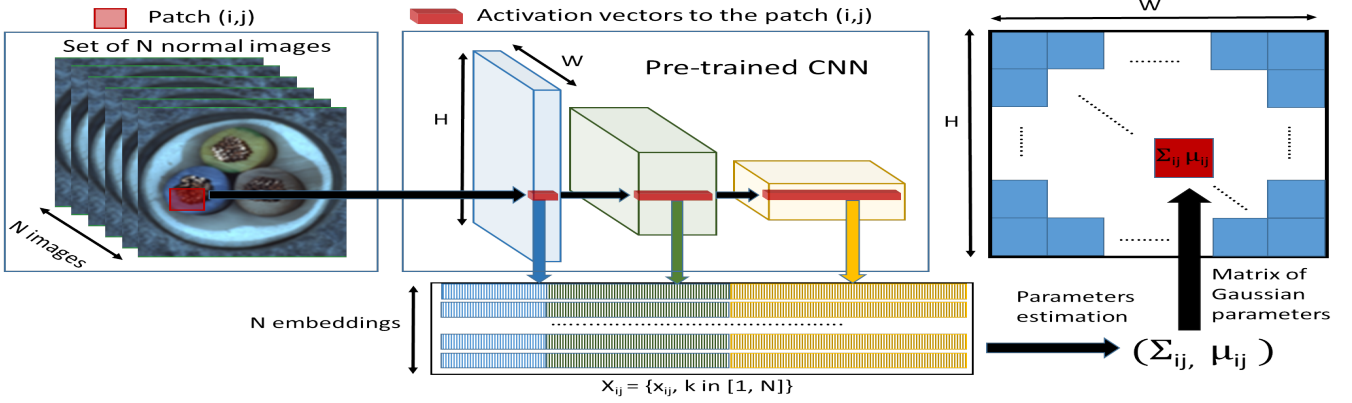


Fig. 2. For each image patch corresponding to position (i, j) in the largest CNN feature map, PaDiM learns the Gaussian parameters (μ_{ij}, Σ_{ij}) from the set of N training embedding vectors $X_{ij} = \{x_{ij}^k, k \in [1, N]\}$, computed from N different training images and three different pretrained CNN layers.



ferent semantic levels of a pretrained CNN.

With this new and efficient approach, PaDiM outperforms the existing state-of-the-art methods for anomaly localization and detection on the MVTec AD [1] and the ShanghaiTech Campus (STC) [8] datasets. Besides, at test time, it has a low time and space complexity, independent of the dataset training size which is an asset for industrial applications. We also extend the evaluation protocol to assess model performance in more realistic conditions, *i.e.*, on a non-aligned dataset.

II. RELATED WORK

Anomaly detection and localization methods can be categorized as either reconstruction-based or embedding similarity-based methods.

Reconstruction-based methods are widely-used for anomaly detection and localization. Neural network architectures like autoencoders (AE) [1], [9]–[11], variational autoencoders (VAE) [3], [12]–[14] or generative adversarial networks (GAN) [15]–[17] are trained to reconstruct normal training images only. Therefore, anomalous images can be spotted as they are not well reconstructed. At the image level, the simplest approach is to take the reconstructed error as an anomaly score [10] but additional information from the latent space [16], [18], intermediate activations [19] or a discriminator [17], [20] can help to better recognize anomalous images. Yet to localize anomalies, reconstruction-based methods can take the pixel-wise reconstruction error as the anomaly score [1] or the structural similarity [9]. Alternatively, the anomaly map can be a visual attention map generated from the latent space [3], [14]. Although reconstruction-based methods are very intuitive and interpretable, their performance is limited by the fact that AE can sometimes yield good reconstruction results for anomalous images too [21].

Embedding similarity-based methods use deep neural networks to extract meaningful vectors describing an entire image for anomaly detection [6], [22]–[24] or an image patch for anomaly localization [2], [4], [5], [25]. Still, embedding similarity-based methods that only perform anomaly detection give promising results but often lack interpretability as it is

not possible to know which part of an anomalous images is responsible for a high anomaly score. The anomaly score is in this case the distance between embedding vectors of a test image and reference vectors representing normality from the training dataset. The normal reference can be the center of a n -sphere containing embeddings from normal images [4], [22], parameters of Gaussian distributions [23], [26] or the entire set of normal embedding vectors [5], [24]. The last option is used by SPADE [5] which has the best reported results for anomaly localization. However, it runs a K-NN algorithm on a set of normal embedding vectors at test time, so the inference complexity scales linearly to the dataset training size. This may hinder industrial deployment of the method.

Our method, PaDiM, generates patch embeddings for anomaly localization, similar to the aforementioned approaches. However, the normal class in PaDiM is described through a set of Gaussian distributions that also model correlations between semantic levels of the used pretrained CNN model. Inspired by [5], [23], we choose as pretrained networks a ResNet [27], a Wide-ResNet [28] or an EfficientNet [29]. Thanks to this modelisation, PaDiM outperforms the current state-of-the-art methods. Moreover, its time complexity is low and independent of the training dataset size at the prediction stage.

III. PATCH DISTRIBUTION MODELING

A. Embedding extraction

Pretrained CNNs are able to output relevant features for anomaly detection [24]. Therefore, we choose to avoid ponderous neural network optimization by only using a pretrained CNN to generate patch embedding vectors. The patch embedding process in PaDiM is similar to one from SPADE [5] and illustrated in Figure 2. During the training phase, each patch of the normal images is associated to its spatially corresponding activation vectors in the pretrained CNN activation maps. Activation vectors from different layers are then concatenated to get embedding vectors carrying information from different semantic levels and resolutions, in order to encode fine-grained and global contexts. As activation maps have a lower