Aligning and Prompting Anything for Zero-Shot Generalized Anomaly Detection

Jitao Ma[1], **Weiying Xie**[1*], **Hangyu Ye**[1], **Daixun Li**[1], **Leyuan Fang**[2]

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
jtm@stu.xidian.edu.cn, wyxie@xidian.edu.cn, wagyr@stu.xidian.edu.cn, ldx@stu.xidian.edu.cn, leyuan_fang@hnu.edu.cn

Abstract

Zero-shot generalized anomaly detection (ZGAD) plays a critical role in industrial automation and health screening. Recent studies have shown that ZGAD methods built on visual-language models (VLMs) like CLIP have excellent cross-domain detection performance. Different from other computer vision tasks, ZGAD needs to jointly optimize both image-level anomaly classification and pixel-level anomaly segmentation tasks for determining whether an image contains anomalies and detecting anomalous parts of an image, respectively, this leads to different granularity of the tasks. However, existing methods ignore this problem, processing these two tasks with one set of broad text prompts used to describe the whole image. This limits CLIP to align textual features with pixel-level visual features and impairs anomaly segmentation performance. Therefore, for precise visual-text alignment, in this paper we propose a novel fine-grained text prompts generation strategy. We then apply the broad text prompts and the generated fine-grained text prompts for visual-textual alignment in classification and segmentation tasks, respectively, accurately capturing normal and anomalous instances in images. We also introduce the Text Prompt Shunt (TPS) model, which performs joint learning by reconstruction the complementary and dependency relationships between the two tasks to enhance anomaly detection performance. This enables our method to focus on fine-grained segmentation of anomalous targets while ensuring accurate anomaly classification, and achieve pixel-level comprehensible CLIP for the first time in the ZGAD task. Extensive experiments on 13 real-world anomaly detection datasets demonstrate that TPS achieves superior ZGAD performance across highly diverse datasets from industrial and medical domains.

Code — https://github.com/majitao-xd/TPS

Introduction

Recently, zero-shot generalized anomaly detection (ZGAD) (Jeong et al. 2023; Gu et al. 2024; Zhou et al. 2023) is a widely applied task in computer vision, with applications ranging from defect detection in industrial product images (Reiss and Hoshen 2023; He et al. 2024; Hu et al. 2024) to tissue lesion detection in medical images (Huang et al. 2024; Liu et al. 2024; Wu and Xu 2024). ZGAD aims to train a single model capable of performing cross-domain anomaly detection without relying on any examples or training data from the target domain. Anomaly detection (AD) consists of two jointly optimized tasks, image-level anomaly classification and pixel-level anomaly segmentation. The purpose of anomaly classification is to determine whether an image contains anomalies or not, and the purpose of anomaly segmentation is to determine which parts of an image are anomalies. Therefore, there is a granularity difference between the two tasks. Existing ZGAD methods are excellent for determining whether an image contains anomalies, resulting in superior anomaly classification. However, these methods fall short in identifying which part of the image is an anomaly, limiting performance of anomaly segmentation.

There have been many studies (Jeong et al. 2023; Zhou et al. 2023; Zhu and Pang 2024) focusing on the powerful visual perception capabilities of CLIP (Chen et al. 2020; He et al. 2020) and applying them to ZGAD tasks with remarkable results. However, existing methods like WinCLIP

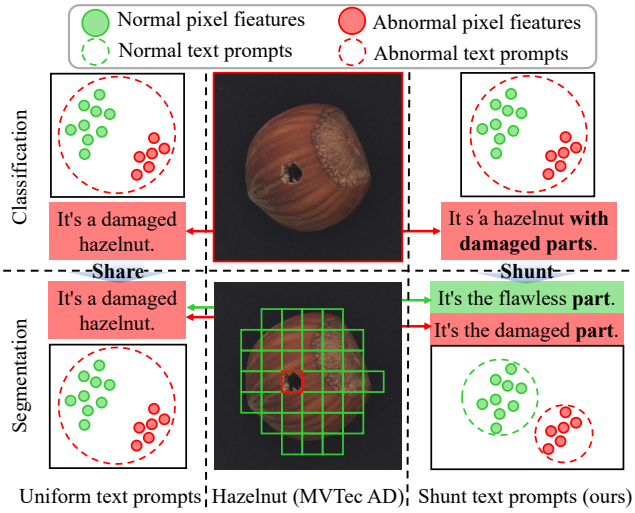

Figure 1: Classification only needs to identify whether there is an anomalous part of the image or not, and segmentation needs to identify which part of the image is anomalous.

*Correspond author.

(Jeong et al. 2023) simply use CLIP as a textual and visual feature extractor, failing to consider the difference between anomaly classification and segmentation tasks, and using only broad text prompts which describe the whole image in all tasks. This results in CLIP output text features that are not well aligned to fine-grained pixel-level visual features. Therefore, they are relatively ineffective in anomaly segmentation. Although on the anomaly classification task, giving a description of whether the image is anomalous or not (*e.g.*, it is a damaged hazelnut) is sufficient. On the anomaly segmentation task, it is better to directly describe which part is anomalous (*e.g.*, this is the damaged part of a hazelnut), as shown in Figure 1.

In order to generate a fine-grained text prompt instead of using broad text prompts to describe the pixel-level features of an image, we propose a novel strategy for generating fine-grained text prompts. Specifically, we design two sets of text prompts for the anomaly classification and segmentation tasks respectively. The text prompts used for anomaly classification not only describe the whole image as anomalous or not, but also add the suffix [*with damaged parts*] to it. For simplicity in calling it, we named it the global text prompts. The text prompts used for anomaly segmentation describe only whether the pixel is anomalous or not, and do not contain information about the class of the object. Considering that the object it describes is a shunt part from the global text prompts, we named it shunt text prompts. This approach enables the model to accurately address both anomaly classification and segmentation tasks. As illustrated in Figure 1, the shunt text prompts exhibit a closer alignment with pixel-level object features, similar to how humans describe abnormal regions within an image (*e.g.*, the recessed area of a metal plate or the diseased section of a liver).

Furthermore, we propose a **T**ext **P**rompt **S**hunt (TPS) model based on CLIP, which jointly optimizes anomaly classification and segmentation by building pathways between textual features of different tasks. On this basis, TPS can learn the relationship between the two sets of text prompts, *i.e.*, the object described by the shunt text prompt is part of the whole image described by the global text prompt. The design helps the anomaly classification and segmentation tasks to guide and facilitate each other. In summary, this paper makes the following main contributions.

- We reveal that the granularity difference between anomaly classification and segmentation tasks is a potential factor affecting the alignment of CLIP text features with visual features. We propose a novel strategy for generating fine-grained text prompts that enables pixel-level comprehensible CLIP for the first time in the ZGAD task.
- We then design the TPS model to jointly optimize both the anomaly classification and anomaly segmentation tasks by learning the potential relationships between text features through pathway modules.
- Comprehensive experiments across 13 datasets in industrial and medical domains demonstrate that TPS achieves superior ZGAD performance in detecting and segmenting anomalies across cross-domain datasets with highly diverse anomaly characteristics.

Relate Work

Few-shot Generalized Anomaly Detection

Several studies have employed VLMs (Bao et al. 2022; Jia et al. 2021; Li et al. 2022; Radford et al. 2021) to address AD challenges, aiming to meet generalization requirements while reducing reliance on training samples (Yang et al. 2024; Wang et al. 2024; Shen et al. 2024). These approaches can be classified as few-shot AD and zero-shot AD based on the required number of samples. WinCLIP (Jeong et al. 2023) utilizes CLIP as a feature extractor, developing a series of text prompts embedded with prior knowledge and applying a multi-scale sliding window technique for initial cross-domain few-shot and zero-shot anomaly detection. However, these prior knowledge-based text prompts often lack the necessary generalization for effective application in other domains. PromptAD (Li et al. 2024b) advances few-shot AD by connecting normal prompts to anomaly suffixes, converting normal prompts into anomaly prompts for learning, and introducing explicit anomaly boundaries to control the distance between normal and anomaly features. InCTRL (Zhu and Pang 2024) employs Multi-layer Patch-level Residual Learning to build contextual residuals, integrating prior knowledge through text prompts for cross-domain few-shot anomaly classification. Despite these advancements, the applicability of few-shot AD is generally less extensive than that of zero-shot AD, particularly in specific contexts.

Zero-shot Generalized Anomaly Detection

MuSc (Li et al. 2024a) explored the implicit normal and abnormal cues within unlabeled images by scoring test images against each other to assign anomaly scores. However, this approach requires test datasets with both normal and abnormal data and receives limitations in medical lesion detection, such as detecting diseased organs in a single patient. AnomalyCLIP (Zhou et al. 2023) introduces object-agnostic learnable text prompts that allow the model to capture general normality and abnormality in the image regardless of the foreground object, allowing it to focus on abnormal image regions rather than object semantics. Despite these advancements, these methods overlook the granularity differences between anomaly classification and segmentation tasks, aligning the global semantics of the image with the anomaly instances using the same text prompts. This misalignment limits the performance of model in handling fine-grained anomaly instances, hindering accurate segmentation of anomalous regions. Simply separating the two tasks and processing them independently can result in information loss. Therefore, we propose extracting finer-grained shunt text prompts from global text prompts, allowing the tasks to be processed separately while preserving their interrelation.

Method

Problem Formulation

The aim of ZGAD is to train a single model capable of performing anomaly detection across target test data from diverse domains without any prior training on the target data. Consequently, the training set is assumed to have a different

distribution from the test sets. We define an auxiliary training dataset $\mathcal{D}_{train} = \{\mathcal{X}_{train}, \mathcal{Y}_{train}\}$ that includes both normal and abnormal classification and segmentation labels. Here, $\mathcal{X}_{train} = \{x_i\}_i^K$ consists of K normal and abnormal images, and $\mathcal{Y}_{train} = \{y_i\}_i^K$, where a value of 0 in y_i denotes normal and a value of 1 denotes abnormal. The test sets $\mathcal{D}_{test} = \{\mathcal{X}_{test}^1, \mathcal{X}_{test}^2, \ldots, \mathcal{X}_{test}^L, \mathcal{Y}_{test}^1, \mathcal{Y}_{test}^2, , \ldots, \mathcal{Y}_{test}^L\}$ comprise images from various application domains and anomaly types. Traditional AD paradigm typically utilize a sample $\mathcal{P} = \{p_{test}^1, p_{test}^2, ..., p_{test}^L\}$ from $\mathcal{D}_{test}$ for training or for reference during testing. While few-shot AD approaches allow the use of a small fraction of the total sample as $\mathcal{P}$, they still fall short of the versatility offered by ZGAD. ZGAD can train an anomaly detection function $f(\cdot)$ using $\mathcal{D}_{train}$ without leveraging $\mathcal{P}$ in any form, assigning higher anomaly scores to the anomalous samples in $\mathcal{D}_{test}$ compared to the normal samples. Formally, this can be expressed as an optimization problem:

$$\begin{aligned} f(\mathcal{X}_{test}^1) &= \mathcal{Y}_{test}^1, \\ s.t. \quad \mathcal{D}_{train} &\rightarrow f(\cdot). \end{aligned} \tag{1}$$

Overview of Our Approach

The excellent generalization of VLMs offer a promising solution for ZGAD. However, anomaly detection typically requires simultaneous classification and segmentation tasks, which have different granularity. The classification task focuses on learning global semantic features of images, while the segmentation task is more concerned with pixel-level features of instance objects, which have different distributions despite their relevance. Consequently, relying on a single text prompt in VLM-based AD limits the effective alignment between image-level visual features and pixel-level patch features.

Our approach, TPS, aims to model normal and abnormal text prompts at a finer granularity, leveraging the generalization capabilities of CLIP to align image and text features across different application domains. CLIP is a VLM consisting of a text encoder $f_t(\cdot)$ and a visual encoder $f_v(\cdot)$, pre-trained in a comparative learning mode using text-image data. An overview of TPS is shown in Figure 2. The proposed shunt text prompts are employed to achieve more precise alignment with pixel-level features. Firstly, the object name is combined with uniform normal and abnormal prefixes to obtain the global text prompt $T_g = \{T_g^n, T_g^a\}$ for anomaly classification, where T_g^n denotes the global text prompt for normal samples and T_g^a denotes the global text prompt for abnormal samples. Then, a fine-grained description is shunted from T_g to obtain the shunt text prompts $T_s = \{T_s^n, T_s^a\}$ for anomaly segmentation, where T_s^n denotes the shunt text prompt for normal samples and T_s^a denotes the shunt text prompt for abnormal samples. By leveraging both visual and textual features, comparative learning is conducted through contrastive loss, alongside the proposed pathway module. The pathway module facilitates the learning of potential holistic-compositional relationships by guiding the shunt text features through global text features and visual features. Finally, the learned relationships are applied to both anomaly classification and segmentation tasks.

Fine-grained Text Prompts Generation Strategy

Commonly used text prompt templates in CLIP, such as *a photo of a hazelnut*, primarily emphasize the semantics of the foreground object. To focus the attention of the model on anomalous instances within these foreground objects, methods like WinCLIP and InCTRL design templates specifying particular types of anomalies, such as *a photo of a hazelnut with cracks*. In contrast, AnomalyCLIP employs learnable templates, extending the range of anomalies that the model can recognize. However, these approaches apply the same text prompt templates for both anomaly classification and segmentation. While this method provides a global perspective suitable for classification, its generality constrains the effectiveness in segmentation tasks, as it lacks the necessary detail to convey precise anomaly information, such as exact location and extent.

In order to accurately prompt the precise location and extent of anomalies, we derive shunt text prompts for fine-grained descriptions of local instances from the global text prompts that describe the entire image. For this purpose, we designed an additional set of more precise prompts, focusing specifically on identifying whether specific parts of an object are damaged. First, we define the global text prompt as:

$$\begin{aligned} T_g^n &= [prefix][object][with\ all\ parts\ flawless], \\ T_g^a &= [prefix][object][with\ damaged\ parts], \end{aligned} \tag{2}$$

where the $[prefix]$ of T_g is optional, and can range from a simple description like *a photo of*, to a more detailed like *a close-up photo of*, or even a learnable embedding. The $[object]$ denotes the target class or an object-agnostic description. The phrase $[with\ all\ parts\ flawless]$ is used to indicate that the text prompt describes the entire context, distinguishing it from the shunt text prompts, and emphasizing global semantic of the foreground object. In particular, the phrase $[with\ damaged\ parts]$ description within T_g enables a more precise delineation of boundaries between normal and abnormal pixels, while providing a semantic basis for subsequent shunting.

Subsequently, we generate fine-grained shunt text prompts that specifically describe both the normal and abnormal parts of the instance:

$$\begin{aligned} T_s^n &= [prefix][flawless][part], \\ T_s^a &= [prefix][damaged][part], \end{aligned} \tag{3}$$

where the $[prefix]$ is of the same type as in T_g, $[flawless]$ and $[damaged]$ are fixed prompts. The $[part]$ component is included to specify that T_s describes a particular part of an object or an instance derived from an object, thereby enabling a more precise alignment of normal and abnormal textual semantics with pixel-level features. This approach also enhances the dependency of T_s on T_g.

Finally, we employ the text encoder to perform equal feature extraction for both sets of text prompts. This design allows for more detailed textual embedding and the enables precise identification of pixel-level anomalies. By accommodating tasks with varying levels of granularity, the text prompts become versatile. Consequently, the generated

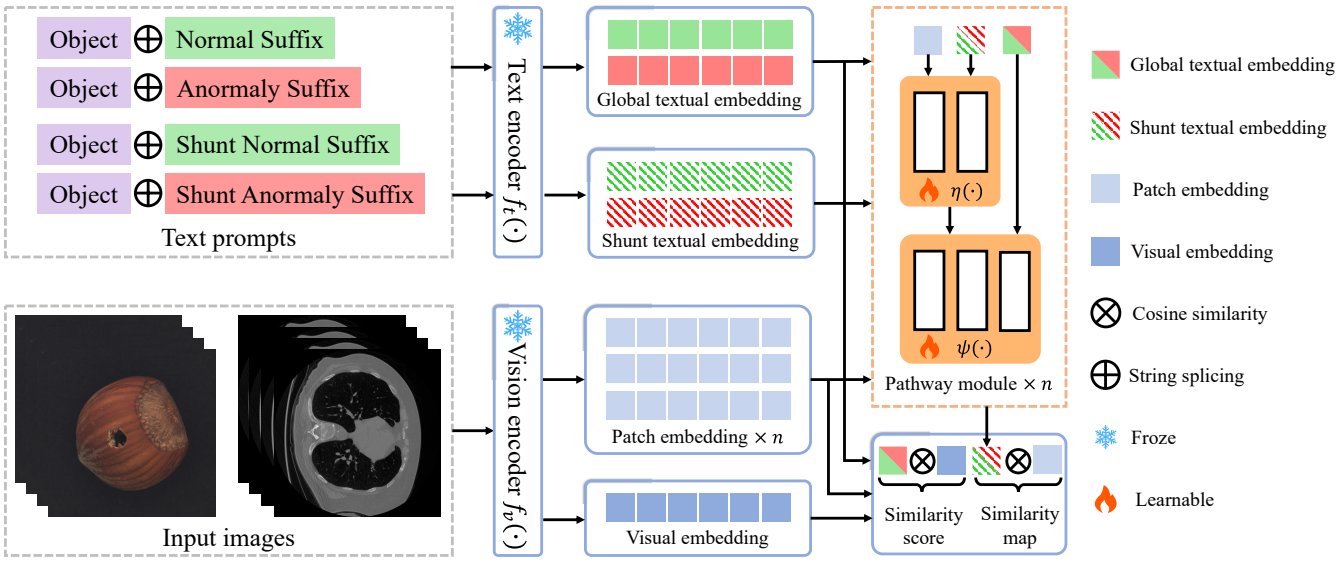

Figure 2: Overview of TPS. To better align patch token embedding with textual embedding, TPS introduces finer-grained shunt text prompts. We then introduce the pathway module to reconstruct the dependencies of the two tasks performed separately. Finally joint tuning is performed to implement ZGAD.

shunted textual embeddings can effectively identify anomalous instances at the pixel level. Furthermore, given that an object may contain multiple instances, using a single $[part]$ surrogate for all instances limits the flexibility of the text prompts. Thus, we adopt the object-agnostic learnable text prompts from (Zhou et al. 2023) as a foundation for replacing the first two components in T_s and T_g.

Pathway Module

To establish the dependency between normal and abnormal instances in the segmentation task and their corresponding objects in the classification task for joint optimization, we introduce a multi-linear shunt pathway learning component within TPS. Typically, textual and visual embeddings extracted by CLIP are used directly to compute anomaly scores via cosine similarity, relying entirely on text prompts and pre-training parameters. This limitation hinders its ability to accurately control the alignment of textual features with visual features on complex objects like PCBs or human tissues. To address this limitation, we incorporate patch embeddings as additional inputs in the pathway module, transforming them into shunt text features customized for each image. Additionally, considering that the CLIP visual encoder consists of a series of block layers that progressively capture visual patterns at different levels of abstraction, the pathway module leverages multi-level patch embeddings to enhance this alignment.

Specifically, assuming the visual encoder consists of N blocks, For a given training image $x \in \mathbb{R}^{H\times W\times 3}$, we extract a series of patch token embeddings through the visual coder $f_v(\cdot)$ according to:

$$\{z_x^i\}_{i\in K} = f_v(x), \tag{4}$$

where $z_x^i \in \mathbb{R}^{L\times D}$ with H, W, and D be the height, width, and dimension of the z_x^i respectively, $L = H \times W$, K is a sequence of numbers less than or equal to N. Subsequently, the textual embeddings are computed as:

$$\{t_g^n, t_g^a, t_s^n, t_s^a\} = f_t(\{T_g^n, T_g^a, T_s^n, T_s^a\}), \tag{5}$$

where t_g^n, t_g^a, t_s^n and t_s^a denotes normal global textual embedding, abnormal global textual embedding, normal shunt textual embedding and abnormal shunt textual embedding respectively, which all have the same dimension as z_x^i. Then, the textual embeddings are fed separately with each patch token embedding into independent pathway modules.

Each pathway module $f_p^i(\cdot)$ consists of a text customization component $\eta^i(\cdot)$ and a dependency building component $\psi^i(\cdot)$. In the i-th pathway module $f_p^i(\cdot)$, t_s and z_x^i are first employed to learn customized shunt text features through $\eta^i(\cdot)$. In order for the patch token embeddings to effectively guide the learning of the shunt text features, $\eta^i(\cdot)$ processes these two embeddings on the basis of Cross Attention according to:

$$\dot{t_s^i} = \eta^i(t_s, z_x^i) = CA_\eta^i(t_s, z_x^i, t_s) + t_s, \tag{6}$$

where $CA_\eta^i(Q, K, V)$ denotes Cross Attention, and the order of input is query, key, value. In $\eta^i(\cdot)$, we employ the patch token embeddings as the key, enabling the model to select the most pertinent visual features based on the text query. These features are not directly utilized for output, instead, they are employed to refine and enhance the text features. This design emphasizes the guiding role of the text during feature fusion, allowing the output features to incorporate visual content while preserving the original textual semantics. This approach ensures semantic consistency of the text and enhances the capture of fine-grained visual information, thus obtaining shunt text features customized for each input image.

Subsequently, $\dot{t}_s^i$ and t_g establish a dependency through $\psi^i(\cdot)$. Similar to $\eta^i(\cdot)$, $\psi^i(\cdot)$ captures the relationship between the shunt textual embeddings and the global textual embeddings through a cross attention according to:

$$\ddot{t}_s^i = CA_\psi^i(\dot{t}_s^i, t_g, \dot{t}_s^i) + \dot{t}_s^i, \tag{7}$$

where $CA_\psi^i(Q, K, V)$ denotes cross attention, and the order of input is query, key, value. Finally, a multi-layer perceptron (MLP) is utilized as a feedforward network to further extract shunt text features. Thus, $\psi^i(\cdot)$ is of the form:

$$\dddot{t}_i^s = mlp(\ddot{t}_s^i) + \ddot{t}_s^i, \tag{8}$$

where $mlp(\cdot)$ denotes the MLP module. In $\psi^i(\cdot)$, we employ the global textual embeddings as the key, enabling the model to selectively integrate relevant information from the global textual embedding at a fine-grained level, thereby enhancing the representation of local textual features. Additionally, shunt text features can incorporate global context information while preserving their fine-grained semantics, thereby effectively capturing the relationship between local and global features. This approach strengthens the dependency of shunt textual embeddings on global textual embeddings, ensuring that shunt textual features remain contextually grounded within the global context for a more comprehensive understanding and processing of anomalies.

Training and Detection

In the training phase, the CLIP weights are frozen and the pathway module is trainable. Specifically, the output from CLIP includes global textual embeddings t_g, shunt textual embeddings t_s, visual embeddings v_x, and patch token embeddings $\{z_x^i\}_{i\in K}$. The pathway module leverages t_g, t_s and $\{z_x^i\}_{i\in K}$ to further refine and extract the shunt textual features $\dddot{t}_{s\,i\in K}^i$. The classification probability for the classification task is then calculated as:

$$C_x = softmax(v_x \odot t_g^n, v_x \odot t_g^a), \tag{9}$$

where $\odot$ denotes element-wise multiplication. The similarity map for the split task is calculated as:

$$S_x = softmax(\sum_{i\in n} z_x^i \odot \dddot{t}_s^{n\,i}, \sum_{i\in n} z_x^i \odot \dddot{t}_s^{a\,i}). \tag{10}$$

To efficiently learn shunt textual features, we jointly optimize two tasks with different granularity, ensuring they are learned from both global and local perspectives. The global context optimization aligns global textual embeddings with visual embeddings of different objects. This helps to efficiently capture the normal and abnormal semantics from the perspective of global features and guide the shunt textual embeddings. The local context optimization focuses shunt textual prompts on fine-grained regions within the intermediate layers of the visual encoder, offering precise descriptions of local anomalies. Consequently, our loss function is expressed as:

$$\mathcal{L} = CE(C_x, y_c) + Focal(Up(S_x), y_s) + Dice(Up(S_x), y_s), \tag{11}$$

where $CE(\cdot)$ denotes cross entropy loss, $Focal(\cdot)$ denotes focal loss (Lin et al. 2017), $Dice(\cdot)$ denotes dice loss (Li et al. 2019), y_c denotes the classification label of the auxiliary training dataset, y_s denotes the segmentation label of the auxiliary training dataset, the $Up(\cdot)$ operation restores the anomaly map to the size of the input image.

In the detection phase, inference is performed with the TPS model weights obtained by minimizing the loss of Eq. 11 through the auxiliary training dataset. Given a test image x, we apply the classification probability C_x as the image-level anomaly score, which tends to be 1 when the anomaly global textual embedding t_g^a is aligned with the visual embedding v_x. For the pixel-level segmentation task, we apply the merged similarity map S_x as the pixel-level anomaly map, which can be formulated as $1 - S_x^n + S_x^a$.

Experiments

Experiment Setup

Implementation details In this paper, we use the publicly available CLIP model (ViT-L/14-336) as our backbone, the code of CLIP for LAION-400M (Schuhmann et al. 2021) and LAION-5B (Schuhmann et al. 2022) scale pre-training is open-scoured by OpenCLIP (Ilharco et al. 2021). It contains 24 layers, which are divided into 4 stages. Each stage is composed of 6 layers. We extract the output of the patch token for each stage as z_x^i, thus $N = 4$, $K = [6, 12, 18, 24]$. We extract the last layer of linear projected class tokens for classification optimization. All test images are scaled to 518×518 and fed into backbone. We fine-tune TPS using the test set of MVTec AD (Bergmann et al. 2019) and evaluate the ZGAD performance on other datasets. Only model tested on MVTec AD is trained with the test set of VsiA (Zou et al. 2022) as an auxiliary dataset. All experiments are performed in PyTorch-1.11.0 with a single NVIDIA RTX 3090 24GB GPU.

Datasets To study anomaly classification and segmentation performance, we conduct experiments on 13 publicly available datasets, covering various industrial inspection scenarios and medical imaging domains (including photography, endoscopy, and radiology) to evaluate the performance of TPS. In industrial inspection, we consider MVTec AD (Bergmann et al. 2019), VisA (Zou et al. 2022), MPDD (Jezek et al. 2021), BTAD (Mishra et al. 2021), SD-saliency-900 (Song, Song, and Yan 2020) and RSDDS-113 (Niu et al. 2020). In medical imaging, we consider brain tumor detection datasets HeadCT (Salehi et al. 2021), BrainMRI (Salehi et al. 2021), Br35H (Ahmed 2020), skin cancer detection dataset ISIC (Codella et al. 2018), colon polyp detection dataset Kvasir (Jha et al. 2020), thyroid nodule detection dataset TN3k (Gong et al. 2021) and lung cancer segmentation dataset MSD (Antonelli et al. 2022). For datasets that do not provide public test set labels, we validate model performance using the training set.

Evaluation metrics and baselines We use Area Under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP) as performance evaluation metrics for anomaly classification. In addition, using AUROC and

Datasets	Category	$\|C\|$	CLIP	WinCLIP	AnomalyCLIP	MuSc	TPS
MVTec AD	Obj & texture	15	74.1, 87.6	91.8, 96.5	91.5, 96.2	**97.8**, **99.1**	96.4, 98.4
VisA	Obj	12	66.4, 71.5	78.1, 81.2	82.1, 85.4	**92.8**, **93.5**	83.3, 85.6
MPDD	Obj	6	54.3, 65.4	63.6, 69.9	**77.0**, **82.0**	61.7, 75.4	73.3, 74.7
BTAD	Obj	3	34.5, 52.5	68.2, 70.9	88.3, 87.3	**96.5**, **93.9**	88.1, 90.0
HeadCT	Brain	1	56.5, 58.4	81.8, 80.2	**93.0**, 91.1	68.6, 58.5	92.3, **91.9**
BrainMRI	Brain	1	73.9, 81.7	86.6, 91.5	90.2, 92.4	88.4, 86.0	**92.4**, **93.8**
Br35H	Brain	1	78.4, 78.8	80.5, 82.2	84.9, 73.8	91.0, 81.7	**96.2**, **96.1**

Table 1: ZSAD classification performance comparison (AUROC, AP). The best performance is marked in **red** and the second best in blue.

Datasets	Category	$\|C\|$	CLIP	WinCLIP	AnomalyCLIP	MuSc	TPS
MVTec AD	Obj & texture	15	38.4, 11.3	85.1, 64.6	91.1, 81.4	**97.3**, **93.8**	94.4, 90.9
VisA	Obj	12	46.6, 14.8	79.6, 56.8	95.5, 86.7	**98.8**, **92.7**	95.5, 90.9
MPDD	Obj	6	62.1, 33.0	76.4, 48.9	96.5, 88.7	**98.2**, 81.0	97.2, **92.1**
BTAD	Obj	3	30.6, 4.4	72.7, 27.3	94.2, 74.8	**98.1**, **81.6**	94.2, 80.9
SD-saliency-900	Texture	1	34.3, 2.4	78.5, 33.1	84.9, 73.8	82.2, 52.2	**89.4**, **75.6**
RSDDS-113	Texture	1	26.7, 2.6	64.3, 30.5	94.8, 83.5	94.2, 67.7	**96.3**, **88.6**
ISIC	Skin	1	33.1, 5.8	83.3, 55.1	89.7, 78.4	77.6, 39.7	**91.2**, **83.7**
Kvasir	Colon	1	44.6, 17.7	69.7, 24.5	78.9, 45.1	74.6, 36.8	**78.8**, **45.2**
TN3k	Thyroid	1	42.3, 7.3	70.7, 39.8	**81.4**, 50.5	77.3, **53.4**	80.2, 47.9
MSD	Lung	1	41.1, 8.7	88.5, 60.8	89.8, 59.9	86.3, 50.5	**93.3**, **71.2**

Table 2: ZSAD segmentation performance comparison (AUROC, AUPRO). The best performance is marked in **red** and the second best in blue.

AUPRO as performance evaluation metrics for anomaly segmentation. The SOTA competing methods include CLIP (Radford et al. 2021), WinCLIP (Jeong et al. 2023), AnomalyCLIP (Zhou et al. 2023), MuSc (Li et al. 2024a).

Quantitative and Qualitative Results

Zero-shot generalized anomaly classification In Table 1, we compare zero-shot generalized anomaly classification results with prior works over 4 industrial defect datasets from very different production lines and 3 medical image datasets of different organs across different imaging devices.

TPS achieves superior ZGAD performance on classification results for most datasets, surpassing the other four methods in both effectiveness and generalization. The relatively weak performance of CLIP can be attributed to its pretraining, which primarily focuses on aligning object semantics rather than anomaly semantics. WinCLIP and AnomalyCLIP use global text prompts to align textual features with both global and local visual features, leading to biased alignment in the classification task. MuSc uses only a visual encoder without the support of text prompts, limiting it in generalisability, and despite its superior results on industrial images, it is relatively poor on medical images. TPS achieves the most superior classification performance on medical images while having the second best performance on industrial images. This is achieved due to our pathway module, which allows for feedback from the segmentation task to inform the semantic information crucial for classification, while performing the classification and segmentation tasks separately. Consequently, the global text embeddings produced by TPS are more effectively aligned with global visual embeddings across diverse target domains.

Zero-shot generalized anomaly segmentation In Table 2, we compare zero-shot generalized anomaly segmentation results with prior works over 6 industrial defect datasets from very different production lines and 4 medical image datasets of different organs across different imaging devices.

Similar to the results of classification, MuSc is more specialized in industrial images, but in due to the fact that medical images of individual patients usually contain only normal or abnormal samples, leading to inapplicability of the MuSc and poor detection results. It is remarkable that methods like AnomalyCLIP and TPS obtain promising ZSAD performance on various medical image datasets, even though they are tuned using a defect detection dataset. Among all these methods, TPS performs best benefit from its robust generalisation and high accuracy resulting from fine-grained alignment. This is particularly evident with medical images featuring complex backgrounds, such as those in the MSD dataset and ISIC dataset. More detailed experimental results are available in Appendix A.

Further validation on medical images Despite achieving excellent performance on medical datasets, TPS performs

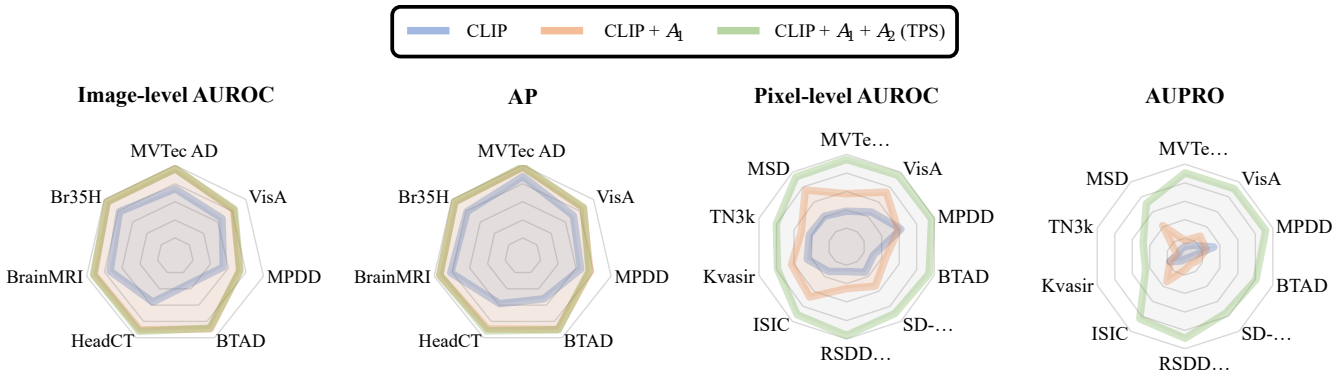

Figure 3: Ablation experiments of A_1 and A_2.

Task	Dataset	TPS (Industrial)	TPS (Medical)
Classification	HeadCT	92.3, 91.9	**99.9**, **99.9**
	BrainMRI	92.4, 93.8	**96.8**, **98.1**
	Br35H	96.2, 96.1	**98.4**, **98.5**
Segmentation	ISIC	91.2, 83.7	**99.6**, **98.1**
	Kvasir	78.8, 45.2	**86.1**, **50.0**
	TN3k	80.2, **47.9**	**86.5**, 44.4
	MSD	93.3, 71.2	**95.0**, **81.9**

Table 3: ZSAD classification (AUROC, AP) and segmentation (AUROC, AUPRO) performance comparison of TPS trained with medical dataset. The best performance is marked in **red**.

weakly on medical datasets relative to industrial datasets. Therefore, we test whether TPS can improve the performance on medical images if it is trained on an auxiliary medical dataset. We train TPS on HeadCT for testing zero-shot generalized anomaly classification performance and on ISIC for testing zero-shot generalized anomaly segmentation performance. The results are presented in Table 3. TPS has shown improved performance in detecting brain tumours in datasets such as HeadCT, BrainMRI and Br35H. This is due to the similarity of the brain images. The performance improvement of TPS on TN3k was not as good as the other datasets, which may be due to the fact that the thyroid is not similar to the other datasets in terms of visual features. Overall, TPS can further improve the ZGAD performance of medical images by training on medical auxiliary datasets.

Ablation Study

In this section, we validate the effectiveness of different high-level modules of our TPS, including shunt text prompts (A_1) and pathway module (A_2). Experiment A_1 uses the global text prompts [*a photo of a normal class with all parts flawless*] and [*a photo of a anomalous class with damaged parts*], shunt text prompts [*a photo of the flawless part*] and [*a photo of the damaged part*], where [*class*] denotes the category of the object being tested. Other than the difference in text prompts and the removal of the pathway module, the other parameters of the A_1 remain consistent with the TPS. Experimental A_2 is equivalent to TPS.

As shown in Figure 3, each module contributes to the remarkable performance of TPS. In particular, the improvement of global text prompts enables CLIP to focus on both the category information and individual parts of the foreground object, thereby improving the detection of anomalous regions. This refinement enhances the alignment of global text prompts for anomaly classification. The shunt text prompts mitigate the impact of complex contexts on segmentation, allowing CLIP to concentrate on the finer-grained features of individual instances. This offer a more precise visual-text alignment strategy for anomaly segmentation. The pathway module creates dependencies between classification and segmentation tasks, guiding anomaly segmentation with global semantic features and providing feedback on instance-level semantic features for classification. Therefore, the introduction of the pathway module substantially improve TPS.

Conclusion

In this paper, we identify the granularity disparity between anomaly classification and segmentation tasks as a key factor impacting their coordination in ZGAD. To address this challenge, we propose the TPS method. Specifically, TPS introduces fine-grained shunt text prompts that are precisely aligned with pixel-level embeddings for the segmentation task. We used improved global text prompts and shunt text prompts as the text inputs for these tasks, respectively, allowing the tasks to be carried out separately and avoiding alignment errors caused by granularity differences. Furthermore, we incorporate pathway modules to reconstruct the dependencies between these independently performed tasks, facilitating mutual learning and enhancing both classification and segmentation performance. Overall, we achieved pixel-level comprehensible CLIP for the first time in the ZGAD task with excellent performance. Extensive experimental results on 13 diverse public datasets, including industrial and medical images, demonstrate that TPS has superior ZGAD performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62322117, 62371365, U24B20136, and U22B2014.

References

Ahmed, H. 2020. Br35h: Brain tumor detection 2020. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature Communications*, 13(1): 4128.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35: 32897–32912.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 168–172. IEEE.

Gong, H.; Chen, G.; Wang, R.; Xie, X.; Mao, M.; Yu, Y.; Chen, F.; and Li, G. 2021. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 257–261. IEEE.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1932–1940.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8472–8480.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Hu, T.; Zhang, J.; Yi, R.; Du, Y.; Chen, X.; Liu, L.; Wang, Y.; and Wang, C. 2024. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8526–8534.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. https://doi.org/10.5281/zenodo.5143773.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; and Skotak, M. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, 66–71. IEEE.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; De Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, 451–462. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, X.; Huang, Z.; Xue, F.; and Zhou, Y. 2024a. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. *arXiv preprint arXiv:2401.16753*.

Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; and Li, J. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855*.

Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; and Ma, L. 2024b. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16838–16848.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, Y.; Zhu, H.; Liu, M.; Yu, H.; Chen, Z.; and Gao, J. 2024. Rolling-Unet: Revitalizing MLP's Ability to Efficiently Extract Long-Distance Dependencies for Medical Image Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3819–3827.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics*, 01–06. IEEE.

Niu, M.; Song, K.; Huang, L.; Wang, Q.; Yan, Y.; and Meng, Q. 2020. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Transactions on Industrial Informatics*, 17(3): 2271–2281.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Reiss, T.; and Hoshen, Y. 2023. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2155–2162.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14902–14912.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shen, Y.; Fu, C.; Chen, P.; Zhang, M.; Li, K.; Sun, X.; Wu, Y.; Lin, S.; and Ji, R. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13193–13203.

Song, G.; Song, K.; and Yan, Y. 2020. Saliency detection for strip steel surface defects using multiple constraints and improved texture features. *Optics and Lasers in Engineering*, 128: 106000.

Wang, M.; Xing, J.; Jiang, B.; Chen, J.; Mei, J.; Zuo, X.; Dai, G.; Wang, J.; and Liu, Y. 2024. A Multimodal, Multi-Task Adapting Framework for Video Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5517–5525.

Wu, J.; and Xu, M. 2024. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11302–11312.

Yang, H.; Pan, L.; Yang, Y.; Hartley, R.; and Liu, M. 2024. LDP: Language-driven Dual-Pixel Image Defocus Deblurring Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24078–24087.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*.

Zhu, J.; and Pang, G. 2024. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17826–17836.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 392–408. Springer.