IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2020, accepted May 8, 2020, date of publication May 27, 2020, date of current version June 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998052

Adaptive Laser Welding Control: A Reinforcement Learning Approach

GIULIO MASINELLI[1,2], (Member, IEEE), TRI LE-QUANG[1], SILVIO ZANOLI[2], KILIAN WASMER[1], (Member, IEEE), AND SERGEY A. SHEVCHIK[1]

[1]Laboratory for Advanced Materials Processing, Swiss Federal Laboratories for Materials Science and Technology (EMPA), 3602 Thun, Switzerland
[2]Embedded Systems Laboratory, Swiss Federal Institute of Technology in Lausanne (EPFL), 1015 Lausanne, Switzerland

Corresponding author: Kilian Wasmer (kilian.wasmer@empa.ch)

This work was supported by the Swiss Federal Laboratories for Materials Science and Technology (EMPA).

ABSTRACT Despite extensive research efforts in the field of laser welding, the imperfect repeatability of the weld quality still represents an open topic. Indeed, the inherent complexity of the underlying physical phenomena prevents the implementation of an effective controller using conventional regulators. To close this gap, we propose the application of Reinforcement Learning for closed-loop adaptive control of welding processes. The presented system is able to autonomously learn a control law that achieves a predefined weld quality independently from the starting conditions and without prior knowledge of the process dynamics. Specifically, our control unit influences the welding process by modulating the laser power and uses optical and acoustic emission signals as sensory input. The algorithm consists of three elements: a smart agent interacting with the process, a feedback network for quality monitoring, and an encoder that retains only the quality critic events from the sensory input. Based on the data representation provided by the encoder, the smart agent decides the output laser power accordingly. The corresponding input signals are then analyzed by the feedback network to determine the resulting process quality. Depending on the distance to the targeted quality, a reward is given to the agent. The latter is designed to learn from its experience by taking the actions that maximize not just its immediate reward, but the sum of all the rewards that it will receive from that moment on. Two learning schemes were tested for the agent, namely Q-Learning and Policy Gradient. The required training time to reach the targeted quality was 20 min for the former technique and 33 min for the latter.

INDEX TERMS Laser welding, laser material processing, reinforcement learning, policy gradient, Q-learning, closed-loop control.

I. INTRODUCTION

Laser welding (LW) is a crucial technology for many industrial sectors, including automotive production, maritime, medical, aerospace, and micromechanics [1]. On the one hand, its advantages are in non-contact processing — avoiding tool wear, ability to process refractory materials, and higher processing rate and joint quality compared to traditional welding processes [2]. On the other hand, LW's main disadvantages derive from the highly complex underlying physical phenomena involved in the process. Thus, despite many developments of this technology, LW still suffers from imperfect quality repeatability, limiting its applications in industrial production requiring high-quality standards.

In the literature, the most commonly reported approach to increase the repeatability of the weld quality is the application of traditional regulators, such as proportional-integral (PI) or proportional-integral-derivative (PID) controllers [3], [4]. These methods allow tracking the desired weld quality using measurements of the surface temperature or the surface shape of the process zone (PZ) as feedback. Unfortunately, since they are based on the linearization of the non-linear welding dynamics, they can only operate in a narrow range of the process parameters. This operating range, moreover, has to be established during a preliminary exhaustive experimental search, which is very time- and material-consuming, making the entire methodology undesirable in an industrial environment.

The associate editor coordinating the review of this manuscript and approving it for publication was Jianyong Yao.

A less common approach, but that is worth investigating, is based on more sophisticated regulators that rely on differential models of the process [5], [6]. But in the case of LW, a reliable model can be complicated to obtain, as it has to take into account many factors that can drastically vary the process, such as the heating and melting dynamics [5].

Nevertheless, a preliminary attempt can be found in Na *et al.* [7], where the authors presented an algorithm that automatically builds a model during the operation using the Hammerstein identification technique.

An example of the actual use of a model-based controller for laser processes was proposed by Song and Mazumder [6], where an experimentally identified model was involved for predictive control of laser cladding — a process that is closely related to LW. This technique heavily relies on its model for the choice of the actions to take according to their impact on the environment evaluated with the model itself. To be specific, a closed-loop process was used to steer the melt pool temperature to a reference temperature profile. In a real-life scenario, unfortunately, this approach has two major drawbacks. First, the temperature of the melt pool is not uniformly distributed over its surface [8]. Second, the optimal temperature profile can vary during the process, as it strictly depends on the geometry, e.g., on the proximity to the edges or the boundaries of the workpiece. Thus, the tracking of a single fixed target has a direct impact on the system performance and so on the desired result.

Similarly, Bollig *et al.* [9] showed promising results by modeling the non-linear process with an Artificial Neural Network and controlling the laser power with a linear model predictive algorithm based on the instantaneous linearization of the neural network itself. In this case, the regulator aimed to track a reference penetration depth detected from the intensity of the plasma's optical emission. However, the experimental calibration curve used to map the measured intensity to the penetration depth may diverge from its real-life values, limiting the application of the same methodology in broader scenarios.

In this context, there is a clear need for a widely applicable, robust, and cost-effective process control system that ensures high-quality standards. In particular, we focus on deep keyhole welding, where the process complexity is even higher compared to other welding regimes, such as conduction welding.

This welding regime is indeed characterized by the co-existence — within a limited volume — of vapor, melt, and plasma phases of the processed material [10]. Moreover, it possesses an extremely complex energy-coupling mechanism that includes Fresnel absorption (due to multiple reflections inside the vapor channel) [11]. These complex phenomena generate many process instabilities, making keyhole welding prone to defects even under constant laser irradiation [10]. Specifically, one of the most critical defects is porosity. Pores are problematic since they are located inside the material and may substantially weaken the mechanical strength of the welding joint [12].

The design of a keyhole LW control system is made all the more challenging by the partial observability of the laser process. In fact, in-depth information of the PZ can only be indirectly obtained either by acoustic emission (AE) sensors or by surface measurements using optical emissions (OE) sensors [12]. Consequently, it is difficult to provide an effective feedback from the process to the control system, since it requires the correlation of the surface measurements with the sub-surface events (e.g., pore formation), which is not a trivial task [12]. Nevertheless, some pilot works in LW monitoring report successes in identifying quality critic momentary events from the corresponding AE and OE signals from the processed zone [13], [14].

The present study starts from the aforementioned preliminary results of process monitoring and focuses on the use of Reinforcement Learning (RL) towards keyhole LW closed-loop control.

RL appears to be an attractive approach since it enables a model-free learning scheme that is capable of solving complex problems and provides high adaptability to specific conditions through active interaction with a given process [15].

Moreover, we take advantage of recent advances in Deep Convolutional Neural Networks (DCNN) developments [16], [17] to derive efficient representations of the laser process from the high-dimensional sensory input — the AE and OE signals from the PZ — and use them to generalize previous experiences to new situations [18]. In our case, indeed, the input data from the sensors do not contain an explicit representation of the physical state of the system, as they are just limited to the optical and acoustic emission. As shown by Mnih *et al.* [18], DCNNs can overcome — and even take advantage of — this condition, allowing the system to learn meaningful position and scale of irregular structures in the data.

Concerning the recent advances of RL, its application towards LW was discussed in Günther *et al*. [19], where a dynamic model substituted the real laser process, and a camera-based system and photodiodes were used for process monitoring. RL was able to efficiently search for strategies for modulating the laser irradiation to compensate for the mentioned process instabilities.

Despite the successes of this work, the efficiency of RL in more complex LW processes remains an open question. To close this gap, we inspected the performance of our methodology in the case of keyhole LW and evaluated its outcomes in terms of the evolution of the weld quality over time during training. Firstly, the AE and OE signatures of the desired weld quality were given to the algorithm, as well as several signatures of undesirable qualities, without any other prior information about the process dynamics. Further search for the optimal process control strategy was carried out in a completely autonomous way. Two RL techniques were investigated in this contribution: Q-Learning [20] and Policy Gradient [21], in order to analyze their strengths and weaknesses in this particular application.

This paper is divided into five sections. Section II describes the experimental setup and the hardware of the control system. Section III describes the developed algorithms, including details on signal dimensionality reduction and the feedback network used for process monitoring. Section IV presents and discusses the results. Finally, Section V concludes this work and gives the perspective of its further developments that would allow LW to operate autonomously and, thus, bringing it closer to the intelligent manufacturing within Industry 4.0 framework [22].

II. EXPERIMENTAL PROCEDURE, MATERIALS, ACQUISITION, AND CONTROL

The experimental setup was similar to the one used in a previous work [14], and therefore just a summary is given in this contribution.

A. EXPERIMENTAL SETUP

A schematic representation of the setup is presented in Fig. 1, along with its picture. The main components were: a laser source, an optical laser head, a workpiece holder — mounted on a moving stage — and an AE sensor.

The laser source was a fiber laser system StarFiber150P (Coherent Switzerland AG, Switzerland), with a maximum output power of 250 W, a wavelength of 1070 nm, and a diameter of the laser spot of 30 μm (within $2w_0$) at the workpiece surface. The source was operated in continuous-wave (CW) mode with the possibility to modulate the output laser power using an external voltage source within a voltage range of 0–5 V. More details are given in Le-Quang *et al.* [23].

The laser experiments were performed in air at atmospheric pressure. To prevent the potential oxidation of the weld, an adequate Ar flow was directed to the PZ via a nozzle. The flow was kept constant at a pressure of 1.5 atm during all experiments. In order to realize line welds, a workpiece was mounted on a linear stage M-663.5U (Physik Instrumente GmbH, Germany), and moved at a constant velocity of 10 mm/s during the process. The movement of the workpiece was synchronized with the laser source so that the irradiation started only when the stage already reached the set velocity.

The aforementioned setup provided the realization of different LW regimes leading to various welding quality [14], [24]–[26], including *no illumination* (laser power $P = 0$ W), *conduction welding*, *keyhole without porosity*, and *keyhole with porosity*.

It must be emphasized that, in terms of process parameters, the weld quality also depends on the velocity of the workpiece. This work, however, was focused on the control of the laser power that, in our setup, can be dynamically modulated via the external voltage generator, as described. Consequently, this process parameter was considered as the sole control variable.

B. SENSORS

The laser head was equipped with a customized optical system that allowed delivering the back-reflected radiations from the PZ to three photodiodes. These sensors are based on Silicon (Si), Germanium (Ge), and InGaAs and are sensitive within the ranges of 450–850 nm, 1000–1200 nm, and 1250–1700 nm, respectively. The Ge sensor was equipped with a narrow bandpass optical filter (FB1070-10, Thorlabs Inc., USA) with a center wavelength of 1070 ± 2 nm to only sense the back-reflected laser radiation from the PZ.

In addition to the optical sensors, an AE sensor PICO (Physical Acoustics, USA) was placed in tight contact with the workpiece, as shown in Fig. 1 (a). The sensor was sensitive within the range 500–1850 kHz. Its purpose is to detect the AE shockwaves generated inside the workpiece during welding.

C. MATERIAL

The workpieces were 2 mm thick plates of titanium alloy (Ti6Al4V, grade 5) with a melting temperature of 1,650 °C. This material was chosen due to its extensive industrial usage, including the medical sector. Additionally, its Heat-Affected Zone (HAZ) can be easily recognized in cross-sections due to the remarkable textural changes [27].

D. REFERENCE QUALITY DEFINITION

To meet the industrial demand for high-quality keyhole welding [12], we defined our reference weld as the one with the highest achievable penetration depth without the presence of pores. In addition to previous experiences [14], [24], several experiments were carried out, taking advantage of the well-controlled welding conditions of our setup that allowed us to reproduce different penetration depths precisely. Each experimental weld was verified by analyzing the cross-sections of the processed workpieces.

Finally, the investigations lead to a reference weld characterized by a laser power of 80 W and a resulting penetration depth of 150 μ m. Every increment in laser power resulted in the introduction of porosity, whereas every decrement corresponded to shallower welds.

E. DATA ACQUISITION AND COMPUTATIONS

In order for the control system to reach a real-time response given the high-dimensional input from the sensors, a combination of specialized hard/software was used. The hardware included a PC equipped with an Intel i7-8750H processor (Intel, USA) that operated at a frequency up to 4.1 GHz, and a Graphics Processing Unit NVidia GTX 2080 Ti (Nvidia, USA).

The signals from all four sensors described in Section II-B were acquired with a high-speed DAQ card Advantech 1840 (Advantech, Taiwan) with four independent input ports for data digitalization. All signals were digitized with a sampling rate of 1 MHz, and their acquisition was triggered when the intensity of the back-reflected laser light detected by the Ge photodiode exceeded a fixed threshold (0.1 V).

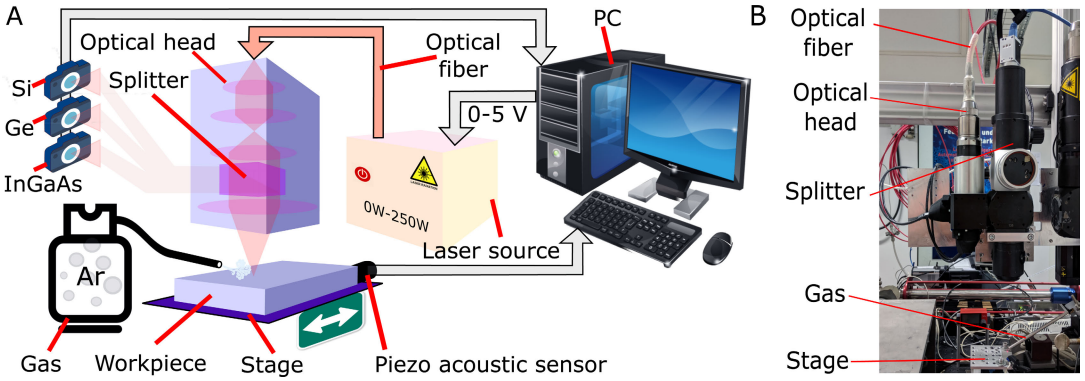

FIGURE 1. (a) Scheme of the experimental setup and (b) its picture. The labels of the individual components in (a) and (b) correspond to each other.

The choice of the Ge sensor as a trigger for the acquisition is based upon the very high intensity of the back-reflected laser radiation at the beginning of the process, when the reflectivity of the workpiece is the highest [23].

To dynamically modulate the laser power, the control signal provided by the RL algorithm was transmitted to the laser source via an external USB unit Advantech 4751L (Advantech, Taiwan). The latter converted the digital values calculated by the RL models into a direct voltage value, which was then delivered to the laser source via a cable connection (see Fig. 1 for details). The time delay between the output from the USB unit and the laser response was experimentally measured to be 0.57 ± 0.25 ms.

The real-time acquisition routine of the input signals using the DAQ board, the data processing in the GPU, and the transmission of the computed control signal to the laser source were carried out with in-house custom-made software. In particular, the data acquisition program was coded in C# in Visual Studio 2017, Community addition. Conversely, the high-level data processing was realized in Python 3.7. Finally, the Deep Learning (DL) library involved was Pytorch (`www.pytorch.org`), version 1.1.0.

III. DATA PROCESSING

The structure of the developed data processing is schematically presented in Fig. 2. The entire control unit consists of three main building blocks: an encoder that processed the data from the measurements to retain only the quality critic events, a smart agent interacting with the welding process, and a feedback network based on a DCNN for quality monitoring.

Before even starting the interaction with the environment, the encoder and the feedback network were trained using a database consisting of 750 signals acquired from previous experiments covering the whole operating range of the laser process. The signals were divided into 5 categories according to the corresponding penetration depth identified with optical inspection of the cross-section of the processed material (more details in Section IV).

A. ENCODER

The encoder was used to reduce the dimensionality of the sensory input of the agent, preserving, at the same time, the structure of the original data while minimizing the computational time. The introduction of this unit was motivated by a resulting simplification of the search of the optimal control law for the smart agent. Indeed, the projection of the high-dimensional input data into a low dimensional latent space allows capturing a "good" parametrization of the signal that focuses only on quality critical events that the user can settle by carefully choosing the training data [28].

To be specific, we based our encoder on a DCNN due to the proven abilities of convolutional networks to explicitly model signals by finding their meaningful degrees of freedom [29], [30]. Indeed, DCNNs also exhibit excellent generative properties [31], which motivates their use as encoders.

Following traditional architectures [30], [32], our DCNN encoder included four convolution layers. Moreover, each convolution was enforced with a batch normalization layer to speed up the training [33]. The activation consisted of a rectified linear unit (ReLU) that is more efficient in multilayer architectures, as it diminishes the gradient vanishing problem [34]. The summarization of the input information is achieved gradually through the convolutional layers by adopting strided convolutions [35].

As stated, the training of the encoder was carried out separately, prior to the interaction with the environment. During training, a decoder with a symmetrical structure was added to process the encoder output. Specifically, in the decoder, the convolutions of the encoder were replaced with their reciprocal transposed convolution. The two models were then trained end-to-end to minimize the mean square error between the training input signals and the output of the decoder [30], [32]. After training, the decoder was removed, thus, keeping the encoder standalone to provide a low dimensional signal representation as input for the smart agent.

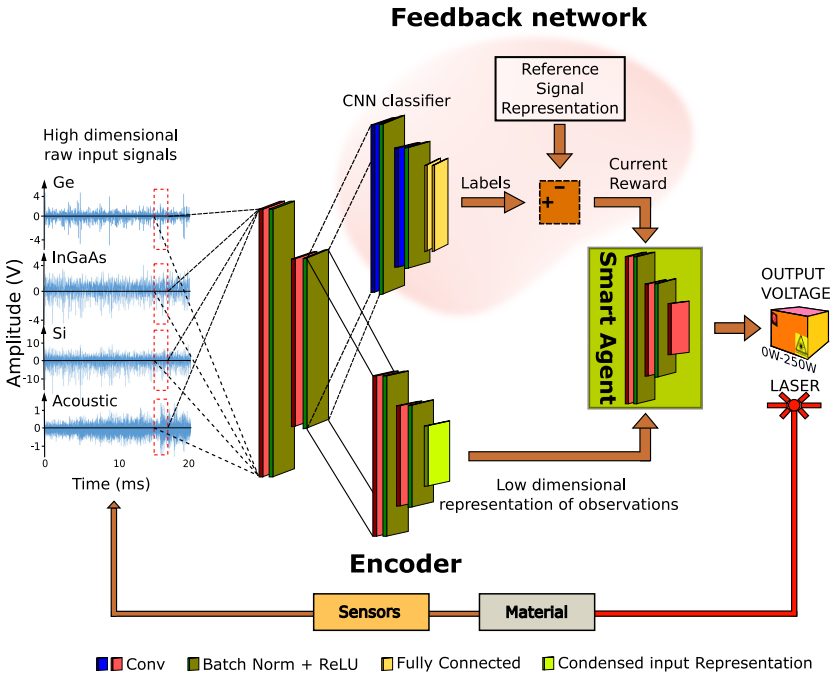

FIGURE 2. Structure of the complete control unit made up of three main building blocks: an encoder that processed the data from the sensory input to retain only the quality critic events, a smart agent interacting with the welding process, and a feedback network based on a convolutional neural network for quality monitoring.

B. FEEDBACK NETWORK

As seen in the introduction, RL is a learning paradigm leading to the design of algorithms that directly interact with an environment and learn via trial and error. Nevertheless, learning by doing is only effective if we can define a notion of reward, something that motivates the intelligent system to behave appropriately. For this reason, the full setup depicted in Fig. 2 included a feedback network based on a DCNN classifier and a summation unit.

This unit is based on our previous work [13], where the AE and OE signals from the PZ were used to identify quality critic momentary events. In this contribution, the output of the classifier is made up of labels that correspond to predefined welding qualities in terms of penetration depth and pore content. The DCNN classifier shares the initial two convolution layers with the encoder, as it is shown in Fig. 2. This detail allows the classifier to reuse the good feature representation learned by the encoder. The final decision on the quality is taken in two fully connected layers that were closed by a softmax layer. In analogy to the encoder, the training of the classifier was carried out prior to the operation of the entire system with the preliminary collected signals database.

To provide the reward signal, the output of the classifier (i.e., the label of the current momentary quality) was compared with the label of the reference signal in the summation unit (see Fig. 2). In case of significant differences, the smart agent is granted with negative rewards; otherwise, positive rewards are assigned (more details in Section IV).

C. SMART AGENT

The final building block is constituted by the smart agent whose purpose is to interacts with the environment — in this case, the laser process — by making actions, i.e., modulation of the laser power. Practically, the agents communicate to the output board that, in turn, delivers the control signal to the laser source (see Section II-E for more details).

The principle of operation is the following: based on the representation of the current sensory input provided by the encoder, the agent chooses an action, which leads to a change in the sensory input, and receives a reward from the feedback network. From this experience — made up by the past sensory input, the executed action, the current input, and the received reward — the agent tries to optimize the outcomes of its actions over time, i.e., to maximize the reward over a defined time horizon.

In our case, the considered time horizon corresponds to the time required to perform a single line weld of 10 mm (1 s in this work, see Section II-A).

In the remainder of the article, we refer to this 10 mm line weld as an episode. Operating in an episodic fashion — i.e., by individually welding lines of 10 mm — permits the algorithm to update its parameters between one line to another, and allows the stage to move in a new unprocessed position to be able to start over. For training the agent, two RL techniques were tested in this study, and their descriptions are given in the next subsections.

D. PARAMETER TUNING

Assuming the use of a conventional RL learning scheme, we can output a single action after the defined sensory input is available, i.e., after a predetermined number of data points is acquired from the AE and OE sensors. Hence, the length of the input window determines the operational frequency, that is, the rate at which the control unit can modify the laser power.

A small window increases the system readiness to adapt to new welding conditions, but, unfortunately, it also raises the sensitivity to noise of the feedback network [14]. In contrast, a large window increases the monitoring accuracy and eases the internal timing constraints, but reduces the number of actions per unit time. In this sense, the window length is crucial, as it is a trade-off between system readiness to react to different stimuli and monitoring accuracy. A good compromise was found by fixing the window length to 20 ms, thus, setting the operating frequency to 50 Hz.

The entire system was also sensitive to multiple other parameters, including the size of the convolutional kernels used in the DCNN and the dimensionality of the encoder output. The adjustments of these parameters were carried out through an exhaustive search, and the final set of parameters was established as follows.

The optimal size of the convolutional kernel used in the very first layer of the feedback network (see Fig. 2) was founded to be 5 ms. Taking into account the given stage velocity and the acquisition rate, the time span of this kernel corresponded to 50 μ m in length of the weld joint, or, equivalently, to a signal sample of 5,000 sampling points obtained from each sensor.

Following the scheme in Fig. 2, the unification of all signals from the sensors in a time interval of 20 ms determines the dimension of the algorithm's input space that amounts to 80,000 data points. As seen before, the agent does not receive this high dimensional input, but its condensed representation from the encoder.

The maximum possible dimensionality reduction achievable in our setup led to low dimensional signals made up of 64 data points for every sensor (from the original 20,000). In our work, it was experimentally established that any further reduction harmed the algorithm's accuracy, provoking higher error rates for the autonomous learning controller.

E. REINFORCEMENT LEARNING

RL is inspired by human and animal behaviors, where the experience/knowledge is acquired through active interaction with the environment by trying to maximize the rewards received [15], [18], [36].

Specifically, RL is the branch of Machine Learning (ML) that aims at designing agents capable of taking, in every moment, the action that maximizes not just the immediate reward, but the sum of all the rewards that will be received thenceforth. The agent chooses actions based on its sensory input that provides a momentary representation of the environment — the so-called states — and tries to optimize the outcomes of these actions over time in terms of reward.

In RL, this concept is formalized through a Markov Decision Process (MDP). MDP is described by a quadruple $\{\mathcal{S}, \mathcal{A}, p, r\}$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces and $p(s_{t+1}|s_t, a_t)$ is the probability of the transition from state $s_t \in \mathcal{S}$ to state $s_{t+1} \in \mathcal{S}$ taking the action $a_t \in \mathcal{A}$. Each change of state is rewarded according to $r(s_t, a_t)$. The strategy of choosing an action a_t given the state s_t is known as policy, and it is indicated by $\pi(a_t|s_t)$ — denoting the probability of selecting the action a_t in state s_t.

The correctness of the choice of the actions is evaluated in terms of the rewards subsequently collected. Concretely, the quality of taking an action a_t given the actual state s_t with the further choice of all remaining actions according to the policy π, can be quantified with the action-value function $Q_\pi(s_t, a_t)$. Given an episode that includes T steps, it is defined as [15]:

$$Q_\pi(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) | s_t, a_t \right], \tag{1}$$

that is the expected total reward from taking the action a_t in state s_t and then following the policy π.

The goal of RL is to approximate the optimal policy π^* that returns, for every state, the best action to take in terms of total reward from that moment on.

One approach consists of estimating the action-value function for π^*. Indeed, in that case, the optimal action a to be taken in state s is the one that maximizes Q_{π^*} for the given state [15]. The different RL algorithms differ in the way $Q_\pi(s, a)$ or, alternatively, the policy parameters are iteratively updated. In this study, we have tested two of the most successful realizations of RL that are Q-Learning and Policy Gradient. Both methods have *pro et contra*, which are discussed in the next two subsections.

F. DEEP Q-LEARNING

Q-Learning is one of the most popular RL algorithms and aims at estimating the Q_{π^*} values for every state — hence the name of the technique.

In the case of high-dimensional state space (e.g., in laser welding), the traditional update methods for the Q_π values become inapplicable as they suffer from the curse of dimensionality [37]. Indeed, those methods require to represent the Q_π values in a tabular form — a table having as many entries as the ordered pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ [15], which is only feasible if the cardinalities of both $\mathcal{S}$ and $\mathcal{A}$ are small.

The concept of DL allows overcoming those limits by using DCNNs to estimate the action-value function [38], exploiting the recent advances in ML where DCNNs proved to be excellent complex function approximators [39], [40].

In our work, the Fitted Q Iteration algorithm (FQI) was used as a basic learning scheme [41], and included the following steps:

(i) using some policy, collect a dataset of transitions:

$$\{(s_t, a_t, s_{t+1}, r_t)\}_{t=1,2,\ldots} \tag{2}$$

(ii) for every transition, compute:

$$y_t = r_t + \gamma \max_a Q_{\pi_\theta}(s_{t+1}, a) \tag{3}$$

(iii) update the parameters θ:

$$\theta \leftarrow \underset{\theta}{\text{argmin}} \left[\sum_t \| Q_{\pi_\theta}(s_t, a_t) - y_t \|^2 \right], \tag{4}$$

where Q_{π_θ} denotes the functional approximator of the function Q_π given by a parametric function with parameters θ. In this contribution, θ represents the weights and biases of a DCNN that takes as input the ordered pair (s_t, a_t) and outputs an estimate of $Q_\pi(s_t, a_t)$. γ is a discount factor $\in (0, 1)$ to weigh less future rewards and more the immediate ones, r_t is the reward collected at time t, and y_t is a momentary target for the computation of the so-called Bellman update in (4) [37].

The minimization problem in (4) can be solved using gradient descent methods. Therefore, it can be addressed using the techniques for loss minimization that are common in DL frameworks [42], [43].

In order to promote the exploration of the state space at the beginning of the training, we have used the so-called epsilon-greedy technique for step (i) of the FQI [15]. This strategy consists in the use of the following policy for the collection of the transitions:

$$\pi(a_t|s_t) = \begin{cases} 1 - \varepsilon, & \text{if } a_t = \underset{a}{\text{argmax}} Q_{\pi_\theta}(s_t, a) \\ \dfrac{\varepsilon}{|\mathcal{A}| - 1}, & \text{otherwise,} \end{cases} \tag{5}$$

where $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$ and $\varepsilon \in (0, 1)$. Following (5), at each timestamp, the algorithm chooses either a random action with probability ε, or the best action according to the actual Q_π estimate with probability $1 - \varepsilon$. As the training progresses, ε is progressively reduced. This procedure encourages the exploration of the environment at the very beginning of the training and the exploitation of the acquired knowledge at the end.

To reduce the oscillations or divergence of the policy, the momentary target y_t and the Q-value $Q_{\pi_\theta}(s_t, a_t)$ were estimated using two separate networks that are known as target network ($Q_{\pi_{\theta t}}$) and Q-network (Q_{π_θ}), respectively [18].

During the interaction with the environment, the parameters of the target network are cyclically updated with the parameters of the Q-network. Additionally, in our study, the Double Q-Learning technique was used [44]. It consists in using the Q-network to evaluate the action to take — using Q_{π_θ} in (5) — and the target network to evaluate the momentary target y_t — using $Q_{\pi_{\theta t}}$ instead of Q_{π_θ} in (3).

The reason was an efficient decorrelation between the noise in the action selection and the noise in the Q-values estimation, which is a common problem for standard Q-Learning realizations [44].

Moreover, to avoid bad local minima and to reduce the correlation between observations, a replay buffer $\mathcal{B}$ was introduced, as in Mnih *et al.* [18]. In particular, during step (i) in FQI, the collected transitions are added to $\mathcal{B}$. During step (ii), we randomly sampled a batch of the accumulated transitions from $\mathcal{B}$ and used those to compute the targets y_t through the target network (see (3)). Finally, the updates of the parameters θ in the Q-network were carried out using (4).

Here one of the key advantages of the introduction of the encoder manifests itself. Indeed, it allows a dimensionality reduction of the input — the reduction factor was 300 in our setup — allowing us to use a bigger buffer $\mathcal{B}$, avoiding the GPU memory saturation.

The advantages and disadvantages of Q-Learning can be explained by the way the targets are computed in FQI. As can be seen in (3), the observed reward in just one transition is used to calculate the targets y_t. In addition, the first term r_t in (3) is significant when the estimation of Q_{π_θ} is inaccurate, as it is a real reward and not an estimation. In contrast, the second term $\gamma \max_a Q_{\pi_\theta}(s_{t+1}, a)$ in (3) is relevant only when the estimation of Q_{π_θ} is reliable, as it is an estimation of the total future reward that is supposed to be higher than the current one.

Consequently, during the Bellman updates (see (4)), the algorithm relies more and more on the actual estimate of the Q-value as soon as it becomes sufficiently large. In Q-Learning, as a result, the strategy of sharply reducing the variance of the estimates (the Q-values) is being adopted, to the detriment of high bias.

G. POLICY GRADIENT

As mentioned above, the main limitation of Q-Learning is the high bias in the estimation of the Q-values. This bias originates from the single-step reward estimator for the targets y_t. The Policy Gradient (PG) approach [15], [45], [46] aims to overcome those limits by evaluating the total reward on an entire episode. Similarly to other RL algorithms, the objective of PG is to find the policy that maximizes the expected total reward in one episode that includes T steps. But contrary to Q-Learning, PG does not try to estimate the optimal Q-values, but the parameters of the policy approximating the optimal policy π^*:

$$\theta^* = \underset{\theta}{\text{argmax}} J(\theta), \tag{6}$$

where

$$J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{T} r(s_t, a_t) \right], \tag{7}$$

and θ stands for the policy parameters. In our case, θ represents the weights and the biases of a DCNN that takes as input the current sensory representation provided by the encoder (see Section III-A) and outputs the action to be taken (e.g., the power of laser irradiation).

In PG, the functional $J(\theta)$ is estimated as:

$$J(\theta) \approx \hat{J}(\theta) = \sum_{t=1}^{T} r(s_t, a_t). \tag{8}$$

The optimization of the objective $J(\theta)$ is carried out by directly differentiating its estimate $\hat{J}(\theta)$ and using gradient ascent to update the parameters as:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \hat{J}(\theta). \tag{9}$$

In particular, the gradient of the objective in (8) is computed as [45], [46]:

$$\nabla_\theta \hat{J}(\theta) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t=1}^{T} r(s_t, a_t). \tag{10}$$

Clearly, the entire approach relies on a single sample estimate of the full expectation (cf. (8)) that, even if unbiased, has a very high variance.

For this reason, even though this method is potentially able to provide better results compared to Q-Learning in terms of the learned policy, it surely requires more learning time.

The implementation of PG was carried out by firstly randomly initializing the parameters of the policy π_θ and then sampling a trajectory (i.e., collecting all the transitions (s_t, a_t, s_{t+1}, r_t) within a single episode). The logarithm of the action probabilities, as well as the rewards collected along the trajectory, were accumulated and used to calculate the policy's gradient according to (10). Finally, the parameters were updated following the direction of improvement indicated by the gradient (cf. (9)).

IV. RESULTS AND DISCUSSION

A. RESULTS

Prior to starting the interaction with the environment, the preparation of the algorithm included two stages, namely: i) collection of the signal database for training the classifier and the encoder, and ii) definition of a reward function.

The first step is motivated by the fact that the classifier and the encoder — to fulfill the role of guiding the smart agent during its learning process — have to learn to recognize, not just the reference quality, but also several other counter-examples.

For this reason, we collected the acoustic and optical signals from multiple weld experiments at various laser power (20, 40, 60, 80, and 120 W).

It must be emphasized that the weld quality depends theoretically not only on the laser power but also on the workpiece velocity and its physical properties such as optical and thermal [10]. But in this work, since the latter factors were invariable, the former one is used to define the weld quality.

The sensors' signals were acquired during three weld experiments at each laser power, then partitioned in samples of 20 ms (see Section III-D, for details), and finally grouped in 5 categories according to the weld quality in terms of penetration depth identified via optical inspection of both surface and cross-section of the workpieces.

Based on the optical inspection, the categories were defined as *insignificant penetration* (achieved with a laser power of 20 W), *poor penetration* (40 W), *medium penetration* (60 W), *highest penetration without pores* (80 W), and *porosity* (120 W). In total, each category consisted of 150 samples.

The second stage concerns the definition of the reward function that determines the reward assignment from the feedback network to the smart agent.

Considering that the agent is designed to act to maximizes the collected rewards in the long run, the engineering of the reward is crucial since it influences the learning process. The reward assigned for every weld quality detected by the classifier used in our experiments is reported in Table 1.

TABLE 1. Rewards assigned for every category detected by the classifier.

Classification	Laser power (W)	Reward
Insignificant penetration	20	-3
Poor penetration	40	-2
Medium penetration	60	1
Highest penetration without pores	80	20
Porosity	120	-1

After the preparation, we let the algorithm interact with the environment in a completely autonomous way without any further interventions. The performance for both Q-Learning and Policy Gradient is shown in Fig. 3, where the red line represents the average values of the rewards obtained in every episode, whereas the shaded area denotes the standard deviation.

The average reward of Q-Learning reached a plateau after approximately 110 episodes, i.e., after performing 110 line welds of 10 mm. Taking into consideration the fact that we wait for 10 s after each line — to permit the agent to update its parameters and to allow the stage to move in a new unprocessed position —, this learning period corresponds to about 20 minutes. In contrast, PG reached a plateau only after 180 episodes (33 minutes). In both cases, additional learning time had little effect in terms of increment of the quality, and it only increased the cost in terms of wasted materials and time.

The dynamics of the agent adaption to the given process can be vividly seen in the evolution of the welds using optical inspections of the surfaces and cross-sections of the processed material. Fig. 4 presents the optical images of the welds corresponding to the first, the 40th, the 80th, and the 110th episode of the Q-Learning training process. To be specific, Fig. 4 (a) shows the light microscope images of the top views of different episodes, whereas Fig. 4 (b), the corresponding cross-sections.

It has to be noted that the results in Fig. 4 show an evolution of the weld quality that is consistent with the increment of the reward observed in Fig. 3. Indeed, in Fig. 4 (a), episode 1 — i.e., beginning of the training — signs of unstable controlled

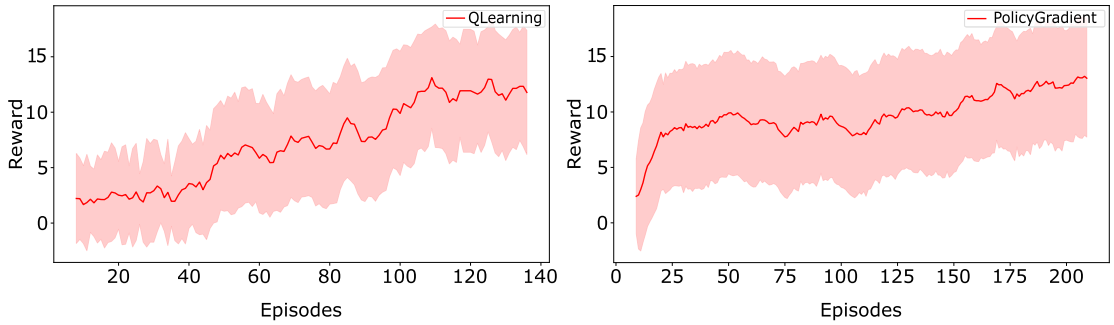

FIGURE 3. Performance in terms of average reward per episode over time for *Q*-Learning and Policy Gradient. The red line represents the average reward over an episode, whereas the shaded area indicates the standard deviation. An episode corresponds to the weld of a line of 10 mm and has a duration of 1s. Between one line to the other, we wait for 10 s to permit the agent to update its parameters and to allow the stage to move in a new unprocessed position.

laser power can be seen on the weld surface. The black marks on the weld correspond to oxidation, which is also an indication of local overheating due to inaccurate laser control leading to a poor weld quality in terms of mechanical properties [12].

This aspect is even more evident from the cross-sections (Fig. 4 (b), episode 1), which is characterized by rapid variations of the weld penetration depth along the line. In this specific case, the local overheating of the material was taking place due to the application of a too high level of laser power generating a highly unstable keyhole that led to the trapping of pores inside the material during the keyhole collapse [10]. The red arrows highlight the pore locations in the magnification in Fig. 4 (b).

After 40 trials, i.e., about 7 min from the beginning of the training (Fig. 4, episode 40), the welds started to be characterized by smoother changes in surface textures and penetration depth.

Confirming the positive trend, significant signs of progress are obtained just after performing other 40 more welds (Fig. 4 (a), episode 80, about 15 min from the beginning), when the texture of the weld surface started to present no perceivable non-uniformities. Nevertheless, some fluctuations in the penetration depth can still be observed (Fig. 4 (b), episode 80).

Finally, a weld comparable to the reference one was only achieved after the completion of other 30 more episodes — see Fig. 4 (a), episode 110 (about 20 min from the start), when the welds began to be characterized by uniform surface texture and constant penetration depth. Fig. 4 (c) also shows the light microscope images of the cross-sections for the trained controlled and reference welds, respectively. As described in Section II-D, the latter was realized after an exhaustive search of the laser parameters and achieved a weld depth of 150 μm, as shown is in Fig. 4 (c), top image. As can be noticed, no measurable differences between the trained controlled weld and the reference one can be found.

Similarly, PG showed identical results apart from a different convergence rate. Indeed, the convergence took about 1.6 times more time compared to *Q*-Learning (see Fig. 3).

B. DISCUSSION

Whether the classifier is of unquestionable fundamental importance as it allows the monitoring of the process, the use of the encoder, on the other side, is debatable. The encoder has indeed some pros and cons that were not obvious before the experiments. As stated in Section III-A, its advantages consist of an effective reduction of the state space dimensionality that potentially simplifies the search of the optimal parameters of the smart agent by capturing a proper parametrization of the signal that can focus only on quality critical events.

In contrast, its drawbacks derive from its output representation, that could not be entirely suited for deriving the dynamics of the system, as its temporal resolution is non-uniform [47]. As a result, the sensitivity of the algorithm to some actions could be reduced and potentially bringing to poor process control.

For the sake of verifying the effectiveness of the encoder, we have also tried to exclude it from the processing pipeline and directly provide the high dimensional raw signals from the sensors as input to the agent.

It resulted in a marginally slower convergence rate in terms of the number of episodes (in the order of tens of episodes), but the two strategies were able to achieve the same results.

We believe that this behavior can be explained by the very first convolutional layer of the agent (see Fig. 2) that, if provided with raw signals, can take over the encoder duty to deliver a good signal representation to the following layers. However, when excluding the encoder, the computations were slowed down due to the more significant input quantities, and we had to increase the time between each episode.

It also has to be mentioned that the present work was realized using a well-controlled laboratory environment and with reliable custom equipment.

These controlled conditions provided a more reproducible laser-material interaction during the welds as they included the processing of always the same material with consistent material properties as well as flat surfaces with identical surface roughness.

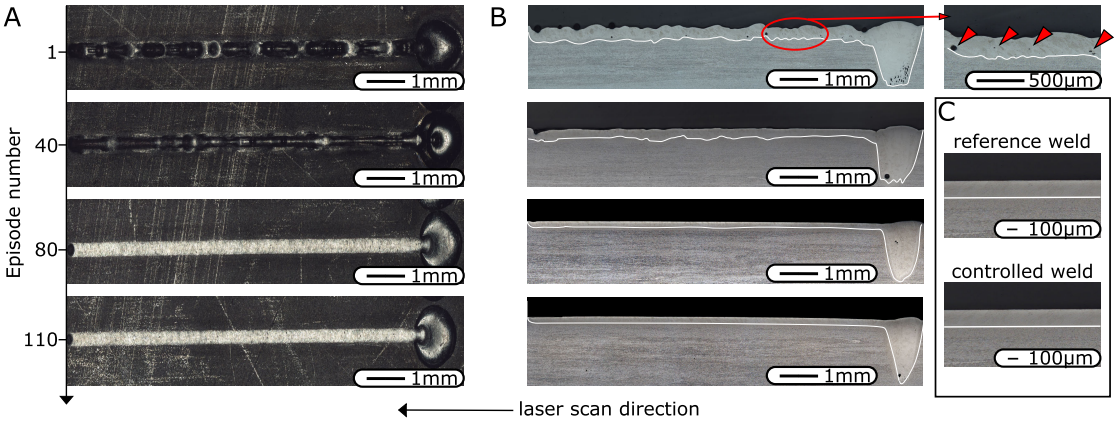

FIGURE 4. Training dynamics of the *Q*-Learning algorithm in terms of welding quality. (a) light microscope pictures of the top view of the welded surface at discrete time points of the algorithm's training; (b) corresponding light microscope pictures of the cross-section of the welds from (a). The magnification for the first episode is shown on the right. The red arrows indicate the pores inside the material; (c) reference weld and controlled weld after the completion of the training procedure. The numbering of the episodes started from the beginning of the training procedure and is indicated on the vertical axis. The arrow at the bottom shows the direction of the laser scan. The white borders denote the boundary of the weld. The deep weld penetration at the beginning of each line constitutes the initial condition from which the algorithm needs to regulate the power.

The well-controlled environment could also be the reason for the small size of the database needed to train encoder and classifier, and this detail may be significantly different in industrial conditions.

V. CONCLUSIONS

This work presents the first results of a study for adaptive closed-loop control of laser welding based on RL applied on a real-life setup.

The developed system includes an encoder that derives efficient representations from the sensory input for the active unit, a feedback network, and a smart agent, which is the active unit itself, that can influence the laser process. The principle of operation is the following: based on the current sensory input provided by the encoder, the agent chooses an action, which leads to a change of its sensory input, and receives a reward — an indirect quality measure of the state the agent ends up in. From this experience — made up by the past sensory input, the executed action, the current input, and the received reward — the agent tries to optimize the outcomes of its actions over time.

In standard RL approaches, the reward signal is provided by the environment and is straightforward to derive. In laser welding, conversely, effective feedback is challenging to provide, as the process is only partially observable since in-depth information of the PZ can be obtained only indirectly from conventional sensors. This reason motivates the introduction of the feedback network: a complete monitoring system based on a DCNN classifier capable of tracking the weld quality in real-time.

In the present work, the control unit was implemented to regulate the output laser power while using the acoustic and optical emission as sensory input. The potential of the system was demonstrated by its capability — without prior knowledge of the process dynamics — to reach a reference weld quality autonomously. The latter was chosen to be represented by the weld with the highest depth achievable without porosity in Ti grade 5 workpiece, to meet the industrial demand for high-quality keyhole welding. This reference weld was determined experimentally and attained a weld depth of 150 μ m without porosity with a laser power of 80 W.

To guide the smart agent, the feedback network and the encoder were trained to recognize not just the reference quality, but also several other counter-examples. For this reason, we collected the acoustic and optical signals from 15 weld experiments at various laser power, namely 20, 40, 60, 80, and 120 W.

The signals were then grouped in 5 categories according to the corresponding weld quality in terms of penetration depth, which were identified via optical inspection of both the surfaces and the cross-sections of the workpieces, and further partitioned in samples of 20 ms. This time span was chosen by taking into consideration the requirement of very high classification accuracy and computation time within the range of 1–5 ms.

After the DCNN classifier and the encoder were trained, the smart agent started its interaction with the laser process by performing line welds with the output laser power being controlled autonomously.

We tested two learning schemes — *Q*-Learning and Policy Gradient — and evaluated their performance both in terms of the evolution of rewards over time, and of the resulting weld quality.

The training time needed for both the algorithms to reach the reference quality was 20 minutes and 33 minutes, respectively. After that time, there was no additional observable increment of weld quality and rewards.

The present results demonstrate the ability of RL to learn a control law for laser welding processes autonomously.

This prospect is highly appealing for the industrial sector as the unit can deal with complex processes without costly simulation and computational tools. Furthermore, the sensor technologies exploited in the present work are commercially available and ready for industrial implementation. It must be emphasized that the proposed framework can also operate with other feedback sensor signals — pyrometer, microphones, or additional photodiodes — making it a rather versatile tool. Further experiments are planned to explore the potential of this approach on more complex conditions, e.g., with surface irregularities or at the interface between two different materials. Additionally, we will increase the number of control variables, including the workpiece velocity and its distance from the laser source. Finally, the RL algorithms will be further enriched with techniques for faster convergence, higher operating frequency, better adaptation under changing materials, and varying noise levels.

REFERENCES

[1] D. Bäuerle, *Laser Processing and Chemistry*. Berlin, Germany: Springer, 1996.

[2] J. R. Berretta and W. Rossi, "Laser welding," in *Encyclopedia Tribology*. Boston, MA, USA: Springer, 2013, pp. 1969–1981.

[3] A. R. Konuk, R. G. K. M. Aarts, A. J. H. I. Veld, T. Sibillano, D. Rizzi, and A. Ancona, "Process control of stainless steel laser welding using an optical spectroscopic sensor," *Phys. Procedia*, vol. 12, pp. 744–751, Jan. 2011.

[4] S. Postma, "Weld pool control in ND: YAG laser welding," Ph.D. dissertation, Dept. Eng. Technol., Univ. Twente, Amsterdam, The Netherlands, 2003. [Online]. Available: https://research.utwente.nl/en/publications/weld-pool-control-in-nd-yag-laser-welding

[5] A. Papacharalampopoulos, P. Stavropoulos, and J. Stavridis, "Adaptive control of thermal processes: Laser welding and additive manufacturing paradigms," *Procedia CIRP*, vol. 67, pp. 233–237, Jan. 2018.

[6] L. Song and J. Mazumder, "Feedback control of melt pool temperature during laser cladding process," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 6, pp. 1349–1356, Nov. 2011.

[7] X. Na, Y. Zhang, Y. Liu, and B. Walcott, "Nonlinear identification of laser welding process," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 4, pp. 927–934, Jul. 2010.

[8] P. A. Hooper, "Melt pool temperature and cooling rates in laser powder bed fusion," *Additive Manuf.*, vol. 22, pp. 548–559, Aug. 2018.

[9] A. Bollig, D. Abel, C. Kratzsch, and S. Kaierle, "Identification and predictive control of laser beam welding using neural networks," in *Proc. Eur. Control Conf. (ECC)*, Sep. 2003, pp. 2457–2462.

[10] M. Courtois, M. Carin, P. Le Masson, S. Gaied, and M. Balabane, "A complete model of keyhole and melt pool dynamics to analyze instabilities and collapse during laser welding," *J. Laser Appl.*, vol. 26, no. 4, Nov. 2014, Art. no. 042001.

[11] X. Jin, L. Li, and Y. Zhang, "A study on fresnel absorption and reflections in the keyhole in deep penetration laser welding," *J. Phys. D, Appl. Phys.*, vol. 35, p. 2304, Sep. 2002.

[12] J. Stavridis, A. Papacharalampopoulos, and P. Stavropoulos, "Quality assessment in laser welding: A critical review," *Int. J. Adv. Manuf. Technol.*, vol. 94, pp. 1825–1847, Feb. 2018.

[13] S. Shevchik, T. Le-Quang, B. Meylan, F. V. Farahani, M. P. Olbinado, A. Rack, G. Masinelli, C. Leinenbach, and K. Wasmer, "Supervised deep learning for real-time quality monitoring of laser welding with X-ray radiographic guidance," *Sci. Rep.*, vol. 10, no. 1, p. 3389, Dec. 2020.

[14] S. A. Shevchik, T. Le-Quang, F. V. Farahani, N. Faivre, B. Meylan, S. Zanoli, and K. Wasmer, "Laser welding quality monitoring via graph support vector machine with data adaptive kernel," *IEEE Access*, vol. 7, pp. 93108–93122, 2019.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[17] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, pp. 1–27, Jan. 2009.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[19] J. Günther, P. M. Pilarski, G. Helfrich, H. Shen, and K. Diepold, "Intelligent laser welding through representation, prediction, and control learning: An architecture with deep neural networks and reinforcement learning," *Mechatronics*, vol. 34, pp. 1–11, Mar. 2016.

[20] C. J. Watkins and P. Dayan, "Technical note: Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.

[21] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 1999, pp. 1057–1063.

[22] L. Bassi, "Industry 4.0: hope, hype or revolution?" in *Proc. IEEE 3rd Int. Forum Res. Technol. Soc. Ind. (RTSI)*, Sep. 2017, pp. 1–6.

[23] T. Le-Quang, S. A. Shevchik, B. Meylan, F. Vakili-Farahani, M. P. Olbinado, A. Rack, and K. Wasmer, "Why is *in situ* quality control of laser keyhole welding a real challenge?" *Procedia CIRP*, vol. 74, pp. 649–653, Jan. 2018.

[24] F. Vakili-Farahani, J. Lungershausen, and K. Wasmer, "Process parameter optimization for wobbling laser spot welding of Ti6Al4 V alloy," *Phys. Procedia*, vol. 83, pp. 483–493, Jan. 2016.

[25] S. Shevchik, T. Le Quang, B. Meylan, and K. Wasmer, "Acoustic emission for *in situ* monitoring of laser processing," in *Proc. 33rd Eur. Conf. Acoustic Emission Test. (EWGAE)*, 2018, pp. 1–10.

[26] F. Vakili-Farahani, J. Lungershausen, and K. Wasmer, "Wavelet analysis of light emission signals in laser beam welding," *J. Laser Appl.*, vol. 29, no. 2, May 2017, Art. no. 022424.

[27] J. Yang, S. Sun, M. Brandt, and W. Yan, "Experimental investigation and 3D finite element prediction of the heat affected zone during laser assisted machining of Ti6Al4 V alloy," *J. Mater. Process. Technol.*, vol. 210, no. 15, pp. 2215–2222, Nov. 2010.

[28] J. Willems, E. Kikken, and B. Depraetere, "Low-dimensional learning control using generic signal parametrizations," *IFAC-PapersOnLine*, vol. 52, no. 29, pp. 280–285, 2019.

[29] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.*, San Francisco, CA, USA: Morgan Kaufmann, 1993, p. 3–10.

[30] J. C. Ye and W. K. Sung, "Understanding geometry of encoder-decoder CNNs," in *Proc. 36th Int. Conf. Mach. Learn., ICML*, Jun. 2019, pp. 12245–12254.

[31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. ICLR*, 2016. [Online]. Available: https://arxiv.org/abs/1511.06434

[32] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, vol. 27, I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver, eds. Washington, DC, USA: Bellevue, Jul. 2012, pp. 37–49.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lile, France, vol. 1, Jul. 2015, pp. 448–456.

[34] H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2684–2691.

[35] R. Ayachi, M. Afif, Y. Said, and M. Atri, "Strided convolution instead of max pooling for memory efficiency of convolutional neural networks," in *Proc. Int. Conf. Sci. Electron., Technol. Inf. Telecommun.* in Smart Innovation, Systems and Technologies, vol. 146, Berlin, Germany: Springer, 2020, pp. 234–243.

[36] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Mach. Intell.*, vol. 1, no. 3, pp. 133–143, Mar. 2019.

[37] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

[38] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: An overview," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, vol. 16, 2018, pp. 426–440.

[39] M. Telgarsky, "Benefits of depth in neural networks," *J. Mach. Learn. Res.*, vol. 49, pp. 1517–1539, Feb. 2016.
[40] P. Petersen and F. Voigtlaender, "Equivalence of approximation by convolutional neural networks and fully-connected networks," *Proc. Amer. Math. Soc.*, vol. 148, no. 4, pp. 1567–1581, Dec. 2019.
[41] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement Learning* (Adaptation, Learning, and Optimization), vol. 12. Berlin, Germany: Springer, 2012, pp. 45–73. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-27645-3_2#citeas
[42] P. Mishra and P. Mishra, "Introduction to PyTorch, tensors, and tensor operations," in *PyTorch Recipes*. New York, NY, USA: Apress, 2019, pp. 1–27.
[43] S. Abrahams, D. Hafner, E. Erwitt, and A. Scarpinelli, *TensorFlow for Machine Intelligence: A Hands-On Introduction to Learning Algorithms*. Santa Rosa, CA, USA: Bleeding Edge Press, 2016. [Online]. Available: https://dl.acm.org/doi/book/10.5555/3125813
[44] H. V. Hasselt, "Double Q-learning," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, vol. 2, 2010, pp. 2613–2621.
[45] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Netw.*, vol. 21, no. 4, pp. 682–697, May 2008.
[46] S. Levine and V. Koltun, "Guided policy search," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, S. Dasgupta and D. McAllester, eds. Atlanta, Georgia: PMLR, Jun. 2013, pp. 1–9. [Online]. Available: http://proceedings.mlr.press/v28/levine13.html
[47] G. Arvanitidis, L. K. Hansen, and S. Hauberg, "Latent space oddity: On the curvature of deep generative models," in *Proc. 6th Int. Conf. Learn. Represent., ICLR*, 2017, pp. 1–16.

GIULIO MASINELLI (Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Bologna, Italy, in 2017, and the M.Sc. degree in electrical engineering (with data science specialization) from the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, in 2019. He is currently pursuing the Ph.D. degree with Swiss Federal Laboratories for Material Science and Technology (EMPA) and EPFL, mainly developing machine learning algorithms for data analysis and industrial automation. His research interests include signal processing and machine learning, with emphasis on deep learning.

TRI LE-QUANG received the B.S. degree in applied physic from Vietnam National University, Ho Chi Minh City, Vietnam, in 2007, the M.Sc. degree in optics from the Friedrich-Schiller-Universität Jena, Germany, in 2013, and the Ph.D. degree in material engineering from the Instituto Superior Tecnico Lisboa, Portugal, in 2017. Since 2017, he has been working as a Postdoctoral Researcher with EMPA, Swiss Federal Laboratories for Materials Science and Technology, Laboratory of Advanced Materials Processing. His research interests include laser material processing, laser technology, and in-situ monitoring.

SILVIO ZANOLI received the B.Sc. degree in electrical engineering from the University of Bologna, Italy, in 2017, and the M.Sc. degree in electrical engineering (with data science and the IoT specialization) from the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, in 2019, where he is currently pursuing the Ph.D. degree in electrical engineering (with data science specialization). His research interests are in signal processing, machine learning, and the IoT with particular attention to low energy solutions.

KILIAN WASMER (Member, IEEE) received the B.S. degree in mechanical engineering from Applied University, Sion, Switzerland, and Applied University, Paderborn, Germany, in 1999, and the Ph.D. degree in mechanical engineering from Imperial College London, Great Britain, in 2003. He joined the Swiss Federal Laboratories for Materials Science and Technology (EMPA), Thun, Switzerland, in 2004, to work on control of crack propagation in semiconductors. He currently leads the Group of Dynamical Processes, Laboratory for Advanced Materials Processing (LAMP). His research interests include materials deformation and wear, crack propagation prediction, and material tool interaction. In the last years, he has focused his work on in situ and real-time observation of complex processes using acoustic and optical sensors in various fields such as in tribology, fracture mechanics, and laser processing. He is in the director committee for additive manufacturing of Swiss Engineering. He is also a member of Swiss tribology, European Working Group of Acoustic Emission (EWGAE), and Swissphotonics.

SERGEY A. SHEVCHIK received the M.Sc. degree in control from the Moscow Engineering Physics Institute, Russia, in 2003, and the Ph.D. degree in biophotonics from the General Physics Institute, Russia, in 2005. He stayed until 2009 as a Postdoctoral Researcher at the General Physics Institute. From 2009 to 2012, he was with the Kurchatov Institute, Russia, developing human–machine interfaces. In 2012 and 2014, he was with the University of Bern, investigating multi-view geometry. Since 2014, he has with Swiss Federal Laboratories for Material Science and Technology (EMPA), working on industrial automation. His current interest is in signal processing.