Generalist Multi-Class Anomaly Detection via Distillation to Two Heterogeneous Student Networks

Hangil Park Yongmin Seo Tae-Kyun Kim
School of Computing, KAIST
{hangil.park, yongmin.seo, kimtaekyun}@kaist.ac.kr

Abstract

Anomaly detection (AD) plays an important role in various real-world applications. Recent advancements in AD, however, are often biased towards industrial inspection, struggle to generalize to broader tasks like semantic anomaly detection and vice versa. Although recent methods have attempted to address general anomaly detection, their performance remains sensitive to dataset-specific settings and single-class tasks. In this paper, we propose a novel dual-model ensemble approach based on knowledge distillation (KD) to bridge this gap. Our framework consists of a teacher and two student models: an Encoder-Decoder model, specialized in detecting patch-level minor defects for industrial AD and an Encoder-Encoder model, optimized for semantic AD. Both models leverage a shared pre-trained encoder (DINOv2) to extract high-quality feature representations. The dual models are jointly learned using the Noisy-OR objective, and the final anomaly score is obtained using the joint probability via local and semantic anomaly scores derived from the respective models. We evaluate our method on eight public benchmarks under both single-class and multi-class settings: MVTec-AD, MVTec-LOCO, VisA and Real-IAD for industrial inspection and CIFAR-10/100, FMNIST and View for semantic anomaly detection. The proposed method achieved state-of-the-art accuracies in both domains, in multi-class as well as single-class settings, demonstrating generalization across multiple domains of anomaly detection. Our model achieved an image-level AUROC of 99.7% on MVTec-AD and 97.8% on CIFAR-10, which is significantly better than the prior general AD models in multi-class settings and even higher than the best specialist models on individual benchmarks.

1 Introduction

Anomaly Detection (AD), is a critical task across a wide range of applications, from industrial quality control to cybersecurity, healthcare, and autonomous systems. The goal of AD is to identify patterns that deviate from the norm, representing potential defects, attacks, or abnormalities. As such, detecting anomalies can prevent system failures, improve security, and ensure quality and reliability in automated processes. One area where AD has been particularly impactful is industrial inspection, where AD systems play a key role in automating the detection of manufacturing defects. In these contexts, anomalies such as scratches, dents, or missing components need to be identified with a high precision to ensure product quality and reduce costs. Despite the success of AD methods in such specialized tasks, a major challenge remains: generalizability across domains.

A major group of recent work in anomaly detection, particularly those based on datasets like MVTec-AD [4], have concentrated on industrial applications. These methods commonly employ techniques such as patch embedding [5, 6, 7, 8], pseudo-anomaly generation [9, 10, 11, 12, 13] and knowledge distillation (KD) [14, 15, 16, 17, 18, 19, 20, 21, 22, 1, 2, 23]. Recent KD-based methods [15, 17, 18] show state-of-the-arts performance on industrial anomaly detection (AD). In these methods, input samples are processed through a pre-trained teacher model, a bottleneck, and a decoder student

Preprint.

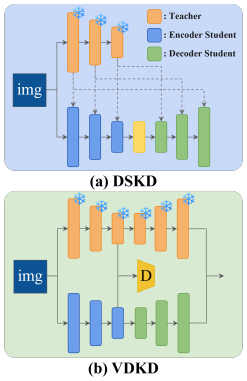


Figure 1: **Prior KD methods with dual students.** (a) DSKD [1], (b) VDKD [2].

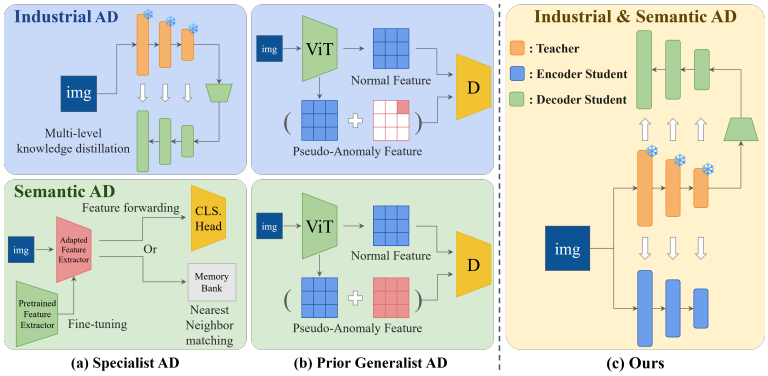


Figure 2: **Comparison among previous anomaly detection methods and ours.** (a) Specialist AD: KD methods(top), major semantic AD methods(bottom), (b) Prior Generalist AD [3], (c) Proposed Generalist AD.

model. The bottleneck and student decoder model are trained to reconstruct the teacher's multi-level features [15]. The KD methods [1, 23, 2] also adopted dual-student models shown in Fig. 1. These methods employ two student models to enhance representation learning. Specifically, they leverage these models for extracting richer feature representations [1], guiding the student to capture global contextual information through an autoencoder, and learning discriminative features using MAE [2]. While the aforementioned methods yield impressive accuracies in detecting specific defect types for industrial AD, they often underperform in broader anomaly detection tasks involving multiple classes or in domains that require semantic understanding of data e.g., CIFAR-10/100.

Recently, a few approaches have aimed to achieve state-of-the-art performance in anomaly detection across both industrial inspection and semantic anomaly datasets [24, 3]. These methods, however, depend on specific parameter choices, backbone networks, or dataset-specific prior knowledge. GeneralAD [3] achieves general anomaly detection via the ViT feature extractor and patch-wise discriminator. Performance of GeneralAD heavily relies on synthetic anomaly generation (See the section E for details). They use different anomaly generation strategies for different domains (e.g., add noise to a random patch feature for industrial AD and add noise to all patches for semantic AD). Additionally, the previous methods primarily focus on one-vs-rest setting, rather than more challenging and generalized multi-class settings.

In this paper, we propose a novel knowledge distillation-based approach that addresses the gap from existing industrial anomaly detection to general multi-class anomaly detection. It exploits an encoder-decoder model for detecting local defects in industrial settings and an encoder-encoder model for semantic anomaly detection, preserving the semantic integrity of input images, as shown in Fig. 2(c). This dual-model framework achieves state-of-the-art performance on both industrial and semantic anomaly detection tasks, in multi-class as well as single-class settings, demonstrating the generalizability and robustness of our method across diverse domains. The main contributions are as follows:

- We propose a novel dual KD-based architecture and learning method for generalist anomaly detection, capable of handling a variety of anomaly types.
- In the encoder-encoder pair, we introduce the class token discrepancy for better detecting semantic-level irregularities. The encoder architecture is advantageous to represent the whole image semantic contexts.
- The decoder student which maps back to image pixels convey local information, thus being better at detecting local defects in industrial settings.
- The proposed method is significantly better than previous general AD methods in multi-class settings, and on par with the best specialist models on each benchmark. Our method achieves SOTA results on six out of eight public benchmarks.
- Our method also demonstrates sample efficiency when only few-shots are available per class.

2 Related work

Industrial Defect Detection. In these applications, anomalies include manufacturing defects such as cracks, scratches, and surface irregularities, which need to be detected with precision to ensure product quality. Recent methods mainly rely on unsupervised learning, where models are trained using only normal data. These methods commonly employ techniques such as patch embedding [5, 6, 7, 8], pseudo-anomaly generation [9, 10, 11, 12, 13, 6], contrastive learning [17, 25] and knowledge distillation (KD) [14, 15, 17, 18, 19, 20, 21, 22, 1, 23, 2]. In the KD based methods, the student model learns to imitate the teacher's feature and is expected to reconstruct well normal samples and struggle for abnormal samples. MKD [14] first suggests KD based anomaly detection by adopting a shallow student encoder model in the distillation framework. RD4AD [15] shows that the reverse distillation framework using a bottleneck and deep decoder student model is a better option for industrial defect detection. UniAD [26] proposed a reconstruction-based multi-class method via a customized ViT architecture. UniAD addresses "identical shortcut" issues in industrial AD, where models tend to learn an identity mapping rather than reconstruction of normal samples. The shortcut reconstructs well for anomalous samples either, leading to a low detection accuracy. Dinomaly [18] also falls to the KD category and and raises concerns that dense, layer-to-layer distillation may cause the student model to overly imitate the teacher's behavior, potentially exacerbating identical shortcut problems, as noted in [27]. Instead, it advocates a looser loss formulation for KD. [1, 23, 2] utilize two different types of student models with knowledge distillation. [1] employs an encoder student and a decoder student, with the teacher distilling knowledge to both. While the encoder student receives the same input as the teacher network, the decoder student takes the output of the encoder student as its input. [23] consists of three components: a teacher, an autoencoder, and a student. The teacher and autoencoder distill local and global information to the student, respectively. [2] is based on the ViT-MAE architecture, where the teacher distills knowledge to the student at both the encoder and decoder levels. For the encoder output, the student is guided by a discriminator that determines whether the feature originates from the teacher, whereas for the decoder output, it is guided using pixel-level labels (e.g., anomaly maps). All methods are, however, tested on industrial AD. In contrast, our proposed method employs two explicit student networks that complement each other, and is demonstrated for semantic as well as industrial AD. Several methods [26, 28, 29] use multi-class settings where a single AD model is trained over all classes of a benchmark than a model per class. [29] applies Mixture-of-Experts framework on intermidiate layers to deal with various input types of multi-class industrial AD. They are yet less applicable to semantic AD tasks where more comprehensive understanding of images is required.

Semantic Anomaly Detection. Semantic AD, often framed as one-class classification problems or Out-of-Distribution detection, tackles cases where anomalies look normal alone but abnormal and normal samples belong to different semantic classes. For example, in CIFAR-10, the class of dog can be defined abnormal against other normal classes e.g., cat.

Unlike the industrial inspection,

semantic AD requires understanding of contextual relations among objects and surroundings. Transformaly [30] achieves SOTA performance on CIFAR-10 by leveraging two feature spaces: one extracted from a pre-trained Vision Transformer (ViT) and another derived from a teacher-student framework. Additionally, [31] introduces the mean-shifted contrastive loss specifically designed for anomaly detection, while FITYMI [10] employs diffusion-based data augmentation to generate pseudo-anomalies for supervised learning. Despite their success in semantic anomaly detection, these methods perform poorly in industrial settings, where normal and abnormal samples often share common semantic features.

In the survey [32, 33] several works [34, 35, 36, 37] are categorised to post-hoc methods that achieve SOTA semantic AD accuracies exploiting the powerful backbone networks.

General Anomaly Detection. Recently, a few methods have been proposed for both industrial inspection and semantic anomaly detection [24, 3]. Uniformaly [24] enhances anomaly detection by eliminating less informative background patches through Back Patch Masking (BPM), followed by anomaly identification using a memory bank based top k-ratio feature matching technique. GeneralAD [3] employs a patch-wise discriminator, which classifies input patch features as normal or anomalous. To train the discriminator, normal and abnormal patch features are required. The normal patch features are extracted from normal samples using a pre-trained ViT, while the pseudo-anomalies

are generated by applying feature distortion to these normal patches. Both methods, however, have limited adaptation to new datasets.

Uniformaly relies heavily on explicit parameters (e.g., k-ratio), and its optimal backbone network varies depending on datasets. GeneralAD's pseudo-anomaly generation strategy is contingent on dataset prior knowledge. In contrast, our method operates without reliance on dataset-specific prior knowledge and is evaluated at more challenging and general multi-class settings.

Adaptation. Also, there are some few-shot methods [38, 39, 40] leveraging pre-trained Visual-Language Models (VLMs) like CLIP [41], which demonstrate superior generalization (via adaptation) capabilities. These methods require manual prompts for each dataset and heavily rely on VLMs. AnomalyCLIP [38], AdaCLIP [39] require auxiliary datasets containing numerous both normal and anomalous images for training and their application is limited to industrial and medical domains. In contrast, we address a novel model architecture and its training strategies that can be applied to any domains, not efficient adaptation techniques.

Relation to Broader Areas. AD also relates to other areas, such as active learning [42, 43] and semi-supervised learning [44, 45]. The core idea of active learning is to select unlabeled samples that maximally improves the model if labeled. Active learning (AL) often picks them up using diversity- or outlier-based strategies and anomalies in AD appear as outliers/different points from normal data. Semi-supervised learner makes use of a large pool of unlabeled data with the small labeled set, improving generalisation by learning the structure of unlabeled data, which also grounds AD from normal data. Often proposed frameworks for SSL are adapted to AL, and vise-versa.

3 KD based generalist multi-class AD

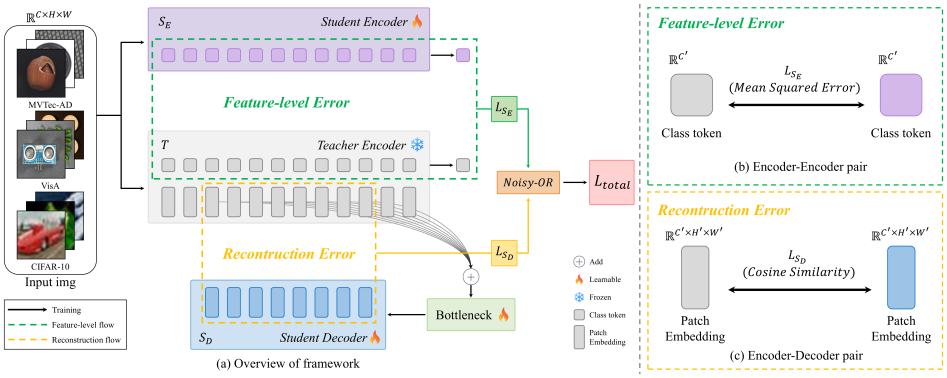


Figure 3: **Framework overview.** Our method consists of three components: a pre-trained teacher encoder, an encoder student, and a decoder student, which form encoder-encoder and encoder-decoder pairs. The encoder-decoder pair is intended to detect local anomalies through the reconstruction error, while the encoder-encoder pair is designed to capture semantic anomalies by leveraging feature-level errors and class tokens. Noisy-OR function is applied to fuse the errors from each pair, enabling an unified anomaly detection framework.

3.1 Method overview

As shown in Fig. 3, our approach combines two teacher-student models based on knowledge distillation (KD) within an ensemble framework. Specifically, we employ an Encoder-Decoder model and an Encoder-Encoder model, both of which share a pre-trained backbone for feature extraction. The encoder student observes the whole input images and captures semantic distributions, while the decoder student reconstructs image pixels or such representations, thus maintaining minor local changes.

Our encoder-decoder pair model is optimized for industrial anomaly detection. This model uses an encoder to extract features from input images and a decoder to reconstruct the encoder features. Anomalies are detected based on discrepancies between the teacher's encoding and the student's reconstruction. This model is effective in identifying structural anomalies, such as defects in industrial products, where patch-level differences are critical. Our encoder-encoder pair model is

designed for semantic anomaly detection. This model is based on an encoder architecture for both the teacher and student networks. The teacher model is pre-trained on ImageNet using DINOv2 [46, 47], while the student model is trained from scratch via distillation on target datasets. This model is capable of identifying anomalies that involve semantic irregularities by leveraging high-level feature representation and class tokenization. The two models are integrated into a shared backbone architecture that provides general feature representations, ensuring both models benefit from high-quality feature embeddings. The models are jointly trained using the Noisy-OR objective.

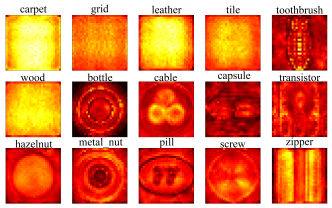
3.2 Encoder-Decoder pair

The Encoder-Decoder model is tailored for industrial anomaly detection, where anomalies typically involve local structural defects that can be detected through patch-level analysis. For the patch-level analysis, we use a ViT architecture for both the teacher and student model.

Dinomaly [18] is adopted for the student model. The teacher model T is an encoder that extracts feature embeddings from input images, while the student model S_D is a decoder that reconstructs the input images from the teacher's feature embeddings. The reconstruction error between the original input and the reconstructed image is used to detect anomalies. The model analyzes anomalies at the patch level, focusing on discrepancies in small, localized regions of images. Large reconstruction errors in specific patches indicate the presence of an anomaly in local areas. Fig. 4 visualises the variance of feature embedding maps from DINOv2. We observe that the variance in feature maps for MVTec-AD dataset displays class specific patterns, whereas the feature maps for CIFAR-10 exhibit no discernible pattern. This supports the effectiveness of our decoder model and its patch-wise features, which is particularly beneficial to industrial AD scenarios. We adopt the loss function from [18], which relaxes the element-wise comparison. This loss takes a large value when the cosine similarity between the features of teacher student is low. It forces the student model to reconstruct well the teacher features for normal input samples in the training set. The loss function for our decoder student S_D is following:

$$L_{S_D}(x) = \frac{1}{2}\sum_{i=1}^{2} 1 - \cos(vec(F_T^i(x)), vec(F_{S_D}^i(x))), \tag{1}$$

where $x \in \mathbb{R}^{C\times H\times W}$ is an input image, vec is the vectorization function, $F_T^i(x) = \frac{1}{4}\sum_{j=4i-1}^{4i+2} f_T^j(x)$, $F_{S_D}^i(x) = \frac{1}{4}\sum_{j=4i-3}^{4i} f_{S_D}^j(x)$, $T : \mathbb{R}^{C\times H\times W} \to \{f^j \in \mathbb{R}^{C'\times H'\times W'}\}_{j\in\{1,\dots,12\}}$, $S_D : \mathbb{R}^{C\times H\times W} \to \{f^j \in \mathbb{R}^{C'\times H'\times W'}\}_{j\in\{1,\dots,8\}}$, $f_T^j, f_{S_D}^j$ is the j-th level patch feature of the teacher and decoder student respectively. The mid-level layers of the teacher are found to be more robust to input noise and global representations, which are not suited to detecting local defects in industrial AD.



(a) MVTec-AD

(b) CIFAR-10

Figure 4: **Variance of feature maps from DINOv2.** Brighter regions indicate larger variances. Normal samples from the training set are used.

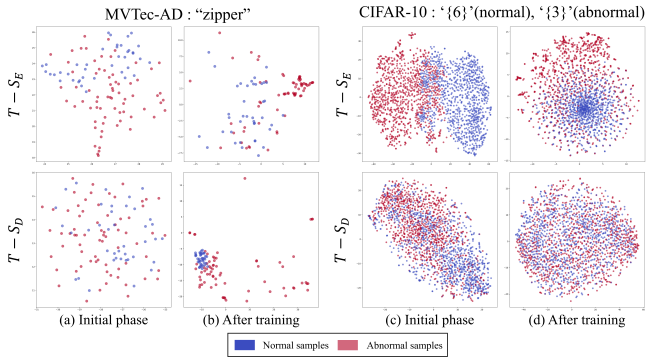


Figure 5: **Feature t-SNE visualization.** We visualize the feature distribution shift during the training phase for MVTec-AD ((a), (b)) and CIFAR-10 ((c), (d)). T-S_E and T-S_D represent the feature difference between the teacher and the encoder student, and the feature difference between the teacher and the decoder student, respectively.

3.3 Encoder-Encoder pair

The Encoder-Encoder model is designed to address semantic anomaly detection, where anomalies manifest as subtle, high-level irregularities within data. The techniques applied in industrial AD, such as alternative data flows through a decoder student model [15, 17, 26, 28, 18], bottlenecks [15, 17, 18], and dropout [18], effectively prevent the student model from learning an identity mapping but are less suitable for semantic AD. These methods convey information reduction, which impacts semantic anomaly detection. To mitigate this issue, we adopt an encoder-based student model, as in [14, 30]. Inspired by the loose loss strategy proposed in [18], we introduce class token-wise distillation, replacing the conventional use of all patch tokens for distillation. Our encoder student model leverages class tokens from each layer of the teacher model, along with the final output class token. Since class token supervision is significantly sparser compared to full patch token supervision, it reduces the likelihood of the encoder student learning an identity mapping.

Both the teacher and student models are based on DINOv2's encoder architecture. The teacher model, pre-trained on ImageNet, generates class tokens that encapsulate the high-level semantic content of input data. The student model, trained from scratch via distillation, learns to replicate these class tokens, embodying the semantic structure of data. Anomalies are detected by evaluating the discrepancy between the class tokens generated by the teacher and those by the student; a substantial discrepancy indicates that the input deviates from the normal patterns learned by the teacher, signaling semantic anomaly. Our loss function for the encoder student is following:

$$L_{S_E}(x) = \frac{1}{m}\sum_{j=1}^{m} ||CLS_T^j(x) - CLS_{S_E}^j(x)||^2, \tag{2}$$

where x is an input image, $T, S_E : \mathbb{R}^{C\times H\times W} \rightarrow \{CLS^j \in \mathbb{R}^{C'}\}_{j\in\{1,\dots,m\}}$, $CLS_T^j, CLS_{S_E}^j : \mathbb{R}^{C'}$ are the j-th level class token of the teacher and encoder student model respectively, m is the number of ViT blocks, including the final class token.

3.4 Learning models

To develop an unified general anomaly detection framework, we integrate the local anomaly score from the Encoder-Decoder model with the semantic anomaly score from the Encoder-Encoder model. These scores are combined using the Noisy-OR function [48, 49] as $P(x) = 1 - \frac{\exp(L_{S_E}(x))}{1+\exp(L_{S_E}(x))}\frac{\exp(L_{S_D}(x))}{1+\exp(L_{S_D}(x))}$, $P(x)$ denotes the probability that x is normal. The probability gets close to 1, only if one of the two model scores is sufficiently small i.e. detecting normal samples. The probability is low if both the models take large values, i.e. detecting anomalies. The total loss function is $L_{total} = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \log(P(x_i)) + (1-y_i)\log(1-P(x_i))\}$

where x_i is the i-th input image, y_i is the binary label for x_i (1 for normal, 0 for abnormal) and n is the number of training samples. The loss function takes the form of binary cross-entropy loss. In our experiments, only normal samples are accessible and used during the training phase, which aligns with the conventional AD settings. The proposed learning framework, however, can be extended to supervised settings where abnormal samples are also available during training. Taking the derivative of L_{total}, we observe the influence of co-training on each model. Let $L_{total}(x)$ be the loss of a normal sample x.

$$\frac{\partial L_{total}(x)}{\partial \theta_{S_E}} = \frac{\partial}{\partial \theta_{S_E}}\{-\log(P(x))\} = -\frac{1-P_{S_D}(x)}{P(x)}\frac{\partial P_{S_E}(x)}{\partial \theta_{S_E}}, \tag{3}$$

where θ_{S_E} denotes the parameters of the encoder student and $P_{S_E}(x) = \frac{1}{1+\exp(L_{S_E}(x))}$, $P_{S_D}(x) = \frac{1}{1+\exp(L_{S_D}(x))}$ is the probability that x is normal predicted by the model S_E and S_D, respectively. If the decoder student predict that x is normal with a high probability, the term $\frac{1-P_{S_D}(x)}{P(x)}$ becomes small, resulting in a reduced gradient for updating the encoder model and vice versa.

Fig. 5 illustrates the t-SNE [50] visualization of feature distributions at different stages of learning, highlighting the discrepancy between the teacher and student models. Since our method identifies anomalies by measuring the difference between the teacher and student models, normal samples converge to the same point, whereas anomalous samples remain dispersed. In the initial phase (Fig. 5(a), (c)), both the discrepancy between the teacher and the encoder student (T-S_E) and the

discrepancy between the teacher and the decoder student (T-S_D) fail to effectively differentiate between normal and anomalous samples, exhibiting similarly scattered patterns for both. As training progresses (Fig. 5(b), (d)), distinct separation patterns emerge. On the MVTec-AD dataset, T-S_E struggles to cluster normal samples into a single point, whereas T-S_D effectively separates normal and anomalous samples by consolidating normal samples into a distinct cluster. Conversely, on the CIFAR-10 dataset, the encoder student (S_E) progressively learns to distinguish the two classes, resulting in a well-defined separation between normal and anomalous instances. These observations underscore the complementary roles of the two student models in different anomaly detection scenarios.

3.5 Inference

To aggregate the anomaly detection results from the two KD models, the anomaly score is defined as $AC(x) = 1 - P(x) = \frac{\exp(L'_{S_E}(x))}{1+\exp(L'_{S_E}(x))} \frac{\exp(L_{S_D}(x))}{1+\exp(L_{S_D}(x))}$, where x denotes the input image, S_E and S_D refer to the encoder student and decoder student models, respectively. Unlike the training phase, we use only the last layer class tokens for measuring the anomaly score of the encoder student $L'_{S_E}(x) = ||CLS^m_T(x) - CLS^m_{S_E}(x)||^2$, where $CLS^m_T, CLS^m_{S_E}$ are the m-th (which is the last) class token of the teacher and encoder student respectively. Although the earlier layers of ViT aggregate features from other input patches, they retain less global structural information compared to the later layers. Note ViT is primarily trained using features from the final layer [51, 52]. The use of only the last class token that captures high-level understanding has been shown more effective for semantic anomaly detection in experiments. By leveraging the strengths of both models through Noisy-OR, our ensemble approach achieves state-of-the-art performance across two distinct domains—industrial and semantic—under both single-class and multi-class settings, demonstrating robust generalization across diverse anomaly detection scenarios.

4 Experiments

4.1 Experiment Setup

Datasets. We conduct experiments on eight benchmark datasets: MVTec-AD [4], MVTec-LOCO [53], VisA [54], Real-IAD [55] for industrial anomaly detection and CIFAR-10/100 [56], Fashion-MNIST [57], View [58] for semantic anomaly detection.

Evaluation Metrics. We use the Area Under the Receiver Operating Characteristic Curve (AUROC) as the metric to evaluate anomaly detection performance. AUROC is widely used for binary classification tasks and is particularly useful for anomaly detection, as it measures the model ability to distinguish between normal and anomalous samples. Higher AUROC scores indicate better performances.

Class settings. To demonstrate the generalization of the proposed method we conduct experiments in two different AD settings; multi-class and single-class. In the **single-class setting**, we train a model for each class. We thus obtain as many models as classes and presume to know which class models to use, as in previous works. On the other hand, in the **multi-class setting**, a model is learned over all classes. The latter does not require much memory and prior-knowledge on classes, which is more general and easy to use in practice. For the multi-class semantic AD, experiments were done using four different splits per dataset, varying the normal class indices according to the settings of UniAD [26]. In each split, half of the classes were designated as normal, while the remaining half were considered abnormal. The detailed class settings, including data splits for each dataset, are provided in Section C.1.

4.2 Results

Industrial Defect Detection. We evaluate our model's performance on industrial anomaly detection using the MVTec-AD, MVTec-LOCO, VisA, and Real-IAD datasets, benchmarking over state-of-the-art methods, including RD4AD [15], SimpleNet [6], DeSTSeg [19], ReContrast [25], UniAD [26], Dinomaly [18], and GeneralAD [3]. For the methods where multi-class AD accuracies are not reported in their original works, we referenced the results from the benchmark studies [59, 18] to ensure comprehensive comparisons.

Table 1: **Anomaly detection results under multi-class setting on on various datasets.** The best results in image-level AUROC(%) are **bold** and the second best results are underlined. If the original studies and the benchmark studies [59] did not provide results for certain datasets, these were marked as unavailable ('-').

	Category	KNN[35]	Transformaly[30]	MSC[31]	RD4AD[15]	SimpleNet[6]	DeSTSeg[19]	ReContrast[25]	UniAD[26]	Dinomaly[18]	GeneralAD[3]	Ours
Industrial	MVTec-AD	89.7	87.8	71.1	94.6	95.3	89.2	98.3	96.5	99.6	56.8	**99.7**
	VisA	81.1	65.0	-	92.4	87.2	88.9	88.9	88.8	98.7	89.7	**98.8**
	MVTec-LOCO	74.8	57.2	-	73.7	81.8	81.2	-	78.7	82.0	82.4	**86.1**
	Real-IAD	-	-	-	82.7	54.9	79.3	82.3	83.1	**89.3**	-	88.7
Semantic	CIFAR-10	92.1	93.6	89.7	-	-	-	-	87.2	71.5	93.9	**97.8**
	CIFAR-100	87.0	85.2	83.7	-	-	-	-	-	65.1	89.7	**92.3**
	FMNIST	88.2	85.6	89.3	-	-	-	-	-	82.5	**92.4**	88.3
	View	65.5	82.2	66.6	-	-	-	-	-	66.6	70.5	**87.3**

Table 2: **Anomaly detection results under single-class settings in image-level AUROC(%).** The best results in image-level AUROC(%) are **bold** and the second best results are underlined. Missing results from the original studies and the benchmark studies [59] for certain datasets are marked as ('-').

	Category	KNN[35]	Transformaly[30]	MSC[31]	RD4AD[15]	SimpleNet[6]	PatchCore[5]	ReContrast[25]	UniAD[26]	Dinomaly[18]	GeneralAD[3]	Ours
Industrial	MVTec-AD	90.2	84.0	87.2	98.4	99.6	99.2	99.5	96.6	99.7	99.2	**99.8**
	VisA	-	-	-	96.0	87.9	94.2	97.5	-	98.9	95.9	**99.0**
	MVTec-LOCO	-	-	-	79.7	77.6	80.3	82.1	-	81.9	84.9	**86.0**
	Real-IAD	-	-	-	-	-	67.2	-	-	**92.0**	-	91.3
Semantic	CIFAR10	97.7	98.3	97.5	86.5	86.5	64.1	84.1	-	90.4	**99.3**	**99.3**
	CIFAR-100	97.0	97.3	96.4	-	-	-	84.0	-	90.3	**98.4**	97.2
	FMNIST	94.2	94.4	95.0	95.0	87.4	77.4	92.4	-	-	**95.2**	94.7
	View	94.6	95.8	95.1	-	76.8	-	-	-	-	95.9	**96.5**

As shown in the upper section of Tab. 1 and Tab. 2, the proposed method achieves consistently high AUROC scores across all four industrial datasets at both single-class and multi-class settings. Notably, we obtain a significant accuracy gain compared to Dinomaly on MVTec-LOCO. The dataset includes both minor defects and logical anomalies which is hard to be dealt with by prior specialist methods. This improvement supports that the two student models in our framework jointly tackles complex and mixed anomaly types. Detailed class-wise evaluation results for MVTec-AD are presented in Tab. 14.

Semantic Anomaly Detection. We evaluate our model on the CIFAR-10/100, Fashion-MNIST, and View datasets for semantic anomaly detection, in comparison to state-of-the-art methods such as KNN [35], Transformaly [30], MSC [31], UniAD [26], Dinomaly [18], and GeneralAD [3]. As shown in the lower section of Tab. 1 and Tab. 2, our approach achieves high AUROC scores across all four semantic anomaly detection datasets. Detailed evaluation results on CIFAR-10 with varying normal indices are presented in Tab. 15. While the CIFAR-10/100 and View datasets capture natural images, FMNIST—a small grayscale image dataset (28x28)—differs significantly from natural images, making it difficult to maintain consistently high performance across diverse semantic anomaly detection datasets. Interestingly, as shown in Tab. 1 and Tab. 2, the methods specializing in industrial anomaly detection generally perform better on FMNIST than on other semantic AD datasets. Note our method demonstrates lower performance variance as well as the highest mean performance across the four semantic AD datasets. The decoder student in our method, which focuses on industrial AD, enhances performance on FMNIST. By leveraging the strengths of both models, our dual approach effectively addresses these complex anomaly types.

Discussions. The experimental results demonstrate that the proposed method outperforms existing SOTA methods in both industrial and semantic anomaly detection. The key to this success lies in the dual-model ensemble, where the Encoder-Encoder model specializes in detecting high-level semantic anomalies, and the Encoder-Decoder model excels at capturing fine-grained structural differences. We also observe that the accuracy of GeneralAD on MVTec-AD drops drastically when applied to multi-class settings. The results show that in GeneralAD the choice of anomaly synthesis strategy is crucial for their performance. On the other hand, our method shows robust performance without any dataset-specific knowledge and different anomaly types at both multi-class and single-class settings.

4.3 Ablation study

To better understand the contributions of each component within our ensemble framework, an ablation study is performed by evaluating the performance of partial models and comparing these results with the comprehensive version of our method. Tab. 3 presents the results, where we evaluate multi-class anomaly detection performance on MVTec-AD, MVTec-LOCO, and CIFAR-10 by varying three experimental conditions. The first two conditions, S_E and S_D, assess the effectiveness of each of the student models in our dual-model framework. The third option, CLS^m, indicates whether the final class token is used for inference. When CLS^m is not selected, we calculate $AC(x)$ using $L''_{S_E}(x)$ instead of $L'_{S_E}(x)$ where $L''_{S_E}(x) = \frac{1}{m-1}\sum_{i=1}^{m-1} ||CLS_T^j(x) - CLS_{S_E}^j(x)||^2$.

The last option, Noisy-OR stands for enabling Noisy-OR ensemble approach. When Noisy-OR is disabled, we simply sum $L'_{S_E}(x)$ and $L_{S_D}(x)$ to aggregate outputs.

Table 3: Effect of each component in the proposed method on MVTec-AD, MVTec-LOCO and CIFAR-10. L_{S_E}: Loss function of encoder student. L_{S_D}: Loss function of decoder student. CLS^m: Using the last class token for inference. Noisy-OR: Using Noisy-OR ensemble approach. The best results are **bold** and the second best results are underlined.

L_{S_E}	L_{S_D}	CLS^m	Noisy-OR	MVTec-AD	MVTec-LOCO	CIFAR-10
✓				96.46	78.09	85.84
	✓			99.64	82.51	71.02
✓		✓		96.58	79.23	97.71
✓	✓		✓	99.65	84.86	83.92
✓	✓	✓		**99.73**	85.65	**97.90**
✓	✓	✓	✓	99.66	**86.10**	97.80

The ablation study results emphasize the significance of combining both the Encoder-Encoder and Encoder-Decoder models. While the Encoder-Encoder model performs effectively on CIFAR-10, its performance on MVTec-AD is lower than that of the Encoder-Decoder model. Conversely, the Encoder-Decoder model excels at industrial anomaly detection but underperforms in semantic tasks. When combined, the ensemble achieves state-of-the-art performance across both datasets. The results also demonstrate that utilizing the last class token for inference notably enhances the encoder student model, particularly on the CIFAR-10 semantic AD dataset. Note that Noisy-OR is particularly effective for MVTec-LOCO which have complex types of anomalies (e.g., logical anomalies). We also report the enhanced performance of previous methods in Tab. 4 to ensure a fair comparison. Our method achieves the best performance among all AD methods when the same backbone network is applied. Furthermore, in Tab. 5, we evaluate the complexity of our method using two different backbones and compare it with GeneralAD to demonstrate the efficiency of our proposed approach. In these experiments, we use the single-class setting for GeneralAD. Our method achieves comparable results to GeneralAD, despite GeneralAD utilizing class knowledge in the single-class setting. Since the single-class setting requires class-wise training, the number of parameters in GeneralAD increases by a factor of n, where n is the number of classes in the dataset. Additionally, our method demonstrates reasonable AD performance even when using a smaller backbone network with fewer parameters and faster inference speed.

Table 4: **Comparison of our method with other state-of-the-art methods using DINOv2 backbone.**

Method	Backbone	MVTec-AD	MVTec-LOCO	CIFAR-10
RD4AD	WideResNet50	94.6	73.7	47.7
	DINOv2-B-reg4/14	98.4	81.6	74.7
Transformaly	ViT-B/16	87.8	57.2	93.6
	DINOv2-B-reg4/14	90.0	82.6	**98.2**
GeneralAD	DINOv2-B-reg4/14	56.8	82.4	93.9
Dinomaly	DINOv2-B-reg4/14	99.6	82.0	71.5
Ours	DINOv2-B-reg4/14	**99.7**	**86.1**	97.8

Table 5: **Complexity comparison between our method and GeneralAD.** n is the number of classes, 15 for MVTec-AD, 5 for MVTec-LOCO and 5 for CIFAR-10.

Method	Backbone	Param (M)	Inference Time(ms)	MVTec -AD	MVTec -LOCO	CIFAR-10
GeneralAD	DINOv2-B-reg4/14	$93.1 \times n$	24.932	99.2	84.9	99.3
Ours	DINOv2-S-reg4/14	58.7	65.592	98.8	81.7	93.3
	DINOv2-B-reg4/14	233.1	93.164	99.7	86.1	97.8

5 Few-shot AD scenarios

In Tab. 6, our approach demonstrates strong performance in few-shot anomaly detection scenarios, where abnormal data is unavailable, and even normal data is scarce. This setting is closely aligned with practical, real-world applications. For the methods that do not report few-shot AD performance in their original works, we refer to the few-shot experiments of [3] for fair comparisons. We evaluate the single-class anomaly detection performance on MVTec-AD by varying the number of shots for training the models. Our method surpasses these baseline approaches by a substantial margin, demonstrating that our model achieves high sample efficiency with limited training samples. As in [39, 38, 40], leveraging the strong prior of of vision-language Models (VLMs) has the potential to further enhance the performance.

Table 6: **Few-shot AD performance on MVTec-AD.** Shots indicates the number of normal data samples used in training.

Shots	1	2	4	8
SPADE [7]	71.6	73.4	82.8	84.0
PaDiM [8]	76.1	78.9	80.5	82.0
PatchCore [5]	84.1	87.2	88.5	92.2
GeneralAD [3]	87.5	91.5	92.8	93.6
Ours	**92.6**	**94.6**	**94.9**	**96.3**

6 Conclusions

In this paper, we propose a novel approach to general anomaly detection that integrates two knowledge distillation (KD)-based teacher–student models: an Encoder–Decoder model for industrial anomaly detection and an Encoder–Encoder model for semantic anomaly detection. By incorporating

dedicated branches for industrial and semantic anomalies, the method supports effective training on heterogeneous datasets and enables domain-agnostic inference without prior knowledge of the test domain. The effectiveness of our dual-model architecture and the contribution of each component are validated through extensive experiments on eight public benchmarks and ablation studies. The proposed ensemble method achieves state-of-the-art performance on both industrial and semantic anomaly detection tasks, while also demonstrating strong generalization in multi-class settings involving complex datasets with mixed anomaly types.

Limitations. The limitations of our method are discussed in Section B.

Acknowledgment

This work was supported by NST grant (CRC 21011, MSIT), IITP grant (RS-2023-00228996, RS-2024-00459749, RS-2025-25443318, RS-2025-25441313, MSIT) and KOCCA grant (RS-2024-00442308, MCST).

References

[1] Liyi Yao and Shaobing Gao. Dual-student knowledge distillation networks for unsupervised anomaly detection, 2024.

[2] Yibo Chen, Haolong Peng, Ke Zhu, and Jianming Zhang. Vdkd: A vit-based student-teacher knowledge distillation for multi-texture class anomaly detection. In *2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 73–78, 2024.

[3] Luc PJ Sträter, Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Generalad: Anomaly detection across domains by attending to distorted features. *arXiv preprint arXiv:2407.12427*, 2024.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, June 2019.

[5] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022.

[6] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, pages 20402–20411, 2023.

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

[8] 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.

[9] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, June 2021.

[10] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees G. M. Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it until you make it : Towards accurate near-distribution novelty detection. In *ICLR*, 2023.

[11] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022.

[12] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021.

[13] 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.

[14] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, pages 14902–14912, June 2021.

[15] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, pages 9737–9746, June 2022.

[16] Jihyun Lee, Hangil Park, Yongmin Seo, Taewon Min, Joodong Yun, Jaewon Kim, and Tae-Kyun Kim. Contrastive knowledge distillation for anomaly detection in multi-illumination/focus display images. In *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5, 2023.

[17] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, pages 24511–24520, June 2023.

[18] Jia Guo, Shuai Lu, Weihang Zhang, and Huiqi Li. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection, 2024.

[19] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *CVPR*, pages 3914–3923, 2023.

[20] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Dual-modeling decouple distillation for unsupervised anomaly detection. *arXiv preprint arXiv:2408.03888*, 2024.

[21] Jie Zhang, Masanori Suganuma, and Takayuki Okatani. Contextual affinity distillation for image anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 149–158, 2024.

[22] 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.

[23] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. pages 128–138, January 2024.

[24] Yujin Lee, Harin Lim, Seoyoon Jang, and Hyunsoo Yoon. Uniformaly: Towards task-agnostic unified framework for visual anomaly detection, 2023.

[25] Jia Guo, shuai lu, LIze JIa, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. In *NeurIPS*, 2023.

[26] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, volume 35, pages 4571–4584. Curran Associates, Inc., 2022.

[27] Chen Liang, Jiahui Yu, Ming-Hsuan Yang, Matthew Brown, Yin Cui, Tuo Zhao, Boqing Gong, and Tianyi Zhou. Module-wise adaptive distillation for multimodality foundation models. *NeurIPS*, 36, 2024.

[28] Xi Jiang, Ying Chen, Qiang Nie, Jianlin Liu, Yong Liu, Chengjie Wang, and Feng Zheng. Toward multi-class anomaly detection: Exploring class-aware unified model against inter-class interference, 2024.

[29] Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. Moead: A parameter-efficient model for multi-class anomaly detection. In *ECCV*, pages 345–361. Springer, 2024.

[30] Matan Jacob Cohen and Shai Avidan. Transformaly - two (feature spaces) are better than one. In *CVPRW*, pages 4060–4069, June 2022.

[31] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *AAAI*, volume 37, pages 2155–2162, 2023.

[32] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[33] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

[34] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. 2022.

[35] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 17–23 Jul 2022.

[36] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pages 4921–4930, June 2022.

[37] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, pages 691–708, Cham, 2022. Springer Nature Switzerland.

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

[39] 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.

[40] 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*, pages 17826–17836, 2024.

[41] 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. PmLR, 2021.

[42] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9583–9592, June 2021.

[43] Razvan Caramalau, Binod Bhatarrai, Dan Stoyanov, and Tae-Kyun Kim. Mobyv2al: Self-supervised active learning for image classification, 2022.

[44] Honggyu Choi, Zhixiang Chen, Xuepeng Shi, and Taekyun Kim. Semi-supervised object detection with object-wise contrastive learning and regression uncertainty. In *British Machine Vision Conference (BMVC)*, 2022.

[45] Minju Kang, Taehun Kong, and Tae-Kyun Kim. Semi-supervised 3d object detection with channel augmentation using transformation equivariance. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 638–644, 2024.

[46] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[47] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

[48] Cha Zhang, John Platt, and Paul Viola. Multiple instance boosting for object detection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NeurIPS*, volume 18. MIT Press, 2005.

[49] Tae-kyun Kim and Roberto Cipolla. Mcboost: Multiple classifier boosting for perceptual co-clustering of images and visual features. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NeurIPS*, volume 21. Curran Associates, Inc., 2008.

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

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[52] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021.

[53] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*, 130(4):947–969, 2022.

[54] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *arXiv preprint arXiv:2207.14315*, 2022.

[55] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, pages 22883–22892, 2024.

[56] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10/100 (canadian institute for advanced research), 2009.

[57] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[58] Intel. Intel Image Classification. `https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data`, 2019.

[59] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. Ader: A comprehensive benchmark for multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024.

[60] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298, 2023.

[61] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.