Are Anomaly Scores Telling the Whole Story? A Benchmark for Multilevel Anomaly Detection

Tri Cao[1], Minh-Huy Trinh[3, 4], Ailin Deng[1], Quoc-Nam Nguyen[1], Khoa Duong[2],
Ngai-Man Cheung[2], Bryan Hooi[1]
[1]*National University of Singapore*, [2]*Singapore University of Technology and Design*,
[3]*University of Science*, [4]*Vietnam National University, Ho Chi Minh City.*

A. Evaluation Metrics

A.1. AUROC metric

Area Under the Receiver Operating Characteristic (AUROC) [7] in the context of Multilevel Anomaly Detection (MAD) is defined as:

$$\text{AUROC} = \frac{\sum_{a=1}^{n} \sum_{x_i \in L_a} \sum_{x_j \in L_0} \mathbb{1}(f(x_i) > f(x_j))}{\sum_{a=1}^{n} |L_a| \cdot |L_0|},$$

where:

- n is the number of severity levels.
- L_0 is the set of normal samples.
- L_a (for $a = 1, 2, \ldots, n$) are the sets of anomalous samples, grouped by severity level.
- $f(x)$ is the anomaly score function.
- $\mathbb{1}(\cdot)$ is the indicator function, equal to 1 if the condition inside is true and 0 otherwise.
- $|L_a|$ and $|L_0|$ are the cardinalities (sizes) of the sets L_a and L_0, respectively.

AUROC achieves a perfect score of 1 if *all the anomaly scores of abnormal samples are greater than all the anomaly scores of normal samples.*

A.2. C-index metric

The C-index [13] is a generalization of the AUROC that can evaluate how well anomaly scores align with the severity levels. In the context of MAD, C-index is defined as:

$$C = \frac{\sum_{a=1}^{n} \sum_{b=0}^{a-1} \sum_{x_i \in L_a} \sum_{x_j \in L_b} \mathbb{1}(f(x_i) > f(x_j))}{\sum_{a=1}^{n} \sum_{b=0}^{a-1} |L_a| \cdot |L_b|},$$

A C-index of 1 corresponds to the best model prediction, achieved when *all samples from higher-severity levels are consistently assigned higher anomaly scores than samples from lower-severity levels or normal samples*. A C-index of 0.5 indicates a random prediction.

A.3. Kendall's Tau-b metric

In our paper, in addition to the C-index, we employ Kendall's Tau [8] to evaluate the consistency between anomaly scores and severity levels. Among its variants, we specifically use Kendall's Tau-b, a stricter version designed to account for tied levels. This makes Kendall's Tau-b essential in settings where samples within the same severity level are expected to have identical anomaly scores for perfect consistency (i.e., $\tau_b = 1$).

Suppose a pair of samples (x_i, x_j) is concordant if it follows the same order in terms of severity levels and anomaly scores. That is, if:

1. The severity level of sample x_i is greater than that of sample x_j, and the anomaly score of x_i is also greater than that of x_j or if
2. The severity level of sample x_i is less than that of sample x_j, and the anomaly score of x_i is also less than that of x_j.

The pair is discordant if it is in the reverse ordering for severity levels and anomaly score, or the values are arranged in opposite directions. That is, if:
1. The severity level of sample x_i is greater than that of sample x_j, but the anomaly score of x_i is less than that of x_j or if
2. The severity level of sample x_i is less than that of sample x_j, but the anomaly score of x_i is greater than that of x_j.

The pair is tied if the severity level of sample x_i is equal that of x_j and/or the anomaly score of sample x_i is equal that of x_j.

Kendall's Tau-b is formulated as:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + X_0)(C + D + Y_0)}},$$

where:
- C: The number of concordant pairs,
- D: The number of discordant pairs,
- X_0: The number of pairs tied *only* on severity levels,
- Y_0: The number of pairs tied *only* on anomaly scores.

Note that pairs where both the severity levels and anomaly scores are tied ($(XY)_0$) are excluded from both the numerator and denominator of Kendall's Tau-b. This exclusion ensures that these pairs do not affect the metric's calculation or penalize the performance.

The Kendall's Tau-b ensures that tied pairs only on anomaly scores or only on severity levels are not ignored but instead reduce the overall Kendall's Tau-b value to reflect the uncertainty caused by ties. Consequently, τ_b will be equal to 1.0 when *the ordering of anomaly scores perfectly corresponds to the ordering of severity levels, and all samples within the same severity level have identical anomaly scores*. This characteristic makes Kendall's Tau-b particularly suitable in contexts where it is essential to ensure that a sample with a higher anomaly score always corresponds to a higher severity level. In contrast, the C-index can achieve a perfect score even when anomaly scores within the same severity level are not consistent, as it only considers pairwise comparisons across levels.

B. Prompts and Design Choices for MLLM-based baselines

B.1. Setup

For all experiments using MLLMs, we set the temperature to 0 to enable deterministic generation. The normal images are sourced from the training set and fixed for each subset.

B.2. Prompt design

We design prompts for MLLM-based baselines with the following key components, each carefully structured to guide the model's performance effectively:
- **Context**: Provides detailed information about the normal reference images and the inference images for comparison. This section ensures the model understands the baseline for normalcy and the target images to evaluate.
- **Task Description**: Clearly defines the objectives of the multilevel anomaly detection task. This includes outlining the model's role, such as identifying deviations between the reference and inference images and explaining the nature of potential anomalies.
- **Severity Levels Description**: Describes the various severity levels of anomalies to guide the model in interpreting and giving corresponding anomaly scores. The prompt specifies the characteristics of each level to standardize interpretation.
- **Format Guidelines**: Specifies the required structure and format for the model's response to ensure clarity and consistency. The output format includes the following components:
  - Anomaly Score: A numerical value indicating the level of severity.
  - Reasoning: A concise yet comprehensive explanation of the detected anomaly, providing insights into the specific features or conditions that led to the classification.

Based on this design, five distinct prompts are provided, covering the following tasks: industrial inspection (MVTec-MAD and VisA-MAD), one-class novelty detection (MultiDogs-MAD), pulmonary imaging analysis (Covid19-MAD), diabetic retinopathy detection (DRD-MAD), and skin lesion detection (SkinLesion-MAD) (see text box below).

C. Zero-shot AD and Few-shot AD using MLLMs

In addition to conducting experiments for few-shot settings on MLLMs (as reported in main paper), we also perform experiments under zero-shot settings. Specifically, in the zero-shot settings, we do not use normal images, aiming to evaluate the

Method	MultiDogs-MAD			MVTec-MAD			VisA-MAD			DRD-MAD			Covid19-MAD			SkinLesion-MAD			Average		
	AUC	Ken	C	AUC	Ken	C	AUC	Ken	C	AUC	Ken	C	AUC	Ken	C	AUC	Ken	C	AUC	Ken	C
RRD [12]	81.67	0.510	78.51	99.51	0.518	80.72	93.99	0.620	88.53	61.83	0.243	63.55	84.36	0.240	63.07	99.53	0.558	82.48	86.81	0.448	76.14
PNI [2]	77.20	0.413	73.11	99.32	0.487	78.91	96.07	0.622	88.64	62.50	0.285	65.94	87.96	0.241	63.12	100.0	0.455	76.51	87.17	0.417	74.37
GPT-4o (zero-shot)	98.67	0.942	98.44	82.16	0.525	75.11	68.70	0.422	68.22	50.25	0.044	50.29	50.95	0.064	50.63	61.33	0.180	55.77	68.68	0.363	66.41
Sonnet (zero-shot)	97.05	0.928	97.12	76.73	0.425	69.00	66.65	0.384	65.72	62.97	0.380	66.53	91.36	0.581	78.56	99.96	0.583	79.61	82.46	0.547	76.09
MMAD-4o	98.44	0.933	95.91	95.98	0.646	83.52	79.34	0.621	77.74	65.80	0.433	67.72	88.07	0.547	76.94	99.43	0.694	85.61	87.85	0.646	81.24
MMAD-Sonnet	97.89	0.936	97.34	90.02	0.557	77.33	76.24	0.561	74.86	65.02	0.403	69.30	92.35	0.601	80.54	99.82	0.610	79.23	86.89	0.611	79.77

Table 1. Multilevel AD performance comparison between zero-shot and few-shot learning on MLLM-based baselines and state-of-the-art conventional baselines across six datasets. The results are averaged across all subsets of each dataset. Higher AUROC (AUC) (%), Kendall's Tau-b (Ken), and C-index (C) (%) values indicate better performance. Across most datasets, few-shot learning significantly outperforms zero-shot learning.

model's capability without reference images. The experimental results presented in Table 1 show that the performance of both Binary AD and Multilevel AD in zero-shot settings drops significantly compared to few-shot settings. This highlights the critical importance of using normal images as references in AD tasks.

D. Output Examples of MLLM-Based Baselines

This section provides examples of the output of MLLM-based baselines across datasets in Table 2, 3 and 4.

E. Number of samples across Severity Levels

We provide detailed number of samples across severity levels for each subset. The specifics are presented in Table 5, 6, 7, and 8, corresponding to the details of VisA-MAD, MVTec-MAD, MultiDogs-MAD, and the three medical datasets.

F. Full Results

We provide detailed results for each subset across severity levels on six datasets. The detailed results for each dataset in the MAD-Bench framework are provided in the corresponding tables: VisA-MAD results are in Tables 9 and 10, MVTec-MAD results are in Tables 11, 12, 13, MultiDogs-MAD results are in Table 14, DRD-MAD results are in Table 15, Covid19-MAD results are in Table 16 , and SkinLesion-MAD results are detailed in Table 17.

The prompt for Industrial Inspection (MVTec-MAD and VisA-MAD)

System prompt:
You are a helpful assistant that responds in detecting anomalies on images in various context.

Instruction:
You are given four industrial inspection images:
Image 1, 2, 3: Belong to the non-defect category of {subset_name}.
Image 4: The inference image to be analyzed.

Your task is to determine an Anomaly Score for the inference image, reflecting the economic impact of the defect. A lower anomaly score suggests minimal economic impact, indicating that the item is likely usable or repairable, while a higher score indicates more severe impact, where the item may be unusable or costly to repair.

Use the following guidelines and ranges to assign the anomaly score:
- If the inference image is consistent with the non-defect category, assign a score of 0.
- If the inference image shows minor defects that are unlikely to impact functionality and may be easily repairable, assign a score between 1 and 30.
- If the inference image has defects suggesting moderate economic impact, where the item is still usable but may require repair, assign a score between 31 and 70.
- If the inference image displays severe defects with high economic impact, suggesting the item may be unusable or expensive to repair, assign a score between 71 and 100.

Output format:
Anomaly Score: [Score]
Reason: [Provide a brief explanation of why this score was assigned]
Please ensure the output strictly follows the format. You must answer the question even if you are not entirely certain about the answer.

The prompt for One-class Novelty Detection (MultiDogs-MAD)

System prompt:
You are a helpful assistant that responds in detecting anomalies on images in various context.

Instruction:
You are given four images:
Image 1, 2, 3: Belongs to the normal class {subset_name}.
Image 4: The inference image to be analyzed.

Your task is to determine an anomaly score for the inference image. This score should indicate how similar the inference image is to the class of the normal images. A lower anomaly score means the inference image is similar to the normal class, while a higher score indicates a greater deviation. Take into account the relationship between the class of the normal images and the class of the inference image. Carefully consider that there may be different levels of relationships between the normal class and the various abnormal classes. For example, some abnormal classes might be more closely related to the normal class (e.g., two different breeds of cats), while others might be completely different (e.g., a cat versus a completely unrelated object).

These relationships should reflect in the anomaly score:
- If the inference image belongs exactly to the same class as the normal images, assign a score of 0.
- If the inference image is similar but not identical to the normal class, assign a score between 1 and 50, with lower values indicating greater similarity.
- If the inference image belongs to an abnormal class but has some level of relationship to the normal class, assign a score between 51 and 75, considering the degree of similarity.
- If the inference image is completely unrelated to the normal class, assign a score between 76 and 100, with higher values indicating a greater dissimilarity.

Output format:
Anomaly Score: [Score]
Reason: [Provide a brief explanation of why this score was assigned]
Please ensure the output strictly follows the format. You must answer the question even you are not sure about the anwser.

The prompt for DRD-MAD

System prompt:
You are a helpful assistant that responds in detecting anomalies on images in various context.

Instruction:
You are given four retinal images:
Image 1, 2, 3: Belongs to the normal retinal category (No DR).
Image 4: The inference image to be analyzed.

Your task is to determine an Anomaly Score for the inference image on a scale from 0 to 100, reflecting the extent of diabetic retinopathy (DR). A lower anomaly score suggests minimal or no DR, while a higher score indicates more severe stages of DR.

Use the following guidelines to assign the Anomaly Score:
- If the inference image shows no signs of diabetic retinopathy, assign a score of 0.
- If the inference image shows mild signs of diabetic retinopathy, assign an anomaly score between 1 and 25.
- If the inference image has characteristics suggesting moderate diabetic retinopathy, assign an anomaly score between 26 and 50.
- If the inference image displays severe diabetic retinopathy, assign an anomaly score between 51 and 75.
- If the inference image reveals advanced proliferative diabetic retinopathy, assign an anomaly score between 76 and 100.

Output format:
Anomaly Score: [Score]
Reason: [Provide a brief explanation of why this score was assigned]
Please ensure the output strictly follows the format. You must answer the question even if you are not entirely certain about the answer.

The prompt for Covid19-MAD

System prompt:
You are a helpful assistant that responds in detecting anomalies on images in various context.

Instruction:
You are given four lung scan images:
Image 1, 2, 3: Belongs to the normal lung category.
Image 4: The inference image to be analyzed.

Your task is to determine an anomaly score for the inference image on a scale from 0 to 100, reflecting the extent of lung damage. A lower anomaly score suggests minimal or no lung damage, while a higher score indicates more severe lung damage.

Use the following guidelines to assign the anomaly score:
- If the inference image is consistent with a normal lung, assign a score of 0.
- If the inference image shows very mild lung damage, such as isolated ground-glass opacities, assign a score between 1 and 15.
- If the inference image shows mild lung damage, with limited areas of ground-glass opacities, assign a score between 16 and 30.
- If the inference image has moderate lung damage, such as more widespread ground-glass opacities, assign a score between 31 and 50.
- If the inference image displays moderately severe lung damage, including partial lung consolidation, assign a score between 51 and 70.
- If the inference image shows severe lung damage, such as extensive consolidation in multiple lung regions, assign a score between 71 and 85.
- If the inference image reveals extreme lung damage, affecting more than 85 percent of the lungs, assign a score between 86 and 100.

Output format:
Anomaly Score: [Score]
Reason: [Provide a brief explanation of why this score was assigned]
Please ensure the output strictly follows the format. You must answer the question even if you are not entirely certain about the answer.

The prompt for SkinLesion-MAD

System prompt:
You are a helpful assistant that responds in detecting anomalies on images in various context.

Instruction:
You are given four dermoscopic images:
Image 1, 2, 3: Belong to the normal skin category.
Image 4: The inference image to be analyzed.

Your task is to determine an anomaly score for the inference image, reflecting its potential danger level in comparison to normal skin. A lower anomaly score suggests that the lesion is benign, while a higher score indicates a more dangerous lesion type, such as precancerous or cancerous.

Use the following guidelines and ranges to assign the anomaly score:
- If the inference image is consistent with normal skin, assign a score of 0.
- If the inference image appears benign but shows minor atypical features, such as a mole (nevus) with slightly irregular borders, assign a score between 1 and 30 to indicate a benign lesion.
- If the inference image has characteristics suggesting it could be precancerous, assign a score between 31 and 70.
- If the inference image displays features strongly associated with malignancy, assign a score between 71 and 100, indicating a high risk for cancer.

Output format:
Anomaly Score: [Score]
Reason: [Provide a brief explanation of why this score was assigned]
Please ensure the output strictly follows the format. You must answer the question even if you are not entirely certain about the answer.

<table>
<tr>
<td>Normal Image
 
Inference Image
  </td>
<td>Subset: Hazelnut in MVTec-MAD.</td>
<td>Severity Level (GT): 3/3. The hazelnut has a large crack, indicating a high economic impact.</td>
</tr>
<tr>
<td>Output from MMAD-4o</td>
<td colspan="2">Anomaly Score: 71.
Reason: The inference image shows a significant hole in the hazelnut, which is a severe defect. This type of damage likely renders the nut unusable or costly to repair, indicating a high economic impact.</td>
</tr>
<tr>
<td>Output from MMAD-Sonnet</td>
<td colspan="2">Anomaly Score: 85.
Reason: The inference image shows a significant defect in the form of a hole or perforation in the hazelnut shell, along with what appears to be some discoloration around the damaged area. This type of damage likely indicates pest infestation or severe physical damage, which makes the hazelnut unsuitable for commercial sale or consumption. Such defects have a high economic impact as the product must be discarded due to food safety concerns and quality standards. Unlike minor surface blemishes, this type of structural damage cannot be repaired or salvaged, hence the high anomaly score.</td>
</tr>
</table>

Table 2. The output of MLLM-based baselines on a sample in MVTec-MAD dataset.

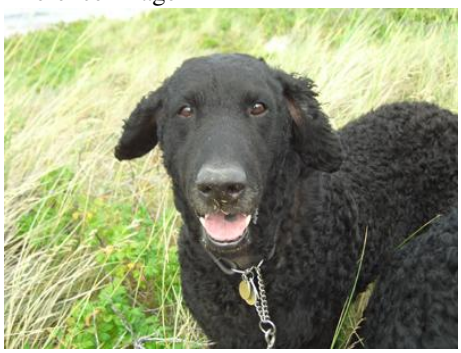
<table>
<tr>
<td>Normal Image
 
Inference Image
  </td>
<td>Subset Name: Golden Retriever in MultiDogs-MAD (Normal images). The inference image depicts a dog of a different breed than the Golden Retriever.</td>
<td>Severity Level (GT): 1/4. The dog is of a different breed than the Golden Retriever but is still a dog.</td>
</tr>
<tr>
<td>Output from MMAD-4o</td>
<td colspan="2">Anomaly Score: 51.
Reason: The inference image depicts a dog that is not a golden retriever but is still a dog, indicating some level of relationship to the normal class. This results in a moderate anomaly score reflecting the similarity.</td>
</tr>
<tr>
<td>Output from MMAD-Sonnet</td>
<td colspan="2">Anomaly Score: 35.
Reason: The inference image shows a Curly-Coated Retriever, which is different from but related to Golden Retrievers shown in the normal images. While it’s not the same breed, it belongs to the same retriever family of dogs, sharing similar characteristics like body structure, general size, and purpose (retrieving dogs). The main differences are the black color and curly coat texture, as opposed to the golden straight coat of Golden Retrievers. The score of 35 reflects that while this is a different breed, it’s still closely related to the normal class, being from the same retriever family, showing similar facial features, body structure, and behavioral characteristics typical of retriever breeds.</td>
</tr>
</table>

Table 3. The output of MLLM-based baselines on a sample in MultiDogs-MAD dataset.

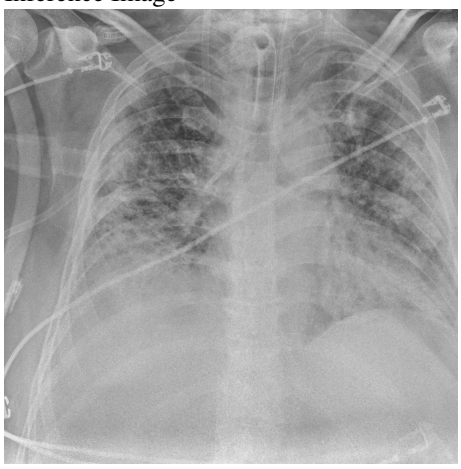
<table>
<tr>
<td>Normal Image
 
Inference Image
  </td>
<td>Dataset Name: Covid19-MAD</td>
<td>Severity Level (GT): 6/6. The inference image shows extreme lung damage, affecting more than 85% of the lung.</td>
</tr>
<tr>
<td>Output from MMAD-4o</td>
<td colspan="2">Anomaly Score: 75.
Reason: The inference image shows extensive consolidation in multiple lung regions, indicating severe lung damage. This is more pronounced compared to the normal lung images provided.</td>
</tr>
<tr>
<td>Output from MMAD-Sonnet</td>
<td colspan="2">Anomaly Score: 75.
Reason: The inference image shows significant abnormalities compared to the normal lung images. There are extensive bilateral ground-glass opacities and consolidation patterns throughout both lung fields, particularly in the mid and lower zones. The lung fields appear much hazier and whiter compared to the clear, dark lung fields seen in the normal images. The presence of medical equipment (tubeslines) also suggests a severe clinical condition. The extent and distribution of the abnormalities affect approximately 75-80% of the visible lung fields, placing this case in the severe category (71-85 range) of lung damage. While severe, it's not in the extreme category as there are still some areas of relatively preserved lung tissue visible.</td>
</tr>
</table>

Table 4. The output of MLLM-based baselines on a sample in Covid19-MAD dataset.

Level	Capsules	Chewinggum	Fryum	Macaroni1	Macaroni2	PCB1	PCB2	PCB3	Pipe_Fryum
Level 0 (training)	542	453	450	900	900	904	901	905	450
Level 0 (test)	60	50	50	100	100	100	100	101	50
Level 1	45	40	6	16	16	14	14	20	39
Level 2	20	8	35	12	9	66	67	60	15
Level 3	20	23	59	35	35	20	19	20	25
Total	687	574	600	1063	1060	1104	1101	1106	579

Table 5. Number of samples of the VisA-MAD dataset across levels.

Level	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal_nut	Pill	Screw	Transistor	Zipper
Level 0 (training)	280	264	245	230	247	209	224	219	391	220	267	320	213	240
Level 0 (test)	28	21	32	33	19	20	58	23	40	22	26	41	60	32
Level 1	17	33	38	36	18	21	23	45	17	22	43	25	10	33
Level 2	19	12	17	31	31	22	22	20	17	23	26	24	10	70
Level 3	34	12	37	17	11	20	11	44	36	25	17	24	20	16
Total	378	342	369	347	326	292	338	351	501	312	379	434	313	391

Table 6. Number of samples of MVTec-MAD dataset across levels.

Level	Bichon Frise	Chinese Rural Dog	Golden Retriever	Labrador Retriever	Teddy
Level 0 (training)	500	500	500	500	500
Level 0 (test)	500	500	500	500	500
Level 1	500	500	500	500	500
Level 2	500	500	500	500	500
Level 3	500	500	500	500	500
Level 4	500	500	500	500	500
Total	2500	2500	2500	2500	2500

Table 7. Number of samples of MultiDogs-MAD datasets across levels.

Level	Covid-MAD	DRD-MAD	SkinLesion-MAD
Leve 0 (training)	703	1000	500
Level 0 (test)	81	700	500
Level 1	96	700	642
Level 2	60	700	327
Level 3	62	700	500
Level 4	90	700	None
Level 5	135	None	None
Level 6	137	None	None
Total	1364	4500	2469

Table 8. Number of samples of three medical datasets across levels.

Method	Binary AD performance				MAD performance		Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C	Level 1	Level 2	Level 3	Whole	Ken	C
	capsules						**chewinggum**					
Skip-GAN [1]	63.11	79.58	71.08	68.86	0.286	67.12	82.55	92.00	96.00	87.97	0.560	83.84
RD4AD [5]	85.26	98.33	98.50	91.45	0.701	91.93	95.30	100.0	100.0	97.35	0.736	94.45
PatchCore [10]	67.26	85.42	94.75	78.00	0.479	78.66	96.65	100.0	100.0	98.11	0.727	93.93
CFLOW-AD [6]	61.41	94.00	87.58	75.24	0.416	74.89	98.00	100.0	100.0	98.87	0.714	93.14
RRD [12]	86.37	99.58	99.50	92.57	0.688	91.12	97.50	100.0	100.0	98.59	0.745	94.99
OCR-GAN [9]	89.48	84.50	80.08	86.10	0.310	68.51	91.80	92.75	94.96	92.93	0.572	84.56
PNI [2]	82.19	94.92	98.58	89.04	0.617	86.88	98.00	100.0	100.0	98.87	0.728	93.97
SPR [11]	67.19	85.08	74.42	73.10	0.311	68.60	35.95	31.75	30.26	33.63	0.213	37.13
IGD [4]	56.22	66.08	68.17	61.35	0.183	60.95	88.90	100.0	98.87	93.38	0.626	87.82
AE4AD [3]	70.83	80.75	67.26	71.27	0.291	67.38	76.09	65.75	68.55	70.68	0.264	65.96
MMAD-4o	70.00	85.00	100.0	80.59	0.726	82.23	93.75	100.0	100.0	96.48	0.832	93.32
MMAD-4o-mini	80.00	80.00	97.50	84.12	0.665	80.40	92.75	96.25	99.13	95.21	0.756	89.79
MMAD-Sonnet	52.22	75.00	82.50	64.71	0.537	67.75	87.32	97.75	99.61	92.48	0.756	89.44
MMAD-Haiku	50.00	50.00	50.00	50.00	nan	50.00	83.75	100.0	100.0	90.85	0.748	85.89
	fryum						**macaroni1**					
Skip-GAN [1]	49.00	92.91	100.0	94.46	0.680	91.15	96.06	90.00	100.0	97.10	0.682	95.34
RD4AD [5]	73.33	90.51	99.97	95.06	0.746	95.15	97.38	97.33	97.63	97.51	0.664	94.11
PatchCore [10]	78.67	92.63	100.0	96.14	0.758	95.86	97.19	98.58	96.66	97.16	0.628	91.74
CFLOW-AD [6]	60.67	83.77	96.98	90.18	0.632	88.27	87.38	91.42	77.14	82.46	0.397	76.38
RRD [12]	77.33	86.29	99.86	93.76	0.731	94.21	91.31	94.75	94.23	93.59	0.597	89.69
OCR-GAN [9]	94.00	86.57	88.61	88.22	0.454	77.45	98.19	98.75	93.37	95.62	0.633	92.04
PNI [2]	85.00	97.89	100.0	98.36	0.759	95.92	97.56	99.42	96.46	97.30	0.605	90.22
SPR [11]	60.00	62.97	75.73	70.32	0.305	68.44	52.50	47.50	71.34	62.02	0.193	62.81
IGD [4]	82.33	77.49	90.81	85.64	0.516	81.24	80.50	78.25	65.43	71.70	0.241	66.03
AE4AD [3]	87.69	63.37	85.67	79.06	0.436	76.39	64.86	79.58	82.06	72.03	0.248	66.50
MMAD-4o	50.00	77.14	96.61	87.00	0.745	87.86	52.62	57.83	82.40	70.16	0.584	70.80
MMAD-4o-mini	54.33	75.26	94.68	85.46	0.688	85.12	49.62	63.17	73.31	65.37	0.423	65.87
MMAD-Sonnet	50.00	70.00	89.83	80.50	0.665	81.37	56.25	58.33	82.86	71.43	0.605	71.92
MMAD-Haiku	50.00	55.71	67.80	62.50	0.382	62.64	49.00	49.00	69.20	60.22	0.396	61.24
	macaroni2						**pcb1**					
Skip-GAN [1]	45.00	53.67	50.46	49.48	0.000	49.99	82.00	87.24	90.05	87.07	0.513	82.35
RD4AD [5]	91.81	80.56	87.40	87.55	0.479	82.21	97.71	97.02	94.20	96.55	0.604	88.08
PatchCore [10]	91.56	77.33	69.77	76.72	0.300	70.21	98.50	98.61	96.70	98.21	0.644	90.57
CFLOW-AD [6]	58.50	46.78	50.94	52.33	0.021	51.39	94.86	93.70	99.05	94.93	0.646	90.68
RRD [12]	92.56	71.67	85.69	85.42	0.447	80.08	94.29	95.50	94.70	95.17	0.612	88.54
OCR-GAN [9]	94.00	90.56	93.00	92.90	0.545	86.66	93.93	93.56	93.65	93.63	0.569	85.87
PNI [2]	98.19	79.56	77.26	83.18	0.364	74.51	99.79	99.61	98.90	99.49	0.656	91.36
SPR [11]	55.56	58.33	53.80	54.95	0.059	53.97	30.14	54.29	41.60	48.37	0.014	49.14
IGD [4]	66.75	70.00	62.46	64.73	0.170	61.43	89.71	90.20	98.85	91.86	0.618	88.92
AE4AD [3]	75.26	73.22	79.25	76.02	0.320	71.53	84.82	95.40	90.64	87.75	0.522	82.87
MMAD-4o	53.12	50.00	61.43	57.50	0.342	57.52	58.82	81.01	71.40	75.98	0.536	73.38
MMAD-4o-mini	38.50	60.61	60.13	54.43	0.108	55.93	60.86	75.00	70.10	72.04	0.448	69.26
MMAD-Sonnet	51.28	47.78	60.00	55.84	0.120	56.20	53.36	76.83	90.95	76.36	0.585	76.58
MMAD-Haiku	48.00	53.56	50.86	50.50	0.033	50.65	49.50	51.03	57.07	52.02	0.166	52.54

Table 9. Full results on VisA-MAD (Part I)

Method	Binary AD performance				MAD performance		Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C	Level 1	Level 2	Level 3	Whole	Ken	C
	pcb2						pcb3					
Skip-GAN [1]	75.57	99.37	100.0	96.16	0.703	94.41	71.68	93.71	94.50	89.47	0.576	85.96
RD4AD [5]	98.93	97.67	90.00	96.39	0.555	85.04	96.68	98.00	90.10	96.16	0.570	85.57
PatchCore [10]	100.0	98.46	91.84	97.42	0.578	86.48	97.82	99.08	97.08	98.43	0.599	87.38
CFLOW-AD [6]	92.43	89.75	86.47	89.50	0.516	82.57	85.20	84.85	85.94	85.14	0.458	78.61
RRD [12]	98.36	93.04	78.95	91.11	0.484	80.56	96.73	96.40	94.46	96.08	0.590	86.82
OCR-GAN [9]	95.71	96.85	98.79	97.06	0.666	92.07	82.97	85.59	88.76	85.70	0.504	81.48
PNI [2]	100.0	99.75	95.84	99.04	0.596	87.66	99.36	99.79	99.01	99.54	0.609	88.03
SPR [11]	48.29	58.40	47.58	54.93	0.058	53.64	51.78	63.20	68.66	62.01	0.184	61.48
IGD [4]	71.93	83.16	93.63	83.58	0.486	80.70	71.88	76.65	88.47	78.06	0.414	75.81
AE4AD [3]	96.66	99.53	90.29	96.31	0.683	93.10	74.26	90.54	71.73	77.01	0.407	75.39
MMAD-4o	9.89	75.96	71.84	75.72	0.490	71.94	67.50	78.33	62.50	73.00	0.475	68.62
MMAD-4o-mini	70.00	69.93	62.68	68.56	0.290	64.06	63.24	80.09	59.95	72.69	0.367	67.36
MMAD-Sonnet	77.71	81.66	73.16	79.49	0.556	74.74	67.50	72.50	67.50	70.50	0.450	66.78
MMAD-Haiku	62.89	65.05	51.21	62.12	0.257	58.36	51.98	56.20	49.50	54.02	0.148	52.87
	pipe_fryum						mean					
Skip-GAN [1]	63.69	73.07	88.00	73.16	0.380	72.49	69.85	84.62	87.79	82.64	0.487	80.29
RD4AD [5]	98.41	100.0	99.84	99.16	0.701	91.54	92.76	95.49	95.29	95.24	0.640	89.79
PatchCore [10]	99.38	100.0	100.0	99.70	0.708	91.91	91.89	94.46	94.09	93.32	0.602	87.42
CFLOW-AD [6]	95.38	100.0	97.92	97.06	0.640	87.92	81.54	87.14	86.89	85.08	0.493	80.43
RRD [12]	99.38	99.87	99.92	99.65	0.688	90.72	92.65	93.01	94.15	93.99	0.620	88.53
OCR-GAN [9]	80.97	97.60	94.32	88.35	0.483	78.61	91.23	91.86	91.73	91.17	0.526	83.03
PNI [2]	99.74	100.0	99.76	99.80	0.662	89.21	95.54	96.77	96.20	96.07	0.622	88.64
SPR [11]	49.49	46.53	37.44	45.11	0.097	44.23	50.10	56.45	55.65	56.05	0.087	55.49
IGD [4]	75.64	90.40	92.40	83.75	0.516	80.56	75.98	81.36	84.34	79.34	0.419	75.94
AE4AD [3]	62.40	63.60	45.79	54.58	0.115	56.79	76.99	79.08	75.69	76.08	0.365	72.88
MMAD-4o	96.15	98.47	99.56	97.67	0.857	94.03	61.32	78.19	82.86	79.34	0.621	77.74
MMAD-4o-mini	85.18	87.87	98.96	90.05	0.756	88.76	66.05	76.46	79.60	76.44	0.500	74.06
MMAD-Sonnet	92.00	96.60	98.36	94.89	0.775	88.97	65.29	75.16	82.75	76.24	0.561	74.86
MMAD-Haiku	69.23	93.33	70.00	74.05	0.479	69.69	57.15	63.76	62.85	61.81	0.326	60.43

Table 10. Full results on VisA-MAD (Part II)

Method	Binary AD performance				MAD performance		Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C	Level 1	Level 2	Level 3	Whole	Ken	C
	carpet						**grid**					
Skip-GAN [1]	29.20	47.18	70.90	54.34	0.243	64.13	72.87	90.08	54.37	72.60	0.150	58.88
RD4AD [5]	100.0	100.0	100.0	100.0	0.587	84.17	100.0	100.0	100.0	100.0	0.616	86.54
PatchCore [10]	97.06	100.0	100.0	99.29	0.663	88.59	96.54	100.0	100.0	97.99	0.659	89.08
CFLOW-AD [6]	100.0	93.61	100.0	98.27	0.446	75.93	90.91	94.84	98.81	93.40	0.482	78.62
RRD [12]	100.0	100.0	100.0	100.0	0.582	83.89	100.0	100.0	100.0	100.0	0.659	89.08
OCR-GAN [9]	94.12	100.0	100.0	98.57	0.286	66.67	97.40	100.0	100.0	98.50	0.550	82.61
PNI [2]	100.0	99.44	100.0	99.85	0.616	85.83	97.40	100.0	100.0	98.50	0.579	84.34
SPR [11]	83.40	100.0	95.06	93.57	0.443	75.79	100.0	100.0	100.0	100.0	0.342	70.28
IGD [4]	64.29	73.87	86.34	77.60	0.383	72.28	68.25	72.62	59.52	67.34	0.127	57.52
AE4AD [3]	5.46	78.20	60.50	51.94	0.171	59.94	66.81	82.54	69.84	70.76	0.268	65.87
MMAD-4o	100.0	100.0	100.0	100.0	0.811	92.10	98.48	100.0	100.0	99.12	0.683	86.80
MMAD-4o-mini	100.0	100.0	100.0	100.0	0.752	87.62	94.23	98.81	100.0	96.41	0.755	89.52
MMAD-Sonnet	100.0	100.0	98.53	99.29	0.744	86.93	96.97	95.83	100.0	97.37	0.728	86.97
MMAD-Haiku	91.18	89.47	94.12	92.14	0.579	78.81	80.30	83.33	100.0	85.09	0.552	75.55
	leather						**tile**					
Skip-GAN [1]	47.70	16.36	10.05	26.77	0.437	24.55	76.68	74.68	85.92	77.81	0.249	64.45
RD4AD [5]	100.0	100.0	100.0	100.0	0.473	77.55	100.0	98.44	100.0	99.42	0.422	74.52
PatchCore [10]	100.0	100.0	100.0	100.0	0.456	76.55	100.0	96.97	100.0	98.88	0.445	75.85
CFLOW-AD [6]	99.84	100.0	100.0	99.93	0.412	73.96	100.0	98.34	100.0	99.39	0.391	72.73
RRD [12]	100.0	100.0	100.0	100.0	0.536	81.21	100.0	99.22	100.0	99.71	0.440	75.57
OCR-GAN [9]	100.0	100.0	100.0	100.0	0.654	88.05	99.49	96.19	96.43	97.66	0.247	64.35
PNI [2]	99.42	100.0	100.0	99.76	0.464	77.01	100.0	99.61	100.0	99.86	0.461	76.77
SPR [11]	99.84	100.0	99.92	99.90	0.425	74.72	99.66	100.0	100.0	99.86	0.398	73.11
IGD [4]	88.49	97.24	87.58	89.74	0.363	71.11	100.0	93.45	100.0	97.58	0.469	77.26
AE4AD [3]	67.68	97.43	73.90	75.68	0.290	66.90	51.52	54.25	57.22	53.68	0.055	53.19
MMAD-4o	99.05	100.0	100.0	99.61	0.782	91.27	100.0	100.0	100.0	100.0	0.623	82.74
MMAD-4o-mini	98.89	100.0	100.0	99.54	0.750	86.30	100.0	87.10	100.0	95.24	0.483	75.17
MMAD-Sonnet	100.0	100.0	100.0	100.0	0.658	82.28	100.0	95.01	100.0	98.16	0.544	78.06
MMAD-Haiku	98.48	99.91	98.44	98.73	0.578	78.38	94.44	74.19	100.0	88.10	0.477	73.71
	wood						**bottle**					
Skip-GAN [1]	83.92	95.76	89.00	90.96	0.451	76.49	60.00	58.41	42.00	53.73	0.078	45.55
RD4AD [5]	100.0	98.47	100.0	99.21	0.548	82.16	99.76	100.0	100.0	99.92	0.602	84.55
PatchCore [10]	100.0	98.13	100.0	99.04	0.491	78.81	100.0	100.0	100.0	100.0	0.580	83.27
CFLOW-AD [6]	100.0	97.28	100.0	98.60	0.474	77.78	100.0	100.0	100.0	100.0	0.526	80.21
RRD [12]	100.0	98.64	100.0	99.30	0.535	81.40	100.0	100.0	100.0	100.0	0.567	82.57
OCR-GAN [9]	87.43	100.0	100.0	96.23	0.688	90.34	97.14	100.0	100.0	99.05	0.458	76.30
PNI [2]	100.0	98.47	100.0	99.21	0.438	75.68	100.0	100.0	100.0	100.0	0.609	84.93
SPR [11]	99.84	100.0	99.92	99.90	0.425	74.72	99.66	100.0	100.0	99.86	0.398	73.11
IGD [4]	100.0	95.93	100.0	97.89	0.461	77.02	98.81	99.77	100.0	99.52	0.453	75.99
AE4AD [3]	85.67	86.08	82.78	85.35	0.372	71.84	95.48	93.18	100.0	96.11	0.441	75.33
MMAD-4o	94.88	95.93	98.33	96.05	0.596	81.31	100.0	100.0	100.0	100.0	0.721	87.66
MMAD-4o-mini	95.18	97.45	99.28	97.11	0.628	83.57	98.57	100.0	100.0	99.52	0.731	89.78
MMAD-Sonnet	100.0	99.49	100.0	99.74	0.580	79.73	86.67	99.55	100.0	95.40	0.639	80.89
MMAD-Haiku	96.78	95.67	97.13	96.27	0.579	75.57	76.19	90.91	92.50	86.51	0.607	79.45
	cable						**capsule**					
Skip-GAN [1]	47.75	32.52	24.45	37.19	0.219	36.55	44.83	94.57	50.59	56.28	0.078	54.56
RD4AD [5]	95.73	99.14	99.22	97.75	0.568	84.98	95.85	100.0	98.32	97.61	0.543	81.89
PatchCore [10]	99.78	99.37	100.0	99.66	0.635	89.10	96.14	100.0	98.81	97.93	0.480	78.17
CFLOW-AD [6]	94.60	95.61	99.69	96.00	0.579	85.62	92.08	99.78	97.73	95.77	0.437	75.65
RRD [12]	99.55	100.0	100.0	99.82	0.591	86.37	98.26	100.0	99.21	98.96	0.545	82.00
OCR-GAN [9]	78.94	94.12	88.87	86.85	0.524	82.28	90.05	100.0	95.95	94.26	0.270	65.87
PNI [2]	98.35	98.90	100.0	98.89	0.676	91.60	99.42	100.0	99.31	99.48	0.360	71.17
SPR [11]	97.90	85.97	99.53	93.53	0.527	82.47	95.94	100.0	97.83	97.45	0.581	84.12
IGD [4]	90.93	93.18	98.28	93.26	0.572	85.22	80.19	99.78	86.46	86.32	0.347	70.40
AE4AD [3]	64.77	88.56	87.77	78.63	0.435	76.75	62.80	99.35	63.44	69.76	0.150	58.78
MMAD-4o	99.06	99.96	100.0	99.60	0.835	93.28	91.88	94.67	96.44	94.24	0.422	72.43
MMAD-4o-mini	93.85	94.59	94.20	94.21	0.647	84.93	88.02	99.57	95.16	93.02	0.492	74.01
MMAD-Sonnet	94.19	97.02	98.98	96.24	0.805	88.66	82.85	81.52	86.41	84.04	0.356	67.59
MMAD-Haiku	60.87	75.00	81.82	70.54	0.519	68.75	61.11	50.00	70.45	62.84	0.269	59.59

Table 11. Full results on MVTec-MAD (Part I)

Method	Binary AD performance				MAD performance		Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C	Level 1	Level 2	Level 3	Whole	Ken	C
	hazelnut						**metal_nut**					
Skip-GAN [1]	100.0	69.56	77.57	81.07	0.218	62.83	43.39	26.09	57.27	42.66	0.084	54.84
RD4AD [5]	100.0	100.0	100.0	100.0	0.560	82.98	100.0	100.0	100.0	100.0	0.469	76.92
PatchCore [10]	100.0	100.0	100.0	100.0	0.504	79.71	100.0	99.60	100.0	99.87	0.455	76.13
CFLOW-AD [6]	100.0	100.0	100.0	100.0	0.611	86.02	98.76	90.91	99.64	96.49	0.430	74.68
RRD [12]	100.0	100.0	100.0	100.0	0.563	83.19	100.0	100.0	100.0	100.0	0.504	78.93
OCR-GAN [9]	99.85	90.15	99.10	97.11	0.482	78.44	73.55	100.0	99.09	91.36	0.212	62.16
PNI [2]	100.0	100.0	100.0	100.0	0.448	76.42	100.0	100.0	100.0	100.0	0.431	74.74
SPR [11]	100.0	94.26	100.0	98.61	0.466	77.49	99.59	86.36	100.0	95.39	0.388	72.31
IGD [4]	93.38	95.29	98.40	96.43	0.636	87.48	79.55	94.86	75.45	83.12	0.222	62.72
AE4AD [3]	99.26	68.82	85.28	84.68	0.319	68.82	48.55	45.45	49.64	47.92	0.026	48.53
MMAD-4o	99.26	98.38	100.0	99.43	0.834	94.05	100.0	100.0	100.0	100.0	0.636	82.89
MMAD-4o-mini	85.07	84.19	98.58	91.80	0.713	88.72	89.15	99.01	91.73	93.31	0.636	82.73
MMAD-Sonnet	94.41	97.57	98.75	97.41	0.825	91.27	95.45	80.43	78.00	84.29	0.363	67.19
MMAD-Haiku	80.15	72.57	92.64	84.73	0.545	79.32	88.64	63.04	56.00	68.57	0.044	48.28
	pill						**screw**					
Skip-GAN [1]	66.55	80.33	96.83	76.70	0.512	80.01	65.85	0.00	0.00	22.55	0.496	21.20
RD4AD [5]	100.0	92.46	100.0	97.72	0.438	75.64	99.41	99.59	95.93	98.33	0.286	66.60
PatchCore [10]	98.03	90.24	99.55	95.97	0.483	78.29	99.90	99.59	98.98	99.50	0.318	68.48
CFLOW-AD [6]	94.19	85.36	96.83	92.04	0.457	76.78	97.07	97.15	95.02	96.42	0.439	75.51
RRD [12]	100.0	96.15	100.0	98.84	0.440	75.77	99.61	100.0	98.78	99.47	0.286	66.60
OCR-GAN [9]	97.50	100.0	98.19	98.39	0.185	60.86	84.68	50.61	74.29	70.06	0.122	57.10
PNI [2]	98.39	88.17	99.77	95.57	0.421	74.69	100.0	100.0	98.48	99.50	0.322	68.71
SPR [11]	92.49	89.94	100.0	93.20	0.495	78.97	98.73	98.98	98.88	98.86	0.420	74.40
IGD [4]	89.18	77.37	93.67	86.49	0.407	73.82	75.32	72.76	60.47	69.60	0.128	57.43
AE4AD [3]	67.62	81.07	96.61	77.42	0.476	77.91	27.90	40.96	61.38	43.20	0.055	53.20
MMAD-4o	97.67	84.62	100.0	94.19	0.498	75.68	96.00	77.08	81.25	84.93	0.480	72.44
MMAD-4o-mini	96.87	88.61	100.0	94.99	0.603	80.87	82.59	80.39	91.41	84.76	0.574	78.97
MMAD-Sonnet	75.22	59.69	87.44	72.94	0.307	63.63	72.39	65.75	74.44	70.88	0.323	65.36
MMAD-Haiku	60.87	50.00	59.73	57.36	0.054	51.63	54.00	52.08	52.08	52.74	0.080	51.22
	transistor						**zipper**					
Skip-GAN [1]	86.33	73.00	52.08	65.88	0.137	58.97	70.83	37.54	60.74	49.89	0.102	43.87
RD4AD [5]	88.17	98.67	99.50	96.46	0.549	85.83	98.30	98.26	99.61	98.45	0.392	73.65
PatchCore [10]	100.0	100.0	100.0	100.0	0.625	90.83	99.15	99.51	100.0	99.47	0.368	72.20
CFLOW-AD [6]	100.0	96.33	84.92	91.54	0.437	78.52	92.71	98.79	99.61	97.22	0.399	74.11
RRD [12]	95.83	99.67	99.17	98.46	0.570	87.24	97.73	98.66	99.80	98.56	0.434	76.19
OCR-GAN [9]	100.0	99.50	96.50	98.12	0.519	83.93	94.32	96.16	93.16	95.25	0.424	75.62
PNI [2]	100.0	100.0	100.0	100.0	0.573	87.41	99.62	99.87	100.0	99.82	0.421	75.41
SPR [11]	100.0	100.0	97.00	98.50	0.524	84.21	99.81	100.0	100.0	99.95	0.484	79.24
IGD [4]	100.0	88.83	86.67	90.54	0.426	77.83	87.03	92.81	98.05	91.91	0.383	73.11]
AE4AD [3]	100.0	85.50	72.17	82.46	0.317	70.72	81.34	82.37	90.04	83.11	0.297	67.91
MMAD-4o	72.92	72.75	96.33	84.58	0.636	80.84	98.48	87.86	96.88	92.02	0.485	75.79
MMAD-4o-mini	67.67	65.67	87.00	76.83	0.478	75.16	98.48	87.14	96.88	91.60	0.358	69.09
MMAD-Sonnet	93.83	82.75	87.83	88.06	0.621	79.69	84.85	69.29	90.62	76.47	0.303	64.35
MMAD-Haiku	61.00	50.00	71.25	63.38	0.336	62.90	69.70	52.14	62.50	58.40	0.010	49.70

Table 12. Full results on MVTec-MAD (Part II)

Method	Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C
	mean					
Skip-GAN [1]	63.99	56.86	55.13	57.75	0.057	53.35
RD4AD [5]	98.37	98.93	99.47	98.92	0.504	79.86
PatchCore [10]	99.04	98.82	99.81	99.11	0.511	80.36
CFLOW-AD [6]	97.15	96.29	98.02	96.79	0.466	77.58
RRD [12]	99.36	99.45	99.78	99.51	0.518	80.72
OCR-GAN [9]	92.46	94.77	95.83	94.39	0.402	73.90
PNI [2]	99.47	98.89	99.83	99.32	0.487	78.91
SPR [11]	97.63	96.82	99.15	97.76	0.451	76.78
IGD [4]	86.82	89.13	87.92	87.67	0.384	72.78
AE4AD [3]	66.06	77.41	75.04	71.48	0.259	65.41
MMAD-4o	96.26	93.66	97.80	95.98	0.646	83.52
MMAD-4o-mini	92.04	91.61	96.73	93.45	0.614	81.89
MMAD-Sonnet	91.20	87.42	92.93	90.02	0.557	77.33
MMAD-Haiku	76.69	71.31	80.62	76.10	0.366	66.63

Table 13. Full results on MVTec-MAD (Part III)

Method	Binary AD performance					MAD performance		Binary AD performance					MAD performance	
	Level 1	Level 2	Level 3	Level 4	Whole	Ken	C	Level 1	Level 2	Level 3	Level 4	Whole	Ken	C
	bichon_frise							**chinese_rural_dog**						
Skip-GAN [1]	92.80	98.22	85.30	99.95	94.07	0.538	80.08	79.79	89.59	85.74	99.64	88.69	0.549	80.70
RD4AD [5]	63.99	73.69	78.60	80.36	74.16	0.311	67.37	54.87	61.76	71.06	80.03	66.93	0.290	66.22
PatchCore [10]	66.82	81.22	79.02	87.93	78.74	0.408	72.78	53.73	74.47	74.92	88.68	72.95	0.396	72.15
CFLOW-AD [6]	93.38	97.57	97.26	95.72	95.98	0.366	70.46	61.95	86.39	89.37	91.08	82.20	0.438	74.49
RRD [12]	75.90	87.30	90.52	96.01	87.43	0.523	79.21	65.95	73.51	85.10	97.59	80.54	0.509	78.47
OCR-GAN [9]	78.46	72.60	69.00	96.58	79.16	0.383	71.40	73.92	75.54	63.97	94.01	76.86	0.339	68.94
PNI [2]	71.16	82.80	78.79	89.51	80.57	0.413	73.07	64.60	77.04	75.50	90.32	76.87	0.402	72.47
SPR [11]	61.83	55.24	79.30	71.25	66.90	0.238	63.30	59.69	54.78	80.94	72.90	67.08	0.249	63.92
IGD [4]	97.67	97.38	96.44	96.92	97.11	0.261	64.60	83.68	92.18	93.70	97.79	91.84	0.489	77.32
AE4AD [3]	56.96	61.07	34.82	49.49	50.59	0.073	45.94	54.29	61.06	33.86	54.16	50.84	0.037	47.93
MMAD-4o	98.00	100.0	100.0	100.0	99.50	0.938	95.89	80.56	99.97	100.0	100.0	95.13	0.908	94.90
MMAD-4o-mini	99.26	99.40	99.40	99.40	99.37	0.622	71.40	75.64	98.80	98.80	98.80	93.01	0.745	81.00
MMAD-Sonnet	99.36	99.99	100.0	100.0	99.84	0.962	98.29	74.53	97.28	100.0	100.0	92.95	0.911	96.34
MMAD-Haiku	94.64	99.77	99.88	98.14	98.11	0.790	89.58	66.04	93.43	97.52	93.34	87.58	0.738	87.15
	golden_retriever							**labrador_retriever**						
Skip-GAN [1]	77.63	88.52	84.17	99.35	87.42	0.526	79.41	66.81	72.34	84.30	98.56	80.50	0.529	79.54
RD4AD [5]	48.12	63.34	72.42	80.57	66.11	0.331	68.50	44.67	59.42	67.04	74.60	61.43	0.281	65.69
PatchCore [10]	58.26	79.66	78.19	89.99	76.53	0.429	73.95	55.54	78.62	75.18	88.30	74.41	0.409	72.84
CFLOW-AD [6]	77.03	94.75	95.91	95.43	90.78	0.491	77.44	69.52	92.76	92.40	93.10	86.94	0.468	76.15
RRD [12]	69.13	82.34	89.01	97.19	84.42	0.538	80.05	55.54	71.06	80.96	93.14	75.17	0.472	76.40
OCR-GAN [9]	69.46	72.48	61.81	93.75	74.38	0.342	69.12	69.33	69.22	55.51	90.52	71.14	0.278	65.54
PNI [2]	60.61	78.41	75.63	89.51	76.04	0.429	73.99	58.47	79.05	73.92	89.60	75.26	0.414	73.15
SPR [11]	54.94	50.44	77.32	78.32	65.26	0.297	66.60	48.21	45.86	68.97	70.89	58.48	0.241	63.44
IGD [4]	89.61	95.63	95.35	97.30	94.48	0.468	76.13	90.53	96.75	96.64	97.20	95.28	0.436	74.38
AE4AD [3]	55.09	60.99	33.96	51.96	50.50	0.052	47.09	47.22	54.46	29.29	45.52	44.12	0.093	44.83
MMAD-4o	98.80	100.0	100.0	100.0	99.70	0.949	96.62	96.56	100.0	100.0	100.0	99.14	0.936	95.43
MMAD-4o-mini	99.22	100.0	100.0	100.0	99.80	0.781	82.73	98.26	100.0	100.0	100.0	99.56	0.786	83.51
MMAD-Sonnet	97.99	99.89	99.89	100.0	99.44	0.952	98.17	92.06	99.82	100.0	100.0	97.97	0.957	98.74
MMAD-Haiku	90.93	98.35	99.68	99.17	97.03	0.790	90.06	78.53	94.48	97.54	96.36	91.73	0.777	89.71
	teddy							**mean**						
Skip-GAN [1]	76.06	79.47	90.15	99.03	86.18	0.564	81.53	78.62	85.63	85.93	99.31	87.37	0.541	80.25
RD4AD [5]	54.06	66.10	74.43	79.77	68.59	0.326	68.23	53.14	64.86	72.71	79.07	67.44	0.308	67.20
PatchCore [10]	58.63	78.79	77.25	88.94	75.90	0.411	72.96	58.60	78.55	76.91	88.77	75.71	0.410	72.94
CFLOW-AD [6]	82.99	94.74	94.88	92.81	91.35	0.394	72.00	76.97	93.24	93.96	93.63	89.45	0.431	74.11
RRD [12]	64.18	77.07	86.42	95.41	80.77	0.509	78.43	66.14	78.26	86.40	95.87	81.67	0.510	78.51
OCR-GAN [9]	74.02	69.05	68.37	94.94	76.60	0.375	70.97	73.04	71.78	63.73	93.96	75.63	0.343	69.19
PNI [2]	63.90	79.67	75.30	90.08	77.24	0.409	72.87	63.75	79.39	75.83	89.80	77.20	0.413	73.11
SPR [11]	59.74	51.80	78.34	65.12	63.75	0.188	60.48	56.88	51.62	76.97	71.70	64.29	0.242	63.55
IGD [4]	92.81	97.26	97.08	98.44	96.40	0.464	75.92	90.86	95.84	95.84	97.53	95.02	0.424	73.67
AE4AD [3]	48.71	57.79	31.07	49.07	46.66	0.065	46.38	52.45	59.07	32.60	50.04	48.54	0.064	46.43
MMAD-4o	95.08	99.90	99.99	100.0	98.74	0.934	96.72	93.80	99.97	100.0	100.0	98.44	0.933	95.91
MMAD-4o-mini	98.35	100.0	100.0	100.0	99.59	0.780	83.12	94.15	99.64	99.64	99.64	98.27	0.743	80.35
MMAD-Sonnet	97.28	99.85	99.95	100.0	99.27	0.899	95.15	92.24	99.37	99.97	100.0	97.89	0.936	97.34
MMAD-Haiku	75.13	94.22	97.88	95.59	90.70	0.739	86.27	81.05	96.05	98.50	96.52	93.03	0.766	88.55

Table 14. Full results on MultiDogs-MAD

Method	Binary AD performance					MAD performance	
	Level 1	Level 2	Level 3	Level 4	Whole	Ken	C
Skip-GAN [1]	52.50	49.33	51.03	52.61	51.36	0.014	50.77
RD4AD [5]	51.96	55.12	66.17	71.94	61.30	0.217	62.14
PatchCore [10]	50.05	54.15	63.27	77.95	61.36	0.259	64.47
CFLOW-AD [6]	50.14	52.10	61.23	77.37	60.21	0.238	63.30
RRD [12]	50.86	55.84	65.77	74.84	61.83	0.243	63.55
OCR-GAN [9]	52.51	55.20	47.89	59.72	53.83	0.054	53.03
PNI [2]	48.09	55.60	66.67	79.66	62.50	0.285	65.94
SPR [11]	47.09	47.15	44.03	47.02	46.32	0.034	48.11
IGD [4]	45.65	48.57	52.24	70.87	54.33	0.170	59.51
AE4AD [3]	48.56	48.17	46.49	55.42	49.66	0.029	51.63
MMAD-4o	49.83	59.86	75.54	77.98	65.80	0.433	67.72
MMAD-4o-mini	49.98	56.50	66.87	80.86	63.55	0.348	66.66
MMAD-Sonnet	47.39	58.13	70.85	83.70	65.02	0.403	69.30
MMAD-Haiku	47.66	48.95	55.20	61.74	53.39	0.125	56.20

Table 15. Full results on DRD-MAD

Method	Binary AD performance							MAD performance	
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Whole	Ken	C
Skip-GAN [1]	68.58	64.48	83.27	91.49	55.52	61.28	68.50	0.006	50.34
RD4AD [5]	79.79	74.50	85.02	81.76	86.21	87.70	83.48	0.219	61.90
PatchCore [10]	72.21	71.98	74.52	74.56	77.75	81.67	76.33	0.200	60.88
CFLOW-AD [6]	68.97	66.65	75.52	72.39	76.27	78.99	74.03	0.180	59.78
RRD [12]	79.35	79.00	80.85	84.08	86.74	89.60	84.36	0.240	63.07
OCR-GAN [9]	78.63	41.79	79.62	63.28	60.57	66.43	65.46	0.038	52.09
PNI [2]	85.23	83.19	87.02	84.81	90.29	92.12	87.96	0.241	63.12
SPR [11]	54.64	41.92	56.87	55.69	59.63	55.69	55.14	0.065	53.51
IGD [4]	78.75	82.62	80.87	82.65	89.20	93.11	85.82	0.312	66.99
AE4AD [3]	68.85	68.75	78.15	71.24	74.75	80.97	74.45	0.189	60.26
MMAD-4o	75.05	80.81	82.28	88.67	93.66	97.02	88.07	0.547	76.94
MMAD-4o-mini	66.17	80.05	75.10	81.69	82.63	86.33	79.58	0.351	66.98
MMAD-Sonnet	77.46	86.60	90.58	95.78	97.11	99.10	92.35	0.601	80.54
MMAD-Haiku	56.65	57.01	59.54	59.60	60.44	61.21	59.42	0.142	54.16

Table 16. Full results on Covid19-MAD

Method	Binary AD performance				MAD performance	
	Level 1	Level 2	Level 3	Whole	Ken	C
Skip-GAN [1]	99.59	99.94	99.40	99.60	0.406	73.65
RD4AD [5]	98.71	98.91	99.24	98.94	0.509	79.60
PatchCore [10]	100.0	99.99	100.0	100.0	0.456	76.57
CFLOW-AD [6]	99.99	99.94	99.99	99.98	0.389	72.63
RRD [12]	99.25	99.57	99.85	99.53	0.558	82.48
OCR-GAN [9]	99.61	99.32	99.58	99.53	0.391	72.76
PNI [2]	100.0	100.0	100.0	100.0	0.455	76.51
SPR [11]	90.57	93.59	90.77	91.31	0.335	69.50
IGD [4]	97.64	96.71	99.03	97.91	0.487	78.34
AE4AD [3]	99.02	98.37	99.79	99.13	0.505	79.42
MMAD-4o	99.49	98.75	99.80	99.43	0.694	85.61
MMAD-4o-mini	99.52	99.90	99.95	99.75	0.653	84.97
MMAD-Sonnet	99.79	99.80	99.87	99.82	0.610	79.23
MMAD-Haiku	99.99	99.67	99.78	99.85	0.479	74.19

Table 17. Full results on SkinLesion-MAD

References

[1] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 13, 14, 15, 16, 17, 18, 19

[2] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni: Industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023. 3, 13, 14, 15, 16, 17, 18, 19

[3] Yu Cai, Hao Chen, and Kwang-Ting Cheng. Rethinking autoencoders for medical anomaly detection from a theoretical perspective. *arXiv preprint arXiv:2403.09303*, 2024. 13, 14, 15, 16, 17, 18, 19

[4] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 383–392, 2022. 13, 14, 15, 16, 17, 18, 19

[5] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 13, 14, 15, 16, 17, 18, 19

[6] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022. 13, 14, 15, 16, 17, 18, 19

[7] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 1

[8] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. 1

[9] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*, 2023. 13, 14, 15, 16, 17, 18, 19

[10] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022. 13, 14, 15, 16, 17, 18, 19

[11] Woosang Shin, Jonghyeon Lee, Taehan Lee, Sangmoon Lee, and Jong Pil Yun. Anomaly detection using score-based perturbation resilience. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23372–23382, 2023. 13, 14, 15, 16, 17, 18, 19

[12] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24511–24520, 2023. 3, 13, 14, 15, 16, 17, 18, 19

[13] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011. 1