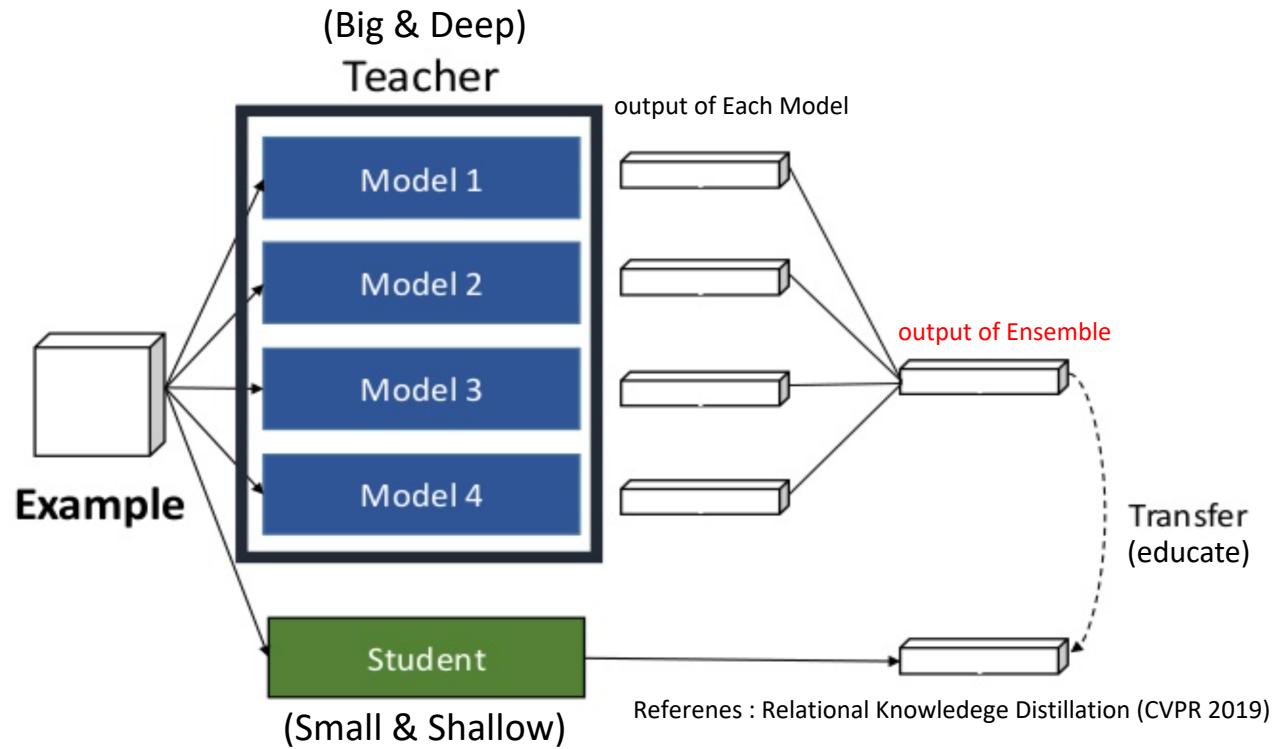
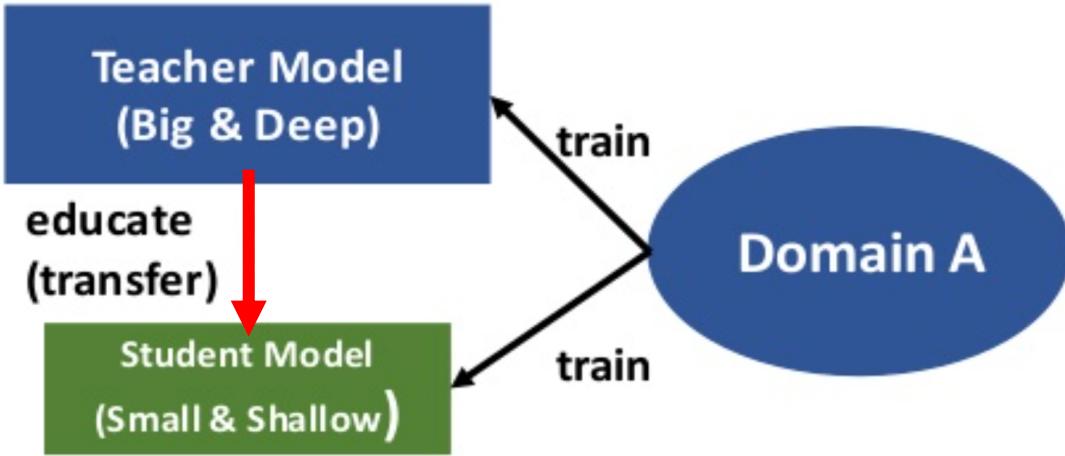


Uniformed students: student-teacher anomaly detection with discriminative latent embeddings

(CVPR 2020)

What is Student-Teacher learning?

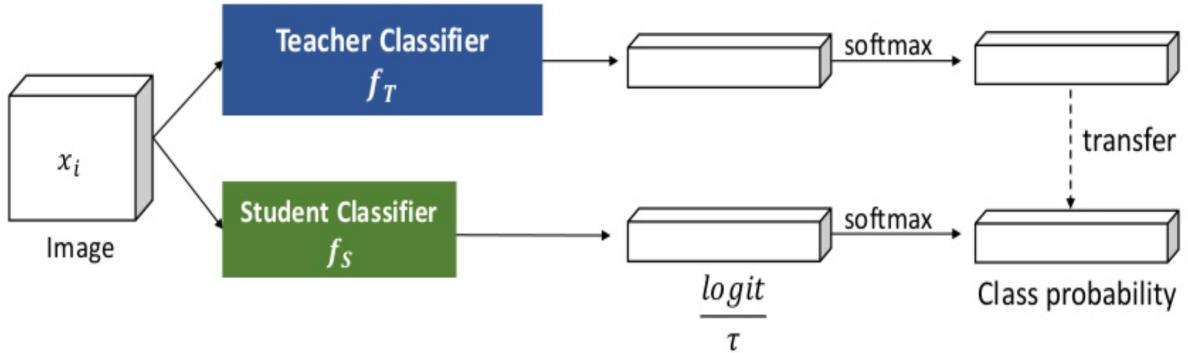


Knowledge Distillation (지식 증류)

Teacher-Student learning은 같은 도메인에서 Teacher network를 pretrain한 다음, 채널수가 적고 얇은 모델인 Student network에 Teacher 지식을 transfer한다. Teacher를 뛰어넘는 성능으로 향상시키기 위한 방법이며 각 모델의 output을 양상불한 결과를 student에 transfer하는 방법으로 student와 teacher와 comparable한 성능을 나타낼 수 있다.

- Distilling the Knowledge in a Neural Network (NIPS 2014)

What is Student-Teacher learning?



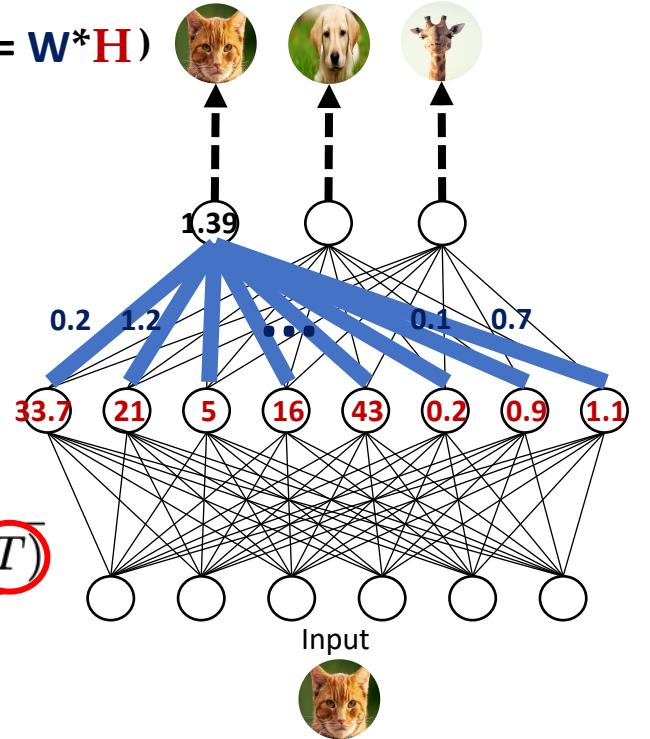
$$\text{Objective: } \sum_{x_i \in \mathcal{X}} \text{KL}\left(\text{softmax}\left(\frac{f_T(x_i)}{\tau}\right), \text{softmax}\left(\frac{f_S(x_i)}{\tau}\right)\right)$$

T (Temperature) : Scaling 역할.

$T = 1$ 이면, 기존 softmax function과 동일. T 가 클수록, 더 soft한 확률분포를 띈다. Temperature는 기존 softmax 함수의 큰 확률은 크게, 작은 확률은 작게 만들어 Hard한 확률 결과를 내는 것을 완화한다. Temperature를 구하는 방법은 따로 존재하지 않고 모든 경우를 실험하여 가장 적합한 T 를 찾는다.

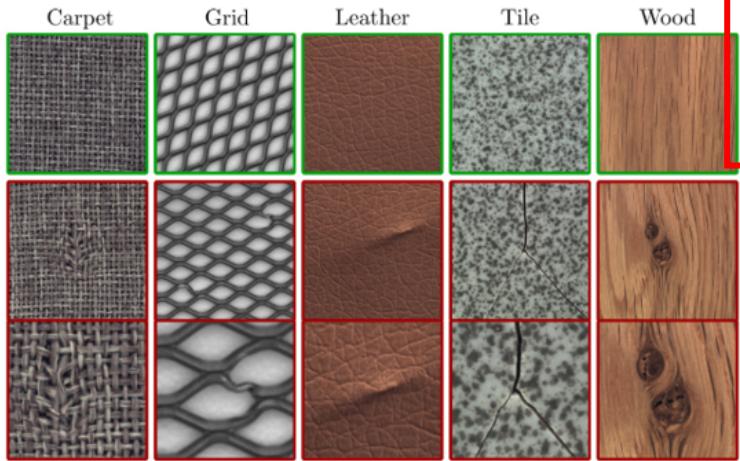


$$O = \text{activation}(\text{Logit} = W^*H)$$

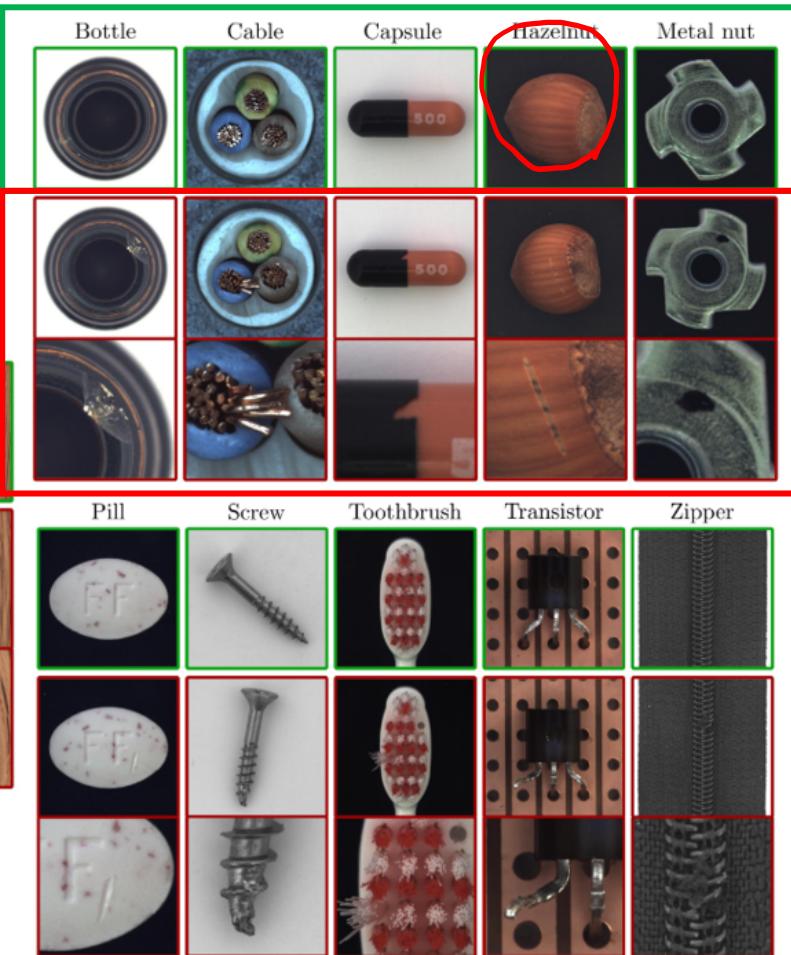


MVTech-AD dataset

Grid, Screw, Zipper = Gray(1ch)



Texture datasets



Objective datasets

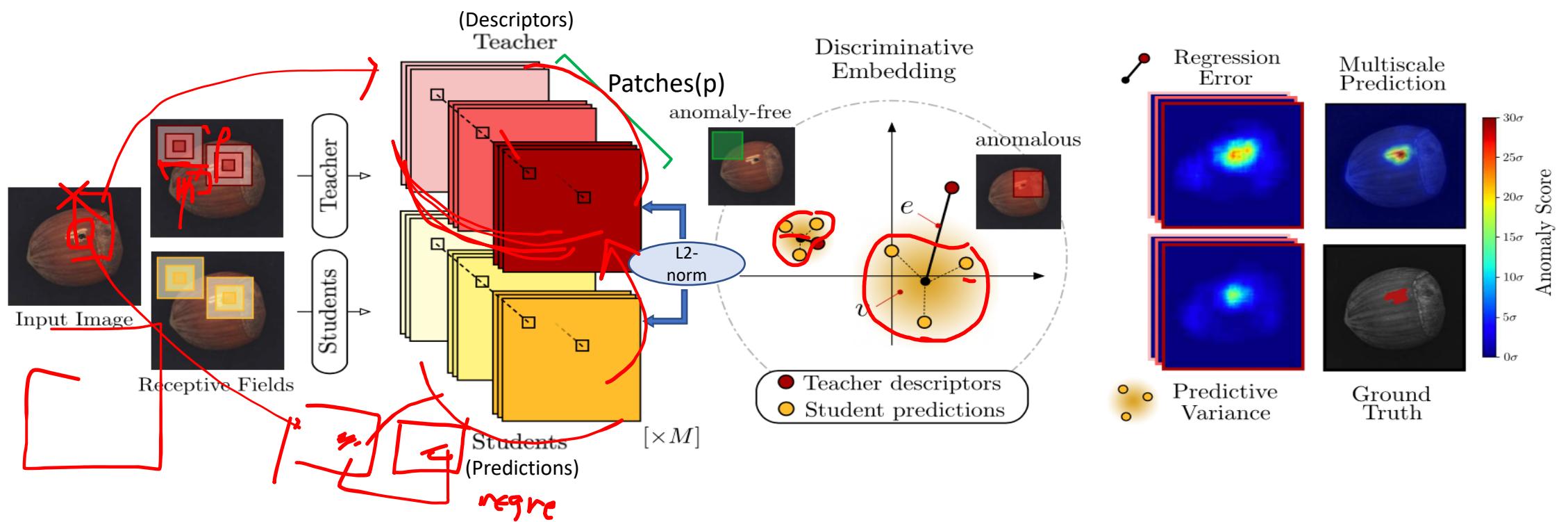
[examples of MVTech-AD dataset]

Anomaly-free(Train / Test)

Anomaly(Test)

Category	# Train	# Test (good)	# Test (defective)	# Defect groups	# Defect regions	Image side length
Textures	Carpet	280	89	5	97	1024
	Grid	264	57	5	170	1024
	Leather	245	92	5	99	1024
	Tile	230	84	5	86	840
	Wood	247	60	5	168	1024
Objects	Bottle	209	63	3	68	900
	Cable	224	92	8	151	1024
	Capsule	219	23	5	114	1000
	Hazelnut	391	40	4	136	1024
	Metal Nut	220	93	4	132	700
	Pill	267	26	7	245	800
	Screw	320	41	5	135	1024
	Toothbrush	60	12	1	66	1024
	Transistor	213	60	4	44	1024
	Zipper	240	32	7	177	1024
Total	3629	467	1258	73	1888	-

Overview of OOD approach Process

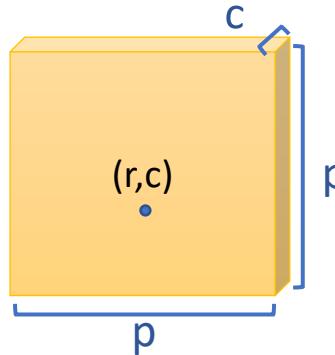


- Step 1.** 결함이 없는(anomaly-free dataset)로 Training으로 줌. 사전 훈련 된 Teacher network(T)는 여러 스케일로 local descriptors (설명자)를 추출.
- Step 2.** Student network의 앙상블은 교사 네트워크의 결과를 regress(회귀)하도록 훈련한다. 그래서 Student는 anomaly-free 이미지들의 descriptors를 예측하는데 전문가가 된다.
- Step 3.** 하지만 Student는 anomalous 영역에서 descriptors를 훈련 중에 관찰하지 않았기 때문에 anomalous 영역은 정확하게 예측하지 못한다. 이로 인해 Student network의 앙상블에서 regression error(회귀 오류)와 predictive variance(예측 분산)이 증가한다.

1. Learning local patch descriptors(Training Teacher T)

1) Knowledge distillation

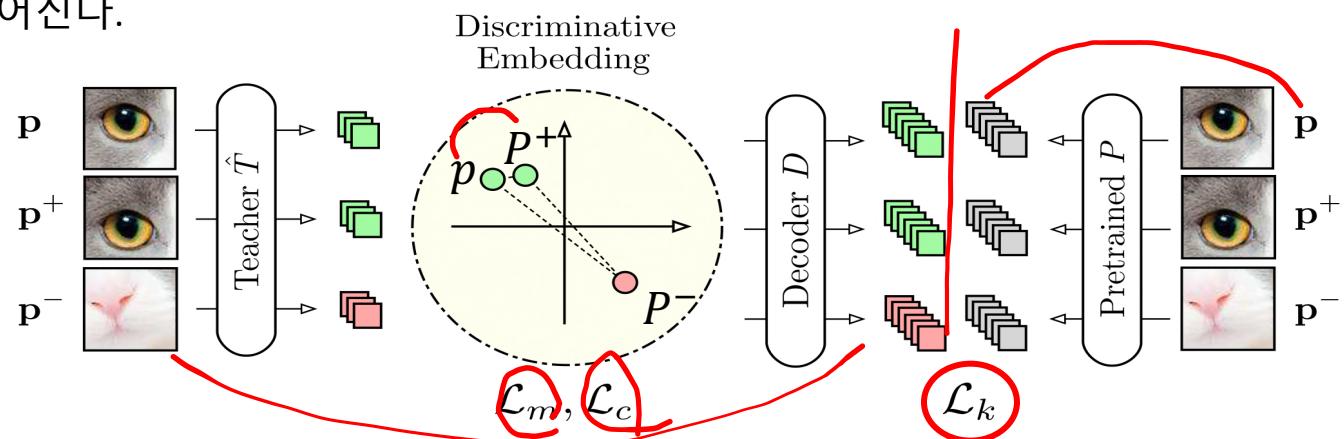
feature extractor는 오직 patch 사이즈의 input을 위한 single feature vectors 혹은 spatially heavily downsample된 feature map을 출력한다. T는 첫번째 학습 네트워크 T^* 를 훈련하여 patch-sized 이미지 $p \in R^{p \times p \times C}$ 를 convolution과 max-pooling layer만 사용하여 차원 d의 demension으로 metric space으로 embedding하여 얻어진다.



Patch-sized image $p \in R^{p \times p \times C}$

Anomaly-free dataset $D = \{I_1, I_2, \dots, I_n\}$

* p 는 image db에서 random crop해서 얻음.



[First loss]

$$\mathcal{L}_k(\hat{T}) = \underbrace{\|D(\hat{T}(p)) - P(p)\|_2^2}_{\text{Term.1}} + \underbrace{\|P(p) - p\|_2^2}_{\text{Term.2}}$$

- Term 1 : D 는 T^* 의 d차원인 output을 pretrain된 Network descriptor의 output 차원으로 decoding한 fully-connected된 네트워크.
- Term 2 : pretrained P에서 나온 output.

결국에는 T^* L_k 를 최소하기 위해 transfer 함.

1. Learning local patch descriptors (Training Teacher T)

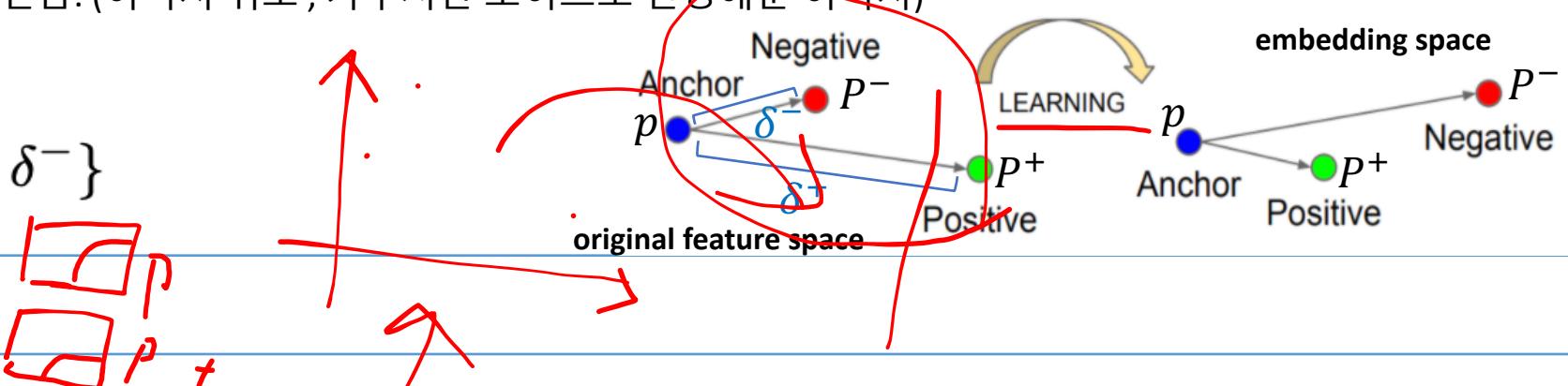
2) Metric learning (Triplet Embedding)

Embedding을 위한 triplet loss는 주어진 데이터 셋에서 선택된 데이터 anchor(p), 그리고 anchor와 동일한 class label을 가지는 positive sample(p^+), 다른 class label을 가지는 negative sample(p^-)로 정의된다. positive 와 negative sample의 특성을 서로 멀어지게 학습시킨다.

p^+ : p 주위의 small random translation으로 얻음. (이미지 휘도, 가우시안 노이즈로 변경해준 이미지)

p^- : 랜덤하게 크롭된 다른 이미지.

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



3) Descriptor Compactness

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

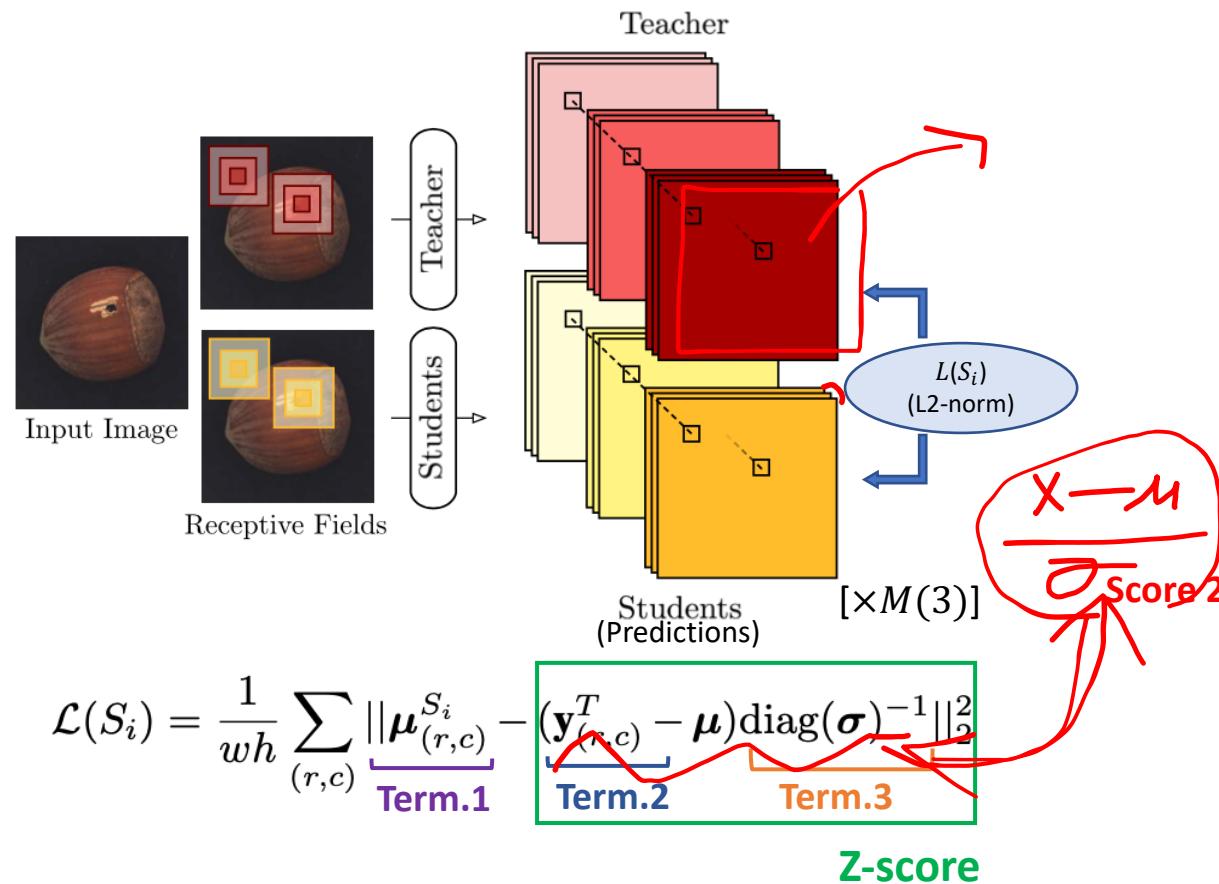
여기서 c_{ij} 는 현재 미니 배치의 모든 descriptors $\hat{T}(p)$ 에 대해 계산된 correlation matrix의 항목.
 p 의 한 미니 배치안의 descriptor의 압축성을 높이고 패치간의 불필요한 중복성을 제거하기 위해 상관관계를 최소화.

Final loss for T

$$\mathcal{L}(\hat{T}) = \lambda_k \mathcal{L}_k(\hat{T}) + \lambda_m \mathcal{L}_m(\hat{T}) + \lambda_c \mathcal{L}_c(\hat{T}) \quad \lambda_k, \lambda_m, \lambda_c \geq 0$$

은 각 loss term을 위한 weight factors.

2. Ensemble of student networks for deep anomaly detection



[Scoring Function for Anomaly Detection]

$$e_{(r,c)} = \|\boldsymbol{\mu}_{(r,c)} - (\mathbf{y}_{(r,c)}^T - \boldsymbol{\mu}) \text{diag}(\boldsymbol{\sigma})^{-1}\|_2^2 \\ = \left\| \frac{1}{M} \sum_{i=1}^M \boldsymbol{\mu}_{(r,c)}^{S_i} - (\mathbf{y}_{(r,c)}^T - \boldsymbol{\mu}) \text{diag}(\boldsymbol{\sigma})^{-1} \right\|_2^2.$$

Score 1. $e(r,c)$: regression error

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

Score 2. $v(r,c)$: each pixel the predictive uncertainty of the Gaussian mixture (variance error)

score 1 + score 2 = Final score

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

- Term 1 : (r,c)에서의 student S_i 를 통해 만들어진 예측 값의 평균.
- Term 2 : (r,c)에서의 학생들이 예측할 teacher의 descriptor . ($\boldsymbol{\mu} \in R^d$: the vector of component-wise means)
- Term 3 : 표준편차 값으로 채워진 대각 행렬의 역행렬. ($\boldsymbol{\sigma} \in R^d$: the standard deviations)

3. multi-scale anomaly segmentation

Anomaly가 Teacher의 receptive field 크기 p 의 작은 부분만 포함하는 경우, **추출된 특징 벡터는 주로 로컬 이미지 영역의 이상이 없는 특성을 설명**한다. 결과적으로 descriptor는 학생들이 잘 예측할 수 있으며 이상 탐지 성능이 저하된다. 이는 입력 이미지를 다운샘플링하여 이 문제를 해결할 수 있다.

다양한 p 값으로 여러 S-T Network ensemble 쌍을 훈련하여 다양한 규모의 이상을 감지 할 수 있다. 각 규모에서 input image와 같은 크기의 anomaly map이 계산된다.

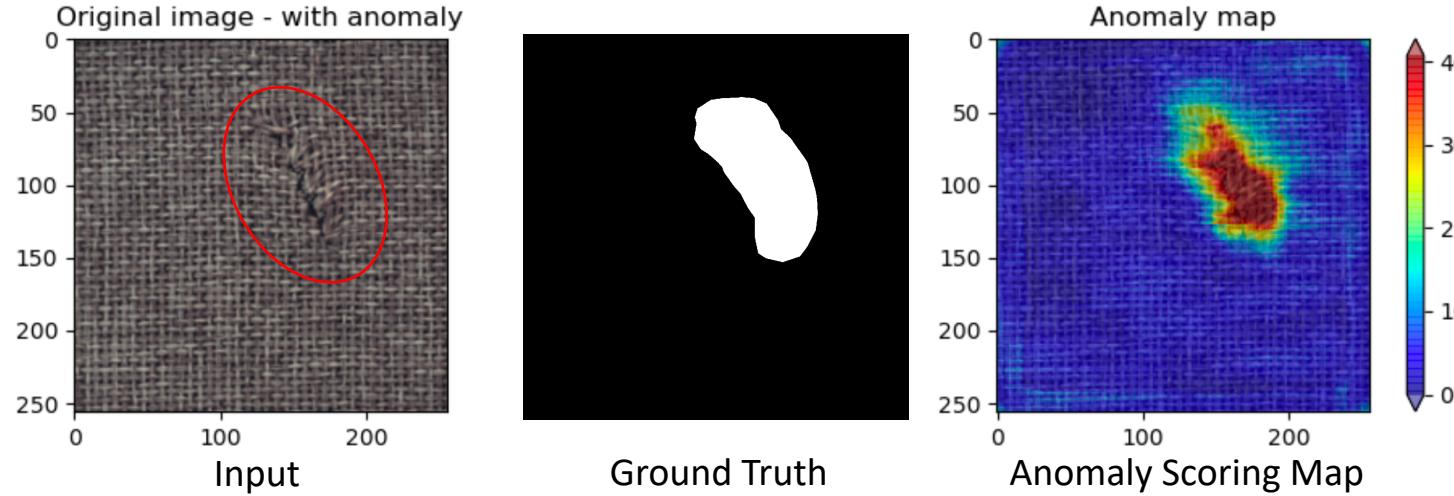
다른 Receptive field를 가진 학생-교사 양상을 쌍을 L 이라고 주면, 각 척도 $|l|$ 의 정규화 된 이상 점수 $e \sim (r, c)$ 및 $v \sim (r, c)$ 는 간단한 평균화로 결합될 수 있다.

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

S-T ensemble pairs with different receptive field

Evaluation Metric

PRO(Per-Region-Overlap) Metric



PRO는 서로 다른 크기의 Ground-Truth 영역에 동일하게 가중치를 부여하는 영역 별 중첩을 기반으로 threshold을 가진 Evaluation Metric이다. 이상 부분을 계산하기 위해 먼저 각 픽셀에 이상에 대한 여부를 binary로 결정한다. Ground Truth 내의 각 연결된 구성 요소에 대해 threshold 이상 영역과의 relative overlap으로 계산된다. 전체 데이터 세트에 대해 픽셀 당 false-positive rate 가 30 %에 도달 할 때까지 증가하는 threshold를 정하고 PRO 곡선 아래 영역을 이상 탐지 성능의 측정 값으로 사용한다.

*FPR (False-Positive Rate) : 원래 이상이 아닌데 이상이라고 판정한 비율.

Experiments

[MVTec Anomaly Detection Dataset]

실험에서는 receptive field 크기가 $p \in \{17, 33, 65\}$ 인 S-T Network에 대해 서로 동일한 네트워크 구조를 사용.
모든 구조는 convolution layer와 max-pooling만 있는 단순한 CNN으로, ReLU activation을 사용.

Layer	Output Size	Parameters	
		Kernel	Stride
Input	$65 \times 65 \times 3$		
Conv1	$61 \times 61 \times 128$	5×5	1
MaxPool	$30 \times 30 \times 128$	2×2	2
Conv2	$26 \times 26 \times 128$	5×5	1
MaxPool	$13 \times 13 \times 128$	2×2	2
Conv3	$9 \times 9 \times 128$	5×5	1
MaxPool	$4 \times 4 \times 256$	2×2	2
Conv4	$1 \times 1 \times 256$	4×4	1
Conv5	$1 \times 1 \times 128$	3×3	1
Decode	$1 \times 1 \times 512$	1×1	1

Category	Ours $p = 65$	1-NN	OC-SVM	K-Means	ℓ_2 -AE	VAE	SSIM-AE	AnoGAN	CNN-Feature Dictionary
Textures	0.695	0.512	0.355	0.253	0.456	0.501	0.647	0.204	0.469
	0.819	0.228	0.125	0.107	0.582	0.224	0.849	0.226	0.183
	0.819	0.446	0.306	0.308	0.819	0.635	0.561	0.378	0.641
	0.912	0.822	0.722	0.779	0.897	0.870	0.175	0.177	0.797
	0.725	0.502	0.336	0.411	0.727	0.628	0.605	0.386	0.621
Objects	0.918	0.898	0.850	0.495	0.910	0.897	0.834	0.620	0.742
	0.865	0.806	0.431	0.513	0.825	0.654	0.478	0.383	0.558
	0.916	0.631	0.554	0.387	0.862	0.526	0.860	0.306	0.306
	0.937	0.861	0.616	0.698	0.917	0.878	0.916	0.698	0.844
	0.895	0.705	0.319	0.351	0.830	0.576	0.603	0.320	0.358
	0.935	0.725	0.544	0.514	0.893	0.769	0.830	0.776	0.460
	0.928	0.604	0.644	0.550	0.754	0.559	0.887	0.466	0.277
	0.863	0.675	0.538	0.337	0.822	0.693	0.784	0.749	0.151
	0.701	0.680	0.496	0.399	0.728	0.626	0.725	0.549	0.628
	0.933	0.512	0.355	0.253	0.839	0.549	0.665	0.467	0.703
	0.857	0.640	0.479	0.423	0.790	0.639	0.694	0.443	0.515
	Mean								

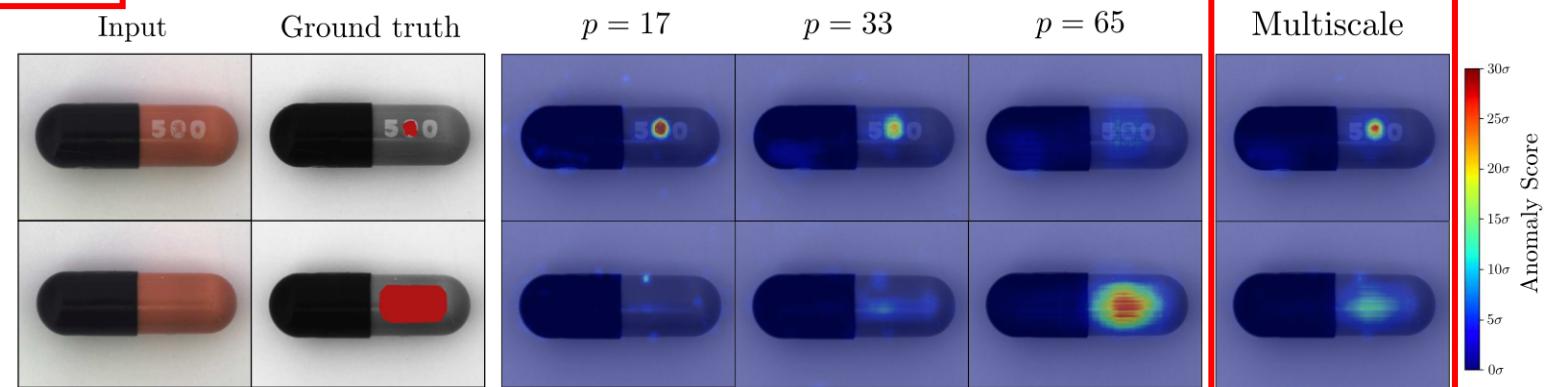
[(왼쪽 table) Outline of network architecture for training teachers T^* with receptive field $p = 65$]

Experiments

[MVTec Anomaly Detection Dataset]

	Category	$p = 17$	$p = 33$	$p = 65$	Multiscale
Textures	Carpet	0.795	0.893	0.695	0.879
	Grid	0.920	0.949	0.819	0.952
	Leather	0.935	0.956	0.819	0.945
	Tile	0.936	0.950	0.912	0.946
	Wood	0.943	0.929	0.725	0.911
Objects	Bottle	0.814	0.890	0.918	0.931
	Cable	0.671	0.764	0.865	0.818
	Capsule	0.935	0.963	0.916	0.968
	Hazelnut	0.971	0.965	0.937	0.965
	Metal nut	0.891	0.928	0.895	0.942
	Pill	0.931	0.959	0.935	0.961
	Screw	0.915	0.937	0.928	0.942
	Toothbrush	0.946	0.944	0.863	0.933
	Transistor	0.540	0.611	0.701	0.666
	Zipper	0.848	0.942	0.933	0.951
Mean		0.866	0.900	0.857	0.914

Multiscale한 결과의 정확도가 가장 높음.



References

- Distilling the Knowledge in a Neural Network / Geoffrey Hinton-Google Inc. (NIPS 2014)
- MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection / MVTech Software (CVPR 2019)
- Uninformed Students: Student–Teacher Anomaly Detection with Discriminative Latent Embeddings / MVTech Software (CVPR 2020)

Etc

- Relational Knowledge Distillation / POSTECH (CVPR 2019)
- L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space (CVPR 2017)
- Deep Transfer Learning for Multiple Class Novelty Detection / Johns Hopkins University (CVPR 2019)
- OCGAN: One-class Novelty Detection Using GANs with Constrained Latent Representations / Johns Hopkins University, AWS AI (CVPR 2019)
- WAIC, but Why? Generative Ensembles for Robust Anomaly Detection / Google Inc. (CVPR 2019)
- Deep One-Class Classification / Humboldt University of Berlin, Germany. (ICML 2018)

Thanks

Do you have any questions?

Email : kka1132@hanyang.ac.kr
Github : <https://github.com/Kyeonga-Kim/>