Multi-Edge Reinforced Collaborative Data Acquisition for Continuous Video Analytics by Prioritizing Quality over Quantity

Lei Zhang[1], Guanyu Gao[1*], Haiyan Yin[2], Huaizheng Zhang[3]
[1]Nanjing University of Science and Technology, Nanjing, China
[2]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[3]Nanyang Technological University, Singapore
leizhang.real@gmail.com, gygao@njust.edu.cn, yin_haiyan@cfar.a-star.edu.sg, huaizhen001@e.ntu.edu.sg

Abstract

Edge computing-based video analytics faces data drift issues due to the occurrence of unseen objects or scenes in ever-changing environments. To maintain accuracy, continuous learning (CL) retrains stale models periodically with newly obtained data. However, it leads to unaffordable costs, as we must keep labeling drift data and retraining models. Regarding this concern, we first investigate video patterns across multiple cameras within an area and reveal significant data redundancies. We find that many of the same objects can be captured by multiple edge cameras or appear many times on the same edges. Our quantitative findings suggest that selecting a subset of high-quality data for CL is preferable over using a larger quantity. Yet, existing efforts for data acquisition have only focused on a single static dataset. These methods are not suitable for multi-edge video analytics scenarios, where videos are captured from multiple sources with non-iid data distribution. Hence, we propose a multi-edge collaborative active video acquisition (AVA) framework to collaboratively learn a reinforced video acquisition strategy to identify informative video frames from multiple edge nodes that best enhance model accuracy, avoiding redundancy across edges. Extensive experiments on three video datasets demonstrate that, our method achieves comparable performance to full-set video training while utilizing only 20% of the data in classification tasks. In object detection tasks, our methods can maintain productive accuracy with a reduction of nearly 70% in training video frames.

Introduction

Deep neural networks (DNNs) have demonstrated extraordinary performances in many video analytics applications, such as traffic monitoring and video surveillance. However, the real-world deployment of DNN models faces data drift issues, which arise when unforeseen objects or scenes appear after model deployment in ever-changing environments, leading to a decrease in recognition accuracy. This challenge is especially prominent in edge computing-based video analytics, where lightweight models are deployed due to the limited processing capacity of edge nodes, making them more sensitive to data drift.

Continuous learning (CL) (Bhardwaj et al. 2022) addresses the challenge of data drift by retraining stale DNN models with newly collected drift data from changing environments. However, adopting continuous learning for video analytics significantly increases financial costs due to the substantial human effort required for labeling drift data, which makes it less feasible for large-scale deployment. For instance, continuous learning for an object tracking model (e.g., YOLOv5) costs over $1,000/hour/video quotas for labeling and training on public clouds[1]. The prevalent approach to minimize the costs of continuous learning is to select a subset of informative data for model retraining. In video analytics, the drift data obtained from video cameras demonstrates considerable redundancy. Given that not every video frame contributes equally to accuracy improvement in model retraining, judiciously selecting the most valuable video frames can yield competitive model performance while significantly reducing overall costs.

Previous research has studied reducing labeling costs while preserving accuracy by designing effective data acquisition strategies to select a subset of training data. These efforts fall into three categories: 1) Uncertainty-based sampling (e.g., (Wang and Shang 2014; Wen, Tran, and Ba 2020; Gal and Ghahramani 2016)) focuses on identifying the most uncertain data instances by posterior probability distribution (Roth and Small 2006) or probabilistic modeling (Gal, Islam, and Ghahramani 2017; Choi et al. 2021); 2) Diversity-based sampling (e.g., (Sener and Savarese 2018; Agarwal et al. 2020; Sinha, Ebrahimi, and Darrell 2019)) involves the selection of a representative subset of instances that could effectively reflect the overall dataset distribution; 3) Hybrid strategies (e.g., (Shen et al. 2023; Parvaneh et al. 2022; Ash et al. 2020)) combine both uncertainty-based and diversity-based sampling methods to strike a balance between informativeness and diversity for the selected subset.

The existing data acquisition strategies are primarily based on modeling data uncertainty and diversity. Yet, the relationship between the uncertainty and diversity of training data and their impact on accuracy enhancement remains intricate. Furthermore, the previous works considered data acquisition from a single static dataset. In the context of multi-edge video analytics, however, video frames are cap-

*Corresponding Author.

[1]https://cloud.google.com/ai-platform/data-labeling/pricing

tured at different moments and locations, leading to non-iid data distribution. Meanwhile, the video data collected from different cameras demonstrate a spatial-temporal correlation. Independently optimizing data acquisition strategy for each edge may lead to redundancies across multiple edges.

To fill these gaps, we propose a multi-edge collaborative active video acquisition (AVA) framework for identifying informative and representative frames from drift videos collected on distributed edges. Given the subtle relationship between video semantics and accuracy improvement, we employ reinforcement learning to design an end-to-end approach that learns the video acquisition policy by considering semantic context across video frames and accuracy improvement feedback. The selected video frames will be labeled within a human-in-the-loop schema (Wu et al. 2022; Wang et al. 2016). To ensure data privacy, the labeled data on each edge is utilized for local model retraining and a federated learning approach is utilized for model aggregation. Furthermore, different edges can learn collaboratively for a distributed data acquisition policy to maximize the overall performance. Our primary contributions are as follows:

- Design a multi-edge collaborative video acquisition framework to reduce the cost of continuous learning in video analytics while maintaining competitive accuracy.
- Propose an end-to-end multi-edge reinforced video acquisition algorithm for learning to identify informative frames across-edge for continuous learning based on video semantic context.
- Conduct extensive experiments to verify our method's effectiveness, which can reduce nearly 60%∼70% training data while maintaining comparable accuracy in both classification and object detection tasks.

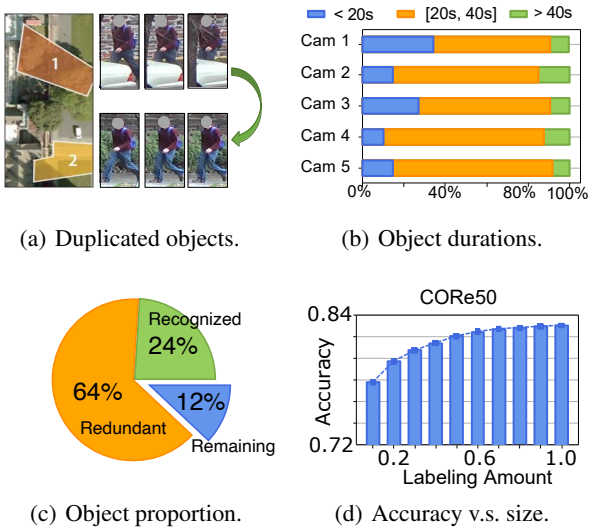
Motivation for AVA

Using all drift data for continuous learning can result in inefficient labeling and training. To validate this, we conduct empirical studies using real-world datasets (Lomonaco and Maltoni 2017) to characterize video patterns.

Observation I: Cross-Camera Redundancy. As illustrated in Fig. 1(a), the same object is often captured by different cameras, indicating a spatial-temporal correlation across different cameras. Thus, if one edge node labels and trains on drifted objects, it may potentially assist other nodes, avoiding redundant labeling and training costs.

Observation II: Temporal Redundancy. Videos frequently contain numerous redundant objects, especially between consecutive frames. Fig. 1(b) depicts the analysis of the real-world DukeMTMC dataset (Ristani et al. 2016), where most objects appear for over 20s in videos. Given the high similarity between these objects, labeling and training all video frames can be wasteful.

Observation III: Well-recognized Objects. Fig. 1(c) depicts the proportion of objects from the CL benchmark (Zhang, Gao, and Zhang 2023). We find that the stale model can accurately recognize nearly 25% of objects in drifted videos, and the well-recognized objects cannot significantly contribute to accuracy improvement during retraining.



(a) Duplicated objects. (b) Object durations.

(c) Object proportion. (d) Accuracy v.s. size.

Figure 1: Statistical analysis from real-world video datasets.

Retraining a model with redundant data not only incurs additional costs but also fails to acquire new knowledge. As illustrated in Fig. 1(d), the training accuracy does not increase linearly with the volume of training data. Instead, the marginal benefit diminishes as more data is included. The question arises: *which video frames should be acquired for labeling and model retraining?* This poses significant challenges, especially in some privacy-sensitive scenarios, such as multi-edge federated learning, where the data collected by different edges cannot be centralized for selection due to privacy concerns. Hence, these statistical findings inspire us to develop a distributed video acquisition approach to eliminate these video redundancies for multi-edge video analytics.

Methodology

Problem Formulation

A multi-edge video analytics problem refers to the task of analyzing and understanding videos captured by multiple edge nodes. Let $c \in \mathcal{C}$ denote each edge node in the set of edge nodes $\mathcal{C}$, i.e., $|\mathcal{C}| \geq 1$. The distribution of videos observed by the node c is denoted as $\mathcal{D}_c$, where $(x, y) \sim \mathcal{D}_c$ is an instance of the observed video from node c. The video data x could further be decomposed as a sequence of video frames, i.e., $x = \{x_1, x_2, ..., x_T\}$, where x_i is the i-th frame in the video and T is the length of the video frame sequence. Note that in practice, the distribution $\mathcal{D}_c$ is different from one edge to another due to various edge environments (e.g., camera views, illuminations). For multi-edge video analytics, the objective is to derive a video model, a.k.a a mapping function $f_\phi : \mathcal{X} \rightarrow \mathcal{Y}$, where ϕ is the parameters for the function $f(\cdot)$, $\mathcal{X}$ is the video data space, and $\mathcal{Y}$ is the label space. The function is trained by minimizing the loss of the model over data obtained from the multi-edge setting, i.e, $\min_\phi \mathbb{E}_{x\sim\mathcal{X}}[\mathcal{L}(f_\phi(x), y)]$, where $x \in \mathcal{X}$ is the video instances and $y \in \mathcal{Y}$ is the corresponding ground-truth labels,

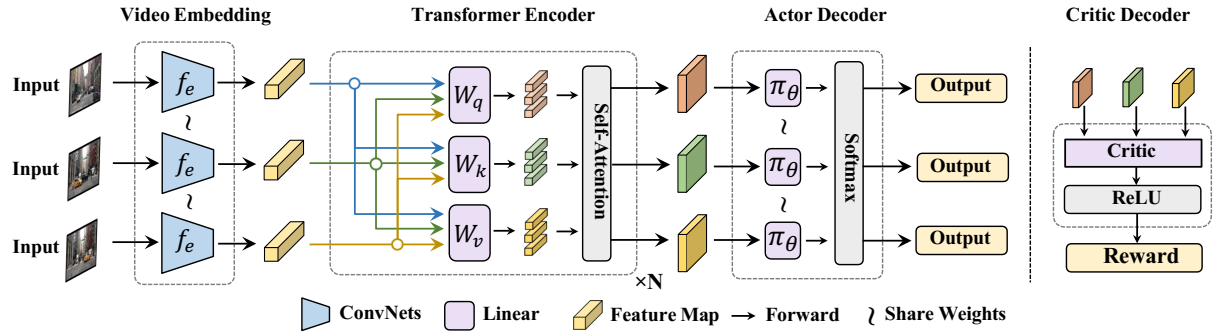


Figure 2: The AVA policy network embeds video data and then decides which frames to acquire.

$\mathcal{L}(\cdot)$ is the loss function (e.g., cross-entropy), and the expectation is taken over the videos x from all the edges.

In reality, labeling and training for multi-edge video analytics are extremely expensive. Furthermore, the videos contain a high degree of redundancy (refer to Section § for statistical evidence), which motivates us to consider a cost-sensitive video analytics challenge. Overall, we aim to achieve the goal of automatic data acquisition of informative video subsets to benefit the model performance gaining, whereas such data acquisition could help alleviate the costly burden of labeling and training data, whilst not dropping the performance compared to training on full data. Formally, we denote such a problem as a multi-edge video acquisition problem. We aim to learn an active video acquisition function $\pi_\theta(\cdot)$ parameterized by θ that could intelligently acquire a subset of videos that are most helpful to the training of $f_\phi(\cdot)$. We denote the AVA selected subset as $\tilde{\pi}_\theta(x)$[2]. In our application, since the cost for labeling and training for each video slice is almost equivalent, we assume the cost is linear to the size of the subset, i.e., $|\tilde{\pi}_\theta(x)|$. To this end, we provide a formal objective in Eq. (1),

$$\min_{\theta,\phi} \sum_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}_c}\Big[\mathcal{L}\Big(f_\phi(\tilde{\pi}_\theta(x)), y\Big)\Big] \quad \text{s.t.} \quad |\tilde{\pi}_\theta(x)| \le \lambda|x|, \tag{1}$$

where for each edge node $c \in \mathcal{C}$, we aim to minimize the expectation of the video model f_ϕ loss trained on video subset $\tilde{\pi}_\theta(x)$. We also ensure that the size of video subset is smaller than full video, i.e., $|\tilde{\pi}_\theta(x)| \le \lambda|x|$, where $\lambda \in (0, 1)$ is the maximum affordable proportion of training videos.

AVA Architecture

We develop the AVA network as an encoder-decoder architecture (shown in Fig. 2). Specifically, the encoder layers are defined as the Transformer-based network (Vaswani et al. 2017), which aims to extract the spatial-temporal visual features across video frames over time. The decoder layers are mainly composed of actor-critic structures (Schulman et al. 2017) to output two groups of predictions. The first group corresponds to the AVA video acquisition probability predicted by the actor. For each frame, it determines whether to acquire the video frame for training the video analytics model. The second group corresponds to the estimated value of the expected reward (Mnih et al. 2016), which is used to compute the policy gradient loss to optimize the AVA policy model. The details of the AVA network are as follows.

Transformer-based Encoder. We adopt the Transformer-based network (Vaswani et al. 2017) as the encoder of our policy network. Transformer incorporates a learnable self-attention mechanism, which allows it to capture spatial-temporal similarities and relationships among input sequences. Formally, we first extract the features of input video sequences as the vector $x \in \mathbb{R}^{T\times d}$ by the pre-trained models (e.g., ResNet, VGG), where T is the length of video sequences and d is the dimension of the features. Then, we adopt the learnable self-attention to capture the spatial-temporal embeddings among video sequences as follows,

$$\mathbf{Attention}(x) := \mathbf{softmax}\left(\frac{x\mathbf{W}_q \cdot (x\mathbf{W}_k)^T}{\sqrt{d}}\right) \cdot x\mathbf{W}_v, \tag{2}$$

where the input sequences $x \in \mathbb{R}^{T\times d}$ are operated by three linear projection metrics $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{d\times d}$. By utilizing the Transformer as the encoder for our policy network, we can effectively discover the importance of each video frame relative to the entire video sequence. Additionally, our Transformer-based encoder in our policy network could be stacked with multiple layers to help compress richer semantic features. In the following, we assume the Attention(·) function has only one layer, whereas in practical implementation, we usually stack more than three layers.

Actor-Critic Decoder. We employ the actor-critic framework as the decoder for our policy network. In our architecture, both the actor and critic networks are based on multi-layer perceptrons (MLPs). The actor network aims to explore decision-making for video acquisition, while the critic network guides the actor's exploration by providing expected rewards (Mnih et al. 2016). Specifically, the actor network for the AVA architecture determines whether a frame in the video sequences should be acquired or not. Hence, the actor network can be defined as a binary decision

[2]The output of the policy $\pi_\theta(x)$ is a probability distribution, whereas $\tilde{\pi}_\theta(x)$ is the video subset sampled from $\pi_\theta(x)$.

for each video frame to output the acquisition probability distribution. We define the actor network π_θ as follows,

$$\pi_\theta(x) := \sigma(\mathbf{Attention}(x) \cdot \mathbf{W}_\alpha + b), \tag{3}$$

where the input sequences $x \in \mathbb{R}^{T\times d}$ are firstly operated by attention-based encoder, and then projected by the linear weight $\mathbf{W}_\alpha \in \mathbb{R}^{d\times 2}$ with an activation function $\sigma(\cdot)$.

The critic network aims to guide the actor network updates by providing expected rewards in reinforcement learning (Mnih et al. 2016). We also define the critic network v based on MLPs as follows:

$$\mathrm{v}(x) := \sigma(\mathbf{Attention}(x) \cdot \mathbf{W}_\beta) \cdot \mathbf{W}_\gamma^T, \tag{4}$$

where the attention-based encoder outputs are projected by the linear weight $\mathbf{W}_\beta \in \mathbb{R}^{d\times 1}$ with activation $\sigma(\cdot)$, and then we aggregate the sequence dimension by a linear weight $\mathbf{W}_\gamma \in \mathbb{R}^{T\times 1}$ to get the final expected rewards. Additionally, the critic network is activated only during the AVA training stages and is removed during the inference stage.

The actor network outputs selecting probabilities of each frame, and we adopt the strategy of probabilistic sampling instead of directly using the Top-1 probability for video frame acquisition. This is because we expect the actor network to increase the exploration of unknown decisions through probabilistic sampling, allowing for the opportunity to try other possible decisions and gain a better understanding of the video environment to find the optimal strategies. Formally, the binary selecting probabilities $\pi_\theta(x)$ of the actor network are sampled under *Categorical* distribution as $\alpha(x) \sim \mathbf{Categorical}\big(\pi_\theta(x)/\tau\big)$, where $\alpha(x) \in \{0,1\}^T$ is the selecting decision for the given video sequences with T length, and τ aims at adjusting the scale of the probability distribution $\pi_\theta(x)$. We experimentally find that a larger τ makes the probability distribution flatter, encouraging exploration of unknown optimal decisions to optimize the AVA network. Conversely, a smaller τ makes the probability distribution more concentrated, favoring the sampling with higher probabilities to increase exploitation of the current optimal decisions (Haarnoja et al. 2018; Hao et al. 2023).

Multi-Edge Reinforcement Learning

Reward Modeling. Unlike many off-the-shelf RL tasks where the reward signal is the golden standard, e.g., game scores in Atari 2600 (Mnih et al. 2015) and winning outcome of the match in AlphaGo (Silver et al. 2016), It is hard to develop informative reward signal for multi-edge AVA. Intuitively, we propose compositional rewards with three key components to effectively facilitate AVA problem-solving. Ideally, the first reward is to reflect the progress of objectiveness as depicted in Eq. (1),

$$R_{obj} = -\mathcal{L}\Big(f_\phi\big(\tilde{\pi}_\theta(x)\big), y\Big) + \lambda\Big(1 - \frac{|\tilde{\pi}_\theta(x)|}{|x|}\Big), \tag{5}$$

where the first component is designed to evaluate the benefit of the acquired video subset $\tilde{\pi}_\theta(x) = \{x_i \mid \alpha(x_i) = 1, i \in \mathbb{N}\}$ to the video model f_ϕ training, and the second component is to constrain the size of acquired subset $|\tilde{\pi}_\theta(x)|$ compared with the size of full-set $|x|$. We introduce the hyperparameter λ to balance these two objectives. We experimentally find that a higher value of λ will make the AVA policy model prefer to select a smaller number of video sequences, whereas a lower value of λ will enforce the AVA policy model to prioritize higher model training performance.

In addition to the performance-related reward, we also introduce two auxiliary rewards to train the AVA policy more efficiently with reward shaping (Hu et al. 2020; Arulkumaran et al. 2017). Basically, the motivation of these two auxiliary rewards is to promote the diversity and representativeness of the acquired subset. The formal definition for these two auxiliary rewards are presented as follows,

$$\begin{aligned} R_{div} =& \frac{1}{T(T-1)} \sum_{i\in\mathcal{I}} \sum_{\substack{j\in\mathcal{I},\\ i\neq j}} \Big(1 - \frac{x_i \cdot x_j^T}{\|x_i\| \cdot \|x_j\|}\Big), \\ R_{rep} =& \frac{1}{T} \sum_{i=1}^{T} \min_{j\in\mathcal{I}} \Big\{1 - \frac{x_i \cdot x_j^T}{\|x_i\| \cdot \|x_j\|}\Big\}, \end{aligned} \tag{6}$$

where R_{div} reflects the diversity of the acquired subset $\{x_i\}_{i\in\mathcal{I}}$ by summarizing the cosine distances for every pair of instances in the subset, and R_{rep} evaluates the representativeness of the video subset $\{x_j\}_{j\in\mathcal{I}}$ with respect to the full-set $\{x_i\}_{i=1}^T$ by accumulating the cosine distances from each full-set instance to the nearest subset instance. The selecting indices of acquired subset are $\mathcal{I} = \{\mathcal{I}_i \mid \alpha(x_{\mathcal{I}_i}) = 1, i \in \mathbb{N}\}$. The compositional rewards work cooperatively, and the reward shaping signals help to overcome the noise in the performance rewards, while providing informative feedback on the quality of AVA policy.

Multi-Edge AVA Policy Training. Our AVA policy network is trained under a collaborative regime. For each node $c \in \mathcal{C}$, it optimizes a shared AVA policy network π_θ by their individual video environment $\mathcal{D}_c$. We define its global objective as $\max_{\pi_\theta} \mathbb{E}\big[\sum_{c\in\mathcal{C}} \gamma_c \cdot R_c\big]$, where R_c is the cumulative reward from the three aforementioned individual rewards at the node c, and $\gamma \in (0,1)$ is the importance weight of each $c \in \mathcal{C}$ in final global reward. Additionally, the importance weight can be estimated as $\gamma_c = |\mathcal{D}_c| / \sum_{i\in\mathcal{C}} |\mathcal{D}_i|$, where $|\mathcal{D}_c|$ is the length of video sequence on the edge node c ($c \in \mathcal{C}$). On distributed edge nodes, we update the AVA policy network by multi-edge gradient synchronization. The policy networks deployed on various edge videos could collect a variety of experiences during continuous learning (Mnih et al. 2016; Espeholt et al. 2018). We enable the edge policy networks to learn collaboratively. Specifically, we first deploy a shared policy network for all AVA edge networks. Then the edge policies evaluate local edge gradients by evaluating the actor-critic loss using their own experience (Mnih et al. 2016). The local gradients are accumulated by the parameter center and the parameters are dispatched to edge nodes. Note that such multi-edge policy update is different from off-the-shelf distributional reinforcement learning, e.g., IMPALA (Espeholt et al. 2018), PPO (Schulman et al. 2017), and Vtrace (Kapturowski et al. 2019), because the latter ones transfer experience tuples to the centralized learner without local gradient inference or data drift.

Experiment

Experimental Setting

Datasets. We utilize 3 video benchmarks, namely, CORe50 (Lomonaco and Maltoni 2017), VisDrone (Zhu et al. 2021), and MOT15 (Leal-Taixé et al. 2015). VisDrone (Zhu et al. 2021) is designed for real-world video object tracking from UAV cameras. MOT15 (Leal-Taixé et al. 2015) is an object tracking benchmark collected from moving and static cameras. CORe50 (Lomonaco and Maltoni 2017) consists of 11 videos in indoor and outdoor settings.

Baselines. We compare with the state-of-the-art data acquisition approaches, namely, Entropy (Wang and Shang 2014), Coreset (Bodó, Minier, and Csató 2011), MDN (Choi et al. 2021), BALD (Gal, Islam, and Ghahramani 2017), BADGE (Ash et al. 2020), CDAL-CS (Agarwal et al. 2020), and ALFA-Mix (Parvaneh et al. 2022). Entropy (Wang and Shang 2014) is the uncertainty-based approach to select the frames with the lowest confidence for training. Coreset (Bodó, Minier, and Csató 2011) and CDAL-CS (Agarwal et al. 2020) are the diversity-based approaches to select a batch of diverse representative frames to represent the whole video for training. MDN (Choi et al. 2021) and BALD (Gal, Islam, and Ghahramani 2017) are the probabilistic approaches, which adopt the mixture density networks and Bayesian approach respectively, to estimate the frame selecting probabilities with contextual information. BADGE (Ash et al. 2020) and ALFA-Mix (Parvaneh et al. 2022) incorporate the uncertainty- and diversity-based methods, considering both predictive uncertainty and sample diversity.

Performance Comparison

Performance on Classification. We evaluate our method's effectiveness in data acquisition for classification accuracy by comparing it to other baselines on the CORe50 video benchmark. In this experiment, there are eight nodes, each handling an individual video sequence to select for model training. Table 1 illustrates that, under a labeling and training budget of about 30% of video frames, our method can outperform other baselines with nearly 2% Top-1 accuracy improvement. Furthermore, this improvement persists across different budget thresholds. As illustrated in Fig. 3(c), a varying video selecting proportion ranging from 10% to 50% continues to exhibit the superior performance of our method over other baselines. Additionally, it's noteworthy from Fig.3(c) that increasing the video selection budget does not lead to a linear increase in model accuracy, aligning with observations in our motivation studies. Hence, it is also suggested to meticulously determine the budgets and expected model accuracy to achieve a better trade-off between labeling/training costs and model accuracy.

Performance on Object Detection. Object detection is widely involved in real-world video analytics applications, therefore, we further conduct experiments on video object detection benchmarks — specifically, VisDrone with UAV cameras, and MOT15 with the KITTI video sequences. In this experiment, we deploy three edge nodes to acquire their respective video sequences for Faster R-CNN model training. Additionally, most baseline methods, except MDN, are not specifically designed for object detection tasks. Hence, to adapt Entropy, we follow (Choi et al. 2021) to estimate the average scores of each frame's predictive bounding boxes. For the Coreset, we follow (Khani et al. 2021) to perform k-center-greedy over its frame-level latent features. For BADGE, we refine its gradient embedding by using a weighted summation of latent features from all predictive bounding boxes. For BALD, we implement each frame's uncertainty as the averaged posterior scores of predictive objects. For CDAL-CS, we follow (Khani et al. 2021) to estimate its contextual diversity by pooling features from the FPN backbone. As depicted in Table 2, with labeling and training budgets of approximately 30% of video frames, our method demonstrates a 3.64% relative gain in average AP accuracy, with improvements of 8.36% and 2.25% in AP_{50} and AP_{75} accuracy, respectively, over the strongest baseline on VisDrone and MOT15 benchmarks. Moreover, as illustrated in Fig.3(a) and Fig.3(b), this advantageous performance can also be observed across different budgets from video acquisition proportion between 10% to 50%.

CORe50-Classification (nodes=8, budget $\leq 20\%$)			
Sampling	**Top-1 (%)**	**Top-3 (%)**	**Top-5 (%)**
Random	59.84±1.77	80.68±0.88	87.69±0.85
Entropy	60.09±1.72	80.57±0.53	87.66±0.61
Coreset	62.17±1.51	82.09±0.88	88.49±0.75
BADGE	60.58±2.15	82.26±0.43	87.09±0.35
BALD	61.05±1.30	81.72±0.27	88.26±0.29
CDAL	60.51±1.13	81.26±0.43	87.83±0.53
Alpha-Mix	61.09±1.26	81.93±0.27	88.37±0.33
Ours	**64.02±0.83**	**83.42±0.26**	**90.28±0.18**

Table 1: The performance comparison with different methods on the CORe50 object classification benchmark.

Findings on Selection Bias. In Table 2, our method demonstrates superior performance across most evaluation metrics, such as AP, AP_{50}, and AP_{75}. However, when considering AP_s and AP_m, specifically concerning small- and medium-sized objects, we notice slight weaknesses against some baselines. This observation suggests that our method may preferentially select video frames containing larger objects, leading to a potential selection bias. To mitigate selection bias, we incorporate the precision of AP_s, AP_m, and AP_l as auxiliary rewards (e.g., R/small, R/medium, R/large) to balance the acquisition of frames with objects of varying sizes. The results, as depicted in Table 2, showcase a notable improvement. For instance, employing the R/small reward could gain improvements of 3.35% in AP_s and 5.78% in AR_s for the VisDrone dataset. In the MOT15 benchmark, we can also witness the improvements of 5.24% in AP_s and 2.28% in AR_s. Furthermore, the introduction of auxiliary rewards R/medium and R/large also shows improvements in the recognition accuracy for medium- and large-sized ob-

VisDrone-Detection (nodes=3, budget $\leq 30\%$)

Sampling	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR_1	AR_{10}	AR_{100}	AR_s	AR_m	AR_l
Random	50.47±0.52	79.17±1.34	57.17±1.28	43.31±1.01	55.72±0.17	36.64±9.10	32.82±0.63	50.89±0.45	56.06±0.67	51.71±1.02	60.55±0.09	39.10±8.26
Entropy	51.76±1.59	80.16±0.36	61.69±2.49	47.15±1.04	56.71±1.08	34.86±3.82	33.30±0.61	51.96±0.62	57.33±0.70	54.72±0.62	61.16±0.95	38.33±3.89
Coreset	53.04±0.49	80.25±0.09	61.66±1.04	44.74±1.10	58.22±0.17	41.68±2.69	33.83±0.20	52.62±0.36	57.59±0.40	52.19±1.17	62.57±0.13	44.76±2.94
MDN	53.14±0.44	81.52±0.19	61.64±1.73	45.47±1.55	59.26±1.32	50.51±6.40	34.89±0.51	55.16±0.54	58.16±0.41	52.13±1.47	65.17±1.50	50.48±4.49
BALD	52.04±1.13	80.64±0.47	61.58±1.60	43.72±0.74	55.93±0.62	40.38±8.77	34.07±0.73	53.64±0.54	57.62±0.71	53.07±0.37	61.04±1.12	40.55±6.19
BADGE	51.30±0.79	79.32±0.27	60.21±1.28	44.69±0.32	55.24±1.17	40.48±2.23	33.56±0.65	51.91±0.84	56.79±0.73	52.36±0.12	60.20±1.34	43.12±2.45
CDALCS	52.70±0.19	79.47±0.24	62.06±0.95	45.52±1.50	57.15±0.44	40.61±2.56	34.06±0.70	52.98±0.10	57.73±0.18	53.15±1.12	61.63±0.39	41.85±1.61
Ours	**56.68±0.60**	**89.88±0.53**	**64.25±2.92**	46.42±0.77	62.66±0.92	64.48±6.32	**39.65±1.18**	**58.07±0.79**	**63.20±0.37**	50.36±1.54	67.97±0.64	66.22±6.68
+ R/small	53.12±0.66	82.53±0.37	62.75±2.74	**49.77±1.12**	50.36±0.53	43.55±8.72	33.29±0.95	53.43±0.85	58.86±0.49	**56.14±1.22**	62.97±0.88	49.60±5.20
+ R/med	54.19±0.73	83.36±0.41	63.03±2.23	47.14±0.85	**64.14±0.79**	63.02±7.41	34.15±0.98	54.30±0.74	59.27±0.81	53.15±1.21	**68.11±1.55**	52.56±4.16
+ R/large	54.42±0.41	85.07±0.35	63.80±2.09	43.20±0.90	61.12±0.22	**66.37±5.57**	35.72±0.74	56.10±0.56	59.12±0.83	49.86±1.64	66.14±0.75	**68.52±7.71**

MOT15-KITTI-Detection (nodes=3, budget $\leq 30\%$)

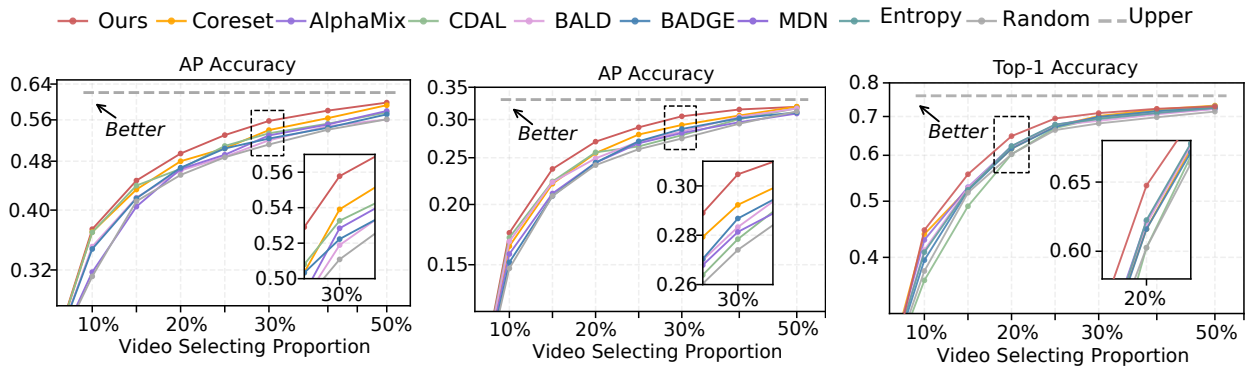
Sampling	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR_1	AR_{10}	AR_{100}	AR_s	AR_m	AR_l
Random	27.31±1.87	62.47±1.06	18.65±3.30	27.14±9.29	36.74±1.04	42.68±5.12	9.92±0.59	39.09±1.21	46.10±1.04	31.87±0.97	41.61±1.48	37.06±1.31
Entropy	28.22±0.55	62.07±0.87	16.50±0.63	23.68±12.13	35.11±10.92	45.59±5.84	9.81±0.21	38.83±0.38	45.43±0.20	31.78±11.30	41.57±10.53	47.80±3.54
Coreset	29.10±0.44	61.84±0.78	17.66±1.08	24.48±10.81	35.30±10.76	47.18±5.41	9.96±0.24	38.49±0.44	45.44±0.11	31.22±10.06	42.09±10.51	46.26±3.83
MDN	29.15±0.42	64.53±0.41	21.07±1.22	28.84±9.95	32.81±10.80	38.15±3.35	10.01±0.31	38.42±0.87	43.44±0.18	36.64±9.30	38.19±9.56	42.66±5.64
BALD	29.38±0.61	65.16±1.09	19.88±0.96	26.76±8.63	38.52±10.84	40.55±8.06	10.31±0.19	41.00±0.61	46.92±0.52	31.21±8.39	44.03±10.63	43.94±3.87
BADGE	28.77±0.68	62.47±1.11	18.51±1.62	26.56±10.17	35.44±10.77	41.02±5.24	9.84±0.37	39.30±0.52	45.54±0.54	31.96±9.26	42.64±9.86	42.03±4.10
CDALCS	28.03±0.50	63.30±0.85	18.19±0.30	27.65±9.45	36.57±11.05	42.61±7.72	9.73±0.24	39.24±0.27	46.65±0.54	33.43±10.34	42.24±11.12	50.10±5.06
Ours	31.91±0.67	64.60±0.62	**26.85±1.67**	24.27±1.38	34.87±7.80	44.51±10.52	10.58±1.15	42.41±1.11	**48.30±0.79**	36.30±8.27	51.89±10.64	56.95±2.35
+ R/small	**32.14±1.44**	66.94±1.15	26.18±2.13	**29.51±4.92**	35.03±5.06	43.72±8.74	**10.66±0.97**	**42.55±1.53**	47.98±0.80	**38.58±4.32**	51.12±12.13	54.91±5.43
+ R/med	31.43±0.81	**68.53**±1.12	23.67±1.47	24.52±3.51	**38.42±6.74**	39.07±8.51	10.12±2.26	42.31±2.31	47.74±0.85	36.92±6.40	**52.21±15.51**	48.74±8.04
+ R/large	30.86±0.92	61.03±1.18	20.28±1.53	24.15±1.40	35.47±7.32	**47.75±9.89**	9.97±1.58	42.38±1.61	46.64±0.73	34.50±5.16	51.07±12.09	**58.42±7.13**

Table 2: The performance comparison of different methods on two video object detection benchmarks.

jects across both VisDrone and MOT15 benchmarks. Hence, our findings suggest that incorporating auxiliary rewards can effectively enhance the object recognition across different sizes in various video analytic tasks.

Eliminating Cross-Camera Redundancy. In Fig. 4, we validate our method's effectiveness in eliminating the cross-camera redundancies. As shown in Fig. 4(a), we employ our AVA policy network to select videos for ResNet18 training, with the given requirement of training accuracy to be higher than 60% (the upper bound with full-set training is 64.3%). Fig. 4(a) depicts that our method significantly reduces the data selecting ratios as more cameras are included. Specifically, with 20 cameras, our method with multi-edge training (Multi-Edge AVA) acquires less than 20% video frames, while maintaining the required 60% training accuracy. In contrast, the single-edge training (Single-Edge AVA) achieves a weaker performance, needing nearly $1.5\times$ training data to achieve comparable training accuracy. Additionally, the advantages persist when training the Vision Transformer network with videos of different camera amounts. We believe this improvement is largely attributed to our design of multi-edge collaborative training for the AVA policy network, enhancing its global knowledge to eliminate cross-camera redundancies, and reduce the selection of similar objects or instances across different camera videos.

Multi-Edge Adaptive Sampling. As shown in Fig. 5, in contrast to naive equal sampling, our AVA policy network excels in performing adaptive sampling for different edge videos, which allows our AVA policy network autonomously adjust the sampling size across distributed edge nodes, and selects the most informative data for model training. We believe this experimental result arises because our AVA policy network prioritizes selecting videos in which the model is uncertain, while reducing the selection of some well-trained edge videos to save on labeling and training budgets. To validate our hypothesis, we further follow (Li et al. 2020) to estimate the gradient norm $\|\nabla f_\phi(\tilde{x})\|_2$ for each node, which can effectively reveal the model's convergence across different edges. Specifically, a higher gradient norm signifies weak fitting to the edge node videos, whereas a lower norm suggests strong fitting to the edge node videos. Fig. 3 shows that if an edge node has a higher gradient norm, the model will have weak fitting on its video data, causing the AVA policy network to select more data for video training. In contrast, the lower gradient norm means strong fitting, thereby, our AVA policy network can intelligently reduce the budgets of data acquisition. These results further demonstrate the effectiveness of our AVA policy network in learning an



(a) VisDrone Benchmark (b) MOT15 Benchmark (c) CORe50 Benchmark

Figure 3: Performance comparison between our method and other data acquisition strategies on three different benchmarks.

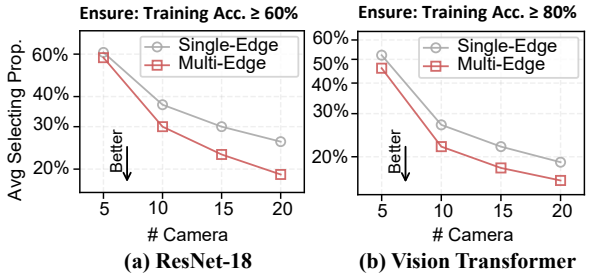


Figure 4: Selecting proportion across camera numbers.

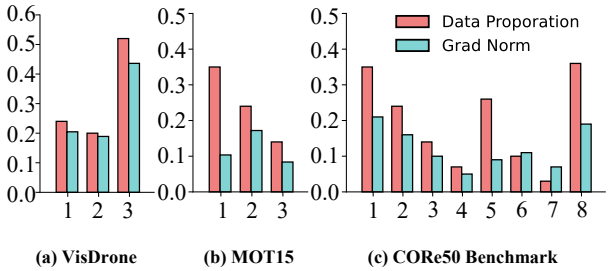


Figure 5: Selecting proportion and model gradient norm.

optimal cross-edge strategy to balance the global trade-off between model accuracy and video training budgets.

Related Work

Uncertainty-based sampling (Roth and Small 2006; Wen, Tran, and Ba 2020; Gal and Ghahramani 2016; Huang et al. 2022; Choi et al. 2021; Gal, Islam, and Ghahramani 2017) primarily selects the most uncertain instances within the posterior probability distribution using measures, such as entropy (Wang and Shang 2014) and margin boundaries (Gal and Ghahramani 2016). Uncertainty can also be estimated with probabilistic modeling, e.g., Ensemble (Wen, Tran, and Ba 2020) introduces the disagreement uncertainty by Monte-Carlo dropout (Gal and Ghahramani 2016), MDN (Choi et al. 2021) estimates aleatoric and epistemic uncertainty with Gaussian mixture models (GMM), and BALD (Gal, Islam, and Ghahramani 2017) estimates the uncertain instances with the highest mutual information (MI) between predictions and posterior Bayesian approximation.

Diversity-based sampling (Sener and Savarese 2018; Agarwal et al. 2020; Sinha, Ebrahimi, and Darrell 2019; Guo, Zhao, and Bai 2022) selects a subset of instances to represent the entire dataset distribution using the algorithms, such as the k-medoid (Kaufman and Rousseeuw 2009) or k-center-greedy (Bodó, Minier, and Csató 2011). For instance, Coreset (Sener and Savarese 2018) greedily queries the representative data centroids to minimize the total distance from other data samples. CDAL-CS (Agarwal et al. 2020) adopts k-center-greedy (Bodó, Minier, and Csató 2011) to query the representative centroids with the varied pairwise contextual diversity. VAAL (Sinha, Ebrahimi, and Darrell 2019) trains a variational auto-encoder (VAE) to separate labeled and unlabeled samples in latent space and query the samples based on the output probability of the discriminator.

Hybrid strategy (Shen et al. 2023; Parvaneh et al. 2022; Xie et al. 2023; Ash et al. 2020) incorporates both uncertain-based and diversity-based sampling to achieve a better trade-off between informativeness and diversity. For instance, ALFA-MIX (Parvaneh et al. 2022) utilizes k-means clustering to refine the queried samples to boost latent representations. BADGE (Ash et al. 2020) represents unlabeled data samples by gradient embedding to measure their uncertainty and query the representative samples by the k-means++ (Arthur and Vassilvitskii 2007).

Conclusion

This paper investigates the data acquisition strategies for reducing the cost of utilizing continuous learning to tackle data drift in video analytics. We develop a multi-edge reinforced video acquisition algorithm that identifies a small and high-quality subset across multi-edge to optimally enhance model improvement. Our method can reduce 60%∼70% of training data while achieving comparable accuracy to achieve cost-efficient continuous video analytics.

Acknowledgements

Haiyan Yin is supported by Career Development Fund (CDF) of the Agency for Science, Technology and Research (A*STAR) (Grant Number: C233312007).

References

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*, 137–153. Springer.

Arthur, D.; and Vassilvitskii, S. 2007. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035.

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bhardwaj, R.; Xia, Z.; Ananthanarayanan, G.; Jiang, J.; Shu, Y.; Karianakis, N.; Hsieh, K.; Bahl, P.; and Stoica, I. 2022. Ekya: Continuous learning of video analytics models on edge compute servers. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 119–135.

Bodó, Z.; Minier, Z.; and Csató, L. 2011. Active learning with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, 127–139. JMLR Workshop and Conference Proceedings.

Choi, J.; Elezi, I.; Lee, H.-J.; Farabet, C.; and Alvarez, J. M. 2021. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10264–10273.

Espeholt, L.; Soyer, H.; Munos, R.; Simonyan, K.; Mnih, V.; Ward, T.; Doron, Y.; Firoiu, V.; Harley, T.; Dunning, I.; et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, 1407–1416. PMLR.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, 1183–1192. PMLR.

Guo, C.; Zhao, B.; and Bai, Y. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, 181–195. Springer.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hao, J.; Yang, T.; Tang, H.; Bai, C.; Liu, J.; Meng, Z.; Liu, P.; and Wang, Z. 2023. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*.

Hu, Y.; Wang, W.; Jia, H.; Wang, Y.; Chen, Y.; Hao, J.; Wu, F.; and Fan, C. 2020. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33: 15931–15941.

Huang, G.; Wang, Y.; Lv, K.; Jiang, H.; Huang, W.; Qi, P.; and Song, S. 2022. Glance and focus networks for dynamic visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4605–4621.

Kapturowski, S.; Ostrovski, G.; Quan, J.; Munos, R.; and Dabney, W. 2019. Recurrent Experience Replay in Distributed Reinforcement Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kaufman, L.; and Rousseeuw, P. J. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Khani, M.; Hamadanian, P.; Nasr-Esfahany, A.; and Alizadeh, M. 2021. Real-time video inference on edge devices via adaptive model streaming. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4572–4582.

Leal-Taixé, L.; Milan, A.; Reid, I.; Roth, S.; and Schindler, K. 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Lomonaco, V.; and Maltoni, D. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, 17–26. PMLR.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 1928–1937.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M. A.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nat.*, 518(7540): 529–533.

Parvaneh, A.; Abbasnejad, E.; Teney, D.; Haffari, G. R.; Van Den Hengel, A.; and Shi, J. Q. 2022. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12237–12246.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, 17–35. Springer.

Roth, D.; and Small, K. 2006. Margin-based active learning for structured output spaces. In *Machine Learning:*

ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17, 413–424. Springer.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shen, M.; Huang, Y.; Yin, J.; Zou, H.; Rajan, D.; and See, S. 2023. Towards Balanced Active Learning for Multimodal Classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 3434–3445. New York, NY, USA: Association for Computing Machinery.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T. P.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nat.*, 529(7587): 484–489.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5972–5981.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, H.; Gong, S.; Zhu, X.; and Xiang, T. 2016. Human-in-the-loop person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 2016, Proceedings*, 405–422. Springer.

Wen, Y.; Tran, D.; and Ba, J. 2020. BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135: 364–381.

Xie, Y.; Lu, H.; Yan, J.; Yang, X.; Tomizuka, M.; and Zhan, W. 2023. Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23715–23724.

Zhang, L.; Gao, G.; and Zhang, H. 2023. Spatial-Temporal Federated Learning for Lifelong Person Re-identification on Distributed Edges. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; and Ling, H. 2021. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7380–7399.