A Survey on Visual Anomaly Detection: Challenge, Approach, and Prospect

Yunkang Cao[1*], **Xiaohao Xu**[2*], **Jiangning Zhang**[3*], **Yuqi Cheng**[1], **Xiaonan Huang**[2], **Guansong Pang**[4], **Weiming Shen**[1†]

[1]Huazhong University of Science and Technology, [2]University of Michigan, Ann Arbor
[3]Youtu Lab, Tencent, [4]Singapore Management University

{cyk_hust, yuqicheng, shenwm}@hust.edu.cn, {xiaohaox, xiaonanh}@umich.edu,
186368@zju.edu.cn, gspang@smu.edu.sg

Abstract

Visual Anomaly Detection (VAD) endeavors to pinpoint deviations from the concept of normality in visual data, widely applied across diverse domains, *e.g.*, industrial defect inspection, and medical lesion detection. This survey comprehensively examines recent advancements in VAD by identifying three primary challenges: 1) scarcity of training data, 2) diversity of visual modalities, and 3) complexity of hierarchical anomalies. Starting with a brief overview of the VAD background and its generic concept definitions, we progressively categorize, emphasize, and discuss the latest VAD progress from the perspective of sample number, data modality, and anomaly hierarchy. Through an in-depth analysis of the VAD field, we finally summarize future developments for VAD and conclude the key findings and contributions of this survey.

1 Introduction

Visual anomaly detection (VAD) stands as a pivotal task spanning diverse domains [Pang *et al.*, 2022], involving the identification of deviations in visual data from established normality. In recent years, we have seen notable progress in this field across multiple domains. For instance, inspecting defects in industrial settings [Bergmann *et al.*, 2019], identifying lesions in medical image analysis [Antonelli *et al.*, 2022; Liu *et al.*, 2023c], and detecting unknown objects in autonomous driving scenarios [Bogdoll *et al.*, 2022]. The significance of VAD extends beyond these specific applications, as its ability to uncover irregularities in visual data contributes significantly to enhancing the overall reliability and safety of various technological systems. Despite great progress, VAD still encounters three predominant challenges as illustrated in Figure 1:

1) Scarcity of Training Data. Practical VAD systems often struggle to amass abundant abnormal samples for training [Bergmann *et al.*, 2019]. In special application scenarios, normal samples may even be inaccessible due to data privacy concerns [Jeong *et al.*, 2023]. The scarcity of training data poses a significant challenge in VAD, requiring the training of models from partially observed samples in order to subsequently detect anomalies in an open-world environment.

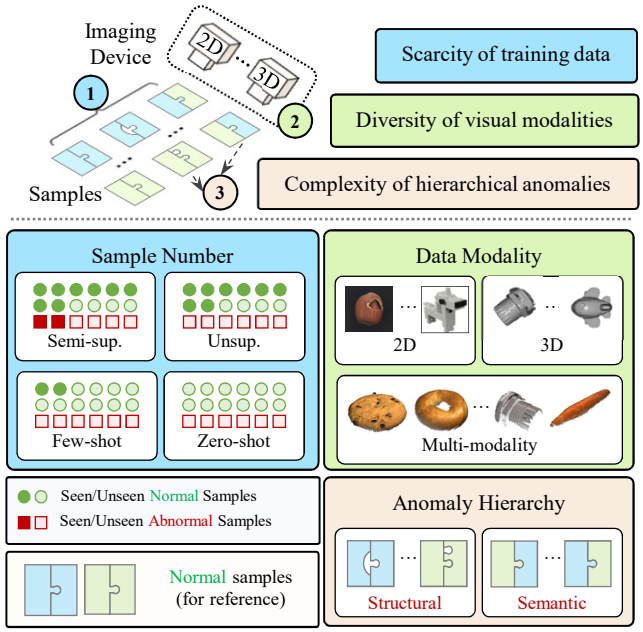

Figure 1: Major VAD challenges (**Top**) and taxonomies (**Bottom**).

2) Diversity of Visual Modalities. VAD systems employ diverse imaging devices to capture visual information, such as color cameras [Bergmann *et al.*, 2019] and radar scanners [Bergmann *et al.*, 2022b]. The utilization of these diverse imaging techniques introduces distinct visual modalities, presenting both utility and complexity in their effective integration.

3) Complexity of Hierarchical Anomalies. Anomalies may manifest in various hierarchies. Some structural anomalies (*e.g.*, visual scratch) can be identified through local regions, while others semantic anomalies (*e.g.*, logical mismatch) demand a higher-level understanding of the normal context [Bergmann *et al.*, 2022a]. It is challenging for VAD models to concurrently possess both fine-grained and global understanding of visual data.

The above challenges continually drive the research frontiers of the VAD field. In this survey, we aim to comprehensively analyze the latest progress made in tackling these chal-

*Equal contribution.
†Corresponding author.

lenges. After providing a concise overview of the background of VAD (Sec. 2), we will review previous achievements and research trends from perspectives corresponding to the aforementioned challenges: 1) sample number (Sec. 3.1), 2) data modality (Sec. 3.2), and 3) anomaly hierarchy (Sec. 3.3). Afterward, we will present some potential and emerging future research directions in Sec. 4. Finally, Sec. 5 will conclude this survey by summarizing key findings and contributions.

2 Background

In this section, we briefly review the background of VAD, encompassing the conceptual definitions and a generic formulation for VAD. Then we outline prominent datasets and metrics used for evaluating VAD methods and finally introduce relevant surveys to clarify the contributions of this survey.

Concept Definition. Here, we formally define concepts of visual data, anomalies, and the VAD task.

1) Visual data is classified into four fundamental categories: *data point*, *entity*, *relation*, and *frame*. i) A data point D represents the smallest discernible element captured by imaging devices, such as a pixel in an image or a point in a point cloud. ii) An entity E is a cohesive set of data points that collectively represent a real-world object. Denoted as $E = \{D_1, D_2, \ldots, D_n\}$, an entity encompasses individual data points that together form a meaningful visual element. iii) A relation function φ takes multiple entities $\{E_1, E_2, \ldots, E_m\}$ as input and combines them to form a visual frame. iv) A frame $F = \varphi(E_1, E_2, \ldots, E_m)$ encapsulates the interrelation between different entities, capturing contextual information within the visual scene.

2) Anomaly concept refers to observations that deviate from the concept of normality [Ruff *et al.*, 2021]. Anomalies in visual data exhibit a hierarchical relationship, where anomalies at lower levels can propagate to higher levels. For instance, errors in individual data points D lead to anomalies in the formation of entities E. This hierarchical structure enables a nuanced understanding of anomalies across different granularities within visual data. Considering the hierarchies of anomalies, we define two types of anomalies: ***i) Structural anomalies*** focus on the integrity of individual data points and their organization within entities, as exemplified in MVTec AD [Bergmann *et al.*, 2019]. These anomalies are particularly insightful for detecting local structure deviations, such as lesions in medical visual data or defects in industrial inspections. ***ii) Semantic anomalies***, on the other hand, encompass deviations at higher hierarchical levels, including the entity, relation, and frame levels. Anomalies at the entity level may involve the misinterpretation of entities within the visual scene, such as unknown objects on the road [Bogdoll *et al.*, 2022]. Relation-level anomalies involve inaccuracies in the contextual connections between entities, as illustrated in MVTec LOCO [Bergmann *et al.*, 2022a]. Frame-level anomalies denote abnormalities in the overall visual acquisition, typically presented as novelty detection [Ruff *et al.*, 2021] or one-class classification [Chen *et al.*, 2022]. Addressing semantic anomalies is vital for understanding the context of visual entities and their interrelations, ensuring accurate and meaningful interpretations of the entire scene.

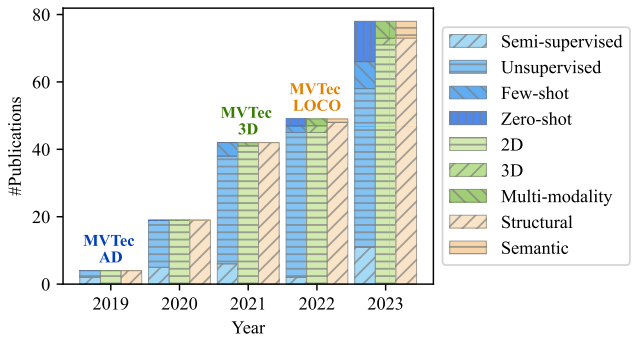

Figure 2: Number of VAD publications regarding taxonomies under the three perspectives in Sec. 3. Blue for Sec. 3.1, green for Sec. 3.2, and orange for Sec. 3.3.

Table 1: Comparison with representative VAD surveys. ✔: Included. +: Partially Included. ✗: Not Included.

Survey	Sample Number	Data Modality	Anomaly Hierarchy
[Ruff *et al.*, 2021]	+	✗	✗
[Pang *et al.*, 2022]	+	✗	✗
[Tao *et al.*, 2022]	+	✗	✗
[Diers and Pigorsch, 2023]	+	✗	✗
[Liu *et al.*, 2023b]	+	+	✗
Ours	✔	✔	✔

3) Visual anomaly detection aims to develop a model for detecting visual anomalies. The task involves a training dataset $\mathcal{F}_{\text{train}}$ comprising both normal frames F_n and abnormal frames F_a, each accompanied by corresponding ground truths expressed as hard labels. The principal goal is to establish a discriminative function $f_\theta : F \to [0, 1]$, parameterized by θ, using $\mathcal{F}_{\text{train}}$ to precisely assign anomaly scores to unlabeled frames in $\mathcal{F}_{\text{test}}$. Recent advancements highlight the necessity for more detailed anomaly scores, extending to data point-level, entity-level, and relation-level assessments.

Scope of This Survey. Notably, VAD has witnessed substantial advancements in the industrial domain compared to other domains. Despite methodological variations across domains, the fundamental principles governing VAD exhibit considerable consistency. Consequently, this survey strategically concentrates on VAD within industrial scenarios as a representative, endeavoring to furnish a meticulous and exhaustive review of the entire VAD landscape. Notably, the semantic anomalies within industrial scenarios predominantly manifest at the relation level, posing challenges and garnering escalating attention.

Datasets & Metrics. Recent advancements in VAD are significantly influenced by several datasets, such as MVTec AD [Bergmann *et al.*, 2019], MVTec 3D [Bergmann *et al.*, 2022b], MVTec LOCO [Bergmann *et al.*, 2022a], and VisA [Zou *et al.*, 2022]. The alignment between predicted probability distributions and true probability distributions provided by these datasets can evaluate the performance of VAD methods. Various metrics [Bergmann *et al.*, 2019; Jeong *et al.*, 2023] are employed for this purpose, encompassing Area Under the Receiver Operating Characteristic curve

(AUROC) and Area Under the Per-Region-Overlap curve (AUPRO), among others.

Comparison to Other Surveys. Recent advancements in VAD methods, particularly tailored for 2D data with structural anomalies, have been substantial, as depicted in Figure 2 [Bergmann *et al.*, 2019]. The prevalence of unsupervised VAD approaches is also discernible. Notably influenced by pivotal milestones such as MVTec AD [Bergmann *et al.*, 2019], MVTec 3D [Bergmann *et al.*, 2022b], and MVTec LOCO [Bergmann *et al.*, 2022a], sub-settings addressing the aforementioned three challenges have demonstrated promising advancements. Examinations of these emerging trends are conspicuously absent in many current VAD surveys, as underscored in Table 1. In contrast, our survey undertakes a more comprehensive exploration, encapsulating the latest developments in VAD from diverse perspectives.

3 Taxonomy

In this section, we review existing methods from three perspectives that correspond to the aforementioned three challenges in Table 2.

3.1 From the Perspective of Sample Number

Given the challenge of data scarcity in practical scenarios, various VAD tasks consider varying numbers of normal and abnormal training samples. The prevalent sub-settings are shown in the top part of Table 2.

Semi-supervised VAD. Semi-supervised VAD methods aim to utilize both normal samples and infrequent observed abnormal samples during training. However, relying solely on a restricted set of seen anomalies may induce overfitting, resulting in poor generalization capacity to novel anomalies [Yao *et al.*, 2023b]. To mitigate the overfitting, DRA [Ding *et al.*, 2022] introduced a disentanglement strategy for anomalies in open-world scenarios. This strategy classifies anomalies into three distinct categories: seen anomalies, pseudo anomalies, and latent residual anomalies. Specific detection heads are trained for individual data types, enabling them to specialize in detecting specific anomalies accordingly. Similarly, PRN [Zhang *et al.*, 2023a] harnesses both seen anomalies and pseudo anomalies to explicitly capture residual features distinguishing anomalies from normal patterns. PRN employs various anomaly generation strategies, considering both seen and unseen appearance variances to create diverse pseudo anomalies. Through learning from these anomalies, PRN constructs multi-scale prototypes and generates more faithful representations for open-world anomalies rather than fixating on seen anomalies. BiaS [Cao *et al.*, 2023d] also disentangles open-world anomalies into seen and unseen anomalies and utilizes two specialists for these anomalies, respectively. Then BiaS proposes a dynamic fusion strategy to fuse the prediction results of these two specialists intelligently. In contrast, BGAD [Yao *et al.*, 2023b] focuses on modeling the normal feature distribution through a flow model, concurrently incorporating seen anomalies to optimize the description boundary of normal features. In summary, the aforementioned methods collectively address overfitting to seen anomalies by introducing diverse pseudo anomalies or by focusing on optimizing the description boundary for normal samples through the incorporation of seen anomalies.

Unsupervised VAD. Unsupervised VAD concentrates on discerning anomalies trained exclusively on normal samples for specific categories [Cao *et al.*, 2023a]. The primary goal is to model the distribution of normal features, typically involving two sub-steps: feature extraction and distribution modeling. Recent advancements predominantly employ pretrained neural networks such as ResNet for feature extraction. Four main schemes for distribution modeling include memory bank, reconstruction, knowledge distillation, and flow-based methods. Memory bank-based methods, exemplified by PatchCore [Roth *et al.*, 2022], directly store features of training normal samples. They then utilize the nearest distance between testing samples and the stored bank to score anomalies. By selecting the most representative features in the training set, the memory bank can be small and representative, ensuring efficient and effective VAD. Reconstruction-based techniques, such as DFR [Shi *et al.*, 2020], and knowledge distillation approaches, exemplified RD4AD [Deng and Li, 2022] and ViTAD [Zhang *et al.*, 2023b], entail the regression of extracted normal features using a secondary trainable network. In the context of reconstruction-based methods, this trainable network is referred to as autoencoders, while in knowledge distillation-based methods, it is denoted as student networks. Since the trainable network is exclusively trained with normal samples, it is expected to produce substantial regression errors for abnormal samples. In contrast, flow-based models [Gudovskiy *et al.*, 2021] utilize a normalizing flow framework to automatically depict the distribution of normal features and explicitly estimate the likelihood of tested features. However, the aforementioned VAD methods often confront an issue where anomaly scores for abnormalities can unexpectedly be low due to imprecise boundary descriptions, attributed to the generalization ability of employed neural networks, termed as over-generalization in CDO [Cao *et al.*, 2023a]. To mitigate over-generalization, some methods like Draem [Zavrtanik *et al.*, 2021] MRKD [Jiang *et al.*, 2023], and DAF [Cai *et al.*, 2023] introduce synthetic anomalies. Consequently, models are not only tasked with regressing normal feature distributions but also with generating substantial regression errors for synthetic anomalies. MemKD [Gu *et al.*, 2023] addresses the over-generalization problem by explicitly storing a memory bank, ensuring that outputs exclusively represent normal features. Similarly, TFA-Net [Luo *et al.*, 2024] proposes to restore normal features explicitly guided by normal templates. These approaches result in significant regression errors when faced with abnormal inputs.

Few-shot VAD. Few-shot VAD concentrates on training the model with a limited amount of normal data [Huang *et al.*, 2022]. However, these few normal samples may not adequately represent the entire normal sample set. Consequently, the model must establish a description boundary by learning from these seen normal samples, intending to simultaneously describe distributions of unseen normal samples while excluding distributions of abnormal samples. This presents a notable challenge. To address this challenge, few-

Table 2: Summary of existing AD methods from three perspectives, addressing the aforementioned challenges. For each sub-setting, we list several representatives. ✔: Involved. +: Partially Involved. ✗: Not Involved. ◯: Disregarded.

Perspective	Taxonomy	Training Set		Modality		Hierarchy		Representative Works	Summary
		F_n	F_a	RGB	3D	Structural	Semantic		
Sec. 3.1 Sample Number	Semi-supervised	✔	+	◯	◯	◯	◯	DRA, PRN, BiaS, BGAD	Preventing Overfitting
	Unsupervised	✔	✗	◯	◯	◯	◯	ST, Draem, PatchCore, RD4AD	Modelling feature distribution
	Few-shot	+	✗	◯	◯	◯	◯	RegAD, GraphCore, FastRecon	Improving feature descriptiveness
	Zero-shot	✗	✗	◯	◯	◯	◯	WinCLIP, APRIL-GAN, SAA	Utilizing external knowledge
Sec. 3.2 Data Modality	2D-aware RGB Image	◯	◯	✔	✗	◯	◯	MVTec AD, Eyecandies, PAD	Optimizing imaging factors
	3D-aware Representation	◯	◯	✗	✔	◯	◯	MVTec 3D, Real3D, CPMF	Learning 3D features
	Multi-modality	◯	◯	◯	◯	◯	◯	AST, BTF, M3DM, EasyNet	Fusing multimodal features.
Sec. 3.3 Anomaly Hierarchy	Structural Anomaly	◯	◯	◯	◯	✔	✗	DSKD, GLCF, EfficientAD	Modelling local structural patterns
	Semantic Anomaly	◯	◯	◯	◯	✗	✔	MVTec LOCO, ComAD, PSAD	Modelling relationships between entities

Table 3: Qualitative comparison of representative VAD methods in Sec. 3.1 in terms of frame-level AUROC.

Dataset	Semi-supervised (10-shot)			Unsupervised					Few-shot (8-shot)			Zero-shot		
	DRA	PRN	BGAD	Draem	PatchCore	RD4AD	TFA-Net	MemKD	RegAD	GraphCore	FastRecon	WinCLIP	VAND	AnomalyCLIP
MVTec AD	96.1	99.4	99.3	98.0	99.2	98.4	98.7	99.6	91.2	95.9	95.2	91.8	86.1	91.5
VisA	-	-	-	88.7	95.1	96.0	88.7	97.6	-	-	-	78.1	78.0	82.1

shot VAD methods primarily emphasize enhancing feature descriptiveness, aiming to make the available few-shot samples a more representative subset. RegAD [Huang *et al.*, 2022] employs registration-based proxy tasks for representation learning, aligning samples of the same category through geometrical transformations, thereby enhancing feature descriptiveness. Another method, GraphCore [Xie *et al.*, 2023], utilizes a vision isometric invariant graph neural network to extract rotation-invariant structural features, particularly beneficial for categories with geometrical transformations. Additionally, FastRecon [Fang *et al.*, 2023] proposes utilizing a few normal samples as a reference to reconstruct their normal versions, achieving anomaly detection through sample alignment. In comparison to RegAD, FastRecon also aims to align few-shot normal samples and testing samples, while accommodating more complex geometrical transformations. In summary, prevailing few-shot VAD methods commonly rely on sample alignment to augment feature descriptiveness, aiming to enhance the representativeness of seen normal samples for the entire normal sample set.

Zero-shot VAD. Zero-shot VAD aims to develop a unified model for detecting anomalies across diverse domains without relying on reference normal samples [Jeong *et al.*, 2023]. While holding significant potential for versatility, this task presents a challenge due to the absence of specific prior information related to the target domains. Existing zero-shot VAD methods tackle this challenge by integrating external knowledge to enhance anomaly detection capabilities. WinCLIP [Jeong *et al.*, 2023], a pioneering zero-shot VAD method, leverages a pretrained visual-language model (VLM) CLIP [Radford *et al.*, 2021]. Through the integration of CLIP, WinCLIP achieves zero-shot VAD by computing the similarity between image patches and normal/abnormal textual captions. Since CLIP has acquired implicit knowledge for distinguishing normality from anomalies through training on extensive datasets with visual-text pairs, the calculated similarities can serve as effective anomaly scores. APRIL-GAN [Chen *et al.*, 2023b], succeeding WinCLIP, tackles the domain gap between CLIP and targeted VAD data. APRIL-GAN proposes adapting CLIP to VAD by training it with annotated auxiliary VAD data, thereby enhancing its suitability for VAD applications. Building on this adaptation scheme, AnomalyCLIP [Zhou *et al.*, 2024] introduces the concept of learning object-agnostic text prompts to overcome the limitations associated with manually designed prompts. Additionally, SAA [Cao *et al.*, 2023c] introduces an ensemble scheme that combines various off-the-shelf VLMs for VAD, providing a means to integrate human expertise into VAD systems. In summary, these zero-shot VAD methods leverage external knowledge, often from off-the-shelf VLMs like CLIP, to facilitate the detection of anomalies in arbitrary categories.

Discussion. This section evaluates the discussed methods from the perspective of sample numbers. As we transition from semi-supervised to zero-shot VAD, the available training samples for specific categories gradually diminish, resulting in a decline in VAD performance, as illustrated in Table 3. While VAD performance reaches saturation with ample samples, the effectiveness of few-shot and zero-shot VAD remains suboptimal. Nevertheless, the shared objective across these sub-settings is to learn from available data and develop a VAD model capable of generalizing to unseen normal samples while detecting abnormal samples. Given the variability in accessible samples within individual sub-settings, their focal points may differ, leading to distinct key motivations. Future efforts could be directed toward establishing a unified VAD framework for all these sub-settings, thus efficiently utilizing available data and consolidating efforts to identify optimal methods for practical applications.

3.2 From the Perspective of Data Modality

In this section, we categorize VAD from the perspective of data modality, as shown in the middle part of Table 2.

2D-aware RGB Image. RGB images play a crucial role in VAD. Datasets like MVTec AD [Bergmann *et al.*, 2019] and VisA [Zou *et al.*, 2022] have curated extensive datasets containing both normal and abnormal images, significantly driving advancements for VAD on RGB images, as Sec. 3.1 elaborates. Despite the commendable performance on this bench-

mark, these datasets typically assume an ideal imaging environment with perfectly aligned objects and optimal illumination. These critical imaging factors are gradually taken into consideration. Illumination, a critical aspect of imaging systems, significantly influences imaging quality [Bonfiglioli *et al.*, 2022]. Adequate illumination conditions can enhance the visibility of anomalies. Eyecandies [Bonfiglioli *et al.*, 2022] accounts for different illumination conditions. Specifically, four lights in various positions are employed to capture multi-illumination images in Eyecandies. A simple autoencoder is then applied to these images, along with depth and normal images [Bonfiglioli *et al.*, 2022]. Considering the pose of objects, most categories in MVTec AD assume perfectly aligned objects, which may not reflect real-world scenarios where objects can be in any pose. PAD [Zhou *et al.*, 2023] introduces a multi-pose VAD dataset and formulates a pose-agnostic VAD task. In summary, there have already been plenty of methods proposed for RGB images, especially for datasets like MVTec AD that are equipped with ideal imaging environments. Recent research has focused on exploring VAD for RGB images in practical imaging environments, considering factors such as unideal illumination and diverse poses.

3D-aware Representation. Geometrical information, often represented as 3D data like point clouds, serves as a direct manifestation of the size and shape of visual entities. Two notable datasets designed for point cloud VAD are MVTec 3D [Bergmann *et al.*, 2022b] and Real3D [Liu *et al.*, 2023a]. These datasets encompass high-resolution point clouds, facilitating the identification of subtle geometrical deviations. Similar to image VAD, point cloud VAD methods can be broadly categorized into two sub-steps [Cao *et al.*, 2023b]: feature extraction and distribution modeling. In contrast to the image domain, where numerous effective off-the-shelf pretrained models can serve as feature extractors, pretrained point cloud neural networks lack robustness [Bergmann and Sattlegger, 2023]. Therefore, a self-supervised learning scheme introduced in [Bergmann and Sattlegger, 2023] aims to construct a more resilient feature extractor for point cloud VAD. Subsequently, a knowledge distillation-based approach is employed for distribution modeling. On the other hand, CPMF [Cao *et al.*, 2023b] transforms point clouds into multi-view depth images, enabling the use of established pretrained image models for point cloud feature extraction. CPMF then incorporates PatchCore [Roth *et al.*, 2022] for distribution modeling. In summary, unlike the abundance of off-the-shelf models for RGB images, there are limited robust models for other modalities. Existing VAD methods that consider the 3D modality usually place special emphasis on learning descriptive features.

Multi-modality. In specific scenarios, employing multi-modality data enhances the comprehensiveness of VAD, such as the coexistence of 3D and RGB modalities. Some methods are explicitly designed to improve the fusion of representations from these modalities. For example, BTF [Horwitz and Hoshen, 2023] straightforwardly concatenates 3D and RGB representations, utilizing them as inputs for PatchCore [Roth *et al.*, 2022]. Building upon BTF, M3DM [Wang *et al.*, 2023] incorporates contrast learning on the two modalities, fostering enhanced synergy between them. Similarly, Shape-Guided [Chu *et al.*, 2023] integrates the two representations guided by shape features. However, these methods heavily rely on pretrained networks, potentially lacking robustness, especially in the context of point cloud networks [Horwitz and Hoshen, 2023]. In contrast, AST [Rudolph *et al.*, 2022] chooses to learn point cloud representation directly from raw data by training an asymmetric teacher-student pair. This pair can process both RGB and 3D data, leading to better integration between the two modalities. Additionally, EasyNet [Chen *et al.*, 2023a] and 3DSR [Zavrtanik *et al.*, 2024] generate synthetic abnormal RGB and point cloud data to train a robust feature extractor for both modalities. In essence, these methods for multi-modality typically concentrate on enhancing the learning and fusion of representations across multiple modalities.

Table 4: Qualitative comparison of representative VAD methods in Sec. 3.2 in terms of frame-level AUROC.

Dataset	2D-aware RGB Image			3D-aware Representation			Multi-modality		
	AST	CDO	M3DM	BTF	CPMF	M3DM	AST	BTF	M3DM
MVTec 3D	88.0	93.8	85.0	78.2	95.2	87.4	93.7	86.5	94.5
Real3D	-	-	-	59.3	62.5	59.4	-	-	-

Discussion. Data modalities for VAD may exhibit variation across targeted scenarios. Among these modalities, RGB and 3D stand out as dominant and extensively explored. In the progression of VAD comprehension, various factors, such as illumination [Bonfiglioli *et al.*, 2022] and pose [Zhou *et al.*, 2023], are considered in conjunction with RGB and 3D modalities. Because of off-the-shelf pretrained models for RGB data, we have witnessed significant progress in 2D VAD. However, due to the scarcity of pretrained models for other modalities, VAD on other modalities remains less promising, as depicted in Table 4. Table 4 further illustrates that employing multiple modalities enables a more comprehensive acquisition of real-world information, consequently leading to improved VAD performance. Future endeavors might be directed toward developing a unified VAD architecture capable of processing all these modalities, thereby enhancing multi-modality feature learning and fusion.

3.3 From the Perspective of Anomaly Hierarchy

Based on the hierarchy of anomalies, the existing VAD methods can be categorized into: structural and semantic anomaly detection, as summarized in the bottom part of Table 2.

Structural Anomaly. Structural anomalies refer to local structural deviations like scratches, distorted shapes, etc. In recent years, there has been a notable increase in the development of VAD methods specifically tailored for addressing structural anomalies. These methods aim to learn fine-grained features that comprehensively describe local structural patterns within visual entities. Almost all the methods mentioned [Cao *et al.*, 2023b; Roth *et al.*, 2022] above focus on structural anomaly detection. A recent noteworthy contribution in this domain is EfficientAD [Batzner *et al.*, 2024], which proposes a lightweight encoder generated through knowledge distillation. This lightweight encoder restricts the receptive field to a relatively small region, enhanc-

Table 5: Qualitative comparison of representative VAD methods in Sec. 3.3 in terms of frame-level AUROC.

Dataset	Structural Anomaly				Semantic Anomaly			
	GLCF	EfficientAD	ComAD	PSAD	GLCF	EfficientAD	ComAD	PSAD
MVTec LOCO	83.8	94.7	90.9	91.6	82.4	86.8	89.4	98.1

ing the modeling of local structures and facilitating both efficiency and effective anomaly detection.

Semantic Anomaly. In contrast to structural anomalies occurring within individual visual entities, semantic anomalies manifest in the relations between multiple entities within a frame. To advance research in this area, datasets like MVTec LOCO [Bergmann *et al.*, 2022a] have been introduced. These datasets typically involve multiple co-occurring visual entities, where their relations may exhibit abnormalities. Various approaches have been proposed to address semantic anomalies. One line of research posits that global information can implicitly capture the relations between entities. For instance, GCAD [Bergmann *et al.*, 2022a] operates on both local-local consistencies and global-global consistencies through two student-teacher pairs. Specifically, one student-teacher pair with a small receptive field compares local structural features in a regression-based manner to identify structural anomalies. Another student-teacher pair, capable of extracting global context for the given frame, similarly compares global semantic features to identify semantic anomalies. Subsequent methods, such as DSKD [Zhang *et al.*, 2024], GLCF [Yao *et al.*, 2023a], and EfficientAD [Batzner *et al.*, 2024], build upon GCAD and also model normal relations between entities through local-local and global-global consistencies. These methods employ various strategies to enhance the understanding of the normal global context, including contextual affinity distillation [Zhang *et al.*, 2024] and local-global alignment [Yao *et al.*, 2023a]. The bottleneck-like architectures in these methods play a crucial role in learning global context. Alternatively, some methods explicitly focus on modeling the relations between entities. ComAD [Liu *et al.*, 2023d] and PSAD [Kim *et al.*, 2024] initially detect individual entities within frames. In the absence of accessible entity-level labels, they typically use cluster-based methods to group similar data points into clusters, thereby detecting entities. Subsequently, they model the relations between entities through strategies like histogram analysis. This approach allows for a more precise description of the relations between entities. In summary, existing methods for semantic anomaly build an understanding of relations between entities either implicitly by learning global context or explicitly by extracting relations between entities.

Discussion. Anomalies can manifest at various hierarchical levels. While early efforts primarily focused on modeling local structural contexts for structural VAD, there has been a gradual shift in the popularity of semantic VAD. Semantic VAD considers abnormal relations among visual entities. Existing methods in semantic VAD predominantly emphasize modeling relations between entities. Some approaches aim to implicitly capture these relations through global context, while others explicitly analyze individual entities and their relations. However, current methods may encounter challenges in understanding normal relations as effectively as humans, particularly for intricate relations involving entity numbers, locations, etc. The implicit learning strategy may fall short in precisely identifying abnormal relations through global context. In contrast, the explicit strategy may better delineate the relations and yield more precise detection results. As depicted in Table 5, explicit methods such as ComAD and PSAD outperform implicit methods like GLCF and EfficientAD in addressing semantic anomalies. Nevertheless, the methods for entity extraction and relation modeling still leave ample room for improvement.

3.4 Other Perspectives

In addition to the previously discussed major perspectives, there are other settings that warrant exploration. For example, noisy VAD posits that practical application labels for training samples may contain errors, potentially impacting the efficacy of conventional VAD methods. SoftPatch [Jiang *et al.*, 2022] suggests denoising data at the patch level by generating soft outlier scores for patches, thereby excluding patches with high outlier scores, *i.e.*, noisy data, for training. Continual VAD [Liu *et al.*, 2024], on the other hand, endeavors to enhance VAD models with gradually accessible novel data. Directly updating VAD models with this data may result in catastrophic forgetting and impose a substantial computational burden. Hence, UCAD [Liu *et al.*, 2024] proposes empowering VAD models with continual learning capacity by establishing a key-prompt-knowledge memory space. Uniformed VAD [He *et al.*, 2024] has also gained recent popularity, aiming to construct a unified VAD model for diverse categories. Unlike zero-shot VAD, which operates without data from targeted categories, uniformed VAD focuses on effectively utilizing samples from specific categories. DiAD [He *et al.*, 2024], as a representative uniformed VAD method, suggests using diffusion models to restore normal references for testing samples. The disparities between testing samples and the restored normal references in the feature space are then utilized to score anomalies. Considering the scarcity of VAD data, some methods [Dai *et al.*, 2024] also endeavor to generate VAD data. In summary, there are numerous variants to consider for practical VAD systems.

4 Future Directions

This survey provides a comprehensive illustration of the evolution of VAD methods from diverse perspectives, highlighting the challenges faced by existing methods. The subsequent discussion will delve into forthcoming trends across the aforementioned varied perspectives.

4.1 Towards Generic VAD

Current literature has focused on developing VAD methods under varying sample numbers, attributed to the diverse accessible samples in different scenarios. Future efforts could be directed toward constructing a generic VAD framework capable of accommodating different sample numbers.

Foundation Model for VAD. Recently, foundation models like GPT4-V(ision) [OpenAI, 2023] and SAM [Kirillov *et al.*, 2023] have demonstrated outstanding generalization abilities, showcasing scalable performance with different sample numbers. These foundation models also present some

efficacy for VAD [Chen *et al.*, 2023b; Cao *et al.*, 2023c; Zhang *et al.*, 2023c]. Advanced techniques, such as prompting learning [khattak *et al.*, 2023] could further enhance the performance of foundation models for VAD. Additionally, training a foundation model specifically tailored for VAD may yield more promising VAD performance. Various pretraining schemes, like contrastive learning [Radford *et al.*, 2021], and sequential modeling [Bai *et al.*, 2023] can be explored. The in-context learning capacity of these foundational VAD models should be taken into consideration, enabling adaptations with limited data without additional training.

Scalable Data for VAD. The availability of large-scale data is crucial for constructing foundation models for VAD. Practical improvements in visual data collection are warranted. On the other hand, anomaly generation can further contribute to scalable data. While some works like DFMGAN [Duan *et al.*, 2023], and AnomalyDiffusion [Hu *et al.*, 2023] have conducted preliminary investigations into anomaly generation, their current generalization ability is insufficient. In the field of image generation, methods like ControlNet [Zhang *et al.*, 2023d] have demonstrated robust generalization abilities and fine-grained control over the generation process. Future endeavors should focus on developing data generation methods capable of faithfully producing anomalies across a wide range of categories, contributing to the establishment of scalable data for VAD.

4.2 Towards Multimodal VAD

Multimodal data can comprehensively reflect the information of visual entities, leading to enhanced VAD performance. Looking ahead, more attention can be directed towards joint imaging parameter optimization and multimodal learning.

Imaging Parameter Optimization for VAD. Conventional public VAD datasets [Bergmann *et al.*, 2019; Bergmann *et al.*, 2022b] commonly assume ideal imaging conditions. However, challenges emerge in complex practical scenarios where given imaging parameters are unideal. While Eyecandies [Bonfiglioli *et al.*, 2022] and PAD [Zhou *et al.*, 2023] have explored the impact of imaging parameters, a comprehensive analysis on optimizing the imaging process is lacking. Imaging parameter optimization strives to automatically optimize the imaging parameters, encompassing aspects like auto-exposure, auto-focus, etc. This technique opens new avenues for acquiring higher-quality data, facilitating improved visualization, and enabling easier detection of anomalies.

Multimodal Learning for VAD. The importance of effective representations in VAD is crucial [Zhang *et al.*, 2023b], particularly in multimodal data [Cao *et al.*, 2023b]. Achieving effective fusion between these modalities is essential for reliable VAD. In contrast to multimodal fusion in other fields [Xu *et al.*, 2023], existing integration methods in VAD appear relatively simplistic, like feature concatenation [Wang *et al.*, 2023; Horwitz and Hoshen, 2023]. Considering multiple modalities for VAD in real-world applications, it is promising to construct a unified architecture capable of multimodal data. In this context, VAD data from different scenarios and modalities can be utilized together to build a unified VAD model, enhancing feature learning and fusion.

4.3 Towards Holistic VAD

While structural VAD has shown promising performance [Cao *et al.*, 2023a], the capability to detect semantic anomalies is crucial for practical VAD systems. From a broader perspective, VAD systems must not only identify anomalies but also establish connections to downstream processes, leading to better overall performance.

Understanding of Relations between Entities. Semantic anomalies, distinct from structural anomalies requiring local structural representations, demand VAD models to genuinely comprehend the relations between entities. While current VAD methods for semantic anomalies [Yao *et al.*, 2023a; Bergmann *et al.*, 2022a] demonstrate reasonable semantic VAD performance, they still fall short of truly understanding normal relations between entities. Foundation models, such as GPT-4V [OpenAI, 2023], exhibit logical reasoning capacity, showcasing a genuine understanding of normal relations between entities [Cao *et al.*, 2023e]. Therefore, incorporating such foundation models for semantic anomaly detection appears promising. Multi-modal inputs and multi-round conversations [Yang *et al.*, 2023] may further enhance the understanding of relations between entities. Additionally, it is straightforward and promising to detect visual entities first and then identify their relations, like ComAD [Liu *et al.*, 2023d]. API-based modeling like ViperGPT [Surís *et al.*, 2023] may enhance this scheme.

Connecting VAD with Downstream Tasks. VAD plays a pivotal role in interconnected systems, particularly in quality inspection pipelines. However, current research often focuses solely on enhancing the isolated perception step, neglecting downstream integration and impact. To optimize effectiveness, VAD must be seamlessly integrated into the broader system workflow. A comprehensive understanding is required of how VAD interacts with and relies on other components, including potential feedback loops. Recent work has initiated exploration in this direction by incorporating VAD outcomes into objectives such as robotic navigation [Wellhausen *et al.*, 2020] and manufacturing processes [Singh *et al.*, 2023]. Moving forward, greater emphasis should be placed on end-to-end system optimization and unified representation learning achieved through the tight integration of VAD and downstream tasks.

5 Conclusion

In this survey, we have delved into recent advancements in Visual Anomaly Detection (VAD). Initially, we underscored three challenges: 1) scarcity of training data, 2) diversity of visual modalities, and 3) complexity of hierarchical anomalies. Subsequently, we furnished background information on VAD and delved into key concepts. Following that, we conducted a comprehensive review of existing VAD methods, with a focus on sample numbers, data modalities, and anomaly hierarchies. Finally, we pinpointed promising directions for future research: generic, multimodal, and holistic VAD. We anticipate that tackling these challenges and pursuing these research directions will propel VAD towards more robust deployment in real-world applications.

References

[Antonelli *et al.*, 2022] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nat. Commun.*, 2022.

[Bai *et al.*, 2023] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint*, 2023.

[Batzner *et al.*, 2024] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *WACV*, 2024.

[Bergmann and Sattlegger, 2023] Paul Bergmann and David Sattlegger. Anomaly detection in 3d point clouds using deep geometric descriptors. In *WACV*, 2023.

[Bergmann *et al.*, 2019] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

[Bergmann *et al.*, 2022a] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*, 2022.

[Bergmann *et al.*, 2022b] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3d-AD dataset for unsupervised 3d anomaly detection and localization. In *VISAPP*, 2022.

[Bogdoll *et al.*, 2022] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. Anomaly detection in autonomous driving: A survey. In *CVPR*, 2022.

[Bonfiglioli *et al.*, 2022] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *ACCV*, 2022.

[Cai *et al.*, 2023] Yuxuan Cai, Dingkang Liang, Dongliang Luo, Xinwei He, Xin Yang, and Xiang Bai. A discrepancy aware framework for robust anomaly detection. *TII*, 2023.

[Cao *et al.*, 2023a] Yunkang Cao, Xiaohao Xu, Zhaoge Liu, and Weiming Shen. Collaborative discrepancy optimization for reliable image anomaly localization. *TII*, 2023.

[Cao *et al.*, 2023b] Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. *arXiv preprint arXiv:2303.13194*, 2023.

[Cao *et al.*, 2023c] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[Cao *et al.*, 2023d] Yunkang Cao, Xiaohao Xu, Chen Sun, Liang Gao, and Weiming Shen. Bias: Incorporating biased knowledge to boost unsupervised image anomaly localization. *TSMC*, 2023.

[Cao *et al.*, 2023e] Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

[Chen *et al.*, 2022] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *AAAI*, 2022.

[Chen *et al.*, 2023a] Rui Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection. *ACM MM*, 2023.

[Chen *et al.*, 2023b] 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. *arXiv preprint arXiv:2305.17382*, 2023.

[Chu *et al.*, 2023] 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 *ICML*, 2023.

[Dai *et al.*, 2024] Songmin Dai, Yifan Wu, Xiaoqiang Li, and Xiangyang Xue. Generating and reweighting dense contrastive patterns for unsupervised anomaly detection. In *AAAI*, 2024.

[Deng and Li, 2022] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022.

[Diers and Pigorsch, 2023] Jan Diers and Christian Pigorsch. A survey of methods for automated quality control based on images. *IJCV*, 2023.

[Ding *et al.*, 2022] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, 2022.

[Duan *et al.*, 2023] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *AAAI*, 2023.

[Fang *et al.*, 2023] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *ICCV*, 2023.

[Gu *et al.*, 2023] Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *ICCV*, 2023.

[Gudovskiy *et al.*, 2021] Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *WACV*, 2021.

[He *et al.*, 2024] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. Diad: A diffusion-based framework for multi-class anomaly detection. In *AAAI*, 2024.

[Horwitz and Hoshen, 2023] Eliahu Horwitz and Yedid Hoshen. Back to the feature: Classical 3d features are (almost) all you need for 3d anomaly detection. In *CVPR*, 2023.

[Hu *et al.*, 2023] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. *arXiv preprint arXiv:2312.05767*, 2023.

[Huang *et al.*, 2022] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratlin, and Yanfeng Wang. Registration based few-shot anomaly detection. In *ECCV*, 2022.

[Jeong *et al.*, 2023] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

[Jiang *et al.*, 2022] Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, WU Kai, Y. Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. In *NeurIPS*, 2022.

[Jiang *et al.*, 2023] Yuxin Jiang, Yunkang Cao, and Weiming Shen. A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. *KBS*, 2023.

[khattak *et al.*, 2023] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023.

[Kim *et al.*, 2024] Soopil Kim, Sion An, Philip Chikontwe, Myeongkyun Kang, Ehsan Adeli, Kilian M. Pohl, and Sanghyun Park. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *AAAI*, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, and Nikhila Ravi et. al. Segment anything. In *ICCV*, 2023.

[Liu *et al.*, 2023a] Jiaqi Liu, Guoyang Xie, Rui Chen, et al. Real3d-ad: A dataset of point cloud anomaly detection. *NeurIPS*, 2023.

[Liu *et al.*, 2023b] Jiaqi Liu, Guoyang Xie, Jingbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *MIR*, 2023.

[Liu *et al.*, 2023c] Mingxuan Liu, Yunrui Jiao, and Hong Chen. Skip-st: Anomaly detection for medical images using student-teacher network with skip connections. In *ISCAS*, 2023.

[Liu *et al.*, 2023d] Tongkun Liu, Bing Li, Xiao Du, Bingke Jiang, Xiao Jin, Liuyi Jin, and Zhuo Zhao. Component-aware anomaly detection framework for adjustable and logical industrial visual inspection. *AEI*, 2023.

[Liu *et al.*, 2024] Jiaqi Liu, Kai Wu, Qiang Nie, Ying Chen, Bin-Bin Gao, Yong Liu, Jinbao Wang, Chengjie Wang, and Feng Zheng. Unsupervised continual anomaly detection with contrastively-learned prompt. In *AAAI*, 2024.

[Luo *et al.*, 2024] Wei Luo, Haiming Yao, and Wenyong Yu. Template-based feature aggregation network for industrial anomaly detection. *EAAI*, 2024.

[OpenAI, 2023] OpenAI. Gpt-4v(ision) system card. 2023.

[Pang *et al.*, 2022] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Roth *et al.*, 2022] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

[Rudolph *et al.*, 2022] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *WACV*, 2022.

[Ruff *et al.*, 2021] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, et al. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

[Shi *et al.*, 2020] Yong Shi, Jie Yang, and Zhiquan Qi. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 2020.

[Singh *et al.*, 2023] Manpreet Singh, Fujun Ruan, Albert Xu, Yuchen Wu, Archit Rungta, Luyuan Wang, Kevin Song, Howie Choset, and Lu Li. Toward closed-loop additive manufacturing: Paradigm shift in fabrication, inspection, and repair. In *IROS*, 2023.

[Surís *et al.*, 2023] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023.

[Tao *et al.*, 2022] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *TIM*, 71:1–21, 2022.

[Wang *et al.*, 2023] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *CVPR*, 2023.

[Wellhausen *et al.*, 2020] Lorenz Wellhausen, René Ranftl, and Marco Hutter. Safe robot navigation via multi-modal anomaly detection. *RAL*, 2020.

[Xie *et al.*, 2023] Guoyang Xie, Jingbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. *ICLR*, 2023.

[Xu *et al.*, 2023] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

[Yao *et al.*, 2023a] Haiming Yao, Wenyong Yu, Wei Luo, Zhenfeng Qiang, Donghao Luo, and Xiaotian Zhang. Learning global-local correspondence with semantic bottleneck for logical anomaly detection. *TCSVT*, 2023.

[Yao *et al.*, 2023b] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*, 2023.

[Zavrtanik *et al.*, 2021] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

[Zavrtanik *et al.*, 2024] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Cheating depth: Enhancing 3d surface anomaly detection via depth simulation. In *WACV*, 2024.

[Zhang *et al.*, 2023a] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *CVPR*, 2023.

[Zhang *et al.*, 2023b] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023.

[Zhang *et al.*, 2023c] Jiangning Zhang, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, and Yong Liu. Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. *arXiv preprint arXiv:2311.02612*, 2023.

[Zhang *et al.*, 2023d] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[Zhang *et al.*, 2024] Jie Zhang, Masanori Suganuma, and Takayuki Okatani. Contextual affinity distillation for image anomaly detection. In *WACV*, 2024.

[Zhou *et al.*, 2023] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad:

A dataset and benchmark for pose-agnostic anomaly detection. *NeurIPS*, 2023.

[Zhou *et al.*, 2024] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

[Zou *et al.*, 2022] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.