Acoustic-Color-Based Convolutional Neural Networks for UXO Classification With Low-Frequency Sonar Data

David P. Williams

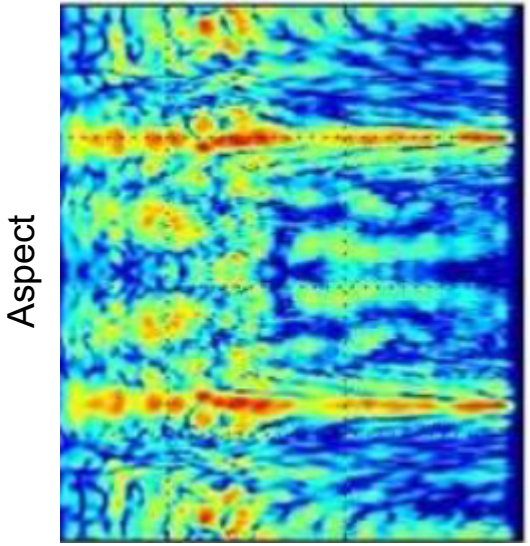
NATO STO Centre for Maritime Research and Experimentation

SERDP • ESTCP SYMPOSIUM
2019 | Enhancing DoD's Mission Effectiveness

Motivation

- Low-frequency sonar response
- Acoustic-color plots
- Robust features

- This work: Investigate whether CNNs can be trained with unconventional input “imagery” like acoustic-color plots for UXO discrimination.

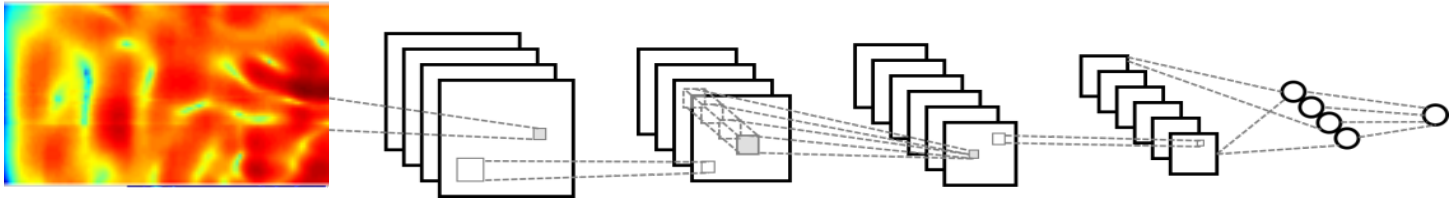


Images courtesy of Steve Kargl (APL/UW)

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

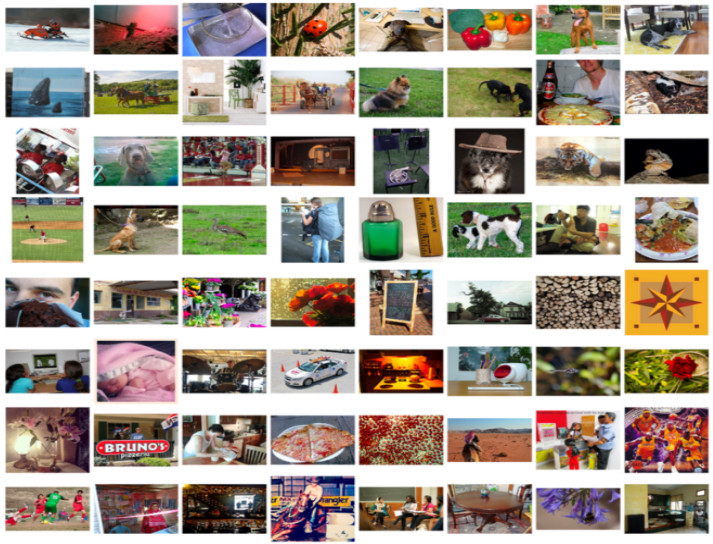
Outline

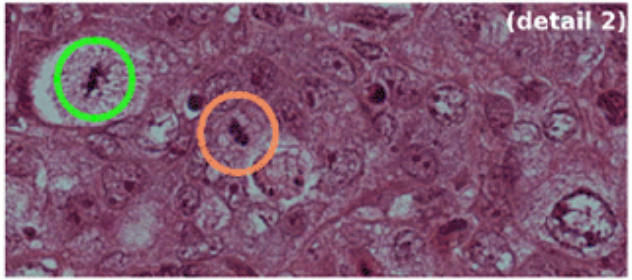
I. Background on Deep Learning for ATR
II. CNNs with Acoustic-Color Data
III. Experimental Results
IV. Conclusions

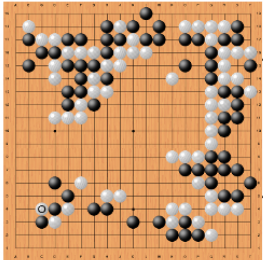


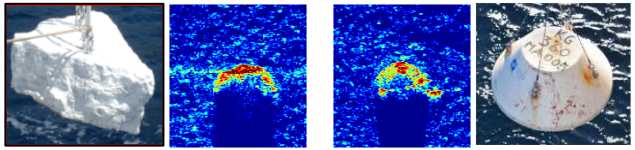
CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Deep Learning in the Wild





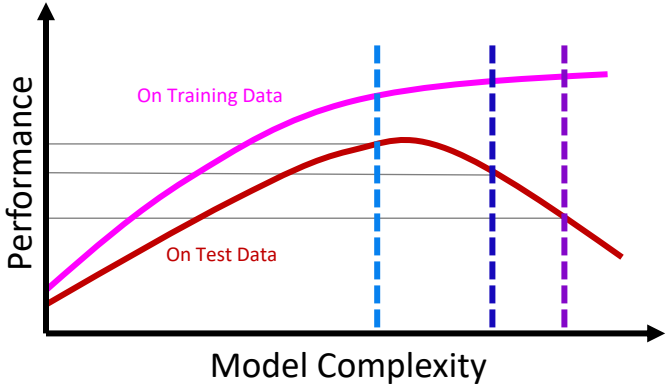


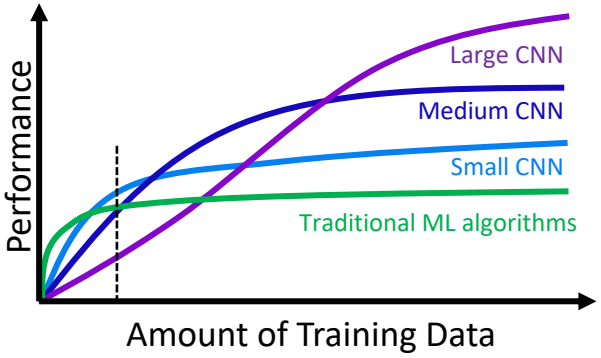


[1] Russakovsky et al., ”ImageNet Large Scale Visual Recognition Challenge,” *IJCV,* 2015.
[2] Ciresan et al., “Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks,” MICCAI, 2013.
[3] Silver et al., “AlphaGo, Mastering the game of Go without human knowledge,” *Nature,* Vol. 550, 2017.
[4] @teenybiscuit

Deep Learning

- Richness of the architecture permits more sophisticated decision surfaces than traditional ("shallow") classifiers
- Traditional shallow classifiers hit performance plateau: beyond a point, more data doesn't improve performance
- DL *continues to improve* as data increases: natural way to leverage huge (and ever-increasing) amounts of experimental data
- Deep Convolutional Neural Networks (CNNs): state-of-the-art for image classification tasks
- **CNN *automatically* learns the most useful bases in which to represent the data – no need to extract predefined features**

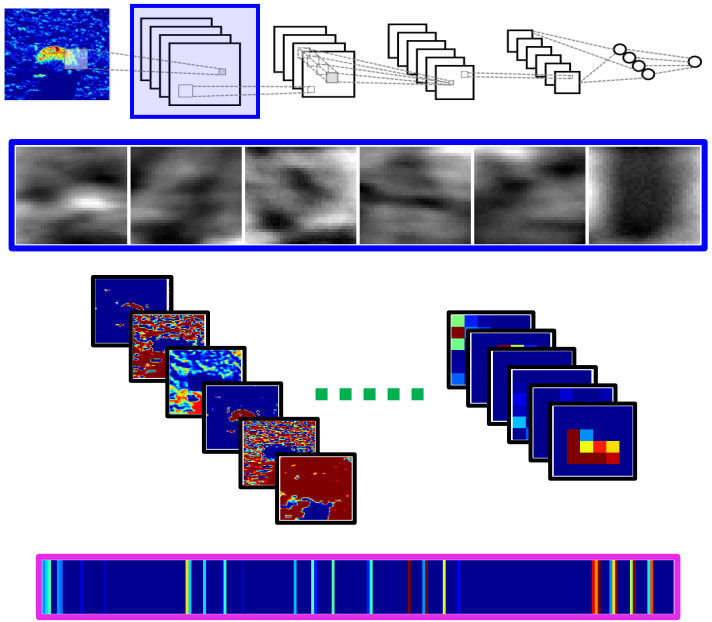




- But: need enough data to avoid overfitting
- Key Ratio: Model Complexity / Training Data
- Sufficient amount of training data to support the model's complexity
- Insufficient amount of training data to support the model's complexity (too many parameters)

Convolutional Neural Networks (CNNs)

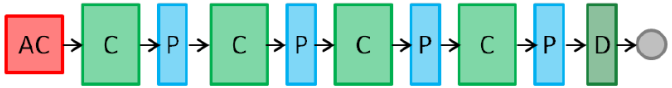
- CNNs are not magic!
  - "Vanilla" CNN: sequences of convolutional layers, nonlinear activations, and pooling
  - Training a CNN simply means learning the filters of the convolutional layers
  - Instead of using pre-defined filters, the CNN *learns* what filters should be from the data
  - These filters transform the input data (imagery) into a new representation space
  - With successive convolutional layers, the level of data abstraction and (highly nonlinear) decision-surface complexity increases
  - At the end of the CNN, have features in a space that should be easily separable by class
- The trick: lots of free parameters to learn, so need lots of data (bigger the CNN, more data required)



❖ *CNN = learning a transformation to map an input image into a new representation space where classes separate*

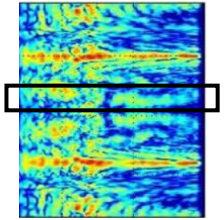
CNNs for Acoustic-Color Data

- Consider 9 CNNs (with 2-12 convolutional layers), common high-level architectures
  - 2/3/**4** convolutional blocks
  - 1/2/3 convolutional layers per block



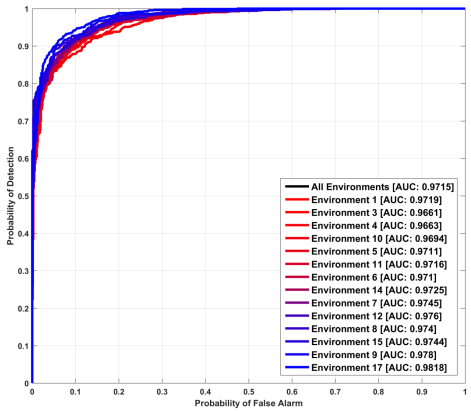
- Average pooling (large factors)
- ReLU activation functions
- 4 filters per convolutional layer
- Filter sizes and pooling factors dependent on input data size
- 4 nodes ("features") at final dense layer
- ~600-4000 parameters to learn per CNN
  - Tiny number compared to typical optical image CNNs
  - Eases training data requirements

- All training *ab initio* ("from scratch")
- **Input data**: size varies depending on data "chunk" and CNN
- Output: probability of belonging to UXO class
- Batch size of 128 (64 per class)
- Loss function: binary cross-entropy
- Learning rate of 0.001 (RMSprop optimization)
- "On-the-fly" data augmentation: reflection in aspect
- Train for 100 epochs
  - 1 epoch is a full pass through training set (~500 batches)
  - Timing: ~1 minute / epoch
- Python with TensorFlow software library +1 GPU (GTX 1080)

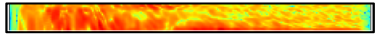


A Note on Performance Metrics

- ATR and machine learning communities evaluate classification performance using metrics like ROC curves, AUC, precision/recall, F_1-score, etc.
  - ***Not appropriate metrics for actual UXO remediation efforts***
- UXO remediation demands very high (at/near unity) probabilities of detection
- Better metric: probability of false alarm at key probabilities of detection
  - e.g., "How many false alarms would we sustain if we ensured that we found all the UXO?"
- Present classification results here mainly in terms of ROC curves and AUC
  - Later will show P_{fa} at key P_d



Example CNN Results

- Target and Reverberation Experiment 2013 (TREX13) data
- Input data for CNNs: Acoustic color plot over 3-30 kHz and X aspect degrees 
- Divide data into disjoint training/test sets ***by object***
- Training set: 67,680 “images” from 47 full AC plots

Test data objects

Object Index	Description
5	5:1 aspect telephone pole section
6	55-gallon drum, water-filled
7	3′ aluminum cylinder
10	panel target
14	scuba tank, water-filled, with stem
17	2′ aluminum cylinder
18	cement block
19	tire
8	155 mm howitzer without collar
12	81 mm mortar
21	steel UXO replica
22	original material UXO
28	155 mm howitzer with collar
29	bullet #2

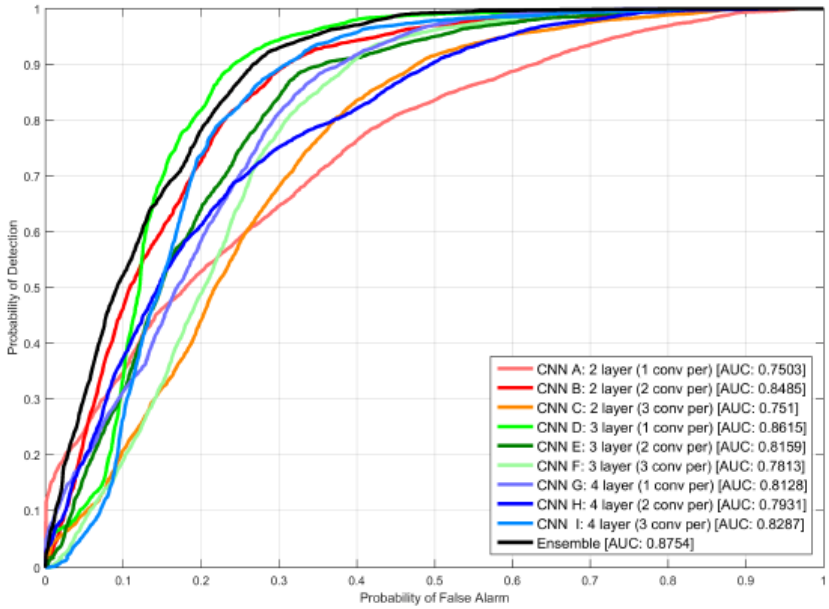


Fig. 5.4. For the 30° input data case, classification performance using the various CNNs.

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Results

	Input Data			
	10°	30°	40°	90°
CNN A	0.7302	0.7503	0.7975	0.7063
CNN B	0.8491	0.8485	0.8007	0.8091
CNN C	0.7826	0.7510	0.7987	0.6273
CNN D	0.8189	0.8615	0.8321	0.8172
CNN E	0.7775	0.8159	0.8005	0.8002
CNN F	0.8496	0.7813	0.7353	0.8227
CNN G	0.8886	0.8128	0.8037	0.8633
CNN H	0.8085	0.7931	0.8395	0.6561
CNN I	0.7620	0.8287	0.7354	0.8313
$\mathcal{E}$(A-I)	0.8534	0.8754	0.8985	0.9198

Range	Input Data			
	10°	30°	40°	90°
[10 m, 15 m)	0.5739	0.5949	0.5573	0.7595
[15 m, 20 m)	0.6859	0.8058	0.7809	0.8290
[20 m, 25 m)	0.9187	0.8874	0.8914	0.9500
[25 m, 30 m)	0.7914	0.8388	0.8871	0.9292
[30 m, 35 m)	0.8863	0.9000	0.9678	0.9877
[35 m, 40 m)	0.9984	0.9988	1.0000	1.0000

Object Index	Description	Input Data			
		10°	30°	40°	90°
5	5:1 aspect telephone pole section	0.865	0.895	0.920	0.940
6	55-gallon drum, water-filled	0.993	0.996	0.999	0.995
7	3′ aluminum cylinder	0.990	0.997	0.998	0.986
10	panel target	0.779	0.791	0.896	0.925
14	scuba tank, water-filled, with stem	0.741	0.842	0.881	0.939
17	2′ aluminum cylinder	0.550	0.535	0.534	0.686
18	cement block	0.787	0.882	0.795	0.839
19	tire	0.920	0.778	0.766	0.678
8	155 mm howitzer without collar	0.781	0.826	0.865	0.924
12	81 mm mortar	0.851	0.815	0.825	0.888
21	steel UXO replica	0.881	0.900	0.910	0.916
22	original material UXO	0.873	0.876	0.888	0.905
28	155 mm howitzer with collar	0.834	0.878	0.920	0.901
29	bullet #2	0.924	0.942	0.967	0.979

Center Aspect	Input Data			
	10°	30°	40°	90°
[−90°, −60°)	0.9237	0.9263	0.9428	0.9197
[−60°, −30°)	0.8155	0.8634	0.8941	0.9271
[−30°, 0°)	0.8431	0.9010	0.9007	0.9680
[0°, 30°)	0.8835	0.8950	0.9518	0.9400
[30°, 60°)	0.8013	0.8180	0.8515	0.8637
[60°, 90°)	0.8585	0.8674	0.8502	0.8964
[90°, 120°)	0.8589	0.8691	0.8517	0.8962
[120°, 150°)	0.8015	0.8183	0.8500	0.8639
[150°, 180°)	0.8828	0.8925	0.9511	0.9383
[180°, 210°)	0.8448	0.9030	0.9024	0.9682
[210°, 240°)	0.8145	0.8622	0.8935	0.9273
[240°, 270°)	0.9222	0.9255	0.9427	0.9201

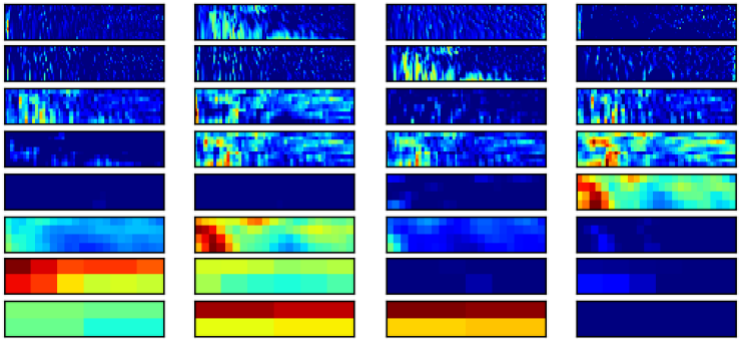
90°

0°

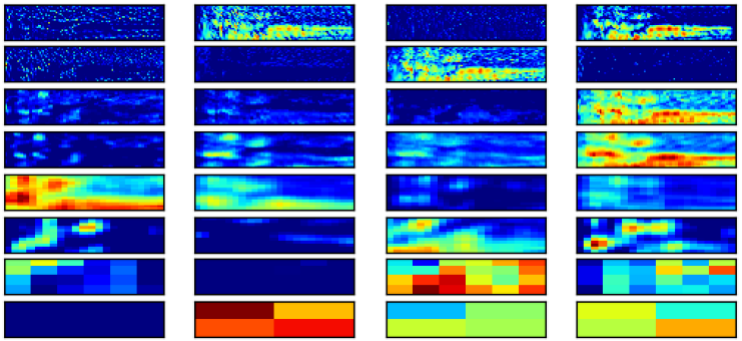
- Discriminatory information exists even in only 10° of data

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

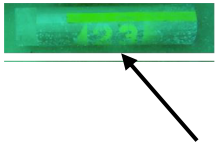
Intermediate Responses



- 3’ cylinder, range 30 m, center aspect 229º
- CNN H, **10º input** (using AC data of 224º -234º)



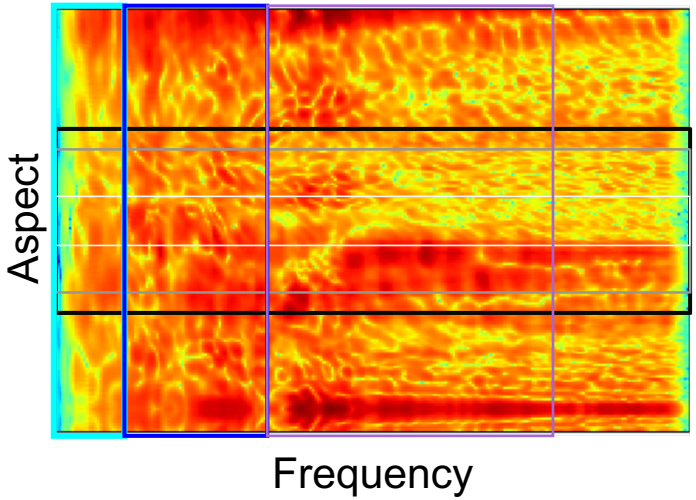
- 3’ cylinder, range 30 m, center aspect 229º
- CNN H, **40º input** (using AC data of 209º -249º)



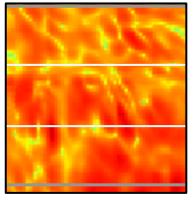
CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Multi-Band, Multi-Aspect

- Rather than using full frequency band to train CNN, use smaller sub-bands
  - Three octaves considered: 3-6 kHz, 6-12 kHz, 12-24 kHz
  - And full band: 3-30 kHz
- Also use sub-apertures
- Forces CNN to find discriminatory clues in small sectors of the acoustic color plot



Input data for 3-6 kHz, 10° CNN



Input data for 6-12 kHz, 30° CNN

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

CNN Design

- Unique CNN (filter sizes and pooling factors) for each sub-band and sub-aperture case
- ~400-1500 parameters in each

CNN Label	Convolutional Blocks	Convolutional Layers Per Block	Column Filter Sizes (pixels) Block 1 Layers Frequency Band (kHz) 3	6	12	Block 2 Layers	Column Pooling Factors
A	2	1	[5]	[8]	[14]	[5]	[f] [5]
B	2	2	[3 3]	[5 4]	[8 7]	[3 3]	[f] [5]
C	2	3	[3 2 2]	[4 3 3]	[6 5 5]	[3 2 2]	[f] [5]

Input: 10°

CNN Label	Row Filter Sizes (Pixels)		Row Pooling Factors	
A	[4]	[4]	[3]	[3]
B	[3 2]	[3 2]	[3]	[3]
C	[2 2 2]	[2 2 2]	[3]	[2]

Input: 30°

CNN Label	Row Filter Sizes (Pixels)		Row Pooling Factors	
A	[8]	[4]	[6]	[6]
B	[5 4]	[4 3]	[6]	[4]
C	[4 3 3]	[4 3 3]	[6]	[2]

Input: 40°

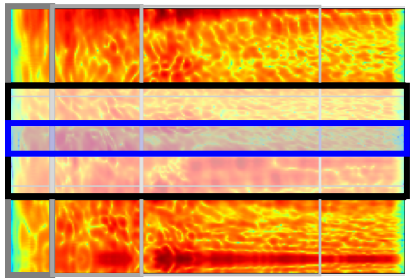
CNN Label	Row Filter Sizes (Pixels)		Row Pooling Factors	
A	[10]	[7]	[6]	[6]
B	[6 5]	[5 5]	[6]	[4]
C	[6 6 6]	[4 4 4]	[6]	[2]

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Results: Sub-bands

Table 5.17. AUC of various cases using CNN A's architecture

θ	ψ	Frequency Band (kHz)				Ensemble
		3-6	6-12	12-24	3-30	
10°	10°	0.7790	0.7970	0.7596	0.8253	0.8596
30°	30°	0.7822	0.7382	0.8365	0.7728	0.8421
40°	40°	0.7877	0.7410	0.7506	0.7807	0.8635
10°	40°	0.8416	0.8429	0.8176	0.8844	0.9033
30°	40°	0.7879	0.7423	0.8434	0.7801	0.8467

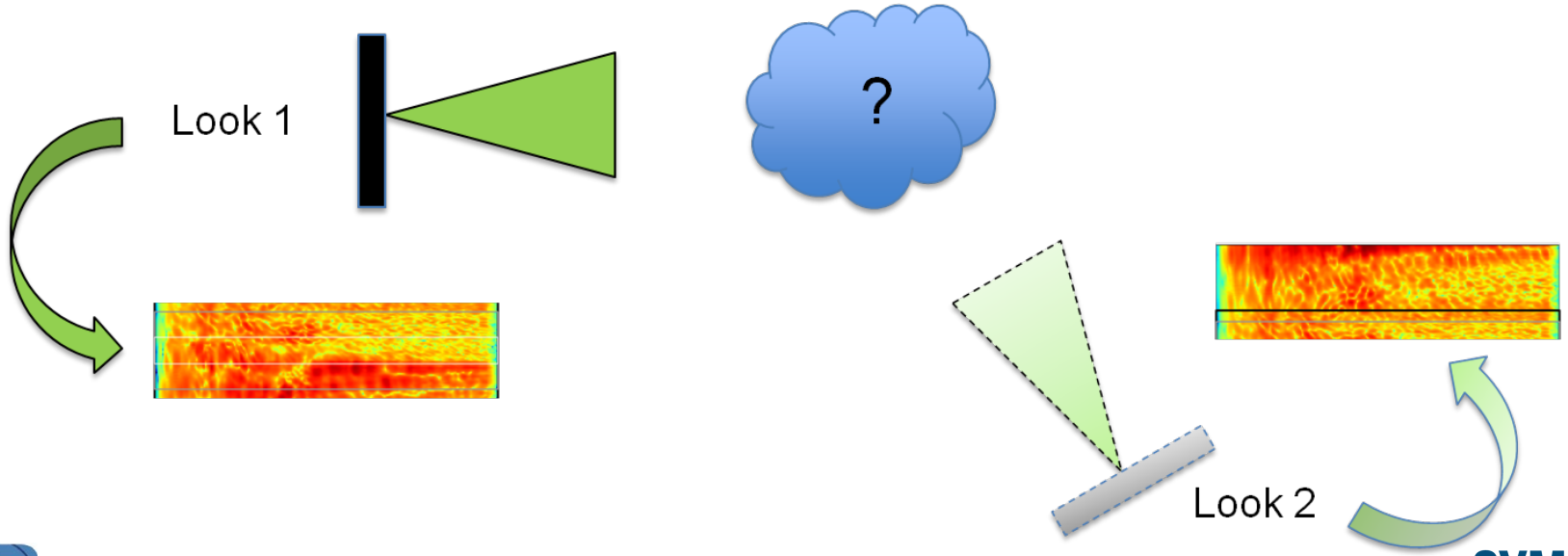


Assume one look provides 40° of data
Θ: CNN input data size
Ψ: Total span of predictions fused

- Considerable discriminatory information exists even in small sectors of the AC plot (e.g., small frequency sub-bands)
- Ensemble of sub-band CNNs outperforms using a single full-band CNN
  - Different clues being leveraged
- Averaging sub-aperture predictions over the span provided by one look outperforms using single full-aperture CNN

Multi-Look Experiments

- Having obtained data from one look of an **unknown object** (and at **unknown orientation**), what is the optimal next look from which to collect data to most improve classification?



CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Results: One Additional Look

- Optimal look separation depends on frequency band and on the span of data provided by one look
  - N.B.: Agnostic to object and **initial-look-aspect**
- Result can inform sensing/re-inspection strategies

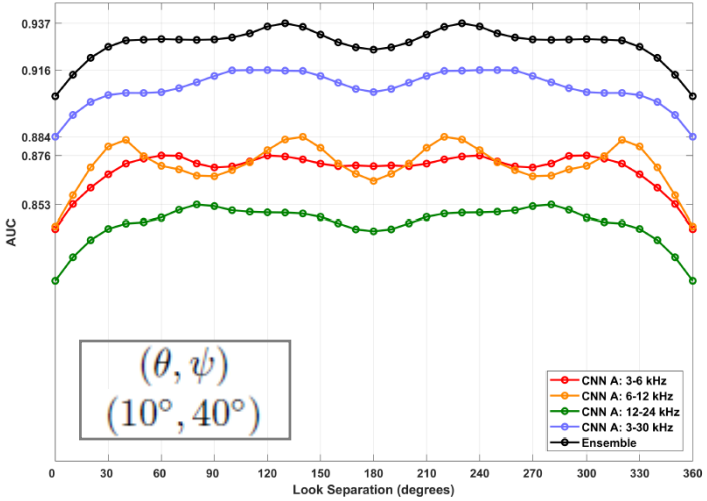


Table 5.18. For CNN A, optimal look separation, from first look, of second look for two-look case

θ	ψ	Frequency Band (kHz) 3-6	6-12	12-24	3-30	Ensemble
10°	10°	120°	140°	80°	150°	140°
30°	30°	80°	120°	140°	140°	120°
40°	40°	130°	130°	40°	100°	120°
10°	40°	60°	140°	80°	120°	130°
30°	40°	80°	120°	140°	150°	120°

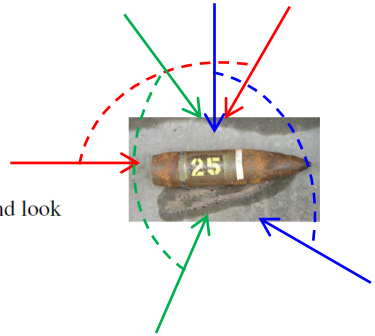
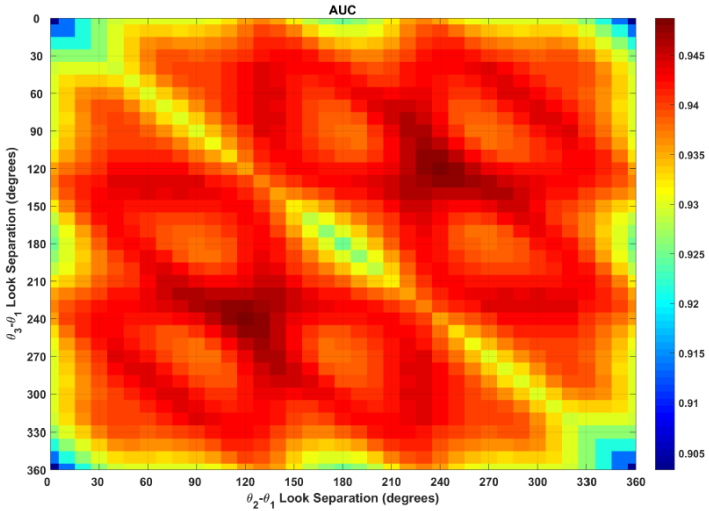


Table 5.19. For CNN A, best two-look AUC using optimal second look

θ	ψ	Frequency Band (kHz) 3-6	6-12	12-24	3-30	Ensemble
10°	10°	0.8385	0.8489	0.8079	0.8733	0.9102
30°	30°	0.8348	0.7720	0.8944	0.8164	0.8827
40°	40°	0.8341	0.7770	0.7954	0.8423	0.9086
10°	40°	0.8757	0.8844	0.8529	0.9155	0.9374
30°	40°	0.8375	0.7760	0.8988	0.8211	0.8853

Results: Two Additional Looks

- Myopic strategy is not always optimal
  - Will select different second look depending on if only one additional look will be obtained, or if two additional looks will be obtained



(θ, ψ)
$(10°, 40°)$

Table 5.20. For CNN A, optimal look separation, from first look, of second and third looks for three-look case

θ	ψ	Frequency Band (kHz) 3-6	6-12	12-24	3-30	Ensemble
10°	10°	50°, 120°	80°, 220°	80°, 160°	80°, 220°	120°, 240°
30°	30°	120°, 240°	120°, 240°	120°, 240°	40°, 80°	120°, 240°
40°	40°	120°, 240°	50°, 130°	40°, 90°	120°, 240°	120°, 240°
10°	40°	60°, 120°	30°, 150°	90°, 220°	110°, 220°	120°, 240°
30°	40°	120°, 240°	120°, 240°	120°, 240°	40°, 80°	120°, 240°

Table 5.21. For CNN A, best three-look AUC using optimal second and third looks

θ	ψ	Frequency Band (kHz) 3-6	6-12	12-24	3-30	Ensemble
10°	10°	0.8633	0.8679	0.8299	0.8950	0.9272
30°	30°	0.8573	0.7932	0.9118	0.8308	0.8955
40°	40°	0.8541	0.7976	0.8138	0.8676	0.9282
10°	40°	0.8896	0.8934	0.8654	0.9271	0.9487
30°	40°	0.8589	0.7965	0.9155	0.8335	0.8971

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Performance with Ensembles: AUC

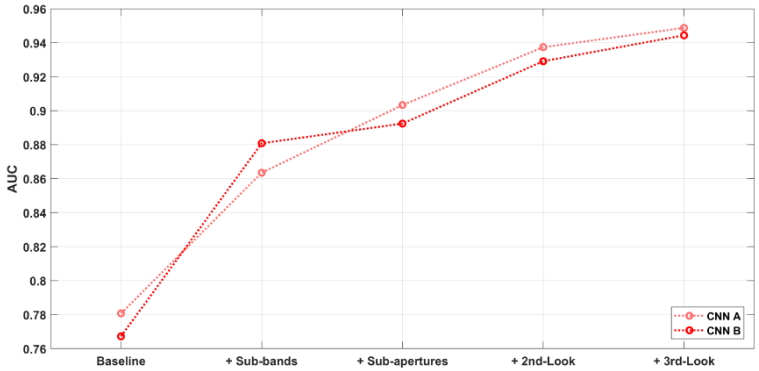
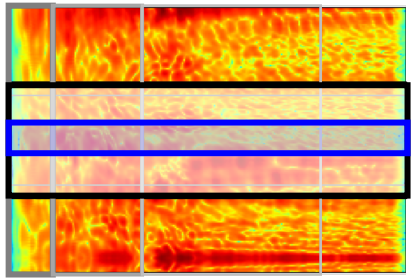
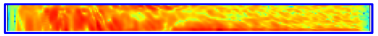
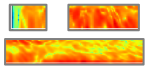


Fig. 5.9. For CNN A and CNN B, the overall AUC as additional representations are included in an ensemble prediction. The additions are cumulative. Refer to the text for details.

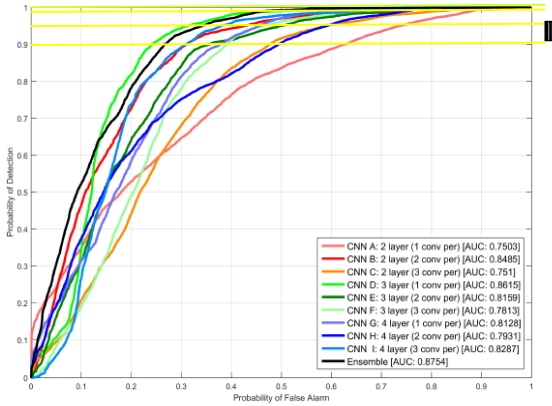


- Baseline: CNN prediction using input data of 10° (all frequencies) 
- + Sub-bands: Average prediction across CNNs for each frequency sub-band 
- + Sub-apertures: Average prediction also across 40° span of data (provided by first look)
- + 2nd-Look: Average prediction also across second set of 40° span
- + 3rd-Look: Average prediction also across third set of 40° span

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Performance with Ensembles: P_{fa} at Key P_d

- AUC is inappropriate metric for actual UXO remediation
- UXO remediation demands very high (at/near unity) probabilities of detection
- Better metric: probability of false alarm at key probabilities of detection



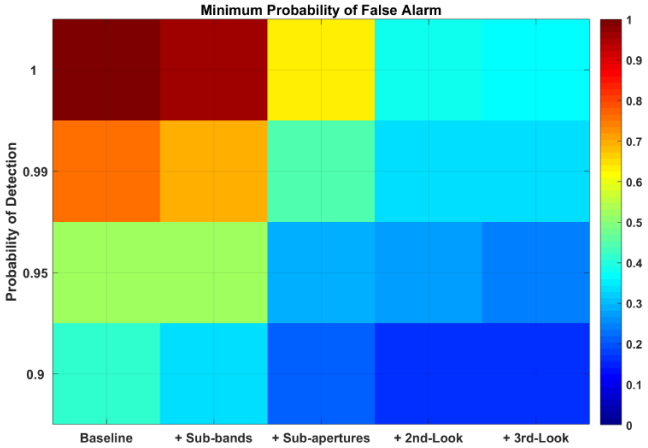
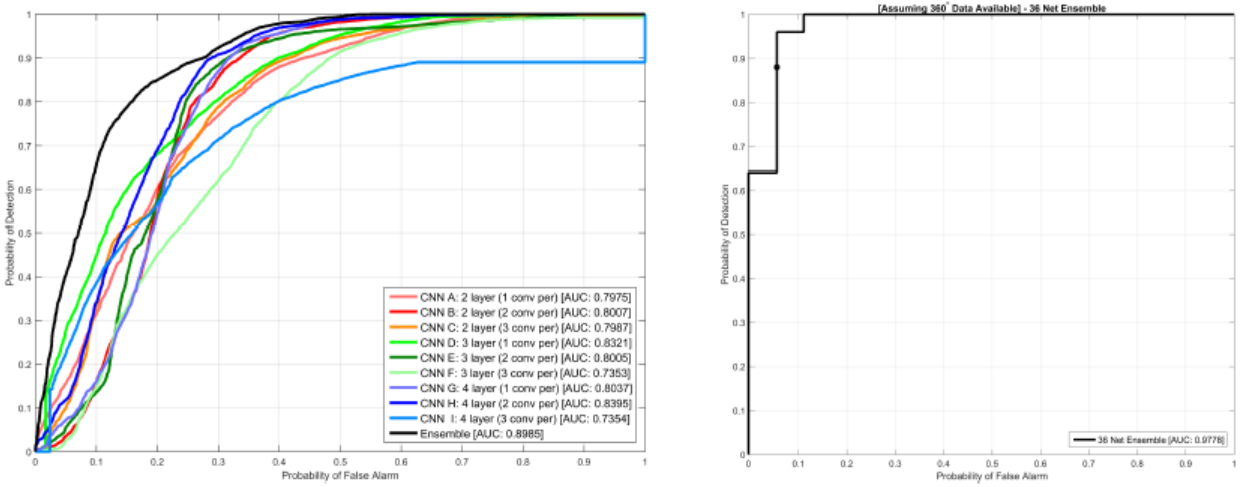


Fig. 5.10. For CNN A, the minimum probability of false alarm that achieves certain key probabilities of detection, as additional representations are included in an ensemble prediction. The additions are cumulative. Refer to the text for details.

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Assuming Full 360° Acoustic-Color Data

- Earlier: making predictions using only a small sector of an acoustic color plot
- UXO remediation less time-sensitive than MCM operations
  - Can execute multiple surveys to obtain full 360° acoustic color plots for each object
- Much easier problem, performance improves significantly

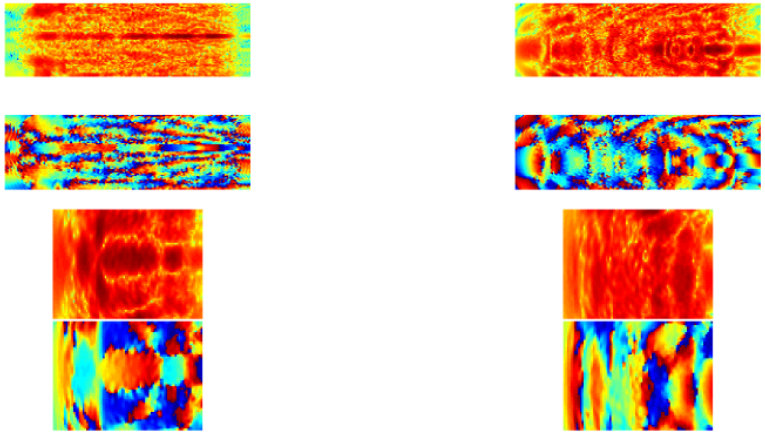


(a) (b)

Figure 3: Classification performance on the TREX13 test data (a) for the 40° input data case using 9 different CNNs, and (b) assuming access to full 360° acoustic-color data and using an ensemble of 36 CNNs. In (b), the operating point on the receiver operating characteristic (ROC) curve corresponding to a 0.5 decision threshold is marked with a dot.

Multi-Representation Acoustic-Color CNNs

- TIERSWAT data
  - Target scattering: APL/UW target-in-the-environment response model
  - Environment scattering: NSWC PC-SWAT



(a) (b)

Figure 4: Input data for the four-representation CNN of a 2:1 cylinder (object 221), at a range of 28 m in environment 10, 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.

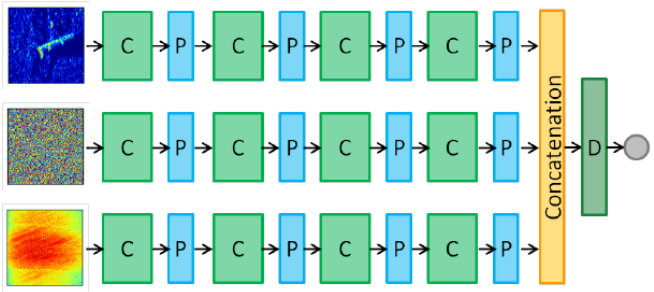


Table 1: Acoustic-color data used in each CNN

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

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Example Result: Multi-Rep. Acoustic-Color CNNs

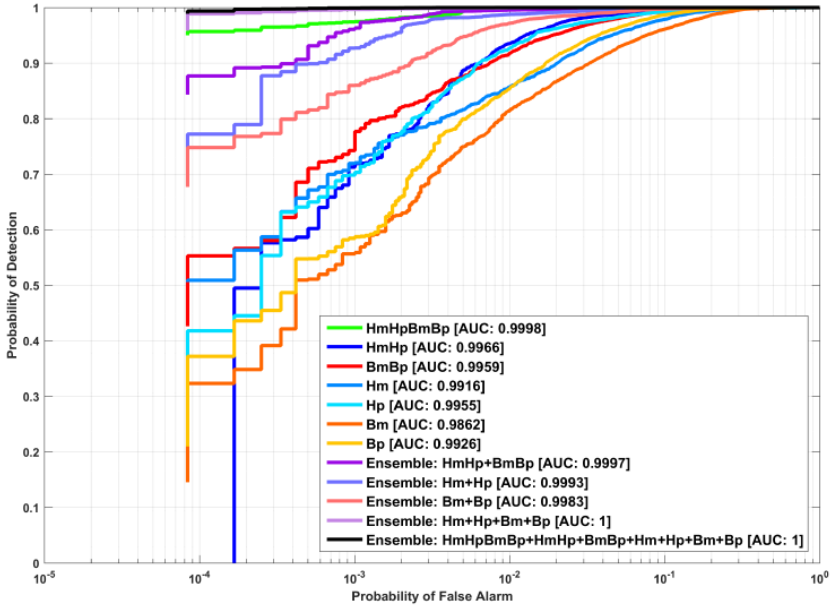


Figure 5: Classification performance of CNNs using different input data representations, when using training data from all environments. (Note the logarithmic horizontal-axis.)

Table 2: AUC for the four-representation HmHpBmBp CNN, but using training data from only certain environments, as a function of test environment

Test Data Environment	Training Data from Environment					
	All	1	5	14	8	17
All	1.000	0.959	0.988	0.984	0.982	0.972
1	1.000	0.993	0.997	0.991	0.984	0.972
3	1.000	0.993	0.994	0.983	0.976	0.966
4	1.000	0.990	0.992	0.981	0.975	0.966
10	1.000	0.980	0.992	0.982	0.978	0.969
5	1.000	0.968	0.994	0.984	0.981	0.971
11	1.000	0.956	0.992	0.990	0.986	0.972
6	1.000	0.952	0.994	0.993	0.989	0.971
14	1.000	0.950	0.993	0.994	0.989	0.973
7	1.000	0.947	0.992	0.994	0.992	0.975
12	1.000	0.947	0.990	0.992	0.993	0.976
8	1.000	0.943	0.988	0.990	0.991	0.974
15	1.000	0.941	0.986	0.991	0.993	0.974
9	1.000	0.940	0.984	0.988	0.991	0.978
17	0.999	0.916	0.962	0.959	0.967	0.982

Intermediate Responses

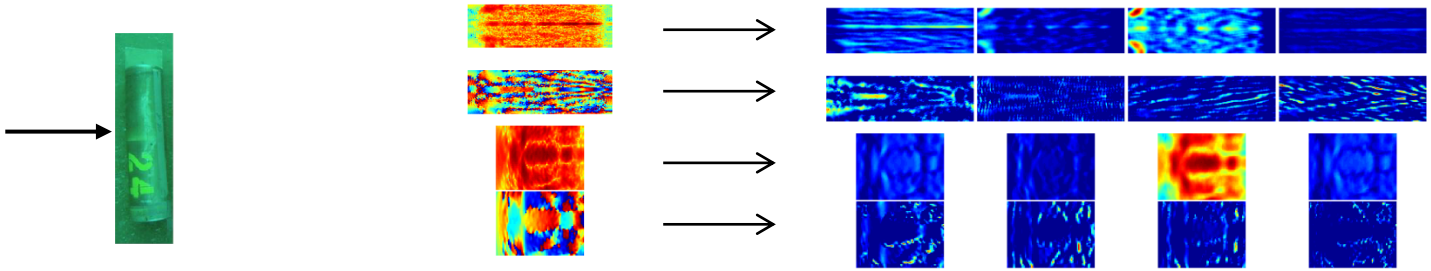


Figure 6: For the input data in Fig. 4(a), a 2:1 cylinder at broadside, the transformed data representations (top to bottom: HF magnitude, HF phase, BB magnitude, BB phase) after the first convolutional layer of the four-representation CNN.

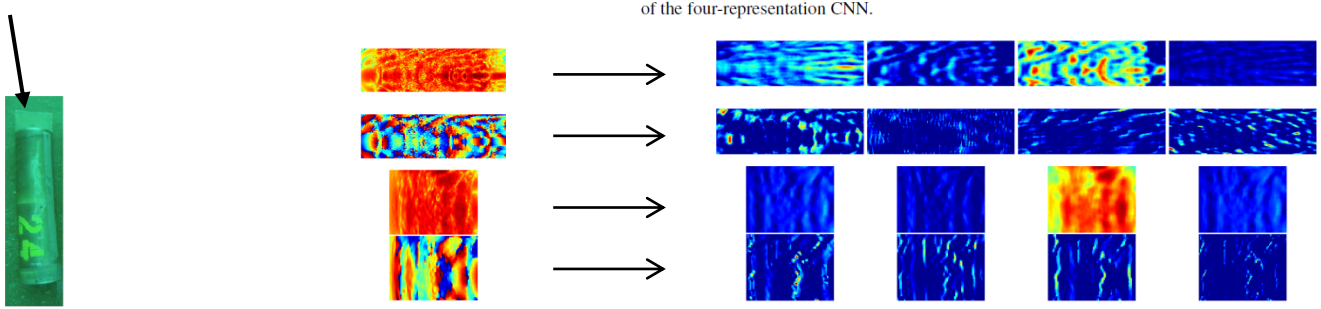


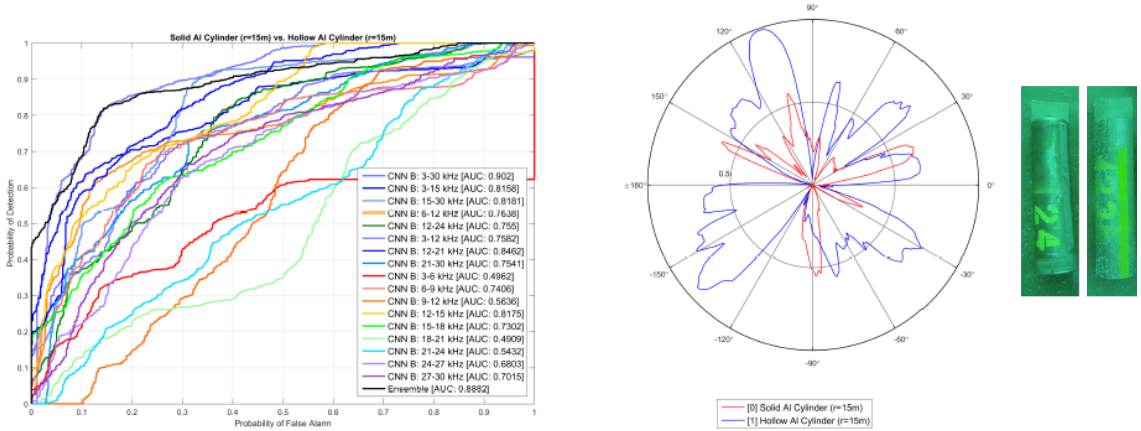
Figure 7: For the input data in Fig. 4(b), a 2:1 cylinder at almost nose-endfire, the transformed data representations (top to bottom: HF magnitude, HF phase, BB magnitude, BB phase) after the first convolutional layer of the four-representation CNN.

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Explainable CNNs

- Using CNNs with acoustic-color data for UXO classification is feasible
- Key remaining unknown: Developing *explainable* CNNs
- Approach: Use developed CNN framework in conjunction with specially controlled experiments to learn principled, explainable features that can be tied directly to the physics involved

Solid Cylinder vs. Hollow Cylinder



(a) (b)

Figure 8: For the binary classification problem attempting to discriminate solid cylinders from hollow cylinders, (a) performance as a function of acoustic-color CNN that uses certain frequency sub-band data, and (b) the predictions as a function of aspect for the 12-15 kHz frequency band case for the two objects at a range of 15 m.

Conclusions

- Using CNNs with acoustic color data is indeed feasible
- Strong classifiers can be obtained from tiny CNNs with only hundreds of parameters to learn
  - Generalize well even to new objects not in training set
  - Still need to test in more challenging scenarios (e.g., new environments)
- Significant performance gains can be achieved by further slicing the acoustic color data into multiple frequency sub-bands and aperture sub-aspects, and from employing multiple looks
- Ensembles from the above, and from employing multiple unique CNN architectures can greatly reduce the false alarm rate at the high P_d values that UXO remediation demands
- Incorporating multiple representations (HF and BB, magnitude and phase of acoustic color) is also promising
- Rich veins for future work:
  - Improving CNN architecture design (filter sizes, pooling factors, etc.)
  - **Tying the uncovered CNN clues back to the physics**
- Acknowledgments:
  - Thanks to Kevin Williams and Steve Kargl at APL/UW for providing the TREX13 and TIERSWAT data
  - This work was funded by SERDP under SEED project MR18-1444

Supplemental Slides

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

CNN Architectures for Acoustic-Color Data

- Consider 9 CNNs (with 2-12 convolutional layers), common high-level architectures
  - 2/3/4 convolutional blocks
  - 1/2/3 convolutional layers per block
- Average pooling (large factors)
- ReLU activation functions
- 4 filters per convolutional layer
- Filter sizes and pooling factors dependent on input data size
- 4 nodes ("features") at final dense layer
- ~600-4000 parameters to learn per CNN
  - Tiny number compared to typical optical image CNNs
  - Eases training data requirements

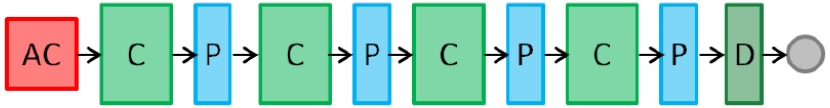


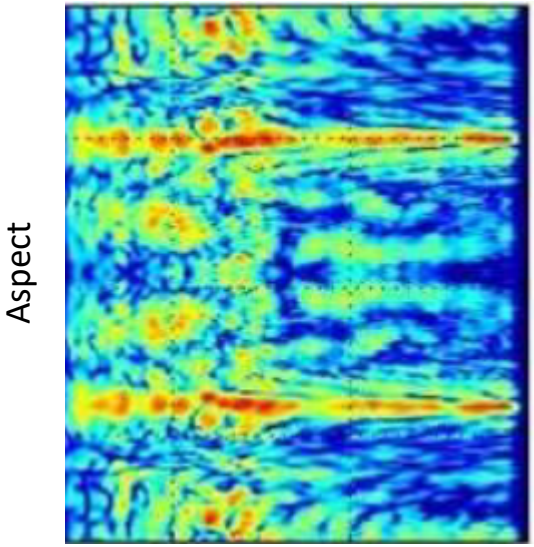
Table 5.3. CNN architectures for 30° input-data (1 pixel represents 0.5° × 100 Hz of data)

CNN Label	Convolutional Blocks	Convolutional Layers Per Block	Filter Sizes in Pixels (Row × Column)	Pooling Factors (Row × Column)
A	2	1	8 × 12 | 4 × 5	[6 × 8] [6 × 8]
B	2	2	5 × 5 4 × 4 | 4 × 6 3 × 5	[6 × 8] [4 × 8]
C	2	3	4 × 6 3 × 6 3 × 6 | 4 × 4 3 × 4 3 × 3	[6 × 8] [2 × 8]
D	3	1	10 × 7 | 9 × 4 | 8 × 6	[2 × 5] [2 × 5] [2 × 5]
E	3	2	6 × 4 5 × 4 | 5 × 5 5 × 5 | 5 × 3 4 × 3	[2 × 5] [2 × 5] [2 × 5]
F	3	3	4 × 5 4 × 5 4 × 4 | 4 × 4 4 × 4 3 × 3 | 3 × 4 3 × 3 2 × 3	[2 × 5] [2 × 4] [4 × 4]
G	4	1	6 × 4 | 5 × 4 | 5 × 5 | 3 × 5	[2 × 4] [2 × 4] [2 × 2] [2 × 2]
H	4	2	4 × 4 3 × 5 | 3 × 5 3 × 5 | 3 × 5 3 × 5 | 2 × 5 2 × 5	[2 × 4] [2 × 2] [2 × 2] [2 × 2]
I	4	3	3 × 5 3 × 5 2 × 4 | 3 × 5 2 × 5 2 × 4 | 3 × 5 2 × 5 2 × 4 | 2 × 3 2 × 3 2 × 3	[2 × 4] [2 × 2] [2 × 2] [1 × 2]

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019

Sonar Data

- Target and Reverberation Experiment 2013 (TREX13)
- Gulf of Mexico, using a rail system
- Frequency band 3-30 kHz
- 27 objects (11 UXO, 16 non-UXO)
- Ranges 10-40 m
- Sandy seafloor
- Resulting acoustic-color (AC) plots
  - 0.5° sampling in aspect
  - 100 Hz sampling in frequency
  - "Images" of size 720 pixels x 271 pixels
  - 90 full (i.e., 360°) AC plots constructed
    - 54 of UXO objects (11 unique objects at various ranges)
    - 36 of non-UXO objects (16 unique objects at various ranges)



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.

CMRE
SERDP • ESTCP SYMPOSIUM
#SerdpEstcp2019