

G²SF: Geometry-Guided Score Fusion for Multimodal Industrial Anomaly Detection

Chengyu Tao¹ Xuanming Cao² Juan Du^{1,2*}

1.The Hong Kong University of Science and Technology 2.The Hong Kong University of Science and Technology (Guangzhou)
ctaoaa@connect.ust.hk, xcao743@connect.hkust-gz.edu.cn, juandu@hkust-gz.edu.cn

Abstract

Industrial quality inspection plays a critical role in modern manufacturing by identifying defective products during production. While single-modality approaches using either 3D point clouds or 2D RGB images suffer from information incompleteness, multimodal anomaly detection offers promise through the complementary fusion of crossmodal data. However, existing methods face challenges in effectively integrating unimodal results and improving discriminative power. To address these limitations, we first reinterpret memory bank-based anomaly scores in single modalities as isotropic Euclidean distances in local feature spaces. Dynamically evolving from Euclidean metrics, we propose a novel Geometry-Guided Score Fusion (G²SF) framework that progressively learns an anisotropic local distance metric as a unified score for the fusion task. Through a geometric encoding operator, a novel Local Scale Prediction Network (LSPN) is proposed to predict direction-aware scaling factors that characterize first-order local feature distributions, thereby enhancing discrimination between normal and anomalous patterns. Additionally, we develop specialized loss functions and score aggregation strategy from geometric priors to ensure both metric generalization and efficacy. Comprehensive evaluations on the MVTEC-3D AD and Eyecandies datasets demonstrate the state-of-the-art detection performance of our method, and detailed ablation analysis validates each component's contribution. Our code is available at <https://github.com/ctaoaa/G2SF>.

1. Introduction

Surface quality inspection is crucial in modern industries, as defects like scratches, dents, or cracks can severely impact product functionality and longevity, requiring automated and intelligent inspection systems for large-scale production. Since defective products are always rare in real-world scenarios, it has led to growing interest in unsuper-

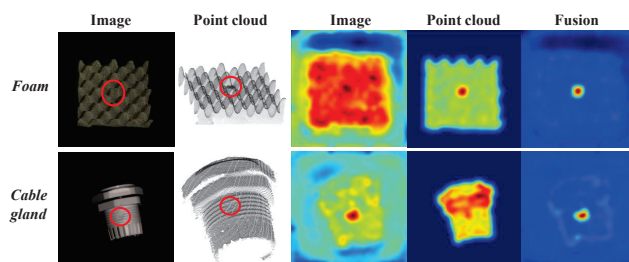


Figure 1. Two representative examples from the MVTEC-3D AD dataset [4], including the original RGB image, 3D point cloud, and anomaly score maps from single modalities (image and point cloud) and multimodal fusion (the proposed G²SF). **Top:** Foam with a ‘cut’ anomaly, which is nearly invisible in the RGB image due to its complex texture. **Bottom:** Cable gland with a ‘thread’ anomaly, which is insufficiently discriminative in the point cloud as it only slightly alters the surface geometry.

vised anomaly detection [3, 4], which uses only nominal (anomaly-free) samples for training, offering a practical solution for a wide range of industrial applications.

In complex industrial scenarios, unimodal anomaly detection methods often fail to comprehensively characterize defects due to the limitations of single data sources [19]. For instance, while 3D point clouds provide rich geometric details, they lack texture and color information in RGB images. By integrating these complementary modalities, multimodal approaches achieve superior performance. Fig. 1 illustrates two representative extreme scenarios from the MVTEC-3D AD dataset [4], where anomalies are predominantly undetectable in one specific modality.

Feature-based anomaly detection approaches have gained popularity due to their flexibility and strong performance [17, 19, 27, 30, 32]. These methods model the distributions of extracted features, obtained by either hand-crafted descriptors or frozen networks, and assign feature-wise anomaly scores. For example, the memory bank-based methods, e.g., PatchCore [27], construct a memory bank of representative prototypes from the whole training dataset and define the anomaly score as the distance to the nearest

*Corresponding author

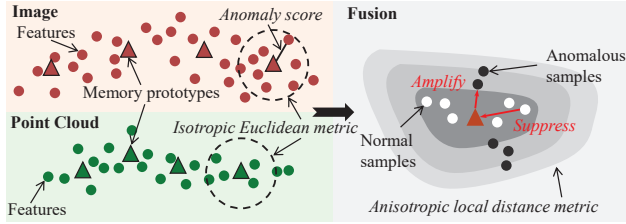


Figure 2. **Left:** The anomaly score in memory bank-based methods, which is defined by the isotropic Euclidean distance to the nearest memory prototype. **Right:** The anisotropic local distance metric learned by G^2SF , suppressing anomaly scores for normal samples while amplifying the discrepancy for anomalies.

prototype in the memory bank.

Despite efforts in multimodal methods to learn intermodal representations or mining crossmodal relationships [1, 11, 32], leveraging the valuable unimodal detection results complementary to the above intermodal information remains underexploited. Current strategies—such as score addition [19], max selection [10], or One-Class Support Vector Machine (OCSVM) [32]—face a critical limitation: unimodal anomaly scores frequently demonstrate insufficient discriminative capacity in practical applications. As visualized by Fig. 1, even normal regions in both point cloud and image data exhibit unexpectedly elevated anomaly scores. These unreliable unimodal assessments consequently compromise the fusion performance. To obtain more discriminative anomaly scores, feature adaptation [26, 30] can be applied that adapts original features into more compact representations beforehand, followed by standard methods, e.g., PatchCore [27], to calculate final anomaly scores. One limitation is that critical details of original features essential for anomaly detection may be over-confidently compromised during first-step adaptation.

This paper re-examines memory bank-based unimodal approaches, revealing a fundamental geometric limitation for their insufficient discriminative capacity: Existing methods compute anomaly scores via isotropic Euclidean distances in local feature spaces centered on memory prototypes, as illustrated in Fig. 2, but critically ignore other local geometric characteristics (e.g., directional patterns) essential for anomaly detection. To address this limitation, we propose Geometry-Guided Score Fusion (G^2SF)—a framework that replaces isotropic distances with a learned anisotropic local distance metric for multimodal anomaly detection. Specifically, we introduce a Local Scale Prediction Network (LSPN) that derives direction-aware scaling factors to characterize local directional distributions of both modalities in the joint feature space. Subsequently, we formulate a fused metric combining these scales with original Euclidean distances from two modalities to amplify normal/anomaly discrepancies, as shown in Fig. 2.

Despite the enhanced discriminative power, our G^2SF

offers two other advantages: (i) Instead of learning a new metric or anomaly detector from scratch, which are often risky in high-dimensional feature spaces [14, 15], perhaps increasing with multiple modalities involved, G^2SF explicitly quantifies how unimodal anomaly scores and local geometric structures in feature spaces contribute to the final metric. Consequently, it enables a progressive evolution from the Euclidean metric to an anisotropic metric, mitigating overfitting and ensuring robust metric learning (see Section 3.4). (ii) As a one-stage anomaly detector, G^2SF maintains complete modality-specific information throughout the detection process (details are discussed in Section 3.3), in contrast to existing feature adaptation frameworks [26, 30] aforementioned above.

The contributions of this paper are summarized as follows:

- We propose a systematic G^2SF framework for industrial multimodal anomaly detection by learning a unified discriminative metric in high-dimensional feature space.
- We propose a novel metric deformation paradigm, evolving from isotropic Euclidean metric to an anisotropic one. Specifically, throughout geometric encoding, we design a new LSPN to predict direction-aware scaling factors, subsequently formulating the final metric.
- We propose critical loss functions to maintain specific geometric properties (e.g., metric consistency and cross-modal correspondence) during metric learning and a novel geometry-informed anomaly scoring strategy, enabling precise multimodal anomaly detection.
- Our method achieves state-of-the-art performance on the MVTec 3D-AD and Eyecandies datasets, with particularly notable superiority in the AUPRO@1% metric.

2. Literature Review

2.1. 2D Image Anomaly Detection

Unsupervised anomaly detection methodologies can be broadly categorized into two paradigms: *reconstruction*-based approaches and *feature*-based techniques. Reconstruction-based methods typically employ generative architectures, e.g., transformers [25] and diffusion models [33], to establish pixel-level reconstruction fidelity for normal images. For inference, anomaly detection is performed by calculating per-pixel discrepancies between input images and their reconstructed counterparts.

Feature-based approaches, in contrast, circumvent pixel-level comparisons by operating on informative handcraft or deep features. The central challenge in this paradigm involves effectively modeling the distribution of normal features. Existing solutions adopt different modeling strategies: PaDiM [12] represents features as multivariate Gaussian distributions across spatial positions, while CFlow-AD [16] employs conditional normalizing flows to capture more

complex probability distributions. PNI [2] further enhances modeling by incorporating positional and neighborhood relationships of features. A notable method is PatchCore [27], which constructs a memory bank of representative normal features through coreset sampling. Owing to its robustness and adaptability, PatchCore has become a foundational method applicable to both 3D [19], multimodal [19, 30, 32], and noisy scenarios [20].

2.2. Multimodal 2D/3D Anomaly Detection

The pioneering MVTEC-3D AD dataset [4] benchmarks this task with various generative networks. For real-time requirements, Easy-Net [9] designs a multi-scale multimodal autoencoder. Feature-based methods prevail in this domain. CFM [11] establishes explicit one-to-one feature mappings to model crossmodal relationships, while 2M3DF [1] extends this by learning intermodal features via multi-view RGB image generation. Similarly, M3DM [32] incorporates mutual information for feature fusion. On the other hand, unimodal anomaly scores-containing essential modality-specific information-also demand dedicated fusion strategies. BTF [19] concatenates 2D/3D features with PatchCore [27], equivalent to linear score fusion. Based on the signed distance function (SDF) for 3D patch representation, Shape-Guided [10] establishes SDF-guided memory banks for fine-grained detection on image modality but relies on simplistic maximum value between image and SDF scores. M3DM [32] employs one-class SVM for data-driven score fusion of the original and fused features. However, the inherent limitations in the discriminative power of unimodal anomaly scores remain unaddressed. Toward this deficiency, LSFA [30] proposes a crossmodal feature adaptation paradigm that learns task-oriented features but risks overlooking the critical details of original features through adaptation. Finally, at data-level, DAUP [23] attributes the issue of indiscriminative point cloud scores to the irregular point density and tackles it by resampling point clouds.

Our G²SF introduces two core advances over these existing approaches: (i) We learn a unified discriminative metric, eliminating suboptimal aggregation from unreliable unimodal anomaly scores; (ii) We introduce a dynamic metric deformation paradigm through isotropic-to-anisotropic evolution, preserving original modality-specific details while enabling precise discrimination in high-dimensional space.

3. Methodology

Our G²SF aims to learn a unified anisotropic local distance metric for multimodal anomaly detection. The overall pipeline is illustrated in Fig. 3.

3.1. Feature Extraction

Our framework accepts a pair of 3D point cloud and RGB image as input. Specifically, we employ two frozen neural

networks (\mathcal{F}_P and \mathcal{F}_R) to extract corresponding features \mathbf{f}_i^P (point cloud) and \mathbf{f}_i^R (image) at location i . Following existing methods [11, 30, 32], we adopt two vision transformers, DINO [7] for images and Point-MAE [24] for point clouds, considering their powerful representational capacity. \mathbf{f}_i^P and \mathbf{f}_i^R have high dimensions, i.e., 1152 and 768, respectively.

3.2. Prototype Retrieval from Memory Bank

Our G²SF builds on the memory bank-based methods for unimodal anomaly detection. For each modality $m \in P, R$, we construct a memory bank \mathcal{M}^m that collects prototypical features from the entire training dataset following PatchCore [27]. Subsequently, for a feature \mathbf{f}_i^m , we can retrieve its $2k + 1$ nearest neighbors $\{\mathbf{m}_{i,j}^m\}_{j=0}^{2k}$ from \mathcal{M}^m . Critically, the Euclidean distance between \mathbf{f}_i^m and the nearest $\mathbf{m}_{i,0}^m$ is typically defined as the anomaly score $s_{i,0}^m$ in these memory bank-based methods [32].

3.3. Geometric Feature Encoding

The main issue that limits the discriminative capability of $s_{i,0}^m$ in Section 3.2 is its isotropic nature, which ignores other critical characteristics valuable for anomaly detection. Nevertheless, local directional consistency can act as an auxiliary property to quantitatively evaluate the normality of target samples: Features aligned with the principal directions of normal samples within the local feature spaces are expected to demonstrate a higher degree of normality.

Therefore, a geometric encoding operator $\mathcal{E}(\cdot)$ that naturally characterizes the first-order local distribution within the local manifold centered at $\mathbf{m}_{i,j}^m$ can be defined as [36]:

$$\begin{aligned} (\mathbf{m}_{i,j}^m, \mathbf{d}_{i,j}^m, s_{i,j}^m) &= \mathcal{E}_{\mathbf{m}_{i,j}^m}(\mathbf{f}_i^m), \\ s_{i,j}^m &= \|\mathbf{f}_i^m - \mathbf{m}_{i,j}^m\|, \quad \mathbf{d}_{i,j}^m = (\mathbf{f}_i^m - \mathbf{m}_{i,j}^m)/s_{i,j}^m, \end{aligned} \quad (1)$$

where $s_{i,j}^m$ represents the Euclidean distance and $\mathbf{d}_{i,j}^m$ denotes the directional vector. Moreover, Eq. (1) is a seamless encoding of \mathbf{f}_i^m . This property enables the preservation of complete details for the subsequent learning process, avoiding comprising valuable information in feature adaptation frameworks [26, 30].

3.4. Anisotropic Local Distance Metric

Based on Eq. (1), we aim to learn a unified metric $l(\mathbf{f}_i^P, \mathbf{f}_i^R, \mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R)$ for multimodal anomaly detection that outperforms the Euclidean distances $s_{i,j}^P$ and $s_{i,j}^R$. While existing local metric learning methods [31, 35] use instance-dependent Mahalanobis distances, they are computationally expensive and unsuitable for our high-dimensional joint feature space (1920 dimensions). From the perspective of spectral analysis, Mahalanobis distances determine the scaling factors along principle directions. Mimicking this desired property, we propose the LSPN that

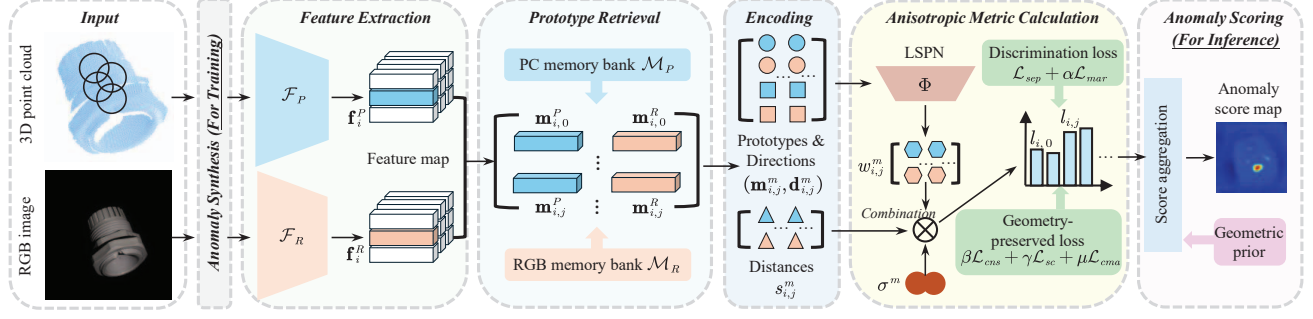


Figure 3. Overall pipeline of the proposed G^2SF . During training, our model takes a pair of 3D point clouds and RGB images as inputs. The anomaly synthesis module (Section 3.7) then injects synthetic anomalies into both modalities. Next, the nearest memory prototypes $\{\mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R\}$ are retrieved from the 3D/2D memory banks $\mathcal{M}_P/\mathcal{M}_R$ (Section 3.2). In Section 3.3, the features $\mathbf{f}_i^P/\mathbf{f}_i^R$ are transformed into geometrically meaningful encodings, represented as triplets: $\langle \text{prototypes } \mathbf{m}_{i,j}^m, \text{directions } \mathbf{d}_{i,j}^m, \text{distances } s_{i,j}^m \rangle$ ($m \in \{P, R\}$). The LSPN Φ learns direction-aware scaling factors $w_{i,j}^m$ from $(\mathbf{m}_{i,j}^m, \mathbf{d}_{i,j}^m, \mathbf{m}_{i,j}^m, \mathbf{d}_{i,j}^m)$, which are combined with the distances $s_{i,j}^m$ and global scaling factors σ^m to compute the anisotropic local distance metric values $\{l_{i,j}\}$, as detailed in Section 3.4. The learning process is supervised by two loss functions: the discrimination loss ($\mathcal{L}_{sep} + \alpha\mathcal{L}_{mar}$, Section 3.5.1) and the geometry-preserved loss ($\beta\mathcal{L}_{cns} + \gamma\mathcal{L}_{sc} + \mu\mathcal{L}_{cma}$, Section 3.5.2). Finally, during inference, a score aggregation strategy is applied, leveraging geometric priors for anomaly scoring, as described in Section 3.6.

leverages a neural network for computational efficiency and implicitly captures shared information across local spaces.

Specifically, LSPN predicts direction-aware scaling factors $w_{i,j}^m$ via a simple Multilayer Perceptron $\Phi(\cdot)$ (MLP) as:

$$[w_{i,j}^P, w_{i,j}^R] = \Phi(\mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R, \mathbf{d}_{i,j}^P, \mathbf{d}_{i,j}^R). \quad (2)$$

Furthermore, the anisotropic metric $l(\cdot)$ is defined in terms of these scaling factors as:

$$l_{i,j} = l(\mathbf{f}_i^P, \mathbf{f}_i^R, \mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R) = \sum_{m \in \{P, R\}} w_{i,j}^m s_{i,j}^m \sigma^m, \quad (3)$$

where σ^m is a *trainable global scaling factor* that describes the influence of modality m . Eq. (3) demonstrates that $l(\cdot)$ incorporates structural priors from unimodal anomaly detection, where $s_{i,0}^m$ explicitly represents the unimodal anomaly score. When initializing Φ with $w_{i,j}^m \approx 1$, the metric $l(\cdot)$ originates as a Euclidean measure in the joint feature space. Although this baseline configuration is marginally effective in unimodal scenarios, it progressively evolves into an anisotropic metric during training under the regularization of $s_{i,j}^m$. This unique property prevents overfitting that would occur when learning a metric entirely from scratch [14, 15]. Furthermore, the supplementary material tracks model performance throughout different training epochs, revealing that G^2SF attains competitive performance within remarkably few training epochs that benefit from the above design.

The architecture of $\Phi(\cdot)$ is illustrated in Fig. 4, which consists of two parallel branches that independently concatenate and encode the prototypes $(\mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R)$ and directions $(\mathbf{d}_{i,j}^P, \mathbf{d}_{i,j}^R)$. $w_{i,j}^m$ is obtained through an activation

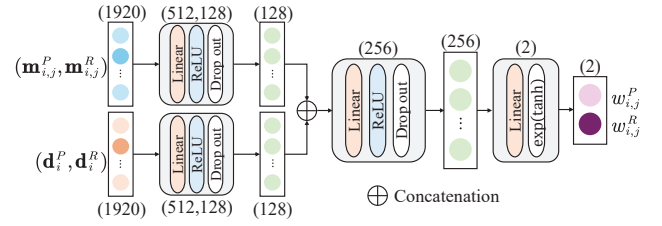


Figure 4. The architecture of LSPN. Each block contains a linear layer, ReLU activation, and dropout. The final layer activation is a $\exp(\tanh(\cdot))$ function that is symmetrical to one and serves as scale suppression/amplification.

$\exp(\tanh(\cdot))$ at the final layer, ensuring $w_{i,j}^m \in [e^{-1}, e^1]$. This constrains the scaling factors to be symmetric around the Euclidean baseline ($w_{i,j}^m = 1$). To enhance the discriminative power of $l(\cdot)$, we aim to reduce scales ($w_{i,j}^m < 1$) for normal samples, while increasing scales ($w_{i,j}^m > 1$) for anomalous samples.

3.5. Loss Functions

To ensure stable training, we employ synthetic anomaly injection in Section 3.7. We denote the label $y_i \in \{0, 1\}$ for $(\mathbf{f}_i^P, \mathbf{f}_i^R)$, where $y_i = 1$ denotes synthetic anomalies. Additionally, we denote $(\cdot)_+$ as $\max(\cdot, 0)$. We deploy the discrimination losses ($\mathcal{L}_{sep}, \mathcal{L}_{mar}$) in Section 3.5.1 to ensure the learned $l(\cdot)$ capable of discriminating normal ($y_i = 0$) and anomalous ($y_i = 1$) samples. Furthermore, the geometry-preserved losses ($\mathcal{L}_{cns}, \mathcal{L}_{sc}, \mathcal{L}_{cma}$) are proposed in Section 3.5.2 to ensure effective metric learning by incorporating desired geometric properties related to metric consistency, local scaling factors, and crossmodal alignments.

3.5.1. Discrimination Loss

We prioritize $l_{i,0}$ to inherit its discriminative capability w.r.t. the nearest prototype $\mathbf{m}_{i,0}^m$.

Separation Loss: The basic separation loss \mathcal{L}_{sep} is to minimize $l_{i,0}$ for normal samples while simultaneously maximizing them for anomalous instances, following the semi-supervised anomaly detection approach [29]:

$$\mathcal{L}_{sep} = \sum_i (1 - y_i) l_{i,0} + y_i (l_{i,0}^{-1} - m_0^{-1})_+, \quad (4)$$

where m_0 prevents excessively large anomaly scores.

Margin Loss: The distribution overlap between normal and anomalous samples is also crucial for anomaly detection [6], which is considered by the margin loss \mathcal{L}_{mar} as:

$$\mathcal{L}_{mar} = \frac{1}{m_u} \sum_i \underbrace{(1 - y_i)(l_{i,0} - m_l)_+}_{\text{Overlap of Normal Part}} + \underbrace{y_i(m_u - l_{i,0})_+}_{\text{Overlap of Anomalous Part}}, \quad (5)$$

where $m_l = \min_{i|y_i=1} l_{i,0}$ is the lower bound of $l_{i,0}$ for anomalous samples and $m_u = \max_{i|y_i=0} l_{i,0}$ is the upper bound for normal samples.

3.5.2. Geometry-Preserved Loss

Since learning a robust and generalizable $l(\cdot)$ in the high-dimensional space for anomaly detection is challenging, despite the architecture design in Section 3.4, we aim to exploit the desired geometric properties of $l(\cdot)$ that benefit our task.

Consistency Loss: Since Eqs. (4) and (5) focus on the discriminative power of $l_{i,0}$, we exploit the geometric consistency of $l_{i,j}$ across different local spaces around $\mathbf{m}_{i,j}^m$, including the following two properties:

- *Proximal consistency:* $l(\cdot)$ is continuous across neighborhood. For adjacent local spaces, i.e., $1 \leq j \leq k$, $l_{i,j}$ is upper bounded by $\eta_{i,j} l_{i,0}$ for $\eta_{i,j} > 1$. It means that the partial gradient of $l(\cdot)$ w.r.t. $(\mathbf{m}_{i,j}^P, \mathbf{m}_{i,j}^R)$ is limited.
- *Distal separation:* For sufficiently large Euclidean distances, $l(\cdot)$ should maintain correspondingly significant values. We mandate $l_{i,j}$ in distant local spaces indexed by $k+1 \leq j \leq 2k$ lower-bounded by $l_{i,0}$. This constraint prevents metric collapse in high-dimensional spaces to avoid incredible values.

Accordingly, we formulate the consistency loss \mathcal{L}_{cns} as:

$$\mathcal{L}_{cns} = \frac{1}{k} \sum_i \underbrace{\sum_{1 \leq j \leq k} (l_{i,j} - \eta_{i,j} l_{i,0})_+}_{\text{Proximal Consistency}} + \underbrace{\sum_{k+1 \leq j \leq 2k} (l_{i,0} - l_{i,j})_+}_{\text{Distal Separation}}. \quad (6)$$

Despite the spatial regularization, \mathcal{L}_{cns} also effectively increases sample diversity within each local space through multi-prototype observations as data augmentation.

Scaling Loss: Despite $l(\cdot)$, we aim to encourage $w_{i,0}^m$ to encode local directional distributions explicitly, that is, suppress/amplify for normal/anomalous samples. To achieve

this, we propose an asymmetric scaling loss:

$$\mathcal{L}_{sc} = \sum_i \sum_{m \in \{P,R\}} (1 - y_i)(w_{i,0}^m - 1)_+ + y_i(e^1 - w_{i,0}^m). \quad (7)$$

\mathcal{L}_{sc} asymmetrically compresses scaling factors of normal samples while confining those of synthetic anomalies near the upper limit e^1 . This design addresses the more pronounced deviations of synthetic anomalies than real-world ones, imposing more penalties on them.

Crossmodal Alignment Loss: Currently, both prototypes and directions are naively concatenated as input to LSPN (Sec. 3.4), ignoring their crossmodal correspondence. To resolve this, we propose a simple strategy by disentangling this property.

Specifically, by permuting indices $\{i\} \rightarrow \{\tilde{i}\}$, we can construct two kinds of negative samples:

$$(\mathbf{m}_{i,0}^P, \mathbf{m}_{\tilde{i},0}^R, \mathbf{d}_{i,0}^P, \mathbf{d}_{\tilde{i},0}^R), (\mathbf{m}_{\tilde{i},0}^P, \mathbf{m}_{i,0}^R, \mathbf{d}_{\tilde{i},0}^P, \mathbf{d}_{i,0}^R). \quad (8)$$

The former breaks holistic prototype coherence. In contrast, the latter focuses on capturing discordant directions while maintaining prototype consistency at a local structural granularity.

During training, we construct the above two kinds with equal probability. Denote $\tilde{w}_{i,0}^m$ ($m \in \{P, R\}$) as their associated scaling factors, we define the crossmodal alignment loss as:

$$\mathcal{L}_{cma} = \sum_i \sum_{m \in \{P,R\}} (e^1 - \tilde{w}_{i,0}^m), \quad (9)$$

which enforces large scales similar to synthetic anomalous samples in Eq. (7).

3.5.3. Overall Loss

Finally, the overall loss \mathcal{L} of our G²SF framework combines all terms as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{sep} + \alpha \mathcal{L}_{mar}}_{\text{Discrimination Loss}} + \underbrace{\beta \mathcal{L}_{cns} + \gamma \mathcal{L}_{sc} + \mu \mathcal{L}_{cma}}_{\text{Geometry-Preserved Loss}}, \quad (10)$$

where α , β , γ , and μ are hyper-parameters balancing the loss components.

3.6. Geometry-Informed Anomaly Scoring

During inference, G²SF computes the anomaly score s_i for each multimodal feature pair $(\mathbf{f}_i^P, \mathbf{f}_i^R)$. Concretely, by propagating features through the geometric encoding operator and LSPN, we obtain the anisotropic metrics $\{l_{i,j}\}_{j=0}^k$ across $k+1$ nearest local spaces, adhering to the proximal consistency constraint in Eq. (6). We apply a min operator to aggregate $\{l_{i,j}\}_{j=0}^k$ to define s_i :

$$s_i = \min\{l_{i,j} | j = 0, 1, \dots, k\}. \quad (11)$$

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
I-AUROC	DepthGAN[4]	0.538	0.372	0.580	0.603	0.430	0.534	0.642	0.601	0.443	0.577	0.532
	VoxelGAN[4]	0.680	0.324	0.565	0.399	0.497	0.482	0.566	0.579	0.601	0.482	0.517
	BTF[19]	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	0.886	0.865
	EasyNet[9]	0.991	0.998	0.918	0.968	0.945	0.945	0.905	0.807	<u>0.994</u>	0.793	0.926
	AST[28]	0.983	0.873	0.976	0.971	0.932	0.885	0.974	0.981	1.000	0.797	0.937
	M3DM [32]	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
	Shape-guided [10]	0.986	0.894	0.983	0.991	0.976	0.857	<u>0.990</u>	0.965	0.960	0.869	0.947
	CFM [11]	0.994	0.888	0.984	<u>0.993</u>	0.980	0.888	0.941	0.943	0.980	<u>0.953</u>	0.954
	2M3DF [1]	0.992	<u>0.969</u>	<u>0.988</u>	0.985	0.981	0.947	0.979	0.942	0.976	0.898	0.966
	LSFA [30]	1.000	0.939	0.982	0.989	0.961	<u>0.951</u>	0.983	0.962	0.989	<u>0.951</u>	0.971
	DAUP [23]	0.996	0.889	0.996	0.998	<u>0.977</u>	<u>0.939</u>	0.983	<u>0.986</u>	0.979	0.960	<u>0.970</u>
Ours	<u>0.997</u>	<u>0.923</u>	<u>0.993</u>	<u>0.967</u>	<u>0.966</u>	0.991	0.994	0.988	0.966	0.922	0.971	
AUPRO@30%	DepthGAN[4]	0.421	0.422	0.778	0.696	0.494	0.252	0.285	0.362	0.402	0.631	0.474
	VoxelGAN[4]	0.664	0.620	0.766	0.740	0.783	0.332	0.582	0.790	0.633	0.483	0.639
	BTF[19]	0.976	0.969	0.979	0.973	0.933	0.888	0.975	0.981	0.950	0.971	0.959
	AST[28]	0.970	0.947	0.981	0.939	0.913	0.906	0.979	<u>0.982</u>	0.889	0.940	0.944
	M3DM [32]	0.970	0.971	0.979	0.950	0.941	0.932	0.977	<u>0.971</u>	0.971	0.975	0.964
	Shape-guided [10]	0.981	0.973	<u>0.982</u>	0.971	0.962	0.978	0.981	0.983	0.974	0.975	<u>0.976</u>
	CFM [11]	0.979	0.972	<u>0.982</u>	0.945	0.950	0.968	0.980	<u>0.982</u>	<u>0.975</u>	<u>0.981</u>	0.971
	2M3DF [1]	0.986	0.983	0.987	0.984	<u>0.968</u>	0.949	0.986	0.971	0.988	0.954	0.975
	LSFA [30]	0.986	0.974	0.981	0.946	<u>0.925</u>	0.941	0.983	0.983	0.974	0.983	0.968
	DAUP [23]	0.976	<u>0.977</u>	0.980	0.960	0.924	0.966	0.981	0.978	0.972	0.980	0.969
	Ours	<u>0.982</u>	<u>0.977</u>	<u>0.982</u>	<u>0.979</u>	0.971	<u>0.976</u>	0.982	0.983	<u>0.978</u>	<u>0.981</u>	0.979

Table 1. I-AUROC and AUPRO@30% results on MVTec-3D AD. Best results in **bold**, runner-ups underlined.

This definition stems from a geometric prior. Specifically, memory bank-based methods calculate anomaly scores using Euclidean distances to their memory banks (Section 3.2), whereas our formulation in Eq. (11) reformulates this as distance measurement to the normal data manifold through the anisotropic metrics $l(\cdot)$. We will compare Eq. (11) with several alternatives in Section 4.3.

Finally, the sample-level anomaly score is obtained by taking the maximum s_i across all feature pairs extracted from a test sample. Following [11, 30, 32], we apply bilinear interpolation and a Gaussian kernel with size 4 to upsample and smooth the score map, ensuring resolution alignment with input dimensions.

3.7. Anomaly Synthesis

Our anomaly synthesis procedure operates following Cut-Paste approaches [9, 22, 34]: First, we generate Berlin noise maps, which are binarized into mask maps to select paste locations. Based on these masks, we then cut corresponding regions from source (normal) 3D point clouds and RGB images from the training dataset. The source point clouds can be other classes or more general 3D models. Subsequently, we apply random augmentations to corrupt the regions - specifically implementing pixel modifications for source images [34], e.g., distortion, brightness, and sharpness changes, while executing random translations for source point clouds. The corrupted regions are finally pasted onto target data. Examples of synthetic anomalous samples are provided in the supplementary material.

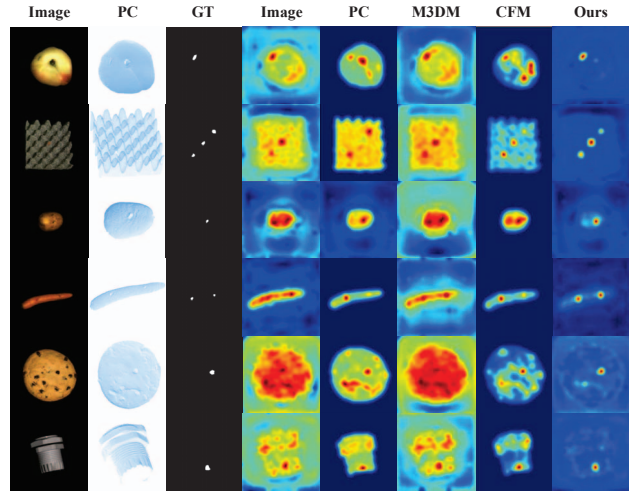


Figure 5. Qualitative comparison on MVTec-3D AD [4]: Visualized results include original RGB images with aligned 3D point clouds (PC) and ground truths (GT), alongside anomaly detection outputs from unimodal methods (PatchCore [27] for image/point cloud) and multimodal benchmarks (M3DM [32], CFM [11]), compared against our proposed G^2SF framework.

4. Experiment

4.1. Experimental Details

Dataset: Experiments are conducted on the MVTec-3D AD [4] and Eyecandies[5] datasets. MVTec-3D AD contains 10 industrial categories (210-361 train, 69-132 test per class), while Eyecandies provides 10 synthetic food items

	Method	I-	P-	AUPRO @30%	AUPRO @1%
		AUROC	AUROC		
MVTec-3D AD	BTF [19]	0.865	0.992	0.959	0.383
	AST [28]	0.937	0.976	0.944	0.398
	M3DM [32]	0.945	0.992	0.964	0.394
	Shape. [11]	0.947	0.996	0.976	0.456
	CFM [11]	0.954	0.993	0.971	0.455
	2M3DF [1]	0.966	-	0.975	-
	LSFA [30]	0.971	0.993	0.968	-
	DUAP [23]	0.970	-	0.969	-
Ours	0.971	0.997	0.979	0.468	
Eyecandies	AST [28]	0.780	0.902	0.744	0.149
	M3DM [32]	0.882	0.977	0.887	0.331
	CFM [11]	0.881	0.973	0.887	0.335
	2M3DF [1]	0.897	-	0.890	-
	LSFA [30]	-	0.974	-	-
	Ours	0.902	0.982	0.898	0.357

Table 2. Mean metrics for all categories on both datasets.

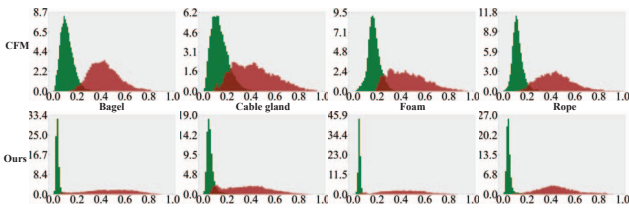


Figure 6. Pixel-wise anomaly score distributions in *Bagel*, *Cable gland*, *Foam*, and *Rope* categories, where the horizontal axis is normalized score (0-1) with vertical axis as the probabilistic density. Green/Brown distributions represent normal/anomalous regions respectively. Compared with CFM [11] (top), our method (bottom) achieves significant distribution compression in normal regions while maintaining anomaly discrimination.

(1k train, 400 test per class). Each sample provides aligned RGB images and 3D point clouds, enabling pixel-wise correspondence between color and spatial coordinates.

Evaluation Metrics: We assess the image-level anomaly detection performance by the Area Under the Receiver Operator Curve (I-AUROC). For evaluating the pixel-wise anomaly segmentation, we adopt the pixel-level AUROC (P-AUROC) and the Area Under the Per-Region Overlap (AUPRO) at 30% integration limits (denoted as AUPRO@30%), the same as existing methods [10, 11, 19, 32]. To better demonstrate the advantages of our method in practical industrial scenarios with lower false positive rates, we also adopted the AUPRO@1% metric (additional AUPRO@5% and AUPRO10% in the supplementary material) used in recent studies [10, 11].

Implementation: We follow a similar procedure of M3DM [32] for feature extraction and preprocessing of the dataset. For \mathcal{L}_{cns} , we set $k = 5$ and $\eta_0 = 1.2$. We train our G²SF with the Adam optimizer [21] with 80 epochs on Linux Nvidia RTX 4090. More details are provided in the supplementary material.

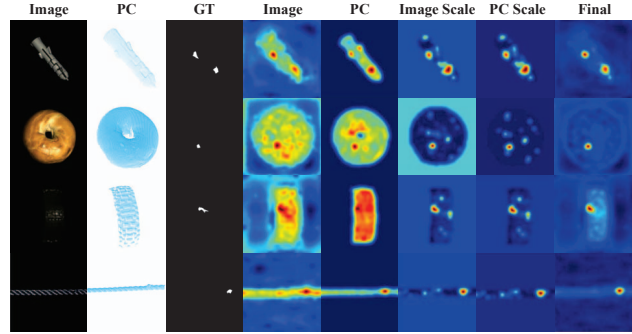


Figure 7. Qualitative result on MVTec-3D AD: Original RGB images, 3D point clouds (PC), ground truth (GT), and score maps of s_i^R (Image), s_i^P (PC), $w_{i,0}^R$ (Image Scale), $w_{i,0}^P$ (PC Scale), and s_i (Final), from left to right. The scaling factors ($w_{i,0}^R$, $w_{i,0}^P$) exhibit false positives in certain normal regions, whereas the final score s_i effectively suppresses these errors through incorporating distance information from s_i^R and s_i^P .

Score	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
s_i^R	0.849	0.988	0.947	0.363
s_i^P	0.854	0.975	0.915	0.319
$w_{i,0}^R$	0.916	0.948	0.916	0.439
$w_{i,0}^P$	0.915	0.984	0.958	0.443
s_i	0.971	0.997	0.979	0.468

Table 3. Analysis of various score metrics on MVTec-3D AD.

4.2. Comparison on Multimodal Benchmarks

We illustrate the results of I-AUROC and AUPRO@30% on MVTec 3D-AD in Table 4, while those of AUPRO@1% and P-AUROC are presented in Table 2. Specifically, for anomaly detection, the proposed G²SF achieves the highest I-AUROC value of 97.1%. Although LSFA [30] matches G²SF in terms of I-AUROC, its AUPRO@30% result for anomaly segmentation is 96.8%, which is lower than G²SF’s 97.9%. Regarding anomaly segmentation, our G²SF outperforms all competitors in three pixel metrics with a +1.2% AUPRO@1% advantage over Shape-guided [10]. Notably, for AUPRO@30% value, although Shape-guided is close to G²SF, its I-AUROC of 94.7% is significantly lower than G²SF’s 97.1%. Overall, G²SF achieves the best performance in both anomaly detection and anomaly segmentation tasks. This performance gain stems from our purpose of learning a discriminative local distance metric. As demonstrated by the qualitative results visualized in Fig. 5, G²SF effectively suppresses overestimated scores in normal regions compared to M3DM [32] and CFM [11], while producing sharper anomaly boundaries. To provide a comprehensive visualization, Fig. 6 illustrates pixel-wise anomaly score distributions in four categories, demonstrating that anomaly scores of normal regions in G²SF shrinkage significantly within a narrow range. This property effectively reduces false positives,

Components	Stepwise Combinations				
	✓	✓	✓	✓	✓
\mathcal{L}_{sep} in Eq. (4)	✓	✓	✓	✓	✓
\mathcal{L}_{cns} in Eq. (6)	✗	✓	✓	✓	✓
\mathcal{L}_{mar} in Eq. (5)	✗	✗	✓	✓	✓
\mathcal{L}_{sc} in Eq. (7)	✗	✗	✗	✓	✓
\mathcal{L}_{cma} in Eq. (9)	✗	✗	✗	✗	✓
I-AUROC	0.868	0.950	0.950	0.961	0.971
P-AUROC	0.994	0.995	0.996	0.997	0.997
AUPRO@30%	0.967	0.972	0.975	0.978	0.979

Table 4. Ablation results of individual loss components within G²SF on MVTEC-3D AD dataset.

Strategy	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
$l_{i,0}$	0.963	0.997	0.977	0.461
max	0.953	0.986	0.960	0.453
mean	0.972	0.983	0.955	0.458
min	0.971	0.997	0.979	0.468

Table 5. Investigation on different score aggregation strategies on MVTEC 3D-AD dataset.

leading to more substantial gains in AUPRO@1% (Table 2). In addition, the results on Eyecandies in Table 2 showcase the best performance of our G²SF over all metrics, particularly with +2.2% AUPRO@1% over CFM [19]. More details are provided in the supplementary material.

4.3. Analysis

Effectiveness of Final Score s_i : Table 3 compares key score metrics for anomaly detection: unimodal scores s_i^m ($m \in \{P, R\}$) (Euclidean distances, see Section 3.2), scaling factors $w_{i,0}^m$ from LSPN in Eq. (2), and our fused score s_i in Eq. (11). As evidenced in Table 3, unimodal anomaly scores s_i^m demonstrate suboptimal performance. This limitation stems from the inherent information incompleteness of individual modalities. Our fused score s_i overcomes these constraints through crossmodal complementarity, achieving significant improvements over unimodal baselines. These quantitative gains are also supported by Figs. 5 and 7.

While $w_{i,0}^m$ characterizes the local directions of normal features within local spaces (Section 3.4), its exclusion of the valuable distances, i.e., s_i^m , leads to critical limitations. Specifically, proximity to memory prototypes ($s_i^m \rightarrow 0$) inherently indicates normality regardless of directional deviations, which is a prior ignored by $w_{i,0}^m$. As evidenced in Fig. 7, the score maps obtained by $w_{i,0}^m$ exhibit noticeable false positives. Using s_i in our G²SF mitigates this issue, which benefits from sufficiently leveraging complementary directional and distance information. The comprehensive results are provided in the supplementary material.

Ablation of Loss Components: Table 4 systematically evaluates the contribution of each loss component through stepwise removal. Using the separation loss \mathcal{L}_{sep} as the

Method	AST [28]	M3DM [32]	CFM[11]	Ours
Frame Rate	4.97	2.12	21.8	6.04

Table 6. Frame rate comparison on MVTEC 3D-AD dataset.

baseline for the metric learning, the addition of the consistency loss \mathcal{L}_{cns} drives an 8.2% I-AUROC improvement, surpassing M3DM [32] across all metrics and achieving parity with CFM [11]. The margin loss \mathcal{L}_{mar} specifically improves pixel-wise localization (AUPRO@30% +0.3%), while the scaling loss \mathcal{L}_{sc} and cross-modal alignment loss \mathcal{L}_{cma} further enhances image-level detection (I-AUROC +1.1/1.0%). These synergistic effects collectively establish G²SF’s state-of-the-art performance, validating the necessity of all proposed loss components.

Investigation of Score Aggregation Strategy: Section 3.6 employs a min operator in Eq. (11) to aggregate metrics $\{l_{i,j}\}$ from adjacent local spaces. We systematically evaluate three alternatives in Table 5: (1) single-space baseline $l_{i,0}$. (2) max operator. (3) mean operator. The min-based strategy achieves superior overall performance (+0.8%/0.7% I-AUROC/AUPRO@1% vs $l_{i,0}$), validating the necessity of multi-space consensus. While max and mean operators incorporate adjacent local spaces, they indiscriminately weight less discriminative $l_{i,j}$ from distant regions, compromising detection precision. The min operator in Eq. (11) prioritizes the most representative local metric through a proper geometric prior as discussed in Section 3.6. The results of all categories of MVTEC-3D AD are summarized in the supplementary material.

Computational Time: Table 6 shows that G²SF detects 6.04 samples per second, sufficient for general industrial requirements.

Analyses of Memory Prototypes and Global Scaling Factors σ_m : Supplementary material provides the sensitivity study about the number of memory prototypes, demonstrating relatively stable performance. Additionally, the investigation on the role of σ_m is included, showing the necessity of learning optimal σ_m instead of fixed ones.

5. Conclusion

In this paper, we propose G²SF for multimodal anomaly detection, which learns a unified anisotropic local distance metric to replace the indiscriminative isotropic Euclidean metrics. This metric is formulated by the direction-aware scaling factors predicted by LSPN and original Euclidean distances in individual modalities. We propose new loss functions to enable robust metric learning, serving for discrimination or geometric property maintenance. A geometric prior-induced anomaly scoring strategy is proposed to enhance anomaly detection and segmentation. Extensive experimental results show our state-of-art performance compared to existing competitive methods.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 72371219 and No. 72001139), and the Guangzhou Municipal Science and Technology Project (No. 2025A04J5288). Corresponding author is Juan Du.

References

- [1] Mujtaba Asad, Waqar Azeem, He Jiang, Hafiz Tayyab Mustafa, Jie Yang, and Wei Liu. 2m3df: Advancing 3d industrial defect detection with multi perspective multimodal fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025.
- [2] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. PNI: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023.
- [3] 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*, pages 9592–9600, 2019.
- [4] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTEC-3D AD dataset for unsupervised 3D anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.
- [5] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the 16th Asian Conference on Computer Vision (ACCV)*, 2022. ACCV.
- [6] Yunkang Cao, Xiaohao Xu, Zhaoge Liu, and Weiming Shen. Collaborative discrepancy optimization for reliable image anomaly localization. *IEEE Transactions on Industrial Informatics*, 19(11):10674–10683, 2023.
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.
- [8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [9] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3D industrial anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7038–7046, 2023.
- [10] Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3D anomaly detection. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.
- [11] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17234–17243, 2024.
- [12] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.
- [14] Jianqing Fan, Yingying Fan, and Yichao Wu. High-dimensional classification. In *High-Dimensional Data Analysis*, pages 3–37. World Scientific, 2011.
- [15] Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(4):745–771, 2012.
- [16] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022.
- [17] Yinghui Guo, Meng Jiang, Qianhong Huang, Yang Cheng, and Jun Gong. Mldfr: A multilevel features restoration method based on damaged images for anomaly detection and localization. *IEEE Transactions on Industrial Informatics*, 20(2):2477–2486, 2023.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [19] Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3D features are (almost) all you need for 3D anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2968–2977, 2023.
- [20] Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022.
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [22] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.
- [23] Hefei Li, Yanchang Niu, Haonan Yin, Yu Mo, Yi Liu, Biqing Huang, Ruibin Wu, and Jingxian Liu. Daup: Enhancing point cloud homogeneity for 3D industrial anomaly detection via density-aware point cloud upsampling. *Advanced Engineering Informatics*, 62:102823, 2024.

- [24] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, pages 604–621. Springer, 2022.
- [25] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, pages 394–406. Springer, 2022.
- [26] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.
- [27] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.
- [28] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2592–2602, 2023.
- [29] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.
- [30] Yuanpeng Tu, Boshen Zhang, Liang Liu, Yuxi Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Self-supervised feature adaptation for 3D industrial anomaly detection. In *European Conference on Computer Vision*, pages 75–91. Springer, 2024.
- [31] Jun Wang, Alexandros Kalousis, and Adam Woznica. Parametric local metric learning for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 25, 2012.
- [32] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023.
- [33] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpn: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.
- [34] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24490–24499, 2023.
- [35] Han-Jia Ye, De-Chuan Zhan, Nan Li, and Yuan Jiang. Learning multiple local metrics: Global consideration helps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1698–1712, 2019.
- [36] Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE*