A Simple Approach and Benchmark for 21,000-Category Object Detection

Yutong Lin[1,2] *, Chen Li[1,2*], Yue Cao[2], Zheng Zhang[2], Jianfeng Wang[2], Lijuan Wang[2], Zicheng Liu[2], and Han Hu[2]

[1]Xi'an Jiaotong University [2]Microsoft
{yutonglin,edward82}@stu.xjtu.edu.cn
{yuecao,zhez,jianfw,lijuanw,zliu,hanhu}@microsoft.com

Abstract. Current object detection systems and benchmarks typically handle a limited number of categories, up to about a thousand categories. This paper scales the number of categories for object detection systems and benchmarks up to 21,000, by leveraging existing object detection and image classification data. Unlike previous efforts that usually transfer knowledge from base detectors to image classification data, we propose to rely more on a reverse information flow from a base image classifier to object detection data. In this framework, the large-vocabulary classification capability is first learnt thoroughly using only the image classification data. In this step, the image classification problem is reformulated as a special configuration of object detection that treats the entire image as a special RoI. Then, a simple multi-task learning approach is used to join the image classification and object detection data, with the backbone and the RoI classification branch shared between two tasks. This two-stage approach, though very simple without a sophisticated process such as multi-instance learning (MIL) to generate pseudo labels for object proposals on the image classification data, performs rather strongly that it surpasses previous large-vocabulary object detection systems on a standard evaluation protocol of tailored LVIS.

Considering that the tailored LVIS evaluation only accounts for a few hundred novel object categories, we present a new evaluation benchmark that assesses the detection of all 21,841 object classes in the ImageNet-21K dataset. The baseline approach and evaluation benchmark will be publicly available at https://github.com/SwinTransformer/Simple-21K-Detection. We hope these would ease future research on large-vocabulary object detection.

Keywords: Large-vocabulary object detection; benchmark; multi-task learning

1 Introduction

Current object detection datasets typically have a limited number of categories, for example, 80 classes of COCO datasets [27], 200 classes of ImageNet-DET [5],

* Equal Contribution. The work is done when Yutong Lin and Chen Li are interns at MSRA.

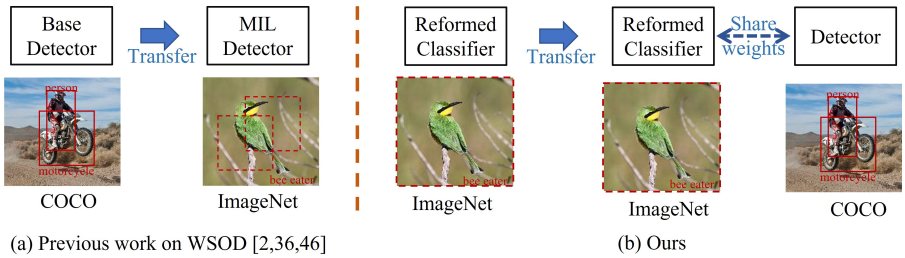

Fig. 1. (a) The knowledge flow of previous works is from a base object detector to image classification data using a weakly supervised object detection (WSOD) framework; (b) The knowledge flow of our approach is from a thoroughly-trained reformulated visual classifier, to a simple multi-task learning framework that combines reformulated classification and object detection.

365 classes of Objects365 [33], 600 classes of OpenImage [23] , 1,203 classes of LVIS [14], 1,594 classes of Visual Genome [21], and so on. Limited by object detection datasets, existing object detection systems typically detect up to a thousand categories or use up to a thousand categories for evaluation.

This paper aims to scale the number of categories in an object detection system up to more than 21,000. We note that existing image classification datasets involve much more object categories, such as the 21,841-category ImageNet-21K image classification dataset[5], which can serve as a complementary source to help achieve the goal of large-vocabulary object detection. Therefore, we are devoted to combining the object detection datasets that have limited object categories, with image classification datasets that have a large number of object categories, for large-vocabulary object detection. In previous works[6,41,47,51], such combination usually starts from a base detector that learns good foreground/background classification and localization capabilities, e.g., a good region proposal network (RPN), and then transfers these capabilities to the image classification data using a multi-instance learning (MIL) framework [2,41] (see Figure 1(a)). We argue that this knowledge flow may be sub-optimal when the number of object categories is large, as the large-vocabulary classification capability is more difficult to be seized at scale. For example, training a good classifier for an ImageNet-1K dataset typically requires long training iterations with strong augmentation, such as traversing about 300 million image samples, however, iterating about 4 million images with weak augmentation is enough to train a good detector on the COCO object detection dataset. In fact, a general foreground/background separation capability that works well to some extent is usually easier to obtain, even when trained on a detection dataset with limited categories, e.g., COCO [13].

Based on this view, we propose to make a reverse information flow (see Figure 1(b)), that we start from a good image classifier and then transfer the gained large-vocabulary classification capability into the object detection data. To enable smooth knowledge transferring, we reformulate the standard image classi-

Table 1. Three experimental setups for the evaluation of large-vocabulary object detection methods based on joining of the object detection and image classification datasets.

Setups	DET dataset			CLS dataset			Evaluation metric
	notation	#cat.	#im.	notation	#cat.	#im.	
S1	LVIS-997-base	720	11K	IN-997	997	1.23M	LVIS-997-novel mAP
S2	COCO	80	11.5K	IN-1K	1,000	1.28M	IN-1K loc. acc.
S3	Object365v2	365	1.7M	IN-21K	21,841	14M	IN-21K loc. acc. LVIS finetune mAP

fication approach, which uses a linear classification head on top of a backbone network, as a special configuration of object detection. In this new formulation, an image is represented by a special RoI corresponding to the entire image (red bounding box with dashed lines), and a heavier RoI classification head like that in an object detection framework is applied on this special RoI to realize image classification. By this reformulation, the two tasks of image classification and object detection are better aligned.

After having gained the large-vocabulary image classification capability by training on the image classification data alone, we employ a simple multi-task learning framework to join the two tasks of image classification and object detection. The image classification and object detection tasks will share the backbone, RPN network, as well as the RoI classification head. The weights of the shared networks are initialized from that of the first step, such that the initial network has owned a strong large-vocabulary capability.

As shown in Figure 1(b), the above two-stage approach does not employ an explicit foreground/background separation mechanism on the image data like that in previous weakly supervised object detection (WSOD) works. This degeneration strategy simplifies training, yet performs surprisingly well in detecting object categories that do not appear in the object detection data. We hypothesize that the sharing of RPN and RoI classification head has been able to help the framework dig out the capabilities required by a large-vocabulary object detector.

We conducted three experimental setups, named S1, S2, and S3, described in Table 1. The first setup S1 aims to evaluate the developed framework with a standard protocol of tailored LVIS. It mainly follows [13] and a concurrent work of [54], which use the 997 intersected categories of the LVIS and ImageNet-21K datasets for experiments. The categories are divided into 720 *base* ("common" and "frequent" of LVIS) and 277 *novel* sets ("rare" of LVIS). The training uses an LVIS-*base* object detection set and an IN-997 image classification set. The evaluation is conducted on LVIS-*novel* set. In the second setup S2, the number of categories for evaluation is extended to 1,000. This setup follows the previous weakly supervised object localization field [53,44,3,50,10] to evaluate the localization accuracy on ImageNet-1K. In this setup, we allow the use of additional COCO object detection datasets to facilitate localization on ImageNet-1K. The third setup S3 allows training of object detectors with the number of categories as

large as 21,000. This setup adopts the Object365v2 object detection dataset [33] and ImageNet-21K image classification dataset for training. To be able to evaluate a large number of object categories, we randomly select 5 validation images for each object category in the ImageNet-21K dataset and annotate all ground-truth bounding boxes on these images. This results in approximately 100,000 images being annotated for evaluation.

On Setup S1, the proposed approach has a much smaller performance gap between the base set and the novel set than recent efforts based on pre-trained visual language models (such as CLIP [30]) [13], or single-stage joint detection/classification [54]. On Setup S2, our approach achieves 75.1% top-1 accuracy with GT category labels known, and 68.3% top-1 accuracy with GT category labels unknown, which are absolutely 6.6% / 13.5% higher than previous best weakly supervised object localization approaches[50]. This indicates the benefits of an additional base detector. On Setup S3, we demonstrate a 21,000-category object detector. All these experiments demonstrate the effectiveness of the proposed method. In addition, as a by-product, the proposed approach shows to learn representations that have better transferability to downstream tasks such as object detection on a standard LVIS, than methods that training on two datasets alone or successively.

We hope our simple approach, along with the new 21,000-category evaluation benchmark, will facilitate future research on large-vocabulary object detection.

2 Related Work

Image Classification is a visual task that assigns category labels to images. This problem has largely driven the development of visual backbone architectures, such as convolutional neural networks [22,34,16] and vision Transformers[8,28]. Image classification datasets can be made with large-vocabulary, for example, the ImageNet-21K dataset [5] contains 21,841 categories; Google's JFT dataset contains 18,291 categories and 3 billion images. These large vocabulary image classification datasets serve as a powerful visual pre-training and semantic concept learning basis for a variety of vision problems. A common practice[34,16] for image classification is to apply a simple linear head on top of the backbone architecture to obtain the classification. In training, it typically[22,8] employs strong augmentations to enhance the networks' invariance property, and a long learning scheduler to train the network thoroughly, such as to distinguish subtle differences between different object categories when the vocabulary size is large.

Object Detection is a vision task that simultaneously localizes objects and performs categorical classification for each object. This is a basic task that provides localized objects for the following additional recognition or analysis aim. Unlike classification datasets, object detection datasets are typically much smaller in terms of the number of classes and images. COCO [27] is the most widely used dataset for evaluating detection methods, with 80 object categories and 115K training images. Object365 [33] / OpenImages [23] scale the number of categories

and images up to 365 / 600 categories and 1.7 / 1.9 million images. LVIS [14] is a re-annotation of COCO images with about 1,200 object categories that, along with Visual Genome [21], are the two largest publicly available object detection datasets on the number of categories. In general, object detection annotations are very expensive, which limit the scale in the number of images and categories.

Weakly Supervised Object Detection (WSOD) and Localization (WSOL) are two problems that learn to use image classification data for object detection and localization, respectively. There have been extensive studies on these topics [36,2,38,37,7,53,35,44,3,48,49,50,29,10]. The WSOD methods [36,2,38,37,7] usually first use unsupervised proposal methods such as EdgeBoxes [55] or Selective Search [42] to generate candidate bounding boxes, and then learns from the image label annotations by multi-instance learning (MIL). The WSOL methods [53,35,44,3,48,49,50,29,10] are mostly based on CAM [53] with the class activation maps as an indicator of the object area. Previous WSOD and WSOL methods are usually evaluated on relatively small datasets, such as COCO/VOC for WSOD and ImageNet-1K for WSOL. In addition, they try to solve difficult detection problems using only image labels without any box annotation, and therefore the accuracy of these systems is often too low for practical use.

This paper studies how image classification and object detection data can be combined to achieve large-vocabulary object detection for a more realistic scenario. There is a more relevant family of work as below.

WSOD with Base Detectors This family of work transfers knowledge in base detectors to aid in the weakly supervised object detection of images with category label annotations. The knowledge transferred from a base detector is either an objectness predictor [6,47], an object proposal [41,51], or a universal bounding box regressor [24].

While most of these efforts are done on small-scale datasets such as Pascal VOC [9] and COCO [27], there are also some works that use these techniques to enable large-vocabulary or open-vocabulary object detection as below.

Large-Vocabulary Object Detection YOLO9000 [31] combines detection and classification data to obtain a 9,000-category object detector. It jointly learns a standard detection loss on regular box annotations and a weakly supervised detection loss that assigns the classification labels to the anchor with the highest prediction score. [43] detects 11K object categories by exploiting semantic relationships between categories. A concurrent work to ours, Detic [54], learns about standard detection and weakly supervised detection similar to YOLO9000, but assigns classification labels to the largest object proposal. Our approach also attempts to transfer knowledge between tasks. However, unlike previous efforts that typically transfer knowledge from base detectors to image classification data, we emphasize the opposite knowledge flow and show that it is very beneficial to transfer the powerful large-vocabulary classifier learnt on the image classification datasets to object detection. In addition, unlike the weakly su-

pervised detection methods, we show a fairly simple multi-task learning that combines classification and detection to already achieve very good performance.

Open-Vocabulary Object Detection Another line is to perform open-vocabulary object detection. Early works expand the classifier of a base detector to be able to handle new categories by an already learnt word embedding [1]. A recent fashion is to use image text contrastive learning, such as CLIP [30], to help extend the classifier in a base detector to open-vocabulary scenario [45,13,25]. Our work is basically complementary to these works, and the text embeddings learnt in CLIP [30] also help to extend our approach to open-vocabulary scenario. We leave this as our future research.

3 Approach

3.1 Image Classification and Object Detection Practices

This paper aims to combine image classification data and object detection data towards large-vocabulary object detection. In this subsection, we review the common practice for image classification and object detection.

Image Classification Practice In image classification models [22,34,16,8,28], the resolution of input images is usually small, such as 224×224. In training, the images go through a series of strong augmentations: random cropping [22], color jittering [22], mix up [46], random erasing [52], and so on, before they are fed into the encoder. The strong augmentations show to be very crucial for image classification training [40,28], probably because a good classifier needs to possess strong deformation invariance. After extracting features with the encoder, classification task usually uses the last layer output of the encoder with average pooling, as the input feature of the classification head. A cross-entropy loss is widely used to drive the training of classification tasks.

Object Detection Practice In the object detection methods [12,32,39], the resolution of input images is usually set high to be able to detect tiny objects, e.g., 800×1333 is a common setting in a widely used baseline detector [15]. In training, it usually employs weak augmentation like random resizing. Similar to image classification, images are also fed into an encoder to extract image features. But in order to detect objects with various scales, feature maps are collected from more than one layer of the encoder, for example FPN [26]. In addition to that, an RoIAlign operator [15] is widely used to extract region features from the image feature maps to maintain the equivariance of region features, instead of the average pooling which usually sacrifice equivariance for invariance. In addition, as the object detector both needs to localize and recognize the objects, both a cross-entropy loss and a bounding box regression loss are adopted in the optimization process.

3.2 A Two-stage Approach

Unlike most of previous works which typically start with a base detector and transfer the knowledge of this base detector to image classification data using a multi-instance learning framework, we argue for making a reverse information flow that transfers knowledge from an image classifier to detection.

The underline reason is that when the number of object categories is large, the large-vocabulary classification capability is very difficult to be seized. In fact, 300 million images need to be traversed to train a good classifier on ImageNet-1K using common vision Transformers [8,28]. To train a good classifier on the large-vocabulary ImageNet-21K dataset, even 4 times longer iterations are required [28]. The training also relies on strong data augmentation to perform well. On the other hand, 4 million images with weak data augmentations have been enough to train a good detector on COCO. The region proposal network (RPN) trained on COCO, which distinguishes foreground objects with the backgrounds, has been general enough for common objects to be also effect beyond the 80 categories annotated in COCO [13].

In this sense, we thus propose a reverse knowledge flow that transfers information from a good image classifier to object detection. There are two stages of training, as shown in Figure 2. In the first stage, a large-vocabulary image classifier is thoroughly trained. There is a reformulation of previous standard image classification approach to be aligned well with the object detection framework. This reformulation facilitates a smooth transfer of the knowledge to the next stage of training. In the second stage, we join the capabilities of image classification and object detection through a simple multi-task learning framework.

In the following, we present the details of these two stages.

Stage I: Image Classification with Reformulation In the first stage, we reformulate the traditional image classification task to make it as close as possible to object detection. In standard practice, the image classification is achieved using a simple linear classification head at the top of the backbone network, while object detection relies on heavier heads for object localization and classification.

To bridge these two tasks, we have two modifications, firstly, treating the entire image area as a proposal RoI to represent the image and performing RoIAlign instead of the previous average pooling operator on this RoI; and secondly, taking the same object classification head in object detection task to replace the traditional linear head. The reformulation is illustrated in Figure 2(a). With this reformulation, we can still maintain the advantages of long/fully trained image classification and strong image augmentations, while have more shared layers and fewer gaps with the object detection task.

Stage II: A Simple Multi-Task Learning Framework to Combine Object Detection and Reformulated Image Classification In the second stage, we perform a joint training framework for image classification and object detection. The training pipeline of this stage is illustrated in Figure 2(b).

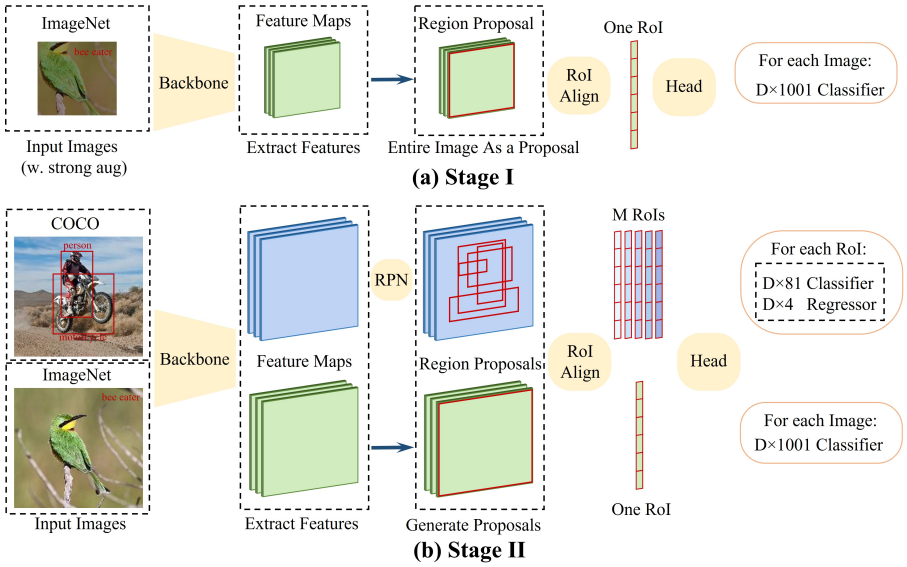

Fig. 2. Illustration on two stages of our approach: image classification with reformulation (stage I); joint object detection and reformulated image classification (stage II).

Firstly, the image processing pipeline is made unified between the object detection and the reformulated image classification tasks. Since the image classifier has been thoroughly trained at the first stage, the image processing pipeline is leaned to be friendly for the detection task: all images are resized to a large resolution such as [800, 1333] as a common settings, and weak augmentations are employed.

Secondly, most modules in architecture of the reformulated image classification are shared to the object detection framework, including the backbone, FPN [26], and the RoI classification head except for the last layer to perform linear classification. The model weights of these components learnt during the first stage are used in initialization, with other new layers randomly initialized.

Since there are no background samples on the image classification branch, we decouple foreground/background classification from object category classification and regard all samples in image classification datasets as foreground. Specifically, we introduce an extra foreground/background classifier on top of the classification head to align it to the classification branch of the object detector,

$$S_i^{\text{fg}} = \text{sigmoid}(f(\mathbf{x}_i)), \tag{1}$$

in additional to the original C-class classifiers,

$$S_i^c = \text{softmax}(g_c(\mathbf{x}_i)), \tag{2}$$

$$\hat{S}_i^c = \text{softmax}(\hat{g}_c(\mathbf{x}_i)), \tag{3}$$

where $f(\cdot) \in \mathbf{R}^{d\times 1}$, $g(\cdot) \in \mathbf{R}^{d\times C}$, and $\hat{g}(\cdot) \in \mathbf{R}^{d\times \hat{C}}$ are linear classifiers of the shared foreground/background classification, category classification on the detection dataset, and category classification on the classification dataset; i is an RoI index; c is a category index; $\mathbf{x}_i$ represents the region feature of RoI i.

The final category score for each foreground object corresponding to category c is:

$$S_i'^c = S_i^{\text{fg}} \cdot S_i^c, \hat{S}_i'^c = S_i^{\text{fg}} \cdot \hat{S}_i^c, \tag{4}$$

and the classification loss is performed on this final score.

In training, the foreground classifier is trained only with detection data, and the category classification classifiers are trained with both detection and classification data. Since image classification data is only used for training category classification, using even the very inaccurate entire image as RoI is sufficient for training the category classifier. In inference, proposals that overlap with objects are typically observed to have high scores on this object category, but only the proposal that is close to the object's ground-truth box will have higher foreground score, resulting in better overall scores.

Inference Method In inference, we remove the image classification branch, except that the final linear classifier layer is remained and put on top of the RoI classification head of the object detection branch. For categories that overlap between object detection and image classification, we empirically use the weighted geometric average of scores from two branches:

$$S_i'^c = (S_i'^c)^{\frac{2}{3}} \cdot (\hat{S}_i'^c)^{\frac{1}{3}}. \tag{5}$$

For the categories appear in only one dataset, we use the score using the corresponding classifier to produce the final category score for RoI i.

4 Experiments

4.1 Experimental Settings

We conduct experiments using three setups as shown in Table 1.

Setup S1 The 997 categories intersected between LVIS [14] and ImageNet-21K [5] are considered, resulting in two tailored datasets noted as LVIS-997 and IN-997. We divide the categories into a *base* set and a *novel* set, respectively, according to its frequency on LVIS dataset. Specifically, the categories belong to "common" and "frequent" are set as *base* classes. The categories belong to "rare" are set as *novel* classes. Using this division, the *base* set involves 720 categories, while the *novel* set involves 277 categories.

In training, the LVIS-997-base and IN-997 training images are used. In evaluation, both the LVIS-997-*base* and LVIS-997-*novel* validation set are used. In addition to the absolute accuracy, the performance gap between LVIS-997-*novel* and LVIS-997-*base* can be a good indicator to evaluate the transferring performance of a method. In this setup, Mask R-CNN [15] is used as the base detector.

Setup S2 The COCO object detection dataset and the ImageNet-1K dataset are used. The COCO dataset contains 118K training images and 80 object categories. The ImageNet-1K dataset contains 1.28M training images and 1000 object categories. The goal of this setup is to evaluate the performance of weakly object localization on ImageNet-1K, which are widely used in previous WSOL literature.

In this setup, we do not map categories of the two dataset, but treat them as independent sets. In evaluation, we only use the linear classification weights trained on the ImageNet-1K dataset. In this setup, Faster R-CNN [32] with FPN [26] is adopted as the base detector.

Setup S3 This setup is used to develop our 21,000-vocabulary detector, as well as building our evaluation benchmark for large-vocabulary object detection.

The Objects365 [33] object detection dataset and the ImageNet-21K image classification datasets are used for training. The Objects365 [33] object detection dataset contains around 1.7 million images, 365 object categories and over 29 million bounding boxes annotations in the training split. The ImageNet-21K image classification dataset [5] has over 14 million images, covering 21,841 categories.

For the ImgaeNet-21K dataset, we first divide the dataset into training split and validation split. Specifically, 5 images of each category are randomly selected for evaluation. For the classes with images less than 25, we sample 20% of the images from each class. As some of the categories are rare in ImageNet-21K (less than 5 images), we exclude them from the validation set. After filtering, we obtain a benchmarking dataset with 101,625 images and 21,077 categories. Then we annotate these images with ground-truth bounding boxes. Given an image and its ground-truth category, we annotate all objects belonging to this category. For each object, the smallest possible box that contains all visible parts of the object is annotated. This benchmarking dataset has been made publicly available to facilitate future research on large-vocabulary object detection.

In addition to evaluating large-vocabulary object detectors, we also verify whether the proposed framework can serve as a good representation pretraining method. To this end, we test its fine-tuing performance on the LVIS dataset [14]. We use the full 1,203 object categories for evaluation, which are distributed in long-tail, that there exists categories that *rare* with less than 10 training samples.

In this setup, Faster R-CNN [32] with FPN [26] is used as the base detector. In evaluation, both the metrics of localization accuracy and mAP are included.

Training and implementation details

– *First stage.* We follow common training recipes for image classification training. For ResNet architectures, we basically follow [17] to use SGD as the optimizer, with a base learning rate of 0.1, and a cosine learning rate scheduler. The weight decay is set as 1e-4, and the training length is set 100 epoch (on ImageNet-1K/997). For data augmentation, the random crop and color

jittering is employed. For vision Transformers such as Swin Transformer, we follow [40,28] and employ all the regularization and augmentation strategies in [28], including RandAugment [4], Mixup [46], CutMix [44], random erasing [52] and stochastic depth [18]. The training length is 300 epochs for ImageNet-1K/997, and 100 epochs for ImageNet-21K.

– *Second stage*. In this stage, we employ a large jittering augmentation[11] with a resolution of 1024 × 1024 resolution and a scale range of [0.1, 2.0]. The random horizontal flipping is also employed. The ratio of detection images and classification images are set as 1:3 in each iteration. The 3× training scheduler is conducted for COCO and LVIS. The learning rate is searched in 3×10^{-4}, 1×10^{-3}, 3×10^{-3} and the weight decay is 0.05. The loss weight of the classification branch is searched in 0.01, 0.1 and 1.0.

4.2 Object Detection on LVIS-997-Novel in Setup S1

Table 2 shows the results of our approach compared to previous methods in Setup S1. We note the previous approaches are built by a different implementation and training settings than ours, and thus the absolute accuracy numbers are not directly comparable, for example, the training length of previous methods are much longer than ours.

Nevertheless, the performance gaps Δ between $\mathrm{AP}^{\mathrm{novel}}_{\mathrm{mask}}$ and $\mathrm{AP}_{\mathrm{mask}}$ can act as a good indicator to evaluate the detectors' transferring ability from *base* categories to *novel* categories. The proposed approach has marginal AP degradations when transferred from *base* categories to *novel* ones, significantly better than previous methods. Compared to the variant without reformulation of classification method, or without Stage I, the performance is all degraded significantly.

These results indicate that our two-stage approach that employs an inverse information flow from a good classifier to object detection is crucial to seize the powerful classification capability.

These results also indicates surprising effectiveness of the simple multi-task learning approach, which has no explicit mechanism to attend to the detailed foreground and background classification on the image data.

Also note our approach is complementary to previous MIL approaches, by applying it afterwards the second stage. We will leave this as our future research.

4.3 Object Localization on ImageNet-1K in Setup S2

We evaluate the proposed method on the ImageNet-1K localization benchmark with the metrics of localization accuracy and GT-Known accuracy. For the localization accuracy, a predicted box is correct when it satisfies the following two conditions. First, the predicted class matches the ground truth category. Second, the predicted box has over 50% IoU with at least one of the ground truth boxes. GT-Known accuracy assumes that the ground truth category is given and only requires the predicted box to have over 50% IoU with ground truth boxes.

As summarized in Table 3, our approach outperforms other WSOL methods by a large margin. Specifically, on the basis of ResNet-50, we achieve the

Table 2. Setup S1 experiments. The comparison of different approaches on LVIS-997-base and LVIS-997-novel. *denotes that a pre-training model using IN-21K dataset, a framework of CenterNet2, and a stronger training recipe is used.

method	backbone	AP_{mask}	AP^{base}_{mask}	AP^{novel}_{mask}	Δ (novel, all)
ViLD-ens. [13]	ResNet-50	25.5	-	16.6	-8.9
Detic [54]	ResNet-50	26.8	-	17.8	-9.0
Detic* [54]	ResNet-50	32.4	-	24.6	-7.8
Ours	ResNet-50	27.6	29.1	22.8	-4.8
Ours (-reform.)	ResNet-50	26.8	28.4	21.8	-5.0
Ours (-stage I)	ResNet-50	15.5	16.3	13.0	-2.5
Ours	Swin-T	30.4	30.9	28.8	-1.6
Ours	Swin-S	35.1	36.5	30.6	-4.5

top-1 localization accuracy of 64.7% and the GT-Known accuracy of 73.8%, also surpassing previous state-of-the-art methods by a gap of 9.9% and 5.3%, respectively. These results indicate that an additional base object detection datasets can significantly benefit the weakly supervised object localization.

Also note by this experiment, we do not aim for fair comparisons, but to encourage the use of this settings for object localization, rather than the previous settings that use only image classification data.

The ablation with the variant that does not use a reformulated image classification model, or the on that does not involve this first stage of training, indicates the importance of our design.

4.4 21K-Category Object Detection in Setup S3

We benchmark the performance of 21,000-vocabulary object detection with the ImageNet-21K localization dataset. In addition to the localization accuracy, we evaluate different methods with the challenging Average Precision (AP) metric.

To test with the AP metric, we allow predicting up to 100 detection boxes per image and adopt a confidence score threshold of 1e-4 for our approach. We compare our approach with Detic [54], which is concurrent with our approach, and is the only paper to develop a detector that can handle more than 20,000 object categories. For Detic, we find that a small confidence score threshold like 1e-6 would still hurt the performance and thus set it as 0.

As shown in Table 4, the proposed method achieves a Top-1 localization accuracy of 19.2% and AP_{50} of 6.9%, revealing the difficulty of discriminating over 21,000 categories. Note that our method outperforms Detic [54] by a large margin by +17.5% on Top-1 localization accuracy and +5.6% on AP_{50}. Comparing to our approach, Detic has a much worse classification performance on the 21,000-category dataset. This indicates it is crucial to have a reverse information flow for large-vocabulary object detection, that a large-vocabulary classifier is well trained first before transferring the knowledge to object detection. We hope that these results can serve as a baseline for future studies.

Table 3. Comparison of joint training model with other state-of-the-art methods on ImageNet-1K validataion set.

Methods	Backbones	Loc. Acc		GT-Known
		Top-1	Top-5	Top-1
CAM [53]	VGG16	42.8	54.9	-
CutMix [44]	VGG16	43.5	-	-
I^2C [50]	VGG16	48.4	58.5	63.9
SPA [29]	VGG16	49.6	61.3	65.1
CAM [53]	InceptionV3	46.3	58.2	62.7
SPG [49]	InceptionV3	48.6	60.0	64.7
I^2C [50]	InceptionV3	53.1	64.1	68.5
SPA [29]	InceptionV3	52.7	64.3	68.3
CAM [53]	ResNet-50	46.2	-	-
CutMix [44]	ResNet-50	47.3	-	-
I^2C [50]	ResNet-50	54.8	64.6	68.5
TS-CAM [10]	DeiT-S	53.4	64.3	67.6
Ours	ResNet-50	64.7	69.9	73.8
Ours (-reform.)	ResNet-50	60.8	62.1	69.9
Ours (-stage I)	ResNet-50	50.4	51.3	71.0
Ours	Swin-T	66.5	72.6	73.9
Ours	Swin-S	**68.3**	**78.5**	**75.1**

Table 4. Results of the joint training model on ImageNet-21K large-vocabulary object localization benchmark. Swin-B is adopted as the backbone for both methods. For Detic[54], we test with the publicly released model Detic_C2_SwinB_896_4x_IN-21K.

Methods	AP_{50}	Loc. Acc	
		Top-1	Top-5
Detic [54]	1.3	2.7	5.2
Ours	6.9	19.2	19.7

4.5 The Proposed Approach as Pre-training

The proposed framework can also serve as a pre-trained model for down-stream tasks. To evaluate this, we fine-tune different pre-trained models on LVIS dataset for 24 epochs using the Faster R-CNN [32] framework. We use the AdamW [19] optimizer, with the learning rate of 1e-4, the weight decay of 0.05 and the batch size of 16. Repeat Factor Sampling (RFS) [14] with a factor of 0.001 is utilized to handle the class imbalance problem in LVIS.

Table 5 summarizes the results of the three methods. Successive training on classification and detection outperforms classification pre-training by a gap of 3.0% on box AP. This may be due to that the classification pre-trained models lack the ability to localize, while successive training models somewhat maintain the classification and localization ability at the same time. However, successive training methods may suffer from the problem of *catastrophic forgetting* [20]. In successive learning, as it would be fine-tuned on the object detection dataset, the model may only focus on categories of the detection dataset and *forget* other

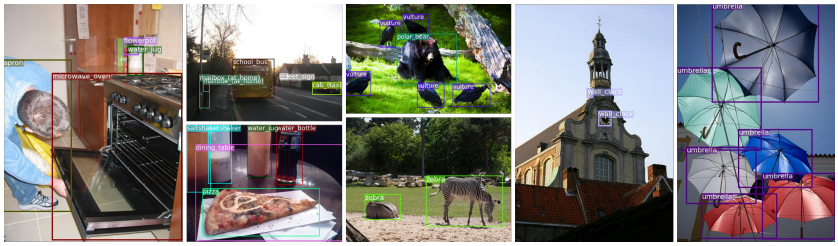

Fig. 3. Visualization of object detection results on LVIS dataset. The rich category semantics learnt by joint training model makes it easy to transfer to other datasets.

Table 5. Comparison of different pre-training methods on LVIS. IN-21K denotes the supervised classification model trained on ImageNet-21K. IN-21K → O365 denotes the successive training of ImageNet-21K classification and Objects365 object detection.

Pre-training	AP	AP_r	AP_c	AP_f
IN-21K	35.8	23.6	36.1	40.9
IN-21K → O365	38.8	26.8	38.9	44.0
Ours	**40.1**	**28.8**	**40.2**	**45.0**

concepts in the classification pre-training. This would limit the model to be transferred to other datasets with different classes.

The proposed joint training method could address this problem because it performs supervised classification and object detection in a joint manner. As shown in Table 5, when compared with the strong baseline of successive training, the proposed method still has a large performance boost of 1.3% and 2.0% on AP and AP_r, respectively. The good performance on *rare* categories verifies that the joint training model could learn richer category semantics.

5 Conclusion

In this paper, we propose a simple two-stage approach for large-vocabulary object detection. Unlike previous approaches which transfers knowledge from a base detector to image classification data, we start from image classification and transfer the large-vocabulary capability to object detection. For better transferring, the image classification problem is reformulated as a special configuration of object detection. The joining of two datasets is conducted using a simple multi-task learning framework in the second stage. Though without sophisticated process to explicitly attend to region proposals on the image data, the proposed multi-task learning approach performs rather strongly using three experimental setups.

We also obtained a 21,000-category object detector, built using a combination of the ImageNet-21K image classification dataset and the Object365v2 object detection dataset. We also built a new benchmarking dataset for the evaluation of object detection on large-vocabularies. We hope the simple baseline approach and evaluation benchmark can ease the future study in this area.

References

1. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 384–400 (2018)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2846–2854 (2016)
3. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. International journal of computer vision **100**(3), 275–293 (2012)
7. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2876–2885 (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
9. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2) (2010)
10. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization (2021)
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2918–2928 (2021)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
13. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
14. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European conference on computer vision. pp. 646–661. Springer (2016)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
23. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
24. Lee, S., Kwak, S., Cho, M.: Universal bounding box regression and its applications. In: Asian Conference on Computer Vision. pp. 373–387. Springer (2018)
25. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. arXiv preprint arXiv:2112.03857 (2021)
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CoRR **abs/2103.14030** (2021), `https://arxiv.org/abs/2103.14030`
29. Pan, X., Gao, Y., Lin, Z., Tang, F., Dong, W., Yuan, H., Huang, F., Xu, C.: Unveiling the potential of structure preserving for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11642–11651 (June 2021)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
31. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
33. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
35. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: International Conference on Computer Vision (ICCV) (2017)
36. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. In: International Conference on Machine Learning. pp. 1611–1619 (2014)
37. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.L.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence (2018)
38. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2843–2851 (2017)
39. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
40. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers distillation through attention. In: International Conference on Machine Learning. vol. 139, pp. 10347–10357 (July 2021)
41. Uijlings, J., Popov, S., Ferrari, V.: Revisiting knowledge transfer for training object class detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1101–1110 (2018)
42. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
43. Yang, H., Wu, H., Chen, H.: Detecting 11k classes: Large scale object detection without fine-grained bounding boxes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9805–9813 (2019)
44. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: International Conference on Computer Vision (ICCV) (2019)
45. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
46. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
47. Zhang, J., Huang, K., Zhang, J., et al.: Mixed supervised object detection with robust objectness transfer. IEEE transactions on pattern analysis and machine intelligence **41**(3), 639–653 (2018)
48. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.: Adversarial complementary learning for weakly supervised object localization. In: IEEE CVPR (2018)
49. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: European Conference on Computer Vision. Springer (2018)
50. Zhang, X., Wei, Y., Yang, Y.: Inter-image communication for weakly supervised localization. In: European Conference on Computer Vision. Springer (2020)
51. Zhong, Y., Wang, J., Peng, J., Zhang, L.: Boosting weakly supervised object detection with progressive knowledge transfer. In: European conference on computer vision. pp. 615–631. Springer (2020)

52. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Computer Vision and Pattern Recognition (2016)
54. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: arXiv preprint arXiv:2201.02605 (2021)
55. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014)