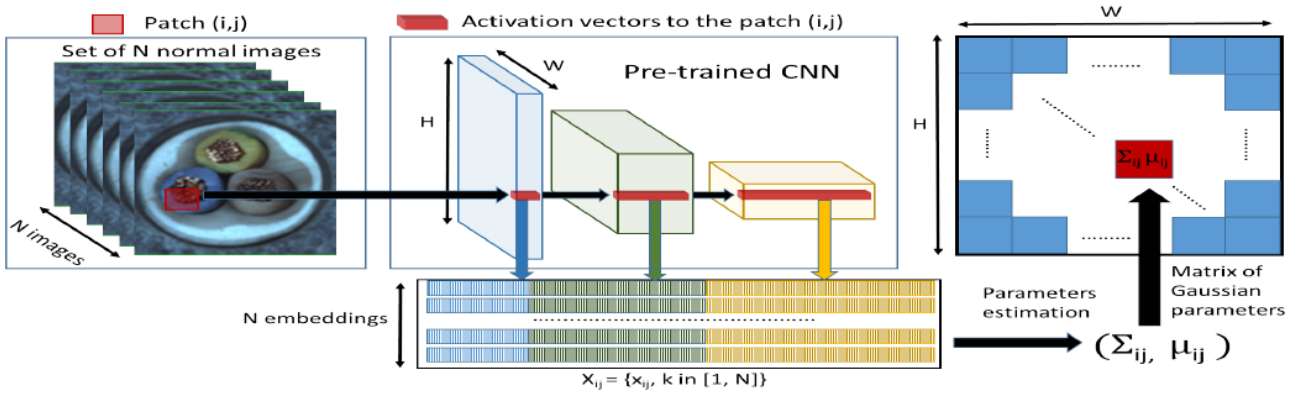
図2. 最も大きなCNN特徴マップ内の位置（i, j）に対応する各画像パッチに対し、PaDiMはNつのトレーニング画像と3つの異なる

ameters (μ_{ij}, Σ_{ij}) from the pretrained CNN layers.



事前学習済みCNNの異なるセマンティックレベル。

この新しい効率的なアプローチにより、PaDiMはMVTec AD [1] とShanghaiTech Campus (STC) [8] データセットにおける異常検出と局所化において、既存の最先端手法を凌駕します。さらに、テスト時における時間と空間の複雑さはデータセットのトレーニングサイズに依存せず、産業応用における利点となります。私たちは評価プロトコルを拡張し、より現実的な条件下でのモデル性能を評価するため、非一致データセット上で評価を実施しました。

II. Related Work

異常検出と局所化手法は、再構築ベースまたは埋め込み類似性ベースの手法に分類されます。

再構築ベースの手法は、異常検出と局所化に広く使用されています。オートエンコーダー（AE）[1]、[9]-[11]、変分オートエンコーダー（VAE）[3]、[12]-[14]、または生成対抗ネットワーク（GAN）[15]-[17]などのニューラルネットワークアーキテクチャは、正常なトレーニング画像のみを再構築するように訓練されます。したがって、異常な画像は適切に再構築されないため検出可能です。画像レベルでは、再構築誤差を異常スコアとして使用する最も単純なアプローチ[10]がありますが、潜在空間[16]、[18]、中間活性化[19]、またはディｃкриミネーター[17]、[20]からの追加情報により、異常な画像をより正確に認識できます。異常を局所化するには、再構築ベースの手法はピクセル単位の再構築誤差を異常スコアとして使用できます[1]または構造的類似性[9]。または、異常マップは潜在空間から生成された視覚的注意マップである可能性があります[3]、[14]。再構築ベースの手法は直感的で解釈可能ですが、AEが異常画像に対しても良い再構築結果を生成する可能性があるため、その性能は制限されます[21]。

埋め込み類似性ベースの手法は、異常検出[6]、[22]-[24]では画像全体を記述する意味のあるベクトルを抽出するために深層神経ネットワークを使用し、異常局所化[2]、[4]、[5]、[25]では画像パッチを記述するために使用します。しかし、異常検出のみを行う埋め込み類似性に基づく手法は有望な結果を示すものの、異常画像のどの部分が異常スコアの高さに寄与しているかを特定できないため、解釈可能性に欠ける場合があります。この場合の異常スコアは、テスト画像の埋め込みベクトルとトレーニングデータセットから正常性を表す参照ベクトルとの距離です。正常な参照は、正常な画像の埋め込みを含むn次元の球の中心[4]，[22]、ガウス分布のパラメーター[23]，[26]、または正常な埋め込みベクトルの全体集合[5]，[24]などです。最後のオプションは、異常局所化において最も優れた結果を報告しているSPADE [5] で使用されています。しかし、テスト時に正常な埋め込みベクトルのセットに対してK-NNアルゴリズムを実行するため、推論の複雑さはトレーニングデータセットのサイズに線形にスケールします。これは、この手法の産業展開を妨げる可能性があります。

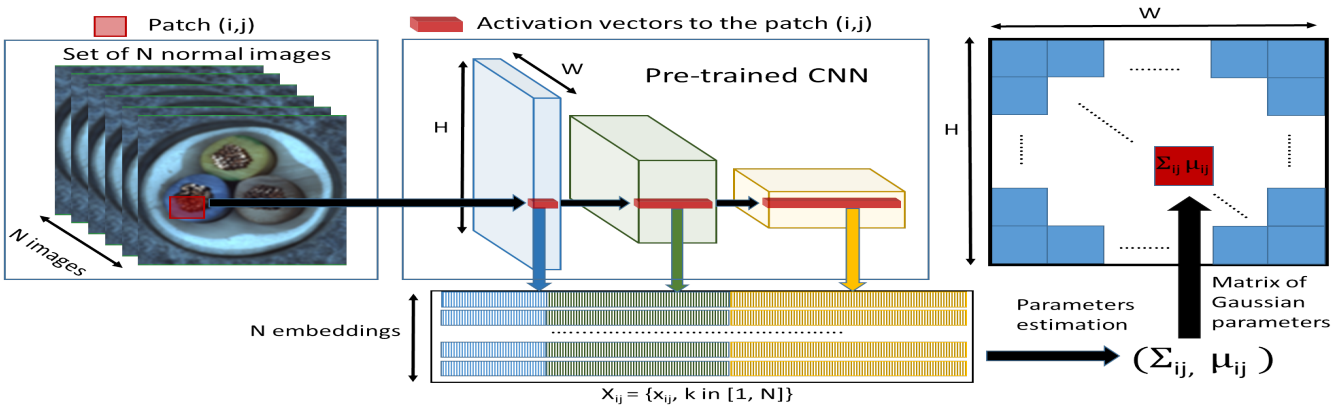
当社の手法であるPaDiMは、異常検出のためのパッチ埋め込みを生成する点で、前述の手法と類似しています。ただし、PaDiMにおける正常クラスは、使用される事前学習済みCNNモデルのセマンティックレベル間の相関関係をモデル化するガウス分布の集合を通じて記述されます。[5]、[23]にインスパイアされ、事前学習済みネットワークとしてResNet [27]、Wide-ResNet [28]、またはEfficientNet [29]を選択しています。このモデル化により、PaDiMは現在の最先端手法を凌駕しています。さらに、その時間複雑度は低く、予測段階ではトレーニングデータセットのサイズに依存しません。

III. パッチ分布モデリング

A. 埋め込み抽出

事前学習済みCNNは異常検出に適切な特徴を出力可能です[24]。そのため、事前学習済みCNNのみを使用してパッチ埋め込みベクトルを生成し、複雑なニューラルネットワーク最適化を回避しました。PaDiMのパッチ埋め込みプロセスはSPADE [5] のもの類似しており、図2に示されています。トレーニングフェーズにおいて、正常画像の各パッチは、事前学習済みCNNの活性化マップにおける空間的に対応する活性化ベクトルと関連付けられます。異なる層の活性化ベクトルを結合することで、異なるセマンティックレベルと解像度からの情報を保持する埋め込みベクトルを生成し、細粒度とグローバルな文脈をエンコードします。活性化マップはより低い

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