POLITECNICO DI TORINO
Repository ISTITUZIONALE

Semantic Novelty Detection via Relational Reasoning

(Article begins on next page)

28 April 2024

Semantic Novelty Detection
via Relational Reasoning

Francesco Cappio Borlino[*,1,2], Silvia Bucci[*,1], and Tatiana Tommasi[1,2]

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

Abstract. Semantic novelty detection aims at discovering unknown categories in the test data. This task is particularly relevant in safety-critical applications, such as autonomous driving or healthcare, where it is crucial to recognize unknown objects at deployment time and issue a warning to the user accordingly. Despite the impressive advancements of deep learning research, existing models still need a finetuning stage on the known categories in order to recognize the unknown ones. This could be prohibitive when privacy rules limit data access, or in case of strict memory and computational constraints (e.g. edge computing). We claim that a tailored representation learning strategy may be the right solution for effective and efficient semantic novelty detection. Besides extensively testing state-of-the-art approaches for this task, we propose a novel representation learning paradigm based on relational reasoning. It focuses on learning how to measure semantic similarity rather than recognizing known categories. Our experiments show that this knowledge is directly transferable to a wide range of scenarios, and it can be exploited as a plug-and-play module to convert closed-set recognition models into reliable open-set ones.

Keywords: Representation Learning, Novelty Detection, Open Set Learning, Domain Generalization, Relational Reasoning

1 Introduction

In the last years, deep learning models have brought significant advances in several computer vision tasks. We can identify two main ingredients as the basis of this widespread success. The first one is the pre-training stage: the possibility to rely on a large set of freely available images allows to learn a representation that is generally helpful to initialize the models. The second component is the optimistic assumption that training and test distributions will perfectly match. Indeed, in real-world conditions, it's much more common to encounter differences between the two, for instance, due to a mismatch among their semantic category sets. This condition is particularly dangerous in safety-critical applications like autonomous driving and healthcare, where previously unseen categories should
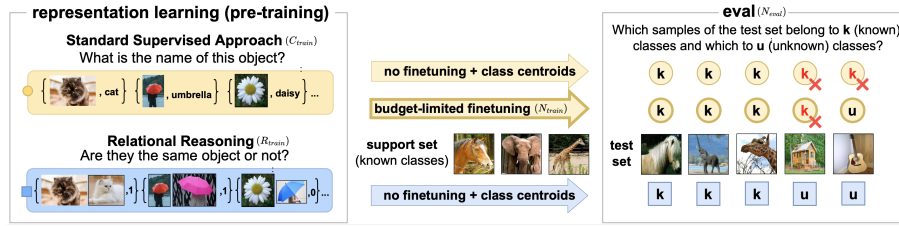
---

[*] equal contributions

Fig. 1. Comparison between standard supervised learning and relational reasoning representation learning. The first aims at recognizing the known object classes, while the second learns a measure of semantic similarity among image pairs. We claim and verify experimentally that relational reasoning is particularly suitable when the final goal is semantic novelty detection. Our pre-trained large-scale relational model can be transferred on semantic novelty detection tasks without the need for a finetuning phase on the known classes of the task at hand.

be reliably detected as *unknown*. Several studies have proposed to improve the learning procedure and make it aware of semantic novelties outside of the training distribution. Existing solutions consist in calibrating the softmax output of deep classifiers [29,48,49], or using generative approaches to synthesize outliers [53,22,71,80,51]. However, a relevant limitation of these techniques is that all of them require to be trained, or at least finetuned, on a reasonably large set of reference data in order to learn what is *known*. In case of limited data access due to privacy concerns, or when dealing with memory and computational constraints (e.g. edge computing), these strategies could be inapplicable.

In this work, we put the spotlight on the pre-training stage. We claim that, rather than considering the usual cross-entropy based classification [28], or self-supervised contrastive learning [10,27], we can exploit ImageNet1k to optimize a relational reasoning objective and obtain a more reliable embedding for novelty detection (see Fig. 1). Specifically, our target is a semantic similarity measure that indicates whether two samples belong to the same class or to different ones. Thus, we focus on learning a representation designed for semantic comparison which does not need further finetuning on the annotated data of the task at hand. It will be enough to compare each test sample with the reference class-prototypes to separate known and unknown categories. Besides being an efficient strategy, our method provides a plug-and-play solution to convert existing closed-set models to open-set ones by including a rejection option for unknown classes.

To summarize, **we focus on Semantic Novelty Detection (SeND) and propose ReSeND, a representation learning approach based on Relational Reasoning that is ready to be used in real-world applications without the need for finetuning**. In particular, our contributions are:

- we conduct a thorough experimental analysis on the ability of several representation learning paradigms to deal with the SeND task, exploring their potentialities and limits;
- we introduce ReSeND and evaluate it on several *intra-* and *cross*-domain scenarios, exploring settings with different ratios of unknown classes in the test

data. An extensive benchmark with several competitors confirms the effectiveness and efficiency of our approach;
- we show how ReSeND can be used as a plug-and-play module on closed-set domain generalization approaches converting them into open-set domain generalization strategies that set the new state-of-the-art.

2   Related Works

Our work relates to three main research areas: representation learning, relational reasoning, and out-of-distribution detection.

Representation Learning makes the difference between classic shallow and modern deep machine learning approaches. The former relies on handcrafted feature representation, while the latter automatically learns to represent the input data through a hierarchy of features during the training process. The literature on this topic is quite extensive [2,26], ranging from the design of neural architectures [31,41,25] to the development of learning paradigms [10,6,19]. The most common approach used to get effective representations from visual data is supervised learning, but recent works have been mainly dedicated to learning representations from unlabeled samples [23,55,37,83,46,10,27,85,11]. They showed how the obtained self-supervised embeddings are able to capture general knowledge of data structure and can be leveraged by a large variety of downstream tasks [42,54,20]. Usually, this happens via a transfer learning procedure that requires finetuning on annotated training data of the final task.

Relational Reasoning is a hallmark of human intelligence and it has been formalized by the machine learning community as learning a function to quantify the relationships between a set of objects. This paradigm has attracted particular attention for the combination of language and vision for scene description [38,67,63]. Other applications are on reinforcement learning [66,57,84], object detection [34], graph networks [1], and few-shot learning [74,86].
Relational reasoning and contrastive learning. Recently, it has been shown that relational reasoning can effectively guide self-supervised representation learning [60], with better results than those of popular contrastive learning strategies [10,32]. On the basis of these results, we can identify one important aspect that makes relational reasoning different from contrastive learning. The latter aims at learning a feature space for individual samples, with the similarity between two samples computed a posteriori using a distance metric; the goal of the former is to construct a representation for sample pairs: the position of a point in the final embedding directly represents the similarity between two samples.

Out-Of-Distribution detection (OOD) studies how to identify whether a given test sample is drawn from the training distribution or not. Both a variation in semantic content and in the visual domain may cause a deviation from the reference distribution. OOD is a wide framework that covers several sub-settings.
OOD subsettings. In *anomaly detection* the training samples belong to a single semantic category and a test sample is considered anomalous both if it contains
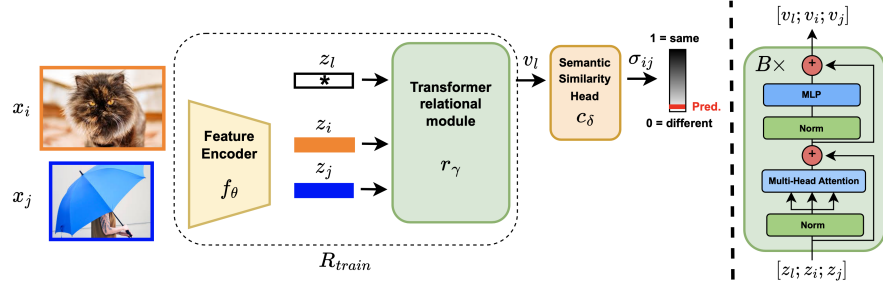
Fig. 2. Schematic illustration of the training phase of ReSeND. The features extracted from a pair of images are provided as input to our relational module. It consists of a transformer encoder that elaborates over a tuple composed of the sample pair and of a learnable label token. The output corresponding to this last token is finally provided as input to a semantic similarity head that predicts the sample resemblance.

a novel class and in case it presents the same known class but with perceptual differences from the training (e.g. local defects, global style). When the training data cover more than one class, the setting is usually indicated as *novelty detection*. As in anomaly detection, the cause of novelty can be either a semantic shift, or domain shift or both [30,81]. We use the name *semantic novelty detection* (SeND) to focus on the first case: models that spot unknown categories in the test while being agnostic to domain variations [56]. *Open-set recognition* extends novelty detection by considering not only a binary identification of known and unknown classes in the test, but also a reliable recognition of the known classes. Usually, this setting is well controlled with training and test data sharing the same visual domain. In *open-set domain generalization* the model should be also robust to the domain shift between train and test data [72].

OOD strategies in literature. From standard classification, we can evaluate whether a test sample is anomalous by applying a threshold on the output score of the top predicted class (maximum-softmax probability, MSP [29]). Improvements over this basic approach have been proposed in [48,33,49]. Instead of the model output, a recent work has shown how the gradients space of neural networks can be used to estimate prediction uncertainty and obtain an OOD scoring function [35]. Generative-based approaches consider the performance of a model trained on reference known classes when reconstructing an input sample. The reconstruction error defines the novelty score [40,15,59]: GAN and flow-based invertible models have been exploited for this purpose [71,80,51]. Some methods synthesize out-of-distribution data [53,22,44] or use external dataset as a source of outlier exposure during training [30,58,9]. A different solution consists in estimating test samples normality by computing their distance from training data using specific embeddings or metrics. [45,68]. A stream of works has also shown the effectiveness of self-supervised representation learning [23,55,37,83,46], and in particular of the contrastive-based strategies for OOD [24,3,75,70,79]. Indeed by removing the focus from the labels, self-supervised models capture analogies and differences among the samples and provide a better way to score similar-

ities. However, training these models needs a non-trivial optimization process with large training batches. Embeddings based on self-attention have been considered as starting point for OOD in [43,73]. Here the powerful transformer architecture ViT [18] pre-trained on ImageNet [17] for classification is finetuned on the training data to then score the test samples via MSP. Still, the risks of finetuning a large model on the training data for OOD were discussed in [16], which highlighted how part of the original knowledge gets lost in this process.

Finally, as also noticed by Huang et al. [36], we underline how most of the existing works on OOD consider experimental analysis on datasets containing only digits or low-resolution images. Combined with the limitation of the existing models described above, it becomes clear the need for novel efficient solutions that can be easily deployed in real-world conditions.

3 Method

3.1 Notation and background

In the semantic novelty detection task, we have two datasets: a *support set* containing labeled samples $\mathcal{S} = \{\boldsymbol{x}^s, y^s\}_{k=1}^K$ drawn from the distribution $p_{\mathcal{S}}$ and a *test set* containing unlabeled samples $\mathcal{T} = \{\boldsymbol{x}^t\}_{h=1}^H$ drawn from the distribution $p_{\mathcal{T}}$. The main difference between $p_{\mathcal{S}}$ and $p_{\mathcal{T}}$ is a semantic shift: it holds $y^s \in \mathcal{Y}_s$ and $y^t \in \mathcal{Y}_t$, with $\mathcal{Y}_s \neq \mathcal{Y}_t$. The two sets of classes can be either completely disjoint $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$, or partially overlapping $\mathcal{Y}_s \subset \mathcal{Y}_t$. In the following we will indicate $\mathcal{Y}_s$ as the *known* classes, while we use the term *unknown* to refer to the test classes $\mathcal{Y}_{t \setminus s}$ not appearing in the support set. Domain shift may contribute to the distribution difference among support and test, causing a variation in the appearance of the samples. Still, the class content remains unchanged. A reliable semantic novelty detector should discriminate between known and unknown samples in the test set while being robust to the domain shift.

Given a test sample $\boldsymbol{x}^t$, the detector D should be able to predict a *score* $\in [0,1]$ that signals whether it is known or unknown with respectively high and low values. Following the traditional strategy, the detector can be formalized as $D : \{C_{train}(\mathcal{I}), N_{train}(\mathcal{S}), N_{eval}(\boldsymbol{x}^t)\}$. At first a good representation is learned by training a classification model C on the samples $(\boldsymbol{x}_i, y_i)_{i=1}^I$ of a large-scale dataset $\mathcal{I}$ as ImageNet1k [17]. The representation is then inherited by the model N which is finetuned on the support set to gather the definition of *normality* from the data. When this training is guided by a simple classification objective, the final evaluation of N on the test is usually performed by MSP: $score = \max_{c \in \mathcal{Y}_s} p(y = c | \boldsymbol{x}^t)$. We highlight that the finetuning process has a computational cost that might not be affordable on edge devices. Moreover, in the long term, its catastrophic forgetting effect reduces the original large-scale knowledge, as well as the ability to anticipate potential semantic anomalies [16]. Thus, carefully designing the representation learning approach and choosing how the pre-trained model should be applied for the downstream task is crucial.

We propose to change the learning paradigm for the semantic novelty detector so that it can be written as $D : \{R_{train}(\mathcal{I}), N_{eval}(\mathcal{S}, \boldsymbol{x}^t)\}$. The first component

R is a representation learning model based on relational reasoning and trained on ImageNet1k. The learned embedding is directly used by an evaluation system to compare each test sample with the support set to obtain its normality score.

3.2   Representation Learning via Relational Reasoning

We consider R composed of a feature extractor f_θ and a relational module r_γ. A pair of samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the reference dataset $\mathcal{I}$ passes first through the feature extractor $(\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i), \boldsymbol{z}_j = f_\theta(\boldsymbol{x}_j))$, and is then fed to the relational module r_γ. The output of this module is the input of the semantic similarity head c_δ which is simply a fully connected (FC) layer. It returns $\sigma_{ij} = c_\delta(r_\gamma(\boldsymbol{z}_i, \boldsymbol{z}_j)) \in [0, 1]$ which represents a semantic similarity measure and can be interpreted as the probability that the two input samples belong to the same category.

The whole representation learning model is trained with a regression objective. Specifically, we assign to each data pair the label $l_{ij} = 0$ if $y_i \neq y_j$ and $l_{ij} = 1$ otherwise, and we minimize the MSE loss:

$$\arg\min_{\theta, \gamma, \delta} \sum_{m=1}^{M} (\sigma_m - l_m)^2 \ , \tag{1}$$

here the index m specifies the pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $i \neq j$ and $x_i, x_j \in \mathcal{I}$.
Despite the ground truth supervision being only at the extremes of the prediction interval, we aim at learning a semantic similarity measure in the continuous range $[0, 1]$. For this reason, the regression loss is particularly suited for the task, but the problem could be also casted as binary classification. In the experimental section we compare the two approaches providing empirical evidences about the beneficial effect of our regression choice.

3.3   Evaluation Process

Starting from the learned embedding, the component N_{eval} of our approach has the simple role of comparing each test sample with the reference support set, without any further training phase. N_{eval} exploits the relational module and provides to it data pairs composed of the feature of each test sample $\boldsymbol{z}^t = f_\theta(\boldsymbol{x}^t)$, and the set of per-class prototypes $\overline{\boldsymbol{z}}_{y^s}^s \ \forall y^s \in \mathcal{Y}_s$ obtained as the average over the samples of each class in the support set. We obtain a vector $\boldsymbol{u}$ of $|\mathcal{Y}_s|$ elements, each corresponding to $c_\delta(r_\gamma(\boldsymbol{z}^t, \overline{\boldsymbol{z}}_{y^s}^s))$ and expressing the similarity of the test sample $\boldsymbol{z}^t$ to one of the known classes. This output is filtered by a softmax function and we apply MSP to get the final normality score: $score = \max(softmax(\boldsymbol{u}))$.

3.4   Relational module

With respect to other standard components of deep neural networks that elaborate on single samples, the peculiarity of the relational module is that it processes

pairs of inputs to provide information on their similarity. Of course, the order of appearance of the two samples should not influence the network output as any good similarity measure needs to be symmetric. Considering its natural permutation invariance and its well known capability of comparing multiple inputs, we implement our relational module through a simple transformer encoder. It consists of B identical blocks, each one composed of a Multi-Head Self-Attention (MSA) and a Multi-Layer Perceptron (MLP), both preceded by Layer-Norm (LN) modules and bypassed by residual skip connections as shown in the right part of Fig. 2. The input feature vectors pair, together with a learnable label token, forms the tuple $[\boldsymbol{z}_l, \boldsymbol{z}_i, \boldsymbol{z}_j]$ which is fed as input to the transformer and passes through all its layers, producing the output sequence $[\boldsymbol{v}_l, \boldsymbol{v}_i, \boldsymbol{v}_j]$. Note that, in this architecture each image represents a single input token to the transformer, as done in [12]. We do not include in our encoder the commonly used positional embeddings as we aim at keeping the permutation invariance. In our implementations we use a ResNet18-based backbone as feature extractor f_θ and select $\boldsymbol{v}_l$ as the output of the relational module r_γ, which is then passed through the head c_δ to produce a semantic similarity score σ_{ij}. In the experimental section we evaluate alternative architectures for our relational module.

4   Experimental Setup

With ReSeND we are proposing a novel strategy fully based on representation learning for OOD. We claim that the embedding space learned via relational reasoning is well suited to detect novel classes simply comparing the test samples with the support set which represents the *normal* reference condition. Since this logic substantially differs from that of previous works in OOD, there are several questions that we need to answer with experimental validations.

Are existing representation learning approaches effective for the SeND task? (see Sec. 5.1) We focus on the data representation learned via a pre-training stage on ImageNet1k. We consider several state-of-the-art learning methods and for all of them, we keep the same prototype-based evaluation strategy used for ReSeND: every class of the support set is identified by averaging on the feature representation of its samples, and the normality score for each test instance is evaluated by measuring the similarity with the nearest known class centroid.

We choose two families of methods. Among the cross-entropy based classifiers we consider this loss applied to **ResNet** [28] and **ViT** [18] architectures, and the data augmentation-based approach **CutMix** [83]. For the contrastive learning techniques we consider the self-supervised methods **SimCLR** [10] and **CSI** [75], as well as their supervised versions **SupCLR** [39] and **SupCSI** [75]. The relation between each test sample and the class prototypes is measured via the Euclidean similarity (inverse of the Euclidean distance [69]) and the cosine similarity, respective for the cross-entropy and contrastive approaches. We highlight that these methods appeared before in the anomaly and novelty detection literature [46,43,64], but their application always involved a training phase on the support set, while here we run them only on ImageNet1k to get their learned

representation. Note also that the different names identify the characteristics of their learning objective, but all of them share the same backbone architecture: ResNet101 [28] with 44M learnable parameters, comparable to the 40M of ReSeND (11M for f_θ, 29M for $r_\gamma + c_\delta$). The only exception is ViT, that we included as an example of Vision Transformer whose usage for OOD was suggested in [43], and for which we use the Vit-Base (86M parameters) implementation from [76].

Is the learned embedding robust to domain variations? (See Sec. 5.2) ImageNet1k contains pictures of real-world objects and it is important to check if the relations encoded in the learned embedding are still relevant when the final goal is to identify novel classes in completely different contexts as for texture images or among sketches. We consider two levels of difficulty. The first is due to a domain difference between the pre-training and the downstream task: the support and test set are drawn from the same domain which is different from that of ImageNet1k. The second is a domain generalization problem and consider also a domain shift between the support and the test set. The support set can be composed of data from a single or multiple domain sources, while the test is from a target domain. We exploits several datasets to perform a thorough analysis.

Textures [13] is a collection of textural images, it consists of 5,640 images organized in 47 categories. We randomly chose 23 categories as known and 24 as unknown. **DomainNet** [61] is a large-scale dataset of common objects from six different domains with 345 object categories. We use this dataset for both intra-domain and cross-domain experiments. For the first case, we used the Natural Language Toolkit [4] to select 50 categories that do not overlap with ImageNet1k classes. We then randomly selected 25 as known and 25 as unknown. **PACS** [47] is composed of four domains and 7 object categories. We follow the known/unknown division proposed in [72] using 6 categories as known and 1 as unknown. **OfficeHome** [77] consists of four domains and 65 categories. We use it in the single-domain generalization experiments by following [5] for the known/unknown category division (25 known and 40 unknown categories). We adopt the same setting of [72] for the multi-source cross-domain experiments. **Multi-Datasets** is a very realistic setting proposed in [72] where the multi-source condition is naturally determined by the use of several datasets as source domains: Office-31 [65], STL-10 [14], Visda2017 [62]. The partial overlap between the source categories, that is simulated for the PACS and OfficeHome benchmarks, in this case is naturally obtained. Here the target domains (Clipart, Real, Painting, Sketch) come from DomainNet.

How does ReSeND compare with state-of-the-art OOD methods? (See Sec. 5.3) Considering that ReSeND does not need access to the support set in the training stage but relies on it during the evaluation, we can measure the time and computational resources it uses in this last stage and provide the same to the training procedure of state-of-the-art OOD methods. We consider the following baselines: **MSP** [29] which uses the standard maximum softmax probability, **ODIN** [48] a simple approach based on input perturbation and temperature scaling, **Energy** [49] that uses an energy score for OOD uncertainty estimation, **GradNorm** [35] which relies on test-time extracted gradients to detect the out-of-distribution

Table 1. Intra-Domain analysis. Best result in bold and second best underlined.

Rep. Learning	Network	Texture		Real		Sketch		Painting	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Cross Entropy	ResNet [28]	<u>0.678</u>	<u>0.892</u>	0.710	0.860	0.553	<u>0.936</u>	0.651	0.926
Cross Entropy	ViT [18]	0.562	0.919	0.696	<u>0.833</u>	<u>0.554</u>	0.952	<u>0.681</u>	<u>0.850</u>
CutMix [83]	ResNet	0.619	0.922	<u>0.721</u>	0.877	0.542	0.943	0.629	0.927
SimCLR [10]	ResNet	0.529	0.942	0.481	0.944	0.502	0.956	0.510	0.956
SupCLR [39]	ResNet	0.534	0.947	0.561	0.899	0.532	0.946	0.532	0.933
CSI [75]	ResNet	0.651	0.906	0.663	0.887	0.514	0.955	0.621	0.910
SupCSI [75]	ResNet	0.652	0.903	0.695	0.875	0.535	0.953	0.652	0.909
ReSeND		**0.691**	**0.859**	**0.780**	**0.805**	**0.623**	**0.917**	**0.735**	**0.829**

samples, the ViT-based approach **OODFormer** [43] and two methods based on tailored metric estimation: Mahalanobis [45] and Gram [68].

Can ReSeND provide unknown detection abilities to closed-set approaches? (See Sec. 5.4) ReSeND does not need any training on the support set and it may work as a plug-and-play module to provide close-set approaches the ability to work in open-set conditions. We focus on the challenging open-set domain generalization (DG) setting presented in [72] and show how ReSeND can enhance existing approaches. Besides **DAML** introduced in [72], we consider the state-of-the-art multi-source closed-set DG method **SWAD** [7], which looks for flat minima in the learning objective function, and two single-source closed-set methods: **SagNet** [52] disentangles shape from style in the image features to reduce the style bias, while **Diversify** [78] synthesizes images with unseen styles.

5   Experiments

Here we report and discuss the results of our experimental analysis. All the evaluations are done on the basis of two metrics. **AUROC** is the Area Under the Receiver Operating Characteristic curve, obtained by varying the normality decision threshold. **FPR95** corresponds to the false positive rate of out-of-distribution examples when the true positive rate of in-distribution examples is at 95%. For the open-set DG experiments we follow [72] and consider also the overall accuracy on the known samples **Acc** and the harmonic mean between the accuracy on known classes and the unknown detection accuracy **H-score**. Implementation[3] details and more experimental analyses are provided in the Appendices A and B. All experimental results are averaged over three runs.

5.1   Intra-Domain analysis

For the intra-domain analysis, we consider the support and test sets drawn from the same visual distribution but showing significant differences from ImageNet1k. In particular, all the testbeds were explicitly designed to avoid semantic overlaps with ImageNet1k: this means that neither known nor unknown classes appear

---

[3] The code is available at `https://github.com/FrancescoCappio/ReSeND`

in its label set. Variation in data type and domain further enlarge the appearance gap. The texture benchmark [13] was already used in [36] and covers a completely different data type with respect to ImageNet1k (objects vs textures). Real, Sketch and Painting benchmarks are obtained from the DomainNet dataset [61] and, differently from Texture, they share the same data type (objects) of ImageNet1k and cover the same (Real) or different (Sketch, Painting) visual domains. In Table 1 we can see that ReSeND obtains the best results showing an excellent knowledge transfer capability. On Texture, the second and third best are respectively Cross Entropy on ResNet and SupCSI, but this ranking is not consistent over all the settings and the performance gap with respect to ReSeND remains evident, especially in the case of Sketch and Painting.

5.2   Cross-Domain analysis

In many real-world conditions, it's impossible to avoid the presence of a visual domain shift between training and test data. This usually increases the complexity of the task at hand. A reliable semantic novelty detection method should disregard the domain shift between the support and the test set, focusing only on the semantic content of the data. We compare ReSeND with the same baselines of the previous section, considering two different benchmarks built from the PACS dataset [47]. Here the support set is composed of images of the source domain, while the target domain is used as test set. In the single-source case (Table 2 top), the Photo domain is always used as source, while the three remaining domains are used as target. The multi-source benchmark (Table 2 bottom) is inherited from [72]: each domain is used in turn as target, with the additional difficulty that the support set is composed by multiple sources that have a partial class overlap (see Fig. 3). We notice that SimCLR is particularly effective when the test domain is sketch, but it is outperformed by other approaches in the remaining settings. On the other hand, ReSeND is able to obtain top results in all benchmarks, showing high robustness to the domain shift, despite not including any tailored strategy designed for bridging it.

5.3   OOD with budget-limited finetuning

As previously discussed, ReSeND doesn't need finetuning on the support set to be used for semantic novelty detection. Hence it is not trivial to make a fair comparison with existing OOD methods for which instead the learning phase on the support set is essential. Nevertheless, we believe that it's important to contextualize ReSeND in the current literature to provide a clearer overview of its performance. With this objective in mind, we focus on the challenging PACS multi-source setting and compare against a number of standard and state-of-the-art OOD methods by letting them learn (refine the original ImageNet1k pretrained model) on the support set for the same time and using the same computational resources exploited by ReSeND in the prediction phase ($\sim$ 30s on 1 GPU for the considered benchmark). For what concerns Mahalanobis [45] and Gram [68], given that they are metric-based methods, the distance between

Table 2. Cross-domain analysis. Top: single-source results, Bottom: multi-source results. We consider the PACS dataset with all the possible combinations of source/target as support/test sets. Best result in bold and second best underlined.

Rep. Learning	Network	PACS **Single-Source**							
		ArtPainting		Sketch		Cartoon		**Avg**	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Cross Entropy	ResNet [28]	0.655	0.940	0.519	0.969	0.546	0.958	0.573	0.956
Cross Entropy	ViT [18]	0.593	0.895	0.595	0.881	0.500	0.953	0.562	0.910
CutMix [83]	ResNet	0.663	0.949	0.372	0.981	0.419	0.980	0.485	0.970
SimCLR [10]	ResNet	0.444	0.984	**0.945**	**0.400**	0.401	0.988	<u>0.597</u>	**0.791**
SupCLR [39]	ResNet	0.500	0.909	0.176	1.000	0.469	0.919	0.381	0.942
CSI [75]	ResNet	0.495	0.987	0.591	0.881	0.433	0.978	0.506	0.949
SupCSI [75]	ResNet	0.546	0.976	<u>0.655</u>	<u>0.819</u>	<u>0.567</u>	<u>0.909</u>	0.589	0.901
ReSeND		**0.828**	**0.668**	0.576	0.981	**0.651**	**0.891**	**0.685**	<u>0.847</u>

Rep. Learning	Network	PACS **Multi-Source**									
		ArtPainting		Sketch		Cartoon		Photo		**Avg**	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Cross Entropy	ResNet [28]	0.575	0.947	0.451	1.000	0.547	0.943	0.361	0.991	0.484	0.970
Cross Entropy	ViT [18]	<u>0.611</u>	<u>0.837</u>	0.566	0.944	0.539	0.904	<u>0.932</u>	<u>0.403</u>	<u>0.662</u>	<u>0.772</u>
CutMix [83]	ResNet	0.604	0.895	0.411	1.000	0.407	0.975	0.655	0.942	0.519	0.953
SimCLR [10]	ResNet	0.461	0.953	**0.933**	**0.663**	0.368	0.995	0.739	0.854	0.625	0.866
SupCLR [39]	ResNet	0.581	0.898	0.100	1.000	0.499	0.909	0.467	0.995	0.412	0.951
CSI [75]	ResNet	0.474	0.984	<u>0.702</u>	<u>0.800</u>	<u>0.560</u>	<u>0.977</u>	0.524	0.946	0.565	0.927
SupCSI [75]	ResNet	0.417	0.984	0.660	0.869	0.323	1.000	0.601	0.946	0.500	0.950
ReSeND		**0.750**	**0.820**	0.685	0.894	**0.660**	**0.854**	**0.963**	**0.181**	**0.765**	**0.687**

Table 3. Comparison with finetuning-based state-of-the-art OOD methods. Best result in bold and second best underlined.

OOD Methods	Fine-Tun.	Eval.	PACS **Multi-Source**									
			ArtPainting		Sketch		Cartoon		Photo		**Avg**	
			AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP [29]	✓	✓	0.617	0.973	0.412	0.998	<u>0.781</u>	**0.767**	0.752	0.905	0.640	0.911
ODIN [48]	✓	✓	0.602	0.977	0.425	0.998	**0.785**	<u>0.774</u>	0.782	0.912	0.649	0.915
Energy [49]	✓	✓	0.583	0.987	0.543	0.996	0.687	0.802	0.845	0.924	0.665	0.927
GradNorm [35]	✓	✓	0.637	0.954	0.514	1.000	0.762	**0.767**	0.851	0.861	0.691	0.896
OODformer [43]	✓	✓	<u>0.703</u>	<u>0.929</u>	0.610	0.973	0.776	0.802	0.732	0.773	<u>0.705</u>	<u>0.869</u>
Mahalanobis [45]	✓	✓	0.596	0.976	0.559	0.933	0.682	0.909	<u>0.861</u>	0.849	0.665	0.916
Gram [68]	✓	✓	0.448	0.962	**0.885**	**0.713**	0.536	0.946	0.838	<u>0.579</u>	0.677	0.800
Mahalanobis [45]	✗	✓	0.596	0.976	0.466	0.981	0.593	0.926	0.808	0.935	0.616	0.954
Gram [68]	✗	✓	0.494	0.960	<u>0.840</u>	<u>0.844</u>	0.494	0.954	0.797	0.981	0.656	0.935
ReSeND	✗	✓	**0.750**	**0.820**	0.685	0.894	0.660	0.854	**0.963**	**0.181**	**0.765**	**0.687**

test samples and the support set can be computed also using a non-finetuned model (although this was not the strategy proposed by the authors). Thus, we tested both the finetuned and not finetuned versions. The results in Table 3 show that ReSeND clearly outperforms all the competitors, which would need a longer training period or more resources in order to converge to a good model. This confirms the role of ReSeND as a powerful tool when semantic novelty detection is performed under restrictive budget constraints.

5.4   Open-set Domain Generalization

The good performance obtained by ReSeND in the analyzed settings suggests that it could be directly and successfully applied in various real-world tasks. We focus on the challenging open-set DG problem that was introduced in [72] (see
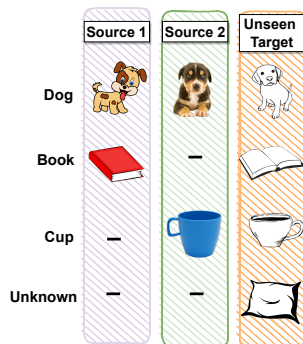
Fig. 3. Open-Set DG setting

Table 4. Open-Set DG experiments.

	Single-Source					
	PACS			Office-Home		
	AUROC	Acc	H-Score	AUROC	Acc	H-Score
ReSeND	0.685	-	-	0.685	-	-
SagNet [52] + MSP	0.643	55.85	48.64	0.699	67.58	59.92
SagNet+ **ReSeND**	**0.700**	55.85	**52.17**	**0.714**	67.58	**61.01**
Diversify [78] + MSP	0.643	52.06	48.12	0.696	70.49	60.03
Diversify+ **ReSeND**	**0.691**	52.06	**51.19**	**0.707**	70.49	**60.77**

	Multi-Source								
	PACS			Office-Home			Multi-Datasets		
	AUROC	Acc	H-Score	AUROC	Acc	H-Score	AUROC	Acc	H-Score
ReSeND	0.765	-	-	0.674	-	-	0.686	-	-
DAML [72] + MSP	0.657	62.85	52.99	0.651	55.28	52.37	0.695	45.90	47.88
DAML+**ReSeND**	**0.722**	62.85	**57.93**	**0.683**	55.28	**54.13**	**0.720**	45.90	**49.96**
Swad [7] + MSP	0.570	60.52	42.85	0.661	53.49	51.06	0.661	47.90	49.10
Swad+**ReSeND**	**0.700**	60.52	**57.05**	**0.682**	53.49	**52.92**	**0.682**	47.90	**50.73**

Table 5. Results obtained by changing the configuration of the relational module. We compare ReSeND with handcrafted feature aggregation strategies for sample pairs.

		PACS - **Multi-Source**								
		ArtPainting		Sketch		Cartoon		Photo		**Avg.**	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ReSeND		0.750	0.820	0.685	0.894	0.660	0.854	0.963	0.181	**0.765**	**0.687**
	Max	0.676	0.899	0.785	0.742	0.616	0.940	0.827	0.786	0.726	0.842
Aggreg.	Sum	0.583	0.976	0.446	0.988	0.514	0.996	0.575	1.000	0.530	0.990
	Concat	0.676	0.842	0.710	0.790	0.635	0.902	0.921	0.438	0.736	0.743

Fig. 3). Multiple source domains are combined together and their different label sets cause some classes to exist in many more domains than other classes. The target is drawn from a different distribution with a large shift with respect to the source, both in terms of style and semantic content. Indeed, the target contains more classes than the source and they should be identified as unknown at test time. Existing closed-set DG methods are able to learn classification models that generalize to the unseen target domain containing the same categories of the source. One simple way to let them reject samples of novel classes is to add a threshold on MSP, considering unknown the samples with uncertain predictions, as done in DAML. We can apply the same technique on SagNet, Diversify and SWAD. Still, the results can take further advantage from a method better suited to spot semantic novelties across domains, as ReSeND.

We consider the source domains as support set and the target as test, running the evaluation procedure of ReSeND to obtain the normality score for each target sample. The obtained values are combined with the MSP produced by each reference method with a simple score averaging as an ensemble strategy. Since the two normality evaluations originate from different input features we aim at leveraging their complementary nature and maximize the final unknown rejection accuracy. The obtained results are shown in Table 4. In all cases, integrating ReSeND with the other methods provides an improvement both in AUROC and

in H-score, with Acc maintaining the exact same values, as ReSeND does not influence predictions on known classes.

6   Further analysis and discussions

Learnable Relational Module. To assess the influence of our design choices for the relational module in ReSeND, we consider alternative strategies to combine the features of sample pairs. Specifically, we evaluate the effect of substituting our transformer-based relational module with hand-designed aggregation functions ($Max/Sum/Concat$), followed by an MLP whose output is fed to the final semantic similarity head. The MLP module is designed to have a similar number of learnable parameters with respect to our transformer-based one. For *Concat* we exploit the feature concatenation as already done in [60]. Note that the *permutation invariance property* of our transformer gets lost by feature concatenation: the order of the images in the pair influences the final predictions.

Table 5 reports the results of this analysis on the PACS multi-source setting. We argue that the superior performance of ReSeND originates from having learned the feature aggregation function rather than relying on a fixed approach imposed a priori. Still, *Max* and *Concat* are able to obtain quite good results (better than what was obtained by the second best in Table 3, OODFormer [43] Avg$_{AUROC}$: 0.705, Avg$_{FPR95}$: 0.869). This is an additional evidence of the effectiveness of the relational reasoning approach for semantic novelty detection.

We remark that an important characteristics of ReSeND is its ability to learn jointly the feature embedding and the semantic similarity metric through an end-to-end training. As highlighted by Sung et al. [74] this is a superior strategy with respect to both methods that learn the feature embeddings but use a fixed similarity measure (e.g. Euclidean) [21], and methods that instead learn a similarity measure on top of a fixed feature representation [50,8].

Regression vs Classification. As mentioned in Sec. 3.2, the relational reasoning learning paradigm can be cast as both a binary classification and a regression problem. We believe the latter is more conceptually appropriate as we want to learn a semantic similarity measure with a continuous value. The alternative solution consists in a binary *same* vs *different* task, in which the prediction for the class *same* could be used as semantic similarity measure. In practice, what really differentiates the two approaches is the trend of the loss function.

In Fig. 4 we represent the loss when varying the probability assigned to the correct class for both the classification cross entropy (CE) and the regression MSE. In both cases a high loss is assigned to a low probability and vice-versa. In the very small and rarely populated region of low probability values ($p \approx 0$), CE is higher than MSE. While the MSE gives more importance through higher loss values to hard samples belonging to the intermediate probability region, the CE focuses more on easy samples ($p > 0.75$) pushing their already high probability values to the same even higher output. The final effect of the CE is a minimization of the difference among the samples, which is not ideal when we want to use the confidence as a semantic similarity metric.
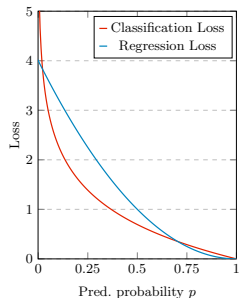
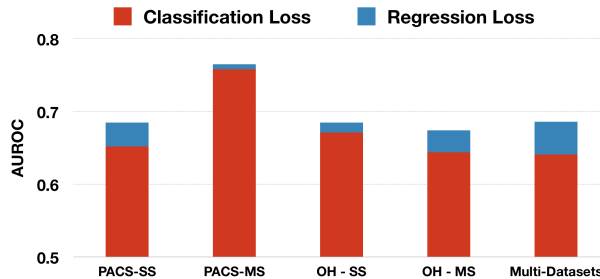**Fig. 4.** Loss trend for the probability of the correct class.

Fig. 5. AUROC comparison with ReSeND trained for classification via Cross Entropy Loss or for Regression via MSE. OH stands for Office-Home. SS and MS indicate respectively the Single- and Multi-Source settings.

We compare the performance obtained by ReSeND with the two different choices for the loss in Fig. 5. We considered all the dataset benchmarks already used for the open-set DG analysis and we show how both the losses provide good results, with the regression outperforming the classification one in all the cases.

7   Conclusions

In this paper we analyzed the problem of semantic novelty detection by extensively studying how traditional representation learning methods can be used for this task. Moreover, we introduced ReSeND a representation learning approach that exploits relational reasoning to model semantic similarity among pairs of samples. ReSeND exploits a basic transformer architecture and, once trained on ImageNet1k, it allows to identify whether a test sample belongs to a known or an unknown category by simply comparing it with the reference support set without the need for finetuning. Our thorough experimental analysis has demonstrated the effectiveness of ReSeND in both intra- and cross-domain settings, and its potential as plug-and-play module to transform closed-set domain generalization approaches into reliable open-set methods with state-of-the-art results.

A trustworthy semantic novelty detection method that is able to prevent wrong annotations by identifying unknown categories without any training time latency is a crucial component in many real-world applications. We believe that our work can pave the way for more research in this direction, focusing on novel paradigms or more advanced architectures for relational reasoning.

Algorithm 1 ReSeND train procedure

Require: $\mathcal{S}, \mathcal{T}, f_\theta : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d, r_\gamma, c_\delta$

  procedure CREATE_PAIRS($\mathcal{S}$)

      $pairs = []$

      for each $(\boldsymbol{x}^s, y^s)$ **in** $\{\mathcal{S}\}$ **do**

         $pairs.append((rand\_same\_class(y^s), \boldsymbol{x}^s, 1))$

         $pairs.append((rand\_diff\_class(y^s), \boldsymbol{x}^s, 0))$

      return pairs

  procedure MAIN()

      for $epoch$ **in** $range(n\_epochs)$ **do**

         $pairs = create\_pairs(\mathcal{S})$

         $shuffle(pairs)$

         for $iter$ **in** $range(iters\_epoch)$ **do**

            $pairs\_batch = next\_batch(pairs)$

            $\boldsymbol{x}_1, \boldsymbol{x}_2, labels = pairs\_batch$

            $\boldsymbol{z}_1 = f_\theta(\boldsymbol{x}_1)$

            $\boldsymbol{z}_2 = f_\theta(\boldsymbol{x}_2)$

            $feats\_pairs = (\boldsymbol{z}_1, \boldsymbol{z}_2)$

            $predictions = c_\delta(r_\gamma(feats\_pairs))$

            $MSE\_loss = \mathcal{L}(predictions, labels)$ ▷ Eq. 1

            Update $\theta, \gamma, \delta \leftarrow \nabla MSE\_loss$

A Implementation details

We start from a standard ResNet-18 [28], pretrained on ImageNet1k [17], which we use as feature extractor f_θ by removing the original final classification layer. Our relational module r_γ has the same structure of the transformer in ViT [18]: we use 4 multi-head self-attention encoder blocks, a number that allows to trade-off performance and time complexity (the number of blocks highly influences the total number of learnable parameters of the network). The features extracted by the backbone are passed through an FC projection layer before entering the transformer. The transformer input sequence is obtained concatenating the learnable label token and the representations of a pair of samples $[\boldsymbol{z}_l, \boldsymbol{z}_i, \boldsymbol{z}_j]$. The output token $\boldsymbol{v}_l$ is then selected and passed through a final FC layer which represents the regression head c_δ.

The transformer procedure is summarized in the following equations:

$$\boldsymbol{z}^0 = [\boldsymbol{z}_l;\, \boldsymbol{z}_i;\, \boldsymbol{z}_j] \tag{2}$$

$$\tilde{\boldsymbol{z}}^b = \mathrm{MSA}(\mathrm{LN}(\boldsymbol{z}^{b-1})) + \boldsymbol{z}^{b-1}, \qquad b = 1 \ldots B \tag{3}$$

$$\boldsymbol{z}^b = \mathrm{MLP}(\mathrm{LN}(\tilde{\boldsymbol{z}}^b)) + \tilde{\boldsymbol{z}}^b, \qquad b = 1 \ldots B \tag{4}$$

$$\boldsymbol{v}_l = \mathrm{LN}(\boldsymbol{z}_l^B) \;. \tag{5}$$

We train our network on ImageNet1k in an end-to-end manner using the MSE loss (Eq. 1 in the main paper) applied to the output of the regression head. Our training procedure uses 13k iterations with a batch size of 4096, where each

---

Algorithm 2 ReSeND eval procedure

Require: $\mathcal{S}, \mathcal{T}, f_\theta : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^d, r_\gamma, c_\delta$

  procedure COMPUTE_PROTOTYPES($\mathcal{S}$)

      $prototypes = zeros((|\mathcal{Y}_s|,))$

      $counters = zeros(|\mathcal{Y}_s|)$

      for each $(\boldsymbol{x}^s, y^s)$ **in** $\{\mathcal{S}\}$ **do**

         $\boldsymbol{z}^s = f_\theta(\boldsymbol{x}^s)$

         $prototypes[y^s] + = \boldsymbol{z}^s$

         $counters[y^s] + = 1$

      for i **in** $range(|\mathcal{Y}_s|)$ **do**

         $prototypes[i] / = counters[i]$

      return $prototypes$

  procedure MAIN()

      $normality\_scores = []$

      $prototypes = compute\_prototypes(\mathcal{S})$

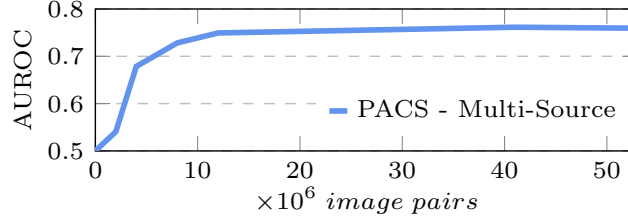   **for each** $\boldsymbol{x}^t$ **in** $\{\mathcal{T}\}$ **do**

         $\boldsymbol{z}^t = f_\theta(\boldsymbol{x}^t)$

         $pairs = (prototypes, \boldsymbol{z}^t.repeat())$

         $predictions = c_\delta(r_\gamma(pairs))$

         $score = max(softmax(predictions))$

         $normality\_scores.append(score)$

Fig. 6. Performance trend increasing the number of image pairs.

element of the batch is an image pair. The learning rate uses a linear warmup for 500 iterations and then is fixed to 0.008. We use LARS optimizer [82] with momentum 0.9 and weight decay $5 \cdot 10^{-5}$. We build image pairs by selecting each image of the dataset as anchor and associating it with a randomly chosen sample with the same label to create *positive* pairs and samples of different labels to create *negative* pairs. All experimental results are averaged over three runs. We summarize in Algorithm 1 and 2 the training and evaluation procedure of ReSeND.

B   Further Analysis

Number of image pairs. Our learning objective is based on the use of image pairs randomly created at training time by coupling samples from the training

dataset. Even if the total number of image pairs that could be formed from ImageNet1k dataset is very high ($\sim 820 \times 10^9$), in Fig. 6 we show that ReSeND converges after having seen a relatively small portion of them.

References

1. Battaglia, P., Hamrick, J.B.C., Bapst, V., Sanchez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G.E., Vaswani, A., Allen, K., Nash, C., Langston, V.J., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., Pascanu, R.: Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 (2018)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE TPAMI **35**(8), 1798–1828 (aug 2013)
3. Bergman, L., Hoshen, Y.: Classification-based anomaly detection for general data. In: ICLR (2020)
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
5. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: ECCV (2020)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
7. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. In: NeurIPS (2021)
8. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: A joint formulation. In: ECCV (2012)
9. Chen, J., Li, Y., Wu, X., Liang, Y., Jha, S.: Atom: Robustifying out-of-distribution detection using outlier mining. In: ECML (2021)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
12. Cheng, Y., Wang, R., Pan, Z., Feng, R., Zhang, Y.: Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In: ACM Multimedia (2020)
13. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
14. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS (2011)
15. Collin, A.S., De Vleeschouwer, C.: Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In: ICPR (2021)
16. Deecke, L., Ruff, L., Vandermeulen, R.A., Bilen, H.: Transfer-based semantic anomaly detection. In: ICML (2021)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
18. 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)

19. Du, Y., Gan, C., Isola, P.: Curious representation learning for embodied intelligence. In: ICCV (2021)
20. Ericsson, L., Gouk, H., Hospedales, T.M.: How Well Do Self-Supervised Models Transfer? In: CVPR (2021)
21. Fontanel, D., Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Boosting deep open world recognition by clustering. IEEE RAL **5**(4), 5985–5992 (2020)
22. Ge, Z., Demyanov, S., Chen, Z., Garnavi, R.: Generative openmax for multi-class open set classification. In: BMVC (2017)
23. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
24. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: NeurIPS (2018)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
26. Goodfellow, I.J., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
27. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
29. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
30. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
31. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (Jul 2006)
32. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)
33. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: CVPR (2020)
34. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018)
35. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. NeurIPS (2021)
36. Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: CVPR (2021)
37. Jenni, S., Jin, H., Favaro, P.: Steering self-supervised feature learning beyond local pixel statistics. In: CVPR (2020)
38. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
39. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS (2020)
40. Kim, K.H., Shim, S., Lim, Y., Jeon, J., Choi, J., Kim, B., Yoon, A.S.: Rapp: Novelty detection with reconstruction along projection pathway. In: ICLR (2020)
41. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: ICLR (2014)
42. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: CVPR (2019)
43. Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V.: Oodformer: Out-of-distribution detection transformer. In: BMVC (2021)

44. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: ICLR (2018)
45. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
46. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: CVPR (2021)
47. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV (2017)
48. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
49. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. NeurIPS (2020)
50. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In: ECCV (2012)
51. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: ICLR (2019)
52. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: CVPR (2021)
53. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: ECCV (2018)
54. Newell, A., Deng, J.: How useful is self-supervised pretraining for visual tasks? In: CVPR (2020)
55. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
56. Oza, P., Nguyen, H.V., Patel, V.M.: Multiple class novelty detection under data distribution shift. In: ECCV (2020)
57. Pan, J., Chen, S., Shou, M.Z., Liu, Y., Shao, J., Li, H.: Actor-context-actor relation network for spatio-temporal action localization. In: CVPR (2021)
58. Papadopoulos, A.A., Rajati, M.R., Shaikh, N., Wang, J.: Outlier exposure with confidence control for out-of-distribution detection. Neurocomputing **441**, 138–150 (2021)
59. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: CVPR (2020)
60. Patacchiola, M., Storkey, A.: Self-supervised relational reasoning for representation learning. In: NeurIPS (2020)
61. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
62. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017)
63. Raposo, D., Santoro, A., Barrett, D.G.T., Pascanu, R., Lillicrap, T., Battaglia, P.W.: Discovering objects and their relations from entangled scene representations. In: ICLR Workshop (2017)
64. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)
65. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV (2010)
66. Santoro, A., Faulkner, R., Raposo, D., Rae, J.W., Chrzanowski, M., Weber, T., Wierstra, D., Vinyals, O., Pascanu, R., Lillicrap, T.P.: Relational recurrent neural networks. In: NeurIPS (2018)

67. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P.W., Lillicrap, T.: A simple neural network module for relational reasoning. In: NeurIPS (2017)
68. Sastry, C.S., Oore, S.: Detecting out-of-distribution examples with Gram matrices. In: ICML (2020)
69. Segaran, T.: Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly (2007)
70. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. In: ICLR (2021)
71. Sensoy, M., Kaplan, L.M., Cerutti, F., Saleki, M.: Uncertainty-aware deep classifiers using generative models. In: AAAI (2020)
72. Shu, Y., Cao, Z., Wang, C., Wang, J., Long, M.: Open domain generalization with domain-augmented meta-learning. In: CVPR (2021)
73. Stanislav Fort, J.R., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. In: NeurIPS (2021)
74. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018)
75. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: NeurIPS (2020)
76. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
77. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
78. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV (2021)
79. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S.M.A., Ronneberger, O.: Contrastive training for improved out-of-distribution detection. arXiv:2007.05566 (2020)
80. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.: Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In: ECCV (2020)
81. Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., Liu, Z.: Semantically coherent out-of-distribution detection. In: ICCV (2021)
82. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv:1708.03888 (2017)
83. 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)
84. Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., Shanahan, M., Langston, V., Pascanu, R., Botvinick, M., Vinyals, O., Battaglia, P.: Deep reinforcement learning with relational inductive biases. In: ICLR (2019)
85. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: ICML (2021)
86. Zhang, H., Koniusz, P., Jian, S., Li, H., Torr, P.H.S.: Rethinking class relations: Absolute-relative supervised and unsupervised few-shot learning. In: CVPR (2021)