BEYOND COSINE SIMILARITY: UNCOVERING ANOMALY INFORMATION IN CLIP'S RAW EMBEDDING GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Extensive experiments following the image-text cosine similarity pre-training paradigm of Contrastive Language-Image Pre-training (CLIP) indicate that CLIP excels at image classification rather than fine-grained anomaly detection; consequently, a prevailing consensus holds that CLIP is anomaly-agnostic. Existing studies have concentrated on downstream training to better adapt CLIP for anomaly detection tasks; in other words, image-text cosine similarity is the computational foundation underlying CLIP-based anomaly detection. In this paper, we investigate the directional information about anomalies in CLIP's raw embedding space and the limitations of cosine similarity with the idea of computing geodesics distance for anomaly shifts. We investigate the raw representation of CLIP and find that anomaly information is already encapsulated in CLIP's raw embedding space, which is mainly formed by class-dependent directional variation. The key reason why CLIP cannot recognize the normality/abnormality in images is that the now-ubiquitous cosine similarity ignores the anisotropy of the normal/abnormal distribution. We demonstrate this through controlled geometric analysis, and use a top-down statistical method as the simplest parameter-free update to uncover the latent anomaly information, making this process interpretable. In extensive experiments, CLIP's latent anomaly information substantially improves AD performance over image-text cosine similarity.

1 INTRODUCTION

Vision-language models (VLMs) (Schuhmann et al., 2022; Alayrac et al., 2022; Radford et al., 2021) break down the barriers between modalities, enabling models to understand image content based on textual prompts. Through contrastive learning on massive image-text pairs, CLIP (Radford et al., 2021) has achieved success in image classification. However, when transferred to anomaly detection, it exhibits anomaly agnosticism, which stems from its pre-training objective of aligning foreground object semantics with text. Guided by the concept of image-text matching, existing CLIP-based AD methods primarily focus on two aspects: one is aligning textual prompts more closely with anomalies, such as the compositional prompt ensemble (Jeong et al., 2023), the object-agnostic anomaly prompts (Zhou et al., 2024), and the dynamic image-conditioned prompts (Cao et al., 2024); the other is extracting finer-grained image representations, such as the multi-window aggregation of patch features (Jeong et al., 2023), the self-correlation attention for enhancing local features (Salehi et al., 2025), and the multi-layer feature pyramid with DPAM alignment (Zhou et al., 2024).

Unlike these data-driven machine learning methods, we posit that normal/abnormal information already resides within CLIP's embedding space. Therefore, it is essential to proactively explore this latent information and release it via top-down statistical methods. Specifically, inspired by the geodesic distance between two points in curved space, we hypothesize that the contrastive loss curves CLIP's embedding space – creating local metric fields that intrinsically determine the representation of anomaly information; however, the now-ubiquitous L2-normalized image-text cosine similarity fails to read this information. Specifically, L2-normalized cosine similarity essentially projects the embedding space onto a unit hypersphere, and then computes the arc length between two vectors. But we cannot guarantee that the intrinsic information of the embedding space is pre-

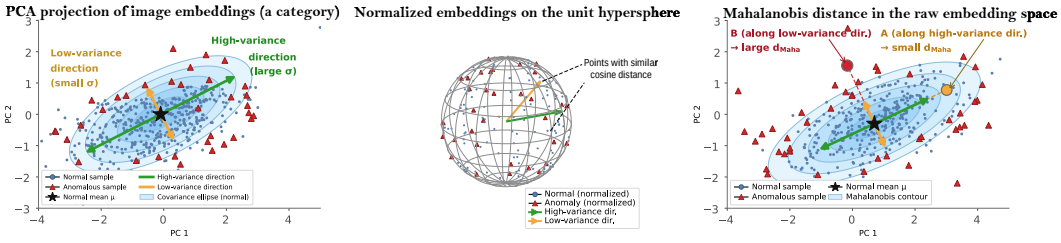
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

served during this process. In a general space, the shortest distance between two points is determined by geodesic distance. Even if the text prompt perfectly aligns with the anomaly direction, the shift of anomalies relative to normal samples in the raw embedding space can hardly be equivalent to the arc length between two L2-normalized vectors. Instead, the true deviation should be measured by the geodesic distance within the raw embedding space. Therefore, the rationale for transferring image-text cosine similarity to anomaly detection remains questionable. Based on existing research into CLIP's image and text embedding spaces' geometry (Liang et al., 2022; Levi & Gilboa, 2025), we are further convinced that L2-normalized cosine similarity fails to accurately measure anomaly deviations. Consequently, we need to re-evaluate CLIP's embedding space and develop the simplest instance model grounded in its representations.

In this paper, we investigate the local geometric structure sculpted by contrastive loss – specifically, positive pairs compress variance along the T^+ axis, whereas negative pairs stretch it along the T^- axis (T^+/T^- is the positive/negative text embedding.). To quantitatively describe this geometric structure and thereby calculate the anomaly shift, we propose the local Gaussian distribution and constant metric assumptions. Under these two assumptions, the distance in the embedding space can be computed using the Mahalanobis distance. Furthermore, we discover that the local embedding geometry differs among different categories (Figure 1(a)), and we generalize the two assumptions above to all classes, which is proved valid in extensive experiments. We find that the directional information in the CLIP embedding space is anisotropic; specifically, it serves to suppress normal noise while amplifying subtle anomaly variations. However, cosine similarity treats all directions equally (i.e., vectors with the same magnitude but different orientations always correspond to the same arc length). We also further explore the micro-level mechanisms of Mahalanobis distance for capturing anomalies. By projecting the embedding shifts of anomalous images onto the eigendirections of the metric, we found that the vast majority of anomaly shifts fall along the directions of high normal variance (Figure 1(c)). Consequently, due to the isotropy of cosine similarity, subtle anomaly signals are easily overwhelmed by massive normal noise. In contrast, the Mahalanobis distance, successfully suppresses normal noise while amplifying subtle anomaly shifts. Moreover, we demonstrate that due to the uncertainty of information loss during the normalization projection, the anomaly detection performance of cosine similarity is highly unstable across different visual encoders (which yield embedding spaces with varying structures). In contrast, the Mahalanobis distance exhibits strong robustness across diverse embedding spaces (details in Table 4).

In summary, our contributions are as follows:

- We reveal that applying CLIP to anomaly detection does not necessarily have to follow its original image-text pair matching paradigm, and we propose the use of CLIP's inherent embedding features to define normality/abnormality.
- Guided by geodesic computation in the embedding space, we reduce the geodesic distance to the Mahalanobis distance, and reveal that anomaly information is mainly encoded in the anisotropic directions of the local metric.
- By leveraging the two assumptions, we effectively model the geometric structure of the embedding space. We provide a comprehensive explanation from both macroscopic and microscopic perspectives, justifying the correctness of our theory.



(a) (b) (c)

Figure 1: (a) illustrates an example of the embedding distribution for a category. (b) shows that L2 normalization projects the multi-dimensional semantic information within the anisotropic embedding space onto an isotropic normalized hypersphere. (c) illustrates the Mahalanobis distance in different directions, and most anomalous deviations fall along the high-variance directions.

2 PRELIMINARY

2.1 CLIP-BASED AD VIA COSINE DISTANCE

CLIP consists of two encoders: a vision encoder $f_{\text{image}}()$ that maps an image I to a raw embedding $\mathbf{z}^{\text{img}} \in \mathbb{R}^m$ and a text encoder that maps a text sequence T to a raw embedding $\mathbf{z}^{\text{txt}} \in \mathbb{R}^m$. For a standard CLIP-based AD, the general pipeline should be: First, let $f_{\text{image}}()$ be a frozen CLIP vision encoder and $f_{\text{text}}()$ the text encoder. Then, for a test image x, we can obtain its corresponding normalized image embedding, the normal/anomaly text embeddings, and the image-text matching anomaly score:

$$\hat{\mathbf{z}} = \frac{f_{\text{image}}(x)}{\|f_{\text{image}}(x)\|_2}, \quad \hat{\mathbf{t}}_n = \frac{f_{\text{text}}(p_n)}{\|f_{\text{text}}(p_n)\|_2}, \quad \hat{\mathbf{t}}_a = \frac{f_{\text{text}}(p_a)}{\|f_{\text{text}}(p_a)\|_2}.$$

$$s(x) = \frac{\exp(\hat{\mathbf{z}} \cdot \hat{\mathbf{t}}_a)}{\exp(\hat{\mathbf{z}} \cdot \hat{\mathbf{t}}_n) + \exp(\hat{\mathbf{z}} \cdot \hat{\mathbf{t}}_a)}. \tag{1}$$

where p_n is a normal-class prompt (e.g., 'a photo of a {class}') and p_a is an anomaly prompt (e.g., 'a photo of a defective {class}').

2.2 MAHALANOBIS DISTANCE AND ANISOTROPY

Starting from the distance between two points in a general space, we simplify the distance formula to the Mahalanobis distance through two assumptions. For a distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and covariance $\Sigma \in \mathbb{R}^{m \times m}$, the squared Mahalanobis distance of a point $\mathbf{z}$ can describe the deviation relative to μ. We let $\mathbf{G} = \Sigma^{-1}$ denote the metric matrix, and Eigendecompose $\mathbf{G} = \sum_k \gamma_k \mathbf{u}_k \mathbf{u}_k^\top$ with $\gamma_1 \geq \cdots \geq \gamma_m > 0$ to analyze the distribution of this deviation along the eigenvectors.

$$d_M^2(\mathbf{z}, \boldsymbol{\mu}) = (\mathbf{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}). \tag{2}$$

$$d_M^2 = \sum_k \gamma_k \underbrace{(\mathbf{u}_k^\top (\mathbf{z} - \boldsymbol{\mu}))^2}_{c_k^2}. \tag{3}$$

- **Large** γ_k: the normal distribution has small variance along $\mathbf{u}_k$ — a **tight direction** where deviations are amplified.
- **Small** γ_k: the normal distribution has large variance along $\mathbf{u}_k$ — a **loose direction** where deviations are suppressed.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

The entire dataset is denoted as $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^n\}$, where n means the number of classes in the $\mathcal{D}$. Define the image space as $\mathcal{X}$ and the text space as $\mathcal{T}$. Then we can show that $f_{\text{image}} : \mathcal{X} \to \mathbb{R}^m$ and $f_{\text{text}} : \mathcal{T} \to \mathbb{R}^m$. where m is the dimension of the embedding space, which varies with different image encoders in CLIP. For a class c, let $\mathcal{D}_{\text{train}}^c = \{x_1, \ldots, x_{N_c}\} \subset \mathcal{X}$ be N_c images known to be normal. Given a test image $x \in \mathcal{X}$, assign an anomaly score $s(x) \in \mathbb{R}$ such that normal and anomalous instances are separated.

3.2 TWO ASSUMPTIONS

Gaussian Assumption: We assume that the normal embeddings of class c follow a multivariate Gaussian distribution

$$\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c).$$

where $\boldsymbol{\mu}_c$ is the mean embedding and Σ_c is the covariance matrix. This is not a claim that CLIP's outputs truly possess Gaussianity. This is a modeling assumption, based on two well-founded reasons:

1. **Maximum entropy principle**: Given the two statistics of mean and covariance, the maximum entropy distribution is Gaussian. We do not utilize higher-order statistics, so Gaussian is the most conservative assumption.

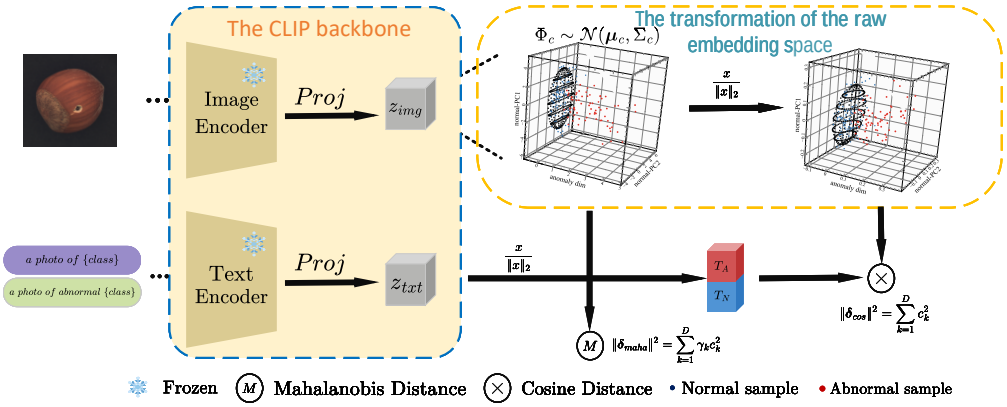


Figure 2: Overview of the proposed pipeline. It illustrates the comparison between CLIP's image-text matching and anomaly information extracted fromthe embedding space via Mahalanobis distance. Through contrastive learning, CLIP maps hazelnut samples into a local metric field, where anomaly features exhibit subtle systematic shifts relative to normal features. The "normal-PC1/2" represents the top two principal component directions of normal noise, while "anomaly dim" denotes the direction from the normal class center to the anomaly class center, as it maximizes the separation between normal and anomaly features.

2. **Predictive power test**: If this assumption were merely a mathematical convenience without correspondence to reality, the predictions derived from it would be rejected by experiments. However, in practice, these predictions are largely confirmed by our experimental results.

Under the Gaussian assumption, Σ_c is a statistical object – the covariance matrix. However, the same Σ_c can be reinterpreted from a geometric perspective (more details in Appendix B). In Riemannian geometry, a metric tensor $\mathbf{G}(x)$ – a positive definite symmetric matrix, is defined on the tangent space at every point x on the manifold. This metric tensor defines the local geometry of the space, providing us with a local ruler for measuring distances. For instance, the squared length of a tangent vector $\mathbf{v}$ is $\|\mathbf{v}\|^2_{\mathbf{G}(x)} = \mathbf{v}^\top \mathbf{G}(x)\mathbf{v}$. For any two points in the space, the distance between them is defined as the infimum of the arc lengths over all smooth curves connecting them. Mathematically, this is formalized as the infimum of the arc length of a parameterized curve $\gamma : [0, 1] \to \mathbb{R}^m$:

$$L[\gamma] = \int_0^1 \sqrt{\dot{\gamma}(t)^\top \mathbf{G}(\gamma(t))\, \dot{\gamma}(t)}\, dt.$$
$$d(x, y) = \inf_{\gamma:\gamma(0)=x,\gamma(1)=y} L[\gamma].$$

In general, $\mathbf{G}(x)$ varies from point to point, and consequently the geodesics connecting two points are curved. This makes distance computation complex. Therefore, we introduce a well-motivated assumption that significantly simplifies the distance calculation (bypassing the need to solve Euler-Lagrange equations) while preserving the essential role of the metric tensor in this process (more details in Appendix A).

Constant Metric Assumption: We assume that the metric tensor is constant in the normal region of class c:

$$\mathbf{G}(x) \equiv \mathbf{G}_c = \Sigma_c^{-1}, \quad \forall x \text{ near the normal region of class } c. \tag{4}$$

This assumption implies that the embedding space is endowed with a class-conditional local metric field, which means the "ruler" we use to measure distances should have different scales in regions corresponding to different categories. Consequently, when measuring distances, we need access to the local geometric information of the embedding space in this regard. Our assumption is demonstrated to be reasonable in practice by multiple experiments. Then, we use the squared d_{Maha} as the anomaly score of a test image x (more details in Appendix A):

$$s(x) = d^2(f_{\text{image}}(x), \boldsymbol{\mu}_c) = (f_{\text{image}}(x) - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1} (f_{\text{image}}(x) - \boldsymbol{\mu}_c). \tag{5}$$

where $\boldsymbol{\mu}_c$ is the mean embedding of the normal class and Σ_c is the covariance matrix of the normal embeddings. The anomaly score is the Mahalanobis distance between the test sample and the normal distribution in the embedding space. This formulation describes the deviation between the test image and normal features in the raw embedding space. Subsequent research has revealed that anomalous deviations in images systematically align with the eigen-directions of the metric tensor, which effectively suppresses normal variations while amplifying anomalous signals through adaptive weight assignment.

3.3 CLIP-BASED AD VIA MAHALANOBIS DISTANCE

Through our measurements of the geometric structures formed by different classes in the embedding space , we find that different classes do not share the same geometric structure. Specifically, each class c has a unique mean embedding $\boldsymbol{\mu}_c$ and a unique covariance matrix Σ_c in the embedding space. For example, we define $\kappa = \frac{\lambda_{\max}(\mathbf{G})}{\lambda_{\min}(\mathbf{G})}$ to quantify the degree of anisotropy. For ViT-L/14, the condition numbers κ_c of 15 classes in the MVTec-AD dataset range from 118 to 1658. Analogous situations are observed in other datasets and other image encoders. Therefore, for class c, given a training set of normal images $\mathcal{D}^c_{\text{train}} = \{x_1, \ldots, x_N\}$, we can compute the mean embedding and covariance matrix of $\mathcal{D}^c_{\text{train}}$ respectively:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N} \sum_{i=1}^{N} f_{\text{image}}(x_i) \in \mathbb{R}^m. \tag{6}$$

$$\hat{\Sigma}_c^{\text{raw}} = \frac{1}{N-1} \sum_{i=1}^{N} \big(f_{\text{image}}(x_i) - \hat{\boldsymbol{\mu}}_c\big)\big(f_{\text{image}}(x_i) - \hat{\boldsymbol{\mu}}_c\big)^\top \in \mathbb{R}^{m \times m}. \tag{7}$$

$$\hat{\Sigma}_c = \hat{\Sigma}_c^{\text{raw}} + \lambda \mathbf{I}_m. \tag{8}$$

When $N < m$ or N is only slightly larger than m, $\hat{\Sigma}_c^{\text{raw}}$ is close to singular. However, on the vast majority of datasets, $\hat{\Sigma}_c^{\text{raw}}$ is singular. Therefore, we employ Tikhonov regularization to ensure the invertibility of $\hat{\Sigma}_c^{\text{raw}}$ (Equation 8). The hyperparameter λ determines the amount of isotropic information injected into $\hat{\Sigma}_c^{\text{raw}}$. Consequently, a large λ would cause the anisotropic information to be overwhelmed by the injected isotropic information. We set the default value of λ to 0.01, at which point the final detection performance approaches saturation as λ decreases.

Under our Gaussian and constant metric assumptions, the $\hat{\boldsymbol{\mu}}_c$ and $\hat{\Sigma}_c$ of $\mathcal{D}^c_{\text{train}}$ can be used as the $\boldsymbol{\mu}_c$ and Σ_c of the normal embedding distribution. Then, we can conclude that $\boldsymbol{\mu}_c = \hat{\boldsymbol{\mu}}_c$ and $\Sigma_c = \hat{\Sigma}_c$ for class c. And the anomaly score of a test image x is given by Equation 5.

3.4 DISPLACEMENT DECOMPOSITION

In this section, we explore the microscopic mechanism of this constraint. For a class c, we can get the mean of normal embeddings $\boldsymbol{\mu}_c$ from Equation 6. For an anomalous image x, it is mapped into an embedding vector $\tilde{\mathbf{z}} = f_{\text{image}}(x)$. Although this vector still falls within the local metric field of the corresponding class, it exhibits an anomalous displacement relative to the normal distribution, and the displacement is:

$$\boldsymbol{\delta} = \tilde{\mathbf{z}} - \boldsymbol{\mu}_c. \tag{9}$$

Meanwhile, we perform an eigendecomposition on the metric tensor $\mathbf{G}_c = \sum_{k=1}^{m} \gamma_k \mathbf{u}_k \mathbf{u}_k^\top$ and project $\boldsymbol{\delta}$ onto the eigenbasis. We aim to obtain the specific characteristics of the anomalous displacement of the foreground object in the image along the eigen-directions. We hypothesize that the anomalous directions of different foreground objects tend to fall on tight directions, enabling the Mahalanobis distance to assign larger weights γ_k in these directions. In contrast, the inherent isotropy of the cosine similarity ($\|\boldsymbol{\delta}\|^2 = \sum_{k=1}^{m} c_k^2$) prevents these anomalous directions from standing out.

$$\boldsymbol{\delta} = \sum_{k=1}^{m} c_k \mathbf{u}_k, \qquad c_k = \mathbf{u}_k^\top \boldsymbol{\delta}. \tag{10}$$

$$d_M^2 = \boldsymbol{\delta}^\top \mathbf{G}_c \boldsymbol{\delta} = \sum_{k=1}^{m} \gamma_k \, c_k^2. \tag{11}$$

To explore the specific distribution of anomalous displacements, we define the stretch ratio $r(\boldsymbol{\delta})$. It represents the normalized ratio of the Mahalanobis distance between the anomalous displacement direction and a random direction in the embedding space, ensuring that the $r(\boldsymbol{\delta})$ depends solely on the direction, independent of the vector's magnitude (details in Appendix C):

$$r(\boldsymbol{\delta}) = \frac{\sum_{k=1}^{m} \gamma_k c_k^2}{\|\boldsymbol{\delta}\|^2 \cdot \frac{1}{m} \sum_{k=1}^{m} \gamma_k} = \frac{d_M^2 / \|\boldsymbol{\delta}\|^2}{\frac{1}{m} \operatorname{tr}(\mathbf{G}_c)}. \tag{12}$$

Interpretation.

- $r(\boldsymbol{\delta}) > 1$: the displacement energy is concentrated in tight directions (small variance).
- $r(\boldsymbol{\delta}) = 1$: the displacement energy is distributed uniformly (isotropic).
- $r(\boldsymbol{\delta}) < 1$: the displacement energy is concentrated in loose directions (large variance).

We find that different visual encoders consistently project the anomalous displacement $\boldsymbol{\delta}$ onto loose directions across all 15 categories in the MVTec dataset. This implies that for anomalous foreground objects, the distribution in the embedding space is concentrated in directions with fluctuation noise. Cosine similarity treats the noise weights in all directions equally, causing the noise to overwhelm the effective information. In contrast, the Mahalanobis distance can suppress this noise through adaptive weight allocation, thereby effectively distinguishing anomalies.

4 Experiments

4.1 Experiment Setup

Datasets and Evaluation Metrics We conducted extensive experiments on 11 open-source datasets, covering images from industrial to medical domains. Specifically, for the industrial domain, we use MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), MPDD (Jezek et al., 2021), BTAD (Mishra et al., 2021), DAGM (Wieler et al., 2007), DTD-Synthetic (Aota et al., 2023), KSDD (Tabernik et al., 2020) and KSDD2 (Božič et al., 2021). For the medical domain, we consider BrainMRI (Chakrabarty, 2019), OCT2017 (Kermany et al., 2018) and HeadCT (Bao et al., 2024), among which OCT2017 and HeadCT are obtained from BMAD (Bao et al., 2024). The competing method shares the same OpenCLIP backbone (ViT-L/14@336px) and preprocessing as ours, but computes anomaly scores via image-text cosine similarity. The anomaly detection performance is evaluated by Area Under the Receiver Operating Characteristic Curve (AUROC). Moreover, we compute average precision (AP) and Maximum F1-score (F1-max) for anomaly detection.

Implementation details. Following the mainstream models (Salehi et al., 2025; Zhou et al., 2024; Cao et al., 2024), we use OpenCLIP with the ViT-L/14@336px architecture as our backbone. The input images are resized to 336×336, following the standard CLIP preprocessing pipeline. All parameters of CLIP remain frozen during the experiments. For the number N_c of normal samples used to compute the mean μ_c and covariance Σ_c, we used 4, 8, 16, 32, 64 and 128. As expected, the anomaly detection performance of the model consistently improves as N_c increases. For the hyperparameter λ, it represents a small amount of isotropic information that we artificially inject to prevent Σ_c from becoming singular. However, if λ is too large, this injected information will overwhlem the original anisotropic information. We set the default λ to 0.01 (Table 1), where the resulting AUROC tends to plateau. All experiments are conducted in PyTorch 2.11.0 with a single NVIDIA GeForce RTX 5060 Laptop GPU.

Table 1: Effect of the λ on the Mahalanobis anomaly score.

λ	0.001	0.003	**0.01**	0.03	0.1	0.3	1.0	3.0	10.0
AUROC (%)	96.68	96.59	**96.38**	95.90	94.93	93.74	92.09	90.34	88.90

4.2 Main Results

Comparison with Cosine Similarity Anomaly Detection. As a controlled comparison, we set up two groups that calculate anomaly scores using cosine similarity and Mahalanobis distance respectively. Specifically, the normal/anomalous text prompts used in the cosine similarity group are "a photo of {class}/defective {class}". Since CLIP is pre-trained for image classification, we evaluate the image-level anomaly classification performance for both paradigms. Experimental results

show that the Mahalanobis distance greatly improves CLIP's anomaly classification capability, respectively achieving Image-AUROC improvements of 53.1% and 44.8% on the industrial dataset BTAD and the medical dataset OCT2017 (details in Table 2). On the vast majority of datasets, the Mahalanobis distance exceeds the performance of the model with image-text cosine similarity. However, the performance of the Mahalanobis distance is less satisfactory on some datasets, such as BrainMRI. We hypothesize that this is because the BrainMRI samples cover multiple brain regions and slices, resulting in the normal distribution splitting into multiple subclusters. This diversity ultimately invalidates the local constant metric and Gaussian assumptions.

Table 2: Comparison on anomaly detection benchmarks. Each cell reports Image-level AUROC, AP and F1-max. In the image-only baseline, the text embedding in the image-text cosine similarity is replaced by the class-mean image embedding $\boldsymbol{\mu}_c$, which is the same mean used in the Mahalanobis distance.

Task	Dataset	$\lvert C\rvert$	OpenCLIP (image-text)			OpenCLIP (image only)			Ours (L/14@336px)		
			AUROC	AP	F1-max	AUROC	AP	F1-max	AUROC	AP	F1-max
Industrial	MVTec	15	79.9	91.4	88.5	87.4	94.0	91.5	96.4	98.4	95.5
	VisA	12	70.6	78.8	76.5	78.0	78.2	79.7	88.9	89.7	84.9
	MPDD	6	60.4	70.2	76.4	70.4	72.5	78.9	88.4	89.5	87.3
	BTAD	3	41.7	56.4	63.6	91.5	91.3	86.9	94.8	98.3	94.4
	DAGM	10	69.7	45.4	47.5	95.2	84.2	80.2	97.4	87.8	85.4
	DTD-Synthetic	12	89.9	96.0	91.8	94.2	97.0	93.3	98.2	99.1	97.4
	KSDD	1	60.4	45.5	46.0	82.0	67.6	67.5	86.2	73.1	73.1
	KSDD2	1	75.3	57.3	53.3	86.4	70.7	63.5	93.6	84.2	76.4
Medical	BrainMRI	1	73.3	84.5	90.6	88.6	94.7	92.7	84.5	90.6	93.6
	OCT2017	1	51.0	78.3	85.7	89.9	96.4	90.7	95.8	98.6	93.7
	HeadCT	1	91.8	93.9	92.4	76.7	84.0	83.0	89.7	92.5	89.9

Experiments under different number N of normal samples. The results in Table 2 were obtained using all normal samples from the training set. To investigate the performance of the Mahalanobis distance under few-shots, we set the number of normal samples N to $\{4, 8, 16, 32, 64, 128\}$ and evaluated our method on the MVTec dataset. As shown in Figure 3, performance improves steadily as N increases, with the most significant gains occurring between $N = 4$ and $N = 32$. It precisely demonstrates that this metric is insensitive to estimation errors in the μ_c and Σ_c for small samples, serving as a universal method for calculating anomaly shifts based on the intrinsic structure of the CLIP embedding space.

Anomaly displacement decomposition. On the MVTec dataset, we projected the anomaly displacement onto the feature directions across 15 categories using different image encoders (detailed in Section 3.4) to observe the micro-characteristics of the anomaly displacement. Results (details in Table 4) show that across all image encoders, the stretch ratio of the anomaly displacement $r(\boldsymbol{\delta})$ consistently remains $r(\boldsymbol{\delta}) < 1$, indicating that the anomaly displacements collectively deviate towards the loose directions. The significant fluctuations in $\rho(r, s_{\text{cos}})$ indicate that cosine similarity is an unstable mechanism highly

Backbone	r_{mean}	$\sigma(r)$	ρ_{cos}	ρ_{maha}
ViT-B/16+	0.362	0.150	0.117	0.785
ViT-L/14	0.320	0.162	0.456	0.799
ViT-L/14@336px	0.322	0.128	0.069	0.796

Table 4: Stretch ratio of anomaly displacement on MVTec, where $\rho_{\text{cos}} = \rho(r, s_{\text{cos}})$ and $\rho_{\text{maha}} = \rho(r, s_{\text{maha}})$.

Table 3: Comparison with CLIP-based anomaly detection methods. Each cell reports Image-level (AUROC, AP, F1-max). Best results are bold; second-best results are underlined for each metric independently.

Task	Dataset	$\lvert C\rvert$	WinCLIP	AnomalyCLIP	AdaCLIP	Crane	Ours (L/14@336px)
Industrial (AUROC, AP, F1-max)	MVTec	15	(91.8, 96.5, 92.9)	(91.6, 96.4, 92.7)	(90.1, 95.7, 92.3)	(93.8, 97.5, 93.8)	(**96.4**, **98.4**, **95.5**)
	VisA	12	(78.1, 81.1, 80.7)	(82.0, 85.3, 80.4)	(75.7, 79.0, 78.5)	(85.3, 87.9, 82.6)	(**88.9**, **89.7**, **84.9**)
	MPDD	6	(63.6, 69.9, 77.5)	(77.5, 82.5, 80.4)	(72.1, 80.2, 77.6)	(81.4, 84.8, 81.6)	(**88.4**, **89.5**, **87.3**)
	BTAD	3	(68.2, 70.9, 67.6)	(88.2, 87.3, 83.8)	(88.7, 93.8, 88.2)	(94.4, 95.9, 91.7)	(**94.8**, **98.3**, **94.4**)
	DAGM	10	(91.8, 79.5, 87.6)	(97.7, 92.4, 90.1)	(96.9, 88.5, 87.7)	(**99.2**, **97.4**, **95.8**)	(97.4, 87.8, 85.4)
	DTD-Synthetic	12	(95.1, 97.7, 94.1)	(93.9, 97.2, 93.6)	(95.3, 97.3, 94.8)	(96.3, 98.5, 95.3)	(**98.2**, **99.1**, **97.4**)
	KSDD	1	(84.3, 77.4, 79.0)	(84.7, 80.0, 82.7)	(97.1, 89.6, 90.7)	(**97.8**, **94.3**, **91.6**)	(86.2, 73.1, 73.1)
	KSDD2	1	(93.5, **94.2**, **86.4**)	(94.1, 84.3, 80.0)	(**95.6**, 89.0, 84.1)	(-, -, -)	(93.6, 84.2, 76.4)
Medical (AUROC, AP, F1-max)	BrainMRI	1	(86.6, 91.5, 86.3)	(90.3, 92.2, 90.2)	(93.5, 95.6, 89.7)	(**95.4**, **96.1**, **93.9**)	(84.5, 90.6, 93.6)
	OCT2017	1	(55.2, 81.1, 82.7)	(69.0, 88.6, 85.6)	(80.8, 92.7, 86.5)	(-, -, -)	(**95.8**, **98.6**, **93.7**)
	HeadCT	1	(81.8, 80.2, 79.8)	(93.4, 91.6, 90.8)	(91.8, 90.6, 84.1)	(**95.3**, **95.7**, **91.1**)	(89.7, 92.5, 89.9)

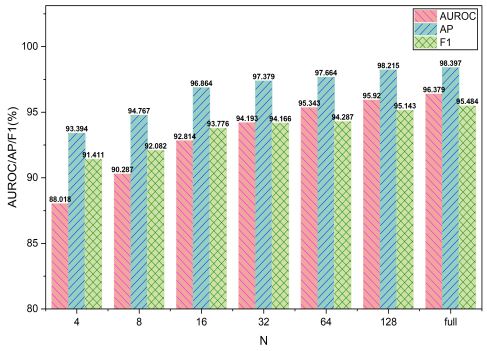


Figure 3: The anomaly detection performance under different number of normal samples N.

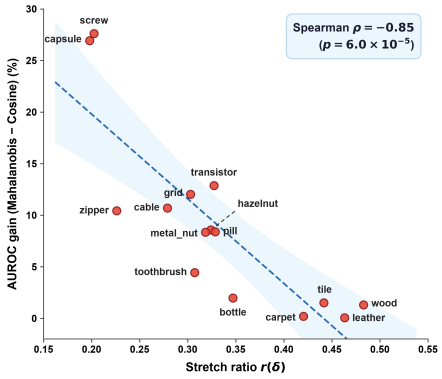


Figure 4: Scatter plot of $r(\boldsymbol{\delta})$ versus the AUROC gain on the MVTec dataset.

dependent on the specific embedding space. Conversely, the stability of $\rho(r, s_{\text{maha}})$ across three different image encoders demonstrates that the Mahalanobis distance is a universal mechanism independent of the embedding space.

We also calculated the Spearman correlation between the AUROC gain ($\text{AUROC}_{\text{Maha}} - \text{AUROC}_{\cos}$) and the stretch ratio $r(\boldsymbol{\delta})$ on the MVTec dataset, as illustrated in Figure 4. The results reveal a strong negative correlation (Spearman $\rho = -0.85$, p=0.00006), which corroborates our earlier assertion that Mahalanobis distance achieves superior anomaly detection by assigning weights to mitigate substantial noise fluctuations

4.3 Ablation Studies

To decompose the Mahalanobis distance, we write the anomaly displacement as $\boldsymbol{\delta} = f_{\text{image}}(x) - \boldsymbol{\mu}_c$ and split it along the unit mean direction $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_c / \|\boldsymbol{\mu}_c\|$ into a radial component $\boldsymbol{\delta}_{\parallel} = (\boldsymbol{\delta}^\top \hat{\boldsymbol{\mu}})\,\hat{\boldsymbol{\mu}}$ and a tangential component $\boldsymbol{\delta}_{\perp} = \boldsymbol{\delta} - \boldsymbol{\delta}_{\parallel} = (\mathbf{I} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top)x$. For the anisotropic terms, Σ is the class covariance and $\Sigma_{\perp} = \mathbf{P}_{\perp}\,\Sigma\,\mathbf{P}_{\perp}$ ($\mathbf{P}_{\perp} = \mathbf{I} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top$) its tangential projection, where $^+$ denotes the pseudoinverse. These quantities define four scoring functions that toggle the radial displacement and the precision matrix (Table 5).

Table 5: Scoring functions in the isolation experiment. "Radial" toggles the radial displacement $\boldsymbol{\delta}_{\parallel}$, and "Aniso." toggles the class-conditional precision matrix Σ^{-1}.

Score	Definition	Radial $\boldsymbol{\delta}_{\parallel}$	Aniso. Σ^{-1}
s_1	$\|\boldsymbol{\delta}_{\perp}\|^2$	×	×
s_2	$\|\boldsymbol{\delta}\|^2$	✓	×
s_3	$\boldsymbol{\delta}_{\perp}^\top \Sigma_{\perp}^{+} \boldsymbol{\delta}_{\perp}$	×	✓
s_4	$\boldsymbol{\delta}^\top \Sigma^{-1} \boldsymbol{\delta}$	✓	✓

Table 6: Ablation experiments of the anomaly information distribution on MVTec and VisA datasets. Each cell reports the mean ± std over classes (%). The abnormal displacement is decomposed into radial $\delta_{\parallel}$ and tangential displacements $\delta_{\perp}$.

Metric	Radial	MVTec		VisA	
		Aniso. ×	Aniso. ✓	Aniso. ×	Aniso. ✓
AUROC	×	87.49±12.17	96.35±4.24	78.23±11.51	88.87±8.20
	✓	87.45±12.21	96.38±4.23	78.11±11.55	88.88±8.22
AP	×	94.03±6.42	98.39±2.10	78.55±14.59	89.70±8.80
	✓	94.01±6.43	98.40±2.10	78.39±14.62	89.70±8.79
F1-max	×	91.55±6.87	95.52±3.84	79.88±9.06	84.89±8.12
	✓	91.51±6.90	95.48±3.89	79.90±9.13	84.87±8.12

To identify which geometric factor carries the anomaly information, we isolate the radial and the anisotropic (directional) components of the Mahalanobis distance in a 2×2 ablation (Table 6). Start-

ing from the L2-normalized cosine distance to the class mean ($s_{\cos}$), enabling the radial component (the norm $\|\mathbf{x}\|$ and the radial displacement $\delta_{\|}$) changes the AUROC by only -0.04 and -0.12 points on MVTec and VisA (both non-significant), whereas enabling the class-conditional anisotropic metric Σ^{-1} yields a large and significant gain of $+8.86$ and $+10.64$ points. The same pattern holds for AP and F1-max: anisotropy contributes $+4.36$/$+3.97$ points on MVTec and $+11.15$/$+5.01$ on VisA, whereas the radial component shifts every metric by less than 0.2 points (all non-significant). This indicates that the anomaly information discarded by cosine similarity resides predominantly in the class-dependent directional (anisotropic) structure, with the radial component contributing only marginally.

5 Related Work

Contrastive Learning. In contrastive learning, models are trained to pull positive pairs closer while pushing negative pairs further apart in the embedding space. This concise paradigm has led to substantial advancements in multiple fields (Chen et al., 2020; Gao et al., 2022). The latent embedding space of contrastive learning is often modeled as a normalized hypersphere (Wang & Isola, 2022). However, the rise of cross-modal models like CLIP (Radford et al., 2021), which align images and text in a shared embedding space, has revealed novel embedding structures, such as the modality gap (Liang et al., 2022) and the narrow cone effect (Schrodi et al., 2025; Liang et al., 2022). Furthermore, research has found that in CLIP's unnormalized embedding space, text and images are actually located on two linearly separable ellipsoidal shells, instead of being distributed on an origin-centered sphere (Levi & Gilboa, 2025).

CLIP-based Zero-shot Anomaly Detection (ZSAD)&Few-shot Anomaly Detection (FSAD). Due to its strong generalization capability, CLIP is frequently used as a pre-trained model for various ZSAD methods. Existing ZSAD methods primarily focus on prompt learning, such as the compositional prompt ensemble in WinCLIP (Jeong et al., 2023), the object-agnostic anomaly prompts in AnomalyCLIP (Zhou et al., 2024), and the static/dynamic prompts in AdaCLIP (Cao et al., 2024). Some FSAD methods exploit a few normal samples to build a normal memory bank, such as VAND (Chen et al., 2023), WinCLIP+ (Jeong et al., 2023), CLIP-Core (Xu et al., 2026).

Recent work utilizing Mahalanobis distance. The Mahalanobis distance has now been applied across a broad range of research. For instance, the core of DCA (Dong et al., 2026) is diversity covariance-aware learning, which efficiently captures the features of different categories, with the Mahalanobis distance serving as the metric within this method. Mahalanobis++ (Mueller & Hein, 2025) proposes normalizing the features before computing the Mahalanobis distance, thereby reducing noise and yielding better OOD detection. Whitened CLIP (Betser et al., 2025) applies an invertible linear transformation to CLIP's latent space so that it can be used directly to compute the likelihood of images and captions. In the field of anomaly detection, CLIP-Core/CLIP-Core+ (Xu et al., 2026) extracts multi-scale patch features from a few samples to construct a normal memory bank and then uses the Mahalanobis distance for anomaly classification and localization.

6 Conclusion

In this paper, we revisit the consensus that CLIP is anomaly-agnostic, and demonstrate that the metric errors induced by the image-text cosine similarity are responsible for this anomaly-unawareness. We discover that CLIP projects images of each category into an embedding space with local features, which can then be leveraged as the foundation for our anomaly detection. We verify that the Gaussian and the constant metric assumptions can accurately describe these local embedding features. Furthermore, without any training, CLIP alone is capable of achieving excellent anomaly classification. We propose using the Mahalanobis distance to measure anomaly shifts in the embedding space, and investigate the underlying microscopic mechanisms explaining why the Mahalanobis distance succeeds in anomaly classification while cosine similarity fails to do so effectively. However, we only demonstrated CLIP's powerful anomaly recognition capability in image classification, without exploring its potential for anomaly segmentation. Further research in this direction is needed in the future. Furthermore, the proposed Gaussian and constant metric assumptions exhibit suboptimal performance on certain datasets. This indicates that a deeper understanding of the distribution of specific categories within the embedding space is still needed in future research.

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation. We have not used generative AI tools for formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, interpreting results, and generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting datasets, supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, summarizing or analysing existing literature, sourcing/searching for information, editing a research paper to improve readability. We have reviewed all AI-assisted work. For example, we tested the LLM-generated code to ensure it aligns with our research methodology and experimental needs. We manually reviewed the original literature to verify the LLM-generated summaries and translations. We verified the LLM-generated figures and tables to ensure data integrity, and the LLM-based editing was limited to the improvement of visual presentation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

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)*, pp. 5564–5572, 2023.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection, 2024.

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)*, 2019.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. In *International Conference on Machine Learning (ICML)*, 2025.

Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

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*, pp. 55–2. Springer Nature Switzerland, September 2024. ISBN 9783031727610.

Navoneel Chakrabarty. Brain MRI images for brain tumor detection, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge, 2023.

Songlin Dong, Zhengdong Zhou, Chenhao Ding, Xinyuan Gao, Alex Kot, and Yihong Gong. Diversity covariance-aware prompt learning for vision-language models. *Pattern Recognition*, 2026. arXiv preprint arXiv:2503.01531.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation, 2023.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 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 (ICUMT)*, pp. 66–71, 2021.

Daniel S. Kermany, Kang Zhang, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip, 2025.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022.

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)*. IEEE, 2021.

Maximilian Mueller and Matthias Hein. Mahalanobis++: Improving ood detection via feature normalization. In *International Conference on Machine Learning (ICML)*, 2025.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Alireza Salehi, Mohammadreza Salehi, Reshad Hosseini, Cees G. M. Snoek, Makoto Yamada, and Mohammad Sabokrou. Crane: Context-guided prompt learning and attention refinement for zero-shot anomaly detection, 2025.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

Matthias Wieler, Tobias Hahn, and Fred A. Hamprecht. Weakly supervised learning for industrial optical inspection. In *DAGM Symposium*, 2007.

Liang Xu, Junle Rao, and Shuyou Lin. Clip-core: A few-sample anomaly detection method for surface defects. *IEEE Access*, 14:20716–20733, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation, 2022.

APPENDIX

A FROM RIEMANNIAN GEODESIC TO MAHALANOBIS DISTANCE

A.1 FORMULATION

Let $\mathcal{M} = \mathbb{R}^m$ be a Riemannian manifold equipped with a metric tensor $\mathbf{G}(x) \in \mathbb{R}^{m\times m}$, symmetric positive-definite at every point $x \in \mathcal{M}$. The squared length of a tangent vector $\mathbf{v} \in T_x\mathcal{M}$ is:

$$\|\mathbf{v}\|^2_{\mathbf{G}(x)} = \mathbf{v}^\top \mathbf{G}(x)\mathbf{v}.$$

For a smooth curve $\gamma : [0, 1] \to \mathbb{R}^m$ parameterized by t, the arc length functional is:

$$L[\gamma] = \int_0^1 \sqrt{\dot\gamma(t)^\top \mathbf{G}(\gamma(t))\, \dot\gamma(t)}\, dt. \tag{A.1}$$

where $\dot\gamma(t) = \frac{d\gamma}{dt}$. The distance between two points $x, y \in \mathcal{M}$ is the infimum of $L[\gamma]$ over all smooth curves connecting them:

$$d(x, y) = \inf_{\gamma:\gamma(0)=x,\, \gamma(1)=y} L[\gamma]. \tag{A.2}$$

After solving the Euler-Lagrange equations, the full geodesic equation is given by:

$$\mathbf{G}(\gamma)\ddot\gamma + \left(\frac{d}{dt}\mathbf{G}(\gamma)\right)\dot\gamma - \frac{1}{2}\dot\gamma^\top \frac{\partial \mathbf{G}}{\partial \gamma}\dot\gamma = 0. \tag{A.3}$$

A.2 CONSTANT METRIC ASSUMPTION

Now, we impose the constant metric assumption:

$$\mathbf{G}(x) \equiv \mathbf{G}_c = \Sigma_c^{-1}, \quad \forall x \text{ in the normal region of class } c. \tag{A.4}$$

Under this assumption, $\mathbf{G}$ does not depend on $\gamma(t)$. Consequently:

$$\frac{d}{dt}\mathbf{G} = 0, \qquad \frac{\partial \mathbf{G}}{\partial \gamma} = 0.$$

Equation A.3 therefore collapses to

$$\mathbf{G}_c\, \ddot\gamma(t) = 0. \tag{A.5}$$

Since $\mathbf{G}_c$ is positive-definite (hence invertible), this implies:

$$\ddot\gamma(t) = 0. \tag{A.6}$$

The geodesic is a straight line. Integrating twice:

$$\gamma(t) = x + t(y - x). \tag{A.7}$$

which clearly satisfies the boundary conditions $\gamma(0) = x$ and $\gamma(1) = y$. Substituting the straight-line geodesic in Equation A.7 into the arc-length functional in Equation A.1 gives

$$\dot\gamma(t) = y - x = \text{constant}$$

and

$$L[\gamma] = \int_0^1 \sqrt{(y - x)^\top \mathbf{G}_c(y - x)}\, dt.$$

The integrand is independent of t, so the integral evaluates trivially:

$$L[\gamma] = \sqrt{(y - x)^\top \mathbf{G}_c(y - x)} \int_0^1 dt = \sqrt{(y - x)^\top \mathbf{G}_c(y - x)}.$$

Since a straight line is the unique global minimizer in a Euclidean (constant-metric) space, the infimum in Equation A.2 is attained:

$$d(x, y) = \sqrt{(x - y)^\top \Sigma_c^{-1}(x - y)}. \tag{A.8}$$

This is the Mahalanobis distance.

B Statistical-Geometric Equivalence

B.1 Gaussian Log-Likelihood

Assume normal-class embeddings follow a multivariate Gaussian:

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c).$$

The probability density function is:

$$p(\mathbf{z} \mid c) = \frac{1}{(2\pi)^{m/2}|\Sigma_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c)\right). \tag{B.1}$$

Taking the negative log-likelihood (up to additive constants):

$$-\log p(\mathbf{z} \mid c) = \frac{m}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_c| + \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c).$$

For a fixed class c, the first two terms are constants. The only term that depends on $\mathbf{z}$ is:

$$-\log p(\mathbf{z} \mid c) \propto \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c) = \frac{1}{2}d_M^2(\mathbf{z}, \boldsymbol{\mu}_c). \tag{B.2}$$

The Mahalanobis distance is the negative log-likelihood of a Gaussian, up to scaling. Points with equal Mahalanobis distance from $\boldsymbol{\mu}_c$ lie on equal-density contours of the fitted Gaussian.

B.2 Dual Interpretation

The covariance matrix Σ_c (and its inverse $\mathbf{G}_c = \Sigma_c^{-1}$) possesses a dual interpretation that bridges the statistical and geometric perspectives:

Perspective	Σ_c	$\mathbf{G}_c = \Sigma_c^{-1}$
Statistical	Covariance of the normal distribution	Precision matrix; appears in the exponent of the Gaussian density
Geometric	Defines the shape of the ellipsoid (variance ellipsoid)	Defines the Riemannian metric on the tangent space

The two views are not competing—they are equivalent descriptions of the same object.

Statistical → Geometric. From the Gaussian density $p(\mathbf{z} \mid c)$, the set $\{\mathbf{z} : d_M(\mathbf{z}, \boldsymbol{\mu}_c) = r\}$ is both an equal-probability contour of $\mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)$ and an isometric surface under the metric $\mathbf{G}_c$.

Geometric → Statistical. Given the metric $\mathbf{G}_c$ and the mean $\boldsymbol{\mu}_c$, the maximum-entropy distribution consistent with these two moments is exactly $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{G}_c^{-1})$.

Formally, the equivalence is:

$$d_M^2(\mathbf{z}, \boldsymbol{\mu}_c) = (\mathbf{z}-\boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\mathbf{z}-\boldsymbol{\mu}_c) = -2\log p(\mathbf{z} \mid c) + \text{const}.$$

This equivalence is what allows us to estimate the metric purely from normal-class data (via the sample covariance) without training any parameters — the normal samples' scatter is the geometry.

C Eigendecomposition and Directional Weighting

C.1 Decomposition of Mahalanobis and Euclidean Distances

Let $\mathbf{G}_c = \sum_{k=1}^m \gamma_k \mathbf{u}_k \mathbf{u}_k^\top$ be the eigendecomposition, with eigenvalues $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m > 0$ and orthonormal eigenvectors $\{\mathbf{u}_k\}$. Any displacement $\boldsymbol{\delta} = \mathbf{z} - \boldsymbol{\mu}_c$ can be expressed in this basis:

$$\boldsymbol{\delta} = \sum_{k=1}^m c_k \mathbf{u}_k, \qquad c_k = \mathbf{u}_k^\top \boldsymbol{\delta}.$$

The squared Mahalanobis distance decomposes as:

$$d_M^2 = \boldsymbol{\delta}^\top \mathbf{G}_c \boldsymbol{\delta} = \sum_{k=1}^m \gamma_k \, c_k^2. \tag{C.1}$$

In contrast, the squared Euclidean distance (and, up to a factor, the cosine distance after L2 normalization) is:

$$d_E^2 = \|\boldsymbol{\delta}\|^2 = \sum_{k=1}^{m} c_k^2. \tag{C.2}$$

The ratio of contributions from direction k differs by a factor of γ_k between the two metrics.

L2 normalization as metric projection. L2 normalization defines a map $\pi : \mathbb{R}^m \to S^{m-1}$:

$$\pi(\mathbf{z}) = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}.$$

Its Jacobian at $\mathbf{z} = \boldsymbol{\mu}$ is:

$$D\pi(\boldsymbol{\mu}) = \frac{1}{\|\boldsymbol{\mu}\|}\left(\mathbf{I}_m - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right). \tag{C.3}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. The matrix $\mathbf{I}_m - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top$ is a projection onto the tangent space $T_{\hat{\boldsymbol{\mu}}}S^{m-1}$—it removes the component parallel to $\boldsymbol{\mu}$.

For a small displacement along eigenvector $\mathbf{u}_k$:

$$\pi(\boldsymbol{\mu} + \boldsymbol{\delta}) \approx \hat{\boldsymbol{\mu}} + D\pi(\boldsymbol{\mu})\boldsymbol{\delta}.$$

The squared cosine distance between $\pi(\boldsymbol{\mu} + \boldsymbol{\delta})$ and $\hat{\boldsymbol{\mu}}$ is:

$$d_{\cos}^2 \approx \frac{1}{\|\boldsymbol{\mu}\|^2}\sum_k c_k^2. \tag{C.4}$$

where $c_k = \mathbf{u}_k^\top \boldsymbol{\delta}$ as before. All directions receive equal weight $1/\|\boldsymbol{\mu}\|^2$, independent of γ_k.

C.2 Stretch Ratio

The squared Mahalanobis distance weights each directional component by its precision:

$$d_M^2 = \boldsymbol{\delta}^\top \mathbf{G}_c \boldsymbol{\delta} = \sum_{k=1}^{m} \gamma_k\, c_k^2.$$

The Euclidean norm (and, up to scaling, the cosine distance after L2 normalization) treats all directions equally:

$$\|\boldsymbol{\delta}\|^2 = \sum_{k=1}^{m} c_k^2.$$

The key difference between two distances is:

$$\frac{d_M^2}{\|\boldsymbol{\delta}\|^2} = \frac{\sum_k \gamma_k c_k^2}{\sum_k c_k^2} \triangleq \bar{\gamma}_{\text{eff}}(\boldsymbol{\delta}). \tag{C.5}$$

This ratio is the effective average precision of the displacement $\boldsymbol{\delta}$—a weighted average of γ_k with weights c_k^2. It encodes the "directional alignment" of a single displacement.

However, this ratio is dimensionful: its absolute value depends on the overall scale of $\mathbf{G}$, as does $\mathrm{tr}(\mathbf{G})$.

Normalization baseline: expectation over random directions. Consider a completely random unit direction $\mathbf{v} \sim \mathrm{Uniform}(S^{m-1})$. By isotropy: $\mathbb{E}[v_k^2] = 1/m$ for all directional components. Therefore, its expected effective precision is:

$$\begin{aligned}\mathbb{E}_{\mathbf{v}}\left[\mathbf{v}^\top \mathbf{G}\mathbf{v}\right] &= \mathbb{E}\left[\sum_k \gamma_k v_k^2\right] = \sum_k \gamma_k \frac{1}{m} \\ &= \frac{1}{m}\sum_{k=1}^{m}\gamma_k = \frac{1}{m}\,\mathrm{tr}(\mathbf{G}) \triangleq \bar{\gamma}.\end{aligned} \tag{C.6}$$

A random direction does not favor any eigenvector—it samples all directions equally, and the resulting average precision is simply the arithmetic mean $\bar{\gamma}$ of all γ_k.

Definition. The stretch ratio is defined as

$$\boxed{\begin{aligned} r(\boldsymbol{\delta}) &= \frac{\bar{\gamma}_{\text{eff}}(\boldsymbol{\delta})}{\bar{\gamma}} = \frac{d_M^2/\|\boldsymbol{\delta}\|^2}{\frac{1}{m}\operatorname{tr}(\mathbf{G})} \\ &= \frac{\sum_{k=1}^{m} \gamma_k c_k^2}{\|\boldsymbol{\delta}\|^2 \cdot \frac{1}{m}\sum_{k=1}^{m} \gamma_k}. \end{aligned}} \tag{C.7}$$

This completes the definition of $r(\boldsymbol{\delta})$. The numerator is the true average precision of the displacement, and the denominator is the random expectation. The ratio is dimensionless, allowing for direct cross-class comparison.