Editors' Suggestion

Neural sensing and control in a kilometer-scale gravitational-wave observatory

N. Mukund,[1,2,3,*] J. Lough,[1,†] A. Bisht,[1] H. Wittel,[1] S. Nadji,[1] C. Affeldt,[1] F. Bergamin,[1]
M. Brinkmann,[1] V. Kringel,[1] H. Lück,[1] M. Weinert,[1] and K. Danzmann[1]

[1] *Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut) and Institut für Gravitationsphysik, Leibniz Universität Hannover, Callinstraße 38, Hannover 30167, Germany*

[2] *Laser Interferometer Gravitational-Wave Observatory (LIGO), Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[3] *National Science Foundation (NSF) AI Institute of Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

(Received 24 April 2023; revised 14 November 2023; accepted 20 November 2023; published 22 December 2023)

Suspended optics in gravitational-wave (GW) observatories are susceptible to alignment perturbations, particularly slow drifts over time, due to variations in temperature and seismic levels. Such misalignments affect the coupling of the incident laser beam into the optical cavities, degrade both the circulating power and optomechanical photon squeezing, and thus decrease the astrophysical sensitivity to merging binaries. Traditional alignment techniques involve differential wave-front sensing using multiple quadrant photodiodes but are often bandwidth restricted and limited by the sensing noise. We present a successful implementation of neural-network–based sensing and control at a GW observatory and demonstrate low-frequency control of the signal-recycling mirror at the GEO 600 detector. Alignment information for three critical optics is simultaneously extracted from the interferometric dark-port camera images via a convolutional neural net–long short-term memory network architecture and is then used for multiple-input–multiple-output control using soft actor-critic-based deep reinforcement learning. The overall sensitivity improvement achieved using our scheme demonstrates the capabilities of deep learning as a viable tool for real-time sensing and control for current and next-generation GW interferometers.

DOI: 10.1103/PhysRevApplied.20.064041

I. INTRODUCTION

GEO 600 is a dual-recycled advanced Michelson interferometer (IFO) with folded arms [1–3], located near Hannover, Germany. With a peak strain sensitivity of about $10^{-22}/\sqrt{\mathrm{Hz}}$ at 1 kHz, the observatory operates in AstroWatch mode [4] and takes astrophysically relevant gravitational-wave (GW) data in the frequency band of 40 Hz to 6 kHz. In April 2020, GEO completed a joint observation with KAGRA [5,6] and searched for transient GW events from neutron-star binaries and generic unmodeled transients [7]. Several technologies pioneered at GEO 600 [8] have been adopted at Advanced LIGO [9,10] and Advanced Virgo [11] and have played a crucial role in advancing GW instrumentation science. One such example is the continuous application of nonclassical sources of light [12] and the demonstration of 6 dB of measured optical squeezing [13] on the key future upgrade goals for GW detectors.

In this work, we present a technique using neural networks (NNs) and demonstrate their capabilities to sense and control the state of the interferometer. The paper is organized as follows. Sec. II describes the existing alignment scheme and its limitations, Sec. III explains the motivations and architecture of the neural sensing, Sec. IV describes the implementation of a deep reinforcement-learning- (RL) based controller, and Sec. V provides the network predictions and improvements to the sensitivity when the trained controller is deployed for low-frequency signal-recycling alignment control.

II. AUTOMATIC ALIGNMENT

The astrophysical sensitivity of a Michelson interferometer can be improved by including two extra cavities, a power-recycling cavity (PRC) and a signal-recycling cavity (SRC), leading to an improved signal-to-noise ratio (SNR) in the readout channel [14]. The PRC at GEO 600 consists of the PR mirror, located at the input port of the

[*] nikhil_m@mit.edu

[†] james.lough@aei.mpg.de

IFO, and the Michelson IFO. With an optical gain of about 800, it is used to enhance the circulating laser power, leading to a reduced level of photon shot noise. Similarly, the SRC is formed by the SR mirror, situated at the output port of the IFO, and the Michelson IFO. It complements the PRC by forming a resonant cavity to enhance the signal sidebands from potential GWs. The microscopic position of the SR mirror also determines the overall frequency response of the cavity. The light that leaks out in transmission of the SR mirror is filtered using the output-mode cleaner (OMC) and is sent to the main photodiode, which is then calibrated to produce the final GW strain data.

The IFO mirrors are suspended as multistage pendulum assemblies to suppress the seismic noise coupling, with the PR mirror having two pendulum stages and the Michelson mirrors and the SR mirror having three. Multistage pendulums at GEO have multiple resonance frequencies centered around 1 Hz and the noise suppression is achieved above the resonance frequency of the pendulum. While this isolation is sufficient within the GW measurement band, the residual pendulum motion around the resonance frequency can cause misalignment of mirrors and long-term drifts, which is detrimental to the required sensitivity of the interferometer. Suboptimal alignment of the incident beam to the OMC leads to intensity fluctuations in the photodiode signal, degrading the overall optical gain and increasing glitches that often mimic the actual GW signal. Such misalignments routinely interfere with the suite of optical-squeezing and thermal-compensation experiments carried out at GEO. Offsets introduced by such drifts are also seen to affect the angular-control loops in Advanced LIGO and Advanced Virgo. Automatic alignment systems are hence critical to attain optimal sensitivity and maintain long lock stretches at current and future-generation GW observatories.

The goal of the autoalignment system is to keep the axis of an incoming beam aligned to that of the cavity axis. In addition, it also keeps the beam spots centered on the mirrors. Angular alignment of the IFO mirrors is primarily carried out using the differential-wave-front-sensing (DWS) technique [15–17] and becomes active once the cavities are "locked" in length using the Pound-Drever-Hall technique [18]. In the DWS technique, phase modulation is imprinted onto the beam incident on a cavity, which is promptly reflected. It is then superimposed over another light field that leaks out of the cavity. This combined light field falls on a pair of quadrant photodetectors that are placed with a relative Gouy phase of 90°. The angle and displacement between the two beams are obtained by taking the difference of the photocurrent (demodulated at the modulation frequency) from the different QPD sections. If the beam spot is off center by one beam radius, then about 86% $(1 - e^{-2})$ of the DWS signal is lost [19]. Hence, there are usually two additional auxiliary centering control loops for DWS, one associated with each quadrant photodetector, that keep the beam spots centered on it. We use additional spot-position control loops to keep these beam spots centered on each mirror. In the transmission port of each mirror is a quadrant photodetector that looks at the position of the beam spot. The cavity mirrors are then actuated directly or in some combination of available external actuators (preceding suspensions) to achieve the desired spot-position control.

The DWS control has a bandwidth of up to 6 Hz, while the centering control loops are the fastest, having up to 1 kHz bandwidth. The slowest is the spot-position control loops, which have less than 0.1 Hz bandwidth. For completeness, we would also like to mention that waist-position and waist-size mismatch between interfering beams are second-order misalignments that are not actively controlled but are controlled by optimal layout design. Despite the autoalignment system, residual mirror misalignments can couple directly to the strain signal or through the interlinked cavities. One well-known mechanism is bilinear noise coupling [20], where the Michelson misalignment couples via the SR longitudinal degree. Such a coupling pathway exists since the PRC and SRC share the Michelson. Error signals for the DWS generated via Schnupp modulation result in radio-frequency (rf) sidebands, which are tens of megahertz offset to the laser (or primary carrier) frequency. Although the OMC suppresses these megahertz sidebands and the higher-order modes by a factor of 100 beyond its optical bandwidth at 2.9 MHz, they still leak into the final photodiode signal, leading to an elevated shot-noise floor and a reduced level of optical squeezing. Decreasing the level of rf sidebands is not viable with the existing system, as it leads to a low SNR error signal, making it harder to control. Additionally for GEO, environmental events such as excessive seismic motion or thermal fluctuations introduce sensing noise, leading to off centering or clipping of the beam on the DWS photodiode, impacting the drift control loops, often requiring a manual inspection. Consequently, the dc position of all the mirrors, notably the SR mirror, has to be tuned once a week for optimal detector sensitivity.

Another alternative to DWS in use at GEO is the dithering scheme. This involves mechanically oscillating the relevant optics at a specific frequency for each degree of freedom (DOF). The transmitted cavity power recorded by a single-element photodiode is then demodulated at the respective frequency to infer the corresponding misalignment. Such a scheme is used, e.g., to align the OMC by dithering one of the beam-directing optics. This scheme has a lower bandwidth (20 mHz) and causes additional jitter on the incident beam, leading to a 0.2-dB loss of squeezing. An alternative scheme based on modulated differential wave-front sensing is currently under commissioning for the OMC alignment [17]. The dithering also enhances the bilinear coupling to the strain if the beam is not well centered on the optic. All these reasons motivate the need for a better solution.

III. NEURAL SENSING

A. Why the dark port is a good witness

The south port of the IFO, referred to as the dark port, is usually kept close to destructive interference but with a slight dc offset of about 5–50 pm. This offset allows about 6 mW of carrier to leak out and about 30 mW of higher-order modes to exit via the dark port. The higher-order modes originate inside the IFO due to mismatch in the interfering beams, which is caused by thermal lensing of the beam splitter, microscopic imperfections on the mirror surfaces, or residual misalignment of the mirrors. Consequently, video-camera images of the dark-port (DP) beam contain much information about the IFO state. It is used for manual prealignment, making the longitudinal lock acquisition easier, and then the wave-front-sensor–based autoalignment systems take over. In the lock, the DP image shows breathing motion corresponding to the residual movement of the suspended optics. A skilled commissioner can often judge some alignment states from this image.

The error signals of several feedback loops can broadly determine the state of the IFO. In particular, the Michelson differential, the PRC, and the SRC-alignment DOFs play a crucial role for GEO. Sensing noise entering the existing DWS-based scheme is often not sufficiently corrected by the current-control loops, leading to pointing drifts and sensitivity degradation. The time scales of these disturbances range from hundreds of milliseconds to a few days and include sources such as temperature variations, seismic disturbances, and optomechanical intracavity cross-couplings. However, since a strong mapping exists between the state of the interferometer and the dark-port image, such disturbances are encoded in their breathing patterns. Once a week, these loop offsets are tuned by commissioners via visual inspection of the camera images. The values are considered optimized when the dark-port image resembles a stable state, often based on recollections from memory.

B. Coherence mapping

Figure 1 reveals the complex way in which the multiple optics imprint their state of alignment on the interferometric dark port. The map is constructed by measuring coherence between the pixel-wise dark-port intensity fluctuations and the temporal variation in the different alignment error signals. C_{xy}, the metric used for computing the coupling, is the magnitude-squared coherence in the (0.1–4.0) Hz band, which contains the alignment information, weighted by the average logarithmic error-point spectra,

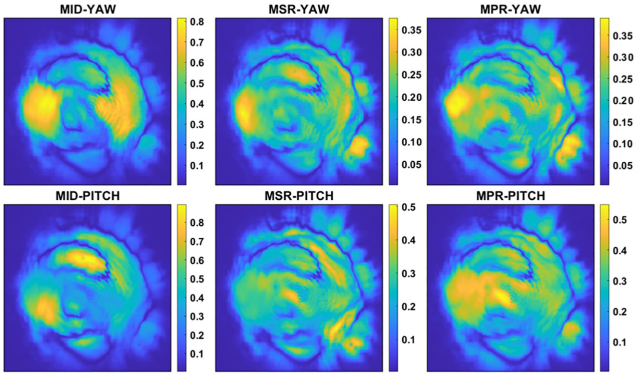

FIG. 1. Coherence maps show the complex coupling of the critical-alignment DOFs to the interferometric dark port. The color bar indicates the coupling strength, estimated using the weighted coherence in the (0.1–4.0) Hz band. The images from left to right depict the Michelson differential arm motion and the motion from the signal-recycling and power-recycling mirrors. The axes represent the dark-port camera pixels. MID, Michelson differential; MSR, signal recycling mirror; MPR, power recycling mirror.

$$C_{xy} = \frac{\int_{f_1}^{f_2} \log_{10}(<y(f)^{\mathrm{asd}}_{\mathrm{norm}}>) \cdot \frac{|P_{xy}(f)|^2}{P_{xx}(f)\,P_{yy}(f)}}{\int_{f_1}^{f_2} \log_{10}(<y(f)^{\mathrm{asd}}_{\mathrm{norm}}>)}. \qquad (1)$$

The high coherence and peculiar spatial spread confirm that dark port contains a treasure trove of information about the IFO. Apart from being a helpful detector-characterization tool, coherence maps can be used to identify potential signals that a trained NN can recover. We find a couple of interesting observations. Angular misalignment causes coupling of the light field into the first-order mode, which is well captured by the appearance of (1,0) and (0,1) Hermite Gaussian mode patterns and is most prominent for the Michelson optics. We see more complex and radially extended structures for signal- and power-recycling optics, which could indicate higher-order spatial modes.

C. Neural-sensor architecture

The neural alignment sensing we intend to do can be formulated as an image-to-time-series regression problem. We chose a CNN-LSTM architecture for this task for a few reasons. While two-dimensional (2D) convolutional neural nets (CNNs) are well suited for analyzing image data with complex spatial representation, long short-term memory networks (LSTMs) [21] excel at temporal modeling and sequence prediction. LSTMs, a specific form of recurrent NNs, use a memory cell that selectively controls the flow of information using input, output, and forget gates. They have also had limited success at linear-system identification tasks, with results comparable to traditional transfer-function estimation [22]. The ability to learn representations in both space and time thus makes the combined deep recurrent convolutional model effective at activity recognition from streaming video data and makes it a good candidate for capturing the underlying system dynamics [23,24].

Our design choice for the CNN architecture is based on transfer learning [25,26], where the initial layers of pretrained networks, fine tuned to extract spatial information at different scales by training on standardized data sets, are reused for a newer task. Transfer learning alleviates the need for training networks from scratch and is useful when the data are limited in size. In particular, we focus on inception-based networks where spatial filters of different scales are convolved in parallel, thus processing information at bigger scales and finer resolution. These networks represent a synergy between classical computer vision and deep architectures and have previously successfully recovered all the GW events listed in the GWTC-1 transients catalog [27]. We use three reference architectures, namely SqueezeNet [28], GoogLeNet [29], and InceptionResNetV2 [30] and compare the respective trade-offs. We chose SqueezeNet as a comparatively lightweight network with 18 layers, making it suitable for embedded devices and low latency inference. The 164-layers-deep InceptionResNetV2 is among the largest pretrained networks and provides high classification accuracy on several benchmark data sets. GoogLeNet is often a good choice when balancing network size and accuracy.

D. Training strategy

Our aim with the CNN-LSTM model is to train the network on sufficient dark-port images and the corresponding DWS alignment signals from a well-tuned interferometer configuration and predict the new error points whenever the detector gets into a misaligned state. If the model is well trained, it should be able to predict the current-loop offset value affected by drifts and then either a human or a controller [either classical PID or RL-based] can set it to the last known "good" state.

Training deep networks is, in general, a time-consuming process and, additionally, we need to find the right hyperparameters to maximize the learning process. We adopt a strategy where we start with SqueezeNet, cut the network just before the final fully connected layers, and add the LSTM layers with ten hidden units to its output. We then freeze the weights of the SqueezeNet layers and let the LSTM layers learn while the combined network is trained to predict the alignment error points from the recorded dark-port images. This configuration makes it easier to determine parameters such as the gradient decay rate, the LSTM hidden units, and the learning rate using minimal computational resources. In the second stage of training, we retrain the entire network comprising the pretrained CNN and newly trained LSTM and fine tune the network weights and biases. This stage of training is carried out on a dedicated A100 graphics processing unit (GPU) cluster. We repeat the process for the other two pretrained networks. Table I compares the root-mean-squared error between the neural-sensor predictions and the actual alignment signals for the six mentioned DOFs.

E. Network quantization

Most often, the learnable parameters of NNs are trained using single-precision floating-point data types. However, the limited dynamic range of these parameters makes it possible to cast them as scaled 8-bit integer data types of fixed length. Such quantization can significantly reduce the memory footprint, improve the inference rate, and lower the power consumption [31,32]. This step would be crucial when the trained networks are deployed at a large scale in GW detectors using embedded devices such as field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), or GPU-accelerated EDGE devices for real-time processing. We use a training data set to calibrate the dynamic range of the weights and biases of the convolutional and fully connected layers and the activations in all the layers. Using a separate validation data set, we quantize to the right data type (single-bit floating point or INT8), ensuring that we cover the range, avoiding overflows but ignoring potential underflows. Table I gives the memory reduction and the improved processing speed, measured in terms of frames per second, and the decrease in accuracy for the three quantized network architectures. We select fine-tuned GoogLeNet LSTM for the rest of our analysis, as it provides a decent trade-off among metrics

TABLE I. Sensor prediction errors for the three neural architectures. We compare the root-mean-square error (RMSE) for the trained LSTM layers with the pretrained CNN and the fine-tuned CNN-LSTM network. The last three columns provide the results after quantizing the combined network from single-bit floating point to INT8 precision.

Model	Pretrained	Retrained	After quantization		
	RMSE	RMSE	Memory compression factor	Frame-rate increase factor	RMSE degradation (%)
SqueezeNet-LSTM	0.135	0.118	3.96	5.9	8.3
GoogLeNet-LSTM	0.122	0.105	3.98	4.6	4.3
InceptionResNetV2-LSTM	0.132	0.096	3.98	4.1	21.9

such as the time for training, the prediction accuracy, and the real-time inference rate.

IV. MODEL-BASED CONTROLLER DESIGN

After the neural sensor is built as described above, we require a suitable controller to close the loop. Designing such controllers with the actual interferometer-in-loop is not encouraged. Doing so reduces the observation time and can lead to undesired behavior such as oscillations in the system that could take a long time to settle down. For example, in its search for the optimal policy, an RL agent can intentionally carry out random action sequences as it tries to strike a balance between exploration and exploitation. We hence follow a model-based approach and design a controller that can utilize the signals from the neural sensor. The original high-fidelity instrument response is approximated by a reduced-order model that sufficiently captures the dominant system dynamics relevant to the controller design. The response of the interferometer is analyzed by randomly perturbing the set points in the SR pitch and yaw DOFs that cover the actuation range of the existing controller. This perturbation leads to variation in the dark-port images and is processed by the CNN-LSTM neural sensor. System identification, mapping set points to neural predictions, is then carried out using subspace-based state-space modeling [33]. State space offers superior performance over transfer-function models due to the ability to include a noise model. During the fitting process, the model order is varied over a reasonable range and the one with the lowest Hankel singular value [34] is selected. The selection of such lower values helps retain the larger energy states, making it possible to have a reduced-order model that preserves most system characteristics. The identified model is further refined using the prediction-error minimization, where the weighted norm of the difference between the measurement and the predicted output of the model is minimized [35]. We have looked at further improvements by adding input and output nonlinearities to the identified state-space model, resulting in a nonlinear Hammerstein-Wiener (NLHW) model. Figure 2 compares both the models on estimation and validation data, where the goodness of fit is given as

$$\mathrm{GOF} = 100\left(1 - \frac{\|y_{\mathrm{model}} - y_{\mathrm{meas}}\|}{\|y_{\mathrm{meas}} - y_{\mathrm{meas}}^{\mathrm{mean}}\|}\right). \tag{2}$$

The NLHW model, with a sigmoid network function representing the nonlinear mapping, only provides a modest improvement in one DOF. Hence, we select the state-space model for the rest of the analysis.

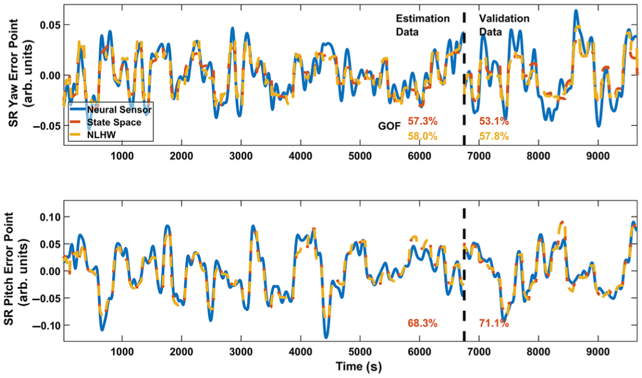

FIG. 2. A comparison of the neural-sensor-inferred system dynamics with reduced-order models. The identified state-space model is used to design the optimal PID controller and train the RL agent.

A. Classical PID-based controller

PID controllers are among the most widely used classical controllers for linear-time-invariant (LTI) systems. They are easy to tune, depend only on the error signal, and are less susceptible to plant variations. They can be designed to ensure closed-loop stability of the plant [36] and also serve as a benchmark while evaluating the performance of the RL-based solutions described in Sec. IV B. These linear controllers, however, need to be separately tuned for each DOF. We use the plant model identified in the previous section and automate the tuning, focusing on reference tracking. Tunable parameters are obtained using H-infinity synthesis by optimizing across the target bandwidth, performance, and robustness requirements [37,38]. However, the presence of an actuator with a limited range introduces nonlinearities and often leads to the well-known integral wind-up [39]. We overcome this using additional anti-wind-up circuity built using a tracking signal and a reference feed-forward. The output for the controller with an error signal $e(t)$, depicted in Fig. 3, is given by

$$u(t) = K_p\, e(t) + K_d\, \frac{d\, e(t)}{dt} + K_r\, r(t) + \int [K_i\, e(t) + K_t\, u_s(t) - K_t\, K_r\, r(t) - K_t\, u(t)]\; dt. \tag{3}$$

where K_p, K_i, and K_d are the usual PID gain coefficients and K_r controls the reference $r(t)$ feed-forward, while the tracking coefficient K_t and the saturated output u_s are part of the modified integral term. As shown in Fig. 3, the derivative term is implemented as a filtered derivative with a gain of N.

B. Deep reinforced controller

RL is an experience-based learning framework that eliminates the need for supervision and subject expertise and attempts to learn to carry out a task-based purely on

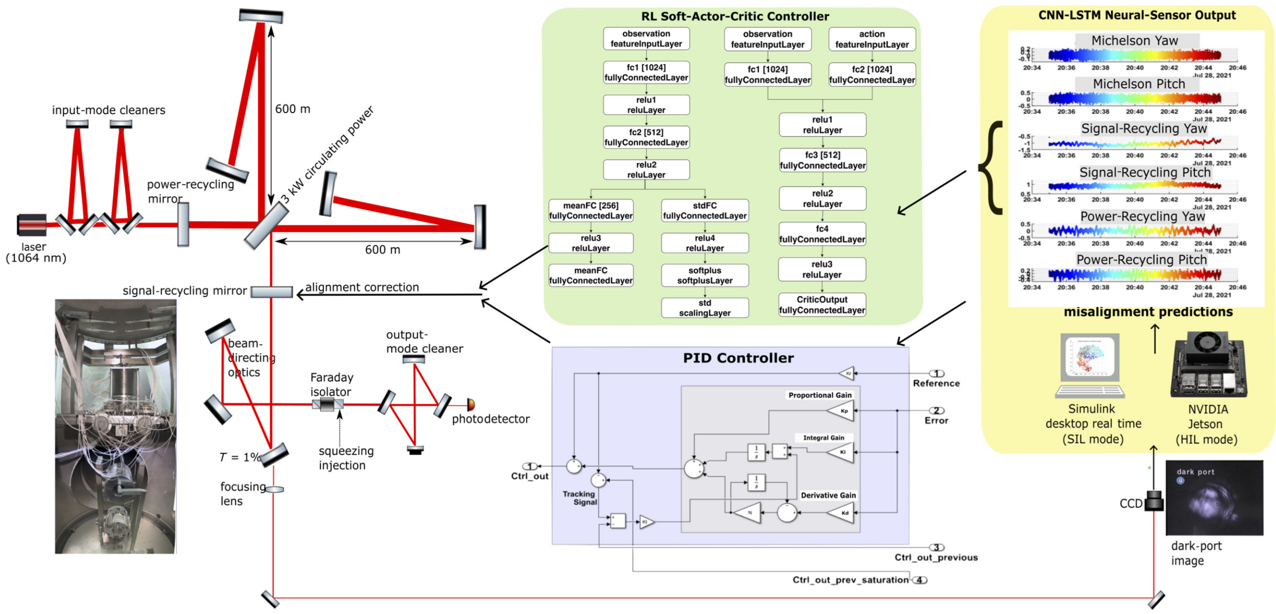

FIG. 3. The simplified optical layout of GEO 600, highlighting the AI-based alignment sensing and control scheme. The CCD captures 2D images of the beam that exits the interferometric dark port. The CNN-LSTM NN simultaneously extracts the pitch and yaw DOFs for the Michelson, signal-recycling, and power-recycling mirrors. The RL agent processes this information and corrects the low-frequency alignment drifts of the signal-recycling mirror, thus improving the astrophysical sensitivity.

its interaction with the system [40]. The notion of a traditional controller is replaced here by an RL agent consisting of a deep NN and a policy-updating algorithm. The former provides high-capacity representations that are easy to generalize, while the latter offers a mathematical formalism for decision making and optimal control. During the training process, the agent observes the current state of the system, interacts with the environment, and considers the new states and the reward, an immediate measure of the goodness or badness of the current action. The agent then tries to learn the optimal policy, or the mapping between states and actions, to maximize the discounted cumulative long-term reward.

1. Soft-actor-critic algorithm

Traditional RL algorithms were thought to be unstable and unpredictable, making them sensitive to hyperparameters and initial conditions. One way to overcome this scenario is to cast RL and the optimal control as a probabilistic inference problem. Soft-actor-critic (SAC) consists of a set of algorithms [41] that utilizes the traditional actor-critic methods [42–45] but ensures maximization of the entropy of the learned policy. The action-value function (or the Q function), which evaluates the quality of the agent's actions, is determined using a pair of critic networks, thus minimizing the over-estimation bias. They are trained using the Bellman equation, which involves an iterative update of the value function whenever a state-action pair is traversed by the agent and is given by

$$\begin{aligned} Q^{\text{new}}(S,A) = Q^{\text{prev}}(S,A) \\ + \alpha\left[\left(R(S,A) + \gamma \max_{A'} Q(S',A')\right)\right. \\ \left. - Q^{\text{prev}}(S,A)\right], \end{aligned} \tag{4}$$

where γ is the discount factor for future rewards and α controls the value-update learning rate for a given state-action (S,A) pair. The actor network representing the policy π is trained using the gradient of the expected return concerning the actions, which is computed using the critic network. By learning a probabilistic regularized "soft" policy trained to maximize both value and policy entropy,

$$\max_{\pi} \mathbb{E}_{\pi}\left[Q(S,A) - \log \pi(A\,|\,S)\right], \tag{10}$$

the agent learns a wide range of behaviors, including stochastic or deterministic behaviors. A comparatively faster learning rate, lower sensitivity to hyperparameters, the ability to reuse past experience, and a balanced trade-off between exploration and exploitation make SAC a good candidate for real-world control problems.

An ideal reward function should guide the agent to the optimal policy. However, creating a suitable reward function is the most critical task in RL training. One goal of this work is to assess the practicality of this approach in designing controllers suitable for GW detectors and probe if a

TABLE II. The reward-function equations used to train the RL agent. The continuous part is built using the linear quadratic regulator cost function. Discrete terms, penalties, and boosts are added to penalize the violation of boundary constraints and emulate final-state constraints.

Cost	$\sum_{j=1}^{\tau}(S_j - S_j^{\text{ref}})^T \mathbb{Q}_j\,(S_j - S_j^{\text{ref}}) + (A_j - A_j^{\text{prev}})^T\,\mathbb{R}_j\,(A_j - A_j^{\text{prev}})$
Penalty	$W_y\,\left((S_j - S^{\min})^2 + (S_k - S^{\max})^2\right) + W_{\text{mvrate}}\,\left((\dot{A}_l - \dot{A}^{\max})^2 + (\dot{A}_m - \dot{A}^{\min})^2\right)$ $\forall\left(S_j < S^{\min}, S_k > S^{\max}, \dot{A}_l < \dot{A}^{\min}, \dot{A}_m > \dot{A}^{\max}\right)$
Boost	$10\sum_{j=1}^{\tau}(3\,\lvert S_j - S_j^{\text{ref}}\rvert < 0.02)^2 + 10\sum_{j=1}^{\tau}(6\,\lvert S_j - S_j^{\text{ref}}\rvert < 0.005)^2$
Reward	$-(\text{cost} + \text{penalty}) + \text{boost}$

set of general guiding principles can help design a reward that leads the agent to the optimal policy. Table II lists the components of the reward function used to train our RL agent. We draw cues from optimal control theory, which aims to operate dynamical systems with minimal controller effort. The continuous portion of the reward can be derived from the corresponding linear quadratic regulator (LQR) cost function. For LTI systems with a quadratic cost function, the LQR provides the optimal gain matrix for state feedback control by solving the Riccati equation of the state-space model. The corresponding cost that drives the state close to the reference with minimal actuator effort is expressed in terms of both the current and reference state (S_j, S_j^{ref}) and the current and previous actuator values (A_j, A_j^{prev}), with $\mathbb{Q}_j$ and $\mathbb{R}_j$ being the respective weight matrices. Such continuous rewards encourage convergence but are prone to local minima and can lead to longer training periods. Adding discrete elements that penalize or encourage the agent increases the probability of finding better states. However, the nonsmooth nature of the resulting loss function can affect the convergence. We discourage boundary-constraint violations from the agent by including a discrete penalty term, where behaviors that drive the states close to the limits ($S^{\min}$ and $S^{\max}$) or increase the controller velocity beyond a threshold value ($\dot{A}^{\min}$ and $\dot{A}^{\max}$) are penalized, with W_y and W$_{\text{mvrate}}$ being the associated weight matrices. We also observe the benefit of including a discrete positive reward when the state is driven close to the reference.

The random initialization of network weights and the entropy-maximization objective associated with the optimal policy learning make the overall convergence rate moderately sensitive to individual simulation runs. Hence, we carry out ten training trials for each network configuration. One usual design decision is to choose between a deeper or wider network. In supervised learning tasks, issues with vanishing gradients make it harder to train deeper networks and are usually overcome using residual connections. However, in the case of the RL agent, the difficulty in training deeper networks arises from the sharpness of the loss-surface curvatures, making them more susceptible to the choice of hyperparameters. Recent studies [46] prefer the wider networks, as they have nearly convex loss surfaces. We observe similar performance improvement with increased network width (see Fig. 4).

C. Multiagent control

Ideally, the designed controllers should be less susceptible to the uncertainties associated with the modeled environment and our limited knowledge of the optimal reward. One way to achieve robustness is by blending in control signals and leveraging the positive aspects of each, such as the low integral error from the PID controller and the faster response of RL. Ensemble learning [47] is another option, where the top-performing agents across the multiple simulations are combined to form the optimal signal by averaging the maximum-likelihood action suggested by each. We report the findings from simulating probable controller combinations in Table III. It includes

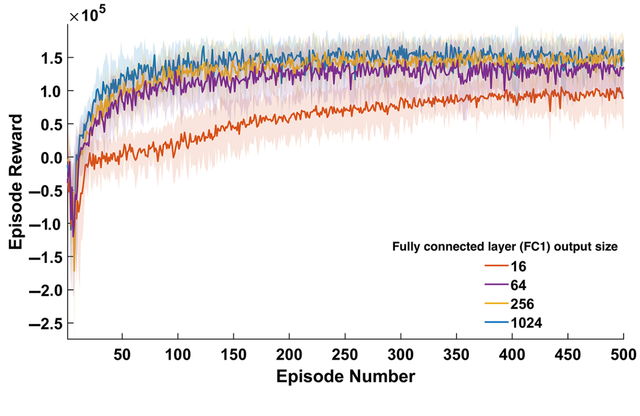

FIG. 4. The average episode reward received for the RL agent with increasing network size. The output size of the first fully connected (FC1) layer is indicated and other FC layers (see Fig. 3 for the network architecture) are proportionally changed. The shaded region gives the standard-deviation errors obtained from ten independent trials.

TABLE III. A comparison of different controllers for alignment reference tracking of the SR mirror.

PID controller active	RL agent active	RL agent type	Relative average reward
Yes	No	· · ·	0.95
No	Yes	Single, suboptimal	0.80
0.7	0.3	Single, suboptimal	0.96
0.5	0.5	Single, suboptimal	0.96
0.5	0.5	Ensemble, optimal	0.96
0.3	0.7	Single, suboptimal	0.94
0.3	0.7	Ensemble, optimal	0.97
No	Yes	Ensemble, optimal	1

a tuned PID controller for each DOF, a single RL agent with an average performance, an ensemble of optimally performing RL agents, and a few combinations where the signals are blended. The ensemble learner achieves the best performance measured in terms of the recovered average reward, where the set points are randomly perturbed across the actuation range. The corresponding bias and variance associated with the 2-DOF reference tracking for each controller configuration are shown in Fig. 5.

V. RESULTS

One of the main objectives of this work has been to evaluate AI-based sensing and control as a viable alternative at GW observatories. We have emphasized the construction of AI strategies that provide superior performance while retaining the advantages of traditional techniques such as ease of design, robustness, and explainable nature. In addition, we have wanted to evaluate whether an RL agent's policy learned from the simulated environment can transfer well when deployed to the real environment. We have focused on optical alignment, commonly encountered across interferometers, and the development of techniques that are easily transferable to other observatories such as Advanced LIGO. In Fig. 6, we present the alignment predictions made by the neural sensor and compare them with the actual measurements for the six critical-alignment DOFs. These signals are generated from unseen dark-port image sequences and we observe a good match in both the time and the frequency domain with actual error points. Compared to the wave-front sensor, the neural sensor is not susceptible to sensing noise, as it directly reconstructs the alignment information from the dark-port light fields and is suitable for long-term use without the need for routine intervention.

In Fig. 3, we show the simplified optical layout of GEO 600, highlighting our AI-based alignment-sensing-and-control scheme. The CCD captures 2D images of the beam that exits the dark port through the 1% transmission port of the beam-directing optic. The CNN-LSTM neural sensor deployed as a quantized network extracts the pitch and yaw DOFs for the Michelson, signal-recycling, and power-recycling mirrors. We finally close the loop by deploying the lighter GoogLeNet-LSTM neural sensor and the ensemble RL agent in real time. The actuation signals from the RL agent provide low-frequency alignment correction of the signal-recycling mirror and the corresponding impact on the detector is described below.

The reasonable metric to assess the impact of the neural scheme is to analyze the GW strain curve and estimate the improvements to astrophysical sensitivity. The nonstationary nature of the noises influencing the detector, primarily the ambient seismic noise, often makes comparison of different time segments difficult. To address this, we analyze several pairs of segments approximately 30 min long, each with the SR mirror optimized manually and using the AI-based controller engaged. We use a previously trained controller, as described in Sec. IV. Figure 7 compares this sensitivity in terms of the farthest luminosity distance for optimally orientated and located coalescing neutron-star binaries (1.4 solar mass each), detectable with a matched-filter SNR of eight or above. The blue trace shows the horizon distance when experienced commissioners fine tune the interferometer by inspecting the dark-port images, while the orange trace depicts the results obtained using the deep RL agent. The first peak in Fig. 7 is when the interferometer is shot-noise limited at high frequencies, a state similar to the training data. The second peak is when the interferometer is operating at a modified state with the injection of squeezed states of light. In both cases, we observe the neural-optimized segments to outperform the manual tuning from the experienced commissioners.

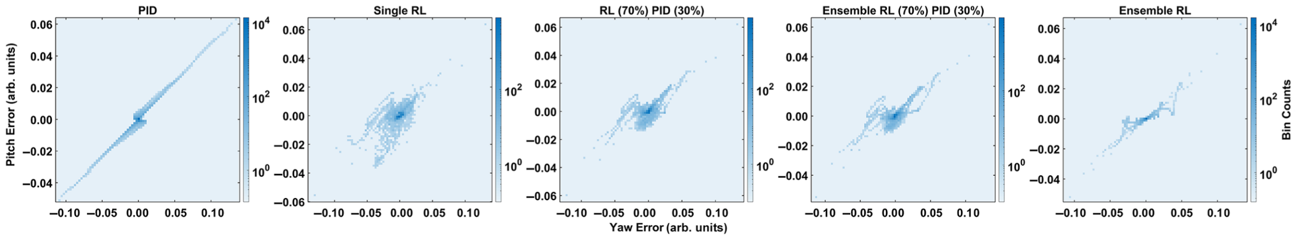

FIG. 5. The bias and variance in 2-DOF reference tracking for the controllers mentioned in Table III.

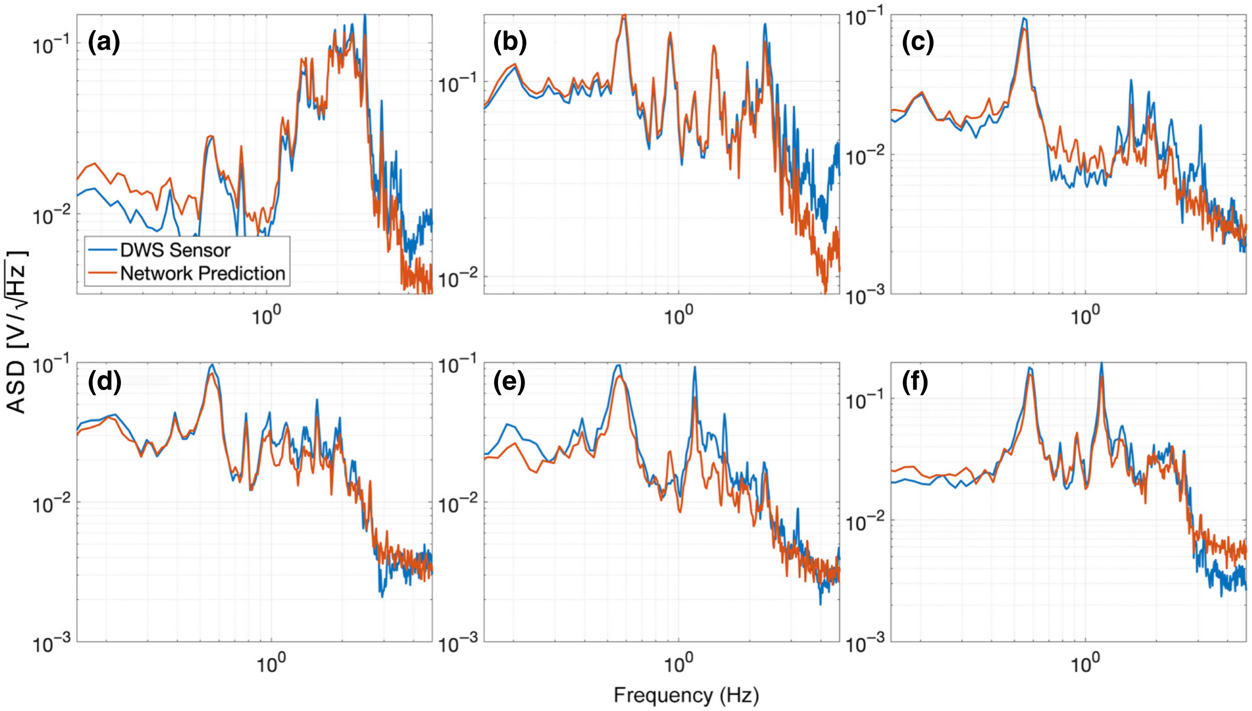

FIG. 6. Comparison alignment predictions for the key optics using the InceptionResNetV2-LSTM network with the measurements from the differential wave-front sensor. (a) Michelson differential yaw, GOF = 67.39; (b) Michelson differential pitch, GOF = 68.10; (c) signal-recycling yaw, GOF = 68.75; (d) signal-recycling pitch, GOF = 69.32; (e) power-recycling yaw, GOF = 51.41; (f) power-recycling pitch, GOF = 63.44.

Keeping such drifts under control helps to reduce the weekly retuning interventions and to increase the chances of acquiring a lock after a lockless event. Lowering the time needed to fine tune the detector periodically leads to more productive use of the commissioner's time and is especially useful when the detector is operated with limited personnel. The technique described above works well without the need for retraining as long as the thermal state of the interferometer remains unchanged. Increasing circulating power inside the arms, for example, would require retuning of the agent's policy. One possibility to overcome this issue is to use different agents for each discrete scenario. Assessing the long-term impact of the deployed infrastructure and exploring the utility of adaptive RL strategies that can internally handle changing system dynamics would be part of future work.

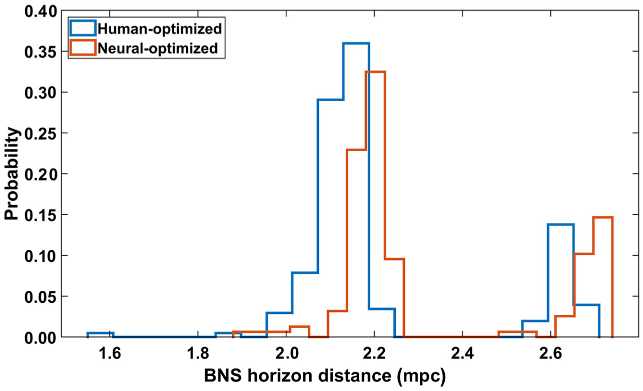

FIG. 7. An astrophysical-sensitivity comparison in terms of horizon distance for coalescing neutron-star binaries of 1.4 solar mass each. The blue trace is when experienced commissioners fine tune the interferometer, while the orange trace provides the results using the deep RL agent. The two peaks represent the scenario with and without applying squeezed states of light.

VI. CONCLUSIONS AND OUTLOOK

With GW detectors becoming more complex with each generation, AI-assisted autonomous sensing and control could play a major role in the operation of interferometers, including automated alignment and multicavity locking. We have developed a deep NN scheme to extract meaningful information about the state of the interferometer and we have reconstructed the alignment error signals using the data from the GEO 600 observatory. We have implemented a control loop using this neural sensor and achieved drift control of the signal-recycling mirror using deep RL, improving overall sensitivity. In this work, machine-learning-based control applied to a kilometer-scale GW interferometer has improved the astrophysical sensitivity. Existing DWS methods use rf sidebands to generate the error signals, which often elevate the photon shot noise at kilohertz frequencies, affecting the GW-sensitive strain signal. An interesting possibility for future work is to fully replace existing DWS-based autoalignment methods with a higher-bandwidth version of the neural scheme presented here.

We have followed a divide-and-conquer approach, deploying different neural architectures and multiple learning strategies for sensing and control. End-to-end learning using a single transformer-based architecture with self-attention [48] could lead to a better flow of gradients and improved predictions. Expanding the RL controller's policy to include a diverse set of tasks would also be desirable if we intend to control multiple subsystems. DeepMind's Gato [49] and Robotics Transformer (RT-1) from Google Brain [50] have recently demonstrated the most promising strides toward artificial general intelligence, enabling multitask learning using a context-based generalized policy. The applicability of such frameworks that combine transformer models with RL strategies looks promising for current and future-generation GW observatories and our work is hopefully the first step in that direction.

VII. TRAINING RESOURCES

The CNN-LSTM neural sensor was built and trained using MATLAB R2022a, while the RL-SAC multiple-input–multiple-output (MIMO) controller was set up and trained using the corresponding Simulink modeling environment. The GPU training was carried out at the Caltech LIGO cluster (AMD EPYC 7763 64-Core, 256-GB RAM) using the NVIDIA A100-80 GB GPU. The NN quantization from single-bit floating point to INT8 data type was carried out for SIL and HIL mode, respectively, using the Intel-MKL deep-learning library and the NVIDIA Jetson Xavier NX.

This paper has LIGO Document No. LIGO-P2200407.

ACKNOWLEDGMENTS

N.M. thanks Thomas Künzel, Franziska Albers, and MathWorks for their technical support. N.M. thanks Kong Chun-Wei, University of Michigan, for his insights about training deep RL controllers. We thank Lisa Barsotti and the LIGO control systems working group for their valuable suggestions. We are grateful for the computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. We thank the GEO collaboration for the development and construction of GEO 600. We also extend our thanks to Walter Graß for his work keeping the interferometer in a good running state. We are grateful for support from the Science and Technology Facilities Council (STFC), the University of Glasgow in the United Kingdom, the Max Planck Society, the Bundesministerium für Bildung und Forschung (BMBF), the Volkswagen Stiftung, the cluster of excellence QUEST (Centre for Quantum Engineering and Space-Time Research), the International Max Planck Research School (IMPRS), and the State of Niedersachsen in Germany.

[1] H. Grote, A. Freise, M. Malec, G. Heinzel, B. Willke, H. Lück, K. A. Strain, J. Hough, and K. Danzmann, Dual recycling for GEO 600, Classical Quantum Gravity **21**, S473 (2004).

[2] H. Lueck, C. Affeldt, J. Degallaix, A. Freise, H. Grote, M. Hewitson, S. Hild, J. Leong, M. Prijatelj, K. A. Strain, B. Willke, H. Wittel, and K. Danzmann, The upgrade of GEO 600, J. Phys. Conf. Ser. **228**, 012012 (2010).

[3] K. L. Dooley, *et al.*, GEO 600 and the GEO-HF upgrade program: Successes and challenges, Classical Quantum Gravity **33**, 075009 (2016).

[4] H. Grote and (for the LIGO Scientific Collaboration), The GEO 600 status, Classical Quantum Gravity **27**, 084003 (2010).

[5] T. Akutsu, *et al.*, Overview of KAGRA: Calibration, detector characterization, physical environmental monitors, and the geophysics interferometer, Prog. Theor. Exp. Phys. **2021**, 05A102 (2021).

[6] K. Somiya, Detector configuration of KAGRA—the Japanese cryogenic gravitational-wave detector, Classical Quantum Gravity **29**, 124007 (2012).

[7] The LIGO Scientific Collaboration, The Virgo Collaboration, The KAGRA Collaboration, R. Abbott, H. Abe, F. Acernese, K. Ackley, N. Adhikari, R. Adhikari, V. Adkins, *et al.*, First joint observation by the underground gravitational-wave detector KAGRA with GEO 600, Prog. Theor. Exp. Phys. **2022**, 063F01 (2022).

[8] C. Affeldt, *et al.*, Advanced techniques in GEO 600, Classical Quantum Gravity **31**, 224002 (2014).

[9] The LIGO Scientific Collaboration, J. Aasi, *et al.*, Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[10] D. V. Martynov, E. D. Hall, B. P. Abbott, R. Abbott, T. D. Abbott, C. Adams, R. X. Adhikari, R. A. Anderson, S. B. Anderson, K. Arai, *et al.*, Sensitivity of the Advanced LIGO detectors at the beginning of gravitational wave astronomy, Phys. Rev. D **93**, 112004 (2016).

[11] F. Acernese, *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2014).

[12] H. Grote, K. Danzmann, K. L. Dooley, R. Schnabel, J. Slutsky, and H. Vahlbruch, First long-term application of squeezed states of light in a gravitational-wave observatory, Phys. Rev. Lett. **110**, 181101 (2013).

[13] J. Lough, E. Schreiber, F. Bergamin, H. Grote, M. Mehmet, H. Vahlbruch, C. Affeldt, M. Brinkmann, A. Bisht, V. Kringel, H. Lück, N. Mukund, S. Nadji, B. Sorazu, K. Strain, M. Weinert, and K. Danzmann, First demonstration of 6 dB quantum noise reduction in a kilometer scale gravitational wave observatory, Phys. Rev. Lett. **126**, 041102 (2021).

[14] B. J. Meers, Recycling in laser-interferometric gravitational-wave detectors, Phys. Rev. D **38**, 2317 (1988).

[15] E. Morrison, B. J. Meers, D. I. Robertson, and H. Ward, Automatic alignment of optical interferometers, Appl. Opt. **33**, 5041 (1994).

[16] E. Morrison, B. J. Meers, D. I. Robertson, and H. Ward, Experimental demonstration of an automatic alignment system for optical interferometers, Appl. Opt. **33**, 5037 (1994).

[17] A. Bisht, M. Prijatelj, J. Leong, E. Schreiber, C. Affeldt, M. Brinkmann, S. Doravari, H. Grote, V. Kringel, J. Lough, *et al.*, Modulated differential wavefront sensing: alignment scheme for beams with large higher order mode content, Galaxies **8**, 81 (2020).

[18] R. W. P. Drever, J. L. Hall, F. V. Kowalski, J. Hough, G. M. Ford, A. J. Munley, and H. Ward, Laser phase and frequency stabilization using an optical resonator, Appl. Phys. B **31**, 97 (1983).

[19] H. Grote, Ph.D. thesis, Hannover Universität, 2003.

[20] N. Mukund, J. Lough, C. Affeldt, F. Bergamin, A. Bisht, M. Brinkmann, V. Kringel, H. Lück, S. Nadji, M. Weinert, and K. Danzmann, Bilinear noise subtraction at the GEO 600 observatory, Phys. Rev. D **101**, 102006 (2020).

[21] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[22] Y. Wang, in *2017 American Control Conference (ACC)*, (IEEE, Seattle, WA, USA, 2017), p. 5324.

[23] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. **39**, 677 (2017).

[24] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, South Brisbane, QLD, Australia, 2015), p. 4580.

[25] S. Bozinovski, Reminder of the first paper on transfer learning in neural networks, 1976, Informatica **44**, 291 (2020).

[26] L. Y. Pratt, J. Mostow, and C. A. Kamm, in *Proceedings of the AAAI Conference on Artificial Intelligence, 9*, (AAAI Press, Anaheim, California, 1991) p. 584.

[27] S. Jadhav, N. Mukund, B. Gadre, S. Mitra, and S. Abraham, Improving significance of binary black hole mergers in Advanced LIGO data using deep learning: Confirmation of GW151216, Phys. Rev. D **104**, 064051 (2021).

[28] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, SqueezeNet: Alexnet-level accuracy with 50× fewer parameters and ¡ 0.5 MB model size, ArXiv:1602.07360 (2016).

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, Boston, MA, USA, 2015), p. 1.

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, in *Thirty-First AAAI Conference on Artificial Intelligence*, (AAAI Press, San Francisco, California, USA, 2017).

[31] M. Courbariaux, Y. Bengio, and J.-P. David, Training deep neural networks with low precision multiplications, ArXiv:1412.7024 (2014).

[32] S. Han, H. Mao, and W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding, ArXiv:1510.00149 (2015).

[33] P. Van Overschee and B. De Moor, N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, Automatica **30**, 75 (1994).

[34] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory* (SIAM, Philadelphia, PA, 1994).

[35] L. Ljung, in *Signal Analysis and Prediction*, (Springer, Boston, MA, 1998), p. 163.

[36] G. F. Franklin, J. D. Powell, A. Emami-Naeini, and J. D. Powell, *Feedback Control of Dynamic Systems* (Prentice Hall, Upper Saddle River, New Jersey, 2002), Vol. 4.

[37] N. Bruinsma and M. Steinbuch, A fast algorithm to compute the H_∞-norm of a transfer function matrix, Syst. Control Lett. **14**, 287 (1990).

[38] P. Apkarian and D. Noll, Nonsmooth H-infinity synthesis, IEEE Trans. Automat. Contr. **51**, 71 (2006).

[39] K. J. Åström and T. Hägglund, *Advanced PID Control* (ISA—The Instrumentation Systems and Automation Society, 2006). https://lup.lub.lu.se/record/535630.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, USA, 2018).

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80, edited by J. Dy and A. Krause (PMLR, Stockholm, Sweden, 2018), p. 1861. https://proceedings.mlr.press/v80/haarnoja18b.html.

[42] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, Adv. Neural Inf. Process. Syst. **12**, 1057 (1999).

[43] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, in *International Conference on Machine Learning*, (PMLR, New York, NY, USA, 2016), p. 1928.

[44] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, High-dimensional continuous control using generalized advantage estimation, ArXiv:1506.02438 (2015).

[45] S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, and S. Levine, Q-prop: Sample-efficient policy gradient with an off-policy critic, ArXiv:1611.02247 (2016).

[46] K. Ota, D. K. Jha, and A. Kanezaki, Training larger networks for deep reinforcement learning, ArXiv:2102.07920 (2021).

[47] M. A. Wiering and H. Van Hasselt, Ensemble algorithms in reinforcement learning, IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **38**, 930 (2008).

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2017), p. 6000.

[49] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, *et al.*, A generalist agent, ArXiv:2205.06175 (2022).

[50] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, Rt-1: Robotics transformer for real-world control at scale, ArXiv:2212.06817 (2022).