Scientific Reports

https://doi.org/10.1038/s41598-026-37667-9

Article in Press

A boosting strategy based on feature mimicking with attention for visual anomaly detection

Received: 1 December 2024

Accepted: 23 January 2026

Published online: 26 March 2026

Cite this article as: Zheng B., Gan Y., Wang L. *et al.* A boosting strategy based on feature mimicking with attention for visual anomaly detection. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-37667-9

Boyuan Zheng, Yi Gan, Lianggang Wang, Xunchao Cong, Chao Hu & Di Wang

A Boosting Strategy Based on Feature Mimicking with Attention for Visual Anomaly Detection

Boyuan Zheng[1,*], Yi Gan[1], Lianggang Wang[1], Xunchao Cong[1], Chao Hu[1], and Di Wang[1]

[1]The 10th Research Institute, China Electronic Technology Group Corporation, Chengdu, 610036, China
[*]corresponding.zhengboyuanneu@163.com

ABSTRACT

Anomaly Detection (AD), referred to as detecting anomalies from images or videos, is commonly considered a one-class classification task (i.e the model is only trained on the normal training data to identify abnormal data during the inference period). A distinguished category of the existing works is the reconstruction-based method where models are trained to reconstruct the inputs and leverage the reconstruction error with the target as an abnormality score. However, without considering global information, these methods may fail due to the generalization capability of the reconstruction model. To tackle this problem, we propose a proxy task of feature mimicking that can be integrated into a wide range of anomaly detection frameworks and utilizes their inherently discriminative hidden-layer features. Moreover, a novel attention module that takes the feature inconsistency matrix generated by the feature-mimicking task as input is presented. The feature inconsistency guided attention module enables the reconstruction-based model to focus on the region or pattern where the global, semantic feature inconsistency is higher. We integrate our method into several state-of-the-art methods for anomaly detection on images and videos. The empirical results show that ou0r method can bring improvement and achieve new SOTA performance on MVTec AD, CUHK Avenue and ShanghaiTech.

Introduction

Due to the wide usage of cameras and surveillance systems, Anomaly Detection (AD) has become an essential and challenging task with the broad application including public security (identifying unexpected events such as criminal activities, car accidents, etc. from surveillance video)[1–11] and industrial quality control (detect and localize defected regions in industrial objects and materials)[12–22]. The defected objects and unexpected events are rare in the real world and the manual annotation is overly time-consuming. Besides, the anomalies (especially in surveillance video) can not be defined previously making it unlikely to collect all kinds of anomalies. Therefore, anomaly detection is usually considered a one-class classification task, where models are only trained on the normal samples while classifying abnormal samples from normal ones during inference time. Since the training sets only contain normal data the AD task is also considered an unsupervised or self-supervised task and the typical solution is to propose proxy tasks that perform well on normal samples while badly on abnormal ones. However, existing methods still face limitations in capturing complex temporal and spatial correlations in multivariate time-series data. Our motivation is to design an anomaly detection framework with good generalisation capabilities that not only exploits self-supervised learning, but also enhances feature representation learning for improved adaptability in the real world.

The proxy tasks for anomaly detection can be roughly classified into two categories: feature-based methods[14,23–25] and reconstruction-based methods[5,15,16,26–28]. Typical feature-based method[29–31] project input images into a high-dimensional feature space where normal and abnormal samples can be distinguished easily according to a probability distribution, distance measurements or supervised classification. Recent studies[6,12,14] introduced knowledge distillation[32] as a proxy task, where knowledge is transferred from deeply pre-trained teacher networks to relatively shallow student networks. Since the model is trained only on the normal data, lower embedding similarities can be expected for abnormal samples. However, feature-based approaches rely on predefined feature distributions, making it difficult to capture subtle anomaly patterns. Therefore, it is important to integrate features at multiple network levels and improve sensitivity to anomaly-related biases through structured representation learning.

Another distinguished category of anomaly detection is reconstruction-based methods, where models are trained to reconstruct the inputs (*e.g.* predicting future frame of surveillance video[1], repair image with pseudo-defects[15]) and leveraging the reconstruction error with the target as an abnormality score. However, such methods may fail due to the excessive generalization ability of the model. For example, an abnormal image may be well reconstructed and classified as normal mistakenly. Recent works following this approach attempt to increase the difficulty of the proxy task to prevent models from possessing excessive generalization ability. To achieve this goal, memory-augmented structure[2,3], image masking strategy[33], multi-task learning[6], etc. are introduced to anomaly detection. However, without considering global information, methods

following this approach may still fail in certain cases inevitably. Therefore, how to effectively combine local and global information in the reconstruction process remains a key issue that needs to be further explored in reconstruction-based methods.

Our work builds upon these insights and introduces a feature-mimicking strategy to address these shortcomings. Unlike prior reconstruction-based approaches that rely solely on pixel-wise loss, our method constrains the student network to replicate intermediate representations of a pre-trained teacher model. This constraint ensures that the model does not overgeneralize on anomalies and maintains a clear distinction between normal and abnormal patterns. By integrating feature-level supervision, we provide a novel framework that balances the advantages of feature-based and reconstruction-based approaches, enhancing interpretability and robustness in anomaly detection.

Specifically, we introduce a general proxy task of feature mimicking which can be integrated into a wide range of reconstruction-based methods. We utilize a teacher-student structure where the teacher network is the pre-trained network from the previous works and the student network is a similarly structured network with the same proxy task as the teacher and an additional task of feature mimicking. The combination of feature-mimicking and original reconstruction-based tasks provides an additional global semantic anomaly criterion and utilizes the middle layer feature of existing methods which can naturally produce discriminative descriptions for normal patterns. Furthermore, a novel attention mechanism based on the feature inconsistency matrix of feature mimicking task is presented. Such an attention module has two important advantages: (*i*) The attention module takes the feature inconsistency matrix between the teacher and the student as input and refines the student feature based on the feature inconsistency which is higher when the input is an abnormal sample. With the guidance of the refined feature, the student network pays more attention to anomalous regions or patterns, making it harder to reconstruct anomalies. (*ii*) The attention model enables the feature inconsistency matrix of the feature-mimicking task to guide the follow-up tasks. This means that our method is not a simple combination but one can effectively influence and guide another. Moreover, the proposed method can be integrated into a wide range of existing works, thus being very universal.

We integrate our method into various state-of-the-art frameworks of anomaly detection[34–41]. A simple but effective way of pre-training the teacher network is proposed to combine with our method. Extensive experiments are conducted on three public benchmarks for image and video anomaly detection: MVTec AD[42], CUHK Avenue[43] and ShanghaiTech[44] and the empirical results shows that the integration with our method brings significant improvement. For example, the anomaly localization average precision (AP) of[15] increases from 68.4% to 74.8% and the improvement is more than 25% for some classes from MVTec AD. Besides, with the combination of our method, we are able to report new state-of-the-art on MVTec AD, CUHK Avenue and ShanghaiTech.

In summary, our contributions can be concluded as follows:

- We introduce a proxy task of feature mimicking which can be integrated into a wide range of anomaly detection frameworks and utilize their inherently discriminative hidden-layer features.
- We propose a novel attention module that can guide the follow-up task with feature inconsistency.
- Extensive experiments are conducted and the empirical results show that our method can improve the performance and achieve state-of-the-art performance on multiple models and benchmarks.

Related Works

Reconstruction-based anomaly detection method. As the anomalies rarely appear in the real world compared with the normal ones. Anomaly detection is usually considered a self-supervised or unsupervised task. An anomaly detection prototype adopts generative models, such as Auto-Encoder (AE)[32], Variational Auto-Encoder (VAE)[5,45], Generative Adversarial Net (GAN)[46], etc. for image reconstruction. The hypothesis these methods hold is model trained on normal data can only successfully reconstruct normal samples but fail for abnormal samples. The hypothesis these methods hold is model trained on normal data can only successfully reconstruct normal samples but fail for abnormal samples. This assumption is sometimes invalid as the excessive generalization capability of the model makes it possible to reconstruct anomalies as well. Recent works attempt to increase the difficulty of the reconstruction-based task to prevent models from possessing excessive generalization ability.[2,3] introduce memory-augmented structure to anomaly detection which memories the normal pattern of the training data, and anomalies are harder to reconstruct during inference. HF2VAD[5] uses a conditional VAE to predict the future frame with the raw images and the reconstructed optical flows of the previous frames. The abnormal frames are hard to reconstruct as the result is heavily influenced by the reconstructed flows. TSGAD[45] uses a graph-attentive variable autoencoder (GA-VAE) to capture the distribution of normal human behavior in pose data for anomaly detection. In addition, some works[15,16,41] train the detection model with pseudo-anomalous samples or adopt other novel strategies for anomaly detection. DRAEM[15] generates pseudo-anomalies by adding random augmented noise to the training data. DMAD[41] addresses the trade-off in anomaly detection by designing a Pyramid Deformation Module (PDM) to measure multi-scale deformations from reconstructed references to original inputs and an Information Compression Module (ICM) to learn prototypical normal patterns. However,

using the fine-grained reconstruction error as the anomaly classifier may fail in certain inevitably without considering the global information. Our work improves the accuracy and effectiveness of anomaly detection by designing special features to mimic the agent task in order to fully integrate global and fine-grained information to jointly guide the judgement of anomaly detection.

Convolutional attention module. Attention module[47–50] is widely used in convolutional networks for its effective improvement and plug-and-play convenience. The typical attention module acquires the attention information with the average or max polling result of the input feature. With the attention information, parts of the input feature are selectively emphasized or suppressed which enables the feature to have a global view. SEnet[47] proposes a channel-attention mechanism for feature recalibration. Given an input feature, SEnet employs a global polling layer for each channel followed by two fully connected (FC) layers and a Sigmoid function to generate channel attention weights. Following this approach, ECAnet[49] emphasizes the influence of adjacent channels by replacing the FC layers in SEnet with a 1-D convolutional layer. Some other works focus on spatial attention and its combination with channel attention. In CBAM[48], the spatial attention weights are calculated by a 2-D convolutional layer with the average and max pooling results of the input feature along the channel axis. SKnet[50] generate the attention refined feature by fusing the attention information produced by different kernel-sized convolutional layers of which the importance is considered to be different. Our work fully combines spatial and channel attention mechanisms to enhance the proxy task of feature imitation through an attention mechanism to increase the complexity of the proxy task, making it possible to greatly improve the generalisation ability of the baseline methods.

Method

In this paper, we present a feature inconsistency-based attention-guided framework for image and video anomaly detection which can be embedded into a wide range of anomaly detection methods. The framework starts with a feature-mimicking task that learns the inherently discriminative features from the pre-trained networks, followed by a feature inconsistency guided attention to focus the model on the regions or patterns where the global semantic feature inconsistency is higher. The core motivation behind our framework is that normal samples typically exhibit high semantic consistency between the feature representations learned by the teacher and the student, while anomalous samples contain semantic patterns that the student—trained only on normal data—cannot reproduce. Unlike pixel-level reconstruction error, this feature-level inconsistency is robust to noise, illumination, and textural variations, and aligns better with human-perceived semantic deviations.

Proxy Task of Feature Mimicking

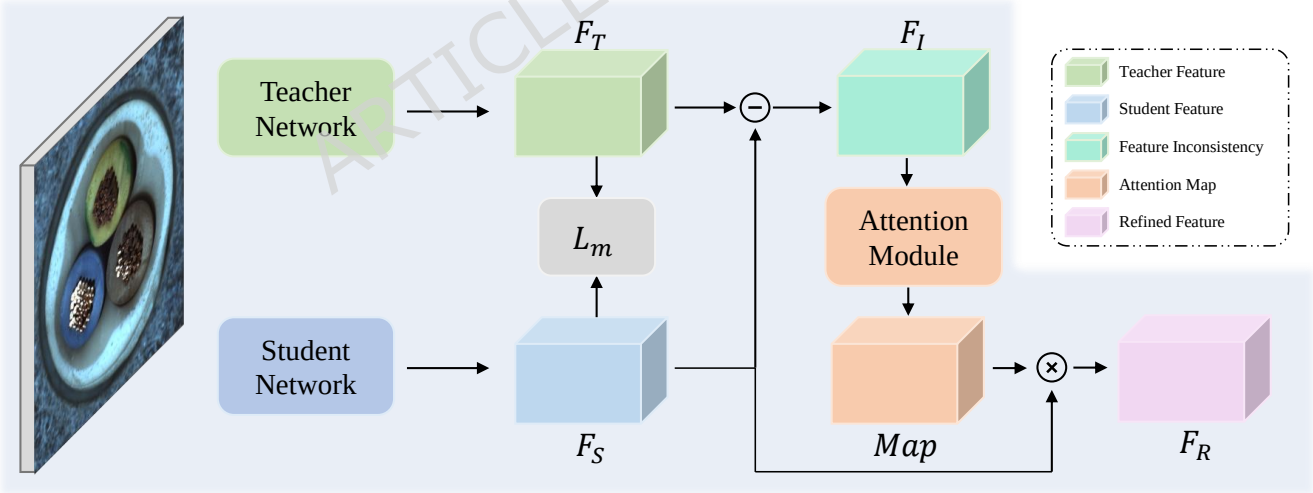

Figure 1. Overview of our method. The student network takes the same architecture and form of follow-up proxy task as the teacher. For each input, the student is trained to mimic the feature produced by the teacher additionally. An attention module performs feature refinement by using the feature inconsistency to selectively emphasize or suppress the feature region of large or small feature inconsistency. With the refined feature, the student network learns to finish the follow-up tasks with the refined feature.

In our framework, a teacher-student structure is adopted where the teacher network is the pre-trained and parameter-fixed network of existing works and the student network is the network of the same structure as the teacher. The middle-layer features produced by the teacher network can inherently describe the normal pattern as the model is trained to complete the

reconstruction-based task on the normal training data, but such features are usually ignored. The student network learns the same task as the teacher but an additional task of feature mimicking. As shown in Figure.1, the student network is trained to mimic the teacher networks with the constraint of L_{fea}. With training on the normal data, the teacher feature Fea_T and the student feature Fea_S tend to be discrepant when the input is anomalous and vice versa for the normal data. The combination of the feature imitation task and the original reconstruction-based task increases the difficulty of the proxy tasks which restricts the generalization capability and enables the model to detect anomalies at different scales.

In this work, the term global information refers specifically to semantic-level globality rather than structural or temporal globality. The feature inconsistency Fea_I provides a global, semantic anomaly detection criteria and the hybrid of the feature inconsistency and the fine-grained reconstruction error improves the overall performance of anomaly detection, especially on video data. The teacher's intermediate representation encodes object-level abstraction learned from the reconstruction task, and the feature inconsistency aggregates semantic divergence across the entire spatial field. This provides a stable global semantic criterion for anomaly detection. We do not claim to model temporal or geometric dependencies; instead, our semantic globality is orthogonal and complementary to these forms of global information. Formally, the feature inconsistency matrix are acquired by the $l1$ loss, as it directly reflect the discrepancy between Fea_T and Fea_S:

$$Fea_I = |Fea_T - Fea_S|. \tag{1}$$

The attention module generates the attention map based on the feature inconsistency matrix and recalibrates the student features. The refined features are used to guide the follow-up task which enables the model to focus on suspected anomalies at the feature level.

Feature Inconsistency Guided Attention

Given a feature inconsistency matrix $X \in \mathbb{R}^{h\times w\times c}$, the attention module generates the attention map M in three ways as shown in Figure.2.

Channel Attention. Following SEnet[47], the input X is reduced to a vector $x \in \mathbb{R}^c$ through a global pooling on each channel. The channel attention map $M \in \mathbb{R}^c$ is calculated by an MLP followed by a Sigmoid function. In the MLP formed of two FC layers, vector x is first squeezed to $\mathbb{R}^{\frac{c}{r}}$ and recovered to $\mathbb{R}^c$.

Spatial Attention. Similar to CBAM[48], the input X is reduced to a tensor $x \in \mathbb{R}^{h\times w\times 2}$ by concating the results of average polling and max pooling. The module generates the spatial attention map $M \in \mathbb{R}^{h\times w}$ using a convolutional layer and a Sigmoid activation layer with tensor x as the input.

Spatio-Channel Attention. Typical attention mechanism generates the attention weight map by the feature itself which means dimension reduction has to be conducted (e.g. channel reduces the dimension along the width and height axis, spatial attention reduce along the channel axis) as the weight of each part is acquired by comparison with the others. Differently, the feature inconsistency matrix inherently represents the abnormal regions or patterns at a feature level, which means a 3-D attention mask can be acquired directly.

Novelty of the Proposed Attention Mechanism. Existing attention modules such as SE[47] and CBAM[48] generate attention weights from the feature itself, aiming to highlight visually salient regions. Our attention mechanism is fundamentally different in both its information source and its functional objective:

- Different input source: Traditional attention takes feature as input. Our attention takes the semantic discrepancy as its sole input.
- Different purpose: Traditional attention enhances salient visual patterns. Our attention highlights regions that cannot be semantically mimicked, i.e., regions of semantic inconsistency.
- Different role in the framework: Our attention is not for improving perceptual quality but is a functional mechanism to amplify anomaly-relevant semantic deviations.

Thus, even though the structural form is simple, this constitutes a functionally new category of discrepancy-guided attention, which does not appear in prior works.

Formally, the spatial-channel attention module takes the feature inconsistency matrix X directly as the input. Subsequently, the spatial-channel attention map $M \in \mathbb{R}^{h\times w}$ is calculated as follows:

$$M = \sigma(D(\delta(E(X)))), \tag{2}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $\delta(\cdot)$ is the ReLU activation, $E(\cdot)$ and $D(\cdot)$ stand for two convolutional layers for encoding and decoding. The input X is first reduced to $\mathbb{R}^{h\times w\times \frac{c}{r}}$ and recovered back with a small-sized auto-encoder structure composed of two convolutional layers. Details setting including the kernel size (k), reduction ratio (r), etc. are discussed in Section 4.6.

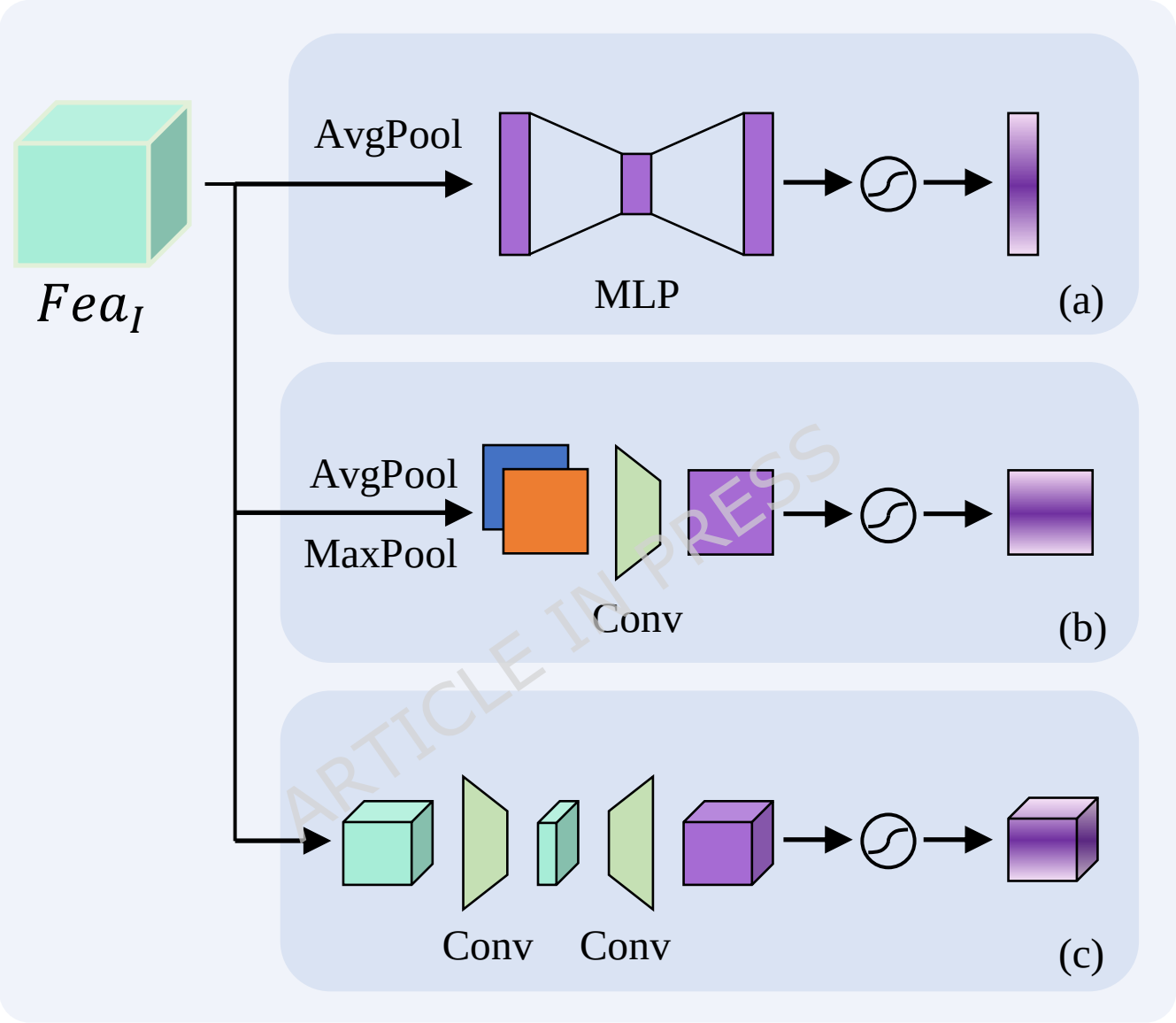

Figure 2. Diagrams of three ways of attention module with the feature inconsistency matrix as the input. (a) Following[47], the Channel Attention (CA) utilizes an average pooling layer, a multilayer perceptron (MLP) composed of two fully connected (FC) layers and a Sigmoid function to output a channel attention map. (b) Similar to[48], the Spacial Attention (SA) utilizes the average and max pooling outputs along the channel axis and forwards them to a convolution layer and a Sigmoid function to produce a spatial attention map. (c) The Spatio-Channel Attention (SCA) map are calculated by a small-scale auto-encoder formed of two convolution layer and a Sigmoid activation layer with the feature inconsistency matrix directly as the input.

Loss Function

We adopt the $l2$ loss instead of $l1$ to constrain the discrepancy between the teacher and student feature as the discrepancy is usually very large at the beginning of the training. Let $T(\cdot)$, $S(\cdot)$, and I denote the teacher network, student network, and the input image or video frame, respectively. We define the feature consistency loss as follows:

$$L_{fea} = \|T(I) - S(I)\|_2^2 = \|Fea_T - Fea_S\|_2^2. \tag{3}$$

As the student network is trained with the proxy task of feature mimic and the original task of the teacher network at the same time. The total loss takes the form as follows:

$$L_{total} = L_{fea} + L_o, \tag{4}$$

where L_o stands for the loss function of the original task of the teacher network.

Anomaly Scoring

The anomaly scores take similar forms as the loss function. The feature inconsistency provides global and semantic anomaly criteria as follows:

$$S_{fea} = \|Fea_T - Fea_S\|_2^2. \tag{5}$$

To combine the original score function with the feature inconsistency score, a hybrid score S_{total} is calculated as follows:

$$S_{total} = \lambda \cdot S_{fea} + (1-\lambda) \cdot S_o, \tag{6}$$

where S_o denotes the original score function of the teacher network, $\lambda \in (0,1)$ is a hyperparameter that controls the importance of the feature inconsistency score with respect to the original score. Specifically, we only calculate the hybrid score when the original score function is based on reconstruction error[3,5]. For the method where the score function is calculated otherwise[7,15], we adopt the original score function instead as the combination with feature inconsistency does not much change the results.

Data Availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Code Availability

You can find our main code in https://github.com/jtkullo/FMABS.

Experiments

Datasets

Avenue, ShanghaiTech, and MVTec AD are valuable benchmarks covering diverse domains of anomaly detection. Each dataset is carefully selected for its unique challenges, facilitating model evaluation in varied environments—ranging from video surveillance and urban settings to industrial contexts. Their broad scope enables a comprehensive assessment of a model's generalization ability and robustness across distinct domains.

Avenue. Avenue is specifically designed for video anomaly detection, comprising 16 training videos and 21 testing videos captured at a resolution of 640×360 using a static camera. It includes 47 types of anomalies, such as running, throwing bags, or walking in the wrong direction. With its diverse anomaly categories and real-world settings, Avenue provides an ideal platform for evaluating methods that need to detect unusual human behaviors in a controlled environment. It strikes a balance between video length and anomaly variety, thereby offering a manageable yet comprehensive test case for video surveillance anomaly detection systems.

Shanghaiech. The ShanghaiTech Campus dataset is one of the largest benchmarks in video anomaly detection, featuring 274,000 training frames and 42,000 testing frames collected from 13 distinct scenes. It includes 130 abnormal events, such as vehicles driving on sidewalks, fights, and other atypical activities. We select this dataset due to its scale and diversity, crucial for testing anomaly detection algorithms in dynamic, real-world urban settings. Its substantial data volume and broad range of anomalies demand robust learning from varied scenes, ensuring reliable predictions across multiple scenarios.

MVTec AD. MVTec AD focuses on industrial image anomaly detection, encompassing 15 categories (5 texture types and 10 object types). The training set contains 3,629 anomaly-free images, while the testing set has 1,725 images, some with anomalies and others without. This dataset is essential for assessing the detection and localization of industrial product surface defects (e.g., scratches, holes, or discoloration). We include it due to its dual challenges at both the object and texture levels, making MVTec AD a critical benchmark for evaluating model performance in industrial quality control applications.

Evaluation Criterions

Video Anomaly Detection. In terms of video anomaly detection, we measure the area under the receiver operation characteristic (AUROC) by plotting the true positive rate (TPR) versus the false positive rate (FPR). In video anomaly detection, TPR and FPR stand for the percentage of frames containing anomalous events that are correctly classified and the percentage of normal frames that are classified as abnormal by error, respectively. High AUROC indicates better performance of anomaly detection.
Image Anomaly Detection. We evaluate the image anomaly detection with the image-level AUROC as a metric. For anomaly localization, we take the average precision (AP) and the pixel-level AUROC as the evaluation metrics. For image-level and pixel-level AUROC, TPR and FPR are computed by the image and the pixels that are correctly or mistakenly classified, respectively.

Implementation Details

For the baselines chosen to combine with our method, we directly use the officially released code and the same hyper-parameter settings (including learning rate, training epochs, etc.) from each method. The hyper-parameter λ form Eq.(6) is determined by a grid search to achieve the optimal results.

Video Anomaly Detection

Baselines. The input data much effect the results of video anomaly detection. The recent state-of-the-art works are mainly based on 3 ways of data pre-processing.[1–3] takes the whole frames as the input and does not modify the input image.[4,5] take a casade-rcnn as an object detector and temporal gradients as a motion detector to extract foreground objects and use the extracted spatio-temporal cubes (STCs) as the input.[6,7] takes the objects detected by YOLOv3 as the input. We choose the sota methods[3,5,7] based on these 3 categories of inputs as the baseline candidates of our method.[3] learns the prototypical patterns of normal data with a memory-guided auto-encoder for anomaly detection.[5] reconstruct the optical flow with a multi-level memory-guided auto-encoder and predict the future frame with the reconstructed flow and previous frame employing a conditional variational auto-encoder.[7] detect anomalies by learning to solve the frame-permutated jigsaw puzzles in terms of spatial and temporal dimensions. For[5] and[7], we use the provided pre-trained networks as the teacher directly. However, the pre-trained networks of[3] are not provided in the official code. So we first trained the network without modifying the released code and use the reproduced network as the teacher. Though using the unmodified code from the official repository, we are not able to reproduce the results as reported, but the numbers are relatively close.
Pre-training of the teacher network. To combine with the proposed method, we present a simple but effective way to train the teacher network. We use an auto-encoder to predict the future frame while mimicking the feature outputted by a certain layer of a network pre-trained on a larger scale dataset at the bottleneck(empirically using block2 of ResNet50 which makes the teacher network to achieve the optimal results for anomaly detection). The combination of two tasks: future frame prediction and feature mimicking enables the teacher network to possess not only a fine-grained but a high-level, semantic description of the normal data.
Ablation Study. Fristly, we present the Frame-level AUC (in%) of some selected baseline before and after combining with our method on CUHK Avenue and ShanghaiTech in Table.1. The integration of our method with[3,5,7] improves the performance of each method, which shows the effectiveness and versatility of our method.

Table 1. Frame-level AUC (in%) of some selected baseline before and after combining with our method on CUHK Avenue and ShanghaiTech. The best result of each dataset are highlighted in bold.

	Method	Avenue	SHTech
Raw Frame	MNAD[3]	83.5	68.5
	MNAD+SCA	84.9	69.7
STC	PKG-Net[38]	93.8	80.2
	PKG-Net (Pre-training)+SCA	**94.3**	79.2
YOLOv3	Jigsaw[7]	92.2	84.3
	Jigsaw+SCA	92.3	**84.5**

What's more, we apply our method at the bottleneck of an auto-encoder as the bottleneck feature commonly has the strongest representation capability among the features. To prove this, we apply our method at different layers of the auto-encoder. We choose the PKG-Net as the baseline and conduct experiments on the CUHK Avenue dataset. As shown in Table.2, though performance improvement is achieved by integrating our method to the auto-encoder at different layers, solely applying our method at the bottleneck achieves the best result.
Comparative Experiments. We present the results of the comparison of our method with state-of-the-art(SOTA) in Table,3.The combination of our method and the pre-trained teacher network reaches 94.3% on CUHK Avenue which is 1.4% higher than

Table 2. Frame-level AUC (in %) on Avenue while integrating our method to an auto-encoder at different locations.

Encoder	Bottleneck	Decoder	AUC
✓			91.7
	✓		**94.3**
		✓	92.1
✓	✓		93.4
	✓	✓	93.2
✓	✓	✓	93.6

Table 3. Frame-level AUC (in%) of SOTA methods on CUHK Avenue and ShanghaiTech.

Method	Venue	Avenue	SHTech
FFP[1]	CVPR18	85.1	72.8
MemAE[2]	pkg19	83.3	71.2
MNAD[3]	CVPR20	83.5	68.5
VEC[4]	MM20	90.2	74.8
HF2VAD[5]	ICCV21	91.1	76.2
SSMTL[6]	CVPR21	91.5	82.4
SSPCAB[33]	CVPR22	92.9	83.6
Jigsaw[7]	ECCV22	92.2	84.3
PKG-Net[38]	MMAsia23	93.8	80.2
SSMTL++v2[37]	CVIU23	91.6	83.8
HSC[40]	CVPR23	93.7	83.4
SDMAE[34]	CVPR24	91.3	79.1
SSMTL++v2[37] + SSMCTB[36]	TPAMI24	91.6	83.6
Ours	Ours	**94.3**	**84.5**

the previous SOTA method. Our method also improves[7] by 0.2% and produces a new SOTA result of 84.5% on ShanghaiTech.

Image Anomaly Detection

Baseline. We choose one of the typical reconstruction-based models (i.e. DRAEM[15]) as the baseline for image anomaly detection. DRAEM utilizes a dual auto-encoder structure to learn a joint representation of anomalous from synthetic anomalies automatically generated on anomaly-free images. We combine the CA of our method with DRAEM for anomaly localization as the SA or SCA may over-magnify the anomalous region which affect the results of anomaly segmentation.

Ablation Study. We present the results on MVTech AD before and after adding our method in Table.4 . The results of the baseline incorporated with SSPCAB[33] are also illustrated for comparison. SSPCAB is a neural block composed of a masked convolutional layer and a channel attention module that can be incorporated into existing anomaly detection methods. Considering the detection results, the spatial-channel attention of our method improves the overall image-level AUROC by 1.2%. The improvement is noteworthy considering the baseline is already very good. The detection results are 0.3% higher than DRAEM+SSPCAB.

In terms of the localization result, the overall AUROC of DRAEM is improved from 97.3% to 98.4%. However, the AUROC metric shows the performance of anomaly localization with bias as the FPR is dominated by the a-prior very large number of normal pixels. Thus the pixel-level AP is a more robust and challenging metric for anomaly localization in which the classes are particularly imbalanced. Our method increases the overall AP of the baseline by 6.4%, from 68.4% to 74.8%. Notably, the AP result on the toothbrush category is improved by 25.0%, from 44.7% to 69.7% and the result on the carpet category is improved by 27.3%, from 53.5% to 80.8%.

Comparative Experiments. We present the results on MVTech AD compare with recent SOTA methods in Table.5. The detection results are 0.3% higher than DRAEM+SSPCAB and competitive even compared with recent SOTA methods. In terms of the localization result, The overall result is even 4.6% higher than the recent SOTA method for anomaly segmentation[16], whose overall result of localization AP is 70.2%.

Details on MVTec AD

The results of DRAEM we reported in the paper are directly copied from the original paper. However, the results are different when we use the pre-trained models from the official repository. For example, the results of localization AP are 63.8% and

Table 4. Comparison of Localization AUC/AP and detection AUC (in%) between baseline, after adding SSPCAB and adding our method. The best result of each category is hightlighted in bold.

	Class	Localization						Detection			
		DRAEM	+SSPCAB	+CA	DRAEM	+SSPCAB	+CA	DRAEM	+SSPCAB	+CA	+SCA
		AUC	AUC	AUC	AP	AP	AP	AUC	AUC	AUC	AUC
Texture	carpet	95.5	95.0	**99.3**	53.5	59.4	**80.8**	97.0	98.2	**99.7**	99.4
	grid	99.7	99.5	99.7	65.7	61.1	**65.9**	99.9	100.0	100.0	100.0
	leather	98.6	99.5	**99.8**	75.3	**76.0**	73.6	100.0	100.0	100.0	100.0
	tile	99.2	99.3	**99.7**	92.3	95.0	**97.0**	99.6	100.0	100.0	100.0
	wood	96.4	96.8	**98.3**	77.7	77.1	**84.0**	99.1	99.5	**100.0**	99.1
Object	bottle	99.1	98.8	**99.3**	86.5	87.9	**90.0**	**99.2**	98.4	96.7	97.8
	cable	94.7	96.0	**96.3**	52.4	57.2	**72.4**	91.8	96.9	94.3	**97.5**
	capsule	94.3	93.1	**95.8**	49.4	50.2	**52.5**	98.5	**99.3**	98.0	97.5
	hazelnut	99.7	**99.8**	99.7	**92.9**	92.6	90.1	100.0	100.0	100.0	99.9
	metal nut	**99.5**	98.9	99.2	96.3	**98.1**	94.1	98.7	100.0	99.2	100.0
	pill	97.6	97.5	**97.7**	48.5	**52.4**	42.8	98.9	**99.8**	93.4	99.5
	screw	97.6	99.8	99.8	58.2	**72.0**	71.2	93.9	97.9	98.7	**99.0**
	toothbrush	98.1	98.1	**99.2**	44.7	51.0	**69.7**	100.0	100.0	100.0	100.0
	transistor	90.9	87.0	**93.8**	50.7	48.0	**55.0**	93.1	92.9	91.1	**98.0**
	zipper	98.8	**99.0**	98.9	81.5	77.1	**83.4**	100.0	100.0	99.6	99.7
	average	97.3	97.2	**98.4**	68.4	69.9	**74.8**	98.0	98.9	98.0	**99.2**

Table 5. Comparison of Localization AUC/AP and detection AUC (in%) with SOTA methods on MVTec AD. The best result of each metric is highlighted in bold.

Method	Venue	AUC_{lo}	AP_{lo}	AUC_{de}
UStudents[12]	CVPR20	93.9	45.5	87.7
PaDim[13]	ICPR21	97.5	55.0	95.5
DRAEM[15]	ICCV21	97.3	68.4	98.0
Reverse Distill[18]	CVPR22	97.8	-	98.5
PatchCore-L[17]	CVPR22	98.2	-	**99.6**
SSPCAB[33]	CVPR22	97.2	69.9	98.9
DSR[16]	ECCV22	-	70.2	98.2
DMAD[41]	CVPR23	98.2	-	99.5
SimpleNet[39]	CVPR23	98.1	-	**99.6**
MGCFT[35]	CVPR24	**98.4**	-	98.2
DRAEM+Ours	Ours	**98.4**	**74.8**	99.2

72.0% for *carpet* and *zipper* while the reported results from the original paper are 53.5% and 81.5%. The overall result of the average localization AUC, AP, and detection AUC are 97.5%, 68.9%, and 98.0% which is even slightly higher than the reported results: 97.3%, 68.4%, and 98.0%.

In the anomaly localization task, we follow the recent works[17, 18] and smooth the results with a Gaussian of kernel width $\sigma = 4$ for a fair comparison. The results of localization AUC and AP are 97.8% and 74.8%, 98.4% and 74.8% before and after using the Gaussian filter, respectively. The DRAEM model contains two auto-encoders (reconstruction and discrimination), and the results of one can influence the other. We only integrate our method into one of the auto-encoder which achieves the optimal results.

Visualization Results

In anomaly detection tasks, accurately identifying and locating anomalous regions is crucial for enhancing system efficiency and reliability. Visualization aims to assess and compare the improvements our proposed strategy brings to the detection accuracy and robustness of existing models. We conducted experiments on various datasets and detection methods, including both image and video anomaly detection. In these experiments, we analyzed the performance of each method across different objects and scenes, focusing on anomaly region localization accuracy, false positive rates, boundary detection, and overall model stability.

In the image anomaly detection task, we compared the performance of the DRAEM method with the improved DRAEM+Ours

method by detecting anomalous images of four different objects (zipper, toothbrush, hazelnut, and cable) from the MVTec AD dataset. The experimental results shown in Figure 3 indicate that the DRAEM method is able to detect anomalous regions in images but struggles with capturing boundaries and details of the anomalies, particularly at the edges, where false positives and diffusion often occur. In contrast, the DRAEM+Ours method significantly improves the localization accuracy and detail recognition by optimizing the DRAEM model. The DRAEM+Ours method more clearly identifies the boundaries of the anomalous regions and effectively reduces both false positives and missed detections. Particularly in the toothbrush and cable detection tasks, it precisely captures the anomalous parts, avoiding the common false highlight issues seen with DRAEM. Additionally, the DRAEM+Ours method demonstrates excellent robustness, consistently identifying anomalies across different images while reducing false positives and maintaining high detection accuracy. Overall, the DRAEM+Ours method outperforms DRAEM, particularly in terms of precision, boundary detection, and robustness, proving its effectiveness in tasks such as surface defect detection and object damage recognition.

In the video anomaly detection task, we visualized the comparative results of two methods. The PKG-Net+Ours method exhibited higher detection accuracy and lower false positive rates than the baseline PKG-Net+Ours model on the 17th video segment of the Avenue dataset, as shown in Figure 4. In the experiment, the anomaly score curve of the PKG-Net+Ours method (in blue) aligns more closely with the true anomalous regions compared to the baseline method (in orange), reducing false positives and improving temporal consistency. This method raised the AUROC to 91.4% (compared to 83.5% for the baseline), demonstrating its superior performance on complex video datasets. The example frames below further highlight the detected anomalous targets, with the anomalous regions marked by red borders, validating the effectiveness of this method in video anomaly detection. Similarly, the Jigsaw+Ours method showed superior anomaly detection performance on the 01_0053 video from the ShanghaiTech dataset, improving detection accuracy and reducing false positives compared to the baseline Jigsaw model, as shown in Figure 5. The experimental results indicate that the anomaly score of the Jigsaw+Ours method more accurately reflects the true anomalous regions, reducing false detections and increasing the AUROC to 95.8% (compared to 90.0% for the baseline). Example frames below demonstrate the detected anomalies, with red borders highlighting the anomalous regions, further confirming the method's effectiveness in complex scenes and enhancing the accuracy and stability of video anomaly detection.

In conclusion, the DRAEM+Ours, PKG-Net+Ours, and Jigsaw+Ours methods all outperform their baseline models in their respective anomaly detection tasks, not only improving detection accuracy but also reducing false positives and enhancing model robustness. This thoroughly validates the broad applicability and effectiveness of the strategy proposed in this paper across different types of anomaly detection tasks.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

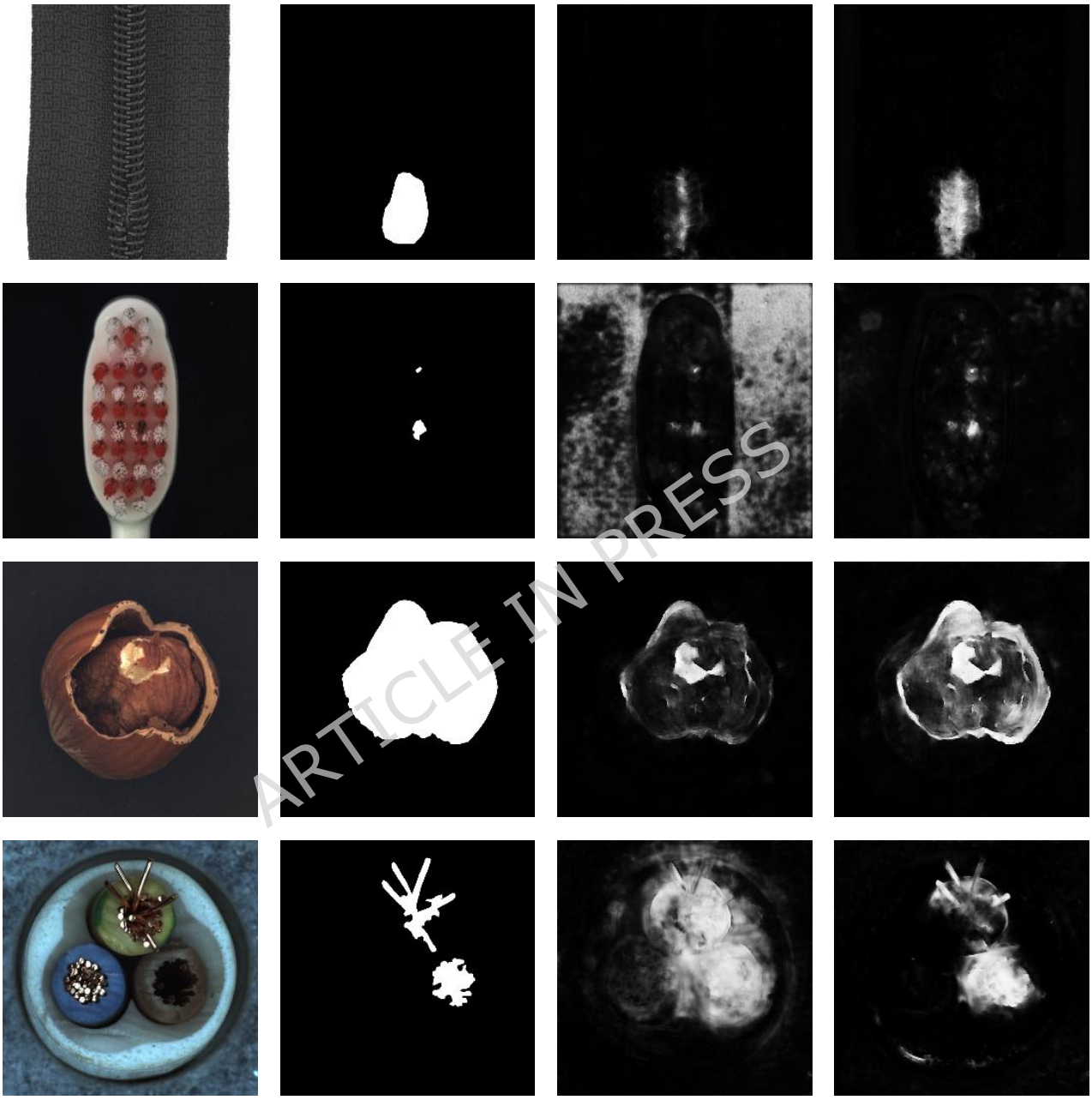

Figure 3. Visualization on (a) anomalous images, (b) ground truth of anomaly localization, (c) anomaly maps generated by DRAEM[15], and (d) DRAEM+Ours method. From top to bottom are: *zipper*, *toothbrush*, *hazelnut* and *cable*.

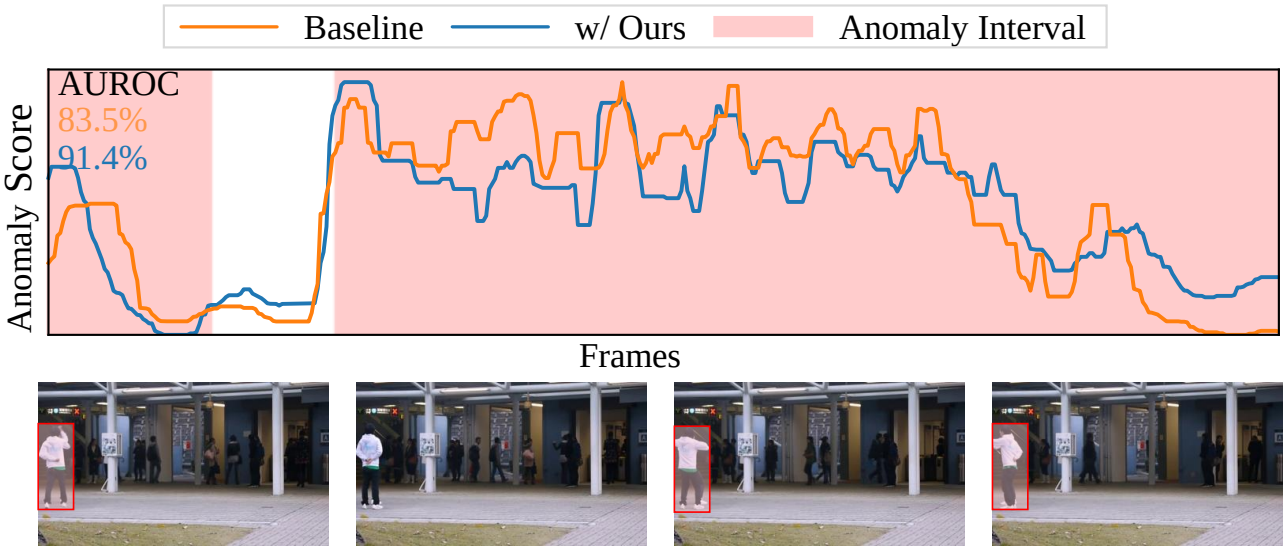

Figure 4. Visualization of video 17 from Avenue. The orange and blue curves denote the frame-level anomaly scores for the PKG-Net and PKG-Net+Ours method. Best viewed in color.

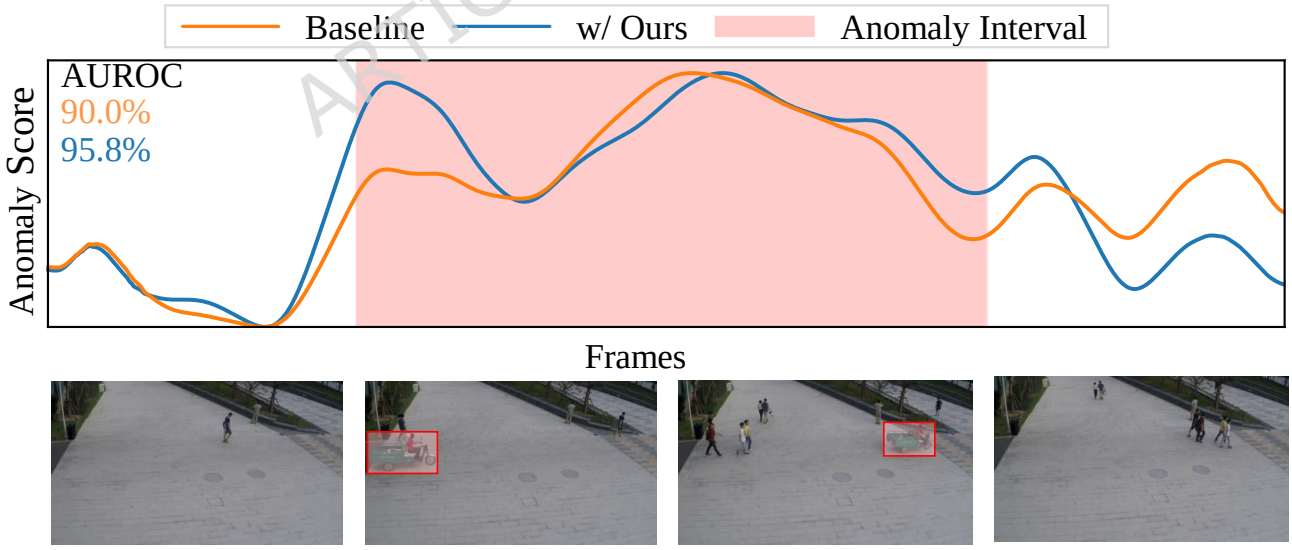

Figure 5. Visualization of video 01_0053 from ShanghaiTech. The orange and blue curves denote the frame-level anomaly scores for the Jigsaw[7] and Jigsaw+Ours method. Best viewed in color.

References

1. Liu, W., Luo, W., Lian, D. & Gao, S. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545 (2018).
2. Gong, D. *et al.* Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1705–1714 (2019).
3. Park, H., Noh, J. & Ham, B. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14372–14381 (2020).
4. Yu, G. *et al.* Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, 583–591 (2020).
5. Liu, Z., Nie, Y., Long, C., Zhang, Q. & Li, G. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13588–13597 (2021).
6. Georgescu, M.-I. *et al.* Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12742–12752 (2021).
7. Wang, G. *et al.* Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, 494–511 (Springer, 2022).
8. Li, C.-L., Sohn, K., Yoon, J. & Pfister, T. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674 (2021).
9. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H. & Rabiee, H. R. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14902–14912 (2021).
10. Aslam, N., Rai, P. K. & Kolekar, M. H. A3n: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. *J. Vis. Commun. Image Represent.* **87**, 103598 (2022).
11. Aslam, N. & Kolekar, M. H. Unsupervised anomalous event detection in videos using spatio-temporal inter-fused autoencoder. *Multimed. Tools Appl.* **81**, 42457–42482 (2022).
12. Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192 (2020).
13. Defard, T., Setkov, A., Loesch, A. & Audigier, R. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, 475–489 (Springer, 2021).
14. Wang, S., Wu, L., Cui, L. & Shen, Y. Glancing at the patch: Anomaly localization with global and local feature comparison. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 254–263 (2021).
15. Zavrtanik, V., Kristan, M. & Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339 (2021).
16. Zavrtanik, V., Kristan, M. & Skočaj, D. Dsr–a dual subspace re-projection network for surface anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, 539–554 (Springer, 2022).
17. Roth, K. *et al.* Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328 (2022).
18. Deng, H. & Li, X. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746 (2022).
19. Morais, R. *et al.* Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11996–12004 (2019).
20. Aslam, N. & Kolekar, M. H. A-vae: Attention based variational autoencoder for traffic video anomaly detection. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–7 (IEEE, 2023).
21. Aslam, N. & Kolekar, M. H. Demaae: deep multiplicative attention-based autoencoder for identification of peculiarities in video sequences. *The Vis. Comput.* **40**, 1729–1743 (2024).

22. Aslam, N. & Kolekar, M. H. Transganomaly: transformer based generative adversarial network for video anomaly detection. *J. Vis. Commun. Image Represent.* **100**, 104108 (2024).

23. Zhang, X., Xu, M. & Zhou, X. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16699–16708 (2024).

24. Costanzino, A., Ramirez, P. Z., Lisanti, G. & Di Stefano, L. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17234–17243 (2024).

25. Freytsis, M., Perelstein, M. & San, Y. C. Anomaly detection in the presence of irrelevant features. *J. High Energy Phys.* **2024**, 1–22 (2024).

26. Miao, J., Tao, H., Xie, H., Sun, J. & Cao, J. Reconstruction-based anomaly detection for multivariate time series using contrastive generative adversarial networks. *Inf. Process. & Manag.* **61**, 103569 (2024).

27. Lai, C.-Y. A., Sun, F.-K., Gao, Z., Lang, J. H. & Boning, D. Nominality score conditioned time series anomaly detection by point/sequential reconstruction. *Adv. Neural Inf. Process. Syst.* **36** (2024).

28. Yu, J. & Do, H. Proximity-based density description with regularized reconstruction algorithm for anomaly detection. *Inf. Sci.* **654**, 119816 (2024).

29. Kwon, G., Prabhushankar, M., Temel, D. & AlRegib, G. Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision*, 206–226 (Springer, 2020).

30. Ruff, L. *et al.* Deep one-class classification. In *International conference on machine learning*, 4393–4402 (PMLR, 2018).

31. Yi, J. & Yoon, S. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision* (2020).

32. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

33. Ristea, N.-C. *et al.* Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13576–13586 (2022).

34. Ristea, N. C. *et al.* Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 15984–15995 (2024).

35. Artola, A., Kolodziej, Y., Morel, J. M. & Ehret, T. Model-guided contrastive fine-tuning for industrial anomaly detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3981–3991 (2024).

36. Madan, N. *et al.* Self-supervised masked convolutional transformer block for anomaly detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):525—-542 (2024).

37. Barbalau, A. *et al.* Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. In *Computer Vision and Image Understanding* (2023).

38. Deng, Z., Chen, D. & Deng, S. Prior knowledge guided network for video anomaly detection. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 1–7 (2023).

39. Liu, Z., Zhou, Y., Xu, Y. & Wang, Z. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 20402–20411 (2023).

40. Sun, S. & Gong, X. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 22846–22856 (2023).

41. Liu, W., Chang, H., Ma, B., Shan, S. & Chen, X. Diversity-measurable anomaly detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 12147–12156 (2023).

42. Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600 (2019).

43. Lu, C., Shi, J. & Jia, J. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, 2720–2727 (2013).

44. Luo, W., Liu, W. & Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, 341–349 (2017).

45. Noghre, G. A., Pazho, A. D. & Tabkhi, H. An exploratory study on human-centric video anomaly detection through variational autoencoders and trajectory prediction. *Proc. IEEE/CVF Winter Conf. on Appl. Comput. Vis.* 995–1004 (2024).

46. Ravanbakhsh, M. *et al.* Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE international conference on image processing (ICIP)*, 1577–1581 (IEEE, 2017).

47. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).

48. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).

49. Wang, Q. *et al.* Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542 (2020).

50. Li, X., Wang, W., Hu, X. & Yang, J. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 510–519 (2019).