Learning to Discover and Detect Objects

Anonymous Author(s)
Affiliation
Address
email

Abstract

We tackle the problem of novel class discovery, detection, and localization (NCDL). In this setting, we assume a *source* dataset with labels for objects of commonly observed classes. Instances of other classes need to be discovered, classified, and localized automatically based on visual similarity, without human supervision. To this end, we propose a two-stage object detection network *Region-based NCDL* (*RNCDL*), that uses a region proposal network to localize potential objects and classify them. We train our network to classify each proposal, either as one of the *known* classes, seen in the source dataset, or one of the extended set of *novel* classes with a constraint that the distribution of class assignments should follow natural long-tail distributions common in the real open-world. By training our detection network with this objective in an end-to-end manner, it learns to classify all region proposals for a large variety of classes, including those that are not part of the labeled object class vocabulary. Our experiments conducted using COCO and LVIS datasets reveal that our method is significantly more effective compared to multi-stage pipelines that rely on traditional clustering algorithms or self-supervised contrastive learning methods and operate on pre-extracted crops. Beyond that, we demonstrate the generality of our approach by applying our method to a large-scale Visual Genome dataset, where our network successfully learns to detect various semantic classes without explicit supervision.

1 Introduction

We tackle novel class discovery and localization in unlabeled datasets, a long-standing problem in computer vision [51, 55, 37, 38, 50]. As we are approaching the limits of well-understood and successful supervised training of object detectors [20, 19, 49, 25], this is becoming an increasingly more important problem, with the task of labeling novel and rare classes that appear in the long tail of object class distribution becoming prohibitively expensive [22].

Object class discovery is a difficult problem, as there may be many possible equally valid attributes for grouping patterns, such as object color or orientation. As shown in [27, 28], *novel class discovery (NCD)* can be well-posed and tackled in a data-driven manner by injecting prior knowledge on *how we wish to group semantic classes* using some degree of supervision. Existing methods [27, 28, 24, 23, 67, 7, 68, 69, 17] all assume curated datasets, where objects of interest are pre-cropped and semantic classes are fairly balanced. In contrast, we tackle joint *novel class discovery and localization (NCDL)*, a task of learning to discover and detect objects from raw, unlabeled, and uncurated data. This is a significantly more challenging problem, as images always contain a mixture of *labeled* and *unlabeled* object classes, and we need to localize as well as categorize them.

To study this problem, we re-purpose the COCO [40] and LVIS [22] datasets. We use 50% of COCO images together with COCO class vocabulary and labels for supervised model training. We use the remaining 50% of images without any labels during the training and evaluate how well our network learns to detect novel classes using the long-tailed LVIS label set. This setting allows us to mimic

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

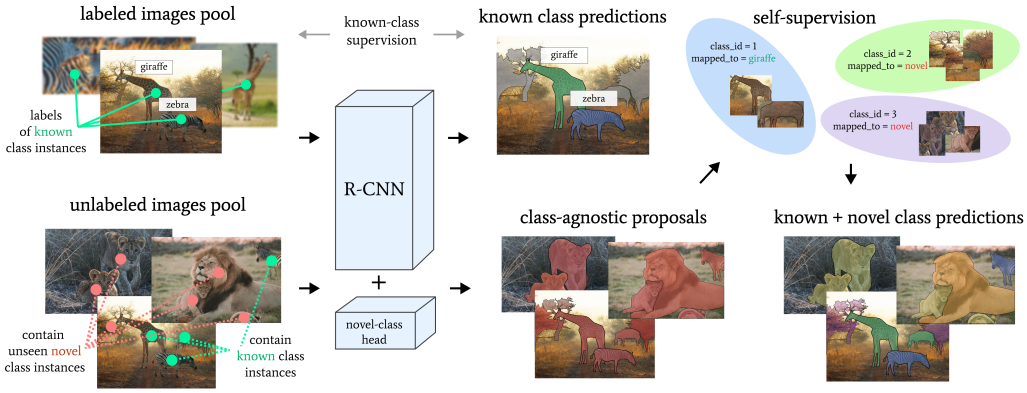

Figure 1: In NCDL problem we operate with labeled images pool containing annotations for *known*, frequently observed semantic classes, and unlabeled images pool that may contain instances of *novel* classes. Our network learns to localize and recognize common semantic classes, as well as categorize novel classes, for which supervision was not given.

real-world scenarios where no supervision is given for the target domain, while a rich set of labels allows us to rigorously study this problem.

We train a two-stage object detector [49, 25] in an end-to-end manner such that (i) we can correctly localize and categorize, *i.e.*, detect *labeled* objects, and, in parallel, (ii) learn feature representations, suitable for categorization of objects of novel categories that would otherwise be classified as the *background* class. We start with supervised training on the labeled, *source* domain, followed by the *self-supervised* training on the target (unlabeled) domain. We transfer knowledge from the source to the target domain by *retaining* weights for class-agnostic modules, such as RPN and bounding box regression, and instance segmentation heads. For classification, we reuse the primary classification head and add a new secondary head, training them jointly with an objective to categorize *every* region proposal consistently under multiple transformations and a constraint that the marginal distribution of class assignments follows a long-tail distribution. By encouraging the network to learn features for a variety of objects via a non-uniform classification prior, we also effectively reduce the bias towards simply classifying regions as one of the *labeled* or *background* classes.

Our end-to-end trainable method significantly outperforms prior art [63] that learns object representations using contrastive learning followed by k-means clustering, while providing categorization for *any* object with a single network forward pass. More importantly, with this work, we lay down the foundations for an exciting new line of research that goes beyond well-established supervised learning for object detection.

In summary, this paper makes the following **contributions**: (i) We study the problem of *novel class discovery* in the wild using standard object detection datasets that do not pre-localize regions of interest manually or assume objects are roughly centered, *i.e.*, photographer bias. (ii) We propose a simple yet effective and scalable method that can be trained end-to-end on the unlabeled domain. Our network is bootstrapped using a labeled dataset to learn features suitable for categorizing future instances. We train feature representations for arbitrary objects using the self-supervised objective. Finally (iii), we reveal that our method can discover and detect novel class instances and outperform prior work, demonstrating the generalization to datasets beyond COCO.

2 Related Work

Object discovery in image and video collections. Discovery of semantic objects that appear in unlabeled video collections is a long-standing research problem [56]. Early methods employ multiple bottom-up segmentations of images [51, 55] to discover commonly occurring patterns using topic modeling, *e.g.*, Dirichlet allocation [6]. Lee *et al.* propose to rely on saliency [37] or to incorporate semantic knowledge about certain semantic classes, for which supervision is available [38]. These methods are evaluated on smaller datasets [21, 54, 16] that contain one or a handful of well-delineated objects. Rubinstein *et al.* [50] demonstrated that discovery can be scaled to internet photo collections.

Beyond images, Kwak *et al.* [36] tackle the discovery of objects in YouTube videos that usually contain a single object exhibiting dominant motion. They build on a two-stage approach: (i) identify the most salient region proposals in terms of appearance and motion, and (ii) co-segment consistently appearing objects across video clips. Video object discovery videos, recorded from a moving platform, were tackled in [45, 46] by first mining video-object tracks by associating region proposal network (RPN) [49] based proposals across time, followed by clustering of video proposals based on pre-trained CNN features. Recent work on open-set panoptic segmentation [30] similarly employs clustering based on RPN features to discover commonly-occurring objects, used as pseudo-labels for learning to segment unlabeled objects. The method by [63] follows the general approach of clustering RPN proposals, however, instead of relying on pre-trained CNN features, they learn representations suitable for clustering using self-supervised objective, trained via contrastive learning.

While most of the foregoing works focus on saliency criteria to significantly narrow down the number of region proposals, [58, 59] resort to combinatorial optimization that can cope with a large set of object proposals to discover objects that frequently appear in image datasets. It has been shown [60] that this approach can be scaled to large image collections, *e.g.*, OpenImages [35].

Novel class discovery. Works from a related Novel class discovery (NCD) field operate under an assumption that given is a labeled data that can be leveraged to train a feature extractor along with learning a clustering criterion to guide a notion of similarity and unlabeled data that is assumed to contain novel class instances only. Hsu *et al.* [27, 28] propose a framework that injects prior knowledge on how we wish to group semantic classes from supervised learning on *labeled* images by transferring knowledge to the *unlabeled* domain via learnable similarity function. The work of Han *et al.* [24] builds on deep clustering [66] and suggests an approach for automatically discovering the number of novel classes. In follow-up work, Han *et al.* [23] suggest an alternative training objective based on pairwise similarity of image features [11] using pseudo-labels for fine-tuning the network and propose leveraging self-supervised pretraining and freezing the backbone to reduce bias towards the known classes. Its follow-up works [67, 32, 69, 68] further boost the performance by introducing additional self-supervised and regularization components. UNO [17] proposes a simple method based on SwAV [10] that bootstraps signal for novel class training using pseudo-labels generated under equipartitioning constraints. ORCA [7] extends the NCD problem and operates under the assumption that unlabeled images can contain instances of known classes as well. Motivated by [57], [11], [61], authors leverage self-supervised initialization of backbone, supervised loss along with unsupervised loss based on pairwise pseudo-labels, and novel dynamic softmax margin [61] to alleviate bias of the known classes. In contrast to our work, the foregoing methods investigate the problem in the image classification setting, assuming objects of interest are localized (pre-cropped), and operate under absence of outliers (*e.g.*, images containing no objects).

Related and complementary research directions. Methods for *generic* [41] or *large vocabulary object detection* [22] focus on detecting/segmenting a large set of a-priori known semantic classes, often by utilizing multiple datasets and (possibly weak) supervision [70, 48, 29]. A special case is *zero-shot learning* [65] and *object detection* [3, 44, 47], where (very weak) supervision for *unseen* classes comes in the form of attributes, or class names, commonly used in a conjunction with language models embeddings to learn a joint embedding space, however, all the target classes are known and pre-determined. In contrast to zero-shot and large-vocabulary detection, we assume no supervision and prior knowledge for object classes that appear in the tail.

Methods for *open-set recognition* [52, 53, 31, 5] and *detection* [44, 15] focus on calibration of per-class uncertainties to minimize the confusion between known and possibly unknown classes.

Open-world recognition methods, as defined by [4, 42] must explicitly recognize *unknown* object instances that were not observed as labeled samples during the model training, and must *continually* update object detectors to recognize these unknown instances. Recent method for *open-world* detection [33] focuses on the *continual learning* aspect of the task: for the evaluation, labels for "novel" classes are fully provided (those are the held-out classes). Our work is orthogonal, and tackles the "missing bit" of the open-world detection pipeline: discovery and semantic grouping of *novel*-class objects that *can appear* in our training data as well, however, are unlabeled. Clusters of discovered *novel* classes could be verified by human annotators and used to update detection models in a continual fashion, as proposed by [4].

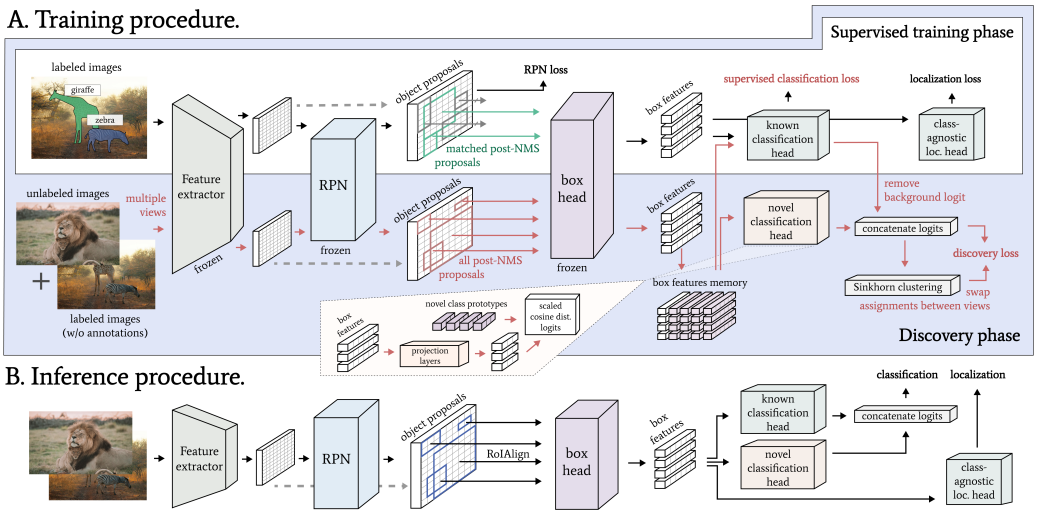

Figure 2: A high-level overview of our network. *(top)* ***During the supervised training phase***, we train our backbone and RPN networks, together with classification head for the *labeled* classes, a class agnostic localization head. ***During the discovery phase*** we freeze all the layers of the network apart from classification head and attach a novel classification head. ***During the inference*** *(bottom)* we perform a standard R-CNN pass, using classification heads of both known and novel categories to predict proposals classes.

3 Novel Class Discovery and Localization

In this section, we detail *novel class discovery and localization (NCDL)*, the task of learning to discover and detect objects from raw, unlabeled, and uncurated data. During the model training, we assume we are given a source (*labeled*) dataset that provides bounding box or segmentation mask level supervision for K object classes. This data can be used to learn features useful for categorization and other aspects of the task, such as bounding box regression and mask segmentation, which can be learned in a class-agnostic manner. In addition, we are given one (or more) *unlabeled* datasets that may contain instances of both previously seen and $N, N > K$ visual object classes that we need to learn to detect in the absence of any labels. Given an arbitrary image from the source or target domain, methods should detect and localize all object instances present in the image, along with predicting their respective semantic class (*i.e.*, a car needs to be detected as the `car` class). As the main contribution of this paper, we propose an end-to-end trainable detection and discovery network that jointly learns discriminative representations for all region proposals, their localization, and categorization alleviating the need for explicit clustering after the network training.

3.1 Region-based Novel Category Discovery and Localization (RNCDL)

We start with a two-stage Faster or Mask RCNN [49, 25]. Such detectors are naturally suitable for NCDL as the region proposal network (RPN) is trained in a class-agnostic manner to propose a variety of regions (anchor boxes) of interest that likely contain objects. Given object-centric feature representations for each candidate region that can be learned via supervised training, one natural approach for discovering *novel* classes would be to cluster such regions in feature space using, *e.g.*, k-means, *c.f.*, [45, 30, 63]. However, in the following, we argue that feature representation learned via supervised RCNN training is not trained to be discriminative with respect to objects of *unlabeled* classes, as every anchor box that does not have sufficient overlap with labeled boxes is classified as *background* class during training. Via this process, feature representations for all regions that do not contain objects of *labeled* classes are effectively *collapsed* and not discriminative for robust categorization of novel classes. Furthermore, such an approach has a strong bias toward classifying novel instances as one of the *labeled* classes. In the following, we discuss our approach for learning to categorize *all* anchor boxes directly on the target (*unlabeled*) dataset in an end-to-end manner.

Supervised training on the source domain. We start with a self-supervised backbone pre-training [26] on ImageNet [14] and train our object detection network on the source (labeled) domain using supervision for frequently occurring objects, such as 80 COCO classes (Fig. 2A, *top*). We use the labeled data to train the shared feature extractor (backbone) together with a class-agnostic region proposal network (RPN), class-agnostic bounding box regression and segmentation heads, as well as a classification head for categorization of *labeled* classes (Fig. 2A). The features learned by the backbone at this stage are trained for categorization of *labeled* objects to semantic classes and the network is incentivized to learn representations that are agnostic to *e.g.*, color, object, orientation.

Discovery on the target domain. During the following self-supervised (discovery) training phase on the target domain we freeze all the layers apart from the primary (known) classification head, remove its *background* class, and attach a secondary (novel) classification head (*i.e.*, categorization head) on top of RoI box features (Fig. 2B). To obtain classification predictions at this phase, we pass RoI features through both *primary* classification head and *novel* classification head, concatenating their logits. We train the classification heads *jointly* using *supervised classification* and *self-supervised classification* losses. The *supervised classification loss* effectively alleviates forgetting the weights for the known classes. Different from standard supervised training, this loss is (i) evaluated based on logits from *both* heads, and (ii) calculated only for positive region proposals, annotation-matched with source dataset, excluding *background* proposals. The *self-supervised classification loss* encourages the network to categorize every region proposal consistently under multiple transformations and a constraint that the marginal distribution of predicted class assignments for the last batches follows a prior long-tail distribution (Fig. 2A, *bottom*). With such a joint training objective, we encourage our network to learn to distinguish a large variety of semantic classes present in the regions sampled by the RPN, instead of collapsing feature representations for unmatched regions into a *background* class, while retaining the capability to classify *known* objects.

Self-supervision via swapping cluster assignments. To learn to discover novel classes we leverage class-agnostic RoI proposals generated by RPN and build on self-labeling. We leverage clustering methods to generate cluster assignments (pseudo-labels) for each of the extracted object proposals in an online fashion and use the resulting labels to train the classification heads. Leveraging standard clustering methods (*e.g.* k-means [43]) to obtain such pseudo-labels was shown to result in unstable training [8, 9]. Instead, we follow [2, 1] and generate cluster assignments under a constraint that (i) the marginal probability distribution of generated assignments follows a specific prior distribution, and (ii) each proposal is classified consistently under multiple transformations.

We first extract RoI proposals and their features from the current batch using RPN and the shared backbone. We forward the resulting features through both known and novel classification heads to obtain logits for all classes. We then follow [2, 1] to generate pseudo-labels for each of the proposals forcing the marginals of resulting current-batch cluster assignments to follow Gaussian distribution. For that, following [2, 1] we optimize a cross-entropy loss between predicted class probabilities and cluster assignments w.r.t. the assignments. This can be posed as a constrained optimization problem and solved as an optimal transport problem using the Sinkhorn-Knopp algorithm [13]. The resulting cluster assignments are then used to calculate cross-entropy loss.

We further extend our approach and enforce the network to output consistent features and classification predictions for proposals under multiple augmentations, similarly to [10]. For that, for each image after generating the proposals coordinates, we first augment the image twice and only then extract the features of the proposals for both of the image views. This way we obtain two sets of features for the same set of proposals. We generate pseudo-labels for each view (set of proposals' features) individually, swap the resulting cluster assignments between views, and calculate the total loss as the average of the losses for both views.

Secondary classification head architecture. As the box features are kept frozen during the discovery phase, we add several non-linear projection layers on top of the box features. This helps to learn to better disentangle feature representation of novel classes and aids the training of novel classification head, as shown experimentally in the ablations section (Sec. 4.2). We employ cosine classification layer, proposed in the context of few-shot learning [18, 12, 62] and related fields [10, 17, 7]. We note that the classification layer size directly implies the expected number of novel classes. As it is problem-specific and requires analysis of desired granularity to match the labels of the target dataset, we perform an analysis of the sensitivity of our method to the layer size in ablations (Sec. 4.2).

Memory module. Generating pseudo-labels of a good quality requires a large set of features to represent the wide variety of object representations present in the data and having a small set of features is likely to result in noisy cluster assignments. As we operate in scenarios with more than a thousand classes to discover, instead of using large batch sizes, we introduce a memory module [26, 10] to our network to store proposals' features from the last batches. For the Sinkhorn algorithm, at each iteration, we use features from both the memory module and current-batch proposals. From the resulting pseudo-labels, we use only those generated for the current-batch proposals for loss calculation and ignore the rest. For each view, we use a separate feature memory module.

Losses. During the supervised phase we use a standard fully-supervised RCNN loss: $loss_{sup} = loss_{RPN} + loss_{box} + loss_{cls} + loss_{mask}$, where each of the components corresponds to the RPN loss, localization branch loss, classification branch loss, and mask branch loss respectively [49, 25]. During the discovery phase, we keep a downscaled classification loss component and introduce a self-supervised loss to train both classification heads: $loss_{disc} = loss_{ss} + \alpha \cdot loss_{cls}$, where $loss_{ss}$ and α are the self-supervised loss and supervised loss scale coefficient respectively. For the self-supervised loss, we use cross-entropy loss on top of pseudo-labels generated online for the current batch of proposals. We show in ablations that keeping the downscaled ($\alpha < 1$) supervised loss in the loop is beneficial and helps to avoid forgetting the known classes in the known-class classification head, while large α may create an unwanted bias towards known classes and harm learning of novel classes.

Inference. During the inference (Fig. 2B) we perform a standard forward RCNN pass [49, 25] based on concatenated logits from both classification heads.

4 Experimental Evaluation

In this section, we discuss our evaluation test-bed, including datasets, evaluation settings, and metrics (Sec. 4.1). In Sec. 4.2 we justify our main design decisions by studying NCDL performance in a well-controlled setting that closely follows a real-world scenario (training on a source *labeled* and target *unlabeled* dataset). We then compare our method's performance to several baselines (Sec. 4.3) and, finally, demonstrate that our method is applicable beyond our evaluation test-bed (Sec. 4.4).

4.1 Evaluation setting

COCO$_{half}$ + LVIS. We re-purpose COCO 2017 [40] and LVIS v1 [22] datasets for running ablations and comparisons with baselines and prior art. We use annotations for 80 COCO classes during the supervised training phase, aiming at further classifying additional $1000+$ LVIS classes. COCO dataset contains $123K$ images with modal bounding box and segmentation mask annotations for 80 classes. The LVIS dataset contains a subset of $120K$ COCO images with annotations for 1203 classes that include all classes from COCO. For our experiments, we follow LVIS training and validation splits, resulting in $100K$ training images and $20K$ validation images. To model the NCDL setup, we further split $100K$ training images in half and use only $50K$ images with 80-class annotations from COCO during the supervised training phase. We refer to this dataset as COCO$_{half}$. We treat the rest of $50K$ images as additional unlabeled data used during the discovery phase.

LVIS + Visual Genome. We perform a large-scale generalization experiment by using annotations for 1203 LVIS classes during the supervised learning (*i.e.*, treating LVIS classes as *labeled*), aiming to learn to discover extra $2700+$ classes from the Visual Genome (VG) v1.4 dataset [34] (*i.e.*, treating VG classes as *unlabeled*). We provide dataset details and splits (that ensure LVIS and VG class vocabularies do not overlap) in the supplementary. We note that annotations provided in VG are not exhaustive per class and many of its classes are abstract or semantically overlapping both within VG and when compared to LVIS. We thus focus on qualitative results, following the common practice, and provide quantitative results in supplementary.

Implementation details. We base our model on Mask R-CNN [25] and FPN [39] implementations from Detectron2 [64]. We give all the implementation details in the supplementary material.

Table 1: Impact of supervised loss strength. We check the results of our method as function of strength of the supervised loss.

Sup. loss coeff.	mAP$_{all}$	mAP$_{known}$	mAP$_{novel}$
0	6.59	16.58	5.77
0.1	6.90	22.26	5.64
0.5	**6.92**	**25.00**	**5.42**
1.0	6.35	25.81	4.75

Table 2: Sensitivity to number of classes. We vary the number of novel classes set during the discovery phase.

# novel classes	mAP$_{all}$	mAP$_{known}$	mAP$_{novel}$
1000	5.42	23.24	3.95
3000	**6.92**	**25.00**	**5.42**
5000	6.24	24.91	4.70

Evaluation metrics. For quantitative evaluation, we follow [24, 23, 17, 63] to match predicted cluster assignments with annotated semantic classes. We follow common practice and report mean average precision (mAP@[.5:.95]) [40]. We detail our evaluation procedure in the supplementary.

Baselines. We compare our end-to-end trainable detector with state-of-the-art Weng *et al.* [63] work, NCD methods ORCA [7], UNO[17], and k-means [43] baseline that operate on cropped object instances. For NCD methods we use labeled instance annotations from COCO$_{half}$ dataset as *labeled* images pool and cropped RPN proposals as *unlabeled* images pool. For proposals extraction we train RCNN on COCO$_{half}$ under the same conditions as our base network, to ensure comparable evaluation conditions. Having trained the NCD classifiers, we use the same RCNN to generate object proposals for the validation images, apply the classifiers on top of the proposals, and continue with the standard RCNN post-processing steps. In addition, we report results of k-means clustering directly applied to FPN-based features. We provide more details on baselines in the supplementary.

4.2 Model ablations

We ablate single and multi-phase training, the impact of keeping supervised loss in the loop, number of novel categories, pseudo-labeling prior, model architecture decisions, and backbone pretraining. In supplementary we provide further results on experiments with hyperparameters for proposals extractor and how class-agnostic heads affect the performance of the supervised baseline.

Single-phase training and unfreezing R-CNN components during discovery. As shown in Table 3 we observe that training the network with both supervised and discovery losses from the beginning, summed up without any scaling, leads to divergence. We hypothesize that this is due to noisy gradients from our self-supervised signal, with R-CNN being very sensitive to the choice of losses and hyperparameters [49]. To further support this claim, we experiment with a standard supervised phase followed by a discovery phase without freezing any of the components. We observe that such a setup also leads to divergence. Unfreezing only the RoI heads during the discovery phase, however, does not lead to divergence, yet reduces the score by 1.32 mAP. We thus keep all the components frozen during the discovery phase, apart from known and novel classification heads.

The strength of the supervised loss. We ablate the strength of the supervised classification loss that is used along with the discovery loss during the discovery phase. We vary the strength from 0 (no supervised loss) to 1 (no downscaling) and provide the results in Table 1. We observe that excluding supervised loss completely degrades the performance of the network for the known classes and benefits the performance of the novel classes. This indicates that there is a tradeoff between performance on known and novel classes, - keeping the full supervised loss introduces too much bias towards the former. We consequently set the strength of the supervised loss to 0.5 in all experiments.

Sensitivity to the number of novel classes. In Table 2 we experiment with varying the number of novel classes we use for the novel classification head. We observe that the best number of classes for the COCO + LVIS setup is 3000, while other choices result in lower scores. We conjecture that such scores are sensitive to the semantic classes defined by the target dataset - with 3000 classes the model learns the granularity and grouping that matches closest to the LVIS classes, but that does not necessarily mean that a finer or coarser granularity categorization is worse. In supplementary we also experiment with attaching and training multiple heads jointly during the discovery phase and observe that scores slightly degrade in such setup but do not observe a significant drop in results.

Prior distribution for Sinkhorn clustering. In Table 3 we experiment with using uniform prior distribution for class-marginals during pseudo-labels generation, as proposed in [9]. We observe

Table 3: Additional ablations. We experiment with removing individual components from the network and their impact on the overall performance.

Method	mAP_{all}	mAP_{known}	mAP_{novel}
RNCDL	**6.92**	**25.00**	**5.42**
Single-phase framework	n/a	n/a	n/a
Unfreeze all layers	n/a	n/a	n/a
Unfreeze RoI heads	5.60	20.55	4.37
Uniform class prior	5.75	24.11	4.23
W/o projection MLP	5.30	26.13	3.58
W/o swapped assignments	6.74	26.02	5.14
W/o memory	2.83	24.19	1.06
W/o MoCo pretraining	4.77	16.92	3.77

Table 4: Comparison with state-of-the-art models. * as per open source code. † adapted to support known classes in target dataset. ‡ randomly initialized novel head.

Method	mAP_{all}	mAP_{known}	mAP_{novel}
Methods that operate on cropped proposals			
k-means [43]	1.33	15.61	0.14
Weng *et al.* [63]*	1.62	17.85	0.27
ORCA [7]	2.03	20.57	0.49
UNO [17]†	2.18	21.09	0.61
Methods that operate on FPN-based features			
k-means [43]	1.55	17.77	0.20
RNCDL w/ random init.‡	1.95	23.51	0.17
RNCDL	**6.92**	**25.00**	**5.42**
Fully-supervised	18.47	39.38	16.74

degradation of the score by 1.17 mAP, which confirms our hypothesis that a non-uniform prior distribution closer resembles a real-world long-tail class distribution.

Embedding projector and swapped assignments. In Table 3 we demonstrate the effect of adding an MLP feature projector on top of RoI box features for novel heads classification. We observe that the performance of the network improves by 1.62 mAP. In the same table, we show that swapping pseudo-labels between proposals under multiple augmentations helps boost the score by 0.18 mAP.

Feature memory module. We demonstrate the importance of memory module in Table 3. We observe that removing the memory module significantly drops the performance by 4.09 mAP, especially for the novel classes. This highlights that for an efficient pseudo-labels generation it is crucial to use a set of features representative of the full feature space that we lack in the absence of memory.

Self-supervised pre-training matters. We compare the results of our framework under different backbone initialization employed for the supervised pre-training stage. For the randomly initialized backbone we set all layers as trainable, while during the MoCo initialization the first two convolutional blocks are kept frozen (*c.f.*, [26]). In Table 3 we observe that MoCo-initialized backbone improves the score by 2.15 mAP over randomly initialized backbone mode.

Proposals extractor hyperparameters. We observe that when extracting the RPN proposals for self-supervised learning during the discovery phase the absence of NMS yields the best results and the optimal number of proposals per image is 50. We thus use such parameters for all the experiments. We provide more details and ablation results in supplementary.

4.3 Comparison to prior work

In Table 4 we compare our method with k-means baselines and three recent state-of-the-art methods that we adapted to our scenario. We present the results for both known (COCO) and unknown (the rest of LVIS) classes. Our method significantly outperforms the previous best NCD method UNO [17], by reaching 4.74 higher overall mAP. We outperform UNO in both the known and especially novel classes, where we reach 4.81 higher mAP. We also significantly outperform approach by Weng *et al.* [63][1]. Our method is the only method to reach over 1 mAP in novel classes, where we reach 5.42 mAP. We believe the main advantages of our method to be the memory module, critical for effective self-supervision, and a backbone trained in an end-to-end manner with the aid of class-agnostic losses, further frozen during the discovery phase. Further, our backbone benefits from high-level semantic features via FPN that is critical for the classification of small objects [39]. We provide more comparison details in the supplementary. We note that as our method heavily relies on the generated region proposals, we are limited by the performance of region proposal network. It also does not update the backbone feature extractor, which could further boost the results, including the performance of localization and segmentation components.

[1]Results re-produced from the code provided by the authors. We note that these numbers are lower than those reported in their paper, however, we used official implementation and consulted the authors of this paper. We provide implementation/reproduction details in the supplementary.

a) COCO → LVIS b) LVIS → VisualGenome

Figure 3: Visualization of predictions for validation images of fully-supervised model and our RNCDL framework. We color the discovered novel classes in red.

a) COCO → LVIS b) LVIS → VisualGenome

Figure 4: Visualization of the common novel classes discovered in LVIS and VisualGenome datasets.

4.4 Cross-dataset Generalization

We analyze the generalization properties of our framework by evaluating the performance of our network trained on LVIS and tested on VG datasets.

We present quantitative results in Table 5. In Figure 4.4 we provide qualitative results with novel classes discovered. In Figure 4.4 we also visualize the 8 common discovered categories. Upon manual examination of the clusters, we observe that many of the novel classes are synonyms or hypernyms of known classes that appear in the LVIS dataset.

Table 5: LVIS → VisualGenome comparisons with a fully-supervised method.

Method	mAP_{all}	mAP_{known}	mAP_{novel}
Ours w/ random init.	2.13	7.71	0.81
Ours	4.46	12.55	2.56
Fully-supervised	4.52	13.72	2.35

5 Conclusions

In this paper, we introduce an end-to-end RNCDL network for the task of novel class discovery, detection, and localization. Our model is a two-stage object detection network that is able to classify both instances of the labeled, known classes and those of unlabeled, novel classes. At the core of our framework is a self-supervision guided by features of region proposals and a constraint for the novel class assignments to follow a long-tail distribution. In our experiments we demonstrate a significant improvement over the previous state-of-the-art. Furthermore, we demonstrate the ability to detect semantic classes without any supervision at a large scale on the Visual Genome dataset.

Broader Impact

The success of deep learning, to a large extent, can be attributed to high-quality labeled datasets. These datasets, for the most part, contain only a fixed number of categories (classes). Deep learning models trained on such datasets can only make predictions for objects that belong to one of the training set categories. However, the world consists of an extremely large number of categories, and enlarging the training set with more labeled data for all the categories is an infeasible problem. Clearly, there is a need of shit towards solutions that are not limited by labeled data supervision. Novel class category discovery has been mostly tackled in image classification setup, but there have been no works that give suitable solutions to the more challenging task of object detection. In this work, we improve compared to previous methods by an order of magnitude.While the problem is still far from solved, we hope that our method makes the first concrete step toward an effective solution, and would inspire others to tackle this challenging problem. We expect that this method will lead to more object detection works that aim to discover instances of novel classes.

References

[1] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020.

[2] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018.

[4] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015.

[5] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[7] Kaidi Cao, Maria Brbić, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[9] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV*, 2019.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020.

[11] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *ICCV*, 2017.

[12] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICML*, 2019.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Akshay Raj Dhamija, Manuel Günther, and Terrance E Boult. Reducing network agnostophobia. In *NeurIPS*, 2018.

[16] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[17] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021.

[18] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018.

[19] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[21] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[23] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020.

[24] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[27] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Deep image category discovery using a transferred similarity function. *arXiv preprint arXiv:1612.01253*, 2016.

[28] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018.

[29] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to Segment Every Thing. In *CVPR*, 2018.

[30] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, 2021.

[31] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, 2014.

[32] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *ICCV*, 2021.

[33] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021.

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[35] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *IJCV*, 128(7):1956–1981, 2020.

[36] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised object discovery and tracking in video collections. In *ICCV*, 2015.

[37] Yong Jae Lee and Kristen Grauman. Object-graphs for context-aware visual category discovery. In *CVPR*, 2010.

[38] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011.

[39] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[41] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *IJCV*, 128(2):261–318, 2020.

[42] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[43] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, 1967.

[44] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018.

[45] Aljoša Ošep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object mining for object discovery from unlabeled video. In *ICRA*, 2019.

[46] Aljoša Ošep, Paul Voigtlaender, Mark Weber, Jonathon Luiten, and Bastian Leibe. 4d generic video object proposals. In *ICRA*, 2020.

[47] Shafin Rahman, Salman Hameed Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *ACCV*, 2018.

[48] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[50] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[51] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[52] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *PAMI*, 35(7):1757–1772, 2012.

[53] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *PAMI*, 36(11):2317–2324, 2014.

[54] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

[55] Josef Sivic, Bryan C. Russell, Andrew Zisserman, William T. Freeman, and Alexei A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.

[56] Tinne Tuytelaars, Christoph H. Lampert, Matthew B. Blaschko, and Wray Buntine. Unsupervised object discovery: A comparison. *IJCV*, 88:284–302, 2010.

[57] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020.

[58] Huy V. Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *CVPR*, 2019.

[59] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*, 2020.

[60] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *NeurIPS*, 2021.

[61] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[62] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020.

[63] Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, and Serena Yeung. Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In *CVPR*, 2021.

[64] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[65] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 2018.

[66] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

[67] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021.

[68] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021.

[69] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, 2021.

[70] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]