入力画像の解像度よりも高い解像度では、多くのピクセルが同じ埋め込みベクトルを持ち、元の画像解像度では重なり合わないピクセルパッチを形成します。したがって、入力画像は （i, j） ∈ [1, W] × [1, H] のグリッドに分割され、ここで W × H は埋め込みを生成するために使用される最大の活性化マップの解像度です。最後に、このグリッド内の各パッチ位置 （i, j） は、上記で説明したように計算された埋め込みベクトル x_{ij} と関連付けられます。

生成されたパッチ埋め込みベクトルには冗長な情報が含まれている可能性があるため、そのサイズを削減する可能性を実験的に検討しました（セクションV-A）。ランダムにいくつかの次元を選択する方法は、古典的な主成分分析（PCA）アルゴリズム [30] よりも効率的であることが判明しました。この単純なランダム次元削減は、トレーニング時間とテスト時間の両方でモデルの複雑さを大幅に削減しつつ、最先端の性能を維持します。最後に、テスト画像のパッチ埋め込みベクトルは、次節で説明する正常クラスの学習済みパラメトリック表現の助けを借りて、異常マップを出力するために使用されます。

B. 正常性の学習

位置 （i, j） における正常画像の特徴を学習するため、まず図2に示すように、N枚の正常トレーニング画像から位置 （i, j） におけるパッチ埋め込みベクトルの集合 X_{ij} = {x_{kij} , k ∈ [[1, N]] } を計算します。この集合が持つ情報を要約するため、X_{ij} が多変量ガウス分布 N （μ_{ij} , Σ_{ij} ） によって生成されたと仮定します。ここで、μ_{ij} は X_{ij} のサンプル平均であり、サンプル共分散 Σ_{ij} は次のように推定されます：

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

正則化項Iは、サンプル共分散行列Σ_{ij} をフルランクかつ逆行列可能な状態にします。最後に、各可能なパッチ位置は、図2に示すガウスパラメータ行列によって表される多変量ガウス分布と関連付けられます。

当社のパッチ埋め込みベクトルは、異なるセマンティックレベルからの情報を保持しています。したがって、推定された多変量ガウス分布 N （μ_{ij} , Σ_{ij} ） も異なるレベルからの情報を捕捉し、Σ_{ij} にはレベル間の相関が含まれます。実験的に示したように（セクション V-A）、事前訓練されたCNNの異なるセマンティックレベル間の関係をモデル化することは、異常局所化性能の向上に役立ちます。

C. 推論：異常マップの計算

[23]、[26] に倣い、テスト画像の（i, j）位置のパッチに異常スコアを付与するために、マハラノビス距離 [31] M（x_{ij} ）を使用します。M （x_{ij} ） は、テストパッチの埋め込み x_{ij} と学習された分布 N （μ_{ij} , Σ_{ij} ） との間の距離と解釈できます。M （x_{ij} ） は次のように計算されます：

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

したがって、マハラノビス距離の行列 M = （M （x_{ij} ） ）_{1<i<W,1<j<H} は異常マップを形成し、計算可能です。このマップでの高スコアは異常領域を示します。画像全体の最終的な異常スコアは、異常マップ M の最大値です。最後に、テスト時において、当手法は K-NN ベースの手法 [4] - [6] , [25] のようなスケーラビリティの問題を抱えていません。なぜなら、パッチの異常スコアを計算するために大量の距離値を計算して並べ替える必要がないからです。

IV. Experiments

A. データセットとメトリクス

メトリクス。局在化性能を評価するため、2つの閾値に依存しないメトリクスを計算します。受信者動作特性曲線（ROC曲線）下の面積（AUROC）を使用し、真陽性率は異常と正しく分類されたピクセルの割合です。AUROCは大きな異常値に偏るため、地域重なりスコア（PRO-score） [2] も採用しています。これは、各接続成分に対して、偽陽性率0から0.3の範囲で正しく分類されたピクセル率の平均値をプロットした曲線から、正規化された積分値をPRO-scoreとします。高いPROスコアは、大規模な異常と小規模な異常の両方が適切に局所化されていることを意味します。

データセット。まず、産業用品質管理における異常検出アルゴリズムの評価を目的とした MVTec AD [1] データセットで、単一クラス学習設定においてモデルを評価します。このデータセットには、約 240 枚の画像からなる 15 のクラスが含まれています。元の画像解像度は700x700から1024x1024です。10のオブジェクトクラスと5のテクスチャクラスが存在します。オブジェクトはデータセット全体で常に中央に配置され、同じ方向に整列されています。これは図1のTransistorとCapsuleクラスで確認できます。元のデータセットに加え、異常検出モデルの性能をより現実的な文脈で評価するため、MVTec ADの改変版であるRdMVTec ADを作成しました。この改変版では、トレーニングセットとテストセットの両方にランダムな回転（-10~+10度）とランダムなクロップ（256×256から224×224まで）を適用しています。このMVTec ADの改変版は、品質管理における異常局在化の実際の使用ケースをより適切に表現する可能性があります。特に、関心対象のオブジェクトが画像内で常に中心に配置されず、整列していない場合です。

さらに評価するため、当社はPaDiMを静止カメラからのビデオ監視を模擬するShanghai Tech Campus （STC） データセット [8] でテストしました。このデータセットには、13のシーンに分割された274,515のトレーニングフレームと42,883のテストフレームが含まれています。元の画像解像度は856×480です。トレーニング動画は通常のシーケンスで構成され、テスト動画には歩行者区域での車両の出現や人々の喧嘩などの異常が含まれます。

B. 実験設定

PaDiMは、ImageNet [32] で事前学習された異なるバックボーン（ResNet18（R18） [27] 、Wide ResNet-50-2（WR50） [28] 、EfficientNet-B5 [29] ）で訓練されます。 [5] と同様に、バックボーンがResNetの場合、パッチ埋め込みベクトルは最初の3層から抽出され、

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 :

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

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