

Appendix for AUPIMO: Redefining Anomaly Localization Benchmarks with High Speed and Low Tolerance

Joao P. C. Bertoldo¹
jpcb Bertoldo@minesparis.psl.eu

Dick Ameln²
dick.ameln@intel.com

Ashwin Vaidya²
ashwin.vaidya@intel.com

Samet Akçay²
samet.akçay@intel.com

¹ Mines Paris, PSL University,
Centre for mathematical
morphology (CMM),
77300 Fontainebleau, France

² Intel

A False positives on normal images

We argue that, in Anomaly Detection (AD), the negative class (normal) is the only well-defined class, and that False Positive Rate (FPR) is a meaningful metric to validate models. The positive class (anomalous) is *not* a well-defined concept because it covers the entire complement of the normal class. As such, it is impossible to cover all types and variations. Based on this principle, we argue that it is problematic to use anomalous samples for model validation (not to be confused with model *evaluation*). For this reason, we propose the validation to depend solely on normal instances, thus based on False Positives (FPs).

For the sake of complementing the discussion, we present an alternative to the (pixel-wise, image-scoped) FPR used in AUPIMO. Counting the number of regions falsely detected as anomalous can be used as meaningful metric to validate (*i.e.* constrain) models. However, such metric is not used in AUPIMO because it is inconvenient to compute, so we propose the FPR as a proxy. Finally, we present visual examples of FP masks at different levels of FPR to provide an intuition of what it represents in practice.

A.1 Rate vs. number of regions

In this section the relation between two (pixel-wise, image-scoped) metrics is analyzed (both measured on normal images at different binarization thresholds of anomaly score maps):

1. False Positive Rate (FPR): the ratio between the number of FP pixels and the total number of pixels;
2. Number of False Positive Regions (NumFPReg): the number of maximally connected FP regions.

To be trusted in real-world applications, an anomaly localization model is expected to find image structures worth the user’s attention. Raising false detections eventually diminishes the user’s interest, so it should happen as rarely as possible. One could assume, for instance, that users eventually investigate detected anomalies manually – or even programatically. From this perspective, we argue that the Number of False Positive Regions (NumFPRReg) is an informative metric in practice because it directly relates to how often a user would investigate FPs, so it is a good estimator of the human cost for using the model (*i.e.* how often one’s time is wasted). A good estimate of the expected NumFPRReg would allow a user to set a threshold based on its operational cost.

However, computing NumFPRReg requires connected component analysis, which has two major drawbacks. First, it is slow to compute, especially on the CPU. Second, some implementations use an iterative process that may not converge in some cases. For instance, the implementation in `kornia` [14] (see `kornia.contrib.connected_components`). The FPR, on the other hand, is fast to compute and, as we show next, can be used as a proxy for the NumFPRReg at low FP levels.

Experiment Anomaly score maps from our experiments were randomly sampled from the set of normal images, upsampled with bilinear interpolation to the same resolution as the original annotation masks, binarized with a series of thresholds, and the NumFPRReg and the FPR were computed for each binary mask. All models and datasets were confounded on purpose because we seek to understand the relationship between FPR and NumFPRReg *in general*, not for a specific model or dataset. Thresholds were chosen such that a series of logarithmically-spaced FPR levels from 10^{-5} to 10^{-3} are covered. A random multiplying factor $\in [0.9, 1.1]$ was added to each target FPR value in this range (like a jitter). Assumptions:

1. Each threshold is interpreted as an operational threshold set to automatically obtain binary masks from an Anomaly Detection (AD) model;
2. Both metrics are computed at the image scope (*i.e.* ratio of pixels and number of regions in each image);
3. In an real-life scenario, the expected values of these metrics would be estimated to describe a model’s behavior to control its operational cost.

Fig. 1a shows a scatter plot of FPR (X-axis, in logarithmic scale) vs. NumFPRReg (Y-axis). NumFPRReg was clipped to the maximum value of 5 and jitter was added to avoid overlapping points. A mean line is displayed in black. The Y-axis values of the mean line are computed as the average NumFPRReg in the bins centered around the pre-set FPR levels.

Fig. 1b shows histograms (counts are numbers of images) of NumFPRReg at three FPR levels: 10^{-5} , 10^{-4} , and 10^{-3} . At each level L , all the points from the scatter in the range $[L/2, 2L]$ are accounted to have a sufficient number of samples. The histograms are normalized to sum to 1. The dashed lines show the sum of the bars’ values from left to right.

Results Fig. 1 shows that the FPR can effectively be used as a proxy for the number of FP regions:

1. FPR and NumFPRReg correlate positively;
2. The majority of images have ≤ 2 regions (more than 90% at FPR 10^{-4} and nearly 100% at FPR 10^{-5});

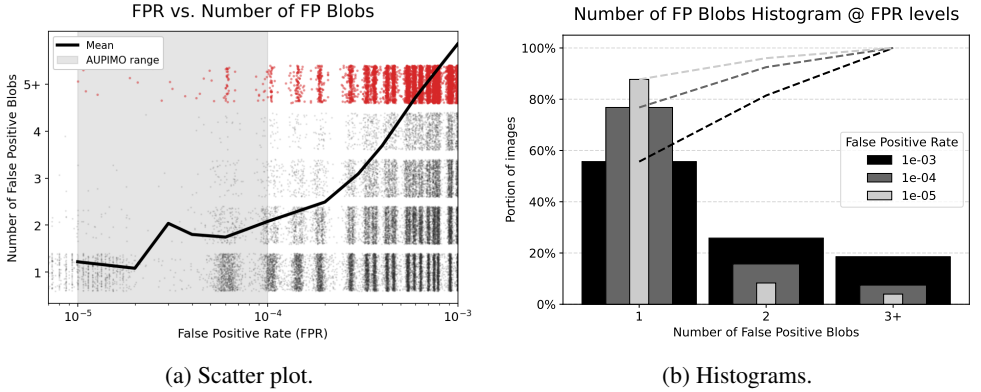


Figure 1: False Positivity. How Image False Positive Rate (ImFPR) relates to the Number of False Positive Regions (NumFPRreg).

3. Inside AUPIMO’s integration range, the average NumFPRreg tends to 1, so the FPR generally equals the relative size of the single FP region in the mask.

In summary, at AUPIMO’s default integration range, the FPR tends to translate to the maximum relative size of FP regions in normal images because they tend to have a single FP region.

As a practical implication, AUPIMO’s bounds can be leveraged to filter out model predictions. For instance, one can ignore detected regions with areas smaller than AUPIMO’s lower bound. Notice in that MVTec Anomaly Detection (MVTec AD)’s datasets do not have anomalies with relative size smaller than 10^{-5} , and very few as small as 10^{-4} (see Appendix B).

A.2 Visual intuition of FP levels

We intend to build an intuition of what Image FPR (F_i) levels visually represent on normal images. The Image FPR on normal images is the relative area covered by an FP mask. As shown in the previous section, with AUPIMO’s low levels of FPR, it further tends to translate to the size of a single FP region.

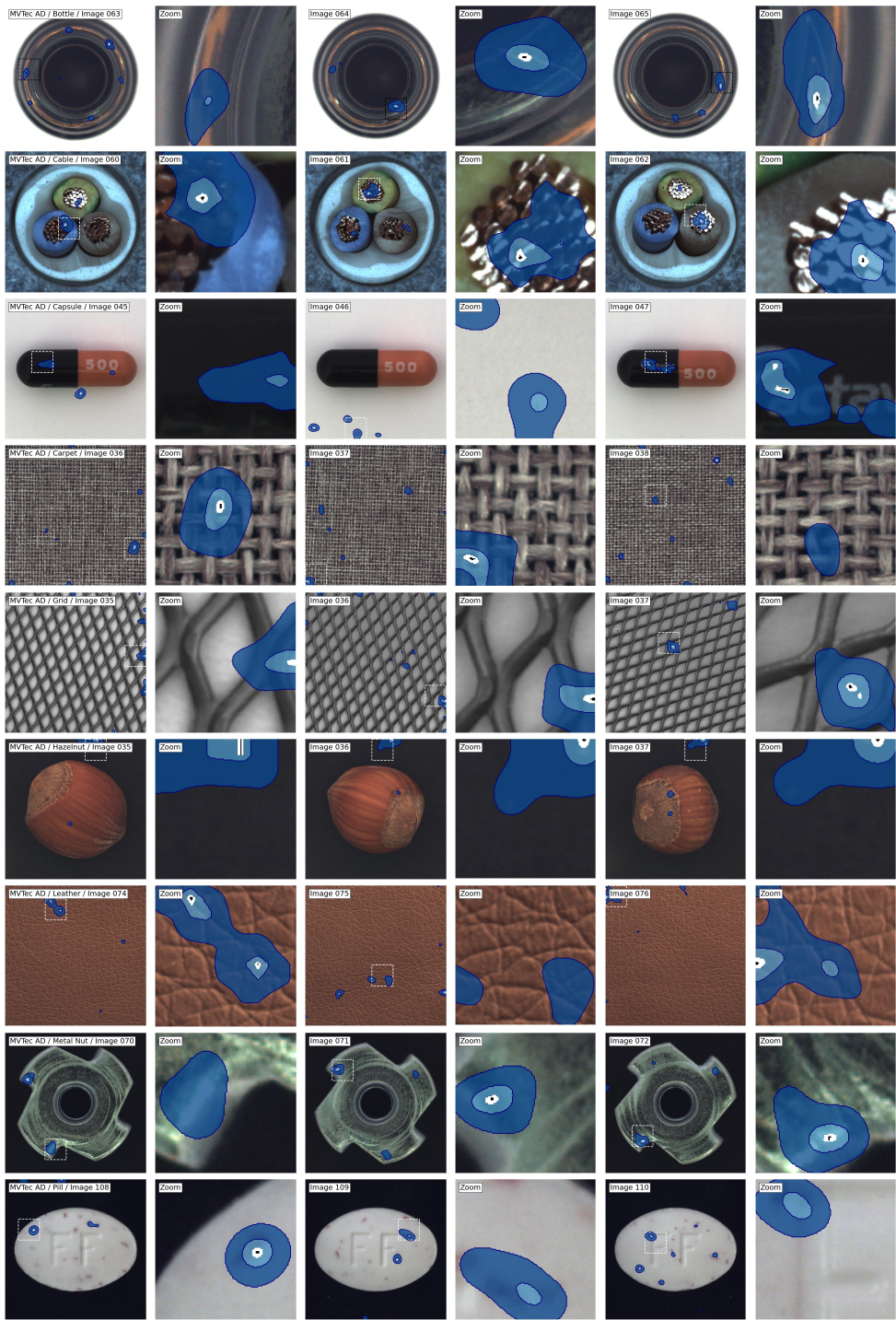
Fig. 2 shows examples of normal images from all the datasets in MVTec AD and Visual Anomaly (VisA) superposed by FP masks. Each dataset is in a row with three samples from the test set. Each image is presented with a zoom on the right (the zoomed area is highlighted in the original image with a dashed rectangle). Each color corresponds to a predicted mask at a given ImFPR level. Color code:

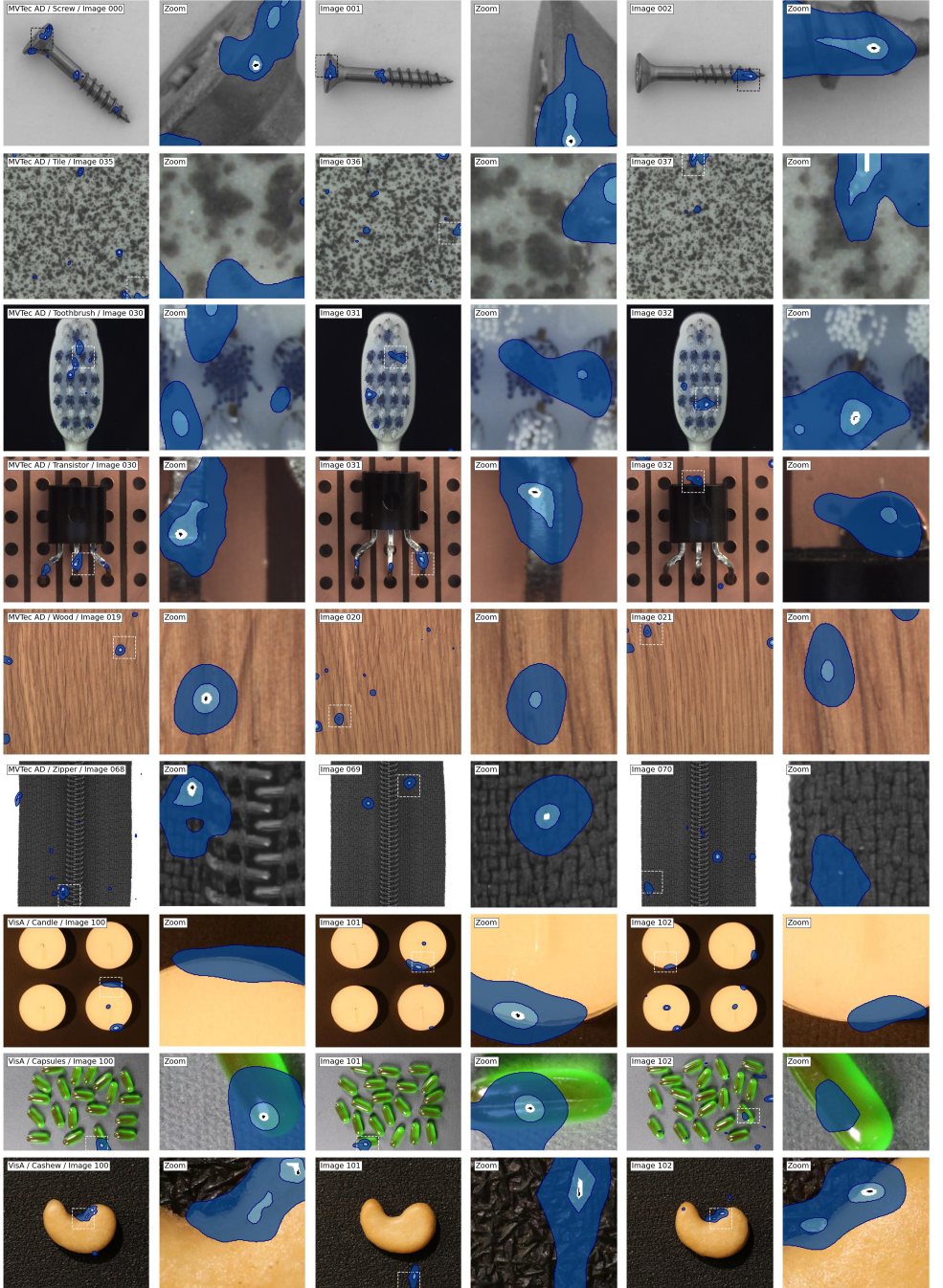
1. Darker blue is $F_i = 10^{-2}$;
2. Lighter blue is $F_i = 10^{-3}$;
3. White is $F_i = 10^{-4}$;
4. Black is $F_i = 10^{-5}$.

The masks are generated from the anomaly score maps produced by a randomly picked model from our benchmark. The different masks in a single image are generated from the same anomaly score map (*i.e.* same model), but different samples may have masks from different models.

Inside AUPIMO’s integration bounds ($10^{-5} \sim 10^{-4}$, *i.e.* between black and white in Fig. 2), FP regions become barely visible at the image scale and generally irrelevant compared to the objects’ structures.

Disclaimer: the *Shared* FPR used in Per-Image Overlap (PIMO) is the *average* Image FPR across all normal images, so it is not to be confused with the Image FPR of a single image. This visual intuition should be understood as an average behavior, not as a strict rule.





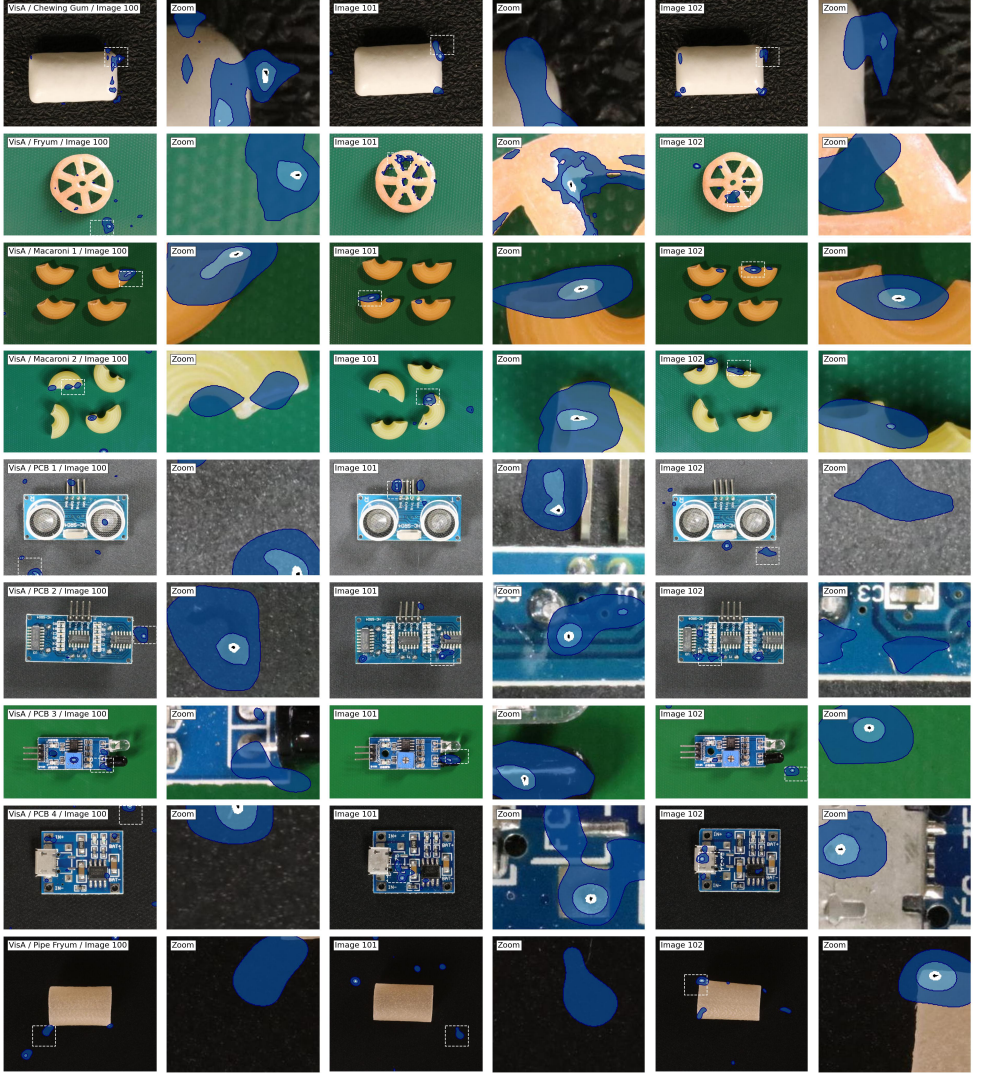


Figure 2: Visual intuition of Image False Positive Rate (ImFPR) levels on normal images. Images are normal samples from the datasets in MVTECAD and VisA. Each color corresponds to a predicted mask at a given ImFPR level: darker blue is 10^{-2} , lighter blue is 10^{-3} , white is 10^{-4} , and black is 10^{-5} .

B Anomaly size

Fig. 3 shows the distributions of the relative region size in ground truth annotations in each dataset from MVTec AD and VisA. Reminder: relative size is the number of pixels in a maximally connected component divided by the number of pixels in the image. Lower and upper whiskers are set with maximum size to 1.5 inter-quartile range (IQR), and fliers (outliers) are shown as gray dots. The gray-shaded span is AUPIMO’s integration range, and the vertical gray line represents the relative size of a single pixel at resolution 256×256 (input size seen by the models in our experiments).

MVTec AD Fig. 3 shows that the size of the anomalies in MVTec AD are generally between 10^{-3} and 10^{-1} . Few cases are as small 10^{-4} . Given this distribution, the AUPIMO scores from our experiments can be interpreted as a (near) FP-free recall. Since (almost) none of the anomalies are as small as the FPR integration range, any prediction with relative size above the integration range is a True Positive (TP). Conversely, one could dismiss any prediction with relative size below the integration range.

VisA The anomalies in VisA are largely biased towards small regions of relative sizes as small as $\sim 10^{-6}$ (*i.e.* a single pixel at resolution 1000×1000). They are so numerous that the actual anomalous regions show as outliers in Fig. 3.

Tiny regions Let “tiny” refer to connected components of relative size smaller than $\frac{1}{256^2}$, which corresponds to a single pixel at resolution 256×256 . In other words, an actual anomaly this small would be seen as a single pixel by the models in our experiments or simply not seen at all. Fig. 4 displays several examples of tiny regions in VisA with zoomed-in views on the right. These regions are meaningless: as Fig. 4 shows, they are often 1-pixel (or “very few”-pixel) regions. They are often near the surroundings of an actual anomaly (e.g. Fryum/Image 048). Extreme cases where completely isolated regions with insignificant size also occur (e.g. Chewing Gum/Image 068 and Macaroni 2/Image 067).

How often and how small are these tiny regions? Tab. 1 shows statistics about the absolute sizes (at original resolution) and the number of tiny regions per image in each dataset from VisA. The right-most plot in Fig. 3 shows VisA’s anomalous region size distribution when discarding the tiny regions.

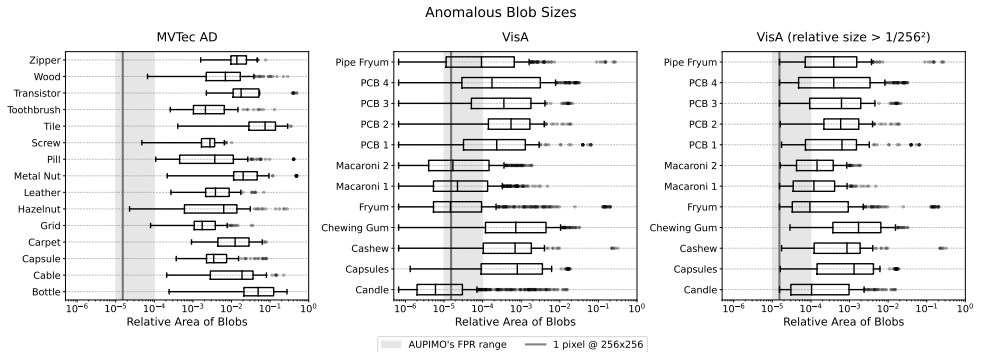


Figure 3: Distribution of relative size of anomalous regions.

Table 1: Statistics from tiny blobs in VisA [28].

(a) Sizes.				(b) Number of regions per image.			
Reg Size (abs) Category	1 - 9	10 - 19	20 - 29	Nb Reg/Img Category	1 - 5	6+	Total
Candle	358	98	20	Candle	5	18	23
Capsules	8	7	3	Capsules	6	1	7
Cashew	10	0	1	Cashew	5	0	5
Chewing Gum	39	1	0	Chewing Gum	6	1	7
Fryum	158	96	22	Fryum	22	13	35
Macaroni 1	114	52	14	Macaroni 1	27	10	37
Macaroni 2	123	54	6	Macaroni 2	21	8	29
PCB 1	19	20	9	PCB 1	10	2	12
PCB 2	11	8	4	PCB 2	10	1	11
PCB 3	20	11	0	PCB 3	8	1	9
PCB 4	32	19	12	PCB 4	10	5	15
Pipe Fryum	44	34	9	Pipe Fryum	17	3	20

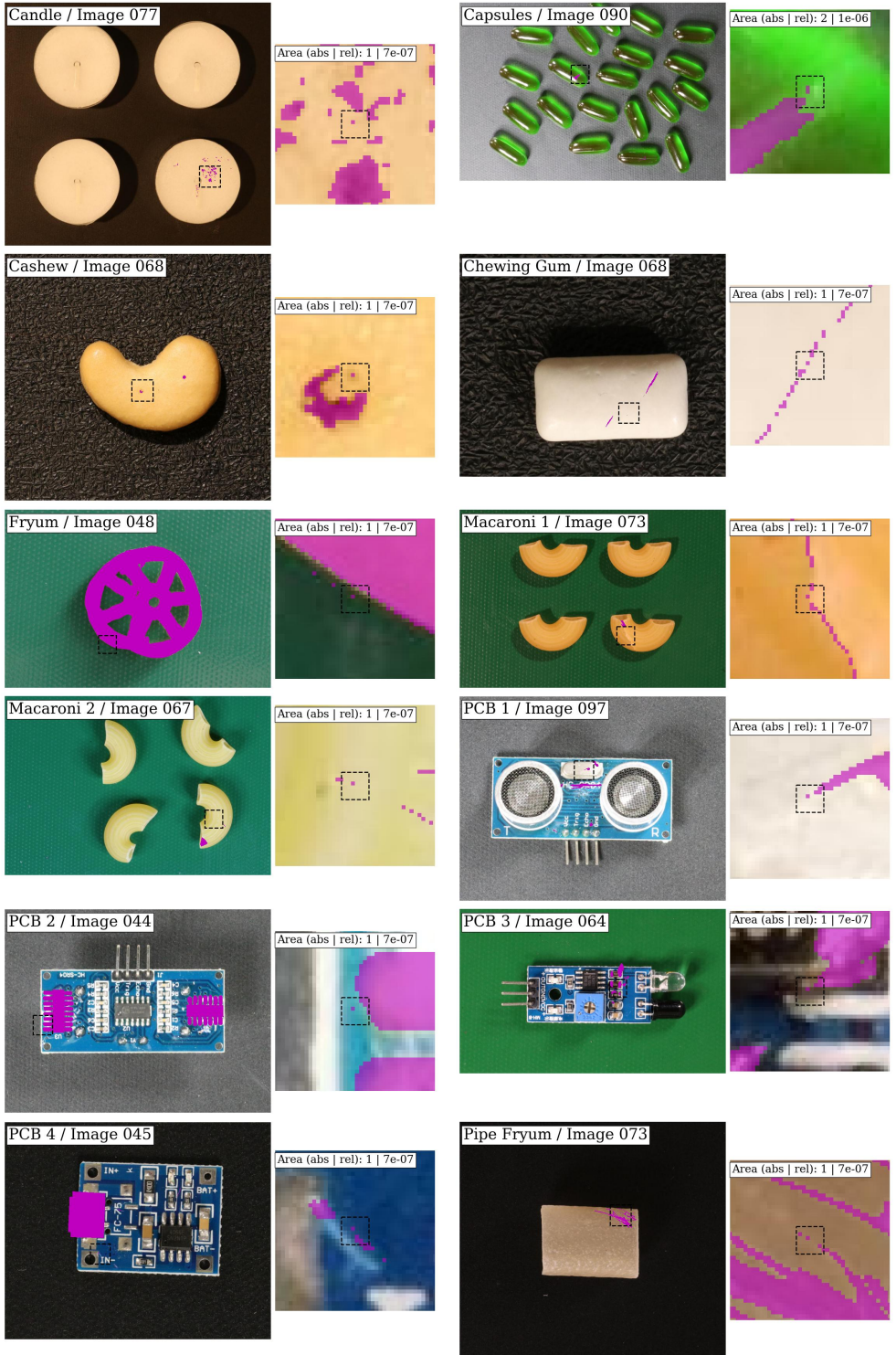


Figure 4: Tiny anomalous regions in VisA.

C Additional results

C.1 Ablation study

Tab. 2 showcases the use of statistical tests in an ablation study of EfficientAD [10] on the dataset MVTec AD / Capsule. The Wilcoxon signed-rank test [13, 14] is used to assess the consistency of performance gain given by different components of the model. The null hypothesis H_0 is that two models A and B are equivalent (average ranks tend to equal), and the alternative hypothesis H_1 is that one of the two models (say, A) is *more often* better than B . No assumption is made about the scores distributions making it robust to outliers [13, 14]. Interpretation: high confidence ($C = 1 - \text{p-value}$) to reject the null hypothesis (*i.e.* low p-value) means that A *consistently* outperforms B .

Table 2: Ablation study (use-case of statistical tests). Layout and model configurations based on Tab. 4 in [10]. At each row a component is added to the model above starting with Patch Description Network (PDN) at top and resulting in EfficientAD at the bottom. C refers to the confidence to reject the null hypothesis ($1 - \text{p-value}$); higher means more confidence on the improvement by adding the new component. Each component generally shows significant improvements, but the bottom right cell is an exception. Pretraining penalty causes a score drop, and the low confidence on the alternative hypothesis confirms that the drop is consistent across images.

Avg. AUPIMO (Diff. [%]; C [%])	EfficientAD-S	EfficientAD-M
PDN	~ 0	~ 0
→ map normalization	22 (+22; 100)	23 (+23; 100)
→ hard feature loss	57 (+35; 100)	59 (+36; 100)
→ pretraining penalty	64 (+7; 100)	57 (-2; 0.02)

C.2 Does AUPIMO correlate with AUROC and AUPRO?

Fig. 5 shows scatter plots of AUROC and AUPRO vs. (cross-image) average AUPIMO. All models and datasets in the benchmark confounded. Notice that the scales of the axes are different for each metric. Both plots seem to show a positive correlation, but one metric is not enough to imply the other. High levels of AUROC and AUPRO do not guarantee high levels of AUPIMO. Conversely, high levels of AUPIMO *tend* to imply higher levels of AUPRO and AUROC (notice the slightly triangular shape of the point clouds).

C.3 Video

In this section we present how AUPIMO can be used in video applications. It must be stressed that this is not a full-fledged video AUPIMO application, but rather a proof of concept. The UCSD Pedestrian dataset [15] was used to illustrate this concept because it has been widely used and cited in the literature, but other datasets like A Day on Campus (ADOC) [16] and Street Scene [17] would also be relevant to this discussion.

A PatchCore [22] model was trained on the normal videos from UCSD Pedestrian dataset at every 2 frames. The model was evaluated with the same procedure than our experiments by ignoring the temporal dimension of the videos and treating all the frames from all the

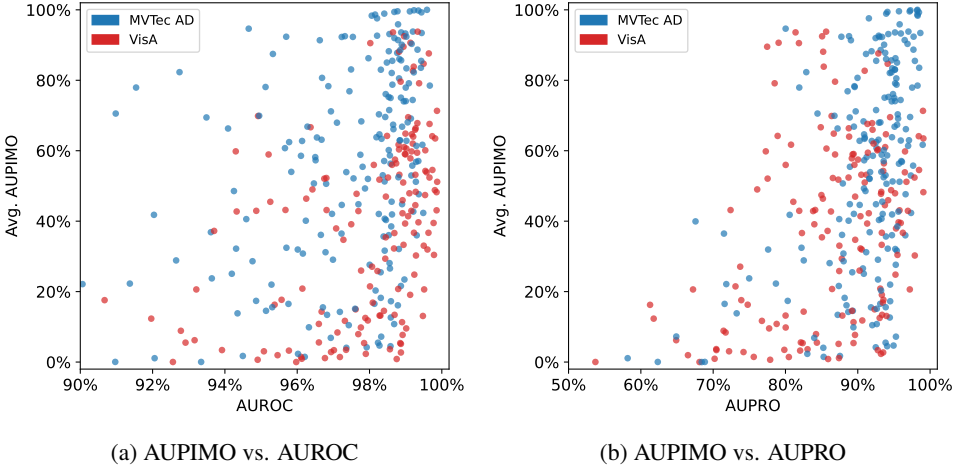


Figure 5: Scatter plots of AUPIMO vs. {AUROC, AUPRO}

videos as a single dataset. In Fig. 6 we show the AUPIMO scores for each frame in the test videos along the time axis (referenced by the frame index). A selection of frames from the video Test006 are shown in Fig. 7.

Notice how AUPIMO’s validation works in practice: the normal frame (175) does not have any visible FP region – *i.e.* anomaly score values above the threshold t_L , corresponding to the lowest FPR level L used in AUPIMO. Frame 61 shows an example case where the image-scoped has limitations: the AUPIMO score is around 50% because there are two indendent anomalous regions in the frame; one of them is well detected by the model, but the other is ignored.. A better modelization for this case would require a more complete annotation where each instance of anomaly is labeled separately, which is not the case in the UCSD Pedestrian dataset.

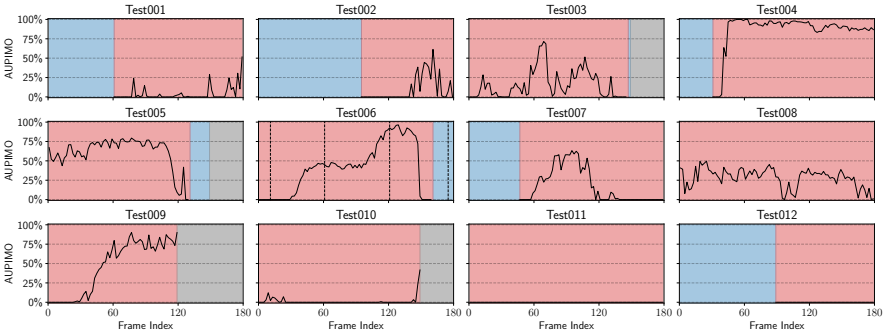


Figure 6: Time vs. AUPIMO in test videos from the UCSD Pedestrian dataset. The x-axis is the frame index in each video and the y-axis is the AUPIMO score at that frame. Blue zones indicate the frame is normal, red zones indicate the frame has an anomaly, and gray zones indicate there is no frame. Vertical dashed lines in "Test006" correspond to the frames shown in Fig. 7.

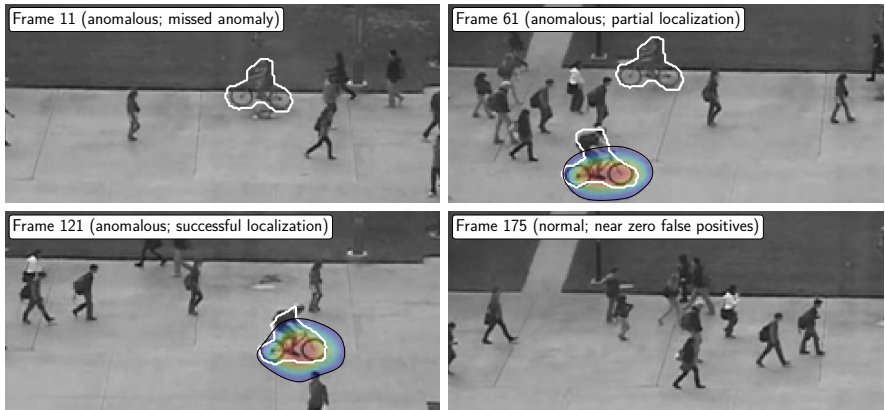


Figure 7: Frames from the video Test006. White contours indicate the ground truth anomalous regions. Black contours correspond to the level sets in each anomaly score map \mathbf{a} at t_L , where $F_{sh}(t_L) = L$. Anomaly scores above below t_L are not shown and above are colored using the JET colormap with local maxima in red and t_L in blue.

C.4 Precision vs. Intersection over Union

Since AUPIMO only concerns recall, our analysis lacks a discussion about the segmentation quality. In this section we aim to mitigate this shortcoming by extending our validation-evaluation framework. Two candidate metrics are considered: the image-scoped precision and the image-scoped Intersection over Union (IoU). As detailed in the next paragraphs, the precision is not suitable for our purposes, so the IoU is chosen to build a *Shared* FPR-based curve and an Area Under the Curve (AUC) score like PIMO and AUPIMO respectively. The anomaly score maps in this section are from PatchCore in the dataset MVTec AD / Metal Nut. We made this restricted choice to simplify the discussion, but similar results are obtained for the other datasets and models.

Fig. 8 shows the precision as a function of binarization thresholds in five images (note: not indexed by F_{sh} like PIMO). The level sets of the anomaly score maps at three thresholds along these curves are shown in black superimposed on the images, which can be compared to the contour from the ground truth annotations in white. The precision curves are not smooth, and optimizing this metric does not correspond to improving the visual aspect of the segmentation. It can be seen that optimizing for precision is not a viable option, as the segmentations tend to have a recall-disaster behavior as the precision increases.

The threshold-vs-precision curves show a breakpoint phenomenon where increasing the threshold generally increases the precision but dramatically decreases the recall at some point. For instance, in image 11 the breakpoint is between 60% and 62% precision; *i.e.* somewhere between their respective contour lines the segmentation switches from being too big to being too small (recall drops from 84% to 6%). In image 67, on the other hand, the breakpoint is between 95% and 98% precision (recall drops from 75% to 8% respectively). Image 102 shows an extreme case of this, where the segmentation is reduced to a nearly invisible region as the precision increases from 60% to 63% (recall drops from 91% to almost 0%).

The IoU curves in Fig. 9 (built in the same way as Fig. 8 described above) are smoother, generally show a global maximum, and the level sets at near-maximum-IoU are more visually stable. As the IoU accounts for a balance between precision and recall, it is a more suitable metric for our purposes.

Fig. 10 shows the Shared FPR vs. IoU curve, which is analogous to the PIMO curve. From this curve, the AUC score is computed like AUPIMO using the same integration bounds (blue area in Fig. 10).

The cross-image average AUC scores were added to the results in our benchmark in Appendix D. Since the paper already contains a large number of figures, we decided to not include the distributions of the IoU scores in the paper, but this promising metric deserves in-depth analysis in future work.

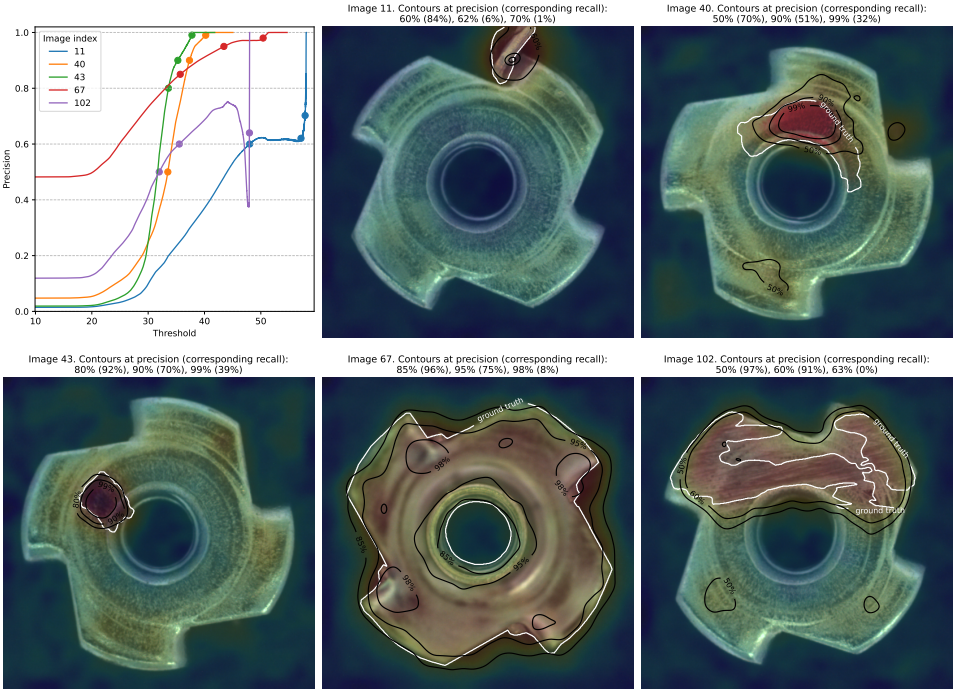


Figure 8: Precision curves and contours at different points along the curves.

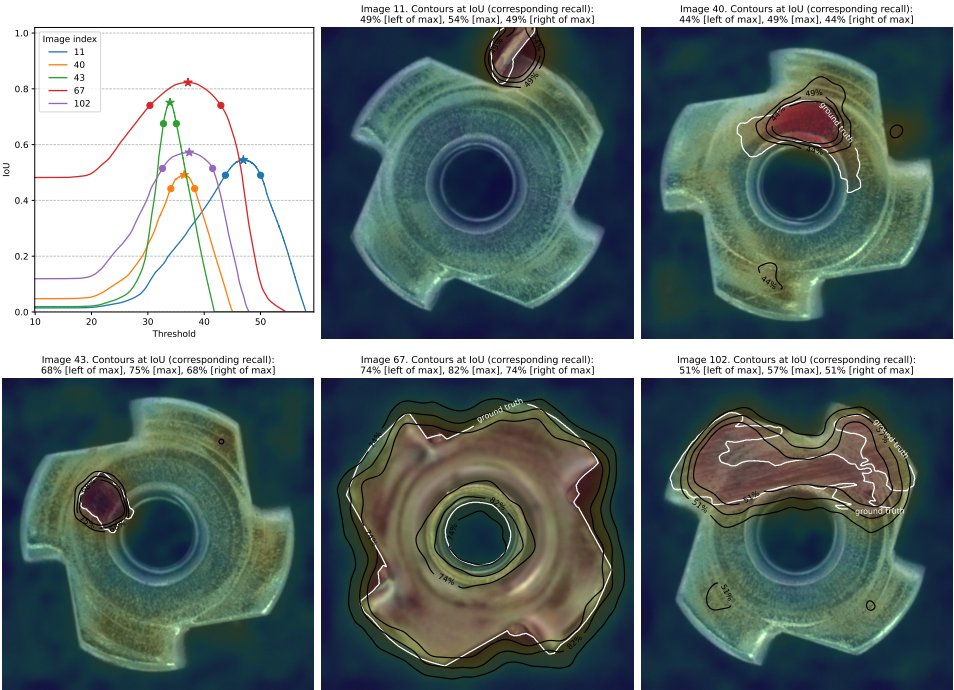


Figure 9: Intersection over union curves and contours at different points along the curves.

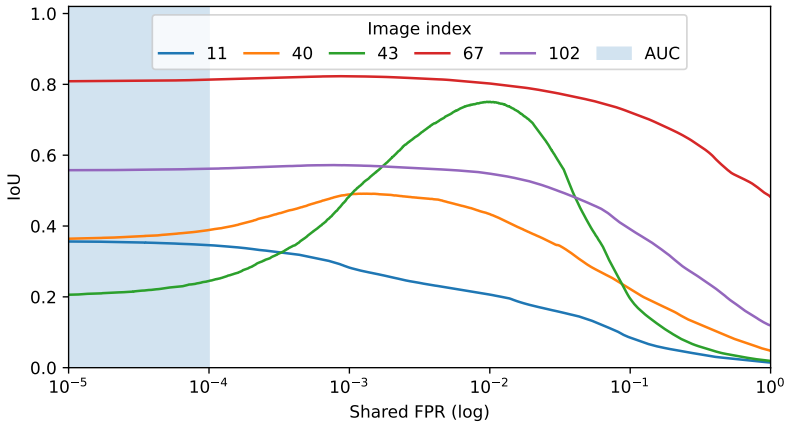


Figure 10: Shared FPR vs. IoU curve.

D Benchmark

In this section we provide additional details about our experiments and results omitted from the main text for brevity. The following paragraphs provide discuss and detail the evaluation guidelines in our experiments and define a standard format to publish AUPIMO scores.

Full resolution Many models typically downsample input images, which conveniently reduces computational costs. However, for a fair and model agnostic evaluation, it is important to use the original resolution as it impacts the decision-making when choosing the most suitable model for a use-case. If a small anomaly is missed due to downsampling, it is desirable to penalize this, while rewarding models that can handle higher resolution. As [2] pointed out, downsampling ground truth masks creates artifacts, leading to inconsistent results across papers. While AUPRO’s computational cost is high at full resolution – especially on CPU – AUPIMO is orders of magnitude faster (see our results in the paper). Our recommendation is to apply bilinear interpolation to upsample anomaly score maps and evaluate at the original resolution in each image.

No crop Center crop has been used to leverage the center alignment of the objects depicted in MVTec AD and VisA. However, this is a prior knowledge, hence we do not apply crop.

Sample selection To avoid biases from cherry-picking qualitative samples, we propose a systematic selection procedure. Select the images whose AUPIMO are closest to the statistics in a boxplot: mean, first/second/third quartiles, and low/high whiskers set with maximum size of 1.5IQR (inter-quartile range). We applied this procedure to select the samples shown in Appendix D. Note that this is applicable to any per-instance score.

Score publication We recommend to publish AUPIMO scores for all images. A standard format is specified below. The field `paths` is optional but recommended. For standard datasets like MVTec AD and VisA, it is a list of paths to the images in the test set with the path truncated to the dataset root directory. The field `num_threshs` is the effective number of thresholds used to compute the AUC, which differs from the number of thresholds used to compute the PIMO curve because only a portion of the curve is used to compute the AUPIMO score.

It is advised to report score distributions (*e.g.* as boxplots and histograms) when possible for a more comprehensive evaluation. All the scores from our experiments are available in this format at github.com/jpcbertoldo/aupimo.

```
{
  "shared_fpr_metric": "mean_perimage_fpr",
  "fpr_lower_bound": 0.00001,
  "fpr_upper_bound": 0.0001,
  "num_threshs": 300,
  "thresh_lower_bound": 0.3342,
  "thresh_upper_bound": 1.1588,
  "aupimos": [0.72107, 0.02415, 0.98991],
  "paths": [
    "MVTec/bottle/test/broken_large/000.png",
    "MVTec/bottle/test/broken_large/001.png",
    "MVTec/bottle/test/broken_large/002.png",
  ]
}
```

D.1 Models

Appendix D.1 lists the models in the benchmark and provides details on the implementation sources and hyperparameters.

We trained and evaluated 13 models from 8 papers listed in Tab. 3. For some models we considered two backbones and selected the (generally) best out of the two to show in the main text of the paper (see column Tab. 3).

We used the following implementations with the same hyperparameters reported in the papers:

- `anomalib` [10] (github.com/openvinotoolkit/anomalib¹) for PaDiM [8], PatchCore [22], and FastFlow [26];
- github.com/gasharper/PyramidFlow for PyramidFlow [14];
- github.com/donaldrr/simplenet for SimpleNet [15];
- github.com/tientrandinh/revisiting-reverse-distillation for RevDist++ [25];
- github.com/mtailanian/uflow for UFlow [24];
- github.com/nelson1425/EfficientAD for EfficientAD [9].

The non-official implementations are the ones from `anomalib` and EfficientAD.

Model	Publication	Backbone	Family	Paper	Implem.
PaDiM	ICPR 21	ResNet18	probability density	✓	<code>anomalib</code>
PaDiM	ICPR 21	WideResNet50	probability density	–	<code>anomalib</code>
PatchCore	CVPR 22	WideResNet50	memory bank	–	<code>anomalib</code>
PatchCore	CVPR 22	WideResNet101	memory bank	✓	<code>anomalib</code>
SimpleNet	CVPR 23	WideResNet50	reconstruction	✓	official
PyramidFlow	CVPR 23	ResNet18	normalizing flow	–	official
PyramidFlow	CVPR 23	–	normalizing flow	✓	official
RevDist++	CVPR 23	WideResNet50	student-teacher	✓	official
FastFlow	arXiv (21)	WideResNet50	normalizing flow	–	<code>anomalib</code>
FastFlow	arXiv (21)	Cait M48	normalizing flow	✓	<code>anomalib</code>
EfficientAD-S	arXiv (23)	WideResNet101	student-teacher	–	unofficial
EfficientAD-M	arXiv (23)	WideResNet101	student-teacher	✓	unofficial
UFlow	arXiv (23)	–	normalizing flow	✓	official

Table 3: Models. Years were abbreviated to the last two digits.

¹Commit 09ad1d4b1e8f634b72f788314275d3aea33815dd.

D.2 Cross-dataset analysis

In this section, the model performances are summarized across all the datasets in MVTec AD and VisA (all confounded) according to

1. AUROC (Fig. 11),
2. AUPRO (Fig. 12),
3. average AUPIMO (Fig. 13),
4. 33rd percentile AUPIMO (Fig. 14),
5. average image-wise rank according to AUPIMO scores (Fig. 15).

Scores In Fig. 11, Fig. 12, Fig. 13, and Fig. 14, each point represents the score in the test set or an AUPIMO statistic (average and 33rd percentile) across the images in the test set. Diamonds are averages across datasets (both collections confounded) or across models. Notice the difference in the X-axis scales.

Percentile 33 score While the average AUPIMO is a useful indicator, we propose the use of the 33rd percentile of AUPIMO scores, denoted P_{33} , for a more rigorous, worst-case evaluation. A P_{33} score of value V indicates that two thirds of the images in the test set have an AUPIMO score of *at least* V . Otherwise stated, a P_{33} score of value V indicates that one third of the images in the test set have an AUPIMO score of *at most* V .

Average ranks Fig. 15 shows the average image ranks according to AUPIMO as points and the average across datasets as diamonds. At each image from a given dataset, ranks are assigned to the models (“which model best detects this specific image?”), and the average is taken across all images from the same dataset. The range of rank values is from 1 (best) to number of models (worst).

Summary table Tab. 4 summarizes the average scores across datasets within each dataset collection (MVTec AD and VisA) and across all datasets (both collections confounded).

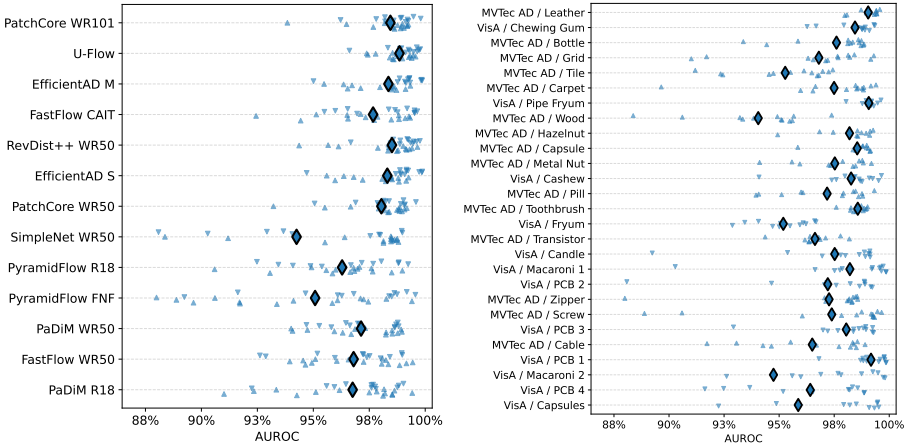


Figure 11: AUROC

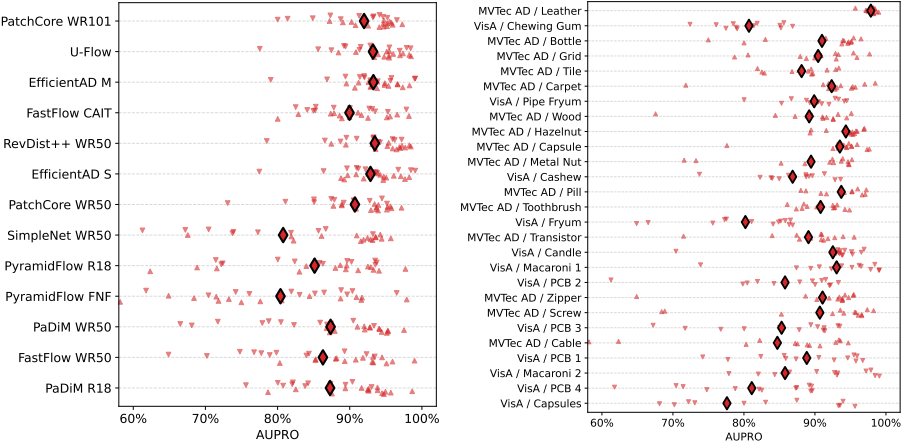


Figure 12: AUPRO

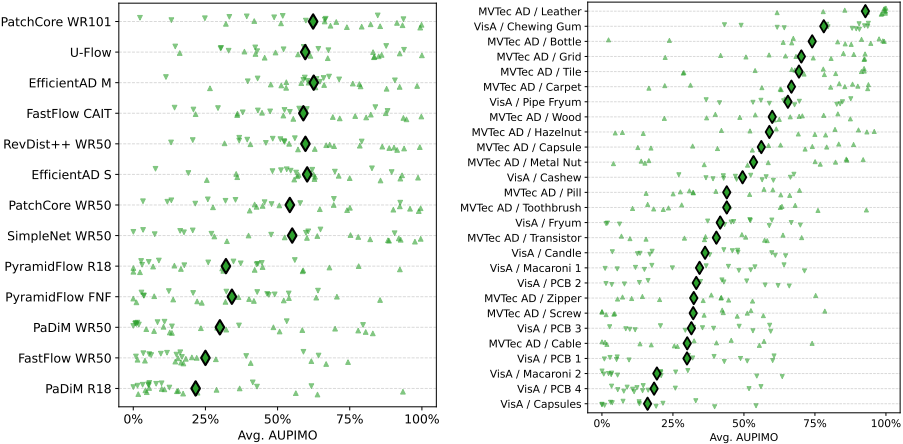


Figure 13: Average AUPIMO

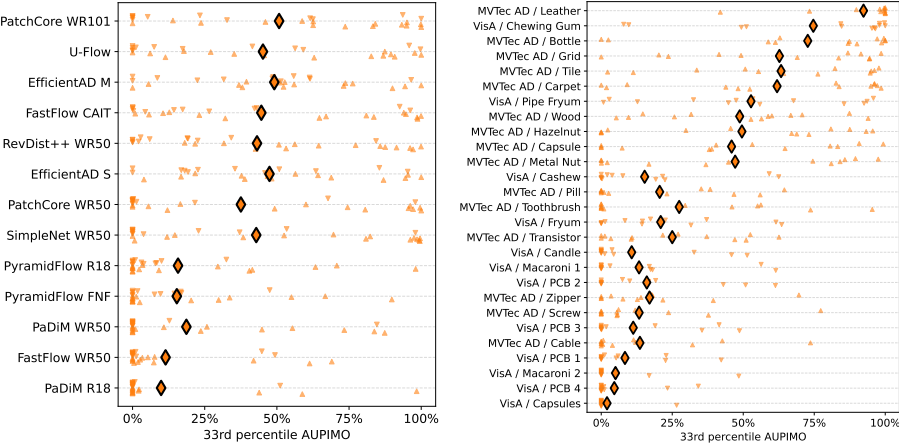


Figure 14: P_{33} AUPIMO

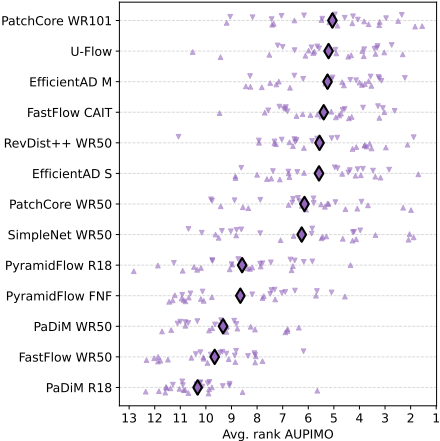


Figure 15: Average rank according to AUPIMO

Model	Dataset Collection	AUROC	AUPRO	AUPIMO	Avg.	P ₃₃	Avg. Rank
PaDiM R18	MVTec AD	96.62	91.58		25.75	14.34	10.5
PaDiM R18	VisA	97.22	81.87		16.42	4.33	10.1
PaDiM R18	All	96.89	87.27		21.61	9.89	10.3
FastFlow WR50	MVTec AD	97.01	90.87		28.49	14.15	10.3
FastFlow WR50	VisA	96.83	80.54		20.65	8.03	8.9
FastFlow WR50	All	96.93	86.28		25.00	11.43	9.7
PaDiM WR50	MVTec AD	97.19	92.57		40.14	27.06	8.9
PaDiM WR50	VisA	97.32	80.81		17.34	8.12	9.9
PaDiM WR50	All	97.25	87.35		30.01	18.64	9.3
PyramidFlow FNF	MVTec AD	94.21	79.10		36.26	19.94	9.4
PyramidFlow FNF	VisA	96.62	82.03		31.55	9.56	7.8
PyramidFlow FNF	All	95.28	80.40		34.17	15.33	8.7
PyramidFlow R18	MVTec AD	96.36	85.81		36.32	23.91	9.0
PyramidFlow R18	VisA	96.53	84.27		26.84	5.55	8.1
PyramidFlow R18	All	96.44	85.13		32.11	15.75	8.6
SimpleNet WR50	MVTec AD	97.13	89.48		71.39	62.78	5.3
SimpleNet WR50	VisA	91.17	69.88		34.66	17.93	7.4
SimpleNet WR50	All	94.48	80.77		55.07	42.84	6.3
PatchCore WR50	MVTec AD	98.01	93.13		67.21	54.95	5.6
PatchCore WR50	VisA	98.26	87.69		38.02	15.74	6.9
PatchCore WR50	All	98.12	90.72		54.24	37.53	6.1
EfficientAD S	MVTec AD	97.96	93.65		64.76	55.16	5.9
EfficientAD S	VisA	98.89	91.90		54.62	37.78	5.2
EfficientAD S	All	98.37	92.87		60.25	47.44	5.6
RevDist++ WR50	MVTec AD	98.23	95.03		71.93	64.93	4.9
RevDist++ WR50	VisA	99.00	91.53		44.30	15.85	6.3
RevDist++ WR50	All	98.57	93.48		59.65	43.11	5.6
FastFlow CAIT	MVTec AD	97.37	90.44		66.79	57.83	5.4
FastFlow CAIT	VisA	98.25	89.37		49.10	28.09	5.4
FastFlow CAIT	All	97.76	89.96		58.93	44.61	5.4
EfficientAD M	MVTec AD	97.96	94.10		66.08	55.97	5.8
EfficientAD M	VisA	99.00	92.25		58.06	40.52	4.6
EfficientAD M	All	98.42	93.28		62.52	49.10	5.2
U-Flow	MVTec AD	98.74	94.89		66.07	56.07	5.4
U-Flow	VisA	99.09	91.14		51.48	31.54	4.9
U-Flow	All	98.89	93.22		59.58	45.17	5.2
PatchCore WR101	MVTec AD	98.35	93.53		73.19	66.12	4.7
PatchCore WR101	VisA	98.70	90.06		48.72	31.58	5.5
PatchCore WR101	All	98.51	91.99		62.31	50.77	5.1

Table 4: Model averages. Scores are in percentages. Ranks range from 1 (best) to number of models (worst).

D.3 Per-model analyses

Fig. 16 shows that current anomaly localization models still are not capable of cracking the datasets from MVTec AD and VisA. Fig. 16b shows the AUPIMO distributions of PatchCore WR101, the model with best cross-dataset average. Even though it is the overall best, it still has a long tail of low AUPIMO scores on several datasets like Grid and Wood, or in some cases it practically fails to detect any anomaly at all, like in Capsules and Macaroni 2. Fig. 16b shows the AUPIMO distributions of the best model per dataset. Even if a user would be willing to select a model per dataset, there is no clear winner, and most datasets from VisA show challenging images that are not detected.

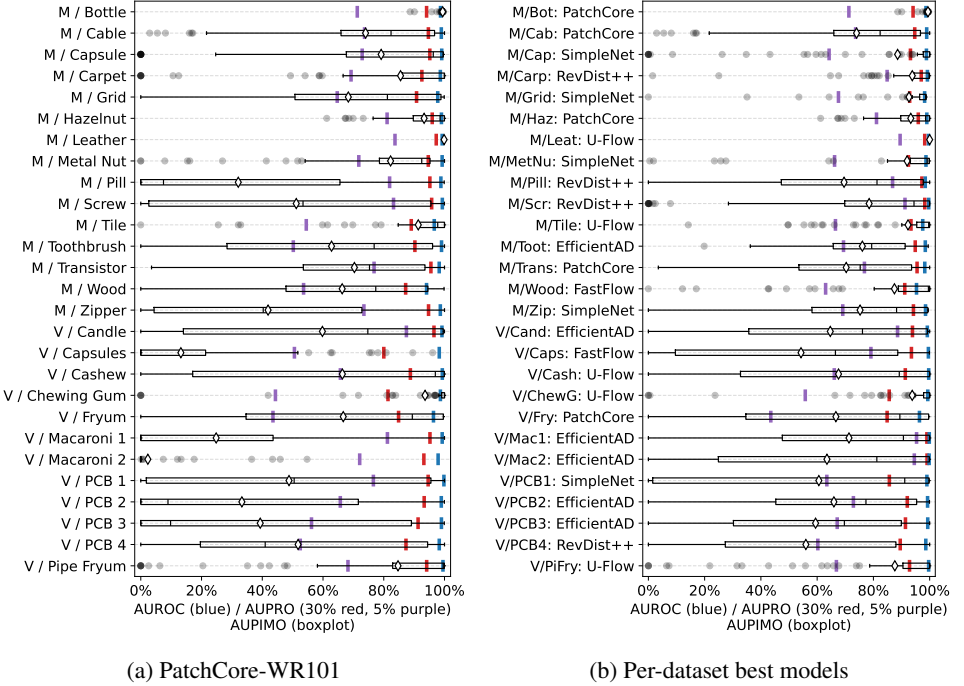


Figure 16: AUPIMO distributions for PatchCore-WR101 (left) and per-dataset best models (right).

D.4 Per-dataset analyses

The following figures are detailed results from the benchmark of all the datasets from MVTec AD or VisA.

- | | |
|------------------------------------|---------------------------------|
| 1. Fig. 17: MVTec AD / Bottle | 15. Fig. 31: MVTec AD / Zipper |
| 2. Fig. 18: MVTec AD / Cable | 16. Fig. 32: VisA / Candle |
| 3. Fig. 19: MVTec AD / Capsule | 17. Fig. 33: VisA / Capsules |
| 4. Fig. 20: MVTec AD / Carpet | 18. Fig. 34: VisA / Cashew |
| 5. Fig. 21: MVTec AD / Grid | 19. Fig. 35: VisA / Chewing Gum |
| 6. Fig. 22: MVTec AD / Hazelnut | 20. Fig. 36: VisA / Fryum |
| 7. Fig. 23: MVTec AD / Leather | 21. Fig. 37: VisA / Macaroni 1 |
| 8. Fig. 24: MVTec AD / Metal Nut | 22. Fig. 38: VisA / Macaroni 2 |
| 9. Fig. 25: MVTec AD / Pill | 23. Fig. 39: VisA / PCB 1 |
| 10. Fig. 26: MVTec AD / Screw | 24. Fig. 40: VisA / PCB 2 |
| 11. Fig. 27: MVTec AD / Tile | 25. Fig. 41: VisA / PCB 3 |
| 12. Fig. 28: MVTec AD / Toothbrush | 26. Fig. 42: VisA / PCB 4 |
| 13. Fig. 29: MVTec AD / Transistor | 27. Fig. 43: VisA / Pipe Fryum |
| 14. Fig. 30: MVTec AD / Wood | |

Each figure contains the following elements:

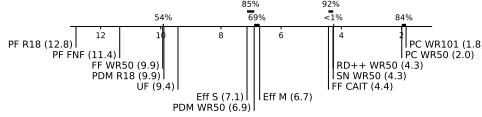
1. a plot with one model per row containing:
 - (a) the AUROC score as a blue vertical line;
 - (b) the AUPRO score as a red vertical line;
 - (c) a boxplot of AUPIMO scores;
 - i. lower and upper whiskers set with maximum size to 1.5 inter-quartile range (IQR);
 - ii. the mean is displayed as a white diamond;
 - iii. fliers are displayed as gray dots;
2. a diagram of (image-wise) average rank according to AUPIMO scores; lower is better; 1 means that the model has the best AUPIMO score at all images;
3. a table comprising two parts:
 - (a) the upper part, in bold, comprises:
 - i. the AUROC scores (in blue);
 - ii. the AUPRO scores (in red);
 - iii. the average and standard deviation AUPIMO score (in black);

- iv. the 33rd percentile AUPIMO score (in black);
 - v. the values in parentheses are the ranks of the models according to the respective score metric in each row;
- (b) the lower part shows the results of pairwise Wilcoxon signed rank tests using AUPIMO scores; each cell shows the confidence to reject the null hypothesis $C = 1 - p$ (where p is the p-value) assuming that the row model is better than the column model as alternative hypothesis; confidence values below 95% (*i.e.* “low confidence”) are highlighted in bold;
4. PIMO curves and heatmap samples from the model with best average AUPIMO rank;
- (a) samples are selected according to the recommendations from the paragraph “Sample selection”;
 - (b) the (2-pixel wide, outer) contour of the ground truth mask is shown in white.
 - (c) heatmaps are colored according to the color scheme described below;

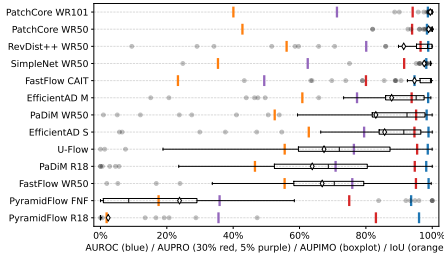
Heatmaps coloring scheme The input images are superimposed by their respective anomaly score map **a**. Coloring rules are linked to the thresholds in AUPIMO’s integration bounds: transparent is for scores below the lowest threshold, blues are for scores between the lowest and the highest thresholds, and reds are for scores above the highest threshold. Darker blue/red tones mean higher scores. The coloring strategy links the heatmaps to the validation-evaluation framework employed in AUPIMO. Transparent heatmap zones are never accounted in the metric because the validation requirement is not respected. Blue zones visually express the average recall measured by the integration in AUPIMO. Additionally, red zones show the model’s local behavior (per-image normalization) within the *valid* score range (*i.e.* scores above the threshold given by the Shared FPR lower bound).

	PF R18	PF FNF	FF WR50	PDM R18	UF	Eff S	PDM WR50	Eff M	FF CAIT	SN WR50	RD++ WR50	PC WR50	PC WR101
AUROC	98.0% (11)	93.0% (13)	98.9% (1)	98.2% (10)	98.7% (6)	98.9% (2)	98.3% (8)	98.8% (4)	94.7% (12)	98.2% (9)	98.8% (3)	98.4% (7)	98.7% (5)
AUPRO	83.0% (11)	75.0% (13)	95.3% (1)	94.7% (15)	95.4% (12)	96.6% (6)	95.1% (13)	93.8% (19)	86.0% (12)	91.5% (10)	96.5% (1)	93.4% (8)	94.1% (17)
AUPRO 5%	93.8% (11)	85.9% (13)	95.3% (1)	95.3% (17)	95.4% (14)	96.5% (12)	95.3% (13)	92.3% (13)	87.4% (19)	91.5% (10)	96.5% (1)	93.4% (8)	94.1% (17)
Avg. AUPIMO	2.3% (13)	23.8% (12)	66.8% (10)	63.8% (11)	67.4% (9)	85.7% (7)	83.0% (8)	87.8% (6)	94.7% (4)	97.6% (3)	91.4% (5)	98.9% (2)	99.4% (1)
Std. AUPIMO	7.9%	11.3%	13.8%	18.8%	16.3%	25.9%	20.0%	24.3%	10.8%	9.8%	18.3%	1.4%	1.9%
P33 AUPIMO	0.0% (13)	2.5% (12)	61.3% (10)	58.8% (11)	62.9% (9)	86.9% (8)	87.4% (7)	90.0% (6)	98.7% (4)	99.5% (3)	97.7% (5)	100.0% (2)	100.0% (1)
Avg. Rank	12.8	11.4	9.9	9.3	9.4	7.1	7.3	6.7	4.4	4.3	4.3	2.0	1.8
Avg. Iou	1.3% (13)	17.3% (12)	55.3% (5)	46.4% (17)	55.5% (4)	62.3% (11)	62.3% (11)	60.8% (12)	23.4% (11)	35.4% (10)	56.1% (1)	42.8% (18)	40.1% (19)
PC WR101 (1.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (2.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (4.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (6.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (6.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (7.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (9.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (9.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (18.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (11.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

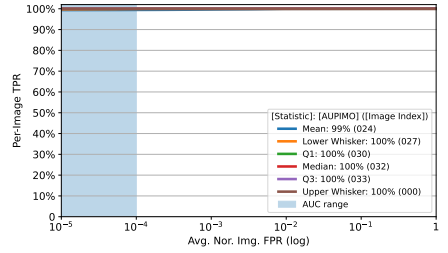
(a) Statistics and pairwise statistical tests.



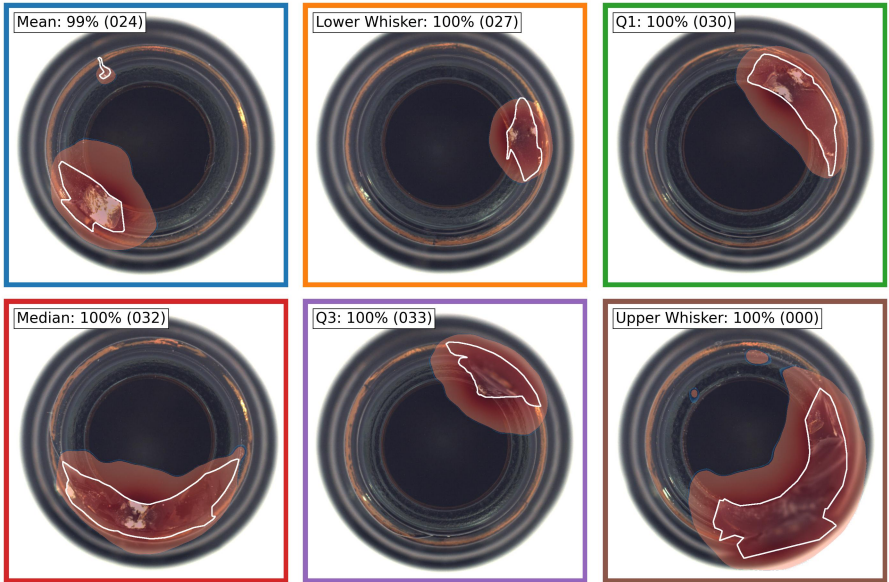
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

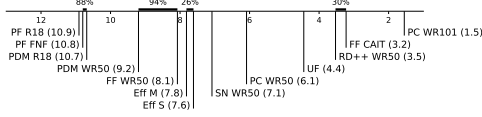


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

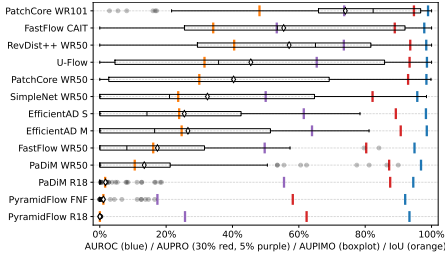
Figure 17: Benchmark on MVTec AD / Bottle. PIMO curves and heatmaps are from PatchCore WR101. 083 images (020 normal, 063 anomalous).

	PF R18	PF FNF	PDM R18	PDM WR50	FF WR50	Eff M	Eff S	SN WR50	PC WR50	UF	RD++ WR50	FF CAIT	PC WR101
AUROC	93.3% (12)	92.1% (13)	94.5% (11)	95.8% (8)	94.9% (10)	98.4% (3)	98.3% (6)	95.7% (9)	98.3% (3)	98.8% (2)	98.4% (5)	97.8% (7)	98.9% (1)
AUPRO	92.3% (12)	90.2% (13)	92.5% (10)	97.5% (8)	90.3% (11)	99.7% (1)	99.7% (1)	99.2% (1)	99.2% (1)	99.4% (1)	99.4% (1)	98.9% (1)	99.7% (1)
AUPRO 5%	25.7% (10)	3.4% (13)	4.3% (11)	21.1% (9)	45.9% (10)	64.9% (3)	61.5% (5)	50.2% (10)	65.3% (1)	73.6% (2)	53.3% (1)	57.1% (1)	77.9% (1)
Avg. AUPIMO	0.6% (13)	1.1% (12)	1.7% (11)	13.4% (10)	17.4% (9)	26.4% (8)	25.5% (8)	32.5% (6)	40.3% (5)	45.5% (4)	57.1% (2)	55.4% (3)	74.0% (1)
Std. AUPIMO	0.1%	3.4%	4.3%	21.1%	45.9%	64.9%	61.5%	50.2%	65.3%	73.6%	53.3%	57.1%	77.9%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.5% (8)	0.6% (9)	3.9% (6)	0.7% (7)	12.1% (4)	11.4% (5)	42.6% (2)	32.1% (3)	73.7% (1)
Avg. Rank	10.9	10.8	10.7	10.2	9.1	7.8	7.8	7.1	6.1	5.5	3.2	3.2	1.5
Avg. Std	0.0% (13)	1.1% (12)	1.6% (11)	10.3% (10)	18.1% (9)	24.7% (6)	23.9% (7)	23.9% (10)	30.9% (5)	33.9% (4)	40.5% (2)	32.2% (3)	42.1% (1)
PC WR101 (1.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (3.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (3.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (6.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (7.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (7.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (7.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (8.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (9.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (10.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (10.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

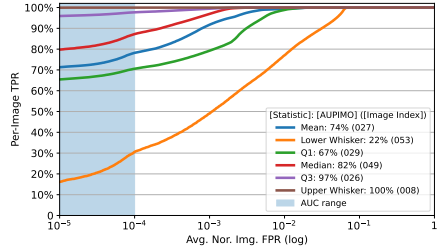
(a) Statistics and pairwise statistical tests.



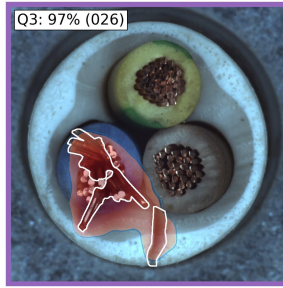
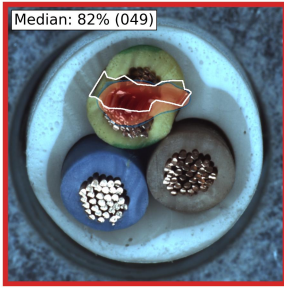
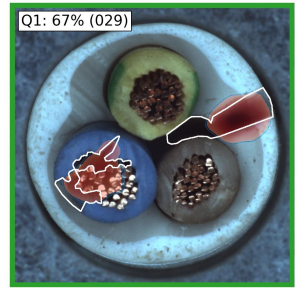
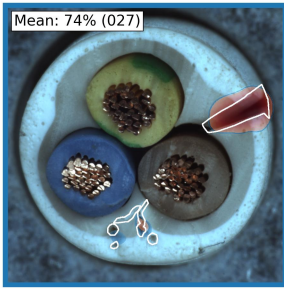
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

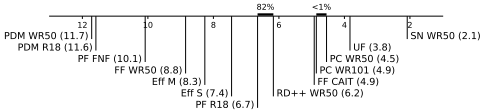


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner. Image index annotated on upper left corner.

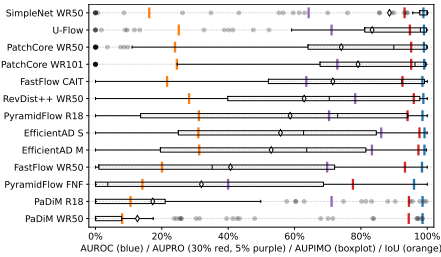
Figure 18: Benchmark on MVTec AD / Cable. PIMO curves and heatmaps are from PatchCore WR101. 150 images (058 normal, 092 anomalous).

	PDM WR50	PDM R18	PF FNF	FF WR50	EFF M	EFF S	PF R18	RD++ WR50	FF CAIT	PC WR101	PC WR50	UF	SN WR50
AUROC	98.8% (9)	98.8% (8)	96.9% (13)	98.4% (12)	99.2% (2)	99.2% (3)	98.6% (10)	98.8% (7)	98.5% (11)	99.1% (3)	99.0% (5)	99.0% (8)	98.9% (6)
AUPRO	94.4% (9)	94.4% (7)	97.6% (13)	99.3% (10)	97.3% (2)	97.3% (3)	94.1% (9)	95.9% (3)	92.5% (12)	95.2% (4)	95.2% (5)	94.4% (6)	93.2% (11)
AUPRO 5%	71.3% (13)	71.3% (13)	89.9% (11)	99.8% (9)	89.3% (12)	89.3% (11)	70.4% (17)	75.3% (11)	63.3% (19)	72.9% (14)	72.9% (14)	71.3% (10)	68.3% (18)
Avg. AUPIMO	12.6% (13)	17.3% (12)	31.9% (11)	40.7% (10)	53.0% (9)	55.8% (8)	58.7% (7)	62.9% (6)	71.3% (5)	79.1% (3)	74.1% (4)	83.4% (2)	88.8% (1)
Std. AUPIMO	26.1%	28.1%	37.3%	36.1%	34.3%	32.9%	30.8%	33.0%	34.1%	31.3%	34.4%	28.1%	28.7%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	7.4% (10)	38.0% (9)	38.9% (8)	40.5% (7)	53.2% (6)	64.7% (5)	81.3% (3)	80.6% (4)	90.8% (2)	99.3% (1)
Avg. Rank	11.7	11.6	10.1	8.8	7.4	7.4	6.2	5.3	4.9	4.9	4.5	3.8	2.1
Avg. Iou	8.0% (13)	10.4% (12)	14.2% (11)	20.1% (10)	31.3% (9)	31.0% (8)	31.5% (7)	28.2% (6)	21.6% (5)	23.5% (4)	24.9% (3)	25.1% (2)	16.2% (1)
SN WR50 (2.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (3.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (4.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (6.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (6.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EFF S (7.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EFF M (8.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF WR50 (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (10.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (11.6)	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

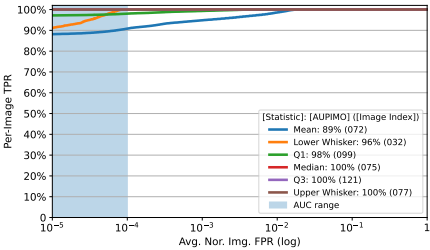
(a) Statistics and pairwise statistical tests.



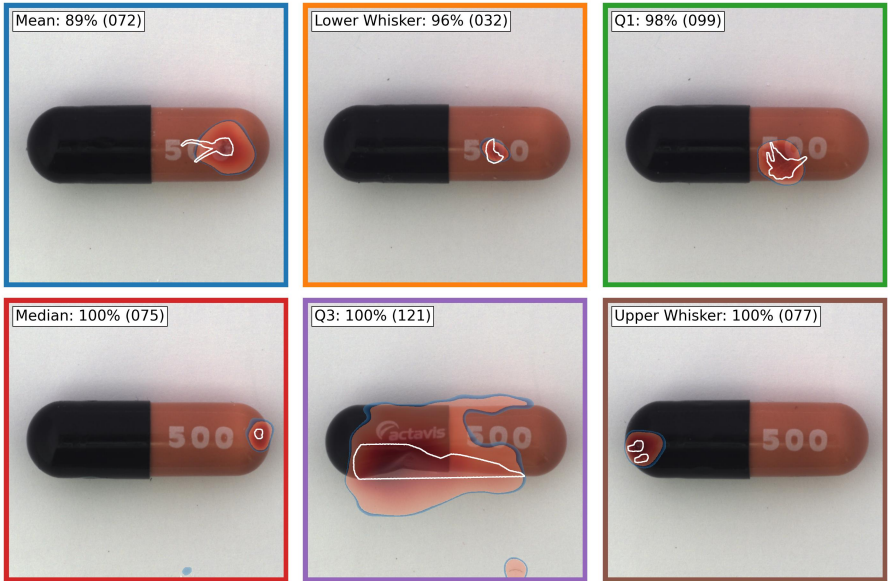
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

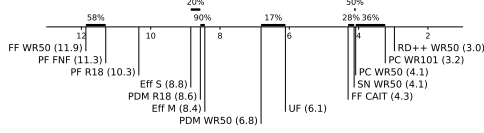


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

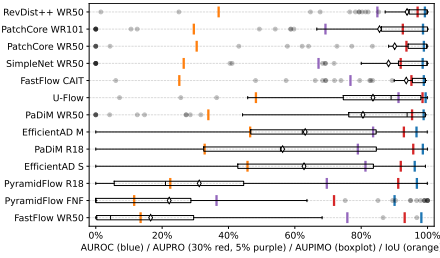
Figure 19: Benchmark on MVTec AD / Capsule. PIMO curves and heatmaps are from SimpleNet WR50. 132 images (023 normal, 109 anomalous).

	FF WR50	PF FNF	PF R18	Eff S	PDM R18	Eff M	PDM WR50	UF	FF CAIT	SN WR50	PC WR50	PS WR101	RD++ WR50
AUDROC	98.1% (1)	90.1% (13)	98.8% (10)	98.1% (12)	98.6% (7)	96.7% (13)	98.8% (5)	99.4% (1)	98.9% (8)	98.5% (8)	98.9% (3)	98.6% (6)	99.2% (2)
AUPRO	93.1% (7)	71.8% (13)	91.2% (12)	91.9% (10)	95.7% (8)	92.9% (8)	95.3% (4)	98.5% (1)	95.1% (9)	91.4% (13)	93.6% (6)	92.5% (9)	97.0% (2)
AUPRO 5%	95.8% (7)	86.4% (13)	91.9% (12)	91.9% (10)	95.7% (8)	93.4% (8)	95.3% (4)	98.5% (1)	95.1% (9)	91.4% (13)	93.6% (6)	92.5% (9)	97.0% (2)
Avg. AUPIMO	16.3% (13)	22.1% (12)	31.2% (11)	62.8% (9)	56.3% (10)	63.1% (8)	80.5% (7)	83.6% (6)	93.6% (2)	88.2% (4)	90.1% (3)	85.5% (5)	93.8% (1)
Std. AUPIMO	12.7%	15.1%	32.6%	29.1%	33.1%	25.1%	28.3%	18.4%	13.9%	26.1%	21.9%	28.4%	14.4%
P33 AUPIMO	1.9% (12)	0.0% (13)	9.1% (11)	48.2% (9)	43.9% (10)	52.3% (8)	83.8% (6)	82.7% (7)	98.6% (1)	98.2% (3)	97.2% (2)	95.0% (5)	96.1% (4)
Avg. Rank	11.9	11.9	10.3	6.8	8.6	8.4	6.8	6.1	4.3	4.1	4.1	3.2	3.0
Avg. iou	13.1% (12)	21.6% (11)	22.2% (11)	45.7% (9)	32.8% (6)	46.3% (2)	33.9% (3)	50.7% (1)	25.7% (10)	26.3% (9)	30.4% (7)	23.6% (8)	37.0% (1)
RD++ WR50 (3.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (3.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (6.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (6.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (8.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (8.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (10.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (11.3)	59%												

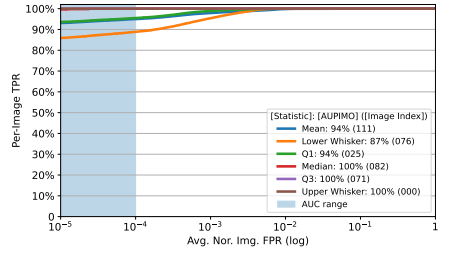
(a) Statistics and pairwise statistical tests.



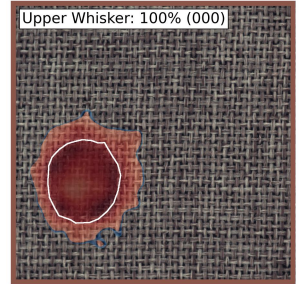
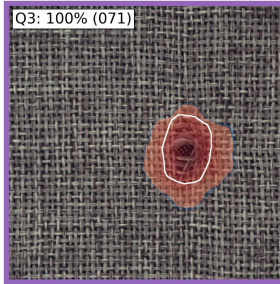
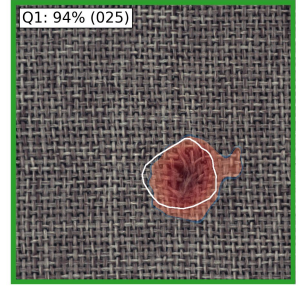
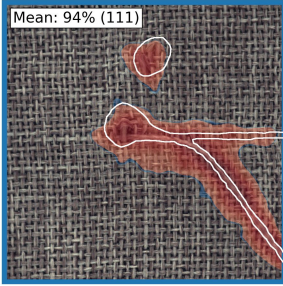
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

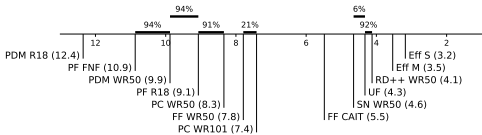


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

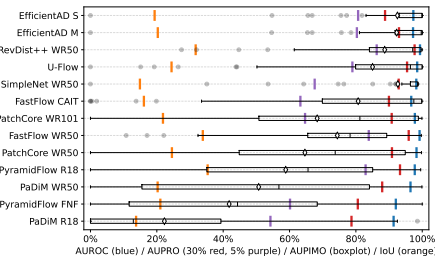
Figure 20: Benchmark on MVTec AD / Carpet. PIMO curves and heatmaps are from RevDist++ WR50. 117 images (028 normal, 089 anomalous).

	PDM R18	PF FNF	PDM WR50	PF R18	PC WR50	FF WR50	PC WR101	FF CAIT	SN WR50	UF	RD++ WR50	Eff M	Eff S
AUROC	91.4% (13)	92.0% (12)	96.5% (11)	97.7% (7)	98.3% (4)	99.2% (2)	97.8% (6)	96.7% (10)	98.2% (5)	98.5% (3)	99.3% (1)	97.2% (9)	97.8% (8)
AUPRO	78.7% (13)	80.4% (12)	87.9% (11)	93.3% (7)	90.9% (7)	95.9% (2)	90.4% (8)	90.0% (9)	92.9% (6)	95.4% (3)	97.7% (1)	93.4% (8)	88.4% (10)
AUPRO 5%	51.5% (11)	50.5% (10)	60.5% (11)	82.4% (7)	81.3% (7)	90.4% (2)	64.1% (10)	63.3% (10)	67.4% (7)	75.3% (3)	86.3% (1)	69.4% (11)	69.4% (11)
Avg. AUPIMO	22.3% (13)	41.8% (12)	50.7% (11)	58.9% (10)	64.4% (9)	74.4% (7)	68.3% (8)	80.7% (6)	92.7% (1)	85.0% (5)	88.7% (4)	92.3% (3)	92.5% (2)
Std. AUPIMO	28.0%	31.4%	33.6%	35.6%	33.3%	20.4%	14.3%	30.2%	17.0%	23.9%	12.1%	15.8%	13.6%
P33 AUPIMO	0.0% (13)	20.3% (12)	23.6% (11)	41.5% (10)	49.2% (9)	69.0% (7)	62.5% (8)	83.6% (6)	98.6% (1)	86.9% (5)	91.5% (4)	94.4% (3)	94.5% (2)
Avg. Rank	12.4	10.9	9.9	9.1	8.3	7.8	8.7	8.5	4.6	4.2	4.1	3.5	3.2
Avg. Iou	31.8% (13)	21.7% (7)	20.2% (9)	35.3% (1)	25.3% (4)	23.9% (2)	21.9% (6)	16.1% (11)	14.4% (12)	22.4% (3)	21.7% (1)	20.3% (8)	19.3% (10)
Eff S (1.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
Eff M (1.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
RD++ WR50 (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
UF (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
SN WR50 (4.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
FF CAIT (5.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
PC WR101 (7.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
FF WR50 (1.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
PC WR50 (8.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
PF R18 (9.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
PDM WR50 (9.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
PF FNF (10.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%

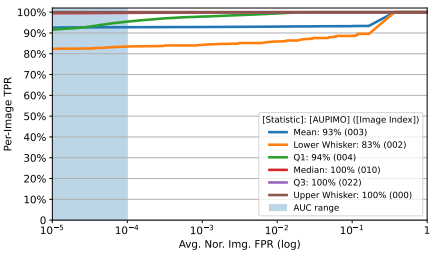
(a) Statistics and pairwise statistical tests.



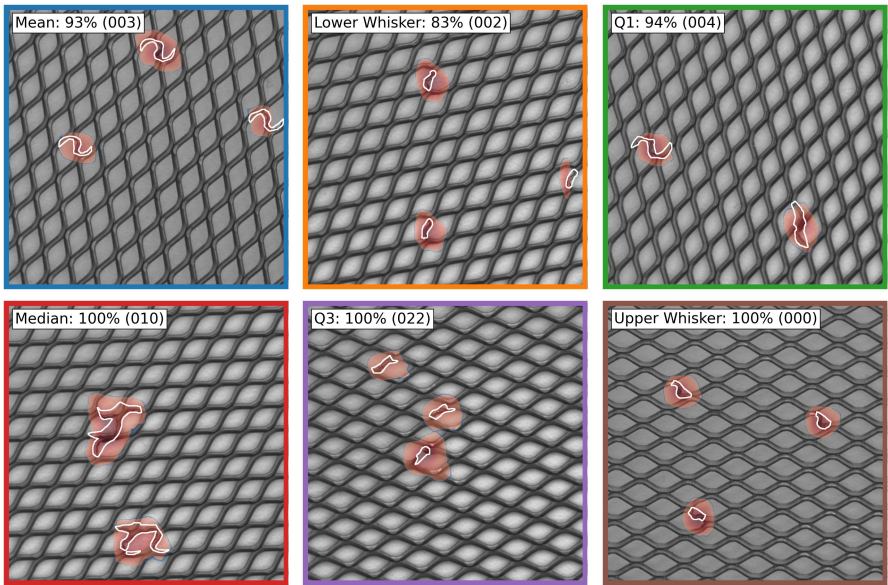
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

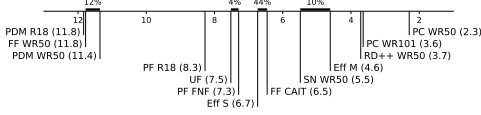


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

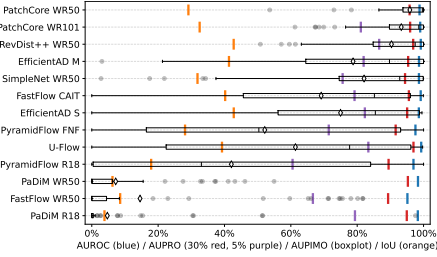
Figure 21: Benchmark on MVTec AD / Grid. PIMO curves and heatmaps are from EfficientAD S. 078 images (021 normal, 057 anomalous).

	PDM R18	FF WR50	PDM WR50	PF R18	UF	PF FNF	Eff S	FF CAIT	SN WR50	Eff M	RD++ WR50	PC WR101	PC WR50
AUROC	98.3% (10)	95.1% (13)	98.3% (9)	97.0% (12)	99.3% (1)	97.6% (11)	98.6% (8)	99.1% (3)	98.0% (7)	98.8% (6)	99.2% (2)	99.0% (4)	98.8% (5)
AUPRO	94.9% (9)	89.3% (12)	95.3% (12)	89.5% (12)	97.0% (12)	91.6% (11)	95.1% (8)	95.6% (10)	94.4% (10)	95.4% (8)	96.6% (1)	96.0% (1)	95.8% (4)
AUPRO 5%	76.3% (8)	66.4% (12)	76.3% (12)	65.5% (11)	83.4% (1)	71.4% (9)	82.3% (1)	82.3% (1)	76.5% (12)	81.3% (1)	86.6% (1)	81.3% (1)	81.3% (1)
Avg. AUPIMO	4.7% (13)	14.5% (11)	7.2% (12)	42.1% (10)	61.4% (8)	52.1% (9)	75.0% (6)	69.1% (7)	82.1% (4)	78.8% (5)	90.3% (3)	93.3% (2)	95.9% (1)
Std. AUPIMO	15.0%	28.0%	11.3%	18.9%	17.7%	17.3%	27.1%	34.1%	21.1%	23.0%	12.9%	9.7%	7.8%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	2.3% (10)	45.6% (8)	29.8% (9)	68.8% (6)	67.0% (7)	80.9% (4)	73.1% (5)	87.5% (3)	93.3% (2)	95.6% (1)
Avg. Rank	11.8	11.8	11.4	6.3	7.5	7.3	6.7	6.5	5.5	4.6	3.7	3.6	2.3
Avg. IQR	3.2% (13)	8.4% (11)	5.2% (12)	17.7% (10)	39.2% (5)	28.1% (9)	42.8% (11)	50.2% (12)	31.9% (9)	41.4% (11)	42.4% (12)	32.2% (8)	23.1% (10)
PC WR50 (2.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (3.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (13.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
Eff M (4.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (6.5)	100%	100%	100%	100%	98%	100%	45%	100%	100%	100%	100%	100%	100%
Eff S (6.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (7.3)	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%
UF (7.5)	100%	100%	100%	100%	100%	4%	100%	100%	100%	100%	100%	100%	100%
PF R18 (8.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (11.4)	99%	13%	100%										
FF WR50 (11.8)	99%												

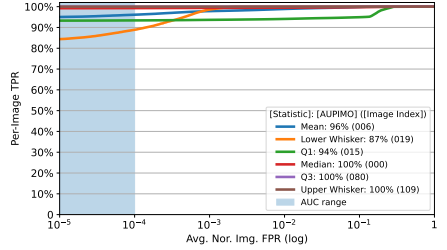
(a) Statistics and pairwise statistical tests.



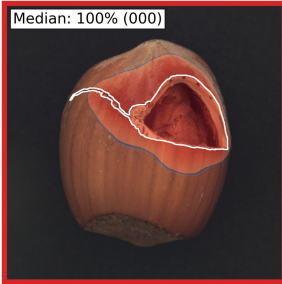
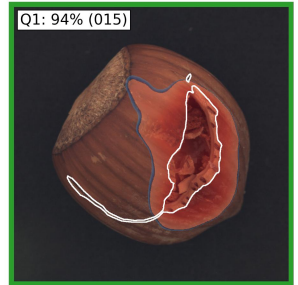
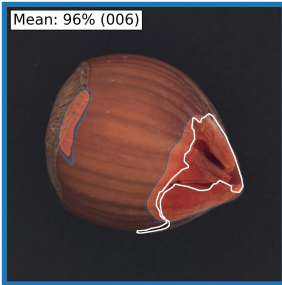
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

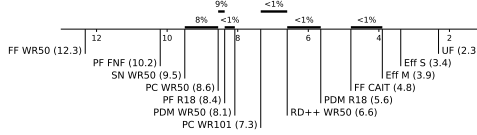


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

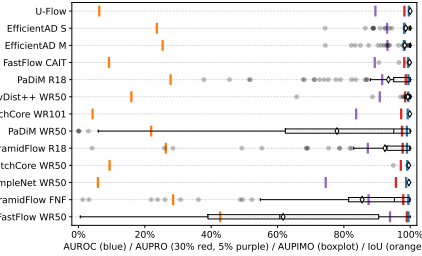
Figure 22: Benchmark on MVTec AD / Hazelnut. PIMO curves and heatmaps are from PatchCore WR50. 110 images (040 normal, 070 anomalous).

	FF WR50	PF FNF	SN WR50	PC WR50	PF R18	PDM WR50	PC WR101	RD++ WR50	PDM R18	FF CAIT	Eff M	Eff S	UF
AUROC	99.3% (2)	99.4% (5)	98.8% (11)	99.1% (8)	99.0% (9)	99.0% (10)	99.2% (7)	99.4% (4)	99.4% (13)	99.4% (6)	98.1% (18)	98.2% (12)	99.8% (1)
AUPRO	99.0% (1)	97.9% (8)	95.7% (13)	97.2% (12)	97.7% (9)	97.5% (10)	97.3% (11)	98.5% (3)	98.7% (17)	98.2% (17)	98.3% (14)	98.2% (16)	98.2% (15)
AUPRO 5%	99.3% (1)	97.4% (8)	95.3% (11)	97.2% (12)	97.2% (9)	97.1% (10)	95.3% (13)	91.3% (4)	91.3% (17)	91.3% (17)	91.3% (17)	91.3% (17)	91.3% (17)
Avg. AUPIMO	61.7% (13)	85.5% (11)	99.6% (4)	99.8% (5)	92.4% (10)	77.9% (12)	99.9% (2)	99.3% (6)	93.4% (9)	99.8% (3)	98.3% (8)	98.7% (7)	100.0% (1)
Std. AUPIMO	29.7%	21.0%	0.0%	0.0%	16.8%	30.3%	0.0%	1.7%	13.1%	1.0%	4.4%	4.0%	0.0%
P33 AUPIMO	41.9% (13)	89.8% (11)	99.6% (8)	99.6% (7)	98.1% (10)	73.3% (12)	99.9% (6)	99.9% (5)	98.4% (9)	99.9% (4)	100.0% (3)	100.0% (2)	100.0% (1)
Avg. Rank	12.3	10.2	9.5	8.4	8.4	8.1	7.3	6.6	5.4	4.8	3.9	3.4	2.3
Avg. Iou	52.7% (11)	28.7% (12)	5.9% (12)	9.4% (9)	25.3% (4)	21.9% (1)	4.2% (11)	15.9% (8)	27.8% (3)	9.2% (10)	23.8% (6)	6.3% (11)	
UF (2.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	95%	100%	100%
FF CAIT (4.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (3.9)	100%	100%	100%	100%	100%	100%	100%	99%	<1%	100%	<1%	100%	
Eff S (3.4)	100%	100%	100%	100%	100%	100%	100%	99%	<1%	100%	<1%	100%	
RD++ WR50 (6.6)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
PC WR101 (7.3)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
PDM WR50 (8.1)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
PF R18 (8.4)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
PC WR50 (9.5)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
SN WR50 (10.2)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	
FF FNF (10.2)	100%	100%	100%	100%	100%	100%	100%	<1%	<1%	100%	<1%	100%	

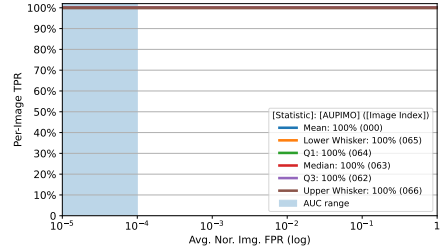
(a) Statistics and pairwise statistical tests.



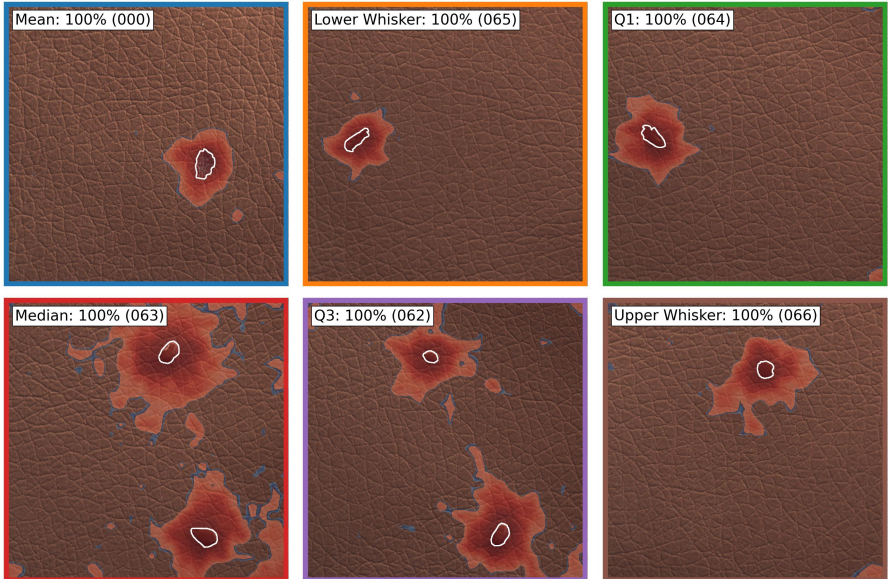
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

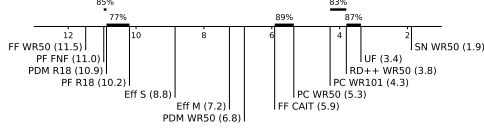


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

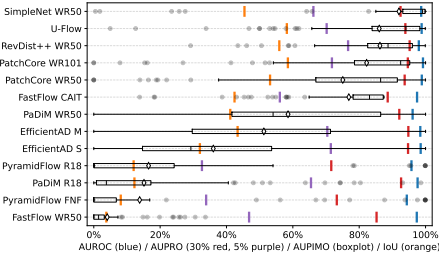
Figure 23: Benchmark on MVTec AD / Leather. PIMO curves and heatmaps are from U-Flow. 124 images (032 normal, 092 anomalous).

	FF WR50	PF FNF	PDM R18	PF R18	Eff S	Eff M	PDM WR50	FF CAIT	PC WR50	PC WR101	RD++ WR50	Uf	SN WR50
AUDROC	97.4% (10)	94.3% (13)	97.6% (8)	95.8% (12)	98.5% (5)	98.4% (6)	96.1% (11)	97.4% (9)	98.9% (2)	99.2% (1)	98.0% (7)	98.8% (3)	98.7% (4)
AUPRO	85.2% (11)	73.2% (12)	82.6% (12)	71.6% (13)	94.7% (3)	94.8% (2)	92.1% (9)	88.6% (10)	93.7% (6)	94.7% (4)	95.0% (1)	93.9% (5)	92.5% (8)
AUPRO 5%	46.8% (9)	33.9% (10)	55.5% (11)	41.4% (13)	70.4% (4)	70.4% (4)	61.1% (10)	56.1% (10)	61.1% (10)	61.1% (10)	61.1% (10)	61.1% (10)	61.1% (10)
Avg. AUPIMO	4.1% (13)	13.8% (12)	15.1% (11)	16.5% (10)	36.0% (9)	51.3% (8)	58.6% (7)	76.9% (5)	75.1% (6)	82.2% (4)	86.3% (2)	86.0% (3)	92.0% (1)
Std. AUPIMO	7.5%	28.3%	23.3%	27.0%	28.6%	48.2%	28.3%	16.3%	25.1%	22.9%	13.5%	19.1%	19.4%
P33 AUPIMO	0.1% (13)	0.0% (13)	2.1% (10)	0.0% (12)	16.8% (9)	36.9% (8)	44.8% (7)	81.0% (5)	74.5% (6)	84.1% (4)	85.8% (3)	89.5% (2)	97.5% (1)
Avg. Rank	11.3	11.0	10.9	10.2	8.8	7.2	6.8	5.9	5.3	4.3	3.8	3.4	1.9
Avg. ICI	3.2% (13)	8.1% (12)	12.2% (10)	12.2% (11)	32.0% (9)	43.4% (8)	51.3% (7)	81.0% (5)	74.5% (6)	84.1% (4)	85.8% (3)	89.5% (2)	97.5% (1)
SN WR50 (1.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Uf (1.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (1.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	84%	87%	100%
PC WR101 (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (5.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (5.9)	100%	100%	100%	100%	100%	100%	100%	100%	89%	100%	100%	100%	100%
PDM WR50 (6.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (7.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (10.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (10.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (11.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

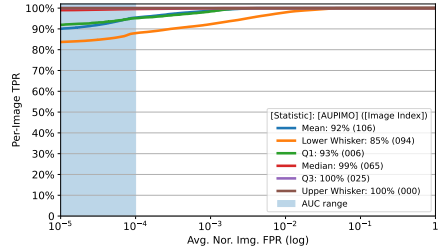
(a) Statistics and pairwise statistical tests.



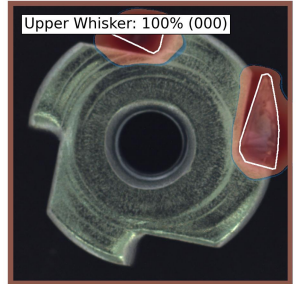
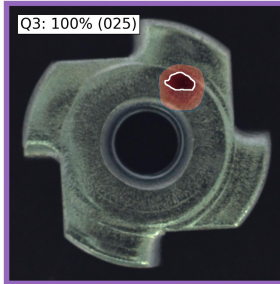
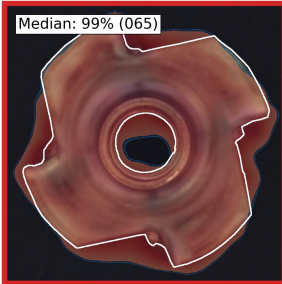
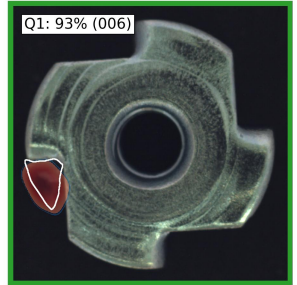
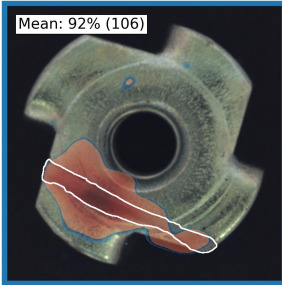
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

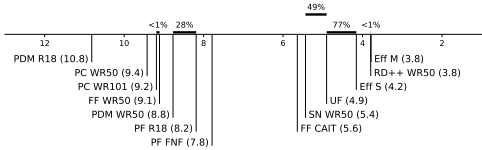


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

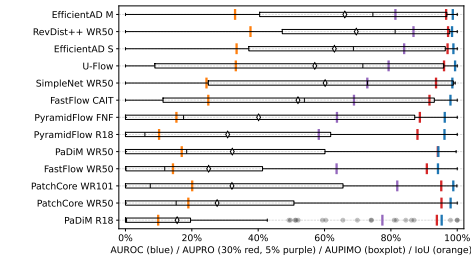
Figure 24: Benchmark on MVTEC AD / Metal Nut. PIMO curves and heatmaps are from SimpleNet WR50. 115 images (022 normal, 093 anomalous).

	PDM R18	PC WR50	PC WR101	FF WR50	PDM WR50	PF R18	PF FNF	FF CAIT	SN WR50	UF	Eff S	RD++ WR50	Eff M
AUROC	95.3% (11)	98.0% (7)	98.8% (3)	94.2% (13)	94.3% (12)	96.2% (10)	96.2% (9)	98.0% (8)	99.3% (5)	99.4% (1)	98.8% (2)	98.4% (6)	98.7% (4)
AUPRO	93.3% (8)	95.5% (6)	95.2% (6)	90.8% (11)	94.1% (7)	88.0% (13)	88.7% (12)	91.4% (10)	93.6% (9)	96.0% (4)	97.1% (2)	97.2% (3)	96.7% (3)
AUPRO 5%	97.4% (9)	97.4% (9)	97.4% (9)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)	93.9% (1)
Avg. AUPIMO	15.6% (13)	27.6% (11)	32.1% (8)	25.1% (12)	32.2% (8)	30.8% (10)	40.1% (7)	52.0% (6)	60.2% (4)	57.1% (5)	63.0% (3)	68.6% (1)	66.1% (2)
Std. AUPIMO	25.6%	31.0%	37.4%	36.1%	34.3%	40.4%	42.5%	39.0%	38.6%	39.3%	33.9%	32.6%	31.3%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	2.6% (7)	1.0% (8)	0.0% (9)	0.0% (10)	23.1% (6)	38.4% (4)	35.2% (5)	46.7% (3)	64.5% (1)	56.0% (2)
Avg. Rank	10.8	9.4	9.2	9.1	8.8	8.2	7.8	5.4	3.4	4.9	4.2	3.8	3.8
Avg. Iou	5.9% (13)	18.9% (8)	20.1% (7)	14.4% (11)	17.0% (9)	10.3% (12)	15.4% (10)	24.4% (6)	31.3% (3)	31.3% (3)	31.3% (3)	31.3% (3)	31.3% (3)
Eff M (3.8)	100%	100%	100%	100%	100%	100%	100%	100%	96%	96%	100%	<1%	<1%
RD++ WR50 (3.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (5.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (5.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (7.8)	100%	100%	99%	100%	98%	98%	100%	100%	100%	100%	100%	100%	100%
PF R18 (8.2)	100%	100%	99%	100%	98%	98%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (8.8)	100%	100%	99%	100%	99%	99%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (9.1)	100%	100%	98%	100%	98%	98%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (9.2)	100%	100%	98%	100%	98%	98%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (9.4)	100%	100%	98%	100%	98%	98%	100%	100%	100%	100%	100%	100%	100%

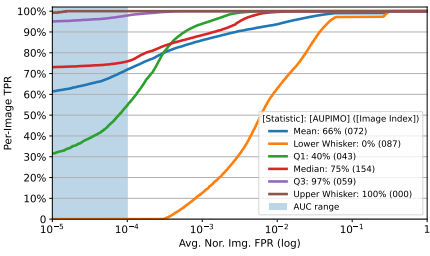
(a) Statistics and pairwise statistical tests.



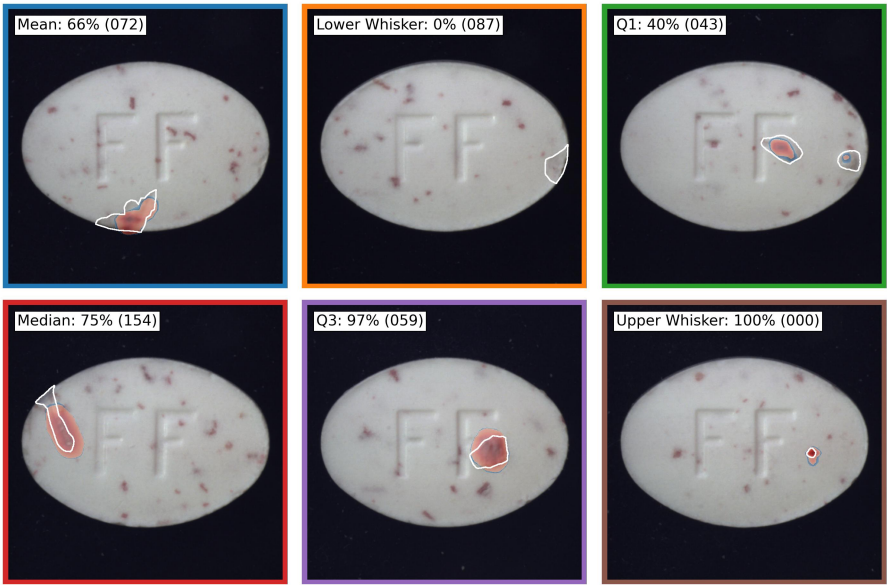
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

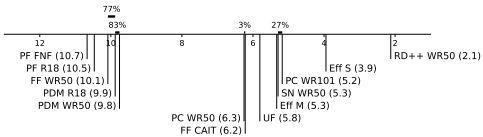


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

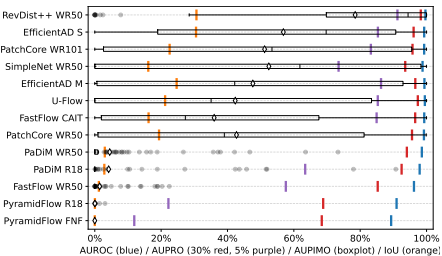
Figure 25: Benchmark on MVTec AD / Pill. PIMO curves and heatmaps are from EfficientAD M. 167 images (026 normal, 141 anomalous).

	PF FNF	PF R18	FF WR50	PDM R18	PDM WR50	PC WR50	FF CAIT	UF	Eff M	SN WR50	PC WR101	Eff S	RD++ WR50
AUROC	89.3% (13)	91.0% (12)	96.2% (11)	97.9% (10)	98.6% (9)	99.2% (7)	99.3% (4)	99.3% (2)	99.4% (3)	98.8% (8)	99.3% (6)	99.3% (5)	99.7% (1)
AUPRO	68.6% (13)	68.8% (12)	85.3% (11)	92.5% (10)	94.1% (8)	95.6% (7)	96.6% (4)	97.4% (2)	96.4% (3)	93.6% (9)	95.6% (6)	96.1% (5)	98.3% (1)
AUPRO 5%	11.9% (13)	2.2% (12)	57.6% (11)	63.3% (10)	72.1% (8)	78.7% (7)	85.3% (4)	86.3% (2)	86.3% (3)	81.2% (9)	83.2% (6)	83.3% (5)	91.2% (1)
Avg. AUPIMO	0.6% (13)	0.6% (12)	1.5% (11)	4.2% (10)	4.3% (9)	42.7% (6)	36.0% (8)	42.3% (7)	47.8% (5)	52.5% (3)	51.2% (4)	56.8% (2)	78.5% (1)
Std. AUPIMO	10.6%	0.3%	4.3%	14.5%	12.1%	38.7%	35.3%	39.3%	42.0%	42.9%	40.3%	37.0%	30.2%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	5.5% (8)	5.7% (6)	6.1% (5)	5.5% (7)	12.0% (4)	23.7% (3)	37.9% (2)	77.3% (1)
Avg. Rank	10.7	10.5	10.1	9.9	9.8	6.2	6.2	5.8	5.3	5.3	5.2	5.9	2.1
Avg. Iou	0.0% (13)	0.0% (12)	1.2% (11)	2.9% (10)	3.0% (9)	19.4% (6)	10.2% (7)	21.2% (5)	24.7% (3)	16.1% (8)	22.6% (4)	35.5% (2)	50.6% (1)
RD++ WR50 (2.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (3.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (5.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	89%	99%	100%	100%
SN WR50 (5.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	93%	100%	100%	100%
UF (5.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	89%	100%	100%	100%
Eff M (5.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	93%	100%	100%	100%
PC WR50 (6.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (6.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (9.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (9.8)	100%	100%	100%	100%	84%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (10.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (10.5)	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

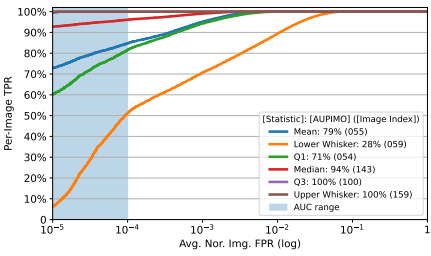
(a) Statistics and pairwise statistical tests.



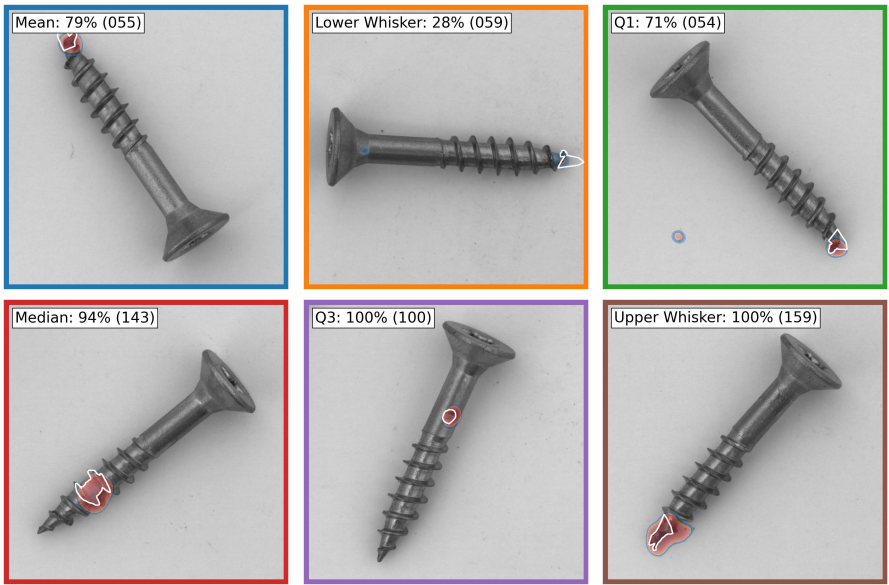
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

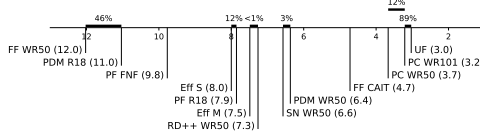


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

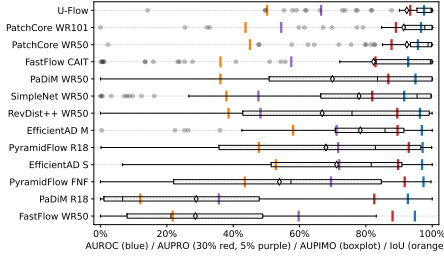
Figure 26: Benchmark on MVtec AD / Screw. PIMO curves and heatmaps are from RevDist++ WR50. 160 images (041 normal, 119 anomalous).

	FF WR50	PDM R18	PF FNF	Eff S	PF R18	Eff M	RD++ WR50	SN WR50	PDM WR50	FF CAIT	PC WR50	PC WR101	UF
AUROC	94.8% (19)	92.7% (12)	97.4% (2)	96.9% (4)	97.1% (3)	98.9% (5)	96.3% (7)	91.3% (13)	94.9% (9)	92.7% (11)	95.7% (8)	96.8% (6)	97.5% (1)
AUPRO	88.9% (18)	82.5% (12)	91.7% (1)	89.7% (4)	92.9% (2)	89.5% (5)	89.5% (6)	81.9% (13)	86.8% (10)	82.9% (11)	87.7% (9)	89.4% (7)	93.4% (1)
AUPRO 5%	89.7% (19)	85.7% (11)	95.5% (4)	92.4% (3)	94.9% (2)	91.5% (5)	89.5% (6)	87.5% (10)	91.5% (8)	87.5% (9)	94.5% (7)	95.5% (1)	95.4% (1)
Avg. AUPIMO	28.6% (13)	28.9% (12)	53.9% (11)	71.2% (7)	68.0% (9)	76.3% (5)	66.8% (10)	77.9% (6)	69.9% (8)	82.3% (4)	92.4% (1)	91.4% (3)	92.3% (2)
Std. AUPIMO	24.0%	21.4%	9.4%	8.2%	8.5%	33.3%	20.8%	11.8%	11.5%	11.3%	13.1%	17.8%	15.7%
P33 AUPIMO	11.3% (12)	2.2% (13)	33.6% (11)	64.3% (7)	63.4% (8)	76.3% (5)	50.7% (10)	82.3% (5)	55.4% (9)	94.7% (4)	96.7% (1)	95.1% (3)	97.8% (1)
Avg. Rank	12.8	11.0	9.8	8.8	7.9	7.5	7.3	6.8	6.4	4.7	3.7	3.2	3.0
Avg. IoU	21.8% (12)	12.5% (13)	43.3% (7)	52.8% (12)	47.8% (11)	58.1% (1)	48.0% (1)	37.9% (9)	38.1% (11)	30.2% (10)	43.7% (5)	50.2% (3)	50.2% (1)
UF (3.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (3.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (1.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (4.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RevDist++ WR50 (6.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EfficientAD M (6.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PyramidFlow R18 (7.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (7.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (7.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (7.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (8.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (9.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (11.0)	46%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

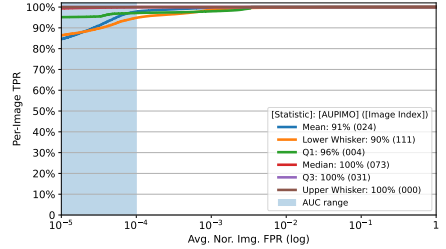
(a) Statistics and pairwise statistical tests.



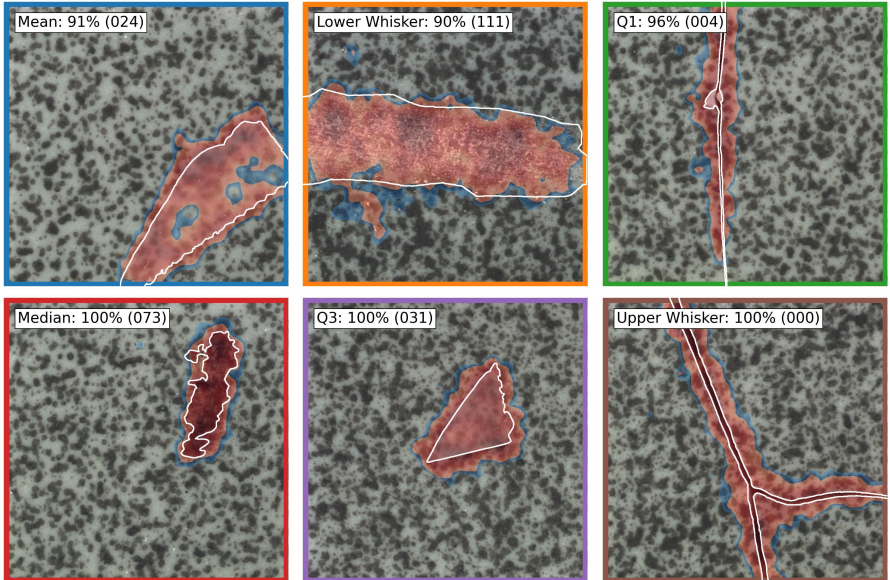
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

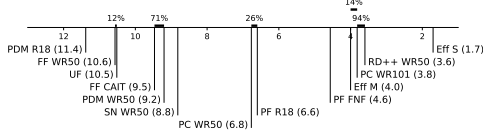


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

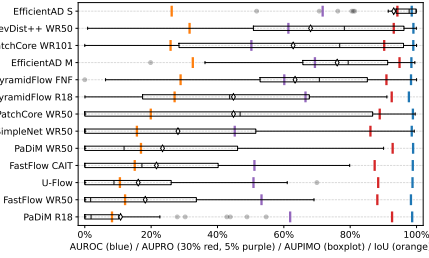
Figure 27: Benchmark on MVTec AD / Tile. PIMO curves and heatmaps are from U-Flow. 117 images (033 normal, 084 anomalous).

	PDM R18	FF WR50	UF	FF CAIT	PDM WR50	SN WR50	PC WR50	PF R18	PF FNF	Eff M	PC WR101	RD++ WR50	Eff S
AUDROC	98.7% (17)	98.3% (12)	98.8% (6)	98.3% (5)	99.0% (13)	98.6% (8)	98.9% (4)	97.7% (13)	98.3% (11)	98.4% (10)	99.4% (12)	99.1% (1)	98.6% (9)
AUPRO	92.6% (15)	88.2% (11)	88.4% (10)	87.4% (12)	92.8% (4)	86.3% (13)	86.9% (9)	82.5% (6)	90.4% (17)	94.9% (13)	90.3% (8)	93.1% (3)	94.2% (2)
AUPRO 5%	83.4% (4)	51.3% (1)	50.5% (10)	51.1% (13)	92.8% (4)	45.2% (13)	45.2% (13)	46.5% (13)	60.3% (17)	69.3% (12)	60.3% (8)	81.4% (3)	81.7% (1)
Avg. AUPIMO	10.8% (13)	18.2% (11)	16.1% (12)	21.6% (10)	23.4% (9)	28.1% (8)	44.8% (6)	44.8% (7)	83.4% (4)	76.1% (2)	62.8% (5)	68.1% (3)	93.1% (1)
Std. AUPIMO	16.4%	24.0%	21.0%	22.8%	29.2%	34.6%	41.1%	29.2%	28.9%	20.6%	36.2%	31.3%	11.0%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.9% (7)	0.0% (10)	0.0% (9)	0.0% (8)	29.6% (4)	55.0% (4)	73.7% (2)	46.0% (5)	56.2% (3)	95.4% (1)
Avg. Rank	11.4	10.6	10.5	9.5	9.2	8.8	6.8	6.6	4.6	4.0	3.8	3.6	1.7
Avg. ICI	8.2% (13)	12.1% (11)	10.6% (12)	15.1% (10)	16.9% (8)	15.4% (9)	15.9% (7)	27.2% (13)	28.9% (13)	32.9% (11)	25.0% (6)	31.4% (3)	24.2% (1)
Eff S (1.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (3.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (3.8)	100%	100%	100%	100%	100%	100%	100%	100%	99%	55%	14%	95%	100%
Eff M (4.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (4.6)	100%	100%	100%	100%	100%	100%	99%	100%	100%	99%	100%	100%	100%
PF R18 (6.6)	100%	100%	100%	100%	100%	100%	96%	27%	100%	100%	100%	100%	100%
PC WR50 (6.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (9.2)	100%	100%	100%	72%	92%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (9.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (10.5)	98%	12%	98%										
FF WR50 (10.6)	100%												

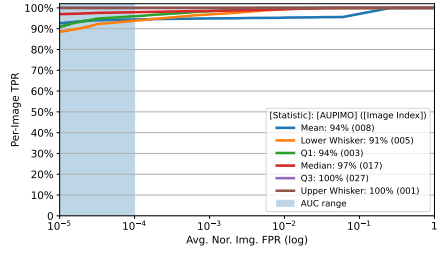
(a) Statistics and pairwise statistical tests.



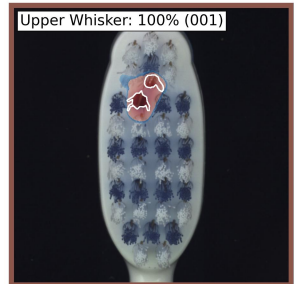
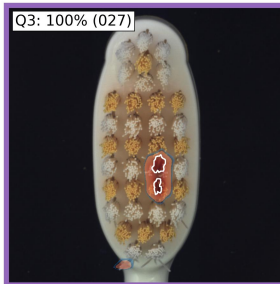
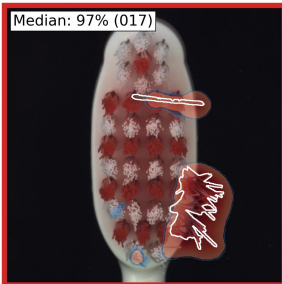
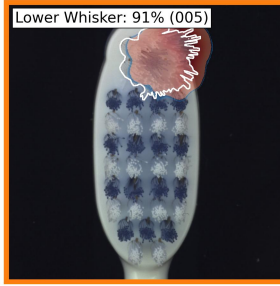
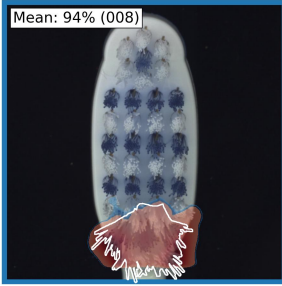
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

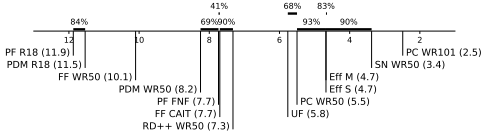


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

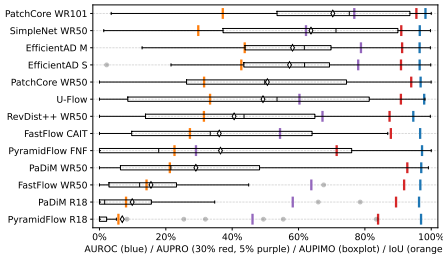
Figure 28: Benchmark on MVTec AD / Toothbrush. PIMO curves and heatmaps are from EfficientAD S. 042 images (012 normal, 030 anomalous).

	PF R18	PDM R18	FF WR50	PDM WR50	PF FNF	FF CAIT	RD++ WR50	UF	PC WR50	Eff S	Eff M	SN WR50	PC WR101
AUROC	86.3% (3)	86.3% (13)	86.7% (7)	87.0% (4)	87.2% (3)	88.7% (8)	94.6% (13)	97.9% (2)	96.8% (6)	96.5% (10)	96.5% (11)	96.3% (9)	98.3% (1)
AUPRO	84.0% (12)	89.4% (9)	91.8% (6)	92.8% (1)	71.3% (11)	87.8% (10)	87.4% (11)	90.9% (8)	94.0% (2)	90.9% (12)	91.2% (10)	91.0% (8)	95.5% (12)
AUPRO 5%	85.3% (10)	88.1% (8)	91.3% (5)	92.8% (1)	71.3% (11)	87.8% (10)	87.4% (11)	90.9% (8)	94.0% (2)	90.9% (12)	91.2% (10)	91.0% (8)	95.5% (12)
Avg. AUPIMO	8.9% (13)	9.9% (12)	15.3% (11)	28.1% (10)	36.3% (8)	36.1% (8)	40.6% (7)	49.3% (6)	50.6% (5)	57.3% (4)	58.2% (3)	63.8% (2)	70.3% (1)
Std. AUPIMO	17.6%	17.2%	15.9%	18.4%	39.3%	29.0%	27.7%	15.0%	18.7%	20.8%	18.2%	10.7%	23.9%
P33 AUPIMO	0.0% (13)	0.0% (12)	3.4% (10)	10.8% (8)	1.3% (11)	10.6% (8)	21.7% (7)	27.1% (6)	37.2% (5)	47.0% (4)	51.0% (3)	52.5% (2)	62.6% (1)
Avg. Rank	11.9	10.1	8.2	7.7	7.7	7.7	5.8	5.3	4.7	4.7	4.7	29.8% (7)	37.2% (1)
Avg. Iou	57.2% (13)	60.9% (11)	14.2% (11)	21.2% (10)	22.6% (9)	27.1% (8)	31.3% (6)	33.3% (4)	31.3% (5)	32.8% (3)	41.7%	51.9% (1)	57.2% (1)
PC WR101 (2.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%
RD WR50 (2.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%
Eff M (3.7)	100%	100%	100%	100%	100%	100%	100%	100%	93%	97%	83%	91%	
Eff S (4.7)	100%	100%	100%	100%	100%	100%	100%	100%	93%	97%	83%	91%	
PC WR50 (5.5)	100%	100%	100%	100%	99%	99%	100%	100%	91%	91%	94%	94%	
UF (5.8)	100%	100%	100%	100%	100%	100%	100%	98%	88%				
RD++ WR50 (7.3)	100%	100%	100%	100%	93%	81%	100%						
FF CAIT (7.7)	100%	100%	100%	100%	89%	42%							
PF FNF (7.7)	100%	100%	100%	100%	70%								
PDM WR50 (8.1)	100%	100%	100%	100%									
FF WR50 (10.1)	100%	100%	100%										
PDM R18 (11.5)	85%												

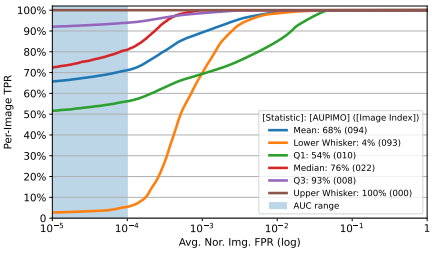
(a) Statistics and pairwise statistical tests.



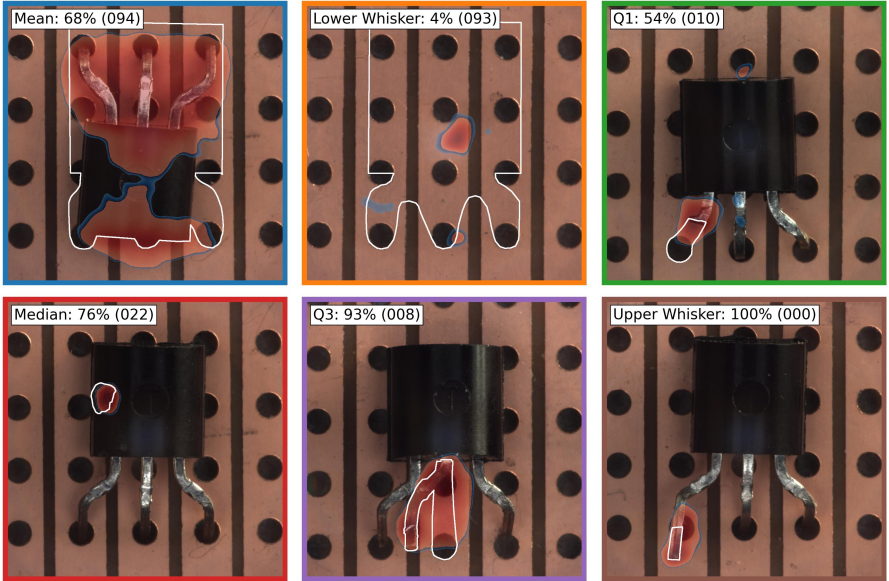
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.



(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

Figure 29: Benchmark on MVTec AD / Transistor. PIMO curves and heatmaps are from PatchCore WR101. 100 images (060 normal, 040 anomalous).

(a) Statistics and pairwise statistical tests.

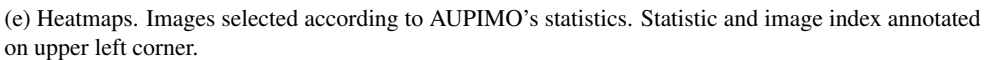
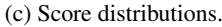
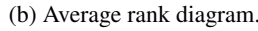
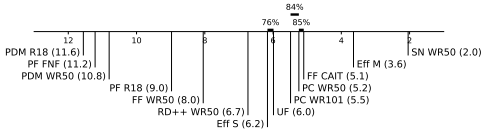


Figure 30: Benchmark on MVTec AD / Wood. PIMO curves and heatmaps are from Fast-Flow CAIT. 079 images (019 normal, 060 anomalous).

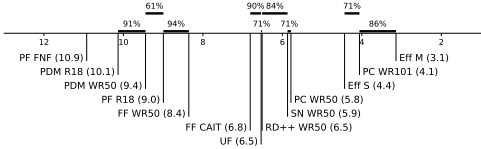
	PDM R18	PF FNF	PDM WR50	PF R18	FF WR50	RD++ WR50	Eff S	UF	PC WR101	PC WR50	FF CAT	Eff M	SN WR50
AUROC	98.2% (8)	98.3% (13)	97.4% (10)	97.2% (12)	98.2% (9)	98.4% (4)	98.4% (5)	98.6% (2)	98.4% (1)	98.3% (7)	97.2% (11)	98.4% (6)	98.5% (13)
AUPRO	93.8% (6)	94.3% (13)	92.0% (10)	88.2% (12)	93.2% (9)	95.4% (2)	93.9% (8)	95.5% (1)	94.7% (3)	93.8% (1)	90.3% (11)	93.4% (7)	94.2% (6)
AUPRO 5%	89.5% (8)	89.3% (11)	87.0% (10)	79.1% (9)	85.2% (7)	87.9% (1)	83.3% (1)	85.1% (1)	87.3% (3)	85.9% (10)	82.5% (12)	85.1% (8)	86.1% (10)
Avg. AUPIMO	5.5% (13)	7.3% (12)	10.3% (11)	14.2% (10)	20.3% (9)	31.6% (8)	35.7% (7)	38.5% (6)	41.9% (5)	42.0% (4)	44.6% (3)	53.0% (2)	75.3% (1)
Std. AUPIMO	14.9%	17.8%	21.5%	16.4%	18.0%	26.4%	24.4%	30.1%	15.3%	15.3%	15.3%	20.8%	10.0%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	3.5% (10)	7.6% (9)	18.1% (8)	21.7% (7)	17.8% (6)	14.0% (5)	16.5% (4)	12.7% (3)	39.5% (2)	69.6% (1)
Avg. Rank	11.8	10.8	9.0	8.0	6.7	4.6	3.2	2.0	1.5	1.2	0.8	0.3	0.1
Avg. Iou	21.8% (13)	5.0% (12)	6.4% (11)	12.4% (10)	15.3% (9)	23.3% (8)	29.0% (7)	29.0% (5)	30.1% (4)	27.4% (3)	23.9% (2)	40.3% (1)	43.2% (1)
SN WR50 (7.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (5.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAT (5.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (5.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (5.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (6.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (6.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (6.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (9.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (10.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (11.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

(a) Statistics and pairwise statistical tests.

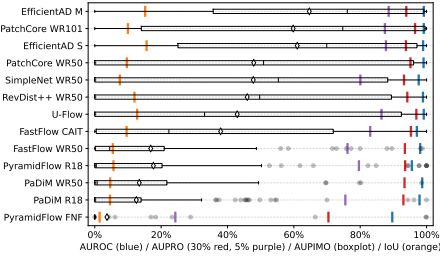


	PF FNF	PDM R18	PDM WR50	PF R18	FF WR50	FF CAIT	UF	RD++ WR50	SN WR50	PC WR50	EFF 5	PC WR101	EFF M
AUROC	89.7% (13)	97.9% (9)	98.7% (7)	95.4% (12)	98.1% (8)	97.1% (11)	99.2% (3)	98.3% (8)	97.5% (10)	99.1% (4)	99.9% (5)	99.2% (2)	99.2% (1)
AUPRO	76.4% (13)	93.0% (12)	93.4% (10)	92.5% (8)	93.5% (9)	95.3% (3)	96.0% (1)	94.2% (5)	93.2% (11)	95.2% (4)	93.8% (7)	96.5% (2)	93.9% (6)
AUPIMO	4.1% (13)	12.5% (12)	13.4% (11)	17.7% (9)	16.9% (10)	37.9% (8)	42.9% (7)	45.9% (6)	47.8% (5)	47.9% (4)	61.0% (2)	59.8% (3)	64.7% (1)
Avg. AUPIMO	3.7% (13)	12.5% (12)	13.4% (11)	17.7% (9)	16.9% (10)	37.9% (8)	42.9% (7)	45.9% (6)	47.8% (5)	47.9% (4)	61.0% (2)	59.8% (3)	64.7% (1)
Std. AUPIMO	15.8%	22.8%	20.8%	30.3%	25.1%	37.5%	42.4%	41.1%	41.4%	43.2%	37.4%	40.9%	36.9%
P31 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	1.2% (9)	4.3% (8)	0.0% (8)	0.0% (7)	3.8% (6)	0.0% (5)	45.9% (2)	32.7% (3)	51.4% (1)
Avg. Rank	10.9	10.1	9.4	9.0	8.4	6.6	6.3	6.5	5.9	5.8	4.4	4.1	3.1
Eff. IoU	1.3% (13)	4.0% (12)	4.7% (11)	3.0% (10)	3.5% (9)	9.0% (7)	12.8% (6)	12.4% (5)	7.6% (4)	9.1% (3)	15.0% (1)	10.0% (2)	15.2% (2)
Eff. F1 (3.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	71%	87%
FF (4.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (5.8)	100%	100%	100%	100%	100%	100%	99%	99%	99%	99%	100%	100%	100%
SN WR50 (5.9)	100%	100%	100%	100%	100%	100%	99%	97%	84%	72%	100%	100%	100%
RD++ WR50 (6.5)	100%	100%	100%	100%	100%	99%	72%	72%	72%	72%	100%	100%	100%
UF (6.5)	100%	100%	100%	100%	100%	100%	90%	72%	72%	72%	100%	100%	100%
FF CAIT (6.8)	100%	100%	100%	100%	100%	100%							
FF WR50 (8.4)	100%	100%	97%	94%	100%								
PF R18 (9.0)	100%	100%	90%	61%									
PDM WR50 (9.4)	100%		91%										
PDM R18 (10.1)	100%												

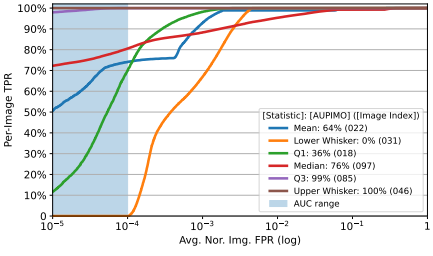
(a) Statistics and pairwise statistical tests.



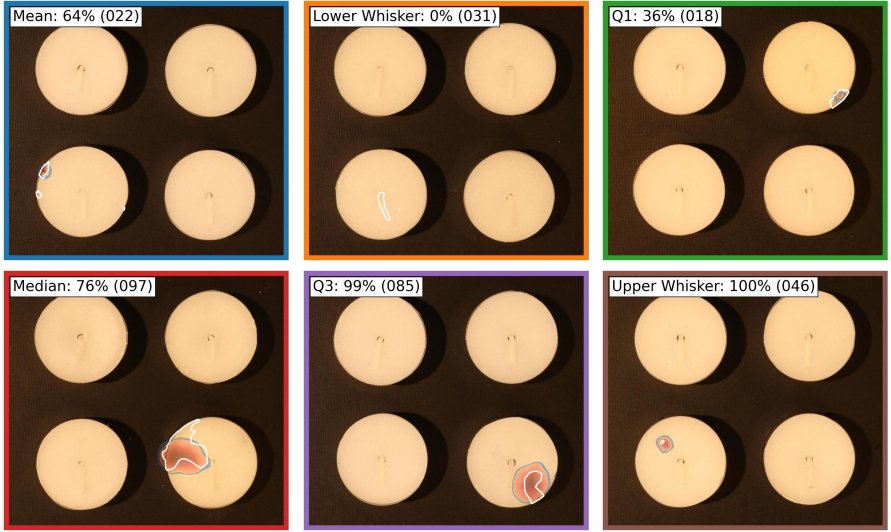
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

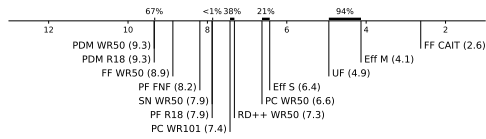


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

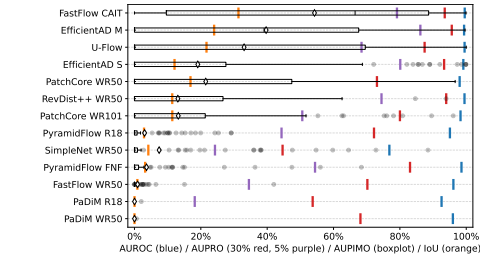
Figure 32: Benchmark on VisA / Candle. PIMO curves and heatmaps are from EfficientAD M. 200 images (100 normal, 100 anomalous).

	PDM WR50	PDM R18	FF WR50	PF FNF	SN WR50	PF R18	PC WR101	RD++ WR50	PC WR50	EFF S	UF	EFF M	FF CAIT
AUROC	99.0% (10)	92.4% (12)	96.1% (9)	96.6% (6)	76.8% (13)	95.1% (11)	98.3% (7)	99.5% (3)	98.0% (8)	99.1% (5)	99.4% (4)	99.5% (1)	99.5% (1)
AUPRO	69.1% (11)	53.7% (12)	70.2% (10)	83.1% (6)	44.6% (13)	72.2% (9)	80.6% (7)	94.0% (2)	73.1% (8)	83.3% (4)	87.5% (5)	95.4% (1)	93.5% (3)
AUPRO++	18.3% (11)	6.4% (12)	14.4% (10)	24.4% (6)	2.2% (13)	14.3% (9)	30.5% (7)	90.5% (2)	23.1% (8)	53.1% (4)	63.8% (5)	86.6% (1)	79.1% (3)
Avg. AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	3.6% (9)	7.4% (8)	3.1% (10)	13.2% (6)	13.1% (7)	21.5% (4)	19.1% (5)	31.0% (3)	39.7% (2)	54.3% (1)
Std. AUPIMO	0.1%	0.2%	4.5%	31.0%	17.4%	6.6%	24.1%	21.8%	32.0%	32.2%	39.8%	37.9%	38.2%
P13 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (8)	0.0% (7)	0.0% (6)	0.0% (5)	0.0% (4)	0.0% (3)	0.0% (2)	26.5% (1)
Avg. Rank	9.3	9.3	8.9	8.2	7.9	7.9	7.4	7.3	6.6	6.4	4.9	4.1	2.6
PC WR101	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (2.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EFF M (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EFF S (6.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (6.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (7.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (7.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (7.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (8.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (9.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

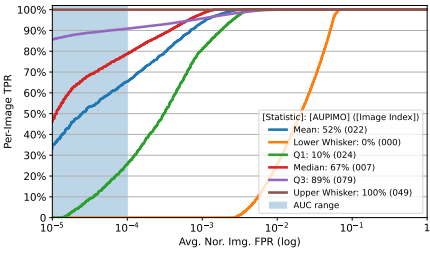
(a) Statistics and pairwise statistical tests.



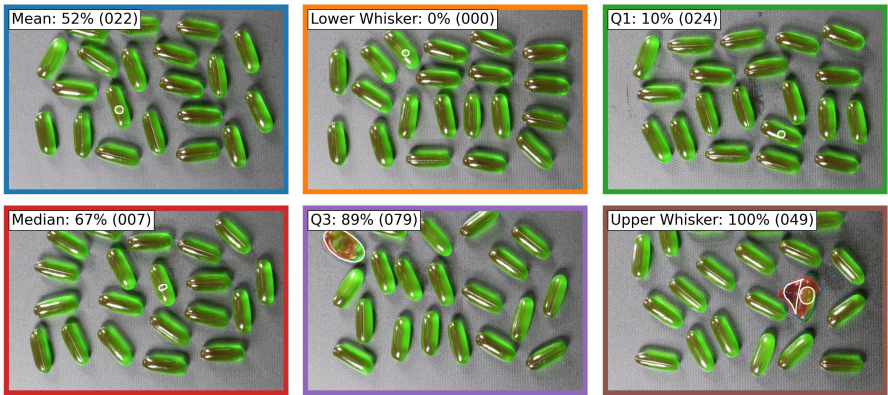
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

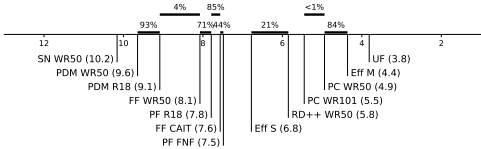


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

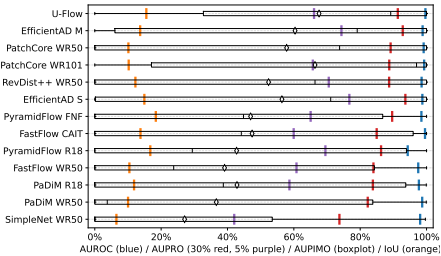
Figure 33: Benchmark on VisA / Capsules. PIMO curves and heatmaps are from FastFlow CAIT. 160 images (060 normal, 100 anomalous).

	SN WR50	PDM WR50	PDM R18	FF WR50	FF R18	FF CAIT	FF FNF	Eff S	RD++ WR50	PC WR101	PC WR50	Eff M	UF
AUROC	98.1% (10)	98.7% (7)	97.7% (11)	97.5% (12)	94.3% (13)	99.5% (2)	98.5% (9)	98.7% (6)	98.5% (8)	99.3% (3)	99.1% (4)	98.8% (5)	99.7% (1)
AUPRO	73.7% (11)	82.3% (12)	82.8% (11)	84.0% (10)	86.3% (8)	85.0% (9)	86.8% (4)	82.6% (11)	88.7% (7)	88.8% (6)	89.2% (5)	92.8% (1)	81.3% (3)
AUPRO 5%	42.2% (11)	52.9% (12)	52.9% (11)	50.8% (13)	68.3% (4)	69.0% (3)	69.3% (7)	76.7% (11)	76.5% (10)	85.7% (1)	85.7% (2)	73.3% (13)	64.3% (10)
Avg. AUPIMO	27.1% (13)	36.6% (12)	42.9% (9)	39.1% (11)	42.7% (10)	47.4% (7)	47.0% (8)	56.4% (5)	52.4% (6)	66.4% (2)	57.8% (4)	60.4% (3)	67.6% (1)
Std. AUPIMO	16.0%	42.0%	43.0%	40.7%	41.5%	44.1%	40.0%	42.7%	44.2%	41.6%	42.8%	43.0%	39.9%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	3.8% (7)	0.3% (9)	7.6% (6)	19.2% (5)	2.1% (8)	62.5% (1)	21.6% (4)	22.4% (3)	59.2% (2)
Avg. Rank	10.2	9.6	9.1	8.1	7.8	7.6	7.5	6.8	5.8	4.5	4.4	4.4	3.8
Mean (001)	6.7% (1)	10.0% (12)	11.8% (8)	10.5% (9)	16.7% (2)	17.1% (5)	10.4% (1)	24.9% (4)	12.5% (7)	10.2% (10)	10.1% (11)	17.7% (6)	12.2% (1)
UF (1.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (4.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (5.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (5.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (6.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF FNF (7.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (7.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF R18 (7.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (9.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (9.6)	99%	94%											

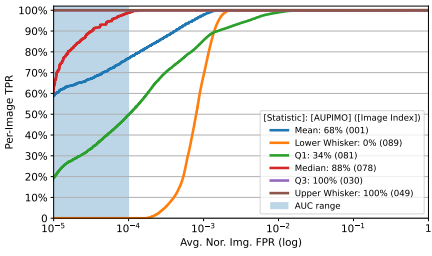
(a) Statistics and pairwise statistical tests.



(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

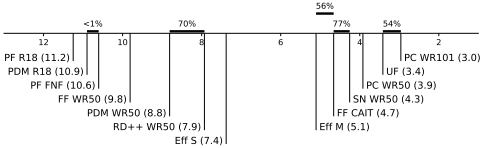


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

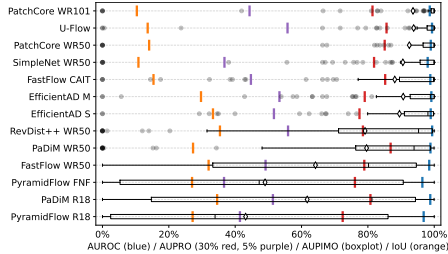
Figure 34: Benchmark on VisA / Cashew. PIMO curves and heatmaps are from U-Flow. 150 images (050 normal, 100 anomalous).

	PF R18	PDM R18	PF FNF	FF WR50	PDM WR50	RD++ WR50	Eff S	Eff M	FF CAIT	SN WR50	PC WR50	UF	PC WR101
AUDOC	95.8% (12)	98.8% (8)	98.4% (13)	98.3% (10)	98.9% (5)	99.4% (1)	98.9% (3)	99.1% (1)	98.8% (7)	98.0% (11)	92.0% (4)	95.4% (2)	98.8% (9)
AUPRO	72.4% (13)	80.8% (7)	76.3% (12)	76.9% (9)	80.9% (1)	78.5% (10)	77.5% (13)	79.1% (8)	85.2% (3)	81.3% (15)	85.1% (6)	85.4% (2)	83.4% (16)
AUPRO SN	51.4% (9)	51.4% (9)	58.6% (11)	49.2% (16)	55.9% (3)	53.7% (2)	53.3% (3)	53.3% (3)	44.7% (17)	36.9% (16)	53.8% (12)	53.8% (12)	44.3% (18)
Avg. AUPIMO	43.2% (13)	61.7% (11)	49.9% (12)	64.3% (10)	79.6% (8)	79.1% (9)	89.3% (6)	90.6% (4)	88.1% (7)	90.5% (5)	92.5% (3)	92.8% (1)	92.8% (2)
Std. AUPIMO	39.2%	38.9%	40.2%	35.2%	30.3%	30.9%	20.4%	19.6%	24.0%	23.2%	22.0%	19.0%	20.6%
P33 AUPIMO	7.9% (13)	51.1% (10)	9.7% (12)	49.3% (11)	84.4% (8)	84.4% (8)	95.3% (7)	96.5% (5)	96.2% (6)	98.0% (3)	97.9% (4)	99.7% (2)	100.0% (1)
Avg. Rank	11.2	10.9	10.4	9.1	8.1	7.9	7.4	5.1	4.7	4.3	3.9	3.4	3.0
Avg. IOM	27.2% (7)	34.0% (12)	27.0% (8)	32.0% (14)	27.3% (6)	35.4% (1)	33.3% (1)	29.7% (3)	15.4% (18)	10.6% (12)	14.1% (10)	13.6% (11)	10.4% (13)
PC WR101 (13.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (3.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (3.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (4.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (4.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (5.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (7.4)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (7.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (8.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM WR50 (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (10.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PDM R18 (10.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

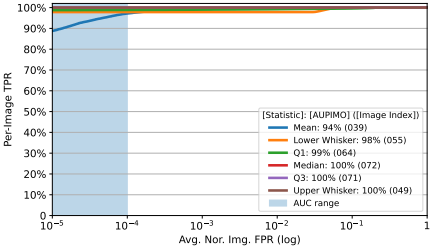
(a) Statistics and pairwise statistical tests.



(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

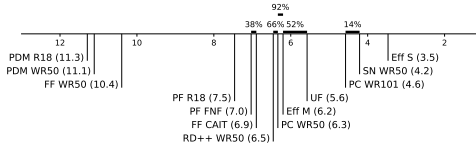


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

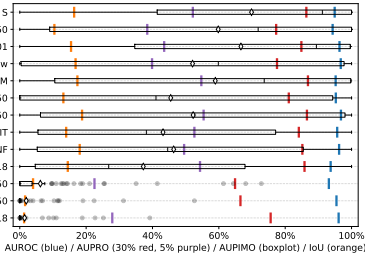
Figure 35: Benchmark on VisA / Chewing Gum. PIMO curves and heatmaps are from PatchCore WR101. 150 images (050 normal, 100 anomalous).

	PDM R18	PDM WR50	FF WR50	PF R18	PF FNF	FF CAIT	RD++ WR50	PC WR50	EFF M	UF	PC WR101	SN WR50	EFF S
AUDROC	86.3% (1)	95.5% (7)	81.2% (13)	93.7% (12)	96.2% (4)	95.7% (6)	96.8% (1)	85.3% (8)	95.2% (9)	96.7% (2)	96.4% (3)	94.3% (11)	94.9% (10)
AUPRO	72.8% (11)	66.5% (12)	65.9% (13)	85.3% (1)	85.2% (1)	84.5% (1)	86.8% (2)	81.3% (10)	86.5% (1)	77.2% (10)	84.9% (6)	77.2% (10)	84.4% (13)
AUPRO 9%	27.9% (10)		22.5% (11)	54.3% (1)	59.5% (1)	52.4% (4)	55.4% (1)	44.3% (2)	54.3% (2)	39.9% (1)	43.5% (7)	38.4% (1)	52.2% (1)
Avg. AUPIMO	1.5% (13)	2.0% (12)	6.2% (11)	37.2% (10)	46.4% (7)	43.2% (9)	52.3% (5)	45.5% (8)	58.9% (4)	52.1% (6)	66.7% (2)	59.8% (3)	69.9% (1)
Std. AUPIMO	5.6%	7.0%	14.5%	34.9%	37.2%	36.4%	40.1%	42.0%	41.2%	41.4%	40.0%	40.9%	36.8%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	8.4% (8)	14.5% (7)	14.3% (8)	22.6% (5)	1.4% (10)	31.6% (4)	17.4% (6)	61.5% (2)	37.1% (3)	63.5% (1)
Avg. Rank	11.1	11.1	10.4	7.5	7.9	6.9	6.5	6.3	5.5	4.5	4.2	4.2	3.5
Avg. Iou	1.3% (13)	1.6% (12)	2.0% (11)	14.5% (7)	18.1% (2)	14.0% (8)	18.8% (1)	13.1% (10)	17.2% (3)	16.8% (4)	19.5% (6)	10.4% (10)	16.4% (1)
EFF S (1.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	53%	97%	
SN WR50 (4.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	35%		
PC WR101 (4.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
UF (5.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
EFF M (6.2)	100%	100%	100%	100%	100%	100%	96%	92%	92%				
PC WR50 (6.3)	100%	100%	100%	98%	79%	94%	66%						
RD++ WR50 (6.5)	100%	100%	100%	100%	100%	97%							
FF CAIT (6.9)	100%	100%	100%	98%	39%	38%							
PF FNF (7.0)	100%	100%	100%	100%									
PF CAIT (6.9)	100%	100%	100%	100%									
PF R18 (7.5)	100%	100%	100%										
FF WR50 (10.4)	100%	100%											
PDM WR50 (11.1)	100%												

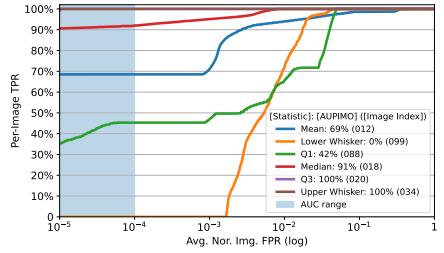
(a) Statistics and pairwise statistical tests.



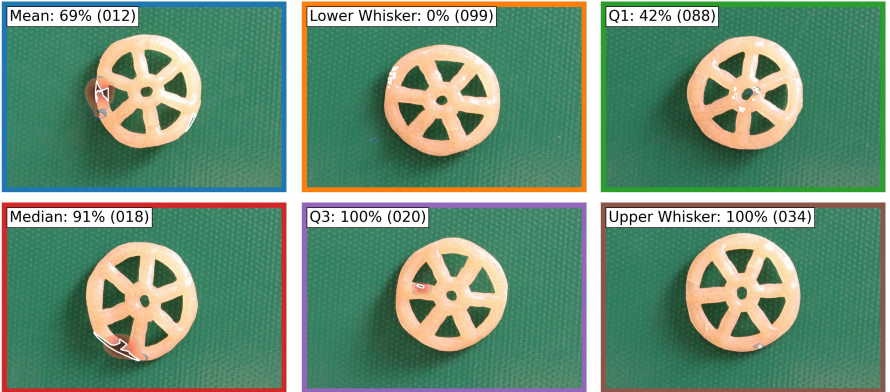
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

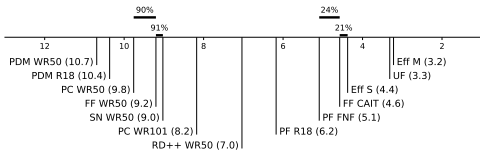


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

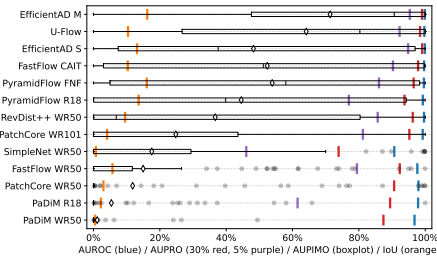
Figure 36: Benchmark on VisA / Fryum. PIMO curves and heatmaps are from EfficientAD S. 150 images (050 normal, 100 anomalous).

	PDM WR50	PDM R18	PC WR50	FF WR50	SN WR50	PC WR101	RD++ WR50	PF R18	PF FNF	PF CAIT	EFF S	UF	EF M
AUDOC	96.8% (12)	97.3% (10)	98.0% (9)	97.6% (11)	98.7% (8)	99.2% (6)	99.4% (5)	99.3% (7)	99.0% (8)	99.7% (4)	99.9% (2)	99.7% (3)	99.9% (1)
AUPRO	87.2% (11)	89.4% (11)	90.4% (10)	92.3% (9)	92.6% (13)	93.2% (17)	94.3% (18)	93.7% (16)	96.3% (15)	97.8% (13)	99.1% (11)	98.4% (13)	99.9% (12)
ASUPRO	81.5% (13)	81.5% (13)	82.4% (12)	85.4% (10)	85.9% (11)	87.3% (14)	87.9% (15)	87.9% (15)	88.3% (15)	89.3% (14)	89.8% (12)	92.3% (13)	93.3% (11)
Avg. AUPIMO	11.1% (13)	15.4% (12)	11.8% (11)	27.2% (9)	15.0% (10)	17.6% (9)	24.8% (6)	16.3% (7)	44.9% (6)	53.9% (3)	92.4% (4)	48.2% (5)	64.1% (2)
Std. AUPIMO	6.0%	15.7%	27.8%	27.2%	30.8%	39.6%	42.0%	41.9%	42.1%	41.9%	40.3%	39.3%	34.7%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (8)	0.0% (7)	3.1% (6)	16.9% (5)	17.4% (4)	18.1% (3)	56.3% (2)	61.5% (1)
Avg. Rank	10.7	10.4	9.8	9.2	9.0	8.2	7.0	6.2	5.1	4.6	4.4	3.3	3.2
Avg. IoU	0.5% (13)	2.2% (11)	3.0% (10)	5.7% (8)	0.7% (12)	4.1% (9)	9.5% (7)	13.8% (11)	16.1% (12)	19.3% (6)	13.1% (4)	10.4% (5)	16.2% (1)
EFF M (3.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	99%
UF (3.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
EFF S (4.4)	100%	100%	100%	100%	100%	100%	100%	99%	97%	3%	21%		
PF CAIT (4.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (5.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (6.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (7.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR101 (8.2)	100%	100%	100%	100%	97%	100%							
FF WR50 (9.0)	100%	100%	100%	91%									
SN WR50 (9.2)	100%	100%	100%										
PC WR50 (9.8)	100%	100%	100%										
PDM R18 (10.4)	100%	99%											
PDM WR50 (10.7)	100%												

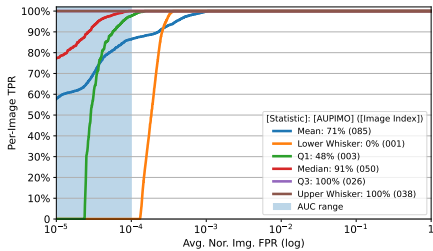
(a) Statistics and pairwise statistical tests.



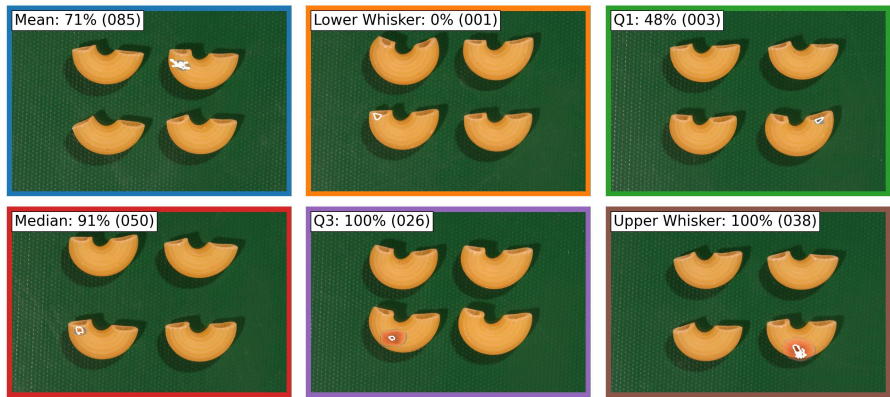
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

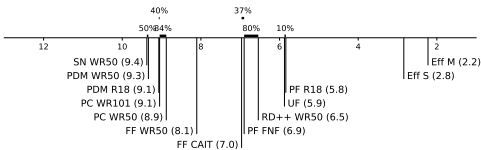


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

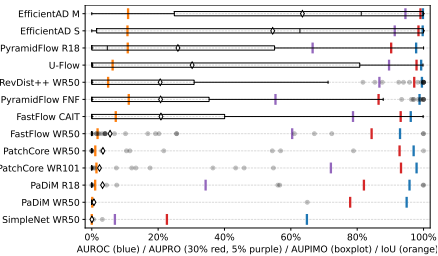
Figure 37: Benchmark on VisA / Macaroni 1. PIMO curves and heatmaps are from EfficientAD M. 200 images (100 normal, 100 anomalous).

	SN WR50	PDM WR50	PDM R18	PC WR101	PC WR50	FF WR50	FF CAIT	PF FNF	RD++ WR50	UF	PF R18	Eff S	Eff M
AUDROC	94.9% (11)	94.9% (11)	95.8% (10)	97.9% (8)	97.0% (8)	92.9% (12)	95.1% (9)	86.8% (18)	99.3% (3)	99.3% (4)	97.8% (7)	99.7% (2)	99.1% (1)
AUPRO	22.8% (13)	22.9% (12)	82.0% (11)	93.3% (5)	92.8% (7)	84.8% (10)	93.3% (6)	86.4% (18)	97.2% (4)	97.9% (3)	98.2% (8)	99.5% (2)	99.3% (1)
AUPRO 90%	7.9% (11)		38.3% (10)	73.3% (8)		65.4% (8)	78.7% (15)	55.4% (18)	88.7% (3)	88.9% (3)	86.4% (7)	91.3% (2)	94.5% (1)
Avg. AUPIMO	0.9%	0.3%	2.3% (10)	2.3% (11)	3.4% (8)	3.8% (8)	20.8% (9)	20.7% (6)	30.6% (3)	30.2% (3)	26.0% (4)	54.5% (2)	61.5% (1)
Std. AUPIMO	0.3%	6.5%	15.9%	9.2%	14.8%	18.2%	31.7%	34.8%	34.3%	41.2%	34.0%	42.8%	39.3%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (8)	0.0% (7)	0.0% (6)	0.0% (5)	0.0% (4)	0.0% (3)	16.9% (2)	48.4% (1)
Avg. Rank	9.4	9.1	8.1	3.1	8.9	8.1	7.0	6.9	6.5	5.9	5.4	2.8	2.2
Avg. IoU	0.0% (13)	0.1% (12)	1.0% (11)	1.4% (10)	1.1% (10)	1.7% (8)	7.2% (5)	11.1% (1)	4.9% (7)	6.3% (8)	10.8% (3)	10.8% (4)	10.9% (2)
Eff M (2.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (2.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (5.8)	100%	100%	100%	100%	100%	100%	100%	96%	98%	88%	100%		
UF (5.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		
RD++ WR50 (6.5)	100%	100%	100%	100%	100%	100%	100%	63%	81%	100%	10%		
PF FNF (6.9)	100%	100%	100%	100%	100%	100%	100%	37%					
PC WR101 (9.1)	100%	100%	100%	100%	100%	100%	100%						
PC WR50 (8.9)	100%	100%	100%	100%	100%	100%	100%						
FF WR50 (8.1)	100%	100%	98%	99%	99%								
PC WR101 (9.1)	100%	97%	40%	84%									
PDM R18 (9.1)	99%	95%											
PDM WR50 (9.3)	96%												

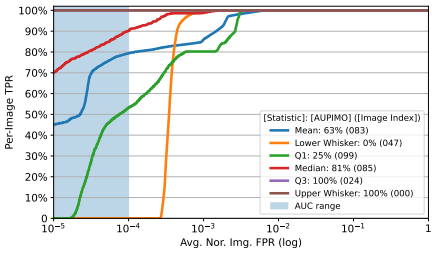
(a) Statistics and pairwise statistical tests.



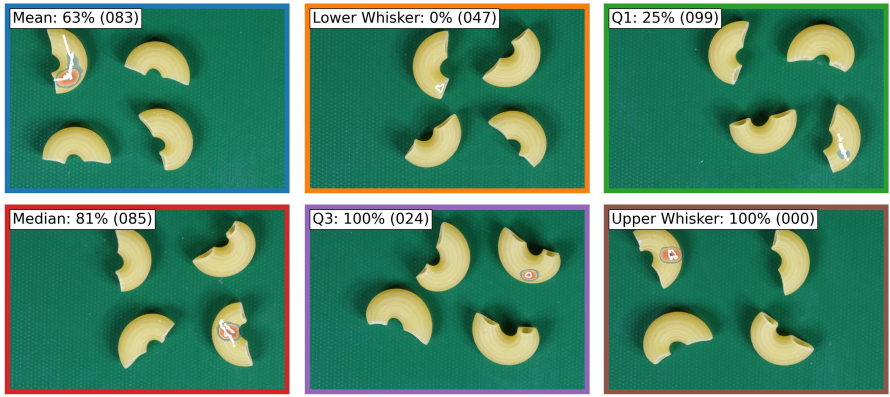
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

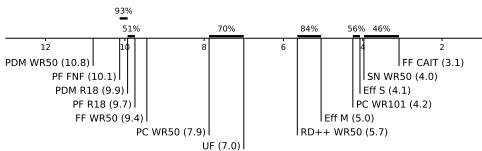


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

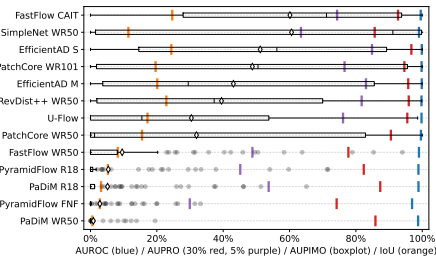
Figure 38: Benchmark on VisA / Macaroni 2. PIMO curves and heatmaps are from EfficientAD M. 200 images (100 normal, 100 anomalous).

	PDM WR50	PF FNF	PDM R18	PF R18	FF WR50	PC WR50	UF	RD++ WR50	Eff M	PC WR101	Eff S	SN WR50	FF CAIT
AUROC	98.8% (12)	96.9% (13)	98.9% (10)	98.8% (11)	99.0% (9)	99.6% (6)	99.8% (5)	99.8% (5)	99.9% (2)	99.8% (3)	99.9% (1)	99.9% (8)	99.5% (7)
AUPRO	85.9% (10)	74.2% (13)	87.3% (11)	84.3% (12)	77.7% (12)	90.6% (7)	95.3% (4)	95.9% (4)	94.6% (1)	94.6% (1)	95.7% (1)	95.7% (10)	92.7% (6)
AUPRO 5%	79.9% (11)	51.7% (10)	81.7% (10)	81.4% (10)	69.8% (1)	88.6% (6)	94.3% (3)	94.7% (3)	91.9% (2)	91.9% (2)	91.9% (2)	91.9% (9)	81.1% (5)
Avg. AUPIMO	8.9% (13)	2.8% (12)	5.2% (11)	5.4% (10)	9.8% (9)	32.0% (7)	30.4% (8)	39.5% (6)	43.1% (5)	48.8% (4)	51.2% (3)	60.6% (1)	60.1% (2)
Std. AUPIMO	3.1%	6.0%	12.3%	14.4%	19.3%	40.4%	14.3%	15.1%	16.9%	42.1%	36.9%	41.7%	15.9%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (8)	1.1% (7)	5.9% (5)	8.6% (4)	5.2% (6)	22.7% (3)	22.9% (2)	42.2% (1)
Avg. Rank	10.8	10.1	9.9	9.7	9.4	7.9	7.8	5.7	5.0	4.2	4.1	2.0	3.1
Avg. Iou	0.4% (13)	2.9% (12)	3.2% (11)	5.3% (10)	8.2% (9)	15.5% (7)	17.7% (8)	22.9% (6)	20.1% (4)	37.4% (3)	24.4% (2)	11.4% (1)	24.2% (1)
FF CAIT (3.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	46%	
SN WR50 (4.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
Eff S (4.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
PC WR101 (4.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	57%	100%	
Eff M (5.0)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
RD++ WR50 (5.7)	100%	100%	100%	100%	100%	100%	99%	100%	85%				
UF (7.0)	100%	100%	100%	100%	100%	70%							
PC WR50 (7.9)	100%	100%	100%	100%	100%								
FF WR50 (8.4)	100%	100%	98%	100%									
PF R18 (9.3)	100%	92%	81%										
PDM R18 (9.9)	100%	93%											
PF FNF (10.1)	99%												

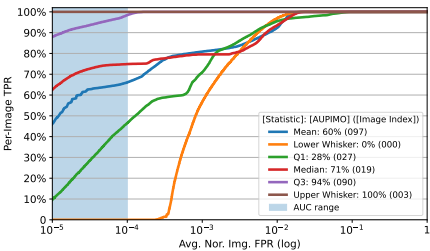
(a) Statistics and pairwise statistical tests.



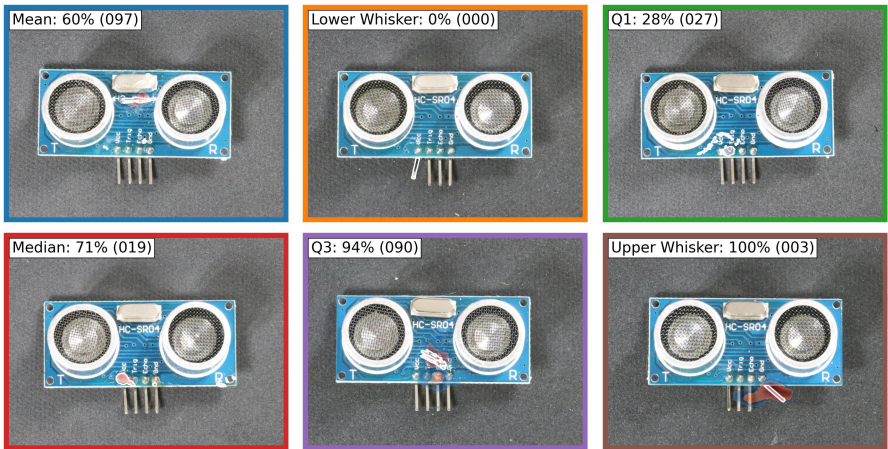
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

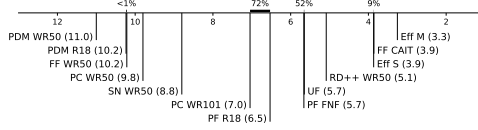


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

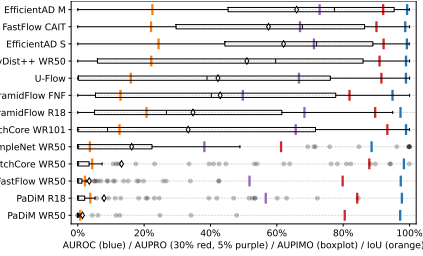
Figure 39: Benchmark on VisA / PCB 1. PIMO curves and heatmaps are from FastFlow CAIT. 200 images (100 normal, 100 anomalous).

	PDM WR50	PDM R18	FF WR50	PC WR50	SN WR50	PC WR101	PF R18	PF FNF	UF	RD++ WR50	Eff S	FF CAIT	Eff M
AUROC	97.2% (11)	97.2% (10)	97.4% (10)	98.3% (9)	98.5% (13)	98.3% (13)	97.3% (10)	94.9% (12)	98.5% (9)	98.9% (4)	99.2% (2)	99.7% (1)	99.3% (1)
AUPRO	80.3% (11)	81.2% (9)	77.8% (12)	97.5% (8)	61.3% (13)	93.3% (1)	89.4% (7)	81.3% (10)	91.3% (3)	91.0% (5)	92.2% (2)	95.0% (6)	92.0% (1)
AUPRO 30%	80.3% (11)	81.2% (9)	77.8% (12)	97.5% (8)	61.3% (13)	93.3% (1)	89.4% (7)	81.3% (10)	91.3% (3)	91.0% (5)	92.2% (2)	95.0% (6)	92.0% (1)
AUPRO 5%	1.4% (13)	7.9% (11)	3.5% (12)	13.2% (10)	16.2% (10)	13.3% (8)	12.7% (7)	42.2% (6)	42.2% (6)	51.3% (4)	62.0% (2)	97.5% (1)	66.0% (1)
std. AUPIMO	6.5%	18.2%	8.4%	27.3%	28.5%	18.4%	10.1%	35.6%	36.9%	37.9%	32.6%	33.4%	34.2%
P33 AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (8)	10.6% (6)	16.7% (5)	7.5% (7)	19.2% (4)	50.8% (2)	43.0% (3)	61.3% (1)
Avg. Rank	11.0	10.2	10.2	9.8	8.8	7.0	6.1	5.7	5.1	3.3	3.3	3.3	3.3
Avg. IoU	0.8% (13)	3.8% (10)	2.2% (12)	4.3% (9)	3.6% (11)	12.4% (8)	20.7% (5)	12.9% (7)	16.4% (6)	22.1% (3)	24.3% (1)	22.1% (4)	22.5% (2)
Eff R13.9	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (13.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%
Eff S (13.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (5.1)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
UF (5.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (5.7)	100%	100%	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%
PF R18 (6.5)	100%	100%	100%	100%	100%	100%	73%	100%	100%	100%	100%	100%	100%
PC WR101 (7.0)	100%	100%	100%	100%	100%	100%	23%	100%	100%	100%	100%	100%	100%
SN WR50 (8.8)	100%	100%	100%	100%	100%	100%	23%	100%	100%	100%	100%	100%	100%
PC WR50 (9.8)	100%	100%	100%	100%	100%	100%	23%	100%	100%	100%	100%	100%	100%
FF WR50 (10.2)	100%	100%	100%	100%	100%	100%	23%	100%	100%	100%	100%	100%	100%
PDM R18 (10.2)	100%	100%	100%	100%	100%	100%	23%	100%	100%	100%	100%	100%	100%

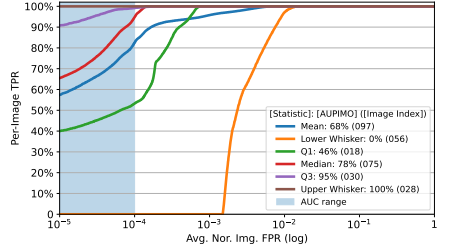
(a) Statistics and pairwise statistical tests.



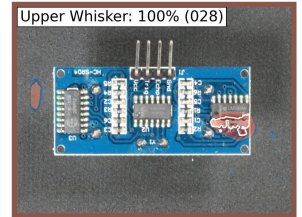
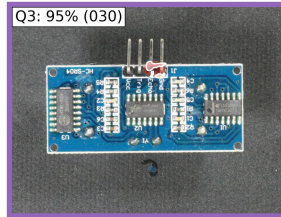
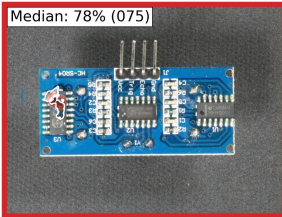
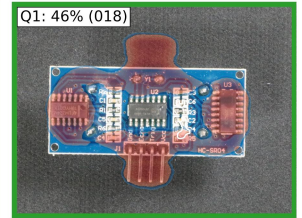
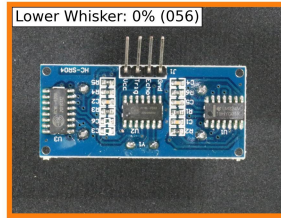
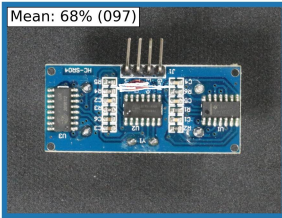
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

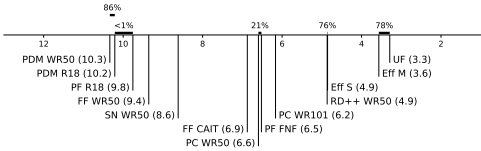


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

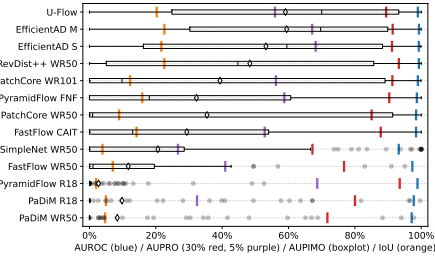
Figure 40: Benchmark on VisA / PCB 2. PIMO curves and heatmaps are from EfficientAD M. 200 images (100 normal, 100 anomalous).

	PDM WR50	PDM R18	PF R18	FF WR50	SN WR50	FF CAIT	PC WR50	PF FNF	PC WR10	RD++ WR50	Eff S	Eff M	UF
AUROC	97.3% (12)	97.4% (10)	98.0% (6)	97.3% (11)	97.3% (13)	97.4% (10)	98.4% (8)	98.7% (7)	97.3% (4)	97.3% (12)	99.3% (1)	99.3% (1)	99.0% (1)
AUPRO	71.3% (12)	80.0% (10)	93.5% (1)	76.7% (11)	67.2% (13)	87.8% (8)	85.0% (9)	90.3% (6)	91.3% (4)	91.3% (12)	91.2% (5)	91.3% (13)	89.3% (7)
AUPRO S	32.4% (8)	32.4% (8)	48.6% (1)	38.9% (10)	28.6% (10)	58.8% (4)	58.8% (4)	58.7% (4)	58.3% (4)	58.3% (4)	58.3% (13)	57.3% (10)	55.9% (15)
Acc. AUPIMO	8.4% (12)	9.8% (11)	2.7% (13)	11.7% (10)	20.6% (9)	35.4% (6)	35.4% (6)	32.2% (7)	38.3% (5)	48.3% (4)	53.2% (1)	59.4% (1)	59.1% (2)
Sid. AUPIMO	21.3%	23.4%	8.2%	19.3%	35.4%	33.8%	44.1%	34.3%	43.2%	38.5%	36.2%	35.0%	36.6%
PI AUPIMO	0.0% (13)	0.0% (12)	0.0% (11)	0.0% (10)	0.0% (9)	0.1% (6)	0.0% (8)	1.7% (5)	0.0% (7)	19.0% (4)	35.3% (1)	48.7% (1)	41.3% (2)
Acc. Rank	10.3	10.2	9.8	9.4	8.6	6.9	6.6	6.5	6.2	4.9	3.4	3.4	3.3
Acc. IoU	2.7% (11)	3.0% (10)	2.0% (13)	7.0% (9)	3.9% (12)	14.2% (6)	8.9% (8)	15.9% (5)	12.3% (7)	22.6% (1)	21.7% (15)	22.6% (12)	20.3% (4)
UF 12.1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff M (3.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Eff S (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RD++ WR50 (4.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR10 (6.2)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (6.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF FNF (6.5)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SN WR50 (8.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF CAIT (6.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PC WR50 (6.6)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FF WR50 (8.4)	99%	96%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PF R18 (9.8)	<1%	<1%	100%										
PF R18 (10.2)	87%												

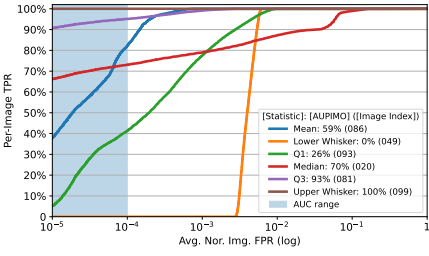
(a) Statistics and pairwise statistical tests.



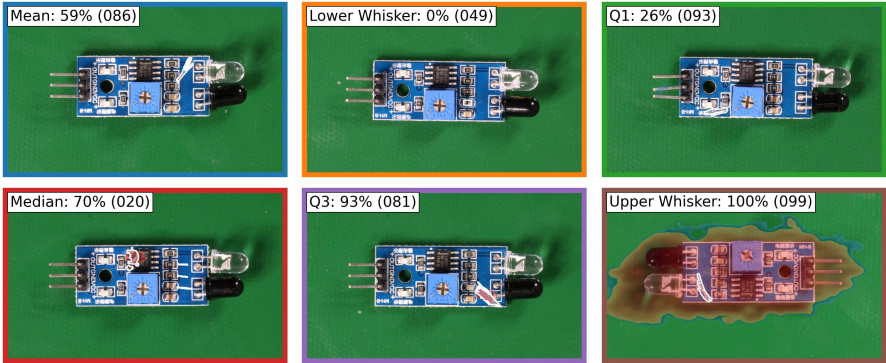
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

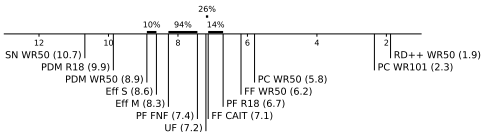


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

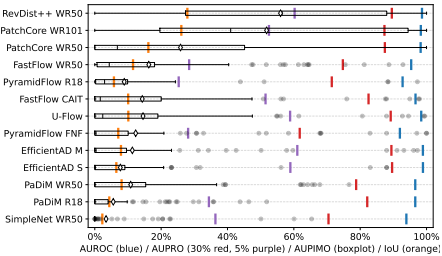
Figure 41: Benchmark on VisA / PCB 3. PIMO curves and heatmaps are from U-Flow. 201 images (101 normal, 100 anomalous).

	SN WR50	PDM R18	PDM WR50	Eff S	Eff M	PF FNF	UF	FF CAIT	PF R18	FF WR50	PC WR50	PC WR101	RD++ WR50
AUDROC	82.1% (11)	86.4% (8)	85.0% (10)	88.5% (1)	88.0% (2)	92.0% (13)	95.4% (4)	86.7% (12)	92.5% (13)	85.4% (10)	88.5% (8)	88.3% (13)	88.5% (13)
AUPRO	70.2% (12)	82.1% (8)	78.8% (9)	89.5% (13)	89.4% (13)	81.8% (13)	89.2% (4)	82.5% (12)	71.4% (11)	74.8% (10)	87.4% (15)	87.3% (6)	89.5% (2)
AUFI	35.4% (1)	55.4% (18)	45.4% (1)	91.0% (13)	91.0% (13)	28.1% (10)	35.8% (4)	31.3% (16)	35.8% (13)	28.3% (9)	28.3% (9)	28.3% (9)	80.5% (13)
Avg. AUPIMO	3.4% (13)	5.4% (12)	10.8% (8)	7.7% (11)	11.3% (10)	12.3% (1)	14.5% (5)	14.3% (6)	8.9% (10)	16.3% (4)	25.8% (3)	51.8% (2)	36.0% (1)
Std. AUPIMO	11.0%	11.3%	19.4%	16.6%	22.6%	24.3%	24.3%	23.8%	16.9%	23.0%	32.9%	37.6%	33.1%
P33 AUPIMO	0.0% (13)	0.0% (11)	0.0% (11)	0.0% (10)	0.0% (9)	0.0% (7)	0.0% (8)	0.0% (6)	0.9% (4)	1.0% (3)	0.3% (5)	23.3% (12)	34.2% (1)
Avg. Rank	10.7	9.9	8.9	8.6	8.3	7.4	7.2	7.1	6.7	6.2	5.1	2.1	1.9
Avg. Iou	4.2% (13)	4.2% (12)	8.1% (7)	8.4% (10)	7.0% (8)	7.1% (8)	10.1% (5)	10.1% (6)	5.4% (11)	11.5% (4)	16.2% (3)	26.1% (2)	27.0% (1)
RD++ WR50 (1.9)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99%	
PC WR101 (2.3)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		
PC WR50 (5.8)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
FF WR50 (6.2)	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%			
PF R18 (6.7)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
UF (7.2)	100%	100%	100%	100%	100%	91%	88%	96%	96%	100%			
PF FNF (7.4)	100%	100%	100%	100%	100%	94%	97%						
Eff R18 (8.1)	100%	99%	100%	100%	100%								
Eff S (8.6)	100%	96%	100%	100%	100%								
PDM WR50 (8.9)	100%	100%											
PDM R18 (9.9)	99%												

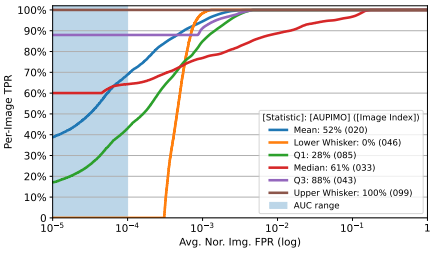
(a) Statistics and pairwise statistical tests.



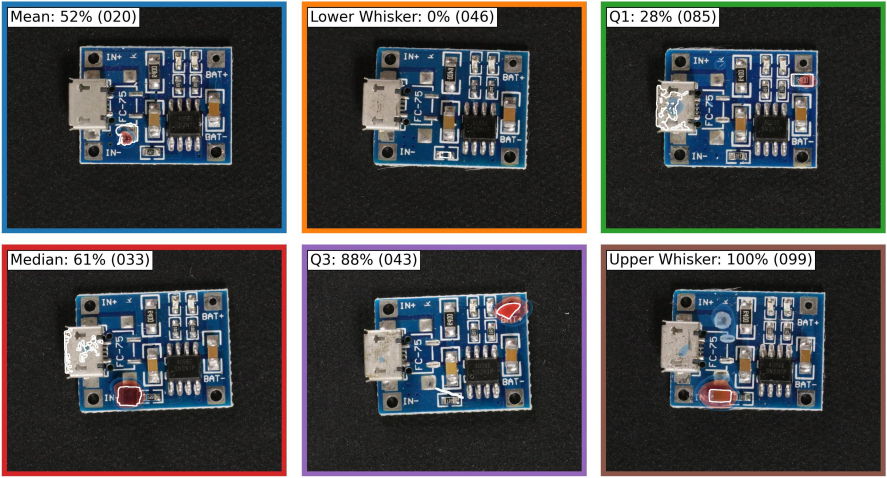
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.

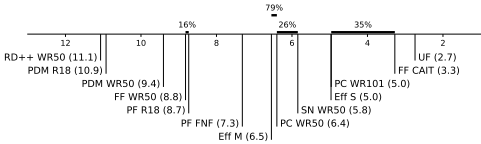


(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

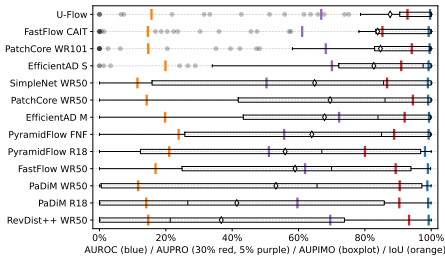
Figure 42: Benchmark on VisA / PCB 4. PIMO curves and heatmaps are from RevDist++ WR50. 201 images (101 normal, 100 anomalous).

	RD++ WR50	PDM R18	PDM WR50	FF WR50	PF R18	PF FNF	Eff M	PC WR50	SN WR50	Eff S	PC WR101	FF CAIT	UF
AUROC	99.2% (4)	99.1% (5)	98.8% (12)	99.0% (10)	98.1% (13)	99.2% (6)	99.3% (3)	99.1% (8)	99.0% (11)	99.2% (7)	99.5% (2)	99.2% (5)	99.6% (1)
AUPRO	93.2% (1)	92.4% (2)	90.5% (17)	90.3% (19)	80.0% (12)	88.8% (10)	92.0% (3)	94.2% (12)	86.9% (13)	91.0% (6)	91.5% (12)	85.2% (12)	92.8% (4)
AUPRO 5%	89.9% (1)	89.7% (2)	82.9% (18)	82.9% (18)	51.1% (10)	89.9% (8)	92.2% (1)	94.2% (12)	86.9% (13)	91.0% (6)	91.5% (12)	85.2% (12)	92.8% (4)
Avg. AUPIMO	36.9% (13)	41.4% (12)	53.2% (11)	59.0% (9)	56.0% (10)	64.0% (8)	67.8% (6)	69.9% (5)	64.9% (7)	82.7% (4)	84.7% (2)	83.9% (3)	87.6% (1)
Std. AUPIMO	37.6%	40.3%	42.3%	35.9%	39.3%	39.8%	35.4%	37.1%	41.3%	36.2%	28.1%	28.6%	25.2%
P33 AUPIMO	2.8% (12)	0.8% (13)	12.8% (11)	44.8% (9)	32.5% (10)	47.5% (8)	55.8% (6)	67.7% (5)	53.4% (7)	85.5% (4)	93.6% (2)	92.5% (3)	95.8% (1)
Avg. Rank	11.1	10.9	9.4	8.8	8.7	7.3	6.5	6.4	5.8	5.0	3.3	2.7	1.3
Avg. IoU	14.7% (8)	14.1% (11)	11.6% (12)	16.9% (5)	21.0% (2)	21.9% (1)	14.6% (13)	14.4% (11)	11.4% (13)	19.3% (3)	11.7% (11)	14.4% (9)	15.2% (6)
UF 12.7	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	91%	100%	
FF CAIT 13.3	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	96%		
PC WR101 15.0	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
Eff S 15.0	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%			
SN WR50 16.8	100%	100%	100%	100%	100%	100%	91%	26%					
PC WR50 16.4	100%	100%	100%	100%	100%	100%	79%						
Eff M 16.5	100%	100%	100%	100%	100%	100%							
PF FNF 17.3	100%	100%	100%	100%	93%								
PF R18 18.7	100%	100%	66%	16%									
FF WR50 16.8	100%	100%											
PDM WR50 19.4	100%	100%											
PDM R18 10.9	98%												

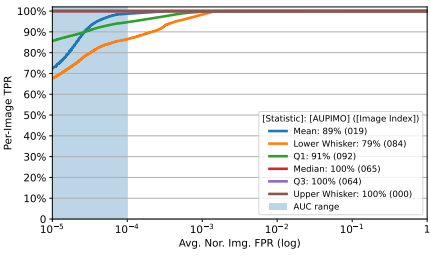
(a) Statistics and pairwise statistical tests.



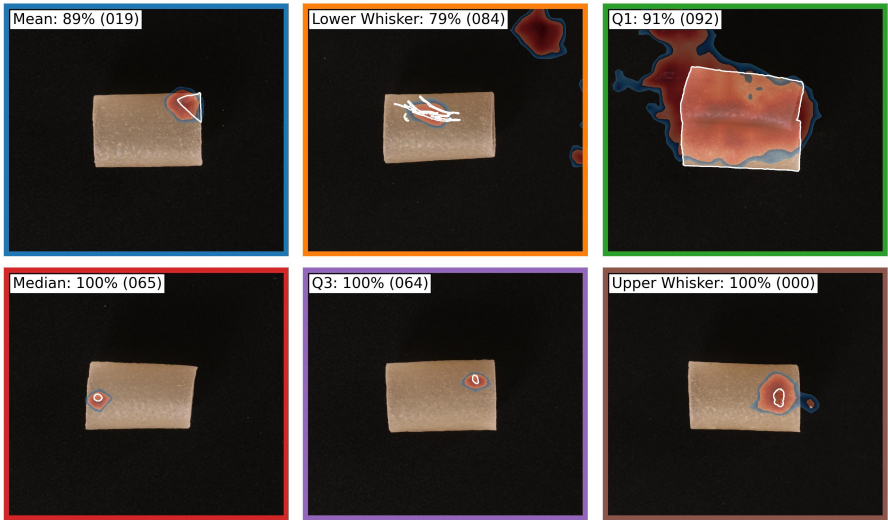
(b) Average rank diagram.



(c) Score distributions.



(d) PIMO curves.



(e) Heatmaps. Images selected according to AUPIMO's statistics. Statistic and image index annotated on upper left corner.

Figure 43: Benchmark on VisA / Pipe Fryum. PIMO curves and heatmaps are from U-Flow. 150 images (050 normal, 100 anomalous).