Revisiting Reverse Distillation for Anomaly Detection

Tran Dinh Tien[1] Anh Tuan Nguyen[1] Nguyen Hoang Tran[1] Ta Duc Huy[1]
Soan T.M. Duong[1,3] Chanh D. Tr. Nguyen[1,2] Steven Q. H. Truong[1,2]
[1]VinBrain JSC, Vietnam [2]VinUniversity, Vietnam [3]Le Quy Don Technical University, Vietnam

Abstract

Anomaly detection is an important application in large-scale industrial manufacturing. Recent methods for this task have demonstrated excellent accuracy but come with a latency trade-off. Memory based approaches with dominant performances like PatchCore or Coupled-hypersphere-based Feature Adaptation (CFA) require an external memory bank, which significantly lengthens the execution time. Another approach that employs Reversed Distillation (RD) can perform well while maintaining low latency. In this paper, we revisit this idea to improve its performance, establishing a new state-of-the-art benchmark on the challenging MVTec dataset for both anomaly detection and localization. The proposed method, called ***RD++****, runs six times faster than PatchCore, and two times faster than CFA but introduces a negligible latency compared to RD. We also experiment on the BTAD and Retinal OCT datasets to demonstrate our method's generalizability and conduct important ablation experiments to provide insights into its configurations. Source code will be available at* https://github.com/tientrandinh/Revisiting-Reverse-Distillation.

1. Introduction

Detecting anomalies is a crucial aspect of computer vision with numerous applications, such as product quality control [4], and healthcare monitor system [21]. Unsupervised anomaly detection can help to reduce the cost of collecting abnormal samples. This task identifies and localizes anomalous regions in images without defect annotations during training. Instead, a set of abnormal-free samples is utilized. Early approaches rely on generative adversarial models to extract meaningful latent representations on normal samples [25, 29, 31]. However, these approaches are computationally expensive, resulting in higher latency and potential performance limitations on unseen data. Other approaches leverage the pre-trained Convolutional Neural Networks (CNNs) [19] backbones to extract comprehensive visual features for anomaly detection systems [3, 23].

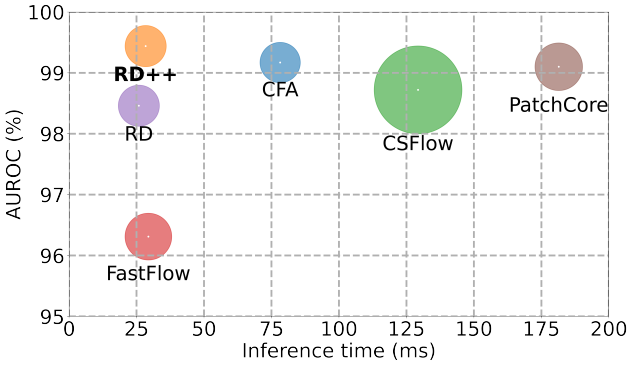

Figure 1. Comparisons of different anomaly detection methods in terms of AUROC sample (vertical axis), inference time (horizontal axis), and memory footprint (circle radius). Our RD++ achieves the **highest** AUROC sample metric for anomaly detection while being **6**× **faster** than PatchCore, **4**× **faster** than CSFlow, and **2**× **faster** than CFA. Additionally, RD++ requires only 4GB of memory for inference, making it one of the least memory-use methods. The test environment was conducted on a computer with Intel(R) Xeon(R) 2.00GHz (4 cores) and Tesla T4 GPU (15GB VRAM).

Alternative approach employs knowledge distillation (KD) [14] based frameworks. For example, Salehi et al. [30] set up a teacher-student network pair, and knowledge is transferred from teacher to student. During training, the student is learned from only normal samples. Thus, it is expected to learn the distribution of normal samples, subsequently generating the out-of-distribution representations when inference with anomalous query [30]. However, Deng and Li [8] point out that the statement is not always accurate due to limitations in the similarity of network architectures and the same data flows in the teacher-student model. To overcome these limitations, they propose a reverse flow, called Reverse Distillation (RD), in which the teacher output is fed to the student through a One-class Bottleneck module (OCBE). The reverse distillation approach achieves competitive performance and also maintains low latency.

Recent research in anomaly detection, such as PatchCore [27], and CFA [20], have achieved state-of-the-art performance in detecting and localizing anomalies. However,

these methods are based on the memory bank framework, which leads to significant latency and makes them challenging to apply in practical scenarios. Our key question: *How can we develop a method that achieves high accuracy and fast inference for real-world applications?*

This paper answers the question from the perspective of reverse distillation. We identify the limitations of the RD approach by examining feature compactness and anomalous signal suppression. We argue that relying solely on the distillation task and an OCBE module is insufficient for providing a compact representation to the student. Furthermore, we do not observe an explicit mechanism to discard anomalous patterns using the OCBE block as the authors claim. To address these concerns, we incorporate RD with multi-task learning to propose RD++, which demonstrates a favorable gain in performance.

The contributions of the proposed RD++ are highlighted as follows:

- We propose RD++ to tackle two tasks. First, feature compactness task: by presenting a self-supervised optimal transport method. Second, anomalous signal suppression task: by simulating pseudo-abnormal samples with simplex noise and minimizing the reconstruction loss.
- We conduct extensive experiments on several public datasets in different domains, including MVTec, BTAD, and Retinal OCT. Results show that our approach achieves state-of-the-art performance on detection and localization, demonstrating strong generalization capabilities across domains. Furthermore, our method's real-time capability is at least twice as fast as its latest counterparts (see Fig. 1), making it a promising method for practical applications.

2. Related works

This section provides an overview of prior approaches to unsupervised anomaly detection. In early literature, generative models such as autoencoders (AE) [18], generative adversarial networks (GAN) [12], and their variants are used to reconstruct normal images from anomalous ones [1, 2, 6, 31]. However, these methods struggle with complex texture reconstruction. Later methods use deep models to improve the quality of reconstructed images [41, 42].

Recently, with the hypothesis that fine-grained visual features deliver revolutionary results in anomaly detection, proposed approaches attempt to learn the representations from nominal samples. A trend in anomaly detection is to use a pre-trained model on an external image dataset to understand the distribution of nominal features. Extracting features from pre-trained networks, i.e., trained on large-scale datasets such as ImageNet [9], is better than processing the image directly in terms of anomaly detection accuracy. Such extracted features are discriminative for normal images, which can be used to approximate distributions of normal features and highlight the difference in defect areas.

With the memory bank usage, PatchCore [27] proposes an algorithm to exploit the association between patches of an image for anomaly detection and presents a way to store the sub-sampled core set of the image. Features extracted from the pre-trained backbone are stored in the memory bank to obtain a patch-level distance between the core set and the sample to detect anomalies. Similarly, CFA [20] detects the adverse effects of biased features from the pre-trained network on anomalous localization and proposes an adaptive solution to the target dataset. They present an approach to obtain discriminant features through metric learning and experimentally verify that the features enable highly complex anomalous localization. CFA's memory bank is compressed independently of the target dataset size, achieving promising performance. However, these methods have disadvantages when training on large datasets due to the need to create memory banks, which require high-cost computational and complex architecture.

Other methods focus on estimating the distribution of normal patterns through a parametric paradigm, i.e., normalizing flow and performing outstanding results. They integrate flow-based sub-networks into their pipelines for better approximate normal feature distributions. By minimizing the loss, i.e., the negative log-likelihood, over normal images during training, flow-based models can map normal image features into the standard target distribution. They use the probability scores to identify and localize anomalies in the image outside the learned distribution. CSFlow [28] proposes a cross-scale normalizing flow to process multi-scale feature maps jointly. It performs density estimation on multiple feature tensors in parallel. The likelihood of these transformed features follows the multivariate standard distribution. Despite high performance, passing multiple tensors into normalizing flow modules significantly increases computational complexity. FastFlow [39] introduces the single-column structure to overcome this limitation. It designs 2D normalizing flow modules with fully convolutional networks followed by the feature extractor for accurate and real-time out-of-distribution feature detection.

Knowledge distillation [14] is utilized as a practical approach for unsupervised anomaly detection. Many teacher-student frameworks are proposed to obtain exact anomaly pixels from the test image. Multiresolution knowledge distillation [30] tries to distinguish unusual features on multi-level features. STFPM [36] adapts a feature pyramid matching mechanism between the teacher network and the student counterpart. The teacher-student model is expected to yield discrepant features on anomalies in inference. RD [8] proposes an interesting reverse distillation approach for anomaly detection, achieving competitive results

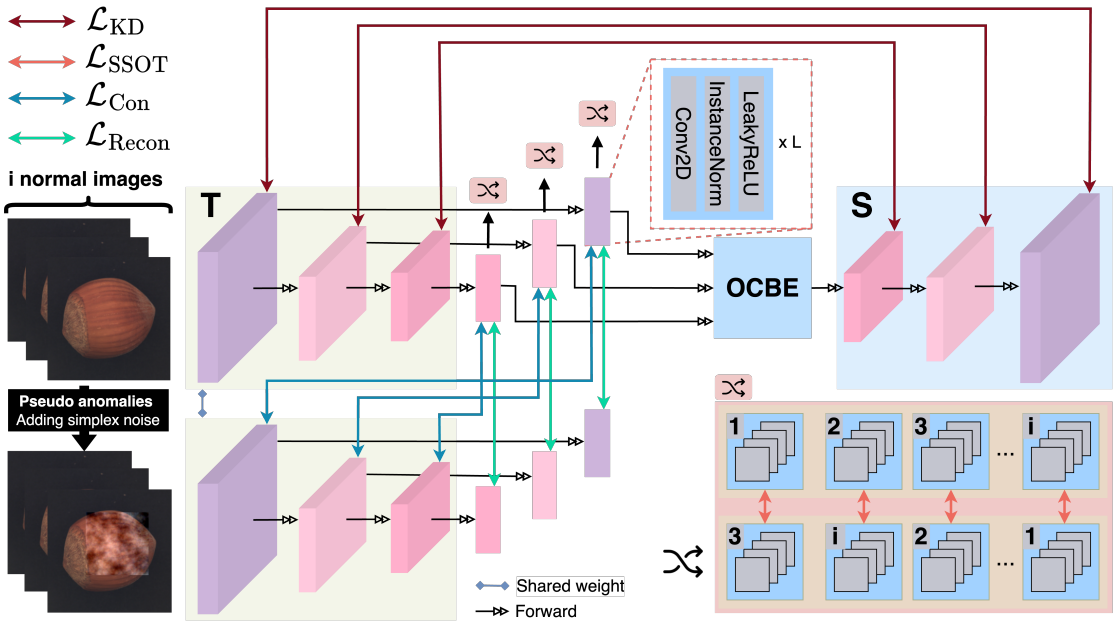

Figure 2. Overview of our RD++ during **training**. First, we integrate a projection layer directly after each intermediate teacher block to provide the student network with a compact, abnormal-free representation. Distillation loss ($\mathcal{L}_{KD}$) introduced in [8] is combined with other multiple loss functions for optimization. For the **feature compactness** task: two loss functions are proposed: Self-supervised optimal transport loss ($\mathcal{L}_{SSOT}$) for projecting the normal feature space to compact representation. Contrast loss ($\mathcal{L}_{Con}$) supports projection layers learning compact embedding by setting projected normal features apart from abnormal features. For the **anomalous signal suppression** task: first, we design a pseudo anomalies mechanism to simulate pseudo-abnormal samples during training, then reconstruction loss ($\mathcal{L}_{Recon}$) is proposed to guide the projection layer to know how to reconstruct the normal feature space from the pseudo-abnormal feature.

in detecting and localizing anomalies. Unlike conventional knowledge distillation, the framework follows the encoder-decoder architecture. The knowledge from the pre-trained teacher model is distilled to the student model in a reverse direction, i.e., have the layers processed back to front.

3. Reverse distillation for abnormal detection

This section summarizes the original reverse distillation (RD) for anomaly detection as proposed in [8]. The RD contains three modules: a fixed pre-trained teacher as the encoder, a trainable one-class embedding (OCBE) module, and a student as the decoder. The heterogeneous encoders' and decoders' reverse direction strategy contributes to the discrepant representations of anomalies. The OCBE module adapts the last block of the Resnet [13] for feature extraction. It is proposed to strengthen the discrepancy for abnormality by condensing the patterns into low-dimensional space and eliminating anomalous signals.

Let ϕ denote the output of the OCBE block. The paired activation correspondence in the T - S model is denoted by $\{f_E^k = E^k(I), f_D^k = D^k(\phi)\}$, where I is the raw image, E^k and D^k respectively define the k^{th} encoder and k^{th} decoder block in the teacher and student model. The T-S model takes the cosine similarity loss as the distillation loss for knowledge transfer. The loss function for optimizing the network (OCBE module + student model) is obtained by the following equation:

$$\mathcal{L}_{KD} = 1 - \sum_{k=1}^{K} \{ \frac{(f_E^k(h,w))^\top \cdot (f_D^k(h,w))}{\|f_E^k(h,w)\| \, \|f_D^k(h,w)\|} \}, \quad (1)$$

where K is the number of feature layers used in training, h, and w denote the height and width of k^{th} feature map.

After training with normal samples, during testing, the vector-wise cosine similarity loss of representations along the channel axis in the teacher-student setting displays a high abnormality score, which indicates abnormal images.

4. Proposed method

The proposed framework RD++ for anomaly detection, inspired by the idea of reverse distillation [8], consists of several modifications for the architecture and the loss function. The overall architecture RD++ is visualized in Fig. 2.

4.1. Pseudo anomalies mechanism

One of our method's primary assumptions is that reverse distillation is effective during inference if the student is re-

Figure 3. Comparison between Simplex noise [37] and Gaussian noise in simulating the pseudo anomalous regions in the image. Simplex noise generates a **more natural** pseudo-anomaly.

stricted from receiving abnormal information. One of the limitations of the training design in RD is that the OCBE module does not have an objective function to prevent abnormal information from being transmitted to the student. It leads to no substantial insurance that the abnormal patterns do not heavily flow to the student when making inferences on anomalous samples. We investigate how to strictly prevent the OCBE module from receiving abnormal patterns. As a result, we integrate the projection layers behind respective blocks in the teacher network and allow all projection layers to take responsibility for restricting the anomalous information flow to OCBE module.

The pseudo-anomalies are simulated during training. A perturbation term is randomly added to the normal images via employing the simplex noise [37]. According to the author, simplex noises are better than gaussian noises at simulating anomalous distributions based on power laws. As shown in Fig. 3, the simplex noises generate more naturally abnormal patterns than gaussian noises.

Algorithm 1 Pseudo-anomaly mechanism for images

Let U[*a*,*b*] $\sim$ Discrete Random Distribution (*a*,*b*)
for epoch = 1,2,...*n* **do**
  for χ_i in normal training-set χ **do**
    Get h_{noise}, $w_{noise} \subset$ U[*a*,*b*]
    Get x_{start}, $y_{start} \subset$ U[*a*-h_{noise},*b*-w_{noise}]
    Randomly generate simplex noise:
    $\epsilon \sim$ Simplex((h_{noise}, w_{noise}), N=6, γ=0.6)
    ξ = np.zeros(χ_i.shape)
    $x_{end} = x_{start} + h_{noise}$
    $y_{end} = y_{start} + w_{noise}$
    $\xi[x_{start}: x_{end}, y_{start}: y_{end}] = \epsilon$
    $\chi_i = \chi_i + \lambda * \xi$ (λ: the degree of adding noises)
    Training process
  end for
end for

4.2. Multiscale projection layers

Projection layers receive the features of their respective teacher's blocks as input and project them into the compact feature representation before feeding into the OCBE module. We design the projection layer by sequentially stacked L Convblocks (Convolution, InstanceNorm [35], LeakyReLU [38]). In experimental settings, we set $L = 4$.

4.3. Training objectives

We propose a combined loss for training RD++. The loss includes three components: (i) self-supervised optimal transport loss, $\mathcal{L}_{\text{SSOT}}$ for learning the compact feature representation between normal samples; ii) reconstruction loss, $\mathcal{L}_{\text{Recon}}$, for recovering normal features from pseudo-abnormal features; and iii) contrast loss $\mathcal{L}_{\text{Con}}$ for further learning the feature compactness. The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{KD}} + \alpha\mathcal{L}_{\text{SSOT}} + \beta\mathcal{L}_{\text{Recon}} + \gamma\mathcal{L}_{\text{Con}}, \tag{2}$$

where α, β, and γ are the positive regularization parameters.

Self-supervised Optimal Transport Loss. Projection layers receive normal features from their respective teacher encoder's blocks and purposely project them into compact feature spaces. We aim to ensure that projected feature representations from normal samples are close to each other. As inspired by Nguyen et al. [24], we consider minimizing the distance between feature embeddings as equivalent to minimizing their probability measures. To achieve this goal, we propose using the de-biased Sinkhorn divergence, a variant of the Optimal Transport distance [7,10,11,26,32]. This distance measure allows us to calculate the spatial discrepancy between two distributions of feature spaces. We train projection layers in a self-supervised manner, ensuring pair-wise feature spaces in a mini-batch of normal images are close by minimizing the de-biased Sinkhorn divergence between their probability measures.

Let $f_{i,k}$ (i=[1,...,m], k=[1,...,K]) be feature output of training sample χ_i at block k^{th} of teacher's encoder, Φ_k be the projection layer at block k^{th} of teacher's encoder, σ be softmax function, π be the transportation plan, and C denotes some ground cost to transport a unit of mass between probability distributions α and β. The optimal transport distance between α and β is:

$$\begin{aligned}\text{OT}_{\varepsilon,\rho}(\alpha,\beta) &= \min_{\pi \geqslant 0} \langle \pi, \text{C} \rangle + \varepsilon\,\text{KL}(\pi, \alpha \otimes \beta) \\ &+ \rho\,\text{KL}(\pi\,\mathbf{1}, \alpha) + \rho\,\text{KL}(\pi^{\intercal}\,\mathbf{1}, \beta) \\ &= \max_{f,g} -\rho\langle \alpha, e^{-f/\rho} - 1 \rangle - \rho\langle \beta, e^{-g/\rho} - 1 \rangle \\ &- \epsilon\langle \alpha \otimes \beta, e^{(f \oplus g - \text{C})/\epsilon} - 1 \rangle,\end{aligned} \tag{3}$$

where ε and ρ are the regularization parameter and the exponent term, respectively. KL is a Kullback-Leibler divergence [15], defined as:

$$\text{KL}(\alpha,\beta) = \langle \alpha, \log \tfrac{\text{d}\alpha}{\text{d}\beta} \rangle - \langle \alpha, 1 \rangle + \langle \beta, 1 \rangle \geqslant 0. \tag{4}$$

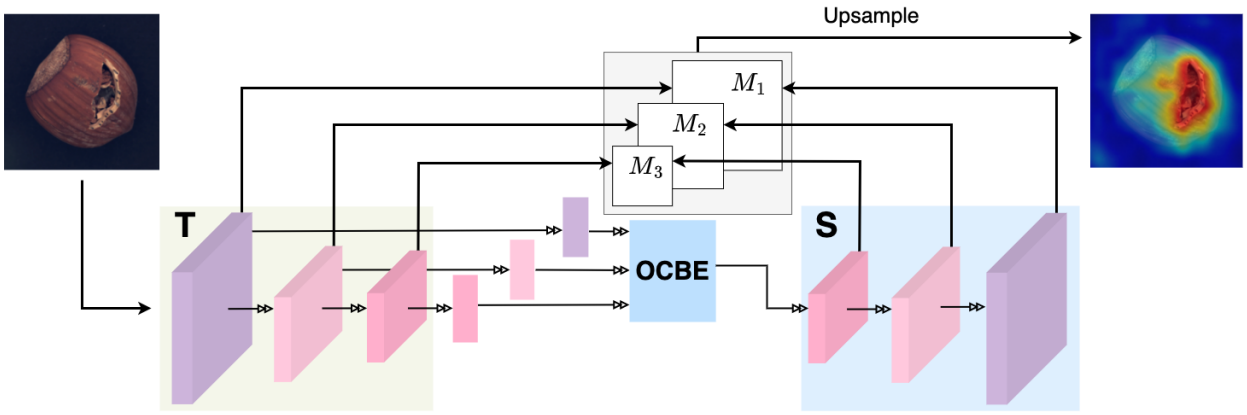

Figure 4. **RD++ inference** procedure for detecting and localizing anomalies on images. The process is almost similar to RD [8]. The only difference is that before being forwarded into the OCBE module, the teacher's output embeddings in blocks are passed to their counterpart projection layer. Since the projection layer is lightweight, the inference time is almost the same as the baseline RD.

The de-biased Sinkhorn divergence between two empirical measures is defined as:

$$\begin{aligned}\mathcal{S}_{\varepsilon,\rho}(\alpha,\beta) =& \text{OT}_{\varepsilon,\rho}(\alpha,\beta) - \tfrac{1}{2}\text{OT}_{\varepsilon,\rho}(\alpha,\alpha) \\ &- \tfrac{1}{2}\text{OT}_{\varepsilon,\rho}(\beta,\beta) + \tfrac{\varepsilon}{2}\|\langle\alpha,1\rangle - \langle\beta,1\rangle\|^2.\end{aligned} \tag{5}$$

We then propose the self-supervised optimal transport loss:

$$\mathcal{L}_{\text{SSOT}} = \tfrac{1}{m}\tfrac{1}{k}\sum_{i,j=1}^{m}\sum_{k=1}^{K} S_{\varepsilon,\rho}(\sigma(\Phi_k(f_{i,k})),\ \sigma(\Phi_k(f_{j,k}))). \tag{6}$$

Reconstruction Loss. We denote $f_{i,k}, \tilde{f}_{i,k}$ are the feature output of encoder block k^{th} of normal image x_i and pseudo-abnormal image $\xi(x_i)$ respectively:

$$f_{i,k} = E^k(x_i), \qquad \tilde{f}_{i,k} = E^k(\xi(x_i)). \tag{7}$$

The reconstruction loss is defined as:

$$\mathcal{L}_{\text{Recon}} = \tfrac{1}{k}\sum_{k=1}^{K}(1 - \cos(\Phi_k(f_{i,k}), \Phi_k(\tilde{f}_{i,k}))). \tag{8}$$

During training, we inject anomalous signals through pseudo-anomaly input images to feature space and encourage the projection layers to learn how to reconstruct normal features from pseudo-abnormal regions. By optimizing this learning objective, we accelerate the projection layer's ability to suppress anomalous information during inference.

Contrast Loss. To strengthen the compact learning of projection layers on normal images, we force the projection layer to concentrate on exploring deeper representations of normal features by pushing away abnormal information from projected normal space. We employ the cosine embedding loss with margin f, and the contrast loss is defined as:

$$\mathcal{L}_{\text{Con}} = \tfrac{1}{k}\sum_{k=1}^{K}\max(0, \cos(\Phi_k(f_{i,k}), \tilde{f}_{i,k}) - f). \tag{9}$$

Algorithm 2 Pseudo-code of RD++ in one epoch training

E, P, O, D: Teacher, projection layer, OCBE, student
f_i, $\tilde{f}_i$: Normal/pseudo-abnormal features at block i
P_i: Projection layer for block i (i is in [1, 2, 3])
Optimizer = Adam($(P_{1,2,3}, O, D)$.parameters())
Load a mini-batch of normal/pseudo-abnormal samples
for $\chi, \xi(\chi)$ in train-dataloader **do**
    Get encoder outputs for normal/pseudo-abnormal images at 3 blocks
    $f_1, f_2, f_3 = E(\chi)$
    $\tilde{f}_1, \tilde{f}_2, \tilde{f}_3 = E(\xi(\chi))$
    Get projected output of normal features
    $\phi_1, \phi_2, \phi_3 = P_1(f_1), P_2(f_2), P_3(f_3)$
    Get projected output of pseudo-abnormal features
    $\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3 = P_1(\tilde{f}_1), P_2(\tilde{f}_2), P_3(\tilde{f}_3)$
    Get feature output of decoder D
    $g_1, g_2, g_3 = D(O(\phi_1, \phi_2, \phi_3))$
    Compute the overall loss
    $\mathcal{L} = \mathcal{L}_{\text{KD}}(f_i, g_i) + \alpha\mathcal{L}_{\text{SSOT}}(\phi_i, \phi_i)$
        $+ \beta\mathcal{L}_{\text{Recon}}(\phi_i, \tilde{\phi}_i) + \gamma\mathcal{L}_{\text{Con}}(\phi_i, \tilde{f}_i)$
    $\mathcal{L}$.backward()
    Optimizer.step()
end for

Inference process. Given an image, the inference procedure is described in Fig. 4.

Table 1. Anomaly detection results in terms of AUROC at image-level on the MVTec dataset [4].

Method	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal nut	Pill	Screw	Toothbrush	Transistor	Zipper	**Avg.**
CSFlow [28]	**100**	99.00	**100**	**100**	**100**	99.80	99.10	97.10	99.60	99.10	**98.60**	97.60	91.90	99.30	**99.70**	98.72
FastFlow [39]	99.40	**100**	99.90	**100**	99.20	**100**	96.20	96.30	99.40	99.50	94.20	83.90	83.60	97.90	95.10	96.31
CFA [20]	97.30	99.20	**100**	99.40	99.70	**100**	**99.80**	97.30	**100**	**100**	97.90	97.30	**100**	**100**	99.60	99.17
PatchCore [27]	98.70	98.20	**100**	98.70	99.20	**100**	99.50	98.10	**100**	**100**	96.60	98.10	**100**	**100**	99.40	99.10
RD [8]	98.90	**100**	**100**	99.30	99.20	**100**	95.00	96.30	99.90	**100**	96.60	97.00	99.50	96.70	98.50	98.46
RD++	**100**	**100**	**100**	99.70	99.30	**100**	99.20	**99.00**	**100**	**100**	98.40	**98.90**	**100**	98.50	98.60	**99.44**

Table 2. Anomaly localization results in terms of AUROC at pixel-level on the MVTec dataset [4].

Method	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal nut	Pill	Screw	Toothbrush	Transistor	Zipper	**Avg.**
FastFlow [39]	99.10	99.20	**99.6**	**96.60**	94.10	98.60	97.20	99.00	98.00	98.80	97.60	96.60	98.00	97.10	98.50	97.87
CFA [20]	**99.28**	98.12	99.37	95.21	91.53	**98.84**	**98.97**	**99.11**	98.85	**99.15**	**98.93**	98.91	98.96	**98.06**	**99.02**	98.15
PatchCore [27]	99.00	98.70	99.30	95.40	95.00	98.60	98.40	98.80	98.70	98.40	97.40	99.40	98.70	96.30	98.80	98.06
RD [8]	98.90	**99.30**	99.40	95.60	95.30	98.70	97.40	98.70	98.90	97.30	98.20	99.60	**99.10**	92.50	98.20	97.81
RD++	99.20	**99.30**	99.40	**96.60**	**95.80**	**98.80**	98.40	98.80	**99.20**	98.10	98.30	**99.70**	**99.10**	94.30	98.80	**98.25**

Table 3. Anomaly localization results in terms of PRO on the MVTec dataset [4].

Method	Carpet	Grid	Leather	Tile	Wood	Bottle	Cable	Capsule	Hazelnut	Metal nut	Pill	Screw	Toothbrush	Transistor	Zipper	**Avg.**
CFA [20]	96.54	94.04	97.43	89.26	90.54	95.76	**94.17**	93.66	95.75	**94.54**	**97.19**	95.23	91.14	**95.35**	95.95	94.44
PatchCore [27]	96.60	96.00	98.90	87.30	89.40	96.20	92.50	95.50	93.80	91.40	93.20	97.90	91.50	83.70	**97.10**	93.40
RD [8]	97.00	97.60	99.10	90.60	90.90	96.60	91.00	95.80	95.50	92.30	96.4	98.20	**94.50**	78.00	95.40	93.93
RD++	**97.70**	**97.70**	**99.20**	**92.40**	**93.30**	**97.00**	93.90	**96.40**	**96.30**	93.00	97.00	**98.60**	94.20	81.80	96.30	**94.99**

5. Experimental results and analysis

5.1. Implementation detail

Experimental settings. We used the WideResNet50 [40] as the backbone in the T-S model, and the image is resized to 256 $\times$ 256. No data augmentation is applied. These settings are widely adopted as the standard configuration for methods comparison on anomaly detection. For methods that used heavy backbones as main configurations, such as FastFlow [39] adapted vision transformer models (DeiT [33], CaiT [34]), we tried our best to reproduce the result on WideResNet50 to ensure a fair and consistent comparison. We used Adam Optimizer [17], with the learning rate set to 0.005 for the student, OCBE module, and 0.001 for projection layers. The weight for distillation loss, SSOT loss, contrast loss, and reconstruction loss is set at 1, 0.2, 0.02, and 0.002, respectively.

Evaluation Metrics. We used the area under the receiver operator curve (AUROC) based on produced anomaly scores to calculate anomaly detection at image-level performance (AUROC sample). Localization performance was evaluated using the AUROC at pixel-level and PRO [5].

5.2. Anomaly detection on MVTec

MVTec [4] includes 15 real-world datasets for anomaly detection, with ten object classes and five textures. MVTec is widely used as the standard dataset for anomaly detection benchmarks. The training data contains 3,629 normal images. The test set contains 1,725 samples with both normal and abnormal images. Per class contains multiple defects for evaluation. The test data also have pixel-level annotations for calculating anomaly localization metrics.

Results. Tables 1-3 show the RD++ outperforms the recent state-of-the-art approaches in all three metrics: AUROC image-level for detection, AUROC pixel-level and PRO metric for localization. Compared to the baseline (RD), across 15 categories, our approach improves the average AUROC image-level by up to 0.98% (Table 1), improves AUROC pixel-level up to 0.44% (Table 2), and PRO metric up to 1.06% (Table 3).

5.3. Anomaly detection on other datasets

To test the generalizability of RD++ method, we conducted experiments comparing it to four other existing methods: FastFlow [39], CFA [20], PatchCore [27], and baseline RD [8] on two datasets, namely BTAD [22] and Retinal OCT [16].

BTAD [22] contains 3 categories of industrial classes with 2,540 samples. The training dataset includes only normal samples. The test data contains both abnormal samples and normal samples.

Retinal OCT [16] contains 84,495 X-Ray images organized into 3 folders (train, validation, test) for 4 categories (NORMAL, CNV, DME, DRUSEN). We trained the model

Table 4. Anomaly detection and localization results in terms of AUROC at image-level/pixel-level/PRO on the BTAD dataset [22].

Method	Class 01	Class 02	Class 03	**Avg.**
FastFlow [39]	**99.40** / **97.10** / 71.70	82.40 / 93.60 / 63.10	91.10 / 98.30 / 79.50	90.97 / 96.33 / 71.43
CFA [20]	98.10 / 95.90 / 72.00	85.50 / 96.00 / 53.20	99.00 / 98.60 / 94.10	94.20 / 96.83 / 73.10
PatchCore [27]	96.70 / 97.03 / 64.92	81.38 / 95.83 / 47.27	99.95 / 99.19 / 67.72	92.68 / 97.35 / 59.97
RD [8]	96.30 / 96.60 / **75.30**	86.60 / **96.70** / 68.20	**100.00** / 99.70 / **87.80**	94.30 / **97.67** / 77.10
RD++	96.80 / 96.20 / 73.20	**90.10** / 96.40 / **71.30**	100.00 / **99.70** / 87.40	**95.63** / 97.43 / **77.30**

Table 5. Anomaly detection results with AUROC at image-level on Retinal OCT dataset [16].

Method	**AUROC sample**
FastFlow [39]	80.40
CFA [20]	98.25
PatchCore [27]	99.70
RD [8]	99.36
RD++	**99.73**

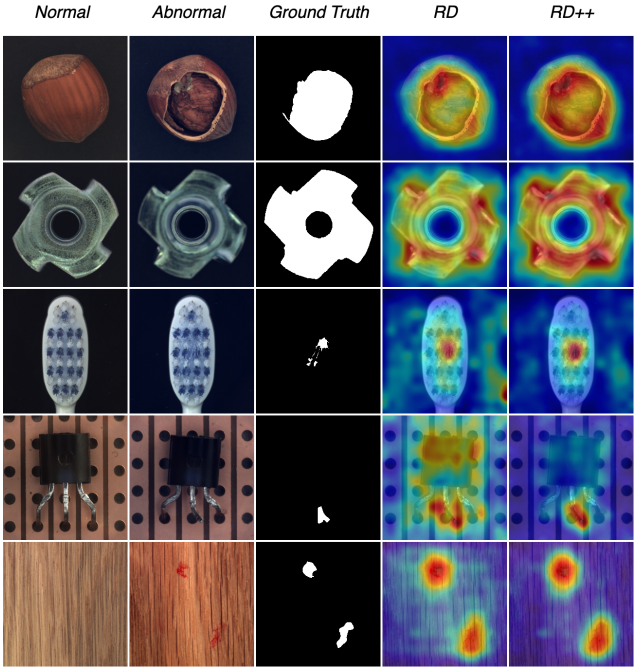

Figure 5. Anomalies in **MVTec** [4] from top to bottom: "crack" on "hazelnut", "flip" on "metal nut", "defective" on "toothbrush", "bent lead" on "transistor", and "color" on "wood". Normal images are included as reference.

on NORMAL categories of the training set and reported the evaluation metrics on the test set.

Results. In the BTAD dataset, RD++ surpasses most state-of-the-art approaches regarding detection and localization results. Table 4 shows that RD++ obtains an improvement of 1.33% and 0.2% for the average AUROC at image-level and PRO metric [5], respectively, compared to the second counterpart RD [8]. For the remainder, i.e., Retinal OCT (Table 5), RD++ achieves 99.73% in AUROC at the image-level, higher than all recent state-of-the-arts.

6. Ablation study

6.1. Training objectives

We investigate the effectiveness of training components and compare results with the baseline.

Table 6. Performance of the proposed RD++ with the configuration of different component sets in the loss on **MVTec** [4]. RD+(SSOT) denotes training RD [8] baseline with $\mathcal{L}_{\text{SSOT}}$.

	AUROC image-level	AUROC pixel-level	PRO
RD	98.46	97.81	93.93
RD + (SSOT)	99.33	98.22	94.95
RD + (SSOT + Con)	99.33	98.24	**95.00**
RD++ (SSOT + Con + Recon)	**99.44**	**98.25**	94.99

Table 6 shows the performances of objective training combinations. While RD [8] reports the effectiveness of condensing features to low dimensional embedding through OCBE module, the analysis proves that the proposed training objective components could significantly improve the model performance. When projection layers are simultaneously trained on self-supervised optimal transport loss, feature representations are much more condensed. We see an improvement in all three metrics. The $\mathcal{L}_{\text{Con}}$ plays as an additional term for condensing normal features. Lastly, the $\mathcal{L}_{\text{Recon}}$ guides the projection layer to alleviate anomalous information throughout reconstructing normal space from pseudo-abnormal information. The more anomalous signal prohibited before forwarding to the student, the stronger discrepancy between Teacher-Student can establish.

6.2. Effectiveness on noise's level

Table 7 shows that more than low noise (λ_{noise} equal 0.1) is needed for solid model performance on **MVTec** [4]. When the amount of noise increases, the PRO metric [5] generally improves with the gradual decrease in anomaly

detection performance. The results show that λ_{noise} equal 0.2 is suitable for generalization in both anomaly detection and localization. We choose λ_{noise} equal 0.5 if we are only interested in per-region-based localization.

Table 7. Importance of noise levels added during training.

λ_{noise}	AUROC image-level	AUROC pixel-level	PRO
0.1	98.98	98.15	94.91
0.2	**99.44**	**98.25**	94.99
0.3	99.25	98.23	94.97
0.4	99.26	98.21	95.02
0.5	99.19	98.08	**95.20**

6.3. What can the student see?

We analyze two essential factors in student input features that play a vital role in the anomaly detection ability of the T/S architecture: (i) feature compactness and (ii) anomalous signal suppression. To elaborate, we analyze the output of OCBE module as this is the input feature for the student.

Feature compactness. We evaluate the compactness of the feature space projected by RD++ versus RD by calculating the pair-wise mean-squared error (MSE) among features of the normal samples. As shown in Fig. 6, RD++ enjoys a much denser feature space while RD has a wider spread of pair-wise distance distribution between the normal samples.

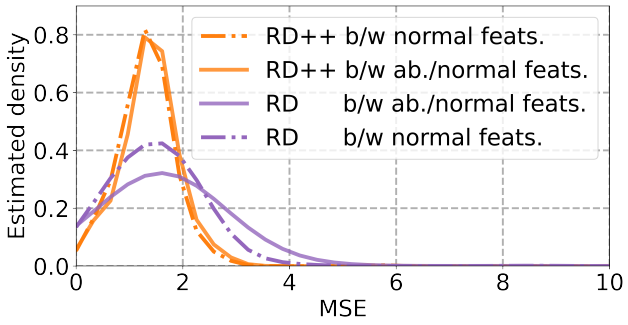

Figure 6. Comparing the normal intra-class and inter-class distance distribution of student's input between the proposed RD++ method and RD [8] baseline on **MVTec** [4] test set.

Anomalous signals suppression. To justify the improvement of our methods over RD on suppressing anomalous signals from the teacher, we calculate the MSE between each of the abnormal samples with every normal sample. The inter-class MSE distribution of RD++ is also narrower than RD. Quantitatively, the variance of the inter-class MSE of RD is 2.235, while RD++ only has a variance of 0.239. The feature distance between a normal and an abnormal sample should be close because the anomalous signals from the abnormal sample are suppressed, making it looks like a normal sample. Accordingly, the wider spread inter-class MSE distribution of RD implies a sub-optimal capability of anomalous signal suppression. Our method, on the other hand, produces a narrower distribution hence showing a better capability.

6.4. Method generalization on different backbones

Table 8 shows that RD++ performs better than RD [8] on different Resnet [13] backbones and also indicates that the deeper and wider network provides stronger detecting anomalies.

Table 8. Quantitative comparison on different backbones on **MVTec** [4].

	AUROC sample		AUROC pixel		PRO	
Backbone	RD	RD++	RD	RD++	RD	RD++
Resnet18	97.90	**98.63**	97.10	**97.64**	91.20	**93.65**
Resnet50	98.40	**99.05**	97.70	**98.17**	93.10	**94.78**
WideResnet50	98.46	**99.44**	97.81	**98.25**	93.93	**94.99**

7. Discussion and conclusion

Limitation. While RD++ shows effectiveness, simulating anomalies by adding noises to random locations in medical images may be suboptimal because specific abnormalities occur in certain areas, such as pneumothorax lesions are mostly concentrated in the inner border of the lungs. The method will be more effective if we incorporate medical knowledge into possible predefining locations where noise should be added.

Conclusion. This paper proposes RD++, inspired by reverse distillation architecture, for anomaly detection. With the proposal for pseudo anomalies mechanism, multiple projection layers integration, and multi-task learning for compact features and abnormal alleviation. Our model obtains competitive accuracy for anomaly detection and real-time inference. Extensive experiments on several datasets demonstrate the effectiveness and generalizability of our proposed RD++ compared to several existing methods for anomaly detection while being simple to implement. We hope the method will be helpful in real applications and pave the way for further advances in this field.

Future work. We aim to apply the method to other tasks besides anomaly detection, such as domain adaptation, where invariant representation is an essential factor. The method can benefit feature invariants through projection layers via compact learning and the ability to recover standard information when domains change. Since projection layers are flexibly integrated into various architectures, we are motivated by the ease of testing the method's effectiveness.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Proc. Asian Conf. Computer Vision*, pages 622–637. Springer, 2018. 2

[2] Haleh Akrami, Anand A. Joshi, Jian Li, Sergül Aydöre, and Richard M. Leahy. A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, 238:107886, Feb. 2022. 2

[3] Jerone T. A. Andrews, Thomas Tanay, Edward J. Morton, and Lewis D. Griffin. Transfer representation-learning for anomaly detection. In *Proc. Inter. Conf. on Machine Learning*, 2016. 1

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 9592–9600, 2019. 1, 6, 7, 8

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 6, 7

[6] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proc. Inter. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019. 2

[7] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269, 2020. 4

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 1, 2, 3, 5, 6, 7, 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2

[10] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proc. Inter. Conf. on Artificial Intelligence and Statistics*, pages 2681–2690, 2019. 4

[11] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proc. Inter. Conf. Artificial Intelligence and Statistics*, pages 1608–1617, 2018. 4

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 8

[14] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2

[15] James M. Joyce. Kullback-leibler divergence. In *Inter. Encyclopedia of Statistical Science*, pages 720–722. Springer, 2011. 4

[16] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 6, 7

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proc. Inter. Conf. on Learning Representations*, 2014. 2

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[20] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022. 1, 2, 6, 7

[21] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 8290–8299, 2018. 1

[22] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: a vision transformer network for image anomaly detection and localization. In *Proc. IEEE Inter. Symposium on Industrial Electronics*, pages 01–06. IEEE, 2021. 6, 7

[23] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by CNN-based self-similarity. *Sensors*, 18(1):209, 2018. 1

[24] Thong Nguyen and Anh Tuan Luu. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proc. AAAI Conf. Artificial Intelligence*, 2022. 4

[25] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: one-class novelty detection using gans with constrained latent representations. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 2898–2906, 2019. 1

[26] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. 4

[27] K Roth, L Pemula, J Zepeda, B Schölkopf, T Brox, and P Gehler. Towards total recall in industrial anomaly detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14298–14308, 2022. 1, 2, 6, 7

[28] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-

based defect detection. In *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, pages 1088–1097, 2022. 2, 6

[29] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 3379–3388, 2018. 1

[30] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 14902–14912, 2021. 1, 2

[31] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019. 1, 2

[32] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019. 4

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. Inter. Conf. on Machine Learning*, pages 10347–10357, 2021. 6

[34] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 32–42, 2021. 6

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: the missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[36] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *The British Machine Vision Conference (BMVC)*, 2021. 2

[37] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 650–656, 2022. 4

[38] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 4

[39] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2, 6, 7

[40] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proc. British Machine Vision Conf.*, pages 87.1–87.12, September 2016. 6

[41] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proc. IEEE/CVF Inter. Conf. Computer Vision*, pages 8330–8339, 2021. 2

[42] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2