Data Invariants to Understand Unsupervised Out-of-Distribution Detection
Supplementary Material

Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila

University of Bern, Bern, Switzerland
{lars.doorenbos,raphael.sznitman,pablo.marquez}@unibe.ch

1 Supplementary Material

1.1 Dataset Details

We briefly describe all datasets used in our experiments. An overview of our experimental set-up is given in Table S1.

CIFAR10 [28]. (In) Small, natural images divided into 10 classes. For *uni-class*, one class forms the in-distribution, with its test set used in the evaluation. For *shift-low-res*, all 50000 training images are used for training when considered in-distribution, and all 10000 test images are used for testing. (Out) The remaining 9 classes are used as OOD for *uni-class*, subsampled to 1000 images.

CIFAR100 [28]. (In) 20 experiments with the training set of one of the semantic superclasses as the in-distribution, with its test set used during evaluation. (Out) Images from the remaining superclasses, subsampled to 500 images.

MVTec [4]. (In) Between 60 and 391 aligned images of 15 different objects and textures. 12-60 images are used as the in-distribution at test time. (Out) 30-141 images of defect objects are used as OOD.

OCT. (In) A collection of 58849 retinal Optical Coherence Tomography images used for training, and 300 for testing. (Out) Corrupted OCT scans built as described in [31].

Chest [59]. (In) The NIH Clinical Center ChestX-ray dataset containing 85524 training images. We use 300 images from the test set during evaluation. (Out) Corrupted X-ray scans as described in [31].

NIH [55]. (In) A collection of 4261 healthy X-ray scans of the NIH Clinical Center ChestX-ray dataset. The healthy test scans are used as the in-distribution during evaluation. (Out) Pathological scans from the same dataset.

DRD [17]. (In) 25809 healthy high-resolution retinal fundus photographs. Healthy test scans are again used during evaluation.
(Out) Retinal fundus photographs depicting 4 different levels of diabetic retinopathy (DR). The level of DR is indicated by a digit next to the method's name (DRD1–DRD4).

SVHN [35]. A dataset consisting of images of house numbers. We only use it as an OOD dataset, where the test set is reduced to 10000 samples.

DomainNet [37]. (In) The train and test images from the first 173 classes are used for training and evaluation respectively (as in [23]). We perform 10 experiments with the real images, and 10 with infographs. (Out) 10 domain-class combinations are used as OOD datasets. We avoid using `Real-B` and `Infograph-B` as OOD in the first and the second group of experiments respectively. All test sets are downsampled to 5000 images.

Table S1: Experimental set-up.

Category	# Tasks	Tasks	# train	# in	# out
uni-class	10	{airplane,automobile,bird,cat,deer, dog,frog,horse,ship,truck}:rest	5000	1000	1000
	20	{aquatic mammals,fish,flowers,food containers,fruit and vegetables, household electrical devices,household furniture,insects, large carnivores,large man-made outdoor things, large natural outdoor scenes,large omnivores and herbivores, medium-sized mammals,non-insect invertebrates, people,reptiles,small mammals,trees,vehicles 1,vehicles 2}:rest	2500	500	500
uni-ano	15	{bottle,cable,capsule,carpet,grid,hazelnut, leather,metal nut,pill,screw,tile, toothbrush,transistor,wood,zipper}:defect	60-391	12-60	30-141
uni-med	1	OCT:corruptions	58849	300	300
	1	Chest:corruptions	85524	300	300
	1	NIH:pathology	4261	677	667
	4	DRD:DRD1-4	25809	500	500
shift-low-res	1	CIFAR10:SVHN	50000	10000	10000
shift-high-res	10	Real A:{Quickdraw A,Quickdraw B,Infograph A, Infograph B,Sketch A,Sketch B, Clipart A,Clipart B,Painting A,Painting B}	61817	5000	5000
	10	Infograph A:{Quickdraw A,Quickdraw B, Sketch A,Sketch B,Real A,Real B, Clipart A,Clipart B,Painting A,Painting B}	14069	5000	5000

1.2 Implementation Details

We provide a short description of all models compared and their implementations. All modes make use of a ResNet-101 and rescale input images to 224×224 unless stated otherwise.

Glow [26] is a generative flow-based model, that allows for the exact computation of the likelihood, which we use as the anomaly score at test time. We use the implementation of [1], and an architecture with three blocks of 32 layers each. Images are resized to 32×32.

IC [51] aims to correct the high likelihood that generative models tend to assign to simple inputs, such as constant color images. To this end, IC computes the ratio between the likelihood of the generative model and a complexity

[1] `https://github.com/y0ast/Glow-PyTorch`

score of the input image. We used the Glow described above as our generative model and the length of the PNG image encoding as the complexity estimate.

HierAD [47] computes the ratio between the Glow generative model likelihood and a general background likelihood consisting of a Glow model trained on the *80 Million Tiny Images* dataset [56], provided at [2]. To make the method fully unsupervised, we do not use their proposed outlier loss during training.

MHRot [20] trains a multi-headed classifier to predict the correct transformation applied to an image. At test time, the classifier's softmax scores are combined for a final OOD score. Models are trained with the default settings until convergence of the validation loss.

DDV [31] aims to build an efficient latent representation by iteratively maximizing the log-likelihood of the low-dimensional latent vectors of the training images. Anomaly scores are given by the negative log-likelihood. We use our own implementation of DDV, following the settings described in its paper, i.e., a latent space of dimensionality 16 and a bandwidth of 10^{-2} [31].

MSCL [40] uses a novel contrastive loss function to fine-tune the final two blocks of a pretrained network, and combines this with an angular center loss for a final score. We used the official implementation with the learning rate set to $5 \cdot 10^{-5}$, as described in the paper, and trained until convergence.

CFlow [18] fits a normalizing flow network to features extracted from a pretrained network at multiple scales, conditioned on spatial information from a positional encoder. Anomaly scores are computed by aggregating the multi-scale likelihoods, upsampled to the original resolution. We again use the default hyperparameters.

DN2 [2] scores outliers by computing the mean distance to its 2 nearest neighbour on features extracted from the penultimate layer of a network pretrained on ImageNet.

SSD [50] uses contrastive learning for self-supervised representation learning. Then, it scores samples by the Mahalanobis distance computed at the last layer. All images were resized to 32×32. We use the default settings described in the official implementation.

MahaAD [41] is the Mahalanobis anomaly detector. Besides the ResNet-101, we also show results with an EfficientNet-b4 as described in [41]. With the ResNets, we resize images to 224×224, while for the EfficientNet-b4 this is 380×380.

1.3 Extended results

In Table S2 to Table S8 we dissect the per-task results from Table 1, reporting the AUC scores for each individual experiment and including some additional methods that were omitted from the main text for clarity.

[2] `https://github.com/boschresearch/hierarchical_anomaly_detection`

Table S2: AUC scores for CIFAR10 experiments of *uni-class*. First published (FP) column contains the dates of first online appearance.
* Our results

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average	FP
OCSVM [49]	63.0	44.0	64.9	48.7	73.5	50.0	72.5	53.3	64.9	50.8	58.5	Dec 1999
AnoGAN [48]	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8	Mar 2017
RCAE [8]	72.0	63.1	71.7	60.6	72.8	64.0	64.9	63.6	74.7	74.5	68.2	Feb 2018
GT [15]	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0	May 2018
Glow* [26]	76.1	44.5	60.3	57.3	43.9	55.1	36.2	46.4	71.0	46.4	53.7	Jul 2018
LSA [1]	73.5	58.0	69.0	54.2	76.1	54.6	75.1	53.5	71.7	54.8	64.1	Jul 2018
DSVDD [42]	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8	Jul 2018
IIC [25]	68.4	89.4	49.8	65.3	60.5	59.1	49.3	74.8	81.8	75.7	67.4	Jul 2018
DIM [21]	72.6	52.3	60.5	53.9	66.7	51.0	62.7	59.2	52.8	47.6	57.9	Aug 2018
OCGAN [38]	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.6	Mar 2019
MHRot [20]	77.5	96.9	87.3	80.9	92.7	90.2	90.9	96.5	95.2	93.3	90.1	Jun 2019
CapsNet [29]	62.2	45.5	67.1	67.5	68.3	63.5	72.7	67.3	71.0	46.6	61.2	Jul 2019
IC* [51]	38.3	62.0	45.5	61.5	48.7	63.9	62.6	63.7	48.4	58.8	55.3	Jul 2019
E3Outlier [58]	79.4	95.3	75.4	73.9	84.1	87.9	85.0	93.4	92.3	89.7	85.6	Sep 2019
DDV* [31]	83.2	58.5	55.4	56.9	61.2	57.9	63.3	57.5	88	71.2	65.3	Oct 2019
DeepIF [36]	-	-	-	-	-	-	-	-	-	-	88.2	Oct 2019
CAVGA-DU [57]	65.3	78.4	76.1	74.7	77.5	55.2	81.3	74.5	80.1	74.1	73.7	Nov 2019
U-Std [5]	78.9	84.9	73.4	74.8	85.1	79.3	89.2	83.0	86.2	84.8	82.0	Nov 2019
InvAE [24]	78.5	89.8	86.1	77.4	90.5	84.5	89.2	92.9	92.0	85.5	86.6	Nov 2019
DROCC [16]	81.7	76.7	66.7	67.1	73.6	74.4	74.4	71.4	80.0	76.2	74.2	Feb 2020
DN2 [2]	92.8	97.8	85.3	85	94.4	92.7	93.1	94.4	95.9	97.3	92.9	Feb 2020
ARAE [44]	72.2	43.1	69.0	55.0	75.2	54.7	70.1	51.0	72.2	40.0	60.2	Mar 2020
GOAD [3]	77.2	96.7	83.3	77.7	87.8	87.8	90.0	96.1	93.8	92.0	88.2	May 2020
MahaAD*$_{\mathrm{RN101}}$ [41]	92.9	96.4	85.8	85	93.8	91.1	94.1	94.8	95.4	96.8	92.6	May 2020
MahaAD*$_{\mathrm{ENB4}}$ [41]	95.1	97.8	92.3	91.6	96.5	**96.8**	**97.6**	96.9	97.4	98.3	96.0	May 2020
HierAD* [47]	47.6	63.4	63.2	59.0	79.2	64.3	77.5	66.4	61.6	59.8	64.2	Jun 2020
CSI [53]	89.9	**99.9**	93.1	86.4	93.9	93.2	95.1	98.7	97.9	95.5	94.3	Jul 2020
Puzzle-AE [45]	78.9	78.1	70.0	54.9	75.5	66.0	74.8	73.3	83.3	70.0	72.5	Aug 2020
PANDA [39]	**97.4**	98.4	93.9	90.6	**97.5**	94.4	97.5	97.5	97.6	97.4	96.2	Oct 2020
ConDA [52]	90.9	98.9	88.1	83.1	89.9	90.3	93.5	98.2	96.5	95.2	92.5	Nov 2020
MKD [46]	90.5	90.4	79.7	77.0	86.7	91.4	89.0	86.8	91.5	88.9	87.2	Nov 2020
SSD [50]	82.7	98.5	84.2	84.5	84.8	90.9	91.7	95.2	92.9	94.4	90.0	Mar 2021
SSL [62]	94.8	96.4	88.3	87.6	92.7	94.2	96.4	94.3	96.1	97.0	93.8	May 2021
MTL [32]	84.3	96.0	87.7	82.3	91.0	91.5	91.1	96.3	96.3	92.3	90.9	Jun 2021
MSCL* [40]	97	98.6	94.6	92.2	97.1	96.4	96.5	97.9	98.4	**98.6**	**96.7**	Jun 2021
OODformer [27]	92.3	99.4	**95.6**	**93.1**	94.1	92.9	96.2	**99.1**	**98.6**	95.8	95.7	Jul 2021
DaA [22]	-	-	-	-	-	-	-	-	-	-	75.3	Jul 2021
CFlow* [18]	69.4	83.5	68	73.9	84.7	77.9	84.4	78.6	80.3	84.2	78.5	Jul 2021

Table S3: AUC scores for CIFAR100 experiments of*uni-class*.
* Our results

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Mean
Glow* [26]	60.7	59.4	25.4	65.7	45.5	66.9	66.1	46.0	46.0	64.8	75.5	51.1	54.0	48.8	50.6	50.2	52.8	50.1	44.1	53.3	53.8
IC* [51]	61.2	53.9	44.4	44.4	48.3	46.4	41.9	51.2	72.0	58.0	48.7	68.3	69.8	51.6	56.1	62.0	62.4	68.8	59.5	48.8	55.9
OC-SVM [49]	68.4	63.6	52	64.7	58.2	54.9	57.2	62.9	65.6	74.1	84.1	58	68.5	64.6	51.2	62.8	66.6	73.7	52.8	58.4	63.1
DAGMM [65]	43.4	49.5	66.1	52.6	56.9	52.4	55	52.8	53.2	42.5	52.7	46.4	42.7	45.4	57.2	48.8	54.4	36.4	52.4	50.3	50.6
DSEBM [64]	64	47.9	53.7	48.4	59.7	46.6	51.7	54.8	66.7	71.2	78.3	62.7	66.8	52.6	44	56.8	63.1	73	57.7	55.5	58.8
DDV* [31]	58.3	58	70.6	75.3	72.2	60.3	65.4	61.4	63.8	72	77	55.5	82.8	53.4	61.4	58.6	51.9	87.5	64.5	72.3	66.1
HierAD* [47]	68.7	59.5	76.5	35.9	59.7	31.6	48.5	59.6	78.4	65.1	76.9	67.6	77.1	55.1	59.1	63.2	69.6	80.1	58.4	57.7	62.4
DVSDD [42]	66	60.1	59.2	58.7	60.9	54.2	63.7	66.1	74.8	78.3	80.4	68.3	75.6	61	64.3	66.3	72	75.9	67.4	65.8	67.0
GOAD [3]	73.9	69.2	67.6	71.8	72.7	67	80	59.1	79.5	83.7	84	68.7	75.1	56.6	83.8	66.9	67.5	91.6	88	82.6	74.5
MHRot [20]	77.6	72.8	71.9	81	81.1	66.7	87.9	69.4	86.8	91.7	87.3	85.4	85.1	60.3	92.7	70.4	78.3	93.5	89.6	88.1	80.1
SSD* [50]	76.5	79.6	88.7	73.4	91.1	72.4	73.9	79.8	80.7	86.0	72.3	79.4	83.1	74.5	87.3	74.4	79.9	90.9	83.3	80.7	80.4
ConDA [52]	82.9	84.3	88.6	86.4	92.6	84.5	73.4	84.2	87.7	94.1	85.2	87.8	82	82.7	93.4	75.8	80.3	97.5	94.4	92.4	86.5
CSI [53]	86.3	84.8	88.9	85.7	93.7	81.9	91.8	83.9	91.6	95	94	90.1	90.3	81.5	94.4	85.6	83	97.5	95.9	95.2	89.6
MKD* [46]	90.3	89.7	90.1	89.9	89.8	90.2	89.7	90.3	90.0	89.5	88.5	90.2	91.0	89.6	89.0	89.8	90.4	88.9	90.1	90.7	89.9
DN2* [2]	88.3	85.6	95.1	95.1	94.4	93.8	94.4	87.3	92.7	91.4	95.8	87.4	88.1	79.3	95.8	78.6	84.1	96.6	91.1	90.4	90.3
PANDA [39]	91.5	92.6	**98.3**	96.6	96.3	94.1	96.4	91.2	94.7	94	96.4	92.6	93.1	89.4	98	89.7	92.1	97.7	94.7	92.7	94.1
MSCL* [40]	**95.8**	**95.2**	97.6	**98.3**	97.1	**96.9**	**98.3**	**94.7**	**97.6**	**97.9**	**97.4**	**96.3**	**94.9**	**91.7**	**98.3**	**92.7**	**93.1**	**98.3**	**97.9**	**97.4**	**96.4**
CFlow* [18]	75.3	67.2	76	76	76.6	71.7	76.5	57.9	79.8	83.7	91.5	70.4	74.3	63.1	71.5	64.8	70.3	90.6	64.9	62	73.2
MahaAD*$_{\mathrm{RN101}}$ [41]	91.9	89.5	96	95.3	94.7	91.1	95.2	89.5	93.6	93.7	95.4	90.6	91.4	84.3	96.7	84.5	87.7	97.1	94.4	92.8	92.3
MahaAD*$_{\mathrm{ENB4}}$ [41]	93.2	92.8	96.7	97.8	**97.2**	95.4	98.0	92.6	95.9	94.9	95.8	93.0	93.0	89.2	97.8	89.1	91.7	97.5	96.2	94.8	94.6

Table S4: AUC scores for *shift-low-res*.
* Our results

	CIFAR10:SVHN
CFlow* [18]	6.6
Glow [47]	8.8
DSVDD [42]	14.5
MKD* [46]	26.8
DDV* [31]	47.9
EBM [14]	63.0
DN2* [2]	57.4
VAEBM [60]	83.0
MSCL* [40]	88.3
TT [34]	87.0
LLRe [61]	87.5
BIVA [19]	89.1
NAE [63]	92.0
HierAD [47]	93.9
IC [51]	95.0
GOAD [3]	96.3
SVD-RND [10]	96.4
MHRot [20]	97.8
DoSE [33]	97.3
CSI [53]	99.8
SSD [50]	99.6
MTL [32]	99.9
WAIC [33]	14.3
WAIC [9]	**100**
MahaAD*$_{\text{RN101}}$ [41]	94.3
MahaAD*$_{\text{ENB4}}$ [41]	96.2

Table S5: AUC scores for *shift-high-res* using `Real-A` as the in-distribution. QD: quickdraw, IG: infograph, SK: sketch, CA: Clipart, PN: Painting. A is the set without semantic shift, and B with semantic shift.
* Our results

	QDa	QDb	IGa	IGb	SKa	SKb	CAa	CAb	PNa	PNb	Mean
MSCL* [40]	33.8	32.9	68.6	67.1	54.7	58.9	58.3	61.3	72.4	75.5	58.3
SSD* [50]	40.3	40.4	69.0	69.6	68.9	73.9	53.1	58.8	77.6	83.3	64.0
MKD* [46]	24.2	23.1	56.6	52.7	47.2	47.3	49.4	47.3	68.6	70.4	48.9
DDV* [31]	87.9	90.9	56	54.6	62.6	64.6	62.1	64.8	52	59.4	64.0
DN2* [2]	50.4	50.8	76.2	74	69.1	73.7	70.7	**74.7**	**79.7**	**85.0**	70.4
MHRot* [20]	71.6	71.6	48.7	50.1	63.8	64.4	60.2	61.5	55.4	57.0	59.7
Glow* [26]	3.2	3.0	54.8	51.0	19.5	20.9	37.1	33.4	66.6	67.0	36.9
IC* [51]	89.9	90.4	66.4	68.8	69.5	68.8	64.4	66.3	55.9	55.7	68.0
HierAD* [47]	**95.5**	**95.7**	36.6	40.6	**84.9**	**82.7**	51.5	58.3	41.6	41.6	61.8
CFlow* [18]	46.6	47	52.2	49.6	48.9	51.1	62.3	62.9	58	57.6	53.6
MahaAD*$_{\text{RN101}}$ [41]	72.9	71.3	**81.6**	**80.8**	64.2	65.5	70.3	70	66	69.2	71.2
MahaAD*$_{\text{ENB4}}$ [41]	79.7	80.4	76.3	76.9	73.8	76.3	**71.0**	73.5	70.5	77.5	**75.6**

Table S6: AUC scores for *shift-high-res* using `Infograph-A` as the in-distribution.
* Our results

	QDa	QDb	SKa	SKb	REa	REb	CAa	CAb	PNa	PNb	Mean
MSCL* [40]	91.9	91.9	83.9	84.3	92.7	92.8	87.3	86.5	**96.3**	**96.2**	90.4
SSD* [50]	35.1	33.5	67.9	69.1	56.7	57.7	69.4	69.3	57.3	58.5	57.3
MKD* [46]	83.0	82.4	81.7	80.4	88.9	91.0	84.5	82.5	95.6	95.2	83.0
DDV* [31]	59.5	72.3	56.3	63.4	69.7	75.4	46.6	54.3	70.3	69.9	63.8
DN2* [2]	75.1	75.7	75.1	76.8	82.7	88.1	80.1	79.5	91.2	92.1	81.6
MHRot* [20]	94.9	95.2	88.5	88.7	87.6	87.9	**89.3**	**89.7**	88.6	89.4	86.7
Glow* [26]	0.7	0.6	12.3	14.0	50.7	49.9	35.3	30.6	69.2	69.5	34.4
IC* [51]	94.1	94.4	64.8	63.5	42.9	44.8	60.3	62.4	46.7	46.8	61.3
HierAD* [47]	**99.8**	**99.8**	**93.8**	**92.7**	83.1	83.3	80.8	83.1	77.6	77.6	84.1
CFlow* [18]	68.8	69	64.9	65.2	74.7	74.9	75.7	75.9	74.5	73.6	71.7
MahaAD*$_{\text{RN101}}$ [41]	92.3	92.1	78.1	77.6	88.1	88.4	81.5	80.3	90.9	91.2	86.1
MahaAD*$_{\text{ENB4}}$ [41]	94.5	94.8	89.5	89.0	**93.6**	**94.7**	87.4	87.1	94.9	95.4	**92.1**

Table S7: AUC scores for *uni-ano*. HN is hazelnut, MN is metal nut, TB is toothbrush and TS is transistor.
* Our results

	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	HN	MN	Pill	Screw	TB	TS	Zipper	Mean
AVID [43]	70	59	58	66	83	88	64	85	86	63	86	66	73	58	84	73
AESSIM [6]	67	69	46	52	83	88	61	61	54	54	60	51	74	52	80	63
AEL2 [6]	50	78	44	77	74	80	56	62	88	73	62	69	98	71	80	71
AnoGAN [48]	49	51	52	51	68	69	53	58	50	50	62	35	57	67	59	55
LSA [1]	74	54	70	70	75	86	61	71	80	67	85	75	89	50	88	73
CAVGA-DU [57]	73	75	71	70	85	89	63	83	84	67	88	77	91	73	87	78
DSVDD [42]	54	59	73	81	87	86	71	69	71	75	77	64	70	65	74	72
VAE-grad [13]	67	83	71	81	89	86	56	86	74	78	80	71	89	70	67	77
GT [15]	46	61.9	82.5	53.9	48.2	74.3	84.8	67.8	33.3	82.4	65.2	44.6	94	79.8	87.4	67.1
Puzzle-AE [45]	65.7	75.4	72.9	65.5	89.5	94.2	87.9	66.9	91.2	66.3	71.6	57.8	97.8	86	75.7	77.6
MKD [46]	79.3	78	95.1	91.6	94.3	99.4	89.2	80.5	98.4	73.6	82.7	83.3	92.2	85.6	93.2	87.7
MSCL* [40]	92.6	53.8	98	97.2	91.2	98.7	88.8	87.4	94.1	85	68.8	63.7	87.5	93.2	96.4	86.4
SSD* [50]	53.4	33.5	61.4	61.9	44.9	78.3	62.7	60.2	62.2	69.4	76.6	59.5	99.8	88.5	74.8	65.8
DDV* [31]	80.3	42	55.1	47.4	46.4	99.7	66.1	77.2	64.2	81	71.9	53.6	64.1	77.8	56	65.5
DN2* [2]	90.3	56.4	98.9	99.2	96.8	99.2	82	84.4	92.9	83.6	69.5	66.4	88.1	91.3	93.8	86.2
MHRot* [20]	47.8	58.9	75	51.2	90.2	82	79.9	59	73.6	75.7	64.9	36.6	86.9	86.5	93.4	70.8
Glow* [26]	72.9	**98.3**	94.1	83.7	96.9	96.6	83.3	67.1	90.5	62.4	84.8	31.8	87.6	88.4	91.3	82.0
IC* [51]	69.7	75.6	94.3	71.2	78.1	96.0	85.8	63.3	64.9	77.0	67.9	29.7	85.8	89.5	54.9	73.6
HierAD* [47]	73.4	95.3	95.5	84.5	97.5	97.3	86.5	70.0	75.0	73.6	74.2	26.2	98.6	92.5	84.1	81.6
SPADE [11]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	85.5
FAVAE [13]	67.1	97	67.5	80.5	94.8	**99.9**	95	80.4	99.3	85.2	82.1	83.7	95.8	93.2	97.2	87.9
AEsc [12]	89	97	89	99	95	98	89	74	94	73	84	74	**100**	91	94	89
DaA [22]	86.6	95.7	86.2	88.2	98.2	97.6	84.4	76.7	92.1	75.8	90	**98.7**	99.2	87.6	85.9	89.5
CFlow* [18]	**99.3**	93.3	**100**	99.2	**98.4**	**99.9**	**95.5**	90.9	**99.7**	**99.5**	**92.3**	83	92.2	93.9	**98.7**	**95.7**
MahaAD*$_{\text{RN101}}$ [41]	79.5	59.6	99.3	**100**	98.2	99.3	91.6	93.8	99.4	93.4	90.6	72.1	98.6	96.1	97.9	91.3
MahaAD*$_{\text{ENB4}}$ [41]	98.6	78.8	99.7	**100**	96.1	99.8	93.5	**97.0**	99.0	93.9	90.3	78.6	96.7	**96.5**	97.7	94.4

Table S8: AUC scores for *uni-med*.
* Our results

	OCT	Chest	NIH	DRD1	DRD2	DRD3	DRD4
IF [30]							44.0
AnoGAN [48]							44.2
DSEBM [64]							43.1
DAGMM [65]							52.0
Glow [26]	44.8	54.6					
GT [3]			79.2				
DSVDD [42]	77.4	66.6	81.8				46.4
DeepIF [36]							74.5
DDV [31]	86.7	79.9	57.7	45.3	48.9	50.2	53.4
GAOCC [54]			83.4				
MemDAE [7]			87.8				
MSCL* [40]	94.1	93.3	81.9	52	**55.8**	68.2	81.1
SSD* [50]	59.4	94.5	74.2	47.5	50.6	54.8	71.4
MKD* [46]	94.9	95.8	**88.0**	53.7	54.6	60.7	75.5
DN2* [2]	94.1	96.9	81.2	**54.4**	55.6	**69.4**	**85.4**
MHRot* [20]	87.7	96.2	81.8	49.0	50.2	52.7	65.3
Glow* [26]	62.3	49.8	65.0	52.2	47.5	54.7	59.5
IC* [51]	83.4	91.6	56.7	47.5	52.1	58.2	66.2
HierAD* [47]	94.3	99.0	79.8	52.1	51.7	57.5	73.5
CFlow* [18]	76.4	81.7	78.7	53.2	55.1	61.3	75.1
MahaAD*$_{\mathrm{RN101}}$ [41]	98	**99.8**	84.6	52.1	52	63.6	79.9
MahaAD*$_{\mathrm{ENB4}}$ [41]	**98.7**	**99.8**	84.2	49.9	55.0	66.3	81.3

References

1. Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2019)
2. Bergman, L., Cohen, N., Hoshen, Y.: Deep nearest neighbor anomaly detection. arXiv preprint arXiv:2002.10445 (2020)
3. Bergman, L., Hoshen, Y.: Classification-based anomaly detection for general data. International Conference on Learning Representations (2020)
4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (6 2019)
5. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4183–4192 (2020)
6. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011 (2018)
7. Bozorgtabar, B., Mahapatra, D., Vray, G., Thiran, J.P.: Salad: Self-supervised aggregation learning for anomaly detection on x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 468–478. Springer (2020)
8. Chalapathy, R., Menon, A.K., Chawla, S.: Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360 (2018)
9. Choi, H., Jang, E., Alemi, A.A.: Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392 (2018)
10. Choi, S., Chung, S.Y.: Novelty detection via blurring. International Conference on Learning Representations (2020)
11. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357 (2020)
12. Collin, A.S., De Vleeschouwer, C.: Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 7915–7922. IEEE (2021)
13. Dehaene, D., Frigo, O., Combrexelle, S., Eline, P.: Iterative energy-based projection on a normal data manifold for anomaly localization. arXiv preprint arXiv:2002.03734 (2020)
14. Du, Y., Mordatch, I.: Implicit generation and generalization in energy-based models. arXiv preprint arXiv:1903.08689 (2019)
15. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: Advances in Neural Information Processing Systems. pp. 9758–9769 (2018)
16. Goyal, S., Raghunathan, A., Jain, M., Simhadri, H.V., Jain, P.: Drocc: Deep robust one-class classification. In: International Conference on Machine Learning. pp. 3711–3721. PMLR (2020)
17. Graham, B.: Kaggle diabetic retinopathy detection competition report. University of Warwick (2015)
18. Gudovskiy, D., Ishizaka, S., Kozuka, K.: Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 98–107 (2022)
19. Havtorn, J.D.D., Frellsen, J., Hauberg, S., Maaløe, L.: Hierarchical vaes know what they don't know. In: International Conference on Machine Learning. pp. 4117–4128. PMLR (2021)
20. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in Neural Information Processing Systems **32** (2019)
21. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
22. Hou, J., Zhang, Y., Zhong, Q., Xie, D., Pu, S., Zhou, H.: Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8791–8800 (2021)
23. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10951–10960 (2020)
24. Huang, C., Ye, F., Cao, J., Li, M., Zhang, Y., Lu, C.: Attribute restoration framework for anomaly detection. arXiv preprint arXiv:1911.10676 (2019)
25. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information distillation for unsupervised image segmentation and clustering. arXiv preprint arXiv:1807.06653 **2**(3), 8 (2018)
26. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems **31** (2018)
27. Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V.: Oodformer: Out-of-distribution detection transformer. arXiv preprint arXiv:2107.08976 (2021)
28. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
29. Li, X., Kiringa, I., Yeap, T., Zhu, X., Li, Y.: Exploring deep anomaly detection methods based on capsule net. arXiv preprint arXiv:1907.06312 (2019)
30. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 eighth ieee international conference on data mining. pp. 413–422. IEEE (2008)
31. Márquez-Neila, P., Sznitman, R.: Image data validation for medical systems. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 329–337. Springer (2019)
32. Mohseni, S., Vahdat, A., Yadawa, J.: Multi-task transformation learning for robust out-of-distribution detection. arXiv preprint arXiv:2106.03899 (2021)
33. Morningstar, W., Ham, C., Gallagher, A., Lakshminarayanan, B., Alemi, A., Dillon, J.: Density of states estimation for out of distribution detection. In: International Conference on Artificial Intelligence and Statistics. pp. 3232–3240. PMLR (2021)
34. Nalisnick, E., Matsukawa, A., Teh, Y.W., Lakshminarayanan, B.: Detecting out-of-distribution inputs to deep generative models using a test for typicality. International Conference on Learning Representations (2019)
35. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)

36. Ouardini, K., Yang, H., Unnikrishnan, B., Romain, M., Garcin, C., Zenati, H., Campbell, J.P., Chiang, M.F., Kalpathy-Cramer, J., Chandrasekhar, V., et al.: Towards practical unsupervised anomaly detection on retinal images. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 225–234. Springer (2019)
37. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1406–1415 (2019)
38. Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2906 (2019)
39. Reiss, T., Cohen, N., Bergman, L., Hoshen, Y.: Panda: Adapting pretrained features for anomaly detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2806–2814 (2021)
40. Reiss, T., Hoshen, Y.: Mean-shifted contrastive loss for anomaly detection. arXiv preprint arXiv:2106.03844 (2021)
41. Rippel, O., Mertens, P., Merhof, D.: Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 6726–6733. IEEE (2021)
42. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. pp. 4393–4402. PMLR (2018)
43. Sabokrou, M., Pourreza, M., Fayyaz, M., Entezari, R., Fathy, M., Gall, J., Adeli, E.: Avid: Adversarial visual irregularity detection. In: Asian Conference on Computer Vision. pp. 488–505. Springer (2018)
44. Salehi, M., Arya, A., Pajoum, B., Otoofi, M., Shaeiri, A., Rohban, M.H., Rabiee, H.R.: Arae: Adversarially robust training of autoencoders improves novelty detection. arXiv preprint arXiv:2003.05669 (2020)
45. Salehi, M., Eftekhar, A., Sadjadi, N., Rohban, M.H., Rabiee, H.R.: Puzzle-ae: Novelty detection in images through solving puzzles. arXiv preprint arXiv:2008.12959 (2020)
46. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14902–14912 (2021)
47. Schirrmeister, R., Zhou, Y., Ball, T., Zhang, D.: Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. Advances in Neural Information Processing Systems **33**, 21038–21049 (2020)
48. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. pp. 146–157. Springer (2017)
49. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C., et al.: Support vector method for novelty detection. In: NIPS. vol. 12, pp. 582–588. Citeseer (1999)
50. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. International Conference on Learning Representations (2021)
51. Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J.F., Luque, J.: Input complexity and out-of-distribution detection with likelihood-based generative models. International Conference on Learning Representations (2019)

52. Sohn, K., Li, C.L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. International Conference on Learning Representations (2021)
53. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
54. Tang, Y.X., Tang, Y.B., Han, M., Xiao, J., Summers, R.M.: Abnormal chest x-ray identification with generative adversarial one-class classifier. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1358–1361. IEEE (2019)
55. Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., et al.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ digital medicine **3**(1), 1–8 (2020)
56. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008)
57. Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: European Conference on Computer Vision. pp. 485–503. Springer (2020)
58. Wang, S., Zeng, Y., Liu, X., Zhu, E., Yin, J., Xu, C., Kloft, M.: Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In: Advances in Neural Information Processing Systems. pp. 5962–5975 (2019)
59. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
60. Xiao, Z., Kreis, K., Kautz, J., Vahdat, A.: Vaebm: A symbiosis between variational autoencoders and energy-based models. In: International Conference on Learning Representations (2020)
61. Xiao, Z., Yan, Q., Amit, Y.: Likelihood regret: An out-of-distribution detection score for variational auto-encoder. arXiv preprint arXiv:2003.02977 (2020)
62. Xiao, Z., Yan, Q., Amit, Y.: Do we really need to learn representations from in-domain data for outlier detection? ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning (2021)
63. Yoon, S., Noh, Y.K., Park, F.C.: Autoencoding under normalization constraints. arXiv preprint arXiv:2105.05735 (2021)
64. Zhai, S., Cheng, Y., Lu, W., Zhang, Z.: Deep structured energy based models for anomaly detection. In: International Conference on Machine Learning. pp. 1100–1109. PMLR (2016)
65. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)