Author's version — provided for personal and academic use. Do not redistribute.

Early Recall, Late Precision: Multi-Robot Semantic Object Mapping under Operational Constraints in Perceptually-Degraded Environments

Xianmei Lei[1]*, Taeyeon Kim[2]*, Nicolas Marchal[1]*, Daniel Pastor[1], Barry Ridge[1], Frederik Schöller[1]
Edward Terry[1], Fernando Chavez[1], Thomas Touma[1], Kyohei Otsu[1], Benjamin Morrell[1] and Ali Agha[1]

***Abstract*— Semantic object mapping in uncertain, perceptually degraded environments during long-range multi-robot autonomous exploration tasks such as search-and-rescue is important and challenging. During such missions, high recall is desirable to avoid missing true target objects and high precision is also critical to avoid wasting valuable operational time on false positives. Given recent advancements in visual perception algorithms, the former is largely solvable autonomously, but the latter is difficult to address without the supervision of a human operator. However, operational constraints such as mission time, computational requirements and mesh network bandwidth can make the operator's task infeasible unless properly managed. We propose the Early Recall, Late Precision (EaRLaP) semantic object mapping pipeline to solve this problem. EaRLaP was used by Team CoSTAR in DARPA Subterranean Challenge, where it successfully detected all the artifacts encountered by the team of robots. We will discuss these results and the performance of the EaRLaP on various datasets.**

I. INTRODUCTION

Multi-robot systems with multi-sensor payloads have facilitated a breadth of applications in recent years, ranging from search-and-rescue operations in which such systems are tasked with autonomously navigating harsh environments to find survivors [1], to potential unmanned exploration of subterranean environments of planets, asteroids and other bodies in our solar system and beyond [2]. These developments have been fueled in part by a significant increase in processor efficiency, allowing for advanced neural network architectures and other complex algorithms to be run in real-time aboard robots with significant size, weight and power (SWaP) limitations [3]. In tandem with such advances, the field is also progressing via the diversification of the underlying mobility platforms beyond traditional wheeled systems and towards solutions that are more adapted to certain types of environments. These include relatively fast-paced legged robots traversing difficult and unknown terrain [4].

Thanks to these developments, a critical point has been reached where it is now possible to run modern, highly capable algorithms for object detection and localization, terrain navigation, hazard avoidance and other purposes on a variety of modestly-sized, agile and low-powered robots, with the goal of semantically mapping harsh environments. However, while such ideas are compelling in theory, they can be hampered in practice by a myriad of challenges when the robots are deployed outside of laboratory settings in real-world scenarios. One such challenge is that even state-of-the-art visual detection and localization algorithms can suffer significant performance drops when faced with uncontrolled scenarios in which perceptual degradation from motion blur, shifting luminosity, sensor failure, occlusion and other ocular hazards, are the rule rather than the exception. Another significant challenge is posed by communications constraints. The low-bandwidth wireless mesh networks that are typically employed by such multi-robot systems demand that limits be placed on the size and frequency of visual observation reports sent by robots to an operational base station.

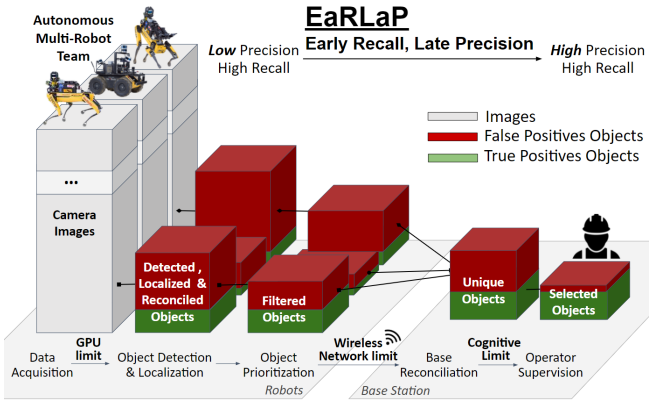

Fig. 1: Early Recall, Late Precision (EaRLaP) semantic object mapping pipeline. Objects detected with high recall in robot-gathered images are localized, then false positives are pruned in multiple stages to respect network constraints before transmission to a base station. After reconciliation of unique objects, an operator reviews detections to achieve high precision.

Depending on the particular application, perceptual degradation may not always be such a concern; however, in tasks involving object detection and localization in which relatively high rates of precision and recall are required, severe performance gaps can emerge. In the search-and-rescue example, it is desirable to both always detect real survivors when they are encountered (high recall) such that all survivors are rescued, and to only report detections that genuinely are survivors (high precision) to avoid hampering rescue efforts. While it is possible to trade off precision for recall, or vice versa, [5] it can be exceptionally difficult to achieve high

This research was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

[1]All authors are with Jet Propulsion Laboratory, California Institute of Technology, United State.

[2]T. Kim is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, South Korea.

*Equal contribution. Corresponding Authors: xianmei.lei@jpl.nasa.gov, taeyeon.k@kaist.ac.kr, marchalnico@gmail.com

©2022 California Institute of Technology. All rights reserved.

Author's version — provided for personal and academic use. Do not redistribute.

rates of both in perceptually degraded circumstances. One possible way to bridge this performance gap is to leverage a visual detection algorithm to achieve high recall and then introduce the supervision of a human operator to make up the precision deficit by identifying true survivors from the detections reported by the algorithms. The downside to this approach is that when the recall is high but the precision is low, the ratio of false positives to true positives that are initially reported might be extremely high and this has the potential to overwhelm both the communications network and the cognitive capacity of the operator in time-critical scenarios.

We present the *Early Recall, Late Precision* (EaRLaP) semantic object mapping pipeline that attempts to optimize human-in-the-loop multi-robot object detection and localization performance under various operational constraints. EaRLaP, illustrated in Fig. 1, initially maximizes recall to ensure that no objects are missed and then siphons a number of detections through a series of filtering stages that gradually prune false positives and compresses information to respect various constraints and achieve high precision. EaRLaP is resilient against uncertainties encountered in the real world (sensor failures, unknown environments, harsh conditions, etc.) and has been tested in extremely challenging situations by the *NASA Jet Propulsion Laboratory (JPL)*. EaRLaP has been integrated as a critical part of the *Networked Belief-aware Perceptual Autonomy (NeBula)* package [6][7] and was used by the JPL team CoSTAR during the final event of the DARPA SubT challenge.

The DARPA Subterranean Challenge (SubT) was a 3-year competition organized by DARPA, that sought novel approaches to rapidly map, navigate, and search underground environments [8]. The goal was to find as many artifacts as possible within a set time limit, and only a very limited number of submissions could be made to the scoring system for evaluation. The artifacts consisted of a fixed set of objects associated with search-and-rescue scenarios, such as backpacks, ropes and helmets, as well as other entities that might not fit the usual object definition, such as RF signals and CO2 gas sources. A critical challenge laid in managing the quantity and quality of the information presented to the operator.

Contribution: EaRLaP is an extension of our previous publication [9] with several new contributions. Firstly, we provide a formalized description of the problem that was previously only informally described in [9] and use this as a basis for defining our proposed solution. Additionally, we introduce new steps in the pipeline to solve the precision issues discussed in [9] and update the existing stages with an enhanced hardware setup and new algorithms to address the reliability issues discussed in [9]. Finally, this work provides a quantitative analysis of our performance during the final event of the DARPA SubT Challenge, where we have demonstrated some of the strongest semantic object mapping performance.

In the remainder, we discuss related work in Sec. II, define the problem formulation in Sec. III, detail the methodology of the EaRLaP approach in Sec. V, and describe results from real in-field experiments and SubT final event competition runs in Sec. VI, before concluding in Sec. VII.

II. RELATED WORK

Perceptually-Resilient CNNs: In perceptually degraded environments, convolutional neural networks (CNNs) suffer from reduced performance due to significant differences between training data and low-quality test images. Combining thermal and visual images can enable a system to be robust to low light conditions and obscurant-filled settings (e.g. dust filled) [10]; however, thermal signals are not distinct after objects have remained in an environment for an extended period. To improve the robustness of a detector that operates in a setting which differs from its training data, [11] tried domain adaptation and [12] used normalizing flows. To overcome challenges due to poor image quality, [13] trained point cloud detectors using adversarial training. These methods are computationally costly and so not compatible with the low latency requirement for our detection system.
Object Localization: To localize an object, a robot must first estimate its own position, which can be done with SLAM if no prior map of the environment is available [14], [15]. The robot then estimates the relative position of the object, which is most commonly done by using a sensor that provides both color and depth information [9]. Stereo cameras and depth sensors such as the Intel® RealSense™ series are popular choices for estimating relative object range from depth images, but they typically only provide accurate depth estimates up to a relatively short maximum distance. An alternative to using depth sensors is to either create a depth map with monocular depth neural networks [16] or to directly predict an object's range with a specialized network [16], however, such monocular depth models do not generalize well to unknown environments. Another approach that can avoid these issues, which we describe in this paper, is to exploit the high accuracy and range of LiDAR sensors as well as good inter-sensor extrinsic calibration and map LiDAR range estimates into RGB camera detection images.

DARPA SubT Challenge: Different solutions for semantic object detection and localization in degraded conditions have been proposed for the DARPA SubT Challenge, where both object detection and communication bandwidth limits are challenging. Team MARBLE [17] used the YOLOv3 architecture [18] for object detection and handled limited bandwidth by reporting each robot's detection results once within range of the communications network. Team CERBERUS [19] used a combination of YOLOv3 and manual confirmation for object detection. One downside of the YOLOv3 model however compared to the pruned YOLOv5m6 model we employed in our pipeline, is that it is not sufficiently lightweight to provide high detection rates for fast traversal. Team CSIRO [20] used the DeNet [21] object detector and tracked and matched object reports within a temporal window in a similar way to what we implemented in our pipeline. Team EXPLORER [22] combined simulation data with real-world data in order to increase training data volume

Author's version — provided for personal and academic use. Do not redistribute.

for their detection models. Our approach, by comparison, solely employed real-world data but emphasized large-scale data gathering in a wide variety of environments which, although costly, can potentially provide more accurate models that generalize more effectively.

III. PROBLEM DESCRIPTION

In this section, we describe a mathematical formulation of the semantic object mapping problem and illustrate how to define the problem as a theoretical optimization problem. We then describe why such optimization is infeasible in closed-form and illustrate the approximated solution constructed by breaking the problem into a series of constrained sub-problems solved by the proposed EaRLaP approach.

A. Semantic Object Mapping Problem (SOMP)

The goal of the semantic object mapping problem is to detect and localize a set of objects using a mobile robot autonomously navigating an unknown environment. The problem is defined using a list of G ground truth object tuples $\mathbf{g} = [\{\mathbf{p}_i^g, l_i^g\}_{i=1}^G]$, named the ground truth set. Given an element of the list $\{\mathbf{p}_i^g, l_i^g\}$, $\mathbf{p}_i^g \in \mathbb{R}^3$ is the 3D position of the ground truth object i and $l_i \in [0, \dots, L]$ is its semantic label, selected from a set of L possible labels. The desired output of the proposed semantic object mapping algorithm is a similar list of S tuples $\mathbf{s} = [\{\mathbf{p}_i, l_i\}_{i=1}^S]$ called the submission set. Each submitted position is compared with the closest ground truth position of the same label to compute the object localization error:

$$e_i(\mathbf{g}, s_i) = \min_{\mathbf{p}_j^g \in \mathbf{g}} |\mathbf{p}_i - \mathbf{p}_j^g|, \quad s.t., \quad l_i = l_j. \tag{1}$$

If an object is correctly classified and its localization error is less than the acceptable limit $e_{\max}$, the object is considered detected and a positive reward $\mathsf{R}_i(e_i(\mathbf{g}, s_i)) = U(e_i \leq e_{\max})$ is received, where $U(\cdot)$ returns 1 when input is true and 0 otherwise. The total reward for the whole submission set is $\mathsf{R}(\mathbf{g}, \mathbf{s}) = \sum_{i=1}^S \mathsf{R}_i(e_i(\mathbf{g}, s_i))$.

The general semantic object mapping problem considered is defined as follows - given S_{max} as the maximum allowed number of submissions, find the best submission set that maximizes the detection reward:

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} \mathsf{R}(\mathbf{g}, \mathbf{s}), \quad s.t., \ |\mathbf{s}| \leq S_{\max}. \tag{2}$$

B. Multi-Robot SOMP

Let's assume that there are R robots connected to a single base station via a mesh network, each using a single camera. Each robot receives an RGB-D measurement $I = [\mathbf{I}_{\text{RGB}}, \mathbf{I}_{\text{D}}, \mathbf{m}]$, where $\mathbf{I}_{\text{RGB}}$ is an RGB image, $\mathbf{I}_D$ is the depth image obtained by a depth sensor (e.g., a LiDAR, stereo camera or RGB-D sensor) and $\mathbf{m}$ is metadata associated with the image (such as the robot's position when the image was captured). Each camera takes an image at rate f for T seconds, creating $2fRT$ images (2 images per timestep for each of the R robots). This defines the vector of all images $\mathcal{I}$. Note that it is a vector and not a set as the ordering of the images can be exploited by the object detection algorithm. We define the function f mapping a vector of images $\mathcal{I}$ to the submission set $\mathbf{s}$ parametrized with $\theta \in \mathbb{R}^{N_\theta}$

$$\mathbf{s} = \{p_i, l_i\}_{i=0}^S = f(\mathcal{I}; \theta). \tag{3}$$

The parameters θ could be learned from a dataset of the form $[\mathcal{I}, \mathbf{g}]$ containing image vectors and associated ground truth positions and labels. An agent learning the solution would need to perform the following optimization:

$$\theta^* = \arg\max_{\theta} \mathsf{R}(\mathbf{g}, \mathbf{s}) \quad \text{s.t.} \ \ \mathbf{s} = f(\mathcal{I}; \theta), \quad |\mathbf{s}| \leq S_{\max}. \tag{4}$$

First, the reward function in (4) is discontinuous and the input space is very high dimensional, which makes it difficult to optimize. Additionally, in practice, and within the confines of the DARPA SubT challenge in particular, the problem is subject to a number of additional constraints and nuances that make the direct optimization of θ infeasible. Thus, in the following Sec. IV, we describe these additional constraints as well as how the general problem can be divided into tractable sub-problems that can be solved by our proposed EaRLaP pipeline approach.

IV. CONSTRAINED SUB-PROBLEM DECOMPOSITION

In the robot-and-base-station scenario, robots must perceive the environment and communicate their findings with a base station and we face trade-offs in deciding how to distribute the computational aspects of this perceptual pipeline between them. Depending on the computational capabilities and network bandwidth, it is necessary to perform some operations on the robot, at the base station, or both. In addition, we exploited the supervision of a human operator to create a final high-confidence submission set $\mathbf{s}$, but the operator can only verify a limited number of detections within the allotted mission time. With these constraints, we decompose the function f previously defined in (3) into three sub-functions - f_r for the robot, f_b for the base station and f_{op} for the operator:

$$\begin{aligned} f(\mathcal{I}; \theta) &= f_{op} \circ f_b \circ f_r(\mathcal{I}), \quad \theta = (\theta_r, \theta_b, \theta_{op}), \\ \mathbf{s} &= f(\mathcal{I}, \theta) = f_{op}\left(f_b\left(f_r\left(\mathcal{I}; \theta_r\right); \theta_b\right); \theta_{op}\right). \end{aligned} \tag{5}$$

In this decomposition, the data $\mathcal{I}$ is first processed on the robot using $f_r(\mathcal{I}, \theta_r) = \mathcal{D}_r$ such that the amount of data $|\mathcal{D}_r|$ being transmitted from the robot to the base station is within the communication bandwidth limit S_b, i.e. $|\mathcal{D}_r| < S_b$. Finding the optimal θ_r thus follows the same manner of SOMP problem (4) using the mathematical formulation:

$$\begin{aligned} \theta_r^* \ &= \arg\max_{\theta_r} \mathsf{R}(\mathbf{g}, \mathcal{D}_r) \\ \text{s.t.} \quad &\mathcal{D}_r = f_r(\mathcal{I}; \theta_r), \quad |\mathcal{D}_r| \leq S_b. \end{aligned} \tag{6}$$

The amount of transmitted data $|\mathcal{D}_r|$ on the base station is assumed to be greater than the maximum number of submissions S_{max}. We further process $\mathcal{D}_r$ on the base station with $f_b(\mathcal{D}_r, \theta_b) = \mathcal{D}_b$ to condense the data into $\mathcal{D}_b$, such that $|\mathcal{D}_b|$ is small enough for a human operator to review within time limit. The maximum size of $\mathcal{D}_b$ is governed

Author's version — provided for personal and academic use. Do not redistribute.

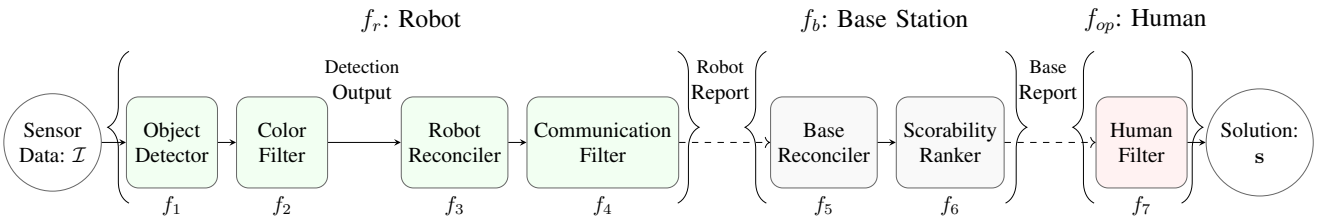

Fig. 2: Hierarchical decomposition of the semantic object mapping problem. f is divided into three sub-functions and these sub-functions are further divided to reach a composition of seven sub-functions, which is the EaRLaP semantic object mapping pipeline.

by the maximum human cognitive load S_{op} or maximum trained human operator cognitive load, which is also assumed to be greater than S_{max}. Once again, finding the θ_b is an optimization problem of the same form as (4).

Finally, the operator is responsible for manually processing $\mathcal{D}_b$ to create the set of submissions $s = f_{op}(\mathcal{D}_b, \theta_{op})$, and for ensuring that $|\mathbf{s}| < S_{max}$. This follows the formulation (4), but f_{op} does not follow a mathematical formulation since θ_{op} represents the operator's human brain process capability. Note that this is still an optimization problem, and the operator undergoes training to optimize the detection performance.

With this high-level decomposition of f in place, we can further decompose the sub-functions f_r, f_b and f_{op} hierarchically into a set of N tractable lower-level sub-functions $f_i(\mathcal{D}_{i-1}; \theta_i)$, each of which permit similar optimization formulations as (6), which are then combined to give a solution to the original problem:

$$f(\mathcal{I}; \theta) = f_N \circ \cdots \circ f_i \circ \cdots \circ f_1(\mathcal{D}_0; \theta_1), \tag{7}$$

where $\theta = (\theta_1, \ldots, \theta_N)$ provides parameterizations for each sub-function, $\mathcal{D}_0 = \mathcal{I}$, and $\mathcal{D}_i = f_i(\mathcal{D}_{i-1}; \theta_i)\ \forall i = 1, \ldots, N$ are outputs after each sub-function. These outputs $\mathcal{D}_i$ contain positions and labels to solve the SOMP but may also include additional data gathered in the previous sub-functions and that will be used in later sub-functions (i.e. images, point clouds, bounding boxes, confidences etc.).

V. EaRLaP Semantic Object Mapping Pipeline

The implementation of (7) that was designed for the EaRLaP pipeline and ultimately deployed in the final events of the DARPA SubT Challenge consisted of $N = 7$ different sub-functions as illustrated in Fig. 2 and detailed below.

f_1 **- Object Detector:** The sensor payloads on the robots contain multiple cameras to cover the entire field of view (FOV), however, this entails that multiple high-rate image streams must be handled by CNN object detectors at a low rate on robot processors with limited computational power. To ensure computational power is used efficiently, an image selection filter samples a subset of images with higher quality. To obtain high-quality detections, in particular for small objects in low lighting environments, high-resolution images [23] must be used. Lightweight models designed for real-time application and high-resolution images, such as the family of YOLO models [24], are suitable for f_1.

Given an RGB image $\mathbf{I}_{\text{RGB}}$, the object detector outputs N_d detections with associated labels, confidences and bounding boxes $\{l, c, \mathbf{B}\}_{i=1}^{N_d} = g_{\text{yolo}}(\mathbf{I}_{\text{RGB}})$. Only the detections with label l and confidence higher than t_l are passed on to the next sub-function.

f_2 **- Color Filter:** After the CNN step, to remove false-positive detections, a function that extracts handcrafted features can be applied if object types contain outstanding features, such as color and shape. These shallow feature filters can be easily trained to reject clear false positives. In the SubT case, the object types and their respective coloring are known and object detection results with unexpected color patterns can be eliminated. To identify the color of an object, the segment of the RGB image $\mathbf{I}_{\text{RGB}}$ inside the detection bounding box $\mathbf{B}$, that is $\mathbf{I}_{\text{RGB}}^{\mathbf{B}}$, is transformed to HSV color space and a mask $\mathbf{M}_l$ is applied which is tailored to the color of each object l. Note that in practice $\mathbf{M}_l$ must accept a wide spectrum of color to account for the different conditions in which we might encounter an object. The masked bounding box is thresholded and the remaining pixels are counted to obtain an object color score percentage p. Each object l has a multiplier $\beta_l > 1$ such that the color score defined by $g_{\text{col}}(l, \mathbf{I}_{\text{RGB}}^{\mathbf{B}}) = \min(\beta_l \cdot p, 1) \in [0, 1]$ returns 1 if even only part of $\mathbf{I}_{\text{RGB}}^{\mathbf{B}}$ is the correct color. To filter detections, the color score is multiplied with the YOLO confidence, and the same YOLO threshold t_l is used to decide if a detection moves on to the next step. β_l is tuned using a representative labeled dataset.

f_3 **- Robot Reconciler:** The global position of a given robot relative to a calibration gate at its starting point is estimated by other system modules [14], [15], however, the detected objects still need to be reconciled and localized relative to the robot. In order to initially eliminate noisy detections, when a detection with label l is received, it is removed if the robot has not moved at least d_{min} meters or rotated α_{min} degrees since the last observation of another object l. The remaining detections must then be reconciled with previously mapped object candidates based on proximity.

Given an observation, those candidates that are not compatible with the bearing of the observation are not considered (a detection from the front camera cannot come from a candidate located behind the robot, for example). Candidates further than a set maximum distance $d_{\max}^{d}$ from the robot are also filtered since detection performance drops significantly beyond that point. If there are multiple candidates after

Author's version — provided for personal and academic use. Do not redistribute.

this initial stage, matching the detection with a candidate close to the robot and the detection direction is favored. After matching the detection to a candidate, the candidate's position and metadata are updated to keep some information about the detection (i.e. its detection confidence and color confidence). If there are no remaining candidates after this process, a new object is created in the map, becoming a new candidate for reconciliation of subsequent object detections.

Through this reconciliation process, we aggregate multiple detections of the same object into what we call a *report*. As we reconcile more observations of the same object instance into a single report, our localization uncertainty decreases and our confidence in this report increases.

The primary means used for robot-relative object position localization are bearing estimates (implemented using the GTSAM library [25]), but the estimates can have high variance, particularly with respect to object range. To refine the range, one could exploit depth images $\mathbf{I}_{\text{D}}$ produced by stereo cameras or RGB-D sensors, however, this can quickly lose accuracy or fail completely at large distances depending on the lighting conditions. Instead, we make use of an alternative approach that transforms LiDAR-derived 3D point clouds into the camera frame in which the object has been detected, re-projects the 3D LiDAR points to 2D image pixels, determines which of those points/pixels lie within the detected object bounding box $\mathbf{B}$, and calculates their mean distance in the camera frame. This approach relies on good LiDAR-camera extrinsic calibration but provides more accurate results at distances outside of the effective range of RGB-D sensors.

Given a calibrated camera and an object detection with associated bounding box $\mathbf{B}$ and estimated range d, we can project $\mathbf{B}$ from 2D to 3D to obtain the estimated object size. If an object of class l has well-defined dimensions (e.g. a backpack of size 30x30x30cm), a size score $g_{\text{size}}(l, \mathbf{B}) \in [0, 1]$ can be created that measures how well $\mathbf{B}$ matches the object size. Similarly to the color score, this must account for a wide range of sizes observed in practice, mainly due to noisy measurements of d and inaccurate $\mathbf{B}$.

f_4 **- Communications Filter:** To limit the quantity of reports sent from the robots to the base station, we only transmit reports in which multiple observations have been reconciled. Additionally, reports for which the median report confidence of all reconciled observations is high are favored.

f_5 **- Base Reconciler:** In practice, the reconciliation of f_3 does not always group the detections perfectly and a robot might send multiple reports of the same object to the base station. Additionally, when multiple robots communicate to the base station, they could report the same objects if they explore the same areas. These reports can be merged into a unique object. A report is grouped into an existing report cluster if it is spatially closer than d^b_{min} to an existing report in that cluster of the same object class. Such distance-based reconciliation assumes that objects of interest are sparsely distributed in the environment. However, false-positive detections can be located close to each other and can be merged into one object during this step. Additionally, when exploring multi-floor buildings, objects on different floors can be reconciled together if d^b_{min} is not chosen wisely.

f_6 **- Scorability Ranker:** In order to mitigate against a time-pressured operator missing true positives while evaluating the reconciled reports on the base station, it is prudent to attempt to rank the reports in terms of their scoring potential. A detection $D = \{l, c, \mathbf{B}\} \in g_{\text{yolo}}(\mathbf{I}_{\text{RGB}})$ produced by running the detector on a given image $\mathbf{I}_{\text{RGB}}$ is assigned a *scorability confidence*

$$g_D(D) = c \cdot g_{\text{col}}(l, \mathbf{I}^{\mathbf{B}}_{\text{RGB}}) \cdot g_{\text{size}}(l, \mathbf{B}). \tag{8}$$

The *scorability of a report* $R = \{D_i\}_{i=1}^{n_d}$ is then computed as

$$g_R(R) = \alpha(n_d) \cdot \frac{1}{n_d} \sum_{i=1}^{n_d} g_D(D_i), \tag{9}$$

where $\alpha(n_d)$ is a function of the number of detections n_d per reconciled report, used to penalize reports with a low number of observations, and the reports are ranked with respect to g_R in descending order. $\alpha(n_d)$ is tuned based on a few factors. The minimum is 2, as two false positives are less likely to be predicted in a row. $\alpha(n_d)$ is further adjusted based on the model's performance on the objects' sizes and the environment's lighting condition. Size score is tuned using a labeled dataset.

f_7 **- Human Filter:** A graphical user interface (GUI) was developed that allows the human operator to control robotic exploration behavior and perform certain actions. Once the detections from f_1 have been reconciled as objects and then ordered using f_6, the operator can review them and manually decide whether to include them in the final submission set $\mathbf{s}$ using the GUI. The operator may also manually make adjustments to the report as necessary in cases of poor reconciliation, etc.

VI. Experimental Results

This section will showcase the results obtained by Team CoSTAR, who implemented EaRLaP to detect and locate objects during the Final Event of the DARPA SubT Challenge.

Although Team CoSTAR detected objects using a wide range of sensors, the EaRLaP methodology presented in Sec. V was motivated primarily by visual detection. This section will discuss the results applied to the following RGB-detectable objects from the competition: backpack, drill, fire extinguisher, helmet, cube, rope and survivor.

Team CoSTAR deployed Husky and Spot mobility platforms: commercially available wheeled vehicles and quadrupeds, respectively. A customized payload that comprised of sensors and computing processors was developed to facilitate high-level autonomy [4]. The payload includes five Intel® RealSense™ D455 cameras, arranged on the robots to provide a near 360° FOV. The RealSense cameras were connected directly to an NVIDIA® Jetson AGX Xavier™ to run custom camera drivers, YOLO object detector (f_1) and the color filtering (f_2). The data was later transmitted to an Intel® NUC computer to perform the reconciliation (f_3) and select the objects to send to the base via wireless communication

Author's version — provided for personal and academic use. Do not redistribute.

(f_4). The robots carried and deployed communication nodes to establish a communication mesh to connect to the base station. The base is a custom-made computer with multiple GPUs and is connected to a monitor displaying the GUI where the operator monitored the mission and interacted with the robots.

Different methods and parameters were selected for each sub-function for the final event of the DARPA SubT Challenge. For the image selector, we used the highest variance of the Laplacian [26] to compute and select the least blurry images in the 25Hz image streaming queue. Team CoSTAR's robots' top speed is 1m/s and we aimed for detections at least once every 25cm. This required the model to run at least 5hz per camera. To achieve this, in f_1, we used a YOLOv5m6 [24] model with input size of 1280x768 and pruned using channel pruning, decreasing the model size by 22.5%. The *SiLU* activations were replaced with *ReLU*, allowing quantization (MinMax strategy) for INT8 inference. Finally, the model was hardware optimized with TensorRT. For communications filter f_4, regardless of the total number of observations reconciled in a report, a maximum of 4 compressed images (brightest, least blurry, highest report confidence, closest) were kept for inspection by the human operator in order to further minimize the amount of transmitted data.

A. Evaluation Criteria

The normative interpretation of precision/recall metrics in a computer vision problem involves, for example, running an object detector on a given image dataset $\mathcal{I} = \{I_i\}_{i=1}^{l}$ to produce a set of detections $\mathcal{D}_\mathcal{I} = \{D_i\}_{i=1}^{m}$ and comparing the D_i to a ground truth set $\mathcal{G}_\mathcal{I} = \{G_i\}_{i=1}^{n}$ to produce true positive (TP), false positive (FP) and false negative (FN) counts such that $TP + FP = m$ and $TP + FN = n$. In this case, n is counted across all images in $\mathcal{I}$. In our scenario, we are not only interested in such *image-based* metrics, but also in *object-based* metrics. To help explain this, we introduce the notion of an environment $\mathcal{E}$ in which the physical ground truth objects $\mathcal{G}_\mathcal{E} = \{G_i\}_{i=1}^{k}$ are positioned and within which all images $I_i \in \mathcal{I}$ are gathered such that $k \leq n$. By reconciling the $D_i \in \mathcal{D}_\mathcal{I}$ detections with respect to the true physical ground truth objects $G_i \in \mathcal{G}_\mathcal{E}$, we can introduce object-based precision/recall metrics such that $TP + FN = k$. This allows us to measure EaRLaP performance when only counting a true positive for each physical ground truth object once and when reconciling detections into object reports across the various pipeline stages.

B. Objectives

When dealing with harsh unknown environments, it is very unlikely for the robots to be able to fully explore them, especially if there are time constraints. One of Team CoSTAR's strategic goals during SubT was, therefore, to aim for the highest possible recall to avoid the operator missing scores for the true positive objects that the robots encountered. Aiming for high recall often comes at the cost of poor precision [5]. Before CoSTAR adopted the EaRLaP pipeline [9], the team (as well as other teams [20]) suffered from low precision. During the 2020 Urban Circuit of the DARPA SubT Challenge, the human operator was unable to identify some true positive objects in a timely manner as they were mixed between hundreds of false positives. Given the many other tasks the operator had to manage (including deploying robots, checking sensor health, setting exploration strategy, etc.), an additional strategic goal was, therefore, to greatly reduce the number of false positives they had to process for scoring submission in order to save them time for these critical tasks and to maximize precision. In this section, we aim to demonstrate using the data from the SubT Final Event and the evaluation criteria in Sec. VI-A, that EaRLaP was indeed effective at achieving these strategic goals.

TABLE I: Number of TP and FP images/reports at the output of different sub-problems.

		Detection Output [1]	Robot Output [2]	Base Output [3]
Preliminary Run (3 robots) (9 true objects)	TP	4574	45	10
	FP	4490	238	50
	time[4]	19h	35min	7.5min
Final Run (3 robots) (4 true objects)	TP	3699	45	4
	FP	3956	260	48
	time[4]	16h	38min	6.5min

See Fig. 2 for the definition of *Detection/Robot/Base Output*

[1] Number of bounding box image detections
[2] Reports where many image detections have been reconciled by the robot
[3] Groups of reports, reconciled on the base station
[4] Estimated time to evaluate each column using an average of 7.5s per image or object report

TABLE II: True Positive (TP) and False Positive (FP) objects at the output of different sub-problems.

		Detection Output	Robot Output	Base Output	Operator Output
Preliminary Run (3 robots) (9 true objects)	TP	9	9	9	8
	FP	278	99	48	0
	Recall	100%	100%	100%	89%
	Precision	3.1%	8.3%	17.24%	100%
Final Run (3 robots) (4 true objects)	TP	4	4	4	4
	FP	479	113	48	0
	Recall	100%	100%	100%	100%
	Precision	0.8%	3.4%	7.7%	100%

See Fig. 2 for the definition of *Detection/Robot/Base Output*. Operator Output is the solution submitted by the operator.

C. Key Results

The competition took place between Sep. 21st–24th, 2021. Over the first three days, competing teams performed in two preliminary runs of 30 minutes each, in which 20 artifacts were hidden. Tables I and II show results from the second preliminary run. Using EaRLaP, Team CoSTAR scored the most artifacts of all teams during this round, finishing at the top of the score table. CoSTAR proceeded into the final run as one of the favorites to win. The run lasted one hour and consisted of 40 artifacts. Unfortunately, in the

Author's version — provided for personal and academic use. Do not redistribute.

final run, our robots encountered challenging conditions and struggled to explore the environment as effectively as in prior runs. Between one Husky and two Spots, only four RGB-detectable objects were encountered. The team utilized more robots with diverse sensory capabilities, thus allowing us to score a total of 13 artifacts and achieving a final rank of 5^{th} place, but data from these robots could not be retrieved.

In Tables I and II, we also show results from some of the most important EaRLaP sub-problems (as illustrated in Fig. 2) using data from both the preliminary and final runs. Table I shows results for the image-based metrics discussed in Sec. VI-A. In order to produce this table, each image or report from each run was manually labeled to be a true positive (TP) or false positive (FP). In the *Detection Output* column we have bounding box detections from our filtered object detector (f_1 + f_2), while in the *Robot Output* column we have reports where multiple detections have been reconciled together on the robots (f_4) and in the *Base Output* column we have groups of reports on the base station (f_6). A human operator could theoretically review the data at any of these sub-problems' outputs (although it might not be a good use of their time) therefore, Table I shows the amount of information that would be exposed to the operator at those stages. This table, however, gives no information on the number of actual objects that the operator would encounter, which is what is truly important for the DARPA SubT challenge and the semantic object mapping problem (i.e. the detector can generate 50 detections of the same backpack).

It is thus necessary to introduce Table II, where we counted the number of unique object instances at the output of the different sub-problems and used the object-based metrics discussed in Sec. VI-A. To create this table, an annotator had to go through all the images or reports (depending on the sub-problem) and keep track of all unique objects already detected to determine whether to add a new object or not. Additionally, the annotator identified the TP as the objects that we are trying to detect, and all the rest was FP. In addition to being directly related to the goal of the DARPA SubT challenge and semantic mapping problem, using object instances also allows us to break the barrier of different data types at the output of different sub-problems. It thus enables us to compare the performance at various stages of the EaRLaP pipeline, with particular attention being paid to the object-based recall and precision.

From Table II, it can be seen that in both the preliminary and final runs we reached our objective since all TP objects encountered by the robots reached the operator. This was possible due to the wide FOV covered by the cameras and the design of our object detector (f_1). During the preliminary run, the operator made an error as they did not recognize a backpack detected at a far distance and in a dark area as a true positive. This may be explained by the fact that this was a practice round and the operator thus was not required to perform at their fullest capacity. This serves as a poignant reminder that although a human's intervention is required to achieve extremely high precision, the operator can also make mistakes. Despite this error, CoSTAR still scored higher than any other team in this particular round.

As expected, the YOLO CNN eliminated a large number of false positives, shown by the *Detection Output* column in Table II. However, the table shows that as we advance through the sub-functions, the precision increases while the recall is kept at 100%. Although each sub-function might only contribute slightly to improving precision, the compounding effect of all sub-problems leads to a 5x increase in precision for the preliminary run and a 9.5x increase for the final run.

When introducing the seven sub-functions in Sec. V, we required that some sub-functions reduce the amount of information such that the detected objects could be transmitted on low-bandwidth wireless networks and displayed to the operator. Table I shows that during the preliminary round the object detector generated more than 9000 detections, but after reconciliation on the robot, all of this information was reduced to less than 300 reports, which is low enough to avoid saturating the communication network. After reconciliation on the base station, this is further reduced to just 60 different objects, which is manageable for a human to manually inspect.

We used an efficient GUI, shown in Fig. 3, which is optimized such that the operator can review all detections rapidly. We estimate that reviewing and accepting or rejecting an object on the GUI takes between 5 and 10 seconds. Table I shows that if a human were to sort through all the detections generated by the neural network, it would take almost 19 hours to do so for the preliminary round, which is only 30 minutes long. Despite a significant data reduction after the robot reconciliation, it would still take a human more than 30 minutes to inspect. However, at the end of our pipeline, we managed to reduce the operator's work duration on this task to 7.5 minutes, which is 25% of the total run time. Therefore, we successfully achieved our goal. Additionally, thanks to the report scorability ranker (f_6), the nine true positive objects are among the first objects displayed to the operator.

VII. CONCLUSION

This paper studied the task of semantic object mapping using human-in-the-loop multi-robot systems and defined a mathematical formulation of the problem. The proposed EaRLaP pipeline decomposes the problem into sub-problems to achieve high precision and high recall under computational, communication network, and human cognitive load constraints, and ultimately to provide high-quality information for human operators. EaRLaP was implemented by being decomposed into seven sub-functions and deployed by Team CoSTAR in the final events of the DARPA SubT Challenge to perform search-and-rescue in perceptually-degraded environments. The performance showed that EaRLaP maintained high recall under the constraints and pruned false positives effectively to allow the human operator to achieve high precision within a time limit. This paper is also a general report of the competition. In future work, we would like to perform ablation studies to evaluate the comparative

Author's version — provided for personal and academic use. Do not redistribute.

Fig. 3: An example of the GUI in the base station shows a candidate submission that contains three artifacts reported from different robots at different times. a) A highlighted report. b) Reports are grouped based on high, median, and low report confidence. c) Reports are ordered by report confidence d) report confidence value and artifact label e) more information on YOLO confidence and color/size scores. f) 3D position of the selected report, where the operator can adjust it before submission.

performance of EaRLaP both with and without the various sub-functions described in Sec. V, as well as performing further field experiments in other scenarios.

References

[1] A. C. Morris, S. Thayer, J. Kuffner, and M. B. Dias, "Robotic Introspection for Exploration and Mapping of Subterranean Environments," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2007.

[2] A. Agha, K. L. Mitchell, and P. J. Boston, "Robotic Exploration of Planetary Subsurface Voids in Search for Life," in *AGU Fall Meeting Abstracts*, vol. 2019, Dec. 2019, pp. P41C–3463. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019AGUFM.P41C3463A

[3] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and Benchmarking of Machine Learning Accelerators," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2019, pp. 1–9.

[4] A. Bouman, M. F. Ginting, N. Alatur, M. Palieri, D. D. Fan, T. Touma, T. Pailevanian, S.-K. Kim, K. Otsu, J. W. Burdick, and A. akbar Aghamohammadi, "Autonomous spot: Long-range autonomous exploration of extreme environments with legged locomotion," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2518–2525, 2020.

[5] M. Buckland and F. Gey, "The relationship between recall and precision," *Journal of the American society for information science*, vol. 45, no. 1, pp. 12–19, 1994.

[6] A. Agha, K. Otsu, B. Morrell, D. Fan, R. Thakker, A. Santamaria-Navarro, S.-K. Kim *et al.*, "Nebula: Team costar's robotic autonomy solution that won phase ii of darpa subterranean challenge," *Journal of Field Robotics*, 2021.

[7] A. Agha, K. Otsu, B. Morrell, D. Fan, S.-K. Kim, M. Ginting, X. Lei, J. Edlund *et al.*, "An addendum to nebula: Towards extending team costar's solution to larger scale environments," *arXiv preprint arXiv:2103.11470*, 2021.

[8] T. Rouček, M. Pecka, P. Čížek, T. Petříček, J. Bayer, V. Šalanskỳ, D. Heřt, M. Petrlík, T. Báča, V. Spurnỳ *et al.*, "Darpa subterranean challenge: Multi-robotic exploration of underground environments," in *International Conference on Modelling and Simulation for Autonomous Systems*. Springer, 2019, pp. 274–290.

[9] E. Terry, X. Lei, B. Morrell, S. Daftry, and A.-a. Agha-mohammadi, in *Robotics: Science and Systems (RSS) Workshop*, 2020.

[10] M. Tsiourva and C. Papachristos, "Multi-modal visual-thermal saliency-based object detection in visually-degraded environments," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–9.

[11] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. Macready, "A robust learning approach to domain adaptive object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 480–490.

[12] N. Marchal, C. Moraldo, H. Blum, R. Siegwart, C. Cadena, and A. Gawel, "Learning densities in feature space for reliable segmentation of indoor scenes," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1032–1038, 2020.

[13] R. DeBortoli, L. Fuxin, A. Kapoor, and G. A. Hollinger, "Adversarial training on point clouds for sim-to-real 3d object detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6662–6669, 2021.

[14] K. Ebadi, Y. Chang, M. Palieri, A. Stephens, A. Hatteland, E. Heiden, A. Thakur, N. Funabiki, B. Morrell, S. Wood *et al.*, "Lamp: Large-scale autonomous mapping and positioning for exploration of perceptually-degraded subterranean environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 80–86.

[15] M. Palieri, B. Morrell, A. Thakur, K. Ebadi, J. Nash, L. Carlone, C. Guaragnella, and A. Aga-mohammadi, "Locus-lidar odometry for consistent operation in uncertain settings," *IEEE Robotics and Automation Letters*, 2020.

[16] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[17] M. Ohradzansky, E. Rush, D. Riley, A. Mills, S. Ahmad, S. McGuire, H. Biggie, K. Harlow, M. Miles, E. Frew *et al.*, "Multi-agent autonomy: Advancements and challenges in subterranean exploration," *Journal of Field Robotics. Under revision*, 2021.

[18] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[19] M. Tranzatto, F. Mascarich, L. Bernreiter, C. Godinho, M. Camurri, S. M. K. Khattak, T. Dang, V. Reijgwart, J. Loeje, D. Wisth *et al.*, "CERBERUS: Autonomous legged and aerial robotic exploration in the tunnel and urban circuits of the DARPA subterranean challenge," *Journal of Field Robotics*, 2021.

[20] N. Hudson, F. Talbot, M. Cox, J. Williams, T. Hines, A. Pitt, B. Wood, D. Frousheger, K. L. Surdo, T. Molnar *et al.*, "Heterogeneous ground and air platforms, homogeneous sensing: Team CSIRO data61's approach to the DARPA subterranean challenge," *arXiv preprint arXiv:2104.09053*, 2021.

[21] L. Tychsen-Smith and L. Petersson, "Denet: Scalable real-time object detection with directed sparse sampling," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 428–436.

[22] S. Scherer, V. Agrawal, G. Best, C. Cao, K. Cujic, R. Darnley, R. DeBortoli, E. Dexheimer, B. Drozd, R. Garg *et al.*, "Resilient and modular subterranean exploration with a team of roving and flying robots," *Submitted to the Journal of Field Robotics*, vol. 2, no. 3, p. 6, 2021.

[23] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 566–583.

[24] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6222936

[25] F. Dellaert, "Factor Graphs and GTSAM: A Hands-on Introduction," Georgia Institute of Technology, Technical Report, Sep. 2012.

[26] R. Bansal, G. Raj, and T. Choudhury, "Blur image detection using laplacian operator and open-cv," in *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*. IEEE, 2016, pp. 63–67.