Interferobot: aligning an optical interferometer by a reinforcement learning agent

Dmitry Sorokin[1,4]
dmitrii.sorokin@phystech.edu

Alexander Ulanov[1]
a.e.ulanov@gmail.com

Ekaterina Sazhina[1,4]
sazhinaekaterina@gmail.com

Alexander Lvovsky[1,2,3]
alex.lvovsky@physics.ox.ac.uk

[1]*Russian Quantum Center, Moscow, Russia* [2]*University of Oxford, United Kingdom*
[3]*P. N. Lebedev Physics Institute, Moscow, Russia* [4]*Moscow Institute of Physics and Technology*

Abstract

Limitations in acquiring training data restrict potential applications of deep rein-
forcement learning (RL) methods to the training of real-world robots. Here we train
an RL agent to align a Mach-Zehnder interferometer, which is an essential part of
many optical experiments, based on images of interference fringes acquired by a
monocular camera. The agent is trained in a simulated environment, without any
hand-coded features or a priori information about the physics, and subsequently
transferred to a physical interferometer. Thanks to a set of domain randomizations
simulating uncertainties in physical measurements, the agent successfully aligns
this interferometer without any fine tuning, achieving a performance level of a
human expert.

1 Introduction

Reinforcement learning is developing explosively nowadays, demonstrating breakthrough results
both in computed environments and physical robotics. It achieved above-human performance in Atari
video games [1, 2], defeated the world champion in the game of Go [3], and beat the best computer
programs in chess and shogi [4]. In robotics, RL has been successful in various manipulation and
locomotion tasks [5, 6] such as solving Rubik's cube [7] and throwing various objects to the goal [8].
Most of the works considering applications of RL to robotics deal with commercial robotic hands
[9, 10] or legged robots [11, 12] in well-defined (sometimes changing [13]) environments. Many
of these applications have been demonstrated in "toy" environments of little practical value. In this
paper, we train an RL agent, which we dub *Interferobot*, to solve a commonly occurring practical
task in experimental optics: alignment of a Mach-Zehnder interferometer (MZI).

The main bottleneck in the application of RL methods to robotic training is data acquisition. Since
RL is heavily based on trial and error, it requires millions of agent-environment interactions to
achieve a stable policy. As a result, the training becomes data inefficient and sometimes even
dangerous (for example, for autopilots). One way to overcome this problem is to (pre-)train an
agent policy in a simulated environment [14]. Training in simulation lets an agent interact with the
environment much faster but introduces a domain gap: bias between simulated and real environment.
Domain randomization can be used to overcome the domain gap and achieve a stable generic

policy [15, 16, 17, 18], sometimes with additional fine-tuning in a real environment [19]. However, generalizing this approach to all possible robotic tasks is an open problem.

In present work, we first create a simulator of an MZI and train an agent's policy in this simulated environment based on the observation of images from a monocular camera. The training is implemented without any domain-specific knowledge about the system and hand-coded features. Subsequently, we transfer the trained policy to a physical setup. We add several types of domain randomization in the training procedure in the simulated environment, which allows the transfer to succeed without fine-tuning (in a "zero-shot" manner).

Robotic tuning of optical circuits and elements is a dream of many experimental physicists. Tabletop optical experiments require the alignment of tens or even hundreds of different optical elements with micrometric precision. For some experiments, this procedure takes a team of skilled specialists many hours and has to be repeated daily. Fortunately, this alignment is a well-defined task with specified quality metrics and is hence amenable to robotics. However, to our knowledge, control of real-world optical setups by RL agents was not developed before.

2   Related work

For relatively simple tasks, it is possible to train a robot's policy directly in the experiment without resorting to a simulated environment [11], [12], [13]. This direct training can be helped with a priori knowledge of the physics and dynamics of the environment. For example in [8], the authors trained an agent to throw arbitrary objects into a target box. To achieve that, they embedded a simple ballistic equation into an agent as a starting point and trained it to learn a correction factor. But most often, direct training is not possible due to the environment's complexity. This is especially the case when the robot uses visual information in its policy. In such cases, the pre-training of a policy in a simulated environment is used, followed by transfer and further fine-tuning in the experiment. This method is extensively studied nowadays and different approaches are used. We will name a few. In [19], the authors trained a robotic hand to grasp various objects. They used a generative adversarial network to transform real-world images to look like images from a simulated environment. An alternative approach has been chosen to learn a dexterous in-hand manipulation [20] and solve Rubik's cube with a robot hand [7]. In these works, a separate neural net was trained in a supervised fashion to extract features from randomized simulated images, which were used as a part of agent's observations. In [17], the authors performed a real-world flight of a quadrotor after training only on simulated images. To achieve that, they trained an agent to predict a probability of collisions using heuristics based on free space detection for initialization. In contrast to these references, we trained the agent directly using images, without a priori knowledge about the environment, nor any additional neural nets for image transformation.

Attempts at automated alignment of optical systems included a 1993 study by NASA [21], in which this task was treated as a supervised (rather than reinforcement) learning problem solvable with a multilayer perceptron. The input data consisted of handcrafted features manually extracted from images and were labeled with the actions by a human expert. The overall performance was below that of a human. Another approach to automated alignment of a Michelson interferometer with Fabry-Perot cavities was pursued by the VIRGO collaboration in 1997 [22]. The actions were purely hand-coded involved no neural networks. The method proved to be easy to implement and reliable, but not easily generalizable to interferometers of other types or even with different parameters. In contrast, we present a general approach that can be applied to a wide range of setups and does not rely on handcrafted features or manually collected datasets.

Finally, we note that the alignment problem is more suitable for RL than black-box optimizers, genetic algorithms and the like. This is because the initial state of a misaligned setup is different for every episode. An RL agent utilizes the experience acquired through training to quickly and successfully align the setup from any initial state.

3   Mach-Zehnder interferometer and its simulator

Interference is a physical phenomenon in which two coherent waves superimpose to produce oscillations of different amplitudes at different points in space. Dependent on whether the two waves oscillate in or out of phase with respect to each other, the interference can be constructive or destruc-
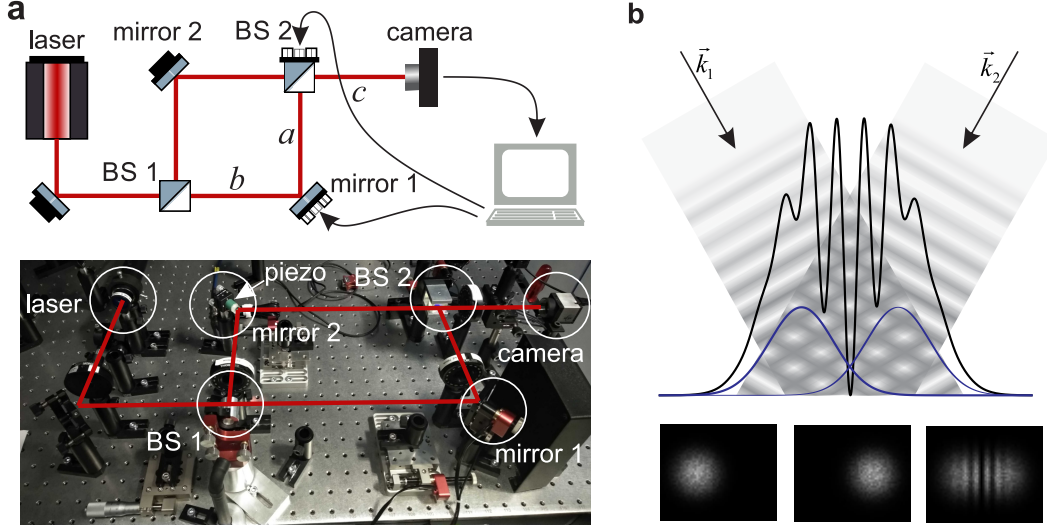
Figure 1: a) Conceptual scheme and experimental setup of the Mach-Zehnder interferometer. The interferometer dimensions are listed in Table 1. b) One-dimensional slice of the interferometric pattern produced by two imperfectly overlapping coherent laser beams with wavevectors $\vec{k}_1$ and $\vec{k}_2$. Wavefronts are shown via shading; blue lines display the intensities of individual beams (not to scale); black line corresponds to the intensity profile of the resulting picture. Corresponding two-dimensional pictures of individual beams and their interference are displayed at the bottom.

tive. In this paper, we consider optical interference of two optical waves — light beams — produced by the same laser source. In this case, the interference manifests itself as a pattern of bright and dark fringes visible by eye.

An interferometer is a physical device whose operating principle is based on the interference effect and is used to precisely measure the relative phase difference between two coherent laser beams. It is a primary tool in both academic and applied experimental optics. For example, the Fabry-Pérot [23] interferometer is used as a high-resolution optical spectrometer, the Michelson interferometer is the main tool of modern gravitational-wave detectors such as LIGO or VIRGO [24, 25] and also used to precisely measure surface roughness [26], Sagnac interferometers are used in modern navigation systems, and the Mach-Zehnder interferometer is a primary tool of modern quantum optical experiments [27, 28].

In this paper, we consider a configuration of the MZI shown in Fig. 1(a). The laser beam is divided into two at a beam splitter (BS 1). The two beams go through different interferometer arms and are subsequently recombined at another beam splitter (BS 2). The superimposed beams are detected at a digital camera. Mirror 1 and BS2 are set up in special optomechanical mounts (Newport [29]), which can be tilted in two planes either manually using knobs or via an electronic actuator. A camera is used to analyze the resulting interference pattern.

The goal of the alignment is to precisely overlap the two beams that propagated through the two interferometer arms, to make both the beam centers and wave vectors coincide. The observations that guide the alignment are the images acquired by the camera. A typical interference picture for two misaligned beams is shown in Fig.1(b). In the area where the beams overlap, spatially-dependent fringes appear due to the difference in the beam directions.

An important part of the observation is the temporal dynamics of the fringes exhibited when the relative phase between the two beams varied. To this end, mirror 2 is mounted on a piezoelectric transducer which can change the optical path length in the upper arm. Voltage is applied to the piezo in a periodic manner; its peak-to peak amplitude corresponds to the phase shift of approximately 2π. For misaligned beams, this motion results in transverse movement of the fringes as seen in Fig. 2(b,c). While the spatial frequency of the fringes is proportional to the absolute difference between $\vec{k}_1$ and $\vec{k}_2$, the direction of their movement provides information about the sign of that difference. For an aligned
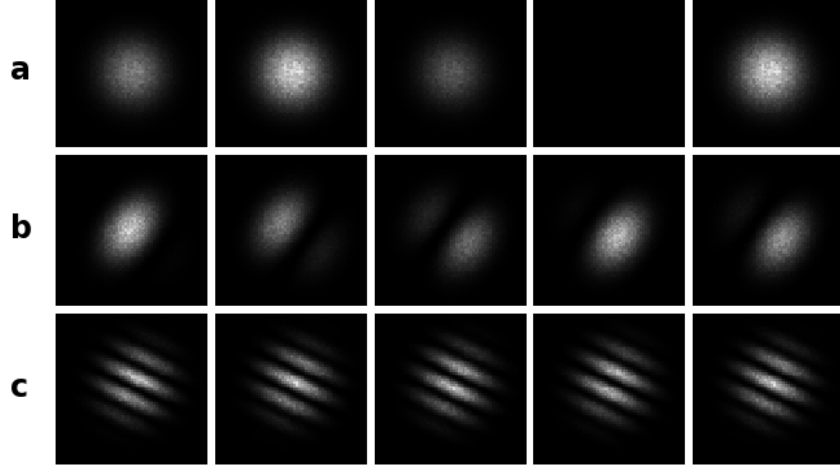
Figure 2: Simulated examples of interferometric patterns. (a) Perfectly aligned beams, visibility = 1; (b) Slightly misaligned beams, visibility = 0.3; (c) Strongly misaligned beams, visibility = 0.0026. Pictures from left to right correspond to different relative phases $[\phi_{\text{piezo}}(t)]$ between the two interferometer arms induced by the movement of mirror 2 .

interferometer, the phase difference of the beams is spatially constant over the camera's sensitive area, so the interference manifests itself as the "blinking" of the bright spot [Fig. 2(a)].

To align the interferometer, one needs to interpret the interference patterns and overlap the two beams accurately. It is achieved by tweaking mirror 1 and BS 2 in two orthogonal planes, either manually or via a computer-controlled interface. Despite its apparent simplicity, this alignment is a tedious and time-consuming task. The reason for this is threefold. First, each movement of either mirror 1 or BS 2 lead to simultaneous change both in the position of the lower beam at the camera and its corresponding wave-vector. Hence by trying to improve the spatial overlap between the beams one can lose their parallelism and vice versa. Second, the pattern measured by the camera is compromised by noise, aberrations and distortions such as dust [Fig. 3(a)]. Third, even advanced optomechanical mounts have limited precision and non-zero hysteresis.

To train Interferobot, we developed a fast and precise simulator of the MZI. In our model, each laser beam is described by a Gaussian transverse profile with a plane wavefront:

$$E(x, y, z) = e^{-\frac{(x-x_0(z))^2 + (y-y_0(z))^2}{r^2}} e^{ik_x x + ik_y y + ik_z z}, \tag{1}$$

where (x, y, z) is the coordinate vector, $(x_0(z), y_0(z))$ the position of the center of the beam, $\vec{k}$ the wave vector and r the beam radius. We neglect the beam divergence due to diffraction and assume paraxial approximation, i.e. $k_z \gg k_x, k_y$, where z is the direction of propagation. The intensity of the interferometric picture produced by the two superimposed beams in the camera plane ($z = 0$) is then $I(x, y, t) = |E_1(x, y, 0) + E_2(x, y, 0)e^{i\phi_{\text{piezo}}(t)}|^2$, where and $\phi_{\text{piezo}}(t)$ is the phase shift induced by the piezo mirror movement.

The traditional alignment quality metric is the visibility defined as

$$V = \frac{\max_t(I_{\text{tot}}) - \min_t(I_{\text{tot}})}{\max_t(I_{\text{tot}}) + \min_t(I_{\text{tot}})}, \tag{2}$$

where $I_{\text{tot}}(t) = \iint_{-\infty}^{+\infty} I(x, y, t) \mathrm{d}x \mathrm{d}y$ is a total optical power incident on the camera; the maximum and minimum are taken over the piezo movement period. The visibility is a real number between 0 and 1. For a fully aligned interferometer [Fig. 2(a)], $\min_t(I_{\text{tot}}) = 0$, so $V = 1$, whereas when the interferometer is misaligned [Fig. 2(c)], $\min_t(I_{\text{tot}}) \approx \max_t(I_{\text{tot}})$, so $V \approx 0$.

To simplify the simulation, we assume that the upper beam propagates exactly along z, so $x_0 \equiv y_0 \equiv k_x \equiv k_y \equiv 0$ for that beam. From Eqs. (1) and (2), we find the visibility to equal

$$V = \exp\left(-\frac{x_0^2 + y_0^2}{2r^2}\right) \exp\left[-\frac{(k_x^2 + k_y^2)r^2}{8}\right], \tag{3}$$

4

where x_0 and y_0 are the coordinates of the center of the lower beam at the camera and k_x and k_y the components of its wave-vector. Assuming that the beam arrives at mirror 1 parallel to the z axis and at the origin of the (x, y) plane, we have

$$k_{x,y}/k = \alpha_{1x,y} + \alpha_{2x,y} \text{ and } x, y = \alpha_{2x,y}c + \alpha_{1x,y}(a + c) \qquad (4)$$

where $k = \sqrt{k_x^2 + k_y^2 + k_z^2} = 2\pi/\lambda$, λ is the wavelength, $\alpha_{1x,y} \ll 1$ and $\alpha_{2x,y} \ll 1$ are the horizontal and vertical angles by which mirror 1 and BS2, respectively, deflect the beam (with respect to the default $90°$ reflection) while a and c are the interferometer dimensions as shown in Fig. 1(a) and listed in Table 1. We used the fact that, in the paraxial approximation, the horizontal and vertical angles of the beam with respect to the z axis are given by k_x/k and k_y/k, respectively.

In the simulation, we represent the camera matrix as a rectangular equidistant grid of 64×64 pixels and calculate the intensity for each pixel. The camera is centered at $(x, y) = (0, 0)$, so the beam that has propagated through the upper arm hits the camera in its center. We simulate 16 interferometric patterns distributed in time over a full period of piezo movement corresponding to different $\phi_{\text{piezo}}(t)$. We then calculate the visibility according to Eq. (2). The fast performance of the simulator is crucial for the training of Interferobot. By using a parallel C++ code, we produce 200 16-frame sets per second. For the training of agents, the simulator has been wrapped into a gym-like interface [30].

Table 1: Setup parameters. The interferometer dimensions are given in millimeters, angles in radians.

parameter	a	b	c	r	$\alpha_{\max(x,1)}$	$\alpha_{\max(y,1)}$	$\alpha_{\max(x,2)}$	$\alpha_{\max(y,2)}$
value	200	300	100	0.95	$5.2 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$2.6 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$

4   Our method

States and observations.   The MZI alignment procedure is treated as a partially observable Markov decision process. The hidden state is the set (x_0, y_0, k_x, k_y), which is not known to Interferobot, while the observation is the set of 16 consecutive 64×64 images taken in one back and forth run of the piezo. Each observation is represented by a (16, 64, 64) tensor.

Actions.   Interferobot has a multi-discrete action space. Each action is a change in one of the beam deflections $\alpha_{(x,y),(1,2)}$ or $'$do nothing$'$. We allow three different magnitudes for these changes: $[0.01, 0.05, 0.1] \times \alpha_{\max}$, where the maximum angle values $\alpha_{\max}$ are listed in Table 1. It lets the agent choose an appropriate magnitude of action dependent on the degree of misalignment. In total, this results in an action space with the dimension $2\,\text{mounts} \times 2\,\text{dimensions} \times 2\,\text{directions} \times 3\,\text{magnitudes} + 1\,'\text{do nothing}' = 25$.

Rewards.   Visibility itself is not optimal as the reward. First, it is a rather sparse signal, as it exponentially decays with the beam distance and angle as per Eq. (3). Second, a small difference in the final alignment quality (e.g. $V = 0.95$ vs. $V = 0.98$) can be of significance for experimental practice. Hence we propose the reward signal $R = V - \log(1 - V) - 1$. In this way, Interferobot is punished for each additional step, but significantly rewarded for high alignment quality.

Episode and reset.   The episode length is set to 100 actions. At the beginning of each episode, the beam angles are reset to random values within $\pm\alpha_{\max}$. The values of $\alpha_{\max}$ for each control are chosen to keep both beams within the frame of the camera and to prevent interference pattern features that are small compared to the size of the pixel. Specifically, the beam centers can be away from each other by about 3 radii and the interference pattern can have no more than 15 fringes.

Examples of interference patterns observed (experimentally) after the reset are shown in the Appendix (Fig. 7).

Randomisation.   To make the learned policy applicable in a physical environment, we applied several domain randomizations inspired by uncertainties in experimental measurements. For each episode, we randomly vary the beam radius by $\pm20\%$, as this parameter is difficult to precisely measure in the experiment. Varying the radius also helps addressing the deviation of the experimental beam profile from precise Gaussian shape.
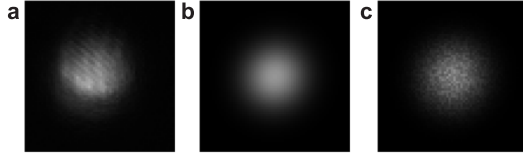
Figure 3: Single laser beams: (a) experimental; (b) simulated Gaussian; (c) simulated with noise. In (a), residual fringes are visible due to a resonator effect in the cover glass of the camera.

The following further randomizations are applied once per time step. First, we randomly scaled the brightness by $\pm 30\%$ to model the variable exposure time of the camera. Second, we added 20% of white noise to each pixel to address the camera detector noise as well as the effects of dust and spurious interference (Fig. 3). The third randomization addresses the challenge of synchronizing the piezo movement with the triggering of the camera. We applied a cyclic shift of simulated interference pattern images within each period of the piezo voltage. In addition, we randomized the "duty cycle", i.e. how the 16 images acquired during the period are distributed between the forward and backward passes of the piezo. The fraction of images acquired during the forward pass was maintained greater than 50% (at least 9 per period), so Interferobot had enough information to distinguish between the two directions.

5 Experiments in the simulated environment

We train the agent in the simulated environment using the double dueling DQN algorithm [31] with a discount factor $\gamma = 0.99$, total number of steps 5×10^6, and replay buffer size 3×10^4. Updates were performed every four steps. The whole training on a NVidia GTX 2060 GPU took about 10 hours. The training curves are shown in Fig. 4. After the training, both the angle and distance between the beams observed at the and of the episode are below the experimental uncertainty (we recall that the beam radius $r = 0.95$ mm and the angle corresponding to a single fringe is 0.6 mrad). One can also see the benefit of the chosen reward function: while the visibility curve plateaus at the end of the training, the return continues to increase making minute policy improvements manifest.
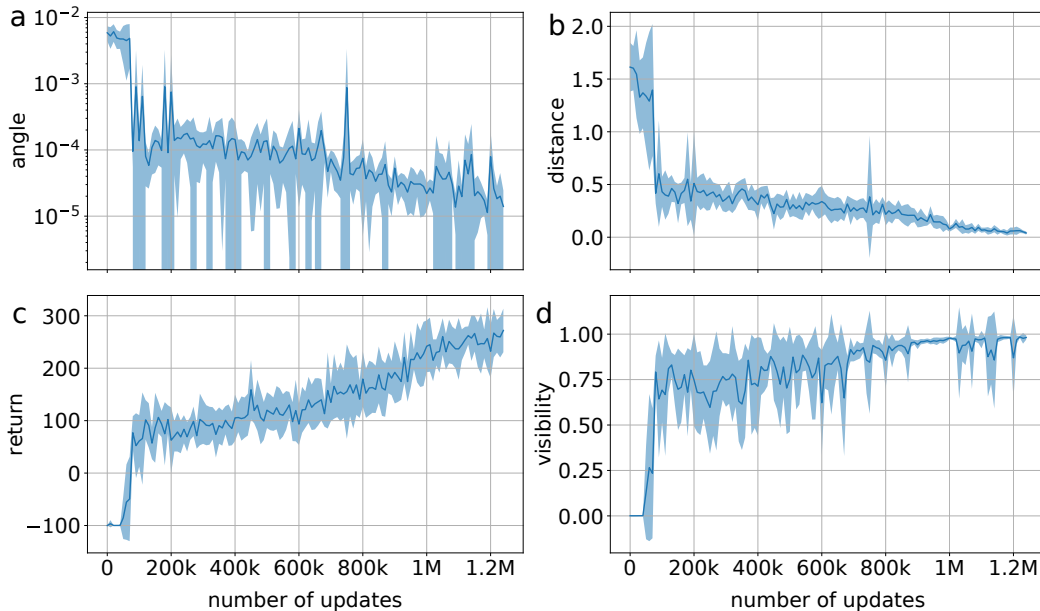

Figure 4: Training in simulation (averaging over 10 training runs). (**a**) Angle and (**b**) distance between beams during training, (**c**) return, (**d**) visibility. The data in (a), (b) and (d) are for the final steps of each episode.

6

6   Experiments on the physical Mach-Zehnder interferometer

The experimental MZI is shown in Fig. 1(a). We used a continuous wave laser with the wavelength $\lambda = 635$ nm, Newport Picomotor mirror mountsan IDS UI-2230SE-M-GL camera with resolution 1600×1200 pixels, and acquisition rate 18 frames per second. The acquired images were re-scaled to 64×64 pixels. A gym-like interface was created to control the experimental setup by the agent trained in the simulated environment. The action of Interferobot consisted in the corresponding command to the mirror controller. The experimental setup generated 0.37 observations per second, corresponding to a single period of the piezo movement.

The trained agent was evaluated automatically without human involvement for 100 episodes of 100 steps each. At the beginning of each episode, the interferometer was reset to a random position as described in the previous section.
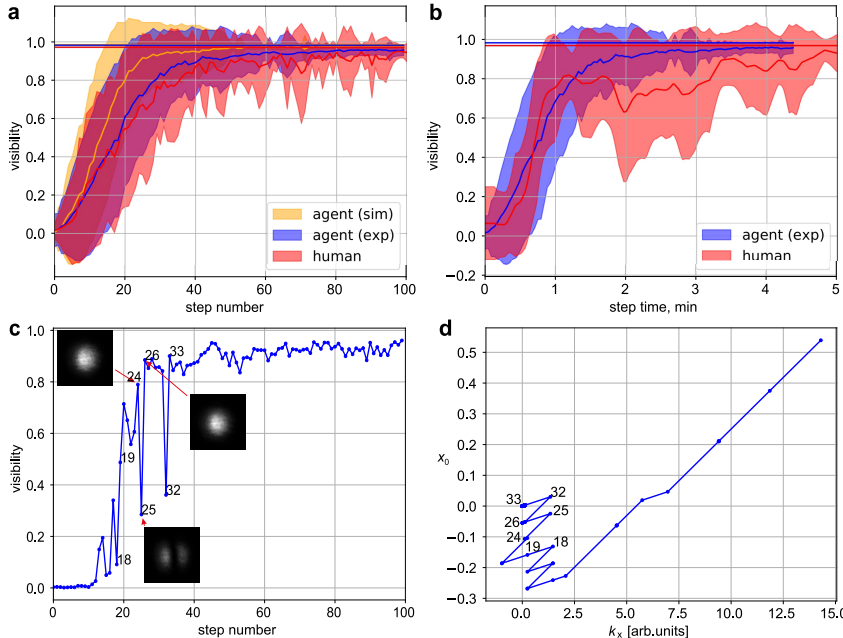

Figure 5: Evaluation results.(**a**) Comparison of Interferobot's performance during an episode in simulation and on the physical setup with a human expert using a keyboard interface. Agent's performance is averaged over 100 evaluations, human's — over 18. (**b**) The same but the human manually aligned the MZI by turning knobs. The human's performance is averaged over 16 episodes in this case. (**c**) Example visibility evolution in a single experimental episode. The interference pattern photos correspond to steps 24, 25 and 26. (**d**) Corresponding trajectory in the phase space (x, k_x) shows gradual convergence to the origin.

In Fig. 5 we present Interferobot's statistics averaged over all evaluation episodes and compare its performance with that of a skilled human expert. The human expert operated in two settings: using a keyboard interface with the same set of actions as the agent and aligning the interferometer manually using mirror knobs, in line with normal experimental practice.

Fig. 5(a) corresponds to the first setting. In addition to the robotic agent operating in the physical environment, we also show the agent's performance in the simulated environment for reference. It is seen that the latter surpasses the former due to the domain gap. However, the robotic agent does outperform the human. As an additional comparison metric, we determine the highest visibility obtained in each episode and average it over all episodes. This value is 0.986 for Interferobot evaluated in the simulation, 0.983 for the experiment, and 0.972 for the human. Note that, while the performance of the human and robot is comparable when measured in time steps, the human works significantly slower in terms of wall time.

Similar results are obtained in the second setting [Fig. 5(b)]. Since, in this evaluation, it is difficult to divide the human's procedure into discrete actions, we plot the time, rather than the number of steps,

along the horizontal axis. In this setting, the mean highest visibility obtained by the human was 0.968. It is seen that the human slightly outperforms the robot in the beginning of each episode, but falls behind when it comes to fine tuning. Initially, when the interferometer is significantly misaligned, it is easy for the human to interpret the fringes and obtain the basic alignment in only a few actions. However, when this is achieved, the human is unable to reliably see a small difference (x_0, y_0) in the beam center positions. To see this difference, the human moves the beams significantly apart, resulting in a temporary visibility drop. Such drop is not present in the first setting [Fig. 5(a)] because significant misalignment requires multiple keyboard action, and the human adopts a different fine-tuning strategy: attempting to move the two optical elements by equal angles in opposite directions and observing whether the visibility has improved.

Interferobot needs to reach, in as few steps as possible, the maximum visibility point, which is located at the origin of the (x_0, y_0, k_x, k_y) space as per Eq. (3). The agent's actions modify $\alpha_{(1,2),(x,y)}$, and hence, according to Eq. (4), are non-orthogonal in that space. This means that the optimal policy will not increase the visibility at every step. Rather, it should work in a zig-zag fashion, repeatedly significantly misaligning mirror 1 and then compensating with BS 2 (or vice versa). This behavior, manifested in significant oscillations of the visibility during the alignment, is seen in Fig. 5(c,d). Note the difference between these deliberate misalignment actions and those by the human expert mentioned above. While the human misaligns the interferometer in order to better observe the beam positions, Interferobot appears to purposefully make zig-zag steps in order to change x_0 or y_0 without modifying k_x or k_y, respectively.

7   Ablation study

Table 2: Ablation study

	visibility	return
All randomizations	**0.96 ± 0.02**	**221 ± 54**
No radius randomization	0.74 ± 0.20	85 ± 69
No brightness randomization	0.91 ± 0.04	178 ± 39
No white noise	0.82 ± 0.07	129 ± 43
No duty cycle and trigger time randomization	0.89 ± 0.07	200 ± 42

We trained and evaluated four additional agents, in each of which one of the domain randomizations listed in Sec. 4 has been removed. Every agent was evaluated on the physical setup for 100 episodes. The evaluation of each agent took around 6 hours and was fully automated. In table 2, we present the mean visibility in the last 20 steps and the mean return, averaged over all episodes. We observe that the most significant benefits are due to the radius randomization and white noise addition.

8   Summary

We have shown that an optical interferometer can be successfully aligned by a robotic RL agent without any prior knowledge of the system's physics. When trained with the randomizations, the agent can be transferred from simulation to an experimental setup without any fine-tuning. The trained agent's performance is comparable with that of a skilled specialist in terms of the achieved visibility, but significantly faster. Our approach can be easily extended to interferometers of various sizes, shapes, and with additional optical elements. In the future, such a method can be applied for fully automated alignment of complex optical setups, which can significantly decrease the amount of routine manual work needed for complex experiments.

9   Links

Videos of the interferometer alignment and software are available via link `https://github.com/dmitrySorokin/interferobotProject`

Broader impact

The direct application of our work is the automation of complex optical experiments. A robotic agent will take over a mundane part of the experimentalist's work, so scientists can concentrate on generating research ideas and analyzing the scientific results. A potential risk is associated with laser safety, which is a primary concern in optical experiments, as laser beams are able to permanently damage the eyesight. While standard precautions (goggles, barriers, etc.) are normally taken to reduce this risk, additional measures may be required to prevent beams from being misdirected outside the plane of the optical table by the robotic agent. This can be achieved, for example, by limiting the angular range of motorized mirror mounts at the hardware level.

Acknowledgments and Disclosure of Funding

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[2] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. Agent57: Outperforming the Atari Human Benchmark. *arXiv e-prints*, page arXiv:2003.13350, March 2020.

[3] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

[4] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv e-prints*, page arXiv:1712.01815, December 2017.

[5] Henry Zhu, Abhishek Gupta, Aravind Rajeswaran, Sergey Levine, and Vikash Kumar. Dexterous Manipulation with Deep Reinforcement Learning: Efficient, General, and Low-Cost. *arXiv e-prints*, page arXiv:1810.06045, October 2018.

[6] Anusha Nagabandi, Kurt Konoglie, Sergey Levine, and Vikash Kumar. Deep Dynamics Models for Learning Dexterous Manipulation. *arXiv e-prints*, page arXiv:1909.11652, September 2019.

[7] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's Cube with a Robot Hand. *arXiv e-prints*, page arXiv:1910.07113, October 2019.

[8] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. TossingBot: Learning to Throw Arbitrary Objects with Residual Physics. *arXiv e-prints*, page arXiv:1903.11239, March 2019.

[9] Eric Jang, Coline Devin, Vincent Vanhoucke, and Sergey Levine. Grasp2Vec: Learning Object Representations from Self-Supervised Grasping. *arXiv e-prints*, page arXiv:1811.06964, November 2018.

[10] Brian Yang, Jesse Zhang, Vitchyr Pong, Sergey Levine, and Dinesh Jayaraman. REPLAB: A Reproducible Low-Cost Arm Benchmark Platform for Robotic Learning. *arXiv e-prints*, page arXiv:1905.07447, May 2019.

[11] Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to Walk via Deep Reinforcement Learning. *arXiv e-prints*, page arXiv:1812.11103, December 2018.

[12] Michael Ahn, Henry Zhu, Kristian Hartikainen, Hugo Ponte, Abhishek Gupta, Sergey Levine, and Vikash Kumar. ROBEL: Robotics Benchmarks for Learning with Low-Cost Robots. *arXiv e-prints*, page arXiv:1909.11639, September 2019.

[13] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation. *arXiv e-prints*, page arXiv:1806.10293, June 2018.

[14] Ofir Nachum, Michael Ahn, Hugo Ponte, Shixiang Gu, and Vikash Kumar. Multi-Agent Manipulation via Locomotion using Hierarchical Sim2Real. *arXiv e-prints*, page arXiv:1908.05224, August 2019.

[15] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. *arXiv e-prints*, page arXiv:1703.06907, March 2017.

[16] Reda Bahi Slaoui, William R. Clements, Jakob N. Foerster, and Sébastien Toth. Robust Visual Domain Randomization for Reinforcement Learning. *arXiv e-prints*, page arXiv:1910.10537, October 2019.

[17] Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real Single-Image Flight without a Single Real Image. *arXiv e-prints*, page arXiv:1611.04201, November 2016.

[18] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. *arXiv e-prints*, page arXiv:1710.06537, October 2017.

[19] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-Real via Sim-to-Sim: Data-efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks. *arXiv e-prints*, page arXiv:1812.07252, December 2018.

[20] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob Mc-Grew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning Dexterous In-Hand Manipulation. *arXiv e-prints*, page arXiv:1808.00177, August 2018.

[21] Arthur Decker, Michael Krasowski, and Kenneth Weiland. Neural-network-directed alignment of optical systems using the laser-beam spatial filter as an example. 09 1993.

[22] D Babusci, H Fang, G Giordano, G Matone, L Matone, and V Sannibale. Alignment procedure for the VIRGO interferometer: experimental results from the frascati prototype. *Physics Letters A*, 226(1-2):31–40, February 1997.

[23] C. Fabri and A. Pérot. On the application of interference phenomena to the solution of various problems of spectroscopy and metrology. *Astrophysical Journal*, 9:87, 1899.

[24] B P Abbott, R Abbott, R Adhikari, and et.al. LIGO: the laser interferometer gravitational-wave observatory. *Reports on Progress in Physics*, 72(7):076901, June 2009.

[25] T Accadia, F Acernese, M Alshourbagy, and et.al. Virgo: a laser interferometer to detect gravitational waves. *Journal of Instrumentation*, 7(03):P03012–P03012, March 2012.

[26] H.C. Kandpal, D.S. Mehta, and J.S. Vaishya. Simple method for measurement of surface roughness using spectral interferometry. *Optics and Lasers in Engineering*, 34(3):139–148, September 2000.

[27] Alexei Ourjoumtsev, Rosa Tualle-Brouri, Julien Laurat, and Philippe Grangier. Generating optical Schrödinger kittens for quantum information processing. *Science*, 312(5770):83–86, 2006.

[28] Demid V. Sychev, Alexander E. Ulanov, Anastasia A. Pushkina, Matthew W. Richards, Ilya A. Fedorov, and Alexander I. Lvovsky. Enlargement of optical Schrödinger's cat states. *Nat. Photonics*, 11:379–382, 2017.

[29] Newport optical mirror mount. `https://www.newport.com/p/8807`.

[30] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. cite arxiv:1606.01540.

[31] Ying Huang, GuoLiang Wei, and YongXiong Wang. V-d d3qn: the variant of double deep q-learning network with dueling architecture. In *2018 37th Chinese Control Conference (CCC)*. IEEE, July 2018.