
Knowledge Distillation based methods for Industrial Image Anomaly Detection

2026 3월 20일

Data Mining & Quality Analytics Lab.

고재영

발표자 소개



❖ **고재영 (JaeYoung Koh)**

- 고려대학교 산업경영공학과 석사과정(2025.03 ~ Present)
- Data Mining & Quality Analytics Lab (김성범 교수님)

❖ **Research Interest**

- Time Series Anomaly Detection (TSAD)
- Industrial Image Anomaly Detection (IAD)

❖ **Contact**

- goh0316@korea.ac.kr

Knowledge Distillation based methods for Industrial Image Anomaly Detection

❖ Background

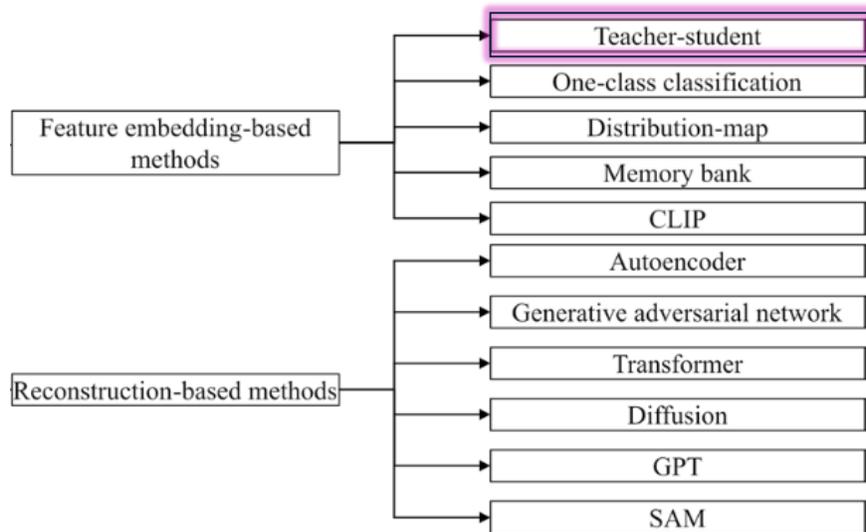
❖ Paper review

- **Uninformed Students:** Student–Teacher Anomaly Detection with Discriminative Latent Embeddings (2020, CVPR)
- **MKD:** Multiresolution Knowledge Distillation for Anomaly Detection (2021, CVPR)
- **RD4AD:** Anomaly Detection via Reverse Distillation from One-Class Embedding (2022, CVPR)
- **AST:** Asymmetric Student-Teacher Networks for Industrial Anomaly Detection (2023, CVPR)

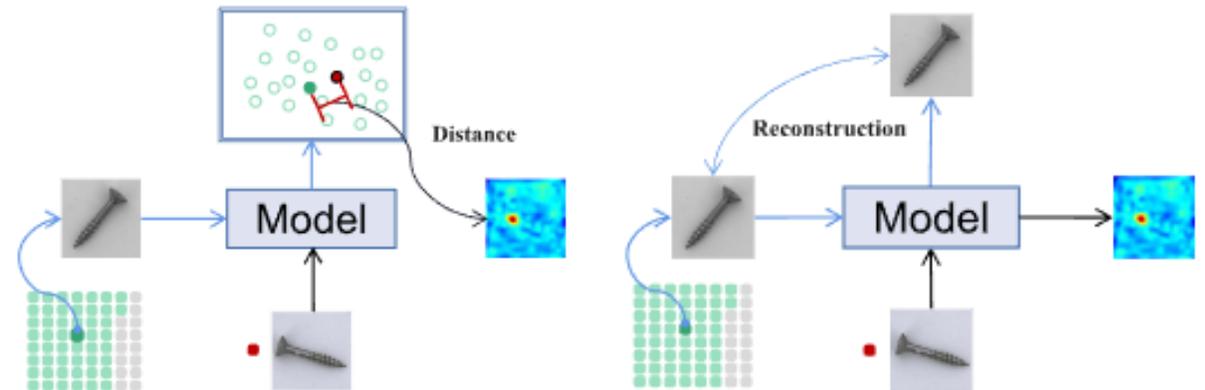
Background

❖ Industrial Image Anomaly Detection

- 산업 이미지는 대개 정상, 비정상 데이터의 불균형 → Unsupervised learning 적합
- 훈련데이터셋은 모두 정상 이미지로 구성, 평가데이터셋은 정상+비정상 이미지
- Unsupervised learning → Feature embedding based, Reconstruction based methods로 구분



[Unsupervised Anomaly Detection]



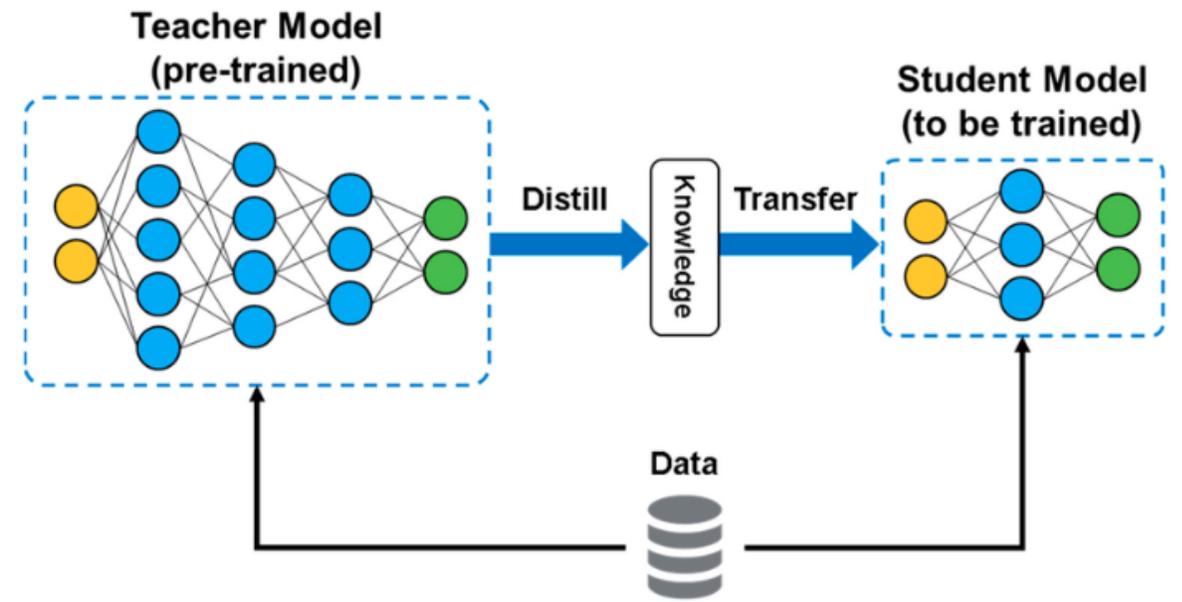
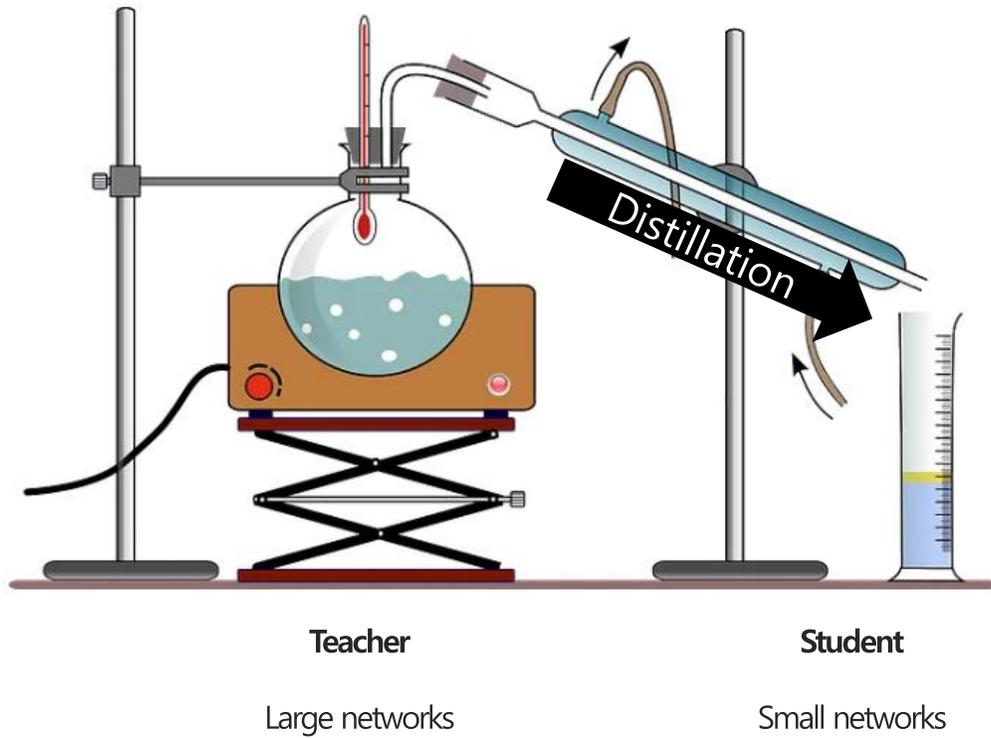
[Feature embedding based methods]

[Reconstruction based methods]

Background

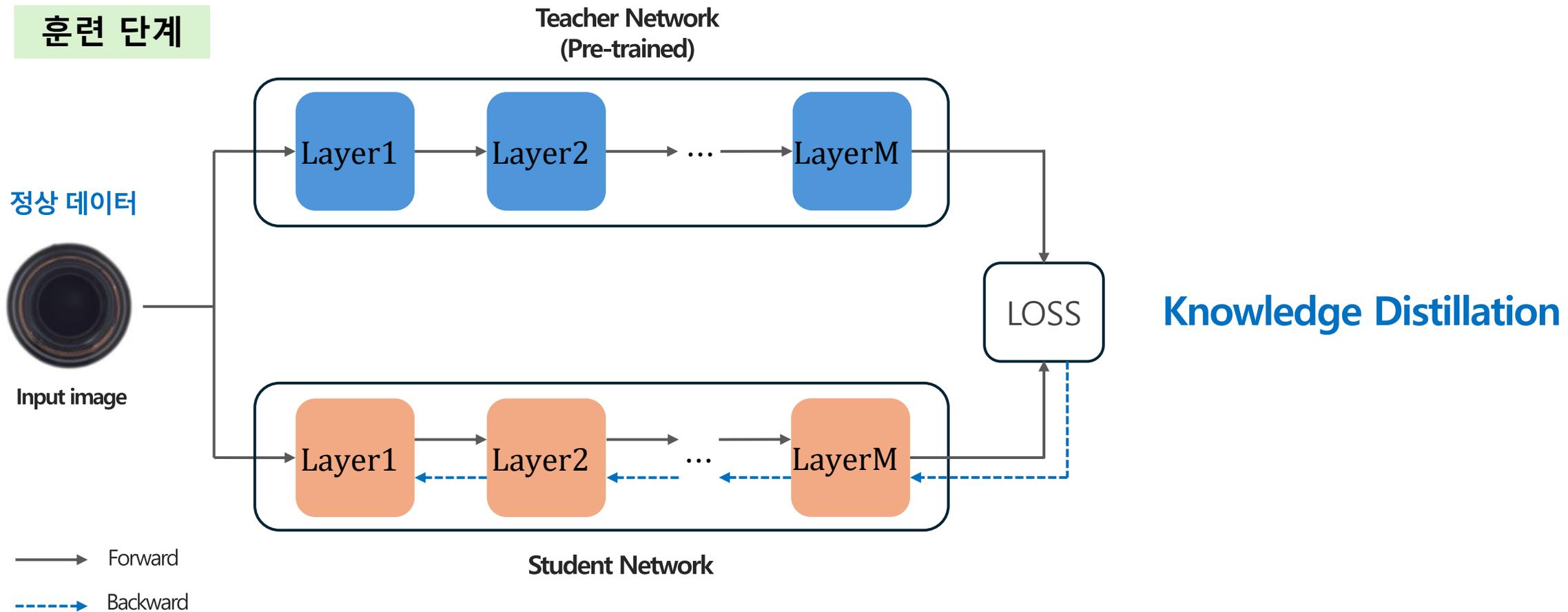
❖ Knowledge Distillation이란?

Distilling the Knowledge in a Neural Network (Geoffrey Hinton, 2015)



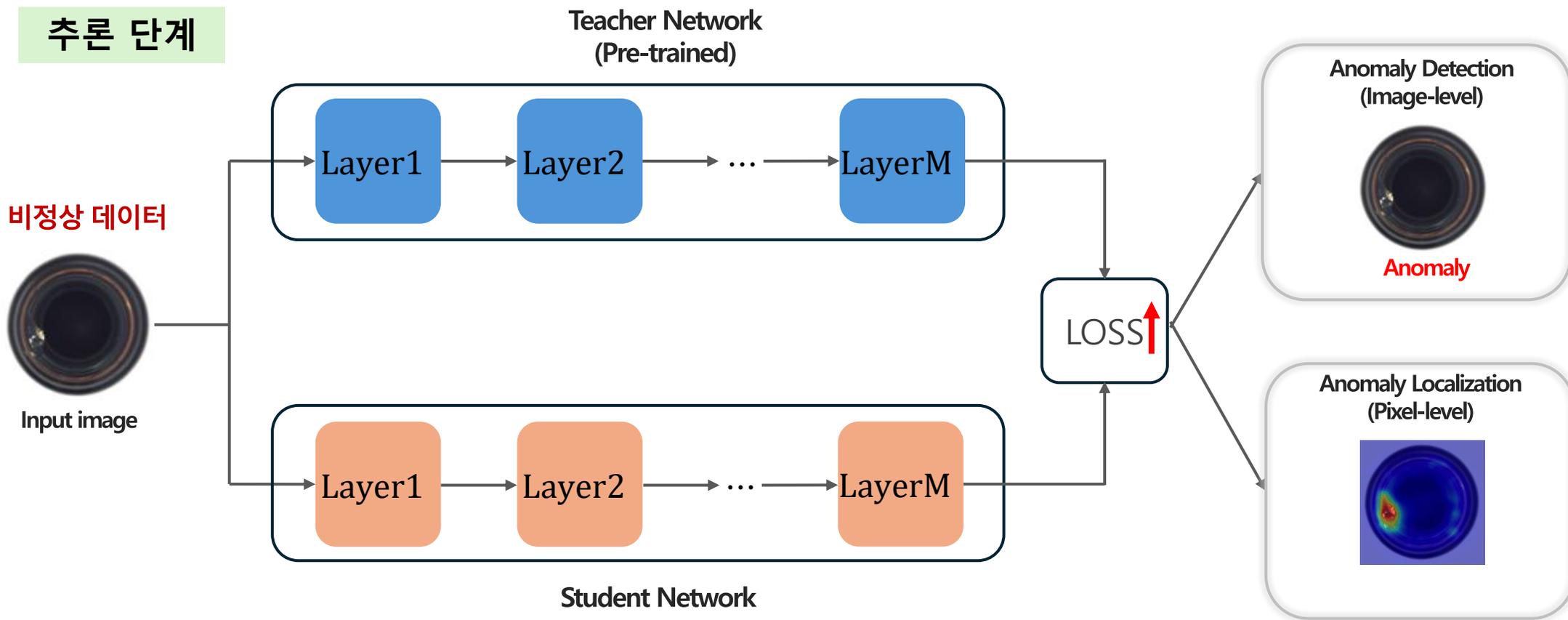
Background

❖ Knowledge Distillation based Image Anomaly Detection



Background

❖ Knowledge Distillation based Image Anomaly Detection



Related DMQA Open Seminar

중요

Introduction to Knowledge Distillation

2020.12.11

Data Mining & Quality Analytics Lab.
발표자 : 황하은

Introduction to knowledge distillation

발표자:  황하은

 2020년 12월 11일

 오후 1시 ~

 온라인 비디오 시청 (YouTube)

[세미나 정보 보기 →](#)

중요

Knowledge Distillation for Anomaly Detection

DMQA Open Seminar

2023.11.03

Data Mining & Quality Analytics Lab.
발표자 : 백민지

Knowledge Distillation for Anomaly Detection

발표자:  백민지

 2023년 11월 3일

 오후 12시 ~

 온라인 비디오 시청 (YouTube)

[세미나 정보 보기 →](#)

논문1. Uninformed Students

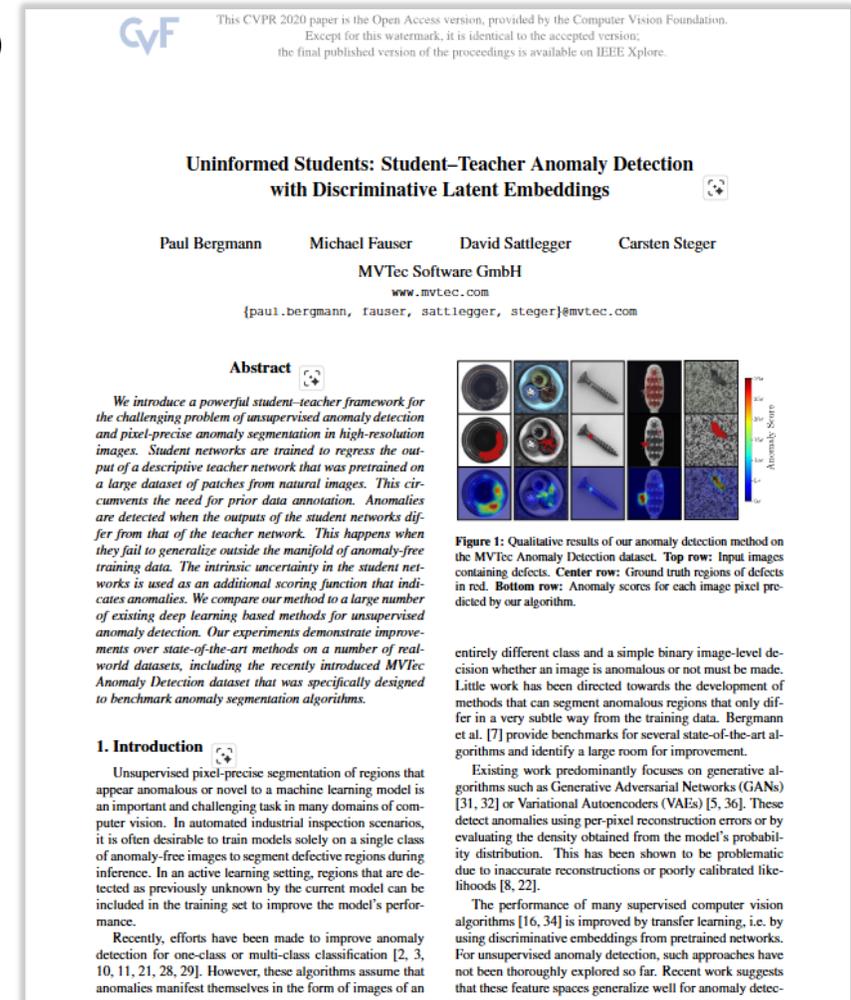
❖ Paper

√. 논문명 : Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings (2020, CVPR)

√. 저자 : Paul Bergmann Michael Fauser David Sattlegger Carsten Steger

MVTecAD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection (2019, CVPR)

√. 인용수 : 1257회 (2026.03.18 기준)



논문1. Uninformed Students

❖ Introduction

√. 기존 연구의 문제점

- **생성 모델(GAN, VAE)의 부정확성**: 픽셀 단위의 단순 재구성 오차에 의존하여 미세한 불량 탐지에 취약함
- 사전 학습된 네트워크와 **얕은 기계 학습** 결합 : 낮은 Capacity로 인한 고차원 데이터 처리에 제약(차원 축소와 서브샘플링 강제 → 정보 손실)

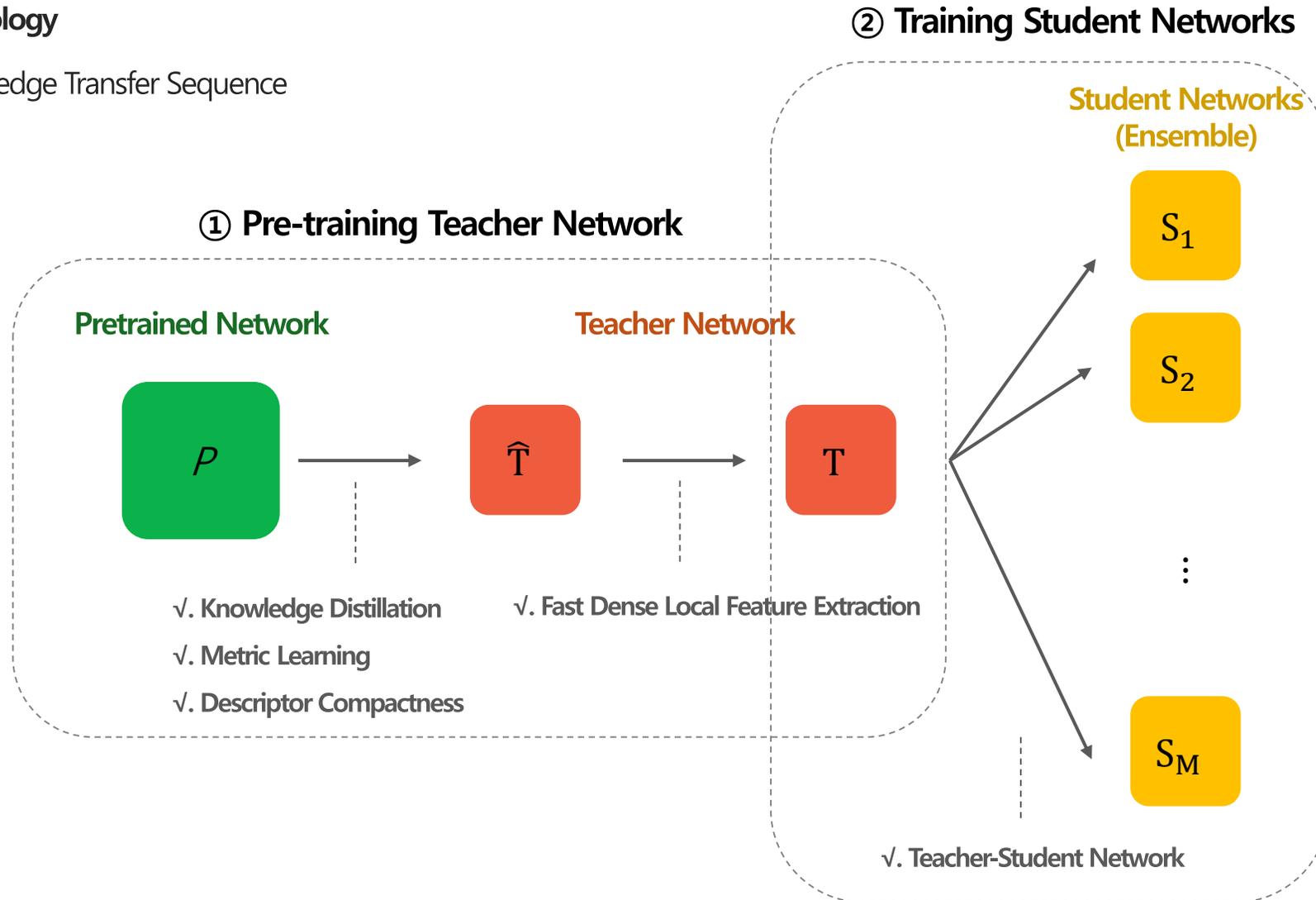
√. Contribution

- **Student-Teacher 프레임 워크**: 비지도 학습 기반의 고해상도 픽셀 단위 이상 분할(Segmentation) 구조 제안
- **효율적인 모방 학습**: 사전 학습된 Teacher의 Local descriptors를 대체 레이블로 활용하여 Student 앙상블 모델을 End-to-End로 학습
- **새로운 이상치 점수 산출**: Student 모델들의 예측 분산(Predictive variance)과 회귀 오차(Regression error)를 결합
- **다중 스케일 확장**: 다양한 크기의 수용 영역(Receptive fields)을 활용해 다양한 크기의 이상치를 탐지

논문1. Uninformed Students

❖ Methodology

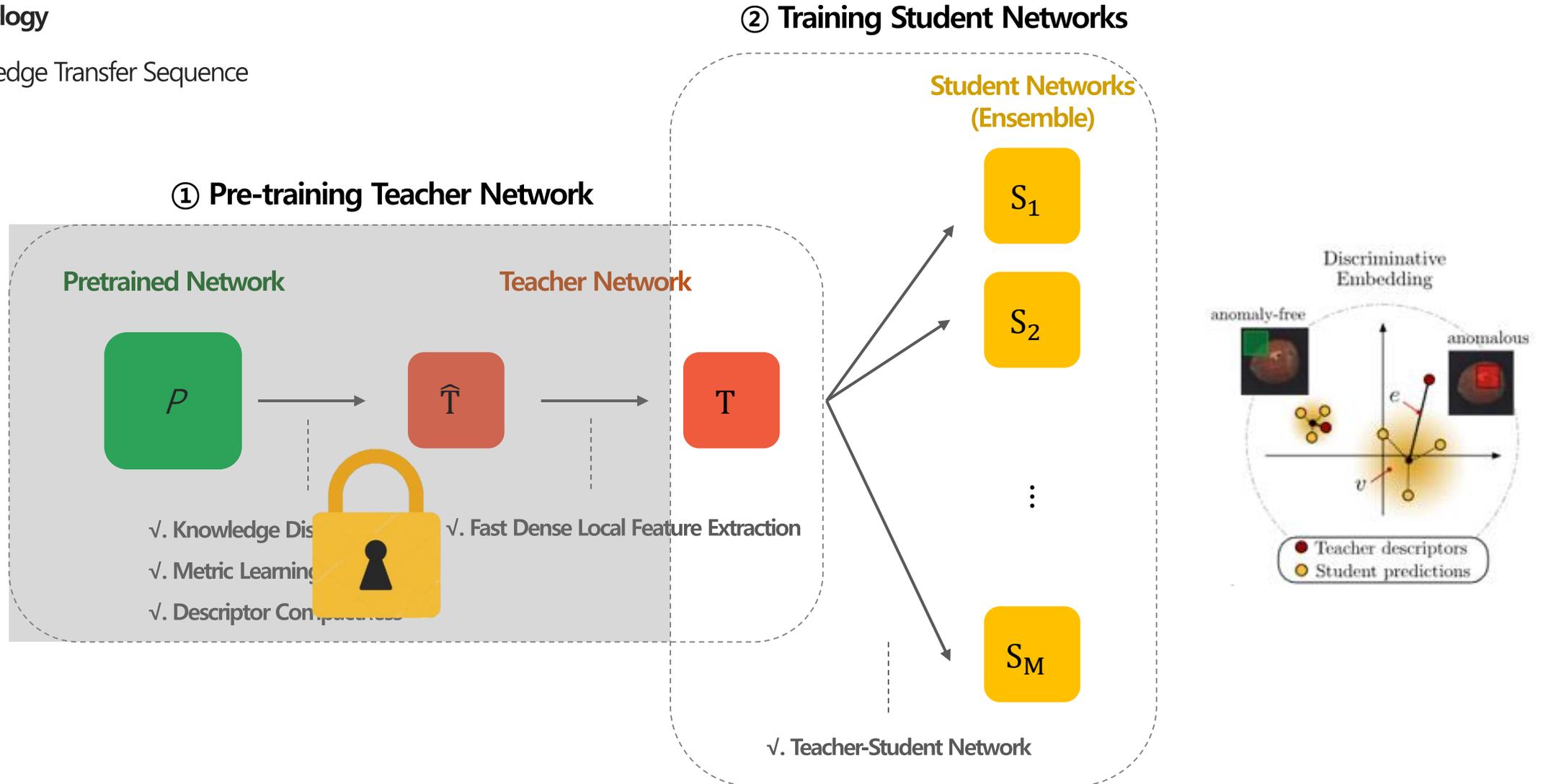
√. Knowledge Transfer Sequence



논문1. Uninformed Students

❖ Methodology

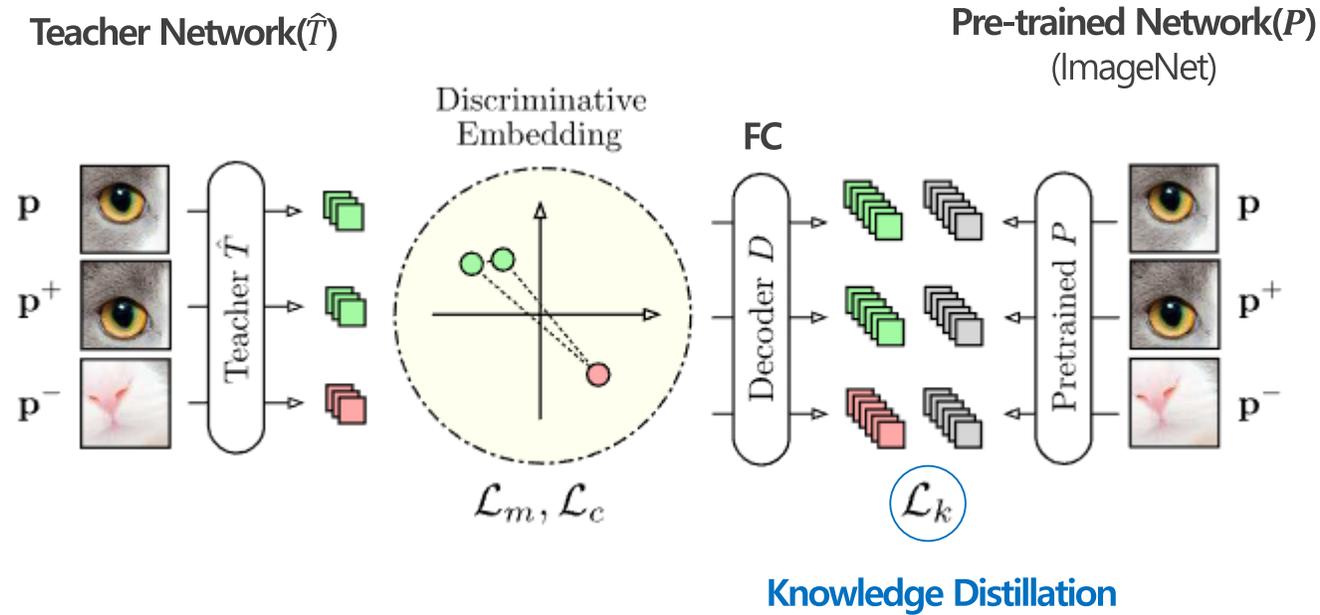
√. Knowledge Transfer Sequence



논문1. Uninformed Students

❖ Methodology

√. Pretraining of the teacher network



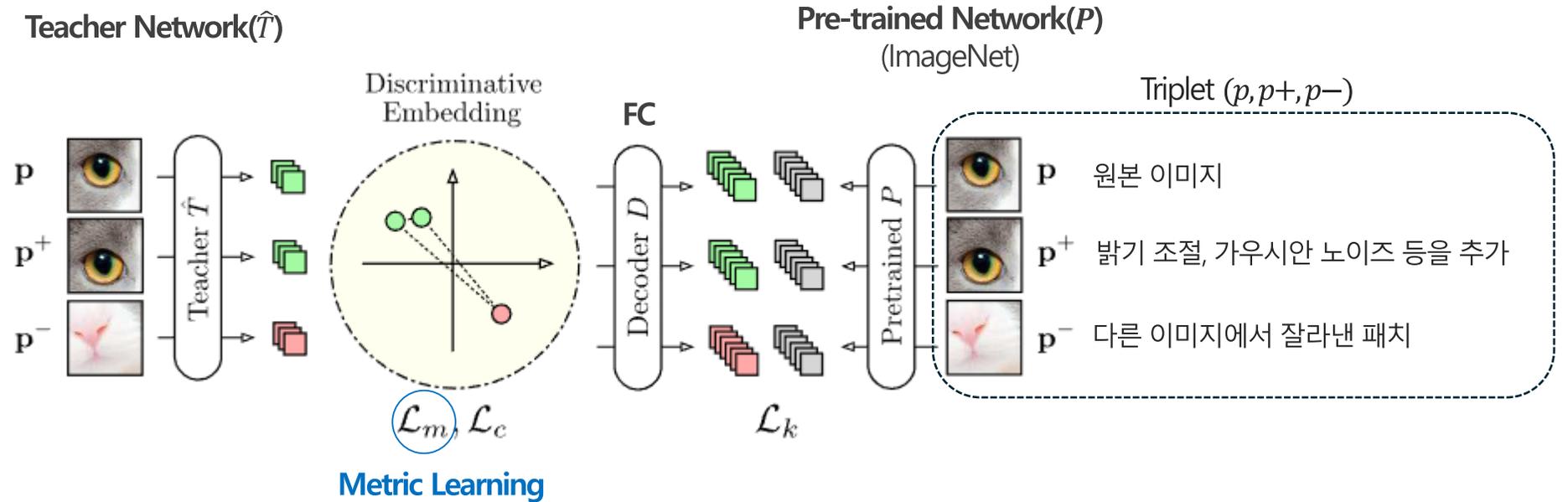
$$\mathcal{L}_k(\hat{T}) = |D(\hat{T}(p)) - P(p)|^2$$

사전 학습 네트워크 P 의 지식을 작고 효율적인 Teacher 네트워크 \hat{T} 로 전달

논문1. Uninformed Students

❖ Methodology

√. Pretraining of the teacher network



$$\mathcal{L}_m(\hat{T}) = \max\{0, \delta + \delta^+ - \delta^-\}$$

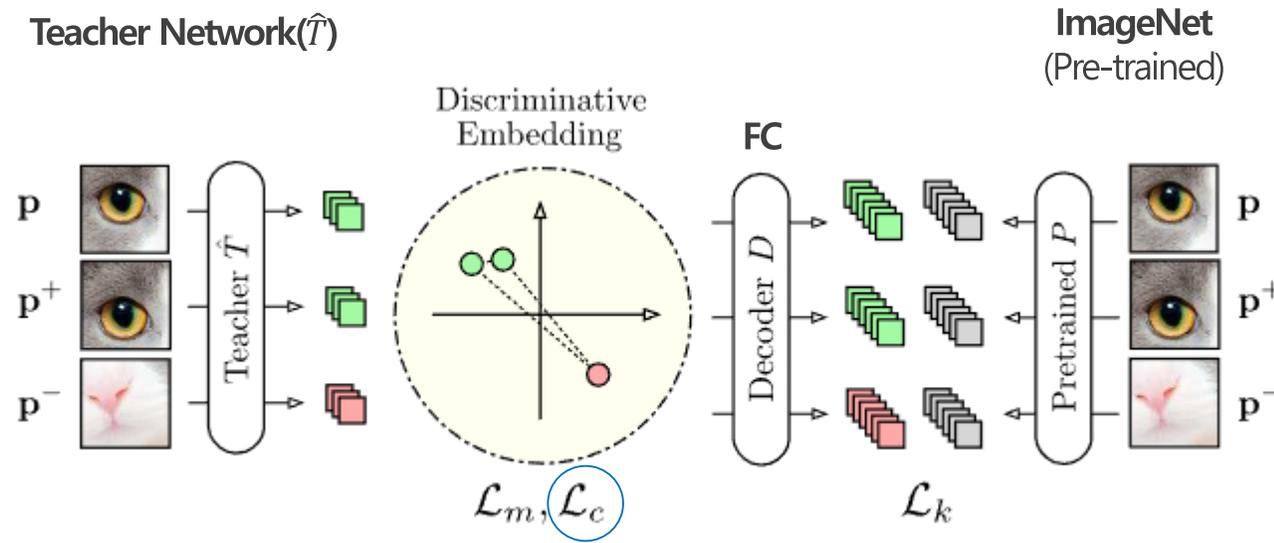
$\delta > 0$: Margin
 $\delta^+ = |\hat{T}(p) - \hat{T}(p^+)|^2$
 $\delta^- = \min\{|\hat{T}(p) - \hat{T}(p^-)|^2, |\hat{T}(p^+) - \hat{T}(p^-)|^2\}$

하나의 기준 패치(Anchor)를 두고, 이와 비슷한 패치는 가깝게, 다른 패치는 멀게 임베딩 공간에 배치하는 학습

논문1. Uninformed Students

❖ Methodology

√. Pretraining of the teacher network



Descriptor Compactness

$$\mathcal{L}_c(\hat{T}) = \sum_{i \neq j} c_{ij}$$

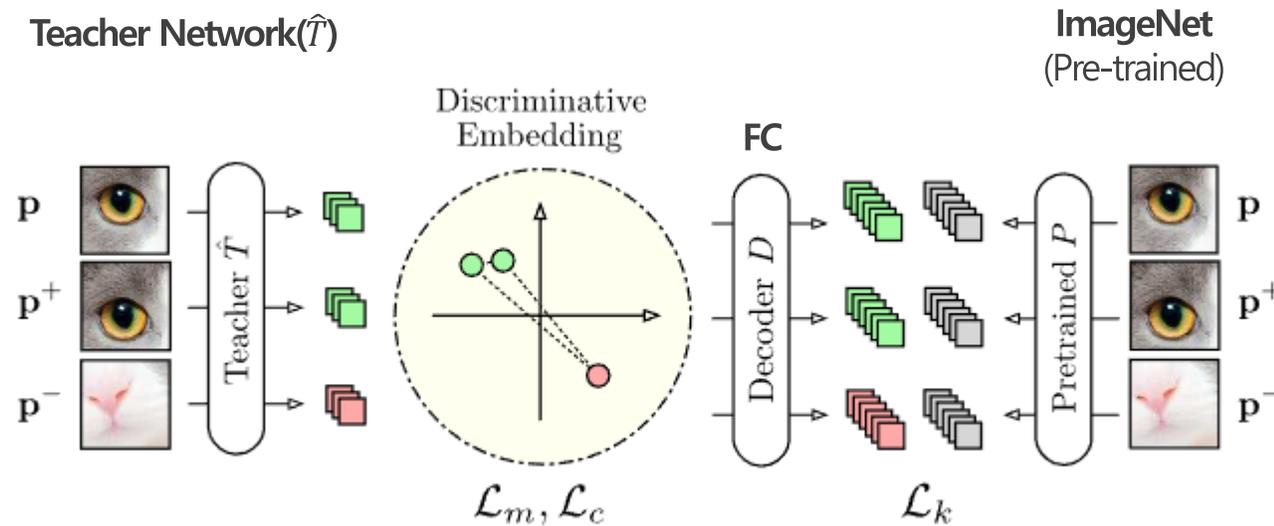
c_{ij} : 미니배치 내 모든 $\text{Descriptor}(\hat{T}(p))$ 간 Correlation Matrix의 원소

미니배치 내의 임베딩 벡터들 간의 상관관계를 최소화하여 불필요한 중복 정보를 제거

논문1. Uninformed Students

❖ Methodology

√. Pretraining of the teacher network



Final Training Loss for \hat{T}

$$\mathcal{L}(\hat{T}) = \lambda_k \mathcal{L}_k(\hat{T}) + \lambda_m \mathcal{L}_m(\hat{T}) + \lambda_c \mathcal{L}_c(\hat{T}),$$

논문1. Uninformed Students

❖ Methodology

√. Knowledge Transfer Sequence

② Training Student Networks

① Pre-training Teacher Network

Pretrained Network



- √. Knowledge Distillation
- √. Metric Learning
- √. Descriptor Compactness

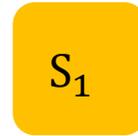


√. Fast Dense Local Feature Extraction

Teacher Network
(Pre-trained)



Student Networks
(Ensemble)



⋮



√. Teacher-Student Network

Teacher와 동일한 네트워크 구조

Training Loss for S_i

$$\mathcal{L}(S_i) = \frac{1}{wh} \sum_{(r,c)} \|\mu_{(r,c)}^{S_i} - (y_{(r,c)}^T - \mu) \text{diag}(\sigma)^{-1}\|_2^2$$

w, h : 입력 이미지의 너비와 높이

$\mu_{(r,c)}^{S_i}$: 픽셀 좌표 (r, c) 에 대한 i 번째 Student Network(S_i)의 예측 값

$y_{(r,c)}^T$: 픽셀 좌표 (r, c) 에 대해 Teacher Network(T)가 출력한 실제 특징 벡터

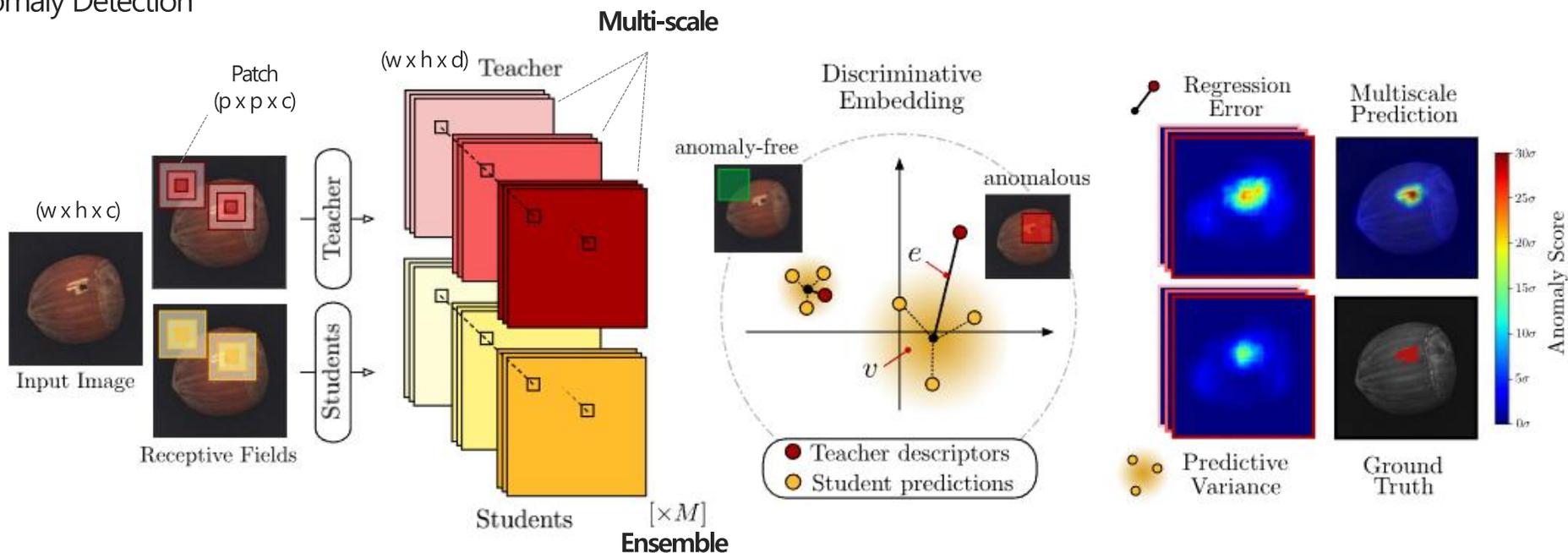
μ, σ : 학습 데이터 전체 특징 벡터의 평균과 표준편차

논문1. Uninformed Students

w, h : 입력 이미지의 너비와 높이
 $\mu_{(r,c)}^{S_i}$: 픽셀 좌표 (r,c) 에 대한 i 번째 Student Network(S_i)의 예측 값
 $y_{(r,c)}^T$: 픽셀 좌표 (r,c) 에 대해 Teacher Network(T)가 출력한 실제 특징 벡터
 μ, σ : 학습 데이터 전체 특징 벡터의 평균과 표준편차

❖ Methodology

√. Anomaly Detection



① 회귀 오차(Regression Error)

M개 Student의 예측 평균이 Teacher의 실제 출력과 얼마나 차이 나는지를 측정

$$\begin{aligned}
 e_{(r,c)} &= |\mu_{(r,c)} - (y_{(r,c)}^T - \mu) \text{diag}(\sigma)^{-1}|_2^2 \\
 &= \left| \frac{1}{M} \sum_{i=1}^M \mu_{(r,c)}^{S_i} - (y_{(r,c)}^T - \mu) \text{diag}(\sigma)^{-1} \right|_2^2
 \end{aligned}$$

② 예측 불확실성(Predictive Variance)

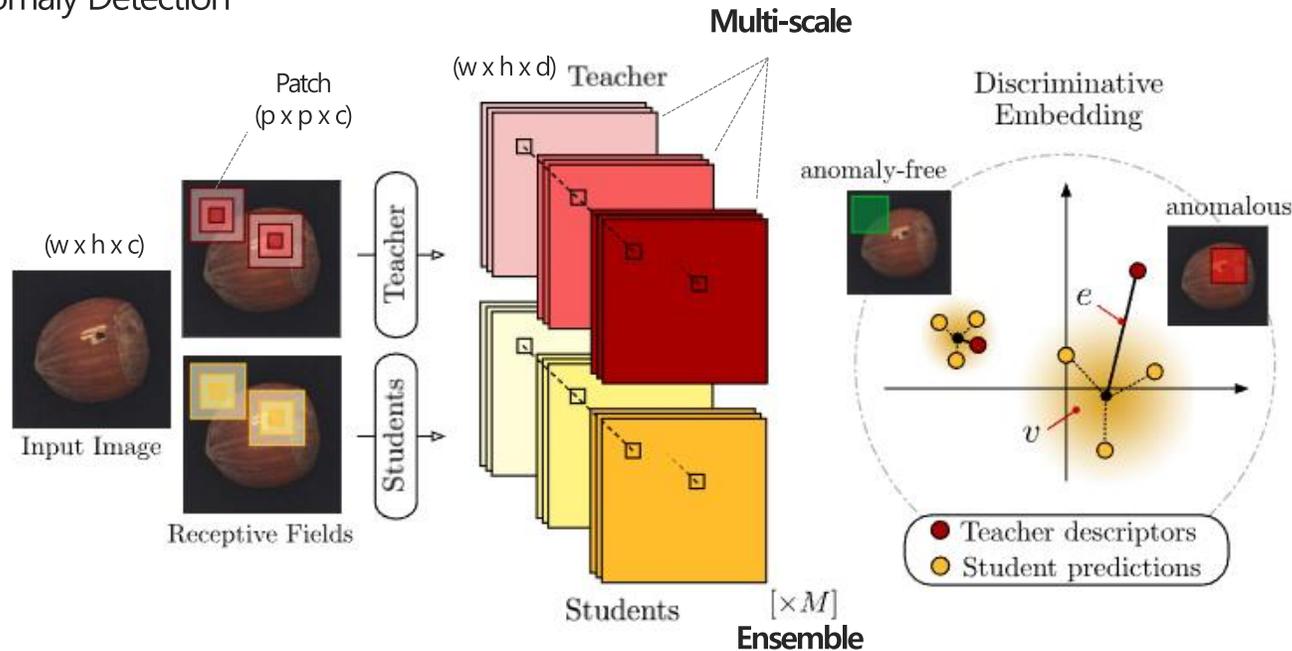
M개의 Student들의 예측 값들이 서로 얼마나 차이가 있는지(분산)을 측정

$$v_{(r,c)} = \frac{1}{M} \sum_{i=1}^M |\mu_{(r,c)}^{S_i}|_2^2 - |\mu_{(r,c)}|_2^2$$

논문1. Uninformed Students

❖ Methodology

√. Anomaly Detection



Final Anomaly Score

$$\tilde{e}(r, c) + \tilde{v}(r, c) = \frac{e(r, c) - e_{\mu}}{e_{\sigma}} + \frac{v(r, c) - v_{\mu}}{v_{\sigma}}$$

From a validation set of anomaly-free images

w, h : 입력 이미지의 너비와 높이

$\mu_{(r,c)}^{S_i}$: 픽셀 좌표 (r, c) 에 대한 i 번째 Student Network(S_i)의 예측 값

$y_{(r,c)}^T$: 픽셀 좌표 (r, c) 에 대해 Teacher Network(T)가 출력한 실제 특징 벡터

μ, σ : 학습 데이터 전체 특징 벡터의 평균과 표준편차

Multi-scale Anomaly Segmentation

$$\frac{1}{L} \sum_{l=1}^L (\tilde{e}_{(r,c)}^{(l)} + \tilde{v}_{(r,c)}^{(l)})$$

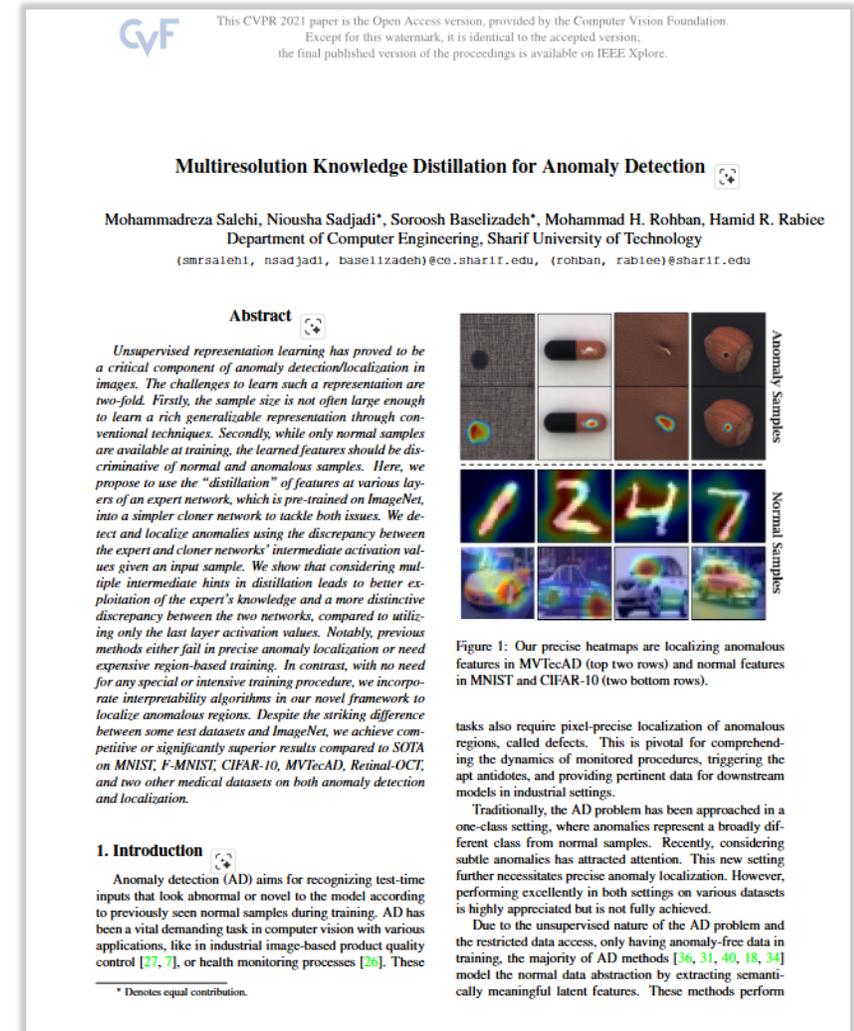
논문2. MKD: Multiresolution Knowledge Distillation for Anomaly Detection

❖ Paper

√. 논문명 : Multiresolution Knowledge Distillation for Anomaly Detection (2021, CVPR)

√. 저자 : Mohammadreza Salehi, Niousha Sadjadi*, Soroosh Baselizadeh*, Mohammad H. Rohban, Hamid R. Rabiee

√. 인용수 : 830회 (2026.03.18 기준)



논문2. MKD: Multiresolution Knowledge Distillation for Anomaly Detection

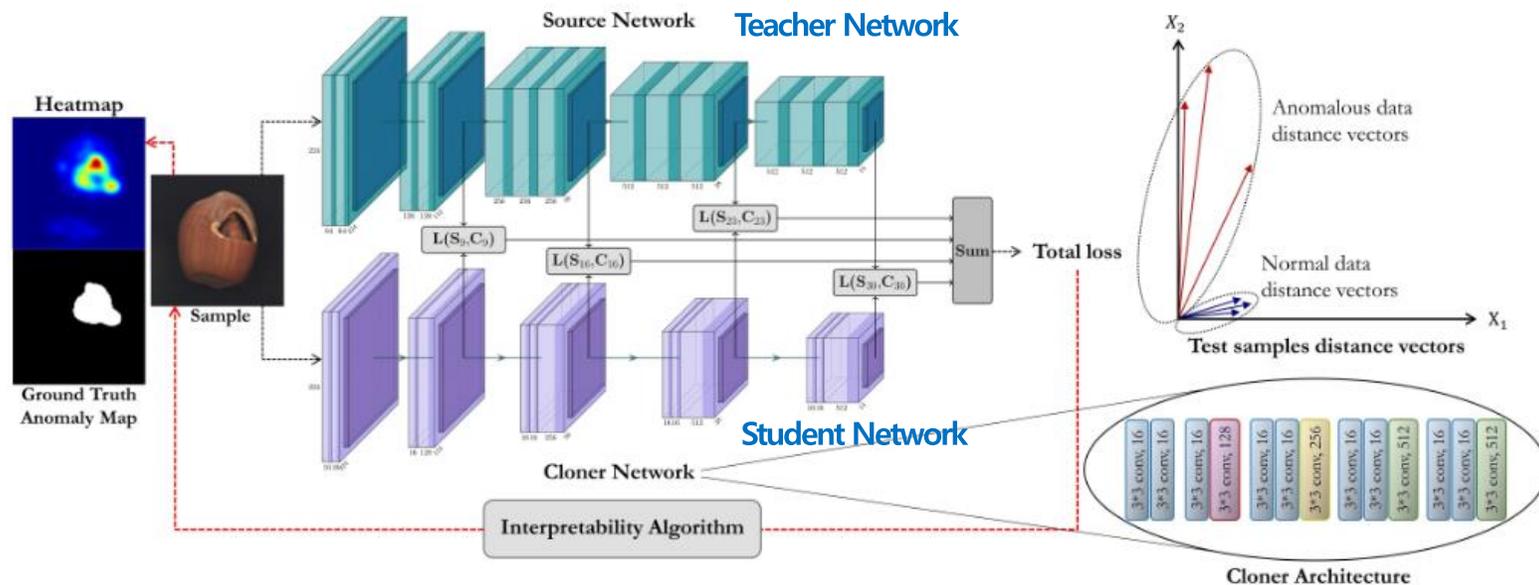
❖ Introduction

√. 기존 연구의 문제점

- **일반화 문제 (Generality Problem):** 단일 클래스 분류(One-class)나 미세 결함 탐지(Defect localization) 중 어느 한쪽에만 치중
- **불완전한 지식 전달:** Teacher 네트워크의 '마지막 레이어'만 모방할 경우, 풍부한 지식을 온전히 활용하지 못하는 문제

√. Contribution

- **다중 해상도 지식 증류 (Multiresolution KD):** Teacher 네트워크의 여러 중간 레이어의 지식을 Student 네트워크로 증류
- **구조적 압축 (Compact Cloner):** Expert(Teacher)보다 단순하고 크기가 작은 Cloner(Student) 네트워크를 설계



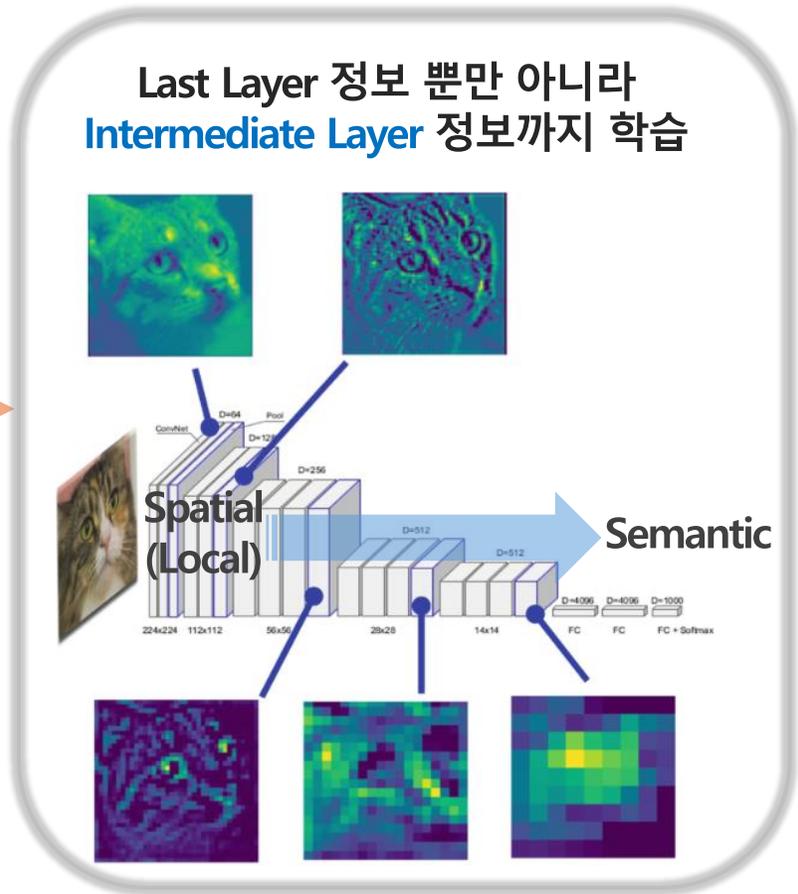
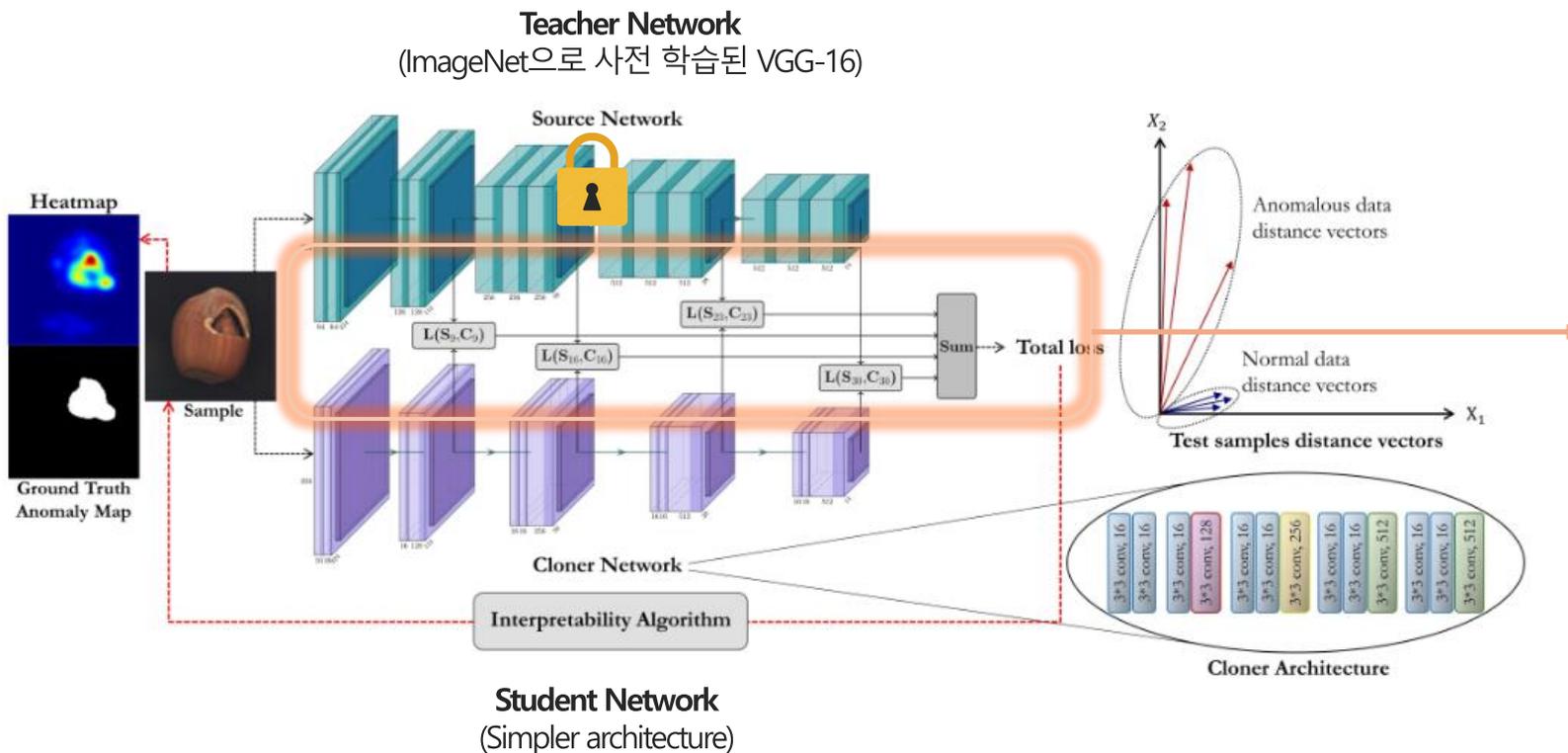
[MKD 프레임워크]

논문2. MKD: Multiresolution Knowledge Distillation for Anomaly Detection

❖ Methodology

√. 학습 과정

- **Teacher Network:** ImageNet으로 사전 학습된 VGG-16 (Frozen)
- **Student Network:** Teacher Network 여러 중간 레이어에서 추출되는 특징을 학습 (Local 정보부터 Semantic 정보까지)

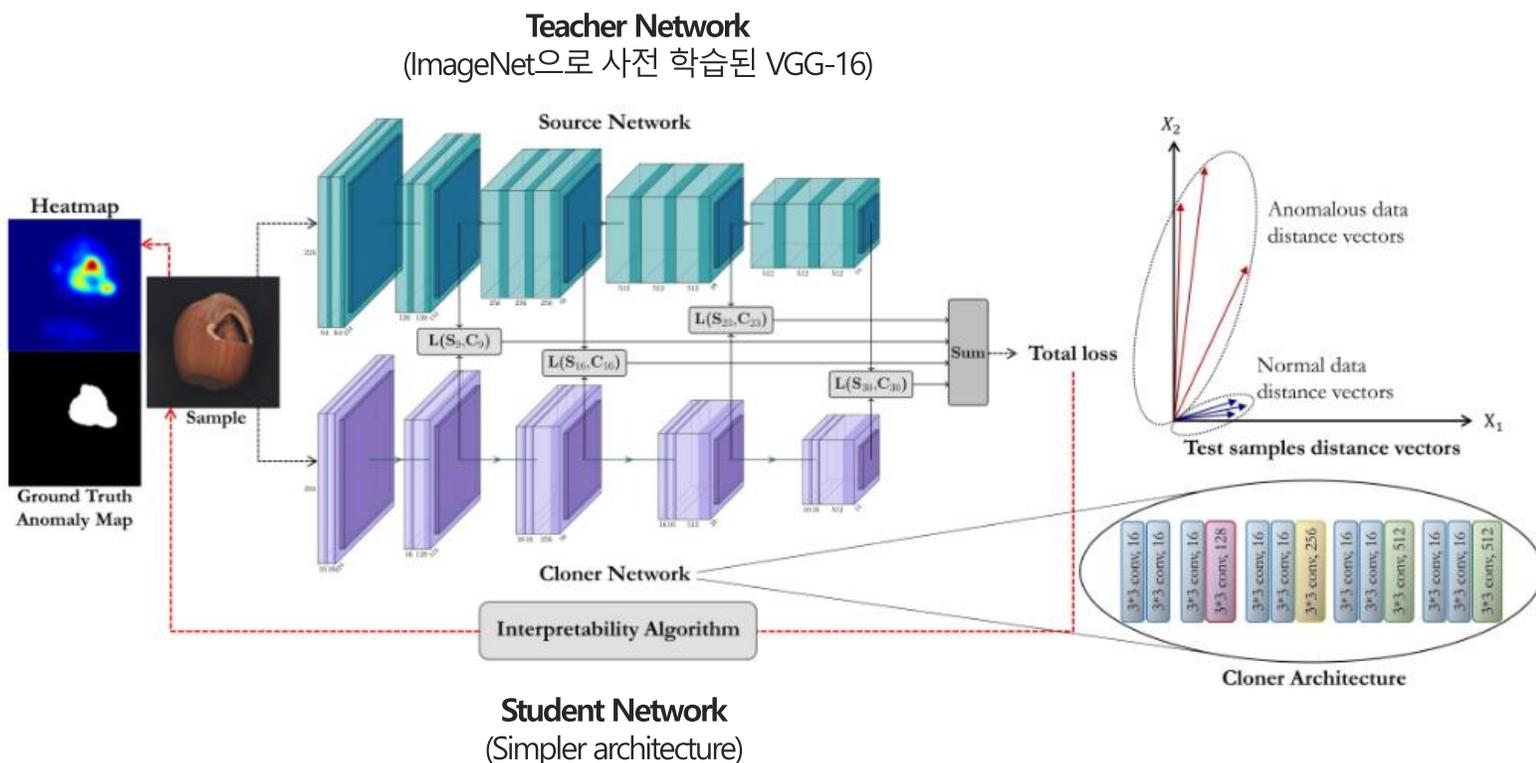


논문2. MKD: Multiresolution Knowledge Distillation for Anomaly Detection

❖ Methodology

√. 학습 과정

- **Teacher Network:** ImageNet으로 사전 학습된 VGG-16 (Frozen)
- **Student Network:** Teacher Network 여러 중간 레이어에서 추출되는 특징을 학습 (Local 정보부터 Semantic 정보까지)



CP_i : i -th Critical layer in the networks
 $a_s^{CP_i}$: the Source activation values of that Critical layer
 $a_c^{CP_i}$: the Cloner activation values of that Critical layer
 N_{CP} : the total number of critical layer
 N_i : the number of neurons in layer CP_i

Knowledge = Value + Direction

① 유클리드 거리 최소화 (거리가 가깝도록)

$$\mathcal{L}_{val} = \sum_{i=1}^{N_{CP}} \frac{1}{N_i} \sum_{j=1}^{N_i} (a_s^{CP_i}(j) - a_c^{CP_i}(j))^2$$

② 코사인 유사도 최대화 (방향이 같도록)

$$\mathcal{L}_{dir} = \sum_i 1 - \frac{\text{vec}(a_s^{CP_i})^T \cdot \text{vec}(a_c^{CP_i})}{|\text{vec}(a_s^{CP_i})| |\text{vec}(a_c^{CP_i})|}$$

$$\mathcal{L}_{total} = \mathcal{L}_{val} + \lambda \mathcal{L}_{dir}$$

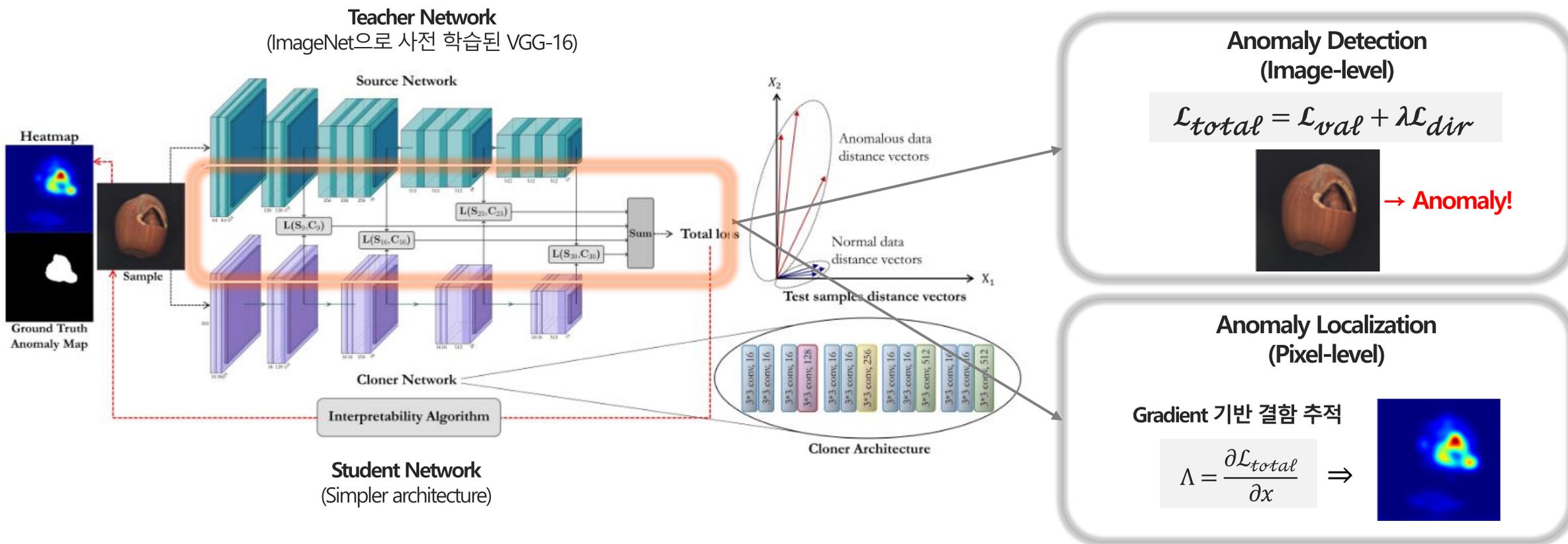
논문2. MKD: Multiresolution Knowledge Distillation for Anomaly Detection

❖ Methodology

√. 추론 과정(Anomaly Detection & Localization)

- **Anomaly Detection (Image-level):** Teacher와 Student의 출력 간 \mathcal{L}_{total} 값을 특정 임계값(Threshold)과 비교하여 판별
- **Anomaly Localization (Pixel-level):** Gradient 기반 결함 추적 & 노이즈 제거 및 Heatmap 정제 (Noise Reduction)

CP_i : i -th Critical layer in the networks
 $a_s^{CP_i}$: the Source activation values of that Critical layer
 $a_c^{CP_i}$: the Cloner activation values of that Critical layer
 N_{CP} : the total number of critical layer
 N_i : the number of neurons in layer CP_i



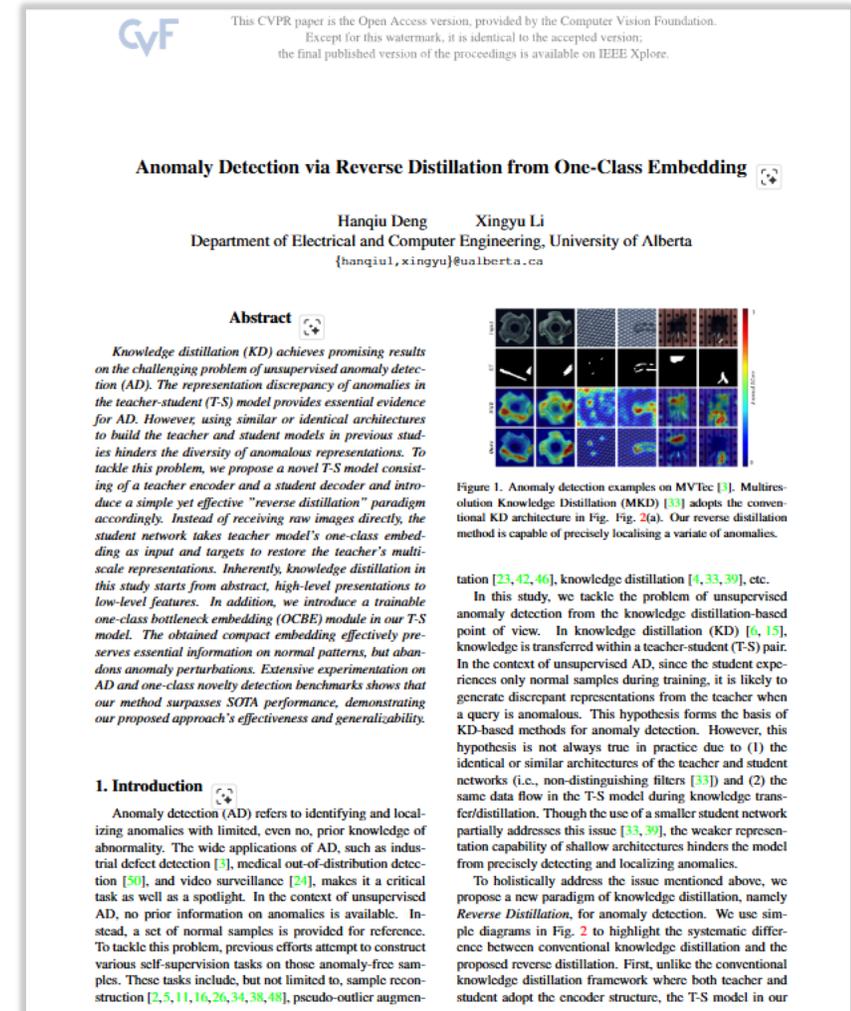
논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

❖ Paper

√. 논문명 : Anomaly Detection via Reverse Distillation from One-Class Embedding (2022, CVPR)

√. 저자 : Hanqiu Deng, Xingyu Li

√. 인용수 : 1117회 (2026.03.18 기준)



논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

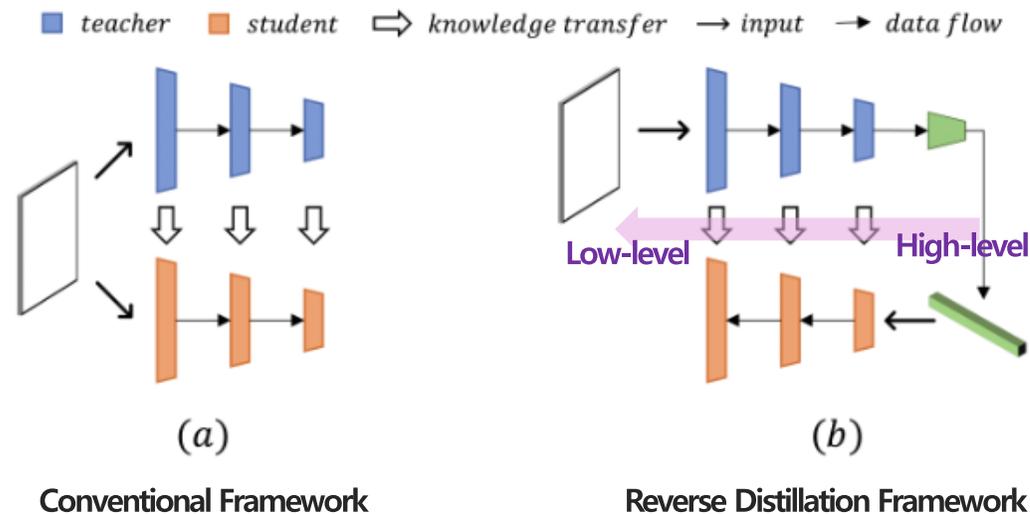
❖ Introduction

√. 기존 연구의 문제점

- **동일하거나 유사한 T-S 네트워크 구조:** 비정상 데이터에 대해서도 두 모델이 비슷하게 특징을 추출해버리는 문제 (Non-distinguishing filters)
- **T-S 두 모델의 동일한 데이터 플로우:** 두 네트워크 모두 원본 이미지를 직접 입력 받기 때문에 비정상 특징에 대한 표현의 다양성이 저해

√. Contribution

- **이질적인 T-S 네트워크 구조:** Teacher는 인코더, Student는 디코더 구조 → 기존의 구조적 유사성으로 인한 문제를 근본적으로 해결
- **역방향 데이터 흐름:** Student는 오직 Teacher의 압축된 임베딩만을 입력으로 받음 → 고차원(추상적) 지식에서 저차원 특징 순으로 지식이 전달
- **OCBE(One-Class Bottleneck Embedding) 도입:** Teacher의 고차원 특징을 압축 → 정상 패턴 정보는 보존, 비정상적인 교란(Perturbation)은 차단

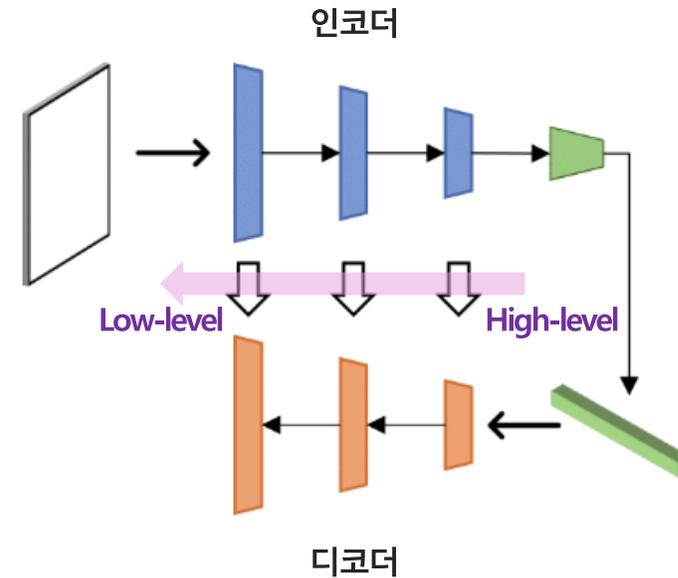
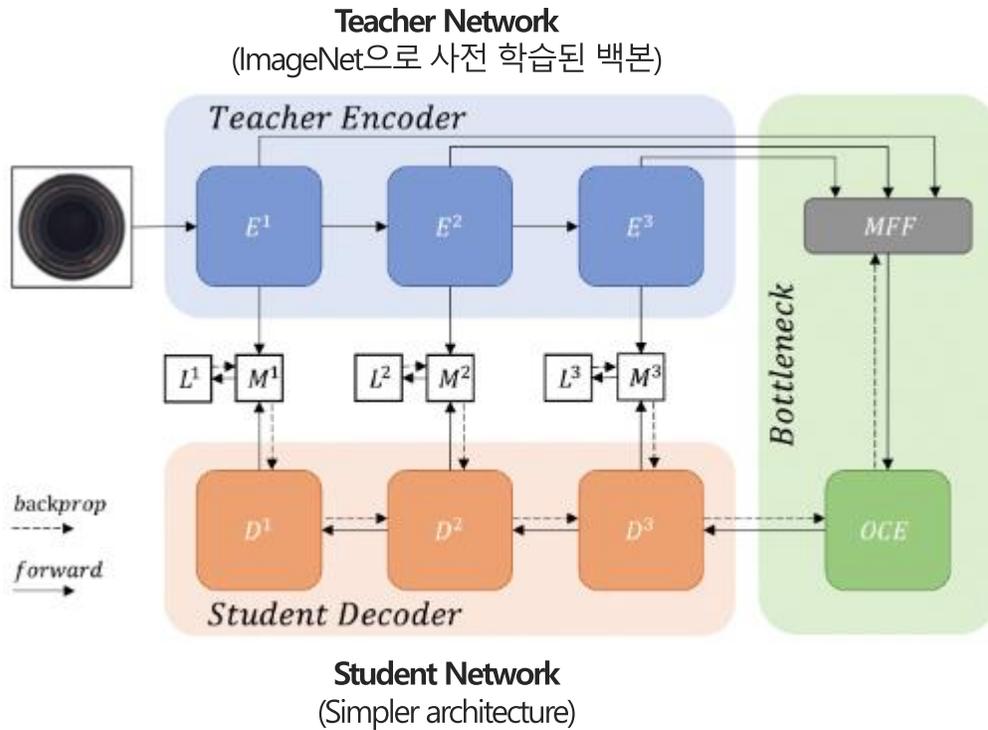


논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

❖ Methodology

√. 전체 프레임워크

- **Teacher 네트워크(인코더 역할)** : 사전 훈련된 WideResNet50 모델을 백본으로 사용 (Frozen)
- **Student 네트워크(디코더 역할)** : Teacher와 대칭적이지만 역전된(symmetrical but reversed) 아키텍처 구조
- **OCBE(One-Class Bottleneck Embedding)** : Multi-scale Feature 정보를 융합하는 **MFF** 블록과 응축을 통해 정보 병목(Information Bottleneck)을 만드는 **OCE** 블록으로 구성

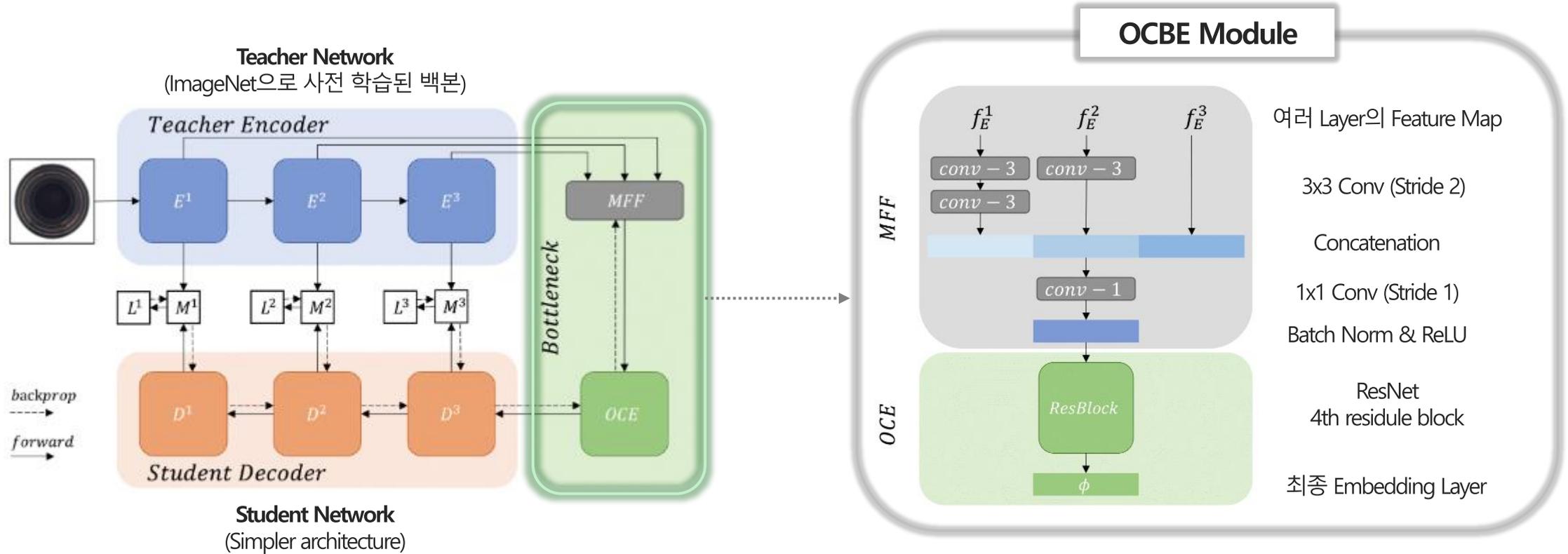


논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

❖ Methodology

√. 전체 프레임워크

- **Teacher 네트워크(인코더 역할)** : 사전 훈련된 WideResNet50 모델을 백본으로 사용 (Frozen)
- **Student 네트워크(디코더 역할)** : Teacher와 대칭적이지만 역전된(symmetrical but reversed) 아키텍처 구조
- **OCBE(One-Class Bottleneck Embedding)** : Multi-scale Feature 정보를 융합하는 **MFF 블록**과 응축을 통해 정보 병목(Information Bottleneck)을 만드는 **OCE 블록**으로 구성

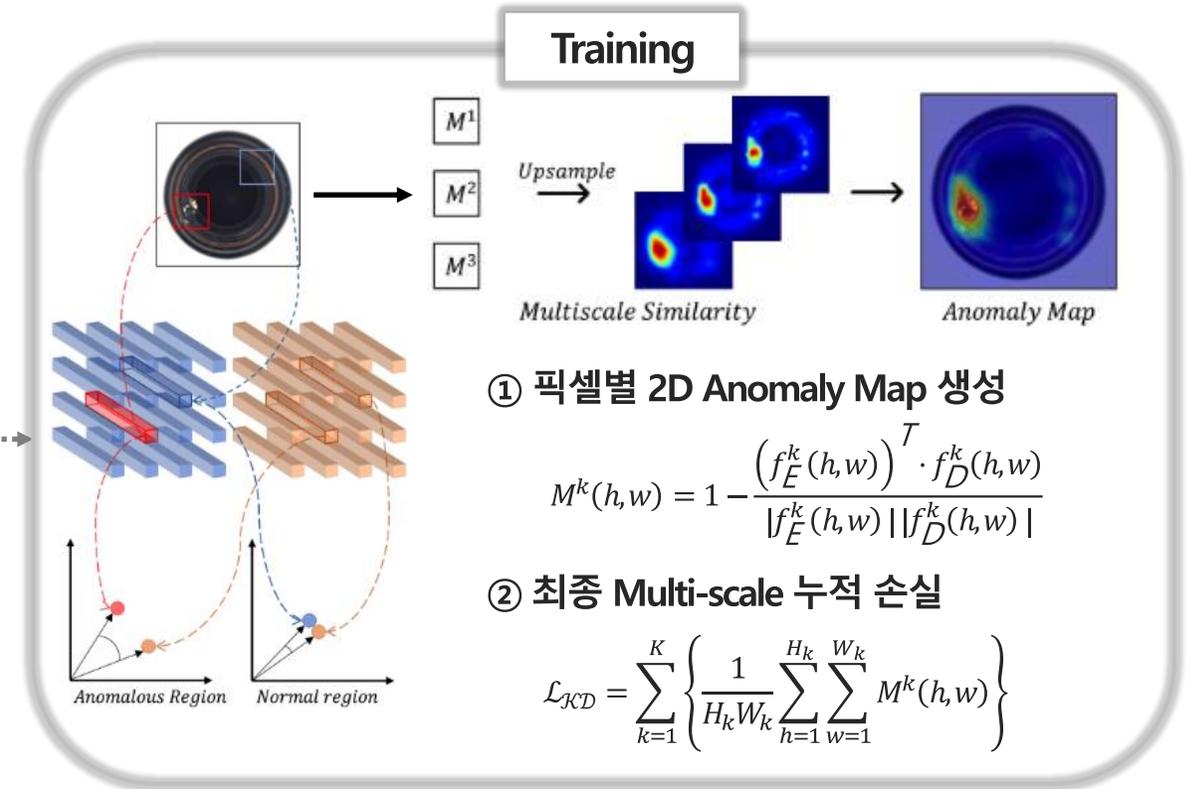
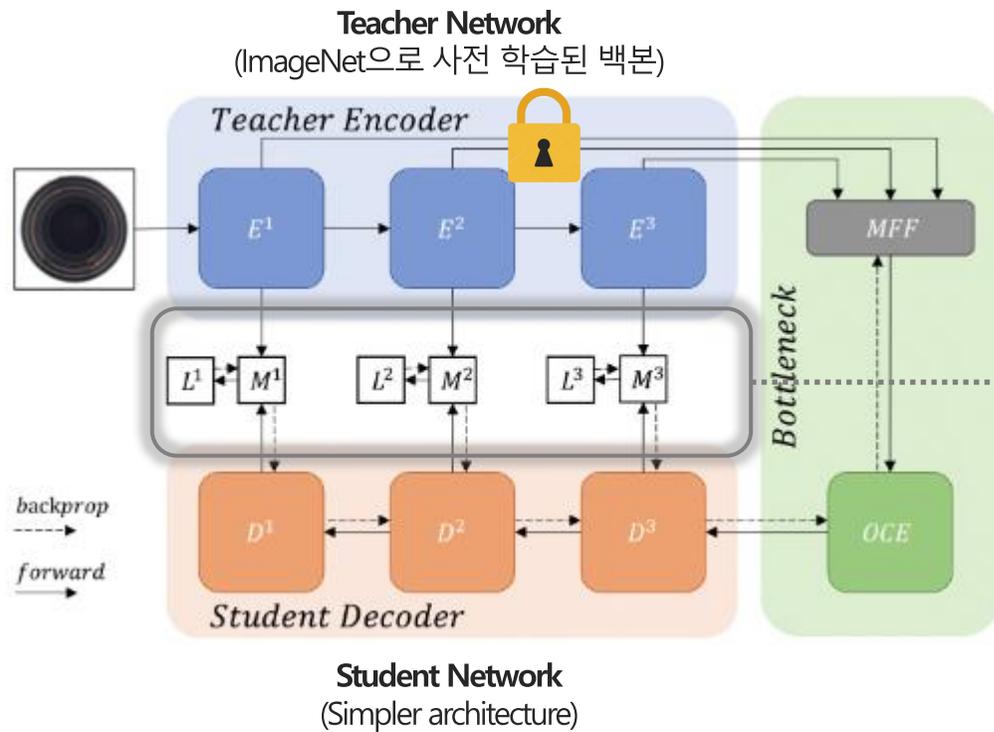


논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

❖ Methodology

√. 학습 과정

- **Teacher 네트워크(인코더 역할)**: 입력 이미지로부터 Multi-scale Feature map 추출 (Frozen)
- **Student 네트워크(디코더 역할)**: Teacher의 출력 정보가 압축된 임베딩을 입력 받아 Teacher의 Feature map을 역방향으로 복원
- **OCBE(One-Class Bottleneck Embedding)**: 정보 병목을 통해 비정상 교란(Perturbation)은 필터링하고 정상 패턴만 Student로 전달

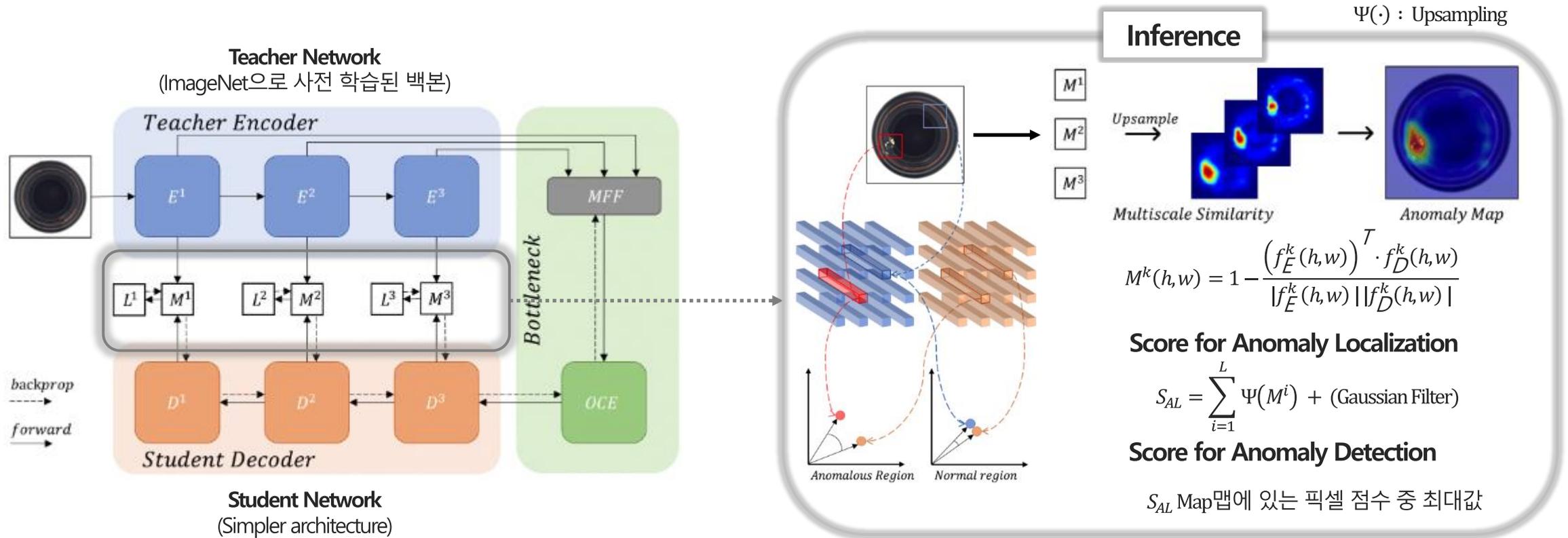


논문3. RD4AD: Anomaly Detection via Reverse Distillation from One-Class Embedding

❖ Methodology

√. 추론 과정

- **Anomaly Detection (Image-level):** 생성된 최종 픽셀 이상치 맵(S_{AL})의 점수 중 최댓값을 추출하여 임계값(Threshold)과 비교 및 판별
- **Anomaly Localization (Pixel-level):** Multi-scale Feature map 간의 코사인 유사도 비교 & 업샘플링 및 가우시안 필터를 통한 최종 Heatmap 정제



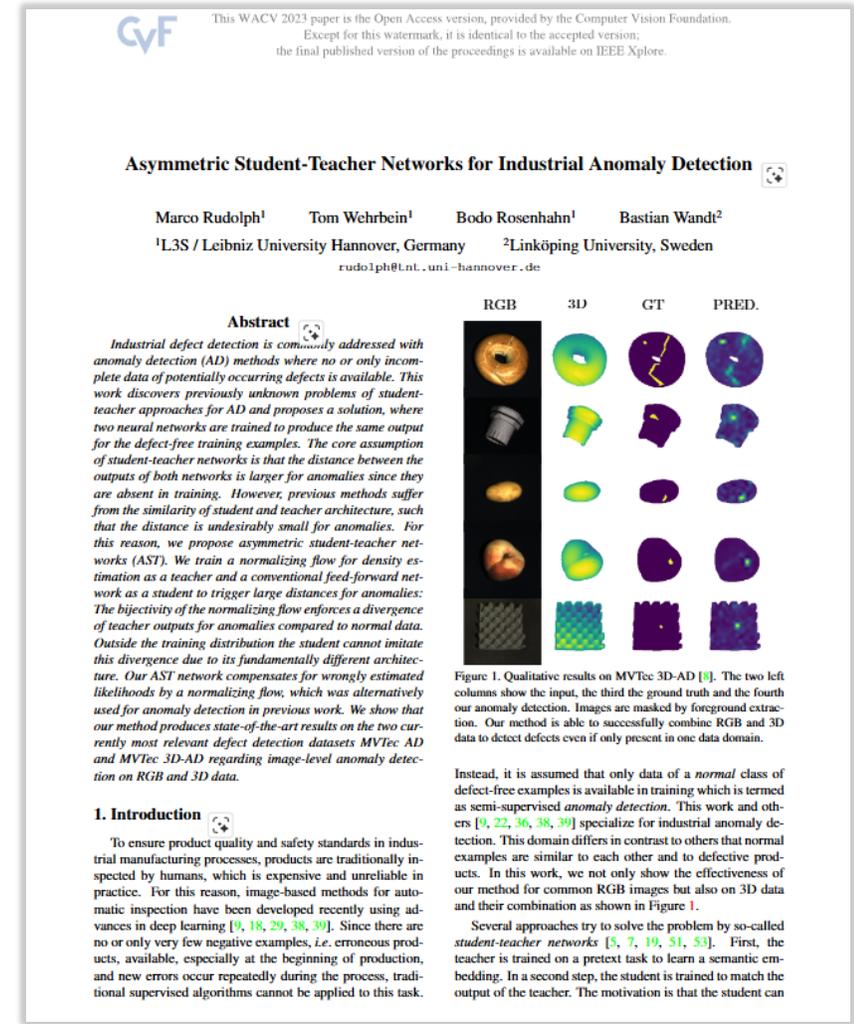
논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Paper

√. 논문명 : Asymmetric Student-Teacher Networks for Industrial Anomaly Detection (2023, CVPR)

√. 저자 : Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, Bastian Wandt,

√. 인용수 : 343회 (2026.03.18 기준)



논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

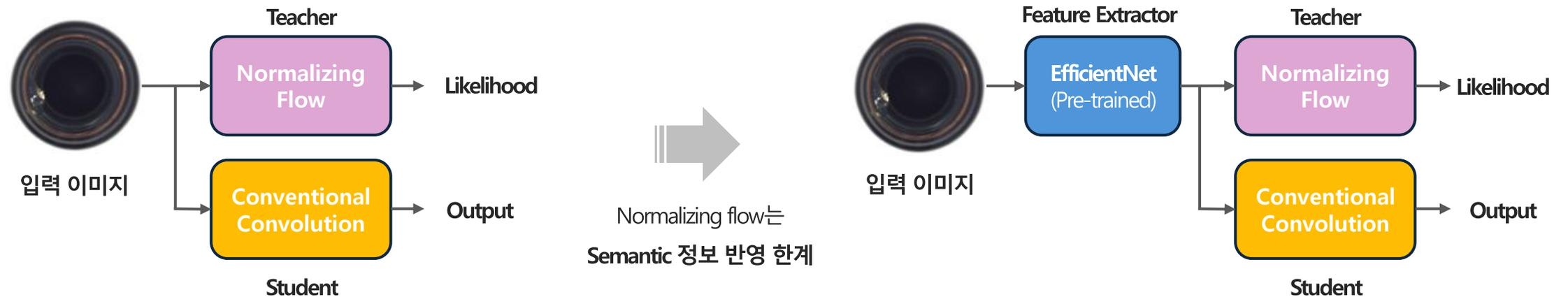
❖ Introduction

√. 기존 연구의 문제점

- ① 동일하거나 유사한 T-S 네트워크 구조: 비정상 데이터에 대해서도 두 모델이 비슷하게 특징을 추출해버리는 문제 (Undesired generalization)

√. Contribution

- ① 이질적인 T-S 네트워크 구조: Teacher는 Normalizing flow, Student는 Conventional Convolutional Neural Networks



논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

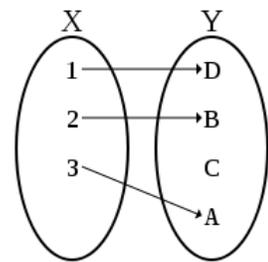
❖ Introduction

√. 기존 연구의 문제점

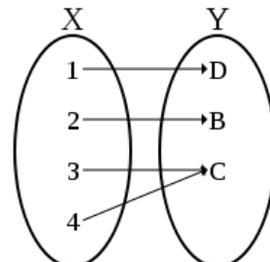
- ① 동일하거나 유사한 T-S 네트워크 구조: 비정상 데이터에 대해서도 두 모델이 비슷하게 특징을 추출해버리는 문제 (Undesired generalization)
- ② 기존 신경망의 비단사성(Non-injectivity): OOD(Out-of-distribution) 데이터 입력이 T-S 출력 사이의 큰 차이를 보장하지 못하는 문제

√. Contribution

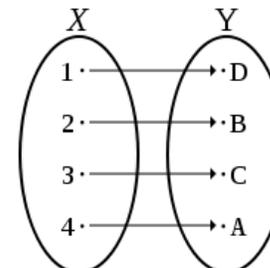
- ① 이질적인 T-S 네트워크 구조: Teacher는 Normalizing flow, Student는 Conventional Convolutional Neural Networks
- ② Normalizing flow의 전단사(Bijective) 특성: OOD 데이터 입력 시 Teacher의 출력이 크게 발산하도록 강제



단사 함수
(injective function)
 $|X| \leq |Y|$



전사 함수
(surjective function)
 $|X| \geq |Y|$



전단사 함수
(bijective function)
 $|X| = |Y|$

$|\cdot|$: 집합의 크기

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Introduction

√. 기존 연구의 문제점

- ① 동일하거나 유사한 T-S 네트워크 구조: 비정상 데이터에 대해서도 두 모델이 비슷하게 특징을 추출해버리는 문제 (Undesired generalization)
- ② 기존 신경망의 비단사성(Non-injectivity): OOD(Out-of-distribution) 데이터 입력이 T-S 출력 사이의 큰 차이를 보장하지 못하는 문제
- ③ Density Estimation의 한계: 아무리 완벽한 밀도 추정 모델이라도 정확한 이상 탐지를 보장하지 못하는 문제 → Likelihood 자체를 이상 점수로 활용하기 어려움

Perfect Density Models Cannot Guarantee Anomaly Detection (MDPI, 2021)

√. Contribution

- ① 이질적인 T-S 네트워크 구조: Teacher는 Normalizing flow, Student는 Conventional Convolutional Neural Networks
- ② Normalizing flow의 전단사(Bijective) 특성: OOD 데이터 입력 시 Teacher의 출력이 크게 발산하도록 강제
- ③ T-S Distance 사용: Teacher가 출력하는 Likelihood에 의존하지 않고 T-S 출력 Distance를 이상 점수로 활용 → Likelihood misestimation 보상

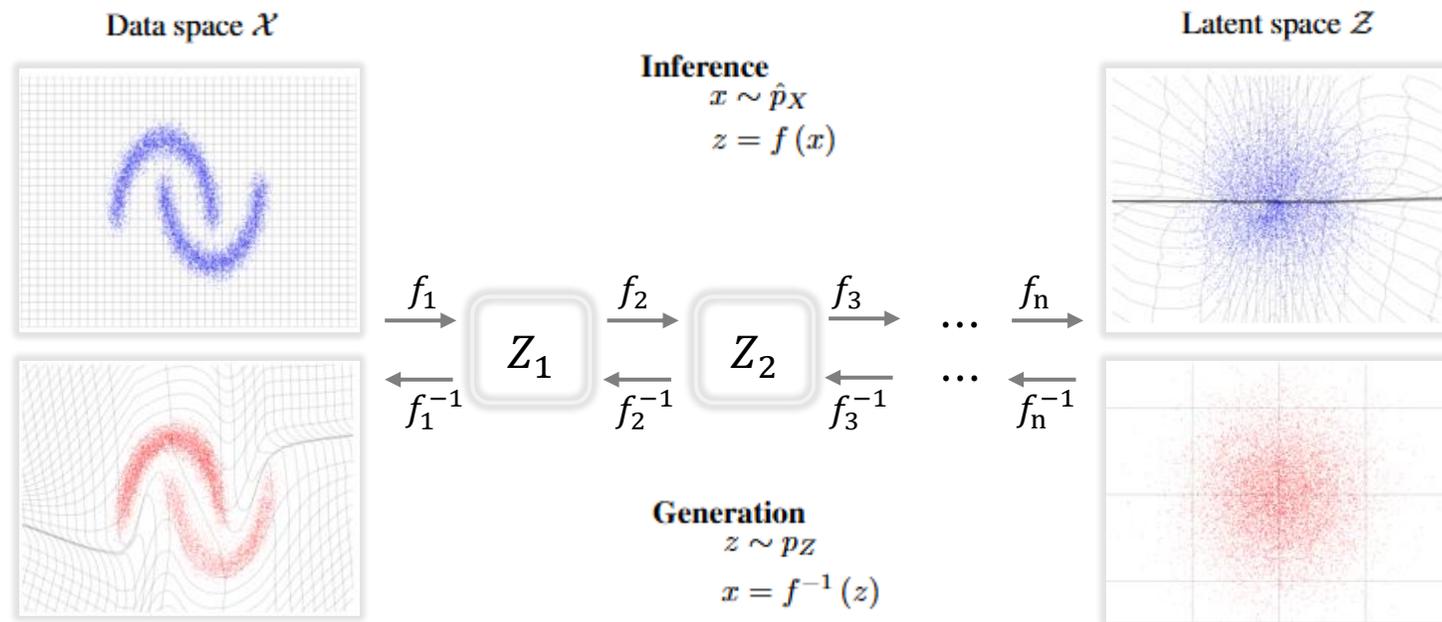


논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Background

√. Normalizing Flow

- Deep Generative Model
- Density estimation → 입력 데이터에 대한 Likelihood를 직접적(**Explicit**)으로 계산 가능(**Tractable**)
- 변수 변환 공식(Change of variables)에 기반한 방법론 → 복잡한 분포 $p_X(x)$ 를 단순한 분포 $p_Z(z)$ 로 **전단사(bijective)** 매핑



출처: DENSITY ESTIMATION USING REAL NVP (ICLR, 2017)

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Background

√. Normalizing Flow

- Deep Generative Model
- Density estimation → 입력 데이터에 대한 Likelihood를 직접적(**Explicit**)으로 계산 가능(**Tractable**)
- 변수 변환 공식(Change of variables)에 기반한 방법론 → 복잡한 분포 $p_X(x)$ 를 단순한 분포 $p_Z(z)$ 로 **전단사(bijective)** 매핑

$$\int p_X(x)dx = \int p_Z(z)dz = 1$$

$$\Leftrightarrow p_X(x)dx = p_Z(z)dz$$

x 가 $f(\cdot)$ 를 거쳐 z 공간으로 변환 시, 부피의 **변화율(Volume Term)**

1차원 수식

$$\Leftrightarrow p_X(x) = p_Z(z) \left| \frac{dz}{dx} \right|$$

$$dz(\text{다차원 부피}) = \left| \det \left(\frac{\partial f(x)}{\partial x^T} \right) \right| dx(\text{다차원 부피})$$

⇓ **다차원으로 확장**

다차원 수식

$$p_X(x) = p_Z(z) \left| \det \left(\frac{\partial z}{\partial x^T} \right) \right|$$

$z = f(x)$

⇔

Jacobian Matrix

$$J = \frac{\partial z}{\partial x^T} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{bmatrix}$$

Maximizing Log-Likelihood Estimation

$$\log(p_X(x)) = \log(p_Z(f(x))) + \log \left(\left| \det \left(\frac{\partial f(x)}{\partial x^T} \right) \right| \right)$$

Log-Likelihood

Normal Distribution

Volume Term

위 목적 함수를 최대화되는 방향으로 학습 진행

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Background

√. Normalizing Flow

- 고차원 데이터에 대해서 Log-Likelihood 계산을 효율적으로 만들어주는 변환 함수(f) 디자인이 핵심
- 변환 함수가 만족해야 하는 두 가지 조건
 - 가역성 (역함수가 존재) = 전단사(Bijective) 성질
 - 효율적인 야코비안 행렬식(Jacobian Determinant) 계산

→ **Coupling layer**를 통해 연산 비용을 획기적으로 단축

여러 Affine Coupling layer 아키텍처 중 Real NVP 채택
Density estimation using real nvp (ICLR, 2017)

① Affine Transformation

입력 벡터 x 를 Channel에 따라 두 부분 분리
출력 벡터 y 역시 동일한 차원의 두 부분으로 분리

$$x = [x_{1:d}, x_{d+1:D}] = [x_1, x_2]$$

$$y = [y_{1:d}, y_{d+1:D}] = [y_1, y_2]$$

Affine Coupling Layer

$$y_1 = x_1$$

$$y_2 = [x_2 \odot \exp(s(x_1)) + t(x_2)]$$

\odot : Element-wise multiplication
 $s(\cdot)$: Scaling을 담당하는 신경망
 $t(\cdot)$: Translation을 담당하는 신경망

② 야코비안 행렬(Jacobian Matrix) 구성

$$J = \frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} I & 0 \\ \frac{\partial y_2}{\partial x_1} & \text{diag}(\exp(s(x_1))) \end{bmatrix}$$

하삼각 행렬
(Lower Triangular Matrix)

※ 이 부분이 신경망 s 와 t 의 복잡한 미분 값을 포함하나 **계산 불필요!**

$$\frac{\partial y_2}{\partial x_1} = \frac{\partial}{\partial x_1} (x_2 \odot \exp(s(x_1)) + t(x_2))$$

③ 행렬식(Determinant) 계산

$$\det(J) = \det(I) \times \det(\text{diag}(\exp(s(x_1))))$$

$$\det(J) = \prod_i \exp(s(x_1)_i)$$

최종 로그 행렬식(Log-Determinant) 도출

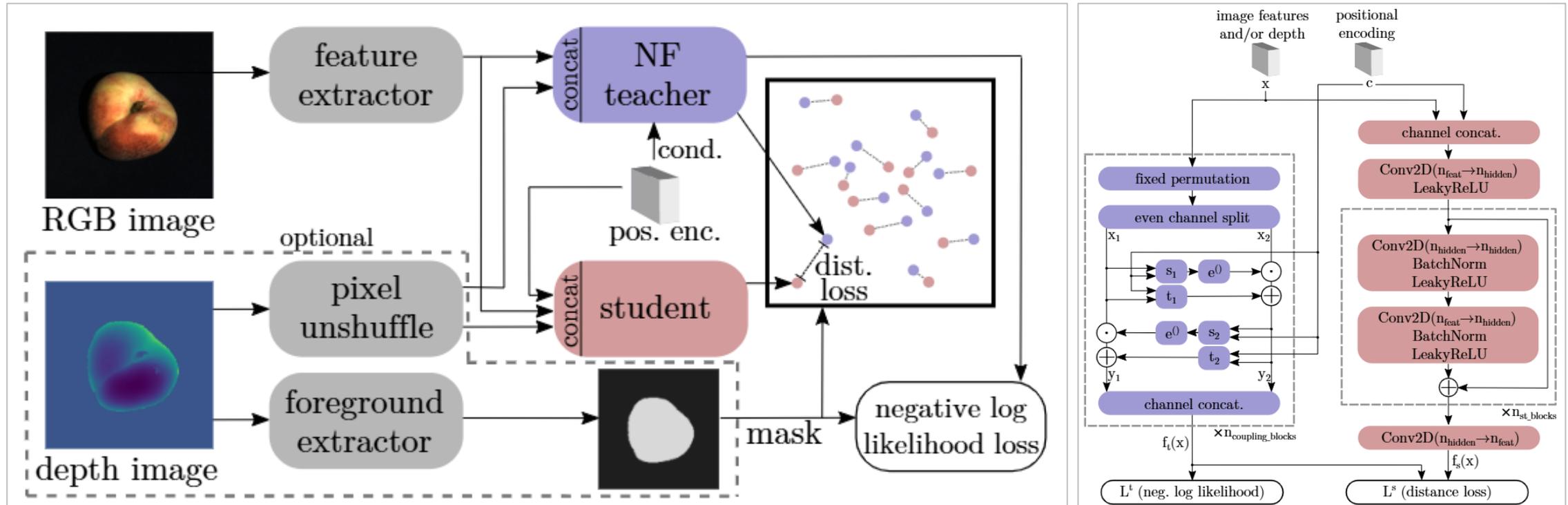
$$\log|\det(J)| = \log\left(\prod_i \exp(s(x_1)_i)\right) = \sum_i s(x_1)_i$$

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Methodology

√. 전체 프레임워크

- 이질적인 T-S 네트워크 구조: Teacher는 Normalizing flow, Student는 Conventional Convolutional Neural Networks
- 2단계 학습 진행(Teacher→Student 순): 정상 데이터로 Teacher 모델 선학습 후 (Frozen), Teacher 모델의 출력 값을 모방하도록 Student 학습
- 3D 이미지도 학습 가능: 깊이 정보가 담긴 depth image 제공 시(선택)



논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Methodology

√. Teacher 학습

- **Teacher Network 구조:** Real NVP 기반 다중 Affine Coupling 블록으로 구성된 Normalizing flow (4개의 Coupling block)
- **학습 목표:** 후속 단계 Student Network를 학습시키기 위한 기준점(Target)을 생성하는 사전 작업(**Pretext Task**)
- **Density Estimation:** 정상 이미지를 Feature Extractor에 통과 시켜 얻은 Feature map을 입력으로 받아, 입력 데이터 분포를 **표준 정규분포로 변환**하도록 학습(Bijectively) $\sim N(0, I)$
- **3D 데이터일 경우(선택):** Pixel-unshuffling을 통해 해상도를 맞춘 Depth Feature map을 RGB Feature map과 결합(Channel-wise)

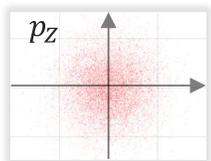
Affine Coupling layer

$$y_2 = x_2 \odot \exp(s_1([x_1, c])) + t_1(x_1, c)$$

$$y_1 = x_1 \odot \exp(s_2([x_2, c])) + t_2(x_2, c)$$

Minimizing Negative Log-Likelihood

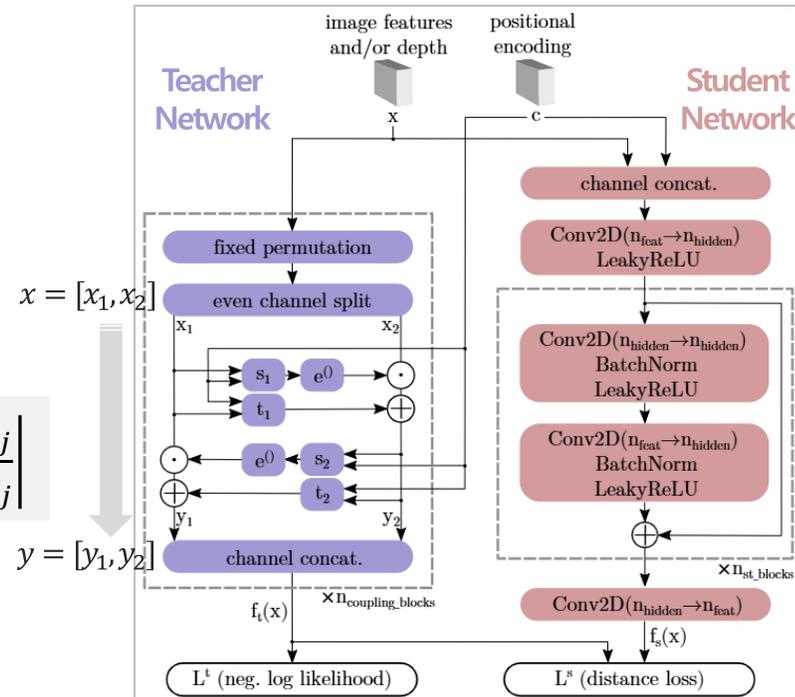
$$\mathcal{L}_{ij}^{\ell} = -\log p_X(x_{ij}) = \frac{|z_{ij}|_2^2}{2} - \log \left| \det \frac{\partial z_{ij}}{\partial x_{ij}} \right|$$



Normal Distribution
Log term

↓ Minimizing

원점 중심으로 모이는 방향으로 학습



논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Methodology

√. Student 학습

- **Student Network 구조:** Fully Convolutional Networks (4개의 Residual block)
- **학습 목표:** 정상 이미지를 입력으로 받아 Teacher Network가 출력하는 기준점(Target)을 모방하도록 학습

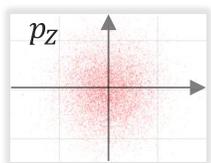
Affine Coupling layer

$$y_2 = x_2 \odot \exp(s_1([x_1, c])) + t_1(x_1, c)$$

$$y_1 = x_1 \odot \exp(s_2([x_2, c])) + t_2(x_2, c)$$

Minimizing Negative Log-Likelihood

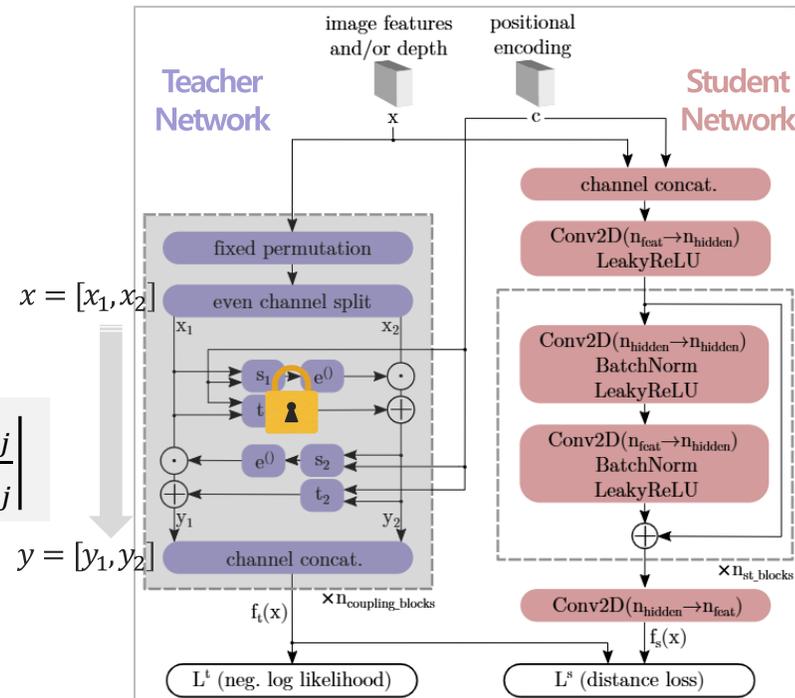
$$\mathcal{L}_{ij}^l = -\log p_X(x_{ij}) = \frac{|z_{ij}|_2^2}{2} - \log \left| \det \frac{\partial z_{ij}}{\partial x_{ij}} \right|$$



Normal Distribution
Log term

↓ Minimizing

원점 중심으로 모이는 방향으로 학습



Minimizing Squared L_2 -distance Loss

$$\mathcal{L}_{ij}^s = |f_s(x)_{ij} - f_t(x)_{ij}|_2^2.$$

Teacher와 Student 출력의 Distance를 줄이는 방향으로 학습

전체 픽셀에 대해 Distance의 평균을 낸 값이 최종 Loss

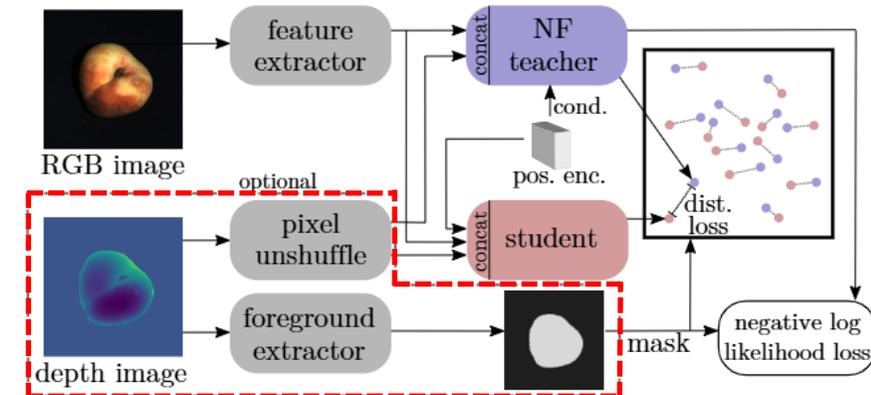
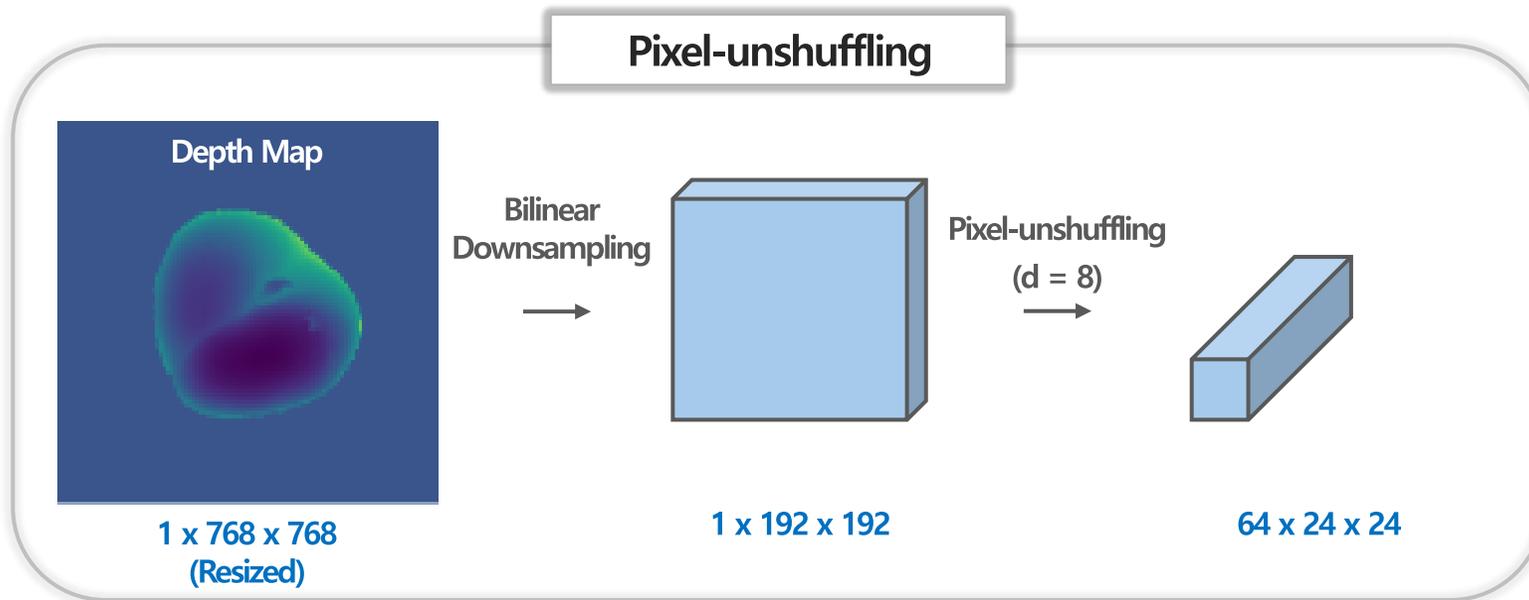
논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Implementation Details

• 3D 데이터 전처리

- ① **사용 데이터:** XYZ 이미지에서 정보량이 적은 X, Y 좌표는 버리고 센티미터 단위의 **Depth(Z) 성분만 사용**
- ② **Pixel-unshuffling:** Depth Map을 Teacher가 출력하는 Feature map과 차원을 맞추는 작업 (Space to Depth)
- ③ **Foreground extractor:** Depth map의 4개 모서리 픽셀을 Interpolating 하여 2D 평면으로 모델링 후 7mm 이상 떨어진 픽셀을 Foreground로 간주
- ④ **Loss Masking:** Foreground는 1, Background는 0으로 하는 **Binary foreground Masking** → Background는 학습 과정에서 무시

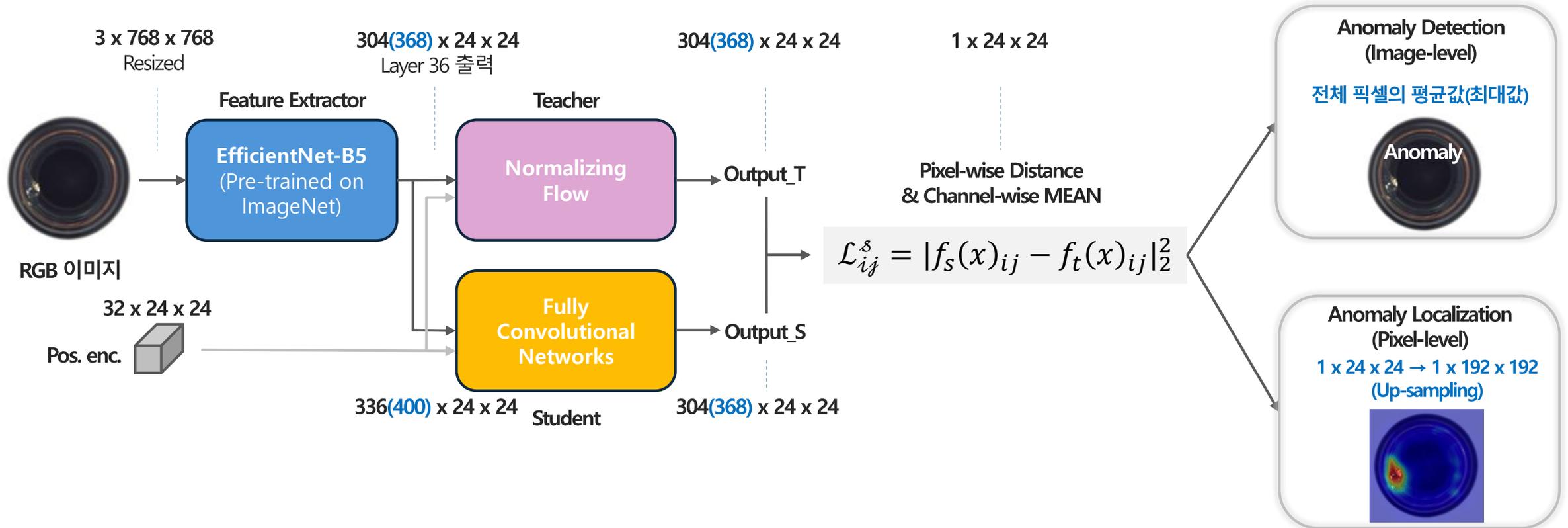


논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Implementation Details

※ (): 3D 이미지인 경우

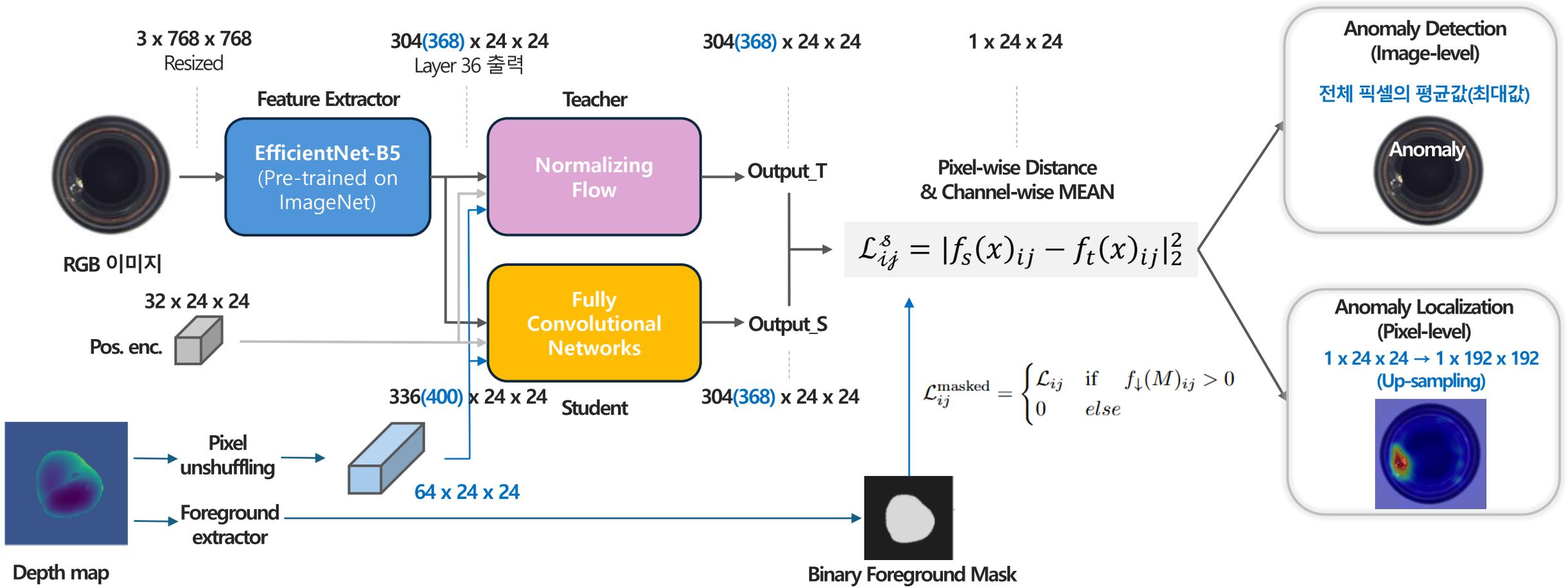


논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Implementation Details

※ (): 3D 이미지인 경우

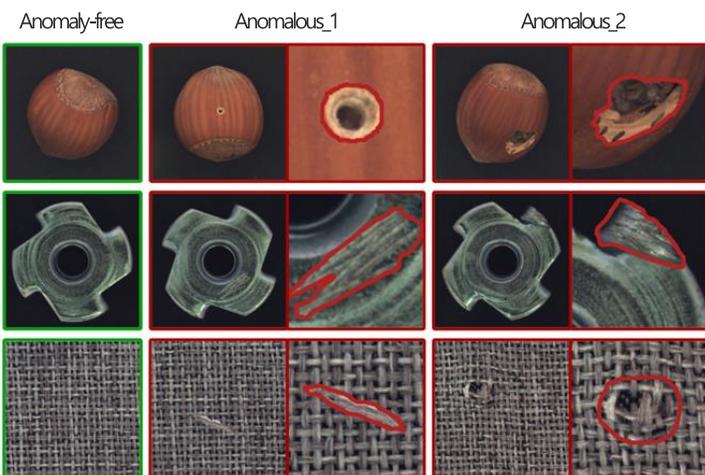


논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Datasets

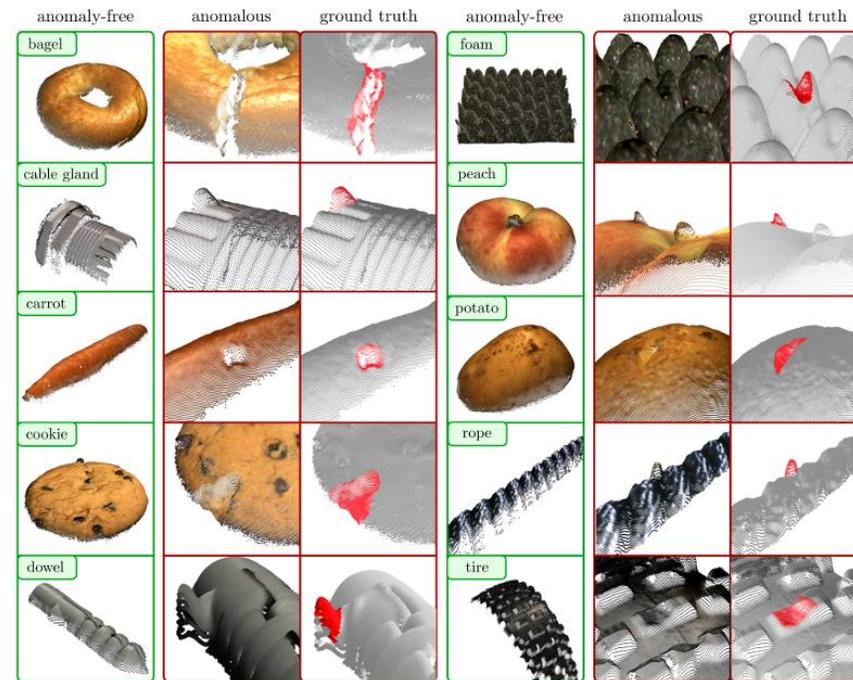
- **사용된 벤치마크 데이터셋:** MVTec AD & MVTec 3D-AD
- **훈련 데이터셋:** 결함이 없는 정상(defect-free) 데이터만 포함
- **평가 데이터셋:** 정상 데이터와 결함이 있는(defective) 데이터 모두 포함
- **레이블:** Anomaly Detection을 위한 Image-level의 레이블 및 Anomaly Localization(Segmentation)을 위한 Pixel-level의 결함 부위 마스크 제공



Displacement(위치 이탈), Cracks(균열), Scratches(긁힘) 등을 포함

Dataset Alias	MVTec AD (MVT2D)	MVTec 3D-AD (MVT3D)
RGB images	✓	✓
3D scans	×	✓
#categories	15	10
image side length	700-1024	400-800
#train samples per cat.	60-320	210-300
#test samples per cat.	42-160	100-159
#defect types per cat.	1-7	3-5

Depth map으로만 식별 가능한 Indentation(패임)이나 RGB로만 식별 가능한 Discoloration(변색) 등의 결함 포함



논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Anomaly Detection (MVT2D)

Category	ARNet [18]	DRÆM [54]	GAN [1]	Rippel [35]	PatchCore [36]	DifferNet [38]	PaDiM [12]	기존 NF 모델		기존 T-S 모델		AST (ours)	
								CFlow [22]	CS-Flow [39]	Uninf. Stud. [7]	STFPM* [51]		
Textures	Grid	88.3	99.9	70.8	93.7	98.2	84.0	-	99.6	99.0	98.1	100	99.1 ± 0.2
	Leather	86.2	100	84.2	100	100	97.1	-	100	99.9	94.7	100	100 ± 0.0
	Tile	73.5	99.6	79.4	100	98.7	99.4	-	99.9	100	94.7	95.5	100 ± 0.0
	Carpet	70.6	97.8	69.9	99.6	98.7	92.9	-	98.7	100	99.9	98.9	97.5 ± 0.4
	Wood	92.3	99.1	83.4	99.2	98.8	99.8	-	99.1	100	99.1	99.2	100 ± 0.0
Avg. Text.	82.2	99.3	77.5	98.5	98.3	94.6	99.0	99.5	99.8	97.3	98.7	99.3 ± 0.08	
Objects	Bottle	94.1	99.2	89.2	99.0	100	99.0	-	100	99.8	99.0	100	100 ± 0.0
	Capsule	68.1	98.5	73.2	96.3	98.1	86.9	-	97.7	97.1	92.5	88.0	99.7 ± 0.1
	Pill	78.6	98.9	74.3	91.4	96.6	88.8	-	96.8	98.6	92.2	93.8	99.1 ± 0.1
	Transistor	84.3	93.1	79.2	98.2	100	91.1	-	95.2	99.3	79.4	93.7	99.3 ± 0.1
	Zipper	87.6	100	74.5	98.8	99.4	95.1	-	98.5	99.7	94.4	93.6	99.1 ± 0.1
	Cable	83.2	91.8	75.7	99.1	99.5	95.9	-	97.6	99.1	78.7	92.3	98.5 ± 0.2
	Hazelnut	85.5	100	78.5	100	100	99.3	-	100	99.6	99.1	100	100 ± 0.0
	Metal Nut	66.7	98.7	70.0	97.4	100	96.1	-	99.3	99.1	89.1	100	98.5 ± 0.2
	Screw	100	93.9	74.6	94.5	98.1	96.3	-	91.9	97.6	86.0	88.2	99.7 ± 0.1
	Toothbrush	100	100	65.3	94.1	100	98.6	-	99.7	91.9	100	87.8	96.6 ± 0.1
	Avg. Obj.	84.8	97.4	75.5	96.9	99.2	94.7	97.2	97.7	98.2	91.0	93.7	99.1 ± 0.03
	Average	83.9	98.0	76.2	97.5	99.1	94.7	97.9	98.3	98.7	93.2	95.4	99.2 ± 0.04

Table 2. AUROC in % for detecting defects of all categories of MVT2D [6] on image-level grouped into textures and objects. We report the mean and standard deviation over 5 runs for our method. Best results are in bold. Beside the average value, detailed results of PaDiM [12] were not provided by the authors. The numbers of STFPM* [51] were obtained by a reimplementation.

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Anomaly Detection (MVT3D)

Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean	
3D	Voxel GAN [8]	38.3	62.3	47.4	63.9	56.4	40.9	61.7	42.7	66.3	57.7	53.7
	Voxel AE [8]	69.3	42.5	51.5	79.0	49.4	55.8	53.7	48.4	63.9	58.3	57.1
	Voxel VM [8]	75.0	74.7	61.3	73.8	82.3	69.3	67.9	65.2	60.9	69.0	69.9
	Depth GAN [8]	53.0	37.6	60.7	60.3	49.7	48.4	59.5	48.9	53.6	52.1	52.3
	Depth AE [8]	46.8	73.1	49.7	67.3	53.4	41.7	48.5	54.9	56.4	54.6	54.6
	Depth VM [8]	51.0	54.2	46.9	57.6	60.9	69.9	45.0	41.9	66.8	52.0	54.6
	1-NN (FPFH) [24]	82.5	55.1	95.2	79.7	88.3	58.2	75.8	88.9	92.9	65.3	78.2
	3D-ST ₁₂₈ [5]📧	86.2	48.4	83.2	89.4	84.8	66.3	76.3	68.7	95.8	48.6	74.8
	AST (ours)	88.1 ± 2.0	57.6 ± 6.9	96.5 ± 1.0	95.7 ± 0.6	67.9 ± 1.1	79.7 ± 1.2	99.0 ± 0.9	91.5 ± 2.1	95.6 ± 0.7	61.1 ± 3.4	83.3 ± 0.8
RGB	PatchCore [36]	87.6	88.0	79.1	68.2	91.2	70.1	69.5	61.8	84.1	70.2	77.0
	DifferNet [38]📧	85.9	70.3	64.3	43.5	79.7	79.0	78.7	64.3	71.5	59.0	69.6
	PADiM [12]*	97.5	77.5	69.8	58.2	95.9	66.3	85.8	53.5	83.2	76.0	76.4
	CS-Flow [39]📧	94.1	93.0	82.7	79.5	99.0	88.6	73.1	47.1	98.6	74.5	83.0
	STFPM [51]*	93.0	84.7	89.0	57.5	94.7	76.6	71.0	59.8	96.5	70.1	79.3
	AST (ours)	94.7 ± 0.7	92.8 ± 1.2	85.1 ± 1.2	82.5 ± 0.8	98.1 ± 0.4	95.1 ± 0.6	89.5 ± 1.1	61.3 ± 2.4	99.2 ± 0.2	82.1 ± 0.9	88.0 ± 0.6
3D + RGB	Voxel GAN [8]	68.0	32.4	56.5	39.9	49.7	48.2	56.6	57.9	60.1	48.2	51.7
	Voxel AE [8]	51.0	54.0	38.4	69.3	44.6	63.2	55.0	49.4	72.1	41.3	53.8
	Voxel VM [8]	55.3	77.2	48.4	70.1	75.1	57.8	48.0	46.6	68.9	61.1	60.9
	Depth GAN [8]	53.8	37.2	58.0	60.3	43.0	53.4	64.2	60.1	44.3	57.7	53.2
	Depth AE [8]	64.8	50.2	65.0	48.8	80.5	52.2	71.2	52.9	54.0	55.2	59.5
	Depth VM [8]	51.3	55.1	47.7	58.1	61.7	71.6	45.0	42.1	59.8	62.3	55.5
	PatchCore+FPFH [24]	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2	92.1	88.6	86.5
	AST (ours)	98.3 ± 0.4	87.3 ± 3.3	97.6 ± 0.5	97.1 ± 0.3	93.2 ± 2.1	88.5 ± 1.4	97.4 ± 1.4	98.1 ± 1.2	100 ± 0.0	79.7 ± 1.0	93.7 ± 0.2

Table 3. AUROC in % for detecting defects of all categories of MVT3D [8] on image-level for 3D data, RGB data and the combination of both. We report the mean and standard deviation over 5 runs for our method. Best results per data domain are in bold. Numbers of listed methods followed by a 📧 are non-published results obtained by the corresponding authors on request. A * indicates that we used a reimplementation. The numbers from PatchCore are taken from [24].

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Anomaly Localization(Segmentation)

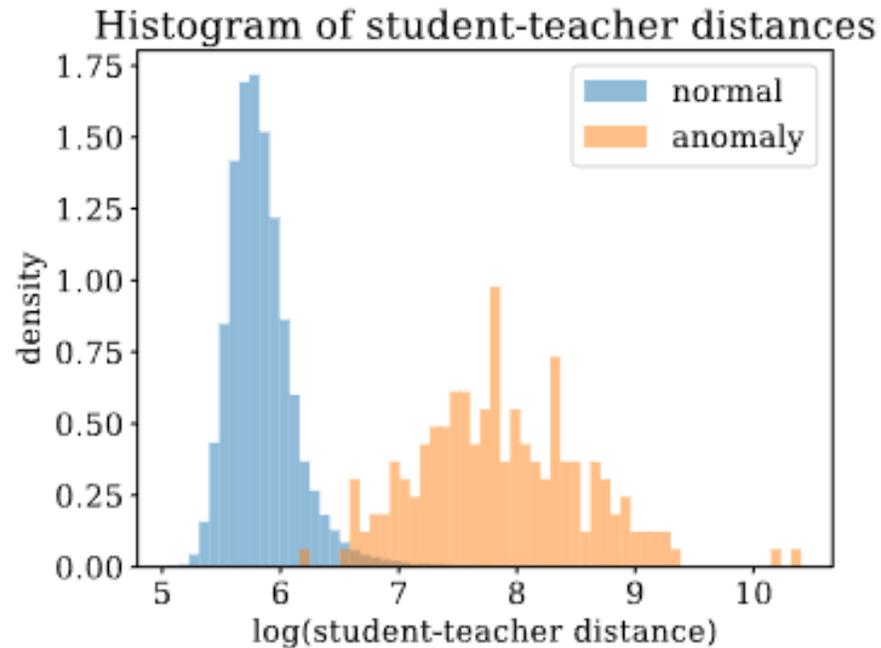
Method	MVT2D	MVT3D (RGB+3D)
AE-SSIM [9]	87.0	-
PatchCore [36]	98.4	-
PatchCore+FPFH [24]	-	99.2
AST (ours)	95.0 ± 0.03	97.6 ± 0.02

Table 4. Anomaly segmentation results measured by the mean pixel-AUROC over all classes and its standard deviation over 5 runs. Despite image-level detection is the focus of this work, our method is able to localize defects for practical purposes with an AUROC of 95% or 97.6%.

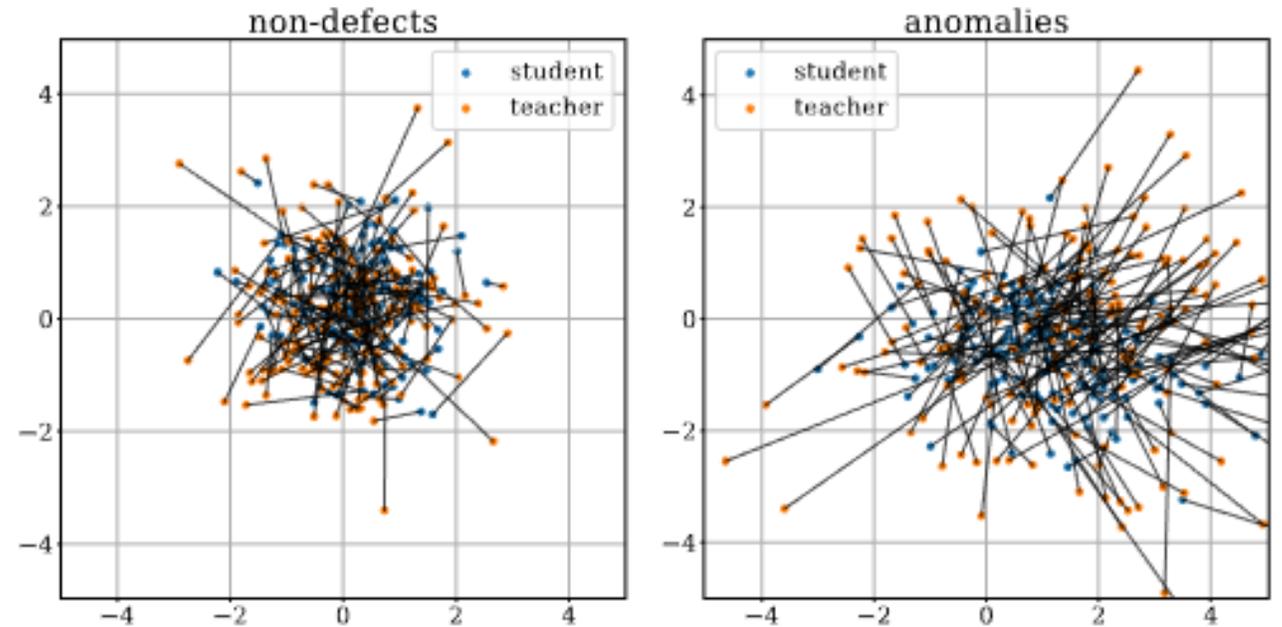
논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Anomaly Localization(Segmentation)



Histogram of AST distances in MVT3D



Random orthographic projections

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Ablation studies

Method	3D	RGB	3D+RGB
Teacher only	82.2	69.8	90.9
NF student (symm.)	81.8	76.0	88.9
NF student (deeper)	81.8	76.7	92.7
AST (ours)	83.3	88.0	93.7

Table 5. Comparison of average detection performance in AUROC percentage on MVT3D of teacher and student-teacher in a symmetric and asymmetric setting. Our proposed asymmetric student-teacher pair outperforms all baselines in all cases.

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Ablation studies

Student Network의 깊이(Residual Block 수)에 변화 따른 비교

성과와 효율성 간의 좋은 균형점

n_{st_blocks}	AUROC [%] ↑	#Params. [M] ↓	inf. time [ms] ↓
1	92.8	26.0	3.4
2	93.3	44.8	6.1
4	93.7	82.6	10.4
8	93.7	151.1	19.8
12	93.8	233.6	29.4
teacher	90.9	3.8	4.5

Table 6. Tradeoff between performance and computational effort on 3D+RGB data of MVT3D. The inference time was measured with a *NVIDIA RTX 1080 Ti*.

논문4. AST: Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

❖ Experiments

√. Ablation study

MVTec 3D-AD 데이터셋에서 여러 설정 변화 따른 비교

input	pos. enc.	mask	teacher	AST
3D	✗	✓	78.4	81.9
	✓	✗	59.4	67.2
	✓	✓	82.2	83.3
RGB	✗	✗	69.3	87.8
	✓	✗	69.8	88.0
	✓	✓	n. a.	n. a.
3D+RGB	✗	✓	90.9	93.8
	✓	✗	66.2	84.0
	✓	✓	90.9	93.7

Table 7. Impact of the positional encoding and the foreground mask on the detection performance of student and teacher on MVT3D. Numbers are given in AUROC percentage. Since masks are obtained from 3D data, there is no mask for RGB.

Conclusion & Future works

❖ Conclusion

√. Teacher-Student 패러다임 진화와 핵심 기술

- 구조적 진화(Architecture) : 동질적 구조(MKD) → 역방향 흐름(RD4AD) → 완전한 비대칭 구조(AST)로 발전
- 표현력의 극대화 (Multi-scale & Feature Fusion) : 사전 학습된 모델을 활용하여 풍부하고 의미론적(Semantic)인 특징 정보를 학습

√. AST 이후 주목할 Teacher-Student 연구 동향

- GCAD(Global Context Anomaly Detection) : '부품 누락', '오배열' 등 구조를 뛰어넘는 논리적 결함(Logical Anomaly) 탐지 연구
- EfficientAD : 무거운 모델 크기와 연산량을 획기적으로 경량화하여 현장 적용성을 높인 실용화 연구

고맙습니다
