ERP: EMPIRICAL REPRESENTATION PROTOTYPING FOR LANGUAGE-FREE ZERO-SHOT ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent language-free zero-shot anomaly detection (ZSAD) methods have shown that textual prompts can be replaced by learnable normal and abnormal visual representations. However, these representations are typically optimized as free parameters through classification and segmentation objectives, serving primarily as discriminative anchors without being explicitly constrained to represent the visual patterns observed in training data. In this work, we revisit how normality and abnormality should be represented in language-free ZSAD and propose Empirical Representation Prototyping (ERP), which replaces optimization-based decision vectors with data-derived visual prototypes. Specifically, we introduce an EMA-based representation prototyping mechanism that continuously aggregates normal and abnormal features from training samples to construct representative visual references. For anomaly segmentation, we further develop an importance-aware prototype construction strategy that emphasizes ambiguous normal patches while preserving spatially diverse normal patterns, preventing prototypes from being dominated by trivial regions. In addition, a cosine-based multi-scale feature aggregation strategy enhances local representations for fine-grained anomaly localization. Without relying on textual prompts, ERP achieves strong performance across 12 industrial and medical anomaly detection benchmarks, demonstrating that empirical visual prototypes provide an effective alternative to freely optimized decision representations for language-free ZSAD.

1 INTRODUCTION

Visual anomaly detection (VAD) (Pang et al., 2021; Ruff et al., 2021) aims to identify abnormal samples or regions that deviate from normal visual patterns and plays an important role in applications such as industrial inspection and medical diagnosis (Bergmann et al., 2019; Fernando et al., 2021). Since collecting exhaustive anomaly annotations is often difficult, zero-shot anomaly detection (ZSAD) has attracted increasing attention for its ability to generalize anomaly recognition beyond the categories observed during training (Xu et al., 2026).

Vision-language models, particularly CLIP (Radford et al., 2021), have substantially advanced ZSAD by transferring semantic knowledge encoded in textual descriptions. Early methods such as WinCLIP (Jeong et al., 2023) employ manually designed normal and abnormal prompts, while subsequent approaches learn object-agnostic (Zhou et al., 2023) or image-conditioned (Cao et al., 2024; Chen et al., 2026) prompts to reduce the dependence on manual prompt engineering. Despite their effectiveness, increasingly sophisticated prompt designs raise a more fundamental question about whether textual representations are necessary for anomaly detection. Recent language-free approaches such as VisualAD (Hou et al., 2026), UniADet (Gao & Wang, 2026), and AnomalyVFM (Fučka et al., 2026) provide an affirmative answer by removing the text encoder and performing anomaly prediction directly in the visual representation space.

However, existing language-free methods retain another unexplored assumption: normality and abnormality are still represented by learnable parameters. Typically, a pair of normal and abnormal prompts is optimized through classification or segmentation losses and subsequently used as visual decision references. Although effective for discrimination, these prompts are freely optimized

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

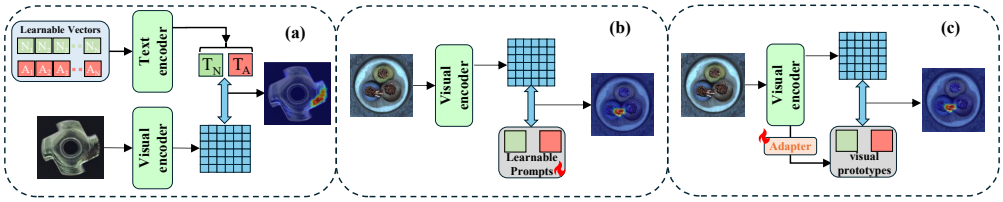

Figure 1: Comparison of different paradigms for zero-shot anomaly detection. Existing VLM-based methods rely on textual representations, while recent language-free approaches replace them with learnable discriminative prompts. In contrast, our ERP constructs data-derived representation prototypes from visual features to explicitly model representative normal and abnormal patterns.

parameters whose positions are determined indirectly by task objectives. They are not explicitly required to represent the empirical normal and abnormal feature distributions observed in the data. This distinction is particularly relevant to zero-shot anomaly detection, where the learned representation must capture category-agnostic visual regularities rather than merely fit source-domain decision boundaries. This raises a natural question: do language-free anomaly references need to be learned as freely optimized parameters at all?

We explore an alternative perspective: rather than optimizing what normality and abnormality should look like, their representations can be directly derived from what the model actually observes. Based on this idea, we propose Empirical Representation Prototyping (ERP), which replaces freely optimized decision vectors with visual prototypes constructed from training features. During training, normal and abnormal representations are continuously accumulated through exponential moving averages, anchoring anomaly references to the empirical feature distribution while retaining the efficiency of prototype-based prediction. Following the task-specific design of anomaly detection (Gao & Wang, 2026), ERP maintains separate prototypes for image-level classification and pixel-level segmentation across multiple feature layers. Unlike learnable prompts or vectors that are directly updated by gradients, these prototypes evolve from transformed visual features and therefore explicitly summarize the normal and abnormal patterns observed in the data.

For pixel-level anomaly localization, naively averaging all normal patches may cause the prototype to be dominated by easily recognized regions while underrepresenting ambiguous normal patterns. To address this issue, we introduce Importance-aware Prototype Construction (IPC), which emphasizes hard normal patches while preserving spatially diverse reliable normal references, yielding more informative prototypes for anomaly localization. We further introduce Cosine similarity-based Multi-scale Feature Aggregation (CMFA), which incorporates multi-scale neighborhood context into patch representations before prototype matching. Together, these designs enable normality and abnormality to be represented by empirical visual prototypes derived from observed features rather than freely optimized decision representations.

Extensive experiments across 12 industrial and medical anomaly detection benchmarks demonstrate that data-derived visual prototypes provide a strong alternative to optimization-based anomaly references. The proposed approach consistently improves language-free ZSAD performance across diverse anomaly detection scenarios while retaining a simple and efficient inference pipeline.

The main contributions of this paper are summarized as follows:

- We propose **ERP**, a language-free ZSAD framework that replaces freely optimized decision representations with data-derived visual prototypes, providing an alternative way to represent normality and abnormality from observed visual patterns.
- We construct normal and abnormal prototypes through exponential moving averages, and further introduce IPC and CMFA to improve prototype quality and local feature representations for fine-grained anomaly localization.
- Extensive experiments on 12 industrial and medical anomaly detection benchmarks demonstrate the effectiveness of data-derived visual prototypes for language-free ZSAD, achieving consistent improvements over existing approaches while retaining a simple inference pipeline.

2 Related Work

Text-Prompt-Based ZSAD. CLIP (Radford et al., 2021), with its strong cross-domain transferability, has been widely adopted in zero-shot anomaly detection. Early methods such as WinCLIP (Jeong et al., 2023), APRIL-GAN (Chen et al., 2023), and CLIP-AD (Chen et al., 2024b) employ handcrafted textual prompts to describe normal and abnormal states. To reduce the dependence on manually designed prompts, AnomalyCLIP (Zhou et al., 2023) introduces object-agnostic learnable prompts. Subsequent works, including AdaCLIP (Cao et al., 2024), CoPS (Chen et al., 2026), and FAPrompt (Zhu et al., 2025), further explore image-conditioned or fine-grained prompt adaptation to improve generalization to diverse anomaly patterns. Beyond prompt learning, MoECLIP (Park et al., 2026) introduces a Mixture-of-Experts (Shazeer et al., 2017) architecture to enhance local representations, while FiLo (Gu et al., 2024) employs multi-scale convolutional operations for fine-grained anomaly localization. Despite their effectiveness, these methods remain dependent on textual representations and often require increasingly sophisticated prompt learning or feature enhancement designs.

Language-Free ZSAD. Recent studies have begun to question whether textual representations are necessary for zero-shot anomaly detection and explore language-free paradigms that operate directly in the visual representation space. AnomalyVFM (Fučka et al., 2026) adapts visual foundation models with LoRA (Hu et al., 2022) and prediction heads, achieving competitive performance without textual prompts. VisualAD (Hou et al., 2026) further demonstrates that text embeddings can be replaced by lightweight learnable visual representations, while UniADet (Gao & Wang, 2026) decouples such representations across feature hierarchies and anomaly detection tasks. Different from these optimization-based approaches, MuSc (Li et al., 2024) performs training-free anomaly detection by exploiting feature correspondences among unlabeled test images. Our ERP follows the language-free paradigm but takes a different perspective on anomaly representation: instead of optimizing normal and abnormal references as learnable parameters, ERP directly constructs them from observed visual features through exponential moving averages, providing a data-derived representation mechanism for modeling normality and abnormality.

3 Method

3.1 Problem Setting

In ZSAD, We aim to learn a model from an auxiliary dataset that generalizes to unseen target domains for detecting abnormal images, without fine-tuning on test data. Let

$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i, \mathbf{G}_i)\}_{i=1}^{N}$$

denote the training set, where $\mathbf{x}_i \in \mathbb{R}^{h\times w\times 3}$ is an RGB image, $y_i \in \{0, 1\}$ is the image-level label ($y_i = 0$ for normal, $y_i = 1$ for abnormal), and $\mathbf{G}_i \in \{0, 1\}^{h\times w}$ is the pixel-level ground-truth anomaly mask, with 1 indicating abnormal pixels.

During testing, we are given a collection of test datasets

$$\mathcal{T} = \{\mathcal{D}_{\text{test}}^1, \mathcal{D}_{\text{test}}^2, \ldots, \mathcal{D}_{\text{test}}^K\},$$

where each $\mathcal{D}_{\text{test}}^j$ contains only unlabeled images $\{\mathbf{x}_k\}$ from a distinct domain, with potentially different object categories and anomaly patterns from those in $\mathcal{D}_{\text{train}}$. No ground-truth labels or masks are available for any test domain.

For any test image $\mathbf{x} \in \mathbb{R}^{h\times w\times 3}$, the model must output:

- an image-level anomaly score $s \in [0, 1]$, indicating the global abnormality of the image;
- a pixel-level anomaly map $\mathbf{A} \in [0, 1]^{h\times w}$, where each entry quantifies the abnormality likelihood of the corresponding pixel.

Higher scores indicate higher probability of being abnormal. The model is trained solely on $\mathcal{D}_{\text{train}}$ and is expected to generalize across diverse target domains without using any unseen data from $\mathcal{T}$.

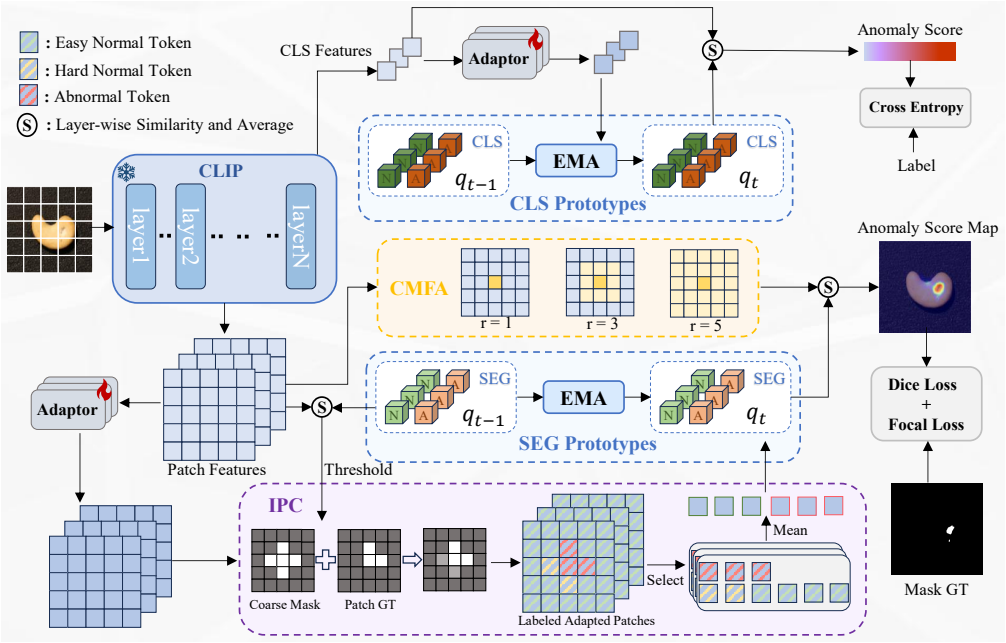

Figure 2: **Framework of the proposed method.** ERP constructs task-specific normal and abnormal prototypes from adapted visual features through exponential moving averages, while IPC selects informative normal patches to improve segmentation prototype construction. During inference, the CMFA aggregates multi-scale contextual information to enhance local representations, and anomaly predictions are obtained through feature-to-prototype matching.

3.2 Framework Overview

As shown in Fig. 2, our ERP consists of a frozen visual encoder and lightweight adapters for extracting and transforming multi-layer visual features. During training, ERP accumulates normal and abnormal adapted visual features through exponential moving average to construct source representation prototypes for image-level classification and pixel-level segmentation. To improve the representativeness of normal segmentation prototypes, IPC selects informative normal patches by emphasizing ambiguous normal patterns while preserving spatially diverse normal adapted visual features with high confidence. During inference, CMFA enhances visual features through cosine-based multi-scale contextual aggregation to incorporate local contextual information. Finally, the image-level anomaly score and pixel-level anomaly map are computed by measuring the similarity between the visual features and the corresponding source representation prototypes.

3.3 Empirical Representation Prototyping

Representation prototypes. Following UniADet (Gao & Wang, 2026), we decouple the source representation prototypes across different layers and tasks. Specifically, given multi-layer visual features extracted from the frozen encoder, lightweight bottleneck-like adapters are applied to obtain task-specific representations. Let Ω denote the set of selected feature layers. For each layer $\ell \in \Omega$, the original class token and patch features are denoted as $\mathbf{c}^\ell \in \mathbb{R}^d$ and $\mathbf{p}^\ell \in \mathbb{R}^{N_\ell \times d}$, respectively, while their adapted counterparts are denoted as $\tilde{\mathbf{c}}^\ell \in \mathbb{R}^d$ and $\tilde{\mathbf{p}}^\ell \in \mathbb{R}^{N_\ell \times d}$. Here, d is the feature dimension and N_ℓ is the number of patches. For each layer ℓ and task branch $b \in \{\text{cls}, \text{seg}\}$, ERP maintains two source representation prototypes corresponding to normal and abnormal patterns, denoted as $\{\mathbf{q}_n^{b,\ell}, \mathbf{q}_a^{b,\ell}\}$, where $\mathbf{q}_r^{b,\ell} \in \mathbb{R}^d$ and $r \in \{n, a\}$ indicates the normal or abnormal state.

Given a visual feature $\mathbf{f} \in \mathbb{R}^d$ and the corresponding normal/abnormal prototype pair $\{\mathbf{q}_n, \mathbf{q}_a\}$, ERP computes the abnormal probabilities based on prototype similarity:

$$\pi_a(\mathbf{f}) = \frac{\exp(\langle \text{norm}(\mathbf{f}), \text{norm}(\mathbf{q}_a) \rangle / \tau)}{\sum_{k \in \{n,a\}} \exp(\langle \text{norm}(\mathbf{f}), \text{norm}(\mathbf{q}_k) \rangle / \tau)}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is the temperature coefficient which is set to 0.07.

During training, only the lightweight adapters are optimized by loss, while the prototypes are updated independently. Specifically, the source representation prototypes are updated using directional exponential moving average:

$$\mathbf{q}_r^{b,\ell} \leftarrow \operatorname{norm}\left(\alpha \operatorname{norm}(\mathbf{q}_r^{b,\ell}) + (1-\alpha)\operatorname{norm}(\boldsymbol{\mu}_r^{b,\ell})\right), \quad r \in \{n, a\}, \tag{2}$$

where $\boldsymbol{\mu}_r^{b,\ell} \in \mathbb{R}^d$ denotes the selected adapted feature mean for branch b, layer ℓ, and state r, and α is the EMA momentum which is defaultly set to 0.97. For the classification branch, $\boldsymbol{\mu}_n^{\mathrm{cls},\ell}$ and $\boldsymbol{\mu}_a^{\mathrm{cls},\ell}$ are calculated from the adapted class tokens of normal and abnormal images within a mini-batch, respectively. For the segmentation branch, the corresponding adapted feature means are constructed by the IPC mentioned later.

Finally, ERP measures the similarity between visual features and source representation prototypes. The final anomaly score map and anomaly score are computed as:

$$\begin{aligned} \mathbf{A}_{\mathrm{src}} &= \frac{1}{|\Omega|}\sum_{\ell\in\Omega} \operatorname{Upsample}\left(\pi_a^{\mathrm{seg},\ell}(\mathbf{p}^\ell)\right), \\ s_{\mathrm{src}} &= \frac{1}{2}\left(\frac{1}{|\Omega|}\sum_{\ell\in\Omega}\pi_a^{\mathrm{cls},\ell}(\mathbf{c}^\ell) + \operatorname{TopKMean}_\gamma(\mathbf{A}_{\mathrm{src}})\right), \end{aligned} \tag{3}$$

where $\mathbf{A}_{\mathrm{src}} \in \mathbb{R}^{H\times W}$ denotes the anomaly score map and s_{src} denotes the anomaly score.Following VisualAD (Hou et al., 2026), γ is set to 0.01.

Importance-aware Prototype Construction (IPC). Naively averaging all normal patches may cause the normal segmentation prototypes to be dominated by easily recognized regions, while ambiguous normal patterns are diluted. To construct more informative segmentation prototypes, we introduce IPC, which selects normal patches according to their anomaly-aware importance. Specifically, the ground-truth mask is first resized to the patch resolution. For image i and patch j at layer ℓ, let $\pi_{i,j,a}^\ell$ denote the abnormal probabilities obtained from Eq. 1, $\tilde{\mathbf{p}}_{i,j}^\ell \in \mathbb{R}^d$ denote the adapted patch feature, and $m_{i,j} \in \{0,1\}$ denote the ground-truth patch label. IPC identifies hard normal patches as

$$\mathcal{H}_i^\ell = \left\{ j \mid m_{i,j} = 0,\ \pi_{i,j,a}^\ell > 0.5 \right\}. \tag{4}$$

In implementation, IPC retains all hard normal patches, and selects the remaining easy normal patches by partitioning the patch map into a 3×3 spatial grid and independently picking the top 1% high confidence ones from each grid, ensuring that reliable normal references are collected from diverse regions rather than concentrated in a single area. The selected normal patches are denoted as $\mathcal{S}_{i,n}^\ell$. The segmentation prototypes are then updated using the selected normal patches and all abnormal patches:

$$\boldsymbol{\mu}_n^{\mathrm{seg},\ell} = \frac{1}{|\mathcal{S}_n^\ell|}\sum_i \sum_{j\in\mathcal{S}_{i,n}^\ell} \tilde{\mathbf{p}}_{i,j}^\ell, \qquad \boldsymbol{\mu}_a^{\mathrm{seg},\ell} = \frac{1}{|\mathcal{S}_a^\ell|}\sum_i \sum_{j:m_{i,j}=1} \tilde{\mathbf{p}}_{i,j}^\ell, \tag{5}$$

where $\mathcal{S}_n^\ell = \bigcup_i \mathcal{S}_{i,n}^\ell$ and $\mathcal{S}_a^\ell$ contains all ground-truth abnormal patches. By emphasizing ambiguous normal regions and preserving abnormal appearances, IPC provides more informative segmentation prototypes, which are subsequently updated through Eq. 2.

Training Objectives Given the image-level label y and pixel-level anomaly mask $\mathbf{G} \in \{0,1\}^{h\times w}$, the lightweight adapters are optimized with image-level classification and pixel-level segmentation:

$$\mathcal{L}_{\mathrm{total}} = \mathrm{CE}(s_{\mathrm{src}}, y) + \mathrm{Focal}(\mathbf{A}_{\mathrm{src}}, G) + \mathrm{Dice}(\mathbf{A}_{\mathrm{src}}, G) \tag{6}$$

Focal Loss (Lin et al., 2017) and Dice Loss (Li et al., 2020) are adopted for anomaly segmentation.

3.4 Cosine-based Multi-scale Feature Aggregation (CMFA)

To improve fine-grained anomaly localization, we propose CMFA, which enhances local patch representations by incorporating multi-scale contextual information before prototype matching. For a

Table 1: Comparison with state-of-the-art ZSAD methods. ✓ and × indicate with and without **text encoder**, respectively. Red and blue denote the best and second-best results, respectively. †Results reproduced from UniADet (Gao & Wang, 2026).

(a) Image-level ZSAD (I-AUROC / I-AUPR).

Dataset ↓	AnomalyCLIP† (✓)	AdaCLIP† (✓)	Bayes-PFL† (✓)	VisualAD (×)	UniADet† (×)	**ERP** (×)
Industrial						
MVTec	91.6, 96.4	90.7, 95.6	92.3, 96.7	92.2, 96.7	92.4, 96.7	95.2, 97.6
VisA	82.0, 85.3	81.7, 84.3	87.0, 89.2	84.7, 87.6	88.0, 89.1	88.2, 90.2
BTAD	88.3, 88.2	89.9, 95.5	93.2, 96.5	94.9, 97.0	96.4, 97.8	95.9, 98.3
DTD	93.9, 97.2	92.7, 96.4	95.1, 98.4	97.5, 99.1	98.1, 99.1	96.1, 98.6
Real-IAD	69.5, 64.3	73.3, 70.2	70.0, 65.8	75.0, 72.0	78.6, 76.7	78.7, 76.1
Mean	85.1, 86.3	85.7, 88.4	87.5, 89.3	88.9, 90.5	90.7, 91.9	90.8, 92.2
Medical						
HeadCT	93.0, 92.6	93.4, 92.2	96.5, 95.5	90.3, 88.0	97.0, 96.7	97.8, 97.9
BrainMRI	90.2, 92.4	94.9, 94.2	96.2, 92.4	96.7, 97.6	95.7, 96.4	97.2, 98.4
Br35H	94.2, 94.2	95.7, 95.7	97.8, 96.2	95.5, 96.2	97.4, 97.1	97.6, 98.0
Mean	92.5, 93.1	94.7, 94.0	96.8, 94.7	94.2, 93.9	96.7, 96.7	97.5, 98.1

(b) Pixel-level ZSAD (P-AUROC / P-AUPR).

Dataset ↓	AnomalyCLIP† (✓)	AdaCLIP† (✓)	Bayes-PFL† (✓)	VisualAD (×)	UniADet† (×)	**ERP** (×)
Industrial						
MVTec	91.1, 34.5	88.3, 39.1	91.8, 48.3	90.8, 41.2	91.8, 42.8	92.2, 46.6
VisA	95.5, 21.3	95.7, 31.0	95.6, 29.8	95.8, 28.4	95.8, 28.0	96.2, 29.2
BTAD	94.2, 45.5	91.6, 42.9	93.9, 47.1	91.1, 43.1	95.7, 42.8	95.9, 49.8
DTD	97.9, 62.6	98.3, 75.2	97.8, 69.9	98.1, 65.5	98.4, 64.3	98.1, 66.2
Real-IAD	95.1, 26.7	96.1, 30.5	96.5, 27.6	96.8, 34.7	96.6, 33.6	96.6, 35.0
Mean	94.8, 38.1	94.0, 43.7	95.1, 44.5	94.5, 42.6	95.7, 42.3	95.8, 45.4
Medical						
ColonDB	81.9, 31.7	79.3, 26.2	82.1, 31.9	76.8, 27.1	84.0, 36.2	85.4, 36.3
ClinicDB	82.9, 38.2	84.3, 36.0	89.6, 53.2	85.0, 37.8	90.1, 55.8	91.2, 55.4
Endo	84.2, 46.6	84.0, 44.8	89.2, 58.6	84.5, 46.0	90.7, 62.5	92.0, 65.4
Kvasir	79.1, 39.8	79.4, 43.8	85.4, 54.2	80.4, 47.5	87.5, 57.5	89.3, 60.5
Mean	82.0, 39.1	81.8, 37.7	86.6, 49.5	81.7, 39.6	88.1, 53.0	89.5, 54.4

patch feature $\mathbf{p}_j \in \mathbb{R}^d$ and its neighboring patch features within an $s \times s$ window, denoted as $\mathcal{N}_s(j)$, CMFA computes center-to-neighbor cosine similarity as attention:

$$a^s_{jk} = \frac{\exp(\langle \text{norm}(\mathbf{p}_j), \text{norm}(\mathbf{p}_k) \rangle)}{\sum_{u \in \mathcal{N}_s(j)} \exp(\langle \text{norm}(\mathbf{p}_j), \text{norm}(\mathbf{p}_u) \rangle)}, \quad \hat{\mathbf{p}}^s_j = \mathbf{p}_j + \sum_{k \in \mathcal{N}_s(j)} a^s_{jk} \mathbf{p}_k. \tag{7}$$

where j, k, u denote patch indices, and s denotes the aggregation window size. The outputs from $s \in \{1, 3, 5\}$ are averaged to obtain the final enhanced patch representations, which are then matched with the corresponding source representation prototypes. In this way, CMFA strengthens locally consistent anomaly responses while suppressing isolated feature noise without introducing additional parameters.

4 Experiments

Datasets. We evaluate the zero-shot anomaly detection performance on 12 public datasets, covering both industrial and medical domains. In the industrial domain, we adopt 5 widely used datasets: MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), DTD-

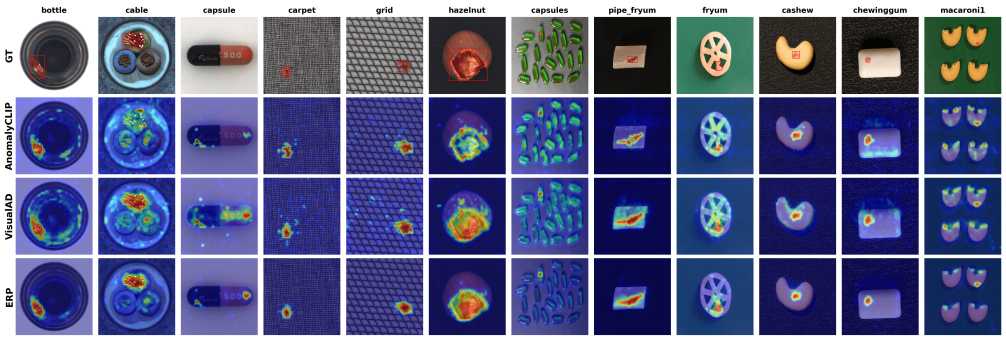

Figure 3: Comparison of anomaly segmentation results between ERP and other works.

Synthetic (Aota et al., 2023), and Real-IAD (Wang et al., 2024). In the medical domain, we include 7 datasets: HeadCT (Salehi et al., 2021), BrainMRI (Lamrani et al., 2022), Br35H (Hamada, 2020), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Tajbakhsh et al., 2015), Kvasir (Jha et al., 2019), Endo (Vázquez et al., 2017). Following recent approaches (Gao & Wang, 2026; Xu et al., 2026; Chen et al., 2024a), we adopt a single industrial dataset as the auxiliary dataset and directly evaluate on the remaining industrial and all medical datasets. By default, VisA is used as the auxiliary dataset; when evaluating VisA itself, we instead use MVTec-AD as the auxiliary dataset.

Evaluation Metrics. We evaluate image-level anomaly classification and pixel-level anomaly segmentation performance using the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR). Higher values indicate better performance.

Implementation Details. We adopt the publicly available pretrained weights of CLIP (ViT-L/14-336) as feature extractors. When using CLIP, input images are resized to 518×518. We extract both global and local visual features from five intermediate layers $\{12, 15, 18, 21, 24\}$.We optimize the model using Adam (Kingma & Ba, 2015) with a learning rate of 0.001 and a batch size of 8. Regarding IPC, the top 1% high-confidence easy normal patches are selected from each region of a 3×3 spatial partition to update the normal representation prototypes of the segmentation branch. Each adapter is implemented as a lightweight bottleneck MLP with a reduction ratio of 8. For cross-domain medical evaluation, prototype matching is performed in the adapter-transformed feature space. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 48GB memory. The reported performance for each dataset is averaged over all object categories.

4.1 Comparisons with State-of-the-Art Methods

Quantitative Comparison. As shown in Table 1, we report results on 12 benchmarks spanning industrial and medical domains. ERP achieves competitive performance on image-level detection and pixel-level AUROC, while showing a more evident advantage on P-AUPR, which reflects the quality of fine-grained anomaly localization. In particular, ERP achieves the highest average P-AUPR on both industrial and medical benchmarks, reaching 45.4% and 54.4%, respectively. Compared with the language-free UniADet, ERP improves the average P-AUPR by 3.1% on industrial datasets (45.4% vs. 42.3%) and by 1.4% on medical datasets (54.4% vs. 53.0%). These results suggest that source representation prototypes provide more accurate pixel-level anomaly localization, supporting the effectiveness of explicitly modeling representative visual patterns rather than relying solely on optimization-based discriminative representations.

Qualitative Comparison. As evidenced by the quantitative results, ERP achieves better performance on the pixel-AUPR metric for anomaly segmentation, indicating a substantial advantage in pixel-level anomaly boundary localization. As shown in Fig. 3, ERP generates compact and well-localized anomaly maps that closely align with the annotated abnormal regions. Benefiting from the source representation prototype mechanism and multi-scale contextual feature aggregation, ERP accurately identifies abnormal pixels while effectively suppressing those false positives. These visual results corroborate that ERP enables more precise anomaly segmentation, which directly accounts for its superior pixel-AUPR performance.

Table 2: Component ablation of ERP on MVTec, VisA, and BTAD. ✓ and × indicate with and without the corresponding module. The best performance in each column is highlighted in bold.

CMFA	IPC	MVTec				VisA				BTAD			
		P-AUR	P-AUP	I-AUR	I-AUP	P-AUR	P-AUP	I-AUR	I-AUP	P-AUR	P-AUP	I-AUR	I-AUP
✓	✓	**92.2**	**46.6**	**95.2**	**97.6**	**96.2**	**29.3**	88.2	**90.2**	**95.9**	**49.8**	**95.9**	**98.3**
×	✓	91.2	44.3	93.4	96.6	95.7	28.2	**88.6**	**90.2**	95.5	49.7	**95.9**	97.7
✓	×	92.0	46.7	94.1	97.1	96.1	28.7	87.9	89.8	95.8	48.4	94.8	98.1

Table 3: Effect of feature adaptation on ERP prototype. ✓ and × indicate with and without adapters . The best performance in each column is highlighted in bold.

Adapter	MVTec				VisA			
	P-AUROC	P-AUPR	I-AUROC	I-AUPR	P-AUROC	P-AUPR	I-AUROC	I-AUPR
✓	**92.2**	**46.6**	**95.2**	**97.6**	**96.2**	**29.3**	**88.2**	**90.2**
×	89.2	37.7	73.1	86.8	94.4	23.7	60.6	66.1

4.2 Ablation Study

Ablation on IPC. We first investigate the effectiveness of IPC for constructing normal segmentation prototypes in Table 2. Without IPC, the model directly aggregates normal patch features, which may cause the resulting prototypes to be dominated by abundant and easily recognized normal regions while underrepresenting ambiguous normal patterns. In contrast, IPC explicitly preserves hard normal patches that are more easily confused with anomalies, while retaining spatially diverse high-confidence normal patches to maintain broad coverage of the normal distribution. As shown in Table 2, removing IPC leads to performance degradation on most metrics across MVTec, VisA, and BTAD. In particular, P-AUPR decreases from 29.3% to 28.7% on VisA and from 49.8% to 48.4% on BTAD, while image-level performance also drops consistently. These results indicate that selectively accumulating informative normal patterns provides more effective prototypes than uniformly aggregating all normal regions.

Ablation on CMFA. We further evaluate the contribution of CMFA . CMFA enhances each patch representation by aggregating multi-scale neighborhood context according to feature similarity, thereby providing more context-aware local representations before prototype matching. Removing CMFA consistently degrades pixel-level anomaly localization, with P-AUPR decreasing from 46.6% to 44.3% on MVTec and from 29.3% to 28.2% on VisA. Similar drops are also observed in P-AUROC and image-level metrics, showing that local contextual aggregation benefits both anomaly localization and overall detection. These results demonstrate that CMFA complements ERP by improving the quality of local features used for prototype matching, particularly for fine-grained anomaly regions.

Ablation on Adapters. We further investigate the contribution of adapters. As shown in Table 3, constructing ERP prototypes with adapted visual features consistently outperforms using the original backbone features on both MVTec and VisA. The degradation is particularly pronounced for image-level detection, where removing the adapter decreases I-AUROC from 95.2% to 73.1% on MVTec and from 88.2% to 60.6% on VisA. Similar reductions are observed in pixel-level localization, especially in P-AUPR. These results demonstrate that task-specific feature adaptation is important for constructing effective prototypes from pretrained visual features.

Ablation on EMA momentum α. We further investigate the effect of the EMA momentum α on image-level performance. As shown in Fig. 4, a small α leads to inferior performance because the prototypes become overly dependent on visual patterns from the current iteration and retain less historical information. As the adapter is optimized, its feature space becomes increasingly task-specific, which may weaken the generalization inherited from the pretrained CLIP ViT. Conversely, when α exceeds 0.97, performance also declines, as early-stage patterns from the randomly initialized adapter receive excessive weight. Therefore, α should balance historical pattern preservation and adaptation to newly learned anomaly-aware representations.

Table 4: Comparison between learnable prompts and ERP prototypes under the same CMFA configuration. The best performance in each column is highlighted in bold.

Representation	MVTec				VisA			
	P-AUROC	P-AUPR	I-AUROC	I-AUPR	P-AUROC	P-AUPR	I-AUROC	I-AUPR
Learnable Prompts	**92.4**	45.6	94.0	97.3	96.1	28.7	**88.6**	**90.3**
ERP Prototypes	92.2	**46.6**	**95.2**	**97.6**	**96.2**	**29.3**	88.2	90.2

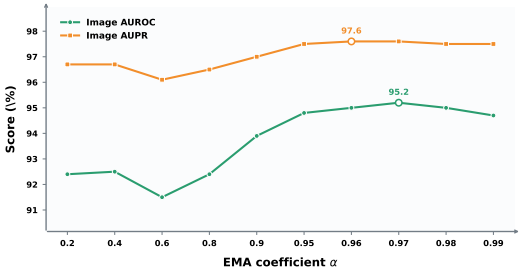

Figure 4: Ablation study of EMA momentum α.

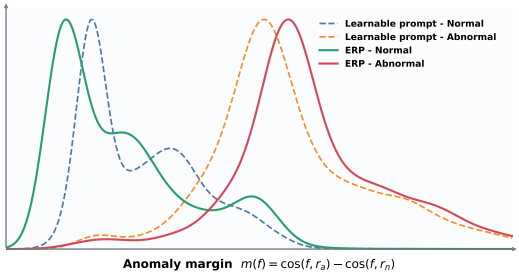

Figure 5: Anomaly-margin distributions for learnable prompts and ERP prototypes.

4.3 Discussion: Empirical Prototypes vs. Learnable Prompts

As shown in Table 4, under the same CMFA configuration, replacing learnable prompts with ERP improves P-AUPR from 45.6% to 46.6% on MVTec and from 28.7% to 29.3% on VisA. We observe that learnable normal and abnormal prompts are much more strongly separated than ERP prototypes, with cosine similarities around -0.9, compared with about -0.2 for ERP prototypes. This suggests that gradient optimization encourages learnable prompts to form a source-domain discriminative decision axis. However, since learnable prompts and visual features are jointly optimized under the same source-domain objectives, such prompt–feature co-adaptation may reshape the feature space toward source-specific discrimination, which does not necessarily transfer well to unseen categories. In contrast, ERP prototypes are updated from evolving visual features rather than independently optimized as free parameters, reducing excessive prompt–feature co-adaptation and potentially providing more transferable anomaly references.

To examine how these different representation mechanisms affect unseen samples, we analyze the anomaly-margin distributions, where $m(\mathbf{f}) = \cos(\mathbf{f}, \mathbf{r}_a) - \cos(\mathbf{f}, \mathbf{r}_n)$. As shown in Fig. 5, ERP yields a larger peak distance and a smaller overlap between the normal and abnormal distributions than learnable prompts. This indicates that, although learnable prompts are more strongly separated at the reference level, such separation does not necessarily translate into better normal–abnormal separability on unseen samples. ERP instead provides references that better preserve discriminative separation when transferred to unseen categories.

5 Conclusion

In this paper, we propose ERP, a language-free ZSAD framework that replaces optimization-based discriminative prompts with data-derived representation prototypes. Specifically, ERP constructs representative normal and abnormal prototypes from source visual features, explicitly modeling the observed visual patterns for both anomaly classification and segmentation. To improve prototype representativeness, IPC selectively emphasizes ambiguous normal patches while preserving spatially diverse normal representations. Furthermore, CMFA enhances local feature representations by incorporating multi-scale contextual information, facilitating anomaly localization. We evaluate ERP on 12 anomaly detection benchmarks spanning industrial and medical scenarios. Our framework demonstrates a prototype-based alternative to learnable discriminative prompts, highlighting the potential of explicitly modeling visual patterns for language-free zero-shot anomaly detection.

AI USE STATEMENT

In this work, we use LLM to draft experimental tables in LaTeX format, as well as to assist in translating text, correcting grammatical errors, and refining adverbs and conjunctions in the article. In the experiments, we utilize LLM to help establish the code framework and fix bugs. All domain concepts, research methods, experiment setup, and analysis of experimental results in this work are independently completed by authors.

REFERENCES

Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600, 2019.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 2024.

Qiyu Chen, Zhen Qu, Wei Luo, Haiming Yao, Yunkang Cao, Yuxin Jiang, Yinan Duan, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. Cops: Conditional prompt synthesis for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8554–8563, June 2026.

Xiaozhen Chen, Xiaokang Zhang, et al. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, 2024a.

Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge tracks 1&2: 1st place on zero-shot AD and 4th place on few-shot AD. *arXiv preprint arXiv:2305.17382*, 2023.

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 17–33. Springer, 2024b. Also available at arXiv:2311.00453.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys*, 54(7):1–37, 2021. doi: 10.1145/3464423.

Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Anomalyvfm – transforming vision foundation models into zero-shot anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Bin-Bin Gao and Chengjie Wang. One language-free foundation model is enough for universal vision anomaly detection. In *arXiv:2601.05552*, 2026.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. *arXiv preprint arXiv:2404.13671*, 2024.

Ahmed Hamada. Br35H: Brain tumor detection 2020, 2020.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21346–21356, June 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, June 2023.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-SEG: A segmented polyp dataset. In *International Conference on Multimedia Modeling (MMM)*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Driss Lamrani, Bouchaib Cherradi, Oussama El Gannour, Mohammed Amine Bouqentar, and Lhoussain Bahatti. Brain tumor detection using MRI images and convolutional neural network. *International Journal of Advanced Computer Science and Applications*, 13(7), 2022.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks, 2020. URL `https://arxiv.org/abs/1911.02855`.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 1–6. IEEE, 2021.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, March 2021. doi: 10.1145/3439950.

Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: Patch-specialized experts for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35534–35544, June 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14902–14912, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-Difference: Self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.