Medical Imaging & Intelligent Reality Lab.

내부 논문 Review Study 발표 : DeSTSeg

2023.08.04

Youngjae Kim

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

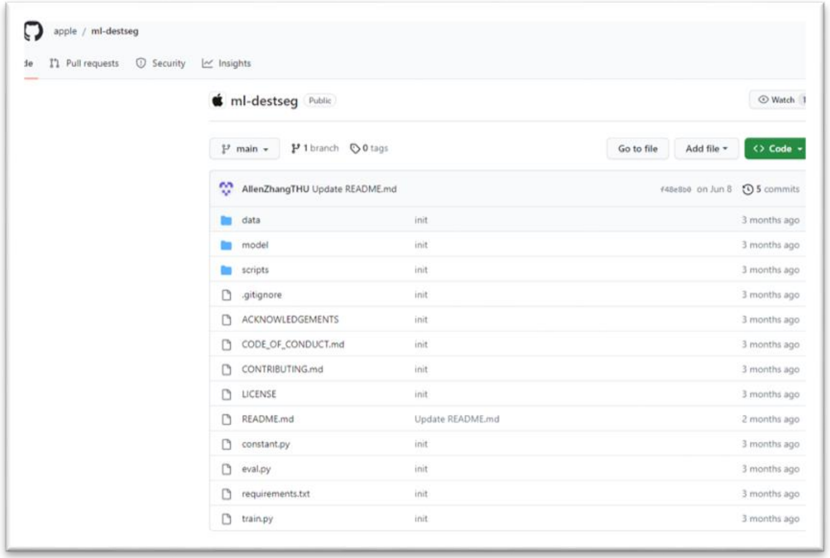
1. About the Paper

DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection

Xuan Zhang[1], Shiyu Li[2], Xi Li[2], Ping Huang[2], Jiulong Shan[2], Ting Chen[1]
[1]Tsinghua University [2]Apple
x-zhang18@mails.tsinghua.edu.cn, {shiyu_li, weston_li, huang_ping, jlshan}@apple.com, tingchen@tsinghua.edu.cn

- **Accepted by CVPR 2023**
- **Tsinghua Univ, Apple**

https://github.com/apple/ml-destseg



2. Introduction & Related Works

- **Importance of Visual Anomaly Detection (AD)**
    1) Industrial inspection
    2) Medical disease screening
    3) Video surveillance (like CCTV)

- **Properties of Anomalous samples**
    1) Types are enormous
    2) Occurs rarely
    3) Impossible to get all possible cases

-> usually, use only "normal" samples to train

2. Introduction & Related Works

- **Student – Teacher Framework (knowledge distillation)**
  - Proven to be effective in AD

Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings

Paul Bergmann

{paul.be

Asymmetric Student-Teacher Networks for Industrial Anomaly Detection

Marco Rudolph[1] Tom Wehrbein[1] Bodo Rosenhahn[1] Bastian Wandt[2]

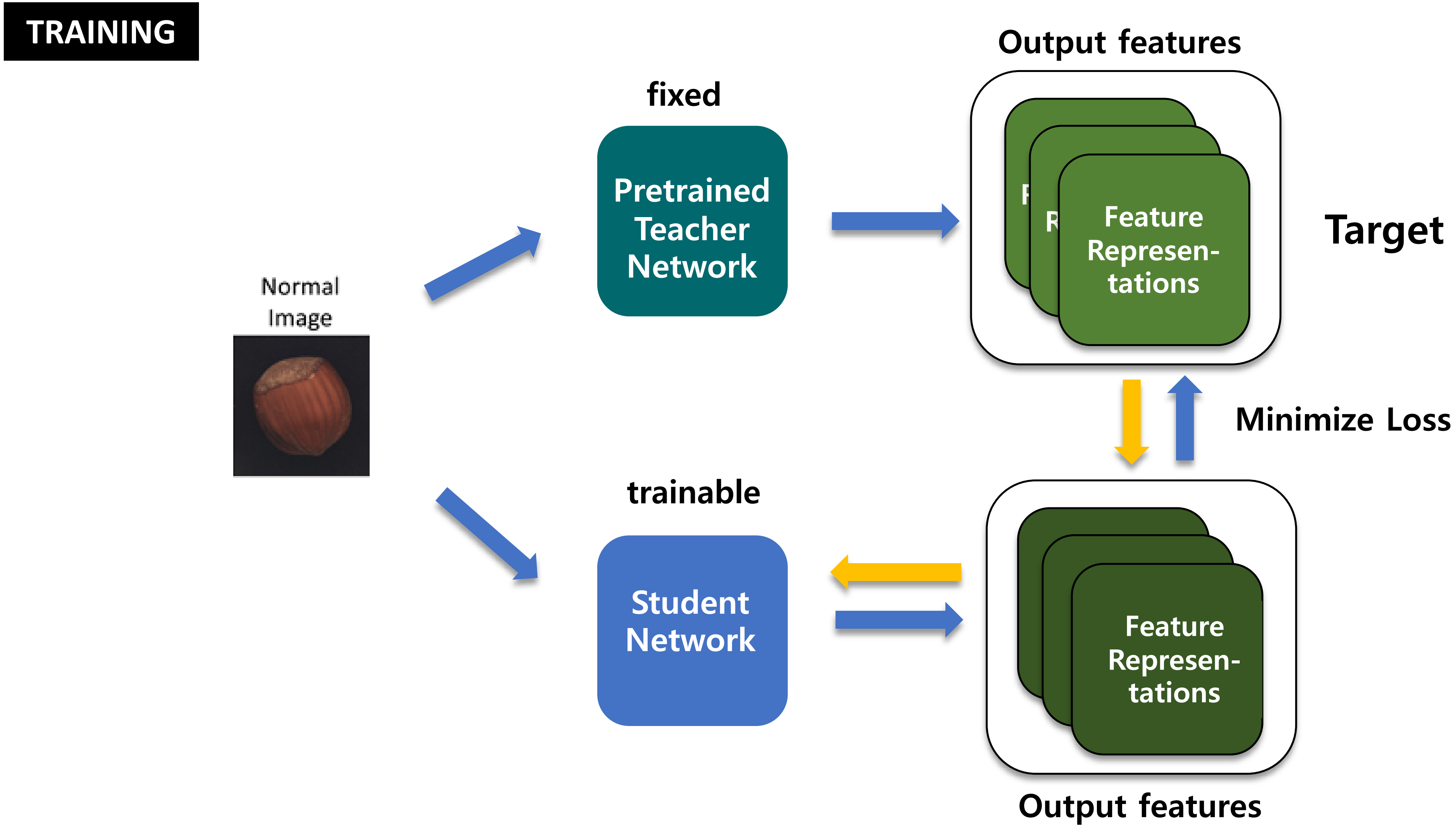
[1]L3S / Leibniz University Hannover, Germany [2]Linköping University, Sweden

rudolph@tnt.uni-hannover.de

- Pretrained on large-scale dataset (ex. ResNet50 on ImageNet)
- Large Models

Pretrained Teacher Network → Trainable Student Network

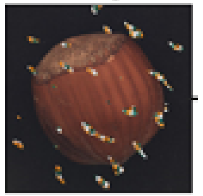
- Mimic feature representations of teacher network
- Small size / but try to keep the performance same


TRAINING
Normal Image
fixed
Pretrained Teacher Network
Output features
Feature Represen-tations
Target
Minimize Loss
trainable
Student Network
Feature Represen-tations
Output features

2. Introduction & Related Works

INFERENCE

Anomalous Image



fixed

Pretrained Teacher Network

outputs

y1

Student Network

y2

Trained only on normal data with teacher

Calculate difference

IF

difference is small

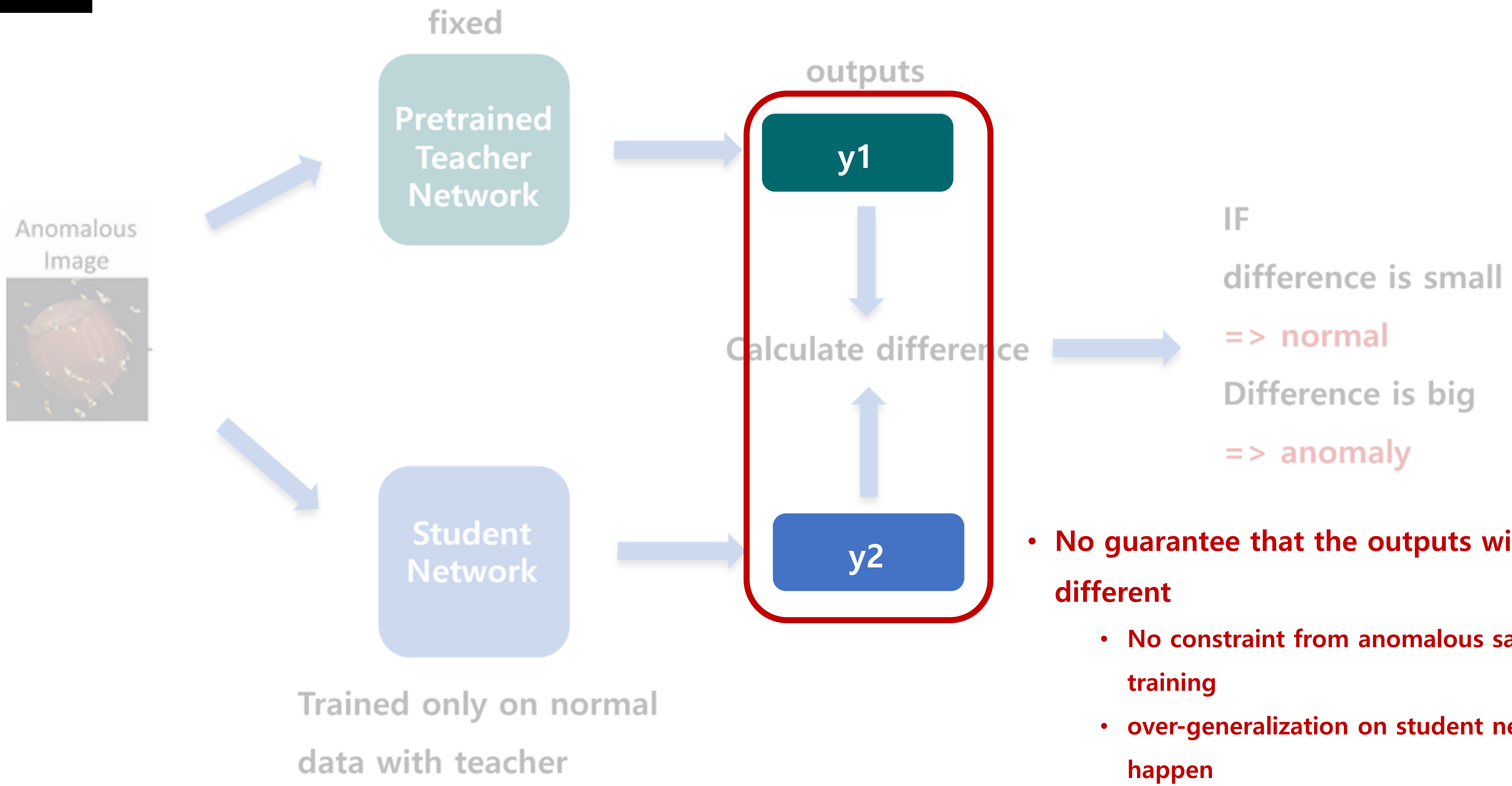
=> normal

Difference is big

=> anomaly

2. Introduction & Related Works

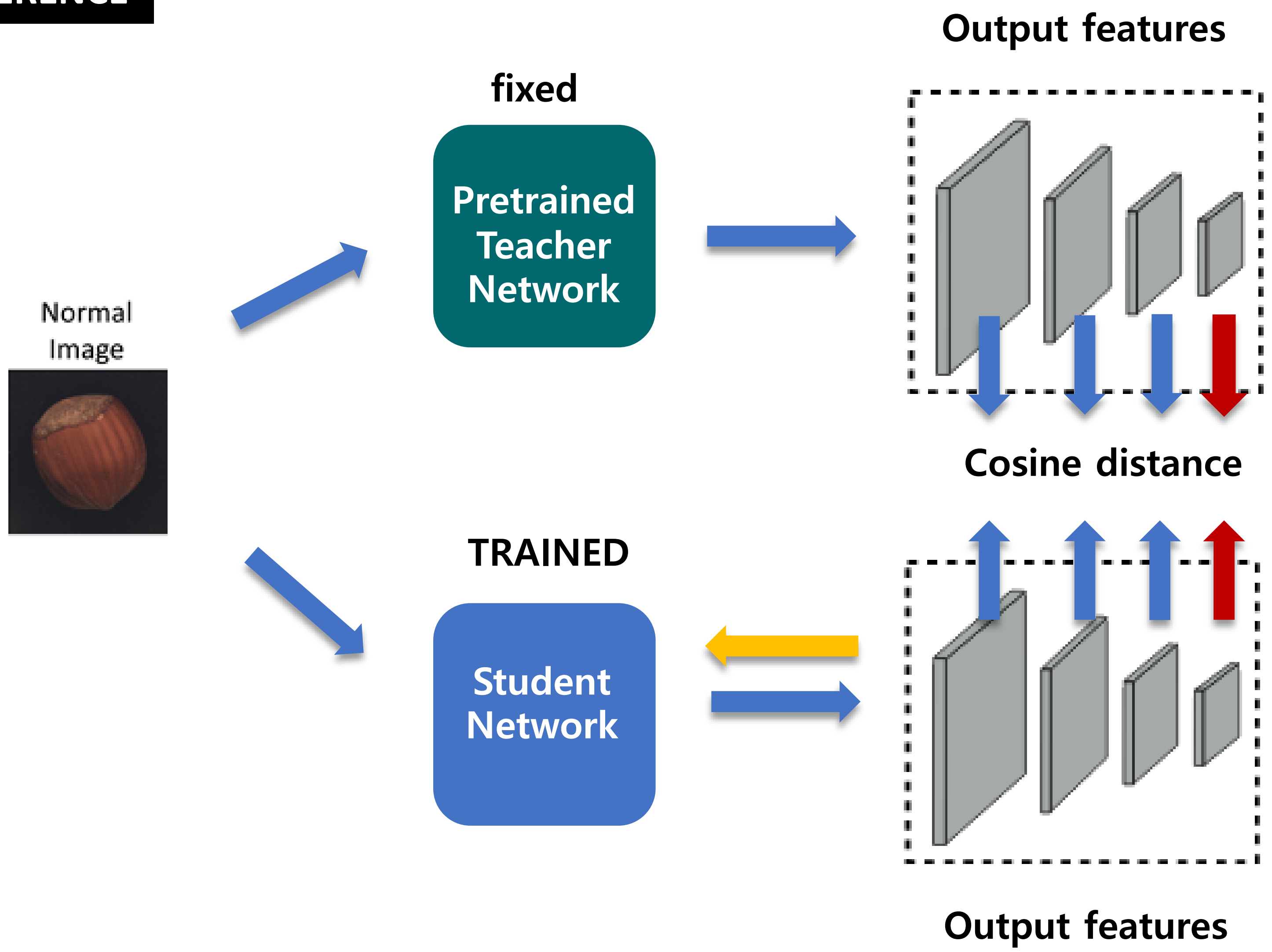
INFERENCE



- No guarantee that the outputs will be always different
  - No constraint from anomalous sample during training
  - over-generalization on student network could happen

2. Introduction & Related Works

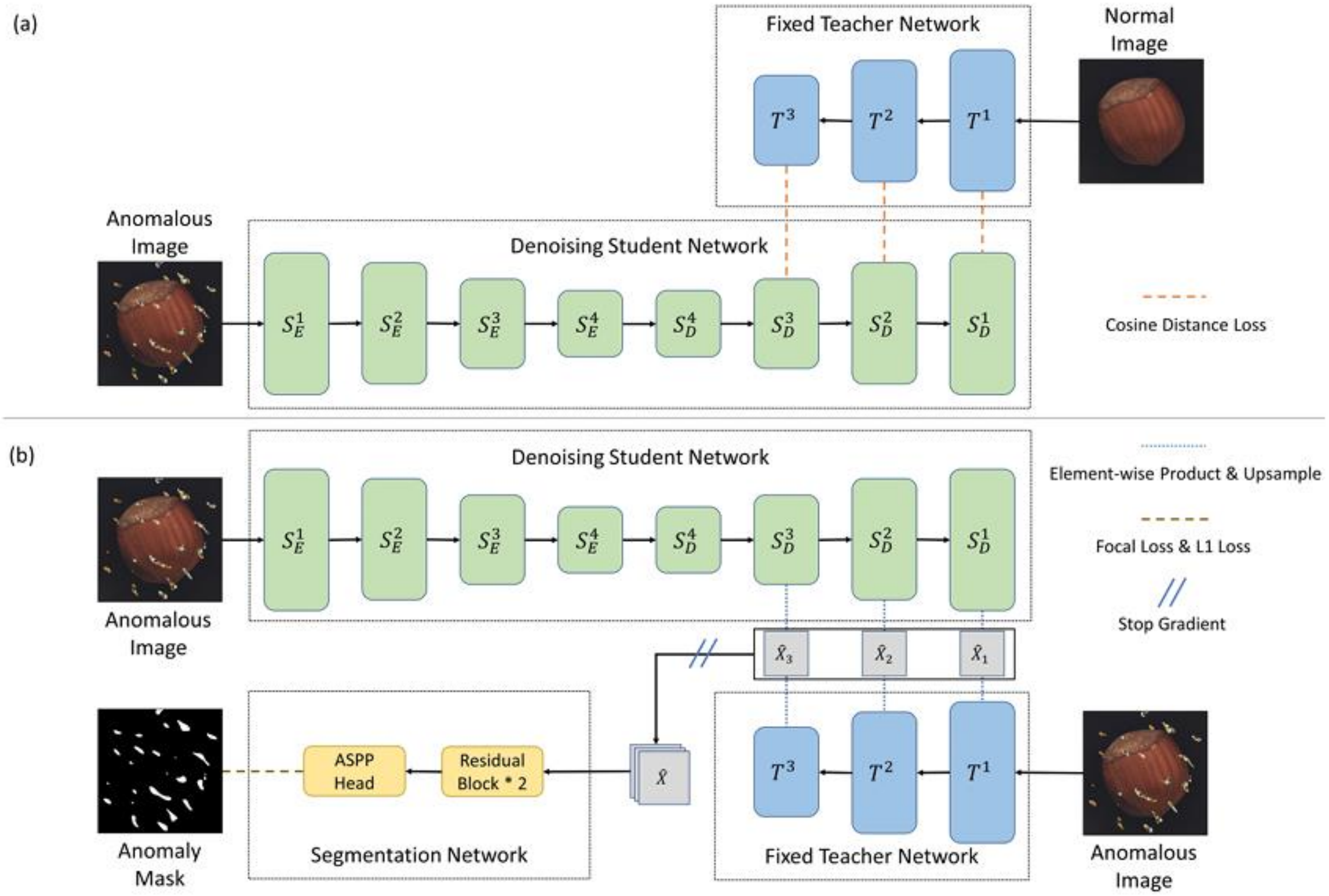
INFERENCE



- Using multi level feature could be suboptimal
  - MVTec AD dataset (category of transistor)
  - 88.4% using only last layer feature representation
  - 81.9% on multi-level features

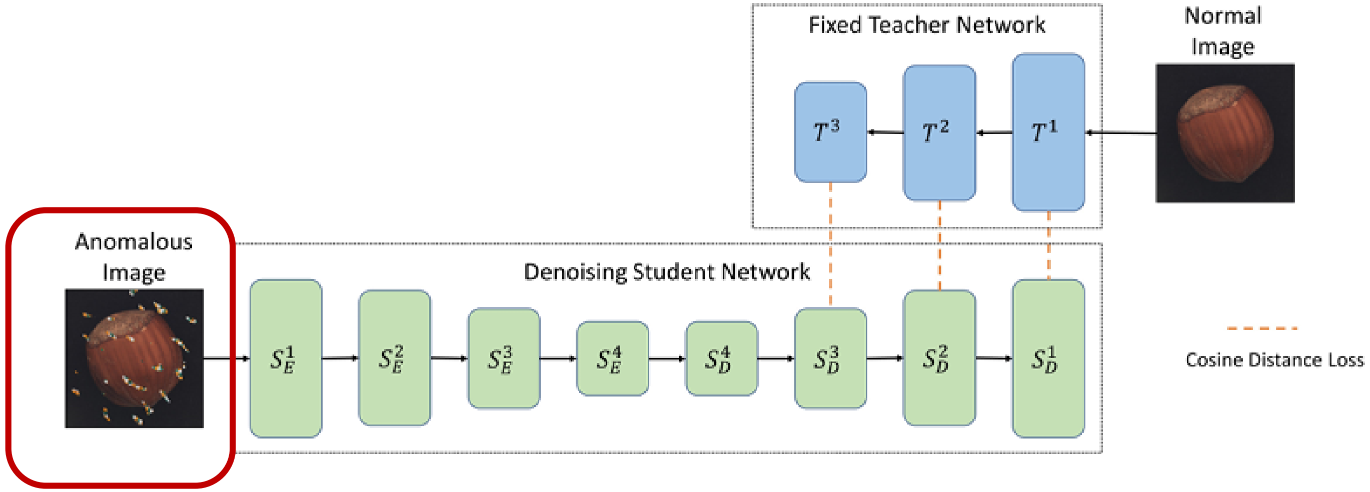
3. Method (DeSTSeg)

DeSTSeg = Denoising S + T + Segmentation



3. Method (DeSTSeg)

(a)



- Generated synthetic anomaly image
  1) Generate random two-dimensional Perlin noise
  2) Binarize to obtain anomaly mask (*M*)
  3) Replace mask region with anomaly-free image (I_a) & arbitrary image from external data source (*A*)
  4) Apply opacity (β) [0.15,1]

$$I_a = \beta(M \odot A) + (1-\beta)(M \odot I_n) + (1-M) \odot I_n \quad (1)$$

$\odot$ means the element-wise multiplication operation.

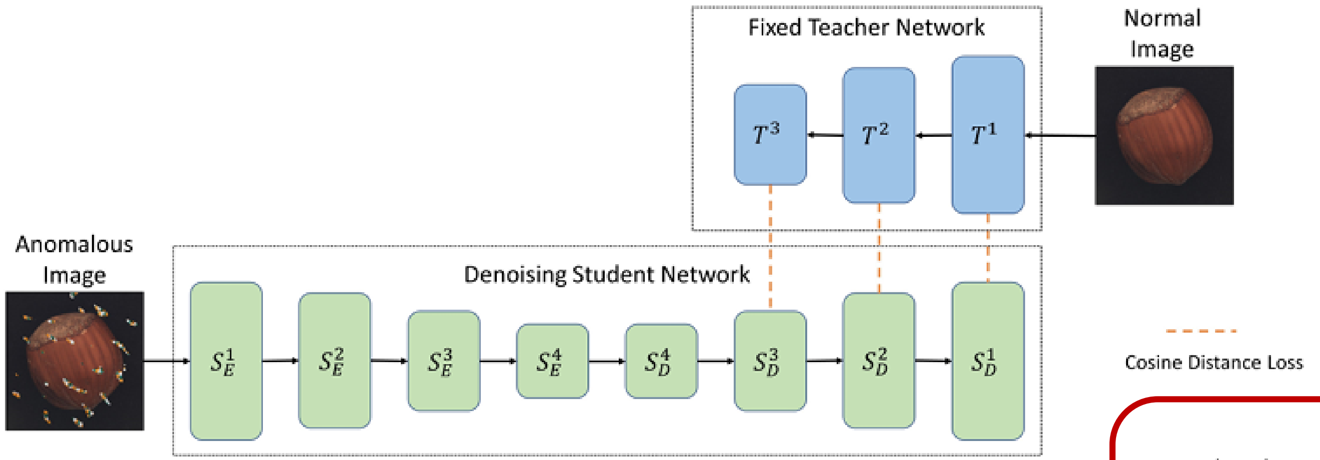
DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection

Vitjan Zavrtanik Matej Kristan Danijel Skočaj
University of Ljubljana, Faculty of Computer and Information Science
{vitjan.zavrtanik, matej.kristan, danijel.skocaj}@fri.uni-lj.si

3. Method (DeSTSeg)

- ImageNet pretrained ResNet18 with final block removed(conv5_x)

(a)



- i , j – spatial coordinates on the feature maps

$$X_k(i,j) = \frac{F_{T_k}(i,j) \odot F_{S_k}(i,j)}{||F_{T_k}(i,j)||_2 ||F_{S_k}(i,j)||_2} \quad (2)$$

Element-wise product between feature maps of S-T

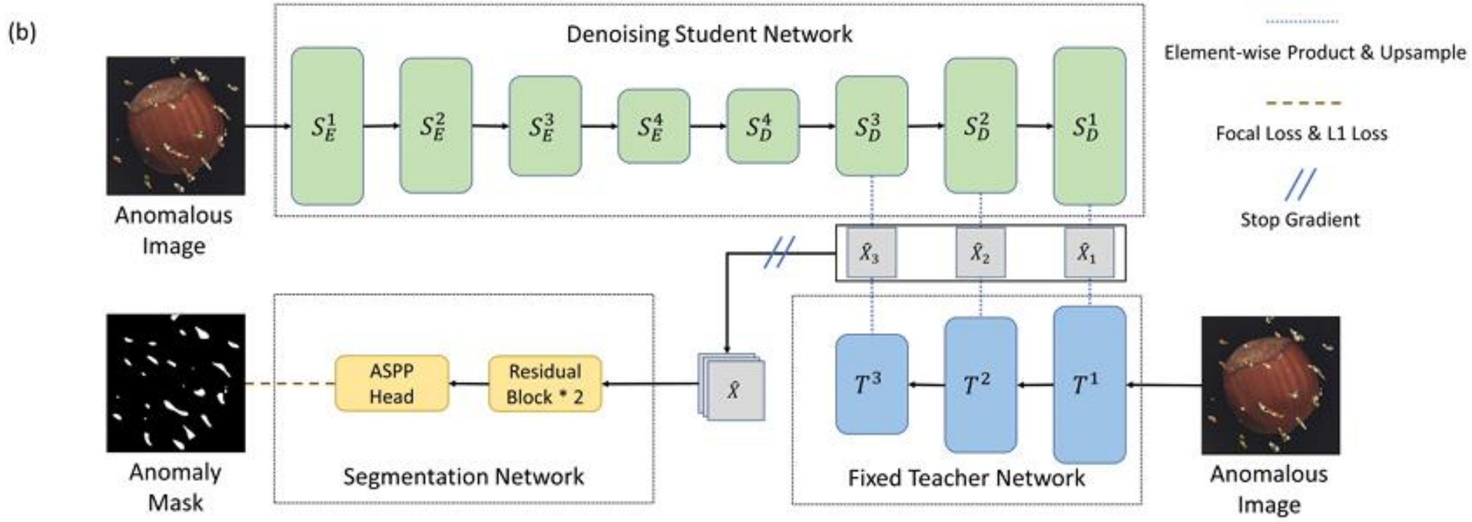
$$D_k(i,j) = 1 - \sum_{c=1}^{C_k} X_k(i,j)_c \quad (3)$$

$$L_{cos} = \sum_{k=1}^{3} \left(\frac{1}{H_k W_k} \sum_{i,j=1}^{H_k,W_k} D_k(i,j) \right) \quad (4)$$

- Encoder – randomly initialized ResNet18
- Decoder – reversed ResNet18

- Train student network to remove noise from anomalous image

3. Method

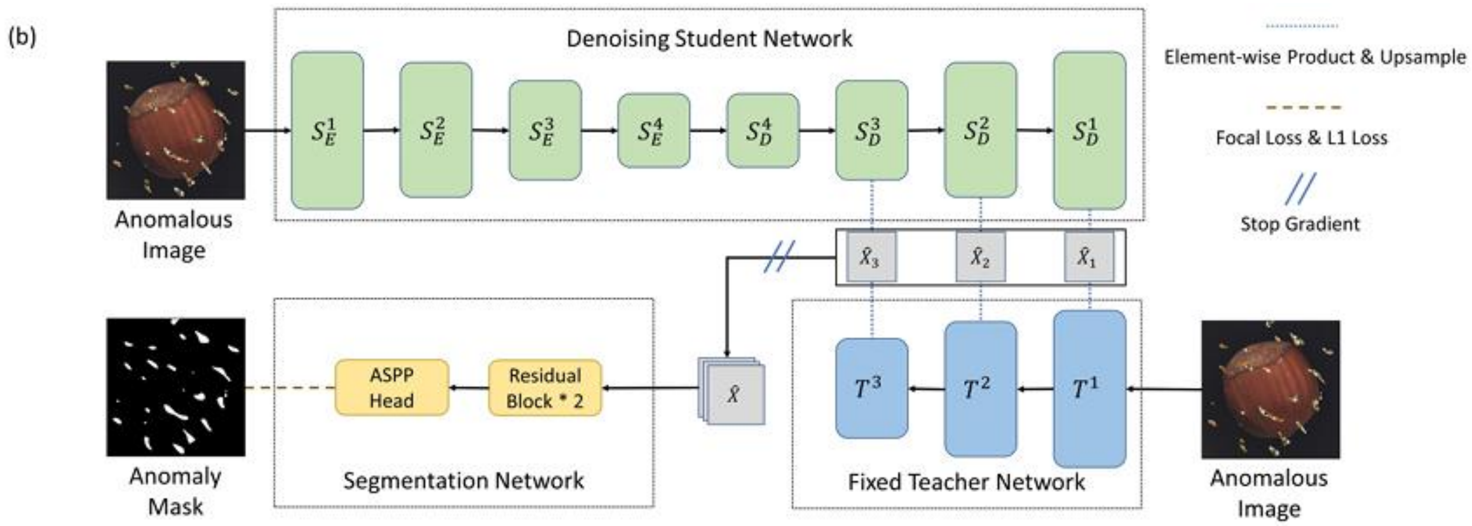


- Freeze Student & Teacher Network
- Anomalous images are used as input for both S & T networks
- Binary anomaly mask (M) which is generated before -> Ground Truth
- X hat 1~3 : similarities of the paired feature maps -> calculated with (2)
- Then, concatenated as X hat -> fed into segmentation network

- Similarities of the feature maps calculated
- Element-wise product between feature maps of S-T
- + Upsample = X hat 1~3

$$X_k(i,j) = \frac{F_{T_k}(i,j) \odot F_{S_k}(i,j)}{||F_{T_k}(i,j)||_2 ||F_{S_k}(i,j)||_2} \qquad (2)$$

3. Method



- Loss for segmentation

$$L_{focal} = -\frac{1}{H_1 W_1} \sum_{i,j=1}^{H_1,W_1} (1 - p_{ij})^{\gamma} \log(p_{ij}) \qquad (5)$$

$$L_{l1} = \frac{1}{H_1 W_1} \sum_{i,j=1}^{H_1,W_1} |M_{ij} - \hat{Y}_{ij}| \qquad (6)$$

$$L_{seg} = L_{focal} + L_{l1} \qquad (7)$$

- **Focal loss (5)**
  1) **Even in anomalous image, majorities are background**
  2) **Focus on the minority category**

- **L1 loss (6)**
  1) **Improve sparsity of the output**
  2) **Segmentation mask's boundaries are more distinct**

3. Method (DeSTSeg)

Problem 1 : No guarantee that the outputs will be different

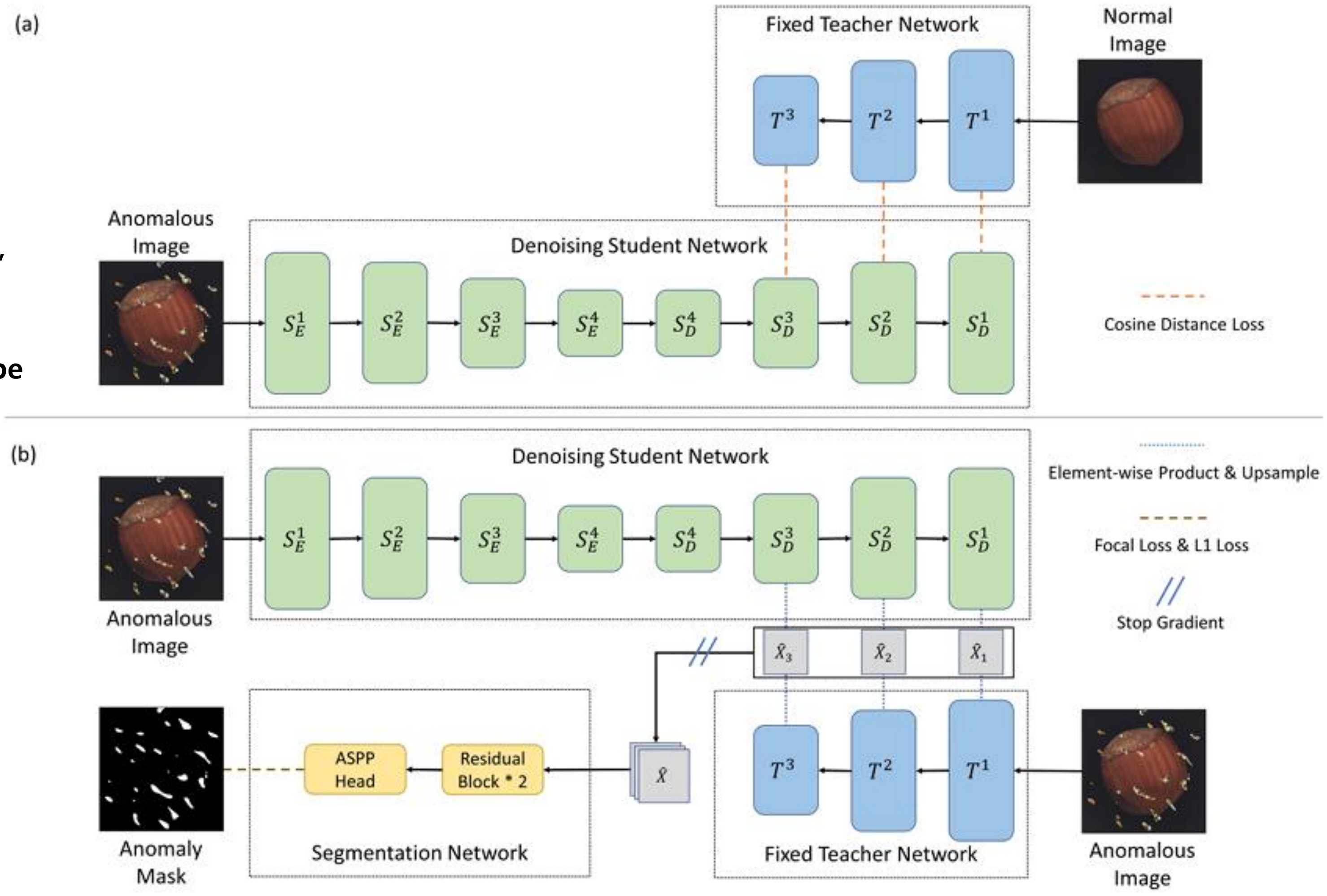
1) Training "Denoising Student Network" with encoder-decoder structure

=> Guaranteed the output features will be different

Problem 2 : Using multi level feature could be suboptimal

2) Training "Segmentation Network"

=> Will use multi-level feature as input of segmentation training

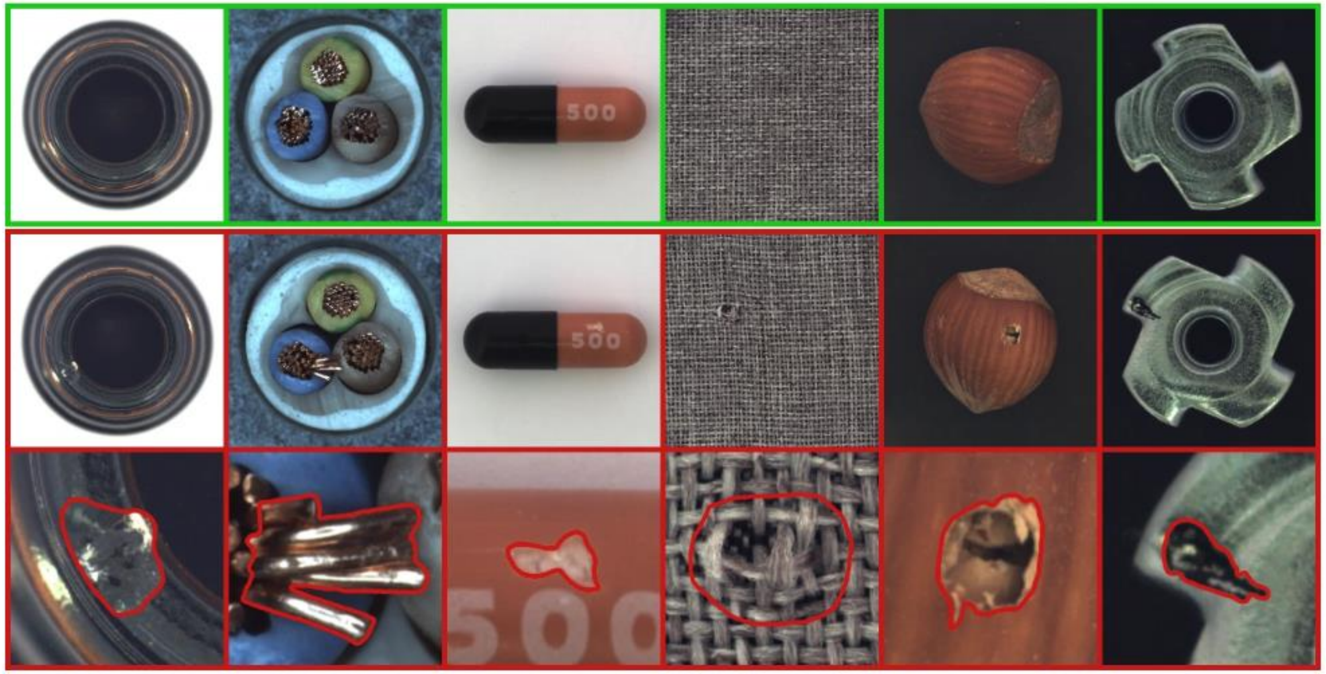


4. Experiments

Dataset

MVTec AD – one of the most widely used benchmarks for anomaly detection and localization

- 15 categories
- Hundreds of normal images for training
- mixture of anomalous and normal images for evaluation



4. Experiments

US [3]	STPM [31]	CutPaste [16]	DRAEM [36]	DSR [37]	PatchCore [24]	Ours
87.7	95.1	95.2	98.0	98.2	98.5	**98.6**±0.4

Table 1. Image-level anomaly detection AUC (%) on MVTec AD dataset. Results are averaged over all categories.

	US [3]	STPM [31]	CutPaste [16]	DRAEM [36]	DSR [37]	PatchCore [24]	Ours
bottle	97.8 / 74.2	98.8 / 80.6	97.6 / -	**99.3** / 89.8	- / **91.5**	98.9 / 80.1	99.2±0.2 / 90.3±1.8
cable	91.9 / 48.2	94.8 / 58.0	90.0 / -	95.4 / 62.6	- / **70.4**	**98.8** / 70.0	97.3±0.4 / 60.4±2.3
capsule	96.8 / 25.9	98.2 / 35.9	97.4 / -	94.1 / 43.5	- / 53.3	**99.1** / 48.1	**99.1**±0.0 / **56.3**±1.1
carpet	93.5 / 52.2	**99.1** / 65.3	98.3 / -	96.2 / 64.4	- / **78.2**	**99.1** / 66.7	96.1±2.2 / 72.8±5.8
grid	89.9 / 10.1	99.1 / 45.4	97.5 / -	**99.5** / 56.8	- / **68.0**	98.9 / 41.0	99.1±0.1 / 61.5±1.6
hazelnut	98.2 / 57.8	98.9 / 60.3	97.3 / -	99.5 / 88.1	- / 87.3	99.0 / 61.5	**99.6**±0.2 / **88.4**±2.2
leather	97.8 / 40.9	99.2 / 42.9	99.5 / -	98.9 / 69.9	- / 62.5	99.4 / 51.0	**99.7**±0.0 / **75.6**±1.2
metal_nut	97.2 / 83.5	97.2 / 79.3	93.1 / -	98.7 / 91.7	- / 67.5	**98.8** / 88.8	98.6±0.4 / **93.5**±1.1
pill	96.5 / 62.0	94.7 / 63.3	95.7 / -	97.6 / 46.1	- / 65.7	98.2 / 78.7	**98.7**±0.4 / **83.1**±4.2
screw	97.4 / 7.8	98.6 / 26.9	96.7 / -	**99.7** / **71.5**	- / 52.5	99.5 / 41.4	98.5±0.3 / 58.7±3.7
tile	92.5 / 65.3	96.6 / 61.7	90.5 / -	**99.5** / **96.9**	- / 93.9	96.6 / 59.3	98.0±0.7 / 90.0±2.5
toothbrush	97.9 / 37.7	98.9 / 48.8	98.1 / -	98.1 / 54.7	- / 74.2	98.9 / 51.6	**99.3**±0.1 / **75.2**±1.8
transistor	73.7 / 27.1	81.9 / 44.4	93.0 / -	90.0 / 51.7	- / 41.1	**96.2** / 63.2	89.1±3.4 / **64.8**±4.0
wood	92.1 / 53.3	95.2 / 47.0	95.5 / -	97.0 / 80.5	- / 68.4	95.1 / 52.3	**97.7**±0.3 / **81.9**±1.2
zipper	95.6 / 36.1	98.0 / 54.9	99.3 / -	98.6 / 72.3	- / 78.5	99.0 / 64.0	**99.1**±0.5 / **85.2**±3.3
average	93.9 / 45.5	96.6 / 54.3	96.0 / -	97.5 / 69.3	- / 70.2	**98.4** / 61.2	97.9±0.3 / **75.8**±0.8

Table 2. Pixel-level anomaly localization AUC / AP (%) on MVTec AD dataset.

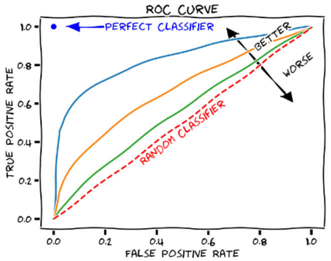
Image-level anomaly detection

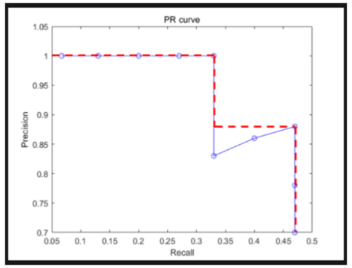
-> discriminate anomaly based on the whole image

-> if there is a anomaly region, the whole image is counted as an anomaly

Pixel-level anomaly detection

-> each pixels are discriminated as normal or anomaly





- AUC (Area under ROC curve)
- AP (Area under PR curve)

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

4. Experiments

Instance-level anomaly detection

-> Calculate overlap area between the object in ground truth mask and predicted mask

IAP = (TP + TN) / (TP + TN + FP + FN)

IAP@k% = if and only if overlapped area is over k%, considered as "detected"

Using different k thresholds, obtained average precision of this curve is called IAP.

	STPM [31]	DRAEM [36]	PatchCore [24]	Ours
bottle	83.2 / 73.3	90.3 / **84.8**	81.8 / 70.1	**90.5**$\pm$1.7 / 82.5$\pm$4.1
cable	54.9 / 17.2	47.0 / 10.8	**69.2** / **50.6**	51.1$\pm$2.5 / 26.7$\pm$3.7
capsule	37.2 / 17.9	**50.7** / 21.4	44.2 / 26.9	49.4$\pm$1.5 / **27.3**$\pm$3.3
carpet	68.4 / 52.2	76.8 / 32.3	64.4 / 43.7	**84.5**$\pm$4.9 / **58.6**$\pm$17.1
grid	45.7 / 21.0	55.5 / 42.3	39.1 / 15.6	**61.6**$\pm$1.8 / **47.4**$\pm$2.9
hazelnut	64.8 / 56.2	**95.7** / **89.0**	63.8 / 52.5	87.7$\pm$1.8 / 77.6$\pm$3.4
leather	46.2 / 24.9	**78.6** / 55.0	50.1 / 30.1	77.5$\pm$1.8 / **65.3**$\pm$3.9
metal_nut	83.4 / 81.7	92.6 / 83.9	90.1 / 84.6	**93.6**$\pm$1.3 / **86.5**$\pm$2.7
pill	72.0 / 45.5	46.9 / 41.5	82.7 / **63.5**	**84.8**$\pm$3.8 / 61.1$\pm$12.4
screw	24.4 / 4.2	**68.8** / **33.0**	38.4 / 16.3	53.6$\pm$3.6 / 8.6$\pm$2.3
tile	62.9 / 55.3	**98.9** / **98.2**	60.0 / 52.1	94.7$\pm$1.8 / 86.5$\pm$3.6
toothbrush	41.9 / 23.4	44.7 / 21.5	40.4 / 22.1	**59.8**$\pm$2.9 / **32.1**$\pm$5.1
transistor	53.4 / 8.5	59.3 / 22.8	69.9 / 36.8	**78.3**$\pm$2.5 / **49.6**$\pm$8.4
wood	56.0 / 35.4	**88.4** / 72.6	59.7 / 35.6	87.8$\pm$2.8 / **76.4**$\pm$3.4
zipper	59.1 / 46.6	78.7 / 67.0	66.0 / 52.4	**90.6**$\pm$2.3 / **80.3**$\pm$4.9
average	56.9 / 37.5	71.5 / 51.7	61.3 / 43.5	**76.4**$\pm$1.0 / **57.8**$\pm$1.8

Table 3. Instance-level anomaly detection IAP / IAP@90 (%) on MVTec AD dataset.

IAP@90 : when 90% of anomaly instances are detected, the value of pixel-level precision

4. Experiments

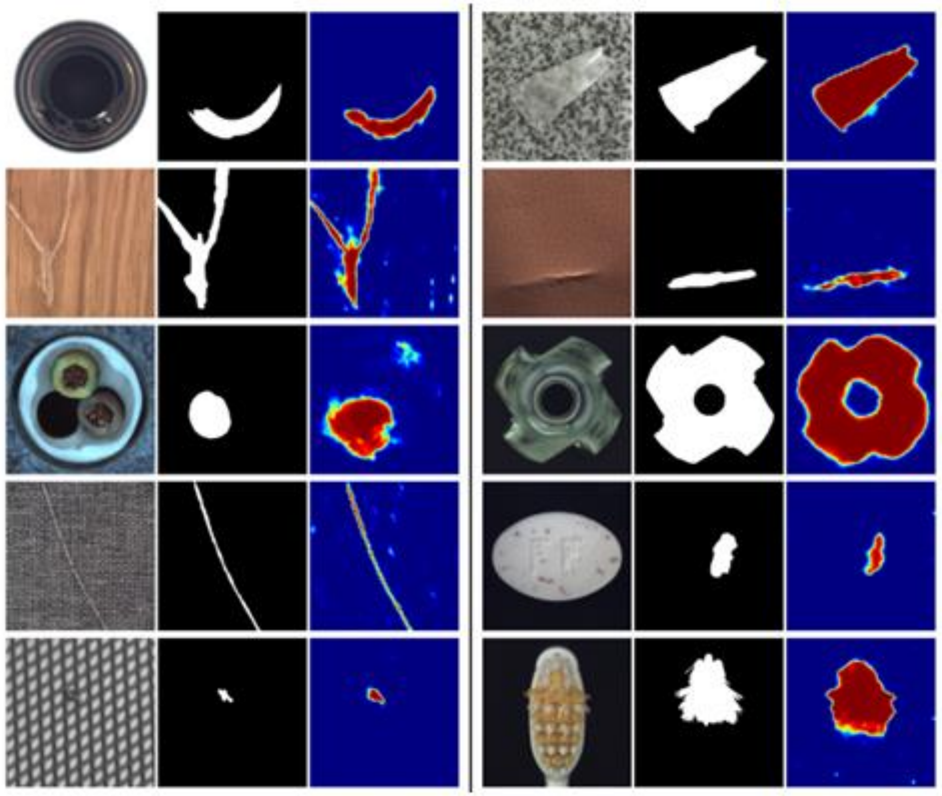


Figure 3. Visualization examples of our method. For each example, left: input image; middle: ground truth; right: prediction map.

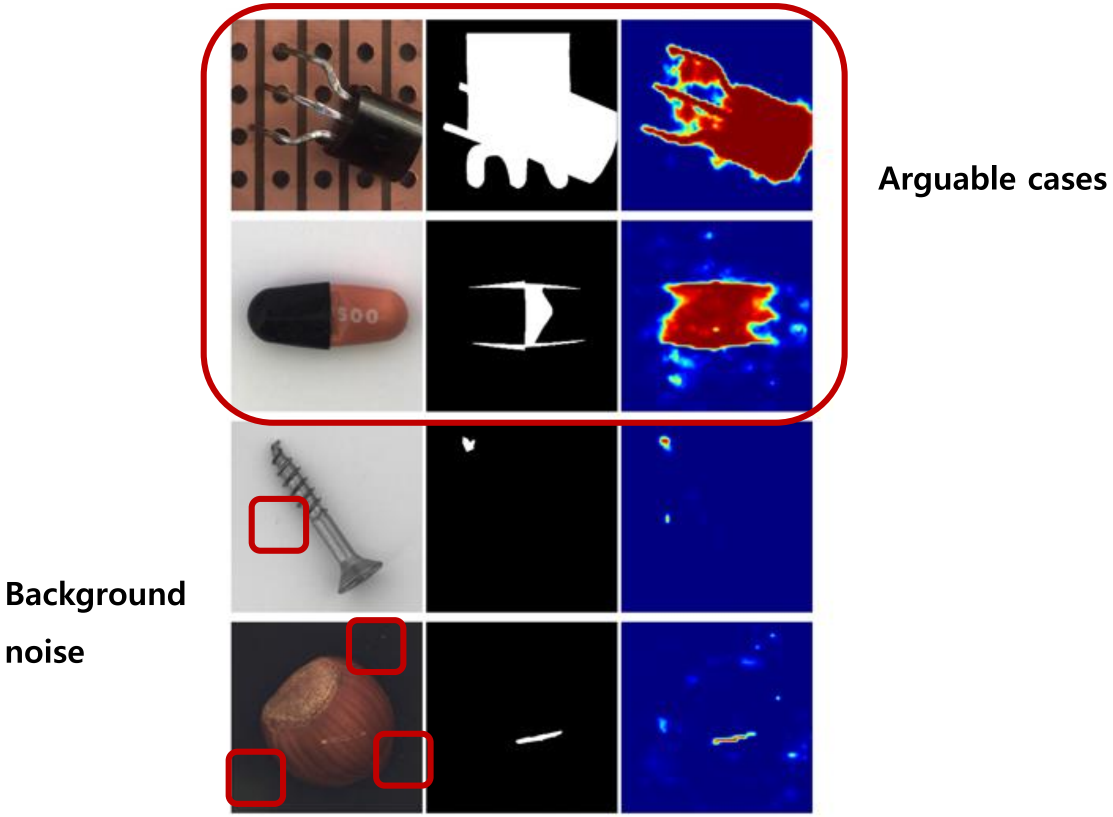


Figure 4. Failure cases of our method. The examples are chosen from transistor, capsule, screw, and hazelnut (from top to bottom). For each example, left: input image; middle: ground truth; right: prediction map.

4. Ablation studies

Exp.	den	ed	seg	img (AUC)	pix (AP)	ins (IAP)
1				94.8	52.9	55.8
2	✓			93.4	49.6	53.9
3		✓		95.4	53.3	57.7
4			✓	97.3	70.1	71.8
5	✓	✓		94.5	54.0	58.5
6	✓		✓	97.3	70.9	72.3
7		✓	✓	97.7	69.7	71.2
8	✓	✓	✓	**98.6**	**75.8**	**76.4**

Table 4. Ablation studies on our main designs: denoising training (den), the encoder-decoder architecture of student network (ed), and segmentation network (seg). AUC, AP, and IAP (%) are used to evaluate image-level, pixel-level, and instance-level detection, respectively. Exp. 1 uses the same architecture of [31], but different training settings to align with Exp. 2∼8.

	img (AUC)	pix (AP)	ins (IAP)
w/o L1 loss	97.9	72.2	74.4
w/ L1 loss	**98.6**	**75.8**	**76.4**

Table 5. Ablation studies on the segmentation loss: AUC, AP, and IAP (%) are used to evaluate image-level, pixel-level, and instance-level detection, respectively.

4. Ablation studies

	img (AUC)	pix (AP)	ins (IAP)
concatenated-ST input	98.0	72.2	72.6
cosine-distance input	98.5	72.0	74.5
DeSTSeg	**98.6**	**75.8**	**76.4**

Table 6. Ablation studies on the input of segmentation network: AUC, AP, and IAP (%) are used to evaluate image-level, pixel-level, and instance-level detection, respectively.

Direct concatenation of feature maps of S-T networks

Computing cosine distance of S-T network's feature map

$$X_k(i,j) = \frac{F_{T_k}(i,j) \odot F_{S_k}(i,j)}{||F_{T_k}(i,j)||_2 ||F_{S_k}(i,j)||_2} \quad (2)$$

$$D_k(i,j) = 1 - \sum_{c=1}^{C_k} X_k(i,j)_c \quad (3)$$

Element-wise product between feature maps of S-T

$$X_k(i,j) = \frac{F_{T_k}(i,j) \odot F_{S_k}(i,j)}{||F_{T_k}(i,j)||_2 ||F_{S_k}(i,j)||_2} \quad (2)$$