2025 하계 세미나

Text Encoder Strategies for Vision-Language Anomaly Detection

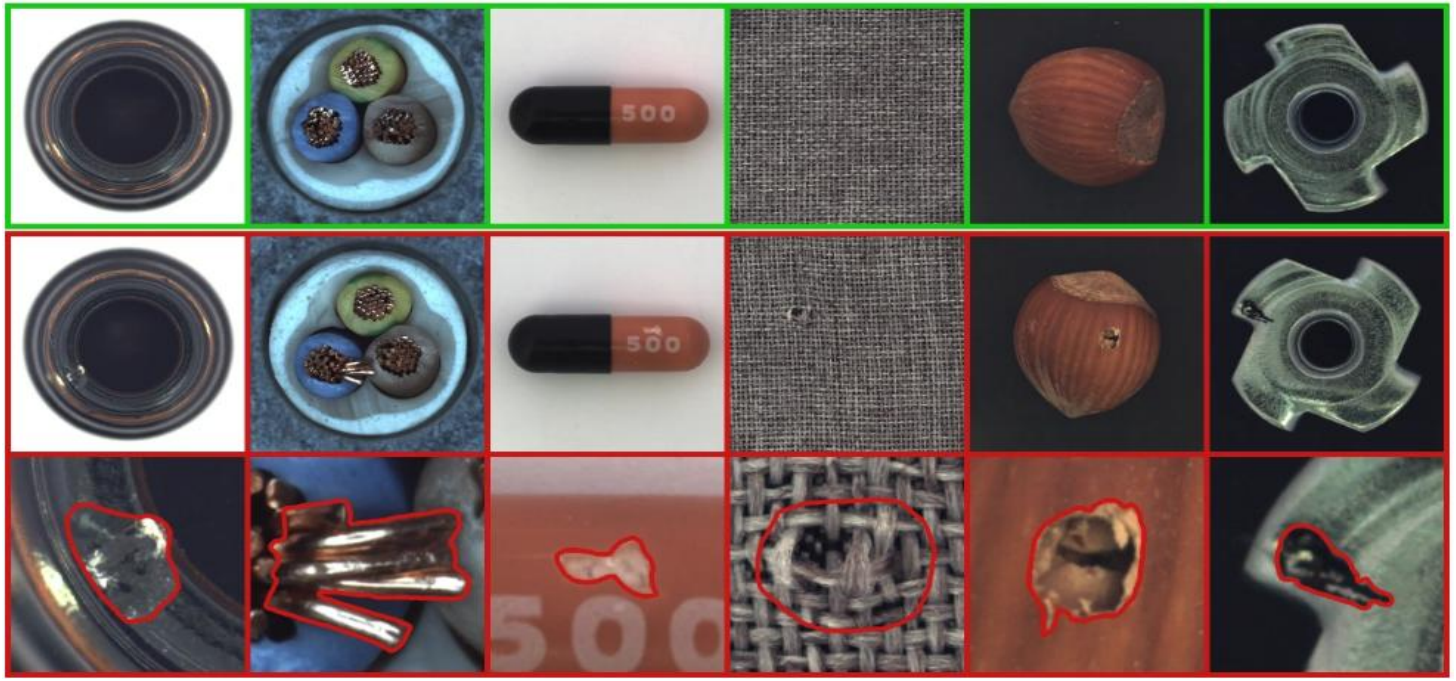
Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
이혜빈

Contents

Background

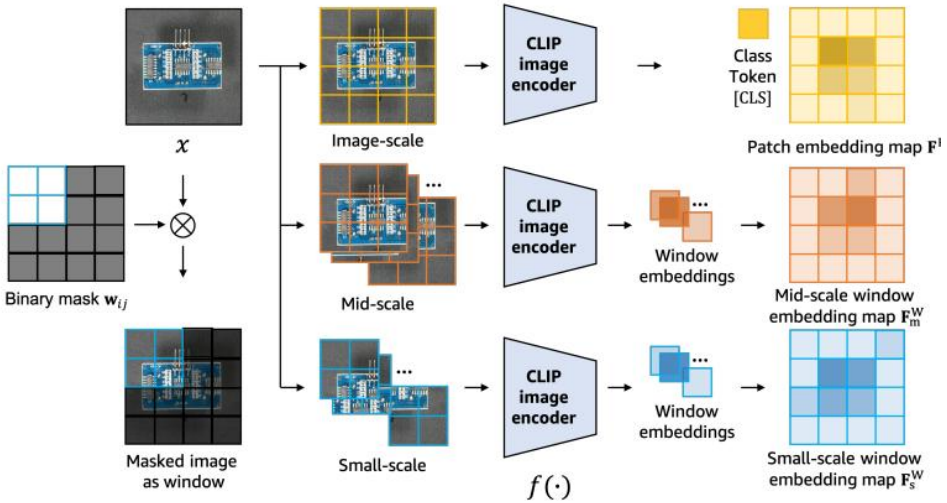
- What is the task?
  - Anomaly detection
    - 정상적인 패턴에서 벗어나는 비정상적인 데이터를 식별하는 기술
    - Vision, 시계열, 금융, 텍스트 등의 분야에서 사용됨
    - 특히, vision anomaly detection은 산업 및 제조, 의료 환경에서 사용됨



<Vision anomaly detection>

Background

- What is the task?
  - Multi-modal anomaly detection (CLIP-based)
    - CLIP
      - Web 규모의 image-text 데이터로 학습한 pretrained model
      - 강력한 일반화 능력 → 다양한 downstream task에서 zero/few-shot 능력 우수함
    - CLIP의 pretrained image-text representation 활용
    - Normal/abnormal text prompt와 image의 similarity 비교를 통한 anomaly 탐지
      - 예) WinCLIP, AnomalyCLIP, AdaCLIP…



<CLIP-based anomaly detection 예시>

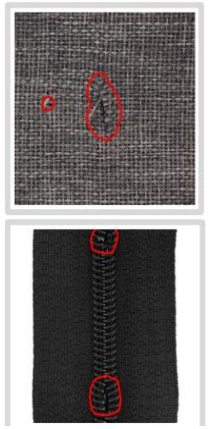
Background

- Why this topic?
  - Unsupervised anomaly detection의 경우, 이미 성능이 포화 상태에 도달함
    - MVTec, VisA 데이터셋 등에서 이미 99% 정도의 성능 달성
    - 기존의 방법론을 중심으로 한 연구는 점차 한계에 도달하고 있음
  - CLIP-based anomaly detection 관련하여 개인 연구 및 논문 작업 진행

AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP (CVPR 2025)

AA-CLIP (CVPR 2025)

- Problem formulation
  - Anomaly Unawareness in CLIP
    - CLIP에서의 평가 방식
      - Visual feature와 normal/anomaly prompt embedding 간의 cosine similarity 계산
      - Anomaly prompt embedding과의 유사도가 더 높으면 anomaly로 판정
    - CLIP text encoder의 한계: anomaly unawareness
      - 실제 응용 시, anomaly 이미지에서도 normal prompt와 더 높은 유사도 보임
      - General한 task에 집중됨 → text space에서 normal, anomaly의 의미가 서로 섞임
      - 클래스에 대한 분류 지식은 있지만 anomaly에 대한 지식이 없음

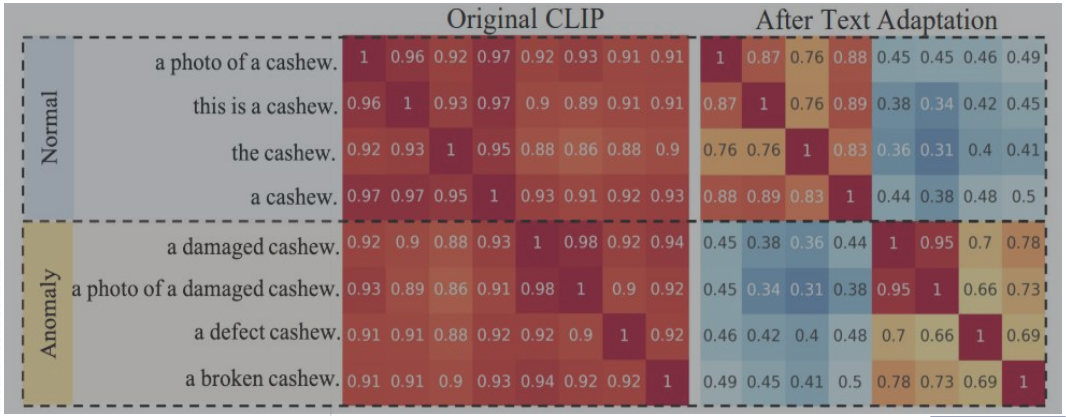


"This is a [ ] carpet."

Semantics	Similarity	Probability$_{\tau=0.01}$
broken	0.18	0.22
normal	**0.19**	**0.78**

"This is a [ ] zipper."

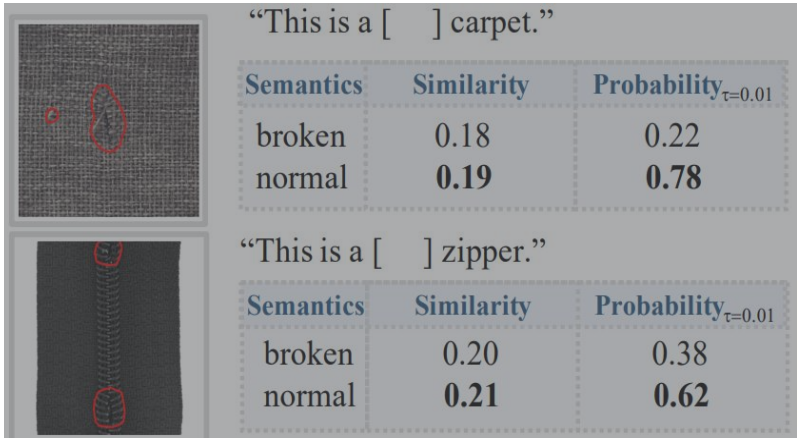
Semantics	Similarity	Probability$_{\tau=0.01}$
broken	0.20	0.38
normal	**0.21**	**0.62**

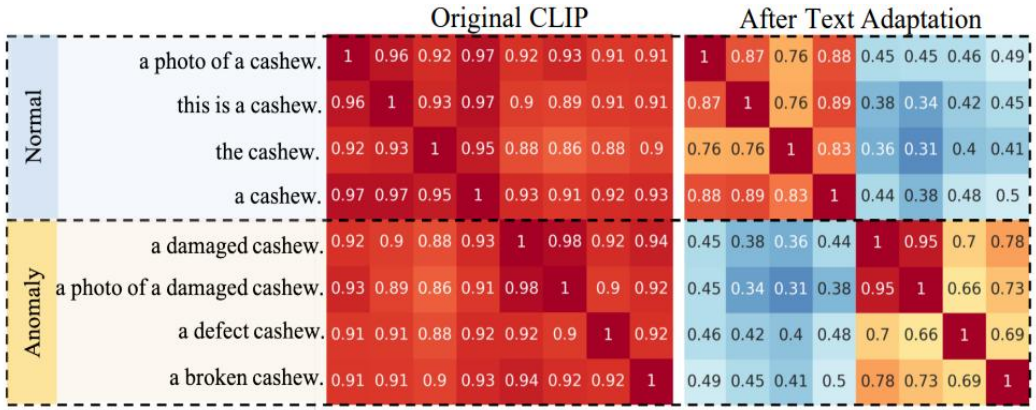


<Anomaly unawareness of CLIP text encoder>

AA-CLIP (CVPR 2025)

- Problem formulation
  - Anomaly Unawareness in CLIP
    - Prior solution: embedding adaptation dilemma
      - Domain adaptation을 통해 anomaly에 대한 지식을 주입함
        - ✓Learnable token으로 학습하거나 adapter를 붙여서 학습
      - 한계
        - ✓CLIP의 장점은 unseen data에 대한 높은 일반화 성능
        - ✓그러나 domain adaptation을 적용할 시, overfitting 가능성이 있음
        - ✓일반화 성능을 유지하면서 anomaly 정보를 반영하는 refinement 과정이 필요함

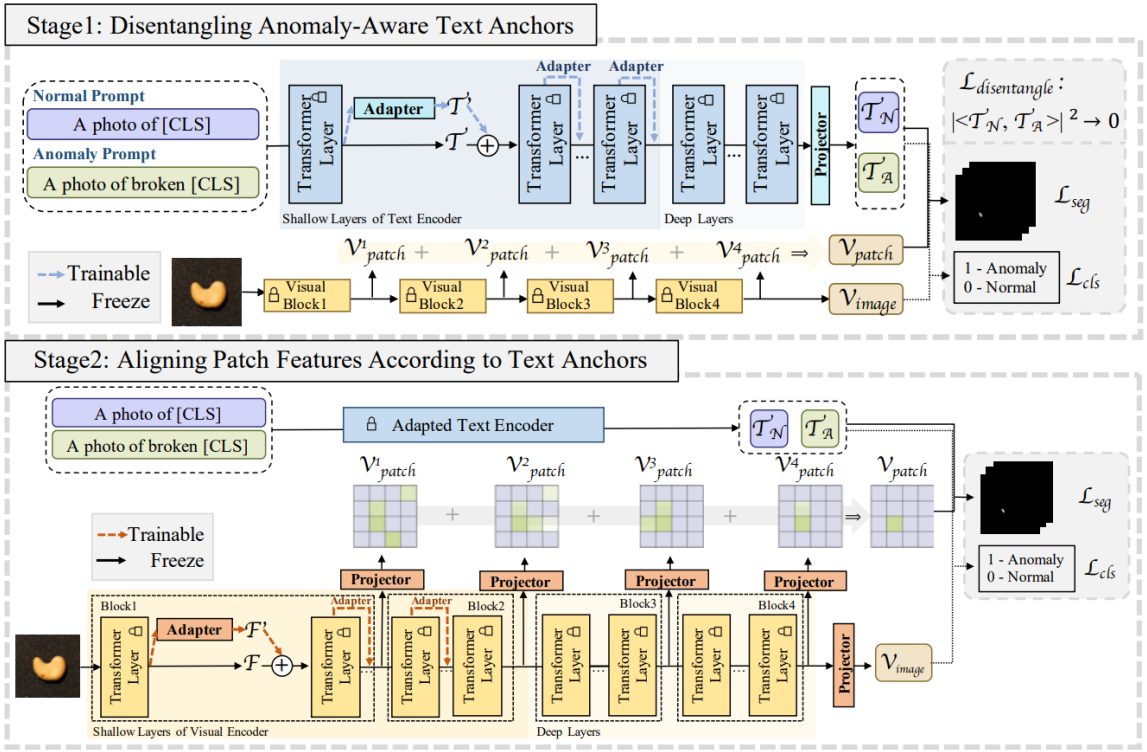




<Anomaly unawareness of CLIP text encoder>

AA-CLIP (CVPR 2025)

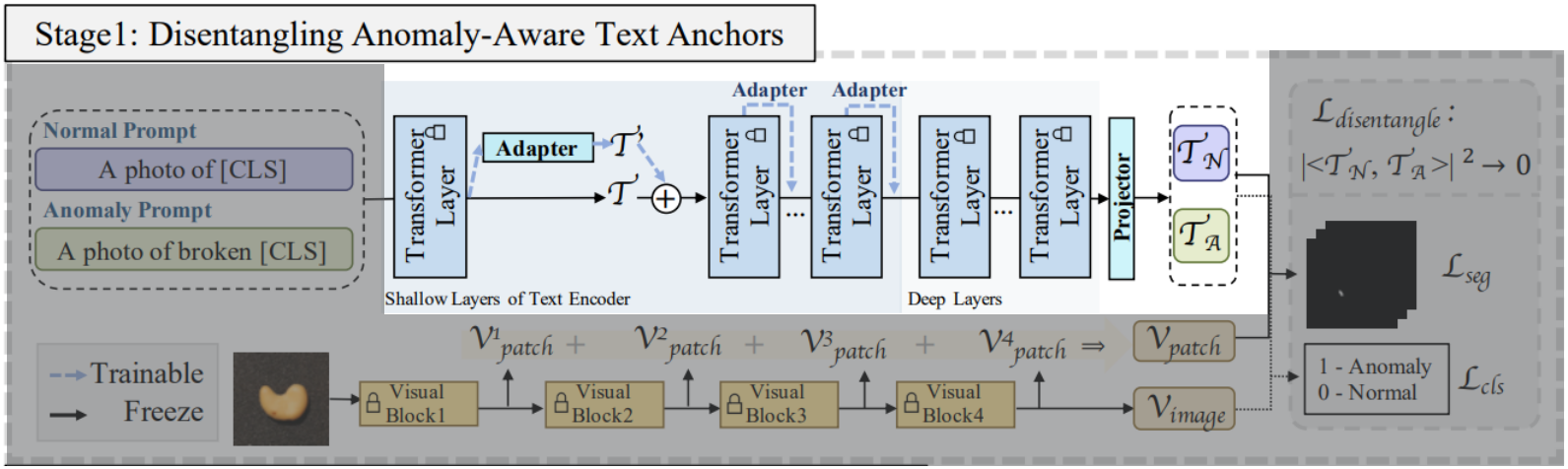
- Method (contribution)
  - CLIP의 일반화 능력 유지하면서 text, visual 공간에서 anomaly 판별 능력 강화
    - Residual adapter: CLIP의 일반적 표현 능력 보존하면서 anomaly detection 능력 보완
    - Stage 1: text space에서 normal, anomaly 의미를 효과적으로 분리함
    - Stage 2: text embedding의 anchor와 patch feature 간의 alignment를 정교하게 조정



<Two-stage pipeline of AA-CLIP>

AA-CLIP (CVPR 2025)

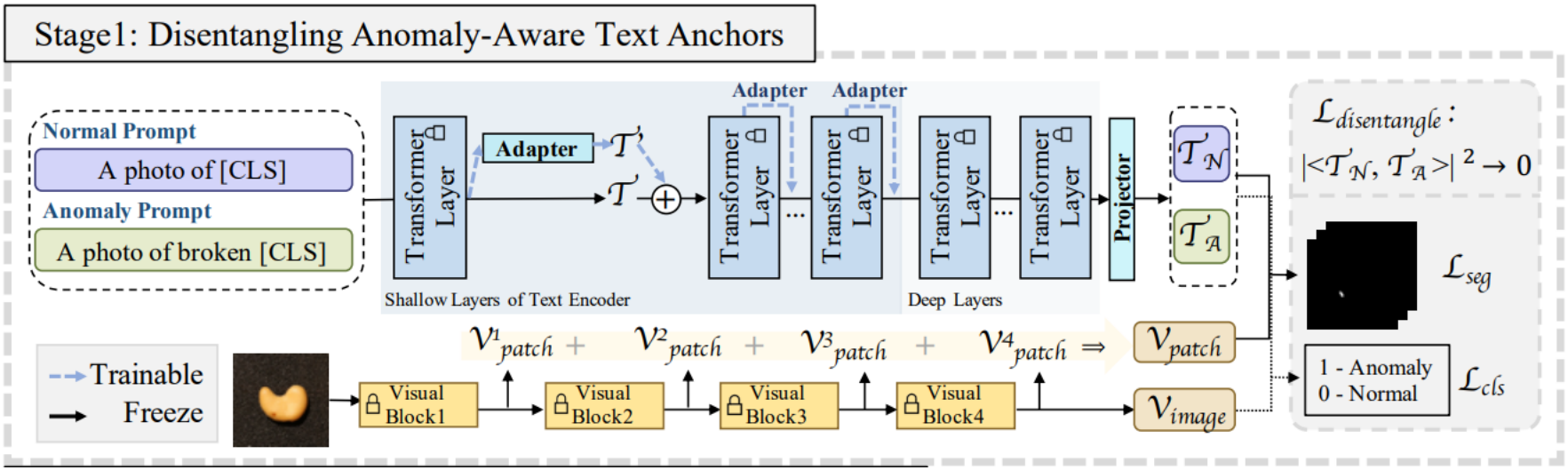
- Method
  - Residual adapter
    - Text encoder와 vision encoder의 shallow layers에 residual adapter를 도입
      - Shallow layer에 적용 → anomaly aware한 정보만 local하게 조정, 일반화 능력 유지
    - $x^i_{residual} = Norm\left(Act\left(W^i x^i\right)\right)$
    - $x^i_{enhanced} = \lambda\, x^i_{residual} + (1-\lambda)x^i$



<Residual adapter of AA-CLIP>

AA-CLIP (CVPR 2025)

- Method
  - Stage 1: Disentangling Anomaly-Aware Text Anchors
    - Vision encoder를 freeze 시킨 상태에서 text encoder의 adapter, projector만을 학습
    - Normal, anomaly를 모두 포함하는 prompt를 사용, embedding의 평균값을 anchor로 정의
      - Normal/anomaly prompt embedding의 anchor: T_n, T_a
      - Text anchor들이 vision feature들과 정렬됨
    - T_n, T_a 간 유사도가 작아지도록 $L_{disentangle}$를 도입 → normal, anomaly 의미가 분리됨



<Stage 1 pipeline of AA-CLIP>

AA-CLIP (CVPR 2025)

- Method
  - Stage 1: Disentangling Anomaly-Aware Text Anchors
    - Vision, text alignment

$$p_{cls} = CosSim(V_{image}, [T_N, T_A])$$

$$p^i_{seg} = CosSim(V_{patch}, [T_N, T_A])$$

    - Classification, segmentation loss

$$L_{cls} = BCE(p_{cls}, y)$$

$$L_{seg} = Dice(p_{seg}, S) + Focal(p_{seg}, S)$$

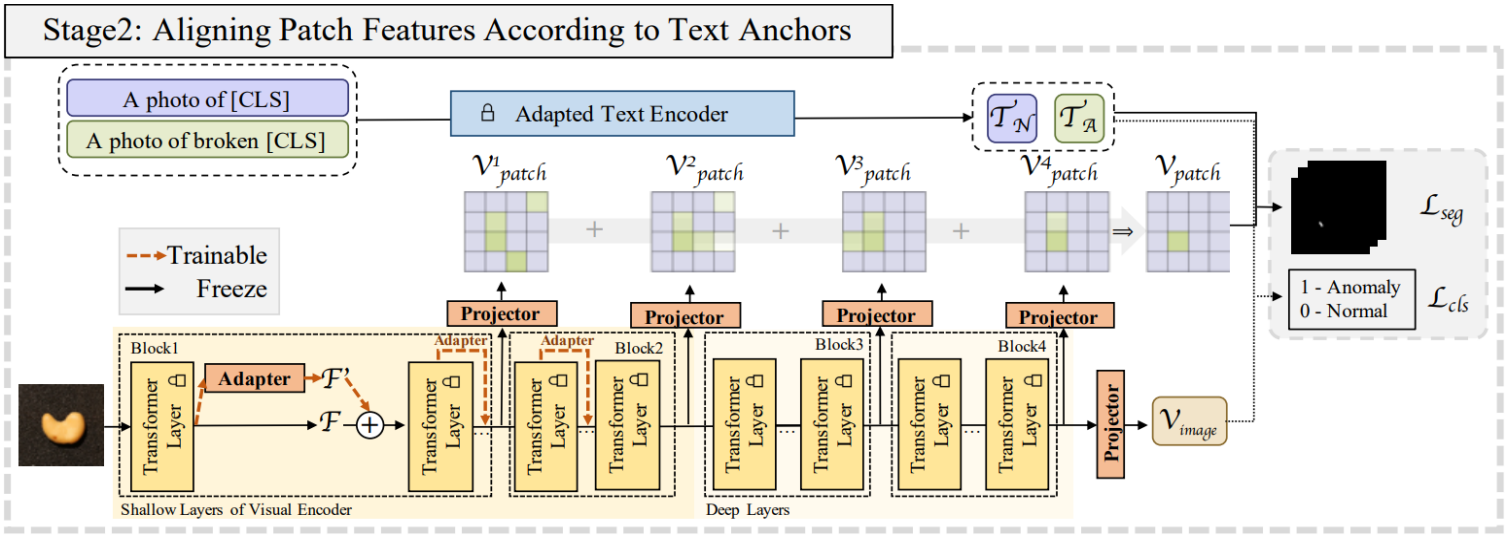
$$L_{align} = L_{cls} + L_{seg}$$

    - Total

$$L_{total} = L_{algin} + \gamma L_{disentangle}$$

AA-CLIP (CVPR 2025)

- Method
  - Stage 2: Aligning Patch Features According to Text Anchors
    - Learnable projector로 text anchor의 채널에 맞춰서 vision feature를 projection
    - 4개의 vit block에서 multi-level vision patch feature인 V_{patch}^{i}를 추출
      - V_{patch}^{i} 는 multi-scale vision feature → segmentation 성능이 향상됨
    - V_{patch}^{i}와 text anchor 간 cosine similarity score를 통해서 patch level prediction map 추출
    - Multi-scale prediction map을 합산하여 최종 prediction map 도출



<Stage 2 pipeline of AA-CLIP>

AA-CLIP (CVPR 2025)

- Experiments
  - Quantitative results
    - 원형 CLIP text encoder를 사용한 방법론
      - CLIP, WinCLIP, VAND, MVFA-AD
    - Learnable prompt 도입한 방법론
      - AnomalyCLIP, AdaCLIP
    - Full shot에서 pixel-level AUROC SOTA 성능 보임

Domain	Dataset	CLIP*	WinCLIP*	VAND*	MVFA-AD	AnomalyCLIP*	AdaCLIP	Ours			
		OpenCLIP	CVPR 2023	CVPRw 2023	CVPR 2024	ICLR2024	ECCV2024	-			
	Available training shots	-	-	full	full	full	full	2	16	64	full
Industrial	BTAD	30.6	32.8	91.1	90.1	93.3	90.8	92.8	94.4	96.5	**97.0**
	MPDD	62.1	95.2	94.9	94.5	96.2	96.6	96.3	96.5	96.3	**96.7**
	MVTec-AD	38.4	85.1	87.6	84.9	91.1	89.9	91.0	91.2	91.6	**91.9**
	VisA	46.6	79.6	94.2	93.4	95.4	**95.5**	93.4	93.8	94.0	**95.5**
Medical	Brain MRI	68.3	86.0	94.5	95.6	96.2	93.9	96.3	96.4	**96.5**	95.5
	Liver CT	90.5	96.2	95.6	96.8	93.9	94.5	97.3	97.7	97.7	**97.8**
	Retina OCT	21.3	80.6	88.5	90.9	92.6	88.5	94.2	95.1	94.4	**95.5**
	ColonDB	49.5	51.2	78.2	78.4	82.9	80.0	83.9	83.5	**84.7**	84.0
	ClinicDB	47.5	70.3	85.1	83.9	85.0	85.9	89.2	87.6	87.8	**89.9**
	Kvasir	44.6	69.7	80.3	81.9	81.9	86.4	82.1	84.6	85.2	**87.2**
	CVC-300	49.9	-	92.8	82.6	95.4	92.9	96.0	**97.4**	96.0	96.4
	Average	49.9	74.7	89.3	88.5	91.3	90.4	92.0	92.6	92.8	**93.4**

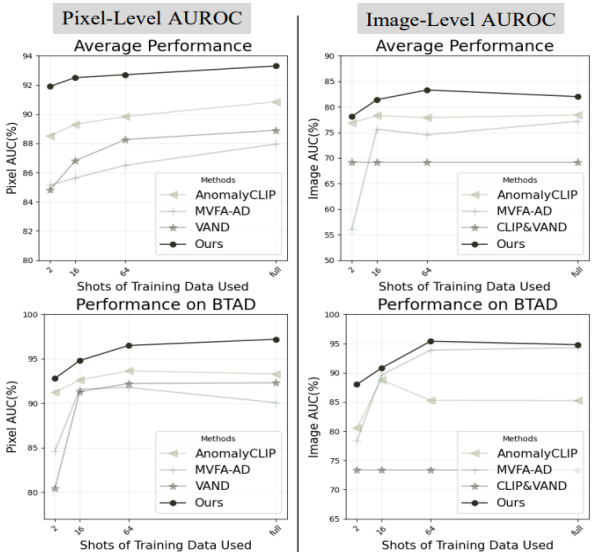
<Pixel-level AUROC of AA-CLIP>

AA-CLIP (CVPR 2025)

- Experiments
  - Quantitative results
    - Image-level에서 2-shot 시, 평균적인 성능, full shot에서 SOTA 성능 보임
    - 데이터 양이 증가할수록 다른 방법론들은 underfitting된 반면 AA-CLIP 성능 향상
  - Implementation detail
    - Backbone: ViT-L/16 기반의 OpenCLIP
    - Multi-scale feature 추출 위해서 vision encoder의 6, 12, 18, 24번째 layer output 사용
    - Image resolution: (518, 518)
    - RTX 3090 1대로 학습

Domain	Dataset	CLIP&VAND*	WinCLIP*	MVFA-AD	AnomalyCLIP*	AdaCLIP	Ours			
		OpenCLIP	CVPR 2023	CVPR 2024	ICLR2024	ECCV2024	-			
	Available training shots	-	-	full	full	full	2	16	64	full
Industrial	BTAD	73.6	68.2	94.3	85.3	90.9	88.0	90.9	94.7	**94.8**
	MPDD	73.0	63.6	70.9	73.7	72.1	63.6	**78.3**	75.7	75.1
	MVTec-AD	86.1	91.8	86.6	90.9	90.0	85.9	89.7	**92.0**	90.5
	VisA	66.4	78.0	76.5	82.1	84.3	78.4	84.0	84.1	**84.6**
Medical	Brain MRI	58.8	66.5	70.9	83.3	80.2	**84.3**	80.4	83.4	80.2
	Liver CT	54.7	64.2	63.0	61.6	64.2	69.4	68.1	69.2	**69.7**
	Retina OCT	65.6	42.5	77.3	75.7	82.7	77.4	81.0	**82.9**	82.7
	Average	68.3	67.8	77.1	78.4	80.6	78.1	81.8	**83.1**	82.5

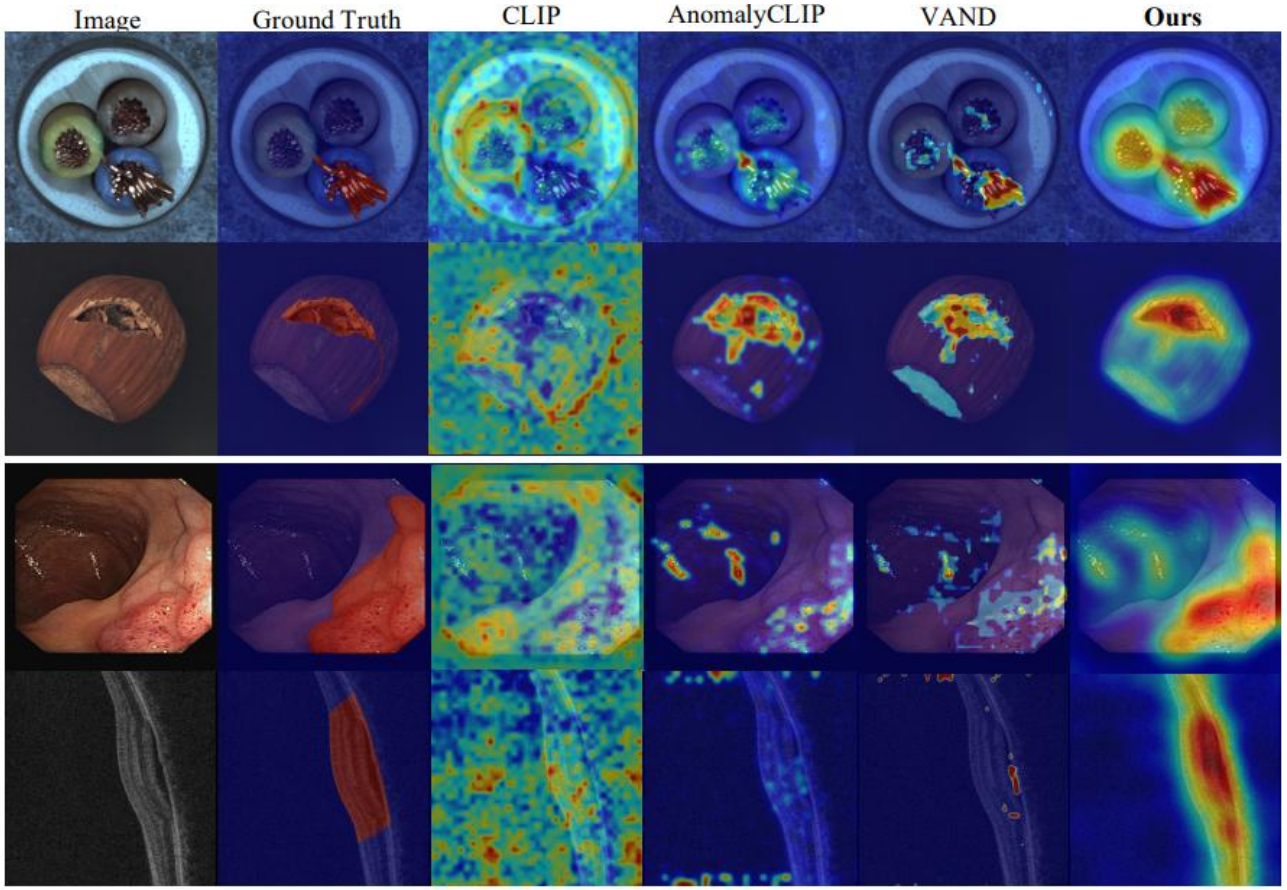
<Image-level AUROC of AA-CLIP>



<데이터양별 성능 비교>

AA-CLIP (CVPR 2025)

- Experiments
  - Qualitative results
    - AA-CLIP이 다른 방법론들에 비해서 false negative 예측이 적고 더 정확함



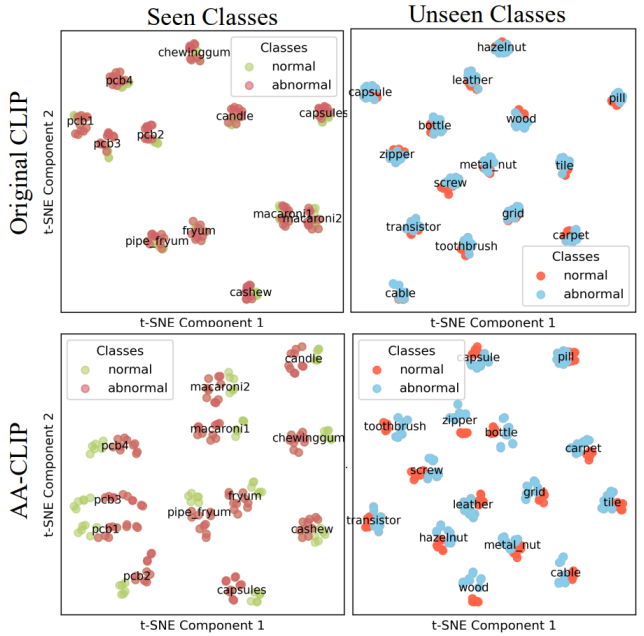
<Visualization of AA-CLIP>

AA-CLIP (CVPR 2025)

- Ablation study
  - Image space
    - Linear projector 적용: CLIP의 일반화 성능 저하 → zero-shot AUROC 성능 저하
    - Residual adapter 적용: 일반화 성능을 유지하면서 anomaly detection 성능 향상
  - Text space
    - Residual adapter 적용: 더 정밀한 semantic 표현력 → zero-shot AUROC 성능 향상
    - Disentangle loss 적용:
      - Image-level에서 성능 향상
      - Normal, anomaly 간의 독립성 확보

		Method	Avg. AUROC	
			Pixel-Level	Image-Level
		CLIP	50.3	69.3
Image	1.	+ Linear Proj. *(VAND [6])*	88.9	69.3
	2.	+ Adapter	48.9(-40.0)	53.4(-15.9)
	3.	**+ Residual Adapter**	91.3(+2.4)	80.7(+11.4)
Text	4.	**+ Residual Adapter**	92.1(+3.2)	82.6(+13.3)
	5.	**+ Disentangle Loss**	92.7(+3.8)	83.3(+14.0)

<Ablation study of AA-CLIP>

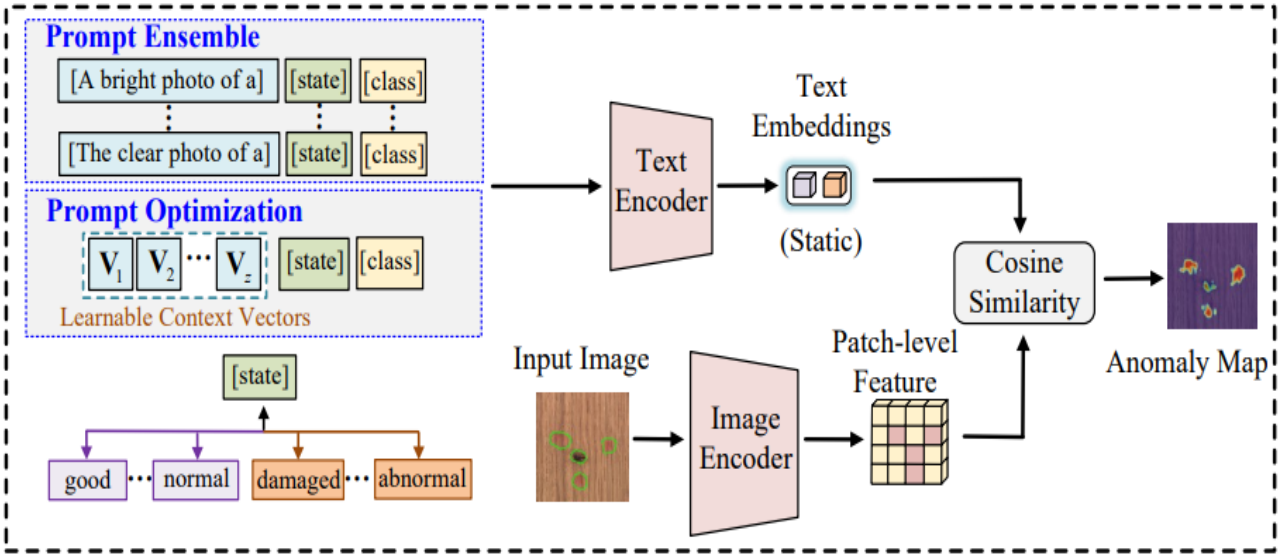


<CLIP vs AA-CLIP t-SNE>

Bayesian Prompt Flow Learning for Zero-Shot Anomaly Detection (CVPR 2025)

Bayes-PFL (CVPR 2025)

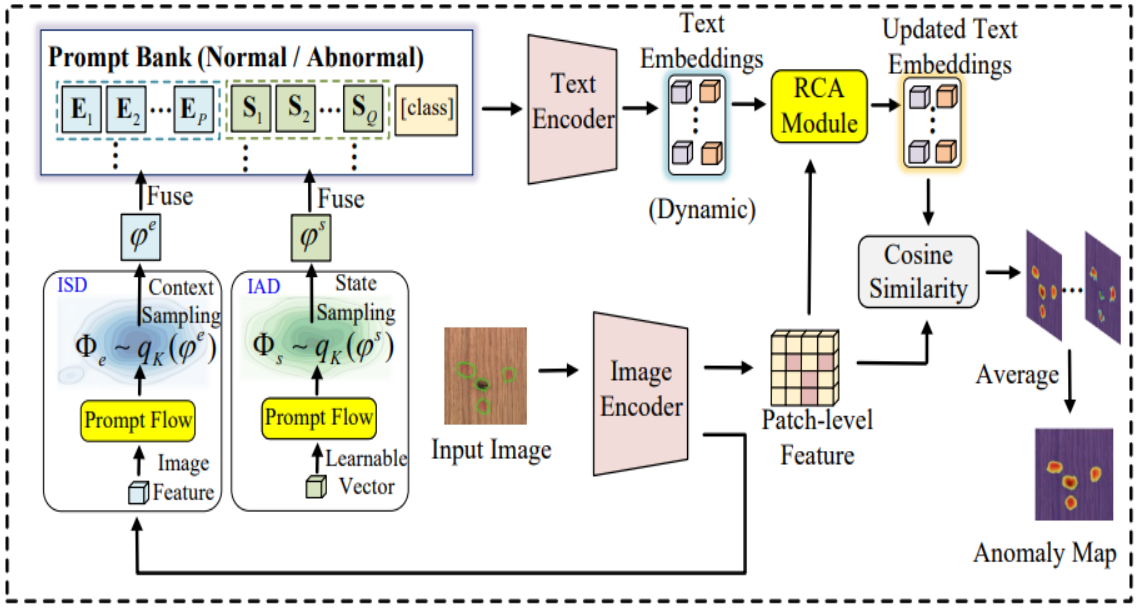
- Problem formulation
  - Zero-shot anomaly detection 방법론의 한계
    - 수동으로 설계된 prompt가 설계자의 지식과 시행착오에 크게 의존함
    - 단일 prompt representation이 복잡한 문맥이나 상태 정보를 포착하기 어려움
    - 제약되지 않는 learnable prompt space의 경우, 새로운 class에 대한 일반화 성능 저하됨



<기존 CLIP 기반 방법>

Bayes-PFL (CVPR 2025)

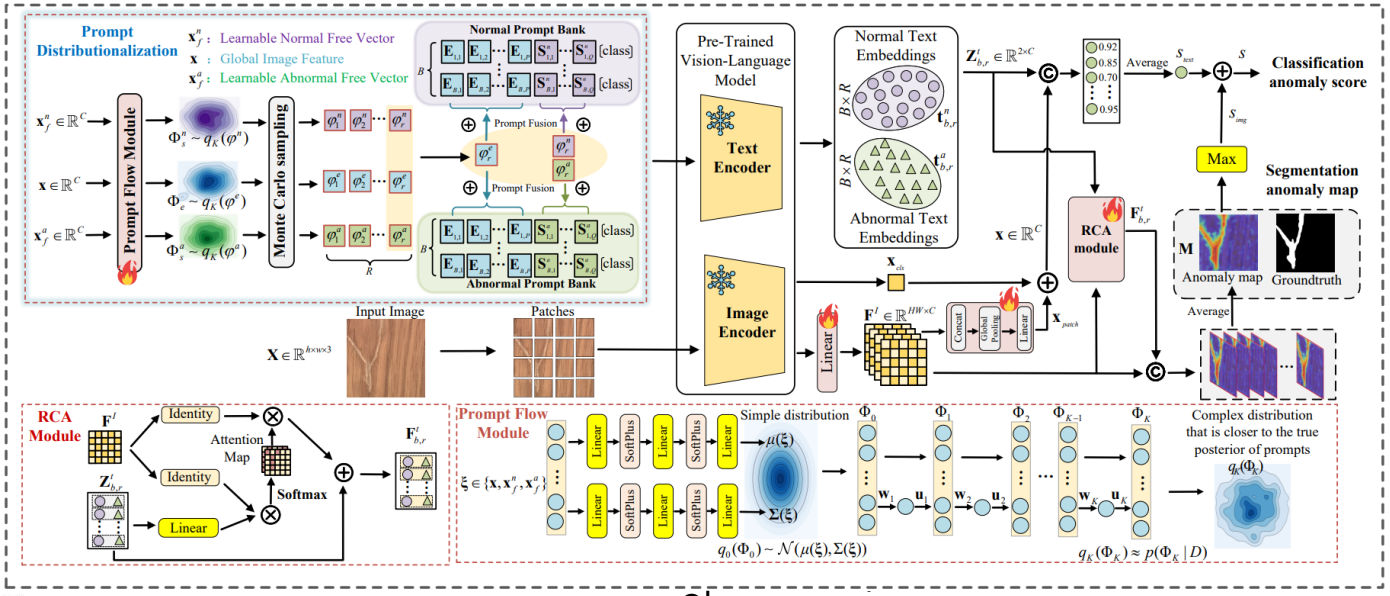
- Bayes-PFL의 contribution
  - Text prompt space를 Bayesian 관점에서 학습 가능한 확률 분포로 모델링
    - Prompt bank: normal/abnormal 클래스에 대한 learnable prompt 저장
    - Prompt flow module: text prompt 내 context, state 의 확률 분포 학습 → 일반화 향상
      - Image agnostic distribution (IAD): normal/abnormal 상태에 대한 의미 분포
      - Image specific distribution (ISD): image context에 따른 의미 분포
    - RCA module: 동적인 text embedding, image feature 간 alignment 향상



<제안하는 Bayes PFL 방법>

Bayes-PFL (CVPR 2025)

- Method
  - Overview
    - Input image X는 patch 단위로 분할 → image encoder로 patch feature 추출
    - Patch feature가 linear layer 거쳐서 vision-text embedding 차원에 맞게 projection됨
    - Prompt flow module이 image, text를 통해 context distribution과 state distribution에 대해 학습
    - Monte carlo sampling 통해서 각 분포 결과가 prompt bank와 결합→ text prompt 생성
    - Text embedding이 RCA module을 통해서 image embedding과 정렬 → anomaly map 도출



<Bayes-PFL의 framework>

Bayes-PFL (CVPR 2025)

- Method
  - Two-class prompt banks
    - Text prompt를 context, state, class 단어로 분해
      - Context: 객체 종류, 재료 등의 정보
      - State: normal/abnormal 정보
    - Context와 state에 대한 token을 learnable token으로 설정

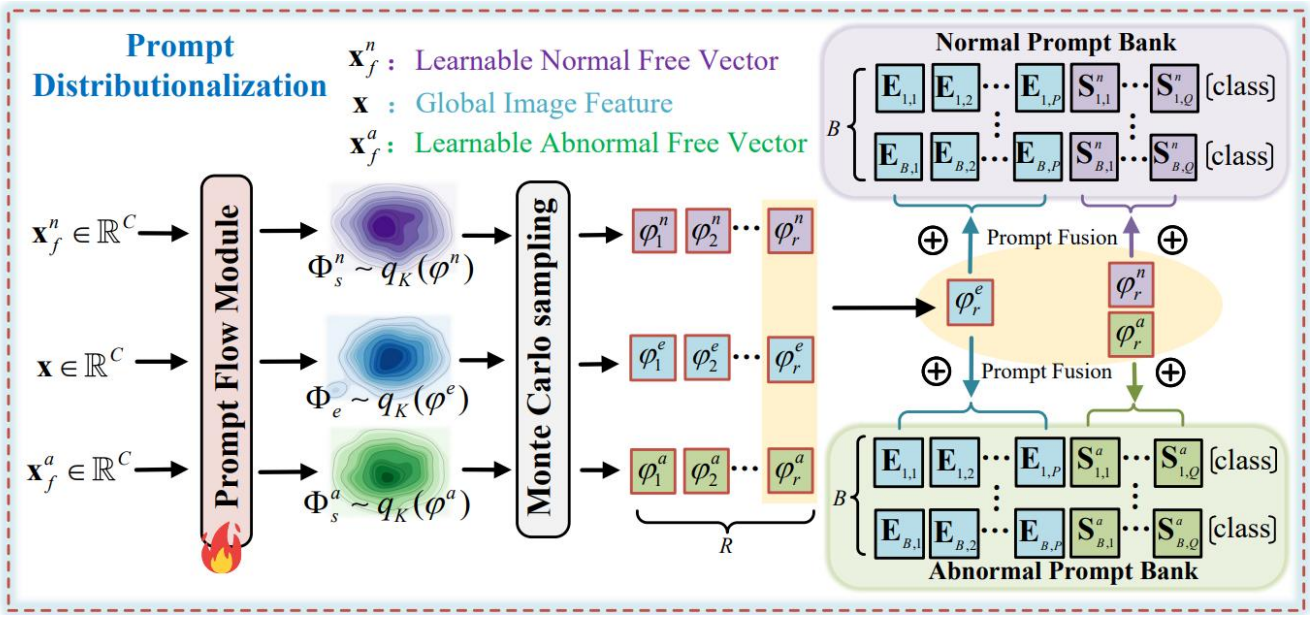
$$g_b^n = [\mathbf{E}_{b,1}][\mathbf{E}_{b,2}]\cdots[\mathbf{E}_{b,P}][\mathbf{S}_{b,1}^n]\cdots[\mathbf{S}_{b,Q}^n][\text{class}]$$

$$g_b^a = [\mathbf{E}_{b,1}][\mathbf{E}_{b,2}]\cdots[\mathbf{E}_{b,P}][\mathbf{S}_{b,1}^a]\cdots[\mathbf{S}_{b,Q}^a][\text{class}]$$

      - g_b^n, g_b^a: normal/abnormal prompt bank 내의 하나의 prompt
      - $[E_{b,i}]$: image의 context 에 대한 learnable token
      - $[S_{b,i}^n]$: image의 normal, abnormal에 대한 learnable token
    - Loss 설정 통해 같은 prompt bank의 embedding 간 직교성 강제 → prompt bank 내 중복 제거

Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Prompt flow module 활용해서 text prompt를 모델링, 생성하는 과정을 표현
    - Distribution에서 샘플링된 결과를 prompt bank와 fusion
    - Prompt space를 ISD, IAD로 제약
      - Unseen class에 대한 모델의 일반화 성능 향상



<Prompt distributionalization>

Bayes-PFL (CVPR 2025)

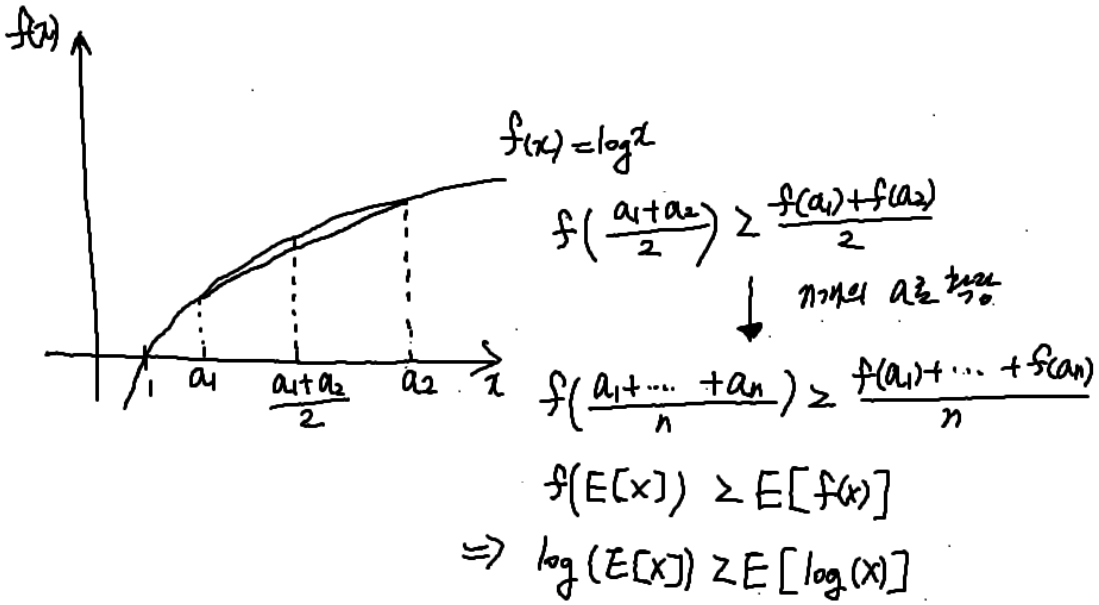
- Method
  - Prompt Distributionalization
    - Bayesian inference
      - $D = \{X, Y_c, Y_S\}$
        - ✓D: 보조 데이터셋
        - ✓X: input image
        - ✓Y_c: image level label (normal, anomaly label)
        - ✓Y_S: pixel level mask
      - $\Phi = \{\Phi_e,\ \Phi_s^n,\ \Phi_s^a\}$
        - ✓Φ_e: context word embedding
        - ✓Φ_s^n: normal word embedding
        - ✓Φ_s^a: abnormal word embedding

Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Bayesian inference
      - Posterior probability

$$p(\Phi|D) = \frac{p(D|\Phi)p(\Phi)}{p(D)}$$

      - 젠슨 부등식



Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Lower bound: $L_e(D)$

$$\log p(D) = \log \int p(D|\Phi)p(\Phi)d\Phi$$

$$\geq E_{q_\gamma(\Phi|D)}[\log p(D,\Phi) - \log q_\gamma(\Phi|D)] = -\mathcal{L}_e(D)$$

∴ P(D)는 계산 불가능 → p(Φ|D)를 learnable distribution $q_r(\Phi)$로 근사

$$q_\gamma(\Phi) = q_\gamma(\boldsymbol{\varphi}^e)q_\gamma(\boldsymbol{\varphi}^n)q_\gamma(\boldsymbol{\varphi}^a)$$

✓각 embedding으로 얻은 분포들이 모두 독립적이라는 가정

✓Lower bound인 $L_e(D)$를 최소화하도록 학습해서 얻게 됨

Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Lower bound: $L_e(D)$ 전개 과정

$$\begin{cases} \log(E[X]) \geq E[\log(X)] \quad -(1) \\ \log P(D) = \log \int P(D|\phi)P(\phi)\,d\phi \quad -(2) \\ E_p[f(x)] = \int f(x)P(x)\,dx \quad -(3) \end{cases}$$

$$\log P(D) = \log \int P(D|\phi)P(\phi)\,d\phi = \log \int \frac{P(D|\phi)P(\phi)}{q_\theta(\phi|D)} q_\theta(\phi|D)\,d\phi \quad (3)\ 적용$$

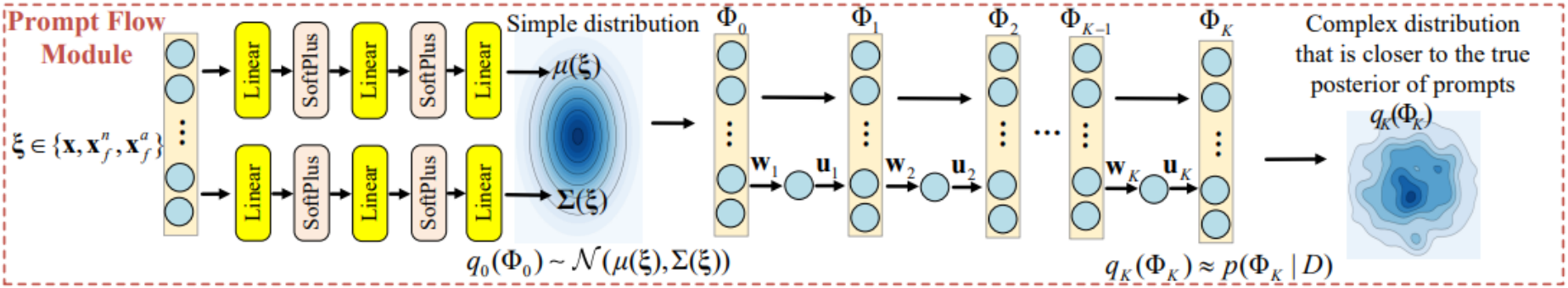
$$= \log E_{q_\theta(\phi|D)}\left[\frac{P(D|\phi)P(\phi)}{q_\theta(\phi|D)}\right] \geq E_{q_\theta(\phi|D)}\left[\log \frac{P(D|\phi)P(\phi)}{q_\theta(\phi|D)}\right] \quad (1)\ 적용$$

$$= E_{q_\theta(\phi|D)}[\log P(D|\phi)P(\phi) - \log q_\theta(\phi|D)]$$

$$= E_{q_\theta(\phi|D)}[\log P(D,\phi) - \log q_\theta(\phi|D)] = -L_e(D)$$

Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Prompt flow module
      - Simple distribution 생성
        - ✓여러 개의 Linear, Softplus layer를 반복 적용
          - Softplus → 항상 양수인 표준편차 σ를 얻음
          - 안정적으로 정규 분포인 simple distribution 생성
      - Distribution refinement
        - ✓Distribution에 linear, tanh 연산 반복 적용 → complex distribution $q_k(\Phi_k)$ 생성
        - ✓$L_e(D)$ 최적화로 인해 $q_k(\Phi_k)$가 $p(\Phi_k|D)$에 가까워짐



<Prompt flow module>

Bayes-PFL (CVPR 2025)

- Method
  - Prompt Distributionalization
    - Prompt flow module
      - Distribution의 밀도 보정

$$\log q_K(\Phi_K) = \log q_0(\Phi_0) - \sum_{k=1}^{K} \log \left|1 + \mathbf{u}_k^{\mathrm{T}} \phi(\Phi_k)\right|$$

✓Distribution을 refinement하면 공간의 모양이 바뀌므로 밀도 변화 반영 필요

✓공간이 커지면 밀도는 낮아지고, 작아지면 밀도는 높아져야 함

✓Jacobian 수식을 이용해서 초기 distribution에서 jacobian 상수를 빼게 됨

      - Input: global image feature인 x경우

✓이미지가 context 의미를 담고 있음

✓Output: Image Specific Distribution (ISD)

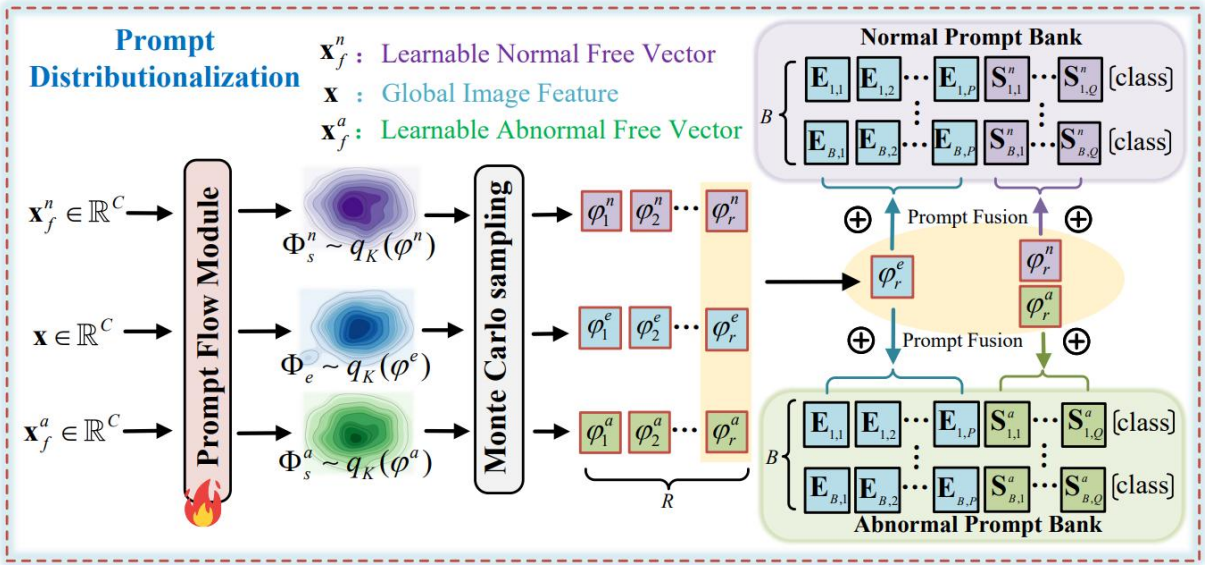
      - Input: trainable vector x_f^n, x_f^a인 경우

✓Normal, abnormal state에 대해 통일된 state semantics를 학습하는 데 사용됨

✓Output: Image Agnostic Distribution (IAD)

Bayes-PFL (CVPR 2025)

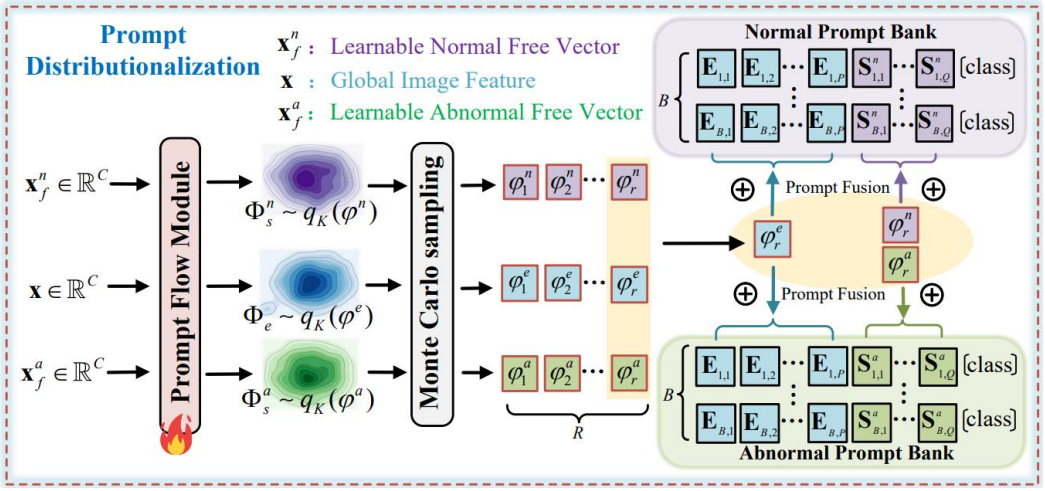
- Method
  - Prompt Distributionalization
    - Prompt sampling and fusion
      - Monte carlo sampling을 사용해서 각 x마다 r개의 vector를 추출
        - ✓Monte carlo sampling
          - 확률 변수들의 각 값을 여러 개 샘플링
          - 대수의 법칙에 따라 샘플의 수가 클수록, 평균 → 분포의 평균
        - ✓Vector: context vector, normal state vector, abnormal state vector



<Prompt distributionalization>

Bayes-PFL (CVPR 2025)

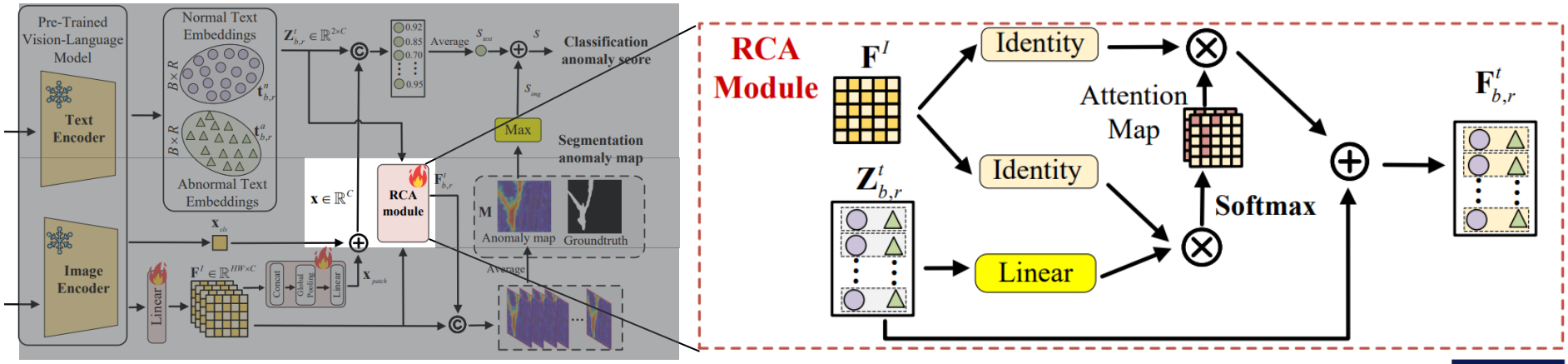
- Method
  - Prompt Distributionalization
    - Prompt sampling and fusion
      - Context vector, state vector, class token을 합쳐서 하나의 prompt로 설정함
        - ✓Normal state text prompt
          - $g^n_{b,r}=[E_{b,1}+\varphi^e_r][E_{b,2}+\varphi^e_r]\cdots[E_{b,P}+\varphi^e_r][S^n_{b,1}+\varphi^e_r]\cdots[S^n_{b,Q}+\varphi^e_r][\text{class}]$
        - ✓Abnormal state text prompt
          - $g^a_{b,r}=[E_{b,1}+\varphi^e_r][E_{b,2}+\varphi^e_r]\cdots[E_{b,P}+\varphi^e_r][S^a_{b,1}+\varphi^e_r]\cdots[S^a_{b,Q}+\varphi^e_r][\text{class}]$
      - $B \times R$ 개의 prompt 생성 → 수작업 감소로 인한 자원 효율성 증대



<Prompt distributionalization>

Bayes-PFL (CVPR 2025)

- Method
  - Residual Cross-modal Attention (RCA) module
    - Text feature
      - Random sampling으로 인해 동적으로 변함
      - IAD, ISD를 모두 갖고 있으므로 global한 feature에 해당함
    - Image feature
      - Patch에 대한 feature이므로 fine-grained한 feature에 해당함
    - Problem
      - 두 modal 간의 modality gap뿐만 아니라 global한 정보, local한 정보 사이의 gap이 존재



<Prompt distributionalization>

Bayes-PFL (CVPR 2025)

- Method
  - Residual Cross-modal Attention (RCA) module
    - Solution

$$F_{b,r}^t = Z_{b,r}^t + softmax(Q_{b,r}\ (F^I)^T/\sqrt{C})\ F^I$$

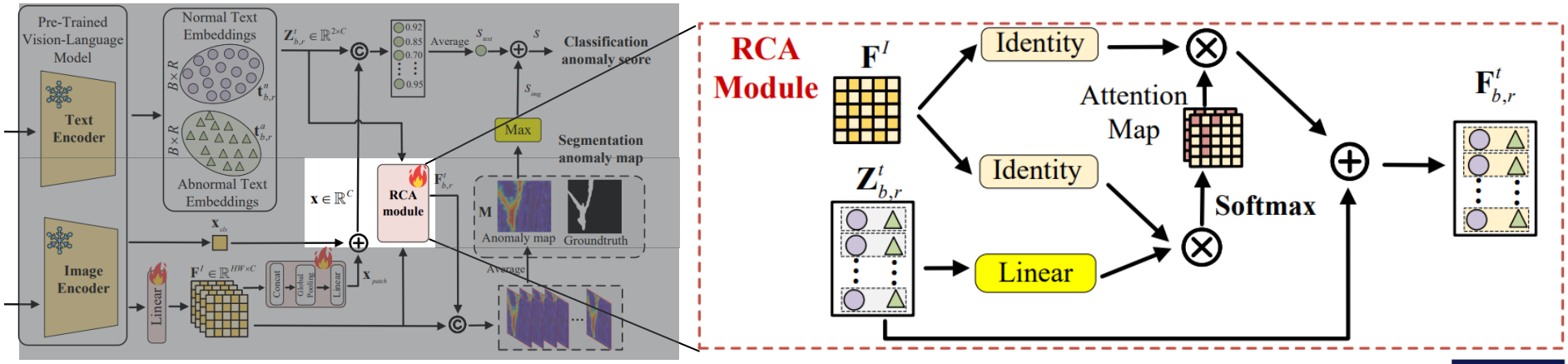
✓$Z_{b,r}^t$: 전체 text embedding, cross-attention 연산에서 query로 설정됨

- 전체 text embedding 의미가 희석되지 않게 residual 값으로 설정됨

✓F^I: patch-level image embedding, cross-attention 연산에서 key, value로 설정됨

✓$F_{b,r}^t$: patch-level에 해당하는 text embedding

Modality 간 gap뿐만 아니라 global, local간 modality gap 줄임 → zero-shot 성능 향상



<Prompt distributionalization>

Bayes-PFL (CVPR 2025)

- Method
  - Anomaly scoring
    - Patch-level anomaly map

$$M_{l,b,r} = softmax(Up(\tilde{F}_l^I \tilde{F}_{l,b,r}^{tT}))$$

$$M = \frac{1}{LBR}\sum_{l=1}^{L}\sum_{b=1}^{B}\sum_{r=1}^{R} M_{l,b,r}$$

✓Patch-level image embedding과 text embedding 각각을 L2 norm

✓L2 norm한 embedding 간 dot product 연산 후, 평균값 계산

    - Image-level anomaly map

$$s_{text} = (\frac{1}{BR})\sum_{b}^{B}\sum_{r}^{R} softmax(\tilde{x}\tilde{Z}_{b,r}^{tT})$$

✓$x = x_{cls} + x_{patch}$

✓Global한 image feature과 global한 text feature 간의 dot product 연산

$$s_{image} = Max(M)$$

$$s_{total} = s_{text} + s_{image}$$

Bayes-PFL (CVPR 2025)

- Method
  - Loss function
    - Orthogonal loss

$L_{ort} = \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \{(< t_{i,1}^{n}, t_{j,1}^{n} >)^2 + (< t_{i,1}^{a}, t_{j,1}^{a} >)^2\}$

✓<,> : cosine similarity 연산

✓각 normal prompt, anomaly prompt들 내에서 중복된 정보가 없도록 유도함

- Prompt flow loss

$$\begin{aligned}\mathcal{L}_p(D) &= E_{q_\gamma(\Phi|D)}[\log q_\gamma(\Phi|D) - \log p(D,\Phi)] \\ &= E_{q_0(\Phi_0)}\left[\log q_0(\Phi_0) - \sum_{k=1}^{K} \log|1 + \mathbf{u}_k^{\mathrm{T}}\phi(\Phi_k)|\right] \\ &\quad - E_{q_0(\Phi_0)}[\log p(\Phi_K)] - E_{q_0(\Phi_0)}[\log p(D|\Phi_K)]\end{aligned}$$

$q_k(\Phi_k)$를 initial distribution인 $q_0(\Phi_0)$에 대한 수식으로 표현

정규화를 위해 initial distribution를 subtract

마지막 term은 data에 대한 loglikelihood

✓Focal loss, dice loss, cross entropy loss 합으로 대체

Bayes-PFL (CVPR 2025)

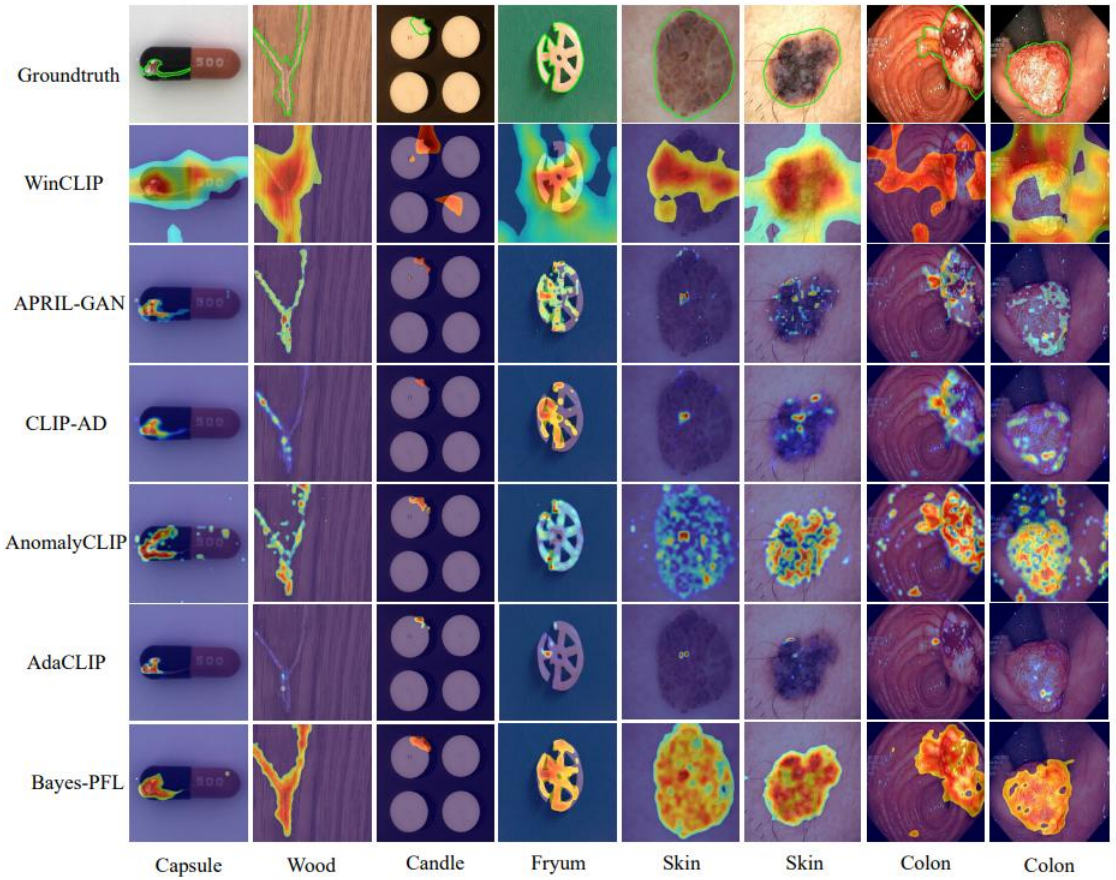
- Experiments
  - Quantitative results
    - Zero-shot AD task의 방법론들과 Bayse-PFL의 성능 비교
      - Metric: I-AUROC, P-AUROC, AUPRO
    - 모든 데이터셋에서 SOTA 성능 달성

Domain	Metric	Dataset	WinCLIP [18]	APRIL-GAN [9]	CLIP-AD [10]	AnomalyCLIP [49]	AdaCLIP [8]	**Bayes-PFL**
Industrial	Image-level (AUROC, F1-Max, AP)	MVTec-AD	(91.8, 92.9, 95.1)	(86.1, 90.4, 93.5)	(89.8, 91.1, 95.3)	(91.5, 92.8, 96.2)	(92.0, 92.7, 96.4)	(92.3, 93.1, 96.7)
		VisA	(78.1, 79.0, 77.5)	(78.0, 78.7, 81.4)	(79.8, 79.2, 84.3)	(82.1, 80.4, 85.4)	(83.0, 81.6, 84.9)	(87.0, 84.1, 89.2)
		BTAD	(83.3, 81.0, 84.1)	(74.2, 70.0, 71.7)	(85.8, 81.7, 85.2)	(89.1, 86.0, 91.1)	(91.6, 88.9, 92.4)	(93.2, 91.9, 96.5)
		KSDD2	(93.5, 71.4, 77.9)	(90.3, 70.0, 74.4)	(95.2, 84.4, 88.2)	(92.1, 71.0, 77.8)	(95.9, 84.5, 95.9)	(97.3, 92.3, 97.9)
		RSDD	(85.3, 73.5, 65.3)	(73.1, 59.7, 50.5)	(88.3, 74.1, 73.9)	(73.5, 59.0, 55.0)	(89.1, 75.0, 70.8)	(94.1, 89.6, 92.3)
		DAGM	(89.6, 86.4, 90.4)	(90.4, 86.9, 90.1)	(90.8, 88.4, 90.5)	(95.6, 93.2, 94.6)	(96.5, 94.1, 95.7)	(97.7, 95.7, 97.0)
		DTD-Synthetic	(95.0, 94.3, 97.9)	(83.9, 89.4, 93.6)	(91.5, 91.8, 96.8)	(94.5, 94.5, 97.7)	(92.8, 92.2, 97.0)	(95.1, 95.1, 98.4)
	Pixel-level (AUROC, PRO, AP)	MVTec-AD	(85.1, 64.6, 18.0)	(87.6, 44.0, 40.8)	(89.8, 70.6, 40.0)	(91.1, 81.4, 34.5)	(86.8, 33.8, 38.1)	(91.8, 87.4, 48.3)
		VisA	(79.6, 56.8, 5.0)	(94.2, 86.8, 25.7)	(95.0, 86.9, 26.3)	(95.5, 87.0, 21.3)	(95.1, 71.3, 29.2)	(95.6, 88.9, 29.8)
		BTAD	(71.4, 32.8, 11.2)	(91.3, 23.0, 32.9)	(93.1, 59.8, 46.7)	(93.3, 69.3, 42.0)	(87.7, 17.1, 36.6)	(93.9, 76.6, 47.1)
		KSDD2	(97.9, 91.2, 17.1)	(97.9, 51.1, 61.6)	(99.3, 85.4, 58.9)	(99.4, 92.7, 41.8)	(96.1, 70.8, 56.4)	(99.6, 97.6, 73.7)
		RSDD	(95.1, 75.4, 2.1)	(99.4, 64.9, 30.6)	(99.2, 90.1, 31.9)	(99.1, 92.0, 19.1)	(99.5, 50.5, 38.2)	(99.6, 98.0, 39.1)
		DAGM	(83.2, 55.4, 3.1)	(99.2, 44.7, 42.6)	(99.0, 83.1, 40.7)	(99.1, 93.6, 29.5)	(97.0, 40.9, 44.2)	(99.3, 98.0, 43.1)
		DTD-Synthetic	(82.5, 55.4, 11.6)	(96.6, 41.6, 67.3)	(97.1, 68.7, 62.3)	(97.6, 88.3, 52.4)	(94.1, 24.9, 52.8)	(97.8, 94.3, 69.9)
Medical	Image-level (AUROC, F1-Max, AP)	HeadCT	(83.7, 78.8, 81.6)	(89.3, 82.0, 89.6)	(93.8, 90.5, 92.2)	(95.3, 89.7, 95.2)	(93.4, 86.5, 92.2)	(96.5, 92.9, 95.5)
		BrainMRI	(92.0, 84.2, 90.7)	(89.6, 85.3, 84.5)	(92.8, 88.7, 85.5)	(96.1, 92.3, 92.3)	(94.9, 90.4, 94.2)	(96.2, 92.8, 92.4)
		Br35H	(80.5, 74.1, 82.2)	(93.1, 85.4, 92.9)	(96.0, 90.8, 95.5)	(97.3, 92.4, 96.1)	(95.7, 89.1, 95.7)	(97.8, 93.6, 96.2)
	Pixel-level (AUROC, PRO, AP)	ISIC	(83.3, 55.1, 62.4)	(85.8, 13.7, 69.8)	(81.6, 29.0, 65.5)	(88.4, 78.1, 74.4)	(85.4, 5.3, 70.6)	(92.2, 87.6, 84.6)
		CVC-ColonDB	(64.8, 28.4, 14.3)	(78.4, 28.0, 23.2)	(80.3, 58.8, 23.7)	(81.9, 71.4, 31.3)	(79.3, 6.5, 26.2)	(82.1, 76.1, 31.9)
		CVC-ClinicDB	(70.7, 32.5, 19.4)	(86.0, 41.2, 38.8)	(85.8, 69.7, 39.0)	(85.9, 69.6, 42.2)	(84.3, 14.6, 36.0)	(89.6, 78.4, 53.2)
		Endo	(68.2, 28.3, 23.8)	(84.1, 32.3, 47.9)	(85.6, 57.0, 51.7)	(86.3, 67.3, 50.4)	(84.0, 10.5, 44.8)	(89.2, 74.8, 58.6)
		Kvasir	(69.8, 31.0, 27.5)	(80.2, 27.1, 42.4)	(82.5, 48.1, 46.2)	(81.8, 53.8, 42.5)	(79.4, 12.3, 43.8)	(85.4, 63.9, 54.2)

<Zero-shot AD task에서의 정량적 성능 비교 >

Bayes-PFL (CVPR 2025)

- Experiments
  - Qualitative results
    - Anomaly 영역 탐지 성능이 뛰어나며, 의료 데이터셋에서 가장 탁월함



<Zero-shot AD task에서의 정성적 성능 비교 >

Bayes-PFL (CVPR 2025)

- Ablation study
  - Components
    - ISD 제거 시, 성능 하락 → 이미지에 조건화된 context 분포 모델링이 중요
    - L_{ort}제거 시, I-AUROC, P-AUROC 1%씩 감소
  - Monte carlo sampling
    - 대수의 법칙을 따라 r이 클수록 성능이 향상됨
    - R이 클수록 성능이 향상되지만 sampling 시 소요되는 시간이 길어짐

		Image-level		Pixel- level	
		AUROC	AP	AUROC	AP
Module Ablation	w/o ISD	89.1	94.5	88.3	46.5
	w/o IAD	90.3	95.2	89.5	47.2
	w/o ISD, IAD	88.6	93.3	87.1	45.6
	w/o RCA	91.8	95.9	89.7	46.1
Loss Ablation	w/o $\mathcal{L}_{ort}$	91.2	96.0	90.2	47.3
Classification Ablation	w/o $\mathbf{x}_{cls}$	89.3	94.4	—	—
	w/o $\mathbf{x}_{patch}$	90.9	95.8	—	—
	w/o s_{text}	89.0	95.1	—	—
	w/o s_{img}	90.9	95.3	—	—
	Bayes-PFL	**92.3**	**96.7**	**91.8**	**48.3**

<Bayes-PFL의 ablation study>

R	Image-level		Pixel-level		Time (ms)
	AUROC	AP	AUROC	AP	
1	89.5±1.1	94.1±1.2	87.1±1.3	45.4±0.6	132.2±1.6
3	90.0±1.0	94.8±0.9	88.2±1.2	46.1±0.6	200.3±1.4
5	91.1±0.8	95.4±0.9	89.5±1.0	46.8±0.5	254.2±1.4
7	92.0±0.7	96.3±0.8	90.9±0.8	47.5±0.5	297.4±1.7
10	92.7±0.4	96.8±0.5	91.8±0.6	48.3±0.4	388.5±1.5
13	92.7±0.4	96.9±0.5	91.9±0.5	48.4±0.4	467.5±1.5
16	92.8±0.3	96.9±0.4	92.0±0.4	48.6±0.3	621.6±1.6
20	92.9±0.2	97.0±0.2	92.1±0.2	48.9±0.1	726.1±1.7

<Monte Carlo Sampling 수에 따른 성능 변화>

1) Ma, Wenxin, et al. "Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip." CVPR, 2025
2) Qu, Zhen, et al. "Bayesian Prompt Flow Learning for Zero-Shot Anomaly Detection." CVPR, 2025

Comparison & Analysis

- Problem
  - CLIP text encoder의 domain specific한 task에서의 성능 저하
- Solution

논문	AA-CLIP[1)]	Bayes-PFL[2)]
Task	Few-shot anomaly detection	Zero-shot anomaly detection
전략 방향	Lightweight Fine-tuning via Adapter	Distribution-based Prompt Sampling
Text Encoder 접근 방식	1. Text encoder, image encoder의 앞 layer에 residual adapter 추가 2. Text encoder 학습 시, disentangle loss를 사용해서 normal, anomaly 의미 분리	1. Normal text, anomaly text, image feature에 대한 각각의 distribution 생성 2. 각 distribution에서 r번 샘플 추출 후, [image embedding][state token][class] 로 prompt bank 구축 3. Prompt를 text encoder에 넣어서 feature 추출
장점	CLIP의 general한 성능을 유지하면서, anomaly detection task에 맞게 학습	Prompt flow module을 통해 context와 normal, abnormal에 대한 공간 모델링
한계	Few-shot보다는 full-shot에 적절한 방법론	-

Thank you ☺