

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  $G = \{F(x_1, p) | p \in P\} \cup \{F(x_2, p) | p \in P\} \dots \cup \{F(x_K, p) | p \in P\}$ . The anomaly score at pixel  $p$ , is given by the average distance between the features  $F(y, p)$  and its  $\kappa$  nearest features from the gallery  $G$ . The anomaly score of pixel  $p$  in target image  $y$  is therefore given by:

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

For a given threshold  $\theta$ , 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  $256 \times 256$  and cropped