Underwater Acoustic Detection and Localization with a Convolutional Denoising Autoencoder

Alberto Testolin and Roee Diamant, *Senior Member, IEEE*

***Abstract*—Detecting and tracking moving targets is a challenging task, which becomes even harder in underwater scenarios due to the extremely low levels of signal-to-noise ratio associated with common acoustic measures. In the context of continuous marine monitoring, a further challenge is provided by the need to deploy computationally efficient methods that guarantee minimum use of power resources in off-shore monitoring platforms. Here we present a novel approach to accurately detect and track moving targets from the reflections of an active acoustic emitter. Our system is based on a computationally- and energy-efficient deep convolutional denoising autoencoder. System performance is evaluated both on simulated and emulated data, and benchmarked against a probabilistic tracking method based on the Viterbi algorithm.**

***Index Terms*—Marine monitoring; Underwater acoustics; Signal detection; Underwater tracking; Denoising autoencoders; Convolutional Neural Networks; Viterbi algorithm.**

I. Introduction

THE importance of continuous long-term biomass monitoring of mobile large bio-fuana like fish, sea turtles, or marine predators is a game-changer in understanding the healthiness and balance of the marine ecosystem [1]. This requires an effective statistical tool for fish aggregation. Surveying natural marine bio-fauna populations have traditionally focused on fish capture; however, modern acoustic methods allow for a more effective detection and quantification of fish abundance and distribution.

Hydroacoustic methods are particularly efficient in time and costs compared to alternative survey methods, and have also important advantages including the capability of scanning the entire water column, coverage of large ranges of size, and non-invasiveness [2]. These features make them a vital component of many resource assessments, which is required for the effective management of fisheries and marine ecosystems. For example, fish biomass and size spectra can be quantified using directional acoustic methods (echo sounders), which may reflect also perturbation of the entire ecosystem [3]. Another well investigated application for active acoustics is the detection of threats to marine infrastructures such as intruding scuba divers [4]. For both applications, omni-directional acoustics that transmits and receives over a wideband frequency band, have been proved a key tool for increasing detection range. Nevertheless, progress on detection of submerged mobile targets through active hydroacoustics in a high clutter environment and in real-time is still a major challenge.

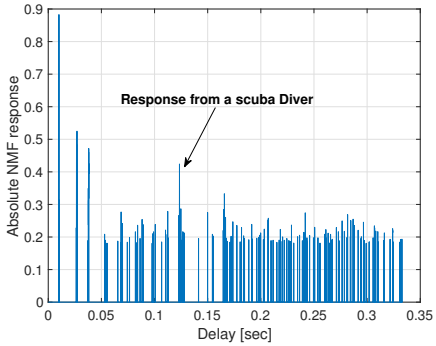

Fig. 1. A single reflection pattern from a scuba diver. The target is almost invisible within the clutter.

Detecting of submerged targets through active acoustics involves detecting the specific target's-based reflection within the reflected signal, which includes stationary reflections from e.g., rocks or chains, as well as reflections from waves or volume scatters, referred to as *clutter noise*. While stationary reflections are relatively easy to identify due to their high reflection power and steady pattern over time, due to the low target strength of marine animals, reflections from the latter are more similar in power to clutter noise. An example for such a reflection from a scuba diver is shown in Fig. 1. Some approaches for active acoustic detection involve an array of receivers to compensate for the high clutter (e.g., [5]). However, this greatly limits the system's setup, since arrays have to be stationary in order to allow directionality. Instead, we focus on the practical setup of a single transceiver that can be deployed from small vessels or even from a kayak. Further, since real-time detection is needed for both online monitoring of marine animals and detection of threats, the computational complexity of the detection system must be limited.

In this paper, we offer a novel machine-learning approach to identify patterns in real-time within a time-delay (TD) matrix formed by concatenating matched filter's outputs sequentially. As presented in the example in Fig. 2 for reflections from a school of Tuna fish, we view these patterns as curved lines in the matrix represented by an image. We identify these lines using a convolutional denoising autoencoder (CDA) [6], whose objective is to produce as output a denoised image containing the target path cleaned from background and clutter noise. To the best of our knowledge, the proposed approach constitutes the first attempt to apply deep learning for identifying targets

A. Testolin is with the Department of Information Engineering and the Department of General Psychology, University of Padova, Italy. Email: alberto.testolin@unipd.it; R. Diamant is with the Department of Marine Technologies, University of Haifa, Israel. Email: roee.d@univ.haifa.ac.il

This work was funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 773753 (SYMBIOSIS). A.T. gratefully acknowledges the support of the NVIDIA Corporation for the donation of a Titan Xp GPU used for this research.

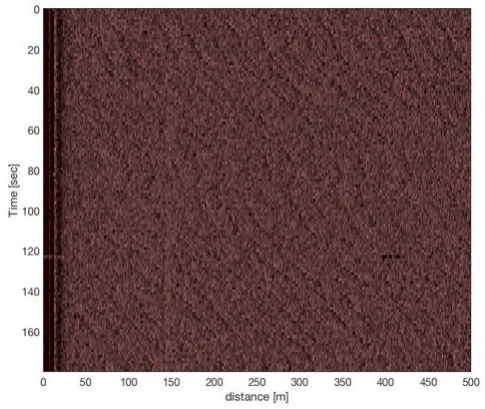

Fig. 2. A TD matrix for reflections from a school of Tuna fish. Each curved line is a reflection from a single fish.

within a reflected acoustic signal. Our results show that even in low signal-to-clutter ratio (SCR), where the reflection pattern from the target is weak, our method yields a favorable trade-off between precision and recall, which exceeds the performance of probabilistic approaches (i.e., the Viterbi algorithm) at a much lower computational complexity, and also allows for a more accurate fine-grained tracking of the target path. Further, since detection and tracking are made per sample within the TD matrix, our approach is scalable with the number of identified targets. Our contribution therefore holds the potential to detect submerged targets using a single transceiver over real-time systems.

The remainder of this paper is organized as follows. In Section II, we describe the state-of-the-art in underwater acoustic target tracking. Our system's model is described in Section III. In Section IV, we provide the details of the convolutional denoising autoencoder. Performance is analyzed in Section V, and conclusions are drawn in Section VI.

II. Related Work

To identify a target within a clutter, the common approach is to emit wideband signals and accumulate matched filter (MF) responses of the received reflections [**?**]. The high processing gain of the wideband signals highlight reflections from the ambient noise, allowing to concentrate on identifying the target within clutter reflections [7]. The reverberation patterns are accumulated in a TD matrix, and tracking is employed to identify the target as a pattern within the clutter [8]. Yet, tracking requires prior knowledge regarding the motion pattern of the tracked target, which may not be available.

An alternative approach is probabilistic detection in the framework of a track-before-detect (TBD) approach. A maximum likelihood TBD was offered in [9] and [10], where the probability of the target's reflections is evaluated and tracking is performed by data association. Dynamic programming is an alternative to evaluate the track of the target by considering the matched filter's output as metrics of likelihoods. Such is the use of the Viterbi algorithm in our recent publication [11] or the use of hidden Markov models for tracking [12]. Yet, while being robust to target types, the complexity of dynamic programming does not allow real-time analysis due to the high number of samples at the matched filter's output. Multi-hypothesis tracking that parameterize the probability density function of the target's reflection is an alternative aimed to reduce the complexity of TBD (see [7] for a survey of such methods). Among these is the method in [13] that combats time-variations in the reflection patterns, and the method in [14] that computes the multi-hypothesis probabilities by an histogram. Yet, for low SCR (as in the case of tracking marine animals or scuba divers with neoprene cover of their tank) results are often poor. Further, the complexity of the solution increases with the number of targets.

A promising framework to overcome the limitations of traditional methods is given by machine learning, which allows to perform pattern recognition efficiently without the need for domain-specific knowledge about the signal characteristics. In particular, deep learning [15] represents the state-of-the-art in most challenging pattern recognition problems such as image classification [16] and speech recognition [17]. The advantage of deep learning over other popular machine learning methods is that the data is encoded using multiple levels of representations, thus allowing for an effective extraction of the relevant features directly from the data. Moreover, once trained deep neural networks are computationally very efficient, since signal processing can be carried out in parallel hardware using basic algebraic operations [18], [19]. Deep learning has been successfully applied also in telecommunication settings [20], [21] and for signal detection under very noisy conditions [22], making it a promising candidate for our challenging underwater scenario.

III. System Model

A. Assumptions and Goals

Our system includes a single transceiver that can be deployed from a small surface vessel or a buoy. For simplicity we assume the transceiver is stationary, although extension to a mobile scenario is straightforward if the motion of the deploying vessel can be measured. The signals transmitted are short and have a narrow auto-correlation response to obtain a high processing gain. Example is the chirp signal, whose cross-correlation is also tolerant to Doppler shift. The signals are emitted periodically. For each emission, the transceiver records the reflecting signal from the channel. To cover large area, we consider an omni-directional transceiver.

For each emission i, the received signal, $r(i)$, is matched with the transmitted signal, s, using the normalized matched filter (NMF),

$$\mathrm{NMF}(i) = \frac{sr(i)^{\mathrm{T}}}{\sqrt{ss^{\mathrm{T}} \cdot r(i)r(i)^{\mathrm{T}}}}, \tag{1}$$

which is used to suppress noise variations within signal s. We filter the output of the NMF to leave non-zero elements in $\mathrm{NMF}(i)$ for only those samples that passes a threshold set by the analysis in [23]. The filtered NMF outputs are then combined to form a TD matrix, M, whose rows are the

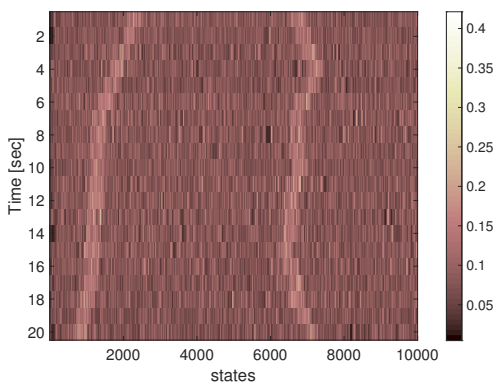

Fig. 3. An emulated TD matrix. Target reflections are marked by the thick curved line. SCR is 5 dB.

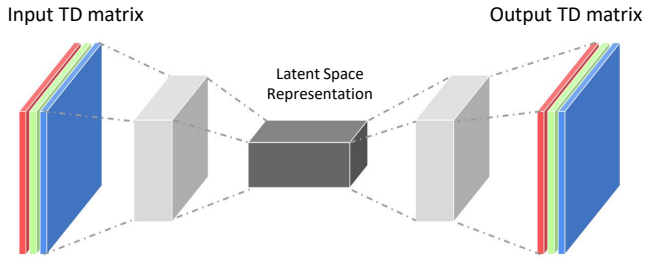

Fig. 4. Graphical representation of a Convolutional Denoising Autoencoder. The noisy TD matrix is given as input, and processed by a hierarchy of convolutional layers that detect increasingly more complex features in the signal in order to build a compressed, latent space representation of the image. The representation is then projected back into the image space by a series of deconvolutional layers, in order to produce a denoised TD matrix.

emission index, and whose columns are the NMF samples. Since, for wideband signals,

$$s * r(i) \approx s * s * h(i) \approx \mathcal{I} * h(i) = h(i), \qquad (2)$$

where $\mathcal{I}$ is the Kronecker delta function, and $h(i)$ is the reverberation channel impulse response for the ith signal, we consider the TD matrix M as a representation of the arrival times and power of the reflecting signal.

Two main tasks are defined:

- *Detection*, which operates image-wise and requires to identify the presence of a mobile target within the reflected pattern of the whole TD matrix. This is the first step within the processing chain to characterize the mobile target, and it is evaluated by measuring the receiver operating characteristics (ROC) curve.
- *Localization*, which operates pixel-wise and requires to accurately identify the target movement by tracking the curved lines in the TD matrix. This is the second step in the processing chain, and it is evaluated by measuring the summed (Euclidean) distances between the real and predicted target position at each time step.

B. Data Sets

To train and validate our system we consider both a *simulated* database and an *emulated* one. The simulated set is totally synthetic. It includes TD matrices of clutter generated from both Gaussian distribution and Beta distribution to represent reverberation noise, while the target is a fixed width line of samples forming a Markov chain. The line's location in the first row is randomly uniformly placed. Then, the following locations within each row are arranged in a random walk fashion while maintaining constraints about the maximum speed of the tracked object. The process then repeats itself for more than one target. The result is a "line" of at least one target spanning across the matrix rows. The synthetic data was mostly used for calibrating the autoencoder architecture and exploring the learning hyperparameters.

The second dataset allows to test the performance of the system under more realistic conditions. The emulation includes both real clutter and real target's-based reflection, both obtained from sea recordings. The target's reflections are identified from roughly 20 sea experiments using the process described in [11] where also the process of identifying stationary reflection is described. The experiments were performed under controlled conditions, where first a fish was caught, then measured, and then released again. Samples which weren't identified as targets constitute clutter. To augment the emulation database, the emulation TD matrices are created from partial target detections, resulting in a dataset with more than 20000 images. The process is similar to that of the simulation, while exchanging the random-type target-like reflections with real detections. The resulting SCR range varied between 20 and -10. An example of such pattern is shown in Fig. 3. Besides images containing targets, both datasets also include a similar number of images containing no targets.

IV. The Convolutional Denoising Autoencoder

A graphical representation of our deep convolutional denoising autoencoder (CDA) is given in Fig. 4. The network receives as input a noisy image (i.e., the TD matrix) and returns as output a denoised version of the same image, where only the target path is present (output = 1) while the noise is suppressed (output = 0). The CDA is composed of four convolutional layers with, respectively, 24, 48, 72 and 96 filters of size 4×4, 6×6, 8×8 and 12×16. After each layer, a pooling layer with pool size 1×2 and stride 1×2 are included. Convolutional layers are followed by four deconvolutional layers of the same size, using nearest neighbor as upsampling function. Rectified linear units were used in all hidden layers, while a logistic activation function was used in the output layer. This allows to interpret the output of the CDA as probability values.

1) CDA training: The autoencoder was implemented in TensorFlow [24]. The network was trained with error back-propagation, using as loss function the cross entropy between network predictions and ground truth binary images containing the real target paths. Due to the unbalanced distribution of active pixels (target positions) compared to clutter or background noise pixels, a weighted cross entropy function was used. Training occurred over mini-batches of size 100 and proceeded for 5000 epochs.

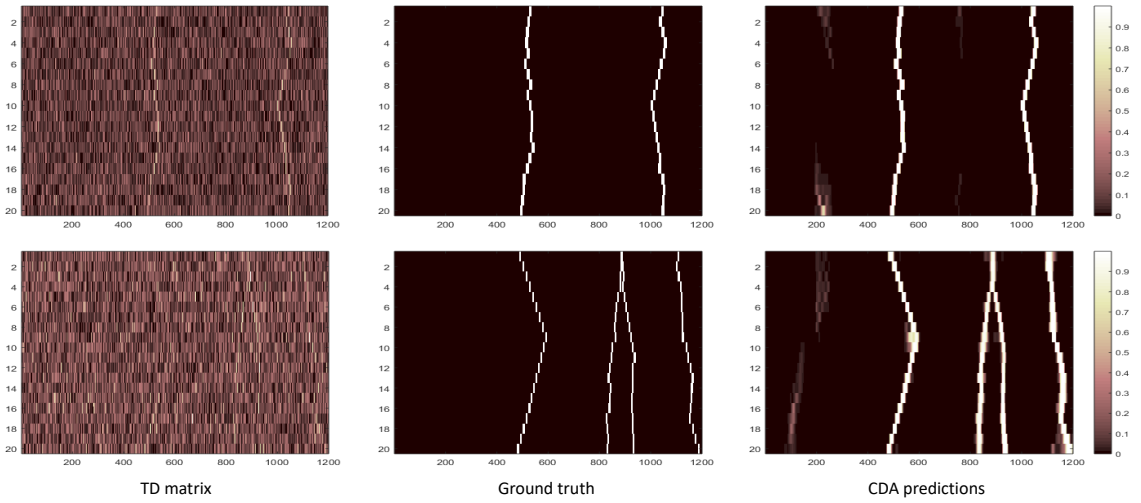

Fig. 5. Examples of noisy and denoised simulated matrices, along with ground truth, for a relatively high SCR (top) and for a low SCR (bottom).

2) CDA testing: The autoencoder was tested on separate test sets, on both tasks defined above. For the simulated dataset, generalization capability was assessed by training the CDA on images with only one target (or no target at all), while the test set included images with up to 10 targets. For the emulated dataset, the test set was created by randomly selecting 50% of the images in the database. For the binary detection task, which is defined image-wise, we computed the sum of pixel-wise predictions and used it as a level of confidence for the presence of a target. As the SCR increases, the network will be less confident in predicting the presence of a path, which will be reflected in a reduced activation of the output neurons. For the localization task, the predicted path directly corresponded to the set of output neurons with higher activation.

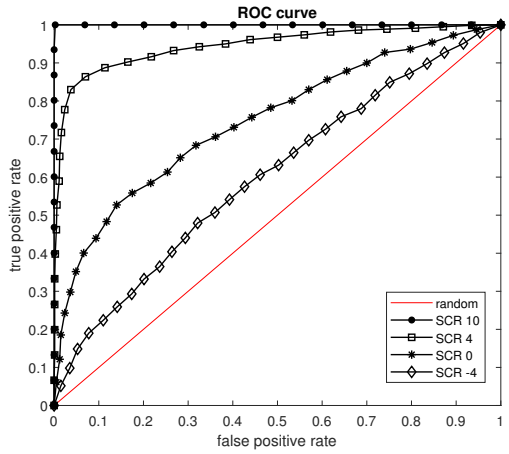

Fig. 6. ROC curves for the CDA at different levels of SCR on the emulated dataset.

V. Results

A. Performance on Simulated Data

For the sake of brevity, for simulated data we only show some examples of denoised TD matrices in Fig. 5. Even from this qualitative assessment it is evident that the CDA is able to accurately reconstruct the underlying paths, even for challenging levels of SCR where the lines in the image are almost invisible to human eye. As expected, predictions become blurred as the signal gets weaker, but overall the CDA output closely matches the ground truth. It is worth noticing that the CDA is able to track multiple targets even if it was only trained on images containing single paths.

B. Performance on Emulated Data

For emulated data, ROC curves for the detection task at four different levels of SCR are shown in Fig. 6. When the signal is clear (i.e., SCR = 10), the CDA detection is almost perfect: the false positive rate is close to zero, while the true positive rate is close to one. As the signal gets weaker performance gradually deteriorates, but it remains fairly good even when the SRC is low. For comparison, in Fig. 7 we show the ROC curves obtained from the Viterbi algorithm for the same levels of SCR. Performance for the highest SCR is still good (though far from ceiling), but all curves flatten out quickly as the SCR decreases. For low levels of SCR (i.e., SCR $\leq$ 0) the Viterbi performance is close to chance level (cfr. with random baseline).

For the accurate localization task, average Euclidean distances of track error for the same levels of SCR discussed above are reported in Fig. 8. In line with detection results, prediction error is indeed very small for the CDA, regardless of the SCR level. On the other hand, the performance of the Viterbi algorithm greatly deteriorates as the signal gets weaker, suggesting that this method cannot be applied when the signal gets too noisy.

VI. Conclusions

In this paper we described a novel application of deep learning for the acoustic detection and localization of moving targets in underwater scenarios. The proposed system is based

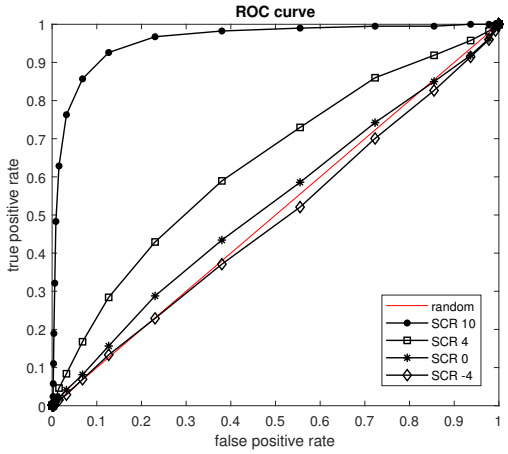

Fig. 7. ROC curves for the Viterbi algorithm at different levels of SCR on the emulated dataset.

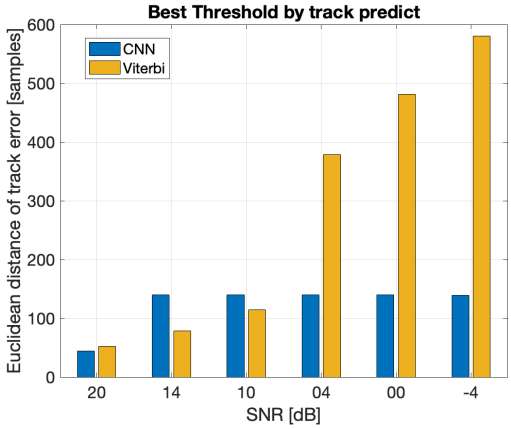

Fig. 8. Average prediction error for different levels of SCR for the localization task on the emulated dataset.

on a convolutional denoising autoencoder, which takes as input a noisy image representing the time-delay matrix of the signal and returns as output a denoised version of the same image, where only the targets' path are highlighted. Our results on both simulated and realistic (emulated) signals showed that this approach is a promising alternative to more traditional methods, such as those based on dynamic programming and the Viterbi algorithm. Besides being much more computationally efficient, our deep learning method turned out to be more accurate both for the detection task as well as for the accurate localization task, especially in high noise conditions.

The next step will be to validate the proposed system on real data collected from a SYMBIOSIS monitoring platform, in order to test the robustness of our method on a greater variety of environmental conditions. Finally, an interesting direction to further improve our system would be to combine deep learning with dynamic programming: the former could be used as an efficient pre-processing step; if targets are detected, the Viterbi algorithm could then be applied directly on the denoised matrix to provide a more precise path tracking.

REFERENCES

[1] T. Letessier, P. Bouchet, and J. Meeuwig, "Sampling mobile oceanic fishes and sharks: implications for fisheries and conservation planning," *Biological Reviews*, vol. 92, no. 2, pp. 627–646, 2017.

[2] A. Bertrand and E. Josse, "Acoustic estimation of longline tuna abundance," *ICES Journal of Marine Science*, vol. 57, no. 4, pp. 919–926, 2000.

[3] M. J. Parsons, I. Parnum, K. Allen, R. McCauley, and C. Erbe, "Detection of sharks with the Gemini imaging SONAR," *Acoustics Australia*, vol. 42, no. 3, p. 0, 2014.

[4] B. Wilson, "Maritime energy security," An ARW Book. IOS Press http://www.enseccoe.org/download/126/maritimeenergysecurity.pdf, 2012.

[5] V. N. Hari, M. Chitre, Y. M. Too, and V. Pallayil, "Robust passive diver detection in shallow ocean," in *OCEANS*, May 2015, pp. 1–6.

[6] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[7] S. Davey, M. Wieneke, and H. Vu, "Histogram-PMHT unfettered," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 3, pp. 435–447, June 2013.

[8] J. Renard, L. Lampe, and F. Horlin, "Sequential likelihood ratio test for cognitive radios." *IEEE Transaction on Signal Processing*, vol. 64, no. 24, pp. 6627–6639, 2016.

[9] P. Willett and S. Coraluppi, "Application of the MLPDA to bistatic sonar," in *IEEE Aerospace Conference*, March 2005, pp. 2063–2073.

[10] W. Blanding, P. Willett, and S. Coraluppi, "Sequential ML for multistatic sonar tracking," in *OCEANS*, June 2007, pp. 1–6.

[11] R. Diamant, D. Kipnis, E. Bigal, A. Scheinin, D. Tchernov, and A. Pinchas, "An active acoustic track-before-detect approach for finding underwater mobile targets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 104–119, 2019.

[12] Y. Wu, P. Chen, F. Gu, X. Zheng, and J. Shang, "HTrack : An efficient heading-aided map matching for indoor localization and tracking," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3100–3110, April 2019.

[13] H. Gaetjens, S. Davey, S.Arulampalam, F. Fletcher, and C. Lim, "Histogram-PMHT for fluctuating target models," *IET Radar, Sonar Navigation*, vol. 11, no. 8, pp. 1292–1301, 2017.

[14] S. Schoenecker, P. Willett, and Y. Bar-Shalom, "Resolution limits for tracking closely-spaced targets," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–1, 2018.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[18] A. Testolin, I. Stoianov, M. De Filippo De Grazia, and M. Zorzi, "Deep unsupervised learning on a desktop pc: a primer for cognitive scientists," *Frontiers in psychology*, vol. 4, p. 251, 2013.

[19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 609–616.

[20] A. Testolin, M. Zanforlin, M. D. F. De Grazia, D. Munaretto, A. Zanella, M. Zorzi, and M. Zorzi, "A machine learning approach to qoe-based video admission control and resource allocation in wireless systems," in *2014 13th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET)*. IEEE, 2014, pp. 31–38.

[21] M. Zorzi, A. Zanella, A. Testolin, M. D. F. De Grazia, and M. Zorzi, "Cognition-based networks: A new perspective on network optimization using learning and distributed intelligence," *IEEE Access*, vol. 3, pp. 1512–1530, 2015.

[22] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, p. 4308, 2014.

[23] R. Diamant, "Closed form analysis of the normalized matched filter with a test case for detection of underwater acoustic signals," *IEEE Access*, vol. 4, pp. 8225–8235, 2016.

[24] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.