First demonstration of neural sensing and control in a kilometer-scale gravitational wave observatory

N. Mukund,[1] 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] K. Danzmann ,[1]
[1]Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut) and Institut für Gravitationsphysik,
Leibniz Universität Hannover, Callinstraße 38, 30167 Hannover, Germany

(Dated: January 18, 2023)

Suspended optics in gravitational wave (GW) observatories are susceptible to alignment perturbations and, in particular, to 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 circulating power and optomechanical photon squeezing, and thus decrease the astrophysical sensitivity to merging binaries. Traditional alignment techniques involve differential wavefront sensing using multiple quadrant photodiodes, but are often restricted in bandwidth and are limited by the sensing noise. We present the first-ever successful implementation of neural network-based sensing and control at a gravitational wave 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 CNN-LSTM network architecture and is then used for MIMO control using soft actor-critic-based deep reinforcement learning. Overall sensitivity improvement achieved using our scheme demonstrates deep learning's capabilities as a viable tool for real-time sensing and control for current and next-generation GW interferometers.

GEO 600 is a dual recycled advanced Michelson interferometer (IFO) with folder arms [1–3], located near Hannover, Germany. With a peak strain sensitivity of about $10^{-22}/\sqrt{Hz}$ 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 the KAGRA detector [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 squeezing [12], and the demonstration 6 dB of measured optical squeezing [13]. Recent studies [14] have also demonstrated its unique capabilities to conduct searches for a specific class of axionic dark matter candidates. In this work, we present another novel 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. I describes the existing alignment scheme and its limitations, Sec. II explains the motivations and architecture of the neural sensing, Sec. III describes the implementation of a deep reinforcement learning (RL) based controller and Sec. IV provides the network predictions and improvements to the sensitivity when the trained controller is deployed for low frequency signal recycling alignment control.

I. 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 in the readout channel [15]. PRC at GEO 600 consists of the PR mirror, located at the IFO's input port, 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 the photon shot noise. Similarly, the SRC is formed by the SR mirror, situated at the IFO's output port, and the Michelson IFO (shown in Figure 7). It complements the PRC by forming a resonant cavity to enhance the signal sidebands from potential GWs. The SR mirror's microscopic position also determines the cavity's overall frequency response. 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. However, the noise suppression is achieved only above the pendulum's resonance frequency, which is close to 1 Hz. While this isolation is sufficient within the gravitational wave 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. Sub-optimal alignment of the incident beam to the OMC leads to intensity fluctuations in the photodiode signal that degrades the overall optical gain and increases the number of glitches that often mimic the true GW signal. Such misalignments also routinely interfere with the suite of optical squeezing and thermal compensation experiments carried out at GEO. Automatic alignment systems are hence critical to attain optimal sensitivity and maintain long lock stretches at the observatory.

The goal of the auto-alignment 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 wavefront sensing technique (DWS) [16, 17] and becomes active once the cavities are 'locked' in length using the PDH 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 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 keep the spot centered.

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 having 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 by optimal layout design. Despite the auto-alignment 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 results in the creation of radio frequency (RF) sidebands which are tens of MHz offset with respect to the laser(or main carrier) frequency. Although the OMC suppresses these MHz 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, environmental events like 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, particularly 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. It involves mechanically oscillating the relevant optics at a specific frequency for each degree of freedom. 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 to align, for example, 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 wavefront sensing is currently under commissioning for the OMC alignment [21]. The dithering also enhances the bilinear coupling to strain if the beam is not well centered on the optic. All these reasons motivate the need for a better solution.

II. NEURAL SENSING

A. Why darkport is a good witness

The south port of the IFO referred to as the darkport, is usually kept close to destructive interference but with a slight DC offset of about 5-50 picometers. This offset allows about six mW of carrier to leak out and about 30 mW of higher-order modes to exit via the darkport. The higher-order modes originate inside the IFO due to mismatch in the interfering beams, which in turn are 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 darkport (DP) beam contain much information about the IFO state. It is used for manual pre-alignment that makes the longitudinal lock acquisition easier, and then the wavefront sensor-based auto-alignment 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, 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. Timescales of these disturbances range from hundreds of milliseconds to a few days and include sources like temperature variations, seismic disturbances, and optomechanical intra-cavity cross-couplings. However, since there exists a one-to-one mapping between the state of the interferometer and the darkport image (see Figure 1), such disturbances are encoded in their breathing patterns. Just before the weekend break, these loop offsets are tuned by commissioners via visual inspection of the camera images. The values are considered optimized when the darkport image resembles a stable state, often based on recollections from memory.

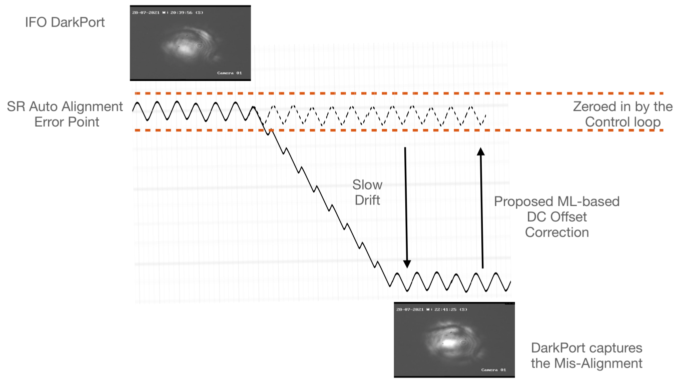


FIG. 1: Representation of the one-to-one mapping between the darkport image and misalignment of the IFO. Existing sensors are unable to detect the DC drift in the alignment control signal, since the feedback loop keeps the error signal zeroed around the setpoint. However, an ML model that infers the actual mirror positions from the darkport images can be used to correct for such drifts.

B. Coherence Mapping

Figure 2 reveals the complex nature in which the multiple optics imprint their state of alignment on the interferometric darkport. The map is constructed by measuring coherence between the pixel-wise darkport intensity fluctuations and the temporal variation in different alignment error signals. $\mathrm{C_{xy}}$, the metric used for computing the coupling is the magnitude-squared coherence in the 0.1-4 Hz band weighted by the average logarithmic error-point spectra,

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

Apart from being a useful detector characterization tool, CoherenceMaps can be used to identify potential signals that a trained neural network can recover. The high coherence and peculiar spatial spread confirm that darkport contains a treasure trove of information about the IFO. 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. For Signal and Power recycling optics, we see more complex and radially extended structures, which could indicate higher-order spatial modes. For all three optics, coherence from pitch seems to be higher than the yaw degree of freedom. This difference is expected since the angle-to-length coupling in suspended optics is more likely to happen via tilt or pitch in comparison to the yaw motion.

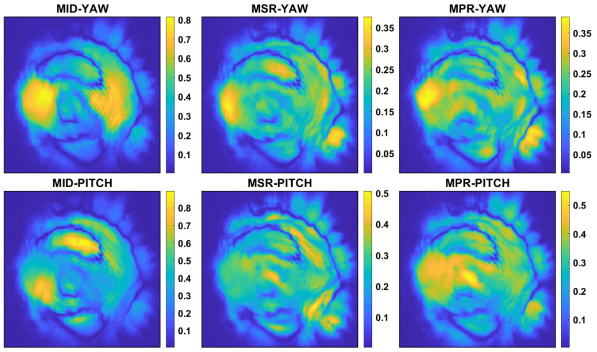


FIG. 2: CoherenceMaps show the complex coupling of critical alignment degrees of freedom to the interferometric darkport. Coupling is estimated using the weighted coherence in the 0.1-4Hz band. The images from left to right depict the Michelson differential arm motion and the motion from the signal recycling and power recycling mirrors.

C. Neural Sensor Architecture

The neural alignment sensing we intend to do can be formulated as an image-to-time-series regression problem. We choose a CNN-LSTM architecture for this task for a few reasons. It is interesting to note that convolutional networks are analogous to non-linear FIR filters, while the recurrent networks behave similarly to non-linear IIR filters [22]. While 2-D convolutional neural nets are well suited for analyzing image data with complex spatial representation and for object detection, long short-term memory networks [23] excel at temporal modeling and sequence prediction. LSTMs, a specific form of recurrent neural networks, make use of a memory cell that selectively controls the flow of information using input, output, and forgets gates. They also had limited success at linear system identification tasks, with results comparable to traditional transfer function estimation [24]. The ability to learn representations in both space and time thus make the combined deep recurrent convolutional models effective at activity recognition from streaming video data and make them a good candidate

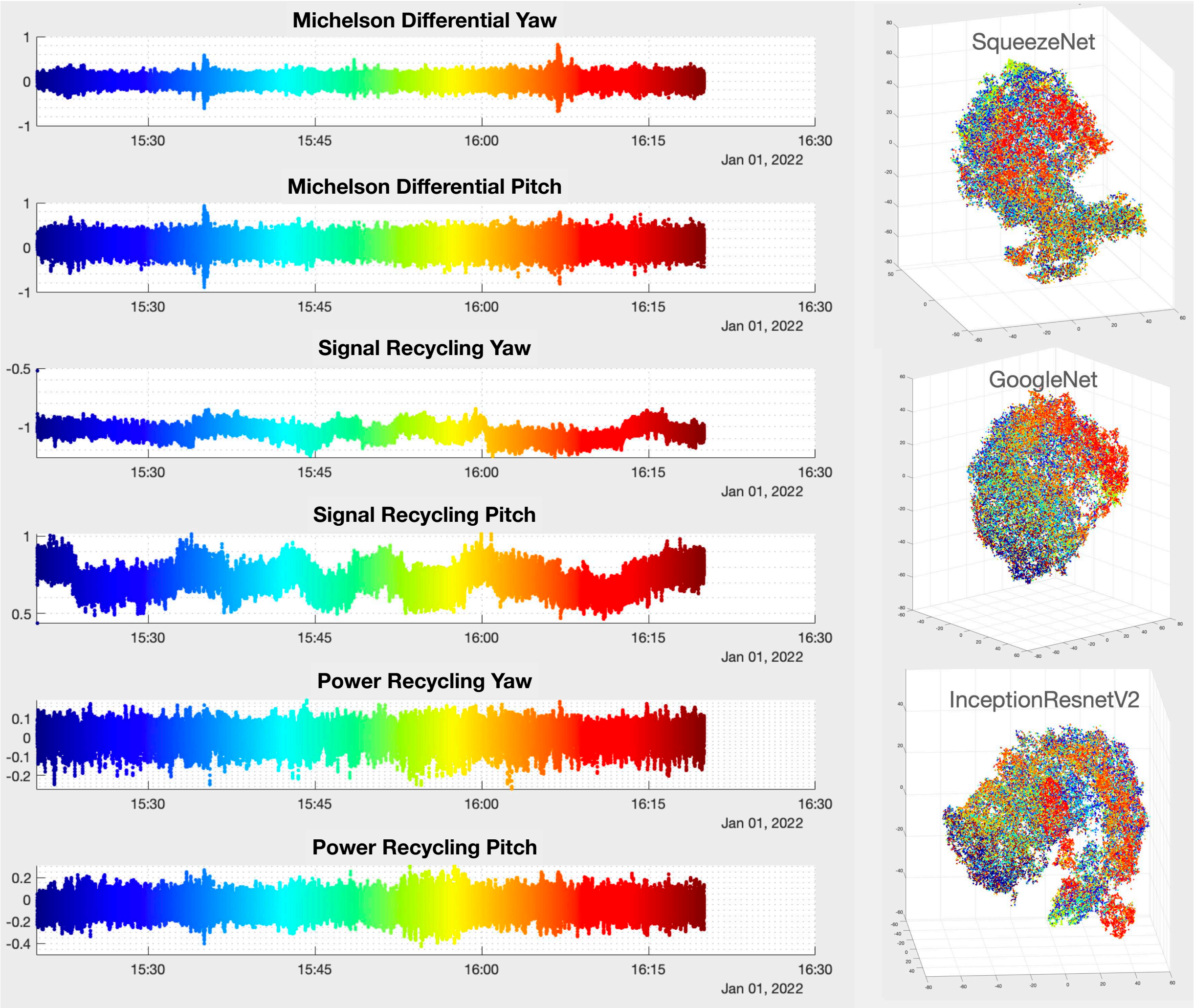


FIG. 3: Time-dependent variation in the one-to-one mapping between the six critical alignment error signals and video frames captured from the interferometric darkport. The higher dimensional features extracted from these darkport images using the pre-trained CNN layers are embedded in three dimensions using the t-stochastic neighborhood embedding technique.

for capturing the underlying system dynamics [25, 26].

Our design choice for the CNN architecture is based on transfer learning [27, 28] where the initial layers of pre-trained networks, fine-tuned to extract spatial information at different scales by training on standardized datasets, are reused for a newer task. Transfer learning alleviates the need for training networks from scratch and is useful when the data is 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 been successful in recovering all the GW events listed in the GWTC-1 transients catalog [29]. In Fig3, we assess their feature extraction capability and compare the response to the temporal evolution of darkport images, primarily influenced by key six alignment degrees of freedom. The extracted features are visualized using the t-SNE (t-distributed Stochastic Neighbor Embedding) [30] technique, which embeds the higher dimensional feature vector to a low-dimensional one while preserving the relative distances between those vectors. We use three reference architectures, namely squeezenet [31], googlenet [32], and inceptionResnetV2 [33] and compare the respective trade-offs. Squeezenet is one of the lightest available networks with 18 layers making it suitable for embedded devices and low latency inference. The 164 layers-deep inceptionResnetV2 is among the largest pre-trained networks and provides high classification accuracy on several benchmark datasets. The inclusion of skip connections [34], in addition, makes these less vulnerable to the typical vanishing and exploding gradient issue of deep networks. Googlenet is often a good choice when we require a balance between network size and accuracy.

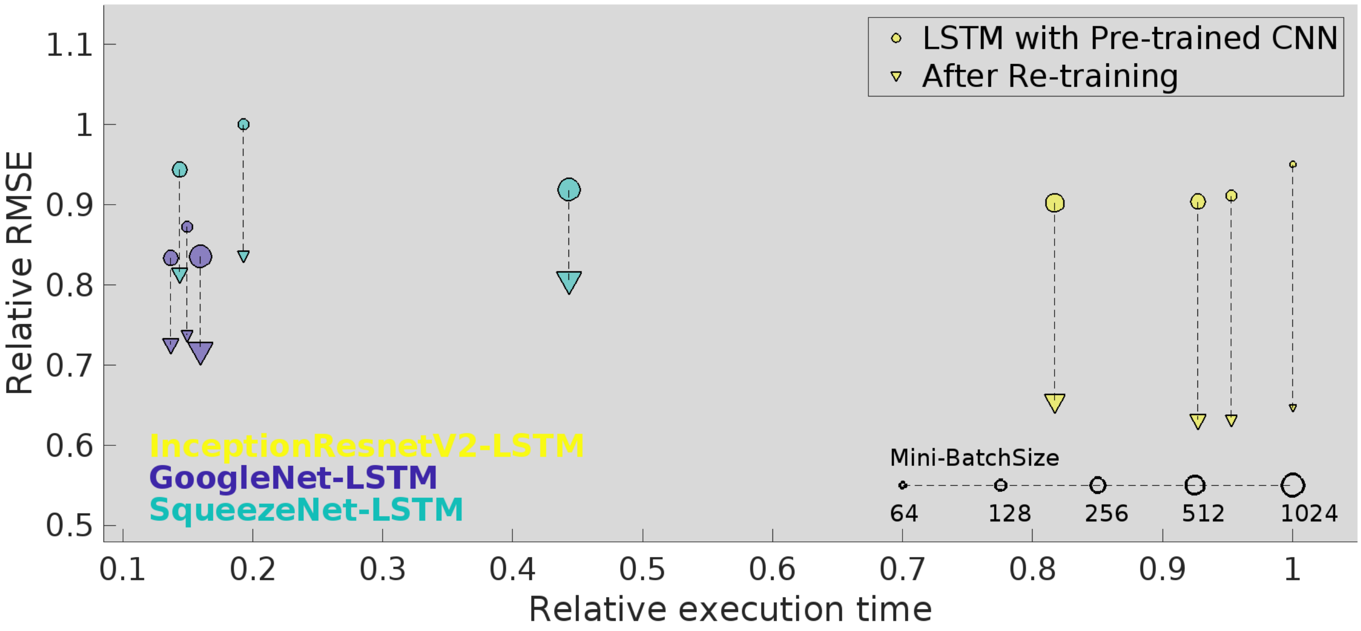


FIG. 4: The following plot assesses the performance of different CNN-LSTM networks across different metrics. These include relative root mean square error, execution time, LSTM with pre-trained CNN vs. results after retraining the whole CNN-LSTM, and the effect of the mini-batch size of data used in training.

D. Training Strategy

Our aim with the CNN-LSTM model is to train the network on sufficient darkport 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 (PID or RL agent) can set it to the last known "good" state, also determined by the model. The corresponding schematic is given in Figure 7. Training deep networks is, in general, a time-consuming process, and additionally, we need also to find the right hyper-parameters 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 to its output. We then freeze the weights of squeezenet layers and let the LSTM layers learn while the combined network is trained to predict the alignment error points from the recorded darkport images. This configuration makes it easier to determine parameters like gradient decay rate, LSTM hidden units, and learn rate using minimal computational resources. In the second stage of training, we retrain the entire network comprising the pre-trained CNN and newly trained LSTM and fine-tune the network weights and biases. This stage of training is carried out on a dedicated A100 GPU cluster. We repeat the process for the other two pre-trained networks. The corresponding results are shown in Fig. 4. A larger mini-batch size generally increases the training time and GPU memory load but results in lower RMSE. We see certain timing trend violations; this could arise from the same GPU running multiple jobs. We select Googlenet-LSTM for the rest of our analysis as it provides a decent trade-off among metrics like time for training and inference, prediction accuracy, and the real-time inference rate.

E. Network Quantization

Most often, the learnable parameters of neural networks 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 [35, 36]. This step would be crucial when the trained networks are deployed at a large scale in GW detectors using embedded devices like FPGAs, 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 dataset, we quantize to the right data type (single or INT8), ensuring to cover the range, avoiding overflows but ignoring potential underflows. Figure 5 gives the memory reduction and the improved processing speed, measured in terms of frames per second, and the relative decrease in accuracy for the three quantized network architectures. Accuracy is given by $(1-\text{RMSE}_{\text{quantized}})/(1-\text{RMSE}_{\text{original}})$, where RMSE

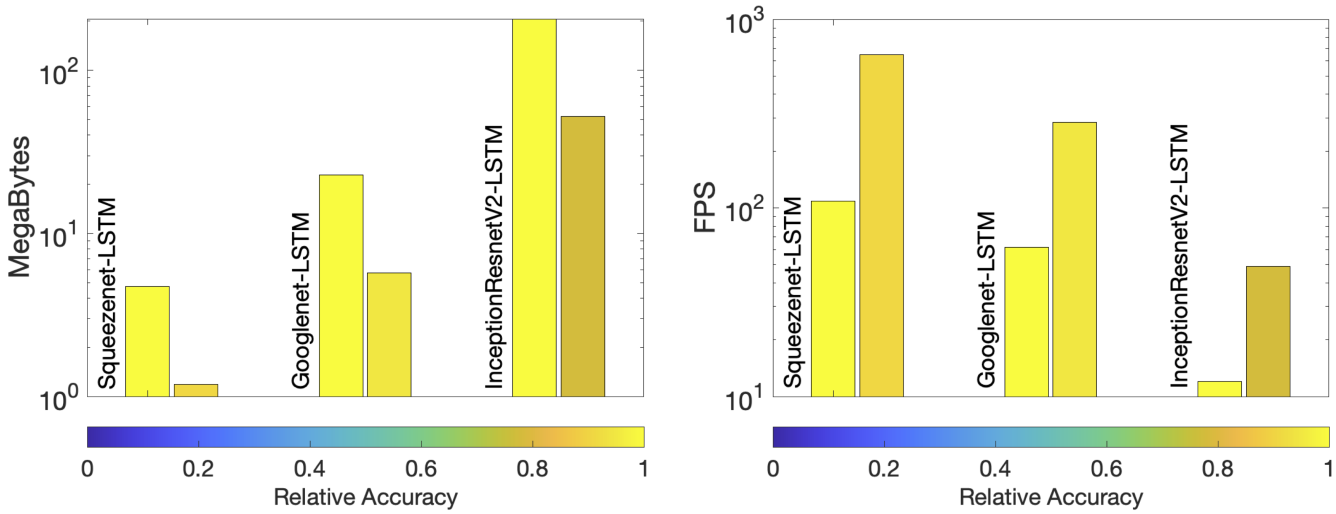


FIG. 5: Figure to the left shows the reduction in memory footprint by quantizing single-precision floating point data types to scaled 8-bit integer data types. The second plot gives the corresponding increase in the achievable inference speed in software-in-loop mode, measured in terms of frames per second. Colorbar indicates the relative accuracy in comparison to the non-quantized network.

is the root mean squared error between the actual error points and the corresponding network predictions and is normalized to the original non-quantized network.

III. MODEL BASED CONTROLLER DESIGN

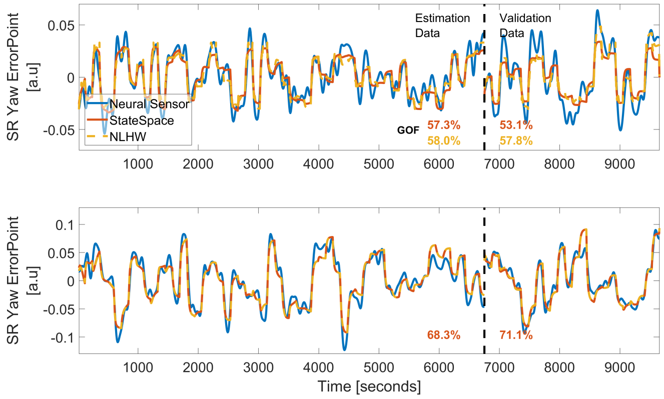


FIG. 6: Comparison of neural sensor inferred system dynamics with reduced order models. The identified statespace model is used to design the optimal PID controller and train the reinforcement learning agent.

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 like oscillations in the system that could take a long time to settle down. When controllers have a lot of parameters, the optimization process can add more delay to the design process. For example, an RL agent can intentionally carry out random action sequences to find the optimal policy and strike a balance between exploration and exploitation. We hence follow a model-based [37] 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 setpoints in SR pitch and yaw degree of freedom that covers the actuation range of the existing controller. This perturbation leads to variation in the darkport images and is processed by the CNN-LSTM neural sensor. System identification, mapping setpoints to neural predictions, is then carried out using subspace-based state-space modeling [38]. State space offers superior performance over transfer function models due to the ability to also 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 [39] is selected. Selecting such lower values helps retain the larger energy states, making it possible to have a reduced-order model that preserves the majority of the 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 model's predicted output is minimized [40]. We looked at further improvements by adding input and output non-linearities to the identified statespace model, resulting in a non-linear Hammerstein-Wiener (NLHW) model. Fig. 6 compares both the models on estimation and validation data, where the goodness of fit is given as,

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

The NLHW model, with a sigmoid network function representing the non-linear mapping, only provides a modest improvement in one degree of freedom. Hence, we select the statespace model for the rest of the analysis.

A. PID Controller

Proportional-Integral-Derivative (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 [41] and also serve as a benchmark while evaluating the performance of the RL based solutions described in the next section. 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 [42, 43]. The presence of an actuator with a limited range however introduces non-linearities and often leads to the well-known integral windup [44]. We overcome this using additional anti-windup 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. 7, is given by,

$$\begin{aligned} u(t) &= \mathrm{K_p}\, e(t) \,+\, \mathrm{K_d}\,\frac{d\,e(t)}{dt} \,+\, \mathrm{K_r}\, r(t) \\ &+ \int \left[\mathrm{K_i}\, e(t) \,+\, \mathrm{K_t}\, u_s(t) \,-\, \mathrm{K_t}\,\mathrm{K_r}\, r(t) \,-\, \mathrm{K_t}\, u(t)\right] dt. \end{aligned} \tag{3}$$

where $\mathrm{K_p}$, $\mathrm{K_i}$, $\mathrm{K_d}$ are the usual PID gain coefficients, $\mathrm{K_r}$ controls the reference r(t) feedforward, while the tracking coefficient $\mathrm{K_t}$ and the saturated output $\mathrm{u_s}$ are part of the modified integral term.

B. Deep Reinforced Controller

Reinforcement learning (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 its interaction with the system [45]. The notion of a traditional controller is replaced here by an RL agent consisting of a deep neural network 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 system's current state, 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 aims is 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 [46] that utilizes the traditional actor-critic methods [47–50] 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,

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

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,

$$\max_{\pi} \mathbb{E}_{\pi}\left[\mathrm{Q(S,A)} - \log \pi(\mathrm{A|S})\right], \tag{5}$$

the agent learns a wide range of behaviors, including stochastic or deterministic behaviors. A comparatively faster learning rate, lower sensitivity to hyperparameters, ability to reuse fast 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

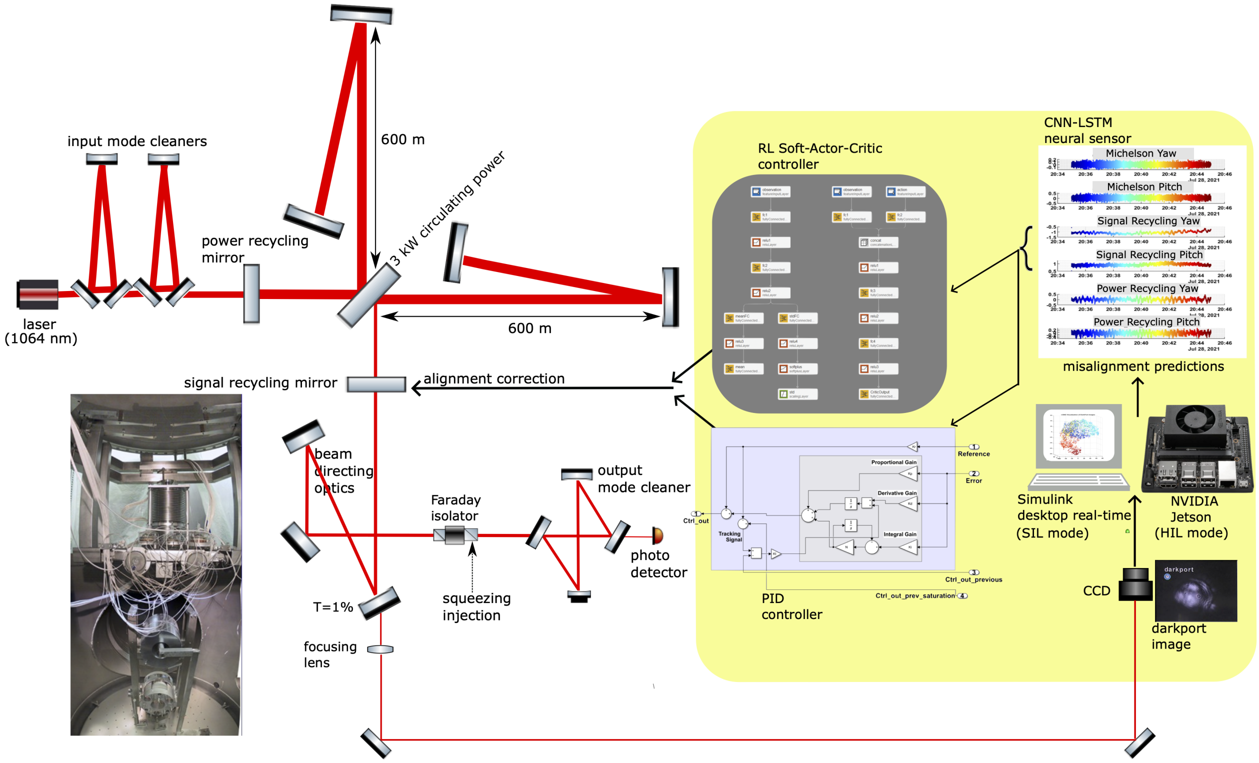


FIG. 7: Simplified optical layout of GEO 600 highlighting the AI-based alignment sensing and control scheme. The CCD captures 2D images of the beam that exit the darkport through the 1% transmission port of the beam-directing optic. The CNN-LSTM neural network simultaneously extracts the pitch and yaw degrees of freedom for the Michelson, signal recycling, and power recycling mirrors. Tuned PID controllers and soft-actor-critic-based reinforcement learning agents process this information and correct the low-frequency drifts of the signal recycling mirror, thus improving the astrophysical sensitivity.

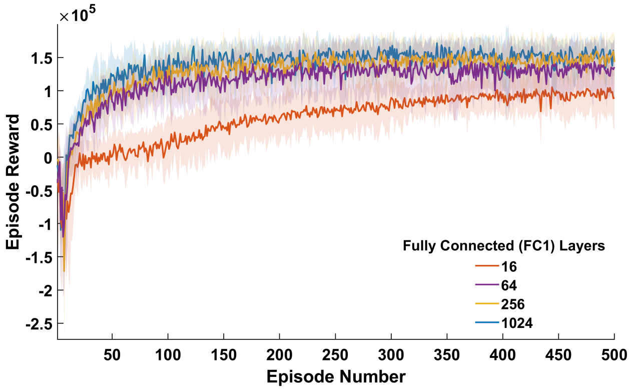
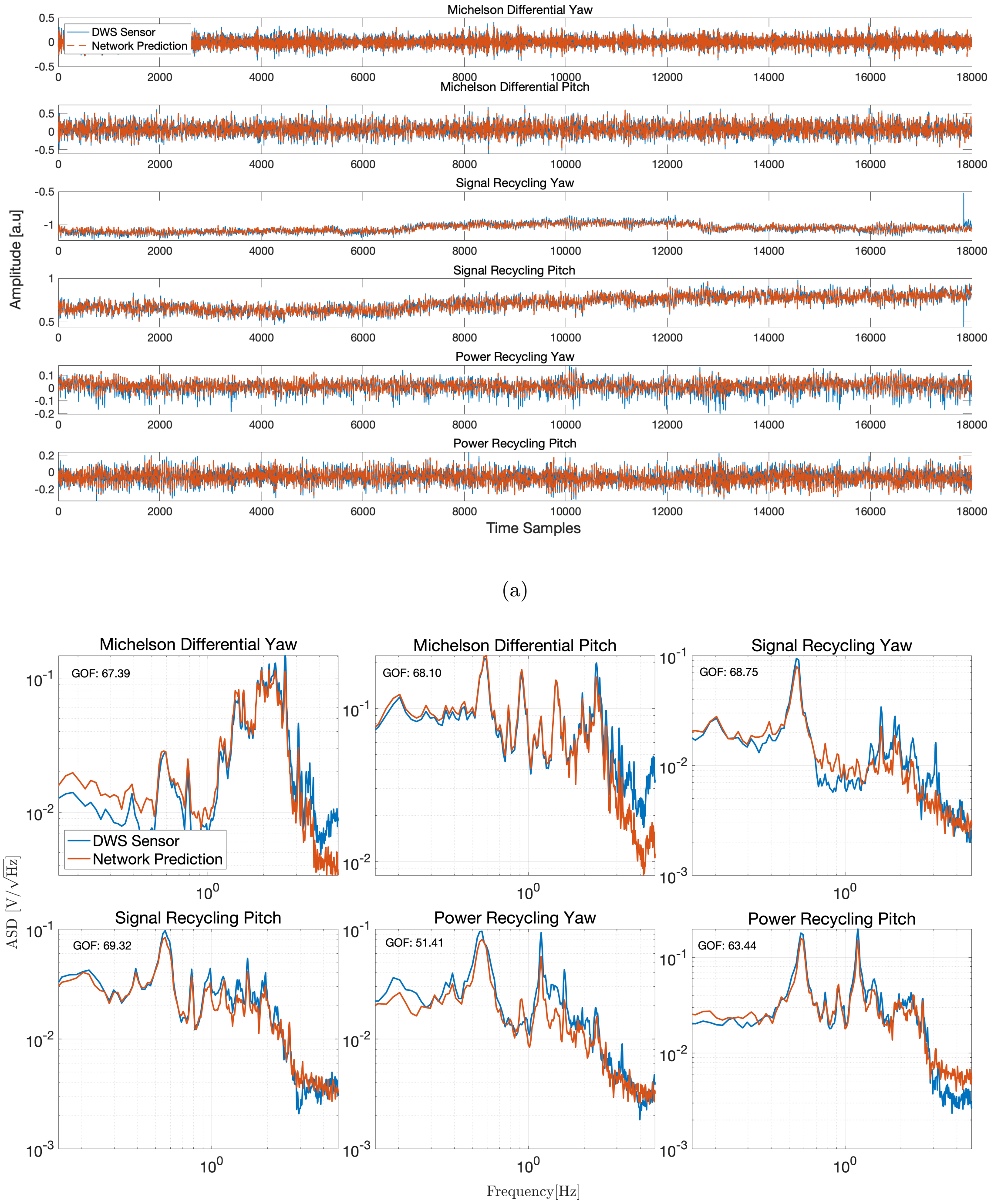


FIG. 8: Average episode reward received for various soft-actor-critic RL agent configurations. The shaded region gives the associated standard deviation errors obtained from ten independent trials.

this work is to assess the practicality of this approach in designing controllers suitable for GW detectors and probe if a set of general guiding principles can help design a reward that leads the agent to the optimal policy. We can 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, 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 states $(\mathrm{S_j}, \mathrm{S_j^{ref}})$, and the current and previous actuator values $(\mathrm{A_j}, \mathrm{A_j^{prev}})$, with $\mathbb{Q}_j$ and $\mathbb{R}_j$ being the respective weight matrices,

$$\mathrm{Cost} = \sum_{j=1}^{\tau} (\mathrm{S_j} - \mathrm{S_j^{ref}})^{\mathrm{T}}\, \mathbb{Q}_\mathrm{j}\, (\mathrm{S_j} - \mathrm{S_j^{ref}}) + \ (\mathrm{A_j} - \mathrm{A_j^{prev}})^{\mathrm{T}}\, \mathbb{R}_\mathrm{j}\, (\mathrm{A_j} - \mathrm{A_j^{prev}}) . \quad (6)$$

Such continuous rewards encourage convergence, but



(a)

(b)

FIG. 9: Comparison of neural network alignment predictions for the key optics in time and frequency domain with the measurements from the differential wavefront sensor.

are prone to local minima and can lead to longer training periods. Adding discrete elements that reward or penalize the agent often increases the probability of finding better states, but the non-smooth nature of the resulting loss function can potentially affect the convergence. We discourage boundary constraint violations from the agent by including a discrete penalty term,

$$\mathrm{Penalty} = \mathrm{W_y}\left((\mathrm{S_j}-\mathrm{S^{min}})^2+(\mathrm{S_k}-\mathrm{S^{max}})^2\right)$$
$$+\,\mathrm{W_{mvrate}}\left((\dot{\mathrm{A}}_\mathrm{l}-\dot{\mathrm{A}}^\mathrm{max})^2+(\dot{\mathrm{A}}_\mathrm{m}-\dot{\mathrm{A}}^\mathrm{min})^2\right). \quad (7)$$

$$\forall\left(\mathrm{S_j}<\mathrm{S^{min}},\ \mathrm{S_k}>\mathrm{S^{max}},\ \dot{\mathrm{A}}_\mathrm{l}<\dot{\mathrm{A}}^\mathrm{min},\ \dot{\mathrm{A}}_\mathrm{m}>\dot{\mathrm{A}}^\mathrm{max}\right),$$

where behaviors that drive the states close to the limits $(\mathrm{S^{min}},\mathrm{S^{max}})$ or increase controller velocity beyond a threshold value $(\dot{\mathrm{A}}^\mathrm{min},\dot{\mathrm{A}}^\mathrm{max})$ are penalized, with $\mathrm{W_y}$ and $\mathrm{W_{mvrate}}$ being the associated weight matrices. We also observe the benefit from including discrete positive rewards when the state is driven close to the reference. These terms have the following form,

$$\mathrm{Boost} = 10\sum_{\mathrm{j}=1}^{\tau}(3\,|\mathrm{S_j}-\mathrm{S_j^{ref}}|<0.02)^2$$
$$+\;10\sum_{\mathrm{j}=1}^{\tau}(6\,|\mathrm{S_j}-\mathrm{S_j^{ref}}|<0.005)^2\,. \quad (8)$$

The final reward used to train the RL-SAC agent and drive it to the optimal policy is constructed using the above three terms as,

$$\mathrm{Reward} \;=\; -(\mathrm{Cost}\;+\mathrm{Penalty})\;+\;\mathrm{Boost}\,. \quad (9)$$

The random initialization of network weights and 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. But as described in the neural sensor section, we usually overcome it by residual connections. However, in the case of the RL agent, the difficulty in training deeper networks arise from the sharpness of the loss surface curvatures, making them more susceptible to the choice of hyperparameters. Recent studies [51] prefer the wider networks as they have nearly convex loss surfaces, and we indeed observe similar performance improvement with an increase in the network width as shown in Figure 8.

C. Multi-Agent 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 PID and the faster response of RL. Ensemble learning [52] 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 designs in Table I. It includes a tuned PID 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 reward, where the setpoints 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 Figure 10.

PID 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

TABLE I: comparison of different controller architecture combinations.

IV. RESULTS

In Figure 9a, we present the time-domain predictions from the retrained InceptionResnetV2-LSTM purely generated from an hour of darkport images and compare them with the measured error points for the six alignment DOFs. The frequency domain plots (see Figure 9b) reveal the network's ability to capture most spectral features. The intentional extra misalignments introduced for the signal recycling pitch and yaw are well recovered.One way to assess the neural system's performance is to directly check the strain curve or the inferred astrophysical sensitivity. A more robust metric is the interferometric optical gain, whose fluctuations directly impact this sensitivity. Optical gain measures the changes in incident light re-

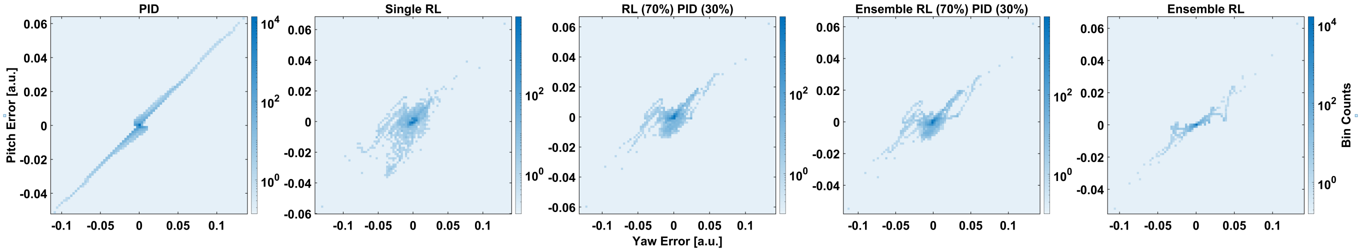


FIG. 10: Comparison of bias and variance in 2-DOF reference tracking for different controller architectures.

ceived on the final DC photodiode per unit change in the differential arm length caused by gravitational waves or terrestrial fluctuations. It is continuously tracked by modeling the optical response of the detector via the injection of calibration lines. Dividing the photodiode signal in the transmission of the output mode cleaner by the optical gain and removing the contribution of the feedback control signals provides the astrophysically relevant strain signal.

Figure 11 compares the optical gain of the IFO under human and ML supervision. In general, the signal recycling mirror introduces frequency dependence to optical gain and the values presented here are those measured below the line width of SRC. The non-stationary nature of the noises influencing the detector, primarily the seismicity, often makes comparing different time segments difficult. To address this, we measure pairs of hour-long segments, each with the SR optimized manually and with the ML in the loop controller. Each pair is then normalized to the manually optimized scenario. For reference, we include specific segments when the IFO is misaligned. Improvements to the optical gain when it is neural-optimized are seen to achieve a performance close to that optimized by experienced commissioners based on their interpretation of the darkport image.

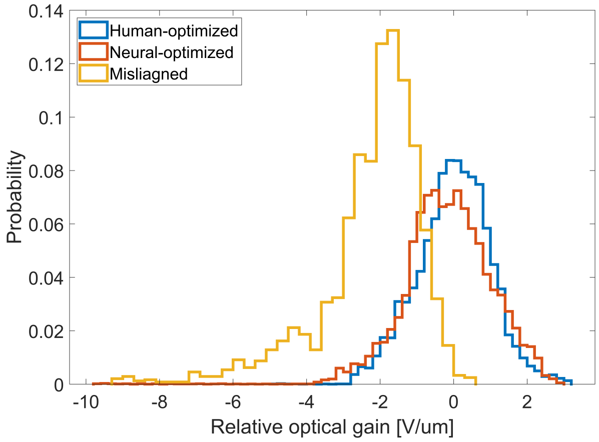


FIG. 11: Optical gain obtained by the neural optimized sensor-actuator compared to one fine-tuned by experienced commissioners. Yellow trace represents a typical misaligned scenario.

V. 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 multi-cavity locking. We developed a deep neural network scheme to extract meaningful information about the state of the interferometer and reconstructed the alignment error signals using the data from GEO 600 observatory. We implemented a control loop that uses this neural sensor and achieved drift control of the signal recycling mirror using deep reinforcement learning, leading to improved overall sensitivity. As far as we know, this work is the first of its kind of control, where machine learning-based control is applied to a kilometer-scale GW interferometer. Radio-frequency sidebands from the existing auto-alignment scheme elevate the shot noise at kilohertz frequencies and affect the strain signal. Fully replacing this method with a higher bandwidth version of the neural scheme presented here is an interesting possibility that is part of future work.

We followed a divide-and-conquer approach, deploying different neural architectures and multiple learning strategies for sensing and actuation. End-to-end learning using a single transformer-based architecture with self-attention [53] 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 sub-systems. DeepMind's Gato [54] and Robotics Transformer (RT-1) from Google Brain [55] have recently demonstrated the most promising strides towards artificial general intelligence, enabling multitask learning using a context-based generalized policy. Applicability of such frameworks that combine transformer models with reinforcement learning strategies indeed looks promising for current and future generation GW observatories, and our work is the first step in that direction.

VI. TRAINING RESOURCES

CNN-LSTM neural sensor was built and trained using MATLAB R2022a, while the RL-SAC MIMO controller were set up and trained using the Simulink modeling environment. 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 neural network 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 NVIDIA Jetson Xavier NX.

VII. ACKNOWLEDGEMENTS

NM thanks Thomas Künzel, Franziska Albers, and MathWorks for their technical support. NM thanks Kong Chun-Wei, University of Michigan, for his insights about training deep RL controllers. We thank the LIGO control systems working group for their valuable suggestions. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and 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 in 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 UK, 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.

VIII. REFERENCES

[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 and Quantum Gravity*, 21(5):S473–S480, feb 2004.

[2] H Lueck et al. The upgrade of GEO 600. *Journal of Physics: Conference Series*, 228(1):012012, may 2010.

[3] K L Dooley, J R Leong, et al. GEO 600 and the GEO-HF upgrade program: successes and challenges. *Classical and Quantum Gravity*, 33(7):075009, mar 2016.

[4] H Grote and (forthe LIGO Scientific Collaboration). The GEO 600 status. *Classical and Quantum Gravity*, 27(8):084003, apr 2010.

[5] T Akutsu et al. Overview of KAGRA: Calibration, detector characterization, physical environmental monitors, and the geophysics interferometer. *Progress of Theoretical and Experimental Physics*, 2021(5), 02 2021. 05A102.

[6] Kentaro Somiya. Detector configuration of KAGRA-the japanese cryogenic gravitational-wave detector. *Classical and Quantum Gravity*, 29(12):124007, jun 2012.

[7] LIGO Scientific Collaboration, Virgo Collaboration, KAGRA Collaboration, R Abbott, H Abe, F Acernese, K Ackley, N Adhikari, RX Adhikari, VK Adkins, et al. First joint observation by the underground gravitational-wave detector KAGRA with GEO 600. *Progress of Theoretical and Experimental Physics*, 2022(6):063F01, 2022.

[8] C Affeldt, K Danzmann, K L Dooley, H Grote, M Hewitson, S Hild, J Hough, J Leong, H Lück, M Prijatelj, S Rowan, A Rüdiger, R Schilling, R Schnabel, E Schreiber, B Sorazu, K A Strain, H Vahlbruch, B Willke, W Winkler, and H Wittel. Advanced techniques in GEO 600. *Classical and Quantum Gravity*, 31(22):224002, nov 2014.

[9] J Aasi, The LIGO Scientific Collaboration, et al. Advanced LIGO. *Classical and Quantum Gravity*, 32(7):074001, mar 2015.

[10] D. V. Martynov, E. D. Hall, et al. Sensitivity of the advanced LIGO detectors at the beginning of gravitational wave astronomy. *Phys. Rev. D*, 93:112004, Jun 2016.

[11] F Acernese et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32(2):024001, dec 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, May 2013.

[13] James Lough, Emil Schreiber, Fabio Bergamin, Hartmut Grote, Moritz Mehmet, Henning Vahlbruch, Christoph Affeldt, Marc Brinkmann, Aparna Bisht, Volker Kringel, Harald Lück, Nikhil Mukund, Severin Nadji, Borja Sorazu, Kenneth Strain, Michael Weinert, and Karsten Danzmann. First demonstration of 6 db quantum noise reduction in a kilometer scale gravitational wave observatory. *Phys. Rev. Lett.*, 126:041102, Jan 2021.

[14] Sander M. Vermeulen, Philip Relton, Hartmut Grote, Vivien Raymond, Christoph Affeldt, Fabio Bergamin, Aparna Bisht, Marc Brinkmann, Karsten Danzmann, Suresh Doravari, Volker Kringel, James Lough, Harald Lück, Moritz Mehmet, Nikhil Mukund, Séverin Nadji, Emil Schreiber, Borja Sorazu, Kenneth A. Strain, Henning Vahlbruch, Michael Weinert, and Benno Willke. Direct limits for scalar field dark matter from a gravitational-wave detector, 2021.

[15] Brian J Meers. Recycling in laser-interferometric gravitational-wave detectors. *Physical Review D*, 38(8):2317, 1988.

[16] Euan Morrison, Brian J. Meers, David I. Robertson, and Henry Ward. Automatic alignment of optical interferometers. *Appl. Opt.*, 33(22):5041–5049, Aug 1994.

[17] Euan Morrison, Brian J. Meers, David I. Robertson, and Henry Ward. Experimental demonstration of an automatic alignment system for optical interferometers. *Appl. Opt.*, 33(22):5037–5040, Aug 1994.

[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. *Applied Physics B*, 31(2):97–105, 1983.

[19] Hartmut Grote. *Making it Work: Second Generation Interferometry in GEO600!* PhD thesis, Hannover U., 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, May 2020.

[21] 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(4):81, 2020.

[22] Boris Kuznetsov, Julian Parker, and Fabian Esqueda. Differentiable iir filters for machine learning applications. 2020.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[24] Yu Wang. A new concept using lstm neural networks for dynamic system identification. In *2017 American Control Conference (ACC)*, pages 5324–5329, 2017.

[25] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[26] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Ha?im Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584, 2015.

[27] Stevo Bozinovski and Ante Fulgosi. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, volume 3, pages 121–126, 1976.

[28] Lorien Y Pratt, Jack Mostow, Candace A Kamm, Ace A Kamm, et al. Direct transfer of learned information among neural networks. In *Aaai*, volume 91, pages 584–589, 1991.

[29] Shreejit Jadhav, Nikhil Mukund, Bhooshan Gadre, Sanjit Mitra, and Sheelu Abraham. Improving significance of binary black hole mergers in advanced LIGO data using deep learning: Confirmation of gw151216. *Phys. Rev. D*, 104:064051, Sep 2021.

[30] t-SNE Matlab. https://www.mathworks.com/help/stats/t-sne.html.

[31] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Training deep neural networks with low precision multiplications. *arXiv preprint arXiv:1412.7024*, 2014.

[36] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[37] Stephen J Kapurch. *NASA systems engineering handbook*. Diane Publishing, 2010.

[38] Peter Van Overschee and Bart De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

[39] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[40] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[41] Gene F Franklin, J David Powell, Abbas Emami-Naeini, and J David Powell. *Feedback control of dynamic systems*, volume 4. Prentice hall Upper Saddle River, 2002.

[42] N.A. Bruinsma and M. Steinbuch. A fast algorithm to compute the h infinity norm of a transfer function matrix. *Systems & Control Letters*, 14(4):287–293, 1990.

[43] P. Apkarian and D. Noll. Nonsmooth h-infinty synthesis. *IEEE Transactions on Automatic Control*, 51(1):71–86, 2006.

[44] Karl Johan Åström, Tore Hägglund, and Karl J Astrom. *Advanced PID control*, volume 461. ISA-The Instrumentation, Systems, and Automation Society Research Triangle Park, 2006.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[46] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[47] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[48] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[49] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[50] Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247*, 2016.

[51] Kei Ota, Devesh K Jha, and Asako Kanezaki. Training larger networks for deep reinforcement learning. *arXiv preprint arXiv:2102.07920*, 2021.

[52] Marco A Wiering and Hado Van Hasselt. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):930–936, 2008.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[55] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.