Proceedings of the Institute of Acoustics

UNDERSTANDING MODEL FIDELITY FOR TRAINING SYNTHETIC APERTURE SONAR IMAGE CLASSIFIERS

T.E. Blanford University of New Hampshire, Durham, New Hampshire, USA
D.P. Williams Pennsylvania State University, University Park, Pennsylvania, USA

1 INTRODUCTION

Convolutional neural networks (CNNs) are increasingly employed for classification tasks in automated target recognition (ATR) algorithms for synthetic aperture sonar (SAS) images. Training ATR algorithms, however, requires many unique observations of the targets of interest. Data available for training is often limited, and acquiring additional training data through experimentation can be prohibitively expensive. Recently, researchers have explored several approaches to increasing the quantity of training data. Machine learning approaches, such as generative adversarial networks[1], attempt to generate additional data by interpolating and extrapolating from the features in the existing data. While this can rapidly generate large quantities of data, it will not generate information in the simulated training data that was not already present. Advances in acoustic modeling suggest that it may be possible to augment ATR training with simulated data. Models, however, inherently are approximations of reality that will contain differences when compared to experimental (i.e., measured) data. Training an ATR algorithm with simulated data that is of insufficient fidelity may limit (or possibly worsen) performance when tested on purely experimental data sets.

Previous work in synthetic aperture radar (SAR) has investigated training neural networks exclusively on simulated data to classify objects in experimental data.[2] While synthetically generated images share similarities with experimentally generated images, there are often clear differences in the appearance of targets and backgrounds. This set of differences is called the distribution gap. Presumably if the simulated and experimental imagery were drawn from the same distribution, the distribution gap would be zero and they would be indistinguishable to a network. Many factors impact the appearance of SAS imagery, and the generating distributions of experimental imagery are likely unknowable and cannot be exactly replicated in simulation. Instead, with increasing fidelity, modeling and simulation can aim to approach the experimental distributions to minimize the distribution gap.

All features in the imagery can contribute to this gap. SAS imagery contains significant complexity resulting from many physical mechanisms. Therefore simulations typically use a combination of multiple models. For example, scattering from targets, scattering from the background, and interactions between the target and the environment are described by different models[3,4,5] and are simulated separately and coherently combined. While simulating a complex scene relies upon multiple models, some features may not require accurate modeling for the purposes of training an ATR algorithm. This paper describes the initial findings of a study to investigate the fidelity of models required so that simulated data may be used interchangeably with experimental data for training CNNs. Using in-air experimentation, a high-fidelity data set was developed with multiple degrees of complexity. A high-frequency sonar signal model was then used to generate complementary simulated data. Careful calibration and co-registration of both

sources of data allowed for superposition of targets and backgrounds between the experimental and simulated data sets. Specific physical features in the data could be individually isolated to identify the sensitivity of a CNN to the fidelity of a particular model. This study was conducted with a range of CNN architectures to identify the relationships between model fidelity and network complexity. A central focus of this study was on the relative fidelity between targets and backgrounds.

2 DATA SET GENERATION

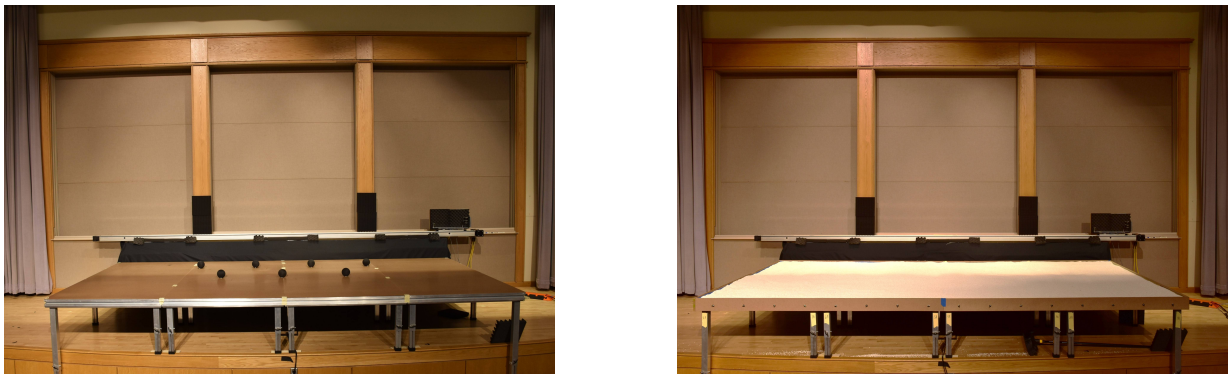
Complementary sets of experimental and simulated data were generated of spherical targets and rough interface backgrounds. In each set, the target scattering and background scattering were produced separately. Precise electroacoustic calibration in the experiment matched the simulated and experimental time series data to a common reference. Due to the tightly controlled motion and calibration, real and simulated data could be interchangeably superimposed as time series or as images. (For example, a composite image of a target on either an experimental or simulated rough interface background could be constructed because all the data is calibrated to the same reference intensity, co-registered, and shares the same spatial sampling pattern.) Each set of data consists of 32 unique scenes.

2.1 Experimental Data

The experimental data set was collected using AirSAS, a synthetic aperture sonar experiment conducted in air.[6] Epistemic sources of uncertainty (uncertainty due to lack of knowledge) that could prevent accurate modeling of the scenes were minimized by tightly controlling and characterizing the experiment. Laboratory experimentation allows for characterization at a level of detail that is challenging or impossible to achieve in a natural field environment. Working in air affords greater experimental control at considerably lower expense than working underwater. While the experimental configuration is similar to underwater SAS systems and scenes, it is not intended to replicate underwater environments in air. Instead, the goal of the in-air experiment was to inexpensively generate a tightly controlled data set that may be accurately modeled and combined with simulated data.

An array consisting of a loudspeaker tweeter with a 1.91 cm (0.75 inch) diameter diaphragm and four microphones was affixed to the moving carriage of a 5 m long linear actuator. The array was advanced over the length of the actuator in 5 mm steps, transmitting a 0.5 ms linear frequency modulated (LFM) downchirp from 30 kHz to 10 kHz at each position. Probes at the edge of the scene measured the air temperature and humidity on a per-ping basis for estimation of the local sound speed. This experimental geometry was designed to produce SAS imagery with approximately 0.9×0.9 cm pixel resolution.[7] Characterization of the noise floor showed in excess of 30 dB of signal to noise ratio (SNR) in the time series data. The background noise is dominated by electrical and ambient environmental noise and is incoherent between pings. The synthetic aperture array gain further increases the SNR upon image reconstruction.

Scenes were established in front of the array and centered along the length of the actuator. Target scenes consisted of seven solid spheres on a flat planar interface of tempered hardboard. The spheres were machined from solid polyurethane and were 10.2 cm (4 inch) in diameter. The seven spheres were arranged in a row of four and a row of three, with 0.75 m between the nominal location of each sphere. After the target collection was complete, the spheres were removed and the hardboard was covered with an approximately 5.1 cm (2 inch) layer of high-density polyetheylene (HDPE) pellets which formed a rough interface. The individual HDPE pellets were irregular in shape and had a maximum dimension of approximately 3 mm ($1/6$ of a wavelength at 20 kHz) which ensured that incident sound waves would be diffusely scattered. Fig. 1 shows photographs of the experimental set-up of the targets and the rough interface.



(a) (b)

Figure 1: Experimental SAS data was collected on spherical targets (a) and a rough interface background (b) separately so that the two could be superimposed in a manner like many simulations. The experiment was conducted in air because it is easier than working underwater and affords tight control, perfect ground truth, and excellent repeatability.

The scenes were randomly perturbed between scans to ensure consecutive collections would not result in near-identical sets of data. The spheres were manually picked up and replaced within markings that defined the target positions, and the surface of the rough interface was swept with a push broom. Sweeping was conducted with strokes in the cross-track direction to prevent the formation of ripples that which would be observable in the imagery.

2.2 Simulated Data

The simulated data set was generated using the Point-based Sonar Signal Model (PoSSM).[3] The model defines a sonar system with projectors and receivers and simulates the bistatic scattering of signals between the projector, a scattering object, and a receiver. It captures both the time-delay associated with propagation and the combined losses from spherical spreading, projector directivity, receiver directivity, and acoustic attenuation. The directivities of the projector and receivers were described by measurements of the beam patterns of the speaker and microphone, respectively, at 10 kHz. The transducer array was moved relative to the targets and backgrounds in the same manner as the experimental configuration. Experimental variability in the ping-to-ping position and sound speed was modeled by statistically characterizing the experimental errors and drawing random realizations from these distributions.

Scattering from the spherical targets was simulated by placing points with unit scattering amplitude at the nominal locations of the centers of the spheres, and using PoSSM to compute the propagation effects of scattering from these points. After execution of PoSSM, the resulting signals were convolved with the far-field scattering function for a rigid sphere[8],

$$f(\theta) = \frac{i}{k} \sum_{n=0}^{\infty} (2n+1) \frac{j_n'(ka)}{h_n'(ka)} P_n(\cos\theta), \tag{1}$$

where i is the imaginary number, k is the wavenumber, a is the sphere radius, j_n' is the first derivative of the spherical Bessel function with respect to its argument, h_n' is the first derivative of the spherical Hankel function with respect to its argument, and P_n is a Legendre polynomial of order n. θ is the angle between the incident and scattered wave and for backscatter is equal to 180°. The method of images was used to

simulate the target-local multipath effects that result from targets lying proud on an interface.[5] Only the first order multipath components (a single interaction with the interface) were included in the simulation.

The background was simulated by discretizing the rough interface into a set of points. Each point was assigned a random position and a scattering amplitude. The scattering amplitude is comprised of a deterministic term based on the differential cross section per unit area per unit solid angle of the rough interface, and a stochastic term that randomizes the phase. The deterministic scattering cross section was described by Lambert's law[9] using a value for Lambert's parameter, $\mu = 0.033$, that was found to match the experimental data. The stochastic term was drawn from a complex circular Gaussian distribution because the scattering from neighboring points was assumed to be incoherent. The background was simulated with multiple densities of points describing the interface.

The scenes were randomly redrawn between consecutive simulations to build an ensemble of data in a similar manner to the experiment. The x and y positions of the target centers were randomly redrawn from a uniform distribution whose endpoints matched the bounds of the experimental target positions. The rough interface was redrawn with new random point positions and stochastic terms. The simulated time series was scaled using the electroacoustic calibration factors of the experiment's transducers. Background noise was not added to the simulation due to the high level of SNR in the experimental data.

3 NETWORK ARCHITECTURE AND TRAINING

The time series data for each scene was reconstructed into imagery using delay and sum beamforming. A 350 pixel $\times$ 350 pixel complex-valued chip, γ, was extracted from the scene and transformed as $\eta = 20\log|\gamma|$. The chip was then normalized as $\zeta = (\eta - \max\eta + 20)/20$. To effectively retain 40 dB of dynamic range, any pixel values of ζ less than -1 were set to -1. The resulting chip was randomly cropped to a 300 pixel $\times$ 300 pixel image before being used as input to a CNN. (A new random crop was applied every time a chip was selected for a training batch.) The normalization eliminates differences in the two data sources due to absolute calibration errors. The dynamic range threshold was applied because experimental data will have noise and background environmental scattering that is not present in simulation. As this difference is not relevant to the study, thresholding was intended to discourage the networks from classifying based solely on low-level background features.

A set of tiny CNNs[10] of similar construction but varying complexity was developed to investigate how network architecture may be sensitive to differences between experimental and simulated data. The CNN architectures began with a pre-pooling layer, which downsamples the input image by a given factor by averaging. Then a series of alternating convolutional blocks (containing one or more convolutional layers) and pooling layers followed, after which there was a single dense layer. (Only networks A and B have more than one convolutional layer per convolutional block; the number of convolutional layers in the other CNNs is equal to the number of convolutional blocks.) There were f filters per convolutional layer, and the final pooling layer ensured that the dense layer also consists of only f nodes. Table 1 summarizes the architecture of each of the CNNs. The data was labeled as belonging to one of two classes based on provenance — experiment or simulation — so the classification task was to predict the provenance of each chip. The CNNs were trained using one pool of chips, and then inference was performed on a disjoint pool of test chips.

Table 1: CNN architecture parameters.

Name	Pre-pooling Factor	Convolutional Blocks	Filters per Layer	Free Parameters
A	1	8	8	7953
B	1	8	4	2169
C	1	5	8	2553
D	1	5	4	701
E	1	4	8	11057
F	1	4	8	2937
G	2	4	8	6193
H	2	4	4	1657
I	4	3	8	4449
J	4	3	4	1249
K	6	3	8	3305
L	6	3	4	997
M	10	2	8	1569
N	10	2	4	641

4 RESULTS

A set of classification experiments was conducted to investigate the relationships between target and background model fidelity through the lens of a CNN. In each experiment, all fourteen networks were used to classify the provenance of image chips as simulated or experimental. The results from the network produced a receiver operating characteristic (ROC) curve. The area under the curve (AUC) provides a summary measure of classification performance. An AUC value of 1 corresponds to a perfect classifier and indicates that the network can easily discriminate the two sources of data, while an AUC value of 0 is the reciprocal of a perfect classifier. An AUC value of 0.5 corresponds to a random classifier and indicates that the two sources of data appear identical to the network. Thus the AUC value is effectively related to the *uncertainty* of the classifier in predicting the provenance of each image chip. As a result, a metric of apparent model fidelity, α, can be defined as the entropy of a Bernoulli process with probability $p =$ AUC,

$$\alpha = -p \log_2 p - (1-p) \log_2 (1-p). \tag{2}$$

This metric is bounded on the interval $\alpha \in [0, 1]$ where $\alpha = 0$ corresponds to a model that the CNN can perfectly distinguish from experiment and $\alpha = 1$ corresponds to a model that is indistinguishable from experiment to the CNN.

4.1 Experiment: Target Model Fidelity

The first experiment investigated how the apparent fidelity of a target model depends on the presence of a background in the image. First, the networks classified chips of spherical targets on a flat planar interface. The image consists of target scattering (and the target local multipath) alone because the backscatter from the background is negligible. Next, both the real and simulated targets were superimposed on measurements of the rough interface. This composite data was “re-chipped” and again classified in terms of the target data’s provenance. Adding the measured background in both sets of data is representative of having a perfect background model in the simulation. (Note that this composite image still lacks features such as shadows that would exist in an experimentally generated image of spheres on a rough interface.

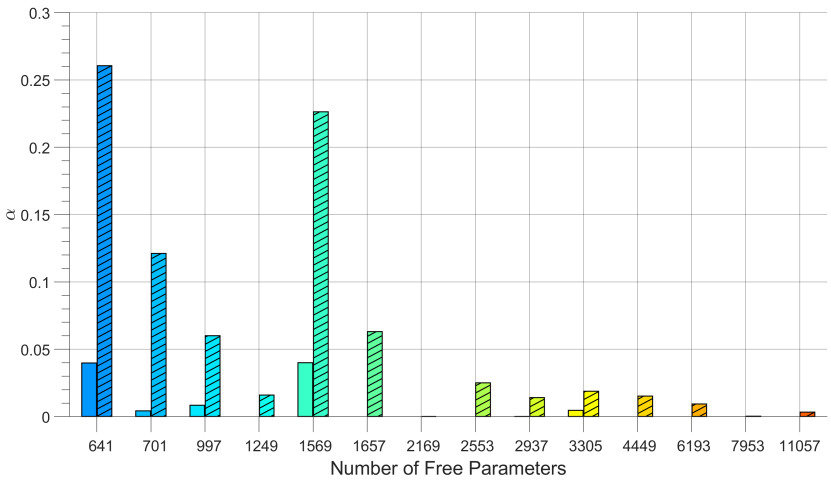


Figure 2: CNNs were used to discriminate experimental and simulated images of a spherical target on an interface. The apparent model fidelity α values are plotted as bars as a function of the number of free parameters in the network. The targets were simulated resting proud on a flat planar interface (solid bars) and superimposed on a measured rough interface (hatched bars). The presence of a background in the image increases the α value for most networks.

Both sources of imagery lack this feature, however, which prevents it from being a discriminating feature in the data.)

Fig. 2 shows the α values for each network in terms of classifying images as experimental or simulated targets. The solid bars correspond to the targets on a flat interface (no background) and the hatched bars correspond to the targets with a measured rough interface background. The networks are very good or perfect classifiers for the target-only images so the α scores are low. This indicates that the target model is not of especially high fidelity for the CNNs, and there likely are one or more features in the appearance of the target that the network can leverage. There are several possibilities for this discrepancy. One is that the polyurethane spheres are not adequately described by the analytic model in (1). Elastic mechanisms and surface roughness are not captured in the analytic model but may be non-negligible. Another source of discrepancy could be the narrowband directivity pattern used in the simulation. The actual directivity of the tweeter varies substantially over the 10-30 kHz band, and this could impact the appearance of the spheres in the imagery.

4.2 Experiment: Background Model Fidelity

A second set of experiments investigated how the apparent fidelity of a background model depends on the presence of a target in the image. Five models for the background were used to simulate the rough interface scattering. The models differed only in the number of points used to describe the interface. The linear density of points ranged from 32 to 512 points per meter (0.55-8.78 points per wavelength at the center frequency). The pixel statistics of the rough interface scattering converge as the number of points increases. A density of at least eight scattering elements per wavelength is often recommended for accurate modeling.[11]

The networks were first used to classify chips of only the background (without a target). These results are shown in Fig. 3a. The bars are arranged from left to right in increasing order of the number of free

parameters in each network. The color of each bar matches Fig. 2. As the density of point scatterers increases, the α values for each network tend to decrease. This indicates that the apparent model fidelity is increasing with higher densities of points. Most α values, however, are less than 0.1 which indicates that there are still clear differences between the real and simulated backgrounds. Next, measurements of spherical targets on a flat planar interface were superimposed on the experimental and simulated backgrounds. This is representative of using a perfect model of the target scattering in the simulation. These α values are shown in Fig. 3b. The presence of the target in the imagery reduces the performance of the networks in classifying the provenance of the background. While the background models did not change, the increase in α values suggests that the model fidelity appears higher for these networks when a target is in the chip.

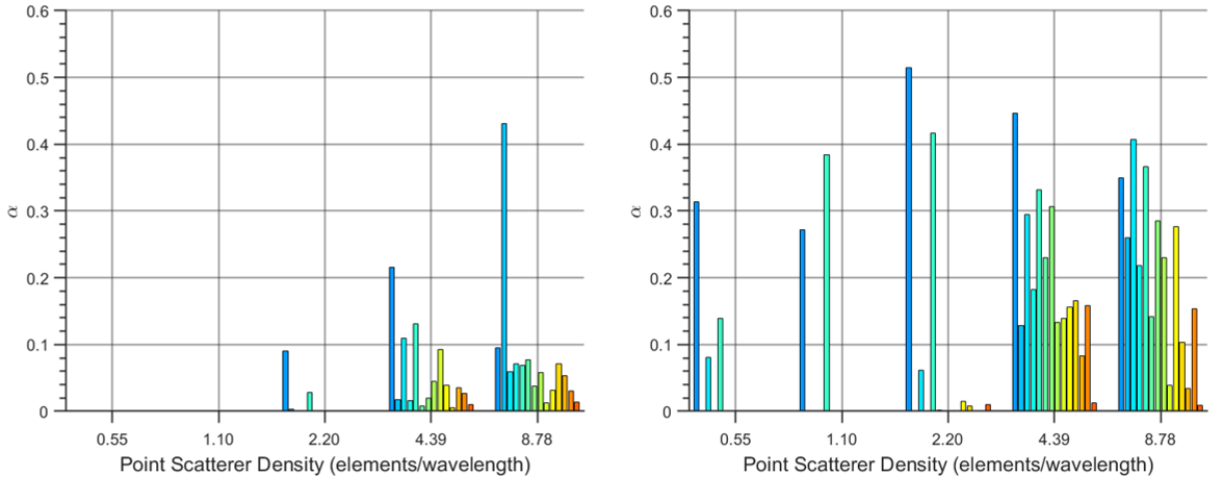


Figure 3: CNNs were used to discriminate experimental and simulated images of rough interface backgrounds. The α values from each classifier are plotted as bars, where the grouping corresponds to simulated data of a particular point scatterer density. The color of each bar matches Fig. 2 and corresponds to a particular network, ordered from left to right in increasing number of free parameters. In (a), the image chips consist solely of a rough interface. In (b), experimental measurements of a spherical target on a flat planar interface are added to the data. Introducing the target increases the apparent fidelity of the background model.

5 CONCLUSIONS

The increasing capability in modeling both target and environment scattering suggests that it may be possible to train ATR algorithms on simulated data. Models are imperfect descriptions of reality, and there will always be differences between experimental and simulated data. SAS imagery contains multiple features from targets, the background, and their interaction, and it is possible that some features require more accurate modeling than others for the purposes of training an ATR algorithm.

This study investigated the performance of CNNs in classifying the provenance of SAS imagery as experimental or simulated. Complementary sets of experimental and simulated time series data were generated using spherical targets and rough interface backgrounds. The two sets of data were reconstructed into imagery using identical algorithms, and CNNs of varying complexity were used to distinguish simulated data from experimental data. A central finding of this initial study is that a CNN's success in discriminat-

ing experimental and simulated data depends on the complexity of the image. A baseline classification performance for each network was established on experimental and simulated image chips consisting of only a rough interface or only a target. The apparent fidelity of the model increased, however, when experimental measurements of targets and backgrounds were superimposed on the baseline data. This superposition increases the complexity of the image, and from the perspective of the network, increases the apparent fidelity of the model. As the fidelity of the model did not change from a physical perspective, this suggests that lower-fidelity models may be suitable in simulations of scenes with higher complexity.

This initial study focused on only two aspects of a SAS image: targets and backgrounds. Future work will extend this study to explore how these trends change with more complicated targets, additional features of the background such as texture, and interactions such as shadowing. By quantifying differences in the distributions of components of the simulated and experimental data, it may be possible to develop predictive metrics for model fidelity. It is important to note that while the indistinguishability of simulated and experimental data may be a sufficient condition for training with simulated data, it is potentially not a necessary condition. Future work will also investigate how binary classification performance depends on model fidelity when networks are trained entirely on simulated data but tested on experimental data.

ACKNOWLEDGMENTS

The research presented in this work was funded by Office of Naval Research (ONR) Grants N00014-22-1–2607 and N00014-23-1–2846.

REFERENCES

1. A. Reed, et al., “Coupling rendering and generative adversarial networks for artificial SAS image generation,” in *OCEANS 2019 MTS/IEEE SEATTLE*, 2019, pp. 1–10.
2. N. Inkawhich, et al., “Bridging a gap in SAR-ATR: Training on fully synthetic and testing on measured data,” *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2942–2955, 2021.
3. D. C. Brown, S. F. Johnson, and D. R. Olson, “A point-based scattering model for the incoherent component of the scattered field,” *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. EL210–EL215, 2017.
4. A. T. Abawi, “Kirchhoff scattering from non-penetrable targets modeled as an assembly of triangular facets,” *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 1878–1886, 2016.
5. S. Kargl, et al., “Scattering from objects at a water–sediment interface: Experiment, high-speed and high-fidelity models, and physical insight,” *IEEE J. of Oceanic Engineering*, vol. 40, no. 3, pp. 632–642, 2015.
6. T. E. Blanford, L. Garrett, J. D. Park, and D. C. Brown, “Leveraging audio hardware for underwater acoustics experiments,” *Proceedings of Meetings on Acoustics*, vol. 46, no. 1, p. 030002, 11 2022.
7. D. W. Hawkins, “Synthetic aperture imaging algorithms: with application to wide bandwidth sonar,” Ph.D. dissertation, University of Canterbury, 1996.
8. E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.
9. D. R. Jackson and M. D. Richardson, *High Frequency Seafloor Acoustics*. Springer, 2007.
10. D. P. Williams, “On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery,” *IEEE J. of Oceanic Engineering*, vol. 46, no. 1, pp. 236–260, 2021.
11. E. Pouliquen, O. Bergem, and N. G. Pace, “Time-evolution modeling of seafloor scatter. I. Concept,” *J. Acoust. Soc. Am.*, vol. 105, no. 6, pp. 3136–3141, 1999.