SERDP Project MR18-1444 (Phase 2)
Final Report

Sonar-Based Deep Learning for Underwater UXO Remediation

David P. Williams
david.williams@cmre.nato.int

NATO STO Centre for Maritime Research and Experimentation (CMRE)

7 May 2021

Version 1.0

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
07/05/2021	SERDP Final Report	

4. TITLE AND SUBTITLE

Sonar-Based Deep Learning for Underwater UXO Remediation - Phase II

5a. CONTRACT NUMBER
W74RDV72490702, W74RDV80675566

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Williams, David P.

5d. PROJECT NUMBER
MR18-1444

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NATO STO Centre for Maritime Research and Experimentation(CMRE)
Viale San Bartolomeo 400
19126 La Spezia (SP)
Italy

8. PERFORMING ORGANIZATION REPORT NUMBER
MR18-1444

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Strategic Environmental Research and Development Program
4800 Mark Center Drive, Suite 17D08
Alexandria, VA 22350-3605

10. SPONSOR/MONITOR'S ACRONYM(S)
SERDP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
MR18-1444

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The primary objective of this project was to develop novel unexploded ordnance (UXO) detection and classification algorithms specifically for volumetric sonar data from two experimental systems, the Sediment Volume Search Sonar and the Multi-Sensor Towbody. Because no automatic target recognition (ATR) algorithms previously existed for these two new systems, the methods developed here addressed a capability gap. The general-purpose detection algorithm that was created exploited the concept of integral images to flag suspicious regions in a given data volume in a fast, computationally efficient manner. The follow-on classification algorithm was based on deep-learning techniques, specifically deep convolutional neural networks that were extended to function with three-dimensional (i.e., volumetric)input data cubes. The developed algorithms were assessed using large sets of SVSS data, and they were also applied to modest amounts of data from the MuST system. Preliminary results showed the promise of the approaches for detecting and classifying both proud and buried targets in measured volumetric sonar data.

15. SUBJECT TERMS

Unexploded Ordnance (UXO), Synthetic Aperture Sonar (SAS), Detection, Classification, Convolutional Neural Networks (CNNs), Sediment Volume Search Sonar (SVSS), Multi-Sensor Towbody (MuST).

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			David P. Williams
UNLCASS	UNLCASS	UNLCASS	UNLCASS	73	**19b. TELEPHONE NUMBER** *(include area code)* +39 0187-527-439

Contents

List of Figures

List of Tables

List of Acronyms

APL-UW	Applied Physics Laboratory - University of Washington
ATR	Automatic Target Recognition
AUC	Area Under the Curve
AUV	Autonomous Underwater Vehicle
BB	Broadband
BRAC	Base Realignment and Closure
CMRE	Centre for Maritime Research and Experimentation
CNN	Convolutional Neural Network
DoD	Department of Defense
ESTCP	Environmental Security Technology Certification Program
FUDS	Formerly Used Defense Sites
GPU	Graphics Processing Unit
HF	High Frequency
LF	Low Frequency
MCM	Mine Countermeasures
MIP	Maximum Intensity Projection
ML	Machine Learning
MR	Munitions Response
MuST	Multi-Sensor Towbody
NATO	North Atlantic Treaty Organization
NSWC-PCD	Naval Surface Warfare Center - Panama City Division
PC SWAT	Personal Computer Shallow Water Acoustics Tool-set
PI	Principal Investigator
PSU-ARL	Penn State University Applied Research Laboratory
ReLU	Rectified Linear Unit
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SAS	Synthetic Aperture Sonar
SEED	SERDP Exploratory Development
SERDP	Strategic Environmental Research and Development Program
SON	Statement of Need
STO	Science and Technology Organization
SVSS	Sediment Volume Search Sonar
SWAT	Shallow Water Acoustics Tool-set
TIER	Target-in-the-Environment Response
TIERSWAT	TIER SWAT
UUV	Unmanned Underwater Vehicle
UXO	Unexploded Ordnance

Keywords

Unexploded Ordnance (UXO), Synthetic Aperture Sonar (SAS), Detection, Classification, Convolutional Neural Networks (CNNs), Sediment Volume Search Sonar (SVSS), Multi-Sensor Towbody (MuST).

Acknowledgments

The author thanks the following:

- the Strategic Environmental Research and Development Program (SERDP) for funding this project;
- Daniel Brown at the Penn State University Applied Research Laboratory (PSU-ARL) for providing the Sediment Volume Search Sonar (SVSS) data;
- Timothy Marston at the Applied Physics Laboratory - University of Washington (APL-UW) for providing the Multi-Sensor Towbody (MuST) data;
- Kevin Williams at the Applied Physics Laboratory - University of Washington (APL-UW) for providing the TIERSWAT data.

Abstract

An unfortunate legacy of former military activities at sites designated for base realignment and closure (BRAC) and at Formerly Used Defense Sites (FUDS) is the contamination of aquatic environments with military munitions. In the United States, more than 400 underwater sites, spanning an area in excess of 10 million acres, potentially contain such munitions. The presence of these munitions is a serious threat to both humans and the environment, so remediation is necessary. But the return of these contaminated waters to public use is contingent upon the analysis and assessment of wide-area and detailed underwater surveys. Therefore, the Department of Defense (DoD) has an express need for the development of technologies that will enable the detection and classification, at high probability, of military munitions found at underwater sites.

The primary objective of this project was to develop novel unexploded ordnance (UXO) detection and classification algorithms specifically for volumetric sonar data from two experimental systems, the Sediment Volume Search Sonar (SVSS) and the Multi-Sensor Towbody (MuST). Because no automatic target recognition (ATR) algorithms previously existed for these two new systems, the methods developed here addressed a capability gap. The general-purpose detection algorithm that was created exploited the concept of integral images to flag suspicious regions in a given data volume in a fast, computationally efficient manner. The follow-on classification algorithm was based on deep-learning techniques, specifically deep convolutional neural networks (CNNs) that were extended to function with three-dimensional (*i.e.,* volumetric) input data cubes. The developed algorithms were assessed using large sets of SVSS data, and they were also applied to modest amounts of data from the MuST system. Preliminary results showed the promise of the approaches for detecting and classifying both proud and buried targets in measured volumetric sonar data.

The work summarized in this report covers only one year of an envisioned four-year project that ended prematurely (due to an organization change of the PI). Nevertheless, the progress made during this abbreviated period provides a solid foundation from which to further this line of research. Because the algorithms were purposely developed to be functional with measured data from existing systems, they should be readily deployable in a short time frame for use in actual remediation efforts. This result can be achieved by executing the remainder of the original project plan, which includes rigorous testing at new SERDP UXO test-bed sites.

The successful culmination of the project's approach should enable the attainment of higher probabilities of detection and classification, at much lower false alarm rates, than is possible with existing approaches. As a result, the application of these machine-learning algorithms to sonar data collected at potentially contaminated underwater sites can guide remediation efforts to effect savings.

1 Executive Summary

1.1 Introduction

An unfortunate legacy of former military activities at sites designated for base realignment and closure (BRAC) and at Formerly Used Defense Sites (FUDS) is the contamination of aquatic environments with military munitions. In the United States, more than 400 underwater sites, spanning an area in excess of 10 million acres, potentially contain such munitions. The presence of these munitions is a serious threat to both humans and the environment, so remediation is necessary. But the return of these contaminated waters to public use is contingent upon the analysis and assessment of wide-area and detailed underwater surveys. Therefore, the Department of Defense (DoD) has an express need for the development of technologies that will enable the detection and classification, at high probability, of military munitions found at underwater sites.

The primary objective of this project was to develop novel unexploded ordnance (UXO) detection and classification algorithms specifically for volumetric sonar data from two experimental systems, the Sediment Volume Search Sonar (SVSS) and the Multi-Sensor Towbody (MuST). Because no automatic target recognition (ATR) algorithms previously existed for these two new systems, the methods developed here addressed a capability gap. The general-purpose detection algorithm that was created exploited the concept of integral images to flag suspicious regions in a given data volume in a fast, computationally efficient manner. The follow-on classification algorithm was based on deep-learning techniques, specifically deep convolutional neural networks (CNNs) that were extended to function with three-dimensional (*i.e.,* volumetric) input data cubes. The developed algorithms were assessed using large sets of SVSS data, and they were also applied to modest amounts of data from the MuST system. Preliminary results showed the promise of the approaches for detecting and classifying both proud and buried targets in measured volumetric sonar data.

The work summarized in this report covers only one year of an envisioned four-year project that ended prematurely (due to an organization change of the PI). Nevertheless, the progress made during this abbreviated period provides a solid foundation from which to further this line of research. Because the algorithms were purposely developed to be functional with measured data from existing systems, they should be readily deployable in a short time frame for use in actual remediation efforts. This result can be achieved by executing the remainder of the original project plan, which includes rigorous testing at new SERDP UXO test-bed sites.

1.2 Objectives

After the successful execution of SEED project MR18-1444 [1], this project was in response to SERDP Statement of Need (SON) "MR19-1444-F1: MR18- 1444 Follow-on," in the Munitions

Response (MR) Program Area. Specifically, the research fell under the topic of "Wide Area and Detailed Surveys," by addressing the need for technologies that would enable the detection and classification, at high probability, of military munitions found at underwater sites.

The primary objective of this project was to develop novel unexploded ordnance (UXO) detection and classification algorithms specifically for volumetric sonar data from two experimental systems, the SVSS developed under SERDP project MR-2545 (PI: D. Brown) and the MuST from ESTCP project MR18-5004 (PI: K. Williams). An auxiliary objective was to explore the use of limited-scope experiments within a CNN framework in order to improve the explainability of the resulting deep-learning classifiers.

Developing the detection and classification algorithms was expected to fill a capability gap, and as a consequence, enable more efficient remediation efforts at contaminated sites.

1.3 Technical Approach

This project made extensive use of data collected by the SVSS [2] and MuST [3, 4] systems to aid in the development and assessment of various ATR algorithms. The systems are complementary in the sense that each focuses on different water depth regimes. The SVSS system is designed to support UXO remediation in shallow water (1-5 m), while the MuST system is intended to operate in deeper water (6-40 m). Both systems employ low-frequency sonar to enable the production of volumetric SAS imagery, with this in turn facilitating the detection of proud and buried objects.

Because the two systems produce roughly comparable data, an effort was made to design flexible algorithms that could function with data from either system (as well as other similar systems developed in the future). Eschewing system-specific algorithms also makes the exploitation of data from multiple systems in a joint manner – *e.g.,* for classifier training – more feasible.

The main algorithms developed for the SVSS and MuST data addressed normalization, detection, and classification. Additional experiments were conducted for limited-scope CNN classifiers using simulated acoustic-color sonar data. Here we provide extremely cursory summaries of these new algorithms.

Given a raw 3-d data cube of beamformed sonar returns, the normalization procedure developed for volumetric sonar data is as follows. First determine the plane of the interface in the image volume for which returns dominate. Then determine and remove the sub-volume of the data cube contaminated by multipath interference associated with the dominant interface. For a given cross-track position and a given distance from the dominant interface, normalize by the median value. Similarly, for a given along-track position and a given distance from the dominant interface, normalize by the median value. Finally, perform a logarithmic transformation.

The object detection task we concerned ourselves with is a classic remote-sensing problem of locating a target signal amid a noisy background [5]. Our proposed algorithm computes a local estimate of the background intensity around each voxel, which are the potential target signals. If the target-to-background ratio exceeds a set threshold, the voxel is flagged. And if a connected volume of such flagged voxels exceeds the minimum size of objects of interest, a discrete alarm is generated. The summed intensity over a rectangular volume needed for the target or background estimates was computed in an extremely efficient manner using integral images [6].

A CNN [7] is a sophisticated classification algorithm that customarily ingests an image as input and produces scalar outputs corresponding to the probabilities of belonging to each

class under consideration (*e.g.,* target and clutter). Our classification approach was based on *3-d* CNNs in which the input data is a 3-d data cube, rather than a 2-d image. The general architecture design largely follows our previous work [8], but extends it to three dimensions. We also chose to employ an *ensemble* of eight CNNs, each of which has a unique network architecture, so that complementary clues may be discovered and exploited by the CNNs. This in turn improves the overall robustness of the classification scheme. Despite the enormous size of the input data cubes, the use of extremely small networks allows the CNNs to be trained with modest computational resources, and vitally, avoids computer-memory constraint issues. A CNN-based classification approach is particularly apropos for this data modality because hand-crafting features is challenging, and CNNs effectively obviate this process.

The aforementioned CNNs were used to discriminate between two general classes of objects, each of which exhibits considerable intra-class diversity in terms of object size, shape, composition, and burial state. As a result, the features that a trained CNN will rely on to make a prediction will be a complex combination that is not easily disentangled. Instead, designing specially controlled experiments can provide a way to learn principled, explainable features that can be tied directly to the wave phenomena of the physics involved. The idea is to develop a CNN classifier in which the two classes are *not* UXO and non-UXO, but rather whether or not a specific object has a certain attribute. As a result, any clues that the CNN uncovers should be due to the single variable that differs. Here we exploited this limited-scope experiment concept (using simulated acoustic-color data) to train CNNs to discriminate air-filled objects from water-filled objects.

1.4 Results and Discussion

This section presents a select set of results of the previously described algorithms.

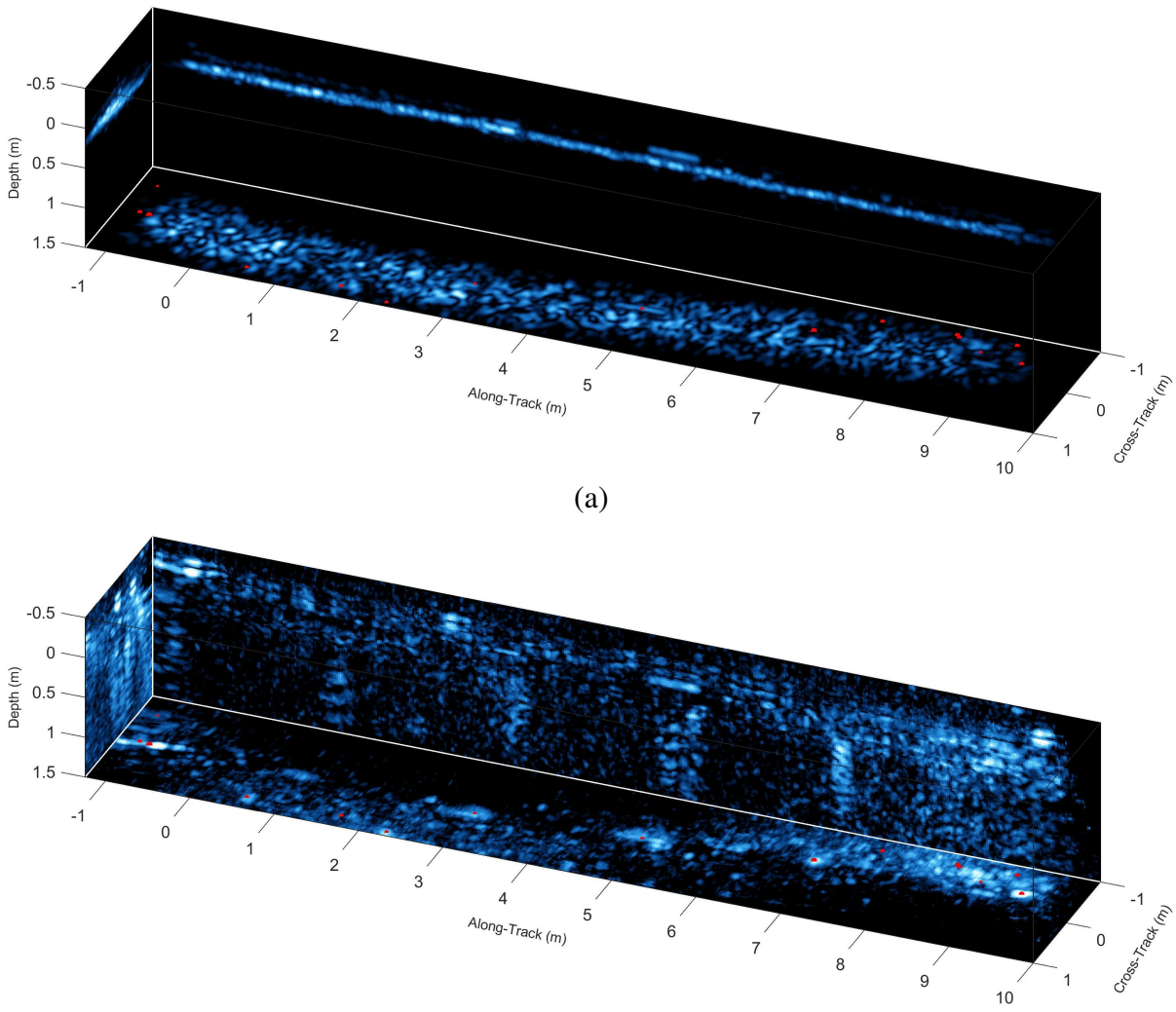
1.4.1 Data Normalization

The challenge of visualizing a 3-d data cube often leads to the use of a 2-d maximum intensity projection (MIP), which collapses the imagery along one of its principal axes by retaining the highest intensity voxels along that axis [9]. A typical data cube from the SVSS system, before and after the proposed normalization algorithm, is shown in Fig. 1.1 as a set of three 2-d MIPs in a common reference frame. In Fig. 1.1(a), the dominant interface return obscures target signatures, whereas in Fig. 1.1(b) it can be observed that the normalization procedure amplifies the signals of interest, including elastic target returns. Although not the principal objective, the normalization method also facilitates human interpretability of the data.

1.4.2 Detection

For the results reported in this work, the SVSS system was used to collect data at three sites in the United States, two distinct locations in the Fosters Joseph Sayers Reservoir in Pennsylvania and one location at the Aberdeen Test Center in Maryland. At the Sayers sites, there existed an upper layer of approximately 8 cm of silt atop a clay base; site "A" had a 1.3 m water depth, while site "B" had a 3.0 m water depth. The Aberdeen location, site "C," featured a *sloping* sediment of sand, resulting in a water depth that ranged from 1.0 m to 2.5 m. The shallow water meant multipath interference was not insignificant.

The sites were reservoirs that could be drained to facilitate target emplacement. So prior to data collection, various man-made objects were deployed, including aluminum cylinders,



(a)

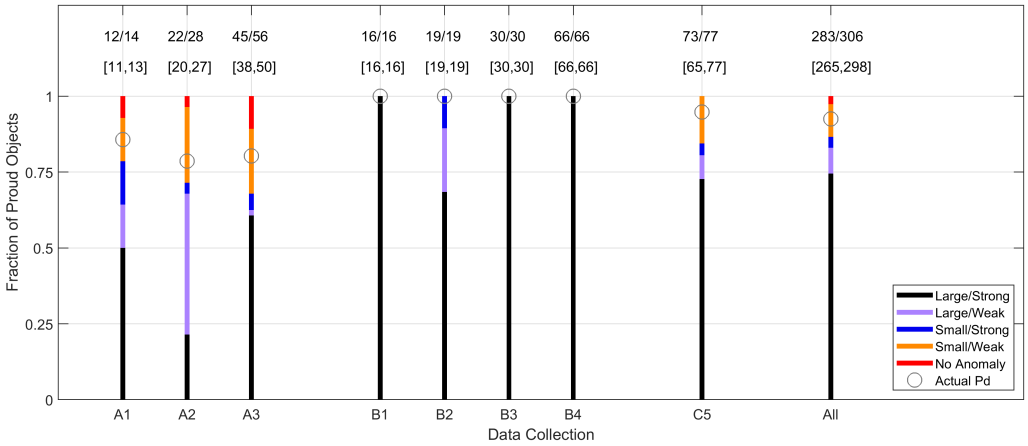
(b)

Fig. 1.1. An example SVSS volumetric scene image displayed as a trio of MIPs, when the data is (a) raw or (b) normalized. Algorithm detections are marked on the depth MIPs with red dots.

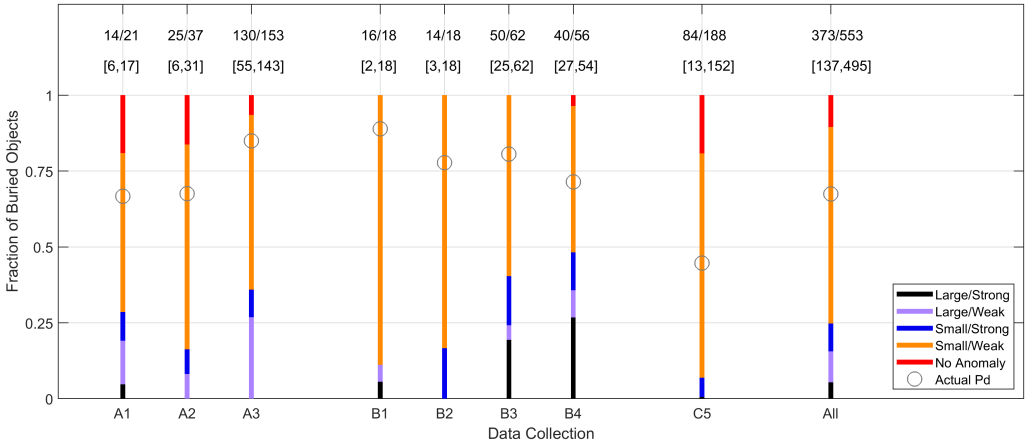
steel pipes, steel shot puts, concrete blocks, and an assortment of munitions with diameters ranging from 20 mm to 155 mm. Some objects were placed proud on the sediment, while others were buried to various depths (up to 60 cm) below the water-sediment interface. Data collections at Sayers took place in 2019 at the following times (nominally labeled "1" through "4," respectively): June, August, early November, and late November. The Aberdeen collection (labeled "5") occurred in March 2020.

The complex physics at play underwater and in the subsediment volume suggest that the collected 3-d data might not always support target detection, regardless of the algorithm employed. Factors such as sediment attenuation, interface scattering levels, the presence of gas bubbles in the sediment, and the relationship between sensor resolution and target dimensions mean that the upper limit on detection capability will likely be less than unity. To assess this possibility, the data from each target opportunity was first visually examined and rated in terms of anomaly size (large or small) and strength (strong or weak) in the imagery. Anomalies that were deemed both small and weak represent a "gray zone" in which detection may or may not actually be feasible.

With these human assessments as a backdrop, the performance of the proposed target detec-



(a)



(b)

Fig. 1.2. Performance of the detection algorithm for each SVSS data collection for (a) proud man-made targets and (b) buried man-made targets, along with the distribution of visual human assessment ratings. Above each bar are the numbers of targets detected vice opportunities, and in brackets the range of targets deemed detectable based on visual human assessment.

tion algorithm at eight distinct data collections, delineated by location and time, are shown in Fig. 1.2 for proud and buried targets. As can be seen from the figure, performance varies considerably across location (*cf.* collection letters), but also across time (*cf.* collection numbers), the latter variation suggesting strong environmental dependence (*e.g.,* water temperature, microbial activity). However, in all cases, the automated detection performance comported with the expected range based on visual inspection of the imagery.

Localized alarm data cubes (displayed as MIPs) extracted from Fig. 1.1 of four targets are shown in Fig. 1.3, along with corresponding object photographs taken during installation. (The 3-d alarm data cubes like in Fig. 1.3(a)-(d) are what would be the inputs to the subsequent CNN classifier.)

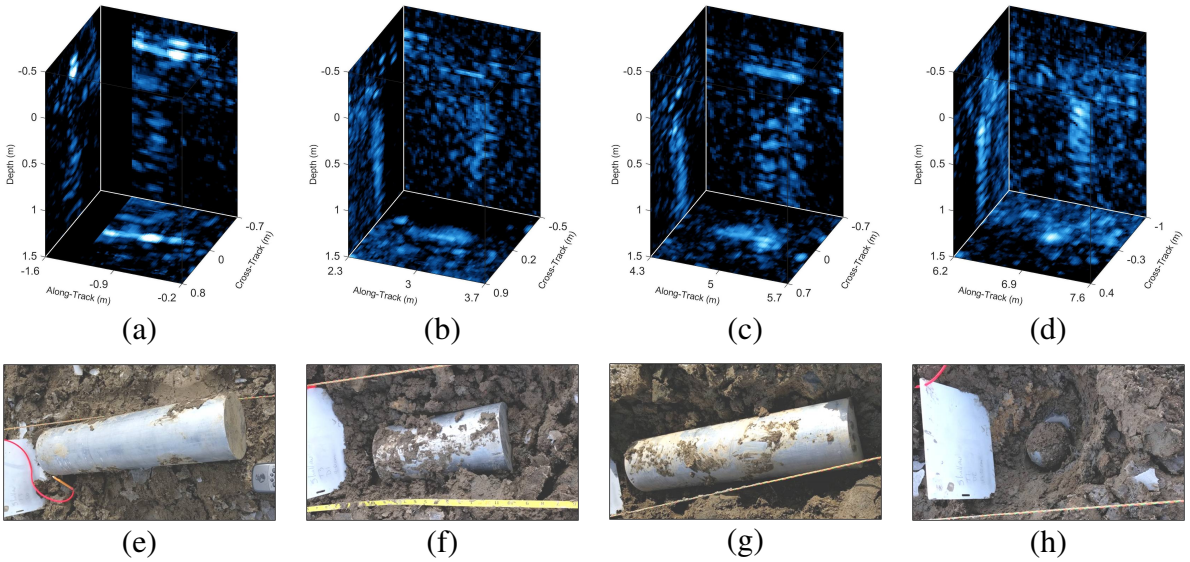


Fig. 1.3. (a)-(d) SVSS alarm cubes (each displayed as a trio of MIPs) of four targets extracted from Fig. 1.1, and (e)-(h) photographs of the objects during installation (pre-burial). The objects, along with the human-assessments of the sonar imagery in parentheses, are: (a) 4:1 solid aluminum cylinder proud (Large/Strong), (b) 2:1 solid aluminum cylinder buried 5 cm (Large/Weak), (c) 4:1 solid aluminum cylinder buried 3 cm (Large/Strong), (d) 10.2 cm diameter steel shot put buried 19 cm (Small/Strong); the cylinders have 15.2 cm diameters.

For the MuST system, we possessed neither ground-truth information nor sufficient amounts of data to make statistically significant detection-performance assessments. But applying the detection algorithm to the modest amounts of data we did have resulted in seemingly reasonable alarms. An example set of such alarms is shown in Fig. 1.4.

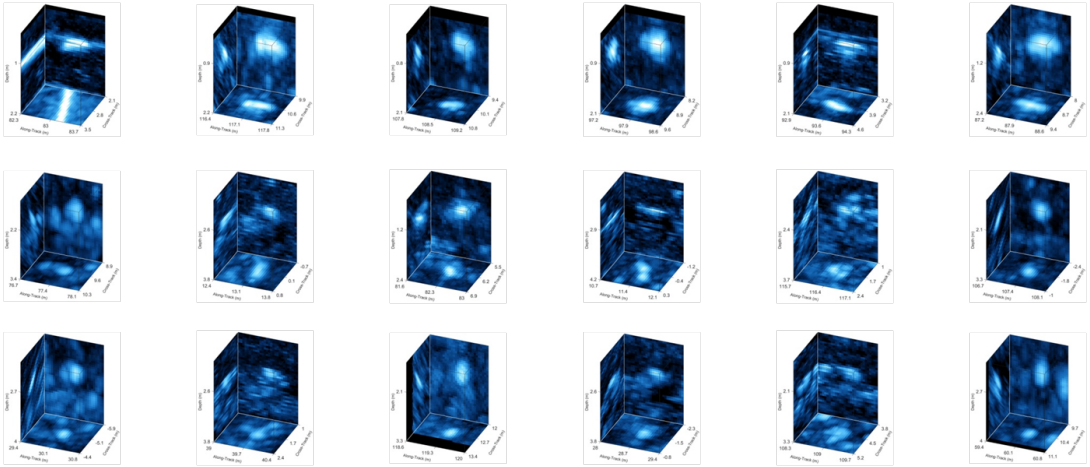


Fig. 1.4. Example alarms generated by the detection algorithm when applied to a MuST scene.

The combined detection performance from pooling the proud and buried targets of the various SVSS data collections is shown in Fig. 1.5; a "score" based on the geometric mean of a size feature and an intensity feature is used to order the alarms. Performance is displayed in terms of receiver operating characteristic-like (ROC-like) curves, where the probability of false alarm is replaced by false alarm rate (per unit volume). It should be noted that these results are for a considerable span of object sizes, some of which are even less than the sensor resolution.

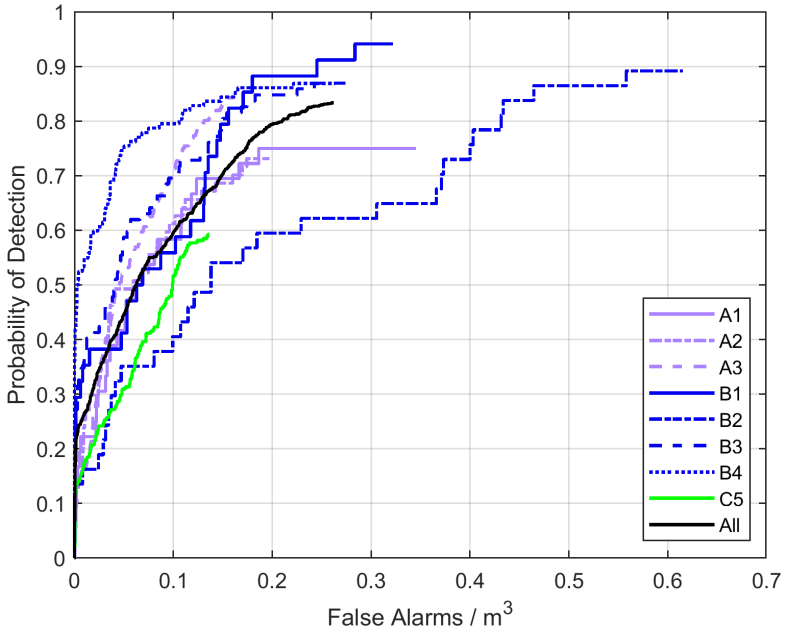


Fig. 1.5. Overall performance of the detection algorithm for each SVSS data collection.

1.4.3 Classification

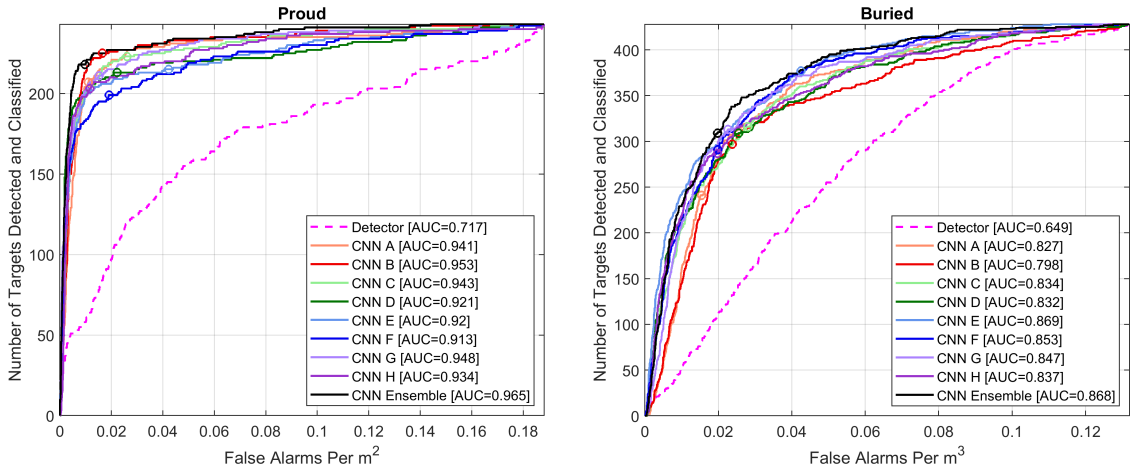
Our main interest in this section is to examine the ability of the CNN-based approach to successfully perform *classification*. Therefore, when presenting performance, the experiments assume that all targets were successfully flagged in the detection stage. Thus, the maximum possible area under an ROC curve (AUC) [10], a scalar summary measure of performance, is always unity. (A perfect classifier would have an AUC of unity.) To obtain the overall probability of both successful detection and correct classification, the vertical axis of the ROC-like curves would need to be scaled by the inverse of the total number of targets present in the scene imagery.

Performance is presented in the form of full ROC-*like* curves, with the abscissa corresponding to the more informative false alarm *rate* instead of the *probability* of false alarm. The probability of false alarm is the probability of incorrectly classifying clutter as a target; the false alarm rate is the number of such incorrect classifications per image area or volume. When considering only proud objects, the false alarm rate is given per image (seafloor) area; when considering only buried objects, the false alarm rate is given per image volume.

Performance of the eight trained CNNs, as well as the ensemble, is shown in terms of ROC-like curves in Fig. 1.6. The AUC (for the corresponding ROC curve) is also provided in the legend. To provide a baseline measure of performance, the performance of the 3-d detection algorithm is also shown. While the full curves are informative, in practice, one must select a single operating point at which to make predictions. Therefore, on each CNN curve, the operating point corresponding to the natural decision threshold of $\tau = 0.5$ is also marked.

As can be observed from Fig. 1.6, the 3-d CNNs greatly outperform the simple baseline detector, as would be expected. But more interestingly, the use of the ensemble of networks proves beneficial and removes the necessity of selecting a single best CNN architecture to employ. The complementary nature of the CNNs, and the unique clues that each uncovers and exploits to make predictions, leads to reduced false alarm rates. As a result, the ensemble approach can directly translate into cost savings during UXO remediation efforts.

It is also worth noting that these classification results are based on using only a single data



(a) (b)

Fig. 1.6. Classification performance on the SVSS test data set in terms of ROC-like curves for (a) only proud objects and (b) only buried objects. The operating point for a $\tau = 0.5$ threshold is marked with a circle.

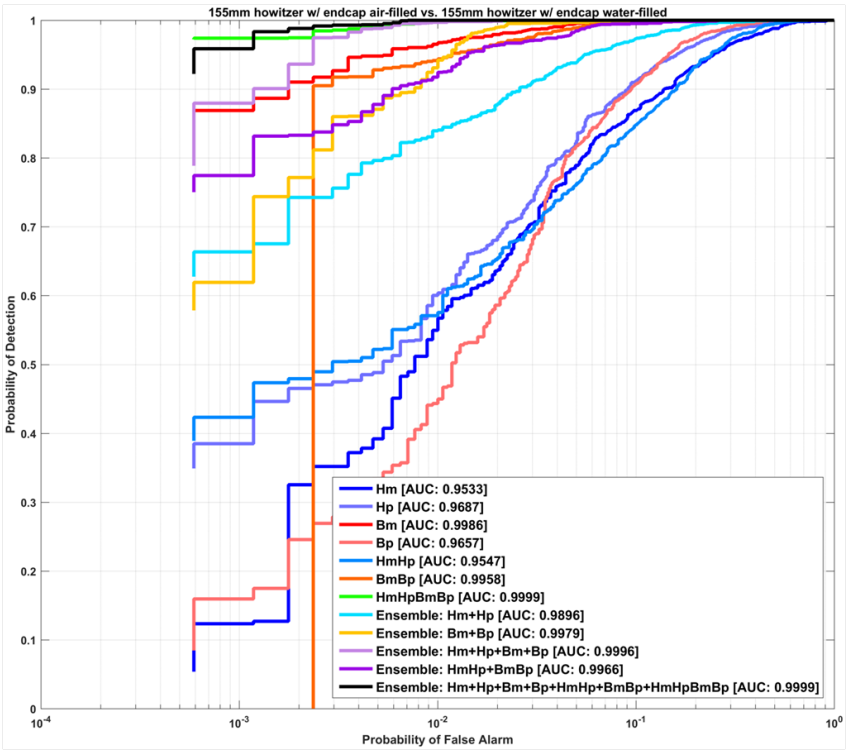
representation, namely the 3-d imagery, as input to the CNNs. A worthwhile avenue to explore is the use of alternative data representations (*e.g.,* acoustic color) in which complementary discriminative clues would be made more accessible to the CNN.

1.4.4 Limited-Scope Classification

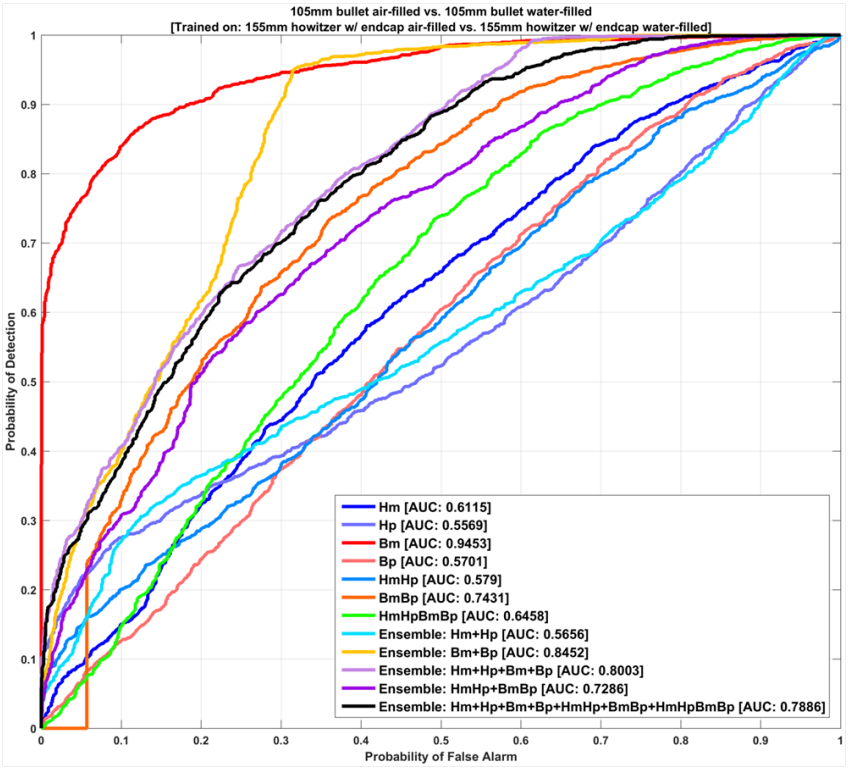
Explainability of classifier predictions can be a useful tool to secure human trust of an algorithm's decision-making process. With an eye toward that long-term goal, experiments were conducted to assess the feasibility of employing CNNs with *simulated* multi-representation acoustic-color sonar data for discriminating air-filled objects from water-filled objects.

CNNs were trained for seven different combinations of input data representations, in terms of frequency band and component (*i.e.,* magnitude or phase). The performance of the CNNs for the different input data representations is presented in Fig. 1.7. Additionally, ensembles that leverage different combinations of the CNNs (by averaging their individual predictions) are also considered. The AUC of each case is shown in the legend.

In Fig. 1.7(a), it can be seen that a CNN trained on any of the representations was able to successfully discriminate the air-filled munitions from the water-filled munitions. However, in this case, the training data and test data – although disjoint – all corresponded to the same object, namely 155 mm munitions. As a result, the features that the CNN learned to rely on when discriminating interior fill might be tied to this specific object. A stronger test of CNN generalization ability is shown in Fig. 1.7(b), where the test objects are 105 mm munitions. In this figure, it can be seen that the CNN trained using broadband magnitude acoustic-color data (the red curve) was still able to reliably classify the test objects' fill. This preliminary result suggests that this CNN indeed leverages attributes associated with the object's interior fill, and more importantly, that these clues are ostensibly present in objects other than the specific type used for training.



(a)



(b)

Fig. 1.7. For different input data representations, classification performance for discriminating air-filled and water-filled objects using 155 mm munitions as training data, and then (a) testing on other 155 mm munitions or (b) testing on 105 mm munitions. (Note the logarithmic horizontal-axis in (a).)

1.5 Implications for Future Research and Benefits

The work summarized in this report covers only one year of an envisioned four-year project that ended prematurely. Nevertheless, the progress made during this abbreviated period provides a solid foundation from which to further this line of research. Because the algorithms were purposely developed to be functional with measured data from existing systems, they should be readily deployable in a short time frame for use in actual remediation efforts. This result can be achieved by executing the remainder of the original project plan, which includes rigorous testing at new SERDP UXO test-bed sites.

The preliminary results already achieved regarding data normalization, detection, and classification were promising. It was demonstrated how the initial normalization plays a vital role in accentuating the difference between target responses and the surrounding environment. This result then facilitated the use of a very computationally efficient detection algorithm, but it also enabled human visualization of the data. The new detection algorithm was capable of identifying and isolating the small fractions of the data cubes that contain information relevant to the UXO remediation problem. Then, the more sophisticated classifier based on deep-learning techniques was shown to provide even better discrimination capability to reduce false alarm rates further. Importantly, it was demonstrated that the CNN-based approach could successfully scale to 3-d data products without incurring computationally prohibitive costs. And finally, the parallel effort to explore limited-scope CNNs to engender explainable classification predictions showed promise using simulated data.

The fundamental limitations on performance imposed by the combination of sensor, target, and environment should be recognized. It is important to identify the regimes in which the physics simply does not support successful detection and classification, regardless of the algorithm employed. In this vein, it is recommended that additional measured data are obtained from different environments so that the developed algorithms can be assessed more fully. The importance of having ground-truth information for training classifiers and evaluating algorithm performance also cannot be emphasized enough. Thus far, only a modest amount of MuST data was available to work with, so evaluating the algorithms on larger MuST data sets should be a priority in the future. Investigating additional alternative data representations beyond the image domain, such as acoustic color, within the CNN context is also a potentially fruitful avenue for further research.

The algorithms presented here addressed a capability gap as they were developed expressly for two new systems, the SVSS and MuST, for which no ATR algorithms previously existed. Provided further refinement and rigorous testing of the algorithms are undertaken successfully, there is great potential for these methods to be leveraged in remediation efforts at contaminated underwater sites. And in the event that they are indeed deployed, fewer resources should be spent investigating harmless clutter and the cost of remediation should decrease substantially.

2 Objective

After the successful execution of SEED project MR18-1444 [1], this project was in response to SERDP Statement of Need (SON) "MR19-1444-F1: MR18- 1444 Follow-on," in the Munitions Response (MR) Program Area. Specifically, the research fell under the topic of "Wide Area and Detailed Surveys," by addressing the need for technologies that would enable the detection and classification, at high probability, of military munitions found at underwater sites.

The primary objective of this project was to develop novel unexploded ordnance (UXO) detection and classification algorithms specifically for volumetric sonar data from two experimental systems, the Sediment Volume Search Sonar (SVSS) developed under SERDP project MR-2545 (PI: D. Brown) and the Multi-Sensor Towbody (MuST) from ESTCP project MR18-5004 (PI: K. Williams). An auxiliary objective was to explore the use of limited-scope experiments within a convolutional neural network (CNN) framework in order to improve the explainability of the resulting deep-learning classifiers.

Developing the detection and classification algorithms was expected to fill a capability gap, and as a consequence, enable more efficient remediation efforts at contaminated sites.

3 Background

An unfortunate legacy of former military activities is the contamination of aquatic environments with UXO. In shallow water, proud and buried munitions pose a particular threat to both humans and the environment, so remediation is necessary. To address this pressing issue, several low-frequency sonar systems – importantly, on mobile platforms – have recently been developed [4, 11, 12]. These downward-looking synthetic aperture sonar (SAS) systems, designed to achieve sediment penetration, provide high-resolution three-dimensional (3-d) *volumetric* imagery below the seafloor, making large-scale buried object detection newly feasible. (Other *side-looking* low-frequency sonars [13, 14] generate only 2-d imagery.)

With the introduction of this new sensor modality, there is now a need for automated detection and classification algorithms that can efficiently process enormous 3-d images to rapidly flag suspicious objects for closer inspection during remediation efforts. Relying on humans to visually assess these new data products is both inherently challenging and inefficient.

The detection of objects of interest in 3-d data cubes is a goal shared across diverse disciplines. For example, previous work with tomographic medical imaging scans [15, 16], video scenes possessing a temporal component [17, 18], hyperspectral data comprising multiple frequency bands [19, 20], and ground-penetrating radar for buried objects on land [21, 22] all exploit 3-d data cubes that can be mined for valuable information. But the particularities of *sonar* sensors and the unique challenges of the underwater environment warrant a new approach specially tailored to the physics involved.

The principal detection algorithm used for volumetric sonar imagery in the literature [11] simply sums the energy at each voxel over multiple pings and compares this quantity to a threshold. A second detection approach [23] for the same system compares moments of the voxel intensities to the local background. But neither approach addresses the complications associated with *near-normal incidence* returns from a strictly downward-looking system. In addition, there have been no classification approaches developed specifically for volumetric sonar data.

In this work, we propose novel detection and classification methods and demonstrate their promise on real, measured sonar data collected by newly developed sonar systems that are operationally capable. The approaches are general in the sense that they can be employed for wide classes of objects of interest, but here they are presented in the context of underwater UXO. The proposed detection and classification algorithms fill a critical capability gap that makes the use of this new class of sensors feasible for real-world UXO remediation operations.

4 Materials and Methods

4.1 Volumetric Sonar Data

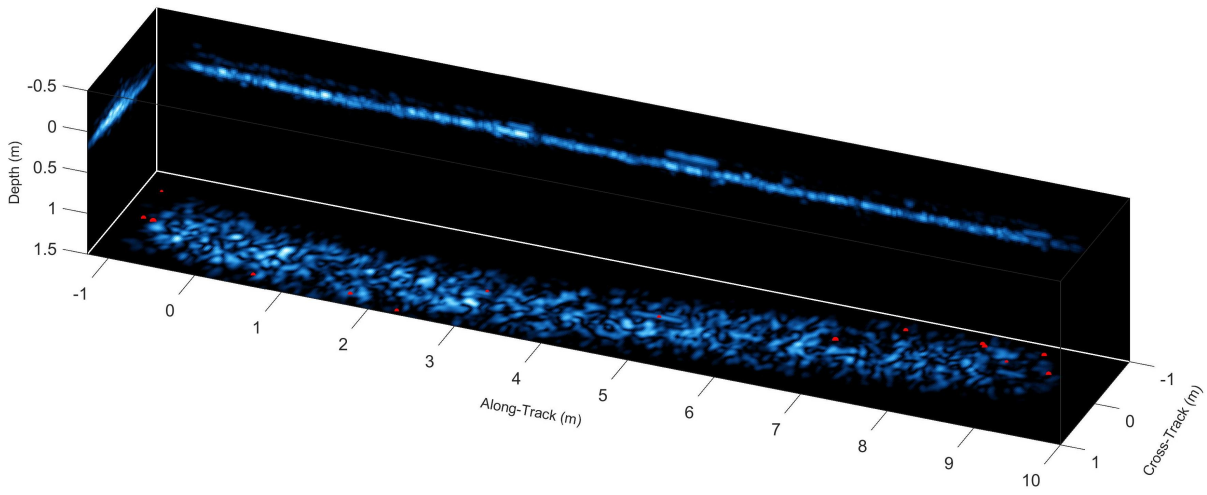
4.1.1 SVSS Data

The SVSS [12] is a recently developed low-frequency sonar system designed to address the UXO remediation problem in shallow water environments (*i.e.,* depths less than 5 m). The experimental system features multiple transmitters and a 2-d receiver array that collectively enable the production of 3-d SAS imagery to facilitate the detection of proud and buried objects. The frequency band of operation is approximately 20-35 kHz. A time-domain back-projection beamformer [24] is used to transform the raw sonar time series returns into 3-d volumetric imagery comprising voxels that span 2 cm in each dimension. Further details of the system can be found in recently concluded SERDP project MR-2545 (PI: D. Brown) [2, 12, 25].

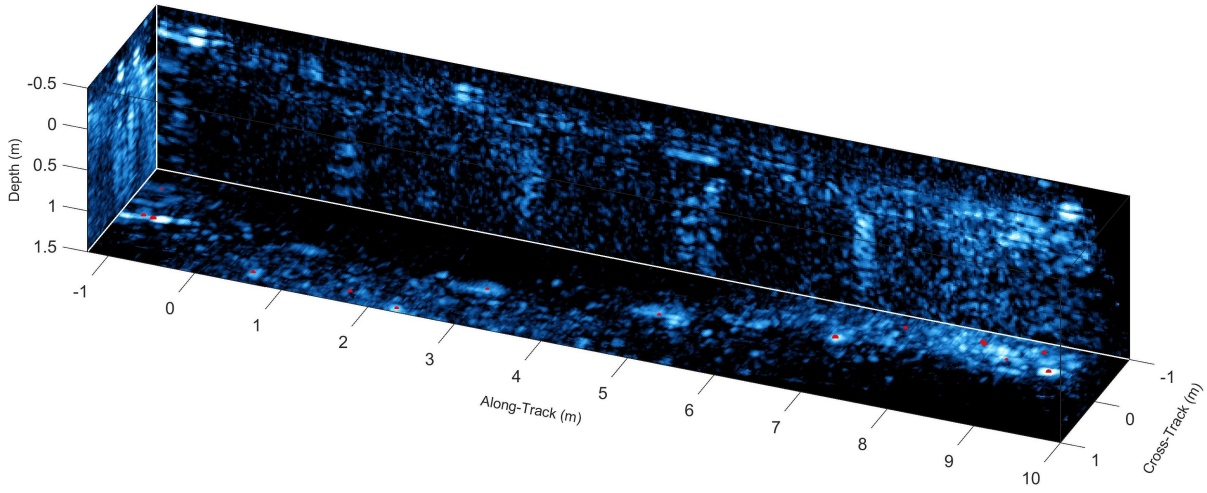
The challenge of visualizing a 3-d data cube often leads to the use of a 2-d maximum intensity projection (MIP), which collapses the imagery along one of its principal axes by retaining the highest intensity voxels along that axis [9]. A typical data cube from the SVSS system, before and after the proposed normalization algorithm described later in Sec. 4.2.1 is shown in Fig. 4.1 as a set of three 2-d MIPs in a common reference frame. In Fig. 4.1(a), the dominant interface return obscures target signatures, whereas in Fig. 4.1(b) it can be observed that the normalization procedure amplifies the signals of interest, including elastic target returns. Localized alarm data cubes (displayed as MIPs) extracted from Fig. 4.1 of four targets are shown in Fig. 4.2, along with corresponding object photographs taken during installation. (The 3-d alarm data cubes like in Fig. 4.2(a)-(d) are what would be the inputs to a subsequent CNN.)

For the results reported in this work, the SVSS system was used to collect data at three sites in the United States, two distinct locations in the Fosters Joseph Sayers Reservoir in Pennsylvania and one location at the Aberdeen Test Center in Maryland. At the Sayers sites, there existed an upper layer of approximately 8 cm of silt atop a clay base; site "A" had a 1.3 m water depth, while site "B" had a 3.0 m water depth. The Aberdeen location, site "C," featured a *sloping* sediment of sand, resulting in a water depth that ranged from 1.0 m to 2.5 m. The shallow water meant multipath interference was not insignificant.

The sites were reservoirs that could be drained to facilitate target emplacement. So prior to data collection, various man-made objects were deployed, including aluminum cylinders, steel pipes, steel shot puts, concrete blocks, and an assortment of munitions with diameters ranging from 20 mm to 155 mm. Some objects were placed proud on the sediment, while others were buried to various depths (up to 60 cm) below the water-sediment interface. Data collections at Sayers took place in 2019 at the following times (nominally labeled "1" through "4," respectively): June, August, early November, and late November. The Aberdeen collection (labeled "5") occurred in March 2020.



(a)



(b)

Fig. 4.1. An example SVSS volumetric scene image, from site A, displayed as a trio of MIPs, when the data is (a) raw or (b) normalized. Algorithm detections are marked on the depth MIPs with red dots.

4.1.2 MuST Data

The SVSS and MuST systems are complementary in the sense that each focuses on different water depth regimes. While the SVSS system is designed to support UXO remediation in shallow water (1-5 m), the MuST system is intended to operate in deeper water (6-40 m). As a result, the cross-track swath of the MuST system is considerably larger. But like the SVSS system, the MuST system also produces 3-d data cubes of comparable resolution.

Detailed information regarding the MuST system can be found in concluded SERDP projects MR-2501 (PI: K. Williams) [3] and MR-2752 (PI: J. Sara) [26], as well as active SERDP/ESTCP projects MR18-5004 (PI: K. Williams) and MR18-1051 (PI: T. Marston) [4], so we refrain from providing more background here.

A modest amount of data collected in September 2019 at Sequim Bay, Washington, was used in this work. A typical data cube from the MuST system, before and after the proposed normalization algorithm described in Sec. 4.2.1, is shown in Fig. 4.3 as a set of three 2-d MIPs in a common reference frame. The protracted extent of the image in the along-track dimension is notable, as this creates an enormous data cube.

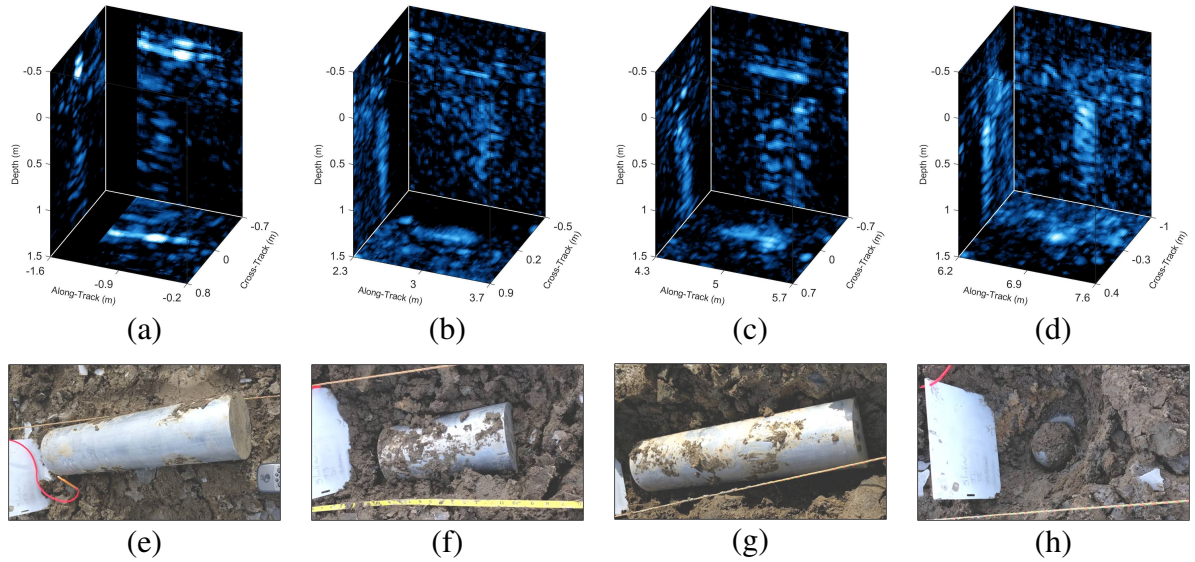


Fig. 4.2. (a)-(d) SVSS alarm cubes (each displayed as a trio of MIPs) of four targets extracted from Fig. 4.1, and (e)-(h) photographs of the objects during installation (pre-burial). The objects, along with human-assessments of the sonar imagery in parentheses, are: (a) 4:1 solid aluminum cylinder proud (Large/Strong), (b) 2:1 solid aluminum cylinder buried 5 cm (Large/Weak), (c) 4:1 solid aluminum cylinder buried 3 cm (Large/Strong), (d) 10.2 cm diameter steel shot put buried 19 cm (Small/Strong); the cylinders have 15.2 cm diameters.

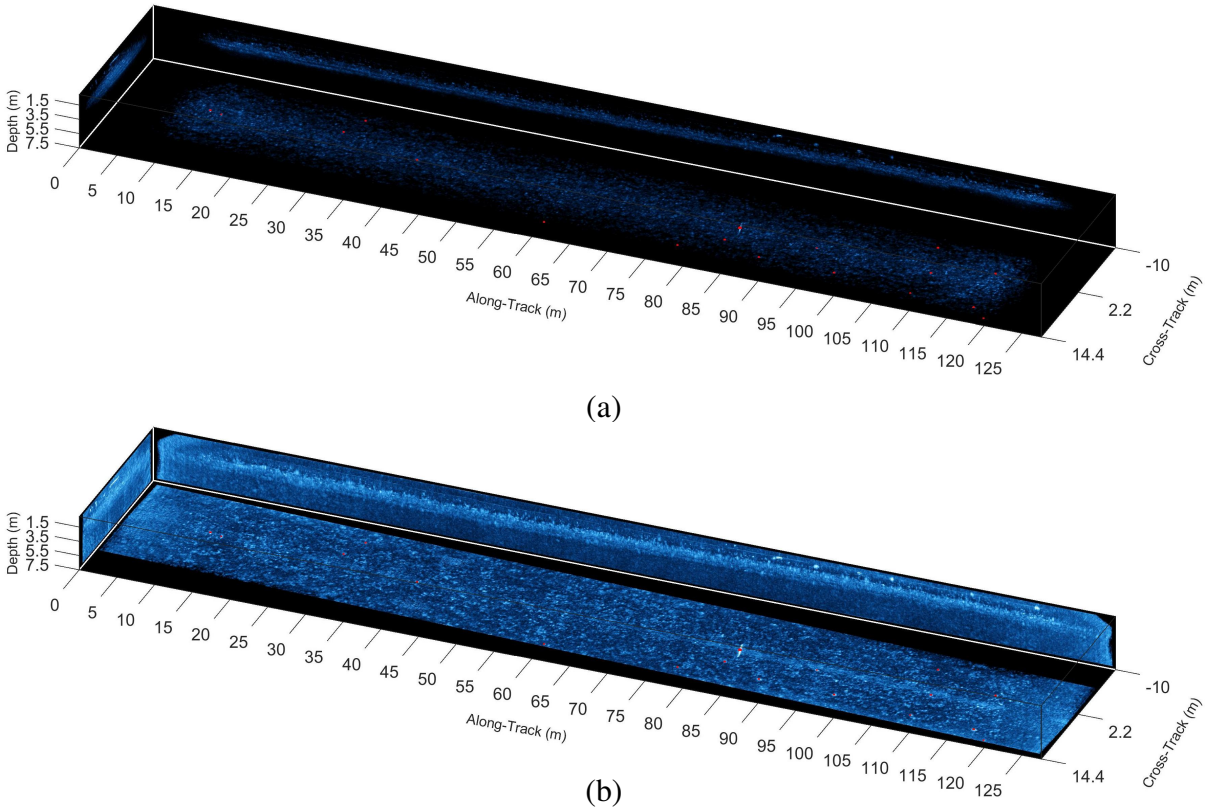


Fig. 4.3. An example MuST volumetric scene image, from Sequim Bay, displayed as a trio of MIPs, when (a) the data is raw or (b) normalized. Algorithm detections are marked on the depth MIPs with red dots.

4.2 Algorithms

4.2.1 Data Normalization

Acoustic waves scatter upon encountering a boundary between two media characterized by a mismatch in acoustic impedance and the roughness of the boundary [27]. The strength of the reflection is determined by the material properties – primarily density and sound speed – of the two media. As a result, the water-sediment interface (*i.e.,* the seafloor) will *not* necessarily produce the strongest backscattered return. For example, in the case of a mud seafloor above a sand substrate, the scattering from the water-mud interface is likely to be lower than the scattering from the deeper mud-sand transition within the sediment itself. But inevitably, there will exist some depth layer that dominates all others, and this necessitates a data normalization pre-processing step.

Suppose there is a raw 3-d data cube of beamformed sonar returns, $\mathbf{A}$, in a coordinate system (x, y, z) defined by the cross-track direction x, along-track direction y, and depth from the sonar platform, z. The normalization procedure proposed for volumetric sonar data is as follows:

1. Determine the plane of the interface in the image volume for which returns dominate. Define z' to be the normal to this plane, with $\hat{\theta}$ the angle between z and z' in the y-z plane.

2. Determine and remove the sub-volume, $\mathbf{V}_M$, of the data cube contaminated by multipath interference associated with the dominant interface.

3. For a given cross-track position x, consider the y-z plane, $\mathbf{A}_x(y, z)$. Compute the median value, $\mu_x(d)$, of pixels located a given distance d from the dominant interface, and divide those pixel values by $\mu_x(d)$. Repeat for each d and x.

4. For a given along-track position y, consider the x-z plane, $\mathbf{A}_y(x, z)$. Compute the median value, $\mu_y(d)$, of pixels located a given distance d from the dominant interface, and divide those pixel values by $\mu_y(d)$. Repeat for each d and y.

5. Convert the normalized data cube to a logarithmic (decibel) scale and truncate the voxel values to a dynamic range of $[0, 40]$.

In the interest of brevity, we refrain from detailing the procedure more formally, but we shall provide insight into the algorithmic choices made. An implicit assumption in the normalization procedure is that any given sediment layer in the imaged volume is approximately uniform in depth, with this justified by the nature of underwater sediment transport processes and diagenesis [28]. We also assume that any bathymetric variation of the dominant interface plane occurs only in the (much longer) along-track direction. That is, the depth of the dominant interface at a given along-track position is taken to be constant. This assumption is reasonable given the relatively narrow sonar swath in the cross-track direction, coupled with the fact that relief changes typically manifest only over longer spatial scales. This simplifies step 1 to determining a single angle $\hat{\theta}$ in the y-z plane characterizing the slope of the dominant interface. In practice this is achieved by applying a Radon transform on the 2-d image slice at the ith cross-track position to determine θ_i, and then taking $\hat{\theta}$ to be the mode of the set comprising all θ_i.

In very shallow water, the presence of multipath scattering resulting from multiple reflections – *i.e.,* from the sonar transmitter down to the dominant sediment interface, up to the

water-air interface, back to the dominant sediment interface, and back to the sonar receiver – can contaminate the data cube [29]. With $m_z = \tan\hat{\theta}$ the slope of the dominant sediment interface, simple geometry dictates that a multipath replica of that interface will manifest with a slope of $2m_z$ [30]. With knowledge of the absolute depth of the water column (*i.e.,* from the air-water interface), one can then determine the region of the data cube where multipath interference will occur via the corresponding equations of the lines (or planes) associated with $2m_z$. The data in the identified multipath-contaminated region is discarded because any authentic signal in the data cube's multipath region will be overwhelmed by the multipath return of the dominant interface.

In steps 3 and 4, cross-track positions are considered before along-track positions in order to first remove the known scattering dependence on grazing angle [31] (and hence cross-track position). The end result is a data cube in which the background is approximately uniform regardless of depth.

4.2.2 Detection

The object detection task we concern ourselves with is a classic remote-sensing problem of locating a target signal amid a noisy background [5]. Common to this class of problems is the need to set two parameters that are tied to the target of interest: a window related to the size of the target, and a threshold related to the relative strength of the target.

In the 1-d case, a simple solution involves estimating the background level and assuming that signals above some threshold indicate a potential target signal. Here we deal with 3-d imagery, but the underlying principles are unchanged. Our proposed algorithm computes a local estimate of the background intensity around each voxel, which are the potential target signals. If the target-to-background ratio exceeds the threshold, the voxel is flagged. And if a connected volume (*i.e.,* "blob") of such flagged voxels exceeds the minimum size of objects of interest, a discrete alarm is generated.

The approach taken in this work is expressly for the UXO application in which the size and shape of objects of interest can vary considerably, from individual bullets or fragments to larger intact shells and munitions. As such, we desire a general purpose detector that is not tied to one specific object, but rather can detect a wide class of objects. Our approach largely extends the Mondrian detector [32] to 3-d imagery. But given the 3-d geometry of the problem, there will be no acoustic shadow clues (as are present in 2-d side-looking sonar images [33, 34]) to exploit. Instead the only insight to leverage is that the man-made objects of interest are assumed to be acoustically harder, producing stronger returns, than the surrounding sediment.

The normalization process applied to the data permits the use of a single threshold for the entire 3-d data cube. In addition, the necessary signal and background estimates can be performed extremely efficiently using integral images [6].

Let $\mathbf{A}$ be a data cube in which targets are to be detected. The integral image $\mathbf{W}$ of $\mathbf{A}$ represents the summed volume of intensities defined by

$$W(x, y, z) = \Sigma_{x' \leq x} \Sigma_{y' \leq y} \Sigma_{z' \leq z} A(x', y', z'). \tag{4.1}$$

The summed intensity over any rectangular volume with corners specified by α^j for $j \in \{0, 1\}^3$ can then be readily computed as

$$U = \Sigma_{j \in \{0,1\}^3} (-1)^{3-||j||_1} W(\alpha^j), \tag{4.2}$$

where $W(\alpha)$ is the integral image at voxel α. Thus, given specified rectangular volumes to be used for the target ("signal") and background ("noise") estimation, the mean intensity of each quantity centered at each voxel in the data cube can be computed quickly.

To ensure that potential target voxels are excluded from the background volume calculations, a third rectangular volume is employed to act as a "guard" window (à la a split window). The target ($\mathbf{T}$), guard ($\mathbf{G}$), and background ($\mathbf{B}$) volumes are concentric, increasing in size. In this work, the sizes (in units of m) of the rectangular volumes are set to

$$\begin{aligned} \boldsymbol{\gamma}(x,y,z) &= [0.12, 0.12, 0.06] && (4.3)\\ \mathbf{T} &= \boldsymbol{\gamma} && (4.4)\\ \mathbf{G} &= 4\boldsymbol{\gamma} && (4.5)\\ \mathbf{B} &= 6\boldsymbol{\gamma}, && (4.6) \end{aligned}$$

with the parameters reflecting the sizes of objects of interest.

Let $T(x,y,z)$, $G(x,y,z)$, and $B(x,y,z)$ represent the summed intensity of the target, guard, and background volumes centered at voxel (x,y,z), respectively. A binary data cube map is then constructed based on the result of the test

$$\left(\frac{n_B - n_G}{n_T}\right)\left(\frac{T(x,y,z)}{B(x,y,z) - G(x,y,z)}\right) \geq \tau_s, \tag{4.7}$$

where τ_s is related to the minimum (relative) intensity of objects of interest, and $n_\cdot$ indicates the number of voxels used in a given sum (to effect mean values). Each connected volume in this map that exceeds τ_v, the minimum volume of objects of interest, is then converted to a discrete alarm.

When using integral images, it should be noted that the voxel values must be non-negative. Additionally, it is imperative that sufficient numerical precision is maintained [35], an issue that is likely to be especially germane when dealing with large data cubes. A simple way to verify that no overflow errors have manifested is to exploit the inversion formula

$$\tilde{A}(x,y,z) = \Sigma_{j\in\{0,1\}^3}(-1)^{2-||j||_1}W(\alpha^j), \tag{4.8}$$

and confirm that the result, $\tilde{\mathbf{A}}$, matches the original data cube $\mathbf{A}$ exactly. In this work, 64-bit precision was required to avoid undesired arithmetic overflow.

4.2.3 Feature Extraction

Volumetric data cubes from sonar surveys can be enormous products, but the vast majority of the content is likely to be benign and irrelevant for the UXO remediation problem. At the same time, human visualization of an entire 3-d data cube is not trivial, and techniques that rely on basic data slices or projections invariably discard considerable information. Therefore, a major role of the detector is to rapidly decrease the amount of data that must be examined further. Indeed, the contacts flagged in the detection stage are expected to be passed on to a more sophisticated classification stage.

But additional processing or data-collection surveys may also be undertaken based on the results of the detection stage. For example, fully-complex 3-d beamforming, a very computationally demanding procedure, can be performed over limited volumes centered on contact locations, with this enabling the generation of acoustic color plots [36] that can reveal informative aspect and frequency-dependent responses associated with the object. Alternatively, additional physical surveys at sea can be undertaken to collect more comprehensive data on detected contacts, such as at new aspects or with different sonar settings (*e.g.,* grazing angle, waveform, depth). For these reasons, it can be valuable to provide an *ordered* list of contacts to prioritize subsequent operations.

To that end, we extract a pair of primitive features that can be combined into an overall detection "score." Specifically, a size feature, f_1, that is the volume of a contact's voxels that exceeded the threshold τ_s is computed. A strength feature, f_2, that is the mean of the contact's $n = 64$ largest voxel values is also computed. The final detection score by which the contacts can be ordered is taken to be the geometric mean of these two features,

$$f = \sqrt{f_1 f_2}, \tag{4.9}$$

as this effectively provides an approximation to the contact's overall intensity (and for the UXO problem, explosive potential or threat).

Other basic shape and orientation features can also be extracted, but we abstain from doing so given that the follow-on classification stage is designed to employ CNNs, for which explicit features are unnecessary.

4.2.4 Classification

Our classification approach is based on 3-d CNNs for volumetric SAS imagery. A CNN-based classification approach is particularly apropos for this data modality because hand-crafting features is challenging, and CNNs effectively obviate this process. Despite the enormous size of the 3-d data involved, it is shown how careful architecture design can make this proposed approach both feasible and successful.

CNN Design

A CNN [7] is a sophisticated classification algorithm that customarily ingests an image as input and produces scalar outputs corresponding to the probabilities of belonging to each class under consideration (*e.g.,* target and clutter). Within these bookends, the architecture of a CNN consists of a sequence of *layers*, each performing a specific mathematical operation, arranged such that the output of one layer is the input to the subsequent layer. This nested functional structure – in conjunction with *nonlinear* functions – enables highly complex decision surfaces. In turn, this is the source of a CNN's rich representational capacity.

In this work, we develop *3-d* CNNs in which the input data is a 3-d SAS data cube, rather than a 2-d image. The general architecture design largely follows our previous work [8], but extends it to three dimensions. More specifically, we develop eight CNNs that share a common architecture of alternating convolutional *blocks* (comprising one or more convolutional layers) and pooling layers. The input to a CNN is assumed to be a 1.22 m$\times 1.22$ m$\times 2.02$ m volumetric SAS image, which for the SVSS data corresponds to a $61 \times 61 \times 101$ voxel data cube, since each SVSS voxel spans 2.0 cm in each dimension. Thus, a single SVSS data cube input to any given network contains approximately 3.75×10^5 voxels. The outputs of the CNN's final layer are the (softmax) probabilities of a 3-d data cube belonging to each class (target or clutter).

Specific details about the designed CNNs are provided in Table 4.1; the information provided is sufficient for recreating the networks exactly. Here, brackets are used to convey the concept of convolutional *blocks*, in which there are multiple convolutional layers in between pooling layers. (The ith set of brackets contains the information about the convolutional layers in the ith block. For example, the first convolutional block of CNN F comprises a convolutional layer with $7 \times 7 \times 12$ filters, followed by a second convolutional layer with $5 \times 5 \times 7$ filters.) The convolutional block construct allows deeper networks, and thus greater complexity, without a proportional increase in the number of parameters to learn. Despite the enormous size of the input data cubes, the use of extremely small networks allows the CNNs to be trained with modest computational resources, and vitally, avoids computer-memory constraint issues.

Table 4.1. Architectures of 3-d CNNs trained

CNN Label	CNN Depth	Conv. Blocks	Conv. Layers Per Conv. Block	Filters Per Conv. Layer	Filter Sizes (Voxels) $[x \times y \times z]$	Pooling Factors $[x \times y \times z]$	Number of Parameters
A	2	2	1	4	$[6 \times 6 \times 6]\ [4 \times 4 \times 5]$	$[8 \times 8 \times 12]\ [4 \times 4 \times 4]$	2157
B	2	2	1	4	$[8 \times 8 \times 12]\ [6 \times 6 \times 7]$	$[6 \times 6 \times 9]\ [4 \times 4 \times 4]$	7117
C	3	3	1	4	$[6 \times 6 \times 6]\ [5 \times 5 \times 5]\ [4 \times 4 \times 5]$	$[4 \times 4 \times 6]\ [2 \times 2 \times 2]\ [2 \times 2 \times 2]$	4161
D	3	3	1	4	$[11 \times 11 \times 12]\ [8 \times 8 \times 7]\ [4 \times 4 \times 5]$	$[3 \times 3 \times 5]\ [2 \times 2 \times 2]\ [2 \times 2 \times 2]$	14273
E	6	3	2	4	$\begin{bmatrix} 5 \times 5 \times 6 \\ 4 \times 4 \times 5 \end{bmatrix} \begin{bmatrix} 5 \times 5 \times 6 \\ 4 \times 4 \times 5 \end{bmatrix} \begin{bmatrix} 5 \times 5 \times 3 \\ 4 \times 4 \times 3 \end{bmatrix}$	$[2 \times 2 \times 4]\ [2 \times 2 \times 2]\ [3 \times 3 \times 3]$	7557
F	6	3	2	4	$\begin{bmatrix} 7 \times 7 \times 12 \\ 5 \times 5 \times 7 \end{bmatrix} \begin{bmatrix} 5 \times 5 \times 8 \\ 4 \times 4 \times 5 \end{bmatrix} \begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix}$	$[3 \times 3 \times 4]\ [2 \times 2 \times 2]\ [1 \times 1 \times 1]$	10525
G	9	3	3	4	$\begin{bmatrix} 4 \times 4 \times 5 \\ 3 \times 3 \times 5 \\ 3 \times 3 \times 4 \end{bmatrix} \begin{bmatrix} 4 \times 4 \times 4 \\ 3 \times 3 \times 4 \\ 3 \times 3 \times 3 \end{bmatrix} \begin{bmatrix} 4 \times 4 \times 4 \\ 4 \times 4 \times 4 \\ 3 \times 3 \times 4 \end{bmatrix}$	$[2 \times 2 \times 3]\ [2 \times 2 \times 2]\ [2 \times 2 \times 2]$	6313
H	12	3	4	4	$\begin{bmatrix} 4 \times 4 \times 3 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \begin{bmatrix} 3 \times 3 \times 4 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix}$	$[2 \times 2 \times 3]\ [2 \times 2 \times 2]\ [1 \times 1 \times 3]$	5141

Each CNN contains either 2 or 3 convolutional blocks; each block contains a specific number of convolutional layers (equal to the number of rows in Table 4.1's filter-sizes column). A given filter always has the same number of voxels in the x and y dimensions (*i.e.,* it is square), and only 4 filters are used in each convolutional layer. All convolutions use a stride of 1, and padding is not used (*i.e.,* in TensorFlow-speak, the padding option is set to '`valid`'). Rectified linear unit (ReLU) activations are used after each convolutional layer, while a softmax activation is used at the output. All pooling layers use average pooling, rather than max pooling. The design of the architecture (and specifically the final pooling layer) ensures that the dense layer always contains 4 nodes. The capacities of the CNNs are intentionally kept so low because the amount of training data available is limited, but also because larger networks would be plagued by computer-memory problems. Collectively, the eight CNNs have only 57244 parameters to learn, which is still several orders of magnitude lower than traditional optical-image CNNs as well as the custom SAS-image CNNs that have been considered in the literature.

At a high-level, the general CNN architecture employed here is not so unusual. However, a few subtle, but key, design choices are a stark break from convention. The most striking deviation is the use of only 4 filters per convolutional layer; almost all CNNs in the literature use many dozens or hundreds or thousands of filters per convolutional layer. Importantly, this self-imposed constraint makes interpretability of the filters and intermediate representations feasible. Additionally, the use of only a *single* dense layer before the output layer is also somewhat unusual. The third uncommon choice is the decision to reduce (via pooling) the size of each feature map in the final convolutional layer to a $1 \times 1 \times 1$ output (*i.e.,* a scalar). Because of this, and the small number of filters, each CNN's dense layer has only 4 nodes. That is, the networks are effectively forced to make predictions from only 4 "features." Although we severely constrain the number of parameters in the models, importantly we do not restrict CNN *depth*. Finally, the retention of a considerable amount of data in the z dimension (*i.e.,* into the sediment) for each input cube enables the CNN to possibly exploit late-arriving elastic scattering phenomena. These clues would likely be inaccessible if 2-d projections (like MIPs) of the data were instead used to train a standard 2-d CNN.

CNN Training

The detection algorithm that operates on 3-d imagery was applied to all of the SVSS scene-level volumetric data collected at the two Sayers sites. The resulting data cubes of alarms from site B were used as training data for the CNNs; the data cubes of alarms from site A were treated as the test set. For the binary-classification experiments in this work, all man-made objects were considered targets, while all other alarms were labeled as belonging to the clutter class. (The small number of UXO alarms present precluded treating only UXO objects as the target class.) The details of the SVSS data sets after the detection stage are summarized in Table 4.2.

Table 4.2. Summary of SVSS sonar data sets after the detection stage

Location	Data Set Usage	Seafloor Area (m^2)	Sediment Volume (m^3)	Number of Clutter	Number of Targets
Site B	Training	15390	23392.8	11029	550
Site A	Test	15630	23757.6	6074	671

CNN training was performed in Python with the TensorFlow [37] software library. Training used an RMSprop optimizer with a learning rate of $\eta = 0.001$, in conjunction with a binary-

cross-entropy loss function. A batch size of $B = 64$ was used, with equal numbers selected from each class to combat the severe class-imbalance of the training data. Data augmentation that respected the geometry of the volumetric sonar data-collection procedure was employed during training; this meant a random cross-track translation $i_{tx} \in [-0.1 \text{ m}, 0.1 \text{ m}]$, along-track translation $i_{ty} \in [-0.1 \text{ m}, 0.1 \text{ m}]$, cross-track reflection $i_{rx} \in \{\pm 1\}$, and along-track reflection $i_{ry} \in \{\pm 1\}$ was applied, "on-the-fly," to each SAS image cube selected for the batch. (To permit these translation operations, the data cubes that comprise the training (and test) sets are slightly larger than the size of the data cubes required as input to the CNNs.) No translation or reflection was performed in depth (*i.e.,* the z dimension). It is worth reiterating that fully-populated 3-d volumetric data cubes – and not 2-d MIPs – are used as input to the CNNs.

Each CNN was allowed to train for 200 epochs, where one epoch was defined as a set of 1000 batches. Each batch was formed by randomly selecting image cubes from the full set of training data. (Thus, every 100 epochs during the CNN training phase, each of the 550 training-set target cubes is seen about 5818 times and each of the 11029 training-set clutter cubes is seen about 290 times, on average.) A validation set, common to all CNNs, was created by randomly selecting 50 targets and 450 clutter from the larger training set. For a given CNN, the epoch for which the model achieved the maximum AUC on the validation set was used to select the final model to evaluate the test sets. Based on this criterion, the number of training epochs actually used to train each CNN is shown in Table 4.3. For reference, training time for these CNNs is about 2 seconds per batch with a single GeForce GTX 1080 GPU. Thus, based on the actual number of training epochs needed, the training time for a single CNN is on the order of only 10 to 100 hours.

Table 4.3. Training epoch at which validation set AUC was maximized

CNN	Training Epoch
A	163
B	145
C	56
D	56
E	53
F	17
G	26
H	56

The convolutional filters learned by each CNN are shown later in Sec. 5.2.2. Sec. 5.2.3 shows the intermediate responses at each layer of each CNN for a single interesting data cube containing UXO.

4.2.5 Limited-Scope Classification

In the previous section, CNNs were used to discriminate between two general classes of objects, each of which exhibits considerable intra-class diversity in terms of object size, shape, composition, and burial state. As a result, the features that a trained CNN will rely on to make a prediction will be a complex combination that is not easily disentangled. (However, it should be noted that the use of very few filters in each convolutional layer does ameliorate this issue.) Instead, designing specially controlled experiments can provide a way to learn principled, explainable features that can be tied directly to the wave phenomena of the physics involved.

The idea is to develop a CNN classifier in which the two classes are *not* UXO and non-UXO, but rather whether or not a specific object has a certain attribute. For example, one experiment could attempt to train a CNN to determine whether, say, an object is filled with air or water. In this case, training data for one class would come from air-filled munitions, while the training data for the other class would come from water-filled munitions. Importantly, all other variables, such as range and environment, would be kept identical. As a result, any clues that the CNN uncovers should, in theory, be due to the single variable that differs (in this example, the interior). If the CNN is able to classify the test data correctly, the features from the dense layer of the CNN would necessarily be linked to the attribute under investigation. In the event that the physics is not fully explainable *a priori*, the intermediate representations uncovered by the CNN can lend insight and improve understanding to the wave phenomena involved. This entire process can also be repeated using frequency sub-bands and aspect sub-apertures, so that there is potentially an entire *set* of features – each with even greater specificity – linked to the attribute.

Provided enough data are available, these controlled comparisons can be conducted for a series of attributes, such as object interior fill, material composition, size (*e.g.,* munition diameter), and burial state. Unique CNNs would be developed for each attribute case. The features derived from the classifiers that can successfully distinguish the two classes under consideration could be retained as explainable features linked to the relevant attribute. When a set of multiple such features was obtained, a new larger CNN that concatenates these controlled-comparison CNNs at the penultimate dense layer could be designed. The end result would be a CNN classifier in which the relative importance, or weight, of each "explainable" feature could be examined for each prediction.

All of the above can also be explored for various CNN architectures, as well as different or multiple data representations, to enable the use of ensembles that were shown to be so powerful in the previous SEED project. But the end result would be that every CNN classifier prediction that is made could be explained by a combination of well-understood features tied to specific attributes, frequency bands, aspects, *etc.*

Our exploration of limited-scope CNN experiments was initiated with simulated sonar data because it provides substantial amounts of training data under the appropriate conditions. (Measured data collected at sea on target fields is typically available only at a few discrete ranges, so fully testing the generalization ability of the trained CNNs is challenging.) More specifically, we attempted to train CNNs to learn to discriminate air-filled objects from water-filled objects.

Data Preparation

This study exploits the TIERSWAT data set, which is simulated uncalibrated sonar data, developed at APL-UW. The data is the simulation of a linear SAS trajectory of a vehicle 5 m above the seafloor. The target scattering signals were generated using APL-UW's target-in-the-environment-response (TIER) model [38], while the scattering from the environment was generated using the Personal Computer Shallow Water Acoustics Tool-set (PC SWAT) simulator [39] of the Naval Surface Warfare Center - Panama City Division (NSWC-PCD).

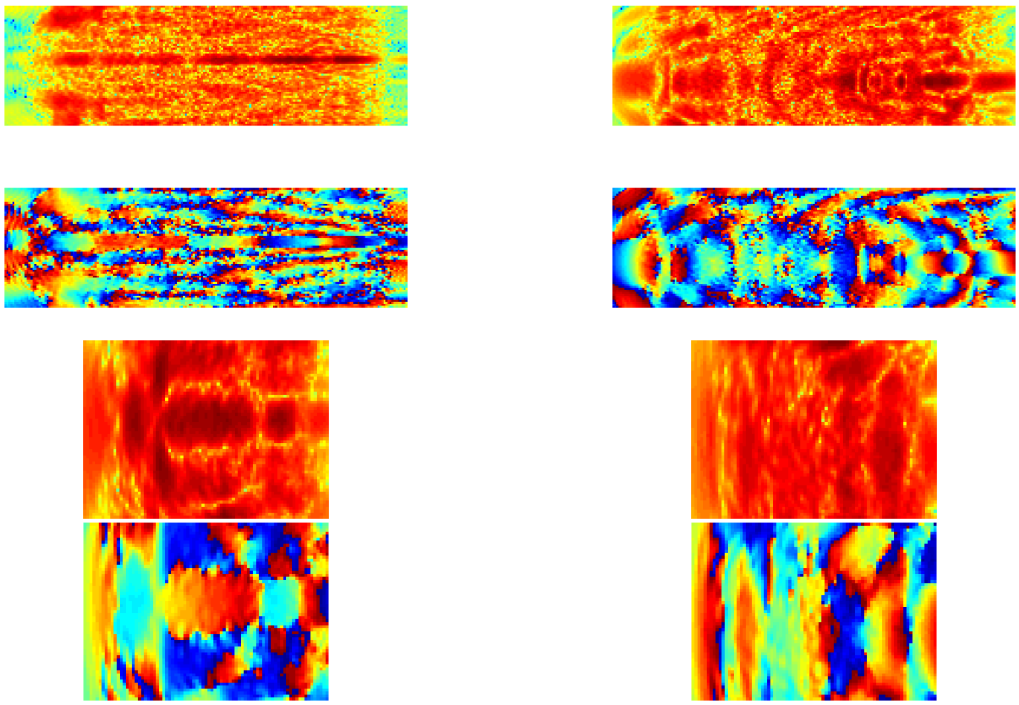
The TIERSWAT data set is dual-band, complex-valued acoustic-color "imagery" that expresses a target's response as a function of frequency and aspect. The high-frequency (HF) band data spans 60° in aspect, over the frequency band 10-30 kHz. The low-frequency, or broadband (BB), data spans 59° in aspect, over the frequency band 2-10 kHz. In either band, one pixel corresponds to 1° in aspect and 100 Hz in frequency. From this data, four acoustic-color data representations are readily available, namely the magnitude and phase of each frequency band's

acoustic-color data. (An object aspect of -90° corresponds to nose-endfire.)

Data sets 1-63 were reserved for training, while data sets 64-126 were used as test data. The difference between the training and test data sets is that jitter was introduced into the object range and aspect (with respect to the sensor) in the latter.

Let $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$ be a raw complex acoustic-color image of an object. The magnitude representation is given by $\mathbf{A}_M = |\boldsymbol{z}|$, and the phase representation is the result of the two-argument arc-tangent function, $\mathbf{A}_P = \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x})$. Each acoustic-color magnitude representation $\mathbf{A}_M$ was normalized as $\mathbf{A}'_M = (2\log_{10}\mathbf{A}_M - 9)/3$ to scale the pixel values approximately into the range $[-1, 1]$. Each acoustic-color phase representation $\mathbf{A}_P$ was normalized as $\mathbf{A}'_P = (\mathbf{A}_P - \pi)/\pi$ to scale the pixel values into the range $[-1, 1]$.

Two examples of the four acoustic-color representations of this data are shown in Fig. 4.4.



(a) (b)

Fig. 4.4. Example TIERSWAT data of a 2:1 cylinder at a range of 28 m and at an orientation of (a) broadside and (b) almost nose-endfire. From top to bottom, the representations are the acoustic color HF magnitude, HF phase, BB magnitude, and BB phase. The horizontal axes correspond to frequency; the vertical axes indicate aspect.

CNN Design and Training

We design a single common CNN architecture for the two different frequency bands of data. The only difference is the size of the convolutional filters, owing to the different input data sizes. The input data for HF representations is 60 pixels $\times$ 201 pixels, while the input data for BB representations is 59 pixels $\times$ 81 pixels. The CNN architecture is a sequence of four sets of alternating convolutional and pooling layers, followed by a fully-connected (dense) layer, and then the output (prediction). Only 4 filters are used in each convolutional layer. ReLU activation functions are used after each convolutional layer, while a softmax activation is used at the output. All pooling layers use average pooling. The number of nodes contained in the dense layer is 4 per input data representation. When multiple input data representations are

used, the features are concatenated at the dense fully-connected layer. Details of the filter sizes and pooling factors are shown in Table 4.4.

Table 4.4. CNN architectures for TIERSWAT input-data (1 pixel represents $1^\circ \times 100$ Hz of data), where aspect is row-wise and frequency is column-wise

Frequency Band	Convolutional Blocks	Convolutional Layers Per Block	Filter Sizes in Pixels (Row × Column)	Pooling Factors (Row × Column)
HF	4	1	5×20 5×16 5×13 3×12	$[2 \times 2]$ $[2 \times 2]$ $[2 \times 2]$ $[2 \times 2]$
BB	4	1	4×8 5×6 5×5 3×5	$[2 \times 2]$ $[2 \times 2]$ $[2 \times 2]$ $[2 \times 2]$

CNN training was performed using the RMSprop optimizer with a learning rate of $\eta = 0.001$, in conjunction with a binary-cross-entropy loss function, for 100 epochs. Each epoch was one full pass through the training data set. A batch size of $B = 128$ was used, with equal numbers selected from each class. Data augmentation was effected by randomly choosing to reflect in the aspect dimension each acoustic-color plot selected for the batch. The final prediction for a test image was taken as the ensemble (mean) of the predictions on the standard and aspect-reversed input data.

The CNNs were trained using data from air-filled or water-filled 155 mm munitions (from data sets 1-63). The trained CNNs were then used to make predictions about the interior fill of test data. The first experiment used data from air-filled or water-filled 155 mm munitions (from data sets 64-126) as the test data. A second experiment used 105 mm munitions (from data sets 64-126) as the test data. The latter experiment represents an opportunity to assess the generalization ability of the trained networks to new object types not seen during the training process.

5 Results and Discussion

5.1 Detection

First, for the SVSS data, we show the performance of the proposed 3-d energy detection algorithm described in Sec. 4.2.2, as well as that of two alternative methods that rely on 2-d image representations of the data. The first of these alternatives, denoted "2-d Elastic," forms a 2-d image by summing the voxel values in the depth dimension at each along-track/cross-track position. The second of these alternatives, denoted "2-d MIP," forms a 2-d image by retaining the maximum voxel value in the depth dimension at each along-track/cross-track position. An equivalent 2-d version of the full 3-d detection algorithm – in which the z dimension of (4.4)-(4.6) is a single voxel – is used to generate alarms for these methods. It is important to note, however, that the 2-d variants lose potentially valuable depth information about the alarms.

In Figs. 5.1-5.3, the detection performance for each of these three approaches is shown for each ground-truth item at three different sites. The nomenclature of the ground-truth items corresponds to that used in SERDP project MR-2545 reporting [2, 25]; detailed information about each object and burial status can be found in Appendix A of [2]. Because the proposed 3-d detector achieved higher detection rates than the 2-d alternatives, all subsequent work used the 3-d approach exclusively.

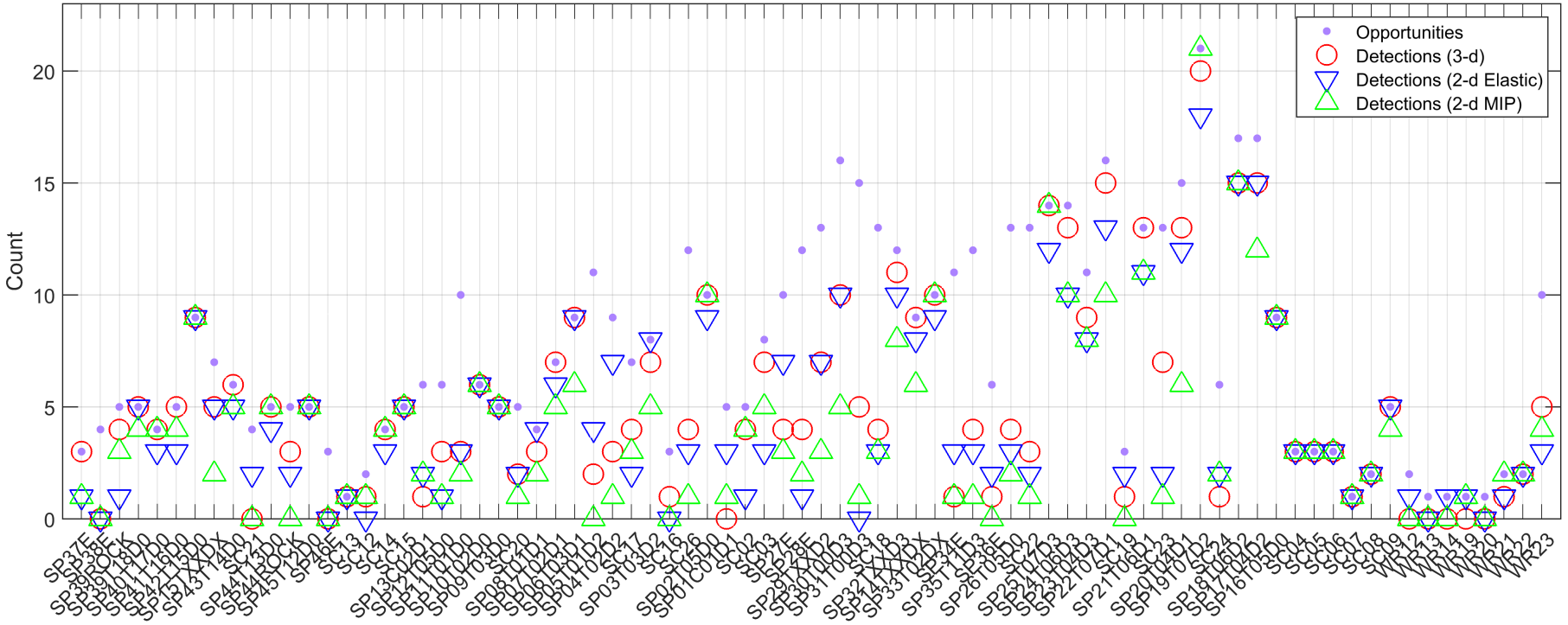


Fig. 5.1. Detection performance as a function of each ground-truth item at site A.

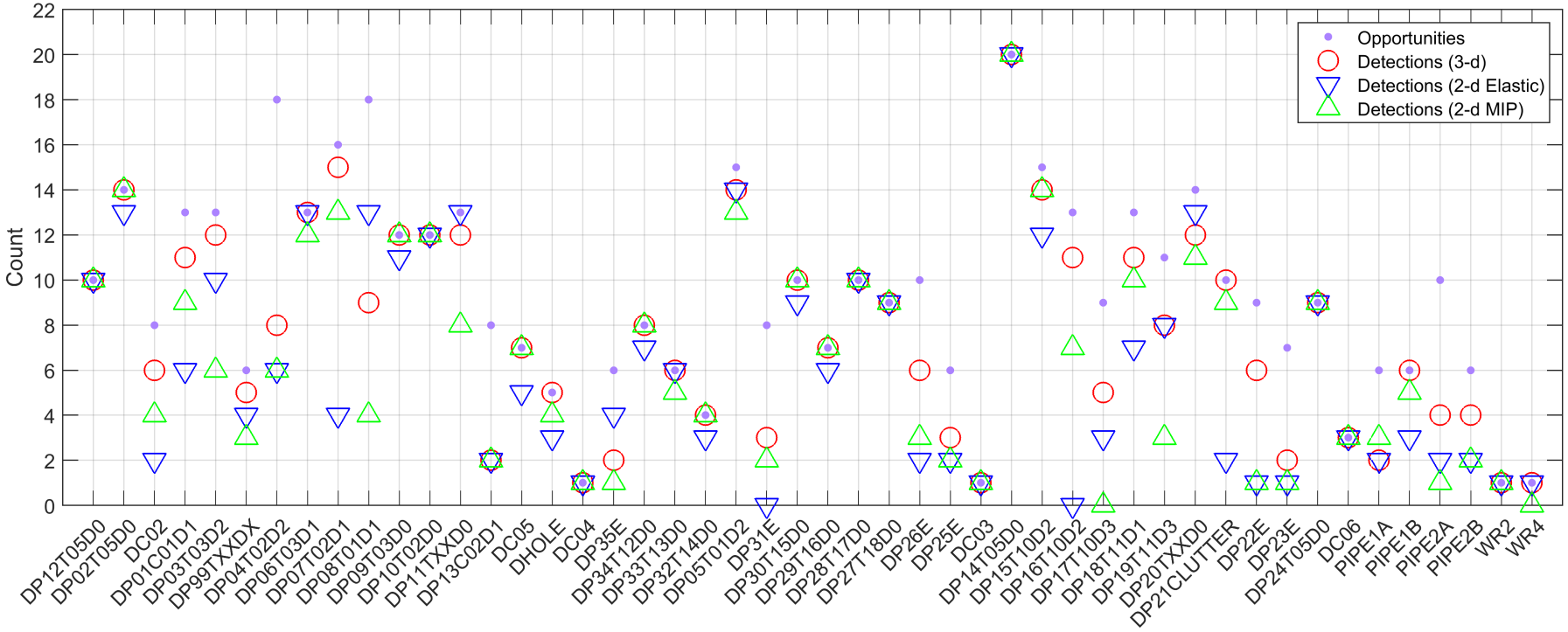


Fig. 5.2. Detection performance as a function of each ground-truth item at site B.

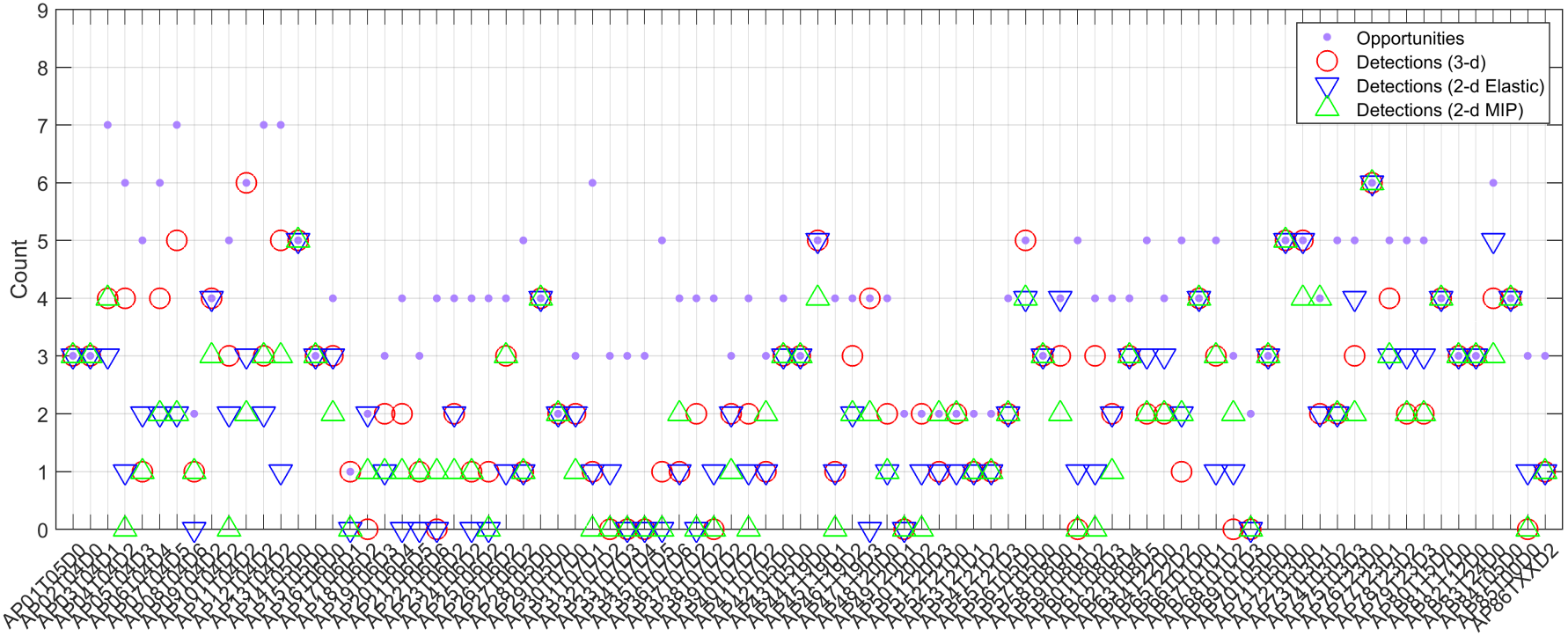
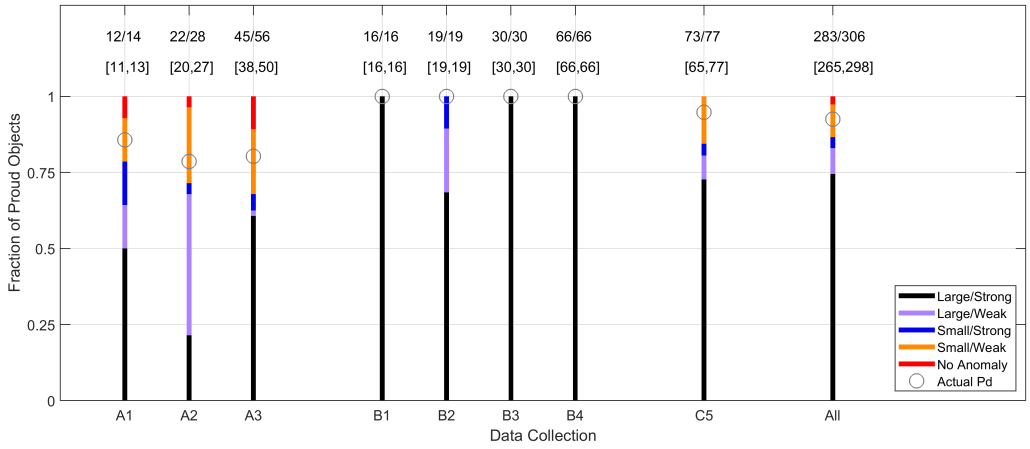


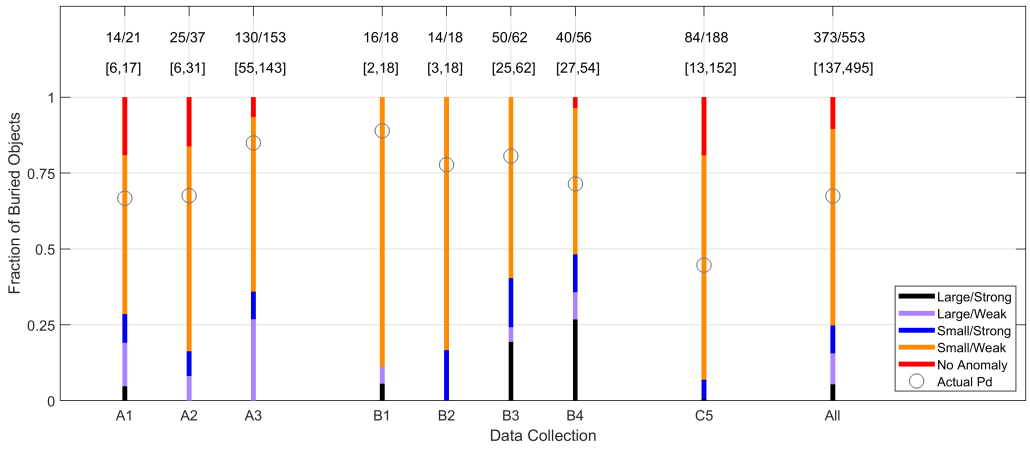
Fig. 5.3. Detection performance as a function of each ground-truth item at site C.

The complex physics at play underwater and in the subsediment volume suggest that the collected 3-d data might not always support target detection, regardless of the algorithm employed. Factors such as sediment attenuation, interface scattering levels, the presence of gas bubbles in the sediment, and the relationship between sensor resolution and target dimensions mean that the upper limit on detection capability will likely be less than unity. To assess this possibility, the data from each target opportunity was first visually examined and rated in terms of anomaly size (large or small) and strength (strong or weak) in the imagery. Anomalies that were deemed both small and weak represent a "gray zone" in which detection may or may not actually be feasible.

With these human assessments as a backdrop, the performance of the proposed target detection algorithm at eight distinct data collections, delineated by location and time, are shown in Fig. 5.4 for proud and buried targets. As can be seen from the figure, performance varies considerably across location (*cf.* collection letters), but also across time (*cf.* collection numbers), the latter variation suggesting strong environmental dependence (*e.g.,* water temperature, microbial activity). However, in all cases, the automated detection performance comported with the expected range based on visual inspection of the imagery.



(a)



(b)

Fig. 5.4. Performance of the detection algorithm for each SVSS data collection for (a) proud man-made targets and (b) buried man-made targets, along with the distribution of visual human assessment ratings. Above each bar are the numbers of targets detected vice opportunities, and in brackets the range of targets deemed detectable based on visual human assessment.

The combined detection performance from pooling the proud and buried targets, and using the score from (4.9), of the various data collections is shown in Fig. 5.5. Performance is displayed in terms of receiver operating characteristic-like (ROC-like) curves, where the probability of false alarm is replaced by false alarm rate (per unit volume). It should be noted that these results are for a considerable span of object sizes, some of which are even less than the sensor resolution.

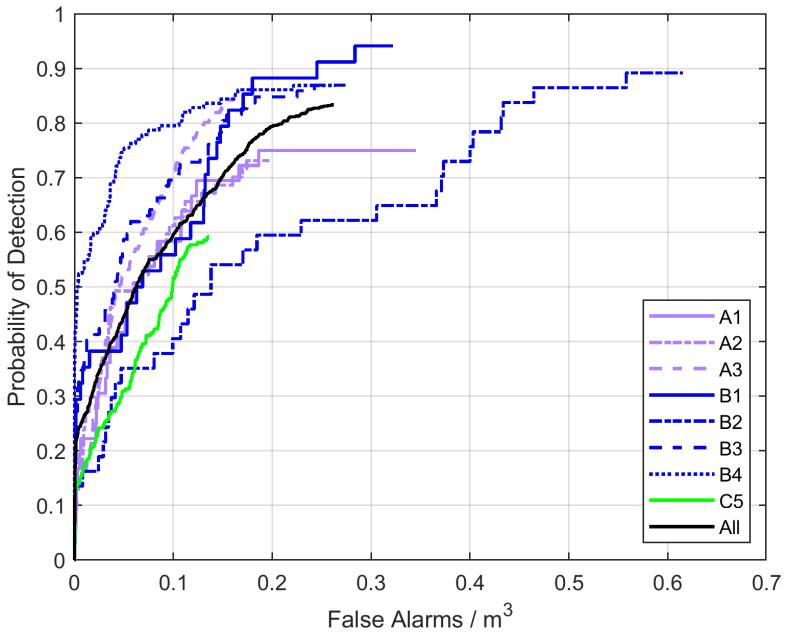


Fig. 5.5. Overall performance of the detection algorithm for each SVSS data collection.

For the MuST system, we possessed neither ground-truth information nor sufficient amounts of data to make statistically significant detection-performance assessments. But applying the detection algorithm to the modest amounts of data we did have resulted in seemingly reasonable alarms. An example set of such alarms is shown in Fig. 5.6.

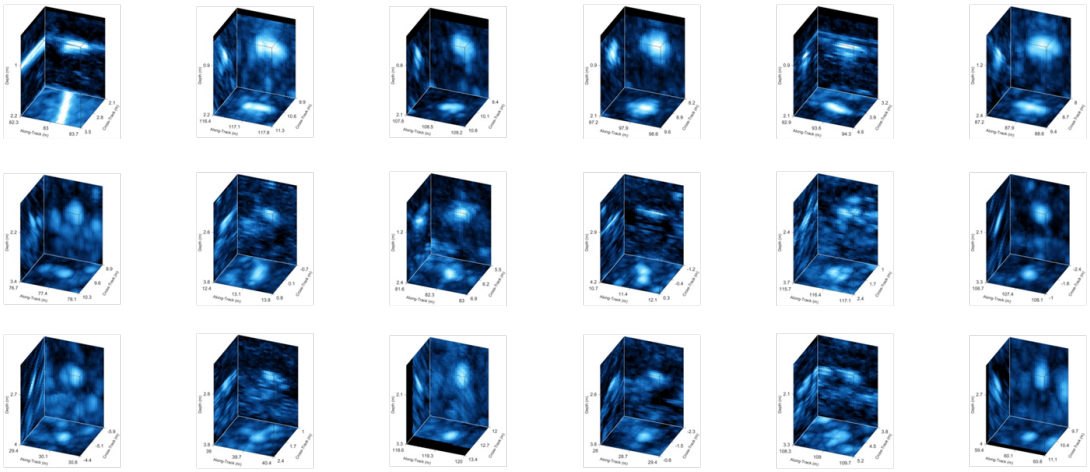


Fig. 5.6. Example alarms generated by the detection algorithm when applied to a MuST scene.

5.2 Classification

Our main interest in this section is to examine the ability of the CNN-based approach to successfully perform *classification*. Therefore, when presenting ROC curves, the experiments assume that all targets were successfully flagged in the detection stage. Thus, the maximum possible area under the ROC curve (AUC) [10], a scalar summary measure of performance, is always unity. (A perfect classifier would have an AUC of unity.) To obtain the overall probability of both successful detection and correct classification, the vertical axis of the ROC-like curves would need to be scaled by the inverse of the total number of targets present in the scene imagery.

5.2.1 CNN Performance

First, we examine the benefit of making CNN predictions based on a set of isometric input data cubes versus using only object-centered cubes. To assess the value of multiple representations of sonar imagery resulting from isometries – *i.e.,* distance-preserving transformations – of each original input data example, we consider a set of 36 affine transformations that do not violate the physics of the sonar-object geometry. This set is formed from the Cartesian product of cross-track translations $i_{tx} = \{-0.1 \text{ m}, 0 \text{ m}, 0.1 \text{ m}\}$, along-track translations $i_{ty} = \{-0.1 \text{ m}, 0 \text{ m}, 0.1 \text{ m}\}$, cross-track reflections $i_{rx} = \{\pm 1\}$, and along-track reflections $i_{ry} = \{\pm 1\}$. The "centered input" case, in which the detected object is well-centered in the data cube, corresponds to $[i_{tx}, i_{ty}, i_{rx}, i_{ry}] = [0 \text{ m}, 0 \text{ m}, 1, 1]$.

Classification performance in terms of AUC for the eight trained CNNs is shown for the test data set in Table 5.1. Specifically, the AUC with and without using isometric input cubes is shown when considering all objects, only proud objects, or only buried objects. Also shown in the table is the ensemble performance, denoted $\mathcal{E}$, which uses, for a given test cube, the mean prediction of the eight CNNs as the final prediction. From the table, it can be seen that the use of isometric inputs consistently improves performance for each *individual* CNN, though not necessarily for the *ensemble* of CNNs. This result is an indication that the diversity engendered by the unique CNN architectures exceeds that which is created by the isometric inputs. Thus, in time-critical applications, one can accelerate the inference phase by obtaining classifier predictions using only the object-centered cube, rather than the full set of isometries. Nevertheless, hereafter, all results correspond to the case using the set of 36 isometries.

Table 5.1. AUC on the test set depending on whether isometric input cubes are used

	Only Centered Input Cubes Used			Isometric Input Cubes Used		
CNN	All Objects	Proud Only	Buried Only	All Objects	Proud Only	Buried Only
A	0.877	0.940	0.820	0.881	0.941	0.827
B	0.865	0.949	0.786	0.874	0.953	0.798
C	0.883	0.943	0.837	0.882	0.943	0.834
D	0.858	0.914	0.821	0.868	0.921	0.832
E	0.883	0.905	0.861	0.895	0.920	0.869
F	0.848	0.873	0.834	0.872	0.913	0.853
G	0.878	0.932	0.838	0.891	0.948	0.847
H	0.855	0.897	0.827	0.873	0.934	0.837
$\mathcal{E}(A\text{-}H)$	0.912	0.965	0.868	0.912	0.965	0.868

Next, performance is presented in the form of full ROC-*like* curves, with the abscissa corresponding to the more informative false alarm *rate* instead of the *probability* of false alarm. The probability of false alarm is the probability of incorrectly classifying clutter as a target; the false alarm rate is the number of such incorrect classifications per image area or volume. When considering only proud objects, the false alarm rate is given per image (seafloor) area; when considering all objects or only buried objects, the false alarm rate is given per image volume.

Performance of the eight CNNs, as well as the ensemble, is shown in terms of ROC-like curves in Fig. 5.7. The AUC (for the corresponding ROC curve) is also provided in the legend. To provide a baseline measure of performance, the performance of the 3-d detection algorithm – which makes predictions based on the geometric mean of a size feature and an intensity feature – is also shown. While the full curves are informative, in practice, one must select a single operating point at which to make predictions. Therefore, on each CNN curve, the operating point corresponding to the natural decision threshold of $\tau = 0.5$ is also marked.

As can be observed from Fig. 5.7, the 3-d CNNs greatly outperform the simple baseline detector, as would be expected. But more interestingly, the use of the ensemble of networks proves beneficial and removes the necessity of selecting a single best CNN architecture to employ. The complementary nature of the CNNs, and the unique clues that each uncovers and exploits to make predictions, leads to reduced false alarm rates. As a result, the ensemble approach can directly translate into cost savings during UXO remediation efforts.

It is also worth noting that these classification results are based on using only a single data representation, namely the 3-d imagery, as input to the CNNs. A worthwhile avenue to explore is the use of alternative data representations (*e.g.,* acoustic color) in which complementary discriminative clues would be made more accessible to the CNN.

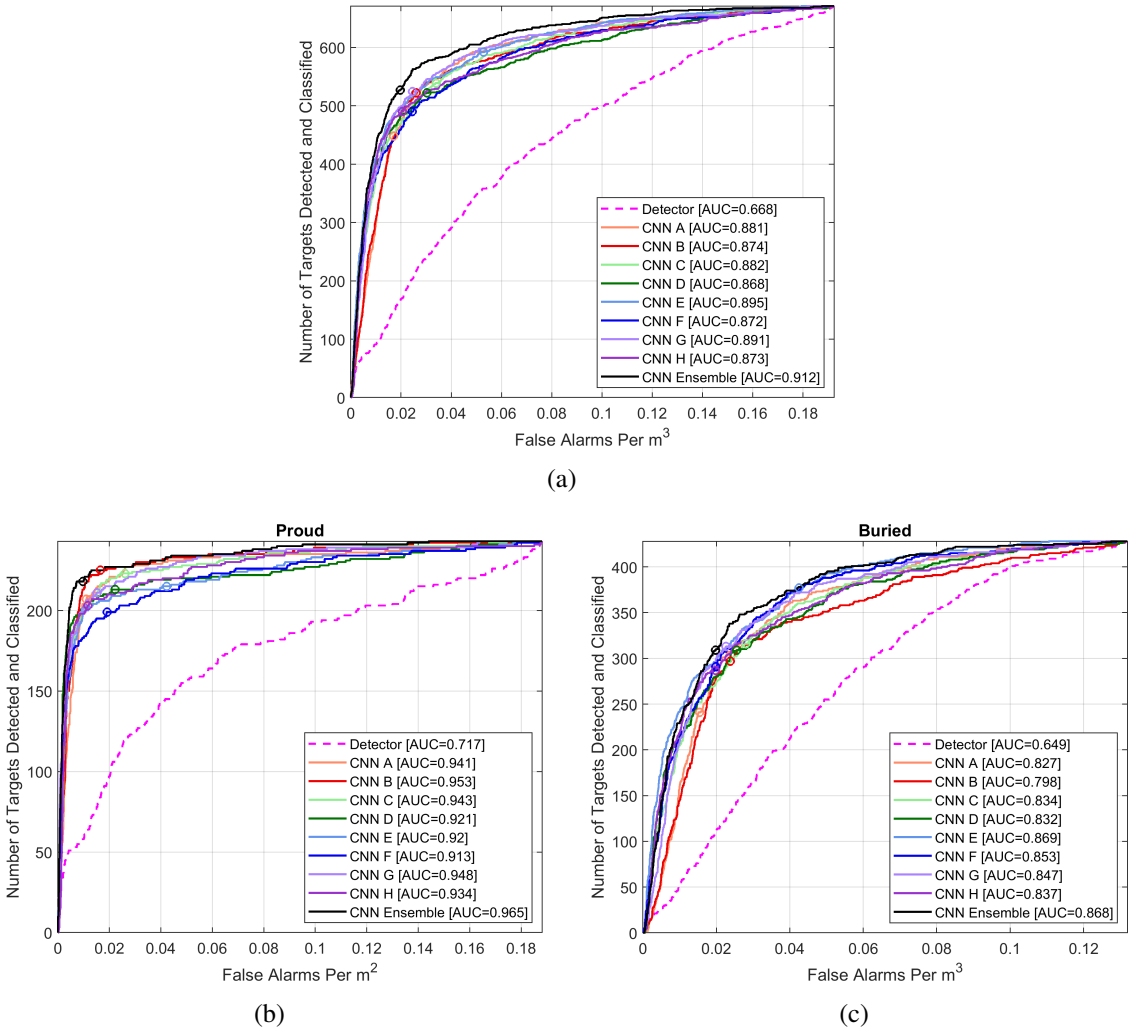
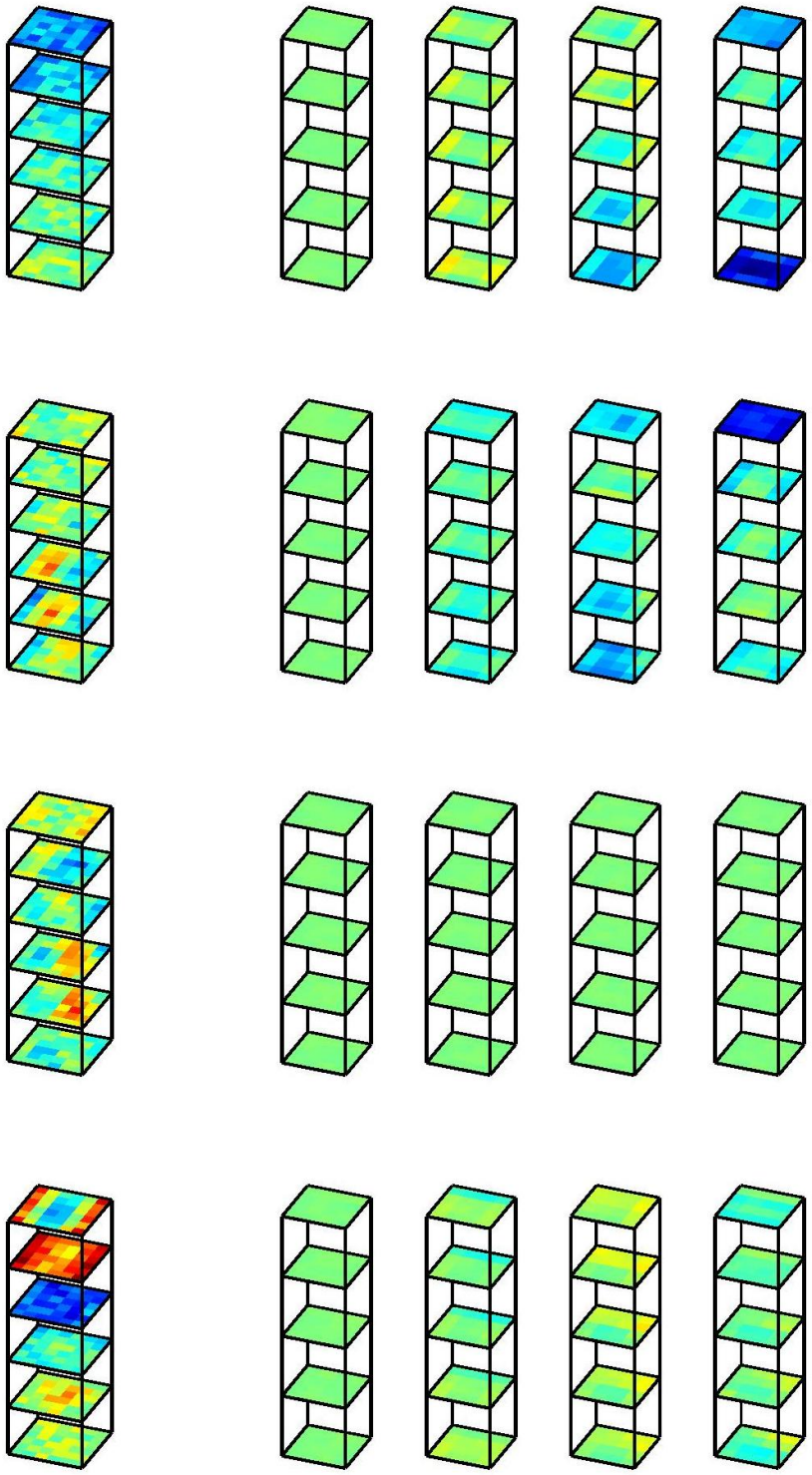


Fig. 5.7. Performance on the SVSS test data set in terms of ROC-like curves for (a) all objects, (b) only proud objects, and (c) only buried objects. The operating point for a $\tau = 0.5$ threshold is marked with a circle.

5.2.2 CNN Filters

To understand the sorts of filters these tiny 3-d CNNs learn, the convolutional filters (without the bias terms) of the eight CNNs when trained on the SVSS data are shown in Figs. 5.8-5.18. Within each sub-figure, a common colorscale is used, in which the color green corresponds to zero, positive values are represented by warmer colors (*i.e.,* reds) and negative values are represented by colder colors (*i.e.,* blues). Each row corresponds to one filter. The 3-d filter cubes are separated in the z dimension (*i.e.,* depth into the sediment) solely to aid in visualization. For the 4-d tensors, filter depth runs horizontally across the page.

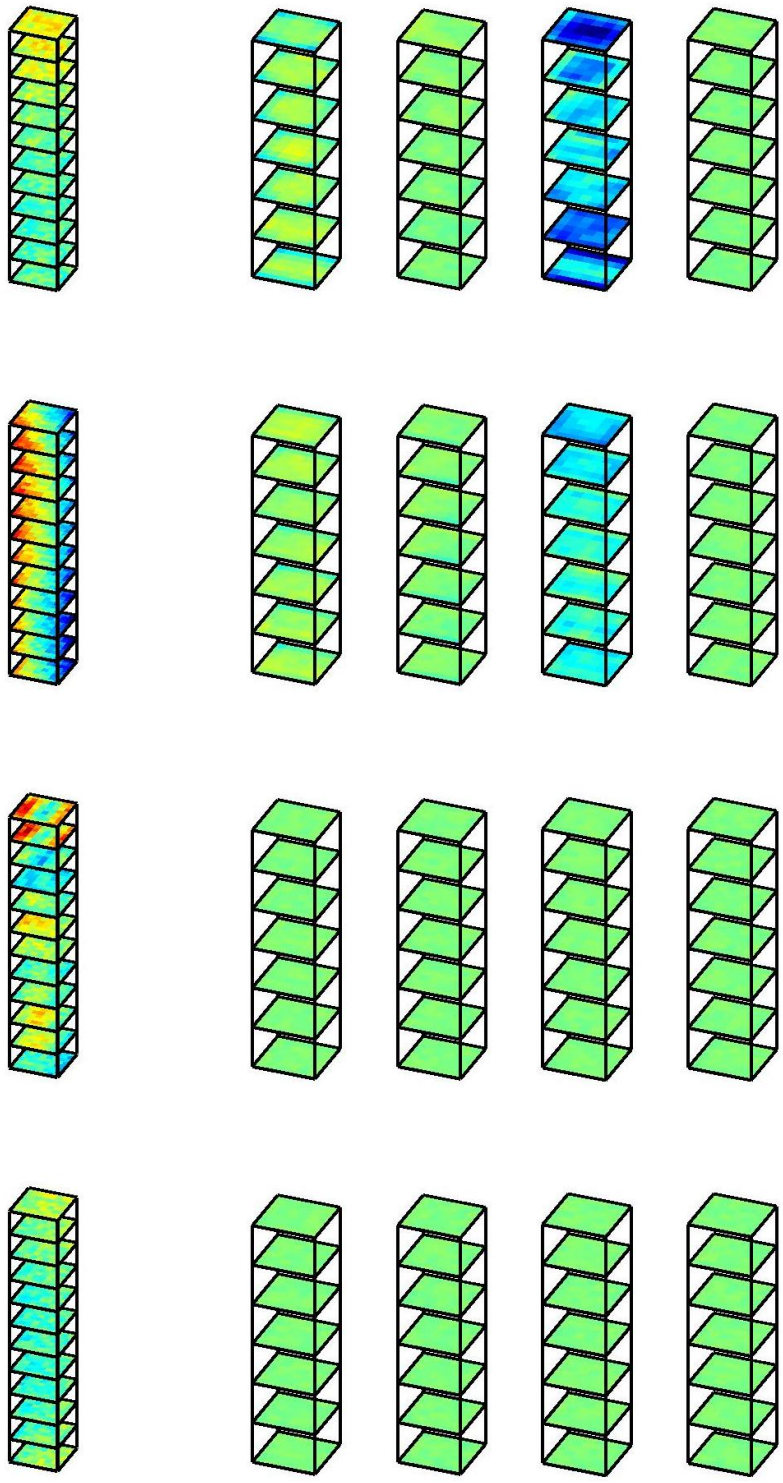
Although general interpretation of the filters is challenging, certain filters are readily understandable. For example, CNN B's second filter of the first convolutional layer is a clear example of a filter that detects specific oriented gradients in 3-d. If hundreds or thousands of filters were employed in each convolutional layer of the CNNs, rapid visualization of the filters would not be feasible.



(a) Convolutional layer 1

(b) Convolutional layer 2

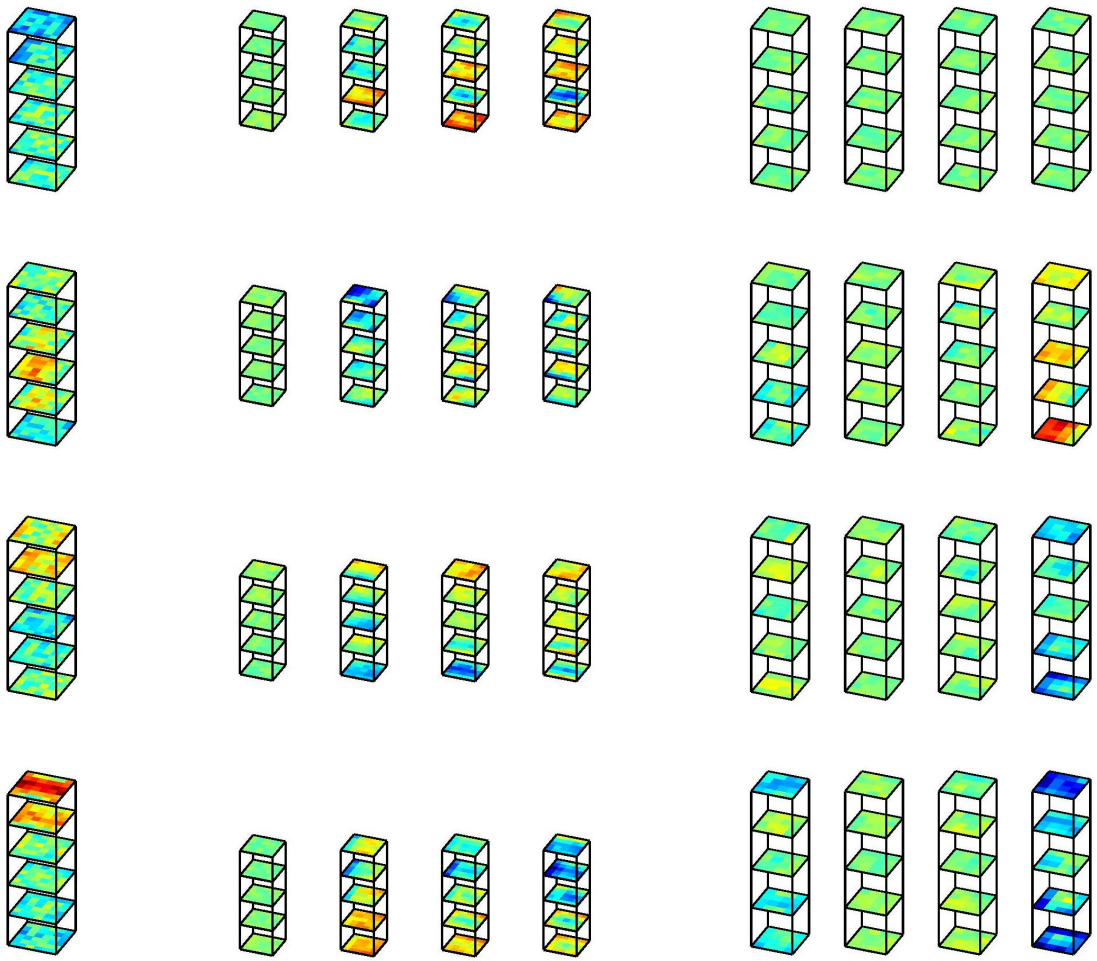
Fig. 5.8. The convolutional filters learned for CNN A using SVSS training data.



(a) Convolutional layer 1

(b) Convolutional layer 2

Fig. 5.9. The convolutional filters learned for CNN B using SVSS training data.

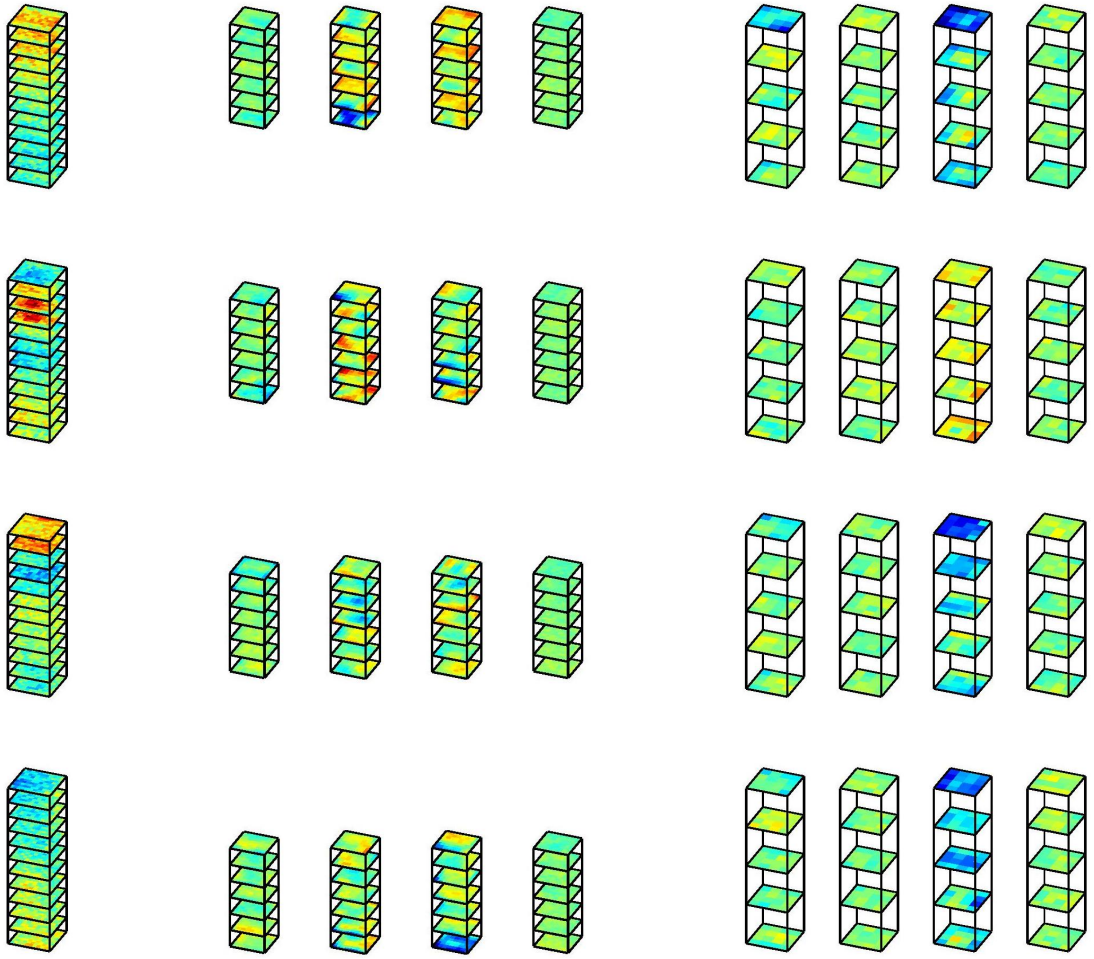


(a) Convolutional layer 1

(b) Convolutional layer 2

(c) Convolutional layer 3

Fig. 5.10. The convolutional filters learned for CNN C using SVSS training data.

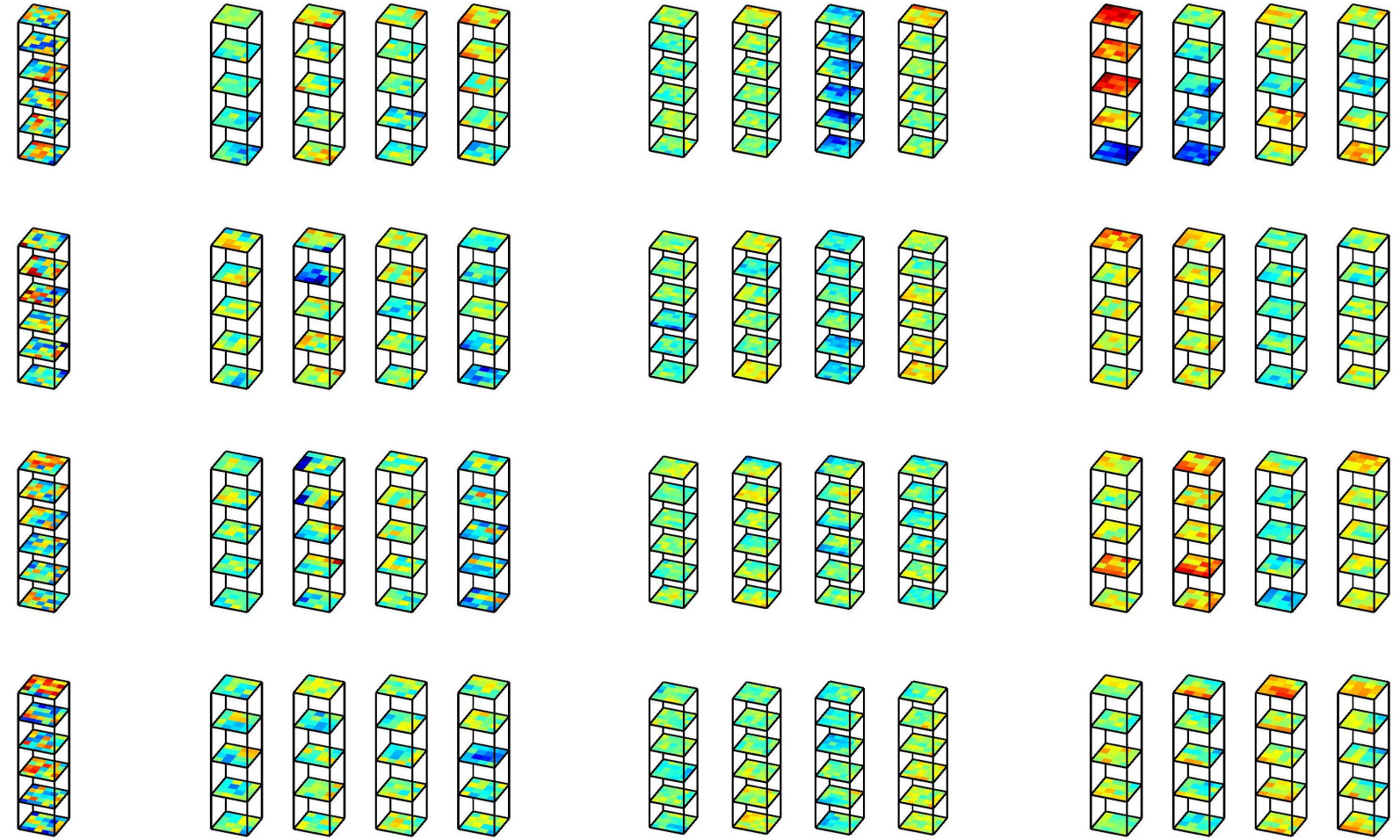


(a) Convolutional layer 1

(b) Convolutional layer 2

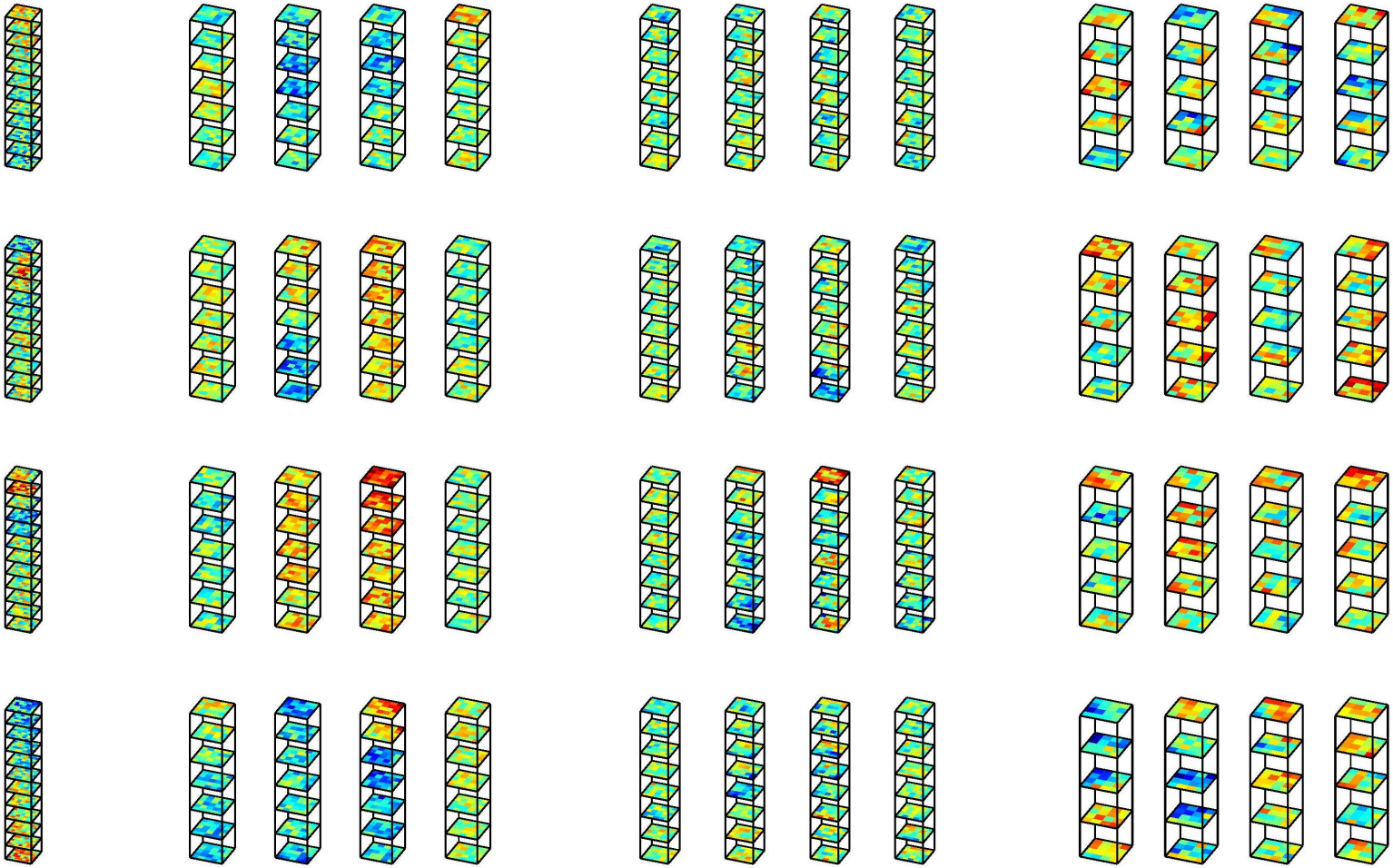
(c) Convolutional layer 3

Fig. 5.11. The convolutional filters learned for CNN D using SVSS training data.



(a) Convolutional layer 1 (b) Convolutional layer 2 (c) Convolutional layer 3 (d) Convolutional layer 4

Fig. 5.12. The convolutional filters learned for CNN E using SVSS training data.



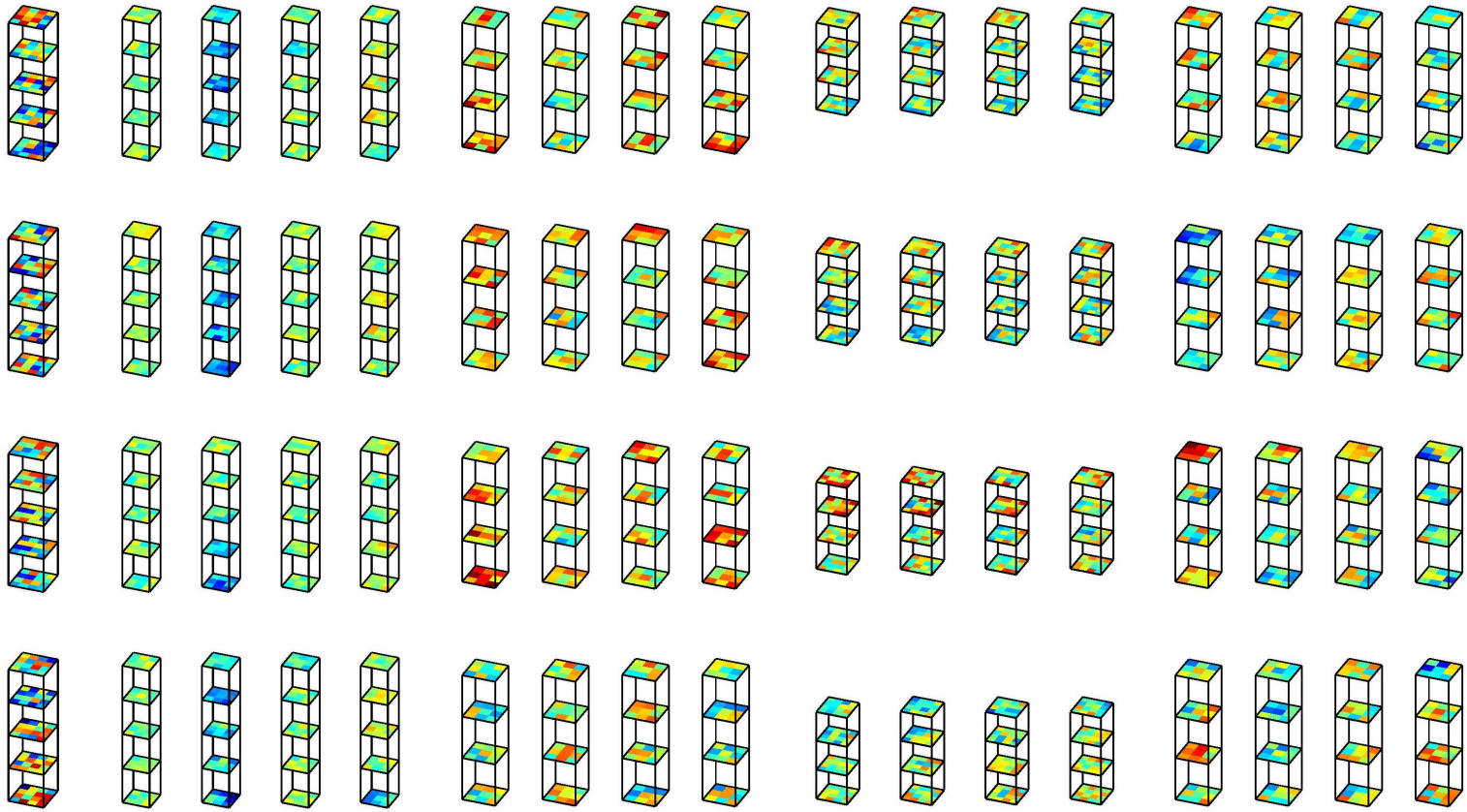
(a) Convolutional layer 1

(b) Convolutional layer 2

(c) Convolutional layer 3

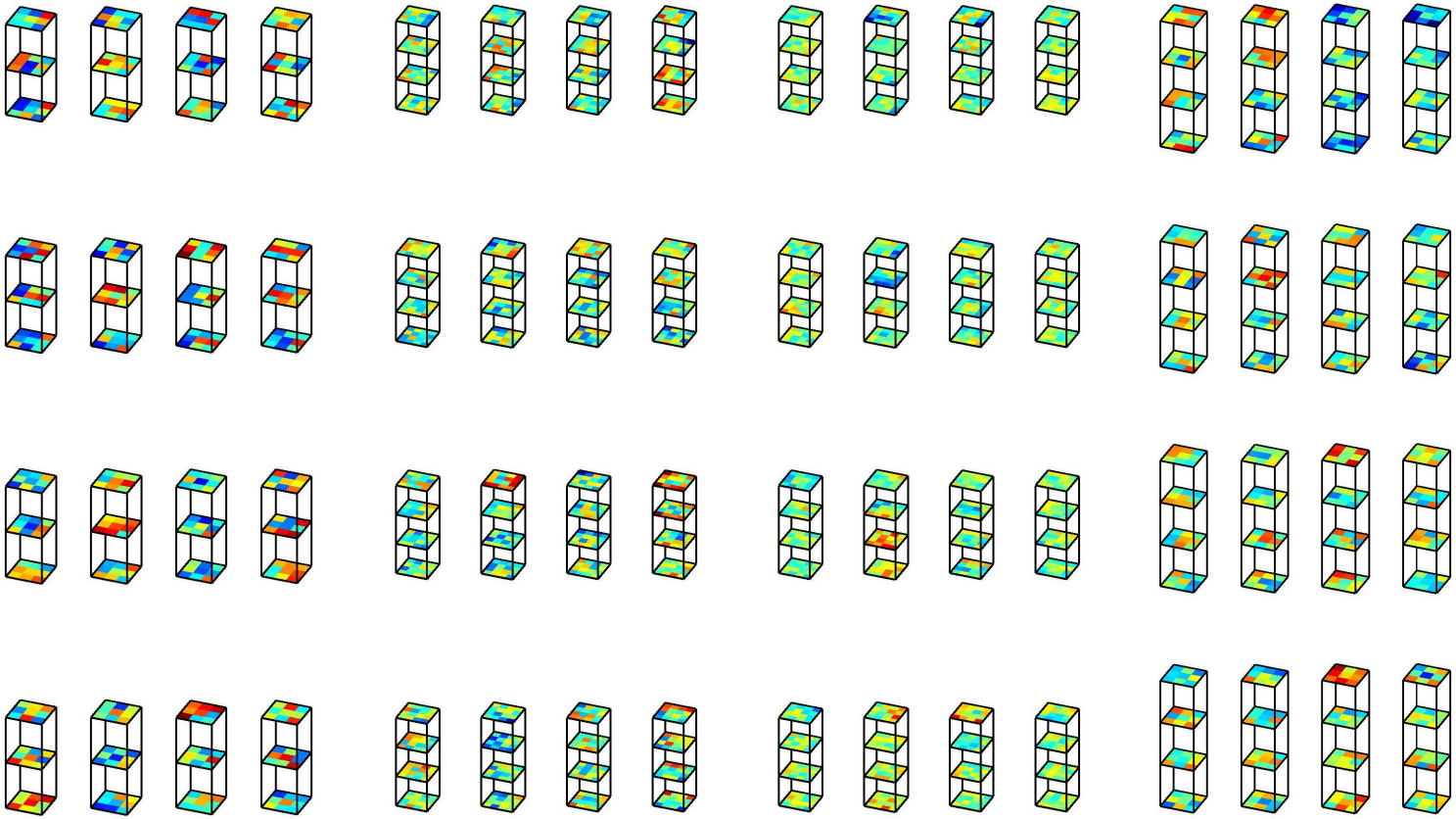
(d) Convolutional layer 4

Fig. 5.13. The convolutional filters learned for CNN F using SVSS training data.



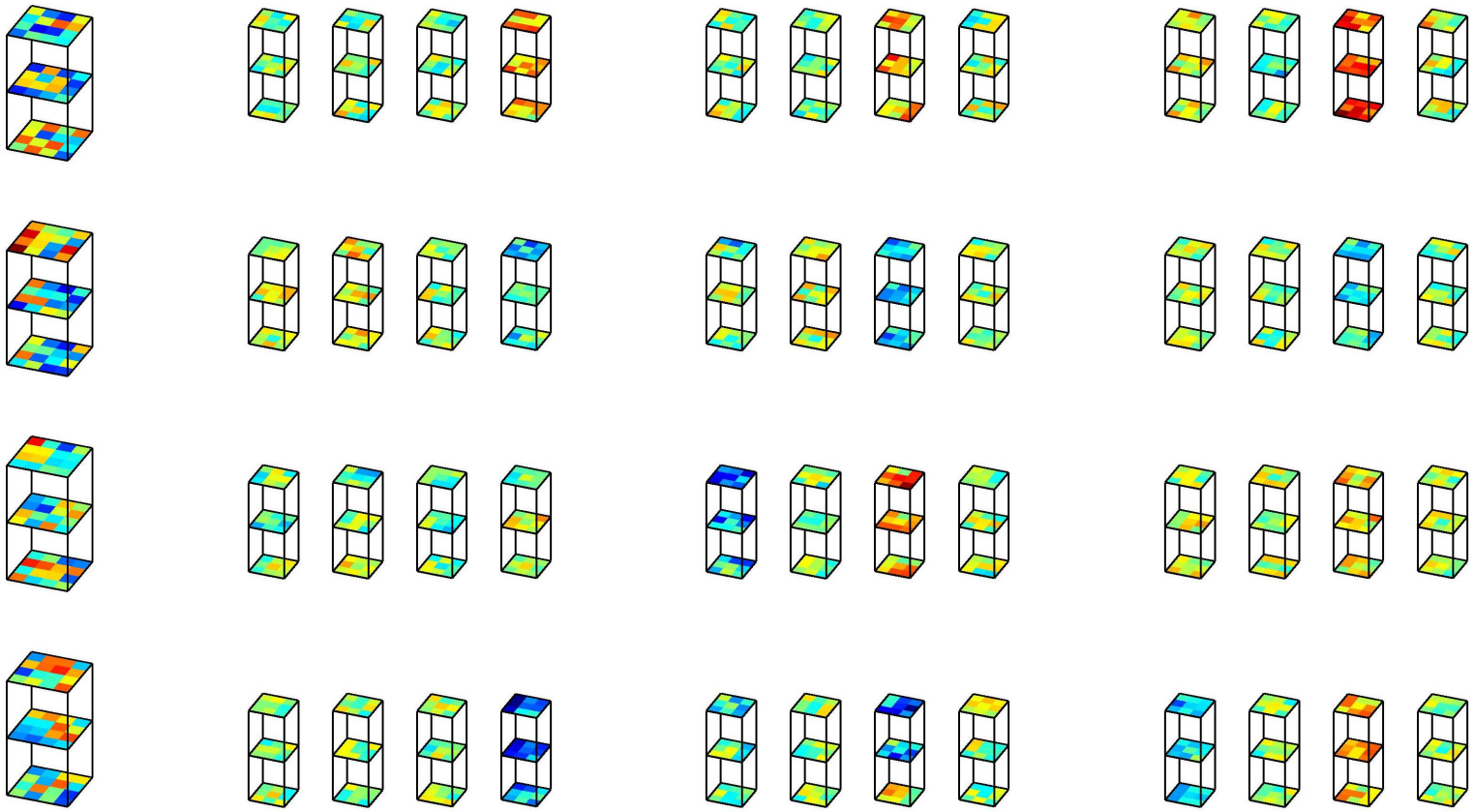
(a) Convolutional layer 1 (b) Convolutional layer 2 (c) Convolutional layer 3 (d) Convolutional layer 4 (e) Convolutional layer 5

Fig. 5.14. The first five layers of convolutional filters learned for CNN G using SVSS training data.



(a) Convolutional layer 6 (b) Convolutional layer 7 (c) Convolutional layer 8 (d) Convolutional layer 9

Fig. 5.15. The last four layers of convolutional filters learned for CNN G using SVSS training data.



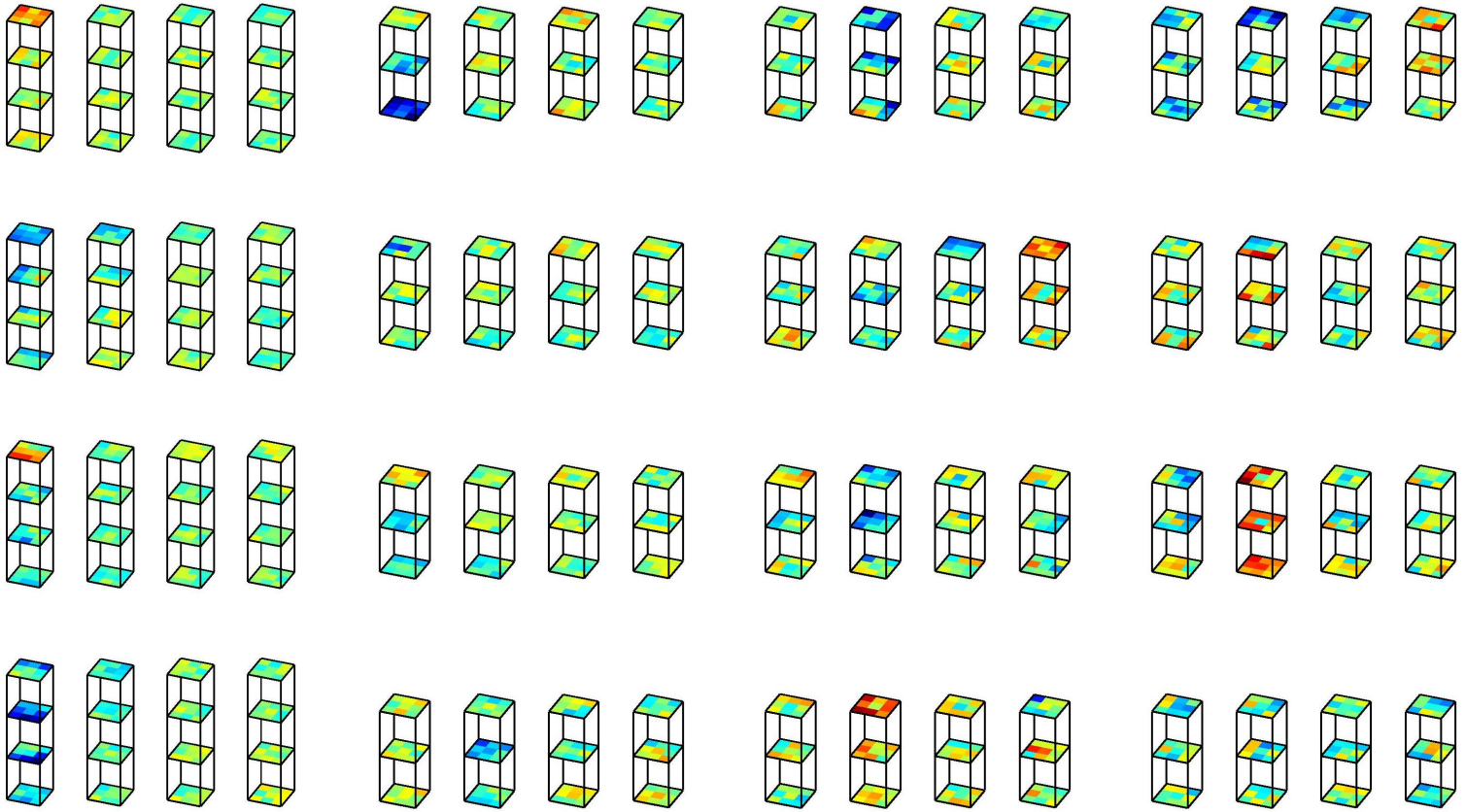
(a) Convolutional layer 1

(b) Convolutional layer 2

(c) Convolutional layer 3

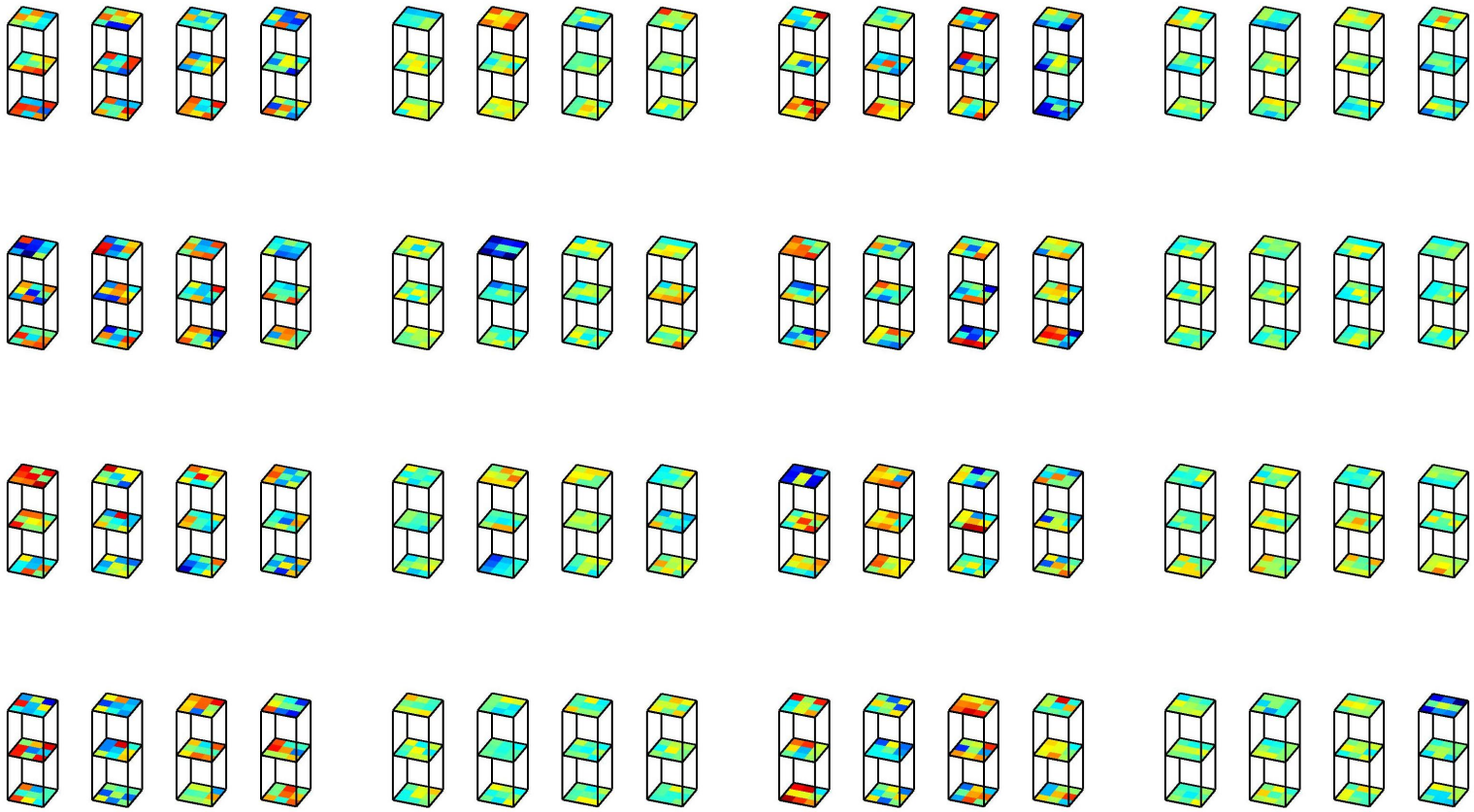
(d) Convolutional layer 4

Fig. 5.16. The first four layers of convolutional filters learned for CNN H using SVSS training data.



(a) Convolutional layer 5 (b) Convolutional layer 6 (c) Convolutional layer 7 (d) Convolutional layer 8

Fig. 5.17. The middle four layers of convolutional filters learned for CNN H using SVSS training data.



(a) Convolutional layer 9 (b) Convolutional layer 10 (c) Convolutional layer 11 (d) Convolutional layer 12

Fig. 5.18. The last four layers of convolutional filters learned for CNN H using SVSS training data.

5.2.3 CNN Intermediate Responses

To illustrate how significantly different clues are exploited by each CNN, we present the intermediate responses at each layer of the eight CNNs for one interesting data cube containing an 81 mm mortar that was buried 10 cm below the water-sediment interface. A photograph of this object (pre-burial) is shown in Fig. 5.19.

Fig. 5.19. Photograph of an 81 mm mortar prior to burial at Sayers site A.

The intermediate responses at each layer of the eight CNNs (trained on SVSS data) for an image cube containing this object are shown in Figs. 5.20-5.27. Because of the difficulty associated with displaying 3-d data, the intermediate responses are shown as a trio of MIPs in a common reference frame. However, it should be remembered that the input data and the intermediate responses are actually fully-populated 3-d volumes.

Examining intermediate responses like these can aid in the interpretation of what clues the filters are exploiting. From the figures, it can be seen that many layers within some of the CNNs perform only subtle transformations that resemble common operations like despeckling and intensity normalization, while others appear to detect oriented gradients or more sophisticated features. CNN D seems to be performing background removal, or equivalently, object segmentation. CNNs E and F appear to compress the elastic energy that initially had spatial extent in the z dimension, and they may perform similar operations with side-lobe energy when it exists in the x and y dimensions. CNN G effectively performs a severe focusing operation, collapsing energy to very localized regions.

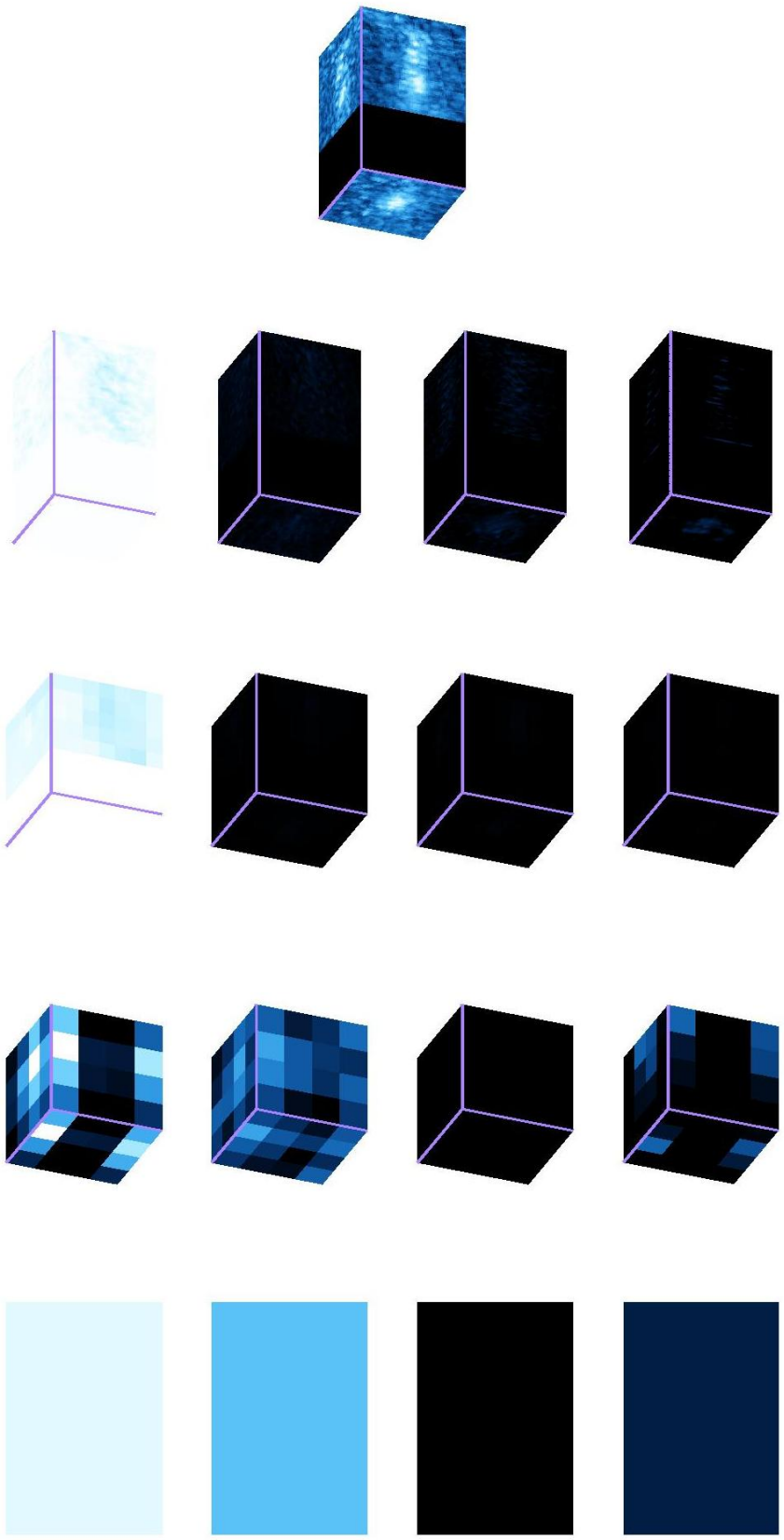


Fig. 5.20. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN A. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.86.

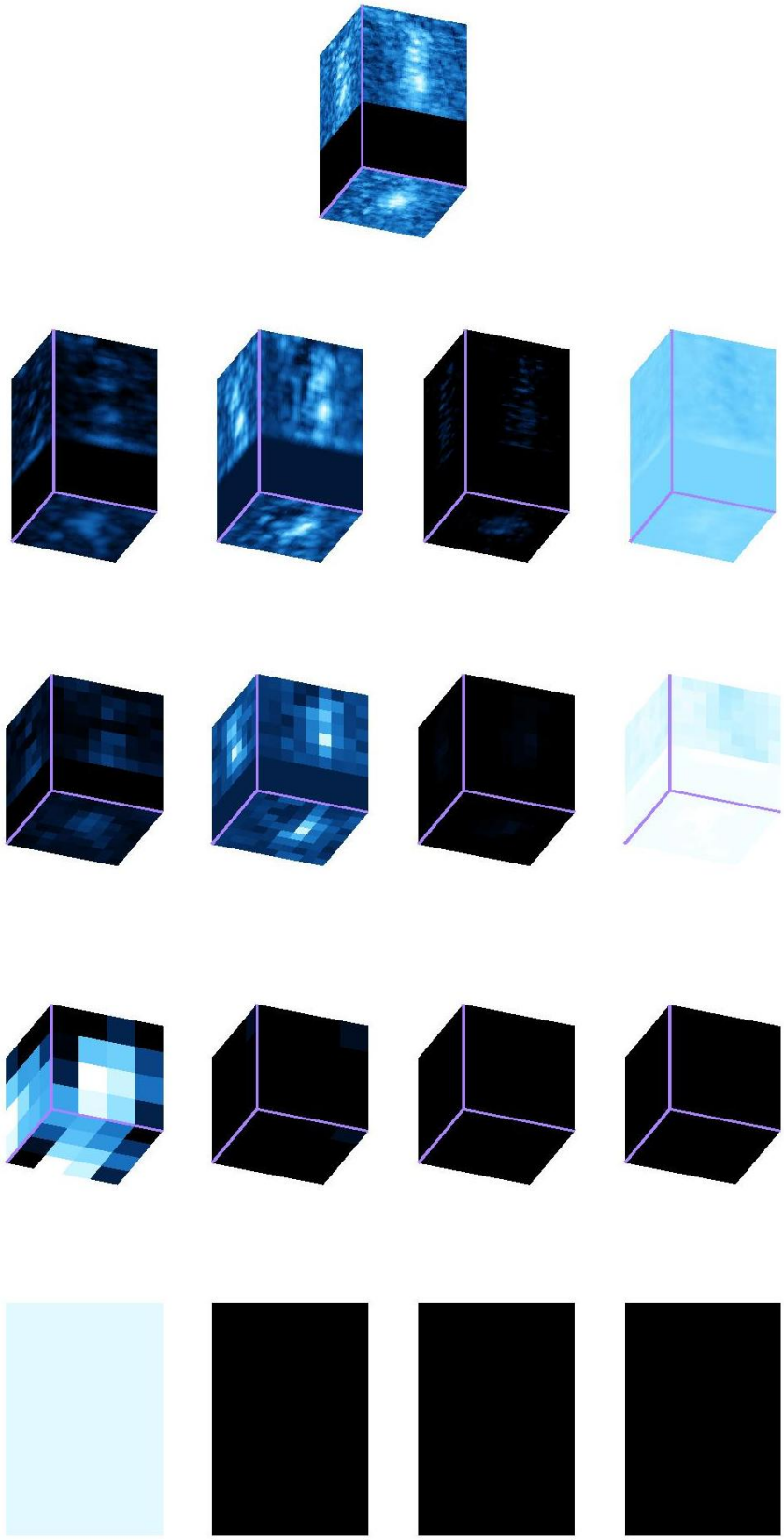


Fig. 5.21. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN B. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.54.

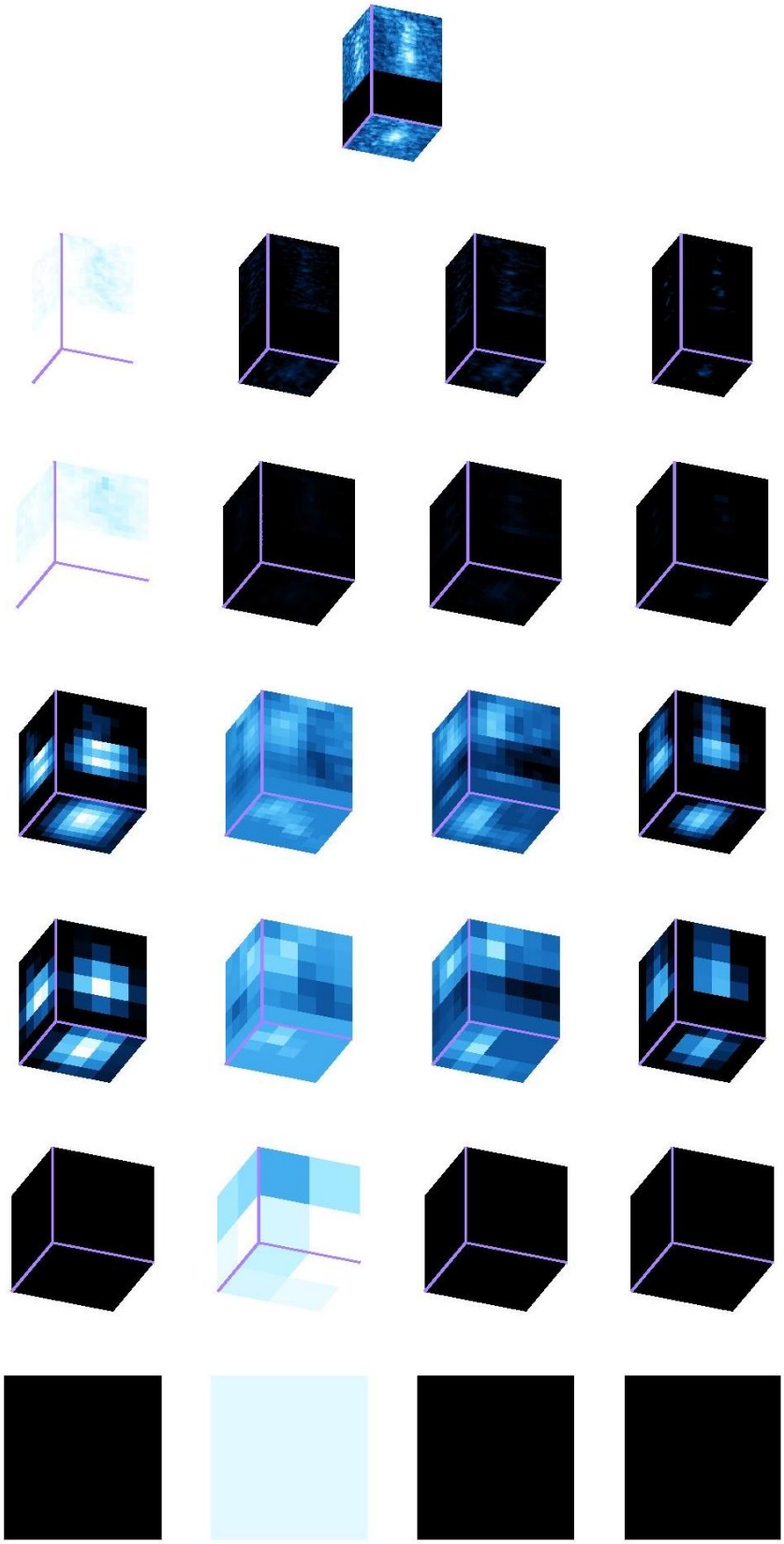


Fig. 5.22. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN C. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.99.

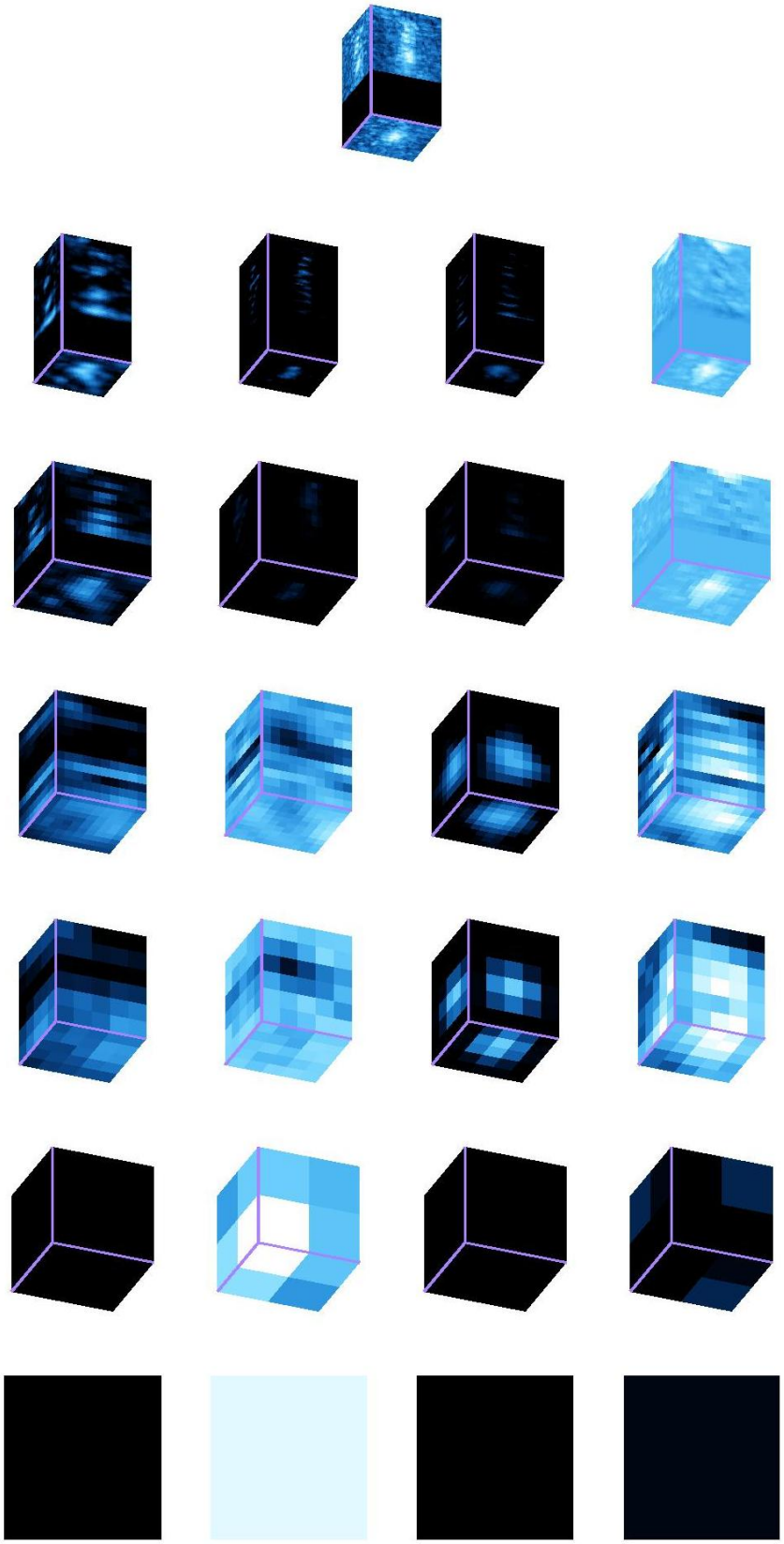


Fig. 5.23. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN D. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.98.

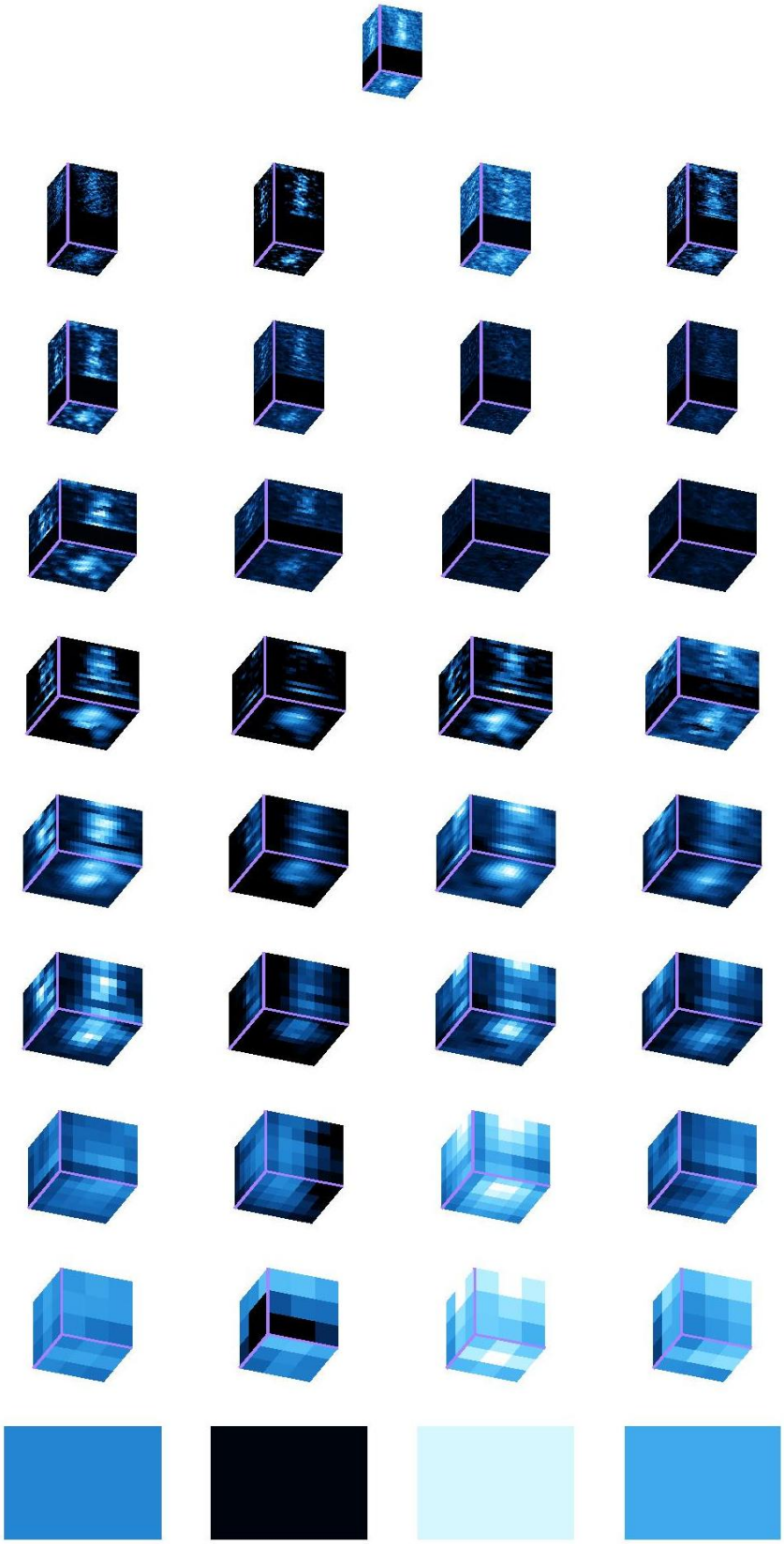


Fig. 5.24. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN E. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.90.

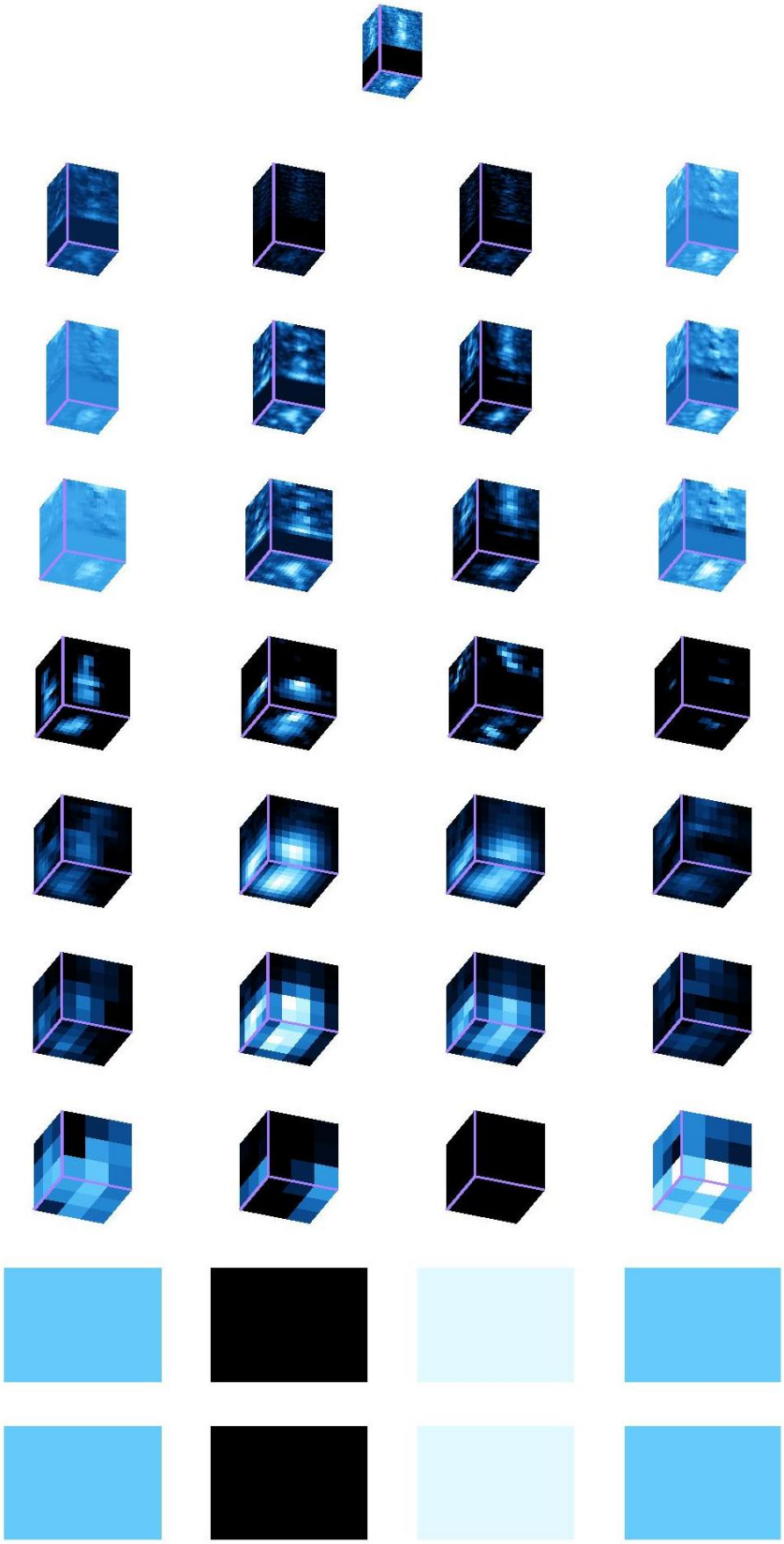


Fig. 5.25. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN F. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.87.

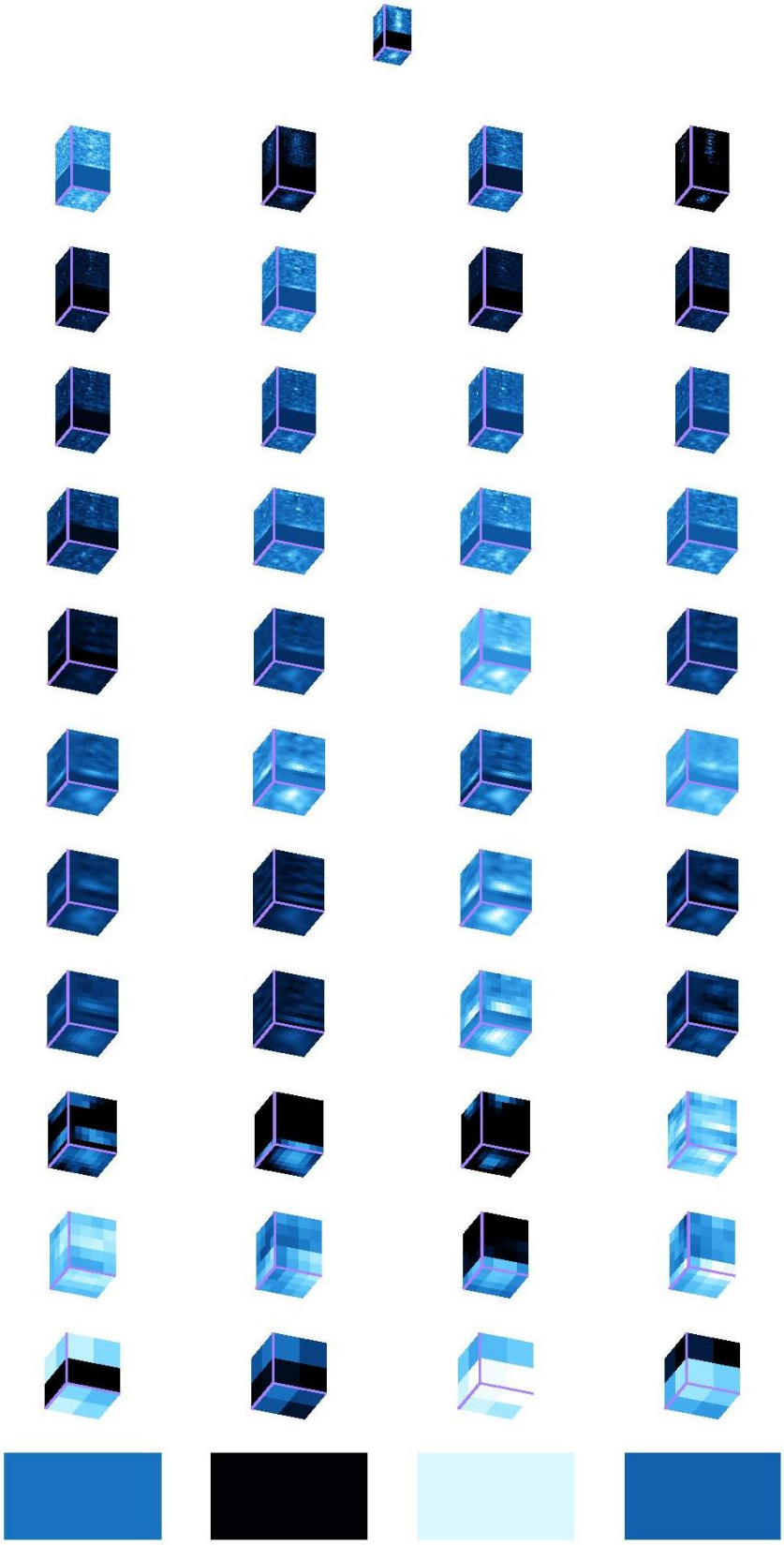


Fig. 5.26. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN G. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.69.

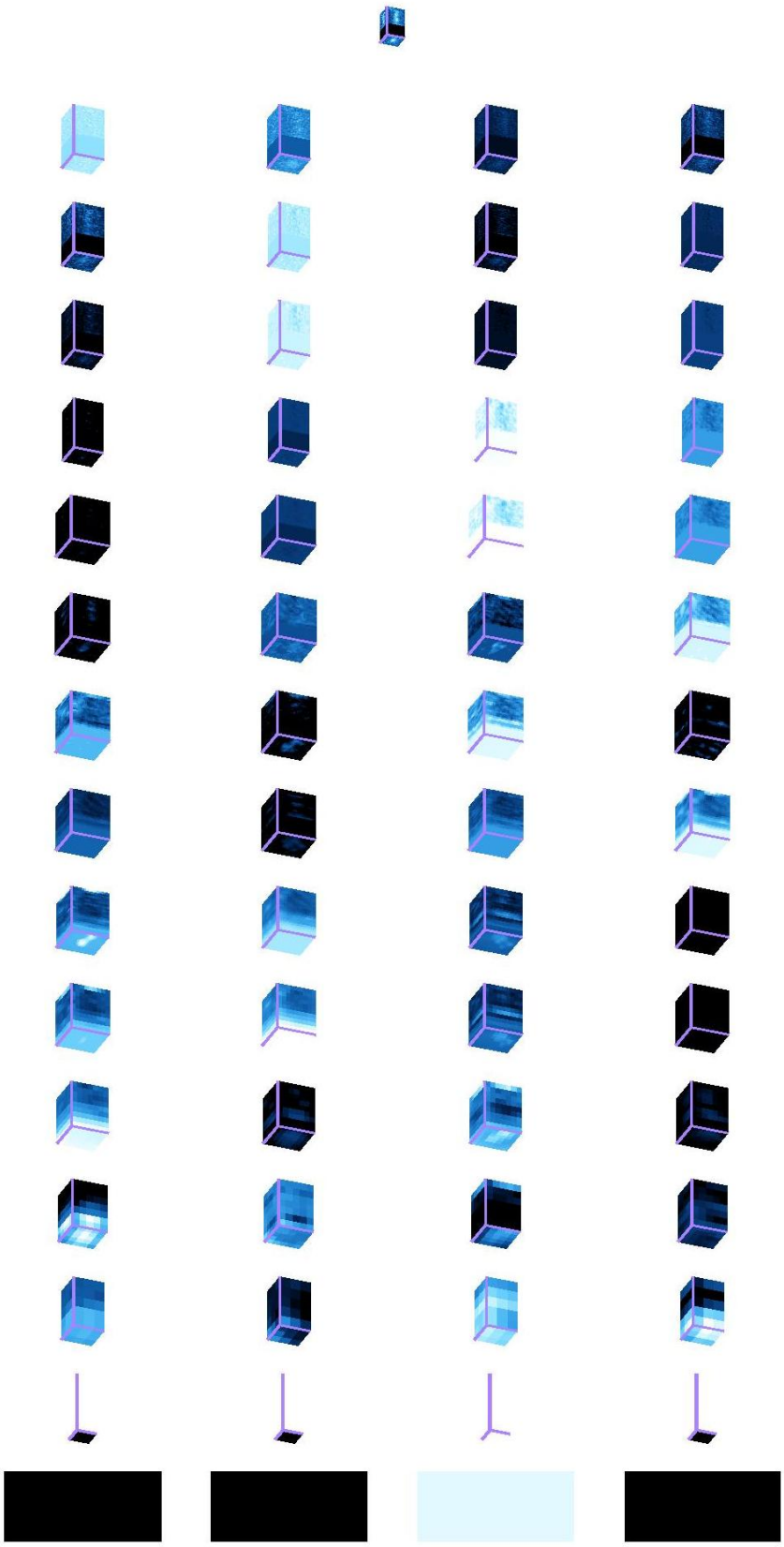


Fig. 5.27. For the data cube in the top row, the intermediate responses (displayed as a trio of MIPs) at each layer of CNN H. Each row corresponds to the output at one layer (convolutional or pooling) of the CNN. The bottom row contains the set of 4 scalar features in the dense layer. With this CNN, the final probability of belonging to the target class was 0.99.

5.3 Limited-Scope Classification

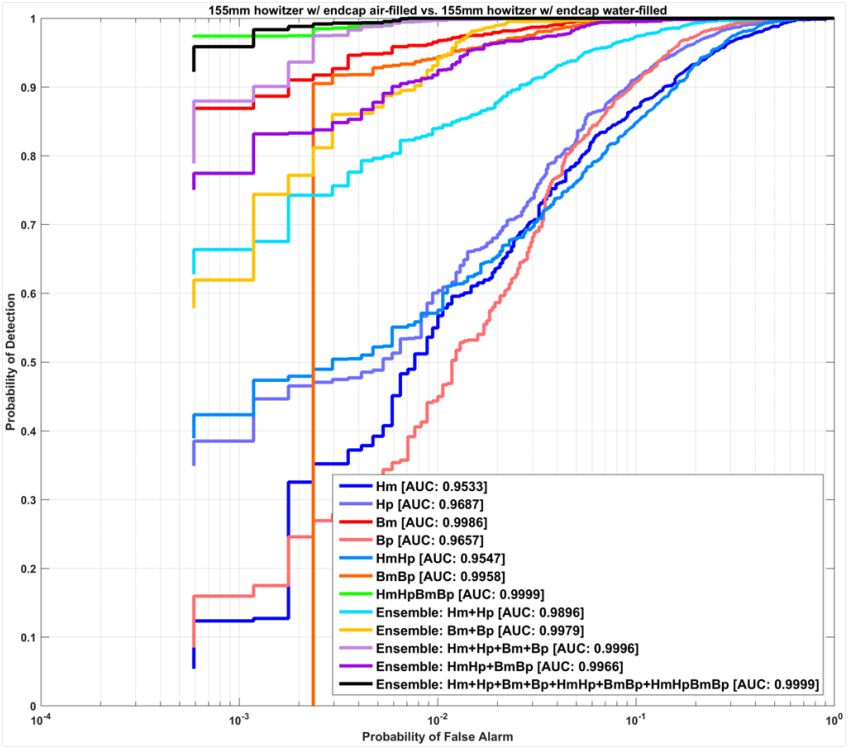
Explainability of classifier predictions can be a useful tool to secure human trust of an algorithm's decision-making process. With an eye toward that long-term goal, experiments were conducted to assess the feasibility of employing CNNs with *simulated* multi-representation acoustic-color sonar data for discriminating air-filled objects from water-filled objects.

CNNs were trained for seven different combinations of input data representations, which are shown in Table 5.2. The performance of the CNNs for the different input data representations is presented in Fig. 5.28. Additionally, ensembles that leverage different combinations of the CNNs (by averaging their individual predictions) are also considered. The AUC of each case is shown in the legend.

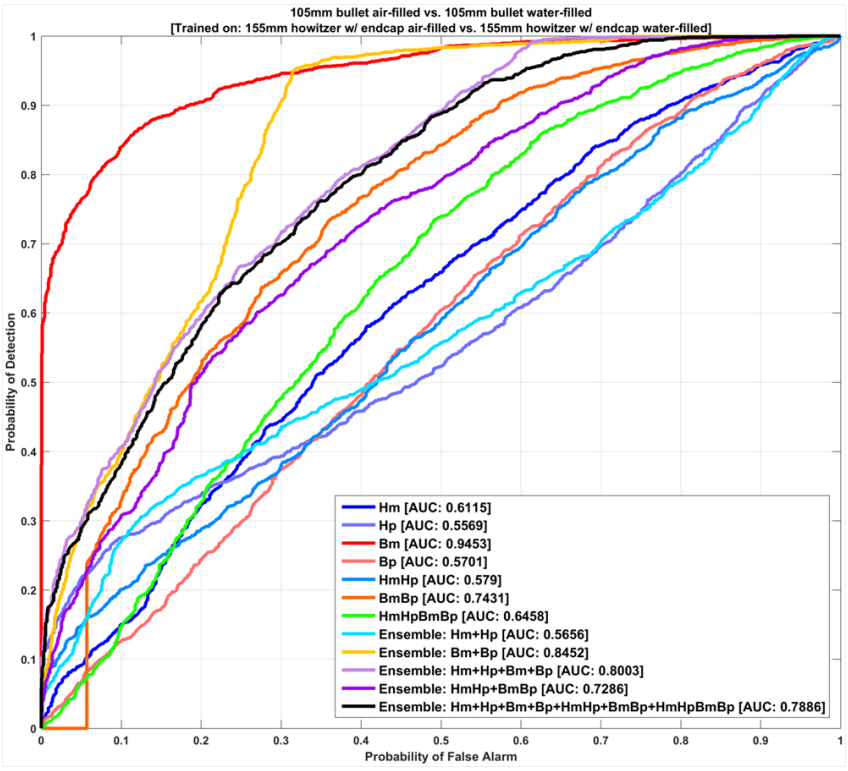
Table 5.2. Acoustic-color data used in each CNN

CNN Label	Number of Representations	Frequency Bands	Representations	Number of Parameters
Hm	1	HF	magnitude	3317
Hp	1	HF	phase	3317
Bm	1	BB	magnitude	1269
Bp	1	BB	phase	1269
HmHp	2	HF	magnitude, phase	6633
BmBp	2	BB	magnitude, phase	2537
HmHpBmBp	4	HF, BB	magnitude, phase	9169

In Fig. 5.28(a), it can be seen that a CNN trained on any of the representations was able to successfully discriminate the air-filled munitions from the water-filled munitions. However, in this case, the training data and test data – although disjoint – all corresponded to the same object, namely 155 mm munitions. As a result, the features that the CNN learned to rely on when discriminating interior fill might be tied to this specific object. A stronger test of CNN generalization ability is shown in Fig. 5.28(b), where the test objects are 105 mm munitions. In this figure, it can be seen that the CNN trained using broadband magnitude acoustic-color data (the red curve) was still able to reliably classify the test objects' fill. This preliminary result suggests that this CNN indeed leverages attributes associated with the object's interior fill, and more importantly, that these clues are ostensibly present in objects other than the specific type used for training. Nevertheless, more extensive experimentation with other objects is still warranted.



(a)



(b)

Fig. 5.28. Classification performance for discriminating air-filled and water-filled objects using 155 mm munitions as training data, and then (a) testing on other 155 mm munitions or (b) testing on 105 mm munitions. (Note the logarithmic horizontal-axis in (a).)

6 Conclusions and Implications for Future Research/Implementation

The work summarized in this report covers only one year of an envisioned four-year project that ended prematurely. Nevertheless, the progress made during this abbreviated period provides a solid foundation from which to further this line of research. Because the algorithms were purposely developed to be functional with measured data from existing systems, they should be readily deployable in a short time frame for use in actual remediation efforts. This result can be achieved by executing the remainder of the original project plan, which includes rigorous testing at new SERDP UXO test-bed sites.

The preliminary results already achieved regarding data normalization, detection, and classification were promising. It was demonstrated how the initial normalization played a vital role in accentuating the difference between target responses and the surrounding environment. This result then facilitated the use of a very computationally efficient detection algorithm, but it also enabled human visualization of the data. The new detection algorithm was capable of identifying and isolating the small fractions of the data cubes that contain information relevant to the UXO remediation problem. Then, the more sophisticated classifier based on deep-learning techniques was shown to provide even better discrimination capability to reduce false alarm rates further. Importantly, it was demonstrated that the CNN-based approach could successfully scale to 3-d data products without incurring computationally prohibitive costs. And finally, the parallel effort to explore limited-scope CNNs to engender explainable classification predictions showed promise using simulated data.

The fundamental limitations on performance imposed by the combination of sensor, target, and environment should be recognized. It is important to identify the regimes in which the physics simply does not support successful detection and classification, regardless of the algorithm employed. In this vein, it is recommended that additional measured data are obtained from different environments so that the developed algorithms can be assessed more fully. The importance of having ground-truth information for training classifiers and evaluating algorithm performance also cannot be emphasized enough. Thus far, only a modest amount of MuST data was available to work with, so evaluating the algorithms on larger MuST data sets should be a priority in the future. Investigating additional alternative data representations beyond the image domain, such as acoustic color, within the CNN context is also a potentially fruitful avenue for further research.

The algorithms presented here addressed a capability gap as they were developed expressly for two new systems, the SVSS and MuST, for which no ATR algorithms previously existed. Provided further refinement and rigorous testing of the algorithms are undertaken successfully, there is great potential for these methods to be leveraged in remediation efforts at contaminated underwater sites. And in the event that they are indeed deployed, fewer resources should be spent investigating harmless clutter and the cost of remediation should decrease substantially.

References

[1] D. Williams, "Cubist-Inspired Deep Learning with Sonar for UXO Detection and Classification," Tech. Rep., SERDP Project MR-1844 Final Report, September 2019.

[2] D. Brown, C. Brownstead, and S. Johnson, "Sediment Volume Search Sonar Development," Tech. Rep., SERDP Project MR-2545 Final Report, forthcoming 2021.

[3] K. Williams, T. McGinnis, V. Miller, B. Brand, and R. Light, "Limited scope design study for multi-sensor towbody," Tech. Rep., SERDP Project MR-2501 Final Report, June 2016.

[4] T. Marston, D. Plotnick, and K. Williams, "Three dimensional fast factorized back projection for sub-sediment imaging sonars," in *Proceedings of IEEE OCEANS*, 2019, pp. 1–6.

[5] H. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*, John Wiley & Sons, 2004.

[6] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.

[8] D. Williams, "On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 1, pp. 236–260, 2021.

[9] J. Wallis and T. Miller, "Three-dimensional display in nuclear medicine and radiology," *Journal of Nuclear Medicine*, vol. 32, no. 3, pp. 534–546, 1991.

[10] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[11] S. Schock, A. Tellier, J. Wulf, J. Sara, and M. Ericksen, "Buried object scanning sonar," *IEEE Journal of Oceanic Engineering*, vol. 26, no. 4, pp. 677–689, 2001.

[12] D. Brown, S. Johnson, I. Gerg, and C. Brownstead, "Simulation and testing results for a sub-bottom imaging sonar," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2019, vol. 36, pp. 1–12.

[13] D. Brown, D. Cook, and J. Fernandez, "Results from a small synthetic aperture sonar," in *Proceedings of IEEE OCEANS*, 2006, pp. 1–6.

[14] A. Hunter and R. van Vossen, "Sonar target enhancement by shrinkage of incoherent wavelet coefficients," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 262–268, 2014.

[15] J. Staal, B. van Ginneken, and M. Viergever, "Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data," *Medical Image Analysis*, vol. 11, no. 1, pp. 35–46, 2007.

[16] B. van Ginneken, A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 286–289.

[17] S. Nadimi and B. Bhanu, "Physical models for moving shadow and object detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1079–1087, 2004.

[18] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 817–825.

[19] C. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, vol. 1, Springer Science & Business Media, 2003.

[20] S. Bourennane, C. Fossati, and A. Cailly, "Improvement of target-detection algorithms based on adaptive three-dimensional filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 4, pp. 1383–1395, 2010.

[21] J. Wilson, P. Gader, W. Lee, H. Frigui, and K. Ho, "A large-scale systematic evaluation of algorithms using ground-penetrating radar for landmine detection and discrimination," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 8, pp. 2560–2572, 2007.

[22] C. Ratto, P. Torrione, and L. Collins, "Exploiting ground-penetrating radar phenomenology in a context-dependent framework for landmine detection and discrimination," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1689–1700, 2011.

[23] D. Sternlicht, J. Harbaugh, A. Shah, M. Webb, and R. Holtzapple, "Buried object classification using a sediment volume imaging sas and electromagnetic gradiometer," in *Proceedings of IEEE OCEANS*, 2006, pp. 1–6.

[24] I. Gerg, "The advanced synthetic aperture sonar imaging engine (ASASIN), a time-domain backprojection beamformer using graphics processing units," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3347–3347, 2016.

[25] D. Brown, C. Brownstead, and S. Johnson, "Sediment Volume Search Sonar Development (Phase II)," Tech. Rep., SERDP Project MR-2545 Interim Report, May 2018.

[26] J. Sara, "Next generation buried object scanning sonar (BOSS) for detecting buried UXO in shallow water," Tech. Rep., SERDP Project MR-2752 Final Report, November 2018.

[27] R. Urick, *Principles of Underwater Sound*, McGraw-Hill, 1983.

[28] D. Jackson and M. Richardson, *High-frequency seafloor acoustics*, Springer Science & Business Media, 2007.

[29] R. Hansen, H. Callow, T. Sæbø, and S. Synnes, "Challenges in seafloor imaging and mapping with synthetic aperture sonar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3677–3687, 2011.

[30] Z. Lowe and D. Brown, "Multipath reverberation modeling for shallow water acoustics," in *Proceedings of the 11th European Conference on Underwater Acoustics*, 2012, pp. 1285–1291.

[31] D. Jackson, "APL-UW high-frequency ocean environmental acoustic models handbook," *Applied Physics Laboratory, University of Washington, Technical Report*, vol. 9407, no. 102, pp. 1499–1510, 1994.

[32] D. Williams, "The Mondrian detection algorithm for sonar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1091–1102, 2018.

[33] G. Dobeck, J. Hyland, and L. Smedley, "Automated detection/classification of seamines in sonar imagery," in *Proceedings of the SPIE International Society of Optics*, 1997, vol. 3079, pp. 90–110.

[34] R. Fandos, A. Zoubir, and K. Siantidis, "Unified design of a feature based ADAC system for mine hunting using synthetic aperture sonar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2413–2426, 2014.

[35] G. Facciolo, N. Limare, and E. Meinhardt-Llopis, "Integral images for block matching," *Image Processing On Line*, vol. 4, pp. 344–369, 2014.

[36] S. Kargl, K. Williams, T. Marston, J. Kennedy, and J. Lopes, "Acoustic response of unexploded ordnance (UXO) and cylindrical targets," in *Proceedings of IEEE OCEANS*, 2010, pp. 1–5.

[37] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[38] S. Kargl, "Acoustic response of underwater munitions near a sediment interface: Measurement model comparisons and classification schemes," Tech. Rep., SERDP Project MR-2231 Final Report, April 2015.

[39] G. Sammelmann, "High-frequency images of proud and buried 3D-targets," *Proceedings of IEEE OCEANS*, pp. 266–272, 2003.

A List of Scientific/Technical Publications

- D. Williams and D. Brown, "New Target Detection Algorithms for Volumetric Synthetic Aperture Sonar Data," *Proceedings of Meetings on Acoustics*, Vol. 40, September 2020.
- D. Williams and D. Brown, "Three-Dimensional Convolutional Neural Networks for Target Classification With Volumetric Sonar Data," to be submitted to *Proceedings of Meetings on Acoustics*, May 2021.