1-2019

The challenge of collaborative IoT-based inferencing in adversarial settings

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Dulanga Kaveesha Weerakoon WEERAKOON MUDIYANSELAGE
Singapore Management University, dulangaw@smu.edu.sg

Kasthuri JAYARAJAH
Singapore Management University, kasthurij.2014@phdis.smu.edu.sg

The Challenge of Collaborative IoT-Based Inferencing in Adversarial Settings

Archan Misra, Dulanga Weerakoon, Kasthuri Jayarajah
School of Information Systems
Singapore Management University, Singapore
Email: {archanm, dulangaw}@smu.edu.sg, kasthurij.2014@phdis.smu.edu.sg

***Abstract*—In many practical environments, resource-constrained IoT nodes are deployed with varying degrees of redundancy/overlap–i.e., their data streams possess significant spatiotemporal correlation. We posit that collaborative inferencing, whereby individual nodes adjust their inferencing pipelines to incorporate such correlated observations from other nodes, can improve both inferencing accuracy and performance metrics (such as latency and energy overheads). However, such collaborative models are vulnerable to adversarial behavior by one or more nodes, and thus require mechanisms that identify and inoculate against such malicious behavior. We use a dataset of 8 outdoor cameras to (a) demonstrate that such collaborative inferencing can improve people counting accuracy by over 8%, and (b) show how a dynamic reputation mechanism preserves such gains even if some cameras behave maliciously.**

I. Introduction

A variety of physical environments, including smart cities and tactical battlefield networks are increasingly being instrumented with large numbers of resource-constrained sensors and IoT devices (e.g., cameras, microphone arrays & environmental sensors). A rising recent trend involves executing inferencing pipelines (to perform increasingly complex tasks, such as object recognition or target localization), *in-situ* and in *real time*, at such edge nodes. There are two salient features associated with these trends:

- Sensors are often deployed with varying degrees of redundant coverage–e.g., cameras in buildings often have partially overlapping fields of view, implying that their sensed data are implicitly spatiotemporally correlated.
- Inferencing increasingly involves the execution of computationally prohibitive machine learning (ML) pipelines (e.g., CNNs for image-based object detection and RNNs for speech recognition). Executing such deep neural networks (DNNs) gives rise to well-known throughput bottlenecks and prohibitive energy consumption.

At present, each such sensor/IoT node performs its inferencing in *isolation*, utilizing the sensory data that it captures. Any fusion of such inferences is performed at a logically higher layer–e.g., fusing such object detection events from multiple sensors to perform tracking of a target of interest. We have recently been advocating the vision of *Collaborative IoT Inferencing*, where the inferencing pipelines of multiple individual devices do not operate independently, but collaboratively adapt in *real time*, based on features and inferences shared by other "correlated" IoT/sensing devices. For example, we shall see (in Section III) how a video sensor node dynamically modifies its people counting pipeline, which combines ML-based people detection with color histogram-based filtering, based on histogram & object coordinates shared by other camera nodes. We strongly believe that such *collaborative ML-based inferencing* will lead to radical improvements in both the *operational efficiency* of the deployed IoT infrastructure and the *dependability/robustness* of the associated inferencing outcomes. On the operational side, such coordination in the inferencing process can minimize unnecessary resource consumption (for example, see [8, 10] for selective activation of sensors in a video monitoring infrastructure). For the inferencing outcomes, such collaboration promotes dependability by overcoming the failure vulnerabilities of individual sensors (e.g., object detection failures due to occlusions in a single camera's FoV).

However, such collaboration has a serious potential pitfall: *it makes individual ML-based pipelines on one device susceptible to inadvertent or malicious errors on other nodes*. For example, in the distributed camera-based sensing infrastructure illustrated in Figure 1, a single camera can deliberately suppress information on detected objects from other collaborating cameras, thereby compromising their ML-based pipelines as well. The goal of this paper is thus two-fold:

- First, demonstrate, via an exemplar, that collaborative, real-time modification of ML-based inferencing pipelines can indeed lead to tangible performance improvement.
- Second, introduce a preliminary approach to tackle the performance degradation that can result from operating in such an adversarial operating environment.

We shall illustrate both of these concepts (the benefits of collaborative ML-inferencing and its vulnerability in adversarial environments) using traces obtained from a multi-camera benchmark dataset [5]. Our aim is to mobilize the attention of the larger IoT/ML community to the problem of making collaborative inferencing dependable and robust in adversarial environments. We emphasize that the video-based "people counting" is used purely as an exemplar: our concepts generalize to different and mixed modalities of sensors, and both deep and shallow ML pipelines.

Key Contributions: We make the following key contributions:

- **Accuracy gains using collaborative multi-camera inferencing:** We describe a model of collaborative inferenc-

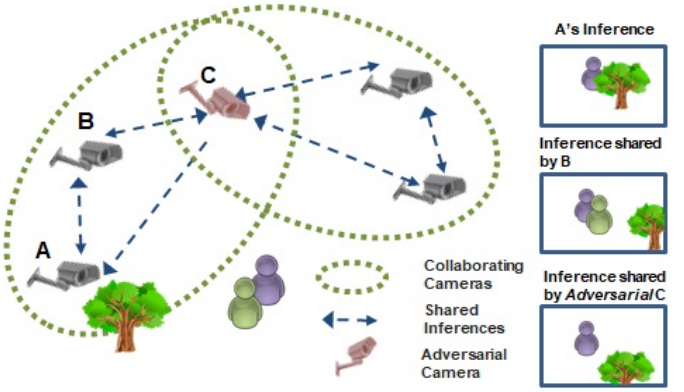


Fig. 1: An Illustrative IoT environment with collaborative cameras.

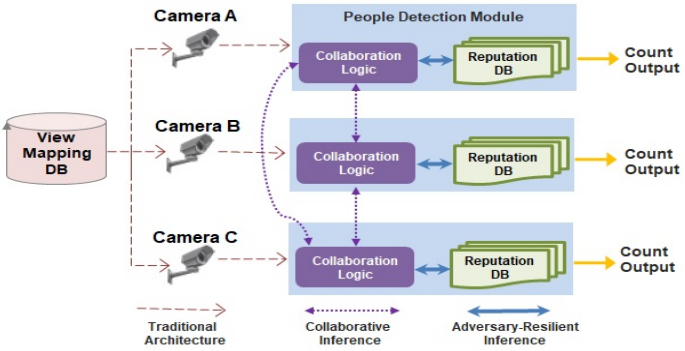


Fig. 2: Overview of the Collaborative (Section III), Adversary-Resilient (Section IV) System Work Flow.

ing for "people counting", appropriate for a networked, multi-camera system system with per-node local processing capabilities. Using the PETS multi-camera benchmark dataset, we show that collaborative inferencing can improve the accuracy, to 75.5%, from a baseline of 68.03% where each camera operates independently. Further, we show that even in the presence of high noise in one or more camera feeds (e.g. perturbing *RGB* values of each image pixel independently), a collaborative system achieves comparable performance gains.

- **Detecting Adversarial/Malicious Nodes:** The accuracy of people counting inference can degrade sharply, if collaborating camera nodes maliciously injects errors–i.e., it deliberately perturbs the histogram values or hides presence of bounding boxes that it shares with neighboring nodes. We propose a measure to capture the reputation of collaborators and using a simulated setting, we show that the measure is able to capture such malicious behavior *fast* – e.g., the reputation score of a malicious camera that lies with 50% probability drops as much by 20% within only $\approx$ 2 minutes of observation (i.e., $\approx$800 frames at 7 FPS), and by 70% when it perturbs reported detections aggressively in addition to hiding detections.
- **Adversary-resilient Collaborative Inferencing:** In this proposed approach, a node continuously updates a reputation score for each neighboring camera based on the observed mutual discrepancy between objects simultaneously identified within an overlapping FoV, and then uses this score to modify its collaborative fusion logic. We show that this approach is able to sustain high inferencing accuracy under such adversarial conditions, achieving F1-score of 73% even with high probability of lying and aggressive perturbations (by a single malicious camera in a 3-collaborator setting) – a 20% improvement over the baseline of a single camera's independent inference.

II. Illustrative IoT environment

To motivate our work, we consider a multi-camera environment illustrated in Figure 1. Each camera has a FoV that has varying degrees of overlap with neighboring cameras–e.g., cameras B and C both observe two individuals and a tree (from different perspectives) concurrently. Each camera also runs an *in-situ* inferencing pipline–for example, performing object detection using a DNN that may be executed on a vision co-processor, such as an Intel^{TM}Movidius device. In addition, for each frame, each camera also shares appropriate features or metadata with nearby/overlapping cameras. It is this *spatiotemporally correlated* information that each camera uses to execute a modified, *collaborative* inferencing pipeline. Such collaboration is likely to offer improved inferencing in various situations–e.g., if one or more objects are occluded in one camera's view (e.g., Camera A is occluded by the presence of the tree and misses observing the "green" person) but visible in the FoV of another camera (e.g., Camera B). The figure also illustrates an adversarial camera, C, which shares misleading or incorrect features/metadata–e.g., it informs cameras A, B, D and E that it can only observe the "purple" individual, deliberately omitting the other ("green") person.

In an adversary-resilient collaborative setting, Camera A who shares/receives inferences from its neighbors B and C, learns from past observations that Camera C is *adversarial* and that B is *trust-worthy*, and makes the correct inference that there are in fact 2 persons in the current frame (although its own view is occluded) by combining its own and B's inferences and disregarding C.

III. Multi-Camera Collaborative Inferencing

We illustrate the work flow of a trust-aware collaborative camera system in Figure 2, which consists of the following two key steps (we defer the discussion of adversary-resiliency to the next section).

Step 1: Mapping between camera views. In the calibration stage, the coordinate mappings between a reference camera's view and its collaborators' views are generated via homography transformation [7]. Such a mapping requires matched points, i.e., points in the real world that are present in both images, as input to homography matrix; these can be extracted either manually or automatically, using a feature matching algorithm such as SIFT [13].

Step 2: People detection. In a non-collaborative, baseline method, each camera runs a people detection algorithm

independently. We use a state-of-the-art deep learning-based detector (SSD) [12] for this. An intermediate output of this stage results in a number of "detections" represented by bounding boxes, each with an associated confidence level. In the final step of the deep network, which is non-maximum suppression (NMS), bounding boxes closely located with significant overlap (computed as the Intersection over Union or IoU $\geq$0.2) are suppressed into a single bounding box, or detection. This output is equivalent to that of a non-collaborative system.

Step 3: Collaborative People Detection The collaborating cameras then send their respective inferences (both before and after the NMS step) to their peer cameras. In this enhanced mode, each camera first established correspondence between its own and each of its collaborators' inferences – the collaborator bounding boxes are transformed to the same coordinate system as the reference camera's using the homographic matrix learned and pairs of bounding boxes are *matched*. To operationalize this, we pose the matching between the two sets of bounding boxes (per frame) as an assignment problem and solve it using the Hungarian algorithm [9] with the cost taken as the distance between the bottom-center coordinates of the bounding boxes. Next, bounding boxes (across cameras) falling close within the same areas are weighed higher in confidence as they are detected by multiple, *trusted* cameras.

IV. Reputation-based Adversary-Resilient Collaboration

The mechanism described above exploits collaboration across the different cameras, but implicitly assumes that the information shared by each camera (the bounding box coordinates and the associated histogram values) are correct. We now extend this collaborative workflow to include a reputation-based mechanism that is resilient to adversarial or malicious behavior.

Step 4: Reputation update. Each camera now maintains a pairwise score of its collaborators' reputation which is based on both (a) whether there exists correspondence between the camera and a collaborator's detections (i.e., whether pairs of matched bounding boxes exist – see Step 3 in Section III) and (b) the content within the matched bounding boxes are similar. If a match is found in (a), then the similarity in content within the boxes (i.e., criteria (b)) is measured as the correlation between their *color histograms*. The reputation score is updated per frame as: $R_{new} = R_{old} + I \times C$, where $I = 1$ if a match is found, and is 0 otherwise, and $C \in [-1, 1]$ represents the correlation value.

If the normalized R_{new} exceeds a specific threshold (T_R), then the reference camera considers this peer camera as a *valid, trustworthy* collaborator for the current frame. If $R_{new} < T_R$, the camera ignores the inputs from this suspicious neighbor. As before, the reference camera combines its own inferences, along with the bounding boxes from its set of trustworthy collaborators, before executing the NMS step.

V. Evaluation

A. Experiment Setting

We use the PETS 2009 dataset [5] which consists of video feeds of 8 synchronous cameras in the outdoors, under varying crowd flow and density settings. The individual cameras record video at an approximate frame rate of 7 FPS and we consider 4 views (views 5-8) with considerable overlap (shown in Figure 3) in our evaluations. The resolution was fixed at 720 $\times$ 576. We processed a total of 3180 frames (i.e., 795 per camera) and consider the camera pertaining to View 005 as the *reference* camera with respect to which we report all our performance results. In this initial effort, we report results for the people detection task. As such, we use the manually annotated ground truth from [18] which provides 2D annotations of 10 persons entering, passing through, staying and exiting the pictured area. The annotations provide 2D bounding boxes for each view and the IDs of persons are consistent across the different views.

We build on the Single Shot Detector (SSD)[1], proposed by [12] for object detection with model trained on the PASCAL VOC dataset [4] and focus only on the "person" object detections. Unless otherwise stated, we run our evaluations at input resolutions of 300 $\times$ 300 (*full* model) and 100 $\times$ 100 (*compressed* model).

B. Performance Metrics

We consider two accuracy metrics, Multiple Objects Detection Accuracy (MODA) [1] and $F-$score (in its usual meaning). For a single camera, given N frames, we denote the set of ground truth bounding boxes by G where G_n^i represents the bounding box of the i^{th} object in the n^{th} frame. Similarly, D_n^i represents the *matched* bounding box (see Section III) of the i^{th} detection of the system in the n^{th} frame. We compute the overlap between every i as the intersection over union (IoU) of the pixels of the two bounding boxes G_n^i and D_n^i . Any IoU less than 0.2 is considered a poor match and is discarded, and the resulting set of matches is denoted by $M_n^{matched}$. Comparing $M_n^{matched}$ and G_n, the $MODA$ accounts for the missed detections and false alarms – for a given frame n, if the number of missed detections (false negatives) are fnn and the number of false alarms are fp_n, then $MODA = 1 - (\alpha \times fn_n + \beta \times fp_n)/|G_n|$. α and β are weights to balance the importance between false negatives and false positives. In this work, we set them to 0.5. We report the average $MODA$ and $F-$score over all N frames, in the next subsection.

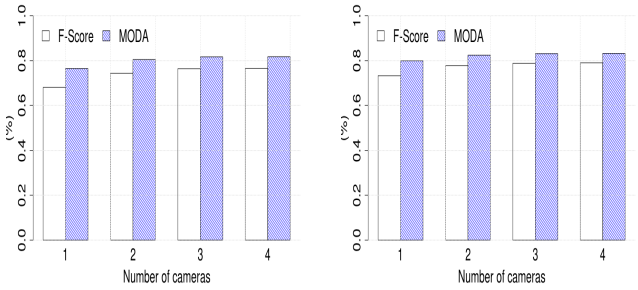
C. Preliminary Findings

Accuracy Improvement with Multiple Collaborating Cameras: In Figure 4, we plot the $MODA$ and $F-score$ on the $y-$axis for (1) a Baseline setting (no. of cameras=1), based solely on the self-inference of the reference camera vs. (2) our proposed collaborative system, combining inferences from 1, 2, and 3 more cooperating cameras ($x-$axis). As anticipated,

[1]Implementation available from https://github.com/weiliu89/caffe/tree/ssd

(a) Reference camera view (b) Collaborative camera 1 (c) Collaborative camera 2 (d) Collaborative camera 3

Fig. 3: Illustrative images from the PETS 2009 benchmarking dataset used in this work.



(a) *Full* model (b) *Compressed* model

Fig. 4: Accuracy of people detection with independent vs. various collaborative camera settings.

we note that the addition of more cameras improves the overall performance (F-Score increasing by 8%, to 75.5% from a baseline of $\sim$ 68%) with the marginal improvement diminishing with each additional increment (F-Score for 4 cameras is only 0.4% than that for 3 cameras), for both the *full* (Figure 4a) and *compressed* (Figure 4b) models.

In most practical situations, video feeds are susceptible to noise (e.g., low light conditions depending on the time of the day, occlusion, etc.). To further understand the utility of collaborative inferencing, we simulate *noisy* conditions by systematically injecting estimation noise to one or more collaborators. Specifically, we perturb the detected bounding boxes of each collaborator, with progressively increasing Gaussian noise (with zero mean and variance varied from 4 to 100). In Figure 5, we plot the performance variation against noise (expressed in SNR, on the decreasing $x-$axis) for the *compressed* model with and without a **single collaborator**. We observe that the performance gains sustain and even at high levels of perturbations (e.g., 9 db SNR $\approx$ 100 variance), the combined inference with even a single collaborator performs better (4% gain in F-score).

Detection of adversarial cameras: Next, we investigate the ability of our reputation score in detecting adversarial cameras in the presence of noise. We simulate an *adversarial* camera which randomly chooses to *hide* a detection (or the corresponding bounding box of a person detected) with a probability, p, and perturbs the content of the detected box with varying intensities (which is operationalized similar to the simulation of noisy conditions in the previous analysis). In Figure 6, we plot the variation of the reputation score ($y-axis$,

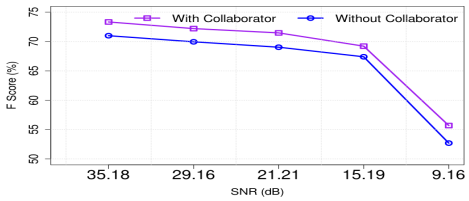
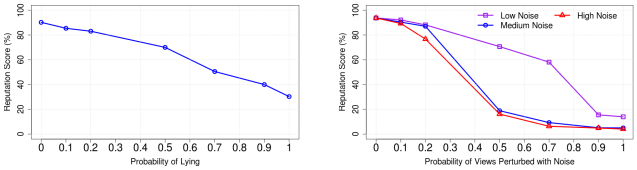


Fig. 5: Impact of noise on people counting accuracy for (a) independent vs. (b) collaborative inference with a single collaborator, with the *compressed* model.

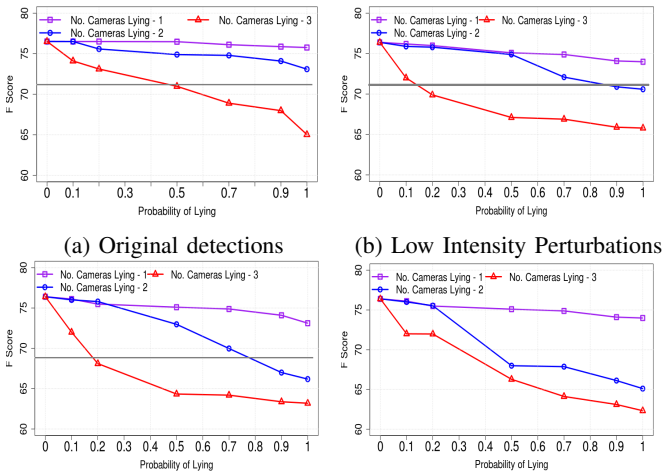


(a) Original detections (b) Perturbed detections

Fig. 6: Variation of the reputation score of a single collaborator with adversarial behavior for (a) randomly *hiding* detections and (b) *hiding* detections and *perturbing* reported detections.

computed at the end of a 2 minute video trace) with increasing value of p, for a single collaborator setting. We repeat the analysis for both cases where the adversary (a) only *hides* detections but does not perturb the content (Figure 6a) and (b) hides as well as perturbs reported detections (Figure 6b) and observe that: (a) the reputation score of a malicious camera shows a sharper decrease with an increased likelihood of lying- e.g., it drops over 20% when $p = 0.5$, and (b) the drop in reputation is significantly larger (e.g., 70% drop for $p = 0.5$) when the adversary perturbs the content with medium–to–high intensity (i.e., 25-100 variance). *Our results thus demonstrate our ability to rapidly isolate and identify a malicious or adversarial camera.*

Resiliency to adversarial cameras: Finally, we evaluate the enhanced reputation-based mechanism (detailed in Section IV), whereby the reference camera ignores metadata from collaborators with reputation scores below a specified T_R threshold. Figure 7 plots the resulting F-Score, as a function of increasing probability of adversarial behavior p, for $T_R = 0.5$ and different levels of perturbation intensity. The baseline accuracy where the reference camera runs its own inference is marked by the solid grey line (as previously seen in Figure 5).



(a) Original detections

(b) Low Intensity Perturbations

(c) Medium Intensity Perturbations

(d) High Intensity Perturbations, Baseline $F-$score 52%

Fig. 7: Performance of independent vs. collaborative inference under differing adversarial conditions.

We make the following key observations:

1) The counting accuracy remains high (F-Score $\approx$ 75%) when an individual camera behaves outright maliciously ($p \rightarrow 1$) in the 4-camera collaboration setting considered, even when the reported detections are highly perturbed (Figure 7d) – the performance gain in this case is $\geq$ 20% (over a baseline of 52% $F-$score observed in Figure 5).
2) As expected, the accuracy drops with an increasing number of adversarial collaborators. Interestingly, even if 2 of the 3 collaborating cameras are adversarial with $p = 0.5$, the accuracy is still quite high (i.e., $\approx$75%, a 4-5% improvement over the independent inference baseline). Of course, when all 3 collaborators are adversarial with $p = 0.5$, the performance is worse by 2% than a non-collaborative baseline (F-Score= 68%).

VI. Related Work

Edge computing technologies are pushing frontiers in enabling real-time analytics systems for situation awareness. Early examples of such systems have been described for various video-based applications and services [15, 16, 17]. Very recently, multi-device cooperation, at the edge, has piqued the interest of the research community (e.g., multi-camera systems [8, 10, 14], cooperative UAV swarms [2, 3], occupant authentication [6]) owing to its advantage of improving accuracy and reducing overheads in dealing with communication with a centralized cloud. As video processing using deep learning pipelines is considered resource-intensive, early efforts in enabling collaboration/cooperation between multi-camera systems explore cost-efficiency without sacrifice in accuracy. Qiu et al.[14] demonstrate the ability to track vehicles across a heterogeneous camera networks consisting of both fixed (e.g., surveillance) and mobile camera. By selectively activating the mobile cameras only to resolve ambiguities whilst much of the heavy-lifting of the video analytics pipeline is performed on the cloud, they achieve high accuracy without overly draining the resource-constrained mobile devices. Further, Lee et al. [10] show that by establishing space-time relationships between views of co-located cameras apriori, significant savings in bandwidth needs can be achieved. They show that by selectively turning off (and on) downstream cameras in the network depending on the moving targets detected by upstream cameras and the respective likelihood of them appearing downstream, the amount of raw footage collected and uploaded to the cloud (for processing) can be reduced as much as by 238 times with a nominal miss rate of 15%. More recently, Jain et al [8] discuss alternative configurations of video analytics pipelines that are triggered by peer cameras that share spatio-temporal correlations between co-located cameras. The authors provide recommendations for cost efficiency (e.g., by reducing redundant processing by cameras sharing overlappnig views) and higher inference accuracy (e.g., cross-camera model refinement).

VII. Discussion

This work, introducing the benefits and challenges of robust collaboration in adversarial environments, needs to be extended to tackle a variety of open issues.

A. Current Limitations

Homography-induced Errors: Our current scheme relies on homographic matching, performed on the 2-D image frames across cameras. In our current evaluation, we ignore estimation errors that arise from such 2-D matching of real-world 3-D coordinates. We will have to enhance the inferencing model, as well as the reputation update mechanism, to explicitly account for such location-dependent homographic matching errors.

Adaptive Reputation Threshold: Our current results are based on a fixed reputation threshold: a camera incorporates the object detection estimates from a peer camera only if its reputation is $\geq T_R (= 0.5)$. In practical deployments, we anticipate the use of a more dynamic threshold, where the right choice of T_R might depend on a variety of deployment factors (e.g., differences in camera fps rates, differences in fraction of overlapping views) as well as contextual conditions (varying crowd density patterns).

Additional Testbeds & Features: Our proposed framework needs to be evaluated and refined under additional settings. We are currently in the process of setting up a 20-30 node distributed camera deployment across 2 buildings on our campus, to help establish performance benchmarks under more-crowded, indoor settings. In addition, our current approach of using histograms may be inadequate in school campuses, where everyone is wearing similar uniforms, and we may need additional features (e.g., observed motion vectors [11]) for more accurate cross-camera matching.

B. Broader Future Work

Our work also needs to explore additional open issues related to the broader problem of multi-device collaborative IoT inferencing.

Autonomic Identification of Collaborative Devices: The experiments presented here involved a small set of IoT devices (cameras) that were set up *a priori* to perform collaborative inferencing. In real-world environments, the set of IoT devices may change dynamically, and the ideal set of collaborating partners may change as well.Accordingly, we will need to develop frameworks that allow one or more IoT devices to first identify the set of devices that can benefit from such collaborative inferencing.

Scalability and Performance Efficiency: We currently evaluated collaboration among a maximum of 4 cameras. However, significant innovations are needed to develop a framework that both scales as the number of individual nodes increases (e.g., when hundreds of cameras are deployed on a university campus) and that is able to perform such adversary identification and inferencing adaptation with low processing overhead. In particular, to enable the adoption of a distributed reputation framework, we are contemplating the use of lightweight cryptographic techniques that allow an individual camera's operational features (e.g., the histograms of the objects that it detects) to be shared in a tamper-proof fashion across multiple nodes without compromising real time processing of video streams.

VIII. Conclusion

In this work, using multiple cameras as an exemplar, we have introduced the notion of real-time collaboration for ML-based inferencing among resource-limited IoT devices. Such collaboration provides several benefits, such as improved accuracy and greater tolerance of noise on individual devices. However, the drawback of such collaboration is greater susceptibility to inadvertent or deliberate failures or false information injected by erroneous or malicious nodes. Through empirical results on the PETS dataset, we show that such adversarial operation can cause the accuracy of camera-based people counting to degrade appreciably (by more than 20%), and then demonstrate that a feature-based dynamic reputation mechanism is resilient to such adversarial attacks. We anticipate that our work will seed greater interest in the community on developing ML-based mechanisms, for both *training* and *inferencing*, that take advantage of the spatiotemporal correlations among different nodes of uncertain fidelity.

Acknowledgment

This material is based on research sponsored in part by the U.S. Army International Technology Center Pacific (ITC-PAC), under Contract No. FA5209-17-C-0006, and partially supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[1] Keni Bernardin, Alexander Elbs, and Rainer Stiefelhagen. 2006. Multiple object tracking performance metrics and evaluation in a smart room environment. In *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*, Vol. 90. Citeseer, 91.

[2] Axel Bürkle. 2009. Collaborating miniature drones for surveillance and reconnaissance. In *Unmanned/Unattended Sensors and Sensor Networks VI*, Vol. 7480. International Society for Optics and Photonics, 74800H.

[3] Xinlei Chen, Aveek Purohit, Carlos Ruiz Dominguez, Stefano Carpin, and Pei Zhang. 2015. DrunkWalk: Collaborative and Adaptive Planning for Navigation of Micro-Aerial Sensor Swarms. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys '15)*.

[4] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vision* 111, 1 (2015).

[5] James Ferryman and Ali Shahrokni. 2009. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE.

[6] Jun Han, Shijia Pan, Manal Kumar Sinha, Hae Young Noh, Pei Zhang, and Patrick Tague. 2018. Smart Home Occupant Identification via Sensor Fusion Across On-Object Devices. *ACM Trans. Sen. Netw.* 14, 3-4 (Dec. 2018).

[7] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

[8] Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, and Joseph Gonzalez. [n. d.]. Scaling Video Analytics Systems to Large Camera Deployments. In *In Proc. of HotMobile*.

[9] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955).

[10] Jongdeog Lee, Tarek Abdelzaher, Hang Qiu, Ramesh Govindan, Kelvin Marcus, Reginald Hobbs, Niranjan Suri, and Will Dron. 2018. On tracking realistic targets in a megacity with contested air and spectrum access. *MILCOM* (2018).

[11] Hanchuan Li, Peijin Zhang, Samer Al Moubayed, Shwetak N. Patel, and Alanson P. Sample. 2016. ID-Match: A Hybrid Computer Vision and RFID System for Recognizing Individuals in Groups. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*.

[13] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. Ieee.

[14] Hang Qiu, Xiaochen Liu, Swati Rallapalli, Archith J Bency, Kevin Chan, Rahul Urgaonkar, BS Manjunath, and Ramesh Govindan. 2018. Kestrel: Video Analytics for Augmented Multi-Camera Vehicle Tracking. In *Internet-of-Things Design and Implementation (IoTDI), 2018 IEEE/ACM Third International Conference on*. IEEE, 48–59.

[15] Mahadev Satyanarayanan. 2017. Edge computing for situational awareness. In *Local and Metropolitan Area Networks (LANMAN), 2017 IEEE International Symposium on*. IEEE, 1–6.

[16] Mahadev Satyanarayanan, Zhuo Chen, Kiryong Ha, Wenlu Hu, Wolfgang Richter, and Padmanabhan Pillai. 2014. Cloudlets: at the leading edge of mobile-cloud convergence. In *2014 6th International Conference on Mobile Computing, Applications and Services (MobiCASE)*. IEEE, 1–9.

[17] Pieter Simoens, Yu Xiao, Padmanabhan Pillai, Zhuo Chen, Kiryong Ha, and Mahadev Satyanarayanan. 2013. Scalable crowd-sourcing of video from mobile devices. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 139–152.

[18] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Song-Chun Zhu. 2017. Cross-View People Tracking by Scene-Centered Spatio-Temporal Parsing.. In *AAAI*. 4299–4305.