Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings Supplementary Material

Paul Bergmann Michael Fauser David Sattlegger Carsten Steger

MVTec Software GmbH

www.mvtec.com

{paul.bergmann, fauser, sattlegger, steger}@mvtec.com

Abstract

We provide the following supplementary material:

- *Additional information on the network architectures for different receptive fields that are used in our experiments.*
- *Additional information on the hyperparameter settings for the conducted experiments and more detailed evaluation results for the MNIST and CIFAR-10 datasets.*
- *Qualitative results of our method on the MVTec Anomaly Detection dataset.*

1. Network Architectures

A description of the network architecture for a patch-sized teacher network $\hat{T}$ with receptive field of size $p = 65$ can be found in our main paper (Table 4). Architectures for teachers with receptive field sizes $p = 33$ and $p = 17$ are depicted in Tables 1a and 1b, respectively. Leaky rectified linear units with slope 5×10^{-3} are used as activation function after each convolution layer.

2. Experiments on MNIST and CIFAR-10

Here, we give details about additional hyperparameters for our experiments on the MNIST and CIFAR-10 datasets. We additionally provide the per-class ROC-AUC values for the two datasets in Tables 2 and 3, respectively.

Hyperparameter Settings For the deterministic ℓ_2-autoencoder (ℓ_2-AE) and the variational autoencoder (VAE), we use a fully connected encoder architecture of shape 128–64–32–10 with leaky rectified linear units of slope 5×10^{-3}. The decoder is constructed in a manner symmetric to the encoder. Both autoencoders are trained for 100 epochs at an initial learning rate of 10^{-2} using the Adam optimizer and a batch size of 64. A weight decay rate of 10^{-5} is applied for regularization. To evaluate the reconstruction probability of the VAE, five independent forward passes are performed for each feature vector. For the One-Class SVM (OC-SVM), a radial basis function kernel is used. K-Means is trained with 10 cluster centers and the distance to the single closest cluster center is evaluated as the anomaly score for each input sample. For 1-NN, the feature vectors of all available training samples are stored and tested during inference.

3. Experiments on MVTec AD

We give additional information on the hyperparameters used in our experiments on MVTec AD for both shallow machine learning models as well as deep learning methods.

Shallow Machine Learning Models For the 1-NN classifier, we construct a dictionary of 5000 feature vectors and take the distance to the closest training sample as anomaly score. For the other shallow classifiers, we fit their parameters on 50 000 training samples, randomly chosen from the teacher's feature maps. The K-Means algorithm is run with 10 cluster centers and measures the distance to the nearest cluster center in the feature space during inference. The OC-SVM employs a radial basis function kernel.

Deep-Learning Based Models For evaluation on MVTec AD, the architecture of the ℓ_2-AE and VAE are identical to the ones used on the MNIST and CIFAR-10 dataset. Each fully connected autoencoder is trained for 100 epochs. We use Adam with initial learning rate 10^{-4} and weight decay 10^{-5}. Batches are constructed from 512 randomly sampled vectors of the teacher's feature maps. The reconstruction probability of the VAE is computed by five individ-

ual forward passes through the network. For the evaluation of AnoGAN, the SSIM-Autoencoder, and the CNN-Feature Dictionary, we use the same hyperparameters as Bergmann et al. in the MVTec AD dataset paper [1]. Only a slight adaption is applied to the CNN-Feature Dictionary by cropping patches of size $p = 65$ and performing the evaluation by computing anomaly scores for overlapping patches with a stride of 4 pixels.

Qualitative Results We provide additional qualitative results of our method on MVTec AD for three objects and three textures in Figure 1. For each category, anomaly maps for multiple defect classes are provided. Our method performs well across different defect types and sizes. The results are shown for an ensemble of 3 students and a multi-scale architecture of receptive field sizes in $\{17, 33, 65\}$ pixels.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.

Layer	Output Size	Parameters	
		Kernel	Stride
Input	33×33×3		
Conv1	29×29×128	3×3	1
MaxPool	14×14×128	2×2	2
Conv2	10×10×256	5×5	1
MaxPool	5×5×256	2×2	2
Conv3	4×4×256	2×2	1
Conv4	1×1×128	4×4	1
Decode	1×1×512	1×1	1

(a) Architecture for $p = 33$.

Layer	Output Size	Parameters	
		Kernel	Stride
Input	17×17×3		
Conv1	12×12×128	5×5	1
Conv2	8×8×256	5×5	1
Conv3	4×4×256	5×5	1
Conv4	1×1×128	4×4	1
Decode	1×1×512	1×1	1

(b) Architecture for $p = 17$.

Table 1: Network architectures for teacher networks $\hat{T}$ with different receptive field sizes p.

Method				0	1	2	3	4	5	6	7	8	9	Mean
OCGAN				0.998	**0.999**	0.942	0.963	0.975	0.980	0.991	0.981	0.939	0.981	0.9750
1-NN				0.989	0.998	0.962	0.970	0.980	0.955	0.979	0.981	0.968	0.971	0.9753
KMeans				0.973	0.995	0.898	0.948	0.960	0.920	0.948	0.948	0.940	0.927	0.9457
OC-SVM				0.980	0.998	0.887	0.944	0.964	0.909	0.949	0.957	0.935	0.940	0.9463
ℓ_2-AE				0.992	**0.999**	0.967	0.980	0.988	0.970	0.988	0.987	0.978	0.983	0.9832
VAE				0.983	0.998	0.915	0.941	0.969	0.925	0.964	0.940	0.955	0.945	0.9535
	L_k	L_m	L_c											
Ours	✓		✓	**0.999**	**0.999**	0.990	**0.993**	**0.992**	**0.993**	**0.997**	**0.995**	0.986	0.991	**0.9935**
Ours	✓	✓	✓	**0.999**	**0.999**	0.988	0.992	0.988	**0.993**	**0.997**	**0.995**	0.984	0.991	0.9926
Ours		✓	✓	**0.999**	**0.999**	**0.992**	0.992	0.988	**0.993**	**0.997**	**0.995**	**0.988**	**0.992**	**0.9935**
Ours	✓			**0.999**	**0.999**	0.989	0.990	0.990	0.990	**0.997**	0.993	0.981	0.989	0.9917

Table 2: Results on the MNIST dataset. For each method and digit, the area under the ROC curve is given. For our algorithm, we evaluate teacher networks trained with different loss functions. ✓ corresponds to setting the respective loss weight to 1, otherwise it is set to 0.

Method				airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck	Mean
OCGAN				0.757	0.531	0.640	0.620	0.723	0.620	0.723	0.575	0.820	0.554	0.6566
1-NN				0.792	**0.860**	**0.746**	0.729	0.815	**0.797**	0.876	**0.836**	0.856	**0.882**	0.8189
KMeans				0.673	0.822	0.665	0.676	0.742	0.746	0.828	0.780	0.817	0.843	0.7592
OC-SVM				0.651	0.785	0.618	0.679	0.733	0.730	0.797	0.760	0.799	0.836	0.7388
ℓ_2-AE				0.747	0.862	0.690	0.698	0.788	0.759	0.849	0.824	0.812	0.869	0.7898
VAE				0.705	0.819	0.605	0.700	0.734	0.731	0.797	0.751	0.801	0.859	0.7502
	L_k	L_m	L_c											
Ours	✓		✓	0.789	0.849	0.734	**0.748**	**0.851**	0.793	**0.892**	0.830	**0.862**	0.848	**0.8196**
Ours	✓	✓	✓	0.784	0.836	0.706	0.742	0.826	0.768	0.870	0.815	0.857	0.831	0.8035
Ours		✓	✓	**0.804**	0.855	0.706	0.709	0.798	0.738	0.860	0.797	0.849	0.824	0.7940
Ours	✓			0.766	0.817	0.715	0.736	0.855	0.763	0.885	0.819	0.838	0.827	0.8021

Table 3: Results on the CIFAR-10 dataset. For each method and class, the area under the ROC curve is given. For our algorithm, we evaluate teacher networks trained with different loss functions. ✓ corresponds to setting the respective loss weight to 1, otherwise it is set to 0.

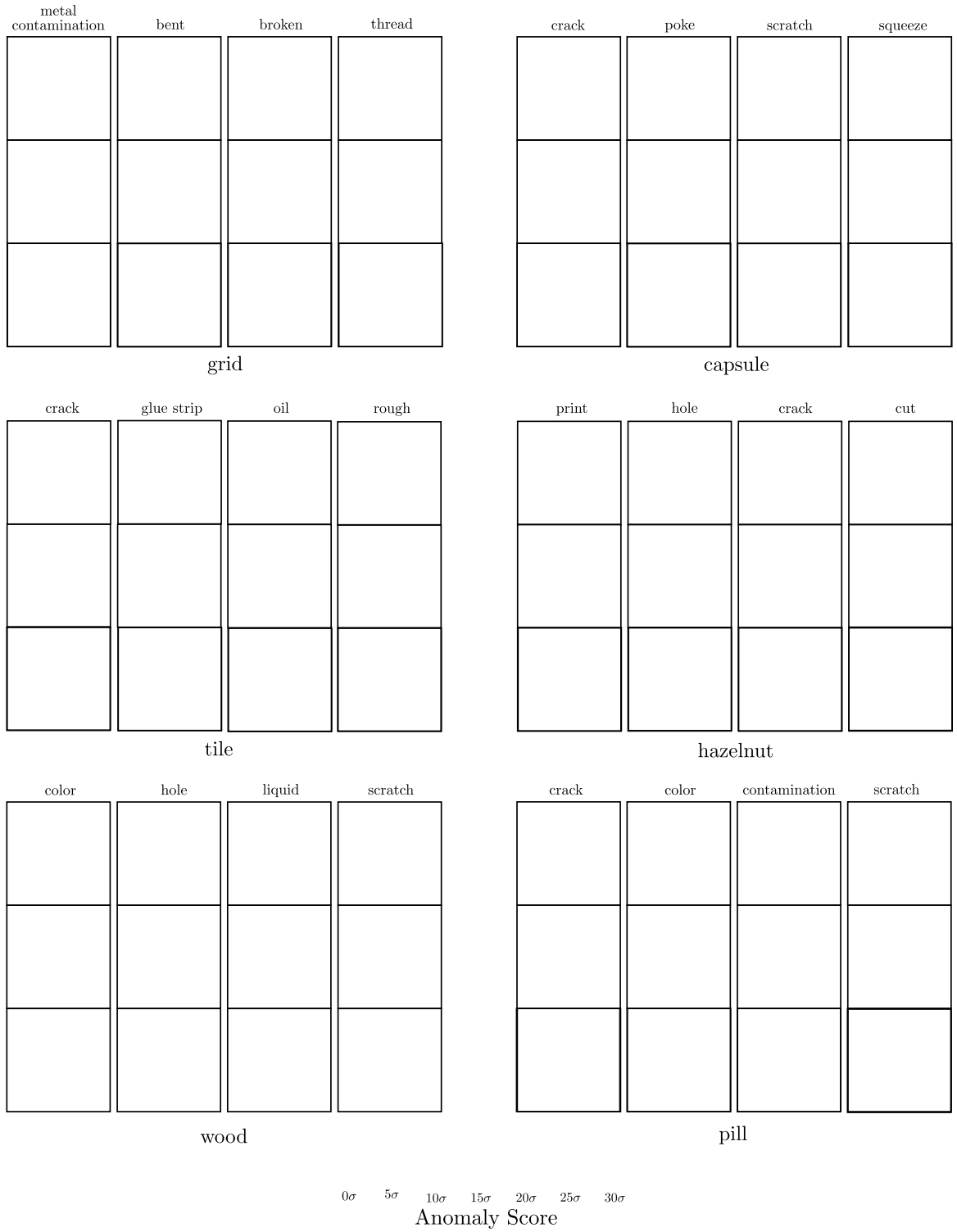

Figure 1: Qualitative results of our method on selected textures (left) and objects (right) of the MVTec Anomaly Detection dataset. Our algorithm performs robustly across various defect categories, such as color defects, contaminations, and structural anomalies. **Top row:** Input images containing defects. **Center row:** Ground truth regions of defects in red. **Bottom row:** Anomaly scores for each image pixel predicted by our algorithm.