레터논문 (Letter Paper)
방송공학회논문지 제31권 제3호, 2026년 5월 (JBE Vol.31, No.3, May 2026)
https://doi.org/10.5909/JBE.2026.31.3.512
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

경계 영역의 반사실적 셔플링을 이용한 효율적 이상탐지

누만 알리 칸[a)], 무수누리 요겐드라 라오[a)], 권 오 설[a)‡]

Efficient Anomaly Detection Using Counterfactual Shuffling on Boundary Regions

Nouman Ali Khan[a)], Yogendra Rao Musunuri[a)], and Oh-Seol Kwon[a)‡]

요 약

산업용 이상 탐지는 정상 이미지만을 사용하여 학습을 수행한다. 따라서 이상 샘플이 없는 상태에서 정상성 주위에 긴밀한 경계를 형성하는 것이 필수적이다. 기존의 합성 기반 모델은 가상의 이상치를 주입하여 이 경계를 정교화하지만, 이는 모델의 복잡도를 높이고 실재하지 않는 인위적 결함을 생성할 위험이 있다. 본 논문은 이러한 제약을 해결하기 위해 특징 공간 내에서 이상치를 직접 생성하는 프레임워크를 제안한다. 적응적 범위는 경계의 조밀도를 예측하며, 기울기 연계 자극 기법은 정상 특징을 학습된 경계 밖으로 밀어내어 암시적 이상치를 생성한다. 또한, 유사도 제약 기반 반사실적 셔플링 모듈은 공간적으로는 불일치하나 국소적으로는 타당한 특징을 추가로 생성한다. 이 두 방식에 의해 생성된 특징들은 모두 학습된 경계 너머로 투영되어 판별기 학습에 활용된다. MVTecAD, WFDD 및 MPDD 데이터셋에 대한 실험 결과, 제안 기법은 정상 데이터만을 사용하고도 이미지 수준의 탐지 및 픽셀 수준의 위치 추정 성능을 유의미하게 향상시켰다.

Abstract

Industrial anomaly detection is trained using only normal images, requiring a tight boundary around normality without anomaly samples. Synthesis-based models can refine this boundary by injecting pseudo-anomalies, but this increases complexity and may introduce unrealistic artifacts. To address these limitations, we propose a synthesis-free framework that produces anomalies directly in feature space. An Adaptive Radius Field predicts boundary tightness, while a gradient-linked nudge pushes normal features outside the learned boundary to create implicit anomalies. A similarity-constrained counterfactual shuffling module further produces spatially inconsistent yet locally plausible features. Both are projected beyond the learned boundary and used to train the discriminator. Experiment on MvtecAD, WFDD and MPDD datasets show improved image-level detection and pixel-level localization only on normal data.

Keyword : Anomaly Detection, Implicit boundary, Feature shuffling, Adaptive radius

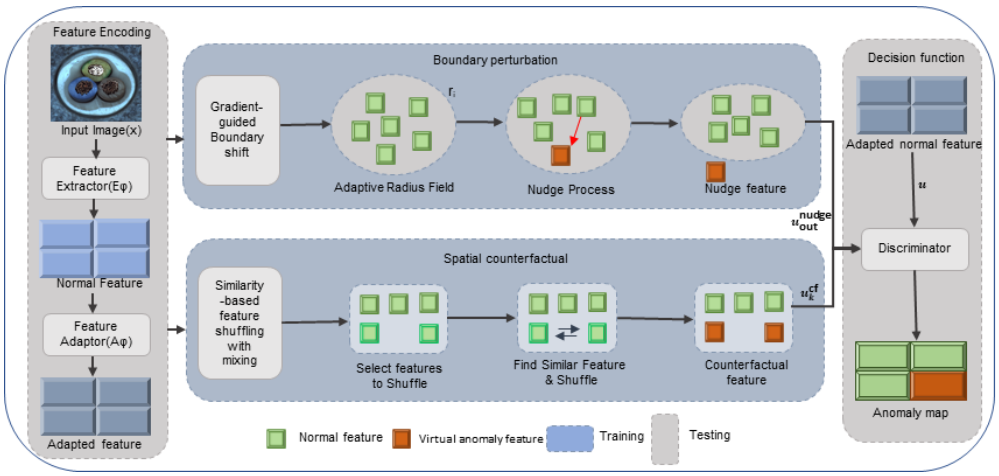

그림 1. 제안한 경계 영역의 반사실적 셔플링 방법의 프레임워크
Fig. 1. Framework of the Proposed Method Based on Counterfactual Shuffling on Boundary Regions

Ⅰ. Introduction

Industrial anomaly detection aims to detect rare and subtle defects using only normal training data [1]. Since accurate normal-boundary modeling is more important than standard classification, patch-based feature modeling has become effective for capturing local texture and structural deviations [2]. However, reconstruction and distribution-based methods may miss subtle anomalies or depend strongly on embedding calibration [3], while synthesis-based methods can improve boundary sharpness but add design complexity and may bias detection toward artificial artifacts [4]. To overcome these limitations, we propose a feature-space anomaly generation method without image synthesis. It includes two mechanisms: an Adaptive Radius Field (ARF) with gradient-linked nudging to push features beyond location-dependent normal boundaries, and similarity-constrained counterfactual feature shuffling to disrupt spatial consistency while preserving local feature statistics.

a) 국립창원대학교 지능로봇융합공학(Dept. of Intelligent Robotics and Convergence Engineering, Changwon National University)
‡ Corresponding Author : 권오설(Oh-Seol Kwon)
E-mail: osk1@changwon.ac.kr
Tel: +82-55-213-3669
ORCID: https://orcid.org/0000-0002-1077-9615
※This research was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2025-00555758) and in part by Korea Electrotechnology Research Institute (KERI) grant funded by the Ministry of Science and ICT (MSIT) (No. 26A01051-01).
· Manuscript April 2, 2026; Revised May 6, 2026; Accepted May 6, 2026.

Ⅱ. Proposed Method

The proposed framework is summarized in Fig. 1, which highlights both the training virtual anomaly generation and the testing scoring path.

We follow [5] and extract intermediate feature maps from a frozen backbone and map them to a task embedding space using a adaptor that inserts a simple adaptor to reduce domain bias. To allow spatially varying boundary tightness, we estimate a positive radius r_{ij} at each patch location (i,j) using an Adaptive Radius Field head. Given the spatial feature F, the head $\phi(\cdot)$ predicts a scalar radius from each local feature. The radius is constrained to be positive by applying the Softplus function and adding a small constant epsilon (ϵ=1e-3) as mentioned in Eq. (1).

$$r_{ij} = softplus\left(\phi(F)_{ij}\right) + \epsilon \tag{1}$$

We generate two complementary types of virtual anoma-

lies using only normal data. Both are applied during training. The first one is a single gradient-linked step used to move a normal feature outward in feature space. The normalized step direction based on the detector loss L_{det} as in Eq. (2).

$$g = \frac{\partial L_{det}}{\partial u}, \quad \tilde{u} = u + \alpha \frac{g}{\| g \|_2} \tag{2}$$

Where g is the gradient of the detection loss with respect to the feature u, and $\tilde{u}$ is the nudged feature obtained by moving u a small step α. The direction from the center is rescaled as shown in Eq. (3).

$$u_{out}^{nudge} = c + (\tilde{u} - c)\frac{t}{\| \tilde{u} - c \|_2} \tag{3}$$

Where u_{out}^{nudge} is the boundary-projected feature obtained by taking the nudged feature, measuring its direction relative to the center *c*, and scaling it to the target distance *t*, so that the resulting feature lies exactly on or just beyond the learned normal boundary. The second one constructs a counterfactual feature map by shuffling patch embeddings across spatial positions within the same image, as defined in Eq. (4).

$$u_k^{cf} = (1 - m_k)u_k + m_k((1 - \lambda)u_k + \lambda u_{j(k)}) \tag{4}$$

The counterfactual feature u_k^{cf} is constructed by shuffling the original feature u_k with another feature $u_{j(k)}$. The binary mask $m_k \in \{0,1\}$ determines whether shuffling is applied at location k, and $\lambda \in [0,1]$ controls the interpolation strength. We train discriminator D to classify normal features as 0 and both virtual anomaly types as 1, with weights w_{nudge} and w_{cf} their contributions during boundary learning as given in the Eq. (5).

$$L_{det} = BCE(D(u),0) + w_{nudge} BCE\left(D\left(u_{out}^{nudge}\right),1\right) + w_{cf} BCE\left(D\left(u_{out}^{cf}\right),1\right) \tag{5}$$

Where BCE denotes the binary cross entropy loss. u_{out}^{nudge} denotes the gradient-nudged feature and u_{out}^{cf} denotes the counterfactual shuffled feature. Boundary-contrastive and compactness loss as in Eq. (6) and Eq. (7).

$$L_{cmp} = E_k\left[\max\left(0,\ d(u_k) - \lambda_c \hat{r_k}\right)\right] \tag{6}$$

$$L_{bcl} = E_k\left[\max\left(0,\ m\hat{r_k} + d(u_k) - d\left(u_{out,k}^{nudge}\right)\right)\right] \tag{7}$$

Where $\hat{r_k}$ is the predicted radius, λ_c is a scaling factor controlling boundary tightness, m is a margin factor, and E_k denotes the expectation. To prevent degenerate radii, we regularize the ARF as shown in Eq. (8).

$$L_{arf} = \beta_{mean}|\bar{r} - r_0| + \beta_{tv}(\|r_{i+1,j} - r_{ij}\|_1 + \|r_{i,j+1} - r_{ij}\|_1) \tag{8}$$

Where $\bar{r}$ predicted radius, r_0 is a target radius, β_{mean} weights of the constraint, and β_{tv} weights of a total variation regularization. The overall training objective is defined as the sum of the detection, compactness, boundary-contrastive, and ARF regularization losses, given in Eq. (9).

$$L = L_{det} + L_{cmp} + L_{bcl} + \gamma L_{arf} \tag{9}$$

Where L is the total training loss composed of the detection loss L_{det}, L_{cmp} is a compactness loss, L_{bcl} is a boundary-contrastive loss that separates normal and virtual anomalies, and L_{arf} is an ARF regularization term weighted by γ=0.1. which jointly enforce normal feature compactness and separation from both nudged and counterfactual negatives. At test time, only the backbone, adaptor, and discriminator are needed as shown in Fig. 1.

Ⅲ. Experimental Results

We evaluate the proposed method on MVTecAD,

표 1. 다양한 데이터셋에 대한 기존 방법과 제안한 방법의 성능 비교
Table 1. Comparison with Existing Methods on MVTecAD, WFDD, and MPDD Datasets(Image AUROC/Pixel AUROC)

Datasets	DSR	BGAD	PatchCore	PatchGuard	Proposed Method
MVTecAD	98.2/95.8	97.9/98.2	99.1/98.1	88.2/92.7	99.6/98.8
WFDD	95.1/87.9	97.1/98.5	96.3/98.1	84.2/94.6	99.7/99.2
MPDD	81.0/76.2	91.8/98.1	93.5/98.9	85.4/93.8	98.0/99.4

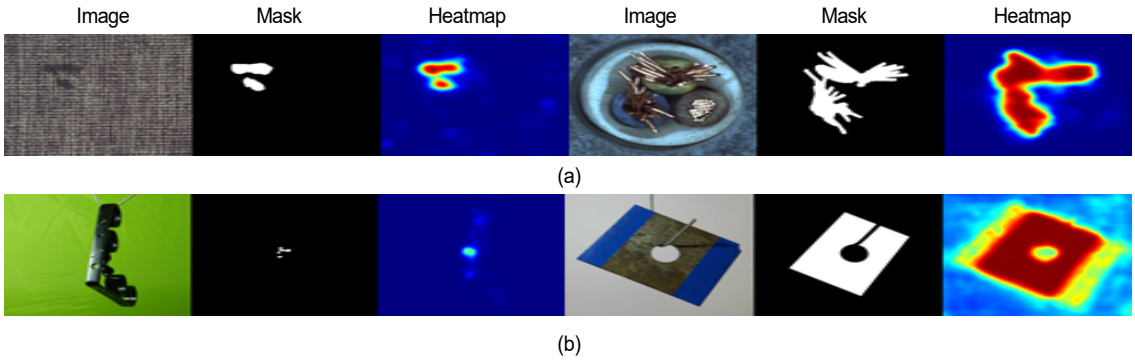

그림 2. 정성적 시각화 (a) MVTecAD dataset (b) MPDD dataset
Fig. 2. Qualitative visualizations on (a) MVTecAD dataset (b) MPDD dataset

WFDD, and MPDD using standard normal-only training images. We report image-level and pixel-level AUROC, while Fig. 2 shows qualitative anomaly maps with the input image, ground truth mask, and predicted heatmap. MVTecAD contains 15 classes, WFDD includes 4 woven-fabric classes, and MPDD focuses on fine-grained metal defects. We use a frozen WideResNet-50 backbone with adaptor, a Softplus-based Adaptive Radius Field head, and a discriminator. Training is performed for 640 epochs with batch size 8 and image size 288 on a single RTX A6000 GPU. During training, two types of virtual anomalies are generated, and the discriminator learns to classify normal features as 0 and virtual anomalies as 1. Table 1 shows that our method achieves the best overall image-level and pixel-level AUROC on all three datasets, demonstrating improved detection and localization.

Ⅳ. Conclusion

We propose a normal-only anomaly detection framework that generates virtual anomalies directly in feature space. It combines two mechanisms an ARF-guided gradient nudge that pushes normal embeddings beyond local boundaries, and counterfactual shuffling that swaps semantically similar but spatially distant features to create hard negatives. Across MVTecAD, WFDD, and MPDD, the method consistently improves performance without synthetic image generation and with efficient inference.

References

[1] I. Song and J. Lee, “Rule-based Zero-shot Video Anomaly Detection Using Object Detection and Semantic Segmentation,” *Journal of Broadcast Engineering*, Vol.29, No.6, pp.1067-1074, November 2024. doi: https://doi.org/10.5909/JBE.2024.29.6.1067

[2] T. Defard, A. Setkov, A. Loesch, and R. Audigier, “Padim: a patch distribution modeling framework for anomaly detection and localization,” *Proceeding of International Conference on Pattern Recognition*, Milan, Italy, pp. 475-489, 2021.

[3] V. Zavrtanik, M. Kristan, and D. Skočaj, “Draem-a discriminatively trained reconstruction embedding for surface anomaly detection,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, Canada, pp. 8330-8339, 2021.

[4] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang, “Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* British Columbia, Canada, pp. 24490-24499, 2023.

[5] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, “Simplenet: A simple network for image anomaly detection and localization,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* British Columbia, Canada, pp. 20402-20411, 2023.