

Towards Better Zero-Shot Anomaly Detection under Distribution Shift with CLIP

-supplementary materials-

Jiyao Gao
jiyao.gao@outlook.com

Chengxin He
cxinhe@foxmail.com

Lei Duan
leiduan@scu.edu.cn

Jie Zuo*
zuojie@scu.edu.cn

College of Computer Science
Sichuan University
Chengdu, China

1 Datasets

In this section, we provide a more detailed description of the public benchmarks we used in our experiments.

MVTec [1] is a widely used unsupervised industrial anomaly dataset that contains samples for training and samples in the testing set. There are 15 categories of which are 5 textures (carpet, grid, leather, tile, wood) and 10 objects (bottle, cable, capsule, hazelnut, metal nut, pill, screw, toothbrush, transistor, zipper) in this dataset. **VisA** [2] is another popular unsupervised industrial anomaly detection dataset, which contains 12 subsets corresponding to 12 different objects. Compared with MVTEC, the VisA dataset is concentrated on more refined anomaly detection tasks, such as the defects on printed circuit boards. Since these two benchmarks only contain samples from a single distribution, to simulate distribution shift in real-world applications, we use `imagecorruption` package to corrupt these samples. Like [2], we generate four types of corrupted samples, which are brightness, contrast, blur, and Gaussian noise. All the corruption's severity is set to 3.

AeBAD [3] is an anomaly detection dataset that focuses on modeling various distribution shifts within real-world industrial scenarios. The distribution shifts are attributed to variations in illumination, viewing angles, and backgrounds. Figure 1, 2 and 3 provide some examples of these datasets.

2 Implementation Details

We provide more information about implementing our method in this section. The foundation CLIP model we used is ViT/B-16+ pre-trained by OpenCLIP. We adopt a prompt

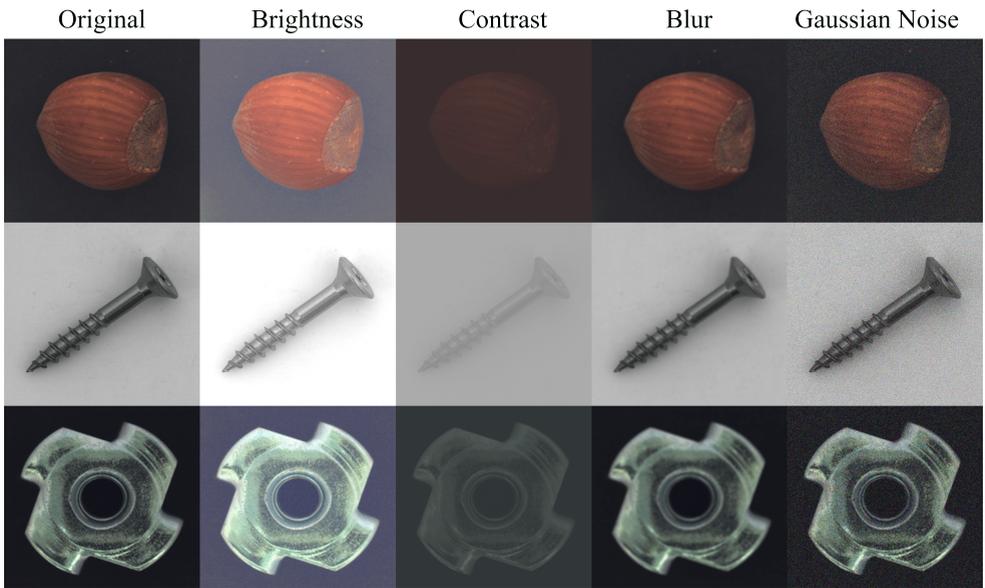


Figure 1: Example images from MVTec.

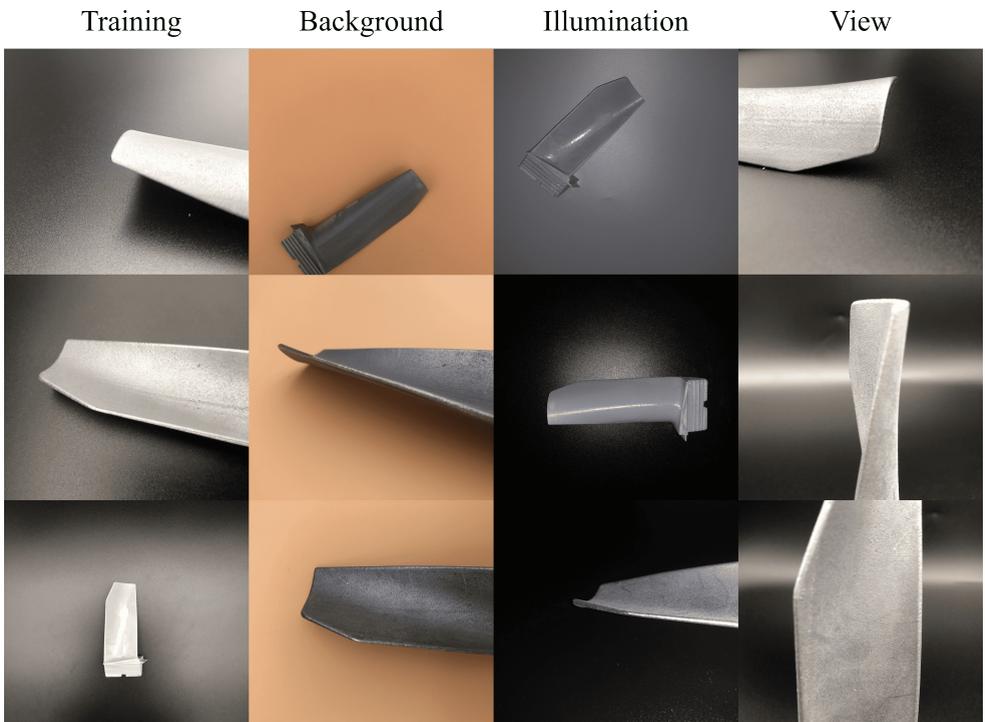


Figure 2: Example images from AeBAD.

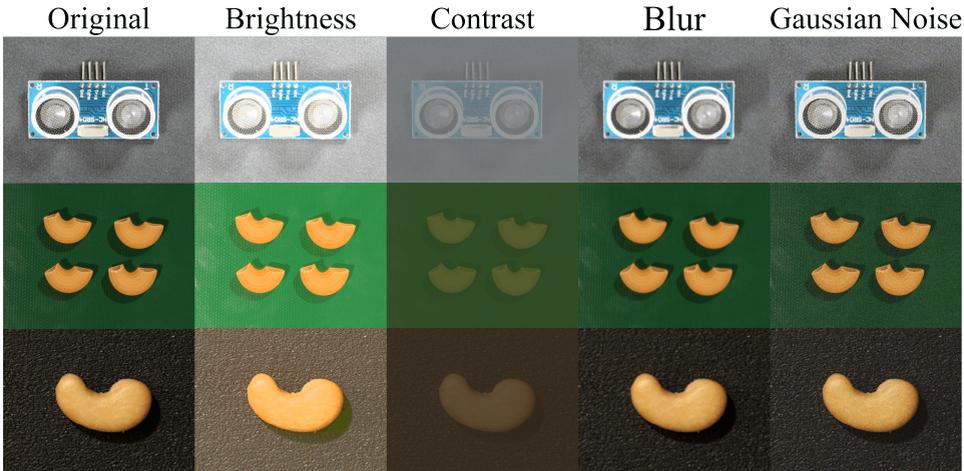


Figure 3: Example images from VisA.

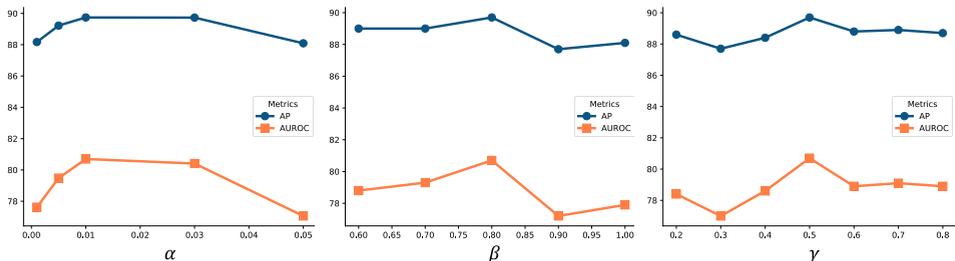


Figure 4: Anomaly detection performance under different balance coefficients on AeBAD.

learning method CoOp [5] to optimize the distribution and diversity words. The length of distribution and diversity words is 2 and 8, respectively. We initialize them by using a Gaussian distribution with 0 mean and 0.02 standard deviation. Then, we use the \mathcal{L}_1 to train each distribution word for 60 epochs. When training the diversity words, we use \mathcal{L}_2 to optimize these words for 100 epochs. An SGD optimizer is used to train the network with an initial learning rate of 0.002 which is adjusted via the cosine decay strategy, and a momentum of 0.9. For the [object] we used in the class prompts, we use `aeroengine blade` for AeBAD and `object` on all the subsets of MVTEC and VisA. As for the prompt template in the \mathcal{L}_{text} , we directly adopted the same prompts used in WinCLIP [9] for a fair comparison.

3 Hyperparameter Analysis

Firstly, we measure some important hyperparameters in our method, including α , β , γ as the balance coefficients in the learning objectives \mathcal{L}_1 and \mathcal{L}_2 . Figure 4 shows the results under the wider values of these hyperparameters. As observed, the best performance is achieved when α equals 0.01, β equals 0.8 and γ is set to 0.5. Our method showed only a maximum

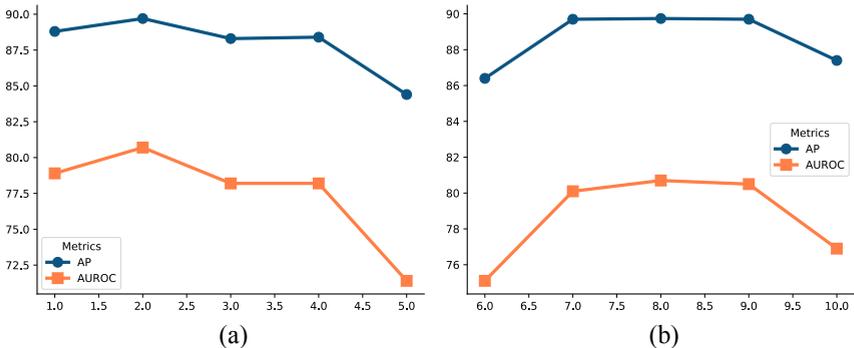


Figure 5: Anomaly detection performance under different lengths of (a) distribution and (b) diversity words on AeBAD.

decrease of 4.5% in performance within this wide range of values, indicating its insensitivity to these hyperparameters. It is worth noting that our model achieves its best only when α is relatively low. We believe it is reasonable, since a larger scale factor α may make the distribution words overfitted to the predefined prompts. This may decrease their diversity and ability to generalize to arbitrary distributions. Also, we study the length of distribution and diversity words in Figure 5 (a) and (b), respectively. We find that as the length of the words increased, there was a decline in the performance of the method, which indicates that a longer length might introduce too much redundant information and compromise our method.

4 Ablation Studies on Knowledge Distillation

This section investigates the performance comparison when using different prompts for \mathcal{L}_{text} . The results are given in Table 1. The first line of Table 1 means we remove the knowledge distillation loss in the first stage, which acts as a baseline method. Since we report the performance of our methods on using prompts designed by WinCLIP (the fourth line in the Table 1), to further investigate the impact of different prompt words on the performance of our method, we replace the prompt words with the template provided by CLIP on ImageNet (denoted by CLIP-AC in Table 1). By comparing the first and the second line of Table 1, we found that the performance of our method is slightly decreased when using ImageNet prompts. Considering that the ImageNet prompt template contains many unreasonable prompts for industrial scenarios like "a weird photo of" or "graffiti of the", we then remove these unreasonable prompts, use a subset of the ImageNet prompts to reproduce the experiment, represented by CLIP-AC* in Table 1 (the third line). With the reduced prompts, our method reaches a satisfactory performance. This finding suggests that the knowledge introduced in \mathcal{L}_{text} should be reasonable and roughly indicate the probable distributions during test time. Introducing prompts that are incapable of industrial scenarios may be harmful to our method.

Methods	AUROC	AP
None	78.5	88.7
CLIP-AC	77.9	88.2
CLIP-AC*	<u>80.5</u>	89.7
WinCLIP	80.7	89.7

Table 1: AUROC and AP on AeBAD with different prompts in the first stage.

Objects	WinCLIP	AnomalyCLIP	Ours
carpet	(99.0,99.7)	(1.0,1.0)	(99.8,99.9)
bottle	(96.3,98.9)	(86.0,95.8)	(97.5, 99.2)
hazelnut	(88.1,93.9)	(85.5,92.7)	(88.7,94.2)
leather	(1.0,1.0)	(99.9,1.0)	(1.0,1.0)
cable	(69.3,79.8)	(65.0,77.6)	(74.9,81.8)
capsule	(68.5,90.6)	(78.8,94.9)	(76.7,94.1)
grid	(96.1,98.7)	(92.4,97.5)	(99.5,99.8)
pill	(68.0,92.4)	(66.2,89.7)	(63.9,91.4)
transistor	(83.0,79.2)	(83.9,80.5)	(81.6, 80.9)
metal nut	(95.1,98.9)	(97.0,99.3)	(93.0,98.2)
screw	(62.2,81.8)	(45.7,75.7)	(50.8,77.3)
toothbrush	(76.7,90.9)	(80.3,92.0)	(85.3,94.4)
zipper	(89.5,96.9)	(89.2,97.0)	(89.2,96.7)
tile	(1.0,1.0)	(1.0,1.0)	(1.0,1.0)
wood	(93.2,97.8)	(99.2,99.8)	(93.2,97.9)
mean	(85.7,93.3)	(84.6,92.8)	(86.3,93.7)

Table 2: Fine-grained AD performance (AUROC and AP) for MVTec Brightness dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
carpet	(96.5,99.0)	(98.0,99.5)	(99.4,99.8)
bottle	(94.6,98.2)	(91.0,97.5)	(92.5,97.7)
hazelnut	(93.6,96.6)	(89.9,94.5)	(91.3,95.3)
leather	(99.8,99.9)	(98.9,99.5)	(99.0,99.7)
cable	(70.4, 82.8)	(59.9,74.3)	(72.1,80.2)
capsule	(65.5,89.5)	(85.2,96.6)	(76.7,94.5)
grid	(97.3,99.1)	(95.4,98.6)	(99.3,99.8)
pill	(87.3,97.3)	(78.8,94.8)	(80.2,96.0)
transistor	(81.5,77.6)	(80.0,79.3)	(76.2,73.4)
metal nut	(87.1,96.8)	(73.7,92.3)	(91.5,97.7)
screw	(65.6,84.2)	(77.6,90.1)	(76.8, 90.9)
toothbrush	(89.4,96.1)	(90.6,96.6)	(86.9,94.6)
zipper	(83.7,95.6)	(84.5,95.3)	(94.5,98.5)
tile	(80.5,93.4)	(99.5,99.8)	(88.9,96.5)
wood	(95.6,98.6)	(89.8,97.2)	(95.0,98.5)
mean	(85.9,93.7)	(86.2,93.7)	(88.0,94.2)

Table 3: Fine-grained AD performance (AUROC and AP) for MVTec Contrast dataset.

5 Fine-grained Anomaly Detection Performance

In this section, we report the performance of three competitive methods (WinCLIP [3], AnomalyCLIP [6] and ours) on each subset of MVTec and VisA, under all four kinds of distribution shifts in Table 2-9.

References

- [1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019.
- [2] Tri Thien Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution

Objects	WinCLIP	AnomalyCLIP	Ours
carpet	(97.6,99.3)	(89.1,96.7)	(98.8,99.6)
bottle	(93.7,98.1)	(93.3,98.0)	(93.9,98.2)
hazelnut	(84.4,92.1)	(84.0,92.4)	(90.4,95.2)
leather	(1.0,1.0)	(99.9,1.0)	(1.0,1.0)
cable	(67.2,81.1)	(80.3,88.2)	(82.3,88.4)
capsule	(62.3,88.0)	(85.4,96.6)	(76.0,94.0)
grid	(98.5,99.6)	(85.4,95.0)	(98.8,99.6)
pill	(77.3,94.2)	(77.0,94.6)	(75.4,93.8)
transistor	(85.7,84.4)	(80.1,78.6)	(80.8,78.3)
metal nut	(93.4,98.5)	(77.7,93.6)	(74.5,93.5)
screw	(65.1,86.2)	(69.5,85.6)	(85.9,95.2)
toothbrush	(89.7,96.2)	(83.6,90.2)	(80.3,92.6)
zipper	(90.6,97.5)	(88.4,96.8)	(92.3,97.9)
tile	(99.7,99.9)	(98.8,99.6)	(99.9,1.0)
wood	(97.5,99.2)	(98.9,99.7)	(97.8,99.4)
mean	(86.8,94.3)	(86.1,93.0)	(88.5,95.1)

Table 4: Fine-grained AD performance (AUROC and AP) for MVTec Defocus Blur dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
carpet	(99.1,97.8)	(93.7,98.2)	(99.9,99.9)
bottle	(96.7,99.0)	(91.0,97.5)	(94.4,98.3)
hazelnut	(92.8,96.3)	(87.9,93.4)	(90.4,95.1)
leather	(99.9,1.0)	(99.5,99.8)	(99.9,1.0)
cable	(78.1,85.6)	(61.5,74.1)	(80.8,86.4)
capsule	(66.0,89.6)	(66.5,90.2)	(79.0,95.2)
grid	(96.4,98.7)	(85.9,94.8)	(96.5,98.9)
pill	(69.4,92.6)	(60.4,89.5)	(64.6,91.0)
transistor	(74.9,72.2)	(68.7,60.1)	(73.0,71.5)
metal nut	(89.0,97.4)	(74.7,92.8)	(82.1,95.8)
screw	(65.5,84.0)	(68.5,84.3)	(71.1,88.6)
toothbrush	(96.1,98.7)	(79.2,92.7)	(91.4,96.9)
zipper	(94.6,98.6)	(82.5,95.2)	(94.7,98.4)
tile	(99.4,99.8)	(99.0,99.6)	(99.9,1.0)
wood	(96.6,98.9)	(96.5,99.0)	(97.2,99.1)
mean	(87.6,94.1)	(81.0,90.7)	(87.6,94.3)

Table 5: Fine-grained AD performance (AUROC and AP) for MVTec Gaussian Noise dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
candle	(92.4,92.7)	(75.2,76.7)	(93.6,93.1)
capsules	(75.4,85.1)	(78.4,87.0)	(76.1,85.1)
cashew	(80.2,91.2)	(72.8,87.3)	(78.5,90.1)
chewinggum	(72.1,87.5)	(86.5,93.7)	(78.4,89.8)
fryum	(72.2,86.0)	(92.2,96.5)	(84.4,92.4)
macaroni1	(72.4,68.9)	(74.6,72.3)	(73.6,70.9)
macaroni2	(62.5,62.1)	(60.0,60.9)	(66.1,68.1)
pcb1	(60.4,64.7)	(65.6,69.1)	(59.4,67.5)
pcb2	(39.8,43.1)	(70.3,70.9)	(39.5,43.3)
pcb3	(57.2,57.4)	(69.0,72.3)	(60.3,63.6)
pcb4	(69.9,68.2)	(89.7,89.3)	(68.7,66.4)
pipe fryum	(59.8,78.3)	(89.0,94.7)	(90.9,95.9)
mean	(67.9,73.8)	(76.9,80.9)	(72.5,77.2)

Table 6: Fine-grained AD performance (AUROC and AP) for VisA Brightness dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
candle	(87.2,89.1)	(82.9,81.8)	(83.4,85.6)
capsules	(63.1,78.7)	(76.6,85.3)	(65.8,79.9)
cashew	(89.2,95.3)	(73.7,88.4)	(87.2,94.0)
chewinggum	(87.7,94.7)	(87.3,94.3)	(89.2,95.2)
fryum	(69.6,85.4)	(85.9,93.6)	(89.1,95.0)
macaroni1	(67.7,69.5)	(71.4,73.4)	(74.2,76.1)
macaroni2	(59.2,59.2)	(59.4,56.4)	(62.5,63.6)
pcb1	(62.4,67.5)	(70.1,74.0)	(58.4,66.2)
pcb2	(60.0,57.4)	(58.4,58.3)	(58.3,58.4)
pcb3	(66.4,70.9)	(52.5,53.1)	(67.0,70.2)
pcb4	(71.2,75.1)	(91.3,92.5)	(80.0,80.8)
pipe fryum	(82.2,91.5)	(90.0,95.0)	(95.3,97.8)
mean	(72.2,77.9)	(75.0,78.8)	(75.9,80.2)

Table 7: Fine-grained AD performance (AUROC and AP) for VisA Contrast dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
candle	(96.0,96.5)	(74.8,74.5)	(96.2,96.5)
capsules	(81.1,88.8)	(81.6,87.9)	(83.5,90.1)
cashew	(92.2,96.6)	(90.7,95.9)	(83.6,92.4)
chewinggum	(92.4,96.9)	(95.8,98.1)	(93.3,97.3)
fryum	(71.6,85.6)	(88.7,94.7)	(83.5,92.5)
macaroni1	(75.2,70.6)	(79.3,79.1)	(73.2,69.8)
macaroni2	(62.9,64.4)	(64.2,61.2)	(63.7,66.5)
pcb1	(71.1,71.3)	(83.3,85.9)	(49.3,57.8)
pcb2	(38.4,41.6)	(47.7,49.8)	(47.1,48.5)
pcb3	(62.2,63.4)	(65.6,71.6)	(70.7,74.4)
pcb4	(80.4,79.0)	(93.3,93.6)	(82.4,84.3)
pipe fryum	(56.5,76.0)	(91.9,96.1)	(86.9,93.9)
mean	(73.3,77.5)	(79.7,82.4)	(76.1,80.3)

Table 8: Fine-grained AD performance (AUROC and AP) for VisA Defocus Blur dataset.

Objects	WinCLIP	AnomalyCLIP	Ours
candle	(90.2,91.0)	(84.0,86.5)	(91.3,92.1)
capsules	(82.7,90.0)	(80.0,87.5)	(78.4,87.1)
cashew	(90.0,95.6)	(72.8,87.5)	(89.7,95.4)
chewinggum	(92.4,96.6)	(84.8,93.5)	(91.9,96.5)
fryum	(75.0,86.5)	(90.6,95.9)	(88.8,94.8)
macaroni1	(82.0,82.4)	(81.6,81.2)	(82.4,82.8)
macaroni2	(62.2,64.7)	(55.3,55.0)	(60.8,63.2)
pcb1	(66.1,67.3)	(76.4,77.3)	(46.2,53.8)
pcb2	(46.9,46.0)	(62.8,62.9)	(60.4,56.5)
pcb3	(59.2,59.2)	(55.5,60.9)	(72.0,75.2)
pcb4	(72.1,69.0)	(93.6,94.2)	(84.6,84.5)
pipe fryum	(73.1,86.4)	(95.1,97.4)	(94.8,97.6)
mean	(74.3,77.9)	(77.7,81.6)	(78.4,81.6)

Table 9: Fine-grained AD performance (AUROC and AP) for VisA Gaussian Noise dataset.

- shift. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6488–6500, 2023.
- [3] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19606–19616. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01878.
- [4] Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 151:103990, 2023.
- [5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.
- [6] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection, 2024.
- [7] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pages 392–408. Springer, 2022.