InCTRL (CVPR 24)

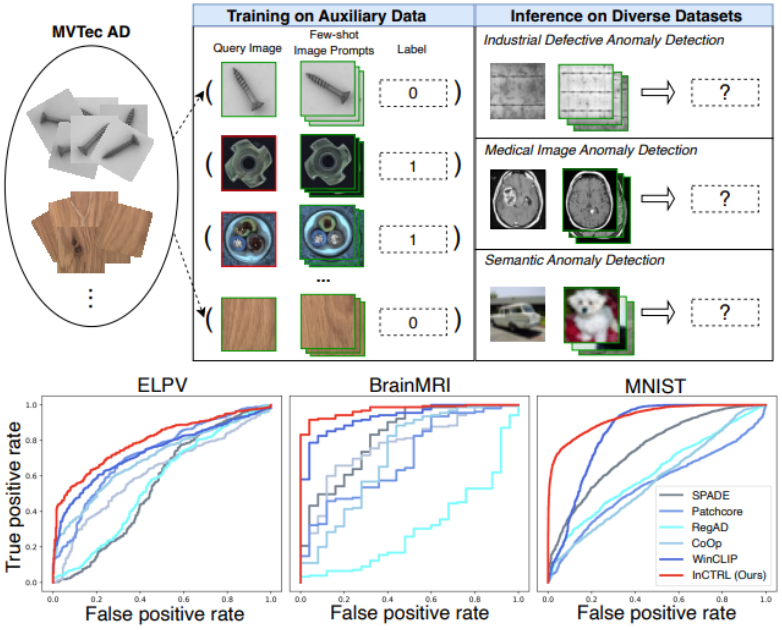
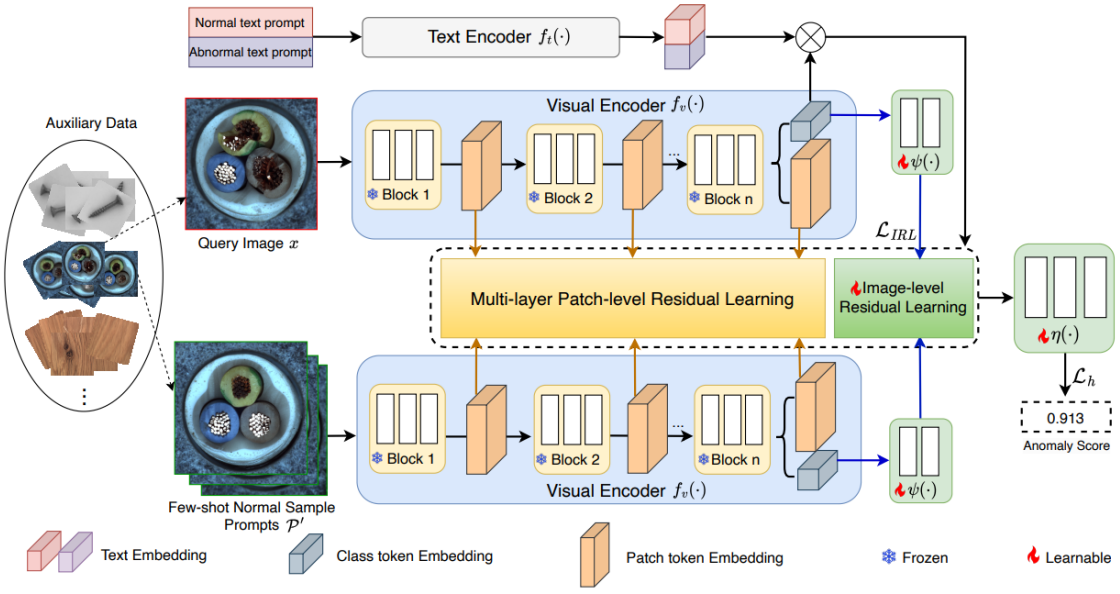


Figure 1. **Top**: An illustration of InCTRL: a one-for-all model using few-shot normal images as sample prompts. **Bottom**: AUROC curves of InCTRL and competing few-shot methods on three different application datasets without any training on the target data.

Auxiliary Data로 학습하여 학습할 때 보지 못한 Diverse Datasets으로 few-shot anomaly detection을 할 수 있는 방법을 제안함.

이 논문에서는 '비정상일수록 정상 데이터(few-normal shots)와 차이(residual)가 클 것'을 가정함.



1. **Multi-Layer Patch-Level Residual learning**

$$\mathbf{M}_x^l(i,j) = 1 - \langle T_x^l(i,j), h(T_x^l(i,j)|\mathcal{P}')\rangle, \qquad (1)$$

where $h(T_x^l(i,j)|\mathcal{P}')$ returns the embedding of the patch token that is most similar to $T_x^l(i,j)$ among all image patches in $\mathcal{P}'$, and $\langle\cdot\rangle$ is the cosine similarity function. The final patch-level residual map $\mathbf{M}_x \in \mathbb{R}^{h\times w}$ is averaged over n layer-wise residual maps:

fine-grained한 residual을 포착하고자 각 스테이지에 대한 임베딩을 활용함. query patch와 해당 patch와 가장 유사한 few-shot patch 의 cosine simialiry를 구하고 1에서 뺌. 이를 통해 local한 수준에서 이상 점수 맵(M)를 확인할 수 있음

$$\mathbf{M}_x = \frac{1}{n}\sum_{l=1}^{n}\mathbf{M}_x^l.$$

각 레이어에서 구한 M을 평균내서 최종 M을 제작함. 해당 맵은 비정상 부분일수록 값이 크고(최대 2), 정상 부분일수록 값이 작음(최소 0)

2. **Image-level Residual Learning**

$$\mathbf{I}_p = \frac{1}{K} \sum_{x'_k \in \mathcal{P}'} \psi(f_v(x'_k); \Theta_\psi), \tag{3}$$

where $\mathbf{I}_p \in \mathbb{R}^{d'}$. Then let $\mathbf{I}_x = \psi(f_v(x); \Theta_\psi)$ be the adapted features of the query image x, the in-context image-level residual features $\mathbf{F}_x$ for x are obtained by performing element-wise subtraction between two feature maps:

$$\mathbf{F}_x = \mathbf{I}_x \ominus \mathbf{I}_p, \tag{4}$$

where $\ominus$ denotes element-wise subtraction. Subsequently, these in-context residual features are fed to an image-level anomaly classification learner $\eta : \mathbf{F}_x \rightarrow \mathbb{R}$, parameterized by Θ_η which is optimized by the binary classification loss:

$$\mathcal{L}_{IRL} = \frac{1}{N} \sum_{x \in X_{train}} \mathcal{L}_b(\eta(\mathbf{F}_x; \Theta_\eta), y_x), \tag{5}$$

where $\mathcal{L}_b$ is a binary classification loss. Focal loss [31] is used by default in our model.

이미지 레벨의 global discriminative한 정보도 이상 탐지에 중요하다고 함. 따라서 Visual Encoder의 class token을 활용하여 few-shot class token들과 학습을 진행함.

이 때, class token은 classification task에 최적화되었으므로, adapter layer로 학습을 진행하는 것임. 학습 방법은, few-shot class token들을 adaper layer에 forward시킨 후 평균지은 feature를 query feature와 뺀 값이 y가 되도록 하는 방법임.

이를 통해 query 이미지와 few-shot 이미지들의 차이가 클수록 1이 되는 것을 기대함.

3. **Fusing Text Prompt-based Prior Knowledge**

Specifically, let $\mathcal{P}_t^n$ be the set of text prompts for the normal class, we use the prototype of the text prompt embeddings to provide a representative embedding of the normal text prompts $\mathbf{F}_n = \frac{1}{|\mathcal{P}_t^n|} \sum_{p_i \in \mathcal{P}_t^n} f_t(p_i)$ where $p_i \in \mathcal{R}^{d'}$; similarly we can obtain the prototype embedding for the abnormality text prompt set $\mathcal{P}_t^a$ by $\mathbf{F}_a = \frac{1}{|\mathcal{P}_t^a|} \sum_{p_j \in \mathcal{P}_t^a} f_t(p_j)$. Then, InCTRL extracts an AD-oriented discriminative feature based on the similarity between the query image x and the two prototypes of the text prompts:

$$s_a(x) = \frac{\exp(\mathbf{F}_a^\intercal f_v(x))}{\exp(\mathbf{F}_n^\intercal f_v(x)) + \exp(\mathbf{F}_a^\intercal f_v(x))}, \tag{6}$$

where $[\cdot]^\intercal$ denotes a transpose operation, and $s_a(x)$ is the probability of the input x being classified as abnormal.

patch, image level에서 residual learning을 하는 것만으로는 해당 이미지가 정상에 해당하는지 비정상에 해당하는지에 대한 semantic한 정보를 획득하기 어렵다고 함.

따라서 normal, abnormal text와 class token간의 유사도를 추가로 고려하여, 입력 query가 정상에 가까운지, 비정상에 가까운지 판단하는 과정을 거침.

4. **In-Context Residual Learning**

In-Context Residual Learning. During training, InCTRL performs a holistic residual learning that synthesizes both patch-level and image-level residual information, augmented by the text prompt-guided features. The holistic in-context residual map of a query image x is defined as:

$$\mathbf{M}_x^+ = \mathbf{M}_x \oplus s_i(x) \oplus s_a(x), \tag{7}$$

where $s_i(x) = \eta(\mathbf{F}_x; \Theta_\eta)$ is an anomaly score based on the image-level residual map $\mathbf{F}_x$ and $\oplus$ denotes an element-wise addition. InCTRL then devises a holistic anomaly scoring function ϕ, parameterized by Θ_ϕ, based on $\mathbf{M}_x^+$, and defines the final anomaly score as:

$$s(x) = \phi(\mathbf{M}_x^+; \Theta_\phi) + \alpha s_p(x), \tag{8}$$

where $\phi(\mathbf{M}_x^+; \Theta_\phi)$ performs a holistic anomaly scoring using patch-, image-level and text prompt-guided features, while $s_p(x) = \max(\mathbf{M}_x)$ is a maximum residual score-based fine-grained anomaly score at the image patch level.

$$\mathcal{L}_h = \frac{1}{N} \sum_{x \in X_{train}} \mathcal{L}_b(s(x), y_x). \tag{9}$$

patch level에서 구한 M과 Image level에서 구한 F_x, 그리고 text level에서 구한 s_a는 모두 비정상일수록 높은 값을 가진다는 특징이 있음. 이 모든 값들을 합쳐서 M+_x를 제작 후 임베딩 레이어를 거쳐서 final anomaly score를 계산함. 이 때, s_p(x)는 M에서 가장 큰 값(가장 anomaly한 값)이며, 이를 더하는 이유는 local한 abnomral region을 고려하여 작은 영역을 간과하지 않기 위함임.

final anomaly score는 y_x가 되도록 학습을 진행함. 이 목적함수 역시 값이 클수록 1에 가깝게 학습시키겠다는 목적을 지님.

$$\mathcal{L}_{InCTRL} = \mathcal{L}_{IRL} + \mathcal{L}_h.$$

마지막으로, 최종 loss는 위와 같음.

Setup	Methods	Industrial Defects					Medical Anomalies		Semantic Anomalies			
									One-vs-all		Multi-class	
		ELPV	SDD	AITEX	VisA	MVTec AD	BrainMRI	HeadCT	MNIST	CIFAR-10	MNIST	CIFAR-10
Baseline (0-shot)		0.855±0.000	0.886±0.000	0.552±0.000	0.812±0.000	0.957±0.000	0.988±0.000	0.970±0.000	0.940±0.000	0.990±0.000	0.606±0.000	0.852±0.000
2-shot	SPADE	0.618±0.007	0.366±0.105	0.470±0.008	0.818±0.031	0.922±0.023	0.952±0.009	0.851±0.022	**0.965±0.004**	0.971±0.003	**0.615±0.068**	0.502±0.035
	PaDiM	0.707±0.058	0.337±0.008	**0.529±0.034**	0.719±0.027	0.890±0.015	0.902±0.046	0.876±0.017	-	-	-	-
	Patchcore	0.840±0.031	0.676±0.003	0.378±0.008	0.841±0.023	0.939±0.012	0.921±0.017	0.913±0.002	0.956±0.001	0.926±0.002	0.482±0.025	0.574±0.015
	RegAD	0.679±0.005	0.173±0.019	0.275±0.035	0.614±0.037	0.837±0.034	0.872±0.065	0.854±0.009	0.913±0.006	0.909±0.003	0.612±0.013	0.672±0.008
	CoOp	0.841±0.020	0.543±0.004	0.443±0.050	0.835±0.019	0.922±0.007	0.923±0.002	0.937±0.014	0.926±0.003	0.911±0.002	0.607±0.009	0.371±0.013
	WinCLIP	**0.849±0.010**	**0.865±0.004**	0.500±0.043	**0.859±0.021**	**0.965±0.007**	**0.989±0.003**	**0.975±0.012**	0.963±0.001	**0.990±0.001**	0.614±0.005	**0.876±0.016**
	Ours (InCTRL)	**0.913±0.008**	**0.917±0.009**	**0.519±0.022**	**0.877±0.016**	**0.969±0.004**	**0.994±0.013**	**0.981±0.013**	**0.975±0.004**	**0.992±0.000**	**0.618±0.012**	**0.899±0.010**
4-shot	SPADE	0.627±0.011	0.385±0.018	0.451±0.031	0.826±0.024	0.924±0.015	0.958±0.017	0.854±0.016	0.966±0.008	0.973±0.002	**0.611±0.053**	0.487±0.047
	PaDiM	0.724±0.067	0.351±0.012	**0.540±0.053**	0.758±0.018	0.909±0.013	0.956±0.011	0.890±0.011	-	-	-	-
	Patchcore	**0.871±0.042**	0.703±0.013	0.377±0.001	0.860±0.016	0.950±0.013	0.945±0.017	0.941±0.009	**0.972±0.002**	0.934±0.003	0.504±0.025	0.606±0.010
	RegAD	0.688±0.018	0.176±0.003	0.294±0.031	0.628±0.034	0.846±0.026	0.900±0.041	0.810±0.028	0.916±0.013	0.908±0.001	0.522±0.085	0.681±0.127
	CoOp	0.867±0.003	0.594±0.014	0.454±0.014	0.842±0.016	0.924±0.008	0.932±0.013	0.957±0.017	0.929±0.002	0.915±0.003	0.611±0.003	0.374±0.012
	WinCLIP	0.864±0.004	**0.868±0.003**	0.513±0.017	**0.875±0.023**	**0.968±0.008**	**0.990±0.001**	**0.974±0.002**	0.971±0.002	**0.990±0.000**	**0.611±0.011**	**0.882±0.009**
	Ours (InCTRL)	**0.916±0.009**	**0.924±0.015**	**0.548±0.016**	**0.902±0.027**	**0.972±0.006**	**0.994±0.013**	**0.984±0.011**	**0.980±0.007**	**0.992±0.004**	**0.620±0.004**	**0.901±0.020**
8-shot	SPADE	0.641±0.018	0.394±0.024	0.427±0.008	0.844±0.031	0.930±0.016	0.962±0.014	0.860±0.019	0.974±0.002	0.976±0.001	0.613±0.035	0.515±0.024
	PaDiM	0.798±0.014	0.384±0.045	0.555±0.031	0.781±0.024	0.927±0.012	0.946±0.007	0.896±0.009	-	-	-	-
	Patchcore	**0.915±0.007**	0.708±0.009	0.389±0.003	0.873±0.022	0.962±0.013	0.957±0.007	0.931±0.006	**0.979±0.001**	0.942±0.002	0.530±0.037	0.635±0.019
	RegAD	0.696±0.015	0.246±0.031	0.314±0.036	0.643±0.032	0.855±0.021	0.908±0.013	0.881±0.014	0.919±0.018	0.911±0.001	0.566±0.048	0.558±0.159
	CoOp	0.905±0.008	0.578±0.001	0.514±0.003	0.848±0.020	0.933±0.007	0.927±0.007	0.965±0.018	0.937±0.004	0.920±0.003	0.610±0.001	0.376±0.003
	WinCLIP	0.897±0.007	**0.865±0.001**	**0.562±0.024**	**0.880±0.021**	**0.973±0.009**	**0.991±0.000**	**0.975±0.003**	0.974±0.001	**0.990±0.000**	**0.616±0.006**	**0.887±0.006**
	Ours (InCTRL)	**0.926±0.006**	**0.925±0.011**	**0.561±0.034**	**0.904±0.025**	**0.977±0.006**	**0.996±0.003**	**0.985±0.005**	**0.989±0.001**	**0.994±0.001**	**0.622±0.008**	**0.912±0.005**

Table 2. AUPRC results(mean±std) on nine real-world AD datasets under various few-shot AD settings. Best results and the second-best results are respectively highlighted in **red** and **blue**. 'Baseline' is a WinCLIP-based zero-shot AD model.

결과는 WinCLIP보다 더 우수한 성능을 보였다고 함. 특히 Medical data에서도 우수한 성능을 보인 것이 놀라움. 하지만 few-shot 데이터들과 유사도 혹은 차이를 이용하므로, zero-shot AD를 하지 못 한다는 점이 아쉬움.