AFRL-RI-RS-TR-2020-203

SELF-DIRECTED LIFELONG VISUAL LEARNING

UNIVERSITY OF MASSACHUSETTS, AMHERST

NOVEMBER 2020

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2020-203 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
TOM RENZ
Work Unit Manager

/ S /
GREGORY HADYNSKI
Assistant Tech Advisor, Computing & Communications Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
NOVEMBER 2020	FINAL TECHNICAL REPORT	MAY 2018 - MAY 2020

4. TITLE AND SUBTITLE

Self-Directed Lifelong Visual Learning

5a. CONTRACT NUMBER
FA8750-18-2-0126

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)

Eric Learned-Miller

5d. PROJECT NUMBER
LLM2

5e. TASK NUMBER
MA

5f. WORK UNIT NUMBER
SS

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
c/o Office of Grant and Contract Administration
Venture Way Center, 100 Venture Way, Suite 201
Amherst MA 01003

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RITB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2020-203

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Our work under Phase I of the DARPA Lifelong Learning Machines (L2M) Program has concentrated on solving lifelong learning problems in the domain of computer vision and machine learning. Our work has focused on solving the general task of "Intelligent Search", which is the combined visual and navigation task of finding an object, described either by name or with a picture, in a new or unknown environment. There are many facets of this problem, including the reasoning of where an object is likely to be based on a visual assessment of the current scene, the ability to navigate quickly and efficiently through unknown environments, the inference of 3D structure from images to aid navigation, and learning policies for efficient navigation. We report on progress for all four of these goals. In addition to these four areas of research being directly related to the larger goal of Intelligent Search, they are all central topics in Lifelong Learning. In each area, the agent is designed so that it continues to improve as it exposed to new distributions of objects, new types of environments, and new types of obstacles. Our results include developing agents that are able to learn models of context to more efficiently find unseen objects, agents that are able to more efficiently navigate in unknown environments, and an extensive and highly detailed 3D scene database with substantially more detail than previous data sets. We detail a large number of scientific publications that describe and support this work.

15. SUBJECT TERMS
Life Long Learning, Continual Adaptive Learning, Surrogate Task Learning, Intelligent Search, Multi-task Learning Inference of 3D structure from Images, Object Location from Image Based Contextual Reasoning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**TOM RENZ**
U	U	U	UU	34	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

TABLE OF CONTENTS

TABLE OF FIGURES

1.0 SUMMARY

Our work under Phase I of the DARPA Lifelong Learning Machines (L2M) Program has concentrated on solving lifelong learning problems in the domain of computer vision and machine learning. Our work has focused on solving the general task of "Intelligent Search", which is the combined visual and navigation task of finding an object, described either by name or with a picture, in a new or unknown environment.

There are many facets of this problem, including the reasoning of where an object is likely to be based on a visual assessment of the current scene, the ability to navigate quickly and efficiently through unknown environments, the inference of 3D structure from images to aid navigation, and learning policies for efficient navigation. We report on progress for all four of these goals.

In addition to these four areas of research being directly related to the larger goal of Intelligent Search, they are all central topics in Lifelong Learning. In each area, the agent is designed so that it continues to improve as it exposed to new distributions of objects, new types of environments, and new types of obstacles.

2.0 INTRODUCTION

Machine learning methods have achieved great strides in the last 10 years, especially since the recent great advances in neural network technology, starting in 2012 with the work of Geoffrey Hinton's group in Toronto. Many of these advances have been at the intersection of computer vision and machine learning. New methods in machine learning have been responsible for large performance improvements in computer vision. At the same time, problems in computer vision have demanded innovation in new neural net architectures to handle the unique data formats (images and video) and the unique scalability requirements of computer vision data sources.

Despite these impressive advances, there are many problems that have continued to stubbornly resist solution. These include the following problem areas:

- the ability for artificial agents to maintain performance on an old problem when learning a new problem
- the ability to continue to improve performance once exposed to more data, even as the amount of data becomes very large
- the ability to develop notions of common sense that allow agents to rapidly adapt to new situations without massive training data
- the ability to learn without explicit labels, or at least with a small number of labels

Our work under Phase I of the DARPA Lifelong Learning (L2M) program has focused on solving these kinds of Lifelong Learning problems in the domain of computer vision and machine learning.

In our proposal for this program, we described a number of investigations aimed at improving lifelong learning. We chose the problem of "Intelligent Search" as a focus point to organize our approach to these problems. By Intelligent Search, we refer to the task of finding an object in a novel environment, specifically when the object cannot be seen directly from the starting point. The object to be found might be described with the name, such as "apple" or "rifle", or via the picture of an example, like the picture of an apple. In this report, we describe several of our advances aimed at improving Intelligent Search by attacking some of the problem areas described above.

3.0 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 HALF&HALF: New Tasks and Benchmarks for Studying Visual Common Sense

The general recognition of objects, people, actions, and scene types has been a core focus of computer vision research [1]. However, now that we have achieved a degree of success in these problems, it is time to define new problems that will spur us to reach the next level of visual intelligence. The development of visual common sense is critical to the development of intelligent agents that can be useful in dynamic, novel environments.

But what exactly is visual common sense? We suggest that the ability to make intelligent assessments of where things might be, when not directly visible, is a critical and ubiquitous capability shared by humans and other intelligent beings and is a fundamental component of visual common sense. Humans regularly demonstrate the ability to make decisions in the absence of explicit visual cue.

Suppose we are in the hallway of an unfamiliar apartment, with three directions to go, and want to find a TV (Figure 1). Which direction should we go? An intelligent agent, leveraging visual clues and reasoning that a TV is more likely to be near a couch, might decide to turn right. In this case, such a choice leads quickly to a TV. This sort of "intelligent search" is a prominent example of visual common sense, and we believe it represents a skill that will be essential in developing intelligent agents.

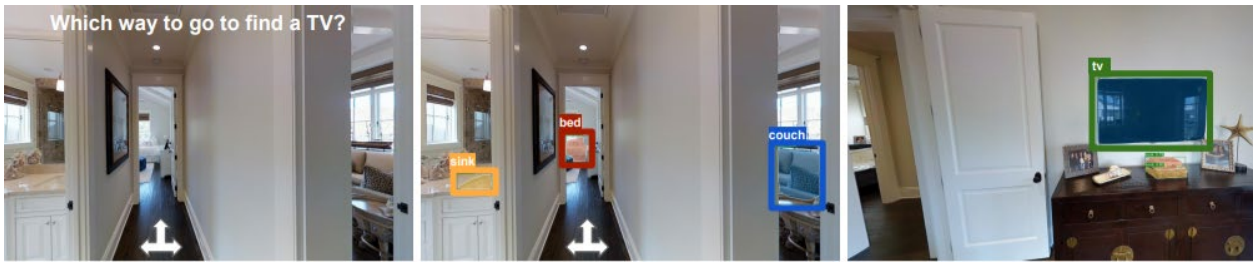

Figure 1: Toy Visualization Demonstrating Intelligent Search

Closely related to our work are earlier efforts on incorporating contextual information for visual prediction [2, 3, 4, 5]. We believe a formal benchmark on such capabilities in the most basic forms can be a valuable addition.

3.1.1 The Half&Half Visual Prediction Tasks

In this work, we formalize the problem of inferring the presence of what we cannot already see in an image. To do this, we rely on the fact that different views of an image depict the same scene. Hence, individual sections can be used as contextual cues for the other section. For this reason, we call these tasks the Half&Half tasks. We define three different Half&Half tasks (Fig.2):

- **Image-to-Label task:** One half of the image is provided, and the task is to infer a categorical label for what is likely to be present in the other half amongst the given K choices (Fig.3).

- **Label-to-Image task:** A target category and a set of K half images are provided. The task is to infer which candidate is most likely to have the target in its other half (Fig.4).
- **Image-to-Image task:** A query image and K image choices are provided, all of them being half images. The task is to infer which of the choices is the most likely to be from the same image as the query (Fig.5).

Figure 2: The Half&Half Visual Prediction Tasks

The three variants of the tasks were inspired by the common-sense reasoning capabilities of intelligent beings under uncertainty. Specifically, an agent trying to find a specific type of object should be able to decide whether the current direction is promising (Image-to-Label). And if not, given observations towards other directions, which one should be preferred (Label-to-Image)? Image-to-Image is modeling an intelligent reasoning capability to directly predict the next visual observations, which can enable an agent to prepare for imminent encounters.

Our hope is that the Half&Half benchmarks, and perhaps their next generations, drive forward the research in designing intelligent agents by training and evaluating such systems for visual "common sense".

3.1.2 The Half&Half Benchmarks

In this section, we describe our three new benchmarks. Each benchmark is constructed to study one of the three variants of the Half&Half tasks introduced in the previous section. We make use of images and annotations from existing datasets originally created for object detection or scene understanding. As we will show in this section, the way we create the benchmarks requires no additional annotations compared to those standard recognition tasks. This allows us to directly make use of large-scale existing datasets.

The Image-to-Label Benchmark

Figure 3: Half&Half Image-to-Label Benchmark Example

Image selection: The benchmark is created using the training and validation images from Microsoft Common Objects in Context, MS-COCO [6]. We consider the left-half of each image as the context for the objects present in the right-half. From the above sets, we sample images that have at least one single object present in its right-half. Furthermore, we discard images whose left-half and right-half contain any overlapping objects. As a design choice, we exclude the "person" category and consider only the remaining 79 categories since we observe that "person" is very common in MS-COCO and has significant co-occurrence with the majority of other categories. In total, we obtain 45,843 images meeting the criteria above.

Problems and splits: Out of all the obtained images, we create a random train/val/test split of 32,000/3,843/10,000 images. Each of the training and validation images are provided with one image (the left-half) and a set of labels from the right-half. From the test images, we form test problems in the form of the Image-to-Label task. Fig.3 shows an example. Five candidate categories are given where only one of them actually appears in the right-half. For the correct candidate choice, we randomly pick one object category that exists in the right-half among the ground truth. For the wrong candidates, we randomly select from all MS-COCO object categories not present in the whole image.

The Label-to-Image Benchmark

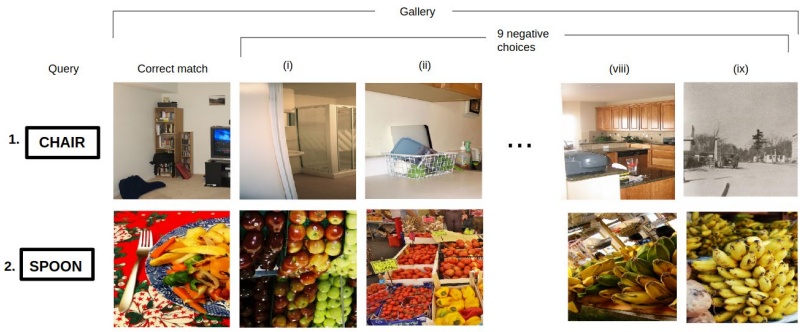

Figure 4: Half&Half Label-to-Image Benchmark Examples

Image selection: Because of the close formulation between the Label-to-Image and Image-to-Label tasks, we can reuse the data collected for the Image-to-Label benchmark, with a few critical modifications. The same set of images, labels, and train/val/test image split are used. The differences only lie in the way the problems are formulated.

Problems and splits: As illustrated in Fig.4, a Label-to-Image problem contains a query label and 10 gallery images. We create one such problem for each of the images in the benchmark using the following steps:

(1) Each (right-half object label, left-half image) pair from Image-to-Label benchmark is sampled as query object and correct candidate.

(2) From the remaining images in the split, images containing the query object label in their right halves are filtered out. The remaining left-half images are then ranked based on the similarity scores with the correct candidate and 9 images are selected randomly as wrong candidates from the top 100.

We follow [7] and use low-level visual features (Generalized Search Tree, GIST and color histogram) for computing the similarity. In total, we obtain 47,370/5,686/10,000 train/val/test problem sets.

The Image-to-Image Benchmark

Figure 5: Half&Half Image-to-Image Benchmark Example

Image selection: Since the Image-to-Image task involves half images as both queries and candidate choices and re-quires no label annotations, we are able to consider any natural images for building the benchmark. We chose to use the SUN360 dataset [8], which offers a large collection of high-resolution 9104×4552 rectangular panorama images. Using panoramas allows us to cut image crops instead of using adjacent halves. By making the two "halves" some distance apart, they will have little overlap in content and offer diverse visual information.

Problems and splits: Among all images available in SUN360, we randomly sample 27,999 training images and 29,142 testing images. Each partition is further divided into a query partition (from which query and correct choices are drawn) and a gallery partition (from which negative choices are drawn). We construct one problem for each image in the query partition: one of its two crops are randomly chosen as the query image and the other one as the correct choice. Nine wrong choices are then randomly sampled from the corresponding gallery partition. A ranking procedure that is the same as the Label-to-Image benchmark is also applied to avoid trivial solutions. In total, we obtain 8,399 training and 8,742 testing problems, which are constructed from the training and testing partitions, respectively.

Evaluation: During testing, the objective is to correctly pick the correct one among the K= 10 choices. The evaluated algorithm is required to pick a top choice for each test problem. Rank-1 accuracy is reported for evaluation.

3.1.3 Methods and Experiments

Image-to-Label

Our task for this benchmark is to identify object categories that are likely to be present in the right-half by observing only the left-half. We formulate this as a multi-category classification problem. We define two types of classifiers: symmetric and anti-symmetric. A symmetric classifier is a standard Convolutional neural Network, CNN trained on the left-half image to predict the labels of objects in the left-half itself, which is equivalent to a traditional classifier. Meanwhile, an anti-symmetric classifier is trained on the same set of left-half images, but with object categories present in the right-half as the target labels.

For training the classifiers, we use the training split provided by the benchmark. Given the set of all left-half images, we train a CNN (ResNet-50 [9] pretrained on ImageNet [10]) to predict the presence of object categories. The last Fully Connected, FC layer is modified to match the 79 object categories we use according to the classifier (symmetric or anti-symmetric). If there are multiple categories for an image, we duplicate the left-half image in the training set and as-sign an individual category to each of the left-half images. We follow this approach so as to maintain consistent behavior with the benchmark-setting, where our candidate list only contains a single correct object category.

Label-to-Image

For the Label-to-Image benchmark, our goal is to rank the candidate images based on the likelihood of containing the given query object. We propose following two methods.

Indirect training: In this case, we use any classifier trained on the Image-to-Label task to compute posterior probabilities for the query object. Based on these posteriors, candidates are ranked. We directly use the anti-symmetric classifier trained for the Image-to-Label benchmark.

Direct training: In the second method, we train a CNN to directly compare the given candidate images conditioned on the query label. We formalize this as a classification problem where, given the candidate set, the objective is to predict the most likely image to contain the query label. For each image in the candidate set, we compute a class score for the query label. To do this, we consider the output of the classification layer of a CNN. We then normalize the class scores across candidate images. This reflects the probability distribution among the candidate images given the query label. Finally, ten candidate images are ranked according to their respective posterior probabilities for the query label. We use the same ResNet-50 model (ImageNet pretrained) as our base classifier.

Image-to-Image

For this task, given a query image and the 10 gallery images, the goal is to find the correct match coming from the same image. As two baselines, we compute the L2 distance of feature vectors from the last fully connected layer of a ResNet-18 [9] between the query image and each of the gallery images. We use models pre-trained on ImageNet [10] and Places365 [11].

In addition to the baselines, we also train networks with two different metric learning techniques. Suppose we have feature vectors, $x \in R^{D\times 1}$, $y \in R^{D\times 1}$, computed from the backbone network. With bilinear metric learning, a similarity score between x and y is computed by $x^{T}Wy$ where $W \in R^{DxD}$ can be trained. We also train networks with symmetric metric learning that learns $L \in R^{DxD}$ in $(Lx)^{T}(Ly)$.

All of our networks are trained using triplet loss [12], where a triplet is generated with a query image, the corresponding correct match, and one of the 9 negatives. Each model is then trained such that a query image is closer to the correct match than to the negative. We first train networks by freezing the parameters in the backbone network. After W and L are learned, we also fine-tune the entire network.

Downstream Task: One-step object search

In this section, we demonstrate how models trained on our benchmarks can be useful for the object search sub-task. Specifically, we investigate an agent's ability to utilize contextual visual cues to guide itself towards to target object. To find a toilet, an intelligent agent should be able to identify and differentiate a washroom with a living room or a kitchen and accordingly move towards it irrespective of whether the toilet is visible.

For this evaluation, we build a small test set using the Active Vision dataset [13], which is originally designed for object instance recognition and navigation in indoor environments. Within the environments provided in Active Vision, we sample locations from where three different types of rooms are visible (Fig. 6).

Figure 6: Example of a Navigation Task Built from the Active Vision Dataset. Query: "toilet"; Correct Answer: (c)

A target category is chosen for an object that is present in one of the given rooms (validated by other views of the environment) but is not visible from the current view. The corresponding room is then designated as the correct answer. Each problem thus consists of one query category and three images. The objective of the agent is to pick which room among the three is the most likely to contain objects belonging to the query category. An example problem is shown in Fig. 6. To increase diversity, we also mix and match images across different environments to obtain some of the problems. In total, we collect 100 such problems.

Lifelong Learning Experiments

To perform the intelligent search task, it is important for the agent to be able to:

- improve performance with exposure to data in new environment (adaptation),
- detect change of environment and build conditional models, depending upon scene type. (change detection).
- should return to previous performance when returning to a previously seen environment and additional exposure in same environment should continue to improve behavior. (Performance recovery).

Hence to quantify an agent's ability to search for objects efficiently, it is critical to evaluate its performance in a Lifelong learning setting. In the following sections, we describe our framework in detail by particularly providing details with respect to:

- Environment design
- Learning and Evaluation scheme

Environment Design

We simulate lifelong learning setting by modifying the Half&Half Image to Label benchmark. Specifically, we divide the training dataset based on discrete scenes types represented by the given image. Here a scene type is defined as a scene with specific spatial correlation structure of objects (frequency of objects, local co-occurrence statistics). We select 'Indoor' and 'Outdoor' as our scene types to represent the changing environment variable over time. To create our environment, we sample 1000 indoor and outdoor images respectively from the training set of the Half&Half Image to Label Benchmark.

We further introduce quantifiable controls for modifying the environment. We do so in the following way:

- **Convex combination of task-dependent data:** In real-world settings, different tasks do not have specific task boundaries. We incorporate this by combining the individual scene-type image sets before performing the learning. Particularly, we linearly combine the original image sets and control the combination using the parameter α where all coefficients are non-negative and sum to 1 (convex combination). For example, let
  - A = Original dataset of scene-type X
  - B = Original dataset of scene-type Y

  Then, we randomly sample from A and B to create A` and B` such that:
  - $A` = \alpha.A + (1 - \alpha).B$
  - $B` = \alpha.B + (1 - \alpha).A$
- **Average number of data points per task:** In a real-world setting, data points are non i.i.d and are observed in a streaming fashion. Based on this, we create mutually exclusive subsets of K images from the individual sets. These subsets represent sub-tasks with same task variable (in our case the scene-types) but different data points. The agent is then sequentially exposed to the individual sub-tasks. For our experiments, we select K=250. While we keep the value of K constant across each sub-task, it need not be same and can be different for each sub-task.

Learning and Evaluation scheme

As demonstrated in the previous sections, Half&Half tasks are good approximation of the 'intelligent search' problem. Hence, for our lifelong learning simulations, we adopt the Half&Half Image-to-Label task as the agent objective. We further select the Anti-symmetric classifier as our agent and follow the same specifications as mentioned before.

Based on standard lifelong learning paradigm, we follow a continuous learning and evaluation cycle, wherein, an agent performs learning on the first task and is then evaluated on it before moving to the new task. Specifically, for each episode (comprising of a new sub-task), we train the agent for 30 epochs and is then evaluated on the test set. Here, the test set is kept same across all tasks, so as to compare learning performance across each episode. For the test set, we use the Half&Half Image-to-Label testing set and evaluate the agent performance based on Rank-1 Accuracy.

3.2 Intelligent Exploration and Navigation

For the agent to conduct intelligent search, it needs to know how to collect visual information [14, 15, 16] and navigate its environment effectively [17]. For example, consider a service robot that is moving around in an open environment without specific goals, waiting for future tasks like delivering a package from one person to another or picking up coffee from the kitchen. It needs to efficiently and constantly gather information so that it is well prepared to perform future tasks with minimal delays. Similarly, consider a search-and-rescue scenario, where a robot is deployed in a hostile environment, such as a burning building or earthquake collapse, where time is of the essence. The robot has to adapt to such new unseen environments and rapidly gather information that other robots and humans can use to effectively respond to situations that dynamically unfold over time (humans caught under debris, locations of fires, and presence of hazardous materials). Having a robot that knows how to explore intelligently can be critical in such scenarios, reducing risks for people while providing an effective response.

Next, we introduce our contributions in the domain of intelligent exploration and navigation.

3.2.1 Occupancy Anticipation for Efficient Exploration and Navigation

In visual navigation, an agent must move intelligently through a 3D environment in order to reach a goal. One of the key factors for success in navigation has been the movement towards complex map-based architectures that capture both geometry and semantics, thereby facilitating efficient policy learning and planning. These learned maps allow an agent to exploit prior knowledge from training scenes when navigating in novel test environments. Despite such progress, state-of-the-art approaches to navigation are limited to encoding what the agent actually sees in front of it. In particular, they build maps of the environment using only the observed regions, whether via geometry or learning. Thus, while promising, today's models suffer from an important inefficiency: to map a space in the 3D environment as free or occupied, the agent must directly see evidence thereof in its egocentric camera.

Our key idea is to anticipate occupancy. Rather than wait to directly observe a more distant or occluded region of the 3D environment to declare its occupancy status, the proposed agent infers occupancy for unseen regions based on the visual context in its egocentric views. For example, in Figure 7, with only the partial observation of the scene, the agent could infer that it is quite likely that the wall extends to its right, a corridor is present on its left, and the region immediately in front of it is free space.

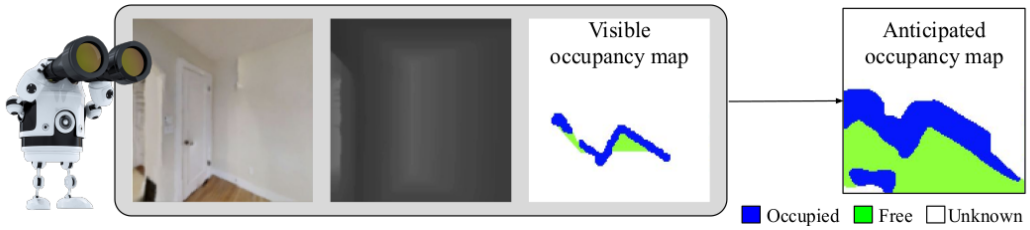

Figure 7: Occupancy Prediction

Such intelligent extrapolation beyond the observed space would lead to more efficient exploration and navigation. To achieve this advantage, we introduce a model that anticipates occupancy maps from normal field-of-view Red, Green, Blue, RGB(D) observations, while aggregating its predictions over time in tight connection with learning a navigation policy.

Furthermore, we incorporate the anticipation objective directly into the agent's exploration policy, encouraging movements in the 3D space that will efficiently yield broader and more accurate inferred occupancy maps [17]. We validate our approach on Gibson and Matterport3D, two 3D environment datasets spanning over 170 real-world spaces with a variety of obstacles and floor plans.

3.2.2 Emergence of Exploratory Look-around Behaviors through Active Observation Completion

Standard computer vision systems assume access to intelligently captured inputs (e.g., photos from a human photographer), yet autonomously capturing good observations is a major challenge in itself. We address the problem of learning to look around: How can an agent learn to acquire informative visual observations? An agent ought to be able to enter a new environment or pick up a new object and intelligently (non-exhaustively) "look around." The ability to actively explore would be valuable in both task-driven scenarios (e.g., a drone searches for signs of a particular activity) and scenarios where the task itself unfolds simultaneously with the agent's exploratory actions (e.g., a search-and-rescue robot enters a burning building and dynamically decides its mission).

Any such scenario brings forth the question of how to collect visual information to benefit perception. A naïve strategy would be to gain full information by making every possible observation—that is, looking around in all directions or systematically examining all sides of an object. However, observing all aspects is often inconvenient if not intractable. Fortunately, in practice, not all views are equally informative. The natural visual world contains regularities, suggesting that not every view needs to be sampled for accurate perception. For instance, humans rarely need to fully observe an object to understand its three-dimensional (3D) shape, and one can often understand the primary contents of a room without literally scanning it. In short, given a set of past observations, some new views are more informative than others. For example, in Figure 8 left, an agent that has observed limited portions of its environment can reasonably predict some unobserved portions (e.g., water near the ship) but is much more uncertain about other portions. Where should it look next? Similarly, an agent inspecting a 3D object (Figure 8 right) having seen a top view and a side view, how must it rotate the mug now to get maximum new information?

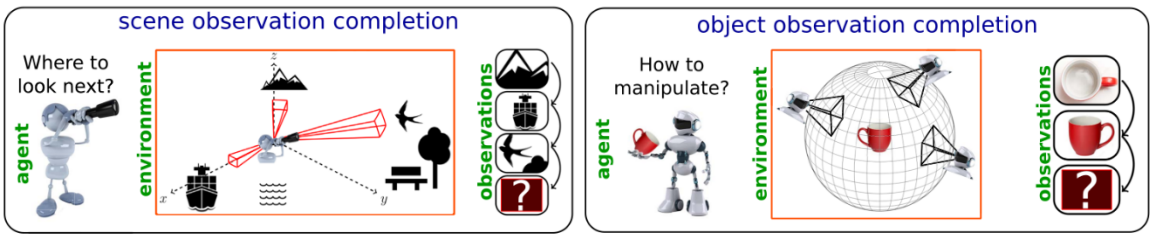

Figure 8: Scene and Object Observation Completion

This fact leads us to investigate the question of how to effectively look around: How can a learning system make intelligent decisions about how to acquire new exploratory visual observations? We propose a solution based on "active observation completion": An agent must actively observe a small fraction of its environment so that it can predict the pixelwise appearances of unseen portions of the environment. Our goal is to learn a policy for controlling an agent's camera motions such that it can explore novel environments and objects efficiently. Critically, we aim to learn policies that are not specific to a given object or scene, nor to a specific perception task. Instead, the look-around policies ought to benefit the agent exploring new, unseen environments and performing tasks unspecified when learning the look-around behavior. To this end, we formulate an unsupervised learning objective based on active observation completion [18]. The main idea is to favor sequences of camera motions that make the unseen parts of the agent's surroundings easier to predict.

3.2.3 An Exploration of Embodied Visual Exploration

A variety of exploration methods have been proposed both in the reinforcement learning and computer vision literature. They employ ideas like curiosity, novelty, coverage, and reconstruction to overcome sparse rewards or learn task-agnostic policies that generalize to new tasks. In this study, we consider algorithms designed to handle complex photorealistic indoor environments where a mobile agent observes the world through its camera's field of view and can exploit common semantic priors from previously seen environments (as opposed to exploration in randomly generated mazes or Atari games).

How do different exploration algorithms work for different types of environments and downstream tasks? Despite the growing literature on embodied visual exploration, it has been hard to analyze what works when and why. This difficulty is due to several reasons. First, prior work evaluates on different simulation environments such as SUN360, ModelNet, VizDoom, SUNCG, DeepMind-Lab, and Matterport3D. Second, prior work uses different baselines with varying implementations, architectures, and reinforcement learning algorithms. Finally, exploration methods have been evaluated from different perspectives such as overcoming sparse rewards, pixelwise reconstruction of environments, area covered in the environment, object interactions, or as an information gathering phase to solve downstream tasks such as navigation, recognition, or pose estimation. Due to this lack of standardization, it is hard to compare any two methods in the literature.

This work [19] presents a unified view of exploration algorithms for visually rich 3D environments, and a common evaluation framework to understand their strengths and weaknesses. To that end, we propose a novel benchmark for consistently evaluating exploration algorithms under common experimental conditions: policy architecture, 3D environments, learning algorithm, and diverse evaluation metrics. We then present a comparative study of four popular exploration paradigms on this standardized benchmark. To enable this study, we introduce new metrics and baselines, and we improve upon some existing approaches to scale well to 3D environments. Specifically, we extend the ideas from reconstruction-based exploration 360° scene exploration approaches to work on general 3D environments. We also introduced a new coverage reward function that improves upon the existing area-coverage method. Our analysis provides a comprehensive view of the state of the art and each paradigm's strengths and weaknesses.

3.3 Multi-modal Learning

Multi-modal perception is important to capture the richness of real-world sensory data for objects, scenes, and events. The sounds made by objects, whether actively generated or incidentally emitted, offer valuable signals about their physical properties and spatial locations.

An intelligent agent needs to leverage both acoustic and visual cues efficiently in order to find objects in its environment [20].

Next, we introduce our contributions in learning how to assign sounds to objects using a visually-guided audio source separation approach.

3.3.1 Co-Separating Sounds of Visual Objects

Learning how objects sound from video is challenging, since they often heavily overlap in a single audio channel. Current methods for visually-guided audio source separation sidestep the issue by training with artificially mixed video clips, but this puts unwieldy restrictions on training data collection and may even prevent learning the properties of "true" mixed sounds. We introduce a co-separation training paradigm that permits learning object-level sounds from unlabeled multi-source videos [21]. Our novel training objective requires that the deep neural network's separated audio for similar-looking objects be consistently identifiable, while simultaneously reproducing accurate video-level audio tracks for each source training pair, see Figure 14.

3.3.2 Self-Supervised Learning

Unlike supervised learning where manual annotations are given to train a model for the target task (e.g. object segmentation [22], pose estimation [23], human-objects interaction [24]), self-supervised image feature learning methods leverage structured information within the data itself to generate labels for representation learning. Hence, it circumvents the need for costly labeled data to learn discriminative features. The learned representation can then be transferred to a downstream task [25] like object identification for robust and accurate performance using few labeled examples.

Self-Supervised Feature Learning by Lifting Views to Viewgrids

We introduce an unsupervised feature learning approach that embeds 3D shape information into a single-view image representation. The main idea is a self-supervised training objective that, given only a single 2D image, requires all unseen views of the object to be predictable from learned features. We implement this idea as an encoder-decoder convolutional neural network [26]. The network maps an input image of an unknown category and unknown viewpoint to a latent space, from which a deconvolutional decoder can best "lift" the image to its complete viewgrid showing the object from all viewing angles, see Figure 9. Our class-agnostic training procedure encourages the representation to capture fundamental shape primitives and semantic regularities in a data-driven manner—without manual semantic labels.

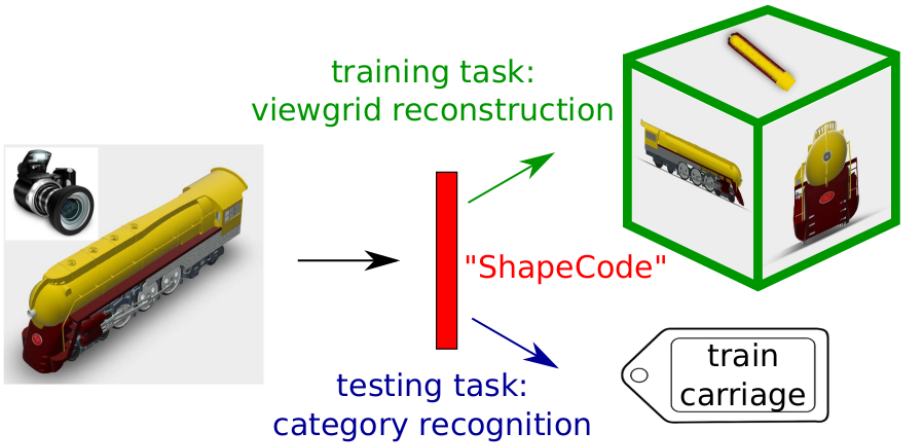

Figure 9: Lifting and Image to a Viewgrid

Visual Sound

Binaural audio provides a listener with 3D sound sensation, allowing a rich perceptual experience of the scene. However, binaural recordings are scarcely available and require nontrivial expertise and equipment to obtain. We propose to convert common monaural audio into binaural audio by leveraging video [27]. The key idea is that visual frames reveal significant spatial cues that, while explicitly lacking in the accompanying single-channel audio, are strongly linked to it. Our multi-modal approach recovers this link from unlabeled video, see Figure 10. We devise a deep convolutional neural network that learns to decode the monaural (single-channel) soundtrack into its binaural counterpart by injecting visual information about object and scene configurations. We call the resulting output 2.5D visual sound—the visual stream helps "lift" the flat single channel audio into spatialized sound.

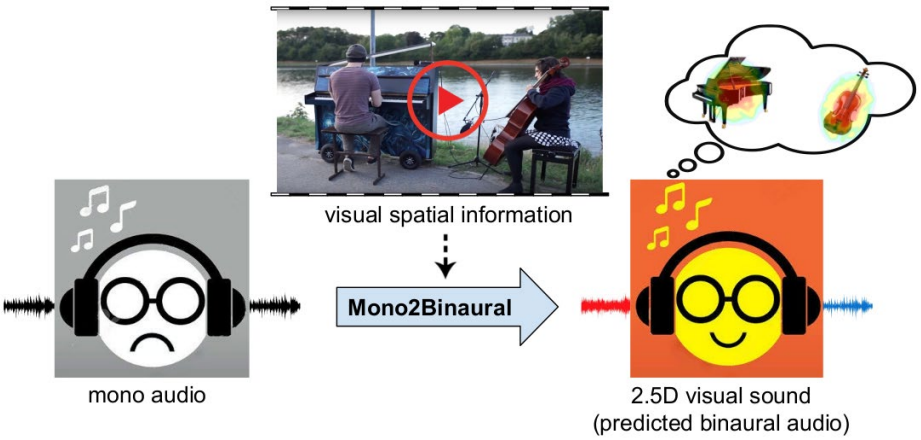

Figure 10: Recovering Sound Sources in Multi-Modal Video

Spatial Image Representation Learning through Echolocation

Several animal species (e.g., bats, dolphins, and whales) and even visually impaired humans have the remarkable ability to perform echolocation: a biological sonar used to perceive spatial layout and locate objects in the world. We explore the spatial cues contained in echoes and how they can benefit vision tasks that require spatial reasoning [28]. First, we capture echo responses in photo-realistic 3D indoor scene environments. Then we propose a novel interaction-based representation learning framework that learns useful visual features via echolocation.

3.4 Additional Research

3.4.1 Adaptation to Changing 3D World

Computer vision systems operate in the 3-dimensional world, and many tasks implicitly or explicitly require reasoning about 3D scenes. Two key tasks in this domain are estimating a depth map for a single image (monocular depth estimation) and estimating surface orientations of scene elements (normal estimation). Depth information is integral to many problems in robotics, including mapping, localization and obstacle avoidance for terrestrial and aerial vehicles, and in computer vision, including augmented and virtual reality. Compared to depth sensors, monocular cameras are inexpensive and ubiquitous, and would provide a compelling alternative if coupled with a predictive model that can accurately estimate depth. Unfortunately, no public dataset exists that would allow fitting the parameters of such a model using depth measurements taken by the same sensor in both indoor and outdoor settings, which is relevant both for general goals of live-long visual learning, and for L2M agents operating in 3D environments that span these settings. Lack of data makes learning to adapt to changes, e.g., due to an agent's transition between indoor and outdoor environments, difficult.

More recently, we have introduced DIODE (Dense Indoor/Outdoor DEpth), a dataset that contains thousands of diverse, high-resolution color images with accurate, dense, long-range depth measurements [29]. DIODE is the first public dataset to include RGBD images of indoor and outdoor scenes obtained with one sensor suite. This is in contrast to existing datasets that involve just one domain/scene type and employ different sensors, making generalization across domains difficult. The dataset is available for download at http://diode-dataset.org, and some samples can be found in Figure 11.

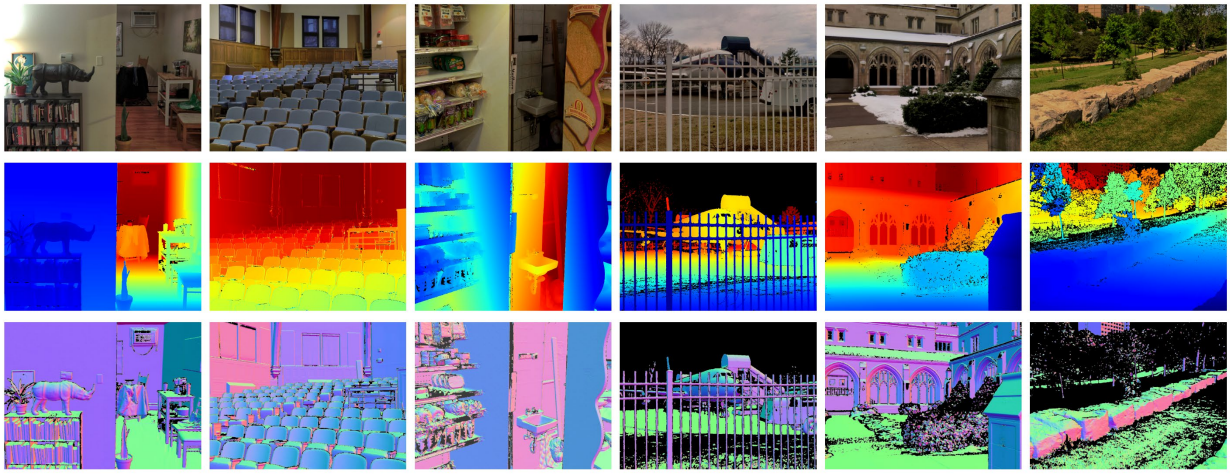

Figure 11: Samples from DIODE. Top: Image, Middle: Depth, Bottom: Surface Orientation

3.4.2 Confidence Estimation for Visual Classification

An important capability of an L2M learning system is safety, in particular, robustness to errors. We have been working on methods for automatic detection of classification errors, aiming to enable a "reject option" with which the classifier, when faced with an instance it cannot classify correctly, can gracefully recover by rejecting the instance rather than making a wrong prediction. Our approach has been based on analysis of stability and invariance of classifier output under image transformations [30]. Most recently, we have proposed a novel approach for classification confidence estimation [31]. We apply a set of semantics-preserving image transformations to the input image, and show how the resulting image sets can be used to estimate confidence in the classifier's prediction.

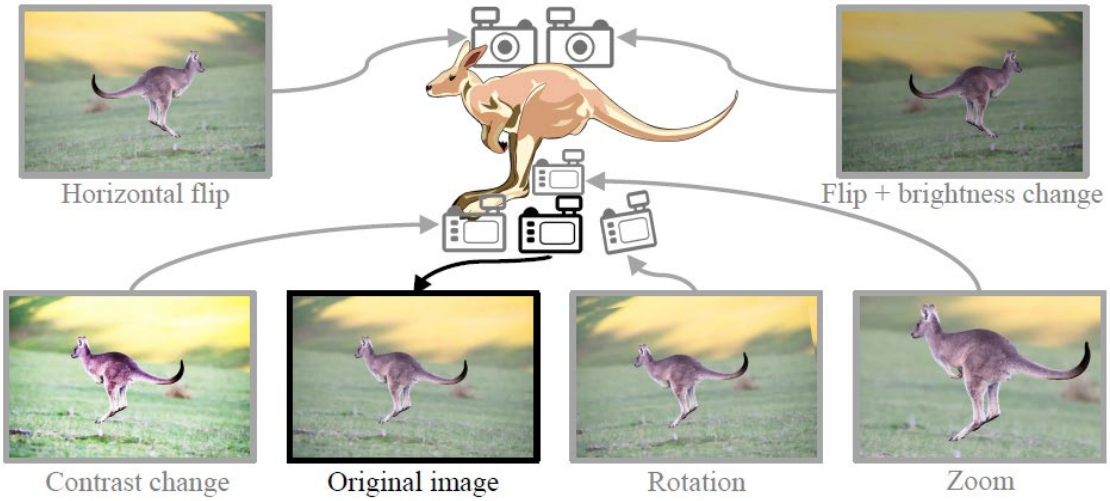

Figure 12: Image Transformation Preserving Content

Our idea is akin to data augmentation, in which a "natural" image is modified to provide a broader coverage of the data distribution for training. Except we apply this augmentation at test time, using a set of semantic preserving image transformations (see Figure 12 for an illustration). The resulting image set is used to estimate the classifier's reliability conditioned on the input image. We find that when the classifier would produce a wrong prediction for the original image, the predictions on the augmented set are less consistent than when the classification would be correct, providing us with a leverage to develop a reject option.

3.4.3 Proxy Tasks for Self-Supervision

As part of our pursuit of self-supervised lifelong learning we have been working on developing proxy tasks that can drive representation learning. One of these tasks is image super-resolution. We have developed state of the art methods for image super-resolution [32], video frame super-resolution [33], and most recently spatial-temporal super-resolution [34].

In the latter, most recent work, we consider the problem of space-time super-resolution (ST-SR): increasing spatial resolution of video frames and simultaneously interpolating frames to increase the frame rate. This is relevant to L2M as a potential proxy task for self-supervised learning, going beyond existing work using purely spatial (pixel) super-resolution for this purpose. Modern approaches handle time and space separately. In contrast, our proposed model

called STARnet super resolves jointly in space and time. This allows us to leverage mutually informative relationships between time and space: higher resolution can provide more detailed information about motion, and higher frame-rate can provide better pixel alignment. The components of our model that generate latent low- and high-resolution representations during ST-SR can be used to fine tune a specialized mechanism for just spatial or just temporal SR.

4.0 RESULTS AND DISCUSSIONS

4.1 Half & Half Results

Recall that the Half & Half benchmark is designed to study how context can affect the recognition and finding of objects that may not be seen, or may be only partially seen. Previously, we defined a number of different paradigms, such as Image-to-Label and Label-to-Image. Below, we give results for each of these paradigms and discuss their significance.

4.1.1 Image-to-Label Task Results

During testing, we evaluate the performance of a model on the test problems based on whether it can pick the right candidate among the five choices, and also the rank that it assigns to the correct candidate. Specifically, benchmark users are required to report:

- Rank-1 Accuracy: $(1/N)\sum_i[[r_i = 1]]$,
- Mean Reciprocal Rank (MRR): $(1/N)\sum_i 1/r_i$. Here N denotes the total number of test samples and r_i is the rank of the correct candidate in a model's output.
- From the trained network, we obtain the posterior probability distribution over all 79 categories in the MS-COCO dataset. We evaluate the performance of our context driven model on the benchmark by computing the ranking of the five candidate categories in the candidate list according to their posterior probabilities.
- Table1 compares symmetric and anti-symmetric classifiers, as well as a Multi-Layer Perceptron, MLP baseline using GIST [35].

Table 1: Evaluations On The Image-to-Label Benchmark.

Classifier	**Rank-1 Accuracy**	**Mean Reciprocal Rank (MRR)**
MLP (GIST)	42.0%	0.635
Symmetric	58.7%	0.707
Anti-Symmetric	74.3%	0.855

4.1.2 Label-to-Image Task Results

Evaluation: During testing, the objective is to correctly pick the correct choice among the K= 10 gallery choices. The evaluated algorithm is required to provide a ranking among the choices for each test problem, and the two evaluation measures, rank-1 accuracy and MRR, are reported.

The direct and indirect classifiers are compared in Table 2. Direct training provides a noticeable gain, which suggests it is beneficial to have a training objective more closely aligned with the actual evaluation protocol.

Table 2: Evaluations On The Label-To-Image benchmark.

Classifier	**Rank-1 Accuracy**	**Mean Reciprocal Rank (MRR)**
Indirect	44.7%	0.624
Direct	46.6%	0.646

4.1.3 Image-to-Image Task Results

Results and comparisons are summarized in Table 3. A few observations can be made: (1) Place365 offers better pretraining compared to ImageNet for our problems; (2) various metric learning techniques all help significantly compared to direct L2 distances; and (3) fine-tuning the back-bone network offers consisting improvements.

Table 3: Evaluations On The Image-To-Image benchmark.

Pre-trained	**Fine-Tuned**	**L2**	**Symmetric Metric**	**Bilinear Metric**
ImageNet	No	47.3%	54.0%	54.1%
ImageNet	Yes	65.3%	64.8%	70.0%
Places365	No	56.2%	63.1%	65.1%
Places365	Yes	67.2%	67.7%	69.0%

4.1.4 One-step Object Search Results

All of the models show significant advantage over chance performance (33%). This validates our initial hypothesis that commonsense knowledge can act as a prior for finding objects under partial or no observability of the target. However, there is still a large gap towards human performance (human performance is at 98.3%). This highlights the importance of future research in this direction.

4.1.5 Lifelong Learning Results for the Half & Half Paradigm

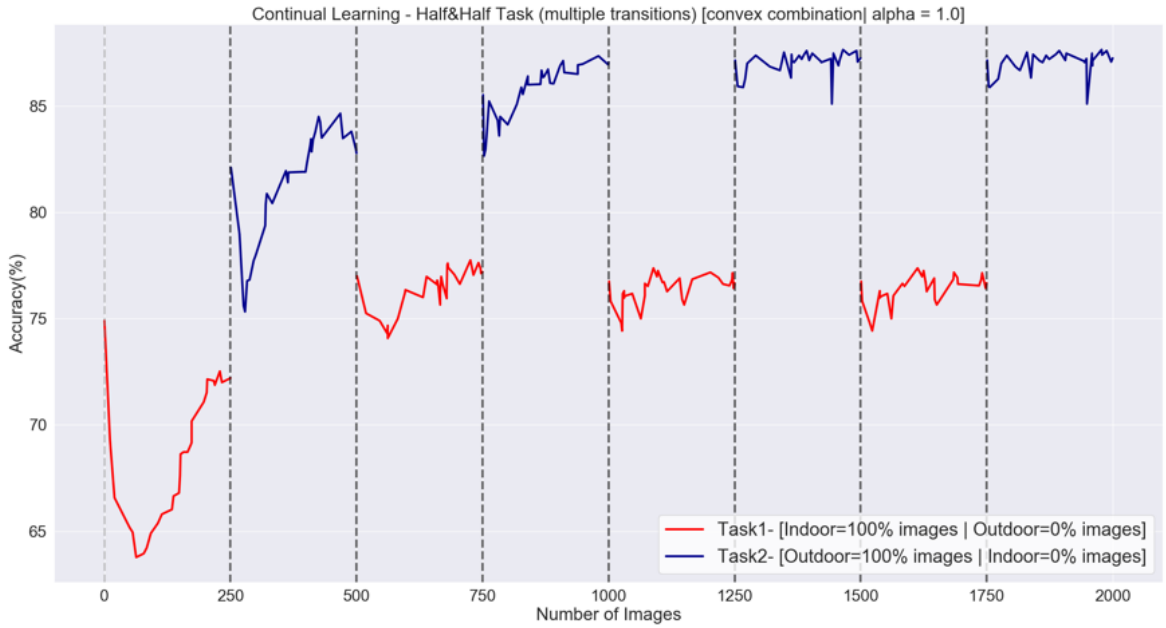

Figure 13: Performance Curve of the Anti-symmetric Classifier Agent

The experimental results of our simulation experiments are summarized in Fig. 13. As can be observed, our anti-symmetric classifier successfully performs positive learning across different tasks. We particularly observe that:

- In the first episode (Task: Indoor; Sub-Task: Indoor subset1), we observe the accuracy increasing.
- As the agent transitions to the new task (Episode 2; Task: Outdoor; Sub-Task: Outdoor subset1), the accuracy drops but recovers quickly. It further improves over the course of learning.
- When transitioning to new episodes, (Tasks keeps alternating; New sub-task in each episodes), the agent performance starts at a lower value but increases quickly. And with each new episode, the drop is lower than previous same task.
- While the overall accuracy increases for the agent across episodes, the trend is different for different tasks (Indoor vs. Outdoor).

4.2 Results for Intelligent Exploration and Navigation

4.2.1 Results for Occupancy Anticipation

We validate our approach on Gibson and Matterport3D, two 3D environment datasets spanning over 170 real-world spaces with a variety of obstacles and floor plans. Using only RGB(D) inputs to anticipate occupancy, the proposed agent learns to explore intelligently, achieving faster and more accurate maps compared to a state-of-the-art approach for neural Simultaneous Location and Mapping, SLAM, and navigating more efficiently than strong baselines. Furthermore, for navigation under noisy actuation and sensing, our agent improves the state of the art, winning the 2020 Habitat PointNav Challenge by a margin of 6.3 Success Path Length, SPL points.

Our main contributions are: (1) a novel occupancy anticipation framework that leverages visual context from egocentric RGB(D) views; (2) a novel exploration approach that incorporates intelligent anticipation for efficient environment mapping, providing better maps in less time; and (3) successful navigation results that improve the state of the art.

- This work [17] is published in the European Conference on Computer Vision (ECCV), 2020.
- Paper and further details: http://vision.cs.utexas.edu/projects/occupancy_anticipation/

4.2.2 Results for Exploratory Look-around Behaviors

Recall that the main idea of this work is to favor sequences of camera motions that make the unseen parts of the agent's surroundings easier to predict. The output is a look-around policy equipped to gather new images in new environments. We demonstrate in experiments that this policy prepares the agent to perform intelligent exploration for a wide range of perception tasks, such as recognition, light source localization, and pose estimation.

- This work [18] is published in the Science Robotics, Vol. 4, Issue 30, 2019.
- Paper: https://robotics.sciencemag.org/content/4/30/eaaw6326

4.2.3 Results for Embodied Visual Exploration

Our key findings in this work include the following:

- In the small and cluttered environments from Active Vision Dataset, the *coverage* and *reconstruction* methods are the strongest paradigms as they prioritize selecting views that maximally increase information.
- In the large and open environments from Matterport3D, *novelty* and *smooth-coverage* approaches are the strongest paradigms as they have smoother reward functions which are easier to optimize in large environments.
- In the diverse Matterport3D testing environments, different approaches tend to dominate on different skills, highlighting the need for diverse evaluation metrics.
- The performance trends among learned approaches remain consistent in noise-free and noisy test conditions, whereas a purely geometric approach like *frontier-exploration* tends to deteriorate rapidly in the presence of sensor noise.
- Our proposed *smooth-coverage* method improves over a prior coverage approach by using a smoother reward function that eases optimization and leads to consistently better performance on a variety of conditions.
- Our proposed adaptation of *reconstruction* successfully explores 3D environments and competes closely with the best methods on most settings.
- Improved memory architectures may be the key to scaling curiosity-driven exploration to visually-rich 3D environments under extreme partial observability.
- An easy-to-implement heuristic *imitation* significantly outperform baselines typically employed, and can serve as a better baseline for future research.

This work [19] is submitted and currently under review.

4.3 Results for Multi-Modal Learning

4.3.1 Result for Co-Separating Sounds of Visual Objects

Our approach disentangles sounds in realistic test videos, even in cases where an object was not observed individually during training, Figure 14. We obtain state-of-the-art results on visually-guided audio source separation and audio de-noising for the MUSIC, AudioSet, and AV-Bench datasets.

- This work [21] was published in the International Conference on Computer Vision (ICCV) 2019.
- Paper and video results: http://vision.cs.utexas.edu/projects/coseparation/

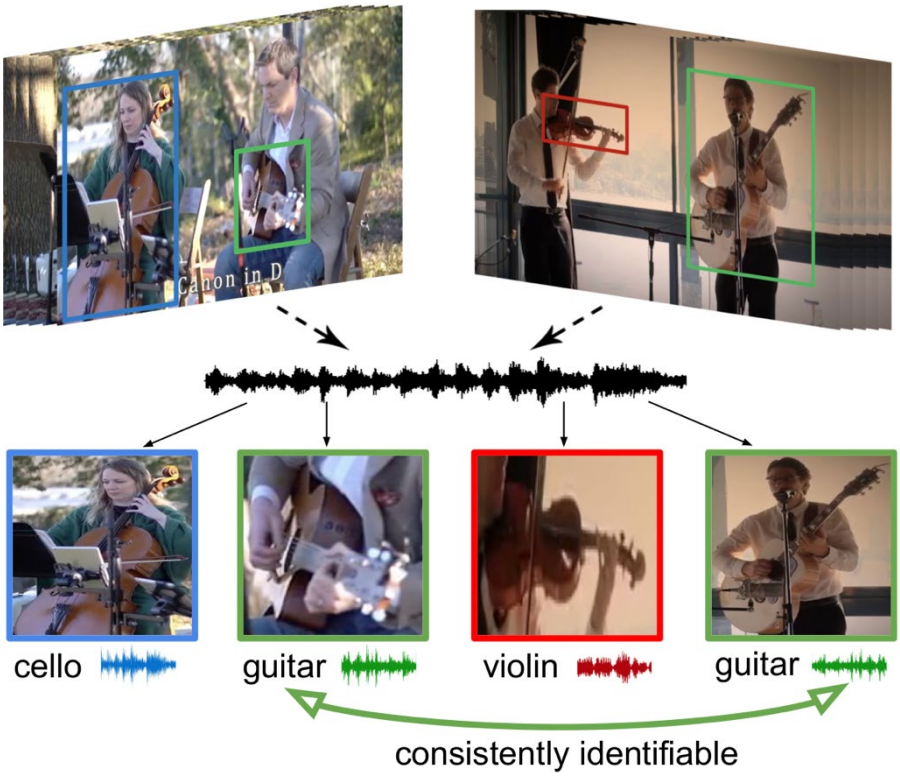

Figure 14: Identifying Sound Sources

4.3.2 Result for Self-Supervised Learning

Lifting to ViewGrids

First we describe our results for self-supervised feature learning by lifting views to viewgrids. Our results on two widely-used shape datasets show 1) our approach successfully learns to perform "mental rotation" even for objects unseen during training, and 2) the learned latent space is a powerful representation for object recognition, outperforming several existing unsupervised feature learning methods.

- This work [26] is published in the European Conference on Computer Vision (ECCV), 2018.
- Paper: https://www.cs.utexas.edu/%7Egrauman/papers/shape-codes-eccv2018.pdf

Visual Sound

In the area of "Visual Sound" we have the following results. In addition to sound generation, we show the self-supervised representation learned by our network benefits audio-visual source separation.

- This work [27] was published in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, and it was recognized as a Best Paper Finalist.
- Paper and video results: http://vision.cs.utexas.edu/projects/2.5D_visual_sound/

Spatial Image Representation Results

We show that the learned image features are useful for multiple downstream vision tasks requiring spatial reasoning—monocular depth estimation, surface normal estimation, and visual navigation—with results comparable or even better than heavily supervised pre-training. Our work opens a new path for representation learning for embodied agents, where supervision comes from interacting with the physical world.

- This work [28] is published in the European Conference on Computer Vision (ECCV), 2020.
- Paper and further details: http://vision.cs.utexas.edu/projects/visualEchoes/

4.4 Results for Additional Research

4.4.1 Adaptation to a Changing 3D World

In early work funded by L2M, we have developed a framework for self-supervised learning driven by depth prediction in videos with nearly linear motion [36]. We further demonstrated an adaptation method that improved depth estimation and other tasks like semantic segmentation when the agent encounters a novel environment (e.g., a new city). Our approach relies on self-supervised, automated fine-tuning in the novel environment.

More recently, we have introduced DIODE (Dense Indoor/Outdoor DEpth), a dataset that contains thousands of diverse, high-resolution color images with accurate, dense, long-range depth measurements [29]. The dataset contains tens of thousands of images with an accompanying depth maps (RGBD images) and very accurate surface normal. The resulting scans exhibit diversity not just between the scenes themselves, but also in the scene composition. Some outdoor scans include a large number of nearby objects compared to existing datasets (like KITTI, where the majority of street scans have few objects near the car), while some indoor scenes include distant objects (e.g., in large meeting halls and office buildings with large atria), in contrast to scenes in other indoor datasets collected with comparatively short-range sensors. As a measure of difficulty induced by the diversity and the environment change between indoor and outdoor scenes, we assessed the ability of the training mean depth to predict depth maps for withheld images, and confirmed that it is drastically worse than for existing datasets, confirming the challenging nature of DIODE.

We expect the unique characteristics of DIODE, in particular the density and accuracy of depth data and above all the unified framework for indoor and outdoor scenes, to enable more realistic evaluation of depth prediction methods and facilitate progress towards general depth estimation methods. We plan to continue acquiring additional data to expand DIODE, including more locations and additional variety in weather and season

4.4.2 Results for Confidence Estimation

We have proposed a novel approach for classification confidence estimation [31]. We apply a set of semantics-preserving image transformations to the input image, and show how the resulting image sets can be used to estimate confidence in the classifier's prediction. We demonstrate the potential of our approach by extensively evaluating it on a wide variety of

classifier architectures and datasets, including ResNext/ImageNet, achieving state of the art performance.

4.4.3 Results for Proxy Tasks for Self-Supervision

We have developed state of the art methods for image super-resolution [32], video frame super-resolution [36], and most recently spatial-temporal super-resolution [34]. Experimental results demonstrate that STARnet improves the performances of space-time, spatial, and temporal video SR by substantial margins on publicly available datasets.

5.0 CONCLUSIONS

In summary, we have made contributions on many fronts. Two of the most significant contributions are our development of the technology to predict where objects are, even when they cannot be seen, and the development of benchmarks to assess this skill. We believe this is a core lifelong learning ability and represents a significant portion of the reasoning ability of agents that we commonly refer to as "common sense". The other major contribution is the development of exploration and navigation behaviors, allowing agents to continue to improve the efficiency with which they move around in novel environments.

Recommendations

We believe the development of these core technologies and the many other contributions described here lay the groundwork for important continuing work in LifeLong Learning, especially as it applies to computer vision and robotics.

The skills that we have focused on can indeed be continually improved as an agent gains more experience, both supervised and unsupervised. Its models of associations and the best strategies for navigation and understanding the 3D world can continue to grow as the agent is exposed to more and more experiences. We believe this is the essence of LifeLong Learning, and is at the heart of this exciting program.

Despite the progress made in this program, there is an enormous amount more to be done. There were interesting discussions about whether the kind of associative learning that we did in this program constituted "LifeLong Learning" or not. As computer vision researchers, we believe that processing of basic sensory inputs, whether they are visual, auditory, somatosensory, or even olfactory, form the critical basis for our "common sense". A more traditional view is that common sense is the knowledge of simple facts, or perhaps the ability to do simple abstract inference and infer things like you would have trouble getting an elephant into your house because it couldn't get through the door.

While these types of inferences no doubt form part of the common conception of common sense, we firmly believe that the bedrock on which these inferences must rest are on fundamental statistical relationships among the stimuli that affect the basic senses such as vision. Therefore, we strongly believe that it will benefit DARPA and the larger research community to continue funding mechanisms whereby researchers can study these basic interactions.

Part of the art of studying these basic interactions, of course, is to come up with tasks that are neither too easy nor too difficult, and demonstrate some aspects of what we consider to be lifelong learning. We believe the tasks of finding objects efficiently by leveraging associative visual cues, and the task of efficiently navigating in new domains both represent important near-term challenges for lifelong learning agents—they are within our grasp, but are not yet solved completely. Of course, we hope to make significant further progress on these directions in Phase II of L2M.

6.0 REFERENCES

1. A. Singh, H. Su, S. Jin, H. Jiang, C. Manjesh, G. Luo, Z He, L Hong, E. G. Learned-Miller & R. Cowell. Half&Half: New Tasks and Benchmarks for Studying Visual Common Sense. In Proc. of the IEEE CVPR Workshops (pp. 1-4), 2019.
2. G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In Proc. ECCV, pages 30–43. Springer, 2008.
3. J. Sun and D. W. Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In Proc. CVPR, pages 1234–1242. IEEE, 2017.
4. A. Torralba. Contextual priming for object detection. IJCV,53(2):169–191, 2003.
5. S. Song, A. Zeng, A. X. Chang, M. Savva, S. Savarese, and T. Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In Proc CVPR, pages 3847–3856, 2018.
6. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In Proc. ECCV, pages 740–755.Springer, 2014.
7. J. Hays and A. A. Efros. Scene completion using millions of photographs .ACM Transactions on Graphics (TOG),26(3):4, 2007.
8. J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In Proc. CVPR, pages 2695–2702. IEEE, 2012.
9. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CoRR, 2015.
10. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In Proc. CVPR, pages 248–255. IEEE, 2009.
11. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. PAMI, 2017.
12. F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proc. CVPR, June 2015.
13. P. Ammirato, P. Poirson, E. Park, J. Kosecka, and A. C. Berg. A dataset for developing and benchmarking active vision. In IEEE ICRA, 2017.
14. D. Jayaraman and K. Grauman. End-to-end Policy Learning for Active Visual Categorization. Transactions on Pattern Analysis and Machine Intelligence (PAMI), Volume 41, Issue 7, pp. 1601-1614, July 2018.
15. D. Jayaraman and K. Grauman. Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 2018.
16. S. Ramakrishnan and K. Grauman. Sidekick Policy Learning for Active Visual Exploration. In Proceedings of the European Conference on Computer Vision (ECCV), Sept 2018.
17. S. Ramakrishnan, Z. Al-Halah, K. Grauman, “Occupancy Anticipation for Efficient Exploration and Navigation,” In Proceedings of the European Conference on Computer Vision (ECCV), August 2020.
18. S. Ramakrishnan, D. Jayaraman, and K. Grauman, “Emergence of Exploratory Look-around Behaviors through Active Observation Completion,” Science Robotics, Vol. 4, Issue 30, May 2019.

19. S. Ramakrishnan, D. Jayaraman, and K. Grauman, "An Exploration of Embodied Visual Exploration," Under review.
20. R. Gao, R. Feris, and K. Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sept 2018.
21. R. Gao and K. Grauman, "Co-Separating Sounds of Visual Objects," In Proceedings of the International Conference on Computer Vision (ICCV), Nov. 2019.
22. B. Xiong, S. Jain, and K. Grauman. Pixel Objectness: Learning to Segment Generic Objects Automatically in Images and Videos. Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2018.
23. Z. Yang, J. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang. Extreme Relative Pose Estimation for RGB-D Scans via Scene Completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, June 2019. (Oral)
24. T. Nagarajan, C. Feichtenhofer, K. Grauman. Grounded Human-Object Interaction Hotspots from Video. ICCV 2019.
25. Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris. SpotTune: Transfer Learning through Adaptive Fine-tuning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, June 2019.
26. D. Jayaraman, R. Gao, and K. Grauman, "ShapeCodes: Self-Supervised Feature Learning by Lifting Views to Viewgrids," In Proceedings of the European Conference on Computer Vision (ECCV), Sept. 2018.
27. R. Gao and K. Grauman, "2.5D Visual Sound." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
28. R. Gao, C. Chen, Z. Al-Halah, C. Schissler, K. Grauman, "VisualEchoes: Spatial Image Representation Learning through Echolocation," In Proceedings of the European Conference on Computer Vision (ECCV), August 2020.
29. I. Vasiljevic et al., DIODE: A Dense Indoor and Outdoor Depth Dataset. Workshop on 3D Scene Understanding, CVPR 2019.
30. Y. Bahat, M. Irani, G. Shakhnarovich, Natural and adversarial error detection using invariance to image transformations, arXiv preprint arXiv:1902.00236.
31. Y. Bahat, G. Shakhnarovich, Classification Confidence Estimation with Test-Time Data-Augmentation, arXiv preprint arXiv:2006.16705
32. M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, CVPR 2018.
33. M. Haris, G. Shakhnarovich, N. Ukita, Recurrent back-projection network for video super-resolution, CVPR 2019.
34. C. Siagian and L. Itti. Rapid biologically inspired scene classification using features shared with visual attention. PAMI,29(2):300–312, Jan. 2007.
35. M. Haris, G. Shakhnarovich, N. Ukita, Space-Time-Aware Multi-Resolution Video Enhancement, CVPR 2010.
36. H. Jiang, G. Larsson, M. Maire, G. Shakhnarovich, E. Learned-Miller, Self-supervised relative depth learning for urban scene understanding, ECCV 2018.

ACCRONYMS

3D	Three Dimensional
CNN	Convolutional Neural Network
DARPA	Defense Advanced Research Projects Agency
DIODE	Dense Indoor/Outdoor DEpth
FC	Fully Connected
GIST	Generalized Search Tree
L2M	Lifelong Learning Machines
MLP	Multi-Layer Perceptron
MRR	Mean Reciprocal Rank
MS COCO	Microsoft Common Objects in Context
RGB	Red, Green, Blue channels
RGBD	Red, Green, Blue, Depth
SLAM	Simultaneous Location and Mapping
SPL	Success Path Length
ST-SR	Space Time – Super Resolution