PASS: 제로샷 이상 탐지를 위한 구조적 프라이어 기반의 의미 편향 보정

김서현, 박현희*
명지대학교 인공지능융합학과
seohy2on, hhpark* @ mju.ac.kr

PASS: Structural Prior-Guided Calibration of Semantic Bias for Zero-Shot Anomaly Detection

Seohyeon Kim, Hyunhee Park*
Dept. of Artificial Intelligence Convergence, Myongji University

요 약

본 논문은 Vision Transformer(ViT) 기반 제로샷 이상 탐지(Zero-Shot Anomaly Detectin, ZSAD)에서 발생하는 의미적 편향과 구조적 정보 소실 문제를 해결하기 위해 PASS(Prior-guided Assessment of Semantic-Structural consistency) 프레임워크를 제안한다. 모델 내부의 어텐션 맵과 패치 임베딩이 객체의 구조적 및 관계적 일관성을 보존한다는 가설에 기반하여, 추가적인 참조 이미지 없이 단일 이미지 내부의 규칙성을 통해 정상성을 정의하는 자기 참조(Self-Reference) 메커니즘을 도입한다. 실험 결과, 제안하는 PASS 프레임워크는 다양한 제조 도메인에서 유효성을 보였으며, 특히 구조적 정보가 중요한 반도체 및 VisA 데이터셋에서 기존 방법론 대비 우수한 탐지 성능을 입증하였다.

Ⅰ. 서 론

산업 제조 현장의 결함 데이터 부족 문제를 해결하기 위해 대규모 Vision-Language Model(VLM)을 활용한 제로샷 이상 탐지(Zero-Shot Anomaly Detection, ZSAD)가 주목받고 있다. CLIP[1]과 같은 VLM은 텍스트와 이미지의 의미적 정렬을 통해 뛰어난 일반화 성능을 보이지만, 이미지를 고차원으로 추상화하는 과정에서 기하학적 정보가 소실되는 '의미적 편향' 문제를 내재하고 있다. WinCLIP[2]이나 CoOp[3]과 같은 기존 연구들은 텍스트 프롬프트 튜닝에 집중하거나 객체 수준의 인식을 강화했다. 그러나 미세한 스크래치나 칩핑(Chipping)과 같이 언어적으로 정의하기 어려운 국소적 구조 붕괴를 포착하는 데에는 한계를 보인다[4].

본 논문은 이러한 의미적 편향을 극복하기 위해, 외부 데이터 없이 이미지 내부의 규칙성을 활용하는 자기 참조(Self-Reference) 기반의 PASS(Prior-guided Assessment of Semantic-Structural consistency) 프레임워크를 제안한다. 본 방법론은 1) 물리적 구조(엣지, 대칭성, 균일성) 붕괴를 정량화하는 SRD와, 2) 패치 간의 관계적 일관성을 그래프로 모델링하는 PRS를 통해 구조적 프라이어를 형성한다. 이를 통해 의미적 임베딩을 보정(Calibration)한다. 이는 의미적 맥락과 기하학적 결함 정보를 상호 보완적으로 통합하여 ZSAD의 정밀도를 개선한다.

Ⅱ. 본 론

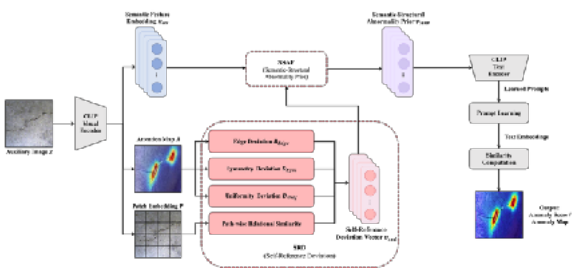

그림 1. PASS 프레임워크

PASS 프레임워크는 ViT 기반 인코더에서 추출된 공간적 어텐션 맵과 패치 임베딩을 분석하여 구조적 이상 징후를 포착하고, 이를 의미 공간에 주입하는 구조로 설계되었다.

2.1 SRD (Self-Reference Deviation)

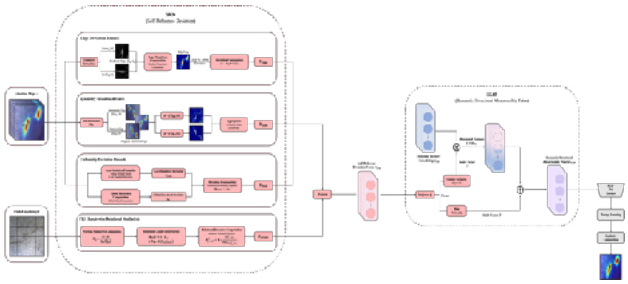

그림 2. Proposed Method

SRD는 단일 이미지 내에서 정상 데이터가 갖는 통계적 규칙성(연속성, 대칭성, 균일성)이 깨지는 지점을 정량화한다.

Edge Structural Deviation 어텐션 맵 내에서 급격한 관계 변화를 포착한다. Sobel 연산자로 추출한 그래디언트 G에 대해, 국소 평균 풀링(AvgPool)을 적용한 평활화 맵과의 잔차를 계산한다. 이를 통해 배경 텍스처 대비 돌출된 엣지 불연속성을 강조한다.

$$D_{Edge} = \left| G - AvgPool(G) \right| \quad (1)$$

Symmetry Structural Deviation 객체 중심 정렬 가정 하에 원본 어텐션 맵과 이를 수평·수직 반전시킨 맵 간의 차이를 통해 비대칭성을 측정한다.

$$D_{Sym} = \frac{1}{2}\left(\left| A - flip_h(A) \right| + \left| A - flip_v(A) \right| \right) \quad (2)$$

구조적 결함이 존재할 경우 완벽한 대칭이 붕괴되므로 해당 위치의 편차 값이 증가한다.

Uniformity Structural Deviation 텍스처의 통계적 분포가 전역적인 경향성에 벗어나는 정도를 측정한다.

$$D_{Unif} = \left| \sigma_{local} - \mu_{\sigma} \right| \quad (3)$$

전역 평균 표준편차와 국소 표준편차 간의 차이를 계산하여, 전체적인 텍스처 분포 대비 이질적인 분산을 갖는 영역을 탐지한다.

2.2 PRS (Patch-wise Relational Similarity)

물리적 특징뿐만 아니라 잠재 공간 내의 관계적 일관성을 평가하기 위해 PRS를 도입한다. 패치 임베딩 간의 양의 상관관계를 보존하는 자기 유사도 행렬을 구축하고, 각 패치의 연결 강도를 차수 중심성으로 측정한다. 전역적 패턴에서 고립된 이질적 패치를 식별하기 위한 관계적 편차 D_{graph}는 다음과 같다.

$$D_{graph}^{(i)}=\frac{\sum_{j=1}^{N} S_{i,j}}{max_k \sum_{l=1}^{N} S_{k,l}} \quad (4)$$

이 값은 주변 문맥과 의미적으로 단절된 결함 패치에서 높은 값을 가지며, 앞서 도출된 SRD 편차들과 결합되어 통합 자기 참조 편차 벡터를 구성한다.

$$v_{srd}=Concat\left[D_{Edge}, D_{Sym}, D_{Unif}, D_{graph} \right] \quad (5)$$

2.3 SSAP (Semantic-Structural Abnormality Prior)

SSAP 모듈은 구조적 편차 벡터 v_{srd}를 활용하여 CLIP의 의미적 임베딩을 보정(Calibration)한다. 이는 VLM이 놓친 기하학적 결함 정보를 의미 공간에 주입하는 과정이다. v_{srd}는 MLP 프로젝터를 통해 투영되어, 채널별 중요도를 조절하는 Scale Factor γ와 잠재 공간 내 위치를 이동시키는 Shift Factor β를 생성한다. 최종적으로 원본 의미 임베딩 v_{cls}는 식 (4)와 같이 특징 변조(Feature Modulation)되어, 의미적 맥락과 구조적 불규칙성 정보를 동시에 포함하게 된다.

$$v_{SSAP}=\gamma \odot v_{cls}+\beta \quad (6)$$

2.3 실험 및 결과

표 1. Image-level Performance (AUROC, AP)

Dataset	Baseline	PASS
AITEX	(74.1, 55.5)	(71.0, 61.1)
VisA	(84.6, 86.8)	(90.7, 91.2)
Semiconductor	(70.6, 82.9)	(80.4, 98.3)

본 논문은 제안 방법론의 유효성 검증을 위해, 미세 결함과 데이터 불균형이 특징인 반도체(Semiconductor) 데이터셋을 포함하여 총 3가지 제조 도메인 데이터셋을 활용하였다. 실험은 목표 도메인에 대한 추가 학습이 없는 제로샷 설정에서 수행되었다. 정량적 성능 평가는 이상 탐지 분야의 표준 지표인 Image-level AUROC와 Average Precision(AP) 척도를 기준으로 측정하였다.

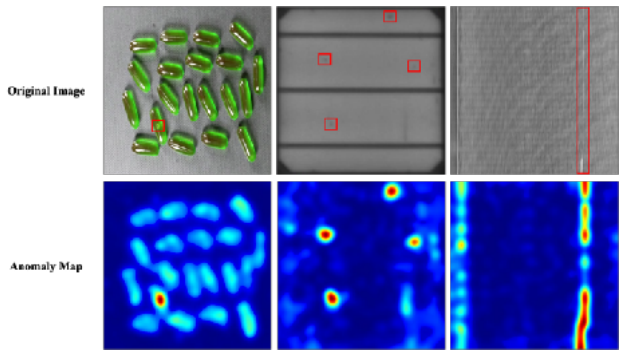

그림 3. Anomaly Map 시각화 예시

실험 결과, 표 1과 같이 PASS는 구조적 정보가 뚜렷한 VisA와 반도체 데이터셋에서 우수한 성능을 보여, SRD 모듈의 미세 결함 탐지 능력을 입증하였다. 반면, 텍스처가 복잡한 AITEX에서는 AUROC가 하락했으나 AP는 상승하여 실질적 정밀도는 개선되었다. 이는 PASS가 물리적 형상에 민감하게 설계되어, 반도체 등 구조적 정합성이 중요한 도메인에서 더욱 강건함을 시사한다. 정성적 분석에서도 배경 노이즈 없이 결함 부위만을 정밀하게 국소화하는 결과를 보여, 의미-구조 정보 보정의 유효성을 확인하였다.

Ⅲ. 결 론

본 논문은 ZSAD의 의미적 편향을 해결하기 위해, 이미지 내부의 자기 참조 정보를 활용하는 PASS 프레임워크를 제안하였다. 제안된 SRD와 PRS는 외부 데이터 학습 없이도 구조적 결함에 대한 민감도를 높였으며, SSAP를 통해 이를 의미 공간과 통합하였다. 실험 결과, 다양한 산업 도메인에서 우수한 일반화 성능을 확인하였다. 향후 연구에서는 VLM의 임베딩 공간 내에서 객체 고유의 기하학적 정합성을 능동적으로 학습하는 구조적 프롬프트 튜닝 방법론으로 발전시켜, 도메인 적응성과 탐지 정밀도를 동시에 강화할 계획이다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025년도 교육부 및 경기도의 재원으로 경기RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-09-A15)

참 고 문 헌

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 139, pp.8748-8763, Jul. 2021.

[2] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation," in Proc. IEEE/CVF Conf, Comput. Vis. Pattern Recognit. (CVPR), pp. 19606-19616, Jun. 2023.

[3] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," Int. J. Comput. Vis. (IJCV), vol. 130, no. 9, pp.2335-2358, Sep. 2022.

[4] S. Kim and H. Park, "Wafer Defect Analysis and NL Report Generation Based on Integrated Severity Estimation and Causal Analysis for On-Site Reliability." In Proc. Korean Institute of Communications and Information Sciences (KICS) Winter Conf., pp.1-4, Dec. 2023.

[5] J. Zhu, Y.-S. Ong, C. Shen, and G. Pang, "Fine-grained Abnormality Prompt Learning for Zero-shot Anomaly Detection," arXiv preprint arXiv:2410.10289, Oct. 2024.