異なるセマンティックレベルからの情報を保持しつつ、局所化タスクに適した十分な解像度を維持します。このアイデアに基づき、EfficientNet-B5を使用する場合、層7（レベル2）、層20（レベル4）、層26（レベル5）からパッチ埋め込みベクトルを抽出します。また、ランダム次元削減（Rd）を適用します（セクションIII-AおよびV-A参照）。私たちのモデル名は、バックボーンと次元削減方法（使用する場合）を示しています。例えば、PaDiM-R18-Rd100は、ResNet18バックボーンを使用し、パッチ埋め込みベクトルに100のランダムに選択された次元を使用するPaDiMモデルです。デフォルトでは、式1から=0.01を使用します。

オリジナル論文で説明されているように、SPADE [5] モデルを再現し、バックボーンとしてWide ResNet-50-2 (WR50) [28] を使用します。SPADEとPaDiMには、[5] で使用されたのと同じ前処理を適用します。MVTec ADの画像を256x256にリサイズし、224x224に中央切り出しを行います。STCの画像には256x256のリサイズのみを適用します。画像と局所化マップはバイキュービック補間を使用してリサイズし、異常マップには [5] と同様にパラメーター $\sigma = 4$ のガウスフィルターを適用します。

私たちは、エンコーダーとしてResNet18を使用し、8×8の畳み込み潜在変数を持つ再構築ベースのベースラインとして独自のVAEを実装しました。各MVTec ADクラスに対して10,000枚の画像を使用し、以下のデータ拡張操作を実施してトレーニングを行います：ランダム回転（-2^{◦}、+2^{◦}）、292×292へのリサイズ、ランダムクロップで282×282に切り出し、最後にセンタークロップで256×256に切り出し。トレーニングは、Adamオプティマイザー [12] を使用し、初期学習率10^{-4}、バッチサイズ32画像で100エポック実施されました。局所化に対応する異常マップは、再構築のピクセル単位のL2誤差に対応します。

V. Results

A. アブレーション研究

まず、PaDiMにおけるセマンティックレベル間の相関をモデル化することの影響を評価し、次元削減を通じて手法を簡素化する可能性を探ります。

層間相関。ガウスモデルとマハラノビス距離の組み合わせは、以前の研究で敵対的攻撃の検出 [26] や画像レベルでの異常検出 [23] に既に採用されています。しかし、これらの方法はPaDiMで実施しているように、異なるCNNのセマンティックレベル間の相関をモデル化していません。表Iでは、ResNet18バックボーンを使用したPaDiMのMVTec ADにおける異常局所化性能を示しています。最初の3層（層1、層2、または層3）のいずれか1層のみを使用する場合と、これらの3モデルの出力合計をアンサンブル手法として用いた場合（最初の3層を考慮するが、層間の相関は考慮しない：層1+2+3）を比較しています。表Iの最終行（PaDiM-R18）は、提案するPaDiMのバージョンで、各パッチの位置は最初の3つのResNet18層とそれらの相関を考慮した1つのガウス分布で記述されます。3つの層のうち、Layer 3を使用した場合がAUROCにおいて最も良い結果を示すことが観察されます。これは、レイヤー3がより高いセマンティックレベルの情報を含み、正常性をより適切に記述するためです。

TABLE I
STUDY OF THE ANOMALY LOCALIZATION PERFORMANCE USING DIFFERENT SEMANTIC-LEVEL CNN LAYERS. RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%) ON THE MVTEC AD.

Layer used	all texture classes	all object classes	all classes
Layer 1	(93.1, 87.1)	(95.6, 86.5)	(94.8, 86.8)
Layer 2	(95.0, 89.7)	(96.1, 87.9)	(95.7, 88.5)
Layer 3	(94.8, 89.6)	(97.1, 87.7)	(95.7, 88.3)
Layer 1+2+3	(95.4, 90.7)	(96.3, 88.1)	(96.0, 89.0)
PaDiM-R18	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

ただし、レイヤー3のPROスコアはレイヤー2よりもやや劣っており、これはレイヤー2の解像度が低く、異常局在化の精度に影響を与えるためです。表Iの最後の2行で示されるように、異なるレイヤーからの情報を集約することで、高いセマンティック情報と高い解像度とのトレードオフ問題を解決できます。モデルLayer 1+2+3が単純に出力を加算するのに対し、当社のモデルPaDiMR18は意味論的レベル間の相関を考慮します。その結果、AUROCで1.1ポイント、PROスコアで1.8ポイント、Layer 1+2+3を上回ります。これは、意味論的レベル間の相関をモデル化することの重要性を確認しています。

TABLE II
STUDY OF THE ANOMALY LOCALIZATION PERFORMANCE WITH A DIMENSIONALITY REDUCTION FROM 448 TO 100 AND 200 USING PCA OR RANDOM FEATURE SELECTION (RD). RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%) ON THE MVTEC AD.

	all texture classes	all object classes	all classes
Rd 100	(95.7, 91.3)	(97.2, 89.4)	(96.7, 90.5)
PCA 100	(93.7, 88.9)	(93.5, 84.1)	(93.5, 85.7)
Rd 200	(96.1, 92.0)	(97.5, 89.8)	(97.0, 90.5)
PCA 200	(95.1, 91.8)	(96.0, 88.1)	(95.7, 89.3)
all (448)	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

次元削減。PaDiM-R18は、各448次元のパッチ埋め込みベクトルの集合から多変量ガウス分布を推定します。埋め込みベクトルのサイズを縮小することで、モデルの計算とメモリの複雑さを軽減できます。私たちは2つの異なる次元削減方法を検討しました。最初の方法は、主成分分析（PCA）アルゴリズムを適用してベクトルサイズを100または200次元へ削減するものです。2つ目の方法は、トレーニング前にランダムに特徴を選択するランダム特徴選択です。この場合、10つの異なるモデルをトレーニングし、平均スコアを算出します。ただし、ランダム性により異なるシード間で結果が変化することはありません。平均AUROCの標準誤差平均（SEM）は常に10^{-4}から10^{-7}の間です。

表IIから、同じ次元数において、ランダム次元削減（Rd）はMVTec ADのすべてのクラスでPCAをAUROCで少なくとも1.3ポイント、PROスコアで1.2ポイント上回ることがわかります。これは、PCAが最も分散の大きい次元を選択するため、正常クラスと異常クラスを区別するのに役立つ次元ではない可能性があるためです [23]。

information from different semantic levels, while keeping a high enough resolution for the localization task. Following this idea, we extract patch embedding vectors from layers 7 (level 2), 20 (level 4), and 26 (level 5), if an EfficientNet-B5 is used. We also apply a random dimensionality reduction (Rd) (see Sections III-A and V-A). Our model names indicate the backbone and the dimensionality reduction method used, if any. For example, PaDiM-R18-Rd100 is a PaDiM model with a ResNet18 backbone using 100 randomly selected dimensions for the patch embedding vectors. By default we use $\epsilon = 0.01$ for the ϵ from Equation 1.

We reproduce the model SPADE [5] as described in the original publication with a Wide ResNet-50-2 (WR50) [28] as backbone. For SPADE and PaDiM we apply the same prepocessing as in [5]. We resize the images from the MVTec AD to 256x256 and center crop them to 224x224. For the images from the STC we use a 256x256 resize only. We resize the images and the localization maps using bicubic interpolation and we use a Gaussian filter on the anomaly maps with parameter $\sigma = 4$ like in [5].

We also implement our own VAE as a reconstruction-based baseline implemented with a ResNet18 as encoder and a 8x8 convolutional latent variable. It is trained on each MVTec AD class with 10 000 images using the following data augmentations operations: random rotation ($-2°$, $+2°$), 292x292 resize, random crop to 282x282, and finally center crop to 256x256. The training is performed during 100 epochs with the Adam optimizer [12] with an initial learning rate of 10^{-4} and a batch size of 32 images. The anomaly map for the localization corresponds to the pixel-wise L2 error for reconstruction.

V. Results

A. Ablative studies

First, we evaluate the impact of modeling correlations between semantic levels in PaDiM and explore the possibility to simplify our method through dimensionality reduction.

Inter-layer correlation. The combination of Gaussian modeling and the Mahalanobis distance has already been employed in previous works to detect adversarial attacks [26] and for anomaly detection [23] at the image level. However those methods do not model correlations between different CNN's semantic levels as we do in PaDiM. In Table I we show the anomaly localization performance on the MVTec AD of PaDiM with a ResNet18 backbone when using only one of the first three layers (Layer 1, Layer 2, or Layer 3) and when summing the outputs of these 3 models to form an ensemble method that takes into account the first three layers but not the correlations between them (Layers 1+2+3). The last row of Table I (PaDiM-R18) is our proposed version of PaDiM where each patch location is described by one Gaussian distribution taking into account the first three ResNet18 layers and correlations between them. It can be observed that using Layer 3 produces the best results in terms of AUROC among the three layers. It is due to the fact that Layer 3 carries higher semantic level information which helps to better describe normality. However, Layer 3 has a slightly worse PRO-score than Layer 2 that can be explained by the lower resolution of Layer 2 which affects the accuracy of anomaly localization. As we see in the two last rows of Table I, aggregating information from different layers can solve the trade-off issue between high semantic information and high resolution. Unlike model Layer 1+2+3 that simply sums the outputs, our model PaDiM-R18 takes into account correlations between semantic levels. As a result, it outperforms Layer 1+2+3 by 1.1p.p (percent point) for AUROC and 1.8p.p for PRO-score. It confirms the relevance of modeling correlation between semantic levels.

TABLE I
STUDY OF THE ANOMALY LOCALIZATION PERFORMANCE USING DIFFERENT SEMANTIC-LEVEL CNN LAYERS. RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%) ON THE MVTEC AD.

Layer used	all texture classes	all object classes	all classes
Layer 1	(93.1, 87.1)	(95.6, 86.5)	(94.8, 86.8)
Layer 2	(95.0, 89.7)	(96.1, 87.9)	(95.7, 88.5)
Layer 3	(94.8, 89.6)	(97.1, 87.7)	(95.7, 88.3)
Layer 1+2+3	(95.4, 90.7)	(96.3, 88.1)	(96.0, 89.0)
PaDiM-R18	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

TABLE II
STUDY OF THE ANOMALY LOCALIZATION PERFORMANCE WITH A DIMENSIONALITY REDUCTION FROM 448 TO 100 AND 200 USING PCA OR RANDOM FEATURE SELECTION (RD). RESULTS ARE DISPLAYED AS TUPLES (AUROC%, PRO-SCORE%) ON THE MVTEC AD.

	all texture classes	all object classes	all classes
Rd 100	(95.7, 91.3)	(97.2, 89.4)	(96.7, 90.5)
PCA 100	(93.7, 88.9)	(93.5, 84.1)	(93.5, 85.7)
Rd 200	(96.1, 92.0)	(97.5, 89.8)	(97.0, 90.5)
PCA 200	(95.1, 91.8)	(96.0, 88.1)	(95.7, 89.3)
all (448)	(**96.3**, **92.3**)	(**97.5**, **90.1**)	(**97.1**, **90.8**)

Dimensionality reduction. PaDiM-R18 estimates multivariate Gaussian distributions from sets of patch embeddings vectors of 448 dimensions each. Decreasing the embedding vector size would reduce the computational and memory complexity of our model. We study two different dimensionality reduction methods. The first one consists in applying a Principal Component Analysis (PCA) algorithm to reduce the vector size to 100 or 200 dimensions. The second method is a random feature selection where we randomly select features before the training. In this case, we train 10 different models and take the average scores. Still the randomness does not change the results between different seeds as the standard error mean (SEM) for the average AUROC is always between 10^{-4} and 10^{-7}.

From Table II we can notice that for the same number of dimensions, the random dimensionality reduction (Rd) outperforms the PCA on all the MVTec AD classes by at least 1.3p.p in the AUROC and 1.2p.p in the PRO-score. It can be explained by the fact that PCA selects the dimensions with the highest variance which may not be the ones that help to discriminate the normal class from the anomalous one [23].