Transfer learning, alternative approaches, and visualization of a convolutional neural network for retrieval of the internuclear distance in a molecule from photoelectron momentum distributions

N. I. Shvetsov-Shilovski* and M. Lein
Institut für Theoretische Physik, Leibniz Universität Hannover, 30167 Hannover, Germany

(Received 2 November 2022; revised 20 January 2023; accepted 24 February 2023; published 14 March 2023)

We investigate the application of deep learning to the retrieval of the internuclear distance in the two-dimensional H_2^+ molecule from the momentum distribution of photoelectrons produced by strong-field ionization. We study the effect of the carrier-envelope phase on the prediction of the internuclear distance with a convolutional neural network and investigate the possibility of reconstruction of the internuclear distance from one-dimensional momentum distributions. We apply the transfer learning technique to make our convolutional neural network applicable to distributions obtained for parameters outside the ranges of the training data. The convolutional neural network is compared with alternative approaches to this problem, including the direct comparison of momentum distributions, support-vector machines, and decision trees. These alternative methods are found to possess very limited transferability. Finally, we use the occlusion-sensitivity technique to extract the features that allow a neural network to make its decisions.

DOI: 10.1103/PhysRevA.107.033106

I. INTRODUCTION

Machine learning focuses on the development of algorithms and methods that are able to learn, i.e., to use data in order to improve automatically through experience [1]. Methods of machine learning are presently widely used in almost all branches of modern science. Strong-field, ultrafast, and attosecond physics that studies the nonlinear processes originating from interaction of strong laser pulses with atoms and molecules [2–7] is no exception. The prediction of the flux of high-order harmonics [8], the prediction of the ground-state energy of an electron confined by a two-dimensional (2D) potential [9], the retrieval of the intensity and the carrier-envelope phase (CEP) of ultrashort laser pulses from frequency-resolved optical gating traces [10] and dispersion scan traces [11], and the efficient implementation of the trajectory-based Coulomb-corrected strong-field approximation [12,13] by using of a deep neural network [14] are examples of machine-learning applications in strong-field physics and related areas. More recently, convolutional neural networks have been used to predict high-order harmonic spectra for model di- and triatomic molecules when the laser intensity, internuclear distance, and orientation of the molecule were given and vice versa, to retrieve molecular parameters from a given high-order harmonic spectrum, see Ref. [15]. Recently a convolutional neural network was used to retrieve the geometric structure of gas-phase molecules from experimentally measured laser-induced electron diffraction images [16]. Furthermore, machine learning was applied to retrieve the internuclear distance in a molecule from a given photoelectron momentum distribution produced by a strong laser pulse [17]. The problems considered in Refs. [16,17] are important for the development of tools aimed at time-resolved molecular imaging, i.e., visualization of molecular dynamics in real time.

Time-resolved molecular imaging requires high resolution in both space and time. Indeed, molecular transformations occur due to motion of atoms on the angstrom scale. Simultaneously, a chemical reaction has a typical duration of less than a picosecond. It would be desirable to apply time-resolved molecular imaging to large molecules, e.g., biomolecules. Eventually, these techniques should allow to study not only the dynamics of the atomic nuclei, but also the electronic dynamics. Important methods for time-resolved molecular imaging include, for example, optical pump-probe spectroscopy, time-resolved electron and x-ray diffraction, and ultrafast x-ray spectroscopy [18].

The advances in laser technologies over the last decades, especially the emergence of tabletop intense optical laser systems and progress in the development of free-electron lasers, have offered possibilities of new methods for real-time molecular imaging. Among these new methods based on strong-field phenomena are laser-induced Coulomb explosion imaging [19–22], laser-assisted electron diffraction [23,24], high-order harmonic orbital tomography [25,26], laser-induced electron diffraction [27–30], and strong-field photoelectron holography [31]. Laser-induced electron diffraction and strong-field photoelectron holography rely on the analysis of momentum distributions of electrons removed by a strong laser pulse. At present these methods are beginning to be successfully applied to dynamical systems, see, e.g., Refs. [32,33]. Therefore, in the near future we can expect experiments aimed at obtaining information about the nuclear motion in a molecule interacting with a strong laser pulse from photoelectron momentum distributions (PMDs). There is no

*n79@narod.ru

 ©2023 American Physical Society

doubt that these forthcoming experiments will bring us closer to the ability to experimentally image the atomic positions in the course of chemical transformations. The tools providing such an ability are expected to revolutionize chemistry, biology, and material science. They will also allow us to get deeper physical insight into a variety of complex physical processes that take place in molecules: ionization, dissociation, high-order harmonic generation, charge transfer, etc. However, the development of these tools based on laser-induced electron diffraction or strong-field photoelectron holography will require the solution of a large number of fundamental physical problems. Before looking for the effects of nuclear motion in the measured PMDs, it is first necessary to analyze the distributions obtained for different internuclear distances with fixed nuclei. More specifically, it is required to get insight into the relation between a given momentum distribution and internuclear distance at fixed nuclei.

In Ref. [17] a convolutional neural network (CNN) was used to retrieve the internuclear distance of a two-dimensional (2D) model H_2^+ molecule from photoelectron momentum distributions generated by a strong few-cycle laser pulse. The momentum distributions were calculated from the direct numerical solution of the time-dependent Schrödinger equation (TDSE). It was shown that a CNN trained on a relatively small number of electron momentum distributions predicts the internuclear distance with a mean absolute error (MAE) below 0.1 a.u. Furthermore, a neural network was able to retrieve both the internuclear distance and the laser intensity from a given photoelectron momentum distribution. Finally, the effect of focal averaging was studied in Ref. [17]. It was shown that a CNN trained on focal averaged momentum distributions also performs well in the retrieval of the internuclear distance.

A number of important questions concerning deep learning for the molecular-structure retrieval remain to be studied. First, it is important to study the effect of the CEP, since the variation of the CEP of a few-cycle pulse affects the PMDs significantly. Second, it deserves to be studied whether the internuclear distance can be retrieved using one-dimensional (1D) momentum distributions. Third, the CNN presented in Ref. [17] shows limited transferability, i.e., it may fail for PMDs corresponding to parameters beyond the ranges of the training data. The transferability problem can be tackled with the transfer learning technique [34].

Furthermore, it is interesting to compare the CNN with alternative methods. The simplest possible method is the direct comparison of a given PMD with a precalculated set of PMDs obtained for various internuclear distances [17]. Support vector machines (see, e.g., Refs. [35–38]) and decision trees [39–41] are further alternatives. Both methods, combined with the histogram of oriented gradients [42], were extensively used for object detection and image comparison before CNNs became widespread. Finally, since a CNN is a "black box," it is interesting to understand how the CNN of Ref. [17] makes its decisions, i.e., to have "a look under the hood" of the neural network. Although this is a difficult task, a number of so-called visualization and explanation methods for deep neural networks have been developed [43–45]. These methods can potentially uncover the features of the PMDs that are used by the neural network to recognize the internuclear distance.

In this paper we address the above-mentioned questions. We apply the CNN to PMDs obtained for three variable parameters: internuclear distance, laser intensity, and CEP. The precise value of the intensity and the CEP are notoriously difficult to control in experiment. This is the motivation to vary these parameters in the simulations and to check the effect on the retrieval problem. Thus, we study the effect of the CEP on the reconstruction of the internuclear distance. We train another CNN to retrieve the internuclear distance based on the 1D momentum distributions. Using the transfer learning technique, we achieve transferability of the CNN and we perform a number of transferability tests. We then compare the machine-learning approach with alternative methods. For the purpose of direct comparison we apply not only the mean squared pixel-wise error, but also the histogram of oriented gradients (HOG) and the scale-invariant feature transform (SIFT) algorithm proposed by Lowe [46]. The image descriptors obtained from the SIFT algorithm are invariant with respect to uniform scaling and orientation changes. They are also partially invariant with respect to affine distortions [46]. Finally, by applying visualization methods, we extract the features that allow the CNN to classify PMDs by internuclear distance.

The paper is organized as follows. In Sec. II we briefly review the architecture of the CNN and the method used for the solution of the TDSE. In Sec. III we study the effect of the CEP on the retrieval of the internuclear distance and train a CNN aimed at reconstruction of the internuclear distance from 1D PMDs. In Sec. IV we use the transfer learning to make the CNN applicable beyond the parameter ranges of the training data set. We compare the application of the CNN to alternative methods in Sec. V, and in Sec. VI we apply the visualization methods. The conclusions are given in Sec. VII. Atomic units are used throughout the paper unless indicated otherwise.

II. MODEL

The architecture of the neural network that we use for prediction of the internuclear distance was discussed in Ref. [17]. We repeat here the most important points to make the presentation self-contained. The same applies to our method for the solution of the 2D TDSE.

A. Architecture and application of convolutional neural network

The architecture of a neural network should be consistent with the format of the data used for training. For the problem at hand the training data are pairs of electron momentum distributions and the corresponding internuclear distances. In terms of machine learning, the distributions and the internuclear distances are the images and labels, respectively. Since we consider ionization of the molecule by a linearly polarized laser pulse, it is natural to assume the aspect ratio of every image as $2:1$ [17].

The images used by the CNN are preprocessed as follows. We calculate the decimal logarithm of the normalized PMD, i.e., $W = \log_{10}(\mathrm{PMD}/\mathrm{PMD}_{\max})$, where $\mathrm{PMD}_{\max}$ is the absolute maximum of the distribution, and we set $W = -5$

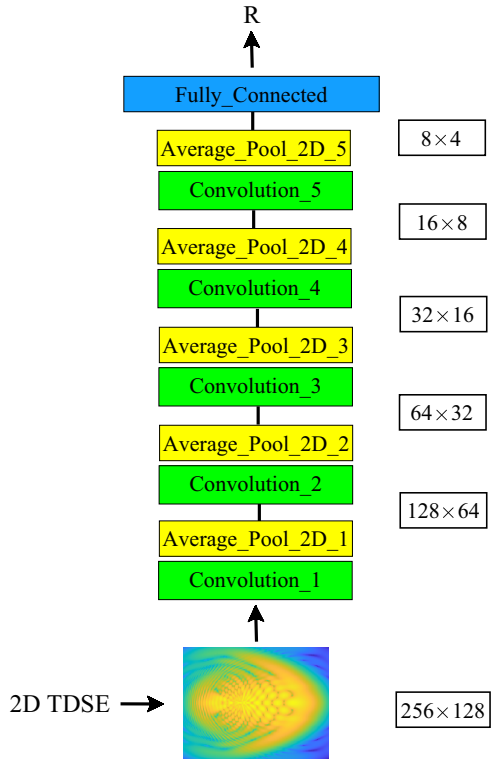

FIG. 1. The architecture of the neural network used for retrieval of the internuclear distance R. The sizes of the image after each average pooling layer are indicated on the right side.

instead of all values smaller than -5. Therefore, we account not only for the low-energy part of the distribution, but also the beginning of its high-energy part. The low-energy part is formed by the electrons that arrive at the detector without recolliding with their parent ions. These electrons are referred to as direct electrons. The direct electrons have energies below $2U_p$ (i.e., momenta below $2\sqrt{U_p}$), where $U_p = F^2/4\omega^2$ is the ponderomotive energy. In contrast to this, the high-energy part of the PMD arises due the rescattered electrons that are driven back to their parent ions by the laser field and rescatter off them by large angles. We select a rectangular part of the image that contains all values of W exceeding -5. By using bicubic interpolation, we downsize this rectangular part of the PMD to the size of 256×128 pixels in order to avoid large matrices, which cause heavy computational costs and slow down the training process. We then rescale all the elements of the resulting matrix to map the maximum value to 255 and the minimum value to zero. The resulting images are given to the neural network.

The architecture of the deep neural network is shown in Fig. 1. The neural network consists of five pairs of nonreducing convolutional layers and reducing average pooling layers. We have found that using a smaller number of these pairs worsens the performance of the CNN. Simultaneously, an increased number does not lead to a significant improvement of the results. Each of the convolutional layers consists of 32 filters, and the size of each filter is 3×3 pixels. The padding needed to make the size of the output of each convolutional layer equal to the size of its input is to be calculated and used by the software in the training process. A convolutional layer convolves its input and produces new images (so-called feature maps [47]). The number of the feature maps is equal to the number of filters. The elements of the filter matrices are the main trainable parameters of the CNN. Following the convolution of their input images, the convolutional layers apply the rectified linear unit (ReLU) activation function, i.e., the piecewise linear function defined as $\mathrm{ReLU}(x) = \max(0, x)$.

The average pooling layers reduce the image size. These layers divide their input images into pooling regions and average the images over each pooling region. The size of the pooling regions is chosen to be 2×2 pixels, and therefore each average pooling layer reduces the image length and height by a factor of two. The output of the last average pooling layer is given to the dropout layer (not shown in Fig. 1). The dropout layer randomly substitutes some fraction of its input values by zeros. In our case this fraction is chosen to be equal to 20%. The presence of the dropout layer helps to prevent overfitting, i.e., the situation in which the CNN learns too many details of the training data, including the noise, and as a result performs poorly on data it has not seen before [48]. The dropout layer is connected to the last layer of our CNN, i.e., a fully connected layer. There are 37 857 trainable parameters in our neural network. These are $3 \times 3 \times 1 \times 32$ of the weights of the first convolutional layer, $4 \times (3 \times 3 \times 32 \times 32)$ weights of the remaining four convolutional layers, $5 \times (32 + 32)$ biases of all these five convolutional layers, 5×32 offsets and 5×32 scale factors of the five batch normalization layers (not shown in Fig. 1) accompanying the convolutional layers and normalizing the input to make the training faster, as well as 384 weights and 1 bias of the fully connected layer. We use the CNN to solve the regression problem, not a classification problem. Indeed, we train a neural network to retrieve the internuclear distance R from a given electron momentum distribution. Therefore, the fully connected layer calculates the output of the neural network—the internuclear distance R (possibly along with any other value of interest). If only R is to be calculated, the fully connected layer multiplies its input (column matrix) by a weight (row matrix) and adds a bias value. The neural network is implemented using the MATLAB package [49].

In the process of training, we minimize the loss function—the measure of deviation between predictions of the CNN and the known labels (internuclear distances) of the training set. The mean squared error (MSE) is used as the measure of the deviation. For minimization we apply the stochastic gradient descent optimizer. We note that the Adam optimization algorithm shows better performance for the problem at hand than the stochastic gradient descent (SGD) optimizer. However, the application of the Adam algorithm instead of SGD for training of the networks discussed here does affect conclusions of the present study. Therefore, in order to ensure consistency with our previous work [17], we use the SGD algorithm for the training process. We split the training data into minibatches, each consisting of 30 images, and use one minibatch for each training iteration. We start the training process with the learning rate $l_r = 10^{-3}$ and decrease the rate by a factor of 10 after 20 training epochs. We find that about 30 epochs are enough for convergence of the loss function. In order to ensure that each PMD creates an unbiased change

in the CNN, we shuffle the training data before each training epoch.

B. Numerical solution of time-dependent Schrödinger equation

We perform our TDSE simulations for a few-cycle linearly polarized laser pulse that is defined through the vector potential

$$\vec{A}(t) = (-1)^{n_p}\frac{F_0}{\omega}\sin^2\left(\frac{\omega t}{2n_p}\right)\sin(\omega t+\varphi)\vec{e}_x. \tag{1}$$

Here F_0 is the field amplitude, ω is the laser frequency, n_p is the number of optical cycles within the pulse, φ is the CEP, and $\vec{e}_x$ is the unit vector in the direction of the x axis (polarization direction). The laser pulse defined by Eq. (1) is present between $t = 0$ and $t = (2\pi/\omega)\cdot n_p$. The electric field is to be calculated from the vector potential (1) as $\vec{F}(t) = -d\vec{A}/dt$.

The velocity gauge TDSE for the 2D H_2^+ molecular ion reads as

$$i\frac{\partial}{\partial t}\Psi(x,y,t) = \left\{-\frac{1}{2}\left(\frac{\partial}{\partial x^2}+\frac{\partial}{\partial y^2}\right) - iA_x(t)\frac{\partial}{\partial x} + V(x,y)\right\}\Psi(x,y,t). \tag{2}$$

Here $\Psi(x,y,t)$ is the wave function and

$$V(x,y) = -\frac{1}{\sqrt{\left(x-\frac{1}{2}R\cos\alpha\right)^2+\left(y-\frac{1}{2}R\sin\alpha\right)^2+a}} - \frac{1}{\sqrt{\left(x+\frac{1}{2}R\cos\alpha\right)^2+\left(y+\frac{1}{2}R\sin\alpha\right)^2+a}} \tag{3}$$

is the soft-core binding potential of the model H_2^+ molecule in the approximation of frozen nuclei. In Eq. (3), a is the soft-core parameter, R is the internuclear distance, and α is the angle between the molecular axis and the polarization direction (orientation angle). We solve the TDSE (2) by applying the Feit-Fleck-Steiger split-operator method, see Ref. [50]. We use imaginary-time propagation to obtain the wave function of the ground state.

The computational grid extending over $x \in [-400, 400]$ a.u. and $y \in [-200, 200]$ a.u. is centered at the origin ($x = 0, y = 0$). Our grids in x and y directions consist of 4096 and 2048 points, respectively. Therefore, we use equal grid spacings for both directions: $\Delta x = \Delta y \approx 0.1953$ a.u. The TDSE is propagated from $t = 0$ (beginning of the laser pulse) to $t = 4t_f$ with the time step $\Delta t = 0.0184$ a.u. Absorbing boundaries are used to prevent unphysical reflections from the boundary of the computational box. More specifically, at every step of the time propagation the wave function is multiplied by the mask,

$$M(x,y) = \begin{cases} 1 & \text{for } r \leqslant r_b \\ \exp[-\beta(r-r_b)^2] & \text{for } r > r_b \end{cases}, \tag{4}$$

where $\beta = 10^{-4}$, $r = \sqrt{x^2+y^2}$, and $r_b = 150$ a.u. [17]. We obtain the electron momentum distributions by using the mask method, see Refs. [27,51].

In order to obtain the training and validation sets of PMDs we solve the TDSE (2) for N randomly chosen internuclear distances and peak laser intensities: $R_k \in [1.0, 8.0]$ a.u. and $I_k \in [1.0, 4.0] \times 10^{14}$ W/cm^2, where $k = 1, ..., N$ [17]. We also calculate focal-volume averaged momentum distributions:

$$\frac{dP}{d^3\vec{k}} = \int_0^{I_0}\frac{dP(I)}{d^3\vec{k}}\left(-\frac{\partial V}{\partial I}\right)dI, \tag{5}$$

where I_0 is the peak laser intensity, $dP(I)/d^3\vec{k}$ is the PMD for a fixed intensity I, and $-(\partial V/\partial I)dI$ is the focal volume element corresponding to the intensity range between I and $I + dI$, see Ref. [52]. Here we assume a Lorentzian distribution of the laser intensity along the polarization direction, and a Gaussian intensity distribution in the orthogonal direction, see, e.g., Refs. [7,53,54]. For such a beam the focal volume element reads as

$$\left(-\frac{\partial V}{\partial I}\right)dI \sim \frac{I_0}{I}\left(\frac{I_0}{I}+2\right)\sqrt{\frac{I_0}{I}-1}\,dI. \tag{6}$$

The integral (5) can be calculated by using the trapezoidal rule. Both the focal volume element and the distribution $dP(I)/d^3\vec{k}$ for a fixed intensity decrease rapidly with I. As a result, about 10 intensity points are sufficient to calculate the integral (5). Since the solution of one 2D TDSE (2) takes usually about 4–8 hours on eight computer cores working in parallel, 40–80 hours are needed to calculate one focal-volume averaged PMD. The calculation of $N_a = 100$ focal-volume averaged distributions requires about one week on a modern computer cluster. Such a set is too small to train a neural network. In order to augment our training data set, we have used 2D interpolation on an irregular grid (see, e.g., Ref. [55]) in the (R, I_0) plane [17]. Since focal-volume averaged momentum distributions are smooth functions of both momentum components, the application of the interpolation technique is well justified. The approach based on the interpolation allows us to create a large set of focal averaged PMDs in a reasonable time.

III. EFFECT OF THE CARRIER-ENVELOPE PHASE OF THE LASER PULSE AND RETRIEVAL OF THE INTERNUCLEAR DISTANCE FROM ONE-DIMENSIONAL MOMENTUM DISTRIBUTIONS

We first study the effect of the CEP on the retrieval of the internuclear distance using the neural network. To this end, we calculate $N = 3000$ electron momentum distributions for N random CEPs $\varphi_k (k = 1, ..., N)$. The corresponding internuclear distances R_k and peak laser intensities I_k are also chosen randomly in the ranges [1.0, 8.0] a.u. and [1.0, 4.0] $\times$ 10^{14} W/cm^2, respectively. Thus, the distributions of the new data set depend on the three random parameters: R, I, and φ. We then split this set of the PMDs into the training and test sets in the ratio 0.75 : 0.25. The 2250 distributions of the training set are used to train a neural network aimed at retrieval of the internuclear distance. This neural network is tested on the 750 PMDs of the test set. We find that the internuclear distance is predicted with an MAE of 0.18 a.u, see Fig. 2(a). This should be compared with the MAE of 0.07 a.u. obtained for the CNN trained on the distributions depending

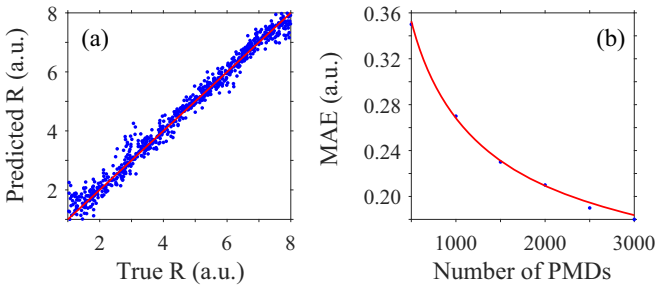

FIG. 2. (a) Plot of predicted vs true internuclear distances illustrating the performance of the CNN for the distributions that depend on the three parameters: laser intensity, internuclear distance, and CEP. (b) The MAE for the internuclear distance retrieved with the neural network as a function of number of images used for training. The blue dots and the red curve correspond to the points obtained in numerical experiments and the fit with a rational function (see text), respectively.

on only two parameters: internuclear distance and peak laser intensity.

Then the following question arises: how many distributions depending on all the three parameters (R, I, φ) are needed so that the network is able to predict the distance R with a MAE less than a certain value., e.g., 0.1 a.u.? In order to answer this question, we train the CNN using different numbers of electron momentum distributions ($N = 500$, 1000, 1500, 2000, and 2500) and calculate the corresponding MAEs. Using this data, we fit the MAE as a function of N: $\text{MAE} = aN^b + c$, where $a = 7.241$, $b = -0.526$, and $c = 0.075$, see Fig. 2(b). By extrapolating this fit, we find that about 50 000 images are needed to achieve an MAE less than 0.1 a.u. We note that this result is similar to the outcomes of the study of Ref. [15]: about 30 000 training samples that depend on three different parameters (internuclear distance, peak laser intensity, and the angle between the molecular axis and the polarization direction) were needed to predict the dipole acceleration with high accuracy [15].

In order to investigate the reconstruction of the internuclear distance using 1D images (curves) we turn to the electron momentum distributions along the polarization direction dP/dk_x (2D image integrated along k_y). We calculate a set of such distributions and use it for training of two neural networks with different architecture. The first neural network is a CNN. It has similar architecture as the neural network shown in Fig. 1. This CNN operates with the figures of the size of 256×1 pixels. Therefore, the size of its convolutional filters is 3×1 pixels, and the size of its pooling regions is 2×1 pixels. For a set of 3000 1D distributions this CNN provides an MAE for the internuclear distance equal to 0.27 a.u. The second neural network has a simpler architecture. It consists of an input layer and four fully connected layers, each followed by a batch normalization layer. Each of the first three fully connected layers has 512×512 weights and 512 biases, and the fourth layer contains 512 weights. The MAE for the internuclear distance obtained by using the second neural network is equal to 0.16. a.u. A change in number of the fully connected layers or in the sizes of these layers does not lead to a substantial decrease of the MAE. We expect slightly worse results when using electron energy spectra, since the nonlinear mapping $E = \vec{k}^2/2$ between the electron energy and momentum is needed to obtain the spectrum. We therefore conclude that the usage of the 1D distributions leads to lower accuracy of retrieval of the internuclear distance. We attribute this to the fact that the 2D distributions contain more information than the 1D PMDs.

IV. TRANSFERABILITY OF THE CONVOLUTIONAL NEURAL NETWORK

We recall that the first neural network of Ref. [17] is trained on a set of distributions calculated for random internuclear distances and fixed (not focal averaged) laser intensities. This CNN shows only limited transferability. We aim to develop new neural networks capable of predicting internuclear distances correctly even for PMDs calculated at parameters that are beyond the corresponding parameter ranges of the training data. To this end, we apply the transfer learning technique to the CNN of Ref. [17]. We freeze all the weights of the original CNN except for those of the layers close to the output layer. Then this pretrained CNN is further trained using a small data set that is outside of the initial training data space. In this way, we modify the CNN to make it applicable to (i) focal-volume averaged PMDs, (ii) larger internuclear distances $8.0 < R < 12.0$ a.u., (iii) nonzero angles between the molecular axis and polarization direction, and (iv) nonzero CEPs.

We begin with the momentum distributions averaged over the focal volume. As in Ref. [17], we use 2D interpolation on an irregular grid to obtain a set of $N = 1000$ averaged momentum distributions from a smaller set consisting of only $N_a = 100$ focal-volume averaged PMDs. We freeze the weights of the first four convolutional layers and retrain the fifth convolutional layer and the fully connected layer. The training is performed for minibatches consisting of 30 images with the learning rate 10^{-2}. We find that the set of $N = 750$ focal averaged PMDs is enough to achieve the MAE 0.15 a.u. for the internuclear distance R, see Fig. 3(a). It should be stressed that this MAE was obtained on the independent test set of calculated (not interpolated) focal-volume averaged PMDs. A CNN trained on 6000 focal-volume averaged PMDs provides the MAE 0.14 a.u. [17]. About 5 min are needed to train the new CNN using 6000 focal averaged PMDs, whereas the application of the transfer learning technique with 750 averaged momentum distributions requires 35–40 s. This applies to a modern PC using a graphic processing unit. Thus, we conclude that for focal averaged PMDs the transfer learning technique has one, but very important, advantage compared to training of a new CNN from the ground up: a smaller training set is needed to achieve the same absolute error. Since the calculation of the training and validation data sets is usually computationally demanding, the application of transfer learning can reduce the computational costs substantially.

We then apply transfer learning to make our CNN applicable to distributions obtained for larger internuclear distances. To this end, we use a set of PMDs obtained for $8.0 < R < 12.0$ a.u. Here we fix the first three convolutional layers and retrain the neural network with the learning rate 10^{-3}, since in the case of larger internuclear distances this provides

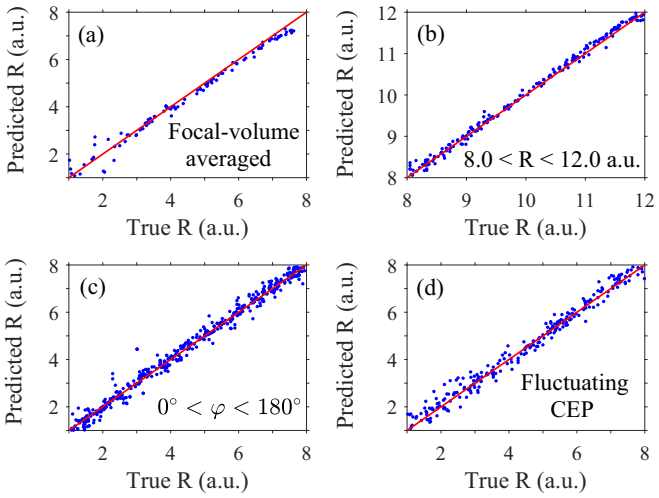

FIG. 3. Plots of predicted vs true internuclear distances illustrating the performance of the initial CNN [see Sec. II A] after application of the transfer learning technique. (a), (b), (c), and (d) correspond to the neural network receiving focal-volume averaged momentum distributions, distributions obtained for $8.0 < R < 12.0$ a.u., distributions calculated for different orientation angles $0^\circ < \alpha < 180^\circ$, and distributions obtained for fluctuating CEP, respectively.

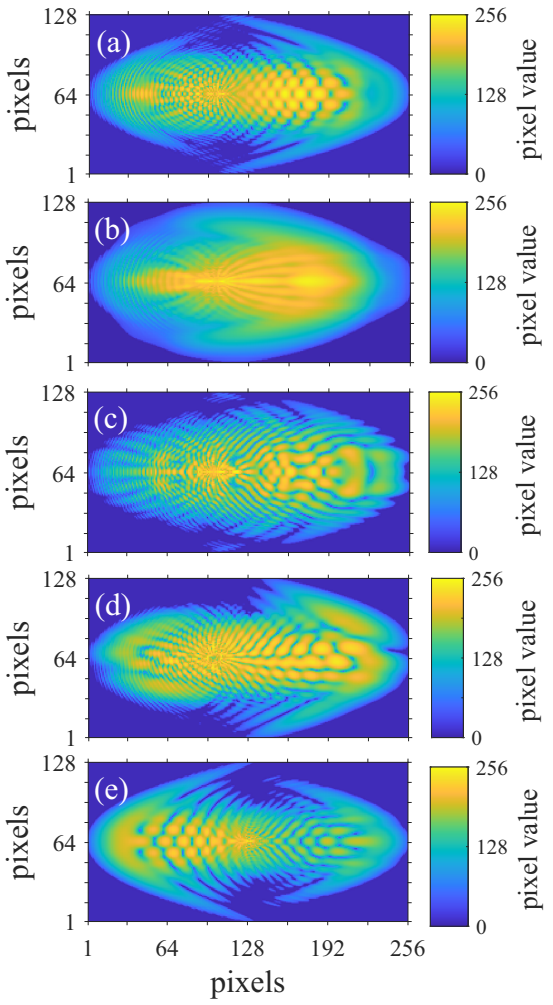

FIG. 4. Electron momentum distributions for ionization of the 2D H_2^+ molecule as they are seen by the neural network, i.e., as 256×128 matrices with element values between 0 and 255. The distributions are obtained from the solution of the TDSE. The laser pulse duration is $n_p = 2$ cycles, the wavelength is 800 nm, and the peak laser intensity is 3.9×10^{14} W/cm^2. (a) corresponds to the internuclear distance $R = 2.0$ a.u., the orientation angle $\alpha = 0^\circ$, and the CEP $\varphi = 0$. (b) shows the focal-volume averaged momentum distribution calculated for the parameters of panel (a). (c) corresponds to the internuclear distance $R = 10.9$ a.u., with the other parameters as in (a). (d) displays the distribution calculated for $\alpha = 158^\circ$ with the rest of the parameters as in (a). (e) corresponds to the CEP $\varphi = 4.50$ rad, with the rest of the parameters as in (a).

better results as compared to the freezing of the first four convolutional layers and the use of the learning rate 0.01. Only about 600 images are needed to obtain the MAE 0.06 a.u. for R, see Fig. 3(b). We note that a new CNN trained on the same set of 600 distributions with large internuclear distances provides the MAE of 0.11 a.u. Thus, a set of 600 distributions is not sufficient to achieve the MAE less than 0.1 by training a new CNN. The transfer learning technique provides better results for the same number of distributions. This could be expected, since more trainable parameters are to be optimized in the training of a new neural network compared to the retraining of only 2–3 layers of the CNN as required by the transfer learning technique.

It should be noted that focal averaging and larger internuclear distances are relatively easy cases for transfer learning. Indeed, the focal-volume averaged PMDs and those corresponding to larger internuclear distances resemble (in terms of the symmetry and the positions of their maxima) the distributions calculated for fixed laser intensities and smaller values of R [compare Fig. 4(a) with Figs. 4(b) and 4(c)]. The situation changes for the PMDs corresponding to nonzero angles between the molecular axis and the laser polarization direction, see Figs. 4(a) and 4(d). It is seen that the distribution of Fig. 4(d) is substantially deformed compared to Fig. 4(a), and this deformation is rather complex. As a result, the transfer learning technique performs worse than in the two previous cases. Indeed, the application of transfer learning using 750 PMDs calculated for random angles $0^\circ \leqslant \alpha \leqslant 180^\circ$ leads to the MAE 0.17 a.u., see Fig. 3(c). Here we again fix the first three convolutional layers and choose the learning rate $l_r = 10^{-3}$. This result should be compared with the MAE of 0.25 provided by a CNN trained on the same set of 750 distributions with nonzero orientation angles. Therefore, transfer learning provides a significant advantage. On the other hand, when 2250 PMDs obtained for nonzero angles are used for training of the new neural network, the corresponding MAE is equal to 0.18 a.u., which is close to the transfer learning result achieved with 750 distributions only. Nevertheless, we made an attempt to improve the outcomes of transfer learning without calculating new momentum distributions. To this end, we have augmented the data set used for transfer learning by rotating the available PMDs by small angles and reflecting them with respect to the polarization direction. However, this does not reduce the corresponding MAE substantially. The

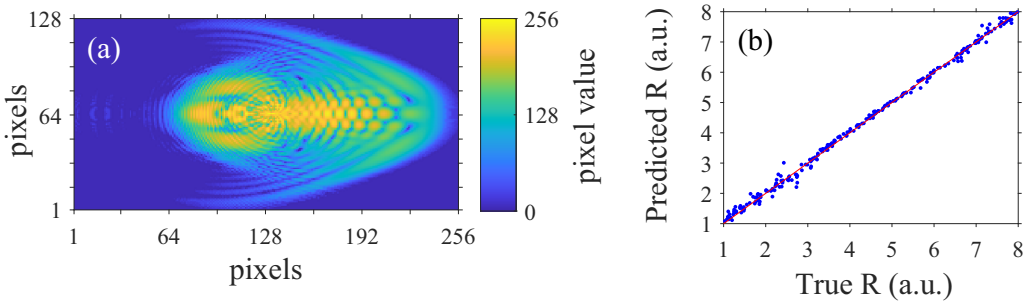

FIG. 5. (a) Electron momentum distribution for ionization of the 2D HeH^+ molecule as it is seen by the neural network. The distribution is calculated from the solution of the TDSE for the internuclear distance $R = 2.0$ a.u., the orientation angle $\alpha = 0°$, and the CEP $\varphi = 0$. The laser parameters are as in Fig. 4. (b) Plot of predicted vs true internuclear distances illustrating performance of the initial CNN after application of transfer learning technique for the distributions obtained in ionization of the HeH^+ molecule.

same applies to a modified approach replacing the last 2–3 layers by new untrained ones and subsequent training of them.

We also apply transfer learning to the case of nonzero CEP. Freezing the first three layers of the neural network, using a set of 750 distributions calculated for random CEPs, and choosing the learning rate 10^{-2}, we obtain the MAE 0.23 a.u., see Fig. 3(d). This result should be compared with the MAE of 0.31 a.u. obtained for a new CNN trained on the same set of 750 momentum distributions, as well as with the MAE 0.18 a.u., corresponding to the training of a new CNN using 2250 PMDs with random CEPs, see Sec. III. We conclude that transfer learning provides some gain even in the case of varying CEP, which, judging from the strong CEP dependence of the PMDs, would seem to be a difficult case [cf. Figs. 4(a) and 4(e)].

Finally, we consider the question of transferability of our CNN to another molecule: HeH^+. In the 2D case this asymmetric molecule can be described by the soft-core potential:

$$V(\vec{r}) = -\frac{1}{\sqrt{(\vec{r}-\vec{r}_1)^2+1/2}} - \frac{1+e^{-\beta(\vec{r}-\vec{r}_2)^2}}{\sqrt{(\vec{r}-\vec{r}_2)^2+1/2}}, \qquad (7)$$

see, e.g., Ref. [56]. In Eq. (7) $\vec{r}_1$ and $\vec{r}_2$ are the locations of the proton and the He nuclei, respectively. For $\beta = 1.063$ the potential (7) reproduces the ionization potential of the 3D HeH^+ molecule equal to 1.66 a.u. at the equilibrium distance $|\vec{r}_1 - \vec{r}_2| = 1.4$ a.u. By solving the TDSE (2) with the potential of Eq. (7), we calculate a set of 1000 electron momentum distributions generated in ionization by the laser pulse (1). An example of the corresponding PMD is shown in Fig. 5(a). We use the same ranges of internuclear distances and laser intensities as for the H_2^+ molecule. At first, we test the CNN of Ref. [17] (trained on H_2^+) on the set of distributions calculated for HeH^+. This leads to an MAE for R equal to 1.7 a.u. In order to make our CNN applicable to the HeH^+ molecule, we again use the transfer learning technique. We freeze the first three layers of the CNN and choose the learning rate $l_r = 10^{-2}$. We find that about 750 images are needed to achieve the MAE of 0.08 a.u. [see Fig. 5(b)], whereas a new CNN trained on 1000 PMDs for the HeH^+ molecule provides the MAE of 0.13 a.u. Therefore, the CNN of Ref. [17] can be used for molecules with asymmetric centers, and the transfer learning technique again provides an advantage.

The question may arise: Why does the transfer learning technique work for distributions of very different shapes that are obtained for different parameters? In order to answer this question, it is necessary to consider that the applicability of the transfer learning technique for a CNN is based on the fact that different convolutional layers learn different features of input images. The bottom convolutional layers, i.e., the layers close to the image input layer, learn the edges of the images and their coarse features. In contrast to this, the top convolutional layers close to the output of the network are responsible for more subtle features of the inner parts of the images, see, e.g., Ref. [43]. Since the boundaries between the areas of zero and nonzero ionization probabilities as well as some other coarse features of the PMDs obtained for different parameters are often similar, the weights of the bottom layers of our initial CNN of Ref. [17] can be used when this CNN is applied to distributions calculated for parameters outside the range of the training data. The smaller details of the distributions of Figs. 4(a)–4(e) and 5(a) are very different. Therefore, the weights of the top layers should be retrained, which is just the essence of the transfer learning technique. In the situations where the coarse details of the corresponding PMDs are close to the coarse details of the distributions used to train the initial CNN [compare, e.g., Fig. 4(a) with Figs. 4(b), 4(c) and 5(a)], transfer learning turns out to be very efficient. In the opposite cases, such as nonzero orientation angle or nonzero CEP [Figs. 4(d) and 4(e)], the transfer learning technique demonstrates worse performance.

V. ALTERNATIVE APPROACHES FOR THE RETRIEVAL OF THE INTERNUCLEAR DISTANCE

An approach based on the direct comparison of a given momentum distribution with a set of the precalculated distributions obtained for different internuclear distances was implemented already in Ref. [17]. The precalculated set of PMDs coincided with the training set used to train the neural network. The MSE was used to compare different images. However, this approach has not been comprehensively compared with the CNN. Only the transferability properties of the direct comparison were discussed in Ref. [17]. The results from the direct comparison approach in Figs. 6(a)–6(c) show that the performance is comparable with the neural network. Indeed, for 2250 images used for comparison and 750 test

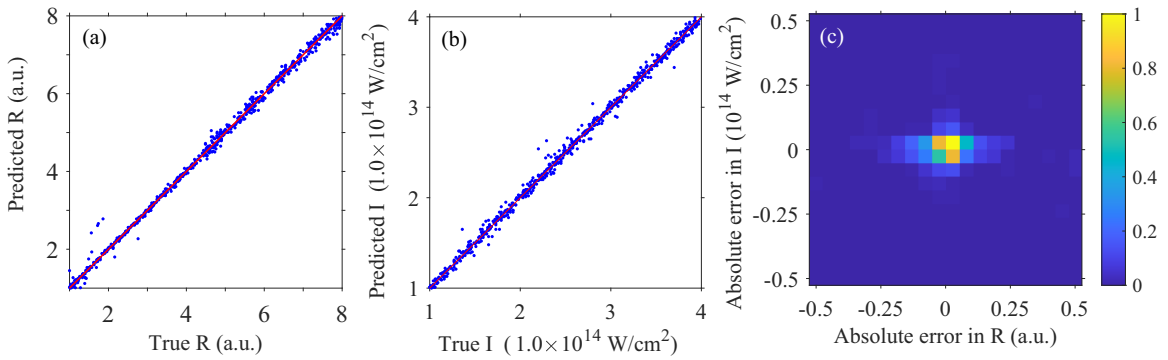

FIG. 6. (a) Plot of predicted vs true internuclear distances in the direct comparison approach. (b) Predicted vs true laser intensities. (c) The joint 2D histogram of absolute errors in the internuclear distance and laser intensity. The histogram is normalized to unity at the maximum value.

images, the MAE for the internuclear distance R and the intensity I is equal to 0.06 a.u. and 0.03×10^{14} W/cm^2, respectively. On the other hand, the CNN trained on the same set of 2250 images to retrieve both the internuclear distance and the laser intensity provides the MAEs for R and I equal to 0.07 a.u. and 0.05×10^{14} W/cm^2, respectively [17].

It may appear natural to employ interpolation for producing more PMDs based on the precalculated ones in order to further improve the accuracy of the direct comparison. However, with an increasing number of interpolated distributions, a limit in the accuracy with which R and I are retrieved will be reached. In our case this limit is reached at about 5250 interpolated PMDs (i.e., 7500 PMDs used for comparison, including 2250 distributions of the precalculated set and 5250 interpolated distributions). The corresponding MAEs for the internuclear distance and laser intensity are 0.055 a.u. and 0.025×10^{14} W/cm^2, respectively.

The direct comparison approach does not necessarily have to be based on the MSE. Any other suitable measure of similarity between two different images can be applied. Here we show the direct comparison using the SIFT algorithm. Although the SIFT is an involved method, for the sake of completeness, here we briefly discuss its main stages (see Ref. [46] for details). At the first stage of the SIFT the so-called scale space of a given image is built. The scale space is produced by the convolution of the given image with the Gaussian kernel $(1/2\pi\sigma^2)\exp[-(x^2+y^2)/2\sigma^2]$ at different scale parameters σ. The original image is then resized to its half size and the procedure is repeated. As a result, several sets of blurred images (octaves) are produced, and the images of each octave are the same size. Four octaves are usually calculated in the SIFT algorithm. The blurred images are required to produce another set of images called difference of Gaussians. Within each octave the difference of Gaussians is calculated by subtracting the image obtained for the larger value of the parameter σ from the neighboring image obtained for the smaller σ. The resulting sets of images corresponding to different octaves are used to find the key points of the initial image.

At the second stage of the SIFT each pixel of the difference of Gaussians is compared with all eight neighboring pixels, as well as with nine pixels of the image with larger σ and nine pixels of the image with smaller σ. The obtained local extrema are the potential key points of the initial image. The generated points that lie along an edge of an initial image or do not have enough contrast are eliminated. The usage of the rest of the points ensures the scale invariance of the SIFT algorithm.

The rotation invariance of the SIFT is provided by the third stage of the algorithm. At this state, a vicinity of each key point is considered and the magnitude and orientation of the gradient in each point of this vicinity is calculated. The obtained values of the magnitude and orientation are used to create a histogram with bins corresponding to different orientation angles. The values that are summed in these bins are the respective gradient magnitudes. The highest peak of the histogram is then chosen, and any other peak exceeding 80% of the highest one is also used to determine the orientation. As a result, a set of key points with the same scale and location but different orientations is found.

Finally, a feature vector (descriptor) of each key point is calculated using the surrounding pixels. This descriptor is based on the gradient orientation and contains 128 values. The rotation dependence of the descriptor is excluded by calculating the difference of each gradient orientation and the orientation of the gradient at the key point. The illumination dependence can be eliminated by using a threshold value and renormalization of the feature vector. When two images are compared using SIFT descriptors, the pairwise Euclidean distances between the feature vectors of the first and the second images is to be calculated. Two feature vectors match if the distance between them is less than a chosen threshold value.

Therefore, we count numbers of matching feature vectors when comparing a given PMD with the distributions of the precalculated set. The distribution with the maximal number of matches is used to determine the quantities of interest, i.e., the internuclear distance and the laser intensity. For the sake of brevity, in this example, as in the other following examples, we discuss the reconstruction of the internuclear distance only. The MAE for the internuclear distance R of the direct comparison based on the SIFT algorithm is 0.054 a.u., see Fig. 7(a).

However, direct comparison is not the only possible alternative to a neural network. Support vector machines and decision trees can also be used for image classification tasks. For two linearly separable sets of points, the support vector machine (SVM) algorithm finds the hyperplane that

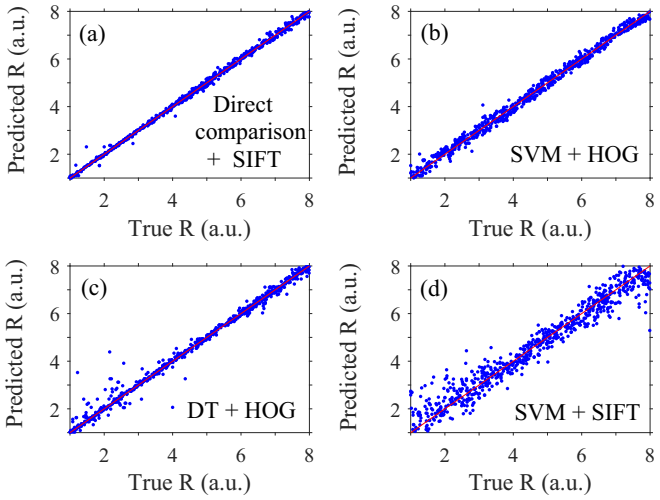

FIG. 7. Plots of predicted vs true internuclear distances illustrating the performance of the alternative approaches to the retrieval of the internuclear distance. (a), (b), (c), and (d) correspond to the direct comparison method employing the SIFT algorithm, the SVM trained using HOGs from different regions of an image, the DT trained on HOGs, and the SVM trained on feature vectors that are extracted with the SIFT algorithm, respectively.

corresponds to the maximum distance from it to the nearest data point on each side. Obviously, the hyperplane obtained with the SVM algorithm ensures the largest separation (margin) between the two sets. For an N-dimensional space, two sets of points are called linearly separable if there exists at least one $(N-1)$-dimensional hyperplane that separates them. For linearly nonseparable classes, the SVM defines some new space and a mapping that transforms the original space to the new space so that an optimal separating hyperplane exists in this new space. Therefore, in the new space the algorithm seeks for a hyperplane separating the transformed points. The mapping used in the SVM algorithm is implicitly defined by the so-called kernel function. The kernel function replaces the scalar product in the minimization problem solved in the SVM (see, e.g., Ref. [37] for details). As a result, the mapping may be nonlinear, and the new space may have more dimensions than the original one.

A decision tree (DT) is a model based on a flowchart-like structure that is used to take decisions. Each internal node of this structure corresponds to some test, and each branch corresponds to the outcome of this test. Finally, each leaf of the tree corresponds to a decision (in our case, a class label). Therefore, the classification rules implemented by a DT are represented by the paths from the root of the tree to its leaves.

However, to the best of our knowledge, the application of the SVMs and DTs to *image regression* problems is not so well documented. Despite this, we use both these methods to retrieve the internuclear distance from PMDs. As a set of features (attributes) that is needed to train the SVM or DT, we use the histogram of oriented gradients (HOG). It is clear that gradients are efficient for finding the edges and corners of an image. The HOG is calculated in the same way as the histogram in the SIFT algorithm, i.e., the magnitudes of the gradient vectors, whose orientation angles correspond to the same bin, are summed up. The size of the cell we used to extract the HOG is 16×32 pixels. The MAEs for the internuclear distance R obtained by using the SVM and DT are equal to 0.13 and 0.10 a.u., respectively [see Figs. 7(b) and 7(c)].

We also combine the SVM with the SIFT algorithm, i.e., we train the SVM using the feature vectors obtained with the SIFT. However, this combination leads to worse results as compared to all other approached discussed here: the corresponding MAE is equal to 0.32 a.u., see Fig. 7(b). We attribute this to the fact that for a significant number of PMDs the SIFT algorithm is not able to extract the sufficient number of key points needed to train the SVM. Indeed, for some of the distributions the SIFT algorithm detects only about 15–20 key points. For these reasons, we do not consider the combination of the SIFT with the SVM as an appropriate tool for retrieval of R.

Finally, it is interesting to perform transferability tests of the approaches discussed here. To this end, we use the same four data sets as in Sec. IV. We apply the direct comparison method employing the SIFT descriptors, as well as the SVM and DTs (both trained using the HOG), to these data sets. It should be stressed that in all these cases the precalculated set (or the training set for the SVM and the DT) is the set of the PMDs obtained for fixed laser intensities (i.e., not focal averaged), $1.0 \leqslant R \leqslant 8.0$ a.u., $\alpha = 0^\circ$, and $\varphi = 0$. The results of these transferability tests are presented in Table 1. For completeness, in Table 1 we also show the results of the transferability tests for the direct comparison based on the MSE: all the corresponding numbers except the MAE for R in the case of the fluctuating CEP were presented in Ref. [17]. We also complement Table 1 with the results obtained by application of the transfer learning technique to the CNN (see Sec. IV). It is seen that all the alternative methods discussed here possess very little transferability. What is even more important, their potential to become more transferable yet still efficient is rather limited: the precalculated (training) data sets are to be considerably increased. Obviously, this will require a lot of calculations. This is a strong argument for applying neural networks to the problem at hand.

VI. VISUALIZATION OF THE CONVOLUTIONAL NEURAL NETWORK

Visualization methods of artificial intelligence are designed to explain and interpret machine learning models. The corresponding new research direction that emerged in the last two decades is often referred to as explainable artificial intelligence. Although the visualization methods are being intensively developed [43–45], there are many open questions in this field of research. Sometimes explanations offered by the visualization methods do not allow to understand what the machine learning model is actually doing. For these reasons, there is even a point of view that it is necessary to stop explaining "black box" machine learning models and use interpretable models instead [57].

Nevertheless, we apply visualization methods to our CNN. Since the vast majority of the visualization methods deal with the CNN designed for classification of images [43,45], we train another neural network that classifies the PMDs into the

TABLE I. The MAE for the internuclear distance R (in a.u.) obtained with the original CNN, the CNN after the application of the transfer learning technique, and by using the alternative methods for four different test sets: focal-volume averaged momentum distributions, distributions for larger internuclear distances, the distributions obtained for nonzero orientation angles, the PMDs calculated for fluctuating CEP, and the PMDs obtained for the HeH^+ molecule. The data in the first line, except the MAE for nonzero CEPs, were already published in Ref. [17].

Method	Averaged	$R \in [8.0, 12.0]$ a.u.	$\alpha \in [0°, 180°]$	$\varphi \in [0, 2\pi]$	HeH^+
CNN	0.83	3.05	0.90	0.89	1.69
CNN + transfer learning	0.15	0.06	0.17	0.23	0.08
Direct comparison + MSE	1.44	5.10	1.37	1.38	1.57
Direct comparison + SIFT	1.19	3.57	1.14	0.71	1.76
SVM + HOG	1.03	2.15	0.99	1.18	1.54
DT + HOG	2.09	2.07	1.54	1.52	2.31

following seven categories: $1.0 \leqslant R < 2.0$ a.u., $2.0 \leqslant R < 3.0$ a.u., ..., $7.0 \leqslant R \leqslant 8.0$ a.u. For simplicity, we use only a quarter of each PMD for the training, namely, the first quadrant of the (k_x, k_y) plane. As any classifying neural network (see, e.g., Ref. [47]), our CNN calculates the probabilities (scores) for a given image to belong to the classes specified above. The classification result is determined by the highest score. The same training and validation data sets (see Sec. II) are used to train this neural network. We aim to understand what features of an image allow the CNN to assign it to a particular class.

To achieve this goal, we apply the occlusion sensitivity method, see Ref. [58]. It is a simple technique that allows us to understand which parts of an image are the most important ones for the CNN to take a classification decision. In this method different parts of the image are occluded with a mask (e.g., gray square) that moves across the image. Simultaneously, the change of the probability score for a specific class is calculated as a function of the occluding mask position. As a result, the occlusion sensitivity map of the image is obtained. This map shows the impact of different parts of the image on the corresponding class score: the parts with high occlusion sensitivity have a positive contribution to the specified class.

The distributions belonging to the same class, e.g., $4.0 \leqslant R < 5.0$ a.u., vary strongly with the peak laser intensity. For this reason, we consider the following intensity intervals: $1.0 \times 10^{14} \leqslant I < 2.0 \times 10^{14}$ W/cm^2, $2.0 \times 10^{14} \leqslant I < 3.0 \times 10^{14}$ W/cm^2, and $3.0 \times 10^{14} \leqslant I \leqslant 4.0 \times 10^{14}$ W/cm^2 for each of the seven classes. Therefore, we have $7 \times 3 = 21$ different alternatives. For each alternative, we randomly choose three PMDs from the validation set and apply the occlusion sensitivity method to these 63 images. We then extract those parts of the chosen PMDs that correspond to values of the occlusion sensitivity larger than 0.2. In doing so, we create a "dictionary" of features corresponding to all seven intervals of the internuclear distance R. Examples of chosen PMDs and the corresponding extracted images are shown in Figs. 8(a)–8(f). The number of 63 images can be further reduced: Many of the resulting images are similar to each other.

As a test, we use this "dictionary" to classify the rest of the PMDs of the validation set. Specifically, we directly compare all the images of our "dictionary" with a given PMD: The image from the "dictionary" that shows the maximal similarity with the corresponding area (areas) of the momentum distribution of interest determines the class of internuclear distances. Here we use MSE for image comparison. The resulting simple classifier, which is based on the application of the occlusion sensitivity method, shows an accuracy of 67%–72%. Here the accuracy is defined as the ratio of correctly classified momentum distributions to the whole number of PMDs used for validation. This accuracy varies depending on the set of distributions that are chosen to extract the relevant features.

The question may arise: What happens if instead of the parts of the PMDs found by the occlusion sensitivity method, we use random parts of the momentum distributions with the same total area as images of the "dictionary"? From Figs. 8(b), 8(d), and 8(f) one might speculate that the applicability of our simple classifier is only a consequence of the large sizes of the extracted parts. However, numerical experiments show that randomly chosen parts with the same total area as the joint area of the occlusion-sensitivity features result in a mean classification accuracy of about only 35%. The stated number is the mean accuracy over a number of

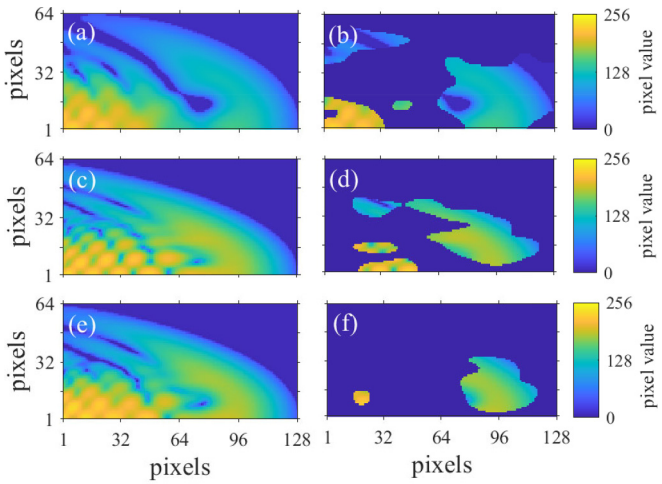

FIG. 8. Electron momentum distributions in the first quadrant of the (k_x, k_y) plane [(a, c, and e)] and their parts extracted by using the occlusion sensitivity method [(b, d, and f)] (see text). (a, b), (c, d), and (e, f) correspond to the laser intensities 2.1×10^{14} W/cm^2, 2.8×10^{14} W/cm^2, and 2.5×10^{14} W/cm^2, respectively. The internuclear distances are (a, b) 4.1 a.u., (c, d) 4.9 a.u., and (e, f) 4.3 a.u. The laser pulse duration and the wavelength are as in Fig. 4, the orientation angle is $\alpha = 0°$, and the CEP is $\varphi = 0$.

tests, since the accuracy obtained in a single test varies in the range between 20% and 46% depending on the choice of random areas.

We note that similar classifiers based on alternative visualization methods, namely, the gradient-weighted class-activation mapping (grad-CAM) [59] and locally interpretable model-agnostic explanations (LIME) [60], give correct predictions for 54% and 59% of the distributions, respectively. It should be stressed that our primitive classifier is not meant to replace the CNN. Nevertheless, its relatively good performance justifies the use of the occlusion sensitivity method. Since the set of features extracted by occlusion sensitivity can be applied to classify images, these features are indeed characteristic for different ranges of internuclear distance. Therefore, these features can be viewed as the ones used by the neural network when making a classification decision.

VII. CONCLUSIONS AND OUTLOOK

In conclusion, we have investigated a number of problems that arise concerning the application of deep learning for prediction of molecular properties and laser parameters from electron momentum distributions. We have studied the effect of the CEP on the retrieval of the internuclear distance. We have shown that it is possible to retrieve the internuclear distance based on the 1D electron momentum distributions. By using the transfer learning technique, we have made our CNN applicable to the PMDs it was not explicitly trained for: focal-volume averaged PMDs and distributions obtained either for internuclear distances, or nonzero angles, or CEPs outside the training range. Furthermore, by applying the transfer learning technique we have made the CNN trained for the case of H_2^+ molecule applicable to momentum distributions corresponding to ionization of another molecule: HeH^+. It is shown that in all these five cases, transfer learning avoids the calculation of large training data sets. These large training sets are needed if, instead of applying the transfer learning technique, we train new CNNs. We have shown that transfer learning for focal-volume averaged distributions, for distributions obtained for large internuclear distances, and for distributions for ionization of the HeH^+ molecule leads to smaller MAEs as compared with the cases of nonzero orientation angles or CEPs. We attribute this to the fact that focal-volume averaged momentum distributions, the distributions calculated for large internuclear distances, and the distributions for HeH^+ are more similar to the PMDs used for training of the initial CNN. Nevertheless, even in the cases of orientation angles and CEPs, the modified CNNs provide MAEs for the internuclear distance of 0.2 a.u. even though no more than 750 images are used for transfer learning.

We have compared the usage of the neural network with alternative approaches: direct comparison of the distributions, SVM, and DT. Moreover, when implementing these alternative methods, we have applied not only the MSE (for direct comparison) or HOG (for SVM and DT), but also the more sophisticated SIFT algorithm used in computer vision. Although some of these alternative approaches, e.g., direct comparison in combination with SIFT, can provide better results than those obtained with the CNN, all these methods have very limited transferability. The only possible way to make the alternative methods more transferable is to significantly enlarge the corresponding precalculated data (for direct comparison) or training data (for SVM and DT). Such an increase involves heavy computational costs.

As a visualization method of machine learning, we have applied the occlusion sensitivity technique to a network trained for the solution of the classification problem: attributing a given momentum distribution to a certain class of internuclear distances. In this way we have extracted the features of the PMDs that allow the CNN to assign a given momentum distribution to one or another class. The appropriateness of the extracted features has been checked by comparing directly a given PMD with the collection of the obtained features, resulting in an accuracy of about 70%. The application of visualization methods to the CNNs trained for the regression problem rather than the classification problem will be the subject of further studies. Also, it will be interesting to apply deep learning for the retrieval of the internuclear distance in the case of moving nuclei. Our present results give us reason to believe that deep learning will lead to significant progress in time-resolved molecular imaging including nuclear motion.

ACKNOWLEDGMENTS

We are grateful to S. Brennecke, F. Oppermann, and S. Yu for stimulating discussions. This work was supported by the Deutsche Forschungsgemeinschaft (Grant No. SH 1145/1-2).

[1] T. Mitchell, *Machine Learning* (McGraw Hill, New York, 1997).

[2] W. Becker, F. Grasbon, R. Kopold, D. B. Milošević, G. G. Paulus, and H. Walther, Above-threshold ionization: From classical features to quantum effects, Adv. At. Mol. Opt. Phys. **48**, 35 (2002).

[3] D. B. Milošević and F. Ehlotzky, Scattering and reaction processes in powerful laser fields, Adv. At. Mol. Opt. Phys. **49**, 373 (2003).

[4] A. Becker and F. H. M. Faisal, Intense field many-body S-matrix theory, J. Phys. B: At. Mol. Opt. Phys. **38**, R1 (2005).

[5] C. Figueira de Morisson Faria and X. Liu, Electron-electron correlation in strong laser fields, J. Mod. Opt. **58**, 1076 (2011).

[6] M. Kitzler and S. Gräfe (Editors), *Ultrafast Dynamics Driven by Intense Light Pulses. From Atoms to Solids, From Lasers to Intense X-rays* (Springer, Cham, 2016).

[7] C. D. Lin, A.-T. Le, C. Jin, and H. Wei, *Attosecond and Strong-Field Physics. Principles and Applications* (Cambridge University Press, Cambridge, 2018).

[8] A. M. M. Gherman, K. Kovách, M. V. Cristea, and V. Toşa, Artificial neural network trained to predict high-harmonic flux, Appl. Sci. **8**, 2106 (2018).

[9] K. Mills, M. Spanner, and I. Tamblyn, Deep learning and the Schrödinger equation, Phys. Rev. A **96**, 042113 (2017).

[10] T. Zahavy, A. Dikopoltsev, D. Moss, G. I. Haham, O. Cohen, S. Mannor, and M. Segev, Deep learning reconstruction of ultrashort pulses, Optica **5**, 666 (2018).

[11] S. Kleinert, A. Tajalli, T. Nagy, and U. Morgner, Rapid phase retrieval of ultrashort pulses from dispersion scan traces using deep neural networks, Opt. Lett. **44**, 979 (2019).

[12] T.-M. Yan, S. V. Popruzhenko, M. J. J. Vrakking, and D. Bauer, Low-Energy Structures in Strong Field Ionization Revealed by Quantum Orbits, Phys. Rev. Lett. **105**, 253002 (2010).

[13] T.-M. Yan and D. Bauer, Sub-barrier Coulomb effects on the interference pattern in tunneling-ionization photoelectron spectra, Phys. Rev. A **86**, 053403 (2012).

[14] X. Liu, G. Zhang, J. Li, G. Shi, M. Zhou, B. Huang, Y. Tang, X. Song, and W. Yang, Deep Learning for Feynman's Path Integral in Strong-Field Time-Dependent Dynamics, Phys. Rev. Lett. **124**, 113202 (2020).

[15] M. Lytova, M. Spanner, and I. Tamblyn, Deep learning and high harmonic generation, Can. J. Phys. **101**, 132 (2023).

[16] X. Liu, K. Amini, A. Sanchez, B. Belsa, T. Steinle, and J. Biegert, Machine learning for laser-induced electron diffraction imaging of molecular structures, Commun. Chem. **4**, 154 (2021).

[17] N. I. Shvetsov-Shilovski and M. Lein, Deep learning for retrieval of the internuclear distance in a molecule from interference patterns in photoelectron momentum distributions, Phys. Rev. A **105**, L021102 (2022).

[18] J. Xu, C. I. Blaga, P. Agostini, and L. F. DiMauro, Time-resolved molecular imaging, J. Phys. B: At. Mol. Opt. Phys. **49**, 112001 (2016).

[19] L. J. Frasinski, K. Codling, P. Hatherly, J. Barr, I. N. Ross, and W. T. Toner, Femtosecond Dynamics of Multielectron Dissociative Ionization by use of a Picosecond Laser, Phys. Rev. Lett. **58**, 2424 (1987).

[20] C. Cornaggia, J. Lavancier, D. Normand, J. Morellec, P. Agostini, J. P. Chambaret, and A. Antonetti, Multielectron dissociative ionization of diatomic molecules in an intense femtosecond laser field, Phys. Rev. A **44**, 4499 (1991).

[21] J. H. Posthumus, L. J. Frasinski, A. J. Giles, and K. Codling, Dissociative ionization of molecules in intense laser fields: A method of predicting ion kinetic energies and appearance intensities, J. Phys. B: At. Mol. Opt. Phys. **28**, L349 (1995).

[22] C. Cornaggia, M. Schmidt, and D. Normand, Laser-induced nuclear motions in the Coulomb explosion of $C_2H_2^+$ ions, Phys. Rev. A **51**, 1431 (1995).

[23] R. Kanya, Y. Morimoto, and K. Yamanouchi, Observation of Laser-Assisted Electron-Atom Scattering in Femtosecond Intense Laser Fields, Phys. Rev. Lett. **105**, 123202 (2010).

[24] Y. Morimoto, R. Kanya, and K. Yamanouchi, Laser-assisted electron diffraction for femtosecond molecular imaging, J. Chem. Phys. **140**, 064201 (2014).

[25] J. Itatani, J. Levesque, D. Zeidler, H. Niikura, H. Pépin, J. C. Kieffer, P. B. Corkum, and D. M. Villeneuve, Tomographic imaging of molecular orbitals, Nature (London) **432**, 867 (2004).

[26] S. Haessler, J. Caillat, W. Boutu, C. Giovanetti-Teixeira, T. Ruchon, T. Auguste, Z. Diveki, P. Breger, A. Maquet, B. Carrè, R. Taïeb, and P. Salières, Attosecond imaging of molecular electronic wavepackets, Nat. Phys. **6**, 200 (2010).

[27] M. Lein, J. P. Marangos, and P. L. Knight, Electron diffraction in above-threshold ionization of molecules, Phys. Rev. A **66**, 051404(R) (2002).

[28] M. Meckel, D. Comtois, D. Zeidler, A. Staudte, D. Pavičić, H. C. Bandulet, H. Pépin, J. C. Kieffer, R. Dörner, D. M. Villeneuve, and P. B. Corkum, Laser-Induced electron tunneling and diffraction, Science **320**, 1478 (2008).

[29] C. I. Blaga, J. Xu, A. D. DiChiara, E. Sistrunk, K. Zhang, P. Agostini, T. A. Miller, L. F. DiMauro, and C. D. Lin, Imaging ultrafast molecular dynamics with laser-induced electron diffraction, Nature (London) **483**, 194 (2012).

[30] M. G. Pullen, B. Wolter, A. T. Le, M. Baudisch, M. Hemmer, A. Senftleben, C. D. Schröter, J. Ullrich, R. Moshammer, C. D. Lin, and J. Biegert, Imaging an aligned polyatomic molecule with laser-induced electron diffraction, Nat. Commun. **6**, 7262 (2015).

[31] Y. Huismans, A. Rouzée, A. Gijsbertsen, J. H. Jungmann, A. S. Smolkowska, P. S. W. M. Logman, F. Lépine, C. Cauchy, S. Zamith, T. Marchenko *et al.*, Time-resolved holography with photoelectrons, Science **331**, 61 (2011).

[32] M. Haertelt, X.-B. Bian, M. Spanner, A. Staudte, and P. B. Corkum, Probing Molecular Dynamics by Laser-Induced Backscattering Holography, Phys. Rev. Lett. **116**, 133001 (2016).

[33] S. G. Walt, N. Ram, M. Atala, N. I. Shvetsov-Shilovski, A. von Conta, D. Baykusheva, M. Lein, and H. J. Wörner, Dynamics of valence-shell electrons and nuclei probed by strong-field holography and rescattering, Nat. Commun. **8**, 15651 (2017).

[34] I. Goodfellow, Y. Benqio, and A. Courville, *Deep Learning* (MIT, Cambridge, 2016).

[35] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).

[36] V. Vapnik, *Statistical Learning Theory* (Wiley, Chichester, 1998).

[37] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)* (MIT, Cambridge, 2002).

[38] I. Steinwart and A. Christmann, *Support Vector Machines* (Springer, New York, 2008).

[39] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, 1984).

[40] J. R. Quinlan, Induction of decision trees, Mach. Learn. **1**, 81 (1986).

[41] L. Breiman, Bagging predictors, Mach. Learn. **24**, 123 (1996).

[42] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. **1**, 886 (2005).

[43] A. Shahroudnejad, A survey on understanding, visualizations, and explanation of deep neural networks, arXiv:2102.01792.

[44] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy **23**, 18 (2021).

[45] M. P. Ayyar, J. Benois-Pineau, and A. Zemmari, Review of white box methods for explanations of convolutional neural networks in image classification tasks, J. Electron. Imaging **30**, 050901 (2021).

[46] D. G. Lowe, Object recognition from local scale-invariant features, Proc. Int. Conf. Comput. Vision **2**, 1150 (1999).

[47] P. Kim, *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence* (Apress, Seoul, 2017).

[48] A. W. Trask, *Grokking Deep Learning* (Manning, Shelter Island, 2019).

[49] MATLAB, version 9.10 (R2021a), MathWorks Inc., Natick, MA (2021).

[50] M. D. Feit, J. A. Fleck, Jr., and A. Steiger, Solution of the Schrödinger equation by a spectral method, J. Comput. Phys. **47**, 412 (1982).

[51] X. M. Tong, K. Hino, and N. Toshima, Phase-dependent atomic ionization in few-cycle intense laser pulse, Phys. Rev. A **74**, 031405(R) (2006).

[52] T. Morishita, Z. Chen, S. Watanabe, and C. D. Lin, Two-dimensional electron momentum spectra of argon ionized by short intense lasers: Comparison of theory with experiment, Phys. Rev. A **75**, 023407 (2007).

[53] A. E. Siegman, *Lasers* (University Science Books, Sausalito, 1986).

[54] S. Augst, D. D. Meyerhofer, D. Strickland, and S. L. Chin, Laser ionization of noble gases by Coulomb-barrier suppression, J. Opt. Soc. Am. B **8**, 858 (1991).

[55] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes. The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, 2007).

[56] N. Eicke, S. Brennecke, and M. Lein, Attosecond-Scale Streaking Methods for Strong-Field Ionization by Tailored Fields, Phys. Rev. Lett. **124**, 043202 (2020).

[57] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. **1**, 206 (2019).

[58] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in *Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, 8689, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Springer, Cham, 2014).

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. **128**, 336 (2020).

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135 (San Francisco, 2016).