The Open-set Push-Pull Loss: A Revisit of Contrastive Learning for Open-set Supervised Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Open-set Anomaly detection (OSAD) can be formulated as a binary classification problem under severe class imbalance and limited knowledge about anomalies. Despite being increasingly used in classification problems, the current formulations of contrastive learning are not well-suited for OSAD. In this work, we address OSAD through a margin-based contrastive loss, the Open-set Push-Pull (OPP) loss. It considers only normal samples as anchors, aggregates them with positives via an attraction term, and pushes anomalies away via a repulsion term. We show that by minimizing OPP, we are equivalently maximizing a lower bound on the anomaly-normal margin on the hypersphere. We also quantify the bounds of this approximation and derive a gradient analysis, which clarifies the role of the temperature τ. Building on this formulation, we propose two variants: (i) the τ-balanced OPP loss that uses two distinct temperatures for the normals and anomalies, directly calibrated through the bound to account for class imbalance, (ii) a regularized OPP loss that saturates attraction and repulsion term to mitigate overfitting to training anomalies. We evaluate different contrastive objectives and our formulation across multiple open-set regimes on the downstream task of AD, on two industrial benchmarks MVTecAD, RealIAD, and finally Mastcam. We further compare against other state-of-the-art supervised AD architectures. Our results show that the OPP loss yields better robustness to increasing openness than standard supervised contrastive objectives and shows promising results when compared to other OSAD methods.

1 Introduction

Anomaly detection (AD) is the task of automatically determining whether a sample deviates from the characteristics shared by normal (or nominal) data. Consequently, a fundamental prerequisite of any AD algorithm is an accurate characterization of normality. As nominal data is typically easier to collect in large quantities than anomalies, much of the AD literature has been developed in the *unsupervised* or *one-class* setting (Schölkopf et al., 2001; Tax & Duin, 1999; Ruff et al., 2020; Zhou, 2018), and more recently, through self-supervised learning (Hendrycks et al., 2019; Hojjati et al., 2024). Alternatively, AD can be viewed as a binary classification problem, with two caveats: class imbalance and lack of knowledge about anomalies.

Despite these challenges, the various efforts to recast AD as a semi-supervised (Ruff et al., 2020; Hendrycks et al., 2019) or a fully supervised problem (Liznerski et al., 2022) demonstrably refine the decision boundary, and outperform unsupervised baselines (Ruff et al., 2021) even with a limited number of anomalies. Standard supervised deep classifiers typically rely on variants of the binary cross-entropy loss. In particular, contrastive learning formulates a per-anchor loss that encourages an *anchor* and its augmented view, termed *positive*, to cluster, while pushing away the remaining samples in the batch, treated as *negatives* (Chen et al., 2020; Wang & Isola, 2020). The label information is further extended to Supervised Contrastive learning (SupCon) by Khosla et al. (2020) in which the anchor is now attracted to multiple positives, namely all the samples sharing the same class label as the anchor. This paradigm offers a principled alternative for supervised AD: contrastive pretraining structures the representation space by grouping normal samples as positives, and separating anomalies as negatives, after which an appropriate anomaly score can be used for the detection.

However, (i) SupCon encourages all samples sharing the same label to cluster in the embedding space, which is not a desirable property for anomalies as they might be of different nature. Lai et al. (2024) attempt to restrict the anchors and positives to normals, however they do not completely decouple normals from anomalies, causing a spurious minimum of the function that is discussed in Appendix A. Moreover, under severe class imbalance (ii) Mildenberger et al. (2025) show that SupCon yields poor embeddings for the minority class, and propose imbalance-aware variants. These propositions will be unified in Section 3. This paper introduces a contrastive objective suited for OSAD, namely the Open-set Push-Pull (OPP) loss. OPP is a soft approximation of a normal-anomaly margin, that uses only normal samples as anchors, and pushes anomalies via a repulsing term. Unlike previous formulations, it decouples the attraction and repulsion functions and uses the same soft approximation using the log-sum-exp operator.

To guard against overfitting to the training subset of anomalies, the evaluation protocol has long relied on the *one vs. rest* (or *leave-one-out*) benchmark on natural images. More recently, there has been a shift to manufacturing and industrial benchmarks such as MVTecAD, and more recently RealIAD (Wang et al., 2024) where anomalies are more subtle, and they share the same semantics as normals. This benchmark is increasingly adopted and recommended in *Open-set Supervised Anomaly Detection* (OSAD) (Ding et al., 2022; Wang et al., 2024). Thus, we focus on this well-established benchmark to align our contributions with OSAD. The objective of our numerical experiments is dual. First, we aim to provide a detailed comparison between six different contrastive objectives and our proposal, at varying levels of *openness* (equation 22). Second, by equipping this representation learner with an anomaly score, we provide additional comparisons with other state-of-the-art supervised AD methods.

Our contributions. To summarize, our contributions are as follows:

- Mainly, we introduce OPP, a contrastive objective tailored for OSAD. Unlike SupCon and related variants, OPP only uses nominal data as anchors, aggregates them with positives and pushes away anomalies via a log-sum-exp function. OPP decouples this attraction and repulsion.
- We show that by minimizing OPP, we are equivalently maximizing a lower bound on the anomaly-normal margin (Proposition 4.1), and further derive a gradient expression which yields a clear interpretation of the temperature τ. Building on this analysis, we propose two variants (i) τ-OPP that uses two different temperatures for normals and anomalies and balances their contribution per-batch using the derived inequality in 4.1, (ii) regularized OPP that adds margin on the max/min similarities in an effort to mitigate for overfitting to training anomalies.
- **OSAD evaluation.** We evaluate our approach on typical experiments done in the field of OSAD. We aim to empirically correlate the quality of the embedding space on the downstream task of OSAD. For this purpose, we provide a detailed comparison of contrastive objectives under varying degrees of *openness* that we rigorously define in equation 22. Results show that an appropriate representation space enable the separation between normals and unseen anomalies. Encouraged by these results, we further compare the best performing OPP loss against state-of-the-art methods in supervised AD.

2 Related works

We first clarify the distinctions between OSAD and other adjacent threads of the literature, namely Out-Of-Distribution (OOD) detection and Open-Set Recognition (OSR) (Salehi et al., 2022). Then, we review various supervised AD architectures. We divide them solely based on the benchmark used to assess their openness.

Related fields. OOD detection differs from OSAD in that it aims at identifying semantically different classes from a pre-defined set in the context of a multi-class classification problem, OSR extends OOD detection to the open-set case, where the aim is to detect and classify OOD samples. Indeed, any OOD detection method can be transposed to AD by considering all the classes as normal.

Towards supervised AD approaches. The classical unsupervised baselines of AD range from distance-based methods (Cover & Hart, 1967), kernel-based methods such as one-class Support Vector Machines

(OC-SVM) (Schölkopf et al., 2001), and support vector data description (SVDD) to their deep counterparts: DeepSVDD (Ruff et al., 2018) trains a neural network to enclose normal data with a minimum-volume hypersphere, and the anomaly score being the distance of a sample to the center. In Ruff et al. (2020), the same authors generalize this objective to the semi-supervised setting, exploiting the few labeled anomalies by repelling them from the center. Later in Liznerski et al. (2022), they propose the Hypersphere Classifier that changes the original DeepSAD loss with a cross-entropy loss and radial basis function. Another paradigm aims to directly learn an end-to-end anomaly scorer, for instance deviation networks (DevNet) (Pang et al., 2019) directly optimizes a scalar Z-score with a neural network trained to push the anomalies into the upper tail through a deviation loss. Liznerski et al. (2022) have indeed evaluated their results using one vs. rest benchmark on natural images, whereas Pang et al. (2019) have not evaluated their generalization to unknown anomalies.

OSAD architectures. Most of the existing methods that explicitly tackle OSAD rely on *Augmentation-based modeling* to compensate for the scarcity of anomalies, and to alleviate the overfitting to known anomalies. Disentangled Representations of Anomalies (DRA) (Ding et al., 2022) extended DevNets with a multi-head architecture, they train three heads (seen anomalies, pseudo-anomalies via CutMix augmentations, and latent residual anomalies via top-K multiple instance learning) to expose the model to various anomalies. RealNet (Zhang et al., 2024) proposes to synthesize realistic anomalies with a diffusion process and selection discriminant pre-trained features. Rather than relying on augmenting the anomalies, Boundary-Guided AD (BGAD) (Yao et al., 2023) learn a separating boundary from a normalizing flow estimate of the nominal samples and optimize a boundary-guided semi-push-pull loss that pulls normal features together while pushing anomalous ones beyond a margin. To combine the strength of these various approaches, Zhu et al. (2024) propose a collaborative framework in which they simulate heterogeneous subsets of the training data to train anomaly models and merge them using collaborative differential learning.

In this work, we tackle the OSAD problem from a representation learning perspective by using a contrastive based approach. As this type of objective has seldom been studied in this particular context, we now detail the variants of contrastive losses that have been applied to related fields.

3 Variants of Contrastive Losses

Our work attempts to tackle OSAD by means of contrastive pre-training with a novel OPP loss. Thus, we first review the standard supervised contrastive objective introduced by Khosla et al. (2020), before narrowing down to works that tackle the challenges of AD. In doing so, we aim, on one hand, to provide a comprehensive framework that facilitates the transition between these various contributions, and more importantly, emphasize the key differences with our proposed formulation.

3.1 Notations

Let $\{\mathbf{X}_i, y_i\}_{i=1}^N$ be a dataset comprising N images $\{\mathbf{X}_i\}$ and their associated labels $\{y_i\}$. We create corresponding *views* by applying N_{views} realizations of a fixed set of transformations T. We define a *multi-viewed* mini-batch $\mathcal{B}$ as the set of indices corresponding to the $(N_{\text{views}} \times N)$ views. In practice, we use $N_{\text{views}} = 2$ and denote $i, i_+ \in \mathcal{B}$ the two views corresponding to the same original sample. Let $g_{\boldsymbol{\theta}}$ be a network, typically the composition of an encoder and a projection head, parametrized by $\boldsymbol{\theta}$, and mapping an input $\mathbf{X}_i \in \mathcal{B}$ to an l_2-normalized representation $\mathbf{u}_i = \frac{g_{\boldsymbol{\theta}}(\mathbf{X}_i)}{\|g_{\boldsymbol{\theta}}(\mathbf{X}_i)\|_2}$ on the unit hyper-sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ (Chen et al., 2020). Let $s_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j$ denote the dot product, which amounts to cosine similarity between normalized representations. In the context of contrastive learning, for a given anchor representation $\mathbf{u}_i$, we consider two index sets drawn from a mini-batch $\mathcal{B}$:

- $\mathcal{P}_i \subseteq \mathcal{B} \setminus \{i\}$: the samples to attract to the anchor, called *positives*
- $\mathcal{N}_i \subseteq \mathcal{B} \setminus \{i\}$: the samples to repel from the anchor, called *negatives*.

The choice of $\mathcal{P}_i$ and $\mathcal{N}_i$ is dependent on the algorithm used. For the OSAD application, we also define $\mathcal{B}_n \subseteq \mathcal{B}$ the set of all normals in a mini-batch, $\mathcal{Q}_i = \mathcal{B}_n \setminus \{i\}$, and $\mathcal{B}_a \subset \mathcal{B}$ the set of all anomalies in

a mini-batch. Note that positives $\mathcal{P}_i$ and negatives $\mathcal{N}_i$ are not equal to normals $\mathcal{Q}_i$ and anomalies $\mathcal{B}_a$ in general.

A broad family of contrastive objectives can be written under the unified form of the per-anchor loss (Lee, 2025),

$$l_i = \Phi_{\text{attract}}\left(\{s_{ij}\}_{j\in\mathcal{P}_i}\right) + \Psi_{\text{repel}}\left(\{s_{ij}\}_{j\in\mathcal{N}_i}\right) \tag{1}$$

the full objective is $\frac{1}{|\mathcal{B}|}\sum_i l_i$ in which Φ_{attract} aggregates similarities to positives and Ψ_{repel} aggregates dissimilarities to negatives. Hereinafter, we refer to the log-sum-exp function as $\text{LSE}\{x_i\}_{i\in[1,\dots,N]}$:

$$\text{LSE} : \begin{cases} \mathbb{R}^N \to \mathbb{R} \\ \{x_i\}_{i\in[1,\dots,N]} \mapsto \log\sum_{i=1}^N \exp(x_i). \end{cases} \tag{2}$$

3.2 Base Contrastive Objectives

NT-Xent. In the self-supervised setting, each anchor i is paired with the single positive i_+. In other words, the set of positives in this case is the singleton $\mathcal{P}_i = \{i_+\}$. Here, the set of negatives is all elements of the batch distinct from i, $\mathcal{N}_i = \mathcal{B}\setminus\{i\}$. The Negative-Temperature-Scaled cross-entropy (NT-Xent) per-anchor loss is defined as (Chen et al., 2020),

$$l_i^{\text{NT-Xent}} = -\frac{\mathbf{u}_i\cdot\mathbf{u}_{i_+}}{\tau} + \text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{B}\setminus\{i\}} \tag{3}$$

where $\tau > 0$ is the temperature parameter which plays a role in hard-mining the negatives: a smaller τ will concentrate on the closest negatives to the anchor point. Wang & Isola (2020) decouple the NT-Xent loss into alignment and uniformity losses, where $\Phi_{\text{attract}}\left(\{s_{ij}\}_{j\in\mathcal{P}_i=\{i_+\}}\right)$ encourages two views of the same sample to cluster in the embedding space, while $\Psi_{\text{repel}}\left(\{s_{ij}\}_{j\in\mathcal{N}_i=\mathcal{B}\setminus\{i\}}\right)$ maximizes the average distances between all samples, leading to features vectors uniformly distributed on $\mathbb{S}^{d-1}$.

Supervised Contrastive Loss (SupCon). Khosla et al. (2020) extend the NT-Xent loss to supervised learning by allowing multiple positives for each anchor, namely all samples in the batch sharing the same label $\mathcal{P}_i = \{j\in\mathcal{B}\setminus\{i\} : y_j = y_i\}$, and by keeping $\mathcal{N}_i = \mathcal{B}\setminus\{i\}$. The per-anchor SupCon loss is defined as

$$l_i^{\text{SupCon}} = -\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i}\log\frac{\exp(\frac{s_{ip}}{\tau})}{\sum_{j\neq i}\exp(\frac{s_{ij}}{\tau})} = -\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i}\frac{s_{ip}}{\tau} + \text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i} \tag{4}$$

$$= -\frac{1}{\tau}\mathbf{u}_i\cdot\bar{\mathbf{u}}_p + \text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}, \tag{5}$$

where $\bar{\mathbf{u}}_p = \frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i}\mathbf{u}_p$. Thus, SupCon combines an average attraction term over same-class positives $\Phi_{\text{attract}}\left(\{s_{ij}\}_{j\in\mathcal{P}_i}\right) = -\frac{1}{\tau}\mathbf{u}_i\cdot\bar{\mathbf{u}}_p$ with a repulsion term over the full batch $\mathcal{N}_i = \mathcal{B}\setminus\{i\}$,$\Psi_{\text{repel}}\left(\{s_{ij}\}_{j\in\mathcal{N}_i}\right) = \text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$.

3.3 Imbalance-Aware Supervised Variants

However, SupCon leads to ineffective representations for long-tailed datasets (Li et al., 2022). Specifically for binary classification, Mildenberger et al. (2025) prove that severe class imbalance yields indistinguishable representations for the two classes, resulting in a poor downstream accuracy.

Supervised Minority (SupMin). Mildenberger et al. (2025) show that under severe class imbalance, the magnitude gradient of l_i^{SupCon} vanishes as the number of samples from the majority class increases, hindering meaningful updates to the majority class. They empirically study this *collapse* through alignment and uniformity metrics originally proposed by Wang & Isola (2020). To address this, they propose a first variant which applies supervision only to the minority class and defaults to the NT-Xent loss for majority samples:

$$l_i^{\text{SupMin}} = \begin{cases} l_i^{\text{SupCon}} \text{ if } i\in\text{Minority} \\ l_i^{\text{NT-Xent}} \text{ if } i\in\text{Majority}. \end{cases} \tag{6}$$

To align their contributions in the context of AD, we treat anomalies as the minority class, and normals as the majority. We propose a variant of SupMin that enforces the alignment of normals using SupCon, while applying the NT-Xent loss to anomalies so that only their augmented views are pulled together. We will refer to this variant as *Supervised Majority* (SupMaj).

$$l_i^{\text{SupMaj}} = \begin{cases} l_i^{\text{SupCon}} \text{ if } i \in \text{Majority} \\ l_i^{\text{NT-Xent}} \text{ if } i \in \text{Minority.} \end{cases} \tag{7}$$

Supervised Prototypes (SupProto). The second proposal of Mildenberger et al. (2025) introduces *class prototypes*. Let $\mathbf{p}_{\text{maj}} \in \mathbb{S}^{d-1}$ denote the prototype of the majority class and $\mathbf{p}_{\text{min}} = -\mathbf{p}_{\text{maj}}$ the minority prototype. Writing $\mathbf{p}_i$ the prototype associated with the anchor i, and $s_{ip} = \mathbf{u}_i \cdot \mathbf{p}_i$, the loss takes the form of:

$$l_i^{\text{SupProto}} = \begin{cases} l_i^{\text{NT-Xent}} - s_{ip}/\tau + \text{LSE}\{s_{ij}/\tau\}_{j \neq i} \text{ if } s_{ip} \geq \delta \\ l_i^{\text{NT-Xent}} \text{ otherwise,} \end{cases} \tag{8}$$

where δ controls the activation of prototype attraction. The intuition is to only attract a sample towards its class prototype when it remains sufficiently far from it, thus preserving some intra-class variability. As for the choice of the prototype, they offer two possibilities: generating a random feature vector $\mathbf{p}_{\text{maj}}$, or learning via gradient descent the unit vector that maximizes the average cosine similarity to the set of feature vectors. In this work, we simply fix $\mathbf{p}_{\text{maj}}$ to the mean of feature vectors in the majority class (normals).

3.4 AD-tailored variants

Mean-Shifted Contrastive loss (MSC). Applying standard contrastive objectives can yield suboptimal representations, specifically when using pre-trained feature extractors: standard contrastive objectives force features to distribute uniformly around the center of the hypersphere, which scatters the normal data and increase its overlap with anomalous samples. To mitigate this, Reiss & Hoshen (2023) propose the Mean Shifted Contrastive loss (MSC) that computes the baseline NT-Xent loss on shifted representations by subtracting the mean of the training features $\mathbf{c} = \mathbb{E}_{\mathbf{X} \in \mathcal{X}_{\text{train}}}[\frac{g_{\boldsymbol{\theta}}(\mathbf{X})}{\|g_{\boldsymbol{\theta}}(\mathbf{X})\|_2}]$. To apply this to OSAD, a supervised setting, we compute the SupCon loss on these shifted features instead.

$$l_i^{\text{MSC}} = -\frac{(\mathbf{u}_i - \mathbf{c}) \cdot (\bar{\mathbf{u}}_p - \mathbf{c})}{\tau} + \text{LSE}\{(\mathbf{u}_i - \mathbf{c}) \cdot (\mathbf{u}_j - \mathbf{c})/\tau\}_{j \in \mathcal{N}_i} \tag{9}$$

Supervised Contrastive OSAD (SC-OSAD). In all the previous formulations of the contrastive objectives, the anchors are sampled from the whole batch. However, the anomalies seen during training may be highly heterogeneous and do not need to form a coherent cluster, so forcing them to collapse in the representative space may hurt generalization to unseen anomalies. For this reason, Lai et al. (2024) propose to restrict the attraction term to normal samples only, while keeping the same denominator as SupCon. Their modification amounts to setting the set of positives as the set of normal samples distinct from i, $\mathcal{Q}_i$:

$$l_i^{\text{SC-OSAD}} = \Phi_{\text{attract}}\left(\{s_{ij}\}_{j \in \mathcal{P}_i = \mathcal{Q}_i}\right) + \Psi_{\text{repel}}\left(\{s_{ij}\}_{j \in \mathcal{N}_i = \mathcal{B} \setminus \{i\}}\right) \tag{10}$$

While this modification prevents anomaly-to-anomaly attraction, the normal samples still contribute simultaneously to Φ_{attract} and Ψ_{repel}. Yeh et al. (2022) refer to this as *negative-positive coupling* (NPC) and show that it hinders the optimization of the contrastive loss especially in small-batch regime. Full details, with simulations on a toy example, of the spurious minimum induced by this choice of Ψ_{repel} in $l_i^{\text{SC-OSAD}}$ is discussed in Appendix A.

To summarize, with the exception of SC-OSAD, very few studies address AD through supervised contrastive learning. Imbalance-aware variants only tackle AD indirectly by mitigating class-imbalance, which motivated our adaptation of SupMaj and SupProto for AD. Since it is closest to our contribution, the limitations of SC-OSAD will be evidenced on a toy example in Appendix A. In light of these limitations, we propose in the following a new contrastive objective: the *Open-set Push Pull* loss. Note that all the aforementioned objectives will be bench-marked against our proposal to evaluate their robustness and generalization to unseen anomalies.

Table 1: Summary of per-anchor contrastive losses. We detail the sets of positives and of negatives used for each loss, as well as their respective attract and repel aggregation functions.

Per-anchor loss	$\mathcal{P}_i$	$\mathcal{N}_i$	Φ_{attract}	Ψ_{repel}
$l_i^{\text{NT-Xent}}$	$\{\mathbf{u}_{i_+}\}$	$\mathcal{B} \setminus \{i\}$	$-\frac{1}{\tau}\mathbf{u}_i \cdot \mathbf{u}_{i_+}$	$\text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$
l_i^{SupCon}	$\{j \in \mathcal{B} \setminus \{i\} : y_j = y_i\}$	$\mathcal{B} \setminus \{i\}$	$-\frac{1}{\tau}\mathbf{u}_i \cdot \bar{\mathbf{u}}_p$	$\text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$
$l_i^{\text{SC-OSAD}}$	$\mathcal{Q}_i = \mathcal{B}_n \setminus \{i\}$	$\mathcal{B} \setminus \{i\}$	$-\frac{1}{\tau}\mathbf{u}_i \cdot \bar{\mathbf{u}}_p$	$\text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$
l_i^{MSC}	$\{j \in \mathcal{B} \setminus \{i\} : y_j = y_i\}$	$\mathcal{B} \setminus \{i\}$	$-\frac{1}{\tau}(\mathbf{u}_i - \mathbf{c}) \cdot (\bar{\mathbf{u}}_p - \mathbf{c})$	$\text{LSE}\{(\mathbf{u}_i - \mathbf{c}) \cdot (\mathbf{u}_j - \mathbf{c})/\tau\}_{j\in\mathcal{N}_i}$
l_i^{OPP} (**ours**)	$\mathcal{Q}_i = \mathcal{B}_n \setminus \{i\}$	$\mathcal{B}_a = \mathcal{B} \setminus \mathcal{B}_n$	$\tau\text{LSE}\{-s_{ip}/\tau\}_{p\in\mathcal{P}_i}$	$\tau\text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$

4 The *Open-Set Push Pull* (OPP) Loss

4.1 Definition of OPP

We recall here that i is the index of the anchor, $\mathcal{B}_n$ denotes the set of normal samples in the batch $\mathcal{B}$, $\mathcal{Q}_i = \mathcal{B}_n \setminus \{i\}$, and $\mathcal{B}_a \neq \emptyset$[1] is the set of anomalous samples in the batch. In this setting, a natural margin to maximize for separation around anchor i is,

$$l_i^{\text{Margin}} = \min\{s_{iq}\}_{q\in\mathcal{Q}_i} - \max\{s_{ia}\}_{a\in\mathcal{B}_a} \tag{11}$$

$\min\{s_{iq}\}_{q\in\mathcal{Q}_i}$ is the normal sample most dissimilar to the anchor, and $\max\{s_{ia}\}_{a\in\mathcal{B}_a}$ is the most similar anomaly to the anchor. A large l_i^{Margin} ensures that normals and anomalies are well separated. We illustrate in figure 1 this margin in $\mathbb{S}^1$.

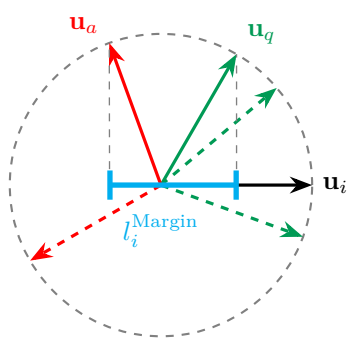

Figure 1: Schematic representation of the i-th term of l_i^{Margin}. The black element $\mathbf{u}_i$ corresponds to the anchor, while red elements $\mathbf{u}_a$ are anomalies and green elements $\mathbf{u}_q$ are from the normal class.

Directly optimizing equation 11 is hard due to the presence of min and max operators which are not differentiable. To address this, a common strategy is to introduce a smooth surrogate. We base our approximation on the log-sum-exp function previsouly defined in equation 2, and define the OPP loss for an anchor i as,

$$l_i^{\text{OPP}} = \tau\text{LSE}\{-s_{iq}/\tau\}_{q\in\mathcal{Q}_i} + \tau\text{LSE}\{s_{ia}/\tau\}_{a\in\mathcal{B}_a}. \tag{12}$$

where $\tau > 0$. The OPP loss retains a direct relationship to the margin in equation 11 through the following inequality.

Proposition 4.1 *$-l_i^{\text{OPP}}$ is a lower bound on the margin*

$$l_i^{\text{Margin}} - \tau \log\left(|\mathcal{B}_a||\mathcal{Q}_i|\right) \leq -l_i^{\text{OPP}} \leq l_i^{\text{Margin}} \tag{13}$$

[1]When implementing this formulation of the loss function, we ensure during training that each batch contains at least one anomaly example using a Batch Sampler.

Sketch of proof: One can recover inequality 13 by using the log-sum-exp inequality $\max\{x_i\}_{i\in[1,N]} \leq \tau\text{LSE}\{x_i/\tau\}_{i\in[1,N]} \leq \max\{x_i\}_{i\in[1,N]} + \tau \log N$ with $\min\{x_i\}_{i\in[1,N]} = -\max\{-x_i\}_{i\in[1,N]}$.

- By minimizing l_i^{OPP}, we are equivalently maximizing a lower bound on the margin. Thus OPP has a clear geometric target.
- The tightness of this approximation is quantifiable and depends on the temperature τ and $\log(|\mathcal{B}_a||\mathcal{Q}_i|)$. When $\tau \to 0^+$, OPP concentrates on the closest anomaly and the furthest normal from the anchor, which results in a tight approximation of the margin but may result in an unstable training using gradient-based descent methods. A large τ results in a looser approximation but a smooth objective function and more stable training.

In the context of AD, anomalies are heterogeneous, thus we restrict the anchors i to be selected from the normal class $\mathcal{B}_n$ to prevent any anomaly-to-anomaly attraction, the full objective to minimize is,

$$\begin{aligned}\mathcal{L}^{\text{OPP}}(\boldsymbol{\theta}) &= \frac{1}{|\mathcal{B}_n|}\sum_{i\in\mathcal{B}_n} l_i^{\text{OPP}} \\ &= \frac{\tau}{|\mathcal{B}_n|}\sum_{i\in\mathcal{B}_n}\Big[\text{LSE}\left(\{-s_{iq}/\tau\}_{q\in\mathcal{Q}_i}\right) + \text{LSE}\left(\{s_{ia}/\tau\}_{a\in\mathcal{B}_a}\right)\Big].\end{aligned} \tag{14}$$

In the formalism of contrastive losses outlined above, l_i^{OPP} uses all normals but i as the set of positives: $\mathcal{P}_i = \mathcal{B}_n \setminus \{i\} = \mathcal{Q}_i$, all anomalies as the set of negatives: $\mathcal{N}_i = \mathcal{B} \setminus \mathcal{B}_n = \mathcal{B}_a$, and the attract and repel aggregate functions $\Phi_{\text{attract}}(\{s_{ip}\}_{p\in\mathcal{P}_i}) = \tau\text{LSE}\{-s_{ip}/\tau\}_{p\in\mathcal{P}_i}$ and $\Psi_{\text{repel}}(\{s_{ij}\}_{j\in\mathcal{N}_i}) = \tau\text{LSE}\{s_{ij}/\tau\}_{j\in\mathcal{N}_i}$. Unlike standard SupCon, OPP uses the log-sum-exp function for both attraction and repulsion terms, and samples anchors from the normal class (similar to SC-OSAD), but with the important distinction that it decouples explicitly positives and negatives into two disjoint sets: normals and anomalies.

4.2 Gradient analysis of the loss

In order to give insight on the behavior of the contrastive loss, many authors provide a gradient analysis (Wang & Liu, 2021) and this derivation enables an explicit relationship between the hyperparameter τ and hard-mining over the negatives. In line with this work, we derive the gradient of the loss w.r.t to the anchor, and defer the remaining derivations w.r.t positives and negatives to Appendix B.

Proposition 4.2 (Gradient with respect to the anchor) *The gradient of l_i^{OPP} w.r.t $\mathbf{u}_i$ has the form:*

$$\nabla_{\mathbf{u}_i} l_i^{\text{OPP}} = \boldsymbol{\mu}_{\mathcal{B}_a}(i) - \boldsymbol{\mu}_{\mathcal{Q}_i}(i) \tag{15}$$

where $\boldsymbol{\mu}_{\mathcal{Q}_i}(i)$ and $\boldsymbol{\mu}_{\mathcal{B}_a}(i)$ are softmax-weighted centroids of positives and negatives:

$$\boldsymbol{\mu}_{\mathcal{Q}_i}(i) = \sum_{j\in Q_i} p_{\text{attract}}(j|i)\mathbf{u}_j, \quad p_{\text{attract}}(j|i) = \frac{\exp\big(-s_{ij}/\tau\big)}{\sum\limits_{k\in Q_i}\exp\big(-s_{ik}/\tau\big)}, \qquad j \in \mathcal{Q}_i \tag{16}$$

$$\boldsymbol{\mu}_{\mathcal{B}_a}(i) = \sum_{j\in\mathcal{B}_a} p_{\text{repel}}(j|i)\mathbf{u}_j, \quad p_{\text{repel}}(j|i) = \frac{\exp\big(s_{ij}/\tau\big)}{\sum\limits_{k\in\mathcal{B}_a}\exp\big(s_{ik}/\tau\big)}, \qquad j \in \mathcal{B}_a \tag{17}$$

Each $p_{\text{attract}}(j|i)$ (resp. $p_{\text{repel}}(j|i)$) are anchor-conditioned probability distributions over the positives set $\mathcal{Q}_i$ (resp. negatives $\mathcal{B}_a$). As $\tau \to 0^+$, $p_{\text{repel}}(j|i)$ focuses on the negatives most similar to i (hard negatives), and $p_{\text{attract}}(j|i)$ focuses on positives with low similarity to i (hard positives), highlighting that the choice of τ controls the hard-mining sensitivity. This behavior is different to that of SupCon, which treats all positives uniformly and only applies hard mining to the negatives. We detail the gradient of SupCon w.r.t $\mathbf{u}_i$ and discuss its differences with the gradient of OPP in Appendix B.

5 Practical extensions

Building on this formulation of the OPP loss, we propose two variants: (i) the τ-balanced OPP loss that uses two distinct temperatures for the normals and anomalies, directly calibrated through the bound in equation 12 to account for class imbalance, (ii) a regularized OPP loss that saturates attraction and repulsion term to mitigate overfitting to training anomalies.

5.1 τ-balanced OPP

Normals and anomalies differ in their numbers and distributions, therefore it may be desirable to deal with the attraction and repulsion terms of the OPP loss differently. By using different values of τ for each term, we instead get the following loss function:

$$l_i^{\tau\text{-OPP}} = \tau_n \text{LSE}\{-s_{iq}/\tau_n\}_{q\in\mathcal{Q}_i} + \tau_a \text{LSE}\{s_{ia}/\tau_a\}_{a\in\mathcal{B}_a}. \qquad (\tau\textbf{-OPP})$$

The key difference with base OPP is that the choice of the tightness of the approximation of l_i^{Margin} can be done independently for each class, also affecting hard-mining differently for each class. We outline a way to chose the values of τ_n and τ_a by considering the relative hard-mining effect desired and the initial temperature hyperparameter of OPP. Let τ be the original OPP temperature and $\lambda = \frac{\tau_n}{\tau_a}$. We use the bound in 13 to estimate and balance the approximations made in base OPP and τ-OPP:

$$\begin{aligned} \tau(\log|\mathcal{B}_a| + \log(|\mathcal{B}_n| - 1)) &= \tau_a \log|\mathcal{B}_a| + \tau_n \log(|\mathcal{B}_n| - 1) \\ &= \tau_a(\log|\mathcal{B}_a| + \lambda \log(|\mathcal{B}_n| - 1)) \end{aligned} \tag{18}$$

which gives

$$\tau_a = \frac{\log|\mathcal{B}_a| + \log(|\mathcal{B}_n| - 1)}{\log|\mathcal{B}_a| + \lambda \log(|\mathcal{B}_n| - 1)}\tau, \quad \tau_n = \lambda\tau_a. \tag{19}$$

Given a choice of τ and λ, these parameters can be automatically computed for each batch during training. This also allows for consistency of the loss amplitude between batches, which can vary in heavily unbalanced cases such as open-set problems.

5.2 Regularization

Negatives (or anomalies in the case of OPP) may concentrate in directions opposite to the mean of the normal data. This behavior can be interpreted as a form of overfitting to the known anomalies. In order to mitigate this effect, we introduce a regularization mechanism that saturates the effect of negatives once they are sufficiently far from positives. Concretely, this is accomplished by defining a hyperparameter δ_a that amounts to the minimum tolerated similarity with the anchor and adding it to LSE operator. This is equivalent to adding a fixed vector $\mathbf{u}_a^{\min}$ to the repel term of the OPP loss, with $\mathbf{u}_i \cdot \mathbf{u}_a^{\min} = \delta_a$. When $s_{ia} < \delta_a$, this fixed term dominates the LSE. Similarly, we define a fixed hyperparameter δ_q which acts as a saturation level for the normal samples. Again, this amounts to adding a fixed vector $\mathbf{u}_q^{\max}$ to the attract term of the OPP loss, with $\mathbf{u}_i \cdot \mathbf{u}_q^{\max} = \delta_q$. These modifications yield the regularized OPP loss hereafter referred to as r-OPP:

$$l_i^{\text{r-OPP}} = \tau_n \text{LSE}\left(\{-s_{iq}/\tau_n\}_{q\in\mathcal{Q}_i} \cup \{-\delta_q/\tau_n\}\right) + \tau_a \text{LSE}\left(\{s_{ia}/\tau_a\}_{a\in\mathcal{B}_a} \cup \{\delta_a/\tau_a\}\right) \qquad (\textbf{r-OPP})$$

In the case of r-OPP, the expression of τ_a and τ_n shown in equation 19 becomes:

$$\tau_a = \frac{\log(|\mathcal{B}_a| + 1) + \log(|\mathcal{B}_n|)}{\log(|\mathcal{B}_a| + 1) + \lambda \log(|\mathcal{B}_n|)}\tau, \quad \tau_n = \lambda\tau_a. \tag{20}$$

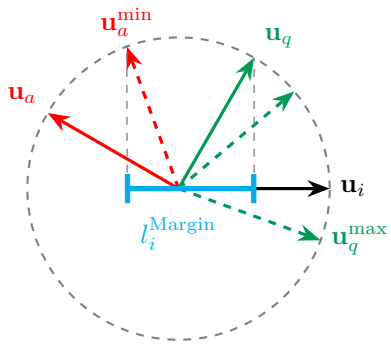

Figure 2: Illustration of the regularization term in r-OPP. $\delta_a = \mathbf{u}_i \cdot \mathbf{u}_a^{\min}$ places a lower bound on the anomalous term of the margin. In the same manner, $\delta_q = \mathbf{u}_i \cdot \mathbf{u}_q^{\max}$ serves as an upper bound on the normal term of the margin.

5.3 Anomaly scoring

Contrastive learning provides a representation-learning framework and can be combined with other downstream tasks. In our case, we want to evaluate the quality of the learned representations on anomaly detection. Our loss shapes the representation space in such a way that normal samples are represented close together and that abnormal representations lie far away from the normal cluster. By making the assumption that the distribution of normal samples used for training is a good estimator for the real normal distribution, we can use the cosine-similarity to measure the distance between the normalized representation of a test sample and the normalized mean representation of normal samples. Thus, we define the anomaly score associated with a test sample $\mathbf{u}_{\text{test}}$ as:

$$s(\mathbf{u}_{\text{test}}) = 1 - \mathbf{u}_{\text{test}} \cdot \frac{\boldsymbol{\mu}_{\text{normal}}}{\|\boldsymbol{\mu}_{\text{normal}}\|_2} \tag{21}$$

where $\boldsymbol{\mu}_{\text{normal}} = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \mathbf{u}_n$ is the average representation of $|\mathcal{N}|$ reference normal samples taken from the training set. We investigate the effect of the composition and cardinality of the set of reference samples $\mathcal{N}$ in the ablation study, and use $|\mathcal{N}| = 64$ random normal samples for the results of tables 4 and 5.

6 Numerical Experiments

This section presents the different experiments carried out in order to compare our proposed methods with (i) other contrastive learning based methods, and (ii) methods taken from the supervised AD literature, in the OSAD setting. We begin by outlining the different experimental settings, before describing the datasets and finally the results we obtain.

6.1 Openness settings

The key aspect of OPP is to leverage a small amount of known anomalies and a large amount of normal data to learn an embedding space that allows for generalization to unseen anomaly classes. To assess this generalization capability, we first quantify the *openness* of an anomaly detection problem as,

$$\mathcal{O} = 1 - \frac{\mathcal{C}_{\text{train}}}{\mathcal{C}_{\text{test}} \times (1 - \log_{10} \frac{n}{N})} \tag{22}$$

where $\mathcal{C}_{\text{train}}$ and $\mathcal{C}_{\text{test}}$ respectively are the number of anomaly classes represented during training and testing, and we have defined the imbalance in the training set as the ratio of the number of training anomalies n to the number of training normal samples N. A completely closed-set problem involves a balanced training set where all anomalies are represented, yielding an openness of $\mathcal{O} = 0$. An typical OSAD problem that involves a small number of known anomaly classes and high class imbalance has an openness that tends to $\mathcal{O} = 1$.

In line with previous OSAD studies (Ding et al., 2022), we perform the evaluation of the different methods on a **general setting** and a **hard setting**, in a *low* and a *high* imbalance case.

General setting. It consists of a scenario comparable to real world conditions, where we have access to examples of all possible classes of anomalies. In this setting, we randomly sample anomalies from all possible anomaly classes and include them in the training set, which initially is only composed of normal samples. These samples are then excluded from the test set.

Hard setting. It is designed to target the open-set capabilities of the evaluated method, by sampling anomalies from a single known anomaly class and testing on normal samples and the held-out anomaly classes unseen during training. The known anomaly class is excluded from testing.

These two settings change the openness through the number of anomaly classes seen during training and testing. The low and high imbalance cases allow further variation of the openness through class imbalance. Choosing the number of anomaly examples to include in the training set for each imblance case is not trivial. In general, the openness depends in particular on $\mathcal{C}_{\text{train}}$, the number of different anomaly classes represented in the anomaly examples samples from the test set. As this is done randomly, we use the expected number of represented anomaly classes to compute the openness. Considering $\mathcal{C}_{\text{test}}$ anomaly classes each consisting of m_i elements, the expected number of classes represented when sampling n elements from the $M = \sum_{i=1}^{\mathcal{C}_{\text{test}}} m_i$ anomalies is given by

$$\mathbb{E}[\mathcal{C}_{\text{train}}] = \sum_{i=1}^{\mathcal{C}_{\text{test}}} \left(1 - \frac{\binom{M-m_i}{n}}{\binom{M}{n}} \right). \tag{23}$$

To choose the number of anomalies to use in the general setting, we use the openness of the MVTecAD datasets with one (high imbalance case) and ten (low imbalance case) anomalies, as was done in previous studies, as a baseline and attempt to harmonize the openness across datasets. The number of known anomalies chosen for each dataset in each case are presented in table 2.

Table 2: Number of anomalous/normal examples included in the training set in both imbalance cases for each dataset. The number of normal samples indicated for the MVTecAD dataset is an average over all its subsets.

Dataset	Low imbalance case	High imbalance case
MVTecAD	10/254	1/254
Mastcam	36/9302	3/9302
RealIAD-Zipper	20/2074	2/2074

6.2 Datasets

We chose defect datasets which, as mentionned above, are better suited to model the OSAD problem than typical classification datasets. For the experiments described hereafter, we use two industrial defect datasets: MVTec AD (Bergmann et al., 2019) and the *Zipper* subset of RealIAD (Wang et al., 2024), referred to as RealIAD-Zipper in the following, as well as Mastcam (Kerner et al., 2020), a collection of multispectral images of the martian surface. The *Toothbrush* MVTec AD subset contains too few images and only a single anomaly class, so we exclude it from this work. The Mastcam images are upscaled and converted to RGB for compatibility with our architecture by assigning channels 2 (805 nm), 0 (527 nm), and 1 (447 nm) to red, green and blue respectively. Appendix C presents a summary of the composition of each dataset as well as images sampled from the normal and anomaly classes in table 10 and figure 8 in respectively.

6.3 Comparison with other contrastive learning based algorithms in the OSAD setting

To ensure consistency with previous contrastive learning studies, we follow the protocol outlined by Chen et al. (2020) and Khosla et al. (2020). It consists of three modules:

Augmentation module. As explained in section 3.1, the first step is to construct a multi-viewed batch through a number (two in our case) of realizations of a fixed set of transformations. We use a combination of random rotations, flips and color jitters as augmentations.

Feature extractor. A feature extractor projects each element of a multi-viewed batch onto the D-dimensional hypersphere. The architecture that we use in this study is a ResNet-18 (He et al., 2016) convolutional feature extractor that produces $D = 512$-dimensional features which we l_2-normalize. We initialize the ResNet-18 using the weights obtained through training on ImageNet.

Projection head. The normalized latent representations are mapped to a smaller, $d = 128$-dimensional space by a single linear layer where they are l_2-normalized. Chen et al. (2020) show that this projection head improves the quality of the representations of the layers before it. The projection head is discarded after training.

We train our architecture for 10 epochs with the Adam optimizer (Kingma & Ba, 2015) using a batch size of 64 for the MVTecAD datasets (due to their smaller size) and 256 for Mastcam and RealIAD-Zipper, and a fixed learning rate of 10^{-4}. The hyperparameters used in the OPP variants were fixed following the protocol outlined in section 7.4 are summarized in table 3. Tables 4 and 5 summarize the results for the general and hard setting respectively. We present the Area Under the Receiver-Operating Characteristic curve (AUROC) achieved by each method on the test set of each dataset. The mean and standard deviations are obtained using 3 (Mastcam and RealIAD-Zipper) or 6 (MVTecAD, due to the smaller dataset size) random seeds.

Table 3: Hyperparameter values used for each dataset.

	MVTecAD	Mastcam	RealIAD-Zipper
τ	0.1	4.0	0.6
λ	0.1	1.0	1.5
δ_a	-0.5	-0.8	-0.2
δ_q	0.7	0.8	0.2

General setting. The OPP loss is consistently among the two best performing losses on all three datasets. Of the three variants of OPP, results tend to favor r-OPP in both imbalance cases, with the exception of the Mastcam dataset in the high imbalance case where OPP/τ-OPP (on this dataset, they are equivalent since $\lambda = 1$) is best. OPP is outperformed only in the high imbalance setting on the Mastcam and RealIAD-Zipper datasets, by SC-OSAD. Despite the downsides of this loss, it highlights the importance of not enforcing similarity between negatives. On top of this, it falls short on the MVTecAD dataset where a smaller batch size was used, confirming that NPC impacts learning more in small-batch regime. Moreover, contrastive objectives that induce clustering of the anomalies perform the worst. Indeed, SupMin, that employs supervision only on the minority class (i.e. anomalies), and SupProto, that attracts anomalies to their class prototype, have the worst results across all datasets. Our variation of SupMin, SupMaj, emphasizes this point as it significantly outperforms SupMin by applying supervision on the majority class and not on the minority. Finally, the Mean-shifted contrastive loss, which originally is applied in a self-supervised settings, gives average results in the supervised case and does not seem to be a strong option in the OSAD setting.

Hard setting. Again, the OPP loss shows consistent performance accross both imbalance cases. In this setting, performance still slightly favors the r-OPP variant, showing that the extensions of OPP generally improve upon the generalization capabilities of base OPP. On the RealIAD-Zipper dataset in the high imbalance case, SC-OSAD and SupMaj demonstrate the best performance, reinforcing the previous conclusion that loose anomaly representations benefit the detection of unknown anomalies more than clustered ones. Fully supervised contrastive losses or variants that employ supervision on the anomaly class perform worse overall and seem to show more variability in performance with respect to the seen anomaly type. This is best illustrated on the Mastcam and MVTecAD datasets in the low imbalance case where the class imbalance plays a less significant role.

Table 4: General setting: AUROC results (mean±std) in %. Best results are in red and second-best results are in blue. Our contributions are highlighted in bold font.

Dataset	$\mathcal{O}(\%)$	SupCon	SC-OSAD	SupMin	**SupMaj**	SupProto	MSC	**OPP**	**τ-OPP**	**r-OPP**
					High imbalance case					
Carpet	94.2	84.40 ± 7.20	80.04 ± 4.86	44.26 ± 13.29	84.40 ± 7.20	35.32 ± 11.34	49.84 ± 6.58	91.54 ± 6.83	92.11 ± 6.54	90.21 ± 8.73
Grid	94.2	48.84 ± 6.43	43.10 ± 4.51	32.68 ± 2.05	48.84 ± 6.43	32.37 ± 2.49	40.32 ± 4.78	50.74 ± 7.25	49.19 ± 4.43	59.51 ± 13.74
Leather	94.1	65.83 ± 13.12	66.55 ± 10.87	52.48 ± 9.00	65.83 ± 13.12	60.03 ± 15.01	53.48 ± 22.77	78.90 ± 14.81	74.62 ± 11.71	80.23 ± 15.77
Tile	94.1	77.79 ± 4.78	76.35 ± 5.66	53.06 ± 3.01	77.79 ± 4.78	53.21 ± 2.73	79.14 ± 5.84	85.91 ± 4.74	85.53 ± 4.26	82.78 ± 5.80
Wood	94.1	80.45 ± 6.81	77.88 ± 6.64	55.93 ± 6.46	80.45 ± 6.81	56.56 ± 7.77	77.48 ± 7.78	90.11 ± 3.53	86.41 ± 2.93	88.63 ± 3.55
Bottle	90.0	99.31 ± 0.35	98.72 ± 0.48	82.73 ± 1.17	99.31 ± 0.35	84.96 ± 1.65	99.60 ± 0.69	99.68 ± 0.15	99.60 ± 0.20	99.52 ± 0.17
Capsule	94.0	61.10 ± 2.60	59.58 ± 2.26	46.93 ± 6.97	61.10 ± 2.60	45.21 ± 4.59	56.86 ± 5.96	57.12 ± 4.89	58.09 ± 4.93	60.55 ± 6.99
Pill	95.8	64.68 ± 3.53	64.72 ± 3.16	65.41 ± 2.10	64.68 ± 3.53	62.08 ± 3.43	64.07 ± 2.42	62.07 ± 8.83	63.35 ± 5.92	67.78 ± 6.04
Transistor	92.5	88.70 ± 1.92	89.23 ± 2.00	69.50 ± 3.23	88.70 ± 1.92	67.70 ± 3.36	75.94 ± 4.08	90.05 ± 2.60	89.02 ± 1.57	90.44 ± 3.71
Zipper	95.8	85.85 ± 1.83	87.09 ± 1.73	91.18 ± 1.26	85.85 ± 1.83	89.76 ± 2.94	83.76 ± 3.54	90.26 ± 2.21	88.65 ± 1.74	88.29 ± 2.95
Cable	96.3	84.93 ± 3.14	84.55 ± 2.18	75.88 ± 1.99	84.93 ± 3.14	76.14 ± 1.79	81.01 ± 4.54	84.16 ± 3.50	84.76 ± 3.69	82.54 ± 4.62
Hazelnut	93.0	94.54 ± 3.00	88.90 ± 3.75	68.41 ± 1.69	94.54 ± 3.00	68.14 ± 2.61	88.07 ± 7.99	94.20 ± 4.64	93.21 ± 5.05	97.57 ± 2.14
Metal-Nut	92.5	79.45 ± 3.41	78.71 ± 1.84	68.40 ± 2.53	79.45 ± 3.41	67.23 ± 3.83	74.51 ± 1.54	83.29 ± 3.31	80.10 ± 4.06	83.62 ± 2.87
Screw	93.3	52.38 ± 6.27	50.88 ± 5.39	32.52 ± 3.68	52.38 ± 6.27	33.98 ± 7.04	46.61 ± 8.02	48.58 ± 4.78	47.70 ± 2.53	48.72 ± 2.95
MVTecAD	93.9	76.30 ± 4.60	74.74 ± 3.95	59.96 ± 4.17	76.30 ± 4.60	59.48 ± 5.04	69.34 ± 6.18	79.04 ± 5.15	78.02 ± 4.25	80.03 ± 5.72
Mastcam	93.0	66.27 ± 4.53	71.20 ± 4.73	43.50 ± 2.09	62.72 ± 0.45	41.63 ± 0.23	60.41 ± 6.17	69.58 ± 1.55	69.58 ± 1.55	62.97 ± 9.00
RealIAD-Zipper	89.3	67.43 ± 0.45	74.59 ± 1.59	30.05 ± 1.99	68.03 ± 1.10	31.57 ± 1.90	61.02 ± 2.98	60.74 ± 4.35	66.77 ± 6.88	69.56 ± 3.89
					Low imbalance case					
Carpet	62.9	95.09 ± 4.16	93.23 ± 2.89	60.09 ± 11.50	98.30 ± 0.88	70.76 ± 15.41	51.86 ± 7.10	93.02 ± 6.74	94.65 ± 5.73	93.32 ± 8.04
Grid	62.2	92.27 ± 8.21	85.12 ± 12.57	32.19 ± 3.46	87.88 ± 14.50	33.20 ± 2.66	47.79 ± 4.74	97.67 ± 1.16	90.61 ± 6.01	94.51 ± 4.31
Leather	62.1	100 ± 0	99.37 ± 0.66	67.10 ± 14.50	99.99 ± 0.02	87.94 ± 4.75	91.16 ± 14.04	100 ± 0	100 ± 0	100 ± 0
Tile	61.5	97.04 ± 1.85	94.06 ± 1.16	59.79 ± 2.99	97.44 ± 1.74	62.51 ± 3.96	89.23 ± 12.42	97.85 ± 1.66	97.28 ± 2.30	96.56 ± 2.39
Wood	61.8	97.60 ± 0.78	95.04 ± 2.20	59.68 ± 10.45	98.26 ± 0.50	65.49 ± 11.29	94.37 ± 6.33	98.46 ± 0.69	98.65 ± 0.60	97.03 ± 1.36
Bottle	57.3	99.73 ± 0.26	99.39 ± 0.37	92.44 ± 1.77	99.65 ± 0.20	91.48 ± 1.13	99.92 ± 0.18	99.61 ± 0.28	99.73 ± 0.16	99.47 ± 0.33
Capsule	61.4	75.52 ± 6.44	71.18 ± 5.73	45.80 ± 3.94	65.20 ± 4.29	45.14 ± 2.29	60.31 ± 11.96	77.15 ± 3.30	70.09 ± 2.38	76.79 ± 3.08
Pill	67.2	84.98 ± 2.73	85.11 ± 2.73	63.59 ± 3.23	79.84 ± 3.33	64.11 ± 2.89	68.40 ± 5.67	75.90 ± 6.37	79.37 ± 3.38	79.93 ± 5.57
Transistor	58.6	89.77 ± 1.90	90.87 ± 2.66	71.13 ± 2.66	90.66 ± 2.75	73.71 ± 2.92	83.43 ± 3.88	89.19 ± 1.27	91.48 ± 1.39	90.68 ± 1.22
Zipper	66.4	87.72 ± 2.12	88.68 ± 2.35	89.03 ± 3.55	89.15 ± 2.67	90.62 ± 2.32	86.64 ± 3.79	93.22 ± 3.11	91.03 ± 3.19	93.21 ± 2.71
Cable	67.8	90.13 ± 1.77	89.09 ± 2.11	77.16 ± 2.15	90.13 ± 0.88	78.25 ± 1.37	82.84 ± 6.68	90.36 ± 1.56	91.41 ± 2.24	90.43 ± 1.42
Hazelnut	63.2	100 ± 0	100 ± 0	84.71 ± 3.78	100 ± 0	78.52 ± 4.16	97.10 ± 4.13	99.91 ± 0.18	99.92 ± 0.17	99.82 ± 0.26
Metal-Nut	59.3	95.65 ± 1.59	88.09 ± 5.59	71.93 ± 6.46	92.0 ± 2.88	72.83 ± 6.97	81.90 ± 3.05	94.24 ± 2.06	93.64 ± 2.17	89.67 ± 3.27
Screw	63.9	70.22 ± 7.30	70.71 ± 7.47	34.30 ± 5.64	73.25 ± 6.03	36.17 ± 5.01	50.00 ± 13.00	70.37 ± 7.68	63.77 ± 3.05	73.75 ± 12.79
MVTecAD	62.5	91.12 ± 2.79	89.28 ± 3.46	64.92 ± 5.43	90.13 ± 2.91	67.91 ± 4.80	77.50 ± 6.93	91.21 ± 2.58	90.12 ± 2.34	91.08 ± 3.34
Mastcam	73.3	83.48 ± 1.83	86.81 ± 2.41	42.19 ± 2.09	80.68 ± 3.27	41.47 ± 2.20	63.69 ± 8.10	84.53 ± 7.30	84.53 ± 7.30	90.22 ± 0.26
RealIAD-Zipper	67.5	83.13 ± 5.10	82.74 ± 7.57	29.89 ± 2.99	82.57 ± 9.57	30.55 ± 4.02	66.64 ± 6.39	86.53 ± 4.52	88.07 ± 2.25	85.71 ± 4.57

6.4 Comparison with supervised AD algorithms in the OSAD setting

In this section, we place the OPP loss within the context of the supervised AD literature. The goal of this comparison is to show that learning an appropriate representation space is enough to generalize to unseen anomaly classes. To do this, we chose to perform this comparison in the hard setting. However, taking too few anomaly examples can be too penalizing and renders comparisons less meaningful. Therefore, we chose the hard setting in the low imbalance case (i.e. the same configuration as in table 5). We restrict our comparison to the Mastcam and RealIAD-Zipper datasets as the subsets of the MVTecAD dataset contain too few images, hindering the quality of comparisons.

Based on the results of the previous experiments, we chose to compare the performance of the r-OPP variant with DevNet, DRA, DeepSAD, HSC and BGAD when applicable (as it needs pixel level anomaly annotations). We use a pre-trained ResNet-18 backbone in all methods (except for BGAD which has its own dedicated architecture). We use the Adam optimizer for all simulations, with a weight decay of 10^{-6}. All models were trained for 10 epochs. r-OPP, DevNet, DeepSAD and HSC were trained using a batch size of 256. DRA and BGAD were trained respectively using a batch size of 48 and 32 as done in their original publications. We use the evaluation protocol of the hard setting and report the results in table 6.

The r-OPP variant is a close second to the state-of-the-art augmentation based method DRA on the Mastcam dataset and behind BGAD and DRA on RealIAD-Zipper. Mastcam is more difficult to model than RealIAD-Zipper for DRA due to the nature of the method used by DRA, which focuses on structural anomalies through augmentations such as CutMix. Though the anomalies included in the Mastcam dataset are also structural, they are more subtle and the quality of the images is significantly worse. On the other hand, since we do not employ any specific anomaly learning and delegate all the discrimination to the feature encoder, r-OPP shows less dependence on the dataset. This highlights a potential complementarity between feature focused and augmentation focused methods.

Table 5: Hard setting: AUROC results (mean±std) in %. Best results are in red and second-best results are in blue. Our contributions are highlighted in bold font.

Dataset	$\mathcal{O}(\%)$	SupCon	SC-OSAD	SupMin	**SupMaj**	SupProto	MSC	**OPP**	**τ-OPP**	**r-OPP**
High imbalance case										
Carpet										
Cut		91.02 ± 3.31	80.53 ± 4.81	40.05 ± 11.58	91.02 ± 3.13	27.89 ± 4.99	47.28 ± 3.91	95.35 ± 2.87	95.68 ± 2.12	95.74 ± 2.04
Hole		87.81 ± 5.18	78.17 ± 4.98	43.53 ± 12.70	87.81 ± 5.18	29.16 ± 9.08	49.00 ± 3.97	92.17 ± 4.56	93.08 ± 3.19	87.74 ± 10.31
Color	94.2	82.59 ± 8.42	78.82 ± 5.84	39.40 ± 14.52	82.59 ± 8.42	23.83 ± 6.87	50.64 ± 5.08	89.69 ± 2.43	91.99 ± 1.70	87.03 ± 3.70
Metal Cont.		81.01 ± 5.98	78.67 ± 3.72	37.94 ± 12.94	81.01 ± 5.98	24.26 ± 5.99	52.77 ± 4.99	81.11 ± 16.14	84.39 ± 14.16	76.94 ± 18.79
Thread		79.39 ± 11.34	77.17 ± 7.76	42.45 ± 12.82	79.39 ± 11.34	30.13 ± 10.55	45.65 ± 4.47	85.07 ± 9.81	86.84 ± 9.09	78.84 ± 15.66
Mean		84.36 ± 6.85	78.67 ± 5.42	40.67 ± 12.91	84.36 ± 6.81	27.05 ± 7.50	49.07 ± 4.48	88.68 ± 7.16	90.39 ± 6.05	85.26 ± 10.10
Metal-Nut										
Bent		83.85 ± 1.54	83.51 ± 2.14	74.48 ± 2.34	83.85 ± 1.54	74.77 ± 2.50	78.29 ± 2.27	84.74 ± 2.67	83.98 ± 1.55	86.84 ± 2.69
Color	94.0	82.28 ± 3.51	82.01 ± 2.76	69.85 ± 3.85	82.28 ± 3.51	71.20 ± 3.71	73.30 ± 3.52	80.23 ± 4.64	81.13 ± 4.79	80.60 ± 6.44
Flip		74.02 ± 5.28	71.92 ± 2.93	58.41 ± 2.77	74.02 ± 5.28	57.45 ± 3.01	61.47 ± 6.96	75.55 ± 3.75	75.30 ± 3.13	75.87 ± 4.51
Scratch		79.19 ± 2.14	75.98 ± 2.54	69.82 ± 2.55	79.19 ± 2.14	68.91 ± 1.92	72.11 ± 3.78	80.12 ± 4.13	79.83 ± 4.72	81.19 ± 3.68
Mean		79.83 ± 3.12	78.36 ± 2.59	68.14 ± 2.88	79.83 ± 3.12	68.08 ± 2.79	71.29 ± 4.13	80.16 ± 3.80	80.06 ± 3.55	81.13 ± 4.33
Mastcam										
Meteorite		58.60 ± 5.87	61.09 ± 4.56	43.38 ± 8.85	59.11 ± 8.83	42.76 ± 6.95	45.26 ± 4.74	57.02 ± 10.79	57.02 ± 10.79	52.42 ± 6.17
Dump pile	97.2	50.05 ± 6.90	56.59 ± 3.69	43.59 ± 8.19	59.43 ± 2.78	45.55 ± 8.06	59.76 ± 7.02	61.40 ± 4.67	61.40 ± 4.67	61.93 ± 2.18
Float		57.39 ± 10.33	56.91 ± 10.48	50.06 ± 6.88	56.20 ± 8.61	51.17 ± 8.29	45.62 ± 10.48	63.16 ± 5.73	63.16 ± 5.73	57.90 ± 3.86
Drill hole		63.10 ± 10.15	59.58 ± 5.87	50.39 ± 5.15	56.65 ± 9.99	48.76 ± 6.11	49.75 ± 7.97	64.28 ± 1.68	64.28 ± 1.68	64.74 ± 5.77
Mean		57.29 ± 8.31	58.54 ± 6.15	46.86 ± 7.27	57.85 ± 7.55	47.06 ± 7.35	50.10 ± 7.55	61.47 ± 5.72	61.47 ± 5.72	59.25 ± 4.50
RealIAD-Zipper										
Missing Part		64.47 ± 4.10	66.71 ± 3.28	33.26 ± 3.18	66.43 ± 4.83	31.79 ± 4.82	64.51 ± 8.52	64.31 ± 9.88	62.71 ± 5.57	63.41 ± 3.40
Deformation	93.9	72.56 ± 2.74	72.45 ± 4.47	37.44 ± 2.95	73.58 ± 3.45	34.99 ± 3.29	59.71 ± 6.18	63.17 ± 2.39	62.10 ± 2.16	63.68 ± 2.69
Damage		73.05 ± 4.08	75.71 ± 4.84	35.15 ± 5.11	72.69 ± 6.44	34.44 ± 2.32	55.47 ± 2.65	70.03 ± 9.93	71.23 ± 8.59	71.87 ± 7.28
Contamination		70.41 ± 1.93	71.05 ± 2.95	37.57 ± 2.32	70.45 ± 2.00	35.60 ± 2.13	65.95 ± 6.20	64.03 ± 1.46	63.30 ± 1.42	64.13 ± 1.92
Mean		70.12 ± 3.21	71.48 ± 3.89	35.86 ± 3.39	70.79 ± 4.18	34.21 ± 3.14	61.41 ± 5.89	65.39 ± 5.92	64.84 ± 4.44	65.77 ± 3.82
Low imbalance case										
Carpet										
Cut		94.68 ± 1.56	93.44 ± 2.97	77.79 ± 5.61	96.19 ± 0.95	73.14 ± 9.13	66.39 ± 14.16	95.61 ± 1.17	95.63 ± 1.41	95.35 ± 1.62
Hole		95.05 ± 1.21	94.37 ± 1.52	80.14 ± 9.02	96.48 ± 0.89	74.27 ± 9.65	55.21 ± 16.38	94.32 ± 1.37	94.54 ± 1.33	95.22 ± 1.68
Color	91.8	85.26 ± 5.97	86.28 ± 8.59	48.26 ± 14.31	93.00 ± 3.41	54.54 ± 24.36	55.64 ± 9.53	89.38 ± 5.42	85.00 ± 6.30	86.78 ± 5.39
Metal Cont.		79.29 ± 11.41	78.51 ± 4.22	49.77 ± 14.31	92.25 ± 4.11	58.18 ± 19.07	49.93 ± 5.40	73.20 ± 14.49	78.31 ± 11.57	79.71 ± 10.90
Thread		98.41 ± 0.37	93.91 ± 4.57	61.86 ± 12.04	98.93 ± 0.41	67.16 ± 13.54	61.60 ± 4.67	98.02 ± 1.07	99.03 ± 0.46	98.29 ± 0.76
Mean		90.54 ± 4.10	89.30 ± 4.37	63.56 ± 11.06	95.37 ± 1.95	65.46 ± 15.15	57.75 ± 10.03	90.11 ± 4.70	90.50 ± 4.21	91.07 ± 4.07
Metal-Nut										
Bent		94.32 ± 2.06	89.77 ± 2.19	77.12 ± 4.38	93.10 ± 2.36	78.00 ± 4.12	69.25 ± 10.04	94.44 ± 1.87	93.33 ± 0.81	90.76 ± 3.58
Color	91.5	89.97 ± 2.47	85.98 ± 1.80	71.23 ± 4.56	85.98 ± 3.67	76.81 ± 4.96	75.87 ± 3.46	91.29 ± 4.45	89.91 ± 4.59	87.26 ± 3.31
Flip		69.16 ± 3.53	71.55 ± 6.24	59.27 ± 6.19	70.04 ± 5.07	60.80 ± 7.49	73.40 ± 2.11	73.87 ± 5.10	73.24 ± 4.36	70.22 ± 4.62
Scratch		90.53 ± 2.13	88.46 ± 2.39	72.89 ± 4.84	90.47 ± 2.32	72.61 ± 5.39	82.07 ± 7.43	90.45 ± 2.62	91.44 ± 2.58	90.05 ± 2.15
Mean		86.00 ± 2.55	83.94 ± 3.16	70.13 ± 4.99	84.90 ± 3.36	72.06 ± 5.49	75.15 ± 5.76	87.51 ± 3.51	86.98 ± 3.09	84.57 ± 3.42
Mastcam										
Meteorite		66.33 ± 3.63	69.98 ± 0.95	41.30 ± 9.70	60.13 ± 3.15	42.44 ± 8.71	54.12 ± 11.83	66.58 ± 4.73	66.58 ± 4.73	68.76 ± 2.46
Dump pile	96.3	74.23 ± 2.70	68.67 ± 2.55	47.00 ± 9.80	67.34 ± 3.50	43.63 ± 10.26	55.86 ± 7.73	67.39 ± 0.74	67.39 ± 0.74	69.48 ± 3.82
Float		53.21 ± 5.81	66.35 ± 3.87	49.97 ± 5.54	64.46 ± 2.72	49.67 ± 10.31	59.71 ± 5.98	66.59 ± 6.38	66.59 ± 6.38	70.50 ± 1.32
Drill hole		71.40 ± 5.46	63.81 ± 8.68	54.35 ± 3.07	59.54 ± 7.28	50.53 ± 5.69	60.52 ± 7.96	69.33 ± 2.18	69.33 ± 2.18	69.58 ± 8.32
Mean		66.29 ± 4.40	67.20 ± 4.01	48.16 ± 7.03	62.87 ± 4.16	45.57 ± 8.74	57.55 ± 8.38	67.47 ± 3.51	67.47 ± 3.51	69.58 ± 3.98
RealIAD-Zipper										
Missing Part		72.34 ± 6.70	71.40 ± 3.14	29.34 ± 5.27	74.90 ± 3.75	32.70 ± 6.55	57.49 ± 3.77	77.77 ± 2.63	80.34 ± 3.82	80.93 ± 1.94
Deformation	91.9	76.20 ± 3.81	71.15 ± 2.15	34.81 ± 4.35	69.04 ± 2.73	34.57 ± 4.17	63.04 ± 8.18	73.46 ± 3.69	71.79 ± 6.22	73.46 ± 3.57
Damage		75.62 ± 2.72	74.76 ± 4.16	35.80 ± 2.73	72.19 ± 1.79	31.36 ± 7.79	58.05 ± 6.80	76.65 ± 1.74	75.54 ± 3.22	74.69 ± 2.61
Contamination		73.75 ± 1.94	70.45 ± 2.11	32.45 ± 1.91	66.05 ± 1.51	36.26 ± 3.66	67.23 ± 7.44	71.94 ± 6.25	72.65 ± 9.60	75.55 ± 7.76
Mean		74.48 ± 3.79	71.94 ± 2.89	33.10 ± 3.57	70.55 ± 2.45	33.72 ± 5.54	61.45 ± 6.55	74.96 ± 3.58	75.08 ± 5.72	76.16 ± 3.97

Table 6: Comparison with the OSAD literature under the **hard setting**. AUROC results (mean±std) in %. Best results are in red and second-best results are in blue.

Dataset	DevNet	DRA	DeepSAD	HSC	BGAD	r-OPP
Mastcam						
Meteorite	53.13 ± 2.13	66.98 ± 0.21	56.75 ± 1.73	60.14 ± 0.22	-	61.97 ± 4.28
Dump pile	50.99 ± 4.64	77.25 ± 7.78	64.93 ± 7.40	64.93 ± 7.38	-	69.48 ± 3.82
Float	51.70 ± 9.77	69.65 ± 8.31	67.78 ± 3.01	69.42 ± 3.14	-	69.79 ± 6.48
Drill hole	48.91 ± 6.53	74.25 ± 0.72	68.03 ± 2.49	66.31 ± 2.06	-	76.58 ± 4.39
Mean	51.18 ± 5.77	72.03 ± 4.26	64.37 ± 3.66	65.20 ± 3.20	-	69.46 ± 4.74
RealIAD-Zipper						
Missing Part	57.89 ± 1.70	93.73 ± 1.99	68.26 ± 6.43	63.74 ± 17.17	97.35 ± 0.53	81.17 ± 0.75
Deformation	54.23 ± 2.16	87.95 ± 4.15	59.50 ± 3.66	62.44 ± 8.71	97.94 ± 0.21	73.74 ± 4.22
Damage	55.91 ± 0.65	91.00 ± 3.68	53.74 ± 1.74	59.04 ± 1.63	97.12 ± 0.28	75.05 ± 0.36
Contamination	46.28 ± 3.55	87.79 ± 1.52	56.34 ± 5.88	55.43 ± 6.15	96.52 ± 0.83	75.49 ± 7.86
Mean	53.58 ± 2.02	90.12 ± 2.84	59.46 ± 4.43	60.16 ± 8.42	97.23 ± 0.46	76.36 ± 3.30

7 Ablation studies

7.1 Impact of the *Push* term

In order to measure the impact of the repelling term in the OPP loss and highlight the importance of even just a few known anomalies, we compare the performance of base OPP with and without the presence of anomalies. This is also equivalent to testing OPP in an openness setting of $\mathcal{O} = 1$. Table 7 presents the results of this ablation in the hard setting on the Mastcam and RealIAD-Zipper datasets. The impact of even just a few known anomalies is made evident by the sharp gain in performance when the first anomaly examples are included during training.

Table 7: Effect of the number of anomalies under the hard setting. AUROC results (mean±std) in % averaged over the different known anomaly classes.

Dataset	Low imbalance case	High imbalance case	No anomalies
Mastcam	67.47 ± 3.51	61.47 ± 5.72	51.44 ± 1.18
RealIAD-Zipper	74.96 ± 3.58	65.39 ± 5.92	61.37 ± 0.40

7.2 Impact of the score function

We test the effect of different numbers of reference samples used to estimate the average normal representation. Figure 3 illustrates this effect by showing the result of varying this number between 32 and 1028 in the hard setting applied to the Mastcam dataset with *Drill Hole* as the known anomaly class. The same models are used as the experiment described in table 5, only inference is performed again. A lower number of reference samples yields worse results on average, with higher variance. Higher numbers have less impact as the quality of the estimate stops improving.

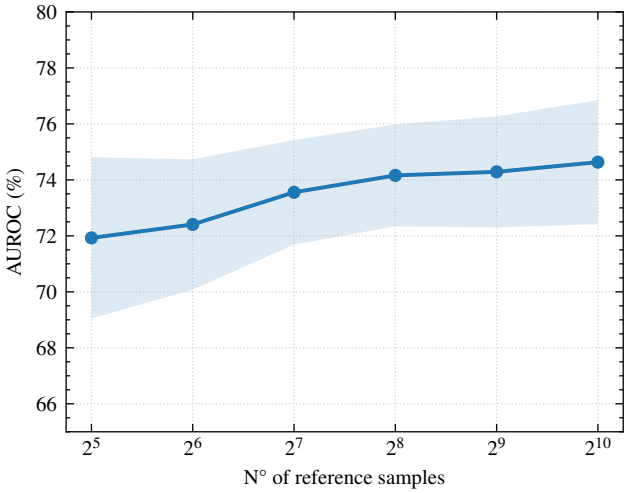

Figure 3: AUROC (%) with an increasing number of reference samples. The blue dots represent the mean of three independent simulations and their standard deviations are represented in blue shading.

7.3 Ablation of the projection head

We investigate the impact of the projection head by comparing the results obtained on the Mastcam and RealIAD-Zipper datasets in the low imbalance case of the hard setting with and without the use of the projection head during training. We report the results of this ablation in table 8. The 15 to 20 percent

decrease in performance when removing the projection head highlights its importance. A likely explanation is that the much higher dimensionality of the latent space significantly hinders representation learning.

Table 8: Ablation of the projection head under the hard setting in the low imbalance case. AUROC results (mean±std) in % averaged over the different known anomaly classes.

Dataset	With projection head	Without projection head
Mastcam	69.58 ± 3.98	51.74 ± 1.52
RealIAD-Zipper	76.16 ± 3.97	61.42 ± 0.27

7.4 Hyperparameter analysis

We perform the analysis of the effect of the hyperparameters of the OPP loss (τ, λ, δ_a and δ_q) on the Mastcam dataset. The objective function used is the AUROC evaluated on a small test set made specifically for this analysis. It is composed of 100 normal samples taken from the Mastcam validation set and of 100 anomalies taken from the eight Mastcam anomaly classes. The amount of anomalies sampled from each class is proportional to the amount available in each class. All 200 samples are removed from the original dataset for all subsequent experiments. To reduce computational cost, the training done prior to this evaluation is done on a training set composed of 2000 out of the 9302 normal training samples and 10 abnormal samples.

We use the Optuna (Akiba et al., 2019) hyperparameter optimization framework to perform this analysis. By default, the hyperparameter sampling is performed using a Tree-structured Parzen Estimator (TPE). We make 120 evaluations of the objective function. The hyperparameter sampling distributions are summarized in table 9. Figure 4 shows the results of the hyperparameter analysis. In general, OPP is only moderately

Table 9: Ranges and distributions of the hyperparameter values used

	min	max	distribution
τ	0.05	5	Log-Uniform
λ	0.1	10	Log-Uniform
δ_a	-1	0	Uniform
δ_q	0	1	Uniform

sensitive to the choice of hyperparameters. τ has the most impact on the objective function, and seems to be favorable to higher values.

8 Conclusion

In this work, we propose to tackle the Open-set Supervised Anomaly Detection problem using a contrastive learning approach. To this end, we introduce the Open-set Push-Pull loss, a supervised contrastive objective that approximates a normal-anomaly margin. It aims at grouping representations of normal samples together while pushing away the representations of anomalies to create a latent space suitable for OSAD. We establish its effectiveness for this application through extensive experiments of varying openness. By comparing it to other contrastive objectives and to previous supervised AD methods, we demonstrate that this representation learning approach is a viable alternative for OSAD problems. The simplicity of this approach allows for future studies to explore more complex approaches, through, for example, different architectures or the combination of feature learning and dataset enhancement methods.

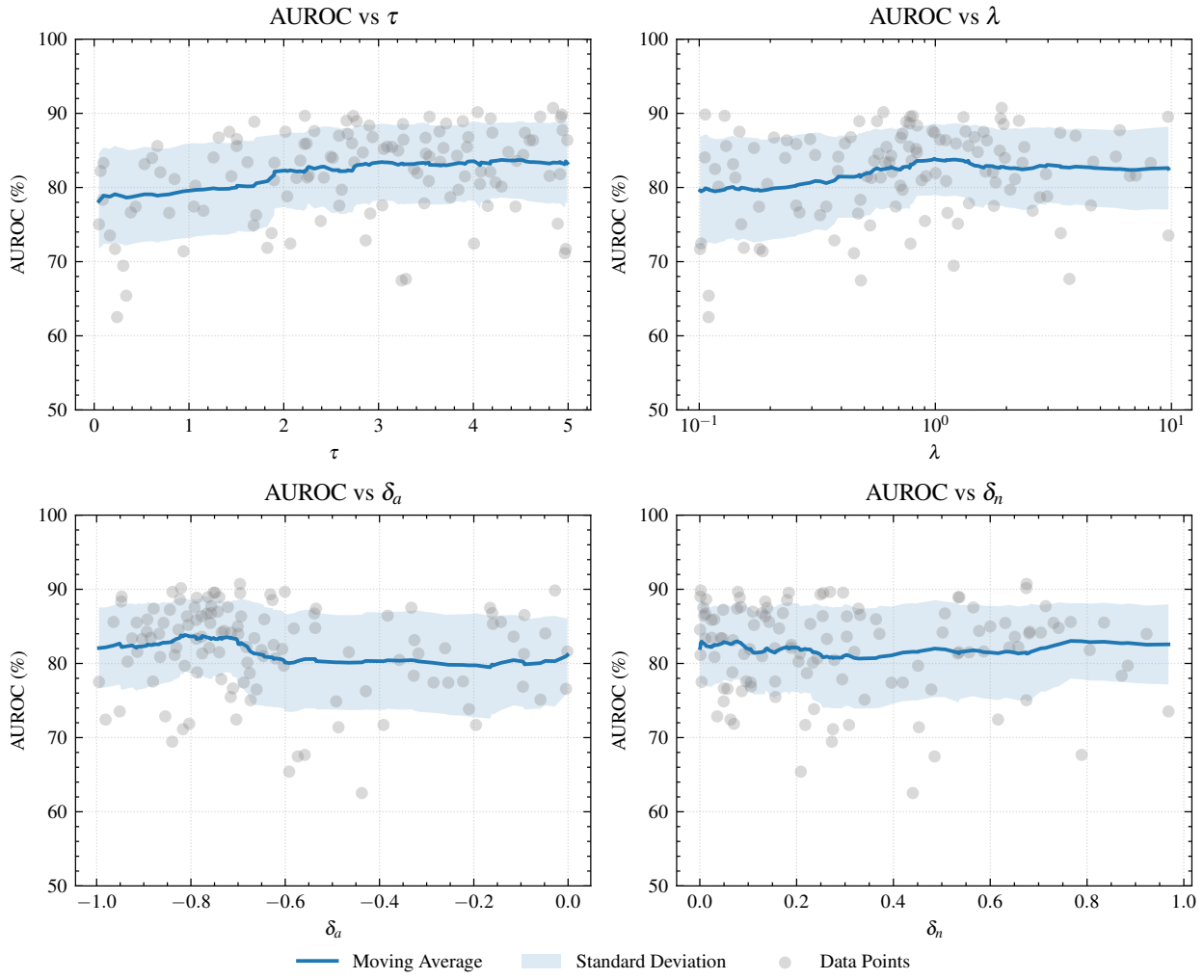

Figure 4: Hyperparameter analysis. The result of each individual simulation is shown as a gray dot. The centered rolling window average and standard deviation are obtained using a window size of 50 points and are shown as a blue line and shading. *Top-left:* AUROC vs τ shows that higher values of τ tend to yield better results. *Top-right:* AUROC vs λ shows degrading results when deviating from $\lambda = 1$. *Bottom-left:* AUROC vs δ_a indicates a slight increase in performance with lower values of δ_a. *Bottom-right:* AUROC vs δ_q displays a negligible effect of the choice of δ_q on the objective function.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

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*, pp. 9592–9600, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7388–7398, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard. Self-supervised anomaly detection in computer vision and beyond: A survey and outlook. *Neural Networks*, 172:106106, 2024.

Hannah R Kerner, Kiri L Wagstaff, Brian D Bue, Danika F Wellington, Samantha Jacob, Paul Horton, James F Bell III, Chiman Kwan, and Heni Ben Amor. Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions. *Data Mining and Knowledge Discovery*, 34(6):1642–1675, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International conference on learning representations*, 2015.

Thomas Lai, Thi Kieu Khanh Ho, and Narges Armanfard. Open-set multivariate time-series anomaly detection. In *European Conference on Artificial Intelligence*. IOS Press, October 2024.

Byeongchan Lee. Understanding self-supervised contrastive learning through supervised objectives. *Transactions on Machine Learning Research*, 2025.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6918–6928, 2022.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy J Franks, Klaus Robert Müller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022, 2022.

David Mildenberger, Paul Hager, Daniel Rueckert, and Martin J Menten. A tale of two classes: Adapting supervised contrastive learning to binary imbalanced datasets. In *Computer Vision and Pattern Recognition Conference*, pp. 10305–10314, 2025.

Guansong Pang, Chunhua Shen, and Anton Van Den Hengel. Deep anomaly detection with deviation networks. In *25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 353–362, 2019.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2155–2162, 2023.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 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*, 2020.

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

M Salehi, H Mirzaei, D Hendrycks, Y Li, MH Rohban, M Sabokrou, et al. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022.

Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 07 2001.

David M. J. Tax and Robert P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199, 1999.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Xiaoyu Yao, Zecheng He, Yanfeng Sun, and Chao Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European conference on computer vision*, pp. 668–684. Springer, 2022.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16699–16708, 2024.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

Jiawen Zhu, Choubo Ding, Yu Tian, and Guansong Pang. Anomaly heterogeneity learning for open-set supervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Illustration of the NPC effect in the SC-OSAD objective

The goal of the following is to illustrate the *negative-positive coupling* (NPC) effect that can be observed in the SC-OSAD loss using a toy example. We apply this toy example on the OPP loss as well to highlight the differences between the two objectives.

We recall the expression of the per-anchor SC-OSAD loss:

$$l_i^{\text{SC-OSAD}} = \Phi_{\text{attract}}\left(\{s_{ij}\}_{j\in\mathcal{Q}_i}\right) + \Psi_{\text{repel}}\left(\{s_{ij}\}_{j\in\mathcal{N}_i}\right) \tag{24}$$

$$= -\frac{1}{\tau}\mathbf{u}_i \cdot \bar{\mathbf{u}}_p + \text{LSE}\{s_{ij}\}_{j\in\mathcal{Q}_i} + \log(1+R_i) \tag{25}$$

wherein $R_i = \frac{\sum_{a\in\mathcal{B}_a}\exp(\frac{s_{ia}}{\tau})}{\sum_{q\in\mathcal{Q}_i}\exp(\frac{s_{iq}}{\tau})}$ and $\bar{\mathbf{u}}_p = \frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i}\mathbf{u}_p$ is the mean representation of the positives. This rewriting of $l_i^{\text{SC-OSAD}}$ directly follows from decomposing $\mathcal{N}_i = \mathcal{Q}_i \cup \mathcal{B}_a$ and $\Psi_{\text{repel}}\left(\{s_{ij}\}_{j\in\mathcal{N}_i}\right) = \text{LSE}\{s_{ij}\}_{j\in\mathcal{N}_i} = \text{LSE}\{s_{ij}\}_{j\in\mathcal{Q}_i} + \log(1+R_i)$. The conflict between the attract term and the LSE in the repel term highlights the NPC effect. To illustrate it, we consider a family of four 2-dimensional vectors on the unit-circle parameterized by their angle with respect to $\mathbf{e}_1$. The normal class is composed of two fixed vectors, $\mathbf{u}_{q1}$ and $\mathbf{u}_{q2}$, and the anchor $\mathbf{u}_i$. The abnormal class is represented by a single vector $\mathbf{u}_a$. Figure 5 depicts this toy example.

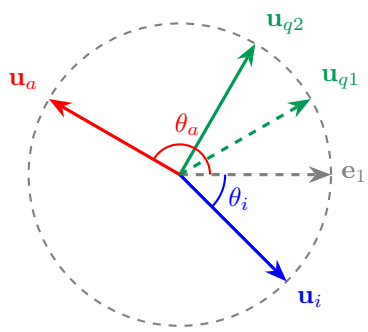

Figure 5: Schematic of the toy example

Figure 6 shows the values taken by $\bar{\Phi}_{\text{attract}} = \frac{1}{|\mathcal{B}_n|}\sum_{i\in\mathcal{B}_n}\Phi_{\text{attract}}$, $\bar{\Psi}_{\text{repel}} = \frac{1}{|\mathcal{B}_n|}\sum_{i\in\mathcal{B}_n}\Psi_{\text{repel}}$ (where $\mathcal{B}_n = \{\mathbf{u}_i, \mathbf{u}_{q1}, \mathbf{u}_{q2}\}$) and $\mathcal{L} = \bar{\Phi}_{\text{attract}} + \bar{\Psi}_{\text{repel}}$ for each $\{\theta_i, \theta_a\} \in [0, 2\pi]^2$ in the case of SC-OSAD. We can see the

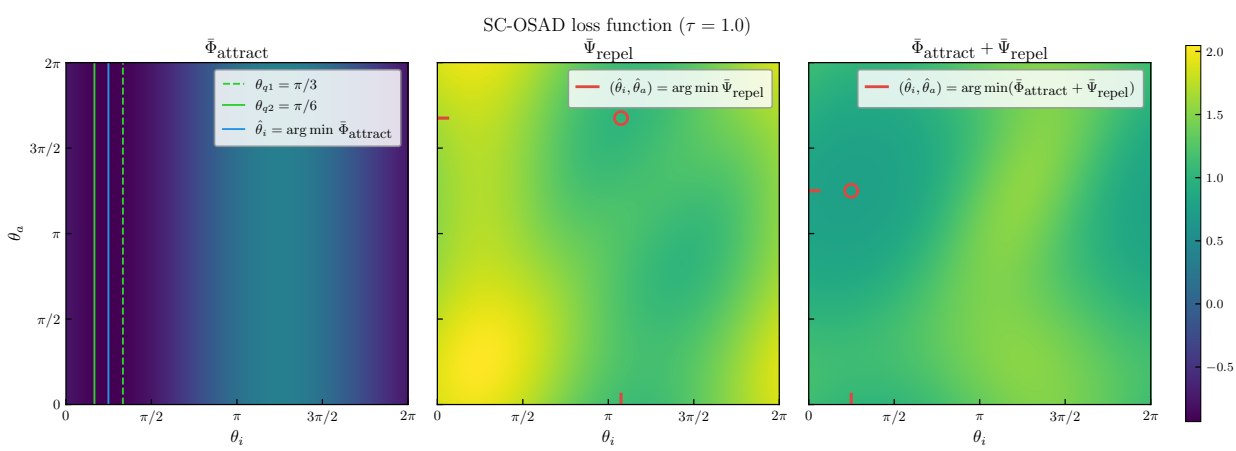

Figure 6: Heatmap of SC-OSAD with respect to θ_i and θ_a.

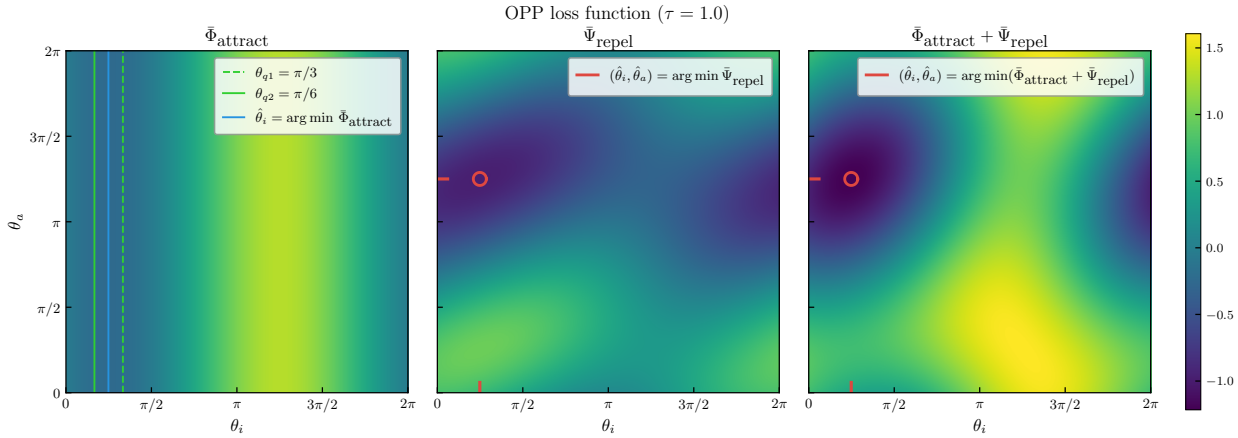

Figure 7: Heatmap of OPP with respect to θ_i and θ_a.

clear minimum of Φ_{attract} at the mean of the two fixed normal vectors in the left plot. The middle plot sheds light on the NPC effect, as Ψ_{repel} has local minimums opposite to $\mathbf{u}_{q1}$ and $\mathbf{u}_{q2}$. As the relative importance of both terms of $\mathcal{L}^{\text{SC-OSAD}}$ vary, this effect can lead to undesired local minima which may hinder the quality of the learned representations. Figure 7 depicts the same experiment with the OPP loss, which does not show any NPC effect.

B Complete derivation of the gradient of the OPP loss

We begin by recalling the expression of the Open-Set Push-Pull loss:

$$\mathcal{L}^{\text{OPP}} = \frac{1}{|\mathcal{B}_n|} \sum_{i \in \mathcal{B}_n} l_i^{\text{OPP}} \text{ with } l_i^{\text{OPP}} = \tau \text{LSE}(\{-s_{iq}/\tau\}_{q \in \mathcal{Q}_i}) + \tau \text{LSE}(\{s_{ia}/\tau\}_{a \in \mathcal{B}_a}) = \Phi_{\text{attract}} + \Psi_{\text{repel}} \tag{26}$$

where $s_{ik} = \mathbf{u}_i \cdot \mathbf{u}_k$. We are interested in the gradient with respect to the representations $\mathbf{u}$ of each sample in $\mathcal{B}_n \cup \mathcal{B}_a$, denoted $\nabla_{\mathbf{u}} \mathcal{L}^{\text{OPP}}$, we decompose it into:

$$\nabla_{\mathbf{u}} \mathcal{L}^{\text{OPP}} = \frac{1}{|\mathcal{B}_n|} \sum_{i \in \mathcal{B}_n} \nabla_{\mathbf{u}} l_i^{\text{OPP}} = \frac{1}{|\mathcal{B}_n|} \sum_{i \in \mathcal{B}_n} \nabla_{\mathbf{u}} \Phi_{\text{attract}} + \nabla_{\mathbf{u}} \Psi_{\text{repel}} \tag{27}$$

We will compute each term separately for anchors $i \in \mathcal{B}_n$, normals $q \in \mathcal{Q}_i$ and anomalies $a \in \mathcal{B}_a$.

$$\begin{aligned}
\nabla_{\mathbf{u}_i} \Phi_{\text{attract}} &= \tau \sum_{q \in \mathcal{Q}_i} \frac{-\frac{\mathbf{u}_q}{\tau} \exp(\frac{-\mathbf{u}_q \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{Q}_i} \exp(-\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} = -\sum_{q \in \mathcal{Q}_i} \mathbf{u}_q \frac{\exp(\frac{-\mathbf{u}_q \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{Q}_i} \exp(-\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} \equiv -\sum_{q \in \mathcal{Q}_i} p_{\text{attract}}(q|i) \mathbf{u}_q \\
\nabla_{\mathbf{u}_q} \Phi_{\text{attract}} &= \frac{-\frac{\mathbf{u}_i}{\tau} \exp(\frac{-\mathbf{u}_q \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{Q}_i} \exp(-\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} = -p_{\text{attract}}(q|i) \mathbf{u}_i \\
\nabla_{\mathbf{u}_a} \Phi_{\text{attract}} &= \mathbf{0}
\end{aligned} \tag{28}$$

where $p_{\text{attract}}(q|i)$ can be interpreted as a probability of attraction between elements i and q. We can see that the effect is mutual: both elements i and q are attracted to each other. Similarly:

$$\begin{aligned}
\nabla_{\mathbf{u}_i} \Psi_{\text{repel}} &= \tau \sum_{a \in \mathcal{B}_a} \frac{\frac{\mathbf{u}_a}{\tau} \exp(\frac{\mathbf{u}_a \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{B}_a} \exp(\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} = \sum_{a \in \mathcal{B}_a} \mathbf{u}_a \frac{\exp(\frac{\mathbf{u}_a \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{B}_a} \exp(\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} \equiv \sum_{a \in \mathcal{B}_a} p_{\text{repel}}(a|i) \mathbf{u}_a \\
\nabla_{\mathbf{u}_q} \Psi_{\text{repel}} &= \mathbf{0} \\
\nabla_{\mathbf{u}_a} \Psi_{\text{repel}} &= \frac{\frac{\mathbf{u}_i}{\tau} \exp(\frac{\mathbf{u}_a \cdot \mathbf{u}_i}{\tau})}{\sum_{r \in \mathcal{B}_a} \exp(\frac{\mathbf{u}_r \cdot \mathbf{u}_i}{\tau})} = p_{\text{repel}}(a|i) \mathbf{u}_i
\end{aligned} \tag{29}$$

where $p_{\text{repel}}(a|i)$ can be interpreted as a probability of repulsion between elements i and a. Thus, since $\nabla_{\mathbf{u}} l_i^{\text{OPP}} = \nabla_{\mathbf{u}} \Phi_{\text{attract}} + \nabla_{\mathbf{u}} \Psi_{\text{repel}}$:

$$\begin{aligned}
\nabla_{\mathbf{u}_i} l_i^{\text{OPP}} &= -\sum_{q \in \mathcal{Q}_i} p_{\text{attract}}(q|i)\mathbf{u}_q + \sum_{a \in \mathcal{B}_a} p_{\text{repel}}(a|i)\mathbf{u}_a \equiv -\boldsymbol{\mu}_{\mathcal{Q}_i}(i) + \boldsymbol{\mu}_{\mathcal{B}_a}(i) \\
\nabla_{\mathbf{u}_q} l_i^{\text{OPP}} &= -p_{\text{attract}}(q|i)\mathbf{u}_i \\
\nabla_{\mathbf{u}_a} l_i^{\text{OPP}} &= p_{\text{repel}}(a|i)\mathbf{u}_i
\end{aligned} \tag{30}$$

The total gradient w.r.t. normals $n \in \mathcal{B}_n$ and anomalies $a \in \mathcal{B}_a$ can therefore be expressed as:

$$\begin{aligned}
\nabla_{\mathbf{u}_n} \mathcal{L}^{\text{OPP}} &= \frac{1}{|\mathcal{B}_n|} \left(-\boldsymbol{\mu}_{\mathcal{Q}_n}(n) - \sum_{i \in \mathcal{Q}_n} p_{\text{attract}}(n|i)\mathbf{u}_i + \boldsymbol{\mu}_{\mathcal{B}_a}(n) \right) \\
\nabla_{\mathbf{u}_a} \mathcal{L}^{\text{OPP}} &= \frac{1}{|\mathcal{B}_n|} \sum_{i \in \mathcal{B}_n} p_{\text{repel}}(a|i)\mathbf{u}_i
\end{aligned} \tag{31}$$

This expression highlights the desired effect of the OPP loss. By minimizing it, the normals are attracted to each other and to their mean as well as being repelled by the anomalies. On the other hand, the anomalies are only affected by a repulsion from the normals. This asymmetry allows the OPP loss to be suited to the OSAD problem.

We can compare the anchor-wise gradients of OPP and of SupCon. The latter can be found in the corresponding paper, and can be expressed as:

$$\nabla_{\mathbf{u}_i} l_i^{\text{SupCon}} = \frac{1}{\tau} \left(-\sum_{p \in \mathcal{P}_i} \mathbf{u}_p X_{ip} + \sum_{a \in \mathcal{N}_i} \mathbf{u}_a P_{ia} \right) = \frac{1}{\tau} \left(\sum_{p \in \mathcal{P}_i} \mathbf{u}_p (P_{ip} - X_{ip}) + \sum_{a \in \mathcal{N}_i \setminus \mathcal{P}_i} \mathbf{u}_a P_{ia} \right) \tag{32}$$

where $P_{ix} = \exp(s_{ix}/\tau) / \sum_{j \in \mathcal{B} \setminus \{i\}} \exp(s_{ij}/\tau)$ and $X_{ip} = 1/|\mathcal{P}_i|$. We find similar elements in both expressions: a grouping term for the positives and a repelling term for the negatives. However, as the set of negatives in the case of SupCon is the entire batch, a repelling term is also applied to the positives creating the NPC effect. Furthermore, the positives are treated uniformly by SupCon ($X_{ip} = 1/|\mathcal{P}_i|$ for every positive p) and the hard mining properties of this loss is only applicable to the negatives through P_{ix}. Finally, the key difference is that every element, normal or abnormal in our application, is taken as an anchor. They are treated in a symmetrical manner, as would be expected in a multi-class classification setting.

C Dataset compositions

We summarize the compositions of the training and testing sets of each dataset in table 10, and show examples of normal and abnormal samples for the datasets used in the hard setting in figure 8.

D Supervised AD baselines

Here, we present the details of the supervised AD baselines used in section 6.4. An element $\mathbf{X}_i$ is assigned the label $y_i = 0$ if it is from the normal class and $y_i = 1$ if it is an anomaly. The features output by the network $g_{\boldsymbol{\theta}}$ with input $\mathbf{X}_i$ are denoted $\mathbf{Z}_i$. In an effort of consistency with section 3, we present these losses in their *per-sample* expression.

DeepSAD. This loss is introduced by Ruff et al. (2020) as a generalization of the unsupervised DeepSVDD method to the semi-supervised AD setting. We have applied it in the supervised AD setting by using a fully labeled dataset.

$$l_i^{\text{DeepSAD}} = (1 - y_i) \left(\|g_{\boldsymbol{\theta}}(\mathbf{X}_i) - \mathbf{c}\|^2 \right) + y_i \frac{1}{\|g_{\boldsymbol{\theta}}(\mathbf{X}_i) - \mathbf{c}\|^2 + \epsilon} \tag{33}$$

Table 10: Composition of the datasets used in this study

Dataset	Training Set	Test Set		N° of classes
	Normal	Normal	Anomaly	
Carpet	280	28	89	5
Grid	264	21	57	5
Leather	245	32	92	5
Tile	230	33	84	5
Wood	247	19	60	5
Bottle	209	20	63	3
Capsule	219	23	109	5
Pill	267	26	141	7
Transistor	213	60	40	4
Zipper	240	32	119	7
Cable	224	58	92	8
Hazelnut	391	40	70	4
Metal Nut	220	22	93	4
Screw	320	41	119	5
Toothbrush	60	12	30	1
MVTec AD	3629	467	1258	73
Mastcam	9302	426	435	8
RealIAD - Zipper	2074	600	2331	4

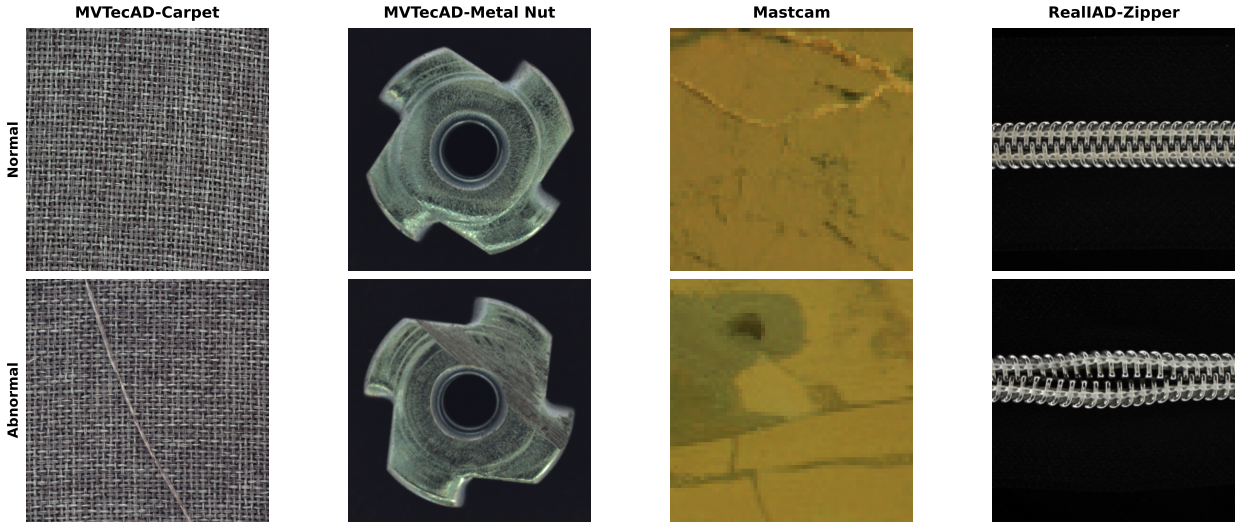

Figure 8: Examples of normal and abnormal samples from each dataset used in the hard setting. The anomaly classes shown are: Carpet/Thread, Metal Nut/Scratch, Mastcam/Drill Hole and RealIAD-Zipper/Deformation.

Hypersphere Classifier (HSC). The most recent supervised AD loss is the Hypersphere Classifier (Liznerski et al., 2022), which reframes DeepSAD as a radial cross-entropy loss with pseudo-Huber loss $l(\mathbf{z}) = \exp\left(-\sqrt{\|\mathbf{z}\|_2^2 + 1} + 1\right)$. HSC minimizes the per-sample loss:

$$l_i^{\text{HSC}} = (1 - y_i) l\left(g_{\boldsymbol{\theta}}(\mathbf{X}_i) - \mathbf{c}\right) - y_i \log\left(1 - \exp\left(-l\left(g_{\boldsymbol{\theta}}(\mathbf{X}_i)\right) - \mathbf{c}\right)\right) \tag{34}$$

HSC therefore concentrates normal embeddings near a fixed center $\mathbf{c} = \mathbf{0}$ and repels the anomalies from $\mathbf{c}$. The implementation of both DeepSAD and HSC was taken from the corresponding github[2] with the center set to the origin $\mathbf{c} = \mathbf{0}$.

DevNet. Pang et al. (2019) proposed an end-to-end score learning framework for supervised AD through a score learner $g_{\boldsymbol{\theta}}$ trained using the deviation loss:

$$l_i^{\text{Dev}} = (1 - y_i)|g_{\boldsymbol{\theta}}(\mathbf{X}_i)| + y_i \max(0, a - g_{\boldsymbol{\theta}}(\mathbf{X}_i)) \tag{35}$$

In this case, the projection head produces a scalar. In practice, the authors of DevNet use a single convolutional layer as their projection head. We reproduced their architecture for the experiments of section 6.4 with the margin a set to 5.

DRA, BGAD. These are standalone methods for which we used the author provided implementations[3,4].

[2] https://github.com/lukasruff/Classification-AD/
[3] https://github.com/choubo/DRA
[4] https://github.com/xcyao00/bgad