TLM-SVDD: DUAL-DRIVEN DEEP LARGE-MARGIN ℓ_p-SVDD WITH VISION TRANSFORMERS FOR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep extensions of the Support Vector Data Description (SVDD) paradigm can jointly learn representations and decision boundaries for anomaly detection. Yet, they typically rely on simplified objectives or use a surrogate objective that does not coincide with the one used to fit the boundary. In this work, we introduce Transformer-based Large-Margin SVDD (TLM-SVDD), a framework that updates the encoder using the exact dual boundary problem to maximize the margin between normal and anomalous samples. By applying Danskin's theorem, we show that at the optimum of the convex dual function, the encoder's hypergradient can be directly derived from a kernel quadratic term evaluated at the optimal dual variables. As such, the encoder update requires no recovered decision thresholds, surrogate losses, or differentiation through a Frank-Wolfe-based solver. Furthermore, we show that this objective is bounded and naturally depends on support vectors. For batch-independent representation learning, we make use of a vision transformer and conduct a controlled ablation study to separately analyze the merits of the backbone architecture for anomaly detection from those of our optimization approach. Through extensive experiments across anomaly detection and long-tailed recognition benchmarks, the proposed TLM-SVDD approach is shown to improve upon convolutional and approximate primal-space predecessors, outperforming state-of-the-art methods in the majority of evaluation settings.

1 INTRODUCTION

Anomaly detection aims to compactly characterizing an observed normal class against an open set of anomalous samples whose test-time characteristics may differ arbitrarily from those seen in training. In practice, a few labeled anomalies may be available during the training stage, but they are typically scarce and are not fully representative of those encountered at test time. Under long-tailed data distributions, the normal class itself may also be severely under-represented. A detector must therefore describe normality tightly while benefiting from even limited anomaly exposure rather than learning a closed-set classifier that only separates the anomalies observed during the training stage.

A very successful group of anomaly detection approaches operate by inferring a boundary enclosing normal observations to discriminate them from any anomalous sample. Among boundary-based approaches, Support Vector Data Description (SVDD) (Tax & Duin, 2004) serves as a de-facto standard and represents normality as the region in a minimal-volume hyper-sphere (possibly in a kernel space). While there are different variants of this methodology Ruff et al. (2018; 2020); Liznerski et al. (2020), the large-margin ℓ_p-SVDD (Rahimzadeh Arashloo, 2024) extends it by explicitly maximizing a margin between normal and anomalous samples, penalizing violations through an ℓ_p-norm slack penalty, and solving the corresponding convex dual optimization task efficiently via the Frank–Wolfe approach (Jaggi, 2013). The more recent deep Large-Margin SVDD (DLM-SVDD) (Saei & Rahimzadeh Arashloo, 2026; Saei & Arashloo, 2026) method removed the fixed-feature restriction of (Rahimzadeh Arashloo, 2024) by alternating between the boundary fitting step and a convolutional neural network (CNN) update stage to optimize a primal margin-violation loss. Although this method achieved state-of-the-art performance outperforming other approaches in a majority of anomaly detection settings, nevertheless, it has certain limitations both in terms of learning and optimization as well as its representational capability. First, the network update step never sees the

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

exact objective function used to optimize the boundary: it employs boundary thresholds recovered from a handful of active support vectors which may change every epoch to define the error, and hence, the network chases a moving target. Furthermore, the softplus surrogate used to stabilize the encoder learning does not exactly match the original ℓ_p-norm slack penalty of the primal large-margin formulation. As the two alternating blocks of optimization use different objective functions, no single objective is guaranteed to decrease. Furthermore, a convolutional backbone is not a perfect host for a kernel-space boundary since batch normalization ties features to batch composition and operating mode (Ioffe & Szegedy, 2015). Consequently, the boundary may be fit on one feature map and differentiated through another. Additionally, post-ReLU features are non-negative and strongly anisotropic, making a single radial basis function (RBF) kernel bandwidth less suitable for their geometry. Last but not least, in standard convolutional networks, global configuration is assembled late, relying on slowly expanding receptive fields and local evidence prone to a texture bias (Geirhos et al., 2018). In contrast, the self-attention mechanism incorporated into transformer architectures makes global context available from the very first layers (Raghu et al., 2021).

In this work, we address these limitations by coupling representation learning directly to the objective used to fit the large-margin SVDD boundary. Rather than updating the encoder through recovered intermediate thresholds and an approximate surrogate loss, for encoder updating, we use the optimal value of the exact same function used in the boundary fitting stage. At the optimum of the dual boundary problem, Danskin's theorem (Mangasarian, 1994) provides the corresponding encoder gradient without requiring additional differentiation through the Frank–Wolfe iterations. As a result, the boundary and representation updates are tied to the same underlying optimization objective. The resulting update is threshold-free, requires no surrogate approximation, and depends only on the samples that support the learned hyper-sphere boundary in the kernel space. In terms of representation, we replace the CNN with a vision transformer (ViT) (Dosovitskiy et al., 2021). As such, layer normalization would be batch-independent, and hence, the boundary and encoder updates operate on the same feature map, unlike a batch-normalized CNN whose features depend on batch composition and train/evaluation operational mode (Ioffe & Szegedy, 2015). Self-attention also provides global interactions from the early layers and avoids relying exclusively on progressively enlarged local receptive fields (Raghu et al., 2021). These properties render the transformer a natural host for the proposed kernel-based update. The resulting Transformer-based Large-Margin SVDD (TLM-SVDD) approach retains the exact large-margin ℓ_p-SVDD boundary formulation of (Rahimzadeh Arashloo, 2024) while adapting its representation through the corresponding optimal-value function. We provide an envelope characterization of this update and a support-vector contrast identity that clarifies its behavior. Our main contributions may be summarized as follows. First, we propose an encoder update scheme based on the exact hypergradient of the optimal large-margin ℓ_p-SVDD objective. At the optimal point of the dual function, Danskin's theorem provides the gradient directly, without the need for differentiating through the Frank–Wolfe solver or introducing thresholds or a surrogate loss. Second, we identify desirable representational properties for applying this update, including batch-independent features and compatibility with a single Gaussian kernel bandwidth. Together with the ability of self-attention to capture global context, these properties motivate the use of a vision transformer with layer normalization and self-attention mechanisms. Third, we show that updating the encoder through the exact dual value allows additional anomaly exposure to improve the learned representation, whereas with a fixed representation, one may only refit the decision boundary. Fourth, the proposed TLM-SVDD framework is evaluated on multiple standard datasets and shown to achieve a superior performance in the majority of experiments under both the standard anomaly detection and long-tailed recognition protocols.

2 Related Work

Data description. The SVDD Tax & Duin (2004), and equivalently, one-class support vector machine (SVM) Schölkopf et al. (2001), enclose normal data in a compact region, with extensions possible to a kernel space. The ℓ_p-norm SVDD (Rahimzadeh Arashloo, 2022) generalizes the linear slack penalty of the conventional SVDD approach to a nonlinear one while the large-margin multiple-kernel ℓ_p-SVDD (Rahimzadeh Arashloo, 2024) adds explicit margin maximization and a Frank–Wolfe-based solver for the dual-based objective function (Jaggi, 2013). This formulation may be considered as the fixed-feature ancestor of the present work. Apart from these studies, there exist other deep variants of the conventional SVDD method that trade boundary fidelity for representation learning such as Deep SVDD (Ruff et al., 2018) and Deep-SAD (Ruff et al., 2020) that simplify the

objective function to a center-distance loss and DASVDD (Hojjati & Armanfard, 2023) that adds an autoencoder, and multi-sphere variants (Xing et al., 2024) that target collapse and multimodal normality. DLM-SVDD (Saei & Rahimzadeh Arashloo, 2026; Saei & Arashloo, 2026) retains the exact large-margin dual introduced in Rahimzadeh Arashloo (2024) in the hyper-sphere boundary learning step while adapting a CNN through a primal approximate loss. The present work addresses the limitations induced by both the approximate surrogate and the CNN backbone structure.

Learning through solving an inner problem. Training a network through a convex program that consumes its features is a bilevel task. Differentiable-optimization and hyperparameter-learning methods (Amos & Kolter, 2017; Franceschi et al., 2018) obtain the outer gradient by unrolling the solver or by implicit differentiation of its optimality conditions as their outer objective depends on the inner solution. Our outer objective is an affine transformation of the inner optimal value, so the envelope (Danskin) theorem for parametric convex programs (Danskin, 2012; Mangasarian, 1994; Bonnans & Shapiro, 2013) applies directly. That is, with a parameter-free feasible set and a network that affects the objective only through the kernel matrix, the hypergradient is obtained from the partial gradient at the optimal dual solution without unrolling or differentiating through the boundary solver. The resulting quadratic term is a support-vector-weighted kernel–target alignment (Cristianini et al., 2001), and the scheme bears similarities with deep kernel learning (Wilson et al., 2016) in terms of training the feature map through the kernel model's own optimal objective rather than through a detached loss.

Transformers for anomaly detection. Beyond the architectural properties noted in Sec. 1, vision transformers (Dosovitskiy et al., 2021; Touvron et al., 2021; Caron et al., 2021) have shown strong pretrained representations for out-of-distribution (OOD) detection as compared with conventional CNN baselines (Fort et al., 2021). Their fine-tuning is delicate in the sense that without large-scale data or heavy augmentation they may converge to local minima (Chen et al., 2021) while their loss landscape may degrade in the small-data regimes (Park & Kim, 2022), motivating the use of a bounded encoder objective and a full-batch update at a small learning rate. Existing transformer anomaly detection approaches rely on reconstruction or distillation (VT-ADL (Mishra et al., 2021), Transformaly (Cohen & Avidan, 2022)) or, similar to the representation-first methods PANDA (Reiss et al., 2021), MSC (Reiss & Hoshen, 2023), CSI (Tack et al., 2020), UNODE (Mirzaei et al., 2024), FIRM (Lunardi et al., 2025) and outlier exposure (Hendrycks et al., 2019), on post-hoc scoring of adapted features. To the best of our knowledge, these earlier transformer-based studies do not train the transformer directly against an explicit margin-maximizing kernel-based boundary.

3 METHODOLOGY

In this section, first, we briefly review the learning approach presented in Rahimzadeh Arashloo (2024), and then, discuss how it may be generalized to enable a direct dual-based learning of a large-margin transformer for improved anomaly detection. We consider one-vs-rest anomaly detection subject to limited anomaly exposure. For each task, the target-class training samples form the normal index set $\mathcal{I}_N$, while the available anomalous samples form index set $\mathcal{I}_A$. The labels are $t_i = +1$ for $i \in \mathcal{I}_N$ and $t_l = -1$ for $l \in \mathcal{I}_A$, collected in $\mathbf{t} \in \{\pm 1\}^n$. A base Vision Transformer with 16×16 patches (ViT-B/16) (Dosovitskiy et al., 2021) with parameters denoted as ω maps an input $\mathbf{x}$ to an ℓ_2-normalized class-token representation $\mathbf{z} = \phi_\omega(\mathbf{x}) \in \mathbb{R}^{768}$. Because layer normalization does not depend on batch statistics, the encoder is batch-independent and both optimization blocks proposed (discussed below) operate on the same representation. Our formulation applies to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with the corresponding Gaussian kernel function $k(\mathbf{z}_i, \mathbf{z}_j) = \exp(-\gamma \|\mathbf{z}_i - \mathbf{z}_j\|_2^2)$ and kernel matrix $\mathbf{K}_\omega$. For a fixed encoder, the large-margin ℓ_p-SVDD problem in (Rahimzadeh Arashloo, 2024) is

$$\begin{aligned} \min_{r,C,\epsilon,\rho} \; & r^2 - \nu\rho^2 + c_1 \sum_{i\in\mathcal{I}_N} \epsilon_i^p + c_2 \sum_{l\in\mathcal{I}_A} \epsilon_l^p \\ \text{s.t. } & \|\Phi(\mathbf{z}_i) - C\|_{\mathcal{H}_k}^2 \le r^2 - \rho^2 + \epsilon_i, \quad i \in \mathcal{I}_N, \\ & \|\Phi(\mathbf{z}_l) - C\|_{\mathcal{H}_k}^2 \ge r^2 + \rho^2 - \epsilon_l, \quad l \in \mathcal{I}_A, \qquad \epsilon \ge 0, \end{aligned} \tag{1}$$

where $p > 1$, $c_1 = \tilde{c}_1/|\mathcal{I}_N|$, and $c_2 = \tilde{c}_2/|\mathcal{I}_A|$. r denotes the radius of the hypersphere while ρ controls the margin. For a fixed encoder and a Gaussian kernel function, the corresponding dual

optimization problem of Eq. 1 is

$$\min_{\boldsymbol{\alpha}\in\mathcal{D}} F_\omega(\boldsymbol{\alpha}) = \bar{c}_1\|\boldsymbol{\alpha}\odot(\mathbf{1}+\mathbf{t})\|_q^q + \bar{c}_2\|\boldsymbol{\alpha}\odot(\mathbf{1}-\mathbf{t})\|_q^q + (\boldsymbol{\alpha}\odot\mathbf{t})^\top\mathbf{K}_\omega(\boldsymbol{\alpha}\odot\mathbf{t}), \tag{2}$$

where $\mathcal{D} = \{\boldsymbol{\alpha} \geq 0 : \mathbf{1}^\top\boldsymbol{\alpha} = \nu,\ \mathbf{t}^\top\boldsymbol{\alpha} = 1\}$, $q = p/(p-1)$, and $\bar{c}_{1,2} = \frac{1}{2}(c_{1,2}p)^{-1/(p-1)}\left(1-\frac{1}{p}\right)$. Note that only the kernel quadratic term in equation 2 depends on the encoder parameters ω.

3.1 Alternating Optimization

The proposed TLM-SVDD approach learns both the encoder as well as the boundary parameters through a novel alternating optimization approach. For this purpose, given the current encoder, the α-step solves the convex dual in equation 2. The resulting dual variables are then kept fixed while the ω-step updates the encoder through the gradient of the corresponding dual function.

3.1.1 α-step: Boundary Update

With ω fixed, equation 2 is solved via the Frank–Wolfe (Jaggi, 2013) algorithm. The constraints separate over the two classes into scaled simplices with $\mathbf{1}^\top\boldsymbol{\alpha}_{\mathcal{I}_N} = \frac{\nu+1}{2}$ and $\mathbf{1}^\top\boldsymbol{\alpha}_{\mathcal{I}_A} = \frac{\nu-1}{2}$, so the linear minimization oracle is a vertex placing all mass on one sample per class:

$$\mathbf{s}^{(\tau)} = \tfrac{\nu+1}{2}\mathbf{e}_k + \tfrac{\nu-1}{2}\mathbf{e}_m, \qquad \boldsymbol{\alpha}^{(\tau+1)} = (1-\lambda_\tau)\boldsymbol{\alpha}^{(\tau)} + \lambda_\tau\mathbf{s}^{(\tau)}, \tag{3}$$

with $k = \arg\min_{i\in\mathcal{I}_N}[\nabla F_\omega(\boldsymbol{\alpha}^{(\tau)})]_i$, $m = \arg\min_{l\in\mathcal{I}_A}[\nabla F_\omega(\boldsymbol{\alpha}^{(\tau)})]_l$, $\lambda_\tau = 2/(\tau+2)$, $\nabla F_\omega(\boldsymbol{\alpha}) = 2(\mathbf{K}_\omega(\boldsymbol{\alpha}\odot\mathbf{t}))\odot\mathbf{t} + \nabla_{\boldsymbol{\alpha}}[\ell_q \text{ terms}]$, and $\mathbf{e}_k$, and $\mathbf{e}_m$ denoting the standard basis vectors. Let $\mathbf{a}^\star = \boldsymbol{\alpha}^\star\odot\mathbf{t}$. The corresponding RKHS center would then be $C^\star = \sum_j a_j^\star\Phi(\mathbf{z}_j)$. For a sample with feature $\mathbf{z}$, the resulting dissimilarity score is then

$$f_\omega(\mathbf{x}) = \|\Phi(\mathbf{z}) - C^\star\|_{\mathcal{H}_k}^2 = 1 - 2\,\mathbf{k}(\mathbf{z})^\top\mathbf{a}^\star + \mathbf{a}^{\star\top}\mathbf{K}_\omega\mathbf{a}^\star, \tag{4}$$

where $\mathbf{k}(\mathbf{z}) = [k(\mathbf{z},\mathbf{z}_1),\ldots,k(\mathbf{z},\mathbf{z}_n)]^\top$. With $S_N = \{i\in\mathcal{I}_N : \alpha_i^\star > 0\}$, $S_A = \{l\in\mathcal{I}_A : \alpha_l^\star > 0\}$ and slacks $\epsilon_j = (\alpha_j^\star/(c_jp))^{1/(p-1)}$, where $c_j = c_1$ for $j\in\mathcal{I}_N$ and $c_j = c_2$ for $j\in\mathcal{I}_A$, complementary slackness yields (Rahimzadeh Arashloo, 2024):

$$r^{\star 2} = \tfrac{1}{2}(A_N + A_A), \quad 2\rho^{\star 2} = A_A - A_N, \tag{5}$$

$$A_N = \tfrac{1}{|S_N|}\sum_{i\in S_N}(f_\omega(\mathbf{x}_i) - \epsilon_i), \quad A_A = \tfrac{1}{|S_A|}\sum_{l\in S_A}(f_\omega(\mathbf{x}_l) + \epsilon_l). \tag{6}$$

3.1.2 ω-step: Encoder Update

The ω-step updates the encoder through the optimal value of the boundary problem inferred in the α-step while keeping $\boldsymbol{\alpha}$ fixed. Let us define

$$F^\star(\omega) = \min_{\boldsymbol{\alpha}\in\mathcal{D}} F_\omega(\boldsymbol{\alpha}), \qquad \mathcal{L}_\omega^{\text{dual}} = -\,\mathbf{a}^{\star\top}\mathbf{K}_\omega\mathbf{a}^\star. \tag{7}$$

The encoder is updated by minimizing $\mathcal{L}_\omega^{\text{dual}}$, or equivalently maximizing the kernel quadratic term $\mathbf{a}^{\star\top}\mathbf{K}_\omega\mathbf{a}^\star$ with respect to ω.

Proposition 1 (Envelope characterization)**.** *Under the strong-duality conditions of the large-margin ℓ_p-SVDD formulation (Rahimzadeh Arashloo, 2024), let $P^\star(\omega)$ denote the optimal value of the primal problem in equation 1 with respect to (r, C, ϵ, ρ) for fixed ω. Then $P^\star(\omega) = 1 - F^\star(\omega)$, and for $p > 1$, whenever $\phi_\omega(.)$ is differentiable, one has:*

$$\nabla_\omega P^\star(\omega) = \nabla_\omega\mathcal{L}_\omega^{\text{dual}}. \tag{8}$$

Proof. We have $P^\star(\omega) = \max_{\boldsymbol{\alpha}\in\mathcal{D}}\{\sum_j\alpha_jt_jk(\mathbf{z}_j,\mathbf{z}_j) - F_\omega(\boldsymbol{\alpha})\} = 1 - F^\star(\omega)$, using $k(\mathbf{z},\mathbf{z}) = 1$ and $\mathbf{t}^\top\boldsymbol{\alpha} = 1$. For $p > 1$ we have $q > 1$, so the ℓ_q terms are strictly convex and the minimizer $\boldsymbol{\alpha}^\star(\omega)$ is unique; since $\mathcal{D}$ is compact and independent of ω, Danskin's theorem (Mangasarian, 1994) yields

$$\nabla_\omega F^\star(\omega) = \nabla_\omega F_\omega(\boldsymbol{\alpha}^\star). \tag{9}$$

in which only the quadratic term depends on ω. Therefore,

$$\nabla_\omega F^\star(\omega) = \nabla_\omega\left[\mathbf{a}^{\star\top}\mathbf{K}_\omega\mathbf{a}^\star\right]. \tag{10}$$

Since $P^\star(\omega) = 1 - F^\star(\omega)$ and $\mathcal{L}_\omega^{\text{dual}} = -\mathbf{a}^{\star\top}\mathbf{K}_\omega\mathbf{a}^\star$, it follows that $\nabla_\omega P^\star(\omega) = \nabla_\omega\mathcal{L}_\omega^{\text{dual}}$. □

Thus, for an exact inner solution, the proposed encoder update is the hypergradient of the optimal primal large-margin ℓ_p-SVDD value. It requires neither recovered boundary thresholds nor differentiation through the Frank–Wolfe iterations.

Proposition 2 (Support-vector contrast). *For any $\boldsymbol{\alpha} \in \mathcal{D}$, $\mathbf{a} = \boldsymbol{\alpha} \odot \mathbf{t}$, and f_ω in equation 4,*

$$-\mathbf{a}^\top \mathbf{K}_\omega \mathbf{a} = \sum_{i \in \mathcal{I}_N} \alpha_i f_\omega(\mathbf{x}_i) - \sum_{l \in \mathcal{I}_A} \alpha_l f_\omega(\mathbf{x}_l) - 1. \tag{11}$$

Proof. By equation 4, $f_\omega(\mathbf{x}_j) = 1 - 2[\mathbf{K}_\omega \mathbf{a}]_j + \mathbf{a}^\top \mathbf{K}_\omega \mathbf{a}$ for a training sample; multiplying by a_j, summing, and using $\mathbf{1}^\top \mathbf{a} = \mathbf{t}^\top \boldsymbol{\alpha} = 1$ gives $\sum_j a_j f_\omega(\mathbf{x}_j) = 1 - \mathbf{a}^\top \mathbf{K}_\omega \mathbf{a}$. Splitting the sum by the sign of t_j yields equation 11. □

With $\boldsymbol{\alpha}^\star$ held fixed during the ω-step, Proposition 2 yields

$$\nabla_\omega \mathcal{L}_\omega^{\text{dual}} = \sum_{i \in S_N} \alpha_i^\star \nabla_\omega f_\omega(\mathbf{x}_i) - \sum_{l \in S_A} \alpha_l^\star \nabla_\omega f_\omega(\mathbf{x}_l). \tag{12}$$

Hence, only support vectors contribute to the encoder update. Normal support vectors are pulled toward the learned description, while anomalous support vectors are pushed away, with strengths modulated by their optimal dual variables. The encoder objective is bounded because $\sum_j \alpha_j^\star = \nu$ and $0 < k \leq 1$, which gives

$$-\frac{\nu^2 + 1}{2} \leq \mathcal{L}_\omega^{\text{dual}} \leq 0.$$

Moreover, for $\nu > 1$, a constant representation gives $\mathcal{L}_\omega^{\text{dual}} = -1$ and therefore cannot attain the optimum of $\mathcal{L}_\omega^{\text{dual}}$. As such, unlike center-distance objectives (Ruff et al., 2018), trivial representation collapse is therefore not a global optimizer of the proposed approach. The proposed alternating optimization approach also motivates the choice of encoder. For consistent optimization and learning, one expects the features used by the α-step to match those differentiated in the ω-step. Layer normalization (typically followed in transformers) provides batch-independent features, unlike batch normalization (commonly applied in CNNs) whose output depends on batch composition and operating mode (Ioffe & Szegedy, 2015). Furthermore, the self-attention mechanism in vision transformers gives the encoder access to global image structure from early layers (Raghu et al., 2021). These desirable attributes led us to use a vision transformer as the host for the proposed dual-driven alternating update. The training alternates between these two blocks across epochs. For this purpose, the most recent encoder defines the kernel for the α-step, after which the recovered dual variables are fixed for the ω-step. The process is repeated until convergence.

3.2 Convergence

For fixed ω, $F_\omega(\boldsymbol{\alpha})$ is convex and continuously differentiable w.r.t. $\boldsymbol{\alpha}$ on the compact feasible set $\mathcal{D}$ for $p > 1$. Under the standard Frank–Wolfe assumptions, the objective value converges to the global optimum of the fixed-ω subproblem when iterated sufficiently, while the Frank–Wolfe gap provides a computable certificate of the optimality (Jaggi, 2013). For the overall alternating scheme, we track the optimal primal value $P^\star(\omega) = 1 - F^\star(\omega)$. By Eq. 7, the ω-step is defined to increase the kernel quadratic term $\mathbf{a}^{\star\top} \mathbf{K}_\omega \mathbf{a}^\star$. By Proposition 1, for an exact inner solution a gradient step on $\mathcal{L}_\omega^{\text{dual}}$ is precisely a gradient step on $P^\star(\omega)$. Assume the encoder iterates remain in a compact set Ω on which $P^\star$ has an L-Lipschitz gradient. A step $0 < \eta \leq 1/L$ then satisfies the standard descent inequality (Nesterov, 2013), giving $P^\star(\omega^{(t+1)}) \leq P^\star(\omega^{(t)})$. In practice, the α-step is only solved to a finite Frank–Wolfe gap rather than exactly. This per-step descent guarantee is therefore an idealization that the finite-gap solver approximates. Since $F_\omega(\boldsymbol{\alpha})$ is continuous on the compact set $\Omega \times \mathcal{D}$, its optimal value $F^\star(\omega)$ is bounded above and below on Ω. So $P^\star(\omega) = 1 - F^\star(\omega)$ is bounded below as well. Under the exact-solution idealization, the sequence $\{P^\star(\omega^{(t)})\}$ is then nonincreasing and bounded below, hence convergent as a bounded monotone sequence. This value-function monotonicity guarantee is not available for the primal-driven scheme of (Saei & Arashloo, 2026), since its boundary and network updates optimize different objective functions rather than a shared optimal-value function.

4 EXPERIMENTS

In this section, we present the results of an experimental evaluation of the proposed approach on multiple standard datasets, and compare the results to the state-of-the-art methods. Additionally, we conduct an ablation study to study the merits of each component of the proposed method.

4.1 EXPERIMENTAL SETUP

Protocols. We evaluate the proposed approach on the most widely used datasets of Fashion-MNIST (Xiao et al., 2017), CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), their long-tailed (LT) variants of CIFAR-10-LT and CIFAR-100-LT Kim et al. (2020), and ImageNet-LT (Liu et al., 2019). Following the standard one-vs-rest protocol, each class is treated in turn as normal and anomalous training samples are drawn from the remaining classes at negative-to-positive ratios of 10%, 50%, or 75% (Rahimzadeh Arashloo, 2024; Saei & Arashloo, 2026). In order to enable a fair comparison against other methods, performance is measured in terms of the area under the receiver operating characteristic curve (AUROC). For CIFAR-10-LT and CIFAR-100-LT, class frequencies decay exponentially with imbalance ratios $\varrho \in \{100, 50, 10\}$ (Kim et al., 2020). ImageNet-LT is evaluated using the same one-vs-rest formulation and the standard Many, Medium, and Few splits. Long-tailed performance is reported using balanced accuracy to enable comparison with the existing approaches. A stratified 20% of each training pool is used for validation.

Implementation. We use ViT-B/16 with the classification head removed and ℓ_2-normalized class-token features. ImageNet-1k supervised initialization is used for Fashion-MNIST, CIFAR-10/100, and their long-tailed variants. Because ImageNet-LT shares the ImageNet-1k label space, we instead use self-distillation with no labels (DINO) initialization (Caron et al., 2021) for the transformer to avoid direct class-label leakage and prevent artificial high performance in the imageNet-LT experiments. All images are resized to 256×256, center-cropped to 224, and normalized using ImageNet statistics. Fashion-MNIST images are replicated across three channels. No data augmentation is used. The hyperparameters are set on the validation sets from $p \in \{1.01, 1.1, 1.2, 2, 5, 10, 100\}$, $\nu \in \{1, 1.2, 1.5, 1.7, 2, 5, 10\}$, and $\tilde{c}_1, \tilde{c}_2 \in \{0.01, 0.1, 1, 10\}$. The encoder is updated using the Adam optimization method with a learning rate of 10^{-5}, cosine decay of 10^{-7}, weight decay of 10^{-4}, gradient clipping at 1, and mini-batches of 16 for accumulating the full-batch gradient. The Gaussian kernel bandwidth is estimated once from the pretrained features as the inverse median pairwise squared distance and is then kept fixed during training. All experiments were run on a single NVIDIA RTX 6000 Pro or RTX 4090 graphics processing unit.

Model selection and inference. For the standard protocol, models are selected based on the validation AUROCs, while for the long-tailed recognition experiments, balanced accuracy at the recovered optimal radius ($r^{\star 2}$) on the validation set is deployed. Training stops after 15 epochs in the case of no validation improvement or when the hyper-sphere radius does not shrink. After model selection, the selected encoder checkpoint is reloaded and the SVDD boundary is refitted from scratch on the corresponding training set so that the final dual variables and recovered boundary correspond to the selected representation. Test samples are then scored using Eq. equation 4, and thresholded decisions use the refitted $r^{\star 2}$ without additional tuning. In the following sections, the *fixed* backbone refers to the case where the pretrained encoder is unchanged (not trained) and only the α-step is applied.

4.2 STANDARD ANOMALY DETECTION

Table 1 compares the proposed TLM-SVDD with representative state-of-the-art one-class, outlier-exposure, and labeled-anomaly methods. The fixed-feature large-margin ℓ_p-SVDD (Rahimzadeh Arashloo, 2024) is reported using its strongest single-kernel configuration to match our single-backbone setting. As may be observed from the table, across all three datasets and all anomaly-exposure ratios, TLM-SVDD outperforms all compared methods in Table 1. For instance, at 10% anomaly exposure, the gains over the previous best performances (reported by DLM-SVDD (Saei & Arashloo, 2026)) are 1.3, 2.0, and 3.0 AUROC points on the Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets, respectively. Across the three exposure levels, the improvements range from 0.5 to 1.3 points on the Fashion-MNIST, 1.5 to 2.0 points on the CIFAR-10, and 2.3 to 3.0 points on the CIFAR-100 datasets. While the previous best performance on the CIFAR-

Table 1: Average test AUROC on the standard one-vs-rest benchmarks. $\star$: pure one-class (no anomalies in training); $\diamond$: external or auxiliary anomaly evidence; OE denotes outlier exposure; percentages are the negative-to-positive training ratio. †: CIFAR-100 results reported under a different protocol in the original paper. "–": not reported.

Method	Fashion-MNIST	CIFAR-10	CIFAR-100
Deep SVDD* (Ruff et al., 2018)	–	0.648	–
PANDA* (Reiss et al., 2021)	0.956	0.962	0.941
MSC* (Reiss & Hoshen, 2023)	–	0.972	0.964
UNODE* (Mirzaei et al., 2024)	0.934	0.969	0.936
FIRM* (Lunardi et al., 2025)	0.968	0.934	0.878†
CSI (Tack et al., 2020)	–	0.943	0.896
DROC-contrastive (Sohn et al., 2021)	0.958	0.925	0.865
Outlier Exposure$^\diamond$ (Hendrycks et al., 2019)	–	0.978	0.879
PANDA (OE)$^\diamond$ (Reiss et al., 2021)	0.918	0.989	0.973
FIRM+OE$^\diamond$ (Lunardi et al., 2025)	0.962	0.976	0.947†
RD+EPHAD$^\diamond$ (Patra & Taieb, 2025)	0.958	0.984	–
Deep-SAD (10%) (Ruff et al., 2020)	0.882	0.860	0.891
Deep-SAD (75%) (Ruff et al., 2020)	0.981	0.925	0.887
Large-margin ℓ_p-SVDD, single kernel (10%) (Rahimzadeh Arashloo, 2024)	–	0.958	0.947
Large-margin ℓ_p-SVDD, single kernel (50%) (Rahimzadeh Arashloo, 2024)	–	0.969	0.953
Large-margin ℓ_p-SVDD, single kernel (75%) (Rahimzadeh Arashloo, 2024)	–	0.972	0.958
DLM-SVDD (10%) (Saei & Arashloo, 2026)	0.976	0.977	0.962
DLM-SVDD (50%) (Saei & Arashloo, 2026)	0.987	0.982	0.971
DLM-SVDD (75%) (Saei & Arashloo, 2026)	0.988	0.984	0.972
TLM-SVDD (10%)	**0.989**	**0.997**	**0.992**
TLM-SVDD (50%)	**0.993**	**0.998**	**0.994**
TLM-SVDD (75%)	**0.993**	**0.999**	**0.995**

10 dataset was already close to saturation, larger gains on the CIFAR-100 corpus indicate a clear benefit on more challenging datasets. The contributions of the backbone, dual update, and anomaly exposure are separately studied in Sec. 4.4.

4.3 Long-Tailed Recognition

Table 2 reports balanced accuracy under the long-tailed protocol. The long-tailed evaluations use the recovered radius $r^{\star 2}$ directly rather than a validation-tuned threshold as the most extreme tail classes contain too few normal validation samples for reliable threshold selection. On the CIFAR-10-LT dataset, TLM-SVDD improves over jointly trained DLM-SVDD by 8.5, 7.6, and 5.4 points at imbalance ratios $\varrho = 100$, 50, and 10, respectively. The corresponding gains on the CIFAR-100-LT database are 7.8, 8.9, and 4.9 points. As may be observed from Table 2, the proposed TLM-SVDD approach obtains the highest balanced accuracy for all datasets and all imbalanced ratios, considerably outperforming the state of the art. On the ImageNet-LT corpus, the proposed TLM-SVDD method improves over DLM-SVDD by 4.6 points on Many, 7.3 on Medium, 6.1 on Few, and 6.1 overall. Convergence is generally slower under the long-tailed protocol than the standard one, typically requiring 20–40 epochs on ImageNet-LT due to its larger, more heterogeneous training pools, compared with at most 25 epochs on the standard CIFAR benchmarks.

4.4 Ablations and Error Analysis

Backbone and encoder update. Table 3 separately studies the effects of the backbone, encoder update scheme, and anomaly exposure. At 10% exposure, replacing the fixed ResNet-50 representation with fixed ViT-B/16 features improves AUROC by 4.2 points on the Fashion-MNIST, 12.8 on the CIFAR-10, and 6.6 on the CIFAR-100 datasets. The transformer representation therefore accounts for a substantial part of the absolute performance, particularly on the CIFAR-10 and CIFAR-100 datasets. The dual encoder update provides an additional gain over the fixed ViT, reaching 4.5–5.0 points on the Fashion-MNIST corpus and smaller but consistent gains on the near-saturated CIFAR benchmarks.

Primal vs. dual update with the same ViT backbone. In Table 4, we compare the two update rules using the same ViT-B/16 architecture and initialization for both update rules for the CIFAR10 dataset. On CIFAR-10 with 10% anomaly exposure, each class is trained with ten independent seeds

Table 2: Balanced accuracy (%) under the long-tailed protocol. ϱ is the imbalance ratio. "–": not reported. The methods in the first group (above the ruler) are multi-class long-tailed classifiers; the methods in the second group (below the ruler) are one-vs-rest SVDD-based detectors. $\star$: strongest single-kernel configuration of large-margin ℓ_p-SVDD (Rahimzadeh Arashloo, 2024). DLM-SVDD-based methods use a ResNet-50 and TLM-SVDD-based methods a ViT-B/16; "joint" denotes joint fine-tuning of the backbone with the boundary.

	CIFAR-10-LT			CIFAR-100-LT			ImageNet-LT			
Method	ϱ=100	ϱ=50	ϱ=10	ϱ=100	ϱ=50	ϱ=10	Many	Medium	Few	All
ResLT (Cui et al., 2022)	82.40	85.10	89.70	49.70	54.50	63.70	63.60	55.70	38.90	56.10
BCL (Zhu et al., 2022)	84.30	87.20	91.10	51.90	56.60	64.90	67.20	53.10	32.90	51.90
DiffuLT (Shao et al., 2024)	84.70	86.90	90.70	51.50	56.30	63.80	63.30	55.60	39.40	56.40
Multi-Granularity (Liu et al., 2025)	87.50	88.16	92.37	54.02	58.13	65.86	71.50	56.80	41.50	60.40
GLMC+SEL (Jian et al., 2025)	85.40	88.57	92.83	56.48	61.13	70.75	68.67	54.37	38.28	57.24
BCE3S (Fan et al., 2026)	90.08	92.55	95.71	59.50	65.23	76.13	68.14	55.90	40.56	57.85
Large-margin ℓ_p-SVDD* (Rahimzadeh Arashloo, 2024)	89.20	90.20	89.60	73.30	78.70	88.70	89.50	81.30	65.10	83.40
DLM-SVDD, joint (Saei & Arashloo, 2026)	86.79	89.11	92.59	75.25	76.84	85.77	92.21	85.00	72.20	86.03
TLM-SVDD	**95.29**	**96.75**	**97.99**	**83.06**	**85.70**	**90.64**	**96.77**	**92.30**	**78.30**	**92.12**

Table 3: Ablation on the standard protocol (AUROC): backbone $\times$ encoder update $\times$ anomaly exposure. "fixed" keeps the pretrained backbone unchanged and learns only the SVDD boundary; "primal" uses the smooth primal margin-violation update of DLM-SVDD (Saei & Rahimzadeh Arashloo, 2026; Saei & Arashloo, 2026); "dual" uses the proposed dual-based encoder update.

	Fashion-MNIST			CIFAR-10			CIFAR-100		
Method	10%	50%	75%	10%	50%	75%	10%	50%	75%
ResNet-50, fixed	0.902	0.901	0.901	0.867	0.868	0.867	0.923	0.923	0.923
ResNet-50, primal	0.976	0.987	0.988	0.977	0.982	0.984	0.962	0.971	0.972
ViT-B/16, fixed	0.944	0.944	0.944	0.995	0.996	0.996	0.989	0.991	0.992
ViT-B/16, dual	**0.989**	**0.993**	**0.993**	**0.997**	**0.998**	**0.999**	**0.992**	**0.994**	**0.995**

using either the smooth primal update of DLM-SVDD (Saei & Arashloo, 2026) or the proposed dual update, with all other settings held constant. The dual update improves every class, increasing the mean AUROC from 0.982 to 0.997. The average gain is 1.5 AUROC points and ranges from 0.5 to 3.4 points across different classes. A paired test on the ten class means gives $t(9) = 5.02$ and $p = 7.2\times10^{-4}$ indicating a highly statistically significant difference between the two sets of ten class mean performances. The dual update also reduces the pooled run-to-run standard deviation from 0.012 to 0.003. This controlled experiment shows that the improvement over DLM-SVDD cannot be explained by the transformer architecture alone and the proposed dual-space learning scheme plays a significant role. Although this controlled comparison focused on a single representative dataset due to space constraints, the backbone-level trend across all three datasets (Table 3) shows the dual update consistently outperforming the primal one.

Effect of anomaly exposure. Table 3 also allows us to see the effect of increasing the amount of labeled anomaly data. TLM-SVDD already performs strongly with only 10% anomaly exposure and improves monotonically as the negative-to-positive training ratio increases from 10% to 50% and then to 75%. The gains are modest because performance is already close to saturation, but the trend shows that the proposed encoder update can make use of additional anomalous samples rather than relying on them only to refit a fixed decision boundary. The fixed-backbone variants, in comparison, change by at most 0.3 AUROC points from 10% to 75% exposure, as additional anomalies are introduced.

Fine-tuning behavior across long-tailed benchmarks. Table 5 reports the fixed and jointly fine-tuned backbones side by side for both SVDD-based detectors under the long-tailed protocol. On CIFAR-10-LT, fine-tuning the ViT backbone (TLM-SVDD vs. TLM-SVDD, fixed) improves balanced accuracy at all three imbalance ratios, by 0.09, 0.68, and 0.90 points at $\varrho = 100$, 50, and 10, respectively. On CIFAR-100-LT, fine-tuning instead reduces balanced accuracy by 1.86, 2.61, and 3.00 points at the same ratios, and the fixed backbone outperforms the jointly fine-tuned one on this dataset. On ImageNet-LT, fine-tuning improves every split: Many, Medium, Few, and overall balanced accuracies increase by 1.0, 2.2, 5.2, and 2.1 points, respectively. Even with this gain, TLM-

Table 4: Matched ViT-B/16 comparison of the primal and dual encoder updates on the CIFAR-10 dataset with 10% anomaly exposure (10 seeds per class for each update rule). Mean and standard deviation (SD) are computed across the ten per-class performances; Δ is the per-class AUROC gain (points) of dual over primal. Columns refer to the CIFAR-10 class indices in standard order (0 airplane, 1 automobile, 2 bird, 3 cat, 4 deer, 5 dog, 6 frog, 7 horse, 8 ship, 9 truck).

Class	0	1	2	3	4	5	6	7	8	9	Mean	SD
Primal	0.984	0.994	0.975	0.958	0.984	0.967	0.993	0.986	0.990	0.988	0.982	0.011
Dual	0.998	0.999	0.997	0.992	0.996	0.992	0.999	0.998	0.998	0.998	**0.997**	0.002
Δ	+1.4	+0.5	+2.2	+3.4	+1.2	+2.5	+0.6	+1.2	+0.8	+1.0	**+1.5**	–

Table 5: Ablation on the long-tailed protocol (balanced accuracy, %): fixed vs. jointly fine-tuned backbones for the two SVDD-based detectors. "fixed" keeps the pretrained backbone unchanged and only learns the hypersphere boundary; "joint" fine-tunes the backbone jointly with the boundary, as reported for these methods in Table 2.

	CIFAR-10-LT			CIFAR-100-LT			ImageNet-LT			
Method	ϱ=100	ϱ=50	ϱ=10	ϱ=100	ϱ=50	ϱ=10	Many	Medium	Few	All
DLM-SVDD, fixed (Saei & Arashloo, 2026)	59.50	65.18	58.84	59.38	66.15	64.70	91.25	88.97	**78.55**	88.43
DLM-SVDD, joint (Saei & Arashloo, 2026)	86.79	89.11	92.59	75.25	76.84	85.77	92.21	85.00	72.20	86.03
TLM-SVDD, fixed	95.20	96.07	97.09	**84.92**	**88.31**	**93.64**	95.75	90.11	73.12	89.97
TLM-SVDD	**95.29**	**96.75**	**97.99**	83.06	85.70	90.64	**96.77**	**92.30**	78.30	**92.12**

SVDD's Few-shot balanced accuracy (78.30) remains within 0.25 points of the fixed DLM-SVDD variant (78.55, Table 5), the only setting in which a fixed backbone is competitive with TLM-SVDD across either detector. In contrast, fine-tuning the primal-space CNN-backbone DLM-SVDD (Saei & Arashloo, 2026) is beneficial only for the Many-shot split and decreases the Medium, Few, and overall results by 4.0, 6.4, and 2.4 points, respectively.

Bandwidth sensitivity. We compare three strategies for selecting the Gaussian kernel bandwidth γ: (1) using a constant preset value, (2) estimating γ once from the pretrained features using the median pairwise squared distance among training samples and keeping it fixed during training, and (3) automatically re-estimating γ from the evolving features at every epoch. Among these, the constant preset value performs the worst because it ignores the feature scale, while per-epoch re-estimation is also inferior and less stable than the one-time estimate. According to equation 12, the encoder is adapted with respect to the current kernel geometry, so re-estimating the bandwidth from the same evolving features introduces an additional varying quantity into the optimization and leads to less stable validation performance across epochs. A single estimate from the pretrained features instead provides a fixed reference scale throughout training, consistent with the observation in (Saei & Arashloo, 2026).

5 Conclusion

We presented the TLM-SVDD approach for anomaly detection, which learns a vision-transformer representation through the optimal-value function of a large-margin ℓ_p-SVDD boundary. At the dual optimum, Danskin's theorem provides the encoder hypergradient without recovered thresholds, surrogate losses, or differentiation through the boundary solver. The resulting encoder objective is bounded, and its gradient depends only on the estimated support vectors. Experiments confirm that both the transformer representation and the proposed dual update scheme contribute to the improvement over previous CNN-based primal approximate methods. In a controlled comparison using the same ViT-B/16 architecture and initialization for both update rules, the dual update improves upon the primal approximate surrogate, highlighting its contribution independent from the architecture. The transformer-based variants also substantially improve performance under severe long-tailed training distributions. Future work may extend the proposed optimal-value encoder update to multiple-kernel and more general kernel-boundary formulations.

ETHICS STATEMENT

This work only uses existing standard benchmark datasets for research evaluation. Fashion-MNIST and CIFAR-10/100 are publicly available, while ImageNet-LT was accessed through the official ImageNet access process and used according to its research terms and conditions. The study is limited to the benchmark settings considered in this paper.

REPRODUCIBILITY STATEMENT

The complete formulation of TLM-SVDD is given in Sec. 3. The boundary and encoder updates are specified in equation 2 and equation 12, and the alternating optimization procedure is described in Sec. 3.1. Propositions 1 and 2 are proved in the main text.

The datasets, data splits, evaluation protocols, backbone initializations, preprocessing steps, selected hyperparameters, hyperparameter search spaces, optimizer settings, stopping criteria, model-selection procedure, and test-time refitting protocol are discussed in Sec. 4.1. The bandwidth-selection analysis and additional optimization behavior are detailed in Sec. 4.4.

An anonymized implementation of TLM-SVDD, including the source code, configuration files, and scripts required to reproduce the reported experiments, is provided as supplementary material. A public GitHub repository will be released upon acceptance.

AI USE STATEMENT

We used generative artificial intelligence (AI) tools in a purely assistive capacity, in the following manner. First, for language editing: grammar correction, improving clarity and concision of already-drafted text, and checking consistency of terminology and notation across sections. Second, for LaTeX assistance: diagnosing and fixing compilation errors, table and float formatting, and cross-referencing in Overleaf.

REFERENCES

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pp. 136–145. PMLR, 2017.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

Matan Jacob Cohen and Shai Avidan. Transformaly - two (feature spaces) are better than one. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4059–4068, 2022. doi: 10.1109/CVPRW56347.2022.00451.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3695–3706, 2022.

John M Danskin. *The theory of max-min and its application to weapons allocation problems*. Springer Science & Business Media, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Weijia Fan, Qiufu Li, Jiajun Wen, and Xiaoyang Peng. Bce3s: Binary cross-entropy based tripartite synergistic learning for long-tailed recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3795–3803, 2026.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

Hadi Hojjati and Narges Armanfard. Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2023. doi: 10.1109/TKDE.2023.3328882.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435, 2013.

Zhongquan Jian, Yanhao Chen, Yancheng Wang, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Supervised exploratory learning for long-tailed visual recognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1870–1880. IEEE, 2025.

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13896–13905, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yuting Liu, Liu Yang, and Yu Wang. Long-tailed classification with multi-granularity semantics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4285–4294. IEEE, 2025.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2537–2546, 2019. doi: 10.1109/CVPR.2019.00264.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020.

William T. Lunardi, Abdulrahman Banabila, Dania Herzalla, and Martin L. Andreoni. Learning in-distribution representations for anomaly detection, 2025.

Olvi L Mangasarian. *Nonlinear programming*. SIAM, 1994.

Hossein Mirzaei, Mojtaba Nafez, Mohammad Jafari, Mohammad Bagher Soltani, Mohammad Azizmalayeri, Jafar Habibi, Mohammad Sabokrou, and Mohammad Hossein Rohban. Universal novelty detection through adaptive contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22914–22923, 2024. doi: 10.1109/CVPR52733.2024.02162.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06, 2021. doi: 10.1109/ISIE45552.2021.9576231.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=D78Go4hVcxO.

Sukanya Patra and Souhaib Ben Taieb. An evidence-based post-hoc adjustment framework for anomaly detection under data contamination. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. URL https://openreview.net/forum?id=FT5eOdWSWN.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Shervin Rahimzadeh Arashloo. lp-norm support vector data description. *Pattern Recognition*, 132:108930, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.108930. URL https://www.sciencedirect.com/science/article/pii/S0031320322004113.

Shervin Rahimzadeh Arashloo. Large-margin multiple kernel ℓ_p-svdd using frank–wolfe algorithm for novelty detection. *Pattern Recognition*, 148:110189, 2024. doi: 10.1016/j.patcog.2023.110189.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 2155–2162, 2023. doi: 10.1609/AAAI.V37I2.25309.

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 (CVPR)*, pp. 2806–2814, 2021. doi: 10.1109/CVPR46437.2021.00283.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, 2018.

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 (ICLR)*, 2020.

Alireza Dastmalchi Saei and Shervin Rahimzadeh Arashloo. Deep convolutional large-margin ℓ_p-svdd for visual anomaly detection, 2026. URL https://arxiv.org/abs/2607.22212.

Alireza Dastmalchi Saei and Shervin Rahimzadeh Arashloo. Deep large-margin lp-svdd with cnn feature learning for novelty detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2376–2380, 2026. doi: 10.1109/ICASSP55912.2026.11463656.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. *Advances in neural information processing systems*, 37:123007–123031, 2024.

Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations (ICLR)*, 2021.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, 2004. doi: 10.1023/B:MACH.0000008084.60811.49.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pp. 370–378. PMLR, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Hong-Jie Xing, Hui-Nan Wu, and Ping-Ping Zhang. Deep multi-sphere support vector data description based on disentangled representation learning. *Pattern Recognition*, 156:110842, 2024. doi: 10.1016/j.patcog.2024.110842.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6898–6907. IEEE, 2022.