This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

From Attraction to Equilibrium: Physics-Inspired Semantic Gravitons for Zero-Shot Anomaly Detection

Yuwen Pan[1] Yuan Wang[2] Shaohui Li[3] Zhi Li[1] Yu Liu[4*] You He[4]
[1]Shenzhen International Graduate School, Tsinghua University
[2]School of Information Science and Technology, University of Science and Technology of China
[3]College of Information Science and Electronic Engineering, Zhejiang University
[4]Department of Electronic Engineering, Tsinghua University
panyuwen@sz.tsinghua.edu.cn, wy2016@mail.ustc.edu.cn,
lishaohui@zju.edu.cn, zhilizl@sz.tsinghua.edu.cn, {liuyu_thu, heyou}@mail.tsinghua.edu.cn

Abstract

Zero-shot anomaly detection (ZSAD) aims to identify unseen anomalies without abnormal supervision, which is essential for open-world scenarios. Recent vision–language models such as CLIP enable anomaly reasoning through shared visual–textual embeddings, but existing methods often rely on coarse prompt fusion, leading to unstable alignment and imprecise localization under domain shifts. To address this issue, we propose the Semantic Graviton Network (SGNet), a physics-inspired framework that models multimodal alignment as an adaptive potential field. We introduce semantic gravitons, learnable dynamic mediators that bridge visual and textual modalities by establishing localized semantic equilibria through attraction and equilibrium forces. A graviton interaction network alternates text-to-graviton and vision-to-graviton coupling to progressively refine multimodal correspondence, while an energy-based potential regularization further stabilizes the interaction process. Extensive experiments on ten industrial and medical benchmarks show that SGNet achieves state-of-the-art performance for zero-shot anomaly detection.

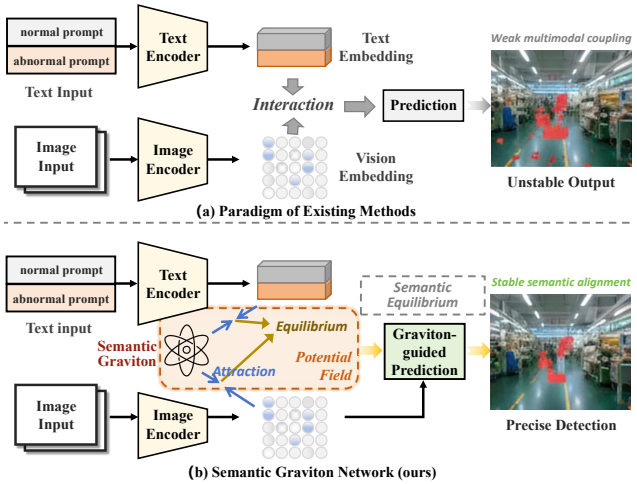

Figure 1. Comparison between existing anomaly detection methods and the proposed Semantic Graviton Network (SGNet). (a) Existing approaches perform late-stage fusion with weak multimodal coupling, leading to unstable correspondence and inaccurate anomaly localization. (b) SGNet introduces a physics-inspired potential field, where learnable *semantic gravitons* mediate between visual and textual modalities through attraction and equilibrium forces, achieving stable alignment and precise detection.

1. Introduction

Anomaly detection aims to detect and localize regions in images that deviate from normal patterns, playing a pivotal role in industrial inspection, medical imaging, and autonomous driving [35]. Unlike conventional image understanding tasks [9, 32], anomaly detection focuses on identifying abnormal patterns that are challenging or impossible to exhaustively enumerate during training, thus placing stringent requirements on model generalization [27]. In recent years, with the rapid development of deep learning [8, 21], anomaly detection methods have evolved from relying on large-scale fully supervised data to semi-supervised [43], weakly supervised [48], and ultimately zero-shot learning approaches [26]. Among these, zero-shot anomaly detection is particularly compelling because it does not require collecting abnormal samples for the target domain and still aims to recognize never-before-seen anomalies. However, achieving precise localization of diverse and complex anomalies without any labeled abnormal data remains a

*Corresponding author.

formidable challenge.

Early anomaly detection methods primarily relied on reconstruction or distribution modeling to separate normal and abnormal regions. Autoencoders [3] and GANs [15] were commonly trained to reconstruct only normal samples, and large reconstruction errors indicated potential anomalies. Distribution-based approaches, such as One-Class SVM [24] and PatchCore [16], focused on modeling the feature space of normal samples and detecting deviations from this distribution. With the development of deep learning, methods such as PaDiM [12] advanced this paradigm by modeling per-pixel embeddings with multivariate Gaussian statistics, enabling more fine-grained localization of subtle anomalies.

Alongside these developments, large-scale pretrained models began showing their strength in zero-shot applications [38], providing a new avenue for anomaly detection. One particularly impactful foundation model is CLIP [36], which jointly learns image and text representations by training on a massive corpus of image–text pairs. CLIP's zero-shot advantage stems from its ability to align visual and linguistic concepts in a shared latent space, enabling tasks like image classification [1] or even segmentation [39] hints without specialized retraining. Recently, methods such as AnomalyCLIP [45] adopted CLIP to match image features against text prompts describing potential anomalies, whereas VCP-CLIP [35] extended this by incorporating learnable visual prompts to further refine both image and text encoders. While these methods advance vision–language anomaly detection by improving prompt design and multimodal alignment, key challenges remain in achieving stable, fine-grained correspondence between text and vision under domain shifts and complex visual conditions.

Despite the promising progress brought by recent vision–language models, the interaction between the two modalities in zero-shot anomaly detection remains loosely organized. Most existing approaches rely on implicit attention or heuristic prompt fusion, which only loosely couple visual and textual representations, as illustrated in Figure 1 (a). Such late-stage fusion often lacks structural constraints, making the learned correspondence fragile under distribution shifts and resulting in unstable image-level reasoning and coarse anomaly localization. To address these challenges, we draw inspiration from physical systems, particularly how particles interact and stabilize within energy fields, and reinterpret multimodal interaction through this perspective, as shown in Figure 1 (b). Instead of heuristic fusion, we model multimodal alignment as an energy equilibrium process within a latent potential field, where visual and textual features dynamically attract and balance each other. This formulation promotes stable and fine-grained semantic correspondence across modalities through the interplay of attraction and equilibrium forces.

Building upon this motivation, we propose the **S**emantic **G**raviton **Net**work (**SGNet**), a physics-inspired framework that models cross-modal interaction as an adaptive potential field. At its core lies the concept of semantic gravitons, a set of learnable tokens that act as dynamic mediators between visual and textual modalities. Each graviton adaptively absorbs linguistic cues and aligns with corresponding visual features, forming localized semantic equilibria that bridge the two modalities in a structured manner. To realize this interaction, we design a Graviton Interaction Network, which alternates text-to-graviton and vision-to-graviton coupling to progressively refine multimodal representations. This process enables stable semantic binding and fine-grained correspondence across modalities. Furthermore, an energy-based potential regularization is introduced to govern the interaction dynamics through attraction and equilibrium forces, ensuring balanced and interpretable multimodal evolution. A graviton-guided prediction head then integrates the aligned features to generate both image-level anomaly scores and pixel-wise anomaly maps in a single forward pass. Together, these components allow SGNet to achieve robust, precise, and interpretable zero-shot anomaly detection across diverse domains.

The main contributions of this work are as follows: (1) We propose the Semantic Graviton Network (SGNet), a physics-inspired framework that models multimodal alignment as an adaptive potential field for zero-shot anomaly segmentation. (2) We introduce semantic gravitons and a unified regularization mechanism that stabilizes cross-modal alignment through attraction and equilibrium forces, yielding fine-grained and interpretable correspondence. (3) Extensive experiments on ten industrial and medical benchmarks show that SGNet delivers state-of-the-art performance with strong generalization and precise localization.

2. Related Work

2.1. Vision-Language Modeling

Vision–language modeling has emerged as a powerful paradigm for learning unified representations across images and natural language [10, 34, 42, 47]. Early systems typically relied on carefully designed multi-stage architectures for captioning, retrieval, or question answering, where textual grounding was achieved through explicit supervision and task-specific design [25, 29, 41]. With the rise of large-scale pretraining, models such as CLIP [36], ALIGN [20], BLIP [23], and ALBEF [22] shifted the field toward contrastive or generative objectives over millions of image–text pairs. These models establish a shared embedding space where visual and linguistic semantics can be aligned in a highly flexible, open-vocabulary manner [33]. Recent efforts have adapted vision–language models to

zero-shot anomaly detection (ZSAD), leveraging CLIP's shared image–text space to express normality and abnormality through prompts [14, 35, 45]. Methods typically refine this alignment via prompt tuning, contextual cues, or patch-level similarity, improving sensitivity to subtle defects. Yet, because VLMs are pretrained for global matching rather than spatial reasoning, their cross-modal interaction remains weakly structured, making localization dependent on heuristic designs and vulnerable to distribution shifts. To address this limitation, we seek a more principled formulation that organizes visual–textual interaction into a stable and fine-grained alignment suitable for ZSAD.

2.2. Zero-shot Anomaly Detection

Zero-shot anomaly detection (ZSAD) has increasingly benefited from large vision–language models such as CLIP [36], which allow models to distinguish normality from abnormality using natural-language descriptions without requiring abnormal training samples. Early CLIP-based approaches like AnomalyCLIP [45] follow a global image–text matching strategy, comparing visual embeddings with prompts that describe normal or defective states. While effective for image-level detection, this global matching offers limited spatial guidance and often struggles to localize fine-grained anomalies. Subsequent works move toward more localized and expressive cross-modal interactions. Patch-level methods such as WinCLIP [17] compute region-wise similarities to enhance anomaly localization, while prompting-based approaches like VCP-CLIP [35] inject visual context into the text space to improve the stability of multimodal alignment. More recent efforts, represented by FE-CLIP [14], enhance the CLIP visual encoder itself through frequency-aware adapters to better capture subtle defect cues. This progression reflects a shift from global similarity matching, to localized reasoning, and finally to representation-level enhancement within the vision–language framework. Despite these advances, existing CLIP-based approaches still lack a principled mechanism to organize and stabilize the interaction between vision and language. Motivated by this gap, we introduce a physics-inspired formulation that models multimodal alignment through an adaptive semantic potential field, enabling learnable semantic gravitons to establish structured and fine-grained correspondence for zero-shot anomaly detection.

3. Method

3.1. Overview

We propose a Semantic Graviton Network (SGNet) that models multimodal alignment as an energy-balancing process within a shared potential field. As shown in Figure 2, given an input image I and a pair of textual prompts, the visual encoder extracts feature maps $F_0 \in \mathbb{R}^{C_v^0 \times H_0 \times W_0}$, while the text encoder produces the embeddings $L = \{L_{nor}, L_{abn}\} \in \mathbb{R}^{2 \times C_l}$, corresponding to the normal and abnormal text prompts. A set of learnable semantic gravitons $G_0 \in \mathbb{R}^{N \times C_l}$ is initialized to represent N latent semantic attractors bridging these two textual states and the visual evidence. Through a graviton interaction network with K stages, each graviton alternately absorbs linguistic cues from $\{L_{nor}, L_{abn}\}$ and visual patterns from F_i, gradually forming a balanced semantic potential field that captures both normality and deviation. The resulting gravitons G_i and fused features X are integrated for anomaly prediction. Specifically, the learned graviton field provides energy-based modulation over the visual embeddings, leading to stable alignment between normal/abnormal semantics and local image regions.

3.2. Graviton Interaction Network

The graviton interaction network (GIN) serves as the core component of SGNet, governing the hierarchical evolution of the learnable semantic gravitons. Each graviton $G_i \in \mathbb{R}^{N \times C_l}$ is regarded as a potential well that adaptively aggregates multimodal cues from text and vision, progressively forming a balanced semantic field through iterative interaction.

At stage i, each graviton interacts with the textual embeddings L through a cross-attention mechanism, acquiring linguistic priors that serve as semantic attractors. The linguistic activation is obtained as:

$$\mathrm{Att}_i^L = \frac{\mathrm{Proj}_g(G_i)\,[\mathrm{Proj}_l(L)]^\top}{\sqrt{C_l}}, \tag{1}$$

$$G_i^L = \mathrm{Norm}\big(G_{i-1} + \mathrm{Softmax}(\mathrm{Att}_i^L)\,\mathrm{Proj}_l(L)\big), \tag{2}$$

where $\mathrm{Proj}_g(\cdot)$ and $\mathrm{Proj}_l(\cdot)$ denote linear projections, and $\mathrm{Norm}(\cdot)$ represents layer normalization. To avoid excessive linguistic bias, we introduce a lightweight text-to-graviton gate that regulates the information flow:

$$G_i^{\mathrm{cross}} = \mathrm{Linear}\big(\gamma(G_i^L) \odot G_i^L + G_{i-1}\big), \tag{3}$$

where $\gamma(\cdot)$ is a two-layer perceptron with ReLU and Tanh activations, and $\odot$ denotes element-wise multiplication. This gating module dynamically rescales the semantic energy injected from the text domain, ensuring that only the most relevant linguistic cues contribute to the graviton potential field.

Given the cross-modally activated gravitons G_i^{cross}, we further incorporate visual grounding through bidirectional attention using the visual feature from the previous stage. Specifically, the fused visual representation F_{i-1} serves as the spatial input that interacts with the gravitons at stage i. Using F_{i-1} enables the gravitons to progressively refine

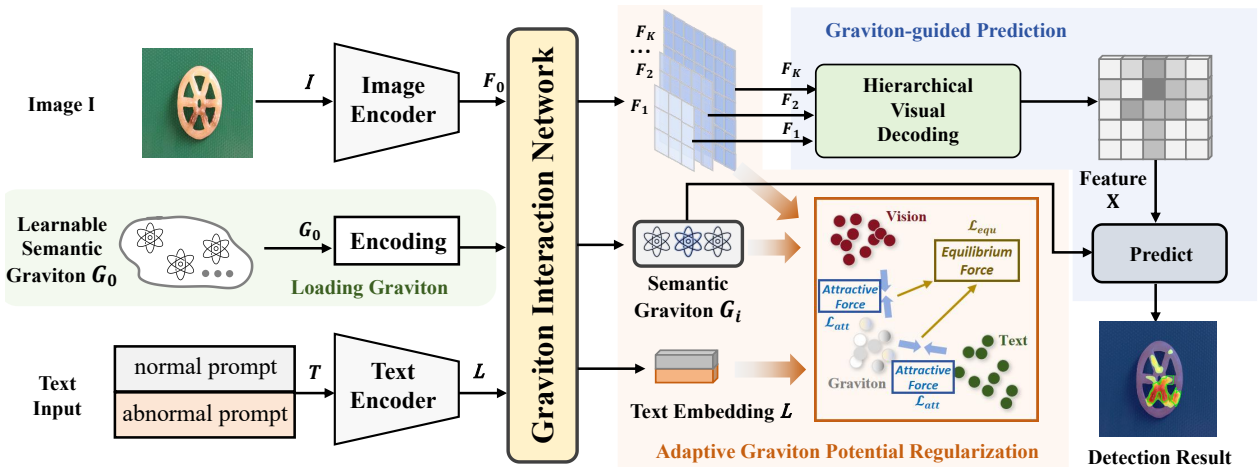

Figure 2. Overview of the proposed Semantic Graviton Network (SGNet). An input image and text prompts are encoded into visual and textual embeddings, which interact with a set of learnable semantic gravitons through the Graviton Interaction Network. Adaptive Graviton Potential Regularization aligns modalities via attraction and equilibrium forces. The refined multi-level visual features are decoded through hierarchical visual fusion, and the gravitons guide the final prediction to produce accurate anomaly detection results.

multimodal alignment in a stage-wise manner, since the feature already encodes the accumulated interactions from earlier layers. The visual attention map is computed as:

$$\mathrm{Att}_i^V = \frac{\mathrm{Proj}_g(G_i^{\mathrm{cross}})\,\mathrm{Flatten}(\mathrm{Proj}_v(F_{i-1}))}{\sqrt{C_v^{i-1}}}. \quad (4)$$

The graviton-updated representation and the fused visual feature at this stage are then obtained as:

$$G_i^V = \mathrm{Softmax}(\mathrm{Att}_i^V)\left[\mathrm{Flatten}(\mathrm{Proj}_v(F_{i-1}))\right]^\top, \quad (5)$$

$$F_i = \mathrm{Norm}\Big(F_{i-1} + \mathrm{Unflatten}\big(\mathrm{Proj}_g(G_i^{\mathrm{cross}})\big)^\top \cdot \mathrm{Softmax}(\mathrm{Att}_i^V)\Big). \quad (6)$$

where $\mathrm{Unflatten}(\cdot)$ restores spatial structure. This update enables the gravitons to absorb increasingly fine-grained visual cues while also modulating the evolving visual representation. Through repeated refinement across stages, the gravitons form a stable semantic bridge that supports subsequent decoding and anomaly prediction.

After completing both linguistic and visual interactions, the gravitons of the next stage inherit multimodal knowledge via:

$$G_i = \mathrm{Norm}\big(G_i^{\mathrm{cross}} + \mathrm{Proj}(G_i^V)\big). \quad (7)$$

This feedback propagation allows higher layers to operate on increasingly stable semantic equilibria, with each graviton representing an evolved semantic attractor that integrates text–vision alignment under the gravitational potential hypothesis.

3.3. Adaptive Graviton Potential Regularization

Each semantic graviton is regarded as a learnable attractor that defines a local potential field over multimodal representations. Instead of globally coupling every graviton to all visual and textual features, we introduce an adaptive mechanism that allows each graviton to selectively influence its semantically relevant regions through attention-guided weighting. This formulation enables the graviton field to evolve into a structured potential landscape, where local energy basins emerge to represent distinct semantic equilibria across modalities.

For the n-th graviton g_n, its relevance to each modality feature is determined by an attention-derived association weight. Let f_v and f_t denote the visual and textual embeddings interacting with the gravitons at each stage. The association coefficients are computed as:

$$a_{v,n} = \frac{\exp(\mathrm{sim}(f_v, g_n)/\tau)}{\sum_m \exp(\mathrm{sim}(f_v, g_m)/\tau)}, \quad (8)$$

$$a_{t,n} = \frac{\exp(\mathrm{sim}(f_t, g_n)/\tau)}{\sum_m \exp(\mathrm{sim}(f_t, g_m)/\tau)}, \quad (9)$$

where $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity and τ controls the sharpness of attention. These coefficients define the responsibility of each graviton for the two modalities, ensuring that only semantically aligned gravitons exert attraction on the corresponding features.

Attraction Potential Force. For each graviton, we define modality-specific energy distributions as $p_v^{(n)} = \mathrm{Softmax}(-\|f_v - g_n\|_2^2)$, $p_t^{(n)} = \mathrm{Softmax}(-\|f_t - g_n\|_2^2)$,

and align them using the 2-Wasserstein distance:

$$\mathcal{L}_{att} = \frac{1}{N}\sum_{n=1}^{N}(a_{v,n} + a_{t,n})\,\mathcal{W}_2\big(p_v^{(n)},\, p_t^{(n)}\big). \tag{10}$$

These energy-induced distributions can be viewed as modality-specific potential fields formed around each graviton. $\mathcal{L}_{att}$ encourages the visual and textual modalities to form isomorphic potential basins around each graviton, aligning not only their positions but also the shape and curvature of their semantic fields. As a result, both modalities converge toward a shared interpretation of each graviton's semantic role, enhancing cross-modal coherence and stabilizing multimodal alignment under distribution shifts.

Equilibrium Potential Force. Beyond topological alignment, we further ensure that the energy magnitudes of the two modalities remain balanced. We measure the free-energy discrepancy between their potential responses:

$$\mathcal{L}_{equ} = \frac{1}{N}\sum_{n=1}^{N}(a_{v,n} + a_{t,n})\,\big|\|f_v - g_n\|_2^2 - \|f_t - g_n\|_2^2\big|. \tag{11}$$

This equilibrium term prevents either modality from dominating the shared potential space, preserving stable multimodal equilibrium while maintaining semantic diversity among gravitons.

The final potential-based regularization combines attraction and equilibrium forces:

$$\mathcal{L}_{grav} = \lambda_{att}\mathcal{L}_{att} + (1 - \lambda_{att})\,\mathcal{L}_{equ}. \tag{12}$$

Through adaptive attention and dual-level energy regulation, each Semantic Graviton develops its own semantic region of influence, forming a self-organized and balanced multimodal potential field.

3.4. Graviton-guided Prediction

After the hierarchical vision decoding produces the feature map X, the semantic gravitons from the final stage are used as adaptive classifiers to generate the anomaly mask.

Hierarchical Visual Decoding. Instead of aggregating the gravitons into a single vector, each final-stage graviton $g_n \in G_K$ acts as an individual classifier. We project each graviton through an MLP to obtain a channel-wise scoring vector:

$$w_n = \mathrm{MLP}(g_n), \quad n = 1, \ldots, N. \tag{13}$$

Each graviton classifier interacts with the decoded feature X to produce a graviton-specific anomaly response: $\hat{M}_n = w_n X^\top$. The final anomaly mask is obtained by averaging the responses of all gravitons: $\hat{M} = \frac{1}{N}\sum_{n=1}^{N}\hat{M}_n$, followed by a sigmoid activation to obtain pixel-wise anomaly probabilities. This formulation allows different gravitons to focus on complementary semantic cues while producing a single coherent mask. A learnable class token interacts with the graviton set through attention, and the resulting embedding is fed to a linear layer with sigmoid activation to obtain the image-level anomaly score $\hat{S}$.

3.5. Training Objective

Let S_{gt} denote the ground-truth image label, where 1 represents an abnormal image and 0 a normal one. The anomaly classification loss is formulated as:

$$\mathcal{L}_{cls} = -\big[S_{gt}\log(\hat{S}) + (1 - S_{gt})\log(1 - \hat{S})\big]. \tag{14}$$

In addition, the segmentation loss $\mathcal{L}_{seg}$ combines focal and dice terms to emphasize boundary precision. The overall optimization objective combines graviton potential regularization, anomaly classification, and segmentation constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{cls} + \lambda_{grav}\mathcal{L}_{grav}. \tag{15}$$

4. Experiments

In this section, we will first introduce the experiment setup of our work in Section 4.1, including the used datasets, implementation details and metrics. Then, we further analyze the main results, including quantitative evaluations and qualitative results in Section 4.2 and ablate the effectiveness of our method in Section 4.3.

4.1. Experiment Setup

Datasets. Following the conventional practice in zero-shot anomaly detection research [14, 35, 46], we conduct comprehensive experiments on ten real-world anomaly detection datasets, encompassing both industrial inspection and medical imaging scenarios to evaluate the generalization ability of our model. For industrial inspection, we adopt MVTec-AD [4], VisA [49], MPDD [18], BTAD [31], DAGM [40], and DTD-Synthetic [2], which cover diverse types of defects including texture inconsistency, surface contamination, and structural deformation. For medical imaging, we employ CVC-ClinicDB [5] and Kvasir [19] for colon polyp detection, and BrainMRI [37] and Br35H [11] for brain tumor segmentation, offering a broad coverage of clinically meaningful anomaly types.

Implementation Details. The publicly available CLIP (ViT-L/14-336) is adopted as the backbone in this study. Following previous works [14, 35, 46], we standardize the input image resolution to 518×518 for all datasets. The GIN layer number K is set to 4, and the graviton number N is set to 20. We set the overall gravitational strength λ_{grav} as 0.6, and the ratio between attraction and equilibrium forces λ_{att} as 0.6, the effects of these hyperparameters are analyzed in detail in our ablation studies. We employ the AdamW optimizer [28] with a weight decay of 0.05. The initial learning rate is set to 0.00005 and scheduled using polynomial

Table 1. Image-level ZSAD results (AUROC) on 8 anomaly detection datasets. The best results are highlighted in **bold**.

Method	Industrial Defects						Medical Anomalies	
	MVTec-AD	VisA	MPDD	BTAD	DAGM	DTD-Synthetic	BrainMRI	Br35H
CLIP[ICML'21] [36]	74.1	66.4	54.3	34.5	79.6	71.6	73.9	78.4
CoOp[IJCV'22] [44]	88.8	62.8	55.1	66.8	87.5	-	61.3	86.0
WinCLIP[CVPR'23] [17]	91.8	78.8	63.6	68.2	91.8	93.2	86.6	80.5
APRIL-GAN[arxiv'23] [7]	86.2	78.0	78.0	73.6	94.4	86.4	89.3	93.1
AnoVL[CVPR'23] [13]	92.5	79.2	68.9	80.3	-	-	-	-
AnomalyCLIP[ICLR'24] [46]	91.5	82.1	77.0	88.3	97.5	93.5	90.3	94.6
AdaCLIP[ECCV'24] [6]	89.2	85.8	76.0	88.6	99.1	95.5	94.8	97.7
AA-CLIP[CVPR'25] [30]	92.0	84.6	78.3	**94.8**	97.8	97.5	84.3	94.2
FE-CLIP[ICCV'25] [14]	91.9	84.6	78.0	90.3	97.5	98.3	94.8	96.8
SGNet (ours)	**93.5**	**85.9**	**80.8**	92.9	**99.2**	**98.7**	**96.4**	**97.8**

Table 2. Pixel-level ZSAD results (AUROC) on 8 anomaly detection datasets. The best results are highlighted in **bold**.

Method	Industrial Defects						Medical Anomalies	
	MVTec-AD	VisA	MPDD	BTAD	DAGM	DTD-Synthetic	CVC-ClinicDB	Kvasir
CLIP[ICML'21] [36]	38.4	46.6	62.1	30.6	30.6	33.9	47.5	44.6
CoOp[IJCV'22] [44]	33.3	24.2	15.4	15.4	17.5	-	17.5	44.1
WinCLIP[CVPR'23] [17]	85.1	79.6	79.6	72.7	87.6	83.9	51.2	69.7
APRIL-GAN[arxiv'23] [7]	87.6	94.2	94.1	60.8	82.4	95.3	80.5	75.0
AnomalyCLIP[ICLR'24] [46]	91.1	95.5	96.5	94.2	95.6	97.9	82.9	78.9
AdaCLIP[ECCV'24] [6]	88.7	95.5	96.1	92.1	91.5	97.9	84.4	-
VCP-CLIP[ECCV'24] [35]	92.0	95.7	96.2	94.1	99.4	98.0	-	-
AA-CLIP[CVPR'25] [30]	91.9	95.5	96.7	97.0	98.8	98.9	**89.9**	87.2
FE-CLIP[ICCV'25] [14]	92.6	95.9	97.0	95.6	98.5	99.0	84.5	79.8
SGNet (ours)	**92.8**	**95.9**	**97.5**	**97.2**	**99.5**	**99.3**	88.5	**87.6**

decay with a power of 0.9. All models are trained for 10 epochs with a batch size of 32. To ensure a fair zero-shot evaluation, we follow the cross-dataset fine-tuning protocol used in previous studies: we fine-tune the model on the test split of MVTec AD and evaluate the Zero-Shot Anomaly Detection (ZSAD) performance on the remaining datasets. Conversely, for MVTec AD, we fine-tune the model on the test split of VisA, thereby avoiding dataset overlap while maintaining domain relevance.

Evaluation Metric. To ensure fair comparison with prior works, we adopt the Area Under the Receiver Operating Characteristic Curve (AUROC) as the evaluation metric.

4.2. Main Results

Quantitative Evaluations. Tables 1 and 2 summarize the quantitative results on eight zero-shot anomaly detection benchmarks, covering both industrial defects and medical anomalies. Across all datasets, our SGNet achieves consistent and substantial improvements over existing vision-language baselines. At the image level (Table 1), SGNet attains the highest AUROC on nearly all datasets, yielding a clear performance gain compared to the previous best model. At the pixel level (Table 2), SGNet demonstrates even stronger advantages. The proposed graviton mechanism enables precise localization of fine-grained anomalies. This confirms that the graviton-guided alignment not only enhances semantic consistency across modalities but also facilitates spatially coherent anomaly segmentation. Overall, these improvements validate the effectiveness of the proposed gravitational potential regularization in achieving more stable and discriminative multimodal representations, thereby boosting both detection performance under zero-shot settings.

Qualitative Results. Figure 3 presents qualitative results across diverse industrial objects and defect patterns. SGNet produces anomaly activations that remain spatially coherent and concentrated around truly abnormal regions, even under subtle defects or cluttered textures. This stability can be attributed to the semantic gravitons, whose potential-field interaction encourages the model to evolve toward a balanced

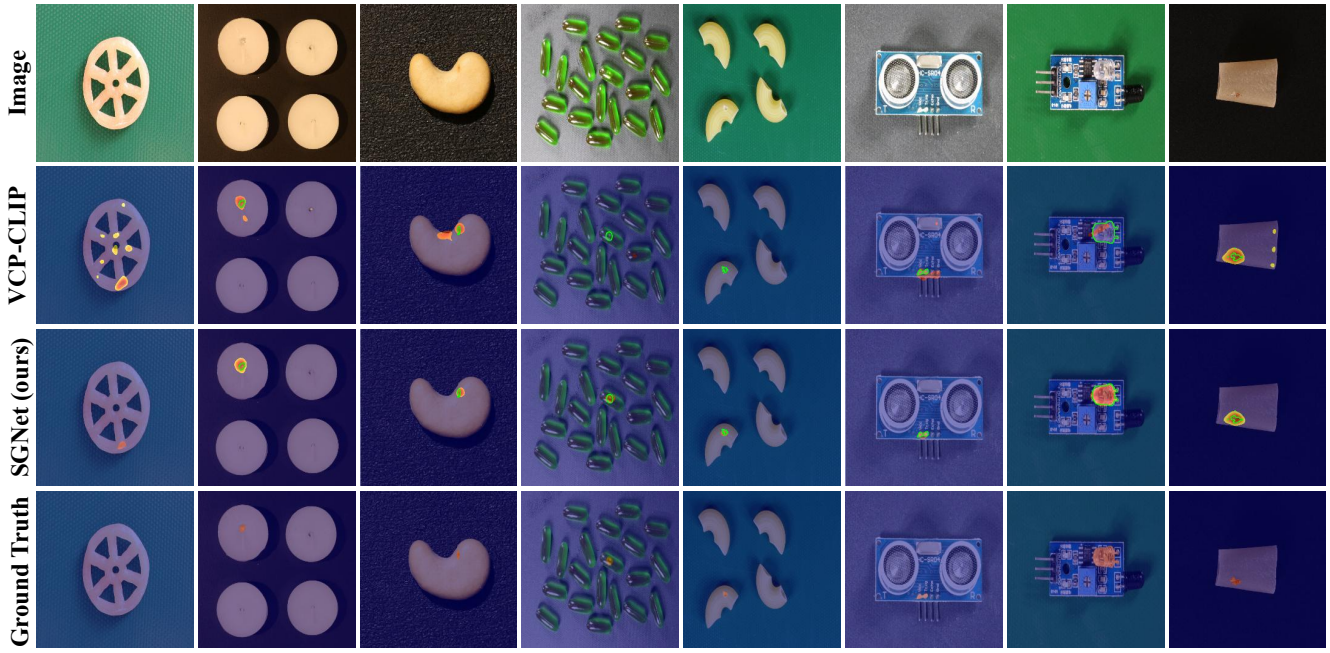

Figure 3. Qualitative comparison on anomaly localization. Green contours denote the ground-truth anomalous regions, while red activations indicate the predicted anomaly masks. SGNet is able to produce accurate and focused localization across diverse categories.

multimodal equilibrium, analogous to how physical systems settle into low-energy stable states. Such alignment allows SGNet to localize anomalies with precision while maintaining robustness across domains, consistent with the quantitative trends observed in our experiments.

4.3. Ablation Study

To better validate the superiority of our method, we conduct ablation experiments on the MVTec-AD and Visa datasets.

How Does the Graviton Interaction Network Enhance Multimodal Alignment? To understand the contribution of the Graviton Interaction Network (GIN), we investigate its role in bridging the visual and textual modalities. As shown in Table 3, without this module, the interaction between modalities remains coarse and globally entangled, leading to weak localization and unstable semantic correspondence. In contrast, introducing gravitons as learnable semantic mediators enables localized and structured communication between vision and language. Each graviton adaptively captures a distinct semantic subspace and acts as an energy node that selectively aligns modality-specific features based on contextual relevance. This design transforms the alignment process from a static global fusion into a dynamic equilibrium interaction, where visual features are gradually drawn toward their semantically consistent textual anchors. As a result, the model achieves more fine-grained alignment and reduced cross-modal ambiguity, which ultimately improves anomaly sensitivity and spatial interpretability.

Why Is Graviton Potential Regularization Important?

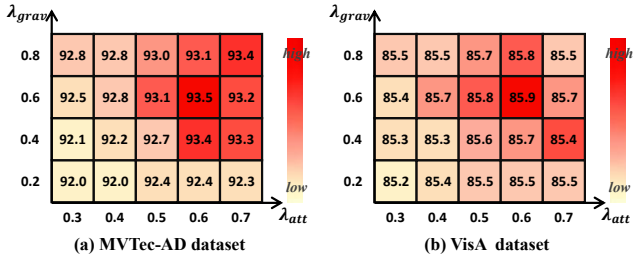

Figure 4. Sensitivity analysis of λ_{grav} and λ_{att} on the MVTec-AD and VisA datasets. Image-level AUROC is reported.

We analyze the effect of graviton potential regularization (GPR), which introduces a physics-inspired energy constraint to stabilize multimodal convergence. As shown in Figure 5, removing GPR leads to scattered or exaggerated activations that either highlight normal textures or miss true anomalies, reflecting a collapse of the graviton field toward visual dominance or textual drift. This behavior is also supported by Table 3, where removing GPR results in clear performance degradation. GPR alleviates this issue by encouraging visual and textual features to gather around shared semantic centers while maintaining balanced energy between modalities. These complementary forces form a self-regulated potential landscape that enhances cross-modal stability and produces more compact and accurate anomaly responses.

How Do the Gravitational Weights Influence Model Be-

Table 3. Ablation of main components of our proposed SGNet on the MVTec-AD and VisA datasets.

Graviton Interaction	Graviton Potential Regularization		MVTec-AD [4]		VisA [49]	
	Attraction Force	Equilibrium Force	Image-level	Pixel-level	Image-level	Pixel-level
✗	✗	✗	91.1	91.8	84.2	95.1
✓	✗	✗	91.8	92.1	85.1	95.4
✓	✓	✗	92.2	92.6	85.2	95.8
✓	✗	✓	92.7	92.5	85.7	95.6
✓	✓	✓	**93.5**	**92.8**	**85.9**	**95.9**

Table 4. Ablation study on the number of Semantic Gravitons N.

Graviton Number N	MVTec-AD [4]		VisA [49]	
	Image-level	Pixel-level	Image-level	Pixel-level
10	92.1	92.4	85.1	95.3
15	93.2	92.7	85.7	95.8
20	**93.5**	**92.8**	**85.9**	**95.9**
25	93.3	92.7	85.6	95.9
30	93.1	92.4	85.5	95.8

havior? We further analyze the impact of the gravitational regularization weights, including the overall potential strength λ_{grav} and the attraction coefficient λ_{att}, where $(1 - \lambda_{\text{att}})$ implicitly controls the equilibrium force. Figure 4 shows the performance variations on the MVTec-AD and VisA datasets. As λ_{grav} increases, the potential field becomes more structured, enforcing consistent vision–language alignment. However, overly large values may over-constrain the optimization, reducing the adaptability of gravitons. A moderate λ_{grav} achieves the best trade-off between structural regularization and semantic flexibility. Similarly, λ_{att} governs the balance between attraction and equilibrium forces. Smaller values yield smoother alignment, while larger ones emphasize convergence toward semantic centers. An intermediate setting establishes an adaptive equilibrium that harmonizes attraction and balance. Overall, the performance remains stable across a wide range of hyperparameter configurations, indicating that the proposed gravitational formulation exhibits inherent self-stabilization and low sensitivity to hyperparameter tuning.

How Does the Number of Semantic Gravitons Influence Performance? The number of semantic gravitons determines how finely the multimodal space is partitioned into semantic attractors. With too few gravitons, the potential field becomes overly coarse, limiting the ability to capture diverse anomaly patterns and weakening semantic alignment. As the number increases, each graviton can specialize in a more distinct semantic region, leading to stronger cross-modal correspondence and more accurate anomaly reasoning. However, excessively large graviton sets introduce redundancy and dilute attention, causing overlapping semantic roles and reduced discrimination. This also suggests

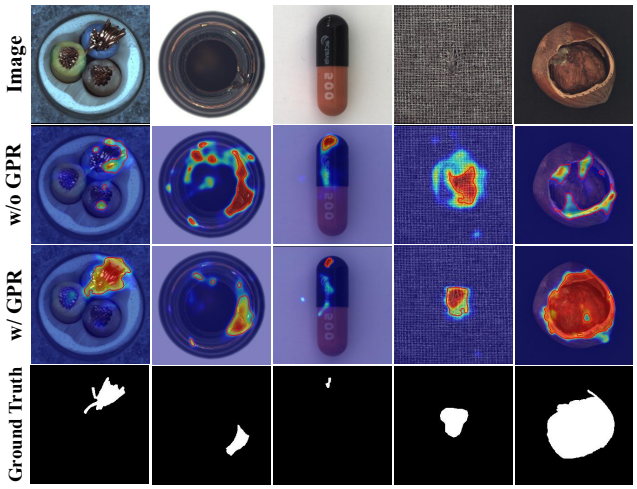

Figure 5. Qualitative effect of Graviton Potential Regularization.

SGNet is not overly sensitive to the choice of N, demonstrating stable behavior across a broad range of configurations.

5. Conclusion

We present Semantic Graviton Network (SGNet), a physics-inspired framework for zero-shot anomaly detection and segmentation. By introducing learnable semantic gravitons for vision–language alignment, SGNet enables stable multimodal convergence and reliable anomaly prediction. It achieves robust anomaly localization and strong generalization across diverse benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62425117 and 62293544, and in part by China Postdoctoral Science Foundation under Grant 2024T170493, 2025M773474, 2025M773484, and in part by the Postdoctoral Innovation Talents Support Program under Grant BX20240178, BX20250413.

References

[1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1348–1357, 2023. 2

[2] 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*, pages 5564–5572, 2023. 5

[3] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023. 2

[4] 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. 5, 8

[5] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[6] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 6

[7] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 6

[8] ZHIXIN CHENG, JIACHENG DENG, XINJUN LI, BAOQUN YIN, and TIANZHU ZHANG. Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2491–2499, 2025. 1

[9] Zhixin Cheng, Jiacheng Deng, Xinjun Li, Xiaotian Yin, Bohao Liao, Baoqun Yin, Wenfei Yang, and Tianzhu Zhang. Ca-i2p: Channel-adaptive registration network with global optimal selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27739–27749, 2025. 1

[10] Zhixin Cheng, Xiaotian Yin, Jiacheng Deng, Bohao Liao, Yujia Chen, Xu Zhou, Baoqun Yin, and Tianzhu Zhang. Adaptive agent selection and interaction network for image-to-point cloud registration. *arXiv preprint arXiv:2511.05965*, 2025. 2

[11] Falak Chhatre, Sudhanva Deshpande, Sidhant Malik, Grace Yan, and Suresh Subramaniam. Brain tumor detection using convolutional neural network. *J. Stud. Res*, 12, 2023. 5

[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. 2

[13] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2(5), 2023. 6

[14] Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21220–21230, 2025. 3, 5, 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[16] Kosaburo Hattori, Tomonori Izumi, and Lin Meng. Defect detection of apples using patchcore. In *2023 International Conference on Advanced Mechatronic Systems (ICAMechS)*, pages 1–6. IEEE, 2023. 2

[17] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 3, 6

[18] 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)*, pages 66–71. IEEE, 2021. 5

[19] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 5

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[24] Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. Improving one-class svm for anomaly detection. In *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, pages 3077–3081. IEEE, 2003. 2

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer, 2020. 2

[26] Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1093–1102, 2024. 1

[27] Zeyi Li, Pan Wang, and Zixuan Wang. Flowgananomaly: Flow-based anomaly network intrusion detection with adversarial learning. *Chinese Journal of Electronics*, 33(1):58–71, 2024. 1

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2

[30] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025. 6

[31] 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)*, pages 01–06. IEEE, 2021. 5

[32] Yuwen Pan, Rui Sun, Naisong Luo, Tianzhu Zhang, and Yongdong Zhang. Exploring reliable matching with phase enhancement for night-time semantic segmentation. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 1

[33] Yuwen Pan, Rui Sun, Yuan Wang, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Purify then guide: A bi-directional bridge network for open-vocabulary semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):343–356, 2024. 2

[34] Yuwen Pan, Rui Sun, Yuan Wang, Tianzhu Zhang, and Yongdong Zhang. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2031–2040, 2024. 2

[35] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024. 1, 2, 3, 5, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3, 6

[37] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 5

[38] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20 (4):447–482, 2023. 2

[39] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2

[40] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007. 5

[41] Jie Xiao, Yinchao Ma, Yuyang Tang, Chuxin Wang, and Tianzhu Zhang. State space models for long-term temporal context in 3d single object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[42] Jie Xiao, Yinchao Ma, Wenfei Yang, and Tianzhu Zhang. Learning adaptive conceptual prototypes for 3d single object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6773–6786, 2025. 2

[43] Fan Zhou, Guanyu Wang, Kunpeng Zhang, Siyuan Liu, and Ting Zhong. Semi-supervised anomaly detection via neural process. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10423–10435, 2023. 1

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 6

[45] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*. 2, 3

[46] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 5, 6

[47] Qiongyi Zhou, Changde Du, Dan Li, Bincheng Wen, Le Chang, and Huiguang He. Interpretable visual neural decoding with unsupervised semantic disentanglement. *Machine Intelligence Research*, 22(3):553–570, 2025. 2

[48] Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (6):2454–2465, 2021. 1

[49] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022. 5, 8