Object Detection with a Unified Label Space from Multiple Datasets

Xiangyun Zhao[1], Samuel Schulter[2], Gaurav Sharma[2], Yi-Hsuan Tsai[2], Manmohan Chandraker[2,3], Ying Wu[1]

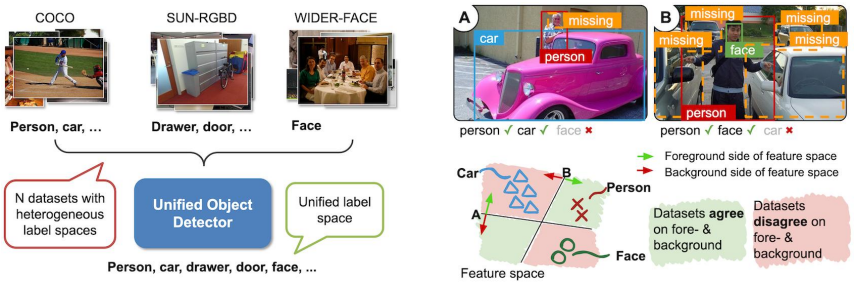
[1]Northwestern University [2]NEC Labs America [3]UC San Diego

Abstract. Given multiple datasets with different label spaces, the goal of this work is to train a single object detector predicting over the union of all the label spaces. The practical benefits of such an object detector are obvious and significant—application-relevant categories can be picked and merged form arbitrary existing datasets. However, naïve merging of datasets is not possible in this case, due to inconsistent object annotations. Consider an object category like faces that is annotated in one dataset, but is not annotated in another dataset, although the object itself appears in the latter's images. Some categories, like face here, would thus be considered foreground in one dataset, but background in another. To address this challenge, we design a framework which works with such partial annotations, and we exploit a pseudo labeling approach that we adapt for our specific case. We propose loss functions that carefully integrate partial but correct annotations with complementary but noisy pseudo labels. Evaluation in the proposed novel setting requires full annotation on the test set. We collect the required annotations[1] and define a new challenging experimental setup for this task based on existing public datasets. We show improved performances compared to competitive baselines and appropriate adaptations of existing work.

1 Introduction

Object detection has made tremendous progress in recent years to become a powerful computer vision tool [10,23,25]. This is driven by the availability of large-scale datasets with bounding box annotations. However, obtaining such data is costly and time-consuming. While multiple publicly available datasets with annotations for various categories already exist, their label spaces are mostly different, which makes a naïve combination of the data impossible [9,15,18]. But such a unification could be crucial for applications that require categories not labeled in some of the datasets, to avoid a costly annotation process. We address this problem and *propose to unify heterogeneous label spaces across multiple datasets*. We show how to train a single object detector capable of detecting over the union of all training label spaces on a given test image (see Fig. 1a).

[1] Project page: http://www.nec-labs.com/~mas/UniDet
This work was part of Xiangyun Zhao's internship at NEC Labs America.



(a) (b)

Fig. 1: **(a)** We train a single object detector from multiple datasets with heterogeneous label spaces. In contrast to prior work [31], our model unifies the label spaces of all datasets. **(b)** Illustration of the ambiguity of background in object detection when training from multiple datasets with different label spaces. Here, only "person" is consistent wrt. both datasets but "car" and "face" are missing in the other one, respectively. Naïve combination of the datasets leads to wrong training signals

Unifying different label spaces is not straightforward for object detection. A key challenge is the *ambiguity in the definition of the "background" category*. Recall that image regions in object detection datasets may contain objects without annotation because the object's category is not part of the label space. These image regions are then considered part of the background. However, their object category may be annotated in a different dataset. For instance, the label space of the COCO dataset [18] has 80 categories but it does not contain "human faces", which is present in many of the COCO images, and also annotated in other datasets like [34]. A pair of images illustrate the point in Fig. 1b. Hence, images and annotations of different datasets can not be trivially combined while preserving the annotation consistency wrt. the union of all their label spaces.

A straightforward attempt to handle the ambiguity of the label space is exhaustive manual annotation, which is very costly in terms of both time and money. Recent works have tried to address a similar problem of expanding the label space of object detectors [22,26,31], which we discuss in Sec. 2. However, none of them can truly unify multiple label spaces and are, thus, not applicable in our setting. In contrast, *we propose to handle the ambiguity directly and unify the label spaces in a single detector*. At each training iteration, we sample images from a single dataset along with the corresponding ground truth. Predicted bounding boxes of the model need to be associated with ground truth in order to compute a loss for object detectors. While positive matches with the ground truth can be assigned a category and a regression target, unmatched boxes, which normally would be assigned the background category, now become ambiguous because of the incomplete label space (see Fig. 1b and Sec. 3.2). To resolve this ambiguity, we leverage a pseudo labeling paradigm, where dataset-specific detectors are trained on individual datasets, thus not suffering the ambiguity, which are then

applied on other datasets that require additional annotation. We then propose in Sec. 3.4 a novel association procedure and loss function that carefully integrate this potentially noisy pseudo ground truth with the actual accurate annotations for the current dataset.

Our proposed task naturally suggests a novel setting for object detection (Sec. 3.5): Given N datasets with heterogeneous label spaces, train a detector capable of predicting over the union of all label spaces, which is evaluated on a test set equipped with annotations for all categories. For this purpose, we choose four existing and challenging datasets (COCO [18], SUN-RGBD [27], VOC [6] and LISA-Signs [19]), mix their respective validation/test sets and collect novel annotations for missing categories. Our results in Sec. 4 show that the proposed training algorithm successfully unifies the label spaces and outperforms competitive baselines based on prior work on this practically relevant and challenging novel task.

2 Related Work

The ultimate goal of our work is to build an object detector that expands its label space beyond what is annotated in a single available training dataset. We propose an algorithmic approach to address this problem, which we contrast to related prior art and alternative approaches in this section.

Manual annotation: The most obvious attempt to expand the label space of an object detector is by manually annotating the desired categories. New datasets are routinely proposed [9,15,16], but these come at high cost in terms of both time and money. Attempts to reduce these costs have been proposed either by interactive human-in-the-loop approaches [24,35] or by relying on cheaper forms of annotations and developing corresponding algorithms [1,20,32]. In the former attempt, one still needs to revisit every single image of the existing dataset for any new category added. The latter attempt is more promising [1,14,29,30]. However, it cannot compete with fully-supervised approaches so far. In contrast, our work shows how to combine multiple datasets with bounding box annotations for heterogeneous sets of categories.

Object detection with bounding box and image-level annotations: Several works try to leverage the combination of object detection datasets and image classification datasets [22,28,33]. Detection datasets typically annotate coarse-grained categories with bounding boxes [6,18], while classification datasets exist with fine-grained category annotation but only on the image level [5]. The main assumption in these works is that there are certain semantic and visual relationships between the categories with bounding box annotation and the ones with only image-level annotation. Mostly, the fine-grained categories are a sub-category of the coarse-grained categories. However, this assumption is a limiting factor because categories without clear visual or semantic relations are hard to combine, e.g., "person" and "car". Our proposed approach assumes bounding box-annotated datasets as input, but has the capability to combine any categories regardless of their visual or semantic relationship.

Universal representations: The recently proposed work by Wang et al. [31] on universal object detection has a similar goal as our work. Inspired by a recent trend towards universal representations [2,13,21], Wang et al. propose to train a single detector from multiple datasets annotated with bounding boxes in a multi-task setup, which we review in more detail in Sec. 3.1. However, the categories in their design are not actually unified, but kept separate. At test time, the object detector knows from which dataset each test image comes from and thus makes predictions in the corresponding label space. In contrast, we propose to truly unify the label spaces from all training datasets with a more efficient network design. Moreover, our experimental setup generalizes the one from [31] as it requests from the detector to predict all categories (unified label space) for any given test image. This is a more challenging and realistic setup since there is no need to constrain each test image to the label space of a single training dataset.

Domain adaptation: Only a few methods have been proposed on domain adaptation for object detection [3,11,12], but they are certainly related as training from multiple datasets is naturally confronted with domain gaps. While this is an interesting direction to explore, it is orthogonal to our contributions because none of these works address the problem of mismatched label spaces across datasets.

Learning from partial annotation: The most relevant work in this context is from Cour et al. [4], who propose loss functions for linear SVMs. Others have tried to use pseudo labeling to address this problem [7]. Most of these works, however, focus on plain classification problems, while we adapt some of these concepts to our specific task for object detection with deep networks in Sec. 3.

3 Training with Heterogeneous Label Spaces

3.1 Preliminaries

Notation: We work with N datasets, $\mathtt{D}_1, \mathtt{D}_2, \ldots, \mathtt{D}_N$ and corresponding label spaces $\mathtt{L}_1, \mathtt{L}_2, \ldots, \mathtt{L}_N$, each $\mathtt{L}_i$ being a set of categories specific to dataset i. In general, we do not constrain the label spaces to be equal, i.e., $\mathtt{L}_i \neq \mathtt{L}_j$ for $i \neq j$ and we allow common categories, i.e., $\mathtt{L}_i \cap \mathtt{L}_j \neq \emptyset$ for $i \neq j$, where $\emptyset$ is the empty set. For example, the label spaces of COCO, VOC and KITTI are different but all of them contain the category "person". We also include the special category "background", b_i, in every dataset $\mathtt{D}_i$ making the complete label space for $\mathtt{D}_i$: $\mathtt{L}_i \cup b_i$. However, the definition of individual "background" categories b_i is different. Thus, merging the dataset-specific background is the main challenge we address below (see also Sec. 1 and Fig. 1b).

Dataset $\mathtt{D}_i$ contains M_i images $I_{i,j}$, with $j = 1, \ldots, M_i$, and their corresponding sets of ground truth annotations $\mathtt{G}_{i,j} = \{g^k_{i,j} = (x_1, y_1, x_2, y_2, c)^k, k = 1, \ldots, |\mathtt{G}_{i,j}|\}$. Each ground truth annotation corresponds to an object present in the image and contains its bounding box coordinates, in image space, and category label $c \in \mathtt{L}_i$. Any region of an image $I_{i,j}$ which is not covered by any bounding box in $\mathtt{G}_{i,j}$ is considered to be in the background category b_i for the respective dataset.

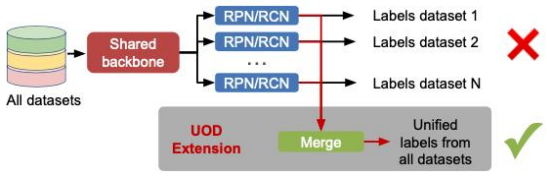


Fig. 2: Design of "universal" object detection [31] and our extension to unify the label spaces (UOD+Merge)

Object detection framework: We use Faster-RCNN [23] together with the Feature Pyramid Network (FPN) [17] as our base object detector. Faster-RCNN takes as input an image and extracts convolutional features with a "backbone" network. It then uses a region proposal network (RPN) which predicts a set of bounding boxes that describe potential (category-agnostic) objects. Further, it has a region classification network (RCN) which classifies each proposal into the defined label space and also refines the localization. See [17,23] for details.

Universal object detection baseline: Wang et al. [31] recently proposed a "universal" object detector (UOD) that is also trained from multiple datasets. This is done by equipping the model with multiple RPN/RCN heads, one for each dataset, and train the model similar to a multi-task setup where each dataset can be seen as one task. While this bypasses the ambiguity in the label spaces, at test-time, UOD needs to know what dataset each test image comes from and can thus activate the corresponding RPN/RCN head. This is not directly applicable in our more realistic setting, because we request to detect all categories from the training datasets on any test image. However, we introduce a simple extension to alleviate this problem, UOD+merge, which acts as a baseline to our proposed approach. It runs all RPN/RCN detector heads on a test image and merges semantically equal categories via an additional non-maximum suppression step (see Fig. 2).

3.2 Unifying label spaces with a single detector

The main goal of this work is to train a single detector with the union set of all label spaces $\mathtt{L}_\cup = \mathtt{L}_1 \cup \mathtt{L}_2 \ldots \cup \mathtt{L}_N$ such that given a new image I, all categories can be detected. This unified label space implicitly also defines a new and unique "background" category $b_\cup$, which is different to all other b_i. The intuitive benefit of a single detector over per-dataset detector heads, as in our UOD+merge baseline, is lower computational costs. The potentially bigger benefit, though, comes from better model parameters which are updated with data from all datasets, although having incomplete annotations.

The corresponding neural network architecture is simple: We use a single backbone and a single detector head with the number of categories equal to $|\mathtt{L}_\cup| + 1$, where $+1$ is for the background category $b_\cup$. We then follow the

standard training procedure of object detectors [23] by taking an image as input, predicting bounding boxes[2], and assigning them to ground truth annotations to define a loss. An association happens if predicted bounding boxes and ground truth sufficiently overlap in terms of intersection-over-union (IoU). Successfully associated bounding boxes get assigned the category and regression target of the corresponding ground truth box. The set of all remaining bounding boxes, which we denote as $\bar{\mathtt{D}}$, would be assigned the background category (without any regression target) in a standard detection setup where ground truth is complete. However, in the proposed setting, where the ground truth label space $\mathtt{L}_i$ of dataset i is incomplete wrt. the unified label space $\mathtt{L}_\cup$, the remaining bounding boxes $\bar{\mathtt{D}}$ cannot simply be assigned the background class $b_\cup$ because it could also belong to a category from another dataset $\mathtt{L}_\cup \setminus \mathtt{L}_i$. To illustrate this case again, consider an object of any category $c \notin \mathtt{L}_1$, which will be treated as part of the background b_1, even though it may be present in images of $\mathtt{D}_1$. However, this category may be annotated in another dataset $\mathtt{D}_2$ and, hence, not be part of background $b_\cup$. As an example, "human faces" are not part of the label space of the COCO dataset [18], but are certainly present in COCO images and other datasets exist with such annotations, like WiderFace [34].

3.3 A loss function to deal with the ambiguous label spaces

The problem we have to address for training a unified detector is the ambiguity of predicted bounding boxes $\bar{\mathtt{D}}$ that are not associated with any ground truth of the given image from dataset i. These bounding boxes either belong to a category not in the label space of the current image, i.e., $\mathtt{L}_\cup \setminus \mathtt{L}_i$, or truly are part of the unified background $b_\cup$.

This problem can be thought of as learning with partial annotation [4], where a given sample can belong to a subset of the actual label space but the actual category is unknown. The only constraint on the true label of the ambiguous detections $\bar{\mathtt{D}}$, without additional information, is that it *does not* belong to any label in $\mathtt{L}_i$, because these categories were annotated for the given image $I_{i,j}$. Thus, the underlying ground truth category belongs to any of $\mathtt{L}_* = (\mathtt{L}_\cup \setminus \mathtt{L}_i) \cup b_\cup$. As suggested in [4], we use this fact to design a classification loss function for ambiguous detections $\bar{\mathtt{D}}$ as

$$\mathcal{L}^{-}_{\text{sum}}(\mathbf{p}, \mathtt{L}_*) = -\log\left(\sum_{c \in \mathtt{L}_*} p_c\right) , \tag{1}$$

where $\mathbf{p} = [p_1, p_2, \ldots] \in \mathbb{R}^{|\mathtt{L}_\cup|+1}$ is the probability distribution for a predicted bounding box over the unified label space $\mathtt{L}_\cup$ and background, with $\sum_{c=1}^{|\mathtt{L}_\cup|+1} p_c = 1$. The loss function (1) essentially is a cross-entropy loss on the sum of the ambiguous categories and can be interpreted as merging all ambiguities into one category. An alternative is to replace the sum inside the log of (1) with a max [4]. An

[2] While predicted boxes can either be anchors (RPN) or proposals (RCN), we only apply our approach on RCN. We did not observe gains when applied on RPN.

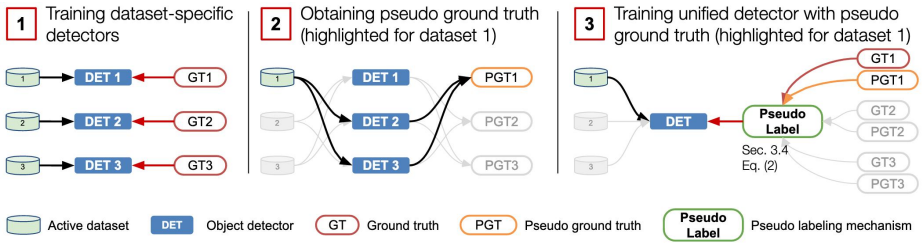


Fig. 3: Overview of our pseudo labeling approach for training a single object detector with a unified label space from multiple datasets: **(1)** We train dataset-specific detectors which do not suffer any label ambiguity. Although we draw three separate detectors, we share the weights of their backbones to encourage better performance across the different datasets. **(2)** We generate pseudo ground truth for each dataset that completes the given ground truth with bounding boxes for categories not present in the current label space. Thus, images from dataset i are put through detectors trained on other datasets $j \neq i$. The figure highlights the generation of pseudo ground truth for the first dataset. **(3)** We train a unified detector from all datasets on a unified label space $\mathtt{L}_\cup$. The figure illustrates one training iteration with an image from the first dataset. The pseudo labeling module generates the training signal for the detector given true ground truth (GT) and pseudo ground truth (PGT) from that dataset

extension to (1) would be to add a minimum-entropy regularization to encourage selectivity instead of spreading probability mass over multiple categories.

While this sounds intuitive and we also use this loss as a baseline in our experiments, it has two flaws: First, one category, the background $b_\cup$, is never among the certainly correct classes but always ambiguous, which may lead to problems during learning. Second, this loss only considers the ambiguity but *does not resolve it*, which can be done with pseudo ground truth as described next.

3.4 Resolving the label space ambiguity with pseudo labeling

Although the loss function (1) in the previous section encodes that one of the ambiguous categories in $\mathtt{L}_*$ is correct, it does not use any prior on the categories. In this section, we propose to adopt a pseudo-labeling approach to impose such a prior by estimating missing annotations. Fig. 3 gives an overview.

Dataset-specific detectors: These are detectors trained on a single dataset $\mathtt{D}_i$ and have no ambiguity in their label space. We use them to augment incomplete annotations on images from different datasets $\mathtt{D}_j$ with different label spaces $\mathtt{L}_j$ ($j \neq i$). We use N dataset-specific detectors for each of the N datasets, which can be trained in different ways. The obvious option is to independently train N detectors, where each of them has its separate feature extraction backbone. Another option is to use a UOD [31] like network where a shared backbone is followed by N separate detector heads.

Since dataset-specific detectors are trained on one dataset $\mathtt{D}_j$ but applied to another one $\mathtt{D}_i$ to augment the annotations, the domain gap between the datasets should be considered. Among the two training options above, we expect the latter to be relatively better since part of the network is shared across domains. Another option for training dataset-specific detectors could be to leverage domain adaptation techniques, e.g. [3,11,12], which we expect to boost performance, but are orthogonal to our contributions and we leave this for future work.

Finally, we note that dataset-specific detectors are only used at train time, but are not required at inference time for our proposed unified object detector.

Training with pseudo labels: During training of our unified detector, each mini-batch contains data only from a single dataset, which enables efficient gradient computation. Suppose we have an image $I_{i,j}$ of dataset $\mathtt{D}_i$ with label space $\mathtt{L}_i$, its ground truth for categories in $\mathtt{L}_i$ are available but those in $\mathtt{L}_* = (\mathtt{L}_\cup \setminus \mathtt{L}_i) \cup b_\cup$ are not. To remedy this, we run all dataset-specific detectors, hence covering $\mathtt{L}_*$, to obtain a set of bounding boxes $\hat{\mathtt{G}}_{*,j}$ of label space $\mathtt{L}_*$ for image j, in which we consider these to be the pseudo ground truth. Each pseudo ground truth box $\hat{g}^k_{*,j}$ has a detection score $\mathcal{S}_{\text{det}}(\hat{g}^k_{*,j})$ associated with it. In the next paragraph, we describe a loss function that leverages this pseudo ground truth as a strong prior to resolve the label space ambiguity.

Loss function with pseudo labels: Pseudo ground truth needs to be used carefully during training because it contains noise owing to (a) the domain gap between datasets and (b) the errors from the detectors. We propose a robust loss function and matching strategy as follows. Given a set of unmatched detections $\bar{\mathtt{D}}$ and all pseudo ground truth boxes $\hat{\mathtt{G}}_{*,j}$, we first compute the IoU similarity $s_{l,k}$ between $\bar{d}_l \in \bar{\mathtt{D}}$ and $\hat{g}^k_{*,j} \in \hat{\mathtt{G}}_{*,j}$ for all l and k. In contrast to standard object detectors, where the closest ground truth for each anchor/proposal is chosen, we keep all pseudo ground truth boxes with sufficiently high IoU similarity (i.e., $s_{l,k} > \tau$). We keep multiple matches with pseudo ground truth to counter the uncertainty of pseudo labeling and average out the potential noise. Suppose for each box $\bar{d}_l$ (not matched to $\mathtt{G}_{i,j}$), we have a set of matched pseudo ground truth boxes $\hat{\mathtt{G}}^{\text{m}}_{*,j} = \{\hat{g}^k_{*,j} | s_{l,k} > \tau \wedge \mathcal{S}_{\text{det}}(\hat{g}^k_{*,j}) > \kappa_{\text{bg}}\}$, where $\mathcal{S}_{\text{det}}(\cdot)$ is the detection score. The threshold κ_{bg} defines a minimum score for a detection to be considered as pseudo ground truth class, and anything below it is considered as background $b_\cup$. If $\hat{\mathtt{G}}^{\text{m}}_{*,j}$ is empty, we assume the ground truth label for $\bar{d}_l$ is "background" and use a standard cross-entropy loss. Otherwise, we employ the following loss on the predicted class distribution $\mathbf{p}_l$ of $\bar{d}_l$:

$$\mathcal{L}_{\text{P}}(\mathbf{p}_l) = \frac{1}{Z} \sum_k \mathit{\Gamma}(\mathcal{S}_{\text{det}}(\hat{g}^k_{*,j})) \cdot CE(\mathbf{p}_l, \text{c}(\hat{g}^k_{*,j})) \ , \tag{2}$$

where $\hat{g}^k_{*,j} \in \hat{\mathtt{G}}^{\text{m}}_{*,j}$ and $\text{c}(\hat{g}^k_{*,j})$ is the category of $\hat{g}^k_{*,j}$. The loss is the sum over the matched pseudo ground truth boxes, weighted by $\mathit{\Gamma}(\mathcal{S}_{\text{det}}(\hat{g}^k_{*,j})) : \mathbb{R} \rightarrow \mathbb{R}$, which decides the importance of the pseudo ground truth. It is normalized by $Z = \max\left(\sum_k \mathit{\Gamma}(\mathcal{S}_{\text{det}}(\hat{g}^k_{*,j})), \epsilon\right)$, where ϵ is a small constant to prevent division-by-zero in case all $\hat{g}^k_{*,j}$ have weight 0. There could be many different ways to

define such a weighting function $\Gamma(\cdot)$, and we analyze various choices in our experiments. One important aspect we find to be crucial for the success of pseudo labeling is to ignore uncertain detections and not mark them as pseudo labels. Soft weighting by the score itself, i.e. $\Gamma(x) = x$ as the identity function, is one possible instantiation of such a weighting function. Another possibility is a hard threshold, i.e. $\Gamma(x; \kappa_{\text{ignore}}) = 1$ if $x > \kappa_{\text{ignore}}, 0$ otherwise. In the second case, the boxes with scores between κ_{bg} and κ_{ignore} are ignored and do not contribute to the loss at all, analogous to common detectors like Faster-RCNN [23].

Note that (2) only defines the classification loss based on pseudo ground truth. We can also add the corresponding regression loss as in standard detectors like [8,23], which we evaluate in our experiments.

3.5 Evaluating a unified object detector

Standard object detection evaluation has typically been limited to a single dataset [10,23]. We propose a more challenging and arguably more realistic experimental setup, where we take N standard datasets with their corresponding train/val splits and *different* label spaces. A unified detector takes the training sets of these N datasets as input and, at inference time, it must be able to predict the union of all training categories on any unseen test image.

For the final evaluation, we mix the validation/test sets from each dataset together into a single larger set of images. The trained model *does not* know from which dataset the images come from, which differs from the setting in [31]. The detector *predicts all categories in all of the images*. Such a setup requires additional annotations for evaluating the test performance. Hence, we collected the annotations for the missing categories of the unified label space for all our test images. Please see the supplementary material for details.

4 Experiments

We now show results on the experimental setup proposed in Sec. 3.5, with various datasets. We evaluate the unified object detectors from Sec. 3 in this novel setting, analyze their different variants, and compare with competitive baselines.

Datasets: Recall that our goal is to train from N datasets with different label spaces, while the evaluation requires predicting over the union of all training categories on any given image. We use publicly available datasets directly for training. Tab. 1 summarizes our dataset choices. We define several specific combinations of datasets that we believe cover a diverse set of category combinations, and one additional setup, that we use for an ablation study. Details for each setting are given in the corresponding sections below. Note that we do not include the recently proposed setting from Wang et al. [31] with 11 datasets for two main reasons: First, the combinatorial problem of annotating missing categories in 11 datasets renders it costly. Second, several datasets in the setup of [31] are completely dissimilar, in a sense that object categories are disjoint from the rest, e.g., medical images vs. usual RGB images, which defeats the purpose of what we want to demonstrate in this paper.

Table 1: Different combinations of datasets that we use in our experimental setup. For each setting, a detector needs to be trained from all datasets predicting the union of their respective label spaces

Setting	Datasets
A	VOC [6] + SUN-RGBD [27] + LISA-Signs [19]
B	VOC [6] + COCO [18] (w/o VOC categories) + SUN-RGBD [27]
Ablation	VOC [6] + COCO [18] (w/o VOC categories)

Implementation details: We implement all our models with PyTorch[3]. The backbone is ResNet-50 with FPN [17], which is pretrained from ImageNet [5]. For each experiment, the network is trained for 50,000 iterations with a learning rate of 0.002. For our unified detector with pseudo labeling, we initialize the backbone with the weights from the ImageNet pretrained model. For the weighting function in (2), our default choice is the hard-thresholding described in Sec. 3.4. We follow the standard detection evaluation as in [31] and report $\text{mAP}^{0.5}$.

4.1 Ablation study

To better understand the importance of the individual components and variations in our proposed method, we first conduct an analysis in a controlled setting.

Experimental setup: We use a special setup where we intentionally remove certain categories, from one dataset, that are available in another dataset. Specifically, we use VOC [6] (2007) and COCO [18] (2017), with (annotations of) VOC categories removed from COCO. In this way, no additional annotation is required for evaluation and we can even monitor pseudo labeling performance on the training set. We evaluate on the combination of both datasets as described in Sec. 3.5 (MIX), but also on the individual ones (COCO and VOC), as well as specifically on the 20 VOC categories on the COCO dataset (V-on-C).

Baselines: We compare with the following baselines:

- **Individual detectors:** The most apparent baseline is to train individual detectors on each of the datasets and run all of them on a new test image. As with our UOD+merge baseline, we apply a merging step as described in Sec. 3.1. Besides the obvious drawback of high runtime, the domain gap between datasets can lead to issues.
- **Universal detectors (UOD)**: Our proposed extension of [31], UOD+Merge, described in Sec. 3.1 and Fig. 2.
- **Unified detector ignoring label space ambiguity**: This baseline ("Unify w/o Pseudo-Labeling" in the tables) trains a single detector head with the unified label space $\mathtt{L}_\cup$ but treats all unmatched detections $\bar{\mathtt{D}}$ as part of the unified background $b_\cup$. It thus ignores the ambiguity in the label spaces.

[3] We use the following code base as the starting point for our implementation: https://github.com/facebookresearch/maskrcnn-benchmark

Table 2: Results of our ablation study on the four different evaluation sets. Please see text for more details on the training sets and the baselines.

Method	VOC	COCO	V-on-C	MIX
(a) Individual detectors + Merge	**81.4**	46.1	*62.0*	46.9
(b) UOD [31] + Merge	81.0	45.6	61.5	46.1
(c) Unify w/o Pseudo-Labeling	79.5	42.6	36.5	43.7
(d) Partial-loss sum [4] (1)	78.9	43.4	39.2	44.1
(e) Partial-loss sum + minimum-entropy	79.1	42.9	38.4	43.9
(f) Partial-loss max [4]	79.9	43.6	36.1	44.6
(g) Pseudo-Labeling (soft-thresh)	80.5	49.6	61.6	51.6
(h) Pseudo-Labeling (w/o regression)	80.9	*50.1*	61.9	*52.0*
(i) Pseudo-Labeling	*81.3*	**50.3**	**62.2**	**52.2**

- **Partial annotation losses**: This baseline implements the unified detector head and tries to resolve the label space ambiguity with the loss function (1) designed for partial annotations [4]. As mentioned in 3.3, we evaluate alternatives to (1): (i) We add a minimum entropy regularization to (1) (sum + ent) and (ii) we replace the sum in (1) with a max function as suggested in [4].
- **Pseudo labeling**: We analyze different variants of the pseudo-labeling strategy. Instead of hard-thresholding in the weighting function $\Gamma(\cdot)$ of the loss (2), which is the default, we also evaluate soft-thresholding (soft-thresh) as described in Sec. 3.4 (paragraph "Loss function with pseudo labels"). Given the inherent uncertainty in pseudo ground truth, we additionally analyze its impact on classification and regression losses by adding a variant where we remove the regression loss for every pseudo ground truth match (w/o regression).

Results: Tab. 2 compares our final pseudo labeling model (Pseudo-Labeling) with various baselines. We make several observations:

First and most importantly, we can observe a clear improvement of pseudo labeling (i) over the most relevant baseline, UOD+Merge (b). A big improvement (46.1 to 52.2) can be observed on the MIX setting, which is arguably the most relevant one since it requests to detect the union of all training label spaces on test images from all domains. We identified a significantly lower number of false positives of pseudo labeling (i) as one of the reasons for this big improvement. The baselines (a-b) suffer more from the domain gap because all (a) or some parts (b) of the detector are trained only for a single dataset. The difference between having individual detectors (a) versus a shared backbone (b) seems negligible, but (b) obviously has a better runtime. Overall, the proposed pseudo-labeling (i) is still the most efficient one in terms of number of operations.

Second, methods that ignore the label space ambiguity (c) clearly underperform compared to other methods, particularly on the sets V-on-C and MIX, where it is important to have categories from one dataset transferred to another.

Third, the loss functions for problems with partial annnotation (d-f) improve only marginally over ignoring the label space ambiguity (c). This is expected to some degree, as we pointed out a potential flaw in Sec. 3.3, which is distinct to the task of object detection. The background category stays ambiguous for all examples while other categories are part of the certain classes at least every N samples, N being the number of datasets.

Finally, comparing different variants of pseudo labeling itself, i.e., (g-i), hard-thresholding (i) outperforms soft-thresholding (g) slightly. Also, using a regression loss on the pseudo ground truth (i) improves compared to (h).

Evaluation of pseudo labels: Our ablation study setting allows us to evaluate the performance of pseudo labeling itself, because ground truth is available. Specifically, we can evaluate the VOC detector on the COCO train set. Using all pseudo detections, we get precision and recall of 0.29 and 0.7, respectively. However, restricting the evaluation to only confident pseudo detections, we get precision and recall of 0.79 and 0.4, demonstrating the importance of thresholding (or weighing) pseudo detections based on confidence as we do in our loss (2).

4.2 Comparing pseudo labeling with an upper bound

In this section, we compare our pseudo labeling approach with an upper bound that has access to the full annotations. We split the COCO train set into two even halves, where each half sees 50% of the images and ignores 40 (different) categories. We train a standard detector on the original COCO train set and our pseudo detector and UOD on the splitted data set. The standard detector (upper bound) obtains 50.2 AP50, our pseudo detector 48.4 and UOD+merge 45.7. With domain gap absent for this experiment, we see that our pseudo labeling achieves competitive results compared to the upper bound.

4.3 Main results

In this section, we present our main results where we compare our proposed unified detectors, (e) and (i) in Tab. 2, with the two most relevant baselines, UOD [31] + Merge (b) and Partial-anno loss based on the loss from [4] (d). We first provide details on the two different settings we chose for evaluation, summarized in Tab. 1 before we show quantitative and qualitative results. Note that for each of the 4 datasets in use, we collect annotations for 500 images with the unified label space.

Setting A combines 3 different object detection datasets from different domains: general (VOC [6]), indoor scenes (SUN-RGBD [27]) and driving scenes (LISA-Signs [19]). The datasets have 20, 18 and 4 categories, respectively, with a few overlapping ones in the label spaces. The unified label space has 38 categories in total (4 overlapping). For the three datasets, we have a total of 1500 images for evaluation. Note that the images from different datasets are mixed for evaluation.

Table 3: **(a)** Results on the two main settings as defined in Tab. 1. **(b)** Performance of pseudo labeling for overlapping classes. All numbers are AP50 [18]

Method \ Setting	A	B
UOD [31] + Merge	59.3	43.7
Partial-anno loss [4]	58.3	41.2
(Ours) Min-Entropy loss	58.7	42.9
(Ours) Pseudo-Labeling	61.1	48.5

(a)

Method \ Class	chair	sofa	tv	bed	toilet	sink
Individ (VOC)	68.6	74.8	81	-	-	-
Unified (VOC)	62.3	76.3	77.6	-	-	-
Individ (COCO)	-	-	-	55.0	66.6	51.5
Unified (COCO)	-	-	-	56.7	68.9	53.7
Individ (SUN)	75.3	68.2	85.1	78.1	67.2	78.0
Unified (SUN)	75.2	64.5	85.5	78.6	66.5	81

(b)

Setting B has three main intentions compared to setting A: First, we increase the label space with the addition of COCO [18]. Second, we increase the number of categories that need to be transferred between different datasets by removing the VOC categories from COCO. Third, we use a more focused set of scenes by using the combination of general scenes (VOC and COCO) and indoor scenes (SUN-RGBD). The datasets have 20, 60 and 18 categories, respectively. The unified label space contains 91 categories. Again, for the three datasets, we have a total of 1500 images, which are mixed together for evaluation.

Quantitative results: Tab. 3a provides the results on the two settings described above. We compare our proposed pseudo labeling method with three competitive baselines: UOD [31] + Merge, partial-annotation-loss [4] and minimum-entropy-loss. In both settings (A and B from Tab. 1), we can observe the same trend. Pseudo labeling significantly outperforms all baselines, e.g. 61.1 vs. 59.3, 58.3 and 58.7, respectively, on setting A, which again verifies the effectiveness of leveraging dataset-specific detectors as prior for training a unified detector. On setting B, we also observe an improvement of +4.8 AP50 points with pseudo labeling over the most relevant baseline UOD+Merge. The improvements are mostly attributed to the reduced number of false positives compared with UOD+Merge. Similar to the ablation study in Sec. 4.1, we see that only using loss functions designed for dealing with partial annotations is not enough to resolve the label space ambiguity for the task of object detection.

Performance of pseudo labeling for overlapping classes: Here, we specifically analyze the performance of pseudo labeling for overlapping classes among different datasets. For setting B, we have: VOC and SUN-RGBD share "chair", "sofa" and "tv-monitor"; COCO and SUN-RGBD share "bed", "toilet" and "sink". To analyze the difference between individual detectors and our unified detector (w/ pseudo labeling), we test on each dataset separately. The results in Tab. 3b show that some categories improve, likely due to increased training data via pseudo labeling, while others get worse, eventually due to domain gap. However, we highlight that our proposed unified detector obtains much better results when

Fig. 4: Some qualitative results of our proposed unified object detector

tested on the mix of all data sets (with a complete label space) compared to individual detectors that may generate false positives on images from other data sets (MIX column in Tab 2).

Qualitative results: Finally, we show a few qualitative results for the settings A and B in Fig. 4. Most notably, we can observe in all examples, the unified detector is able to successfully predict categories that are originally not in the training dataset of the respective test image. For instance, the second image in the top row in Fig. 4 is from the LISA-Signs dataset, which does not annotate the category "person". This category is successfully transferred from other datasets.

5 Conclusions

We introduce a novel setting for object detection, where multiple datasets with heterogeneous label spaces are given for training and the task is to train an object detector that can predict over the union of all categories present in the training label spaces. This setting is challenging and more realistic than standard object detection settings [23], as well as the recent "universal" setting [31], since there can be both (i) unannotated objects of interest in training images, as well as, (ii) the original dataset of the test image is unknown. We propose training algorithms that leverage (i) loss functions for partially annotated ground truth and (ii) pseudo labeling techniques, to build a single detector with a unified label space. We also collect new annotations on test image to enable evaluation on the different methods in this novel setting.

Acknowledgements This work was supported in part by National Science Foundation grant IIS-1619078, IIS-1815561.

References

1. Bilen, H., , Vedaldi, A.: Weakly Supervised Deep Detection Networks. In: CVPR (2016) 3
2. Bilen, H., Vedaldi, A.: Universal representations: The missing link between faces, text, planktons, and cat breeds. arXiv:1701.07275 (2017) 4
3. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain Adaptive Faster R-CNN for Object Detection in the Wild. In: CVPR (2018) 4, 8
4. Cour, T., Sapp, B., Taskar, B.: Learning from Partial Labels. JMLR **12** (2011) 4, 6, 11, 12, 13
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009) 3, 10
6. Everingham, M., Gool, L.V., Williams, C.K.I., Winn., J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV **88**(2), 303–338 (Jun 2010) 3, 10, 12
7. Feng, L., An, B.: Partial Label Learning with Self-Guided Retraining. In: AAAI (2019) 4
8. Girshick, R.: Fast R-CNN. In: ICCV (2015) 9
9. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019) 1, 3
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017) 1, 9
11. Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive Domain Adaptation for Object Detection. In: WACV (2020) 4, 8
12. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In: CVPR (2018) 4, 8
13. Kalluri, T., Varma, G., Chandraker, M., Jawahar, C.: Universal Semi-Supervised Semantic Segmentation. In: ICCV (2019) 4
14. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: ContextLocNet: Context-aware Deep Network Models for Weakly Supervised Localization. In: ECCV (2016) 3
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982 (2018) 1, 3
16. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A Composite Dataset for Multi-domain Semantic Segmentation. In: CVPR (2020) 3
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: CVPR (2017) 5, 10
18. 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: ECCV (2014) 1, 2, 3, 6, 10, 13
19. Møgelmose, A., Trivedi, M.M., Moeslund, T.B.: Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey. IEEE Transactions on Intelligent Transportation Systems **13**(4), 1484–1497 (Dec 2012) 3, 10, 12
20. Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V.: Training Object Class Detectors With Click Supervision. In: CVPR (2017) 3
21. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NIPS (2017) 4
22. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: CVPR (2017) 2, 3
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: NIPS (2015) 1, 5, 6, 9, 14

24. Russakovsky, O., Li, L.J., Fei-Fei, L.: Best of Both Worlds: Human-Machine Collaboration for Object Annotation. In: CVPR (2015) 3
25. Singh, B., Davis, L.S.: An Analysis of Scale Invariance in Object Detection - SNIP. In: CVPR (2018) 1
26. Singh, B., Li, H., Sharma, A., Davis, L.S.: R-FCN-3000 at 30fps: Decoupling Detection and Classification. In: CVPR (2018) 2
27. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In: CVPR (2015) 3, 10, 12
28. Uijlings, J., Popov, S., Ferrari, V.: Revisiting knowledge transfer for training object class detectors. In: CVPR (2018) 3
29. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection. In: CVPR (2019) 3
30. Wan, F., Wei, P., Jiao, J., Han, Z., Ye, Q.: Min-Entropy Latent Model for Weakly Supervised Object Detection. In: CVPR (2018) 3
31. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards Universal Object Detection by Domain Attention. In: CVPR (2019) 2, 4, 5, 7, 9, 10, 11, 12, 13, 14
32. Wang, Z., Acuna, D., Ling, H., Kar, A., Fidler, S.: Object Instance Annotation With Deep Extreme Level Set Evolution. In: CVPR (2019) 3
33. Yang, H., Wu, H., Chen, H.: Detecting 11K Classes: Large Scale Object Detection without Fine-Grained Bounding Boxes. In: ICCV (2019) 3
34. Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: A Face Detection Benchmark. In: CVPR (2016) 2, 6
35. Yao, A., Gall, J., Leistner, C., Gool, L.V.: Interactive Object Detection. In: CVPR (2012) 3