Journal of *Marine Science and Engineering*

MDPI

Article

Adaptive Control for Underwater Simultaneous Lightwave Information and Power Transfer: A Hierarchical Deep-Reinforcement Approach

Huicheol Shin [1,2], Sangki Jeong [1], Seungjae Baek [1,2] and Yujae Song [3,*]

[1] Maritime ICT & Mobility Research Department, Korea Institute of Ocean Science and Technology, Busan 49111, Republic of Korea; shc0305@kiost.ac.kr (H.S.); jeongsk313@kiost.ac.kr (S.J.); baeksj@kiost.ac.kr (S.B)
[2] Marine Technology and Convergence Engineering, University of Science and Technology, Busan 49111, Republic of Korea
[3] Department of Robotics Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
* Correspondence: yjsong@yu.ac.kr

Abstract: In this work, we consider a point-to-point underwater optical wireless communication scenario where an underwater sensor (US) transmits its sensing data to a remotely operated vehicle (ROV). Before the US transmits its data to the ROV, the ROV performs simultaneous lightwave information and power transfer (SLIPT), delivering both control data and lightwave power to the US. Under the considered scenario, our objective is to maximize energy harvesting at the US while supporting predetermined communication performance between the two nodes. To achieve this objective, we develop a hierarchical deep Q-network (DQN)–deep deterministic policy gradient (DDPG)-based online algorithm. This algorithm involves two reinforcement learning agents: the ROV and US. The role of the ROV agent is to determine an optimal beam-divergence angle that maximizes the received optical signal power at the US while ensuring a seamless optical link. Meanwhile, the US agent, which is influenced by the decision of the ROV agent, is responsible for determining the time-switching and power-splitting ratios to maximize energy harvesting without compromising the required communication performance. Unlike existing studies that do not account for adaptive parameter control in underwater SLIPT, the proposed algorithm's adaptive nature allows for the dynamic fine-tuning of optimization parameters in response to varying underwater environmental conditions and diverse user requirements.

Keywords: simultaneous lightwave information and power transfer (SLIPT); reinforcement learning; underwater optical wireless communication; adaptive control

Citation: Shin, H.; Jeong, S.; Baek, S.; Song, Y. Adaptive Control for Underwater Simultaneous Lightwave Information and Power Transfer: A Hierarchical Deep-Reinforcement Approach. *J. Mar. Sci. Eng.* **2024**, *12*, 1647. https://doi.org/10.3390/jmse12091647

Academic Editor: Cameron Johnstone

Received: 7 August 2024
Revised: 6 September 2024
Accepted: 12 September 2024
Published: 14 September 2024

1. Introduction

Underwater optical wireless communication (UOWC) is a cutting-edge technology that uses light to transmit data through water, enabling high-speed and reliable communication in underwater environments. It has gained significant importance in various scientific and industrial applications, including underwater sensing, environmental monitoring, underwater robotics, and offshore exploration. Compared to traditional acoustic-based communication, UOWC offers several advantages, such as higher data rates, wider bandwidths, lower latencies, and immunity to electromagnetic interference. These advantages make it a promising solution for fulfilling the increasing demands of underwater communication systems [1–3].

However, despite the numerous advantages of UOWC, it faces several challenges that can significantly impact its performance. A primary issue is signal attenuation and fading, which arise due to the inherent properties of water, such as absorption, scattering, and turbulence. Another critical challenge is the misalignment between the transmitter

and receiver, often caused by factors like water currents or the movement of underwater vehicles, leading to degraded link quality and reduced communication range. To address these challenges, various studies have been conducted, focusing not only on evaluating UOWC system performance considering optical signal attenuation and fading [4–7], but also on mitigating the effects of unpredictable misalignment between the transmitter and receiver [8–11].

Meanwhile, in recent times, the development of simultaneous lightwave information and power transfer (SLIPT) techniques [12] has garnered significant importance in the field of UOWC, leading to the emergence of underwater SLIPT. SLIPT is an extended concept of simultaneous wireless information and power transfer, called SWIPT [13,14], to the optical domain, utilizing light signals for both data transmission and power transfer. SLIPT offers the unique capability to not only transmit data but also provide power simultaneously, thus addressing the challenges of power supply in underwater environments. The ability to simultaneously transfer data and power represents a significant technological advancement in underwater communication and holds great promise for facilitating reliable and sustainable operations in challenging underwater environments. The work of [15] introduced an overview of various SLIPT techniques in the time, power, and spatial domains. Moreover, it presented two underwater proof-of-concept demonstrations of time-switching (TS) SLIPT. In [16], an underwater SLIPT system was designed that consisted of a laser diode (LD)-based transmitter and a multi-element receiver with a single-photon avalanche diode and a solar panel. In [17], the authors investigated closed-form expressions for energy harvesting (EH), bit error rate, and spectral efficiency (SE) over log-normal turbulence channels under different underwater SLIPT methods. Optimization problems were then formulated for each method, and the optimal TS and power-splitting (PS) ratios were determined. The work of [18] presented the constellation design for an optimized color-shift keying system to maximize the minimum distance between the constellation points while mitigating the total received current constraint to optimize communication performance. In [19], the evaluation of communication link performance and charging speed was conducted under an actual experimental setup of an underwater SLIPT system. For the expansion of the UOWC range, a dual-hop structure with an underwater SLIPT was introduced based on the TS method [20]. Subsequently, expressions for the average BER at the target node and the harvested energy by the relay node were derived over underwater attenuation channels. The work of [21] considered a cooperative non-orthogonal multiple-access-assisted uplink UOWC system based on SLIPT. In particular, in the process of performance evaluations, various practical assumptions, including misalignment at the relay node, were reflected.

However, despite these research achievements in the field of UOWC, it is worth noting that the existing works on underwater SLIPT, including [15–21], have not addressed the adaptive control of TS and PS ratios in combination with the beam-divergence angle, considering changes in the underwater environment. This aspect is crucial because the adaptive control of these parameters plays a vital role in providing seamless communication service while maximizing EH in dynamic and time-varying underwater environments. In real-sea conditions, unlike on land, the UOWC channel is subject to a range of external factors such as water currents, salinity, and temperature fluctuations, which can cause rapid and unpredictable changes in channel characteristics. These challenges make it particularly difficult to guarantee consistent and reliable communication performance in underwater environments, further underscoring the importance of adaptive control strategies.

1.1. Contributions

We highlight our contributions in this work as follows:

- In this study, our objective is to develop an online algorithm for UOWC that adaptively determines the TS and PS ratios of SLIPT as well as the beam-divergence angle to maximize EH while ensuring seamless communication performance between a remotely operated vehicle (ROV) and an underwater sensor (US) with SLIPT capabilities. To carry out the ROV missions set in this study, we consider a hybrid UOWC system

that utilizes both LD and light-emitting diode (LED) technologies. LD-based UOWC is employed for control data and power transmission from the ROV to the US via SLIPT, whereas LED-based UOWC is used for sensing data transmission from the US to the ROV.

- To address the challenges of this communication scenario, we propose a hierarchical deep Q-network (DQN)–deep deterministic policy gradient (DDPG) algorithm. This algorithm involves two reinforcement learning (RL) agents: the ROV agent and the US agent. The role of the ROV agent is to determine the beam-divergence angle that maximizes the received optical power at the US node while ensuring a seamless optical link. On the other hand, the US agent, influenced by the decisions of the ROV agent, is responsible for determining the TS and PS ratios that maximize the EH without compromising the required communication performance.
- Through extensive simulations, we demonstrate that the proposed algorithm successfully maximizes the EH while maintaining the predetermined communication requirement at the US. The adaptive nature of the algorithm allows it to dynamically adjust the system parameters in response to changing underwater environmental conditions and sensor requirements, therefore enabling efficient and sustainable energy transfer and communication in underwater environments.

1.2. Organization

The rest of this paper is organized as follows. In Section 2, we formally present our underwater UOWC scenario between the ROV and US. Section 3, we introduce a hybrid TS and PS SLIPT technique and its corresponding performance metrics. In Section 4, we first formulate an optimization problem to achieve our objective and then propose an online learning algorithm (i.e., hierarchical DQN–DDPG algorithm) to solve the problem in real time. In Section 5, we provide an evaluation of the performance of our proposed algorithm based on extensive simulations. Finally, conclusions are drawn in Section 6.

2. System Model

2.1. Network Model

We consider a three-dimensional (3D) underwater communication network in which a US communicates with the ROV as shown in Figure 1. More specifically, the US is fixed onto the seafloor and measures, at regular intervals, a variety of underwater environmental data depending on its purpose. On the other hand, the ROV can conduct many shallow and deep underwater missions, such as marine science and oil and gas extraction missions, which would otherwise be very difficult or dangerous for humans to do, even if diving in a submersible or submarine. In these applications, the motions of the ROV are guided either by a human pilot on a surface support vessel through an umbilical cable that provides power and telemetry or by an automatic pilot system [22]. This study assumes that the ROV is controlled by a human pilot through an umbilical cable and that it has two missions: (1) collecting sensing data measured by the US and (2) wirelessly transferring power to the energy-deprived US for battery charging. To support these ROV missions, this study adopts hybrid LD–LED-based UOWC. The modems for this hybrid UOWC are installed at the bottom of the ROV and at the top of the US, respectively, to align their beams for optical links. More specifically, to simultaneously transmit both power and control data (e.g., wake-up and communication completion data) from the ROV to the US, we employ LD-based SLIPT. For such missions, adopting LD-based communication is reasonable because it is an effective method for transferring power with high efficiency compared to that of LED-based UOWC. On the other hand, to transmit underwater sensing data collected by the US to the ROV, we employ LED-based communication. This is because LED-based communication can support reliable data transmission over a relatively large FOV, even when the ROV and US are not perfectly aligned owing to various factors in the water. As illustrated in Figure 1, the specific ROV operation procedure for achieving these missions is as follows:

1. First, the ROV is launched into the ocean from the support vessel using a launch and recovery system (LARS). Once in the water, the ROV moves to the location where the US is located. At this location, the ROV performs SLIPT to not only transmit control data (e.g., wake-up or communication completion data) but also transfer power.
2. Perceiving the control data and power, the US proceeds to transmit its collected sensing data to the ROV via LED-based UOWC. Although LED-based UOWC may have a relatively lower data rate compared to that of LD-based UOWC, it still provides a sufficient data rate (e.g., more than Mbps [23]) to transmit the sensing data with high reliability.
3. Once the data reception process is complete, the ROV is retrieved and brought back to the support vessel for recovery, which is facilitated by the LARS.

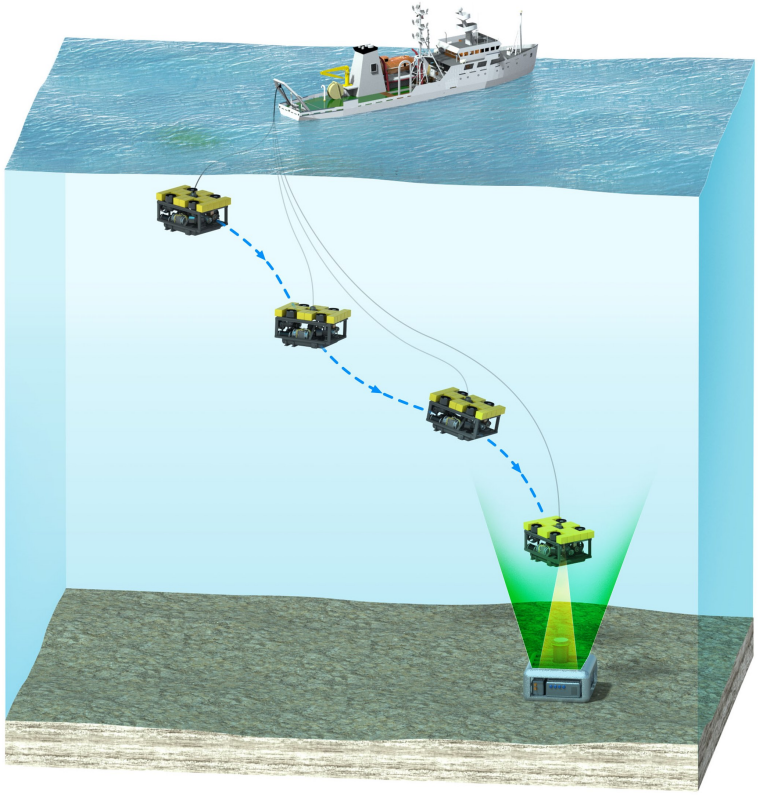


Figure 1. UOWC scenario between an ROV and a US with SLIPT capabilities.

2.2. Signal Model

We consider UOWC that is based on intensity modulation and direct detection (IM/DD), in which the light intensity is modulated as an information-bearing signal, and information is recovered at the receiver side by measuring the intensity of the received light [24]. Under IM/DD, the information bits are modulated via M-ary pulse amplitude modulation (M-PAM), where M denotes the modulation level.

Let T be the time duration of a data frame consisting of M-PAM symbols, such that the symbol interval can be expressed as $T_s = T/M$. We denote the M-PAM symbol as x. Since M-PAM differentiates information solely based on signal amplitude without incorporating phase information, each M-PAM symbol can be geometrically represented as one of the one-dimensional signal points with possible values:

$$\frac{(m-1)A}{M-1}, \quad m = 1, 2, ..., M, \tag{1}$$

where $A \in [0, (I_{\max} - I_{\min})/2]$ is the peak amplitude, and $I_{\max}$ and $I_{\min}$ denote allowable the maximum and minimum input bias currents, respectively. The instantaneous emitted optical intensity signal can be expressed as follows:

$$P_{\text{tx}} = \delta(x + B), \tag{2}$$

where δ denotes the slope efficiency of LD. Meanwhile, $B = I_{\max} - A$ is the DC bias which plays a role in guaranteeing that the resulting signal power is non-negative [25].

When an optical signal is transmitted in an underwater environment, the optical signal undergoes both path loss and fading induced by turbulence. Thus, the instantaneous received optical power P_{rx} at the receive node can be expressed as

$$P_{\text{rx}} = hP_{\text{tx}} = h_{\text{AL}}h_{\text{GL}}h_{\text{F}}P_{\text{tx}}, \tag{3}$$

where h is the underwater channel coefficient which is affected by the 3D position of ROV, and it includes the attenuation loss h_{AL}, geometrical loss h_{GL}, and fading h_{F}. Specific definitions of these terms are presented in the next subsection. The solar panel converts the optical intensity into an electrical current. The received electrical signal can be given as follows [17]:

$$i_{\text{RX}} = rh\delta B + rh\delta x + n = I'_{\text{DC}} + I_{\text{AC}} + n, \tag{4}$$

where r is the solar panel responsivity, and n is the additive white Gaussian noise (AWGN) with zero mean and variance of σ^2. In (4), the first term (i.e., I'_{DC}) and second term (i.e., I_{AC}) in the right equation are the AC and DC components of the received signal, respectively.

2.3. Underwater Optical Channel Model

In this subsection, we describe each channel component of the underwater channel coefficient defined in (3). In UOWC, the path loss experienced by a transmitted optical signal can be characterized by two components: attenuation loss and geometrical loss. The attenuation loss is caused primarily by the absorption and scattering of light in the water medium, whereas the geometrical loss arises as a result of the transmitted beam spreading and propagating between the ROV and US.

First, for the computation of the attenuation loss in water, many past studies have commonly adopted the Beer–Lambert (BL) formula [26], which assumes that the ROV and US are aligned perfectly. However, under the considered UOWC scenario, misalignment between the ROV and the US is unavoidable because unpredictable shaking and movement of the submerged ROV might occur due to various external factors, even when hovering. To include this issue, the inclination angle θ_0, which refers to the angular difference between the center of the transmit node's optical signal and the receiving node, is modeled as the Gaussian random variable with a mean of $\bar{\theta}_0$ and variance of $\sigma^2_{\theta_0}$ [27]. Therefore, the attenuation loss of an underwater optical signal, accounting for misalignment, can be expressed as

$$h_{\text{AL}} = \exp\left\{-c(\lambda)\frac{d}{cos(\theta_0)}\right\}, \tag{5}$$

where d is the distance between the ROV and the US and $c(\lambda)$ is the extinction coefficient which is defined as the summation of the absorption coefficient $a(\lambda)$ and scattering coefficient $b(\lambda)$, i.e., $c(\lambda) = a(\lambda) + b(\lambda)$. The absorption and scattering coefficients are affected by the wavelength of light and the type of water. Specific values of the absorption and scattering coefficients for different water types at a specific wavelength are presented in [9].

On the other hand, the geometrical loss of an underwater optical signal, accounting for the misalignment, can be expressed as [28]:

$$h_{\text{GL}} = \begin{cases} \frac{A_r cos(\theta_0)}{2\pi d^2[1-cos(\theta)]}, & \theta \geq \theta_0 \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where A_r is the aperture area of the selected optical receiver. It should be noted that h_{GL} can have a positive value only when the beam-divergence angle θ at the transmitter is equal to or greater than the inclination angle θ_0 (i.e., $\theta \geq \theta_0$).

Another important factor affecting the UOWC channel is underwater turbulence, which arises from refractive-index fluctuations caused by variations in the salinity and temperature of the water medium [6]. These turbulence-induced fluctuations result in fluctuations in the received signal intensity, commonly referred to as fading. In particular, in vertical underwater links, the salinity and temperature gradients change with depth. To model the channel such that this characteristic is considered, the underwater vertical link can be approximated as a series of non-mixing layers, each with different properties [7]. Nevertheless, since this work considers short-range communication (i.e., short link distance) between the two nodes for efficient SLIPT, a single underwater vertical link is considered. As such, under the assumption of weak turbulence, which is typically observed for short link distances [29], h_{F} can be modeled with a log-normal (LN) distribution, and its probability density function (PDF) is given by

$$f_{h_{\mathrm{F}}}(h_{\mathrm{F}}) = \frac{1}{h_{\mathrm{F}}\sqrt{2\pi(4\sigma_x^2)}} \exp\left(-\frac{(\ln(h_{\mathrm{F}}) - 2\mu_x)^2}{2(4\sigma_x^2)}\right), \tag{7}$$

where μ_x and σ_x are the mean and variance, respectively, of the log-amplitude coefficient $X = 0.5\ln(h_{\mathrm{F}})$. Because the fading coefficient does not change the average power, its amplitude should be normalized, i.e., $E[h_{\mathrm{F}}] = 1$, such that $\mu_x = -\sigma_x^2$ [30]. The variance of σ_x can be computed as

$$\sigma_x^2 = 0.25\ln\left(1 + \sigma_h^2\right), \tag{8}$$

where σ_h^2 is the scintillation index achieved under spherical waves. Under the assumption of quasi-static fading, h_{F} remains constant during the symbol interval.

3. Underwater Hybrid Time Switching–Power Splitting (TS–PS) SLIPT

Various SLIPT methods, such as AC–DC separation, TS, PS, and hybrid TS–PS SLIPT methods, have been introduced in [17]. Among them, we adopt the hybrid TS–PS SLIPT method, which, as the name implies, is a combination of the TS and PS methods. More specifically, we denote $T_{\mathrm{EH}} \leq T$ and $T_{\mathrm{ID}} \leq T$ as the time durations allocated to EH and information decoding (ID), respectively, within the given time duration of a data frame. Because $T_{\mathrm{EH}} + T_{\mathrm{ID}} = T$, these durations are defined by $T_{\mathrm{EH}} = (1-\tau)T$ and $T_{\mathrm{ID}} = \tau T$, respectively, where $\tau \in [0,1]$ denotes the time-switching factor. The hybrid TS–PS method consists of two phases and its brief procedure is described as follows:

1. In the first phase (referred to as EH mode), with a duration of T_{EH}, only EH is conducted (similar to the TS method). In this phase, there is no need to transfer information, such that DC bias can be set to its maximum value (i.e., $B = I_{\max}$) to maximize EH, which leads to $A = 0$. The AC component in this phase is blocked by an inductor, such that only the DC component (i.e., I'_{DC}) passes through the EH block as illustrated in Figure 2.
2. In the second phase (referred to as PS mode) with a duration of T_{ID}, the receiver conducts the PS method to perform EH and ID at the same time. For this, the received signal power is split into two streams using the power-splitting factor $\rho \in [0,1]$. As a result, $(1-\rho)i_{\mathrm{RX}}$ and ρi_{RX} are dedicated for EH and ID, respectively. Through the suppression of the AC or DC component, the inputs to the EH and ID blocks can be presented as $(1-\rho)(I'_{\mathrm{DC}} + \varsigma\bar{I}_{\mathrm{AC}})$ and $\rho(\bar{I}_{\mathrm{AC}} + n)$, respectively, where ς is the AC–DC conversion efficiency [31], and $\bar{I}_{\mathrm{AC}} = rh\delta\mathrm{E}[x] = rh\delta A/2$ is the average of the AC component.

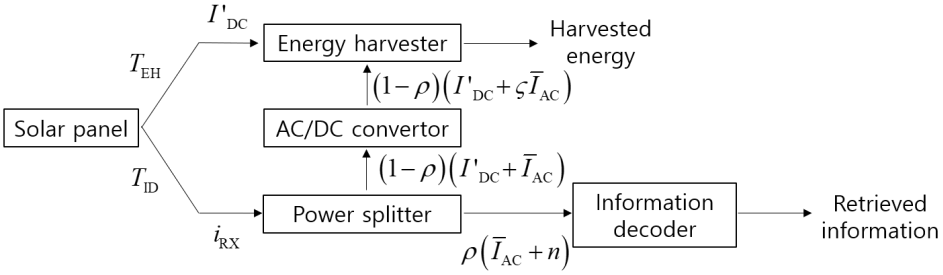


Figure 2. Receiver structure for hybrid TS–PS SLIPT method.

3.1. Performance Metric 1: Energy Harvesting

The maximum output power of a solar panel in the EH mode can be achieved at its maximum power point (MPP) [32] and is given as follows [33]:

$$P_{\mathrm{MPP}} = F I_{\mathrm{DC}} V_{\mathrm{OC}}. \tag{9}$$

In (9), V_{OC} is the open circuit voltage which is calculated as $V_{\mathrm{OC}} = V_t \ln(1 + I_{\mathrm{DC}}/I_0)$, where V_t is the thermal voltage and I_0 is the dark saturation current [34]. Furthermore, I_{DC} is defined as $I_{\mathrm{DC}} = I'_{\mathrm{DC}} + \iota\varsigma\bar{I}_{\mathrm{AC}}$, where ι (which takes a value of 0 or 1) is used to indicate whether or not AC component is used for EH, i.e., if only the DC component is used for EH, then ι is set to zero. Meanwhile, F is the fill factor defined as $F = I_{\mathrm{MPP}} V_{\mathrm{MPP}}/I_{\mathrm{DC}} V_{\mathrm{OC}}$, where I_{MPP} and V_{MPP} are the optimal values of MPP voltage and MPP current, respectively. The optimal values of I_{MPP} and V_{MPP} can be obtained automatically using dynamic tracking techniques for a given temperature and irradiance [35]. Based on (9), we can obtain the energy harvesting condition on I_{DC} by multiplying T_{EH} with P_{MPP} as follows

$$E = T_{\mathrm{EH}} P_{\mathrm{MPP}} = T_{\mathrm{EH}} F I_{\mathrm{DC}} V_t \ln\left(1 + \frac{I_{\mathrm{DC}}}{I_0}\right). \tag{10}$$

3.2. Performance Metric 2: Spectral Efficiency

Similar to [36], a low bound of SE (bps/Hz) for IM/DD systems conditioned on I_{DC} is expressed as follows:

$$\eta \geq \frac{T_{\mathrm{ID}}}{T}\left[\frac{1}{2}\log_2\left(1 + \frac{e}{2\pi}\beta\right)\right], \tag{11}$$

where β denotes the average electrical signal-to-noise ratio (SNR). Since the AC component carries the information, β can be expressed as follows:

$$\beta = \frac{\bar{I}_{\mathrm{AC}}^2}{\sigma^2} = \frac{(rh\delta \mathrm{E}[x])^2}{\sigma^2}, \tag{12}$$

where $\sigma^2 = N_0/2T_s$ is the noise variance and N_0 is the noise power spectral density.

4. Proposed Algorithm

In this study, the US conducts underwater SLIPT with the objective of maximizing EH while satisfying the SE requirement for control data reception by jointly controlling the TS and PS ratios. In this point, we emphasize that the EH, as well as SE at the US, is obviously affected by the received electrical signal power, which can vary with the beam-divergence angle chosen by the ROV.

Thus, to achieve this objective, we aim to jointly optimize not only the beam-divergence angle at the transmitting node (i.e., ROV) but also the TS and PS ratios at the receiving node (i.e., US) by solving the following optimization problem:

$$\begin{aligned} &\max_{\tau,\rho,\theta} E(\tau,\rho,\theta) \\ &\text{s.t. } \eta(\tau,\rho,\theta) \geq \eta_{th}, \\ &\quad 0 \leq \tau,\rho \leq 1, \\ &\quad \theta_{\min} \leq \theta \leq \theta_{\max}. \end{aligned} \tag{13}$$

To solve problem (13), we can decompose the problem into two subproblems. This is because when determining the beam-divergence angle, the TS and PS ratios do not have an impact. Conversely, when determining the TS and PS ratios, the beam-divergence angle does have an effect. More specifically, the first subproblem (P1) is to determine the beam-divergence angle (i.e., θ) of the ROV to maximize the received electrical signal power (3) at the US while maintaining a seamless connection between the two nodes:

$$\begin{aligned} &\text{(P1)} \ \max_{\theta} i_{\mathrm{RX}}(\theta) \\ &\text{s.t. } \theta_{\min} \leq \theta \leq \theta_{\max}. \end{aligned} \tag{14}$$

On the other hand, the second subproblem (P2) is to simultaneously determine both the TS and PS ratios (i.e., τ and ρ) at a given $i_{\mathrm{RX}}(\theta)$ to maximize EH while supporting the SE requirement at the US:

$$\begin{aligned} &\text{(P2)} \max_{\tau,\rho} E(\tau,\rho|i_{\mathrm{RX}}(\theta)) \\ &\text{s.t. } \eta(\tau,\rho) \geq \eta_{th}, \\ &\quad 0 \leq \tau,\rho \leq 1. \end{aligned} \tag{15}$$

Since each subproblem should be solved on different sides (i.e., (P1) and (P2) on the ROV and US sides, respectively), we refer to the ROV and US as agents, each of which solves the subproblem on its side.

4.1. ROV Agent: Beam-Divergence Angle Adaptation

The role of the ROV agent is to choose the beam-divergence angle θ at every time slot by learning the inclination angle θ_0 between the ROV and US. We note that the inclination angle between the two nodes can vary at every time slot because of the unpredictable shaking and movement of the submerged ROV due to various external factors in the ocean. To solve problem (P1) for an underwater environment, we can define a Markov decision process (MDP) for the ROV agent, which can be represented as a tuple $\left(\mathbf{S}^{\mathrm{ROV}}, \mathbf{A}^{\mathrm{ROV}}, r^{\mathrm{ROV}}\right)$ where $\mathbf{S}^{\mathrm{ROV}}$, $\mathbf{A}^{\mathrm{ROV}}$, and r^{ROV} refer to the state space, action space, and reward for the ROV agent.

We denote $a^{\mathrm{ROV}}(t) = \theta(t) \in \mathbf{A}^{\mathrm{ROV}}$ as an element in action space $\mathbf{A}^{\mathrm{ROV}}$ which represents a set of discrete beam-divergence angles in degrees that the ROV can choose at time:

$$\mathbf{A}^{\mathrm{ROV}}(t) = \{\theta_{\min}, \theta_{\min} + \alpha, ..., \theta_{\max} - \alpha, \theta_{\max}\}, \tag{16}$$

where $\theta_{\min}$ and $\theta_{\max}$ are the minimum and maximum beam-divergence angles supported by the optical modem installed on the ROV, respectively, and α is the gap between two consecutive beam-divergence angles.

We also denote $\mathbf{S}^{\mathrm{ROV}}(t)$ as the state space at time slot t, which includes various information that affects the action selection of the ROV agent:

$$\mathbf{S}^{\mathrm{ROV}}(t) = \left\{\mathbf{s}_{\mathrm{his}}^{\mathrm{ROV}}(t-1), \theta_0(t-1), \theta_{\mathrm{gap}}(t-1)\right\}, \tag{17}$$

where $\mathbf{s}_{\mathrm{his}}^{\mathrm{ROV}}(t-1)$ contains historical information on the actions and rewards experienced by the ROV agent, $\theta_0(t-1)$ refers to the inclination angle at time slot $t-1$, and $\theta_{\mathrm{gap}}(t-1)$ refers to the difference between $\theta(t-1)$ and $\theta_0(t-1)$. When defining $\mathbf{s}_{\mathrm{his}}^{\mathrm{ROV}}(t-1)$, we

adopt the concept of a sliding window of size l at each time slot to limit the size of the state space as the learning progresses. Thus, $\mathbf{s}_{\text{his}}(t-1)$ can be expressed as follows:

$$\mathbf{s}_{\text{his}}^{\text{ROV}}(t-1) = \left\{a^{\text{ROV}}(t-l), r^{\text{ROV}}(t-l), ..., a^{\text{ROV}}(t-1), r^{\text{ROV}}(t-1)\right\}. \tag{18}$$

For the reward function of the ROV agent, we adopt the received electrical signal power defined in (4) under the chosen action as follows:

$$r^{\text{ROV}}(t) = i_{\text{RX}}(\theta(t)). \tag{19}$$

The reward data obtained at the US are fed back to the ROV during the transmission of sensing data from the US to the ROV via LED-based UOWC.

4.2. US Agent: SLIPT Adaptation

Given the received electrical signal power $i_{\text{RX}}(\theta(t))$ affected by the action of the ROV agent, the US agent aims to maximize EH at the US while guaranteeing the SE requirement for control data transmission. For this purpose, we can also define an MDP for the US agent, consisting of a tuple $\left(\mathbf{S}^{\text{US}}, \mathbf{A}^{\text{US}}, r^{\text{US}}\right)$ where $\mathbf{S}^{\text{US}}$, $\mathbf{A}^{\text{US}}$, and r^{US} refer to the state space, action space, and reward for the US agent.

Unlike the ROV agent, which chooses only one discrete value (i.e., beam-divergence angle), the US agent needs to determine two different continuous values, i.e., the TS and PS ratios, such that the action space can be defined as follows:

$$\mathbf{A}^{\text{US}}(t) = \{\tau, \rho\}. \tag{20}$$

The state space for the US agent includes historical information on the actions and rewards experienced by the US agent as well as current channel quality between the two nodes as follows:

$$\mathbf{S}^{\text{US}}(t) = \left\{\mathbf{s}_{\text{his}}^{\text{US}}(t-1), h(t)\right\}. \tag{21}$$

Similar to that done for the ROV agent, the concept of a sliding window of size l is adopted to define $\mathbf{s}_{\text{his}}^{\text{US}}(t-1)$ as follows:

$$\mathbf{s}_{\text{his}}^{\text{US}}(t-1) = \left\{a^{\text{US}}(t-l), r^{\text{US}}(t-l), ..., a^{\text{US}}(t-1), r^{\text{US}}(t-1)\right\}. \tag{22}$$

To achieve the objective of the ROV agent, we define a reward function, which is influenced by the chosen action set (i.e., $\{\tau, \rho\}$):

$$r^{\text{US}}(t) = \begin{cases} E(\tau, \rho|\theta) & , \ \eta \geq \eta_{th} \\ \eta(\tau, \rho|\theta) - \eta_{th} & , \text{ otherwise.} \end{cases} \tag{23}$$

The aforementioned reward indicates that if the SE requirement is satisfied, the reward is set to the EH at the US. Otherwise, the reward sets the difference between the achieved and required SE to prevent the achieved value from becoming smaller than the required value.

4.3. Proposed Algorithm

To obtain the solutions of such two MDPs, this study proposes a hierarchical DQN–DDPG-based online learning algorithm that determines not only the beam-divergence angle at the ROV but also the TS and PS ratios at the US in real time. The conceptual structure of the proposed hierarchical DQN–DDPG-based online learning algorithm is illustrated in Figure 3.

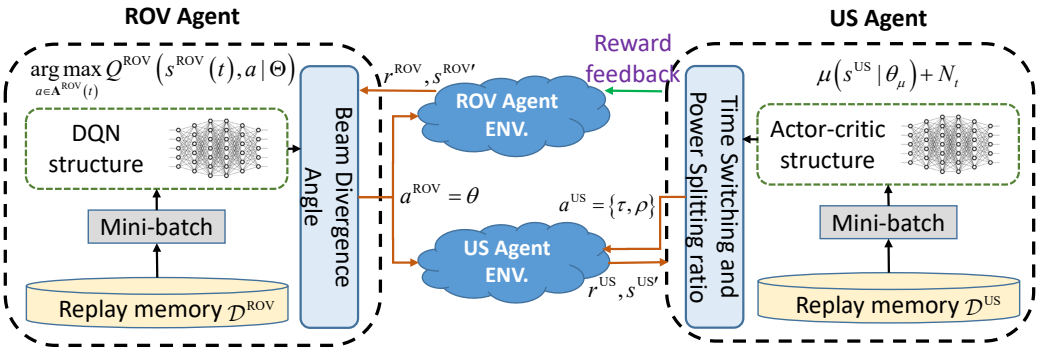


Figure 3. Structure of the proposed hierarchical DQN–DDPG-based online learning algorithm.

First, since the role of the ROV agent is to determine the discrete beam-divergence angle, it adopts the DQN algorithm [37]. At each time t, the ROV agent constructs state $s^{\mathrm{ROV}}(t) \in \mathbf{S}^{\mathrm{ROV}}$, for which it needs to gather historical information on its reward from the US through feedback. As mentioned previously, although such feedback data are transmitted together with the sensing data, LED-based UOWC from the US to the ROV can offer a sufficient data rate (e.g., more than Mbps) to transmit them. After constructing $s^{\mathrm{ROV}}(t)$, the ROV agent makes a decision to choose the beam-divergence angle based on the ε-greedy algorithm. The ROV agent chooses its action by the following equation with probability $1-\varepsilon$:

$$a^{\mathrm{ROV}}(t) = \underset{a \in \mathbf{A}^{\mathrm{ROV}}(t)}{\arg\max}\, Q^{\mathrm{ROV}}\left(s^{\mathrm{ROV}}(t), a|\Theta\right), \tag{24}$$

where $Q^{\mathrm{ROV}}\left(s^{\mathrm{ROV}}(t), a\right)$ is the Q-function achieved by action a in state $s^{\mathrm{ROV}}(t)$, and Θ is a set of weights for the deep neural network (DNN) of the ROV agent. With probability ε, it randomly chooses its action in the action space $\mathbf{A}^{\mathrm{ROV}}(t)$. At each time slot, the weights of the DNN are updated using the mean-squared error (MSE) loss function as follows:

$$\mathcal{L}^{\mathrm{ROV}}(\Theta) = \underset{e^{\mathrm{ROV}} \sim \mathcal{D}^{\mathrm{ROV}}}{\mathbb{E}}\left[y^{\mathrm{ROV}} - Q^{\mathrm{ROV}}\left(s^{\mathrm{ROV}}(t), a^{\mathrm{ROV}}(t)|\Theta\right)\right]^2, \tag{25}$$

where $\mathcal{D}^{\mathrm{ROV}}$ denotes the experience replay buffer for the ROV agent which contains tuples $e^{\mathrm{ROV}} = \left(s^{\mathrm{ROV}}, a^{\mathrm{ROV}}, r^{\mathrm{ROV}}, s^{\mathrm{ROV}\prime}\right)$. Meanwhile, y^{ROV} is the target value for updating Θ, and is expressed as follows:

$$y^{\mathrm{ROV}} = r^{\mathrm{ROV}} + \gamma \max_{a} Q^{\mathrm{ROV}}\left(s^{\mathrm{ROV}\prime}, a|\tilde{\Theta}\right), \tag{26}$$

where γ denotes the discount factor, and $\tilde{\Theta}$ denotes the set of weights for the target network. Algorithm 1 describes the DQN algorithm implemented by the ROV agent.

On the other hand, unlike the ROV agent, which chooses a discrete value as an action, the role of the US agent is to determine two contribution values (i.e., TS and PS ratio) with the same bound (i.e., $0 \le \tau, \rho \le 1$) as an action. For this, we adopt the DDPG algorithm [38], which is one of the representative deep RL (DRL) algorithms for finding a continuous action vector. Let $Q^{\mathrm{US}}\left(s, a|\theta_Q\right)$ be a critic network with weights θ_Q that estimate the Q-function. Additionally, let $\mu\left(s|\theta_\mu\right)$ be the actor network with weights θ_μ which specifies the current policy by deterministically mapping states to a specific action. Then, the gradient of the accumulated discounted reward (denoted as J) can be expressed as follows:

$$\nabla_{\theta_\mu} J = \underset{e^{\mathrm{US}} \sim \mathcal{D}^{\mathrm{US}}}{\mathbb{E}}\left[\nabla_{\theta_\mu} \mu\left(s|\theta_\mu\right)|_{s=s^{\mathrm{US}}} \nabla_a Q^{\mathrm{US}}\left(s, a|\theta_Q\right)|_{s=s^{\mathrm{US}}, a=\mu\left(s^{\mathrm{US}}\right)}\right], \tag{27}$$

where $\mathcal{D}^{\mathrm{US}}$ denotes the experience replay buffer for the US agent, which contains tuples $e^{\mathrm{US}} = \left(s^{\mathrm{US}}, a^{\mathrm{US}}, r^{\mathrm{US}}, s^{\mathrm{US}\prime}\right)$. Similar to the ROV agent, the US agent updates its Q-function by minimizing the MSE loss function as follows:

$$\mathcal{L}^{\mathrm{US}}(\theta_Q) = \mathop{\mathbb{E}}_{e^{\mathrm{US}} \sim \mathcal{D}^{\mathrm{US}}} \left[\left(y^{\mathrm{US}} - Q^{\mathrm{US}}\left(s^{\mathrm{US}}, a^{\mathrm{US}} | \theta_Q\right) \right)^2 \right], \tag{28}$$

where y^{US} is the target value for updating Q^{US}, and is also expressed as follows:

$$y^{\mathrm{US}} = r^{\mathrm{US}} + \gamma Q^{\mathrm{US}}\left(s^{\mathrm{US}\prime}, \mu\left(s^{\mathrm{US}\prime} | \tilde{\theta}_\mu\right) | \tilde{\theta}_Q\right), \tag{29}$$

where $\tilde{\theta}_Q$ and $\tilde{\theta}_\mu$ are the sets of the weights of the target network with respect to Q^{US} and μ, respectively. Algorithm 2 explains the DDPG algorithm implemented by the US agent.

In Algorithm 2, $\mathcal{N}$ is the noise process for constructing the exploration policy, φ is a predetermined value for the repetitive initialization of $\mathcal{N}$, and ω is the weight for soft target updates, and X is the number of samples in the mini-batch.

To facilitate a clear understanding of Algorithms 1 and 2 described earlier, we have provided flow charts for each algorithm, as shown in Figure 4.

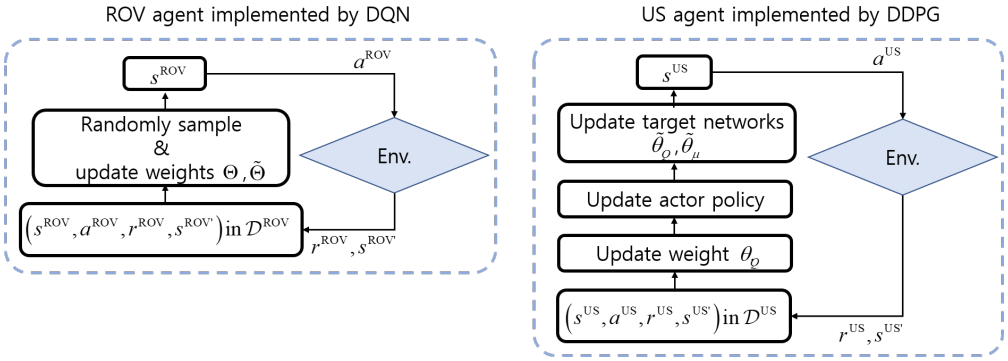


Figure 4. Flow charts of Algorithms 1 and 2.

Algorithm 1: DQN algorithm for determining the beam-divergence angle at the ROV agent.

Establish a DQN with weights Θ and a target DQN with weights $\tilde{\Theta}$, and experience replay buffer $\mathcal{D}^{\mathrm{ROV}}$ at the ROV.
Initialize Θ and set $\tilde{\Theta} \leftarrow \Theta$
for $t = 1$ to T **do**
  The ROV agent construct the state space s^{ROV}.
  ROV agent chooses action a^{ROV}, through ϵ-greedy policy, under which the agent selects action $a^{\mathrm{ROV}} = \arg\max_{a \in \mathbf{A}^{\mathrm{ROV}}} Q^{\mathrm{ROV}}\left(s^{\mathrm{ROV}}, a | \Theta\right)$ with probability $1 - \epsilon$, and randomly selects action a^{ROV} in action space $\mathbf{A}^{\mathrm{ROV}}$ with probability ϵ.
  Executes an action a^{ROV}, and then observes reward r^{ROV} and the new state $s^{\mathrm{ROV}\prime}$.
  Store $e^{\mathrm{ROV}} = \left(s^{\mathrm{ROV}}, a^{\mathrm{ROV}}, r^{\mathrm{ROV}}, s^{\mathrm{ROV}\prime}\right)$ in $\mathcal{D}^{\mathrm{ROV}}$.
  The ROV agent randomly samples a mini-batch from $\mathcal{D}^{\mathrm{ROV}}$, and updates weights Θ.
  In every predetermined time interval, the ROV agent updates the target DQN with $\tilde{\Theta} \leftarrow \Theta$.
end

Algorithm 2: DDPG algorithm for determining TS and PS ratios at the US agent.

Establish critic network $Q^{\mathrm{US}}\left(s^{\mathrm{US}}, a^{\mathrm{US}}|\theta_Q\right)$ with weights θ_Q, actor network $\mu\left(s|\theta_\mu\right)$ with weights θ_μ, and experience replay buffer $\mathcal{D}^{\mathrm{US}}$ at the US.
Establish the target networks of critic and actor networks with weights $\tilde{\theta}_Q$ and $\tilde{\theta}_\mu$.
Set $\theta_z^{Q'} \leftarrow \theta_z^Q$ and $\theta_z^{\mu'} \leftarrow \theta_z^\mu$
Initialize a random process $\mathcal{N}$ for action exploration.
for $t = 1$ to T **do**
  The US agent construct the state space s^{US}.
  Select action $a^{\mathrm{US}} = \mu\left(s^{\mathrm{US}}|\theta_\mu\right) + \mathcal{N}_t$ regarding the current policy and exploration.
  Execute a^{US}, and observe reward r^{US} and new state $s^{\mathrm{US}\prime}$.
  Store $\left(s^{\mathrm{US}}, a^{\mathrm{US}}, r^{\mathrm{US}}, s^{\mathrm{US}\prime}\right)$ in reply buffer $\mathcal{D}^{\mathrm{US}}$.
  The US agent randomly samples a mini-batch of X transitions $\left(s^{\mathrm{US}}, a^{\mathrm{US}}, r^{\mathrm{US}}, s^{\mathrm{US}\prime}\right)$ from $\mathcal{D}^{\mathrm{US}}$.
  Set $y^{\mathrm{US}} = r^{\mathrm{US}} + \gamma Q^{\mathrm{US}}\left(s^{\mathrm{US}\prime}, \mu\left(s^{\mathrm{US}\prime}|\tilde{\theta}_\mu\right)|\tilde{\theta}_Q\right)$. Then, updates the weights of critic θ_Q by minimizing the following loss: $\mathcal{L}^{\mathrm{US}} = \frac{1}{\mathrm{X}}\sum\left(y^{\mathrm{US}} - Q^{\mathrm{US}}\left(s^{\mathrm{US}}, a^{\mathrm{US}}|\theta_Q\right)\right)^2$.
  Update the actor policy using the sampled policy gradient: $\nabla_{\theta_\mu} J \approx \frac{1}{\mathrm{X}}\sum \nabla_{\theta_\mu}\mu\left(s|\theta_\mu\right)|_{s=s^{\mathrm{US}}} \nabla_a Q^{\mathrm{US}}\left(s, a|\theta_Q\right)|_{s=s^{\mathrm{US}}, a=\mu\left(s^{\mathrm{US}}\right)}$.
  Update the target networks: $\tilde{\theta}_Q \leftarrow \omega\theta_Q + (1-\omega)\tilde{\theta}_Q$ and $\tilde{\theta}_\mu \leftarrow \omega\theta_\mu + (1-\omega)\tilde{\theta}_\mu$.
  if $T \,\%\, \varphi = 0$ **then**
    Initialize a random process $\mathcal{N}$ for action exploration.
  end
end

4.4. Verification for Online Operation via Computational Complexity Analysis

We analyze the time computational complexity of the proposed hierarchical DQN–DDPG algorithm using big O notation denoted by O[·].

In the training stage, the ROV agent first executes Algorithm 1 (i.e., beam-divergence angle decision algorithm), which is based on the DQN consisting of two DNNs having the same structure, e.g., a Q-network and a target network. Let L^{DQN} and m_l^{DQN} as the number of layers of the DNNs and the number of neurons in the l-th layer among L^{DQN} layers. According to [39], the computational complexity of each training step in the DNN using a fully connected network can be presented as $O\left[\mathrm{Y}^{\mathrm{DQN}}\sum_{l=1}^{L^{\mathrm{DQN}}-1} m_l^{\mathrm{DQN}} m_{l+1}^{\mathrm{DQN}}\right]$, where $\mathrm{Y}^{\mathrm{DQN}}$ is the mini-batch size of the DQN. Thus, total training computational complexity of Algorithm 1 is $O\left[2T_{\mathrm{CV}}\mathrm{Y}^{\mathrm{DQN}}\sum_{l=1}^{L^{\mathrm{DQN}}-1} m_l^{\mathrm{DQN}} m_{l+1}^{\mathrm{DQN}}\right]$, where T_{CV} is the number of time slots until performance of the hierarchical DQN–DDPG algorithm converges. Next, the US agent executes Algorithm 2 (i.e., TR and PS ratios decision algorithm), which is based on the DDPG network consisting of two DNNs with a different structure, e.g., an actor network and a critic network. Assuming that that the actor and critic networks contain L^{ACT} and L^{CRIC} fully connected layers, respectively, total training computational complexity of Algorithm 2 is $O\left[T_{\mathrm{CV}}\mathrm{Y}^{\mathrm{DDPG}}\left(\sum_{l=1}^{L^{\mathrm{ACT}}-1} m_l^{\mathrm{ACT}} m_{l+1}^{\mathrm{ACT}} + \sum_{l=1}^{L^{\mathrm{CRIC}}-1} m_l^{\mathrm{CRIC}} m_{l+1}^{\mathrm{CRIC}}\right)\right]$, where m_l^{ACT} and m_l^{CRIC} are the numbers of neurons in the l-th layer among L^{ACT} and L^{CRIC} layers, respectively: $\mathrm{Y}^{\mathrm{DDPG}}$ is the mini-batch size of the DDPG network. As a result, in the training

stage, total computational complexity of the proposed hierarchical DQN–DDPG algorithm is $O\left[2T_{\mathrm{CV}}\mathrm{Y}^{\mathrm{DQN}}\sum_{l=1}^{L^{\mathrm{DQN}}-1}m_l^{\mathrm{DQN}}m_{l+1}^{\mathrm{DQN}}\right]+O\left[T_{\mathrm{CV}}\mathrm{Y}^{\mathrm{ACT}}\left(\sum_{l=1}^{L^{\mathrm{ACT}}-1}m_l^{\mathrm{ACT}}m_{l+1}^{\mathrm{ACT}}+\sum_{l=1}^{L^{\mathrm{CRIC}}-1}m_l^{\mathrm{CRIC}}m_{l+1}^{\mathrm{CRIC}}\right)\right]$.

In the test stage, the computational complexity of the proposed hierarchical DQN–DDPG algorithm in each time slot can be dramatically reduced to $O\left[\sum_{l=1}^{L^{\mathrm{DQN}}-1}m_l^{\mathrm{DQN}}m_{l+1}^{\mathrm{DQN}}\right]+O\left[\left(\sum_{l=1}^{L^{\mathrm{ACT}}-1}m_l^{\mathrm{ACT}}m_{l+1}^{\mathrm{ACT}}\right)\right]$. This is because once the performances of the networks finally converge, we do not need iterations for updating the target network for DQN and the critic network for the DDPG network with reply buffers, respectively. This indicates that the proposed algorithm is capable of being implemented in real-time operations.

5. Simulation Results

To assess the validity of the proposed algorithm, we conduct numerical simulations to evaluate the performance of the proposed hierarchical DQN–DDPG algorithm and then compare it with those of a number of already existing UOWC algorithms. For simulations, a UOWC scenario between the ROV and the US is considered in which the ROV is randomly shaking even when hovering. As aforementioned, the shaking of the ROV is affected by the inclination angle θ_0 which is modeled as a Gaussian random variable with a mean of $\bar{\theta}_0$ and variance of $\sigma_{\theta_0}^2$.

5.1. Simulation Parameters

For numerical simulations, we set static parameters as follows: $\theta_{\min}=3^\circ$, $\theta_{\max}=5^\circ$, $\varsigma=1$, $T=1$ s, $\bar{\theta}=2.5^\circ$, $\sigma_{\theta_0}=0.5$, and $d=10$ m, respectively. Also, the size of sliding window l for $\mathbf{s}_{\mathrm{his}}^{\mathrm{ROV}}(t-1)$ and $\mathbf{s}_{\mathrm{his}}^{\mathrm{US}}(t-1)$ is set to 3. Other static system and channel parameters adopted for our simulations are from [17], which are summarized in Table 1.

Table 1. List of static network and channel parameters [17].

Parameter	Value
Time duration of a data frame T	1 s
Distance between ROV and US	10 m
Receiver Aperture diameter A_r	0.2 m
Extinction coefficient $c(\lambda)$	0.15 (clean ocean)
Solar panel responsibility r	0.6 A/W
Slope efficiency of LD δ	1.33 W/A
Maximum input bias current $I_{\max}$	1.2 A
Minimum input bias current $I_{\min}$	0.2 A
Fill factor F	0.75
Thermal voltage V_t	0.025 W
Dark saturation current I_0	10^{-9} A
Noise power spectral density N_0	10^{-19} W/Hz
Mean of inclination angle $\bar{\theta}_0$	0.0436 rad
Standard deviation of inclination angle σ_{θ_0}	0.0087 rad

Regarding the learning environment, the DQN structure for the ROV agent is a fully connected neural network with two hidden layers containing 68 neurons in each layer. Other learning hyperparameters for the DQN are summarized in Table 2.

Table 2. List of DQN hyperparameters.

Hyperparameter	Agent
ϵ for ϵ-greedy	0.01
Mini-batch size	64
Optimizer	Adam
Activation function	Relu
Learning rate	10^{-4}
Experience replay buffer size	2000
Discount factor	0.99
Considered time slots for $\mathbf{s}_{\text{his}}(t)$	2

On the other hand, the DDPG structure for the US agent is a fully connected neural network with two hidden layers, where the first and second hidden layers contain 512 and 256 neurons, respectively. Other learning hyperparameters for DDPG are summarized in Table 3. These two DRL agents are implemented using Keras Python libraries with a TensorFlow backend.

Table 3. List of hyperparameters for DDPG networks.

Parameter	Value
Mini-batch size	64
Experience replay buffer size	10^6
Discount factor	0.99
Learning rate of actor	10^{-4}
Learning rate of critic	3×10^{-4}
Soft update rate of target parameters	2×10^{-1}

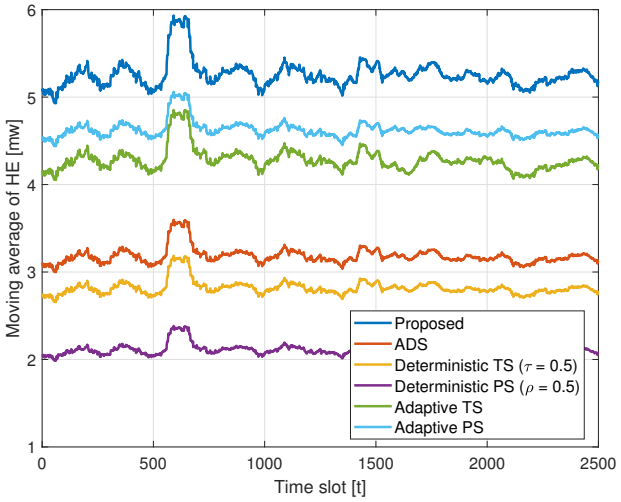
5.2. Benchmark Algorithms

For the performance comparison, the proposed algorithm is compared with these already existing SLIPT algorithms: AC–DC separation (ADS) method [40], TS method [41], and PS method [42]. In the ADS algorithm, the AC (e.g., $I_{\text{AC}} + n$) and DC (e.g., I'_{DC}) components of the received signal (4) are separated by the capacitor and inductor and are then used for ID and EH at the receiver, respectively. Since ID and EH are conducted simultaneously at the receiver, we set $T_{\text{ID}} = T_{\text{EH}} = T$. In the TS method, the receiver switches only in time between the ID and EH modes by a factor of τ. In other words, the TS method is equivalent to executing only the first phase (i.e., EH method) of the hybrid TS–PS method described in Section 3. On the other hand, in the PS method, the received electrical power i_{RX} is split into two streams with a factor of ρ, i.e., $(1-\rho)i_{\text{RX}}$ and ρi_{RX}, which are used for EH and ID, respectively, during the time duration of a data frame ($T_{\text{ID}} = T_{\text{EH}} = T$). In other words, the PS method is equivalent to executing only the second phase (i.e., PS method) of the hybrid TS–PS method.

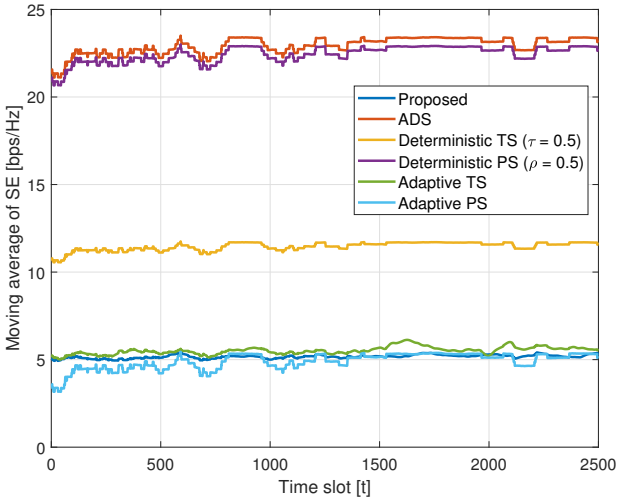
5.3. Simulation Results

Figure 5 compares the performance of the proposed algorithm with those of existing algorithms at a given $\eta = 5$ [bps/Hz]. In the figure, the deterministic TS and PS methods

execute the TS and PS algorithms with deterministic TS and PS ratios (e.g., $\tau = \rho = 0.5$ in this simulation). By contrast, the adaptive TS and PS methods refer to the DRL algorithm, which adaptively determines only the TS and PS ratios, respectively. For a fair comparison, we set the state space and reward function of the adaptive TS and PS methods to be the same as those of the US agent of the proposed algorithm. Regarding the performance metrics for comparison, we employ the moving averages of EH and SE with window sizes of 100. From Figure 5, it can be observed that the proposed algorithm achieves at least an 11% improvement in EH performance while meeting the communication requirements at the US, compared to benchmark SLIPT algorithms. This is because, by learning a time-varying underwater environment, the proposed algorithm can conduct online and adaptive control of the optimization parameters (e.g., TS and PS ratios and beam-divergence angle) in (13) to achieve our objective. Furthermore, although both the adaptive TS and PS methods can fulfill the SE requirement, the EH performance of the adaptive PS method is better than that of the adaptive TS method under the given SE requirement. The difference in EH performance between the two methods can vary according to changes in the SE requirement at the US. By contrast, in the case of the deterministic TS and PS methods, they cannot fulfill the SE requirement, which limits the utilization of these algorithms in the considered network scenario.



(**a**) Moving average of HE



(**b**) Moving average of SE

Figure 5. Performances of the proposed and existing algorithms.

Figure 6 presents the performance of the proposed algorithm with respect to variations in the SE requirement at the US, demonstrating that the proposed algorithm can fulfill a variety of SE requirements. Moreover, it is observed that as the SE requirement increases, the EH performance at the US decreases. This is because, to achieve a higher SE requirement, more time or more power should be allocated, which leads to a decrease in the amount of EH.

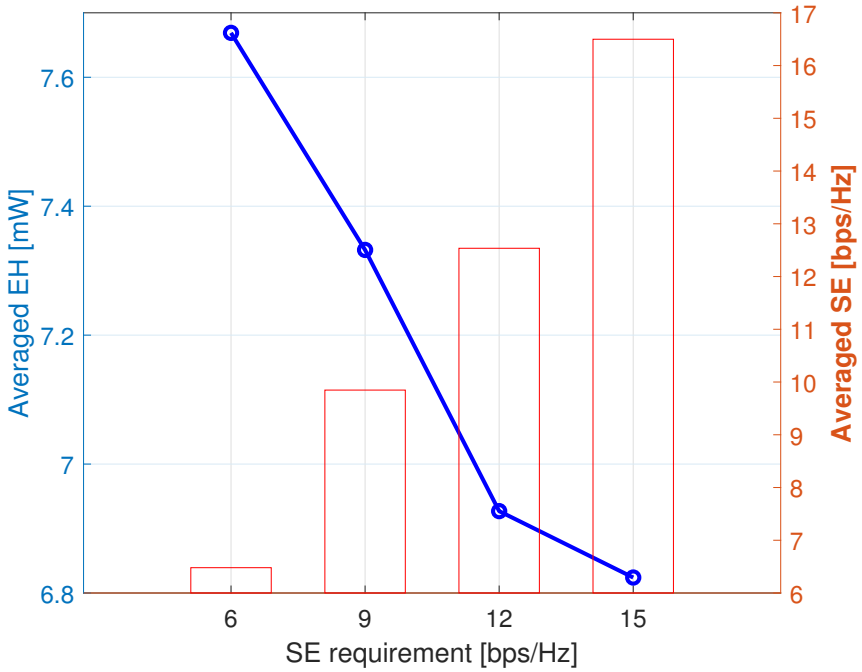


Figure 6. Performance of the proposed algorithm according to a change in SE requirement.

Figure 7 presents a performance comparison between the proposed algorithm when both the ROV and US agents are considered (hereafter referred to as the proposed algorithm with two agents) and the proposed algorithm when only the US agent is considered (hereafter referred to as the proposed algorithm with only the US agent) under given $\theta = \theta_{\min}$ and $\theta = \theta_{\max}$, respectively. From Figure 7, it can be observed that the proposed algorithm with two agents exhibits the best EH performance while guaranteeing SE requirements. By contrast, the constrained case, i.e., the proposed algorithm with only the US agent under given $\theta = \theta_{\min}$, exhibits severe performance degradation. This is because, at $\theta = \theta_{\min}$, cases wherein the chosen beam-divergence angle is smaller than the inclination angle (i.e., $\theta < \theta_0$) frequently occur due to irregular shaking or movement, which results in $i_r = 0$. Moreover, although the proposed algorithm with only the US agent under given $\theta = \theta_{\max}$ supports a stable and better EH performance compared with that under given $\theta = \theta_{\min}$, its performance is still worse than that of the proposed algorithm with two agents. This is because although the maximum beam-divergence angle may have an advantage for a seamless connection, it offers the worst SNR performance when the link is connected. This result demonstrates the validity of the adaptive control of the beam-divergence angle at the ROV in the proposed algorithm.

Furthermore, to check whether the beam-divergence angle chosen at the ROV implementing the proposed algorithm is adaptively controlled depending on the degree of misalignment, we conduct simulations that measure the average of the beam-divergence angles chosen at the ROV agent for 10,000 time slots with respect to changes in the mean of inclination angle, i.e., $\bar{\theta}_0$. In Figure 8, it can be observed that as the degree of misalignment between the ROV and US is growing larger (i.e., increase in $\bar{\theta}_0$), the ROV chooses a wider beam-divergence angle. For example, when there is a slight misalignment (i.e., $\bar{\theta}_0 = 2$), the ROV chooses the minimum beam-divergence angle, i.e., $\theta = \theta_{\min} = 3$. By contrast, as the change in misalignment becomes larger, a wider beam is selected, and eventually,

the widest beam, i.e., $\theta = \theta_{\max} = 5$, is selected. This tendency is reasonable because the proposes algorithm sets the beam-divergence angle as narrow as possible to maximize the SNR under the assumption of no change in the misalignment; however, in the opposite case, it sets the beam as wide as possible to prevent disconnection between the two nodes.

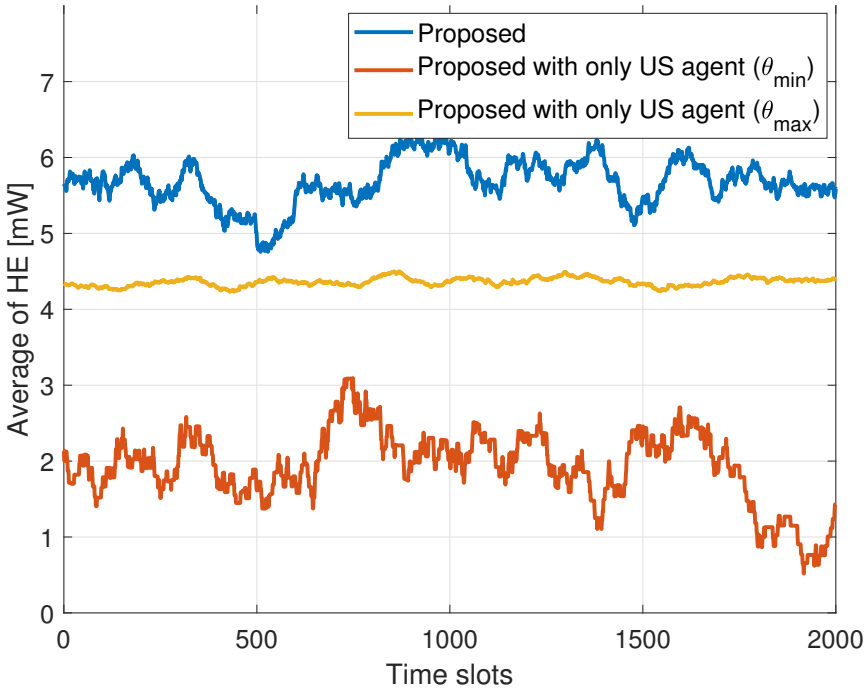


Figure 7. Performance comparison between the proposed algorithms.

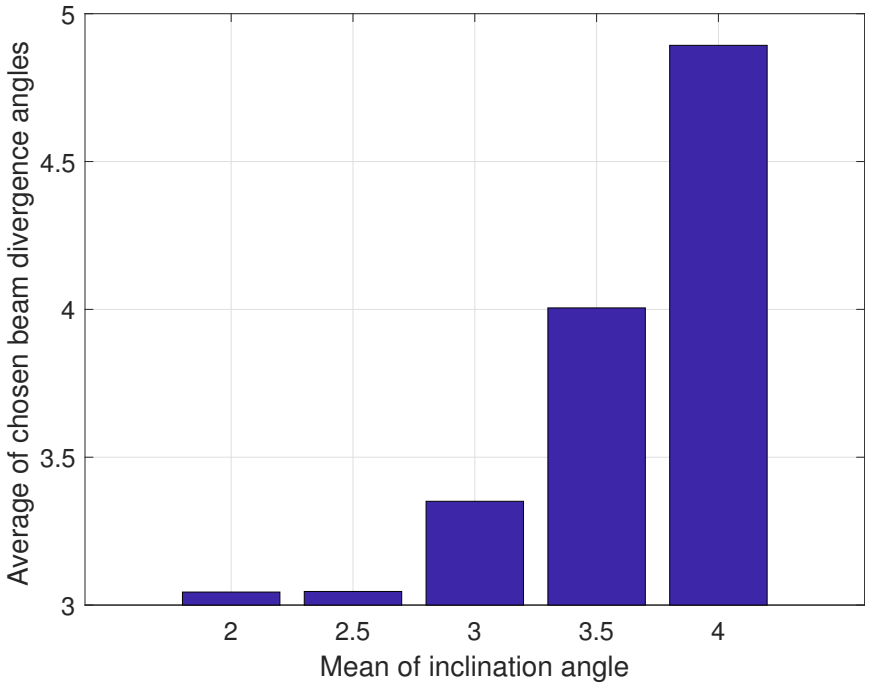


Figure 8. Averaged beam-divergence angle according to a change in the mean of inclination angle.

6. Conclusions

This work studied an adaptive control mechanism for a UOWC between the ROV and a US endowed with SLIPT capabilities. The primary goal is to maximize EH at the US while sustaining a predefined SE performance level between the two nodes. To address this

objective, we proposed a hierarchical DQN–DDPG-based online algorithm that involves two RL agents: the ROV agent, which optimizes the beam-divergence angle to enhance the received optical power at the US while maintaining an uninterrupted optical link, and the US agent, which determines the TS and PS ratios to maximize EH without compromising the SE requirements. Extensive simulation results demonstrated the effectiveness of the proposed algorithm, achieving at least an 11% improvement in EH performance while meeting the communication requirements at the US, compared to benchmark SLIPT algorithms. The adaptability of the algorithm to dynamically adjust optimization parameters in response to varying underwater environmental conditions and user requirements enhances the integration of energy transfer and communication in underwater contexts. Furthermore, the exploration of additional communication performance requirements within the proposed optimization framework will be addressed as part of future research.

Author Contributions: Conceptualization, Y.S.; Validation, S.J. and S.B.; Investigation, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2024 Yeungnam University Research Grant, South Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UOWC	Underwater optical wireless communication
SLIPT	Simultaneous lightwave information and power transfer
SWIPT	Simultaneous wireless information and power transfer
TS	Time-switching
LD	Laser diode
EH	Energy harvesting
SE	Spectral efficiency
PS	Power splitting
ROV	Remotely operated vehicle
US	Underwater sensor
LED	Light-emitting diode
DQN	Deep Q-network
DDPG	Deep deterministic policy gradient
RL	Reinforcement learning
3D	Three-dimensional
LARS	Launch and recovery system
IM/DD	Intensity modulation and direct detection
PAM	Pulse amplitude modulation
AWGN	Additive white Gaussian noise
BL	Beer-Lamber
LN	Log-normal
PDF	Probability density function
ID	Information decoding
MPP	Maximum power point
SNR	Signal-to-noise ratio
MDP	Markov decision process
DNN	Deep neural network
MSE	Mean-squared error
DRL	Deep-reinforcement learning
ADS	AC–DC separation

Glossary

Explanation of Key Symbols			
Symbols	**Explanation**	**Symbols**	**Explanation**
M	Modulation level	T	Time duration of data frame
T_s	Symbol interval	x	M-PAM symbol
A	Peak amplitude	$I_{\max}$	maximum input bias current
$I_{\min}$	minimum input bias current	P_{tx}	Instantaneous emitted optical intensity signal
δ	Slope efficiency of LD	B	DC bias
P_{rx}	Instantaneous received optical power	h	Underwater channel coefficient
h_{AL}	Attenuation loss	h_{GL}	Geometrical loss
h_{F}	Fading	i_{RX}	Received electrical signal
r	Solar panel responsivity	n	Additive white Gaussian noise
I'_{DC}	AC component of received signal	I_{AC}	DC component of received signal
θ_0	Inclination angle	d	Distance between ROV and US
$c(\lambda)$	Attenuation coefficient	$a(\lambda)$	Absorption coefficient
$b(\lambda)$	Scattering coefficient	A_r	Aperture area
θ	Beam-divergence angle	X	Log-amplitude coefficient
σ_h^2	Scintillation index	T_{EH}	Time duration of EH
T_{ID}	Time duration of ID	τ	time-switching factor
ρ	power-splitting factor	ς	AC-to-DC conversion efficiency
P_{MPP}	Maximum output power	V_{OC}	Open circuit voltage
V_t	Thermal voltage	I_0	Dark saturation current
ι	Indicator for use of AC component in EH	F	Fill factor
I_{MPP}	optimal value of MPP voltage	V_{MPP}	optimal value of MPP current
β	Average electrical SNR	N_0	Noise power spectral density
η	Low bound of SE	S^{ROV}	State space of ROV
S^{US}	State space of US agent	A^{ROV}	Action space of ROV agent
A^{US}	Action space of US agent	r^{ROV}	Reward of ROV agent
r^{US}	Reward of US agent	$\theta_{\min}$	Minimum beam-divergence angle
$\theta_{\max}$	maximum beam-divergence angle	t	Time slot
s_{his}	Historical information	l	Sliding window size
Θ	Set of weights	$\mathcal{L}$	Mean-squared error loss function
$\mathcal{D}$	Experience replay buffer	y^{ROV}	Target value of ROV agent
y^{US}	Target value of US agent	γ	discount factor

References

1. Kaushal, H.; Kaddoum, G. Underwater Optical Wireless Communication. *IEEE Access* **2016**, *4*, 1518–1547. [CrossRef]
2. Schirripa Spagnolo, G.; Cozzella, L.; Leccese, F. Underwater Optical Wireless Communications: Overview. *Sensors* **2020**, *20*, 2261. [CrossRef] [PubMed]
3. Shihada, B.; Amin, O.; Bainbridge, C.; Jardak, S.; Alkhazragi, O.; Ng, T.K.; Ooi, B.; Berumen, M.; Alouini, M.S. Aqua-Fi: Delivering Internet Underwater Using Wireless Optical Networks. *IEEE Commun. Mag.* **2020**, *58*, 84–89. [CrossRef]
4. Johnson, L.J.; Green, R.J.; Leeson, M.S. Underwater optical wireless communications: Depth-dependent beam refraction. *Appl. Opt.* **2014**, *53*, 7273–7277. [CrossRef]

5. Sahu, S.K.; Shanmugam, P. A study on the effect of scattering properties of marine particles on underwater optical wireless communication channel characteristics. In Proceedings of the OCEANS 2017, Aberdeen, Scotland, 19–22 June 2017; pp. 1–7.
6. Elamassie, M.; Miramirkhani, F.; Uysal, M. Performance Characterization of Underwater Visible Light Communication. *IEEE Trans. Commun.* **2019**, *67*, 543–552. [CrossRef]
7. Elamassie, M.; Uysal, M. Vertical Underwater Visible Light Communication Links: Channel Modeling and Performance Analysis. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6948–6959. [CrossRef]
8. Gabriel, C.; Khalighi, M.A.; Bourennane, S.; Léon, P.; Rigaud, V. Misalignment considerations in point-to-point underwater wireless optical links. In Proceedings of the MTS/IEEE OCEANS, Bergen, Norway, 10–14 June 2013; pp. 1–5.
9. Shin, H.; Kim, S.M.; Song, Y. Learning-Aided Joint Beam Divergence Angle and Power Optimization for Seamless and Energy-Efficient Underwater Optical Communication. *IEEE Internet Things J.* **2023**, *10*, 22726–22739. [CrossRef]
10. Shin, H.; Baek, S.; Song, Y. Multidimensional Beam Optimization in Underwater Optical Wireless Communication Based on Deep Reinforcement Learning. *IEEE Internet Things J.* **2024**, *11*, 28623–28634. [CrossRef]
11. Romdhane, I.; Kaddoum, G. A Reinforcement Learning based Beam Adaptation for Underwater Optical Wireless Communications. *IEEE Internet Things J.* **2022**, *9*, 20270–20281. [CrossRef]
12. Guo, Y.; Xiong, K.; Gao, B.; Fan, P.; Ng, D.W.K.; Letaief, K.B. Max-Min Fairness in Rate-Splitting Multiple Access-Based VLC Networks With SLIPT. *IEEE Internet Things J.* **2024**. [CrossRef]
13. Zhang, R.; Xiong, K.; Lu, Y.; Ng, D.W.K.; Fan, P.; Letaief, K.B. SWIPT-Enabled Cell-Free Massive MIMO-NOMA Networks: A Machine Learning-Based Approach. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 6701–6718. [CrossRef]
14. Zhang, R.; Xiong, K.; Lu, Y.; Fan, P.; Ng, D.W.K.; Letaief, K.B. Energy Efficiency Maximization in RIS-Assisted SWIPT Networks with RSMA: A PPO-Based Approach. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 1413–1430. [CrossRef]
15. Filho, J.I.d.O.; Trichili, A.; Ooi, B.S.; Alouini, M.S.; Salama, K.N. Toward Self-Powered Internet of Underwater Things Devices. *IEEE Commun. Mag.* **2020**, *58*, 68–73. [CrossRef]
16. Ammar, S.; Amin, O.; Alouini, M.S.; Shihada, B. Energy-Aware Underwater Optical System With Combined Solar Cell and SPAD Receiver. *IEEE Commun. Lett.* **2022**, *26*, 59–63. [CrossRef]
17. Uysal, M.; Ghasvarianjahromi, S.; Karbalayghareh, M.; Diamantoulakis, P.D.; Karagiannidis, G.K.; Sait, S.M. SLIPT for Underwater Visible Light Communications: Performance Analysis and Optimization. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 6715–6728. [CrossRef]
18. Kogo, T.; Kozawa, Y.; Habuchi, H. Chlorophyll concentration-based CSK constellation point design for underwater SLIPT with priority on communication performance. In Proceedings of the International Symposium on Wireless Personal Multimedia Communications (WPMC), Okayama, Japan, 14–16 December 2021; pp. 1–6.
19. Majlesein, B.; Guerra, V.; Rabadan, J.; Rufo, J.; Perez-Jimenez, R. Evaluation of Communication Link Performance and Charging Speed in Self-Powered Internet of Underwater Things Devices. *IEEE Access* **2022**, *10*, 100566–100575. [CrossRef]
20. Ye, K.; Zou, C.; Yang, F. Dual-Hop Underwater Optical Wireless Communication System with Simultaneous Lightwave Information and Power Transfer. *IEEE Photonics J.* **2021**, *13*, 1–7. [CrossRef]
21. Palitharathna, K.W.; Suraweera, H.A.; Godaliyadda, R.I.; Herath, V.R.; Ding, Z. Lightwave Power Transfer in Full-Duplex NOMA Underwater Optical Wireless Communication Systems. *IEEE Commun. Lett.* **2022**, *26*, 622–626. [CrossRef]
22. Aguirre-Castro, O.A.; Inzunza-González, E.; García-Guerrero, E.E.; Tlelo-Cuautle, E.; López-Bonilla, O.R.; Olguín-Tiznado, J.E.; Cárdenas-Valdez, J.R. Design and Construction of an ROV for Underwater Exploration. *Sensors* **2019**, *19*, 5387. [CrossRef]
23. Wei, W.; Zhang, C.; Zhang, W.; Jiang, W.; Shu, C.; Qiao, X. LED-Based Underwater Wireless Optical Communication for Small Mobile Platforms: Experimental Channel Study in Highly-Turbid Lake Water. *IEEE Access* **2020**, *8*, 169304–169313. [CrossRef]
24. Mizukoshi, I.; Kazuhiko, N.; Hanawa, M. Underwater optical wireless transmission of 405nm, 968Mbit/s optical IM/DD-OFDM signals. In Proceedings of the OptoElectronics and Communication Conference and Australian Conference on Optical Fibre Technology, Melbourne, Australia, 6–10 July 2014; pp. 216–217.
25. Dimitrov, S.; Sinanovic, S.; Haas, H. Signal Shaping and Modulation for Optical Wireless Communication. *J. Lightw. Technol.* **2012**, *30*, 1319–1328. [CrossRef]
26. Mobley, C.D.; Gentili, B.; Gordon, H.R.; Jin, Z.; Kattawar, G.W.; Morel, A.; Reinersman, P.; Stamnes, K.; Stavn, R.H. Comparison of numerical models for computing underwater light fields. *Appl. Opt.* **1993**, *32*, 7484–7504. [CrossRef] [PubMed]
27. Eroğlu, Y.S.; Yapıcı, Y.; Güvenç, I. Impact of Random Receiver Orientation on Visible Light Communications Channel. *IEEE Trans. Commun.* **2019**, *67*, 1313–1325. [CrossRef]
28. Celik, A.; Saeed, N.; Shihada, B.; Al-Naffouri, T.Y.; Alouini, M.S. End-to-End Performance Analysis of Underwater Optical Wireless Relaying and Routing Techniques Under Location Uncertainty. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 1167–1181. [CrossRef]
29. Korotkova, O.; Farwell, N. Light scintillation in oceanic turbulence. *Waves Random Complex Media* **2012**, *22*, 260–266. [CrossRef]
30. Navidpour, S.M.; Uysal, M.; Kavehrad, M. BER Performance of Free-Space Optical Transmission with Spatial Diversity. *IEEE Trans. Wirel. Commun.* **2007**, *6*, 2813–2819. [CrossRef]
31. Sandalidis, H.G.; Vavoulas, A.; Tsiftsis, T.A.; Vaiopoulos, N. Illumination, data transmission, and energy harvesting: The threefold advantage of VLC. *Appl. Opt.* **2017**, *56*, 3421–3427. [CrossRef]

32. Kyritsis, A.; Papanikolaou, N.; Tatakis, E.C. A novel Parallel Active Filter for Current Pulsation Smoothing on single stage grid-connected AC-PV modules. In Proceedings of the European Conference on Power Electronics and Applications, Aalborg, Denmark, 2–5 September 2007; pp. 1–10.
33. Li, C.; Jia, W.; Tao, Q.; Sun, M. Solar cell phone charger performance in indoor environment. In Proceedings of the IEEE 37th Annual Northeast Bioengineering Conference (NEBEC), New York, NY, USA, 1–3 April 2011; pp. 1–2.
34. Zainal, N.A.; Ajisman; Yusoff, A.R. Modelling of Photovoltaic Module Using Matlab Simulink. *IOP Conf. Ser. Mater. Sci. Eng.* **2016**, *114*, 012137. [CrossRef]
35. Esram, T.; Chapman, P.L. Comparison of Photovoltaic Array Maximum Power Point Tracking Techniques. *IEEE Trans. Energy Convers.* **2007**, *22*, 439–449. [CrossRef]
36. Wang, J.B.; Hu, Q.S.; Wang, J.; Chen, M.; Wang, J.Y. Tight Bounds on Channel Capacity for Dimmable Visible Light Communications. *J. Light. Technol.* **2013**, *31*, 3771–3779. [CrossRef]
37. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
38. Lillicrap, T.P. Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, PR, USA, 2–4 May 2016; pp. 1–14.
39. Li, C.; Xia, J.; Liu, F.; Li, D.; Fan, L.; Karagiannidis, G.K.; Nallanathan, A. Dynamic Offloading for Multiuser Muti-CAP MEC Networks: A Deep Reinforcement Learning Approach. *IEEE Trans. Veh. Technol.* **2021**, *70*, 2922–2927. [CrossRef]
40. Xu, K.; Shen, Z.; Wang, Y.; Xia, X.; Zhang, D. Hybrid Time-Switching and Power Splitting SWIPT for Full-Duplex Massive MIMO Systems: A Beam-Domain Approach. *IEEE Trans. Veh. Technol.* **2018**, *67*, 7257–7274. [CrossRef]
41. Kim, S.M.; Won, J.S. Simultaneous reception of visible light communication and optical energy using a solar cell receiver. In Proceedings of the International Conference on ICT Convergence (ICTC), Jeju, Republic of Korea, 14–16 October 2013; pp. 896–897.
42. Jalbert, J.; Baker, J.; Duchesney, J.; Pietryka, P.; Dalton, W.; Blidberg, D.R.; Chappell, S.; Nitzel, R.; Holappa, K. A solar-powered autonomous underwater vehicle. In Proceedings of the MTS/IEEE Oceans, San Diego, CA, USA, 22–26 September 2003; Volume 2, pp. 1132–1140.