resolution than the input image, many pixels have the same embeddings and then form pixel patches with no overlap in the original image resolution. Hence, an input image can be divided in a grid of $(i, j) \in [1, W] \times [1, H]$ positions where WxH is the resolution of the largest activation map used to generate embeddings. Finally, each patch position (i, j) in this grid is associated to an embedding vector x_{ij} computed as described above.

The generated patch embedding vectors may carry redundant information, therefore we experimentally study the possibility to reduce their size (Section V-A). We noticed that randomly selecting few dimensions is more efficient that a classic Principal Component Analysis (PCA) algorithm [30]. This simple random dimensionality reduction significantly decreases the complexity of our model for both training and testing time while maintaining the state-of-the-art performance. Finally, patch embedding vectors from test images are used to output an anomaly map with the help of the learned parametric representation of the normal class described in the next subsection.

B. Learning of the normality

To learn the normal image characteristics at position (i, j), we first compute the set of patch embedding vectors at (i, j), $X_{ij} = \{x_{ij}^k, k \in [\![1, N]\!]\}$ from the N normal training images as shown on Figure 2. To sum up the information carried by this set we make the assumption that X_{ij} is generated by a multivariate Gaussian distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$ where μ_{ij} is the sample mean of X_{ij} and the sample covariance Σ_{ij} is estimated as follows :

$$\Sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathrm{x_{ij}^k} - \mu_{\mathrm{ij}})(\mathrm{x_{ij}^k} - \mu_{\mathrm{ij}})^{\mathrm{T}} + \boldsymbol{\epsilon} I \qquad (1)$$

where the regularisation term $\boldsymbol{\epsilon} I$ makes the sample covariance matrix Σ_{ij} full rank and invertible. Finally, each possible patch position is associated with a multivariate Gaussian distribution as shown in Figure 2 by the matrix of Gaussian parameters.

Our patch embedding vectors carry information from different semantic levels. Hence, each estimated multivariate Gaussian distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$ captures information from different levels too and $\boldsymbol{\Sigma_{ij}}$ contains the inter-level correlations. We experimentally show (Section V-A) that modeling these relationships between the different semantic levels of the pretrained CNN helps to increase anomaly localization performance.

C. Inference : computation of the anomaly map

Inspired by [23], [26], we use the Mahalanobis distance [31] $M(x_{ij})$ to give an anomaly score to the patch in position (i, j) of a test image. $M(x_{ij})$ can be interpreted as the distance between the test patch embedding x_{ij} and learned distribution $\mathcal{N}(\mu_{ij}, \boldsymbol{\Sigma_{ij}})$, where $M(x_{ij})$ is computed as follows:

$$M(x_{ij}) = \sqrt{(x_{ij} - \mu_{ij})^T \Sigma_{ij}^{-1} (x_{ij} - \mu_{ij})} \qquad (2)$$

Hence, the matrix of Mahalanobis distances $M = (M(x_{ij}))_{1<i<W, 1<j<H}$ that forms an anomaly map can be computed. High scores in this map indicate the anomalous areas. The final anomaly score of the entire image is the maximum of anomaly map M. Finally, at test time, our method does not have the scalability issue of the K-NN based methods [4]–[6], [25] as we do not have to compute and sort a large amount of distance values to get the anomaly score of a patch.

IV. Experiments

A. Datasets and metrics

Metrics. To assess the localization performance we compute two threshold independent metrics. We use the Area Under the Receiver Operating Characteristic curve (AUROC) where the true positive rate is the percentage of pixels correctly classified as anomalous. Since the AUROC is biased in favor of large anomalies we also employ the per-region-overlap score (PRO-score) [2]. It consists in plotting, for each connected component, a curve of the mean values of the correctly classified pixel rates as a function of the false positive rate between 0 and 0.3. The PRO-score is the normalized integral of this curve. A high PRO-score means that both large and small anomalies are well-localized.

Datasets. We first evaluate our models on the MVTec AD [1] designed to test anomaly localization algorithms for industrial quality control and in a one-class learning setting. It contains 15 classes of approximately 240 images. The original image resolution is between 700x700 and 1024x1024. There are 10 object and 5 texture classes. Objects are always well-centered and aligned in the same way across the dataset as we can see in Figure 1 for classes Transistor and Capsule. In addition to the original dataset, to assess performance of anomaly localization models in a more realistic context, we create a modified version of the MVTec AD, referred as Rd-MVTec AD, where we apply random rotation (-10, +10) and random crop (from 256x256 to 224x224) to both the train and test sets. This modified version of the MVTec AD may better describe real use cases of anomaly localization for quality control where objects of interest are not always centered and aligned in the image.

For further evaluation, we also test PaDiM on the Shanghai Tech Campus (STC) Dataset [8] that simulates video surveillance from a static camera. It contains 274 515 training and 42 883 testing frames divided in 13 scenes. The original image resolution is 856x480. The training videos are composed of normal sequences and test videos have anomalies like the presence of vehicles in pedestrian areas or people fighting.

B. Experimental setups

We train PaDiM with different backbones, a ResNet18 (R18) [27], a Wide ResNet-50-2 (WR50) [28] and an EfficientNet-B5 [29], all pretrained on ImageNet [32]. Like in [5], patch embedding vectors are extracted from the first three layers when the backbone is a ResNet, in order to combine