Large Class Separation is Not What You Need for Relational Reasoning-Based OOD Detection

Large Class Separation is not what you need for Relational Reasoning-based OOD Detection

Lorenzo Li Lu[1 [0009−0001−9718−9422]], Giulia D'Ascenzi[1 [0009−0003−3238−0300]], Francesco Cappio Borlino[1,2 [0000−0002−8507−0213]], and Tatiana Tommasi[1,2 [0000−0001−8229−7159]]

[1] Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2] Italian Institute of Technology, Italy
{lorenzo.lu, giulia.dascenzi}@studenti.polito.it,
{francesco.cappio, tatiana.tommasi}@polito.it

Abstract. Standard recognition approaches are unable to deal with novel categories at test time. Their overconfidence on the known classes makes the predictions unreliable for safety-critical applications such as healthcare or autonomous driving. Out-Of-Distribution (OOD) detection methods provide a solution by identifying semantic novelty. Most of these methods leverage a learning stage on the known data, which means training (or fine-tuning) a model to capture the concept of *normality*. This process is clearly sensitive to the amount of available samples and might be computationally expensive for on-board systems. A viable alternative is that of evaluating similarities in the embedding space produced by large pre-trained models without any further learning effort. We focus exactly on such a fine-tuning-free OOD detection setting.
This works presents an in-depth analysis of the recently introduced relational reasoning pre-training and investigates the properties of the learned embedding, highlighting the existence of a correlation between the inter-class feature distance and the OOD detection accuracy. As the class separation depends on the chosen pre-training objective, we propose an alternative loss function to control the inter-class margin, and we show its advantage with thorough experiments.

Keywords: Out-Of-Distribution Detection · Cross-Domain Learning · Relational Reasoning

1 Introduction

In recent years, Deep Neural Networks have seen widespread adoption in multiple computer vision tasks. Still, standard recognition algorithms are typically evaluated under the *closed-set* assumption [27], limiting their prediction ability to the same categories experienced at training time. As most real-world scenarios are very different from the well-defined and controlled laboratory environments, an agent operating in the wild will inevitably face data coming from unknown distributions, thus it should be able to handle novelty which is a task of utmost importance for safety-critical applications. In this regard, Out-Of-Distribution

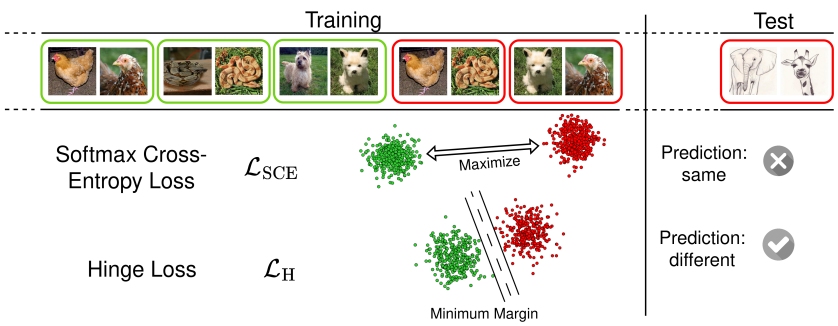


Fig. 1: Schematic overview of a relational reasoning-based OOD method that exploits different training losses. Our work empirically demonstrates that controlling, and in particular reducing, the distance between *same* and *different* classes improves semantic novelty detection.

(OOD) detection techniques have gained considerable attention as they enable models to recognize when test samples are *In-Distribution* (ID) with respect to the training ones, or conversely *Out-Of-Distribution* (OOD). Specifically, *Semantic Novelty Detection* [27] refers to the *open-set* case in which the distribution shift originates from the presence of unknown categories in the test set, together with the known *normal* ones already seen during training. Many techniques have been proposed for this task [8,16,18,23]. However, they typically need a significant number of reference known samples for the model to learn the concept of *normality* through either training from scratch or at least a fine-tuning phase. While such approaches generally lead to good performance, they can also be problematic for low-power edge devices with limited computational resources, and anyway become unfeasible if the amount of known data is scarce or their training access is restricted for privacy reasons.

Recently, two studies proposed techniques that enable OOD detection without fine-tuning [1,19]. Both rely on pre-trained models whose data representation can be easily exploited to perform comparisons and identify unknown categories, avoiding further learning effort. In terms of training objectives, they share the choice of moving away from standard classification and highlight the importance of analogy-based learning to better manage open-set conditions. Specifically, ReSeND [1] proposed a new relational reasoning paradigm to learn whether pairs of images belong or not to the same class. Instead, MCM [19] inherits the CLIP model trained on vision-language data pairs to promote multi-modal feature alignment via contrastive learning.

In this work, we are interested in the single modality case to further evaluate its potential and limits. More precisely, we examine the connection between inter-class distance in the features embedding produced via relational reasoning and the ability to perform semantic novelty detection in that space. As highlighted in [11], training objectives that enforce a stronger inter-class separation may cause the learned representations to be less transferable. Thus, we present an extensive analysis of relational reasoning performed with various loss func-

tions that have different control on class separation. Our findings indicate that avoiding to maximize inter-class separations provides more generalizable features, improving the performance of the pre-trained model on the downstream semantic novelty detection task (see Fig. 1). Building on this conclusion, we design a tailored hinge loss function that provides direct control of class separation and increases the OOD results of the relational reasoning-based model.

Finally, we observe that certain OOD detection methods based on classification pre-training and originally intended to be used via fine-tuning may skip the latter learning phase [14,24]. Hence, these approaches can serve as a fair benchmark reference for relational reasoning-based methods.

To summarize, our key contributions are:

- We discuss and evaluate how the feature distributions originating from the use of different pre-training objectives affect the capability of a relational reasoning model for OOD detection;
- We introduce an alternative loss function that provides better control of class-specific feature distributions;
- We run a thorough experimental analysis that demonstrates the advantages of the proposed loss, considering as a reference also the powerful but costly k-NN-based OOD detector [24], re-casted for the first time to work in the fine-tuning-free setting.

2 Related Work

Out-Of-Distribution detection is the task of determining whether test data belong to the same distribution as the training data or not. A distributional shift may occur due to a change in domain (*covariate shift*) or categories (*semantic shift*). We expect a trustworthy model to detect whether a sample belongs to a new category regardless of the visual domain, thus our main focus is on semantic novelty detection. The first baseline for this task was proposed by Hendrycks et al. [8], who suggested that the Maximum Softmax Probability (MSP) score produced by a classification model trained on ID data should be higher for ID test samples than for OOD ones. Several other approaches followed the same post-hoc strategy enhancing ID-OOD separation, via temperature scaling [16], or by focusing on energy scores [18] and network unit activations [23]. A different family of techniques uses distance metrics to identify OOD samples in the feature space learned on the ID data [14,24].

OOD detection without fine-tuning. All the OOD detection solutions described in the previous paragraph require training (or at least fine-tuning) on nominal samples in order to learn the concept of normality. However, this learning phase requires a sizable amount of ID data and computational resources, making it expensive and impractical for many real-world applications. Additionally, fine-tuning can hurt the generalization of learned representations [13] as it is susceptible to *catastrophic forgetting* [10]: the model may overfit the fine-tuning dataset, losing the knowledge previously learned on a much larger one. Only some recent work has started addressing this problem, proposing solutions that

do not require a fine-tuning stage to perform OOD detection [1,19]. In particular, [1] suggests substituting the standard classification-based pre-training task with relational reasoning, which directs the network's focus on the semantic similarity between two input images to predict a normality score. This pretext objective is less domain- and task-dependent than classification and leads to an embedding space with great transfer capabilities. On the other hand, [19] leverages CLIP [22] to perform zero-shot OOD detection, thus exploiting two modalities (vision and language) rather than one. Finally, we point out how distance-based strategies such as [14,24] could also be used without performing the fine-tuning stage, although this aspect was not addressed in their original works.

Pre-training loss functions and transfer learning. The possibility to easily inherit and reuse pre-trained models for novel tasks is certainly one of the more appreciated characteristics of deep learning. As discussed above, such a procedure may be relevant even for OOD detection applications in which the representation learned on a large-scale dataset is leveraged to evaluate sample similarity. Still, only a few works have analyzed how the exact choice of the pre-training objective influences the transferability of the extracted knowledge. An implicit hypothesis is that models that perform well on the pre-training task also perform well on the downstream one. However, this is not always the case [12]: for instance, some regularization techniques that provide an improvement on the pre-training task produce penultimate layers features that are worse in generalization. This phenomenon has been described as *supervision collapse* [5]. The R^2 metric introduced in [11] to evaluate intra-class compactness and inter-class separation provides a way to shed light on this behavior: the most advanced strategies to increase accuracy on the pre-training task lead to a greater class separation which however is associated with reduced knowledge transferability. As the use of pre-trained models without fine-tuning for OOD detection is still scarcely explored, we find it relevant to perform an analysis of the role of different pre-training objectives for this downstream task. Specifically, we focus on relational reasoning-based OOD detection performed via different loss functions.

3 Relational reasoning for OOD detection

We consider a set of labeled samples $\mathcal{S} = \{\boldsymbol{x}^s, y^s\}$ that we call *support set*, and a set of unlabeled ones $\mathcal{T} = \{\boldsymbol{x}^t\}$ called *test set*. They are drawn from two different distributions and present a category shift, besides also a potential domain shift. The support label set $\mathcal{Y}_\mathcal{S}$ identifies *known* categories. The target label set $\mathcal{Y}_\mathcal{T}$ includes both known and unknown semantic categories: $\mathcal{Y}_\mathcal{S} \subset \mathcal{Y}_\mathcal{T}$. The goal of an OOD detector is to identify all the test samples whose categories do not appear in the support set (i.e., which are *unknown*). Traditional methods require a training or fine-tuning stage on the support set $\mathcal{S}$, while in the fine-tuning-free scenario the support set is only accessed at evaluation time.

In ReSeND [1], the authors presented a relational reasoning-based learning approach specifically designed for OOD detection. The model is trained on sample pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ drawn from a large-scale object recognition dataset and learns

to distinguish whether the two images belong to the same category ($l_{ij} = 1$, if $y_i = y_j$) or not ($l_{ij} = 0$, if $y_i \neq y_j$). This task can be cast as binary classification or regression. In both cases, the model learns how to encode in an embedding space the samples' semantic relationship $\boldsymbol{p}_m = r(\boldsymbol{z}_i, \boldsymbol{z}_j)$, where the index m ranges over all the possible sample pairs, and $\boldsymbol{z} = \phi(\boldsymbol{x})$ represents the features extracted via an encoder ϕ from the image $\boldsymbol{x}$. Then, the last network layer converts this information into a scalar similarity value that is compared to the ground truth l_m with a chosen loss function. At inference time, the support set samples are grouped according to their category and their representation is averaged to get per-class prototypes $\bar{\boldsymbol{z}}_y^s$ for $y = \{1, \ldots, |\mathcal{Y}_\mathcal{S}|\}$. Each test sample $\boldsymbol{z}^t = \phi(\boldsymbol{x}^t)$ is then compared with every prototype to get the corresponding similarity score. Finally, the vector collecting all the $|\mathcal{Y}_\mathcal{S}|$ elements is filtered by a softmax function on which MSP is applied to get the final normality score.

In this framework, by observing the embedding space produced by the penultimate layer of the network, we expect to see pairs of samples of the training dataset organized into two clusters representing the broad *same* and *different* concept classes. Once trained on the large-scale ImageNet-1K dataset, this embedding can be used for OOD detection on a variety of domains without fine-tuning, so its generalization ability is crucial.

4 Relational reasoning and class separation

4.1 Class compactness and separation

In order to analyze the learned feature space we focus on the separation between the *same* and *different* classes described above. In particular, we leverage the R^2 index introduced in [11]. This metric is based on the ratio between the average within-class and average global cosine distance for the considered feature vectors, providing a relative measure of the sparsity of the representation of each class in the embedding space. Specifically, the index value is given by:

$$R^2 = 1 - \bar{d}_{within}/\bar{d}_{total} \tag{1}$$

$\bar{d}_{within} = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \sum_{j=1}^{M_k} \frac{1-\text{sim}(\boldsymbol{p}_i^k, \boldsymbol{p}_j^k)}{K M_k^2}$, $\bar{d}_{total} = \sum_{h=1}^{K} \sum_{k=1}^{K} \sum_{i=1}^{M_h} \sum_{j=1}^{M_k} \frac{1-\text{sim}(\boldsymbol{p}_i^h, \boldsymbol{p}_j^k)}{K^2 M_h M_k}$

where the indices $i, j \in \{1, \ldots, M_k\}$ now range on the pairs of samples $\boldsymbol{p}^k$ which belong respectively to the $K = 2$ classes. The relative distance is measured via the cosine similarity: $\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^T \boldsymbol{b}/(\|\boldsymbol{a}\|\|\boldsymbol{b}\|)$. The right part of Fig. 3 gives an idea of what high and low R^2 values mean in terms of class separation.

4.2 Relational Reasoning Loss Functions

In the following we review some of the most common loss functions used for binary problems. In all the loss equations we use σ to refer to the score produced as output by the network for a sample pair $\boldsymbol{p}$, while the ground truth label is l.

Binary Cross-Entropy. The Cross-Entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{m=1}^{M}\sum_{k=1}^{K} t_{m,k}\log(\hat{t}_{m,k}) \tag{2}$$

where K is the number of categories, while $t_{m,k}$ and $\hat{t}_{m,k}$ are respectively the target value and the predicted probability of the class k for the sample m. In particular $t_{m,k}$ will assume the value 1 for the ground truth class of the sample ($k = l_m$) and 0 for all the other categories (one-hot encoding). In the binary case (i.e., when $K = 2$), such loss function can be expressed as:

$$\mathcal{L}_{\text{BCE}} = -\sum_{m=1}^{M}\left(t_{m,1}\log(1-\hat{t}_{m,2}) + t_{m,2}\log(\hat{t}_{m,2})\right) \tag{3}$$

where $\hat{t}_{m,2}$ is generally obtained by applying the logistic sigmoid function to the model output score ($f(\sigma) = 1/(1+e^{-\sigma})$).

Impact on the class separation: this loss is non-zero even for correctly classified samples. As a result, the intra-class compactness and inter-class separation keep increasing for the whole training procedure.

Softmax Cross-Entropy. The categorical Cross-Entropy loss that is generally adopted for multi-class problems, is obtained by using the Cross-Entropy in Eq. (2) after having applied the softmax function to the model output scores ($f(\sigma)_k = e^{\sigma_k}/\sum_{c=1}^{C} e^{\sigma_c}$). Considering that the labels are one-hot, the overall summation will contain for each sample only the term corresponding to its ground truth label, so we can write the Softmax Cross-Entropy as:

$$\mathcal{L}_{\text{SCE}} = -\sum_{m=1}^{M}\log\frac{e^{\sigma_{m,l_m}}}{\sum_{k=1}^{K} e^{\sigma_{m,k}}} \tag{4}$$

where $\sigma_{m,k}$ is the score corresponding to the class k for the sample m and l_m represents its ground truth label. In the binary case we suppose $k, l \in \{1, 2\}$.

Impact on the class separation: As in the previous BCE case, this loss is non-zero even for correctly classified samples. It has been shown that the consequent trend of growing intra-class compactness and inter-class separation leads to miscalibrated classifiers providing overconfident predictions [26,20].

Focal Loss. A possible solution for the miscalibration issue mentioned above is to adjust the penalty assigned to a sample based on the network's confidence in predicting its true class [20]. This can be accomplished with the Focal Loss [17]. Starting from the Cross-Entropy formulation (see Eq. 2), such loss function can be expressed as:

$$\mathcal{L}_{\text{focal}} = -\sum_{m=1}^{M}\sum_{k=1}^{K} t_{m,k}(1-\hat{t}_{m,k})^{\gamma}\log(\hat{t}_{m,k}) \tag{5}$$

where γ is a hyperparameter controlling the rescaling strength.

(a) Compressed sigmoid (b) Hinge loss trend

Fig. 2: (a) Increasing c in the MSE compressed sigmoid transforms it into a Heaviside step function: the loss is zero when the output score has the correct sign. (b) H loss trend for positive ($l_m = 1$) and negative ($l_m = -1$) pairs ($\delta = 1$).

Impact on the class separation: by varying γ, it's possible to tune the magnitude of the rescaling, effectively bringing the loss value for correctly classified samples near zero and therefore mitigating the class separation tendency.

MSE with a compressed sigmoid. In ReSeND [1] the problem of separating the same and different classes was formalized as a regression task by using the MSE loss computed between the ground truth $l_m \in \{-1, 1\}$ and the output provided by a sigmoid rescaled on the $[-1, 1]$ range and with a modified slope, controlled by a factor c (see Fig. 2 (a)):

$$\mathcal{L}_{MSE} = \sum_{m=1}^{M} (\hat{s}_c(\sigma_m) - l_m)^2 \qquad \text{with} \quad \hat{s}_c(\sigma_m) = \frac{2}{1 + e^{-c\sigma_m}} - 1 \tag{6}$$

Impact on the class separation: by varying the value of c, it is possible to tune the penalty associated with different scores σ. Specifically, for higher c values, the sigmoid function will be more horizontally compressed: as a consequence samples already correctly classified receive a loss value that is almost zero decreasing the need for further class separation.

4.3 Controlling class separation: Hinge Loss for relational reasoning

As it is clear that class separation is crucial for the problem, we introduce a loss function that allows us to precisely tune it in a simple and straightforward way.

Let's start from the output of the last layer which is a scalar score σ_m and can be positive or negative, indicating the corresponding two classes. We can simply set a threshold at zero and fix a margin δ around it, within which even correct predictions pay a penalty. The loss will cancel out for $\sigma_m > \delta$ on positive samples and $\sigma_m < -\delta$ for negative ones, but would grow linearly if a negative score is assigned to a positive sample and vice-versa (see Fig. 2 (b)).

In this way we keep the two classes separated (which is crucial to retain the model's discriminative power), but the margin is limited and fixed to δ. This formulation corresponds to a hinge loss applied on the scalar score σ_m:

$$\mathcal{L}_H = \sum_{m=1}^{M} \max(0, \delta - l_m \sigma_m) \qquad \text{with} \quad l_m \in \{-1, 1\} \tag{7}$$

5 Experiments

5.1 Experimental protocol

Our experimental analysis presents an extensive benchmark of fine-tuning-free OOD methods. All of them consist of a pre-training phase on ImageNet-1K [4] with a different objective, followed by a distance-based OOD prediction protocol. For ReSeND [1] the pre-training task is relational reasoning (same vs different) executed with all the loss functions described in the previous Section. The other competitors exploit either supervised classification or self-supervised objectives, with both cross-entropy-based approaches (ResNet[7], ViT[6], CutMix[28]), and contrastive strategies (SimCLR [2], SupCLR [9], CSI [25], SupCSI [25]). We also evaluate Mahalanobis [14] and k-NN [24]. We emphasize that the k-NN approach has never been previously evaluated in a fine-tuning-free setting. We include it in our comparison despite its potentially higher computational cost, as it involves comparing the test sample with each support set instance (which must be stored in memory), rather than with a single prototype per class.

Unless otherwise specified, we always adopt a ResNet-101 backbone, as it includes a comparable number of parameters to ReSeND (44M and 40M, respectively). We publish the code, together with implementation details and additional results in our project page [3].

We adopt two different experimental set-ups, by following [1]. The *intra-domain* setting is designed to evaluate the OOD detection ability of a model when there is a purely semantic distribution shift between the support and the test sets. It is built upon the DomainNet [21] and DTD [3] datasets. In the *cross-domain* setting the support and test set are sampled from different domains so we can evaluate the ability of the OOD methods to focus on semantics and disregard other visual appearance discrepancies. Rather than using the limited PACS dataset [15] as done in [1], we propose a novel benchmark built on top of DomainNet [21]. This choice allows for more statistically significant results.

Following common practice, we report results in terms of Area Under the Receiver Operating Characteristic curve (AUC) and FPR@TPR95 (FPR), which indicates the false positive rate value when the ID true positive rate is 95%.

5.2 Impact of the training objective

We evaluate the impact on ReSeND of various loss functions. As the learning objective shapes the structure of the feature space, we can investigate how the distribution of the data in the learned embedding relates to the final OOD performance. For this analysis we focus on the intra-domain setting. The average AUC on the four datasets, along with the corresponding R^2 value, are reported in the scatter plot in the left part of Fig. 3. Detailed per-dataset results can be found on our project page. The results clearly highlight a general trend in which a higher inter-class separation is associated with a lower OOD detection

[3] https://github.com/lulor/ood-class-separation

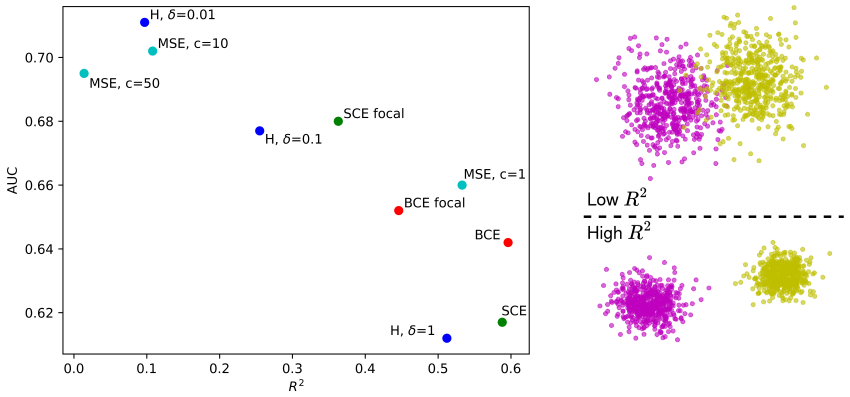
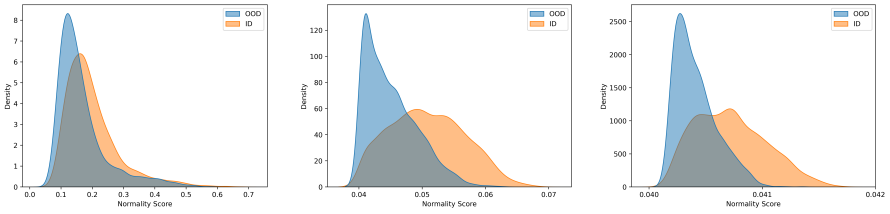


Fig. 3: Analysis of the OOD performance of a relational reasoning-based model trained with different loss functions (average *intra-domain* benchmarks results). The scatter-plot on the left shows that lower OOD results are generally associated with higher R^2 values, which means a stronger class separation as in the point distribution shown in the right-bottom part. On the other hand, more generalizable features and higher OOD performance are obtained with lower R^2 values corresponding to minimal class separation as in the right-top part.

performance. This behavior is even more evident when focusing on a specific loss and looking at how the results change by varying its hyperparameter value (e.g. when changing δ for our H loss or c for the MSE). Of course, there is a limit in the performance gain that can be reached by reducing the inter-class separation: after a certain point, the features start losing their discriminative power. For example with $c \geq 50$ for the MSE case, the training becomes less effective and the OOD detection performance starts slowly decreasing. We can conclude that for relational reasoning it is important to choose a learning objective that allows for a precise margin control. Only the proposed $\mathcal{L}_H$ loss satisfies this condition. Its hyperparameter δ represents a geometrical margin, and when a training sample meets the margin condition, the sample loss no longer affects the learning process. This behavior avoids the overconfidence typical of the standard softmax cross-entropy as highlighted by the normality score distributions represented in Fig. 4: the normality score values provided by the SCE loss are generally higher and have a larger range than those provided by the $\mathcal{L}_H$ loss (see the horizontal axis), but at the same time they provide a weaker ID-OOD separation.

5.3 Intra-Domain and Cross-Domain OOD Results

Intra-Domain analysis. In this setting support and test sets only differ in terms of semantics. Still, with respect to the pre-training dataset (ImageNet-1K), there may be a domain shift of varying magnitude (smaller for the Real case, larger for the others). In Table 1 we collect the results of the original ReSeND formu-



(a) SCE (b) H, δ=0.1 (c) H, δ=0.01

Fig. 4: Normality Score distributions on the intra-domain Real setting for ReSeND pre-trained with different loss functions. We can see how the hinge loss with low margin pushes the model to provide more conservative scores, which are very close to each other (check the horizontal axis' scale) but more discernible.

Table 1: Intra-domain setting. Best result in bold and second best underlined

Model	**Texture**		**Real**		**Sketch**		**Painting**		**Avg**		**Avg**
	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	n.comp ↓
ResNet	0.672	0.897	0.710	0.863	0.554	0.939	0.649	0.919	0.646	0.904	25
ViT	0.537	0.937	0.701	0.829	0.553	0.955	0.673	0.853	0.616	0.894	25
CutMix	0.605	0.925	0.722	0.876	0.544	0.944	0.627	0.929	0.625	0.919	25
SimCLR	0.526	0.942	0.475	0.943	0.489	0.955	0.508	0.959	0.500	0.950	25
SupCLR	0.588	0.921	0.496	0.956	0.481	0.953	0.514	0.959	0.520	0.947	25
CSI	0.627	0.898	0.695	0.850	0.513	0.960	0.613	0.912	0.612	0.905	25
SupCSI	0.662	0.896	0.716	0.864	0.521	0.957	0.640	0.902	0.635	0.904	25
Mahalanobis	0.656	0.911	0.744	0.850	0.590	0.928	0.710	0.857	0.675	0.886	25
ReSeND	0.684	0.847	0.782	0.777	0.610	0.934	0.721	0.826	0.699	0.846	25
ReSeND-H	0.688	0.885	0.798	0.755	0.638	**0.898**	0.719	0.826	0.711	0.841	25
k-NN (k=1)	**0.774**	**0.840**	**0.843**	**0.596**	**0.640**	0.914	**0.800**	**0.757**	**0.764**	**0.777**	4100

lation ($\mathcal{L}_{MSE}$, with $c = 10$), its version based on the hinge loss that we name ReSeND-H ($\mathcal{L}_H$, with $\delta = 0.01$) as well as the ResNet baseline and several reference approaches. We observe that ReSeND-H obtains a small but meaningful improvement across most of the considered settings, particularly in the Real and Sketch ones. The k-NN method from [24], applied without fine-tuning, achieves the best performance among all the considered techniques. We highlight how this result comes with a significant cost in terms of memory usage. Specifically, we calculated the average number of comparisons (n.comp) per test sample needed at evaluation time and reported the corresponding value in the last table column. Indeed, this introduces an important scalability limitation.

Cross-Domain analysis. In this setting train and test data differ in semantic content and in visual style. From the results in Table 2 we can see how, despite using the whole support set rather than just the class prototypes, the k-NN method does not achieve the same performance advantage exhibited in the intra-domain case. Indeed, relying on all the support samples appears misleading. As a consequence, this method is less robust to domain shifts compared to ReSeND and ReSeND-H, which instead shows similar performance to the corresponding one in the intra-domain setting.

Table 2: Cross-domain setting. Best result in bold and second best underlined.

Model	Real-Paint.		Real-Sketch		Paint.-Real		Paint.-Sketch		Sketch-Real		Sketch-Paint.		Avg		Avg
	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	n.comp ↓
ResNet	0.596	0.949	0.539	0.938	0.627	0.922	0.546	0.941	0.533	0.929	0.524	0.940	0.561	0.937	25
ViT	0.627	0.921	0.526	0.931	0.618	0.901	0.524	0.946	0.568	0.945	0.591	0.924	0.576	0.928	25
CutMix	0.585	0.940	0.533	0.944	0.630	0.915	0.534	0.949	0.550	0.939	0.530	0.950	0.560	0.940	25
SimCLR	0.499	0.965	0.486	0.949	0.465	0.961	0.489	0.956	0.496	0.961	0.419	0.966	0.476	0.960	25
SupCLR	0.507	0.966	0.471	0.959	0.468	0.962	0.469	0.957	0.524	0.968	0.463	0.965	0.484	0.963	25
CSI	0.585	0.942	0.531	0.943	0.689	0.863	0.503	0.953	0.552	0.867	0.448	0.942	0.551	0.918	25
SupCSI	0.586	0.943	0.492	0.963	0.658	0.898	0.473	0.957	0.490	0.963	0.434	0.973	0.522	0.949	25
Mahalanobis	0.612	0.945	0.564	0.938	0.646	0.943	0.577	0.928	0.577	0.912	0.564	0.919	0.590	0.931	25
ReSeND	**0.666**	0.912	0.572	0.934	0.727	0.878	0.566	0.942	**0.705**	0.860	**0.659**	0.911	**0.649**	0.906	25
ReSeND-H	0.639	0.938	**0.583**	**0.919**	0.720	0.864	**0.590**	**0.899**	0.679	0.895	0.637	0.914	0.641	0.905	25
k-NN (k=1)	0.662	**0.902**	0.560	0.934	**0.754**	**0.781**	0.584	0.908	0.666	**0.836**	0.627	**0.900**	0.642	**0.877**	4800

6 Conclusions

In this work, we focused on the OOD detection task considering methods that do not need a fine-tuning stage on the ID data in order to detect semantic novelties. We analyzed how different learning objectives influence the performance of a relational reasoning-based solution to this problem, showing that a lower inter-class separation leads to better generalization. Exploiting this finding we proposed to use a tailored hinge loss function that provides better results than the original method implementation. At the same time, we pointed out how a previously unexplored fine-tuning-free k-NN strategy for OOD detection provides unexpectedly good accuracy at the cost of a higher computational effort. Still, it may fail when the support and the test set are drawn from different visual domains.

Acknowledgments This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.
Computational resources were provided by IIT HPC infrastructure.

References

1. Cappio Borlino, F., Bucci, S., Tommasi, T.: Semantic novelty detection via relational reasoning. In: ECCV (2022)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
5. Doersch, C., Gupta, A., Zisserman, A.: Crosstransformers: spatially-aware few-shot transfer. In: NeurIPS (2020)

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. ICLR (2017)
9. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS (2020)
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
11. Kornblith, S., Chen, T., Lee, H., Norouzi, M.: Why do better loss functions lead to less transferable features? In: NeurIPS (2021)
12. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: CVPR (2019)
13. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: ICLR (2022)
14. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
15. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV (2017)
16. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: ICCV (2017)
18. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. NeurIPS (2020)
19. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: NeurIPS (2022)
20. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H., Dokania, P.K.: Calibrating deep neural networks using focal loss. In: NeurIPS (2020)
21. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
23. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021)
24. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: ICML (2022)
25. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: NeurIPS (2020)
26. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: ICML (2022)
27. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334 (2021)
28. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)