Semi-Supervised Object Labeling on Video Data with Collaborative Classification and Active Learning

Bruno Padilha[a] and João Eduardo Ferreira[b]
Institute of Mathematics and Statistics (IME-USP), University of São Paulo, São Paulo, Brazil

Keywords: Active Learning, Out-of-Distribution Classification, Collaborative Image Classification, Big Data Labeling.

Abstract: Streaming applications in video monitoring networks generate datasets that are continuously expanding in terms of data amount and sources. Thus, given the sheer amount of data in these scenarios, one big and fundamental challenge is how to reliably automate data annotation for downstream tasks such as object detection, image classification, object tracking among other functionalities. In this work, we propose a novel active learning strategy based on multi-model collaboration able to self-annotate training data, providing only a small initial subset of human verified labels, towards incremental model improvement and distribution shifts adaptation. To validate our approach, we collected approximately 50,000 hours of video data sourced from 193 security cameras from University of São Paulo Monitoring System (USP-EMS) during the years 2021-2023, totaling 7.3TB of raw data. For experimental purposes, this work is focused on identification of pedestrians, cyclists and motorcyclists resulting in 3.5M unique objects labeled with accuracy between 92% to 96% for all cameras. Time-stamped data along with our incremental learning method also facilitate management of naturally occurring distribution shifts (e.g., weather conditions, time of the year, dirty lenses, out-of-focus cameras). We are currently working to release this dataset in compliance with local data privacy legislation.

1 INTRODUCTION

State-of-the-art deep learning models for image classification rely on large volumes of annotated data (e.g. (Yu et al., 2022), (Srivastava and Sharma, 2024) and (Kirillov et al., 2023)). Once upon a time, obtaining data for machine learning was costly and difficult to come by due to technology restrictions in availability of sensors (i.e. cameras, social networks, signal detectors, etc...), data storage and processing power. Nowadays, data is cheaper, easier to come by and being produced at an accelerating pace. On the other hand, annotating data for supervised learning remain expensive once human generated labels are still pervasive in many successful training strategies. In spite of recent advancements in object tracking algorithms, pre-trained models and other tools that can assist humans to speed up data annotation (Ashktorab et al., 2021)(Li et al., 2021a), the cost is still high and can increase faster than linear with the dataset size (Kokilepersaud et al., 2023).

Some large annotated datasets (e.g. ImageNet (Ridnik et al., 2021), COCO (Lin et al., 2014), OpenImages (Kuznetsova et al., 2020), SA-B1 (Kirillov et al., 2023)) share similar classes also found in many other domains. One low-cost way to leverage pretrained models is known as Transfer Learning (TL) (Zhuang et al., 2021) and consists in fine-tuning a pretrained model with a much smaller dataset of a target domain. It is feasible providing the two domains (original and targeted) share good amounts of general object attributes and data distributions are not too far apart. However, it is harder to find such large datasets for niche domains that would allow us to apply TL in more domain specific downstream tasks (e.g. medical images (Hesamian et al., 2019), manufacturing quality control (Zuo et al., 2022), agricultural applications (Li et al., 2021b)). Furthermore, we will show that there is no guaranties that TL will generalize well even when the target domain contain classes supposedly known to by the pre-trained model, culminating in introduced noise as false positives and or negatives. In object detection, target objects are annotated with bounding boxes (i.e the minimal area rectangle encompassing the object). Misplaced or missing boxes in training data directly affects the model performance (Murrugarra-Llerena et al., 2022).

[a] https://orcid.org/0000-0002-6668-1529
[b] https://orcid.org/0000-0001-9607-2014

Padilha, B. and Ferreira, J. E.
Semi-Supervised Object Labeling on Video Data with Collaborative Classification and Active Learning.
DOI: 10.5220/0013707800004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 2: KEOD and KMIS*, pages 247-256
ISBN: 978-989-758-769-6; ISSN: 2184-3228

Another approach to alleviate the burden of annotating a new dataset is known as Active Learning (AL) (Ren et al., 2021b). In AL, the main concern is how to attain the best possible performance from a model with a minimal amount of annotated data. In other words, data can labeled in small amounts to train a model in an iterative manner and the previous acquired knowledge is leveraged to devise a query strategy to label new samples. One such strategy is based on the model uncertainty (He et al., 2019b) of each class to identify samples contains most novelty. If the model is too certain for a given sample, there is no novel information regarding the underlying distribution. On the other hand, too much uncertainty could mean the sample is too far away from the known distribution so far or it is just noise. In our experiments we have observed the optimal uncertainty region to effectively use AL, that is in the vicinity of the decision boundary, is somewhere in between depending on how much knowledge the model accumulated. To make the best of both TL and AL as annotation tools, we employ a pre-trained model (Yolov8 (Jocher et al., 2023) on COCO(Lin et al., 2014)) to initiate our dataset with a collection of detected objects. Due to varying degree of generalization from camera to camera, noise samples will be produced and AL comes into play to separate good samples from bad ones.

In this work, we present a novel method based on the combination Transfer Learning with Active Learning to reliably annotate large amount of objects in video data in a semi-automated manner. It is composed by teams of binary classifiers whose decisions are made **collaboratively** by voting and consensus policies. Initially, a small subset (e.g 1000 samples per class) of a raw dataset is randomly sampled for human verification and possible correction of wrong labels and discarding of noisy samples (e.g. partial objects or pieces of background) to train one weak classifier per class, what we named *experts*. Data is partitioned in a one vs. rest (OvR) fashion to minimize the odds of teams yielding arbitrary high confidence outputs for far out-of-distribution samples (Nguyen et al., 2015) (Hein et al., 2019). This strategy allows these teams of experts to self-annotate data relying on partial acquired knowledge in order to incrementally expand the training set (i.e. select new batches of 1000 samples) and eventually, after a few iterations, **converging towards the true underlying distribution**. Other than this incremental learning strategy, the proposed method also concerns continual monitoring to guide experts updating and cope with distribution shifts.

We demonstrate the effectiveness of our method with classification experiments on real data for the classes *cyclist*, *motorcyclist* and *pedestrian*. The classification data contains $\sim$3.5M unique object samples extracted from 50.000h of raw video footage from 193 cameras from USP-EMS (Electronic Monitoring System at University of São Paulo)(Ferreira et al., 2018) collected between the years of 2021-2023. In all experiments, the teams of experts were able to consistently learn the cameras distributions, reaching up to 96% in accuracy, while using only a few thousand of automatically labeled samples per class. Intending to better understand the challenges of learning with AL and the teams of experts, we opted to approach it as a classification problem leaving added complexities of detection and segmentation problems for future works.

The rest of the paper is organized as follows. Section 2 provides an overview of related works on Active Learning and automated data labeling. In Section 3, we present important details of our incremental learning strategy. Experiments are described in Section 4. In Section 5 we highlight the limitations of our method and present some concluding remarks.

2 RELATED WORK

2.1 Expansible Dataset

The new dataset we are going to build is continuously sourced by video footage from security cameras monitoring open public areas in the dependencies of the University of São Paulo (USP). The most common classes of objects found in this environment are vehicles, including cars, buses and trucks, persons, bicycles, motorcycles, animals (e.g dogs, wild birds) and an assortment of static objects such as trees, traffic signs, benches, etc. Due to these classes being present in both ImageNet and COCO datasets, we proceeded to evaluate the last iteration of YOLO (Jocher et al., 2023) object detector pre-trained on both dataset on videos from USP cameras in order to crop out samples for classification. Not surprisingly, detection effectiveness depends on the camera and also on the class as shown in Table 1. In the scope of this work, we consider only the classes motorcyclist, cyclist and pedestrian. These classes have many visual attributes in common and will further challenge our semi-automated labeling method. Furthermore, detection of motorcyclist and cyclist is done by object composition as presented in (Nardi et al., 2022).

2.2 Active Learning

The main concept behind Active Learning (AL) is to start training a model with only a small fraction of reliable (i.e. human verified) labeled data. Then, this partial learned knowledge can be employed to discover which unlabeled samples will contribute the most to improve the model this time. The process repeats until the model meets some performance criteria. Another key concept is how to apply a model to query the unlabeled set. According to (Ren et al., 2021b), AL strategies can be categorized into membership query synthesis, stream sampling and pool sampling. In the context of deep learning, the first one is usually related to sample generation, for example with GANs or VAEs models, and request it to be human labeled. The second one is suitable for storage and computing limited devices in which is there is no access to at least a sizable portion of the unlabeled set. In the last one, pool sampling, model knowledge is used to rank unlabeled data based either on sample diversity (Agarwal et al., 2020) or uncertainty (Hwang et al., 2024) (Kokilepersaud et al., 2023).

Also based on uncertainty sampling, (Hwang et al., 2024) and (Kokilepersaud et al., 2023) are closer to our approach. Just like ours, the former proposes a new multi-camera city scenes dataset. They show with empirical experiments that, contrary to previous beliefs, the cost of human labeling increases faster than linear with dataset size. Regarding sampling and model training, they leverage optical flow motion relationships between consecutive frames to label entire sequences of unique objects. The model is a YOLOv5 object detector and unlabeled sequences are ranked by entropy based on the softmax probabilities of this model. On the other hand, we opted to shift from object detection to image classification on cropped out objects to use simpler models (resnet-18) working collaboratively to account for potential overconfidence in softmax and out-of-distribution robustness. The latter, (Kokilepersaud et al., 2023), is specific to mitigate overconfidence in AL scenarios. Their approach is based on a augmented MixUP (Zhang et al., 2017) training strategy and a overconfidence sensitive ranking function. We employ label smoothing (Müller et al., 2019) to tackle overconfidence combined with uncertainty voting policies towards a similar purpose. Nonetheless, our approach favors interpretability.

2.3 Out-of-Distribution Detection

Out of distribution (OOD) detection is the ability of a model to recognize data samples that deviate from data distribution of learned representations. It has gained more attention recently as an increasing number of applications are being developed to work with data generated in real-time. Contrary to static datasets, the knowledge obtained from live data is much more subject to distribution shifts, which essentially demands classification models to operate on data containing OOD samples. According to (Winkens et al., 2020), OOD detection problems are more challenging when OOD samples are near the in-distribution ones (near-OOD) than when they appear farther away (far-OOD). Expansible datasets contemplates both scenarios once new *unlabeled samples* can contain novel knowledge (near-OOD), valuable to improve classification, or simply detrimental noise (far-OOD).

One simple approach the far-OOD case is a method know as the Mahalanobis Distance (MD)(Lee et al., 2018). It is a function that computes the distance of a point to a known distribution. In Deep learning, the features map of a deep layer of a classification model trained on N classes is used, along with the in-distribution training set, to fit N class conditional Gaussian distributions. The means vector and a covariate matrix is then used to compute the confidence and uncertainty scores for test inputs. Several authors ((Winkens et al., 2020), (Ren et al., 2021a), (Denouden et al., 2018), (Podolskiy et al., 2021)) have been trying to improve the MD method in the near-OOD case. However, these proposals rely on large and consolidate datasets with reliable annotations (e.g. CIFAR-10, CIFAR-100, ImageNet-21k), which initially is not available when building a new dataset. Furthermore, according to (Maciejewski et al., 2022), estimating multivariate normal densities for limited feature maps (i.e a single layer) and for an insufficient number of samples (1000 to 5000 per class) makes Mahalanobis distance-based methods ineffective for near-OOD data.

Pre-trained Vision Transformers (ViT) (Han et al., 2022) have been shown to be more robust to distribution shifts, less prone to learn spurious correlations (short-cut learning) and to present better results when fine-tinning on smaller datasets when compared to Convolutional Neural Networks (CNN). (Fort et al., 2021) coupled a ViT pre-trained on ImageNet-21k with MD to show it can improve near-OOD detection, at least in benchmark datasets as CIFAR-10 and CIFAR-100. As presented in (Zhang and Ranganath, 2023), robustness to spurious correlations plays an important role in OOD detection for real word datasets once, for example, objects captured in a shared scene will inevitably contain many confound features in common (e.g. same background). Further-

more, the approach proposed by (Fort et al., 2021) is still subject to some of the limitations of MD methods as discussed in (Maciejewski et al., 2022).

Contrastive Learning (CL) (Schroff et al., 2015)(Chen et al., 2020) has also been explored in OOD detection tasks. In (Tack et al., 2020), for example, a self-supervised CL based RestNET-18 is trained with rotated samples to serve as negative examples in the contrastive loss function. As stated by the authors, and demonstrated in (Chen et al., 2020), the most appropriate shift transformations for CL are dataset dependent, that is, not invariant to data distribution and may not work well for expansible datasets. More recent approaches (Sun et al., 2022) (Mou et al., 2022) combine CL with nearest neighbors algorithms attempting to provide alternative OOD distance functions. Arguing that CL may not be sufficient to learn proper in distribution representations for OOD, (Li et al., 2023) propose a ViT based model trained on Masked Image Modeling (MIM). For the OOD metric, they conclude the Mahalanobis distance gives the best results. Experiments did not account for data containing varying amounts of noisy samples or incorrect labels, which are common in expansible datasets.

2.4 Overconfidence and Uncertainty Estimation

Most modern deep learning models are based on the softmax to compute probabilities and the cross-entropy as the loss function. However, these probabilities can be overestimated and do not represent true likelihood (Guo et al., 2017). Moreover, training data with one-hot labels may lead to cross-entropy overfitting to labels before it actually overfit to data (Zhang et al., 2017). Both issue lead to overconfident models and hinders uncertainty estimation. As mitigating measures, (Kristiadi et al., 2020) proposes an adversarial training technique to enforce low confidence for far out-of-distribution data. Mixup (Zhang et al., 2017) combines samples and labels from different classes to provide soft targets. Label smoothing (Müller et al., 2019), which we have employed in our method, has been proven (Zhang et al., 2021) to be a effective regularization technique to soften one-hot labels and mitigate overconfidence.

Despite the great advances in related works presented in items 2.2, 2.3, 2.4 of this section, automated or semi-automated annotation in expansible datasets remains one of the big challenges to be overcome for data labeling in video monitoring networks. More concretely, the new datasets are going to be built in continuously sourced by video footage from security cameras monitoring open public areas such as in the dependencies of the University of São Paulo (USP). The most common classes of objects found in this environment are vehicles, including cars, buses and trucks, persons, bicycles, motorcycles, animals (e.g dogs, wild birds) and an assortment of static objects such as trees, traffic signs, benches, etc. Due to most of these classes being present in both ImageNet and COCO datasets, we proceeded to evaluate the last iteration of YOLO (Jocher et al., 2023) object detector pre-trained on both dataset on videos from USP cameras in order to crop out samples for classification. Not surprisingly, detection effectiveness depends on the camera and also on the class as shown in Table 1. In the scope of this work, we consider only the classes motorcyclist, cyclist and pedestrian. These classes have many visual attributes in common and will further challenge our semi-automated labeling method. Furthermore, detection of motorcyclists and cyclists is done by object composition as presented in (Nardi et al., 2022).

3 PROPOSED METHOD

Creating a new annotated dataset is a laborious assignment that, in spite of many tools to generate automated annotations (Adnan et al., 2021), ultimately requires human expertise to at least assert the quality of annotations (CAI Li and Yang-Yong, 2020).Noise is unavoidable when collecting data from real world sources. The magnitude and type of noise present in our generated datasets will depend on how well the object detection tool generalizes to each camera. In this work we are concerned with two types of noise: noisy labels (Jiang et al., 2020) and noisy images. Noisy labels arrive from confusion between classes during detection (e.g cyclist vs motorcyclist), while noisy images are false positives containing mostly pieces of background or heavily occluded objects.Table 1 illustrates the varying amount of noise in the raw objects dataset depending on the data source. For each chosen camera, a random sample of a 1000 images per class is selected for human validation. Both noisy images and noisy labels are removed. For example, we consider the sample for camera S5-24 to be low noise for all classes. On the other hand, while the sample for camera S5-15 has moderate noise for the classes motorcyclist and cyclist, it is more severe for the person class. Our proposed method is able to discard both types of noise as far OOD samples and, as long as good cropped images are being produced, learning is feasible once we can always obtain more data for low efficiency cameras.

Table 1: Number of correct classified samples (human validated) of a random selection of a thousand images per class (motorcyclist, cyclist and pedestrian) from the raw dataset produced by YOLOv8 on video footage from several USP cameras.

camera	motorcyclist (1000)	cyclist (1000)	person (1000)	total (3000)
S5-12	807	843	662	2312 (77.0%)
S5-13	682	825	847	2354 (78.4%)
S5-15	752	755	405	1912 (63.7%)
S5-24	962	958	912	2832 (94.4%)
S10-08	346	599	491	1436 (47.8%)

Small datasets sampled from real world sources, lets say about 1000 samples per class, even in the absence of label or sample noise as defined earlier usually are insufficient to properly learn a domain distribution. However, this limited dataset do contain some knowledge to train weak classifiers. Furthermore, in case we can segment this dataset per data source (i.e per camera), we can train a team of n independent binary weak classifiers (one for each class) that can collaboratively reach a voting-based consensus to identify near-OOD samples, which we empirically show that are the ones containing novel information regarding the distribution of this dataset segment. A batch of these near-OOD samples, automatically labeled by the team, can be selected based on the team joint confidence to increment the training set and expand the model knowledge. Because we start with weak classifiers, we increase the training set in small increments to avoid absorbing too much noise. After only a few rounds of increment-and-train, we have a *team of experts* for that distribution. Our solution should not be confused with ensemble learning (Zhang and Ma, 2012). Contrary to the later approach, member of our committee evolve from weak classifiers at training phase to full fledge independent models.

3.1 Model Architecture

Figure 1 depicts our proposed architecture for a team of classifiers. All members of this team, also called *branches*, are ResNet-18 (He et al., 2016) paired with CBAM (Woo et al., 2018) attention layers. Branches are binary classifiers specialized in a single class, that is, they individually decide if a given sample belongs (true) or do not belong (false) to that class. Input data for training is split in a One-vs-Rest (OvR) fashion where, for each branch, the true class contain only samples of a specific class (e.g. motorcyclist) and the false class is composed of a combination of samples from all other classes (e.g cyclist + pedestrians). In our experiments, this training data arrangement has been demonstrated to be a reliable way to approximate the pseudo-class *I don't know* when decisions are made collaboratively by the members of a classification team. Moreover, in order to mitigate the over-confidence problem (Hein et al., 2019) that may occur when training ResNets with piece-wise linear activation functions (e.g. Relu and variants), models are trained with label smoothing (Szegedy et al., 2015). In summary, we have the following hyper-parameters:

- Implemented in PyTorch framework
- Three classes (Motorcyclist, Cyclist and Pedestrian)
- Max epochs: 30
- Batch sizes: Train = 64, Test = 16
- Early stopping to avoid overfitting
- 5-fold validation
- Kaiming weights initialization(He et al., 2015)(He et al., 2019a)
- Weighted Cross-Entropy with label smoothing as the loss function
- Gradients calculated with SGD (learning rate = 0.1, momentum = 0.9, weight_decay = $5 * 10^{-4}$)
- Cosine Annealing learning rate scheduler(Loshchilov and Hutter, 2016)

3.2 Team Consensus

When in evaluation mode, individual decisions are combined in the voting module. One simple yet effective voting strategy is the *consensus*, meaning all branches must agree on one class. For example, Figure 2 illustrates the results of the evaluation of a picture containing a motorcyclist for which the "motorcyclist" branch voted true while the other two branches, "cyclist" and "pedestrian", voted false, thus reaching an agreement for classifying this image as motorcyclist. On the other hand, Figure 3 illustrates a case of no consensus for which both the "cyclist" and the "pedestrian" branches voted true. In this case, the team as a whole could not decide and the verdict is "I don't know". The confidences of each branch will be used as thresholds to decide what images should be considered for expanding the training set for the next iteration.

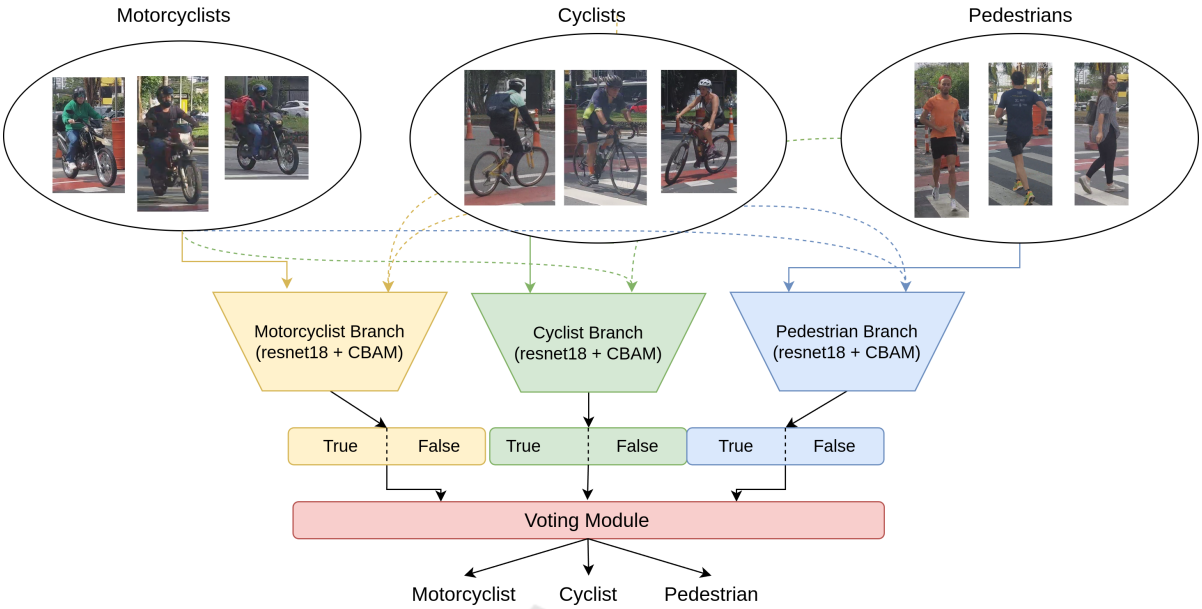
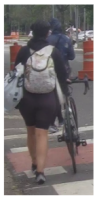


Figure 1: Classification team architecture.

	Branch		
	Motorcyclist	Cyclist	Pedestrian
Vote	True	False	False
Confidence	96.7%	93.1%	95.1%

Prediction: Motorcyclist

Figure 2: Evaluation of a picture containing a motorcyclist for which the ”motorcyclist” branch voted true while the other two branches, ”cyclist” and ”pedestrian”, voted false, thus reaching an agreement.



	Branch		
	Motorcyclist	Cyclist	Pedestrian
Vote	False	True	True
Confidence	92.3%	73.8%	85.5%

Prediction: *"I don't know"*

Figure 3: Team was unable to reach a consensus and the final verdict is ”I don’t know”.

4 EXPERIMENTS

The main source of data for the following experiments is USP-EMS (Electronic Monitoring System)(Ferreira et al., 2018). It contains hundreds of security cameras to monitor USP dependencies in eight campuses in the state of São Paulo, Brazil. The footage used in this work was collected during the years of 2021 and 2022, spanning varying seasons, weather and times of the day. Each source video is one hour long. These cameras were hand picked in close collaboration with campus security department to reflect regions and times of biggest traffic movement, cyclist concentration, street crossing and some intersections prone to intercurrences.

In order to avoid manual annotating data from scratch, we leverage transfer learning by feeding raw video data to a pre-trained object detector and tracker for objects cropping. For this purpose, the most suitable tool we found was a combination of YOLOv8 (Jocher et al., 2023) pre-trained on COCO (Lin et al., 2014) dataset, the strongSort object tracking algorithm (Du et al., 2023) and our custom cyclist and motorcyclist detection algorithm (Nardi et al., 2022). This combination, named YOLO+, produced enough data containing our three chosen target classes (pedestrian, cyclists and motorcyclists) with varying levels efficiency of noise depending on the camera. The tracking algorithm is necessary to group together unique objects as much as possible to optimize data usage at training time and to alleviate the oversampling effect (Mohammed et al., 2020). Samples are organized hierarchically by camera, video file of origin, class and unique object. The final dataset for image classification is a 3-sampled view (i.e. three samples per unique object) of the original data.

4.1 Incremental Learning

The incremental learning process in our proposed architecture starts with a small human validated portion of the targeted expansible dataset assuming la-

Table 2: Difference in terms of recall and average confidence (true class only) of a classification team trained on the initial human validated dataset (1000 images per class) vs. trained on a random sample of the same size from the raw output as produced by YOLO+, and the team of experts after six rounds for camera S10-08

S10-08	recall			average confidence		
dataset origin	motorcyclist	cyclist	pedestrian	motorcyclist	cyclist	pedestrian
raw	65.3%	48%	66%	58.4%	61%	44.3%
human	91.5%	88%	88%	89.3%	88.1%	85.7%
experts	96.3%	95.8%	93.6%	95.1%	94.7%	93.2%

bels, if present, cannot be relied upon and an variable amount of noise is present. We fixed the size of this initial dataset in one thousand images per class once this quantity was sufficient to bootstrap the incremental learning in all evaluated cameras. Furthermore, it takes on average only 12 minutes of human supervision per class to select that amount from a slight larger random batch (about 2000 samples) from raw cropped data, which itself is pre-classified by YOLO+. As presented in Table 1, the raw data produced for camera S10-08 contains a large amount of noise, to the point of precluding learning. After cleaning up this noise, a massive improvement was observed as displayed in Table 2, further improved after the incremental learning. Similar improvements were observed for all fourteen cameras selected as data sources. Due to space limitation, these results were not included in this paper.

Once the initial dataset is selected, we proceed to the first round of incremental learning. Table 3 presents the evolution of branches training after six rounds of increment-and-train. The dataset is split into 80% for training and 20% for validation. Evaluation metric is the recall for the individual *voting classes* True and False. The numbers i1 ... i5 represent the new set of thousand images per class as selected by the team through the consensus mechanism. For the selection strategy, we evaluate on the remaining samples in the raw dataset for that specific camera and sort the output of the consensus by confidence, considering only the top 25% values (the upper quartile) as candidates. Among the candidates, selection of the next thousand samples, is random once it demonstrates to be beneficial to mitigate inductive biases introduced by human selecting samples in step i0 (Kaltenpoth and Vreeken, 2023), which if is not addressed, may propagate to subsequent iterations and potentially degenerate some models. When visually inspecting the self-selected samples less and less noise is observed, reflecting the improvement in recall numbers and providing compelling evidence that, in fact, our learning strategy converges towards the true underlying distribution of these expansible datasets.

4.2 Limitations and Assumptions

This work was initially developed to address a real-world demand to produce reliable annotations for non-stop growing datasets of objects extracted from security cameras in USP-EMS. In this scenario, we are able to define individual cameras as local domains, thus producing local datasets. The concept of near/far OOD in this limited view of the world, although based on real-word data, is more self-behaved than applying the same concept without imposing any restriction on global data (from all cameras sources). In order to overcome this limitation, we are currently working on an improved version of our approach based on concepts and techniques proposed in some of the works presented in Section 2.

5 CONCLUSION

One main challenge when building novel datasets is how to reliably annotate data for supervised learning. Expansible datasets continuously sourced by real-world live sensors, in our case security cameras, takes this challenge even further due to the increased significance of distribution shifts and the inexorable presence of noise. When training classification models to evaluate new incoming data, these changes in data distribution manifest itself in the form of near-OOD samples, while noise is mostly concentrated in the far-OOD ones. Based on this premise, our active learning based strategy along with the consensus algorithm enables us to build highly accurate collaborative models, starting with weak classifiers that require only small subsets of human verified data to bootstrap training. The teams of experts have been demonstrated to be a simple yet effective approach to auto-annotate the sheer amount of data found in expansible datasets.

Table 3: Recall results of Incremental learning for six rounds on camera S10-08 (Praça R) with a team containing three members (motorcyclist, cyclist and pedestrian). Only the data for the initial iteration (i0) is human certified. After every training round, the team jointly evaluates, through the consensus mechanism, the available raw data for that camera to select near OOD samples to increment the training set (1000 images increments)

Camera S10-08 (Praça R): Branches Recall

	Motorcyclist		Cyclist		Pedestrian	
	true	false	true	false	true	false
i0 (human)	91.5%	92.2%	88%	92%	88%	94.2%
i0+i1	93.5%	94%	93.7%	95.1%	92.2%	94.3%
i0+i1+i2	93.6%	90.2%	94.8%	94.9%	93.1%	92%
i0+i1+i2+i3	97.6%	93.4%	96.6%	95%	91.1%	94.1%
i0+i1+i2+i3+i4	95.6%	94%	96%	96.7%	93%	96.9%
i0+i1+i2+i3+i4+i5	96.3%	95.1%	95.8%	96.9%	94.6%	96.4%

ACKNOWLEDGEMENTS

This work was supported by The São Paulo Research Foundation, FAPESP (grant number. 2020/06950-4) and The National Council for Scientific and Technological Development (CNPq) - CNPq Research Productivity Scholarship Program.

REFERENCES

Adnan, M. M., Rahim, M. S. M., Rehman, A., Mehmood, Z., Saba, T., and Naqvi, R. A. (2021). Automatic image annotation based on deep learning models: A systematic review and future challenges. *IEEE Access*, 9:50253–50264.

Agarwal, S., Arora, H., Anand, S., and Arora, C. (2020). Contextual diversity for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 137–153. Springer.

Ashktorab, Z., Desmond, M., Andres, J., Muller, M., Joshi, N. N., Brachman, M., Sharma, A., Brimijoin, K., Pan, Q., Wolf, C. T., et al. (2021). Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27.

CAI Li, WANG Shu-Ting, L. J.-H. and Yang-Yong, Z. (2020). Survey of data annotation. *Journal of Software*, 31(2):302.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., and Vernekar, S. (2018). Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*.

Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., and Meng, H. (2023). Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*.

Ferreira, J. E., Antônio Visintin, J., Okamoto, J., Cesar Bernardes, M., Paterlini, A., Roque, A. C., and Ramalho Miguel, M. (2018). Integrating the university of são paulo security mobile app to the electronic monitoring system. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1377–1386. IEEE.

Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110.

He, K., Girshick, R., and Dollár, P. (2019a). Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, T., Jin, X., Ding, G., Yi, L., and Yan, C. (2019b). Towards better uncertainty sampling: Active learning

with multiple views for deep convolutional neural network. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 1360–1365. IEEE.

Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019). Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50.

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596.

Hwang, Y., Jo, W., Hong, J., and Choi, Y. (2024). Overcoming overconfidence for active learning. *IEEE Access*.

Jiang, L., Huang, D., Liu, M., and Yang, W. (2020). Beyond synthetic noise: Deep learning on controlled noisy labels. In *International conference on machine learning*, pages 4804–4815. PMLR.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8.

Kaltenpoth, D. and Vreeken, J. (2023). Identifying selection bias from observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8177–8185.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kokilepersaud, K., Logan, Y.-Y., Benkert, R., Zhou, C., Prabhushankar, M., AlRegib, G., Corona, E., Singh, K., and Parchami, M. (2023). Focal: A cost-aware video dataset for active learning. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1269–1278. IEEE.

Kristiadi, A., Hein, M., and Hennig, P. (2020). Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Li, J., Chen, P., He, Z., Yu, S., Liu, S., and Jia, J. (2023). Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11578–11589.

Li, K., Li, G., Wang, Y., Huang, Y., Liu, Z., and Wu, Z. (2021a). Crowdrl: An end-to-end reinforcement learning framework for data labelling. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 289–300. IEEE.

Li, L., Zhang, S., and Wang, B. (2021b). Plant disease detection and classification by deep learning—a review. *IEEE Access*, 9:56683–56698.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Loshchilov, I. and Hutter, F. (2016). SGDR: stochastic gradient descent with warm restarts. *CoRR*, abs/1608.03983.

Maciejewski, H., Walkowiak, T., and Szyc, K. (2022). Out-of-distribution detection in high-dimensional data using mahalanobis distance-critical analysis. In *International Conference on Computational Science*, pages 262–275. Springer.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

Mou, Y., He, K., Wang, P., Wu, Y., Wang, J., Wu, W., and Xu, W. (2022). Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for ood intent discovery. *arXiv preprint arXiv:2210.08909*.

Müller, R., Kornblith, S., and Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, 32.

Murrugarra-Llerena, J., Kirsten, L. N., and Jung, C. R. (2022). Can we trust bounding box annotations for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4813–4822.

Nardi, E., Padilha, B., Kamaura, L., and Ferreira, J. (2022). Openimages cyclists: Expandindo a generalização na detecção de ciclistas em câmeras de segurança. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 229–240, Porto Alegre, RS, Brasil. SBC.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Podolskiy, A., Lipin, D., Bout, A., Artemova, E., and Piontkovskaya, I. (2021). Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. (2021a). A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021b). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Srivastava, S. and Sharma, G. (2024). Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1236–1248.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Tack, J., Mo, S., Jeong, J., and Shin, J. (2020). Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al. (2020). Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning*, volume 144. Springer.

Zhang, C.-B., Jiang, P.-T., Hou, Q., Wei, Y., Han, Q., Li, Z., and Cheng, M.-M. (2021). Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L. H. and Ranganath, R. (2023). Robustness to spurious correlations improves semantic out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15305–15312.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Zuo, C., Qian, J., Feng, S., Yin, W., Li, Y., Fan, P., Han, J., Qian, K., and Chen, Q. (2022). Deep learning in optical metrology: a review. *Light: Science & Applications*, 11(1):39.