Stream-Based Active Distillation for Scalable Model Deployment

Dani MANJAH[1] Davide CACCIARELLI[2] Mohamed BENKEDADRA[3]
Baptiste STANDAERT[1] Gauthier ROTSART DE HERTAING[1] Benoît MACQ[1]
Stéphane GALLAND[4] and Christophe DE VLEESCHOUWER[1]
[1]Université catholique de Louvain [2] Technical University of Denmark
[3] Université de Mons [4] Université de technologie de Belfort Montbéliard
{dani.manjah, baptiste.standaert, gauthier.rotsart}@uclouvain.be
dcac@dtu.dk mohamed.benkedadra@umons.ac.be

Abstract

This paper proposes a scalable technique for developing lightweight yet powerful models for object detection in videos using self-training with knowledge distillation. This approach involves training a compact student model using pseudo-labels generated by a computationally complex but generic teacher model, which can help to reduce the need for massive amounts of data and computational power. However, model-based annotations in large-scale applications may propagate errors or biases. To address these issues, our paper introduces Stream-Based Active Distillation (SBAD) to endow pre-trained students with effective and efficient fine-tuning methods that are robust to teacher imperfections. The proposed pipeline: (i) adapts a pre-trained student model to a specific use case, based on a set of frames whose pseudo-labels are predicted by the teacher, and (ii) selects on-the-fly, along a streamed video, the images that should be considered to fine-tune the student model. Various selection strategies are compared, demonstrating: 1) the effectiveness of implementing distillation with pseudo-labels, and 2) the importance of selecting images for which the pre-trained student detects with a high confidence.

1. Introduction

Deep Neural Networks (DNNs) are effective for object detection in images, but their predictive power comes at a high cost. The training of highly performant DNNs is based on high-performance cloud servers with a large-scale data set. This requires (i) a large workforce to prepare the data set or implementation of training (ii) as well as a significant investment in time and money. These data, time, and hardware costs create a barrier for most practitioners in terms of transition from theory to practice [5]. Furthermore, a single investment in resources to create large general-purpose models, regardless of their size, is no longer sufficient. Without retraining, these models cannot be robust with respect to the **stochastic** and **ever-evolving** environments. In the example of Closed-Circuit Television (CCTV) monitoring traffic on the city scale, there is no data set large enough to cover all aspects of every urban landscape [35]. Therefore, a *scalable, efficient, and recurrent* retraining is necessary to reduce costs and avoid **under-performing** systems.

Knowledge Distillation (KD) is a promising technique that enables the creation of lightweight but powerful models. The process assumes that for the same data set, large models (that is, *teachers*) have higher knowledge capacity than smaller models (that is, *students*). The teacher, typically a pre-trained or very large generic model (e.g., YOLOv8x6[1]), can transfer its knowledge (i.e., pattern recognition mechanisms) to students without significant model degradation. However, recourse to other models for labeling could lead to confirmation bias, a phenomenon that refers to noise accumulation when the model is trained using incorrect predictions for semi-supervised or unsupervised learning [2]. Furthermore, an immediate rebound effect of the scheme is

[1]There is no official paper available for this deep learning model. For the latest information, please visit the official repository: https://github.com/ultralytics/ultralytics.

the multiplication, on scale, of the number of models to be trained. The inference costs could become significant. Additionally, if the teacher model runs on a cloud-based platform, there may be additional costs associated with its usage, such as hourly usage fees or data transfer costs. This could be mitigated by using Active Learning (AL), which aims to identify the most informative examples for labeling. The importance of sampling has been first formulated in [4] as the problem of developing KD methods that are query-efficient and robust to labeling inaccuracies due to teacher imperfection (i.e., *confirmation bias*). The method developed in [4] was designed for a pool-based setting, which represents an offline scenario where a pool of unlabeled data points is made available to the learner. We claim that, in many real-world applications, a large number of unlabeled samples arrive in a streaming manner, **making it impossible to maintain all of the data in a candidate pool**. To the best of our knowledge, there is no framework supporting the development of AL methods that are query-efficient and robust to labeling inaccuracies in stream-based settings. **The contributions of this paper are the following:**

1. Formulate Stream-Based Active Distillation (SBAD) as the problem of developing AL methods that are both query-efficient and robust to labeling inaccuracies in stream-based settings.
2. Demonstrate the benefits of the proposed scheme for large-scale video-based object detections on a public dataset [26].
3. Establish simple but effective baselines to train a YOLOv8n student from a YOLOv8x6 teacher.
4. A code to reproduce the experiences and the framework available at https://github.com/manjahdani/SBAD/.

2. Related Work

2.1. Knowledge Distillation

KD is a method that involves training a smaller model to imitate the performance of a larger model. The main objectives of this technique are to prevent a decrease in the model's performance when it operates on a data set that is distributed differently than the source domain, referred to as Unsupervised Domain Adaptation (UDA), and to produce lightweight models suitable for the storage and computational capacities of miniaturized devices, referred to as Model Compression (MC) applications. In this study, we use a technique called *Self-training with knowledge distillation*, which was introduced by [6]. This technique trains a student model using pseudo-labels generated by a teacher model, which is beneficial when the labeled data is limited but we have access to a large sample of unlabeled data. Furthermore, the aforementioned distillation scheme does not need a direct access to the teacher. Yet, it may also propagate errors or biases.

In addition, we will discuss two additional techniques of interest in the following paragraphs: online distillation and context-aware distillation.

Online Distillation. This approach involves training a smaller student model to mimic the output of a larger teacher model on a per-example basis. In [13], the authors designed an online knowledge distillation scheme to perform real-time human segmentation in sports videos. Experiments show the ability of the model to adapt to contextual variations. Online distillation is also employed in [24] to adapt a low-cost semantic segmentation model to a target video where the data distribution is not necessarily stationary.

Context-aware Distillation. The works in [19, 28] attempt to exploit the contextual characteristics of the scene to develop effective KD. They directly worked on the distillation scheme to develop more specialized students. For example, [19] added a temporal dimension such that the student learns the variations in the intermediate features of the teacher over time, taking into account the redundancies of the frames within a CCTV stream.

2.2. Active Learning

AL is a sampling approach that selects the most informative data points to minimize the number of labels required for model training [33]. AL can be divided into three macro scenarios: synthesis of membership queries, pooled AL, and streamed AL [7]. The majority of approaches in deep AL have focused on the pool-based scenario, where the learner selects the most useful data points from a closed set of unlabeled observations. The stream-based AL scenario for object detectors has not been investigated. Moreover, AL assumes the availability of a perfect oracle, where the true label of a data point is revealed when queried. However, this assumption does not hold in a KD framework, where the pseudo-labels provided by the teacher may be incorrect.

Active Learning for Image Classification. AL strategies for pool-based classification can be categorized into uncertainty-based or diversity-based approaches [36]. Uncertainty-based strategies estimate model uncertainty using techniques such as Monte Carlo dropout [18] or ensemble networks [23], while entropy and margin-based sampling strategies are also widely employed [29]. Task-agnostic methods, such as Learn loss [38], use a loss prediction module to estimate data points that are likely to be wrongly predicted. Among diversity-based strategies, Core-set [32] is one of the most popular, using a K-center Greedy algorithm to locate a set of representative data points. Cluster-Margin [14] combines uncertainty and diversity, while DRMRS [16] takes into account the margin and diversity. BADGE [3] balances uncertainty and diversity using a $k-$MEANS++ seeding algorithm on gradients obtained from the last layer of the network. CDAL [1] replaces the Euclidean distance with the pairwise contextual diversity in the greedy K-center algorithm used in the Core-set. CLUE [25] performs uncertainty-weighted clustering to identify target instances that are uncertain according to the model and diverse in feature space. VAAL [34] uses a Variational Autoencoder (VAE) to map instances into a latent space, which is then fed into a discriminator that learns to differentiate between labeled data and unlabeled samples.

Active Learning for Object Detection. AL approaches to object detection can be classified into black-box and white-box methods [30]. Black-box methods do not depend on the underlying network architecture and use informativeness scores, such as the confidence obtained from the softmax layer, while white-box methods are dependent on the architecture of the underlying network. The Minmax approach, which selects the least confident images among the unlabeled pool, is a popular black-box method [30]. Ensemble methods have also been used for object detection-oriented AL [17, 31]. Query strategies based on localization tightness and stability [21], mixture density networks [12], and a unified box regression and classification metric [39] have also been proposed. MIAL [40] is a multiinstance framework that filters out noisy instances to bridge the gap between instance-level and image-level uncertainty. PPAL [37] is a two-stage algorithm that includes difficulty-calibrated uncertainty sampling and category-conditioned matching similarity. [20] proposed to cluster the unlabeled observations into groups based on the frequency domain values and to use different sampling rates for each group.

2.3. Challenges of Stream-based Active Distillation

The importance of sampling has been first formulated in [4] as the problem of developing KD methods that are both query-efficient and robust to labeling inaccuracies due to the imperfection of the teacher (i.e., *confirmation bias*). Their methods provide a theoretical guarantee that the scheme leads to queries where the teacher provides the correct labels. However, this approach has been developed in a pool-based setting where the student has access to the entire information pool. In contrast, in stream-based scenarios, techniques such as diversity-based strategies, clustering, or pairwise distance matrices may not be feasible, especially in contexts where the spatio-temporal correlation among the data is significant. Another aspect is that, due to the complexity of the student model, uncertainty techniques relying on Monte Carlo dropout or Learn loss modules may not be viable options.

3. Problem Statement

Let $\theta_{student}^{general}$ define a compact general pre-trained model learning the distribution $\mathcal{D}$ of a data stream $\mathcal{X}$. We assume a spatio-temporal correlation among the data. The student is equipped with SELECT (I_t), a rule that determines whether an image I_t should be selected to fine-tune the student model, using the pseudo-label predicted by a universal but imperfect model $\theta_{teacher}^{general}$. The objective is to train a high-performing student by querying the minimum number of teacher pseudo-labels. In this work, the pseudo-labels consist of bounding boxes generated by $\theta_{teacher}^{general}$ for each selected image. We assume a large-scale setting (e.g., city-scale deployment of CCTV, monitoring of large construction sites) and affordable hardware. Therefore, the selected frames and their associated pseudo-labels, which constitute the training set $\mathcal{L}$, must not exceed a maximum training frame budget per student B, i.e., $|\mathcal{L}| \leq B$. Furthermore, efficient SELECT strategies are necessary to ensure the scalability of our stream-based active distillation (SBAD). Indeed, if a selection rule takes longer than the frame rate to make a decision, a temporary buffer will be required to store recently seen images until the decision is made. This would increase the system resource requirements for data storage and processing, which is not scalable.

Algorithm 1 SBAD Framework

Require: a pre-trained student model $\theta^{general}_{student}$, a general purpose teacher model $\theta^{general}_{teacher}$, a training frame budget B and a SELECT strategy.
Ensure: $B \geq 1$
  $\mathcal{L} \leftarrow \emptyset$ ▷ Selected frames and their pseudo-labels
  $t \leftarrow 0$ ▷ Timestamp
  while $|\mathcal{L}| \leq B$ **do**
    Observe current frame I_t
    if SELECT(I_t) **is** TRUE **then**
      $\{b_i^{pl}\}_t \leftarrow \theta_{teacher}(I_t)$ ▷ Pseudo-labels
      $\mathcal{L} \leftarrow \mathcal{L} \cup (I_t, \{b_i^{pl}\}_t)$
    end if
    $t \leftarrow t + 1$
  end while
  return update($\theta^{general}_{student}$, $\mathcal{L}$)

Figure 1 provides a visual illustration of the SBAD framework. During the sampling phase, the SELECT rule is used to identify the most informative samples. The selected frames are then pseudo-labeled by the teacher model and used to fine-tune the student models. Once the fine-tuning is complete, specialized models could be optionally evaluate using a test-set with ground truth $\mathcal{T} := \{I^{test}, \mathbf{b^{gt}}\}$. Note that this step is not necessary for SBAD, but in real-life scenarios, it could be seen as a sanity check if you have access to a test-set.

4. Methodology

In the context of stream-based active learning, single-pass evaluation of data points is often addressed by applying a threshold to certain informativeness scores [8–11, 15, 27]. However, this approach has not been tested in online active distillation tasks for object detection. In this paper, we investigate the effectiveness of thresholding algorithms based on the confidence of the base student model $\theta^{general}_{student}$ for the SBAD framework. At round t, when the student model $\theta^{general}_{student}$ observes an image I_t, $n \geq 0$ objects are detected, which are defined by the bounding boxes b_{it} and confidence scores c_{it}. According to [30], a unique confidence score $C(I_t)$ can be obtained for I_t using:

$$C(I_t) \coloneqq \max_i c_{it}$$

This means that the confidence of each image is approximated by the highest confidence score among the objects detected in that image. Using this confidence metric, we can then apply a threshold Δ to the confidence scores of the incoming frames. The general structure of the top confidence threshold sampling scheme is presented in Algorithm 1. To estimate the threshold Δ for selecting the most informative frames, we introduce a warm-up phase where the student model $\theta^{general}_{student}$ observes the incoming frames for a period of length w without querying any image and without storing anything other than a single scalar representing the confidence scores $C(I_t)$ at the image level, where $t = 1, ..., w$. At the end of the warm-up phase, the student model estimates an $(1-\alpha)$-upper percentile on the distribution of confidence scores, where α represents the desired sampling rate. In other words, the threshold Δ is chosen so that:

$$\mathbb{P}(C(I_t) \geq \Delta) = \alpha,$$

and the frames to pseudo-label and fine-tune $\theta^{general}_{student}$ correspond to a ratio of α frames out of the total number of frames.

While in traditional AL, the focus is on querying images that the student model is least confident about, this approach may not be optimal for stream-based object-detection KD scenarios. The least confident images often correspond to very hard examples that may not be informative enough for the student model in the early rounds of AL when it has not been fine-tuned for the specific scene. Additionally, selecting images with high uncertainty for pseudo-labeling may lead to confirmation bias as the pseudo-labels may not align with the ground truth due to the imperfection of the teacher model $\theta^{general}_{teacher}$ as an oracle. This is why, in our work, we propose to let the student model $\theta^{general}_{student}$ query the most confident frames. Ideally, by doing so, the student will sample informative examples that the teacher model can accurately pseudo-label. These examples will contribute best to the student's fine-tuning while avoiding frames that are too uncertain to be used in the initial stages AL.

5. Experiments

5.1. Experimental Settings

Dataset. We evaluated the effectiveness of the SBAD approach using the Watch and Learn Time-lapse (WALT) data set [26], which comprises 12[2]

[2]We tested two out of twelve cameras and produced extra annotations to evaluate our techniques. Detailed information about this process and the dataset are available in our GitHub repository.

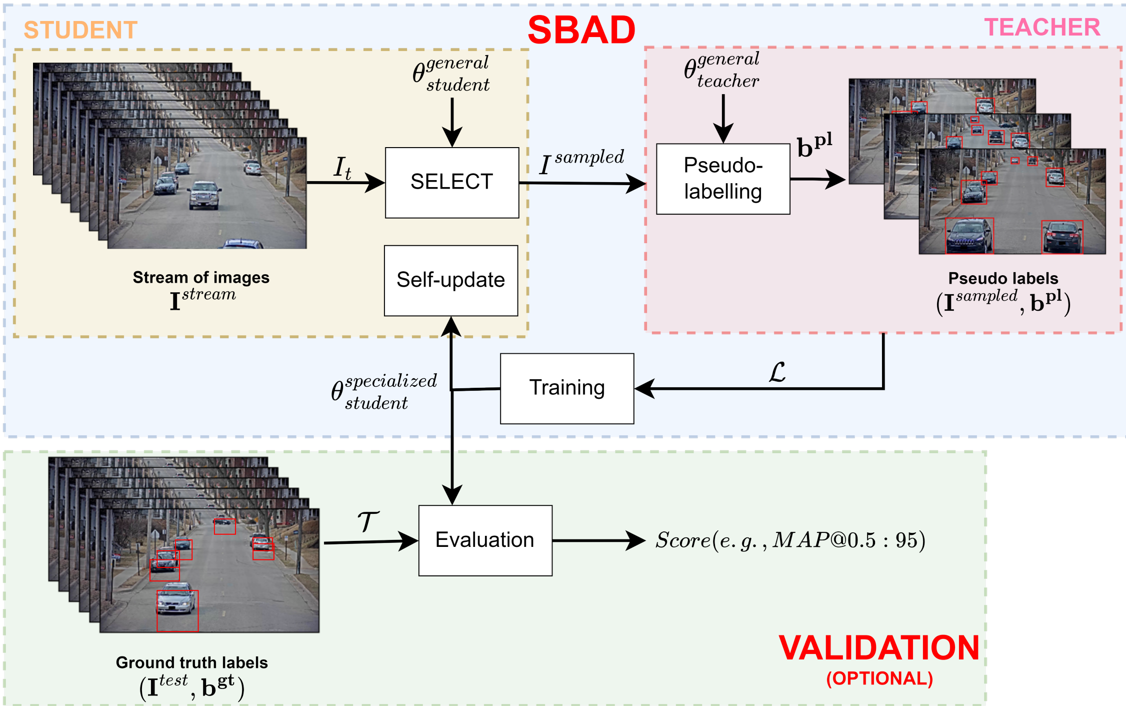

Figure 1. SBAD pipeline: sampling, fine-tuning and evaluation.

cameras that capture an urban environment. This data set offers a diverse range of spatial and temporal settings, with varying viewpoints and lighting conditions, including both day and night settings. By testing our approach on this realistic data set, we assess its performance in real-world scenarios.

Distillation implementation. In line with the principles of data distillation proposed by [6], we employ a large and complex teacher model, YOLOv8x6 (261.1 GFLOPs), to generate pseudo-labels. These labels are then used to train several smaller student models, YOLOv8n (8.7 GFLOPs), with less architectural complexity. Both networks are initially pretrained on the COCO dataset [22]. The student models are re-trained for 100 epochs with a batch size of 16 and a learning rate (LR) of 0.01. The learning rate is adjusted for each epoch with a change factor (LF) of 0.01 using Equation 1. The budget of the SBAD framework is determined by the number of pseudo-labels used for fine-tuning, which ranges from 25 to 250 in our experiments.

$$LR = \left(\frac{1 - LR}{epochs}\right) \times (1 - LF) + LF \quad (1)$$

Methods. Due to the lack of prior research on the SBAD problem in object detection, there are no baselines to compare with. To explore the effectiveness of the confidence-based thresholding algorithm, we used different baselines. First, a naive N-*First* approach has been implemented, where the student models are fine-tuned by simply taking the first N images observed from each camera. A second baseline is given by a *random* sampling approach, where a number $s \sim U(0,1)$ is generated for each incoming frame, which is queried only if $s \geq 1 - \alpha$. A third baseline is given by a more active learning-oriented *least confidence* approach, where similarly to the case of the highest confidence, we impose a threshold on the confidence score at the image level. The main difference is that the threshold Δ is estimated by taking the α-lower percentile from the warm-up set $\mathcal{W}$.

In our experiments, both α-lower and α-higher methods used $\alpha = 10\%$. However, it is important to note that this choice was influenced by the frame rate and the length of the data stream recorded for each week. Although smaller values of α may yield better performance, they would need to span a longer data stream as we become more selective in terms of selecting only the most confident frames. There-

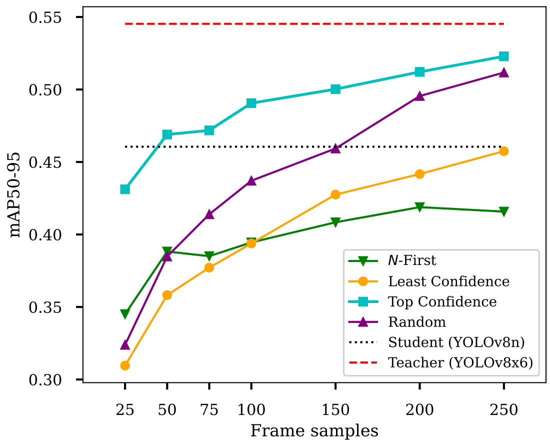

Figure 2. Learning curves obtained on the first two cameras of WALT. Results show that increasing the number of frames used for fine-tuning improves the student model's performance, approaching that of the teacher model with 250 frames. However, using only a small number of frames may lead to overfitting and poor performance on balanced evaluation sets. Top confidence thresholding is more effective than least confidence-based methods for stream-based active learning, highlighting the importance of avoiding highly uncertain images during fine-tuning.

fore, the choice of α should be based on a balance between performance and the length of the data stream required to select the desired number of frames.

5.2. Experimental Results

Figures 2. and 3. shows the learning curves obtained using stream-based active learning techniques on the WALT dataset. Our analysis can be approached from two perspectives. Firstly, from a knowledge distillation standpoint, we observe how the student model's performance improves as we use more frames for fine-tuning. In particular, we found that the mAP50-95 score approaches that of the teacher model when 250 pseudo-labeled frames are used. However, we also noticed that the student's performance deteriorates significantly when only a small number of frames are used for fine-tuning, which could be attributed to overfitting due to the limited number of images presented to the network. In addition, if the model is fine-tuned on images biased towards a specific time of day, such as only night or day, it may perform poorly on the balanced test set used for evaluation. Furthermore, as depicted in Figure 4, choosing highly uncertain images for pseudo-labeling may lead to incorrect labels due to the teacher's own bad prediction.

From an active learning perspective, the performance achieved with the *top confidence threshold* algorithm is significantly better than that obtained using the least confidence-based method. This highlights the importance of fine-tuning the model with highly certain images, especially when the model has not yet been specialized for the scene.

5.3. Limitations

The present work has three limitations. Firstly, the maximum budget is limited to 250 due to the frame rate and length of the data stream. Second, our approach was only evaluated on the WALT data set, and its generalizability to other data sets remains to be investigated. Third, the reduced number of heuristics may limit the effectiveness of the approach, and further exploration of different methods or combinations of methods could be a fruitful research direction. Additionally, exploring other deep neural network architectures, such as Transformers or Mask-RCNN, could also enhance the approach.

6. Conclusion

This paper proposes SBAD to bridge the gap between large-scale and affordable deep learning models while adapting to changing environments. This framework enables the scalable deployment of deep learning models under tight budget constraints.

The framework evaluates the informativeness of each frame, accounting for teacher imperfections in a KD scheme. Experiments demonstrate that traditional AL strategies may not be optimal for KD. Future research could explore alternative sampling strategies and distillation mechanisms to improve performance.

Acknowledgments

This work was partially funded by Win2WAL (#1910045) and Trusted AI Labs. We also thank *Openhub* for providing the equipment.

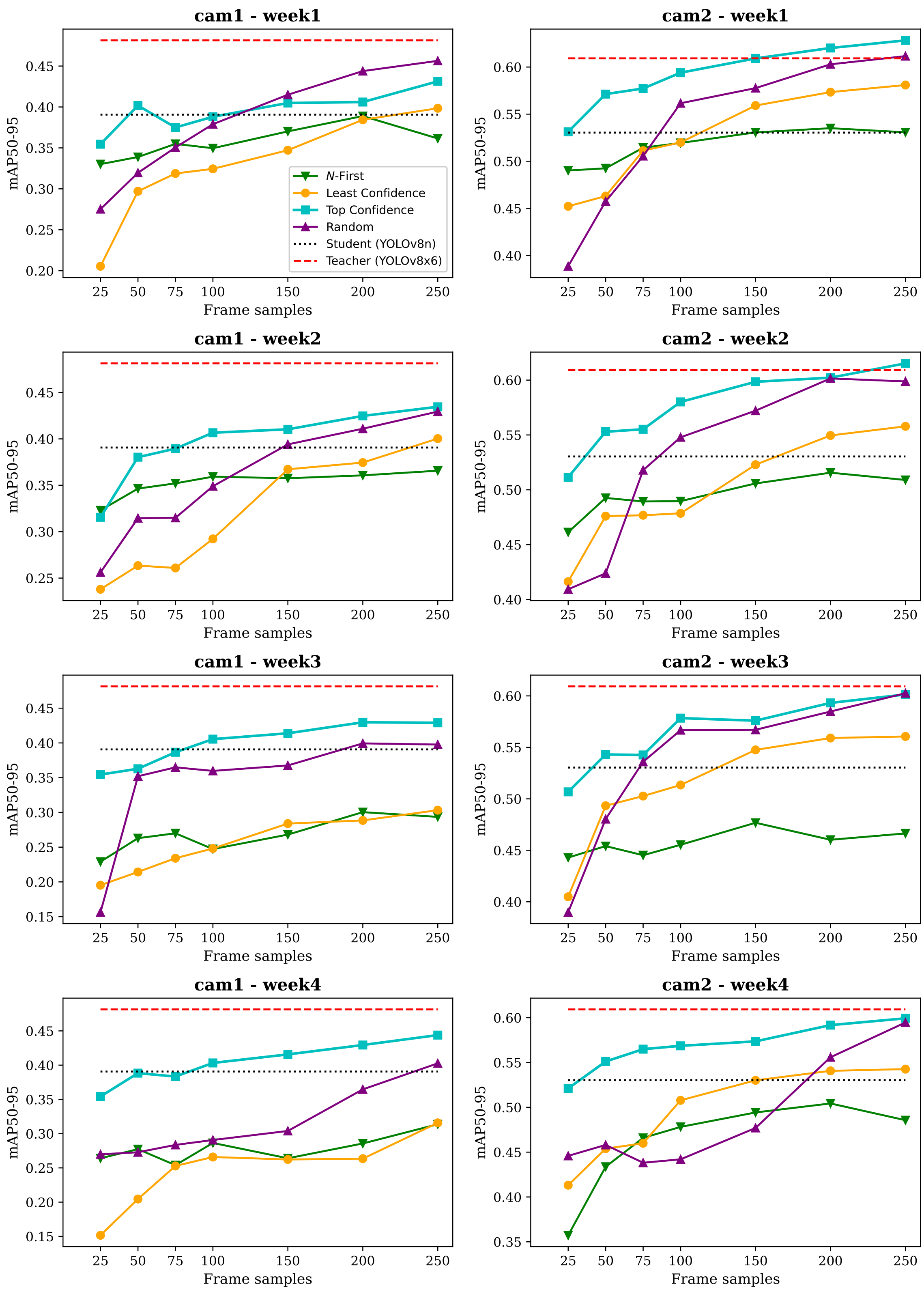

Figure 3. Weekly analysis on the first two cameras of WALT.

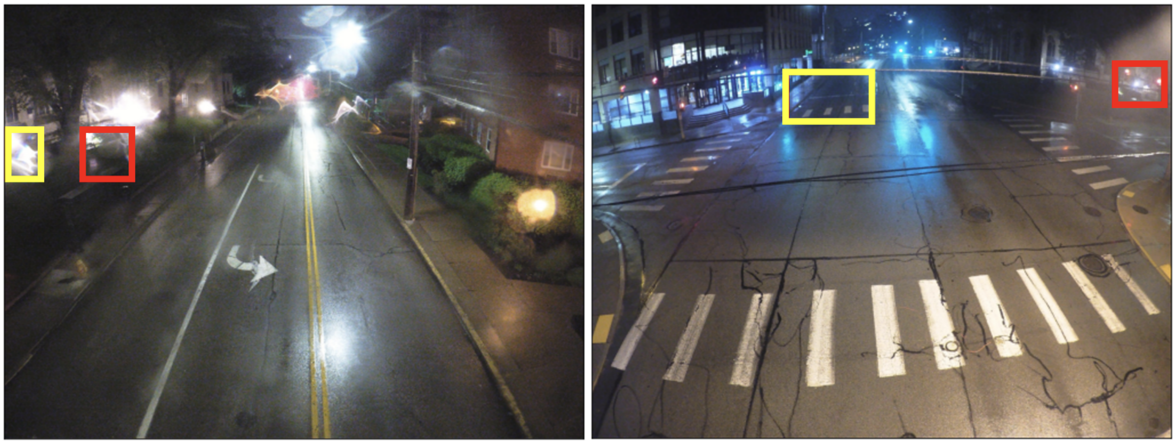

Figure 4. Two difficult examples (one for each camera) that lead to *confirmation bias*: when the student requests highly uncertain images based on its predictions (in yellow), wrong pseudo labels are revealed (in red).

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision (ECCV) 2020*, 8 2020. 3

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 1

[3] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *2020 International Conference on Learning Representations*, 6 2019. 3

[4] Cenk Baykal, Khoa Trinh, Fotis Iliopoulos, Gaurav Menghani, and Erik Vee. Robust active distillation. *arXiv preprint arXiv:2210.01213*, 2022. 2, 3

[5] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *CoRR*, abs/2106.05237, 2021. 1

[6] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA, 2006. Association for Computing Machinery. 2, 5

[7] Davide Cacciarelli and Murat Kulahci. A survey on online active learning. *arXiv preprint arXiv:2302.08893*, 2023. 2

[8] Davide Cacciarelli, Murat Kulahci, and John Tyssedal. Online active learning for soft sensor development using semi-supervised autoencoders. In *ICML 2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022. 4

[9] Davide Cacciarelli, Murat Kulahci, and John Sølve Tyssedal. Stream-based active learning with linear models. *Knowledge-Based Systems*, 254:109664, 10 2022. 4

[10] Davide Cacciarelli, Murat Kulahci, and John Sølve Tyssedal. Robust online active learning. *arXiv preprint arXiv:2302.00422*, 2023. 4

[11] Andrea Castellani, Sebastian Schmitt, and Barbara Hammer. Stream-based active learning with verification latency in non-stationary environments. In *Artificial Neural Networks and Machine Learning 2022*, 4 2022. 4

[12] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clément Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. *CoRR*, abs/2103.16130, 2021. 3

[13] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. Arthus: Adaptive real-time human segmentation in sports through online distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[14] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. In *Conference on Neural Information Processing Systems*, 7 2021. 3

[15] Sanjoy Dasgupta, Adam Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *Lecture Notes in Computer Science*, volume 10, 12 2005. 4

[16] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S. Shankar Sasrty. A convex optimization framework for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 209–

216. Institute of Electrical and Electronics Engineers Inc., 2013. 3

[17] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *30th IEEE Intelligent Vehicles Symposium*, 2019. 3

[18] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. 3

[19] Amirhossein Habibian, Haitam Ben Yahia, Davide Abati, Efstratios Gavves, and Fatih Porikli. Delta distillation for efficient video processing, 2022. 2

[20] Wei Huang, Shuzhou Sun, Xiao Lin, Dawei Zhang, and Lizhuang Ma. Deep active learning with weighting filter for object detection. *Displays*, page 102282, 1 2022. 3

[21] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision (ACCV) 2018*, 1 2018. 3

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5

[23] Salman Mohamadi, Gianfranco Doretto, and Donald A Adjeroh. Deep active ensemble sampling for image classification. In *16th Asian Conference on Computer Vision (ACCV 2022)*, 2022. 3

[24] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3573–3582, 2019. 2

[25] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *International Conference on Computer Vision (ICCV) 2021*, 2020. 3

[26] N. Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9356–9366, June 2022. 2, 4

[27] Carlos Riquelme, Ramesh Johari, and Baosen Zhang. Online active linear regression via thresholding. In *31st AAAI Conference on Artificial Intelligence*, 2017. 4

[28] Daniel Rivas, Francesc Guim, Jordà Polo, Pubudu M Silva, Josep Ll Berral, and David Carrera. Towards automatic model specialization for edge video analytics. *Future Generation Computer Systems*, 2022. 2

[29] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning (ECML*, 2006. 3

[30] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. *29th British Machine Vision Conference(BMVC)*, 2018. 3, 4

[31] Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *2020 IEEE Intelligent Vehicles Symposium*, 2020. 3

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018*, 8 2017. 3

[33] Burr Settles. Active learning literature survey. *Computer Sciences Technical article, University of Wisconsin–Madison*, 2009. 2

[34] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 3

[35] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166, 2021. 1

[36] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 3

[37] Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. *http://arxiv.org/abs/2211.11612*, 2022. 3

[38] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5 2019. 3

[39] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 3

[40] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4 2021. 3