特定の正常画像との一致が失敗する可能性がある ii） 小規模なデータセットや複雑な変動を経験する対象の場合、テスト画像とあらゆる点で類似する正常なトレーニング画像を見つけることができないため、偽陽性検出が発生する可能性がある iii） 画像差分の計算は、使用される損失関数に非常に敏感である。

上記の問題を解決するため、私たちはマルチイメージ対応手法を提案します。関連するテスト画像と通常のトレーニング画像のピクセル位置 p ∈ P ごとに、特徴抽出器 F（x_{i} ， p） を使用して深層特徴を抽出します。特徴抽出器の詳細は第3.4節で説明します。K近傍のすべてのピクセル位置における特徴のギャラリーを構築します。G = {F （x_{1} ， p）|p ∈ P } ∪ {F （x_{2} ， p）|p ∈ P } } .. ∪ {F （x_{K} ， p）|p ∈ P } } 。ピクセルpの異常スコアは、特徴F （y， p）とそのギャラリーG内のκ近傍の特徴との平均距離によって与えられます。したがって、ターゲット画像yにおけるピクセルpの異常スコアは次のように与えられます：

$$d(y,p) = \frac{1}{\kappa} \sum_{f \in N_\kappa(F(y,p))} \|f - F(y,p)\|^2 \tag{3}$$

閾値 θ に対して、ピクセルが異常と判定されるのは d（y， p） > θ の場合、つまり K 個の最近傍正常画像において密接に対応するピクセルが見つからない場合です。

3.4 特徴量ピラミッドマッチング

密接な対応によるアライメントは、画像の正常な部分と異常な部分を決定する有効な方法です。効果的なアライメントを行うためには、マッチングのための特徴量を決定する必要があります。以前の段階と同様に、当手法は事前訓練された深層ResNet CNNの特徴量を使用します。ResNetは特徴量のピラミッドを生成します。画像ピラミッドと同様に、初期の層（レベル）はより高い解像度の特徴量を生成し、文脈情報をより少なくエンコードします。後段の層は、より多くの文脈をエンコードするが、空間解像度が低い特徴を生成します。効果的なアライメントを行うため、特徴ピラミッドの異なるレベルの特徴を用いて各位置を記述します。具体的には、最後のMブロックの出力から特徴を結合します。Mの異なる値の結果は実験セクションに示されています。私たちの特徴量は、細粒度の局所特徴とグローバルな文脈の両方をエンコードします。これにより、ターゲット画像とK ≧ 1の正常画像間の対応関係を検出することができ、画像の明示的なアラインメントを必要としないため、技術的に困難で脆弱なアプローチを回避できます。私たちの方法はスケーラブルで実践的に展開しやすいです。セクション4で示すように、私たちの方法はサブ画像異常セグメンテーション精度において最先端の性能を達成しています。

3.5 実装の詳細

すべての実験において、ImageNetで事前訓練されたWide-ResNet50 × 2特徴量抽出器を使用しています。MVTec画像は256 × 256にリサイズされ、クロップされました

Readable

consist of, alignment to particular normal images may fail ii) for small datasets or objects that experience complex variation, we may never in fact find a normal training image which is similar to the test image in every respect triggering false positive detections iii) computing the image difference would be very sensitive to the loss function being used.

To overcome the above issues, we present a multi-image correspondence method. We extract deep features at every pixel location $p \in P$ using feature extractor $F(x_i, p)$ of the relevant test and normal training images. The details of the feature extractor will be described in Sec. 3.4. We construct a gallery of features at all pixel locations of the K nearest neighbors $G = \{F(x_1, p)|p \in P\} \cup \{F(x_2, p)|p \in P\}\}.. \cup \{F(x_K, p)|p \in P\}\}$. The anomaly score at pixel p, is given by the average distance between the features $F(y, p)$ and its κ nearest features from the gallery G. The anomaly score of pixel p in target image y is therefore given by:

$$d(y, p) = \frac{1}{\kappa} \sum_{f \in N_\kappa(F(y,p))} \|f - F(y, p)\|^2 \tag{3}$$

For a given threshold θ, a pixel is determined as anomalous if $d(y, p) > \theta$, that is, if we cannot find a closely corresponding pixel in the K nearest neighbor normal images.

3.4 Feature Pyramid Matching

Alignment by dense correspondences is an effective way of determining the parts of the image that are normal vs. those that are anomalous. In order to perform the alignment effectively, it is necessary to determine the features for matching. As in the previous stage, our method uses features from a pre-trained deep ResNet CNN. The ResNet results in a pyramid of features. Similarly to image pyramids, earlier layers (levels) result in higher resolution features encoding less context. Later layers encode lower resolution features which encode more context but at lower spatial resolution. To perform effective alignment, we describe each location using features from the different levels of the feature pyramid. Specifically, we concatenate features from the output of the last M blocks, the results for different numbers of M is shown in the experimental section. Our features encode both fine-grained local features and global context. This allows us to find correspondences between the target image and $K \geq 1$ normal images, rather than having to explicitly align the images, which is more technically challenging and brittle. Our method is scalable and easy to deploy in practice. We will show in Sec. 4 that our method achieves the state-of-the-art sub-image anomaly segmentation accuracy.

3.5 Implementation Details

In all experiments, we use a Wide-ResNet50 $\times$ 2 feature extractor, which was pre-trained on ImageNet. MVTec images were resized to 256×256 and cropped

Readable