Sea-CLIP: Mining Semantic-Aware Representations for Few-Shot Anomaly Detection with CLIP

Supplementary Material

We include the following content in this supplement material.

- Implementation Details.
- Hyperparameter Study.
- Ablation Study on VisA.
- Additional SD Generated Samples.
- Additional FSAD Qualitative Results.
- Limitations.

6. Implmentation Details

Our proposed Sea-CLIP is based on the implementation of OpenCLIP [39]. Specifically, we use the ViT-B-14 pre-trained on the LAION400M [46] as the backbone. In the preprocessing procedure, we follow the previous work [32] to conduct a channel-wise normalization using pre-computed mean $[0.48145466, 0.4578275, 0.40821073]$ and standard deviation $[0.26862954, 0.26130258, 0.27577711]$. Additionally, the bicubic resize operation is used to unify the input image resolution into 240×240 for the training and inference. Secondly, the pre-trained DINOv2 used to obtain the semantic-aware representation is also based on ViT-B-14, and its pre-trained weights are publicly available at Link1. More specifically, we use features from the "token" facet from the last (12 th) layer from DINOv2. The CLIP and DINOv2 visual features we use in the PM module have sizes of $15 \times 15 \times 896$ and $15 \times 15 \times 384$, respectively. We concatenate these two features and feed them to the proposed Anomaly Matching Decoder (AMD), which has 3 transformer encoder blocks. It is worth mentioning that we modify `sigmoid` into

$$\sigma(x) = \frac{1}{1 + e^{-\alpha x}},$$

in which α is set as 10. This hyperparameter helps the training converge faster while overall FSAD performance remains. For the object-agnostic prompt, we set the length of trainable tokens in both normal and abnormal prompts to 12. Our source code will be made publicly available upon acceptance.

The proposed Sea-CLIP is trained end-to-end, with learning rates for AMD and learnable tokens set at 0.05 and 0.002, respectively. We use 4 learnable tokens for both normal and abnormal samples. In Eq. 11, λ_{cls}, λ_{focal}, and λ_{dice} are set 1, 0.01, and 0.01, respectively. Lastly, to reduce such randomness, we report the mean over 5 random seeds for each measurement.

Hyperparameter	PRO
Number of Learnable Tokens	
2	89.8
4	**90.8**
8	90.6
16	90.1
DINOv2 Feature Layer	
Block 9	90.1
Block 10	89.1
Block 11	90.5
Block 12 (Final)	**90.8**

Table 5. Parameter analysis on MVTec (4-shot PRO).

For the data augmentation, we use Stable Diffusion V1-5 to generate 150 abnormal images, and its pre-trained weights are publicly available at Link2. Specifically, Tab. 8 illustrates 15 textual prompts that describe the anomaly pattern. Then, we use each of these prompts with 10 binary masks randomly generated via Bezier curves as inputs to SD. Consequently, we generate 150 anomaly images for each normal image for training.

7. Hyperparameters Study

Tab. 5 reports analysis on two key hyperparameters. (a) We set the number of learnable tokens as 2,4,8,16, SeaCLIP achieves 89.8, 90.8, 90.6, and 90.1 PRO, respectively. (b) When we use DINOv2 features from the last four blocks, SeaCLIP achieves 90.1, 89.1, 90.5, and 90.8 PRO, respectively. These results confirm our choices are robust — 4 learned tokens and features from 12th, *e.g.*, the last block, DINOV2.

8. Ablation Study on VisA

Tab. 6 shows our ablation study on VisA, indicating our design choices are consistently effective. The trends are even more pronounced on the challenging VisA set, especially for the PRO metric, underscoring the importance of our semantic-aware design for complex objects.

9. Additional SD Generated Samples

Fig. 10 shows additional training samples generated by SD. For example, we generate anomaly `cable` with different textures and spatial locations. We believe it is important to

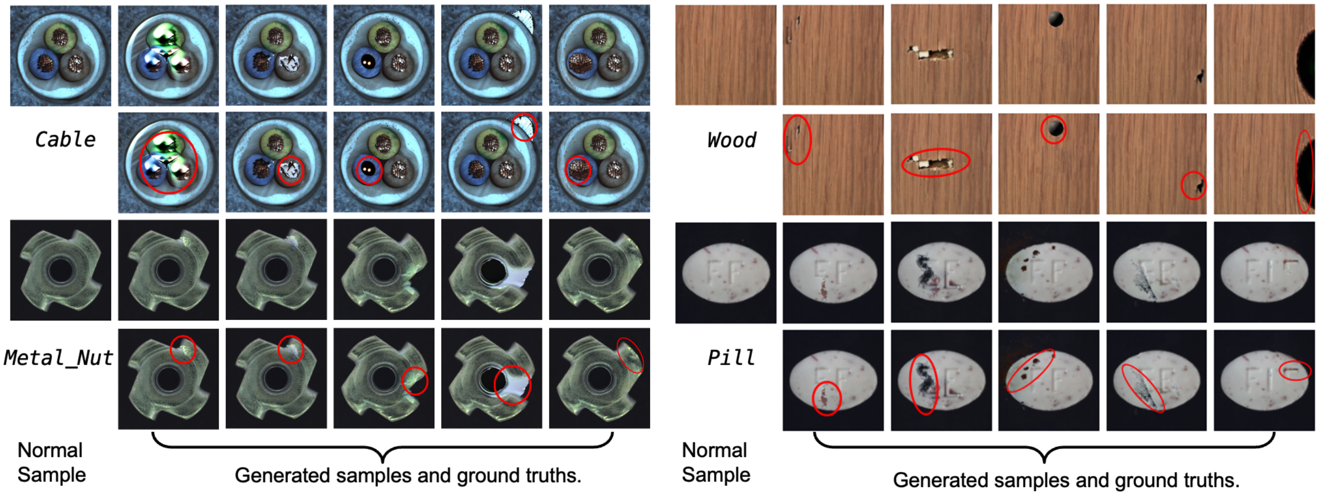
Cable
Wood
Metal_Nut
Pill
Normal Sample
Generated samples and ground truths.
Normal Sample
Generated samples and ground truths.

Figure 10. Additional self-blending generated samples.

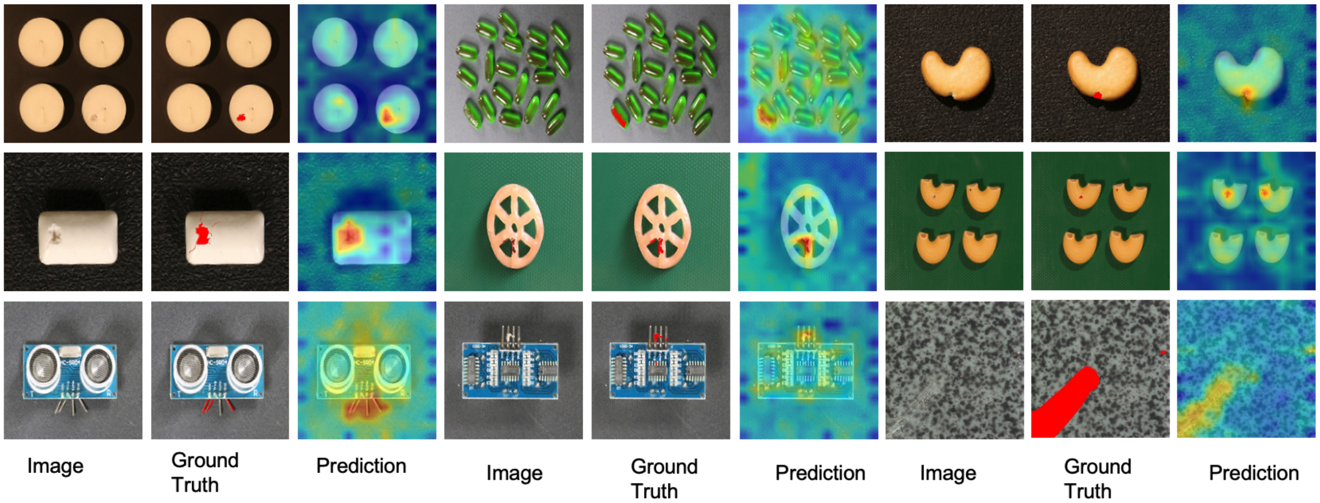

Figure 11. Additional qualitative FSAD results.

PM		L-Token	AMD	VisA	
SM	RA			$\mathrm{AUC_i}$	PRO
	✓			88.9	86.4
✓	✓			89.0	87.0
✓	✓	✓		91.6	87.2
✓	✓		✓	90.4	87.6
✓	✓	✓	✓	**93.1**	**88.1**

Table 6. Ablation study on VisA with a 4-shot AD setup. [Key: PM: Patch Matching Module, SM: S-match, L-token: Learnable tokens, AMD: Anomaly Matching Decoder.].

use such variations in anomaly patterns for training, since our Sea-CLIP contains an AMD that is a much more complicated model than the adapters used in CoOp or Maple.

Method	MVTec		VisA	
	image	pixel	image	pixel
PDT+WinCLIP	91.8	85.1	78.1	79.6
PDT+VV-CLIP	90.5	86.7	77.2	82.9
CoOp	81.5	87.8	72.6	85.5
CoCoOp	60.6	52.1	61.6	72.3
Maple	66.8	64.9	60.5	61.5
DS-adCLIP	**94.9**	**95.4**	**87.0**	**96.5**

Table 7. FSAD Results of different methods. [Key: PDT: predefined textual template.]

	Pre-defined Descriptive Anomaly Pattern Prompts
1	`a blue anomaly rocky region.`
2	`a black round anomaly region.`
3	`a serious broken area.`
4	`a region with scattered black objects.`
5	`a damaged area.`
6	`a weird green textures region.`
7	`a black hole.`
8	`a red rust area.`
9	`a cracked gray surface.`
10	`a fractured gray rocky texture.`
11	`a worn-out faded fabric texture.`
12	`a region looks like torn white paper material.`
13	`a serious white worn out faded texture.`
14	`a large, serious, worn-out white faded texture.`
15	`a region with dark, dirty stains.`

Table 8. We use pre-defined textual prompts describing the anomaly pattern.

10. Additional FSAD Qualitative Results

We visualize the predicted anomaly segmentation map in Fig. 11, in which Sea-CLIP identifies anomalous regions accurately. For example, the first two objects, *i.e.*, `cable` and `capsules`, have anomaly regions with small spatial sizes, which can be challenging for FSAD algorithms; however, Sea-CLIP identifies them accurately In addition, this visualization includes both rigid and flat objects, on both of which our proposed Sea-CLIP generates remarkable anomaly localization performance.

Please note that the previous FSAD methods' results, including SPADE [12], PaDiM [43], PatchCore [43], and WinCLIP [29], are cited in PromptAD [32].

11. Limitations

The first limitation is that Sea-CLIP relies on DINOv2 for the semantic-aware representations, while we do not have rigorous verification on which self-supervised method can produce the best semantic-aware representations. Also, it is interesting to see how it performs when it goes from image to video domain, where semantic information is less important. Moreover, our work is optimized via SD generated samples, which, however, can cause misinformation about real object appearances. Users can detect these attacks using existing deepfake detection methods [23–26, 50, 67] or by training one using the images generated by Sea-CLIP.