Modeling the Distribution of Normal Data in Pre-Trained Deep Features for Anomaly Detection

Oliver Rippel*
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany
Email: oliver.rippel@ifb.rwth-aachen.de

Patrick Mertens*
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany

Dorit Merhof
Institute of Imaging & Computer Vision
RWTH Aachen University
Aachen, Germany

Abstract—Anomaly Detection (AD) in images is a fundamental computer vision problem and refers to identifying images and/or image substructures that deviate significantly from the norm. Popular AD algorithms commonly try to learn a model of normality from scratch using task specific datasets, but are limited to semi-supervised approaches employing mostly normal data due to the inaccessibility of anomalies on a large scale combined with the ambiguous nature of anomaly appearance.

We follow an alternative approach and demonstrate that deep feature representations learned by discriminative models on large natural image datasets are well suited to describe normality and detect even subtle anomalies. Our model of normality is established by fitting a multivariate Gaussian to deep feature representations of classification networks trained on ImageNet using normal data only in a transfer learning setting. By subsequently applying the Mahalanobis distance as the anomaly score we outperform the current state of the art on the public MVTec AD dataset, achieving an Area Under the Receiver Operating Characteristic curve of $95.8 \pm 1.2\%$ (mean $\pm$ SEM) over all 15 classes. We further investigate why the learned representations are discriminative to the AD task using Principal Component Analysis. We find that the principal components containing little variance in normal data are the ones crucial for discriminating between normal and anomalous instances. This gives a possible explanation to the often sub-par performance of AD approaches trained from scratch using normal data only. By selectively fitting a multivariate Gaussian to these most relevant components only we are able to further reduce model complexity while retaining AD performance. We also investigate setting the working point by selecting acceptable False Positive Rate thresholds based on the multivariate Gaussian assumption.

I. INTRODUCTION

Anomaly Detection (AD) relates to identifying instances in data that are significantly different to the norm [1], [2]. Correspondingly, AD in images aims at finding irregularities in images and poses a fundamental computer vision problem with various application domains ranging from industrial quality control [3] to medical image analysis [4]. In general, AD tasks are defined by the following two characteristics:

- Anomalies are rare events, i.e. their prevalence in the application domain is low.
- Anomaly appearance is not well-defined (i.e. anomalies types are ambiguous).

Together, these characteristics result in AD datasets that are heavily imbalanced, often containing only few anomalies for model verification and testing.

As a consequence, AD algorithms often focus on semi-supervised learning approaches, where a model of normality is established based on normal data only [3]–[5]. While small dataset sizes predestine the capitalization of pre-training on large-scale databases such as ImageNet [6], only little research is performed to explore this potential [7]–[9]. Instead, methods focus on learning feature representations from scratch, often in reconstruction-based approaches [4], [10].

As our main contribution, we demonstrate the effectiveness of pre-trained deep feature representations transferred to the AD task. By fitting a multivariate Gaussian to normal data of deep features learned by ImageNet training and using the Mahalanobis distance [11] as the anomaly score, we are able to outperform the prior state of the art on the public MVTec AD dataset [3]. We additionally gain insight into and explain the discriminative nature of pre-trained deep features by means of Principal Component Analysis (PCA). Here, we find that principal components that retain little variance in normal data are highly discriminative to the AD task, indicating that learning these features from scratch may be difficult using normal data only. We further show that the working point can sensibly be set based on choosing an acceptable False Positive Rate (FPR) under the multivariate Gaussian assumption. Here, retaining only highly variant principal components decreases FPR at the cost of AD performance. These results demonstrate that there should be a clear focus on leveraging pre-trained deep feature representations in future AD research.

II. RELATED WORK

In recent years, a large body of research has been published in the field of AD. Therefore we provide an extensive overview of AD techniques in the following, focussing on methods applied to image data. We further categorize the approaches into whether they leverage pre-trained deep feature representations in a transfer learning approach or are learned from scratch.

A. Learning AD from Scratch

Learning useful representations from scratch in a semi- or unsupervised manner is a major research field on its own.
Out of the multitude of ways of learning such representations, autoencoder-based approaches are the most popular in AD.

Here, autoencoders (AEs) try to learn the identity function in a semi-supervised manner using a given set of exclusively normal training images. The learned identity function is constrained, whereas the model first has to compress the input image to a low dimensional embedding, and subsequently has to reconstruct the input image based on this embedding. It is argued that the overall model cannot represent anomalous image structures, reconstructing a plausible normal image instead. An AE trained until convergence can then be used for AD in different ways:

Anomalous images can be detected by comparing the input test image with its reconstruction yielded by the model. There have been various proposals employing this reconstruction for AD [10], [12], [13]. While the results of reconstruction-based AD approaches are intuitive to understand, they suffer from two drawbacks: (I) The reconstruction has to be post-processed in order to yield an image-level anomaly score, thus increasing the complexity of the method and (II) the decoder part introduces additional computational overhead.

Embeddings learned by AEs are also utilized in many AD frameworks. Common approaches try to model the distribution of normal data in the AE embedding in a generative way using variational AEs [14] that are oftentimes trained using an adversarial objective [15]. Alternatively, classical shallow ML methods such as k-Nearest Neighbor (k-NN) or one class Support Vector Machine (oc-SVM) [16] are also applied to embeddings learned by an AE [17]. More recently, Ruff et al. have initialized their proposed Deep Support Vector Data Description (Deep SVDD) using pre-trained AEs [5], [18].

Hybrid approaches also exist, where anomaly scores are generated by combining measures proposed on the embeddings with reconstruction errors [19]–[21], enhancing model performance at cost of further increased complexity.

B. Transfer Learning AD with Deep Feature Representations

Less extensively studied than semi-supervised feature learning methods, AD has also been performed by using deep representations learned by large-scale ImageNet training in a transfer learning setting for both anomaly segmentation and image-level AD.

While there has been recent success in adapting deep feature representations for anomaly segmentation [7], [22], [23], these proposals compute features patch-wise to yield the pixel-wise output. As a consequence, receptive fields are limited, feature complexity is rather low and there is an implicit assumption that the anomalies fit inside one patch. Further, segmentations have to be aggregated to yield image-level AD. Regarding image-level AD, Christiansen et al. [24] repurpose deep AlexNet [25] and VGG [26] features for agricultural anomalous object detection. While they also fit a multivariate Gaussian to deep feature representations and use the Mahalanobis distance as an anomaly measure, they evaluate their model using a small in-house dataset only. Further, in their use-case anomalous instances deviate significantly in appearance from the normal class, and benchmarking against other AD approaches is not performed. Also, they do not investigate the properties of the pre-trained feature representations that make them suitable to AD. Andrews et al. successfully fit an oc-SVM to deep representations learned by VGG on ImageNet for AD in X-Ray scans of containers [8]. Bergman et al. [27] and Cohen et al. [9] evaluate a k-NN using L2-distance on ResNet [28] features. Here, Cohen et al. [9] report an average Area Under the Receiver Operating Characteristic curve (AUROC) of 85.5% on the public MVTec AD dataset with $k = 50$ and average-pooled features extracted from the last convolutional layer of a Wide-ResNet50-2. Except for these and a 1-NN approach with different normalizations in surveillance videos [29], little notice has been given to employing deep features in AD for classifying full images.

While not directly used as an AD algorithm, Lee et al. also model the data distribution of in-distribution data by means of multivariate Gaussian for Out-Of-Distribution (OOD) detection [30]. Contrary to AD, OOD determines whether a given query image is part of the in-distribution dataset (i.e. the dataset used for training) or OOD. Using a small subset of anomalies for fine-tuning, they apply the linear combination of Mahalanobis distances computed at various depths of a pre-trained ResNet [28] to a test image. The test image is additionally pre-processed to maximize Mahalanobis distance by means of performing a single gradient ascent step to implicitly evaluate Probability Density Function (PDF) around the test image. They further show that discriminative deep classifiers employing softmax learn the same posterior distribution as generative classifiers under a Linear Discriminant Analysis (LDA) assumption (i.e. Gaussian Discriminant Analysis with class-tied covariance). They expand on this and argue that the pre-trained features of the deep softmax classifier may also follow the class-conditional Gaussian distribution of the generative classifier. While they do not give a theoretical proof for this, the performance of a generative classifier based on pre-trained features is verified experimentally. They also show that prior unseen classes may be easily integrated into the generative classifier by introducing a new class to the LDA (compute new mean and update joint covariance). This finding is the motivation for our work, where we apply pre-trained deep feature representations to the AD task in a transfer learning setting.

III. MODELING NORMAL DATA DISTRIBUTION IN DEEP FEATURE REPRESENTATIONS

Based on the findings of Lee et al. [30], we hypothesize that pre-trained deep representations can also be successfully applied to the AD task. Similar to the class-incremental learning approach, we establish a model of normality using normal data only and omit any fine-tuning of the learned model.

Such a model is the multivariate Gaussian, which is defined as

$$
    \varphi_{\mu, \Sigma}(x) := \frac{1}{\sqrt{(2\pi)^D |\det \Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.
$$

(1)
Here, $D$ is the number of dimensions, $\mu \in \mathbb{R}^D$ is the mean vector and $\Sigma \in \mathbb{R}^{D \times D}$ the symmetric covariance matrix of the distribution. $\Sigma$ must be positive definite.

Under a Gaussian distribution with mean $\mu$ and covariance $\Sigma$, a distance measure between a particular point $x \in \mathbb{R}^D$ and the distribution is called the Mahalanobis distance and defined as

$$M(x) = \sqrt{(x - \mu)^\top \Sigma^{-1} (x - \mu)}.$$  \hfill (2)

Introduced by Mahalanobis in 1936, $M(x)$ is a useful measure of uncertainty of a sample [11]. This interpretation stems from the fact that the Mahalanobis distance uniquely determines the probability density $\varphi_{\mu, \Sigma}(x)$ of an observation. When $x$ is sampled from the Gaussian distribution, $M(x)^2$ is chi-squared distributed with $k = D$ degrees of freedom. This $\chi^2$-distribution with $k$ degrees of freedom is the sum of $k$ conditionally independent standard normal random variables. Its PDF is given as

$$f_k(x) = \begin{cases} \frac{x^{k/2-1} e^{-x/2}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$  \hfill (3)

Here, $\Gamma(s)$ is the gamma function for $s > 0$. The Cumulative Distribution Function (CDF) of the $\chi^2$-distribution is calculated as

$$F_k(x) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$  \hfill (4)

with the lower incomplete gamma function $\gamma(s, x)$.

A. Covariance Estimation

As the true distribution of the novel data in the deep feature spaces is unknown, the covariance matrix $\Sigma$ needs to be approximated from observations $x_1, \ldots, x_n \in \mathbb{R}^D$ with the sample covariance

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top.$$  \hfill (5)

Here, $\bar{x}$ denotes the empirical mean of the observations.

However, the sample covariance matrix is only well-conditioned when the number of dimensions $D$ is much lower than the number of samples $n$. If $\frac{D}{n}$ is non-negligible, the covariance estimate becomes unstable, and when $D > n$, $\hat{\Sigma}$ becomes singular and hence is not invertible. To solve this problem the concept of shrinkage has been proposed for sample covariance estimation and will be used to estimate the sample covariance in this work.

Shrinkage is defined as a linear combination of empirically estimated covariance matrix and the (scaled) identity matrix $I_D$,

$$\hat{\Sigma}_{\text{shrunken}} = (1 - \rho)\hat{\Sigma} + \rho \frac{\text{tr}(\hat{\Sigma})}{D} I_D$$  \hfill (6)

with shrinkage intensity $\rho$. Thus, $\rho$ regulates the influence of the empirical estimator on the final matrix, preferring the well-conditioned identity matrix for larger $\rho$. It can be seen as a bias-variance tradeoff between the biased, invariant identity estimate and the unbiased high-variance empirical covariance.

By minimizing the expected squared error $E[\|\hat{\Sigma}_{\text{shrunken}} - \Sigma\|^2]$ to the true covariance, Ledoit, Wolf et al. obtain a closed form solution for the amount of shrinkage that allows optimal selection of $\rho$ given an unstable estimate of $\hat{\Sigma}$ [31].

B. Setting the Working Point

In the case of assuming an underlying multivariate Gaussian, the working point can be estimated based on probabilities. The idea is that if a specific Mahalanobis distance corresponds to a probability $p$ of seeing a normal sample, this matches the expected True Negative Rate (TNR) of a detector thresholded at that distance. $1 - p$ can be seen as the allowed probability of falsely-labeled normal instances, i.e. the FPR.

For a (multivariate) Gaussian the probability of seeing a sample with a Mahalanobis score less than $t$ with $t > 0$ is given by the CDF $F_D$ of the chi-square distribution as

$$1 - \text{FPR} = P(M < t) = P(M^2 < t^2) = F_D(t^2) = \frac{\gamma\left(\frac{D}{2}, \frac{n}{2}\right)}{\Gamma\left(\frac{D}{2}\right)}.$$  \hfill (7)

Solving for $t$, the AD threshold is obtainable using the inverse CDF for any desired FPR.

$$t = \sqrt{F_D^{-1}(1 - \text{FPR})}.$$  \hfill (8)

IV. Experiments and Results

First, we assess the suitability of deep features extracted at various stages of a pre-trained classifier model for AD. We employ EfficientNet, which achieves state-of-the-art accuracy on ImageNet classification [32], as well as ResNet [28], a commonly applied model in research, as architecture variants. We extract features at the end of every model block “level” to assess which feature level gives the best performance. Here, “level” is defined as in [28], [32] (cf. Appendix Table VI for EfficientNet-B0). We argue that class probabilities are too application-specific and therefore make use of the features before the final mapping in the highest level. As feature probability maps may contain spatial dimensions in earlier levels, aggregation is necessary. We choose simple average pooling to reduce the complexity of our approach, but it should be noted that dedicated aggregation procedures may be an avenue of future research, especially for smaller anomalies. To increase repeatability of our work, we utilize pre-trained models provided publicly by others\(^1\). The overall approach is depicted in Fig. 1.

We now compare our approach to two different assumptions: (I) When assuming a fixed-variance univariate Gaussian distribution, the anomaly score reduces to the simple $L_2$-distance to the mean of the training set. (II) When assuming a feature-

\(^1\)Torchvision for ResNet variants and model weights hosted by Melas for EfficientNet variants [33]

Table I  Feature level AUROC (± SEM) scores in percent for EfficientNet-B4 using different normal distributions.

| Level | $L_2$ Mean (SEM) | SED Mean (SEM) | Mahalanobis Mean (SEM) |
|-------|-----------------|----------------|-----------------------|
| 1     | 44.5 (4.8)      | 51.6 (5.7)     | 60.3 (6.1)            |
| 2     | 47.3 (5.3)      | 48.1 (5.1)     | 62.0 (6.4)            |
| 3     | 58.1 (5.9)      | 59.2 (6.3)     | 71.1 (5.4)            |
| 4     | 59.7 (4.7)      | 61.5 (5.1)     | 75.6 (5.5)            |
| 5     | 62.6 (4.8)      | 66.1 (5.0)     | 82.1 (4.6)            |
| 6     | 71.7 (4.4)      | 74.3 (4.3)     | 89.1 (3.1)            |
| 7     | 82.9 (4.3)      | 85.1 (4.0)     | 96.7 (1.0)            |
| 8     | 83.2 (3.7)      | 85.2 (3.4)     | 95.5 (1.1)            |
| 9     | 83.3 (3.7)      | 87.8 (3.0)     | 93.1 (1.7)            |
| Sum   | 75.3 (4.5)      | 79.5 (6.6)     | 94.8 (1.6)            |

Fig. 1. Anomaly Detection using pre-trained deep feature representations. Fitting a multivariate Gaussian to features extracted from every level of an ImageNet pre-trained model and subsequently applying the Mahalanobis distance as anomaly score followed by their unweighted summation yields a simple yet effective Anomaly Detection algorithm. The figure depicts this procedure for the EfficientNet-B0 architecture.

## Independent Univariate Gaussian

An anomaly score can be defined with the standardized Euclidean distance (SED):

$$S(x) := \sqrt{\sum_{d=0}^{D} \frac{(f_d(x) - \bar{f}_d)^2}{s_d^2}}.$$  \hfill (9)

Here, $s_d$ is the (empirical) standard deviation of the $d$-th feature in the training set, and $\bar{f}_d$ is the mean.

To increase robustness of our evaluation, we perform a 5-fold evaluation over the original training dataset of each MVTec category, where we compute the necessary characteristics for each fold, respectively, and apply the scores to the test set of MVTec AD. To fully assess the capability of the approach, we compute and compare the AUROC metric, a commonly employed measure for binary classification problems. It focuses on overall possible performance by neglecting the task of finding the working point [34]. We report mean ± SEM AUROC performance over all categories and folds in percent. In addition to the feature level performances, we also report the AUROC performance yielded by summing distance scores over all levels. Further, note that the evaluation employs full feature spaces of the pre-trained models, i.e. no feature reduction is performed.

Assessing the performance of the three different normal distributions with their respected anomaly scores for EfficientNet-B4 in Table I, the following two observations can be made: (I) The multivariate Gaussian is best suited for AD due to its high and robust performance (with an AUROC of 96.7% ± 1.0% for level 7). (II) Deeper feature representations are more suitable for AD in a transfer learning setting. This is congruent with findings reported by [8], and reasons for this may be found in the increased abstraction level that is necessary to conclusively describe the distribution of normality. However, performance saturates (and even starts to decline) in higher levels (level 8 and 9) in case of the multivariate approach.

Comparing model architectures, features extracted from ResNet models yield worse performance as indicated by the lower average AUROC of 89.0% ± 3.0% for the best level 4 and 88.2% ± 4.0% for the sum predictor in ResNet-34 (cf. Appendix Table VII). The increased AD performance of EfficientNet may be attributed to its efficient architecture (i.e. higher ImageNet accuracy per trainable weight) and the output range of the Swish activation function [35]. In fact, SED score calculation with features extracted after the ReLU activation used in ResNet often failed as the activations for normal data are clipped to zero for some features.

Compared to OOD [30], no learned, linearly weighted sum of feature-level anomaly score is required to achieve strong performance. In fact, average AUROC of 94.8% ± 1.6% is achieved by simple equal weighting. While Hsu et al. [36] show that this linear weighting of feature level distributions is also not strictly necessary for OOD, their OOD approach still relies on input preprocessing by means of gradient ascent. It should also be noted that OOD, although similar to AD, still ultimately pursues a slightly different objective.

We also evaluate the influence of model complexity on AD performance of deep features in a transfer learning setting and apply the proposed method to all EfficientNet variants.

Analyzing performance across model complexities, it can be seen that features learned by less complex variants of EfficientNet (i.e. B0–B3, cf. Table II) perform worse in a transfer learning AD setting. Further, it can be seen that the performance saturates eventually, and even degrades for EfficientNet-B7. This could indicate that more complex EfficientNet variants start to overfit on ImageNet and no longer learn features that generalize well to new domains/use cases. A similar effect is observed in our evaluations with Mahalanobis distance on ResNet architectures (cf. Appendix Table VII).
the issue of choosing a working point, the multivariate

B. Choosing a Working Point Solely on FPR

While our evaluation has focussed on AUROC, neglecting
the issue of choosing a working point, the multivariate

Gaussian assumption also offers a theoretical framework for
selecting the working point by choosing an acceptable FPR out
of the box (cf. (8)). Note that a target FPR cannot be easily
set for the sum mode where the Mahalanobis distances of
feature-level multivariate Gaussians are added. Therefore,
we restrict our evaluations to level 7 features of two different EfficientNet variants, choosing EfficientNet-B0 for its low model complexity and EfficientNet-B4 for its high AD performance at medium complexity. We assess effects of performing no compression, 99% PCA and 0.01% NPCA compression. Here, we compare target FPR based on (8) to the FPR achieved on the test set, also reporting the TPR yielded by that working point and overall AUROC. We also assess the potentially beneficial effect of augmentations in order to to artificially enlarge small datasets for more robust covariance estimation. Augmentations are selected per MVTec category to avoid accidental transformation of normal to anomalous data (details can be found in the Appendix Fig. 2). We artificially increase each dataset’s size by aggregating over 100 epochs.

Performing the experiments, we observed that augmentations were essential to enable setting the working point that completely failed otherwise (e.g. FPR of 99.8% and TPR of 99.9% were achieved 3σ for EfficientNet-B0 at no compression). Looking at Table IV, it can be observed that PCA decreases FPRs yielded on the test set, whereas NPCA increases the FPRs. Therefore, PCA and NPCA behave inverse to each other, and PCA compression may prove useful in providing robust estimates of achieved FPRs at the cost of reduced AD performance. Furthermore, even with artificially enlarged datasets, a sensible setting of the FPR based on training data is possible only for the smallest model, EfficientNet-B0. This indicates the curse of dimensionality, as complex models require increasingly more data to avoid overfitting on noise present in the training data (bias-variance tradeoff). Thus, experiments should be reevaluated on larger AD datasets to confirm our findings. Still, setting the working point by means of a FPR can be realized via this theoretical framework. This is novel, as purely empirical approaches dominate the current literature (i.e. setting working point based on a hold-out validation set before applying to the test set).

### Table II

| Level | EfficientNet-B0 | EfficientNet-B1 | EfficientNet-B2 | EfficientNet-B3 | EfficientNet-B4 | EfficientNet-B5 | EfficientNet-B6 | EfficientNet-B7 |
|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|       | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      | Mean ± SEM      |
| 1     | 56.8 ± 6.0      | 55.7 ± 6.0      | 59.9 ± 6.1      | 60.1 ± 6.3      | 60.3 ± 6.1      | 61.5 ± 6.2      | 62.1 ± 6.3      | 61.0 ± 6.2      |
| 2     | 62.3 ± 5.7      | 58.2 ± 6.0      | 59.5 ± 5.7      | 62.0 ± 6.2      | 62.0 ± 6.4      | 63.7 ± 6.5      | 63.2 ± 6.4      | 61.6 ± 7.0      |
| 3     | 68.4 ± 6.0      | 67.8 ± 6.1      | 68.4 ± 5.9      | 70.1 ± 6.2      | 71.1 ± 5.4      | 69.5 ± 6.2      | 70.6 ± 5.7      | 71.2 ± 5.8      |
| 4     | 73.8 ± 5.4      | 73.6 ± 5.7      | 75.2 ± 5.2      | 73.5 ± 5.6      | 75.6 ± 5.5      | 76.5 ± 5.4      | 75.1 ± 5.9      | 76.8 ± 5.2      |
| 5     | 79.1 ± 5.3      | 81.0 ± 4.8      | 82.7 ± 4.9      | 82.1 ± 5.1      | 82.1 ± 4.6      | 83.9 ± 4.3      | 81.7 ± 4.8      | 82.4 ± 4.8      |
| 6     | 86.1 ± 4.1      | 87.1 ± 3.9      | 89.1 ± 3.6      | 91.2 ± 2.8      | 89.1 ± 3.1      | 89.0 ± 3.0      | 88.1 ± 3.1      | 87.5 ± 3.6      |
| 7     | 92.5 ± 2.3      | 95.3 ± 1.4      | 95.5 ± 1.4      | 96.4 ± 1.2      | 96.7 ± 1.0      | 96.9 ± 1.2      | 96.7 ± 1.1      | 96.3 ± 1.6      |
| 8     | 92.2 ± 2.5      | 94.7 ± 1.5      | 94.7 ± 1.6      | 94.8 ± 1.7      | 95.5 ± 1.1      | 96.2 ± 1.2      | 95.7 ± 1.0      | 95.7 ± 1.3      |
| 9     | 91.3 ± 3.0      | 93.4 ± 2.0      | 93.1 ± 2.0      | 92.8 ± 2.1      | 93.1 ± 1.7      | 93.0 ± 1.9      | 93.4 ± 1.5      | 92.2 ± 1.8      |
| Sum   | 90.6 ± 3.2      | 93.3 ± 2.2      | 93.6 ± 2.1      | 94.0 ± 2.2      | 94.8 ± 1.6      | 95.2 ± 1.6      | 95.3 ± 1.2      | 94.2 ± 1.8      |
C. Comparison with State of the Art on MVTec

Finally, we compare the performance of our proposed AD approach with state-of-the-art AD algorithms on the MVTec dataset. Our evaluation comprises a semi-supervised reconstruction approach using a convolutional AE, a fully-supervised AD classifier as well as an oc-SVM fit to the pre-trained feature representations of EfficientNet. We further compare to the current state-of-the-art performance reported in literature on MVTec, taking the corresponding values directly from the linked sources.

While the fully-supervised classifier can not be deployed in practice to AD problems, it serves as an upper bound of what can be achieved by AD algorithms. Here, we fine-tune a pre-trained EfficientNet-B4, EfficientNet-B2 as well as ResNet-18 and ResNet-34 variants per category. For data splits, we still perform a 5-fold evaluation, but no longer adhere to the original MVTec splits, as there are no anomalies present in the train datasets. Instead, we pool both train and test set and stratify splits to maintain identical anomaly prevalence in all folds. We compute AUROC on a val set split from the train set to select the best model state. The best model is then applied to the unused test set. For training, we select and apply the same augmentations per-category as used to artificially enlarge dataset size (Appendix Fig. 2). We use a batch-size of 64 for ResNet, 16 for EfficientNet and train using the Adam [37] optimizer with an initial learning rate of 0.0001 employing the binary cross-entropy loss function.

For the AE, we choose the ResNet-18 for the encoder and an inverted ResNet-18 for the decoder part (i.e. every operation of the encoder should be inverted by the decoder). For the upsampling operations we employ pixel shuffle operations as introduced by Shi et al. [38] to reduce checkerboard artifacts which would be present otherwise. The latent dimension of the bottleneck is set to 32 and yields proper reconstruction of the normal class in all categories. We also generally employ the same augmentations as before, but disable noise augmentations (cf. Appendix Fig. 2). Batch-size, learning rate and optimizer are the same as for the supervised classifier, and based on preliminary experiments the $L_2$-distance is chosen for the reconstruction error. As stated before, an aggregation of the residual image to image-level information is necessary for reconstruction-based approaches. While the threshold employed for ROC calculation is set on the pixel level, we perform connected component analysis and label a test image as defective only if it contains a connected component at least as big as the smallest anomaly present in the test dataset. Note that by extracting the minimal anomaly size from the test dataset, knowledge about the process is introduced to the AE approach, increasing complexity of the procedure.

To enhance comparability, we also apply augmentation to covariance estimation and compute features across 100 epochs of normal training data per split. To demonstrate the general applicability of our approach, we evaluate the proposed approach in sum mode over all feature levels of EfficientNet-B4, as we can achieve comparable performance and further reduce complexity by omitting feature level selection.

For the oc-SVM, we fit a RBF-kernel model to every feature level using normal data only and aggregate the predicted anomaly score over all levels to yield a sum score similar to the proposed pipeline.

Assessing performance results, it becomes apparent that multivariate Gaussian estimation on pre-trained deep features vastly outperforms the state of the art on MVTec, achieving 10% higher average AUROC than the next best model SPADE (cf. Table V). Notably, SPADE as proposed by Cohen et al. [9] also leverages pre-trained deep feature spaces in combination with further lifting the performance of the proposed pipeline.

### Table III

| Level | No Compression | PCA 99% Mean | PCA 95% Mean | NPCA 1% Mean | NPCA 0.1% Mean | NPCA 0.01% Mean |
|-------|----------------|--------------|--------------|--------------|----------------|-----------------|
|       | Mean SEM       | Mean SEM     | Mean SEM     | Mean SEM     | Mean SEM       | Mean SEM        |
| 1     | 60.3 6.1       | 50.8 6.1     | 45.8 5.4     | 64.1 6.4     | 67.8 6.6       | 69.9 6.2        |
| 2     | 62.0 6.4       | 53.3 5.9     | 48.6 5.7     | 67.6 6.3     | 68.0 5.9       | 67.3 5.7        |
| 3     | 71.1 5.4       | 65.5 6.5     | 59.7 6.2     | 71.6 4.9     | 68.1 4.2       | 65.7 3.7        |
| 4     | 75.6 5.5       | 69.5 6.1     | 63.2 6.4     | 76.1 5.1     | 73.1 4.6       | 69.4 4.0        |
| 5     | 82.1 4.6       | 76.2 5.3     | 66.6 6.5     | 82.5 4.0     | 78.7 3.6       | 72.3 3.7        |
| 6     | 89.1 3.1       | 83.3 4.8     | 77.3 5.7     | 90.2 2.5     | 88.2 2.4       | 83.6 2.9        |
| 7     | 96.7 1.0       | 93.4 2.1     | 87.1 4.0     | 96.1 1.0     | 94.5 1.3       | 89.6 2.5        |
| 8     | 95.5 1.1       | 91.9 2.1     | 88.6 3.1     | 94.8 1.2     | 93.8 1.4       | 90.6 2.3        |
| 9     | 93.1 1.7       | 91.3 2.1     | 88.6 2.9     | 93.3 1.6     | 91.2 2.1       | 86.3 3.0        |

### Table IV

| $n$ | Target FPR | EN-B0 PCA 99% | EN-B0 NPCA 1% | EN-B4 PCA 99% | EN-B4 NPCA 1% |
|-----|------------|---------------|---------------|---------------|---------------|
|     | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR |
| 1   | 31.7 | 17.9 | 77.8 | 30.4 | 89.1 | 47.4 | 94.0 | 66.1 | 98.4 |
| 2   | 4.6  | 10.1 | 71.2 | 19.3 | 84.2 | 29.8 | 88.9 | 56.0 | 97.3 |
| 3   | 0.3  | 5.8  | 66.1 | 13.5 | 80.2 | 18.9 | 83.2 | 44.7 | 95.9 |
| 4   | 6×10$^{-3}$ | 1.5 | 61.1 | 9.3 | 76.3 | 11.3 | 78.5 | 35.7 | 94.0 |
| 5   | 6×10$^{-5}$ | 0.7 | 57.7 | 6.5 | 72.7 | 6.8 | 74.4 | 28.2 | 91.6 |

### Table V

AUROC

The expected FPR and achieved FPR / TPR in percent per multiple $n \cdot \sigma$ for EfficientNet level 7 features under different compression modes. AUROC values are also reported.
with deep \( k \)-NN and \( L_2 \) loss. Also, NPCA slightly improves robustness of the method, leading to an increased average AUROC score and a decreased \( \text{SEM} \). Furthermore, performance is mostly comparable to fully-supervised fine-tuning of pre-trained classifiers but at times slightly worse (cf. per-category ROC score and a decreased \( \text{SEM} \)).

We stress again that fully-supervised AD cannot be realistically applied to most AD problems, and only serves as an upper limit of achievable AUROC performance.

### V. Discussion

We have demonstrated the benefits of using ImageNet pre-training for general-purpose AD in images. In particular, we showed that the multivariate Gaussian assumption of high correlation in pre-trained deep features is crucial to attain state-of-the-art AD performance (cf. Table I). Therefore, the generative assumption proposed by Lee et al. [30] holds also in a transfer learning setting for AD in images semantically different to the training dataset and can be leveraged to improve AD. Our PCA analysis revealed that discriminative components contain little overall variance in normal data and should thus be difficult to learn. We therefore agree with Bergman et al. [27] that pre-trained feature spaces should be adopted for AD in a transfer learning approach instead of learning features from scratch using normal data only. We further expand upon this and conclude that the multivariate Gaussian assumption is crucial to fully recapitulate the characteristics of pre-trained deep feature representations and to realize their potential.

While good performance has been achieved on the MVTeC AD dataset, we expect that with increasing semantic distance to natural images (e.g. images of the medical domain), out-of-the-box AD performance will decrease. In such scenarios, pre-trained models could be used as starting points for fine-tuning on target domains, e.g. by employing the deep SVDD proposed by Ruff et al. [5]. The multivariate Gaussian assumption can be easily integrated here, as the hypersphere could be initialized by transforming the features with the inverse Cholesky decomposition of the estimated covariance.

While the unimodal multivariate Gaussian assumption proved to be crucial for the performance of our AD approach, AD may also occur in a multi-modal context [42]. We therefore plan to extend the presented AD approach to multi-modal normal distributions (e.g. by fitting a multivariate Gaussian Mixture Model (GMM) to the deep features of normal data) on a complex multi-modal in-house fabric dataset.

The multivariate Gaussian further offers a framework to set the working point by determining an acceptable FPR. However, while augmentations and low model complexities were shown to alleviate the mismatch between desired and achieved FPR, an evaluation on even larger AD datasets is required. Further, modifications to the model architecture may prove beneficial to enforce a normal distribution in deep feature representations, as we have seen from the difference in performance between ResNet and EfficientNet features. Here, Self-Normalizing Neural Networks (SNNs) provide a basis for learning Gaussian-distributed features [43]. Also, class-wise distributions may be explicitly constrained to follow Gaussians in deep feature spaces during ImageNet pretraining.

It should be noted that while the FPR may be selected, Probably Approximately Correct (PAC) style TPR guarantees, as required for security-critical use-cases, cannot be given by this approach and are a different field of research. Here, Liu et al. [44] achieved PAC-style guarantees for the TPR requiring well-defined anomaly distributions to do so.

### VI. Conclusion

We have achieved state-of-the-art performance on the public MVTeC AD dataset using deep feature representations extracted from classifiers pre-trained on ImageNet. Our approach is simple yet effective and consists of fitting a multivariate Gaussian to normal data in the pre-trained deep feature representations, using Mahalanobis distance as anomaly score. We have further investigated the reason behind the effectiveness of our approach. Using PCA, we reveal that principal components containing only little variance in normal data are ultimately those necessary for discriminating between normal and anomalous images. We argue that these features are difficult to learn from scratch using normal data only, and propose to instead use feature representations generated by large-scale discriminative training in a transfer learning setting assuming multivariate Gaussian distributions. Future research in image AD should focus on (I) increasing the generalizability of pre-trained features, (II) fine-tuning of transferred representations using the small available datasets as well as (III) extending the approach to AD tasks with multi-modal normal data distributions.

| Approach                | Architecture     | Mean     | SEM  |
|-------------------------|------------------|----------|------|
| GeoTrans [39] (source: [40]) | Wide-ResNet      | 67.2     | 4.7  |
| GANomaly [41] (source: [40]) | DCGAN           | 76.1     | 1.6  |
| ITAE [40]               | Custom           | 83.9     | 2.8  |
| SPADE [9]               | Wide-ResNet50-2  | 85.5     |      |

**Table V**

Comparison to the state of the art reported in literature. We report AUROC (=SEM) scores in percent. The AE approaches map-mean and CCA stand for score map mean and connected component analysis, respectively. Mahalanobis and oc-SVM approaches are summed over all feature levels. The highest AUROC amongst non-fully supervised methods is boldfaced.
### APPENDIX

#### TABLE VI
**EfficientNet-B0 baseline network** *(source: [32])*

| Stage | Operator          | Resolution | #Channels | #Layers |
|-------|-------------------|------------|-----------|---------|
| i     | $F_i$             | $H_i \times W_i$ | $C_i$ | $L_i$ |
| 1     | Conv3x3          | $224 \times 224$ | 32       | 1       |
| 2     | MBConv1, k3x3    | $112 \times 112$ | 16       | 1       |
| 3     | MBConv6, k3x3    | $112 \times 112$ | 24       | 2       |
| 4     | MBConv6, k5x5    | $56 \times 56$  | 40       | 2       |
| 5     | MBConv6, k3x3    | $28 \times 28$  | 80       | 3       |
| 6     | MBConv6, k5x5    | $14 \times 14$  | 112      | 3       |
| 7     | MBConv6, k5x5    | $14 \times 14$  | 192      | 4       |
| 8     | MBConv6, k3x3    | $7 \times 7$    | 320      | 1       |
| 9     | Conv1x1 & pool & FC | $7 \times 7$    | 1280     | 1       |

#### TABLE VII
**Feature level AUROC (± SEM) scores in percent for ResNet architectures with Mahalanobis distance**

| Level | ResNet-18 | | | ResNet-34 | | | ResNet-50 | | |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|       | Mean     | SEM       | Mean     | SEM       | Mean     | SEM       | Mean     | SEM       | Mean     |
| 1     | 66.6     | 6.4       | 67.5     | 6.8       | 68.0     | 5.9       |          |           |          |
| 2     | 71.6     | 6.4       | 71.9     | 6.1       | 73.7     | 6.4       |          |           |          |
| 3     | 78.6     | 5.0       | 79.4     | 4.8       | 81.3     | 4.9       |          |           |          |
| 4     | 86.7     | 4.1       | 90.4     | 3.6       | 89.0     | 4.3       |          |           |          |
| 5     | 88.3     | 2.7       | 89.0     | 3.0       | 86.9     | 3.8       |          |           |          |
| Sum   | 86.4     | 4.0       | 88.2     | 4.0       | 87.3     | 4.5       |          |           |          |
Fig. 2. Augmentation pipeline. Columns indicate mutually exclusive weighted random choices. Probability of activation shown below the lines.
### TABLE VIII
AUROC scores in percent for all models per MVTec AD category. The AE approaches map-mean and CCA stand for the image-level aggregation by averaging or connected component analysis respectively. ResNet and EfficientNet are abbreviated as RN and EN. Mahalanobis approaches and oc-SVM are averaged over all levels to provide the simplest approach. The highest AUROC score per row is highlighted in bold.

| Score | Architecture | MSE | AE | Pre-trained Classifier | Mahalanobis (ours) | oc-SVM |
|-------|--------------|-----|----|------------------------|--------------------|--------|
|       |              | Map-Mean | CCA | All Features | NPCA 1% |
|       |              | RN-18 | RN-18 | RN-34 | EN-B0 | EN-B4 | EN-B4 | EN-B0 | EN-B4 |
| Textures |              |       |       |       |       |       |       |       |       |
| Carpet |              | 65.9  | 80.5  | 97.5  | 96.8  | 97.2  | 98.9  | 100.0 | 100.0 | 61.1  | 89.2  |
| Grid   |              | 80.9  | 92.9  | 90.2  | 89.3  | 97.8  | 98.4  | 81.7  | 89.7  | 23.7  | 44.7  |
| Leather|              | 46.0  | 90.5  | 99.8  | 98.4  | 98.8  | 99.1  | 99.7  | 100.0 | 75.8  | 86.7  |
| Tile   |              | 55.4  | 75.7  | 97.7  | 99.1  | 99.2  | 97.8  | 99.8  | 99.8  | 91.6  | 95.7  |
| Wood   |              | 91.8  | 91.4  | 98.1  | 95.0  | 95.6  | 99.6  | 98.6  | 99.6  | 92.5  | 79.1  |
| Objects |              |       |       |       |       |       |       |       |       |       |       |
| Bottle |              | 79.5  | 58.3  | 91.1  | 90.8  | 92.2  | 92.0  | 95.5  | 95.0  | 78.3  | 81.0  |
| Cable  |              | 74.5  | 76.7  | 92.4  | 93.0  | 89.0  | 96.3  | 93.8  | 95.1  | 66.1  | 73.0  |
| Capsule|              | 90.0  | 91.1  | 98.6  | 98.8  | 98.4  | 99.8  | 99.6  | 99.1  | 83.5  | 81.7  |
| Hazelnut |            | 58.2  | 64.6  | 95.6  | 95.4  | 94.6  | 96.8  | 94.7  | 94.7  | 73.4  | 77.3  |
| Metal Nut |            | 80.1  | 58.0  | 86.0  | 83.3  | 89.8  | 93.6  | 88.4  | 88.7  | 66.7  | 69.0  |
| Pill   |              | 95.7  | 93.7  | 85.0  | 88.8  | 90.3  | 95.2  | 85.4  | 85.2  | 19.5  | 31.0  |
| Screw  |              | 94.2  | 100.0 | 82.0  | 86.9  | 80.7  | 85.4  | 96.4  | 96.9  | 90.5  | 86.3  |
| Toothbrush |            | 85.9  | 79.1  | 92.4  | 89.7  | 90.0  | 94.1  | 96.3  | 95.5  | 82.0  | 83.8  |
| Transistor |           | 86.8  | 88.8  | 97.9  | 98.3  | 98.8  | 97.9  | 97.8  | 97.9  | 91.7  | 94.7  |
| Mean   |              | 78.8  | 81.8  | 93.3  | 93.4  | 94.1  | 96.3  | 95.2  | 95.8  | 73.0  | 78.1  |
| SEM    |              | 4.1   | 3.4   | 1.4   | 1.3   | 1.4   | 1.0   | 1.5   | 1.2   | 6.1   | 4.7   |