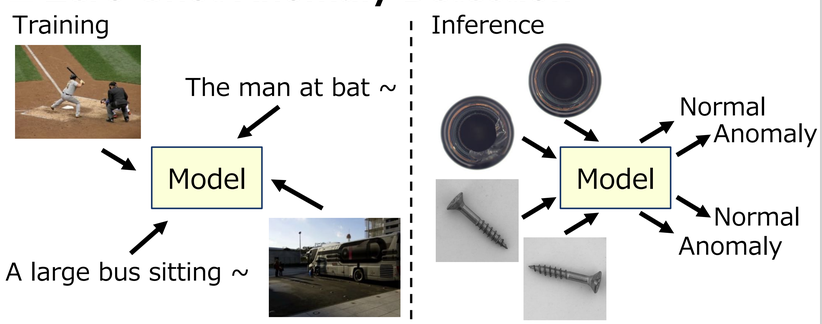
Random Word Data Augmentation with CLIP for Zero-Shot Anomaly Detection

Masato Tamura
Big Data Analytics Solutions Lab., Hitachi America, Ltd.
masato.tamura@hal.hitachi.com

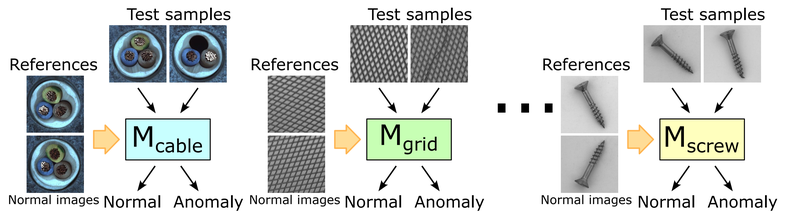
Zero-Shot Anomaly Detection



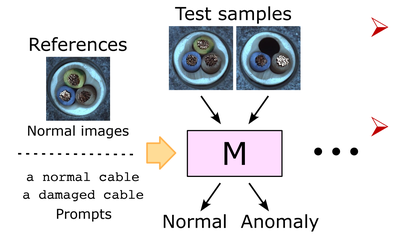
- Anomaly detection (AD) models are trained without target object information and may not be trained for AD.
- During test time, normal and anomaly images of any objects are fed into the trained models.

Motivation

Prior Work [1][2]



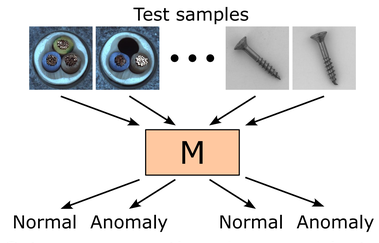
Vanilla AD [1]



- Prior work [1] requires category-specific detectors to detect anomalies.
- Recent work [2] is a category-agnostic method, but target category information is required during test time.

Category-agnostic known object AD [2]

Our Solution

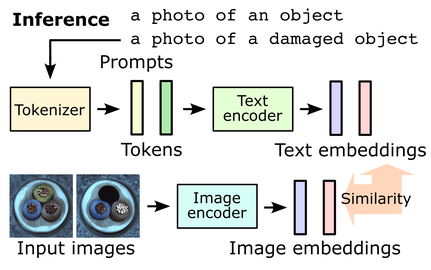


- Our method trains a category-agnostic anomaly detector with CLIP and random word data augmentation.
- The trained detector can be used without category information during test time.

Category-agnostic unknown object AD

Proposed Method

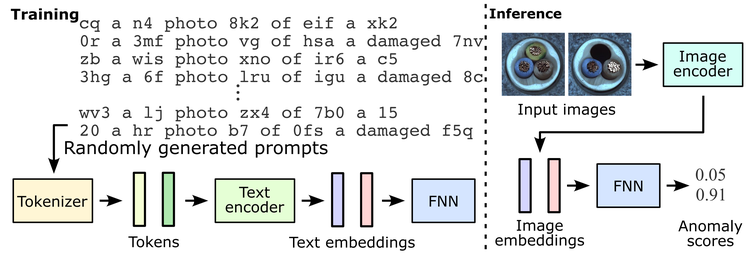
Prompt-guided AD



Prompt-guided AD

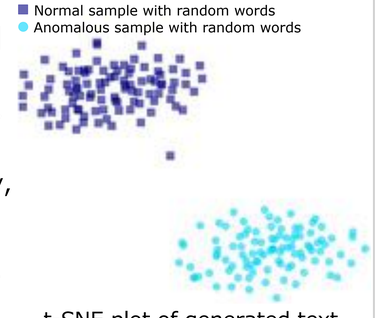
- Normal and anomaly prompts (e.g., "a photo of a normal object" and "a photo of a damaged object") are transformed into tokens $t^{(n)} \in \mathbb{Z}^{C_t}$ and $t^{(a)} \in \mathbb{Z}^{C_t}$ with the tokenizer.
- The tokens are transformed into text embeddings with the text encoder $f_{tenc}(\cdot)$ in CLIP as $e^{(n,t)} = f_{tenc}(t^{(n)})$ and $e^{(a,t)} = f_{tenc}(t^{(a)})$.
- An input image $I \in \mathbb{R}^{3\times H\times W}$ is transformed into an image embedding with the image encoder $f_{ienc}(\cdot)$ in CLIP as $e^{(i)} = f_{ienc}(I)$.
- With the normalized embeddings $\bar{e}^{(n,t)} = \frac{e^{(n,t)}}{|e^{(n,t)}|}$, $\bar{e}^{(a,t)} = \frac{e^{(a,t)}}{|e^{(a,t)}|}$, and $\bar{e}^{(i)} = \frac{e^{(i)}}{|e^{(i)}|}$, an anomaly score $s_{pr} \in [0,1]$ is obtained using the softmax as $s_{pr} = \frac{\exp(\bar{e}^{(a,t)} \cdot \bar{e}^{(i)}/T)}{\exp(\bar{e}^{(n,t)} \cdot \bar{e}^{(i)}/T) + \exp(\bar{e}^{(a,t)} \cdot \bar{e}^{(i)}/T)}$.

AD with random word data augmentation



Our method

- Normal and anomaly prompt templates "[w_0] a [w_1] photo [w_2] of [w_3] [n] [w_4]" and "[w_5] a [w_6] photo [w_7] of [w_8] [a] [w_9]" are prepared. At the locations of "[n]" and "[a]", words of normal and anomaly are inserted, respectively, in the same way as the prompt-guided AD. At the locations of "[w_i]", randomly generated words are inserted.
- The prompts generated with random words are transformed into text-embedding pairs $\mathcal{E} = \{(e_i^{(n)}, e_i^{(a)})\}_{i=1}^{N_p}$ with the tokenizer and text encoder in CLIP. The embedding pairs are used for training an FNN.
- The t-SNE plot shows the generated embeddings have diversity.



t-SNE plot of generated text embeddings

Experimental Results

Comparison against State of The Art

Setup	Method	MVTec-AD [3] AUROC	MVTec-AD [3] AUPR	MVTec-AD [3] F_1-max	VisA [4] AUROC	VisA [4] AUPR	VisA [4] F_1-max
0-shot (Object unknown)	CLIP	91.5	95.7	92.0	76.5	80.5	78.1
	Ours	91.0	95.4	92.2	78.1	81.3	79.8
	CLIP + ours	**92.2**	**96.0**	**92.8**	**78.2**	**81.5**	**79.9**
0-shot (Object known)	WinCLIP	91.8	**96.5**	92.9	78.1	81.2	79.0
	CLIP	92.6	96.3	93.0	76.3	80.4	78.8
	CLIP + ours	**93.0**	96.4	**93.1**	**79.8**	**82.8**	**79.9**
1-shot	WinCLIP	93.1	96.5	93.7	**83.8**	85.1	83.1
	CLIP + ours	**93.3**	**96.7**	**94.0**	83.4	**85.8**	**83.6**
2-shot	WinCLIP	**94.4**	**97.0**	**94.4**	84.6	85.8	83.0
	CLIP + ours	94.0	96.9	94.1	**85.6**	**87.5**	**84.1**
3-shot	WinCLIP	**95.2**	**97.3**	**94.7**	**87.3**	**88.8**	84.2
	CLIP + ours	94.5	97.1	94.4	86.6	88.4	**84.5**

Comparison without multiple crops on the MVTec-AD dataset in the zero-shot known-object setup & model complexities/speeds.

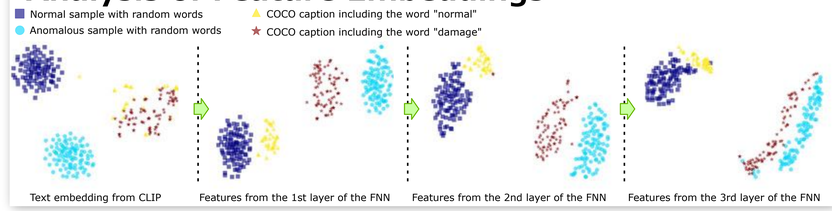
Method	Prompt ens.	AUROC	AUPR	F_1-max	#Params	#MACs	Latency (ms)
CLIP		89.8	95.4	92.1	**77.81M**	**105.4G**	**18.0**
WinCLIP	✓	90.8	96.1	**92.5**	**77.81M**	205.9G	41.9
Ours		89.6	95.5	91.5	78.11M	**105.4G**	18.3
CLIP + ours		**91.0**	**96.2**	**92.5**	78.11M	**105.4G**	18.4

Ablation Study

AUROCs with various word pairs in the zero-shot unknown-object setup. The left values are the results of CLIP, and the right values are those of CLIP + ours.

	"a damaged"	"a broken"	"a defective"	"an anomalous"
"an"	91.5/**92.2**	**87.5**/86.1	79.4/**85.7**	67.6/**73.7**
"a normal"	89.3/**90.5**	87.3/**88.6**	81.8/**84.5**	69.1/**71.9**
"a good"	88.4/**89.6**	86.1/**87.0**	80.6/**86.3**	68.6/**73.0**
"a flawless"	88.5/**90.3**	85.7/**86.0**	77.7/**84.5**	68.8/**75.8**

Analysis of Feature Embeddings



Conclusion

- A novel approach for zero-shot AD is proposed, which can be applied to the case where anomalous samples of unknown objects must be detected.
- Our method achieves competitive performance without any prompt ensemble.
- Extensive experiments show the potential use case.

References

[1] Li *et al.*, CVPR2021
[2] Jeong *et al.*, CVPR2023
[3] Bergmann *et al.*, CVPR2019
[4] Zou *et al.*, ECCV2022