Contextual Affinity Distillation for Image Anomaly Detection

Jie Zhang[1] Masanori Suganuma[1,2] Takayuki Okatani[1,2]
[1]Graduate School of Information Sciences, Tohoku University [2]RIKEN Center for AIP
{jzhang,suganuma,okatani}@vision.is.tohoku.ac.jp

Abstract

Previous studies on unsupervised industrial anomaly detection mainly focus on 'structural' types of anomalies such as cracks and color contamination by matching or learning local feature representations. While achieving significantly high detection performance on this kind of anomaly, they are faced with 'logical' types of anomalies that violate the long-range dependencies such as a normal object placed in the wrong position. Noting the reverse distillation approaches that are under the encoder-decoder paradigm could learn from the high abstract level knowledge, we propose to use two students (local and global) to better mimic the teacher's local and global behavior in reverse distillation. The local student, which is used in previous studies mainly focuses on accurate local feature learning while the global student pays attention to learning global correlations. To further encourage the global student's learning to capture long-range dependencies, we design the global context condensing block (GCCB) and propose a contextual affinity loss for the student training and anomaly scoring. Experimental results show that the proposed method sets a new state-of-the-art performance on the MVTec LOCO AD dataset without using complex training techniques.

1. Introduction

The task of anomaly detection(AD) and localization aims to identify whether an image is normal or anomalous and localize the anomalies [5, 29]. It has a wide range of real-world applications including industrial inspection of products [2, 4]. As anomalous samples rarely appear in manufacturing product lines and the unpredictable nature of anomalies, most of the efforts are paid to unsupervised AD methods, in which we have only anomaly-free samples for training.

Recent studies showed that using intermediate features of a deep pre-trained model is representative enough to achieve state-of-the-art performance [26]. Knowledge distillation [15] is one of the most effective ways to achieve this goal. Recent knowledge distillation-based AD approaches

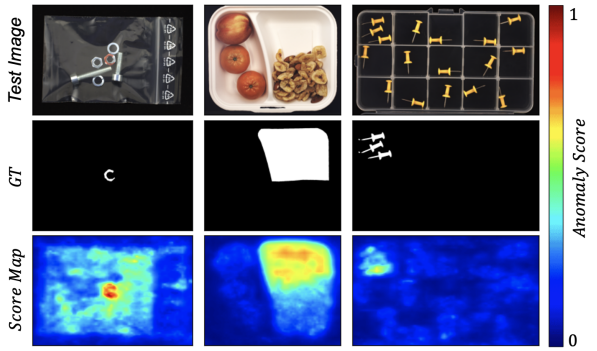

Figure 1. Examples from MVTec LOCO [1]. We show one structural(the left one) and two logical(the right two) anomaly samples with our detection results.

[1, 3, 8, 30, 34] try to transfer the knowledge of normal samples from a teacher which is pre-trained on a large-scale natural image dataset, *e.g.*, ImageNet [9] into a student model. The use of per-pixel [8, 34] or local patch-based regression loss [1, 3] further improves the fine-grained knowledge transfer and AD performance. The teacher model acts as a knowledgeable feature extractor that could extract representative feature embeddings for both normal and anomalous samples, while the student is trained exclusively on anomaly-free samples that it is expected to only mimic the teacher's behavior for normal features. During inference, the anomaly scores are derived from the discrepancy between student and teacher features. As anomalies could be of any size and at any abstract level, using multi-layer features from the teacher could better cover more types of anomalies.

Previous unsupervised anomaly detection and localization datasets focus on concise scenes where each image consists of only one product object, *e.g.*, a capsule or one kind of texture. They contain only the type of anomalies termed *structural* anomalies [1], such as cracks and scratches. The above-mentioned methods are effective for them. However, there are different types of anomalies in complex scenarios with global contextual constraints, termed *logical* anoma-

lies. These are anomalies related to long-range dependencies, such as a specific object in an incorrect position or a missing object. Figure 1 shows structural and logical anomaly examples from MVTec LOCO [1] dataset. The first image of a screw bag contains an example of a structural anomaly, while cereals missing in the breakfast box and two additional pushpins in one compartment are logical anomalies. The aforementioned methods struggle with logical anomalies because deep high-semantic level features from a pre-trained model exhibit source domain bias [28] and lack precise low-level information [8].

To enhance the detection of both structural and logical anomalies, we propose a dual-student knowledge distillation framework (DSKD), leveraging the concept of reverse distillation [8]. Unlike conventional ensemble methods [10, 17, 27] where each model assumes identical roles, we explicitly divide the student models into two models: *local* and *global* students. The local student aims to reconstruct the low-level features of the teacher model, primarily focusing on detecting structural anomalies. Conversely, the global student is trained to harness global contextual information, thereby improving logical AD. Recognizing the inherent challenge of detecting logical anomalies, we incorporate a trainable module, named a global context condensing block, in the global student. This block seeks to effectively condense global information derived from teacher features [1, 21]. The training paradigm of our method is based on the reverse distillation [8], where the teacher acts as an encoder and the students play the role of decoders for the feature reconstruction.

We also introduce a new loss function, termed contextual affinity loss, designed to enhance the global student's ability to capture global contextual information. This is achieved by calculating the cosine similarity between each feature vector of the global student and the entire set of feature vectors from the teacher. These cosine similarity maps are then transformed into probability distributions using a softmax function. We then minimize the discrepancy between the probability distributions of the global student and the teacher. Notably, our approach is distinct from methods like pair-wise distillation [22], which treat all features uniformly. In contrast, our method captures vital contextual information across the entire image.

We conduct extensive experiments on public unsupervised AD datasets and achieve state-of-the-art performance. Our contributions are threefold:

1. We present a novel dual-student knowledge distillation framework. The local student concentrates on precise local feature reconstruction, while the global student focuses on capturing global contextual information. With these distinct roles, our framework enhances detection capabilities for both structural and logical anomalies.

2. We propose the global contextual condensing block and contextual affinity loss, further enforcing the global contextual learning ability.

3. We demonstrate the effectiveness of our method, achieving state-of-the-art performance on standard public datasets.

2. Related Works

We briefly review recent research on unsupervised AD as well as related knowledge distillation works on *supervised* dense prediction tasks. The recent works on AD could be classified into three prototypes: generative models, anomaly synthesis-based methods, and methods leveraging features extracted by pre-trained networks.

Generative models aim to reconstruct normal samples from the encoded feature space. Autoencoders (AEs) and Generative Adversarial Nets [12] are popularly used for sample reconstruction. These models are trained exclusively on normal images. Since the input image is encoded into a compact feature space to only keep the most useful information, the unseen anomalies are expected to be abandoned during inference and thus reconstruct the anomaly-free images for anomaly samples [4]. However, deep models could generalize well to anomaly patterns and fail to detect anomalies. To overcome this issue, normal representation searching [31] in the encoded continuous feature space, iterative reconstruction approaches [7] and memory-guided autoencoders [11, 24] are proposed to limit the model's generalization ability.

Anomaly synthesis-based methods [18, 25, 37] focus on addressing the issue of the lack of anomaly samples so as to train the models in a supervised manner. However, the detection ability is heavily affected by the synthesizing strategies, and their performance shows a strong bias to the synthesized kind of anomalies [14, 38]. To generate more realistic anomalies, DSR [38] tries to generate near-in-distribution low-level anomalies from a vector-quantized feature space. However, it is still challenging for high semantic-level anomaly generation.

There is also a lot of attention paid to employing pre-trained models to extract representative features. The kNN-based approaches [6, 28] construct a feature gallery for normal representations and derive anomaly scores by computing the distances between input and its nearest neighbors in the feature space. They suffer from computational complexity [8, 35] and can not utilize high-semantic level features well as they as more source-domain biased [28].

The knowledge distillation-based approaches try to transfer knowledge of normal samples to student networks. US [3] distills knowledge from a pre-trained teacher network to an ensemble of students for each patch scale. MKD [30] directly distills multi-level features into one compact

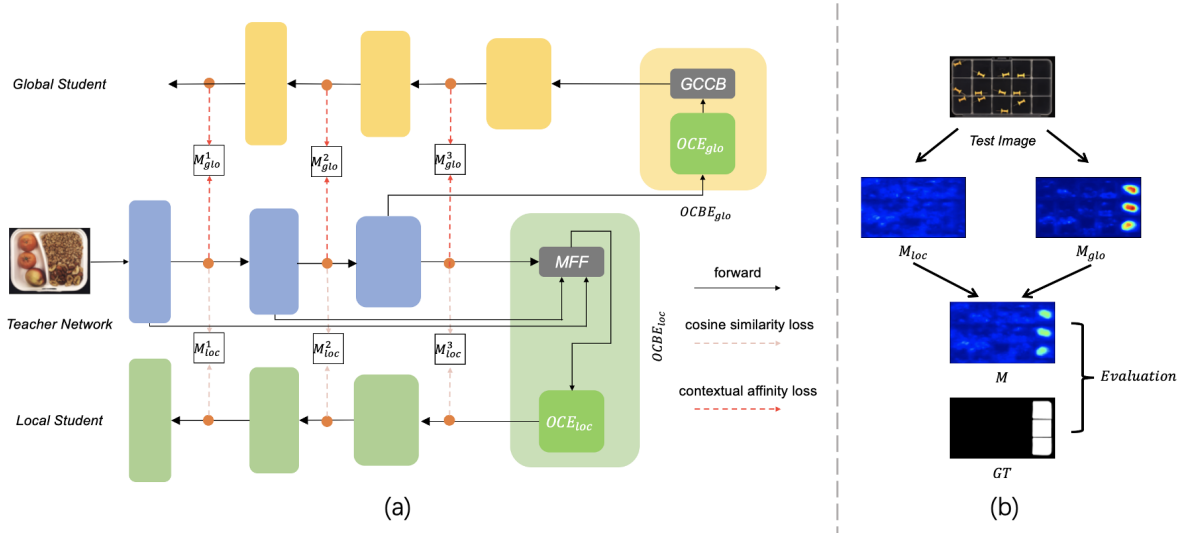

Figure 2. Overview of the dual-student knowledge distillation framework. (a) Our model employs a pre-trained teacher encoder as the feature extractor T, a local student for accurate low-level feature learning, and a global student to capture global contextual information. During training, the students can only learn to mimic the teacher's behavior for normal samples. (b) Anomaly scoring. Firstly, the multi-scale score maps from each student are accumulated into one single scale-normalized map separately. Then the two normalized score maps are added together to get our final detection results.

student model. STFPM [34] uses a vector-wise cosine similarity loss for both student training and anomaly scoring. The reverse distillation [8] proposed the encoder-decoder architecture to distill the knowledge from a bottleneck feature space. These methods are capable of learning local features or patches but are likely to ignore global contextual constraints. GCAD [1] designed a two-branch framework based on US [3] for both structural and logical AD but it is still a two-step distillation framework where the teacher is trained with a deep pre-trained model and a large number of cropped image patches from ImageNet [9]. To ensure training stability, multi-step training and skip connections with linearly decreased weights are added to the student.

There are also some knowledge distillation methods applied to dense prediction tasks such as semantic segmentation [23, 39] and object detection [20] trying to transfer the knowledge to a compact student network via fully exploring the rich information within the intermediate features from the teacher. MIMIC [19] samples features from feature regions for object detection. Pair-wise knowledge distillation [22] was proposed to distill the structured knowledge from the feature. Channel-wise knowledge distillation [32] converts the features of each channel into probability distributions leveraging the prior that the activations from each channel tend to encode specific scene categories. It is pointed out that strictly applying the per-pixel loss which means each pixel or correlation is treated equally may enforce overly strict constraints on the student model and lead to sub-optimal solutions [32]. However, the most important guidance for training the student model comes from the ground truth labels that are not available for unsupervised AD. Also, in conventional knowledge distillation applications where only the student model is deployed after training, the structured knowledge is computed or evaluated *separately* for each student and teacher feature map, while both the teacher and student are used for knowledge distillation-based AD methods. We distinguish our proposed contextual affinity loss from prior arts that the contextual affinity for student features is computed using both student and teacher features for better guidance and to make training stable.

3. Proposed Method

3.1. Problem Formulation

Given a set of anomaly-free training images $\mathcal{S}^t = \{I^t_1, ..., I^t_{n_t}\}$ and a validation set $\mathcal{S}^v = \{I^v_1, ..., I^v_{n_v}\}$ that consists of also anomaly-free images, we aim to detect if a test image from the test set $\mathcal{S}^q = \{I^q_1, ..., I^q_{n_q}\}$ is anomalous or not, and also localize the defected area if it is anomalous.

3.2. Overview of the Dual-student Framework

As shown in Fig 2, our DSKD consists of five parts: a deep neural network pre-trained on ImageNet as the fixed teacher T to extract multi-level representative features, a one-class bottleneck embedding module $OCBE_{loc}$ for the local student, a local student decoder S_{loc}, a $OCBE_{glo}$ for the global student that contains a global context condensing block $GCCB$, and a global student decoder S_{glo}. The

$OCBE_{loc}$ is designed for fusing multi-level features into a compact feature space followed by a local student decoder S_{loc} to reconstruct the feature representations, especially low-level features accurately. The first three modules compose the reverse distillation [8] which is effective for structural AD. To better capture global contextual correlations which we expect to have the benefit of logical AD, we additionally design the S_{glo} with a $OCBE_{glo}$ that can keep the most condensed contextual information, and the output is then decoded by the global student decoder S_{glo}. Since the teacher T is pre-trained on a large natural image dataset, it is expected to extract representative features for both normal and anomaly inputs. However, the two students are trained solely on anomaly-free samples, and they fail to mimic the teacher's behavior for either low-level structural anomalies or global logical anomalies during inference. The pixel-level anomaly scores are computed by comparing the decoded features from both students and the teacher features. The local student is primarily responsible for structural AD and the global student pays more attention to the global contextual constraints.

3.3. Local Knowledge Distillation

Different from the conventional discriminative paradigm where the student is also a feature extractor, reverse distillation [8] is in a generative manner to reconstruct the features extracted by the teacher. The student decoder receives dense encoded features and decodes the features from high-semantic levels to low-semantic levels. The dense feature space is likely to abandon unseen anomalous feature representations at inference to encourage feature discrepancies for anomalies. We use the reverse distillation [8] method for accurate feature reconstruction. Given an image I, the output of the first three residual stages of a pre-trained WideResNet50 [36] T extracts multi-layer intermediate features $F_T^l \in \mathbb{R}^{h_l \times w_l \times c_l}$, where $l \in \{1, 2, 3\}$. The $OCBE_{loc}$ encodes the features into the embedding ϕ_{loc}. The local student S_{loc} then generates the corresponding feature maps $F_{S_{loc}}^l$ from ϕ_{loc}. The student decoder has a symmetrical architecture with the teacher T while the input is high abstract feature representations and the down-sampling operations used in original ResNets [13] are replaced by up-samplings. The vector-wise cosine distance is used as the loss function for training the local student. A $2-D$ anomaly score map could be obtained at each layer scale

$$M_{loc}^l = \mathbf{1} - \frac{F_T^l \cdot F_{S_{loc}}^l}{\|F_T^l\| \|F_{S_{loc}}^l\|} \tag{1}$$

The final loss for training the local student is

$$\mathcal{L}_{loc} = \sum_{l=1}^{3} \left\{ \frac{1}{h_l \cdot w_l} \sum_{i=1}^{h_l \cdot w_l} M_{loc}^l \right\} \tag{2}$$

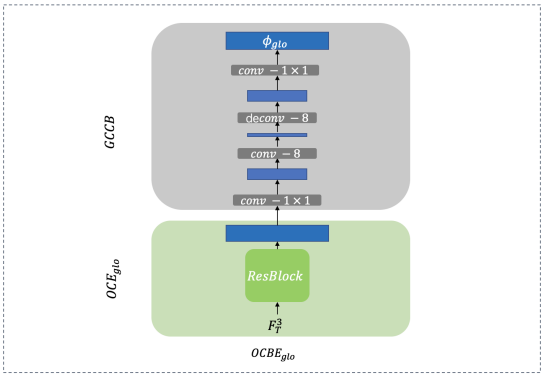

Figure 3. The one-class bottleneck embedding module for global student and the global context condensing block. We use the 4-th residual stage of ResNet as trainable OCE_{glo}.

3.4. Global Contextual Affinity Distillation

Although the encoder-decoder architecture naturally can keep the most important information and the multi-scale feature distillation paradigm could take both low-level information and high-level information into account, the student still cannot learn globally. Furthermore, the $OCBE_{loc}$ fuses low-level features into the final embedding space, which is beneficial for accurate low-level feature reconstruction but decreases the global contextual learning ability. We design the global context condensing block to keep the most important global information as shown in Fig. 3. It is realized by compressing the high-semantic level feature F_T^3 into a one-dimensional feature space with g channels and then restoring it to the original feature size. The output ϕ_{glo} of $GCCB$ is then decoded by a student S_{glo} that has the identical architecture as S_{loc}.

To further encourage the global student to better learn the global contextual information, different from the per-pixel cosine similarity loss used for the local student, we propose the contextual affinity loss for the global student. To learn the local feature embedding $f_{S_{loc},i}^l$, the local student S_{loc} can only learn from $f_{f_T,i}^l$, which is a benefit for accurately reconstructing local features. However, it fails to learn the contextual information from the whole image. For example, if a normal feature appears in the wrong position, the local student can't detect it as anomalous since the feature itself is a representation of a normal structure. We are inspired to propose the contextual affinity loss aiming to enable the student to learn a local feature $f_{S_{glo},i}^l$ from the whole feature map. For a feature vector $f_{T,i}^l$ from a feature map extracted by the teacher F_T^l, we first compute the cosine similarity between $f_{T,i}^l$ and all feature vectors to get a similarity list

$A^l_{T,i} = [a^{t,l}_{i,1}, ..., a^{t,l}_{i,h_l \cdot w_l}]$, where

$$a^{t,l}_{i,j} = \frac{f^l_{T,i} \cdot f^l_{T,j}}{\|f^l_{T,i}\| \|f^l_{T,j}\|} \tag{3}$$

We define the *contextual affinity* for a feature vector in the whole feature map as the probability distribution $P^l_{T,i} = [p^{1,l}_{T,i}, ..., p^{h_l \cdot w_l, l}_{T,i}]$ where

$$p^{i,l}_{T,i} = \frac{\exp(\frac{a^{t,l}_{i,j}}{\mathcal{T}})}{\sum_{j=1}^{h_l \cdot w_l} \exp(\frac{a^{t,l}_{i,j}}{\mathcal{T}})} \tag{4}$$

where $\mathcal{T}$ is the temperature. By converting the similarity list into a probability distribution, the scales of the contextual affinity for each feature vector are normalized, and the large spatial similarity relations that we believe are the most important elements are paid more attention to, while the less similar relations are ignored. By using a small $\mathcal{T}$, the probability distribution becomes harder, which means we only focus on a small portion of spatial relations in the feature map. Similarly, for the student feature embedding, it is intuitive to compute the corresponding contextual affinity probability distribution $P^l_{S_{glo},i}$ for $f^l_{S_{glo},i}$ *within the student feature map* $F^l_{S_{glo}}$ and minimize the difference between $P^l_{T,i}$ and $P^l_{S_{glo},i}$. However, in this case, the optimization of one feature vector $f^l_{S_{glo},i}$ is coupled with all the feature vectors in $F^l_{S_{glo}}$, making the optimization difficult [16]. For unsupervised knowledge distillation where we only have the knowledge distillation training target, we experimentally found the model can't converge. Considering that we also use the teacher model for inference, we compute the contextual similarity list $A^l_{S_{glo},i}$ and affinity probability distribution $P^l_{S_{glo},i}$ using the corresponding student feature embedding $f^l_{S_{glo},i}$ and the whole *teacher feature map* F^l_T, where

$$a^{s_{glo},l}_{i,j} = \frac{f^l_{S_{glo},i} \cdot f^l_{T,j}}{\|f^l_{S_{glo},i}\| \|f^l_{T,j}\|} \tag{5}$$

We then use KL divergence to evaluate the discrepancy between the contextual affinity distributions from the teacher and global student

$$\mathcal{KL}(P^l_{T,i}, P^l_{S_{glo},i}) = \mathcal{T}^2 \sum_{j=1}^{h_l \cdot w_l} p^{j,l}_{T,i} \cdot \log \left[\frac{p^{j,l}_{T,i}}{p^{j,l}_{S_{glo},i}} \right] \tag{6}$$

The $p^{j,l}_{T,i}$ in Equation (6) could be interpreted as a weighting factor. Large $p^{j,l}_{T,i}$ values that indicate the spatial relations with high similarities are paid more attention to, while the KL divergence tends to neglect less similar relations. Note that since each feature vector always has the largest similarity with itself, it is not a contradiction with accurately reconstructing the feature vector. The high-similarity relations are the guiding *signposts* for training the student feature vector from the global context. During inference, the global student fails to capture the global contextual information for logical anomalies. Similarly, the final loss for training the global student is

$$\mathcal{L}_{glo} = \sum_{l=1}^{3} \left\{ \frac{1}{h_l \cdot w_l} \sum_{i=1}^{h_l \cdot w_l} \mathcal{KL}(P^l_{T,i}, P^l_{S_{glo},i}) \right\} \tag{7}$$

3.5. Pixel and Image Anomaly Scoring

Following Equation (1) and Equation (6), we could get anomaly score maps M^l_{loc} and M^l_{glo} for the l-th layer from the local and global student. Each element in the score map indicates the feature or contextual affinity discrepancy. To get precise multi-scale AD and localization, we first upsample each score map to the image resolution and conduct element-wise addition for each student. The final score map for an input image I is the combination of the two students' normalized detection results

$$\begin{aligned} M(I) &= \frac{M_{loc} - \mu_{loc}}{\sigma_{loc}} + \frac{M_{glo} - \mu_{glo}}{\sigma_{glo}}, \\ M_{loc} &= \sum_{l=1}^{3} \Psi(M^l_{loc}), M_{glo} = \sum_{l=1}^{3} \Psi(M^l_{glo}) \end{aligned} \tag{8}$$

Where Ψ is the bilinear up-sampling operation, μ and σ are the mean and standard deviation values, respectively. They are computed on the validation set $\mathcal{S}^v$ or the training set $\mathcal{S}^t$ if $\mathcal{S}^v$ is not available. The image-level anomaly score is derived by choosing the maximum score from the final score map. We apply a Gaussian filter before image-level anomaly scoring to remove local noises.

4. Experimental Results

4.1. Experimental Settings

Datasets. We use two public datasets for unsupervised anomaly detection and localization: MVTec LOCO AD [1] and the modified MVTec AD [2]. The recently introduced MVTec LOCO AD covers both structural anomalies and logical anomalies. The dataset consists of five object categories, providing 1,772 anomaly-free images for training, 304 for validation, and 1,568 for testing. Each test image is either anomaly-free or has at least one structural or logical anomaly, with pixel-level annotations. The MVTec AD dataset features 15 distinct object or texture categories. Although the majority of these samples are either anomaly-free or exhibit structural anomalies, 37 test images are specifically identified as logical anomaly samples and are split out as a test subset for logical anomalies [1].

Model training. All images are resized to 256×256 resolution. We follow the one-model-per-category setting of

Table 1. *Anomaly Localization* results on MVTec LOCO AD dataset [1]. The area under the sPRO curve is computed up to an average false positive rate of 0.05. We report the mean scores for structural and logical anomalies. The best scores are in bold.

Method	Breakfast Box	Screw Bag	Pushpins	Splicing Connectors	Juice Bottle	Mean
AE	0.189	0.289	0.327	0.479	0.605	0.378
VAE	0.165	0.302	0.311	0.496	0.636	0.382
MNAD [24]	0.080	0.344	0.357	0.442	0.472	0.339
VM	0.168	0.253	0.254	0.125	0.325	0.225
f-AnoGAN [31]	0.223	0.348	0.336	0.195	0.569	0.334
SPADE [6]	0.372	0.331	0.234	0.516	0.804	0.451
US [3]	0.496	0.602	0.523	0.698	0.811	0.626
RD [8]	0.326	0.568	0.597	0.702	0.840	0.607
PatchCore-25 [28]	0.510	0.577	0.504	0.731	0.794	0.623
GCAD [1]	0.502	0.558	0.739	**0.798**	**0.910**	0.701
DSKD (Ours)	**0.568**	**0.627**	**0.825**	0.767	0.865	**0.730**

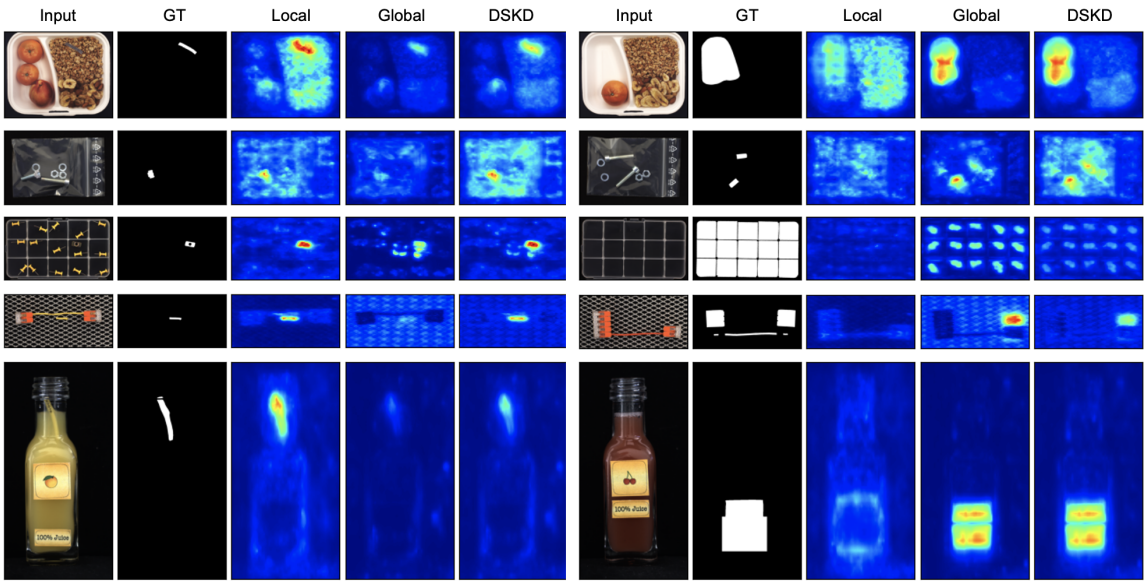

Figure 4. Qualitative examples for each category. Structural anomalies from top to bottom: "contamination" on cereals, "broken" screw head, "contamination" in one compartment, "additional" short cable, and "contamination" on the juice. Logical anomalies from top to bottom: "missing" one nectarine and one tangerine, one short screw "replaced" by a long screw, "missing" pushpins for all compartments, "wrong connector pair", and "misplaced" bottom label. We show the detection results from the local student which is identical to RD [8], the global student, and the final detection results.

previous studies. The two students are trained simultaneously. For each student, we use the same training configuration as [8]. We use Adam optimizer using $\beta = (0.5, 0.999)$ with a fixed learning rate 0.005. Each student is trained for 200 epochs with the same batch size of 16. The channel dimension g of $GCCB$ is set to 1024 by default and the temperature $\mathcal{T}$ is set to 1.

Evaluation metrics. We use the area under the receiver operating characteristic (AUROC) score, a threshold-free metric, for image-level AD evaluation. The AUROC is also appropriate for evaluating structural AD. However, logical anomalies, such as a missing object, present challenges in pixel-wise annotation and segmentation. Therefore, as recommended in [1], we adopt the saturated per-region overlap (sPRO) metric—a generalized version of the PRO metric from [2]—to assess anomaly localization performance.

4.2. Results on LOCO

We compare our proposed method against autoencoders including a vanilla autoencoder (AE), a variational autoencoder (VAE), a memory-guided autoencoder (MNAD) [24], f-AnoGAN [31], Variation Model (VM) [33], Uninformed Students (US) [3], SPADE [6], Reverse Distillation (RD) [8] and Global Context Anomaly Detection (GCAD)

Table 2. The image-level *anomaly detection* AUROC scores on MVTec LOCO AD dataset. The best scores are in bold.

Method	Structural AD	Logical AD	Mean
AE	0.565	0.581	0.573
VAE	0.548	0.538	0.543
MNAD	0.702	0.600	0.651
VM	0.589	0.565	0.577
f-AnoGAN	0.627	0.658	0.642
SPADE	0.668	0.709	0.689
US	**0.883**	0.664	0.773
RD	0.867	0.669	0.768
PatchCore-25	0.855	0.759	0.807
GCAD	0.806	**0.860**	0.833
DSKD (Ours)	0.869	0.812	**0.840**

[1]. The same data augmentations are used as GCAD [1] throughout our experiments.

We begin by presenting the anomaly localization results in Table 1. Our proposed method achieves an average score of 0.73 over five categories. Notably, on items like the breakfast box, screw bag, and pushpins—where most competing methods achieve only modest scores due to complex contextual logical constraints—our method surpasses their performance significantly.

Table 2 shows the results for image-level AD. While many existing methods have demonstrated strong performance in structural AD—particularly the knowledge distillation-based US [3] and RD [8] with their patch-based and per-pixel training targets—their performance significantly declines in logical AD. Notably, GCAD [1], which builds upon US [3] to enhance logical AD, does achieve the highest score in this domain. However, its performance in structural AD is compromised. Our proposed method significantly boosts logical AD capabilities without compromising the structural AD strengths of RD [8], setting a new benchmark with a score of 0.84.

In Fig. 4, we provide qualitative visualization results for each category. Each category showcases a structural anomaly image on the left and a logical anomaly image on the right. The local student excels in detecting low-semantic level structural anomalies but struggles with capturing long-range dependencies. In contrast, the global student effectively learns global contextual constraints but underperforms in fine-grained local structural AD. By employing the DSKD, our method is equipped to detect both types of anomalies.

4.3. Results on the Modified MVTec AD

We report the anomaly detection and localization results on the modified MVTec AD dataset in Table 3. The results show that some of the existing methods perform well

Table 3. Experimental results on the modified MVTec AD [2]. We report the image-level AUROC scores / the normalized AU sPRO scores with an integration limit of 0.05. The best results are in bold and the second-best results are with underlines.

Method	Structural	Logical	Mean
AE	0.761/0.337	0.718/0.224	0.740/0.281
VAE	0.766/0.336	0.737/0.215	0.751/0.276
MNAD	0.709/0.294	0.427/0.032	0.568/0.163
VM	0.690/0.240	0.679/0.069	0.684/0.155
f-AnoGAN	0.751/0.290	0.751 0.231	0.751/0.261
SPADE	0.898/0.632	0.906/0.647	0.902/0.640
US	0.936/0.762	0.747/0.417	0.842/0.590
RD	**0.986/0.793**	0.914/0.477	0.950/0.635
PatchCore-25	0.985/0.790	**0.998**/0.619	**0.992**/0.705
GCAD	0.871/0.716	0.991/**0.863**	0.931/**0.789**
DSKD (Ours)	0.955/0.755	0.906/0.649	0.931/0.702

on structural AD, while still showing the ability for logical AD. This is because of the concise nature of the modified MVTec AD dataset and the limited number and type of logical anomaly samples. The SPADE [6] achieves a 0.906 image-level AUROC score which is even higher than structural AD. An underlying assumption is that SPADE uses the high-semantic level feature, *e.g.*, the output of the 4-th residual block of a pre-trained ResNet for image-level anomaly scoring. RD [8] also achieves a high score of 0.914 because it benefits from the compact one-class bottleneck embedding space that also contains high-semantic level information. PatchCore achieves SOTA image-level AD results with the help of using locally aware patch features. Similar to the results on LOCO, GCAD [1] is capable of logical AD, at the cost of an obvious performance drop for structural AD. The proposed DSKD showed better logical and average anomaly localization performance than RD [8].

We visualize two logical anomaly samples from the transistor and cable category in Fig 5, where one example is better detected by the local student while the remaining one is better detected by the global student.

Limitations. Although our proposed method achieves the second-best overall performance, we observed similar limitations with results on the LOCO dataset. Our global student could capture global logical constraints but is not sensitive to small-sized defects and ambiguous anomalies violating both low-level and long-range dependencies. In Fig. 5, a blue cable replaced by a green one may also be defined as a kind of color contamination. However, the global student could identify a missing object or an object in the wrong place while identifying the right position.

4.4. Ablation Studies

We investigate the effectiveness of the dual-student architecture and the contextual affinity loss and assess the

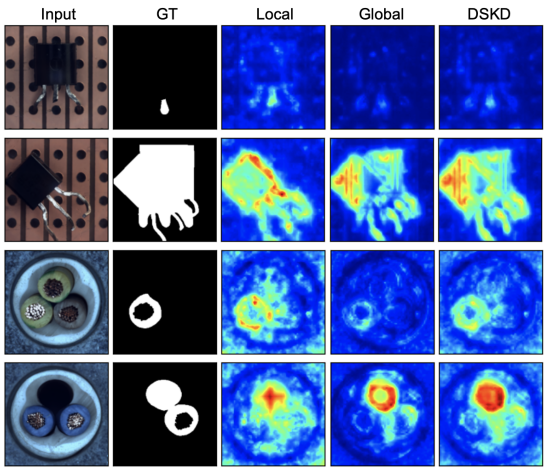

Figure 5. Qualitative results of logical anomaly detection on the modified MVTec AD dataset.

Table 4. Anomaly detection results using different student architecture and loss. Pi means per-pixel cosine similarity loss and CA means our proposed contextual affinity loss.

Model	Loss Type	Structural	Logical	Mean
Local	Pi	0.739	0.474	0.607
Local	CA	**0.752**	0.507	**0.630**
Global	Pi	0.547	**0.693**	0.620
Global	CA	0.336	0.640	0.488

Table 5. Anomaly detection results with different loss combinations for the DSKD.

$\mathcal{L}_{loc}$	$\mathcal{L}_{glo}$	Structural	Logical	Mean
Pi	Pi	0.748	0.675	0.711
CA	CA	0.752	0.678	0.708
Pi	CA	**0.754**	**0.707**	**0.730**

sensitivity of hyperparameters. The performance of a single student trained with different losses is reported in Table 4. Benefiting from the RD [8] architecture which has a low-level feature bias, and our contextual affinity learning scheme, the local student is capable of logical AD and yields the best overall performance. The global student trained with per-pixel cosine similarity is enhanced for better low-level feature reconstruction which in turn improves both structural and logical AD performance.

Table 5 gives qualitative comparisons of our DSKD trained with different loss pairs. Although the global student trained with per-pixel loss outperforms the one trained with contextual affinity loss, however, the DSKD design releases the constraint of accurate low-level feature reconstruction for the global student and encourages the global student to focus on global contextual information.

Table 6. Mean detection results with different g dimension values. "w/o" means $GCCB$ is not used and for $g = 2048$ channels, we do not use $conv1 \times 1$ layers to downsample and upsample the channel dimensions.

g	w/o	512	768	1024	1280	2048
AU sPro	0.607	0.714	0.721	**0.730**	0.720	0.729

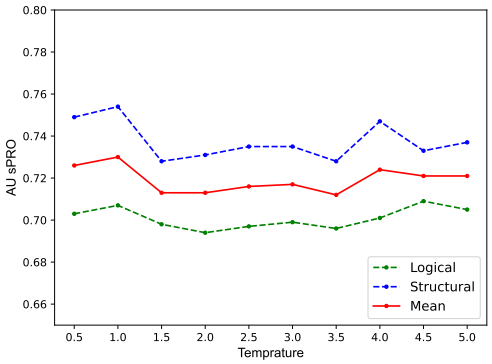

Figure 6. Impact of temperature $\mathcal{T}$.

We also investigate the impact of $GCCB$ along with its channel dimensions g. The results are shown in Table 6. The use of $GCCB$ improved the performance by a large margin and performed well with various g values.

The results with different $\mathcal{T}$ are shown in Fig. 6. A large $\mathcal{T}$ makes the distribution softer and covers wider relations. Although it may confront the low-level feature reconstruction ability, our method is stable for a wide range of $\mathcal{T}$.

5. Conclusion

We proposed the dual-student knowledge distillation framework and contextual affinity loss for structural and logical anomaly detection. The local student aims for accurate low-level feature reconstruction and the global student learns global context. The proposed contextual affinity loss further enhances capturing long-range correlations. The use of both teacher and student networks for *unsupervised* anomaly detection at inference enables us to compute the contextual affinity loss for the student using both teacher and student features, decoupling the training of each student feature vector. Experiments showed that the proposed method outperformed previous studies and achieved SOTA performance on public benchmarks.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 23H00482 and 20H05952.

References

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. 1, 2, 3, 5, 6, 7

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 5, 6, 7

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 1, 2, 3, 6, 7

[4] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 1, 2

[5] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 1

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2, 6, 7

[7] David Dehaene, Oriel Frigo, Sébastien Combrexelle, and Pierre Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. *arXiv preprint arXiv:2002.03734*, 2020. 2

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 1, 2, 3, 4, 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 *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[10] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 2

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. 2

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

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

[16] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 5

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[18] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 2

[19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017. 3

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[21] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020. 2

[22] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 2, 3

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[24] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 2, 6

[25] Masoud Pourreza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, and Mohammad Sabokrou. G2d: generate to detect anomaly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2003–2012, 2021. 2

[26] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. 1

[27] Lior Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33(1):1–39, 2010. 2

[28] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 6

[29] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. 1

[30] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on 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. 2, 6

[32] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 3

[33] Carsten Steger, Markus Ulrich, and Christian Wiedemann. *Machine vision algorithms and applications*. John Wiley & Sons, 2018. 6

[34] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for unsupervised anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021. 1, 3

[35] 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

[36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 4

[37] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2

[38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *European Conference on Computer Vision*, pages 539–554. Springer, 2022. 2

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3