (***Dataset***), which is the setting of our main experiment. Furthermore, we divide each dataset into two equal sub-datasets and conduct training on these sub-datasets respectively (***SubDataset***). The results are shown in Table 5. Experiments with dataset as the unit achieve better results compared to modality and sub-dataset, indicating that not only the number of units is important, but also the amount of data in each unit. There is a tradeoff between the two.

Study on zero/few-shot part. In the few-shot setting, we demonstrate the results of using Transferable Visual Prototypes alone (***TVP***), Visual Discrepancy Inference alone (***VDI***), and a combination of both, as shown in Table 6. It's worth stating that the TVP-only results are consistent with zero-shot setting. When only using Visual Discrepancy Inference, the model can achieve detection performance with an AUROC of 95.7% on MvTec AD and 89.7% on VisA in the two-shot setting. The model can achieve the best results when both Transferable Visual Prototype and Visual Discrepancy Inference are used simultaneously.

5. Conclusion

In this work, we introduce the task of cross-modal anomaly detection and propose a method featuring three core components: the *Transferable Visual Prototype*, which directly learns normal/abnormal semantics in the visual space; a *Prototype Harmonization strategy* that adaptively utilizes these prototypes from various modalities for inference on an unknown modality; and a *Visual Discrepancy Inference* for enhancing performance when a few normal images are available under the few-shot setting. Extensive experiments conducted on ten datasets across four modalities substantiate the effectiveness of our approach.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.U23B2060, No.62088102), and the Key Research and Development Program of China (No.2020AAA0108305).

References

[1] 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. 5

[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 Workshops*, pages 4042–4053, 2024. 1

[3] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–138, 2024. 2

[4] 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. 5

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 2

[6] 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

[7] J. Bhaskar, V. Nidheesh Kumar, L. Vineetha Naidu, R. Divya Reddy, T. Bhargavi, and Sumathi D. Paddy leaf disease detection. In *2024 10th International Conference on Advanced Computing and Communication Systems*, pages 2031–2036, 2024. 6

[8] 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. 5

[9] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 1

[10] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72, 2024. 1, 2, 6

[11] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 1, 2, 6

[12] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 1, 2, 3

[13] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 1, 2, 3

[14] 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. 2

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 4

[16] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 3, 6

[17] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 1, 2, 4, 6

[18] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pages 66–71, 2021. 5

[19] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 4

[20] Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1093–1102, 2024. 1, 2, 3, 4, 6

[21] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 2, 4

[22] Lihao Liu, Angelica I. Aviles-Rivero, and Carola-Bibiane Schönlieb. Contrastive registration for unsupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):147–159, 2025. 1

[23] Majid Memari, Mohammad Shekaramiz, Mohammad A. S. Masoum, and Abdennour C. Seibi. Data fusion and ensemble learning for advanced anomaly detection using multispectral rgb and thermal imaging of small wind turbine blades. *Energies*, 17(3), 2024. 6

[24] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics*, pages 01–06, 2021. 5

[25] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys*, 54(2):1–38, 2021. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 1, 4, 6

[27] Khalid Raza and Nripendra K Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9):1059–1077, 2021. 1

[28] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 4

[29] 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. 1, 2, 3, 6

[30] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022. 1

[31] Vishaal Udandarao. Understanding and fixing the modality gap in vision-language models. *Master's thesis, University of Cambridge*, 2022. 2, 4

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[33] Jie Yang, Ruijie Xu, Zhiquan Qi, and Yong Shi. Visual anomaly detection for images: A systematic survey. *Procedia computer science*, 199:471–478, 2022. 1

[34] 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. 2

[35] 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, 2022.

[36] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. 2

[37] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 2

[38] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16699–16708, 2024. 2

[39] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 1

[40] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 1, 2, 4, 6

[41] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 1, 4

[42] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3, 6

[43] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 5