Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Deep learning framework for analyzing birefringence imaging by incorporating optical polarization overlap in stress-induced ferroelectric SrTiO

Hirotaka Manaka, Shoutarou Katayama, Soichiro Honda & Yoko Miura

To cite this article: Hirotaka Manaka, Shoutarou Katayama, Soichiro Honda & Yoko Miura (07 Oct 2025): Deep learning framework for analyzing birefringence imaging by incorporating optical polarization overlap in stress-induced ferroelectric SrTiO, Science and Technology of Advanced Materials: Methods, DOI: 10.1080/27660400.2025.2568376

To link to this article: https://doi.org/10.1080/27660400.2025.2568376

Publisher: Taylor & Francis & The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group
Journal: *Science and Technology of Advanced Materials: Methods*
DOI: 10.1080/27660400.2025.2568376

Deep learning framework for analyzing birefringence imaging by incorporating optical polarization overlap in stress-induced ferroelectric $SrTiO_3$

Hirotaka Manaka[a], Shoutarou Katayama[a], Soichiro Honda[a], and Yoko Miura[b]
[a]Graduate School of Science and Engineering, Kagoshima University, Korimoto, Kagoshima 890-0065, Japan; [b]National Institute of Technology, Suzuka College, Shiroko-cho, Suzuka, Mie 510-0294, Japan

ABSTRACT

Optical microscopy is vital in many scientific fields, and various super-resolution techniques have been developed to overcome the resolution limit that restricts the separation of spatially mixed light. However, conventional methods inherently cannot resolve overlapping optical polarization (OP) components, limiting the “polarization resolution” in polarized light microscopy. Instead of quantitatively evaluating “polarization resolution,” this study aims to reliably separate intrinsic OP states based on consistent clustering results that are robust to variations in the spatial receptive field (SRF) size. We integrate statistical analysis, machine learning, and deep learning to evaluate overlapping OP states in temperature-dependent birefringence imaging of the stress-induced ferroelectric $SrTiO_3$ under an external force of 231 MPa. A long short-term memory (LSTM) network is used to extract temperature-dependent features from sequential image data, which effectively captures subtle changes in structural and ferroelectric phase transitions. A 3D convolutional autoencoder (3DCAE) learns spatial relationships between adjacent pixels from these temperature-dependent features, addressing OP overlap at different spatial scales based on different SRF sizes. Although the 3DCAE output considerably depends on the SRF size, clustering results obtained via temperature series forest (Tsf) analysis are highly consistent. This robustness indicates that the extracted OP states reflect physically meaningful spatial distributions rather than convolution artifacts. The proposed sequential analytical framework successfully reconstructs intrinsic OP distributions while balancing local and global structural features, providing a robust foundation for OP-sensitive imaging in materials science.

KEYWRODS

LSTM; temperature series principal component analysis; 3D convolutional autoencoder; temperature series forest; birefringence image; stress-induced ferroelectricity; $SrTiO_3$

E-mail: manaka@eee.kagoshima-u.ac.jp

1. Introduction

Optical microscopy is an essential tool that is widely used in various fields such as biology, medicine, and materials science for microstructural observations. Figure 1(a) shows that the optical resolution of a microscope is the minimum distance at which two distinguishable adjacent points in an image—a limit imposed by the diffraction of light[1, 2]. In conventional optical microscopy, the optical resolution is typically limited to about half the incident wavelength, i.e., about 200 nm for visible light. To overcome this diffraction limit, several super-resolution microscopy (SRM) techniques have been developed; these techniques can be categorized into three types. (1) Hardware-driven techniques that improve resolution by physically modifying optical systems, e.g., near-field scanning optical microscopy (NSOM) and stimulated emission depletion (STED) microscopy[3, 4]; (2) techniques that integrate both hardware and software approaches, combining optical engineering with computational techniques, e.g., photoactivated localization microscopy (PALM), stochastic optical reconstruction microscopy (STORM), and structured illumination microscopy (SIM[5–7]; (3) software-driven super-resolution techniques that rely primarily on computational algorithms, e.g., single-molecule microscopy (SMM)[8, 9]. More recently, deep learning-based methods such as convolutional neural networks (CNNs) and 3D convolutional autoencoders (3DCAEs) have been explored to reconstruct high-resolution information from low-resolution images[10–14].

These SRM techniques offer considerably improved spatial resolution; however, they do not directly address a fundamental limitation of polarized light microscopy (PLM), which is separating overlapping optical polarization (OP) components. Super-resolution methods reconstruct missing spatial details, but they do not explicitly separate mixed OP components. As shown in Figure 1(b), OP resolution requires distinguishing overlapping OP states rather than merely refining spatial details at the pixel level[15]. This fundamental difference requires an approach that extends beyond traditional SRM methods. To address this from a software-driven perspective, the concept of "polarization resolution" (Figure 1(b)) is introduced herein; this concept refers to the ability to distinguish spatially overlapping OP components within the same spatial region. Unlike "optical resolution," which determines the minimum distinguishable distance between adjacent points, "polarization resolution" refers to the ability to separate coexisting OP states at a given location. OP properties critically affect absorption, scattering, and birefringence and provide valuable information about anisotropy, crystal structure, and internal stress/strain that cannot be captured via conventional optical microscopy[16]. However, Figure 1(a) shows that overlapping OP components can obscure the OP distribution. Contrast in conventional PLM is primarily due to birefringence: when light passes through an optically anisotropic sample, different refractive indices for each OP component cause phase retardation. The transmitted light is split into ordinary and extraordinary rays that travel along the optical principal axes of the sample, regardless of the incident OP. The phase retardation and alignment of these components produce image contrast, but PLM does not adequately account for spatially overlapping OP components.

In radio frequency (RF) systems, electromagnetic polarization (EP) discrimination techniques are well established to separate mixed EP components[17–19]. However, these techniques are primarily designed for signal separation rather than spatial resolution improvement. Spatial resolution in RF systems is primarily determined by antenna directivity and is therefore less relevant to polarization separation. In contrast, PLM requires the precise spatial separation of OP components because their direct mixing affects the accuracy of analysis. Traditional PLM techniques assume that increasing spatial resolution will lead to better OP separation; however, this assumption is flawed because OP components are considerably affected by overlap, particularly at domain boundaries and edges. Unlike RF systems, where EP discrimination is primarily a matter of signal separation, OP mixing in PLM occurs due to intrinsic spatial overlap, which results in the loss of OP-specific information. Thus, conventional RF approaches are not directly applicable to PLM, and a new computational strategy for "polarization resolution" is required.

Solving the spatial overlap of OP components in PLM is difficult via hardware modifications alone. To address this issue, a deep learning—based approach is proposed herein for OP resolution—aware separation. Unlike conventional super-resolution techniques that rely on pixel interpolation, CNNs are expected to extract complex spatial OP relationships between adjacent pixels. 3DCAE also enables temperature (T)-dependent feature learning to track OP dynamics under varying T such as thermal fluctuations and phase transitions[20–25]. Software-based image analysis on conventional PLM images can be broadly categorized into those based on pixel-by-pixel and clustering methods that capture group-level trends. In particular, OP patterns often show continuity or gradual changes across adjacent pixels, reflecting the underlying material properties; therefore, clustering is an effective tool rather than pixel-by-pixel analysis. Typical clustering methods are particularly well-suited for analyzing birefringence images, where ground-truth labels are usually unavailable. Additionally, semi-supervised and self-supervised learning approaches can leverage limited external labels or self-generated pseudo-labels to improve learning accuracy. According to statistical causal inference, clustering algorithms can easily identify treated and untreated groups and reveal hidden regularities[26]. In this study, we focus on the clustering-based approach because the spatial receptive field (SRF) of the 3DCAE learns spatial dependencies over multiple pixels via convolutional operations. By capturing these spatial relationships, the model can extract more hierarchically structured OP features that are critical for distinguishing overlapping components. Thus, clustering is particularly effective because it groups pixels based on shared OP features rather than treating them independently. This reduces the ambiguity caused by overlapping OP components, resulting in highly accurate and robust separation. SRF refers to the region of input data that is processed by a given neuron or filter at any given time. A small SRF size captures fine-grained, localized details, making it sensitive to small-scale structures. In contrast, a large SRF size integrates information over a larger region, thereby allowing the network to extract more global features. By investigating the effect of SRF size variations on the robustness of OP feature extraction and clustering stability, its scale dependence is elucidated herein—a factor largely overlooked in conventional PLM analysis. Using this approach, this study advances OP analysis techniques using deep learning-based frameworks that can perform multiscale spatial feature extraction and dynamic OP tracking along the T axis. We establish the concept of "polarization resolution" as a new analytical perspective, paving the way for more accurate characterization of materials under stress and strain conditions and overcoming the inherent limitations of conventional PLM.

However, no standardized method currently exists to quantitatively assess "polarization resolution" or to effectively separate overlapping OP components. Therefore, rather than quantitatively assessing the "polarization resolution", we focused on achieving a robust and consistent identification of intrinsic OP states. We particularly aimed to ensure that clustering results remain stable and reliable despite variations in the SRF size used during analysis. To this end, we developed an integrated analytical framework that strategically combines statistical analysis, machine learning, and deep learning techniques. Section 2 covers the detailed methodology and a summary of previous results supporting our analysis strategy. To validate the proposed framework, we applied our approach to T-dependent birefringence image datasets of $SrTiO_3$—a well-known quantum paraelectric[26–30]. By analyzing its structural and ferroelectric phase transitions under an external force, we demonstrated the robustness of clustering results under different analytical conditions, including SRF sizes and preprocessing parameters, in identifying OP distributions. Using this approach, OP features across different spatial scales and temperature variations were extracted robustly. These findings provided new insights into material characterization under stress and strain conditions.

2. Dataset for analysis

To analyze the T dependence of birefringence images, the image analysis method that considers "polarization resolution" is developed. The variables for quantifying birefringence are defined. When incident light is irradiated on an anisotropic optical material, the difference in refractive indices in the plane perpendicular to the propagation vector of the light causes birefringence. These two refractive indices, defined as the directions of the optical principal axes of the refractive ellipsoid, are n_1 (slow axis) and n_2 (fast axis), with $n_1 \geq n_2 > 1$. Birefringence (Δn) and retardance (δ) are defined as follows:

$$\Delta n \equiv n_1 - n_2, \tag{1}$$
$$\delta = \Delta n \times t, \tag{2}$$

where t is the optical path length through the sample[16, 31]. The fast-axis direction (ψ) corresponds to the orientation associated with n_2. δ is generally proportional to the stress/strain in the material due to the photoelastic effect. In addition, spontaneous polarization generates strain, which further increases δ. In such cases, ψ reflects the direction of stress/strain and spontaneous polarization[32–34]. Examining the T dependence of Δn allows us to capture the effects arising from stress/strain and the emergence of spontaneous polarization simultaneously. Therefore, the $\Delta n(T)$ curves are an essential physical quantity in our analysis.

Herein, data structure analysis is performed using the dataset of birefringence imaging[27], and the measurement conditions required for analysis are summarized. Birefringence images were obtained using a WPA-100 ellipsometer (Photonic Lattice, Inc.) for accurately characterizing the OP using the Stokes parameters (S_1, S_2, S_3) on the Poincaré sphere[16]. A polarimeter with a 6W LED white light source and circular OP at $S_3 \simeq -1$ was also used. The wavelength (λ) was monochromated using three types of filters ($\lambda = 523$, 543, and 575 nm). To obtain δ and ψ images simultaneously at a 384×288 pixel, the polarizer and waveplate generated using autocloning techniques were placed between the CCD camera and objective lens. The Stokes parameters (S_1, S_2, S_3) at the pixels were obtained using a four-detector method[35, 36].

The value of δ is the length of the arc connecting the OP states of the incident and transmitted light on the Poincaré sphere[31]. One rotation of the Poincaré sphere is equal to 360°

, which is equal to λ, based on which, the rotation angle of the arc has degrees of freedom for clockwise and counterclockwise rotation. Thus if the intrinsic rotation angle of δ is between 180 $^\circ$ and 360°, then it cannot be distinguished from the rotation angle measured from the opposite direction. If δ exceeds 360°, the number of revolutions cannot be determined. The circumference of the Poincaré sphere is different for each λ[37]; therefore, the number of revolutions and the direction of rotation (clockwise or counterclockwise) can be distinguished from three different λ. However, as verifying this method at the pixel level is computationally expensive, a new rotation angle θ was introduced[29]. This computational cost can be reduced by setting the shortest path between the initial and final OP states. Therefore, the period of θ is 360°, but it satisfies the conditions $\theta[0^\circ,180^\circ]$ and $\delta = |l \pm \theta / 360^\circ| \times \lambda$ nm ($l = 0,1,2,\cdots$). On the Poincaré sphere, the direction of the rotation axis in the S_1S_2 plane is defined as the rotation angle ϕ measured counterclockwise from the $+S_1$ axis. The period of ϕ is 180°, and the fast-axis direction is expressed as $\psi = (\phi \pm 90^\circ)/2$. Many features can be extracted using only θ and ϕ, without specifying “l“ and “$\pm$“[29, 30].

The fundamental physical properties of the quantum paraelectric $SrTiO_3$ have been well studied. Under stress-free conditions, cubic-to-tetragonal phase transition occurs at $T_c = 105$ K and the relative permittivity increases rapidly in this T region[38–40]. However, the ferroelectric state is suppressed by large quantum fluctuations and the quantum paraelectric state is realized. When electric fields or external forces are applied, quantum fluctuations are suppressed and ferroelectric transition occurs[41–45]]. We observed the ferroelectric state using birefringence imaging measurements under electric fields and external forces[27,28,46,47]. When a strong electric field of 5.2 kV/cm is applied in [001], the structural phase-transition temperature is constant ($T_c \simeq 105$ K), whereas the ferroelectric phase-transition temperature is $T_F \simeq 51$ K. Below T_F, spontaneous polarization rapidly increases δ and a 90° rotation of ψ. However, the electric field-induced ferroelectric state does not spread over the entire substrate, such that its coexistence with the paraelectric state can be realized. In contrast, when an external force of 231 MPa is applied along [001], T_c increases up to ~ 123 K and ferroelectric phase transition occurs at $T_F = 20$–30 K. Figure 2 shows the spatial distributions of δ and ψ for 575 nm at 14.1 K. These images are 302×140 pixel data extracted from the 384×288 pixel raw data. As shown in Figure 2(a), $\delta(14.1K)$ distribution has an unexpected value in the small rectangular region on the right that spans several pixels because $\delta = (1 - \theta / 360^\circ) \times \lambda$ and $\psi = (\phi + 90^\circ)/2$ are assumed across the substrate from the measurements at three different λ. These results show that a stripe structure is clearly visible, particularly in ψ rather than in δ. This is because when the external force is applied at room temperature, dislocations occur and slip planes are generated that extend in $\langle 111 \rangle$. The analysis of T dependence of δ on a pixel-by-pixel basis shows that the ferroelectric state is realized over the entire substrate contrary to that observed the electric field experiments. However, as the stress is unevenly distributed, the value of T_F is locally high and ψ deviates significantly from 90° in stress-concentration regions[48].

The dataset (3λ, $1S_1$, $1S_2$, $1S_3$, $3,362T$, 42,280 pixels) was obtained in previous experiments by continuously decreasing T from 300.0 K to 14.1 K at a rate of -0.25 K/min[27, 28]. Based on the Poincaré sphere, (S_1, S_2, S_3) can be reduced to (δ, ψ). After further converting from (δ, ψ) to (θ,ϕ), to account for their periodicity, they were rewritten using circular statistics based on four variables $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi)$ [29, 49]. Here, $\overline{R}$ is the mean resultant length expressed as follows:

$$\overline{R}_\theta = \frac{\sqrt{\left(\sum_{i=1}^{m}\cos\theta_i\right)^2 + \left(\sum_{i=1}^{m}\sin\theta_i\right)^2}}{m}, \tag{3}$$

$$\overline{R}_\phi = \frac{\sqrt{\left(\sum_{i=1}^{m}\cos 2\phi_i\right)^2 + \left(\sum_{i=1}^{m}\sin 2\phi_i\right)^2}}{m}, \tag{4}$$

where m is the number of data points. By including $\overline{R}$ averaged over $m = 42,280$ pixels at each T, T-series analysis is stabilized as it reflects the variability within the image. Figure 3 shows previously reported K-shape multivariate clustering results[30], wherein data were grouped into four clusters based on the shapes of 12 T-dependent variables such as $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi) \times 3\lambda$ from 130.9 K to 14.1 K with -0.34 K intervals[50–52]. This result clearly shows the stripe structures, which were classified as follows. In clusters E1 and E2, the δ value is large due to stress concentration, whereas in clusters E3 and E4, the stress is evenly distributed. The distribution of ψ shows that E1 and E3 form one group and E2 and E4 form another. A comparison of Figures 2 and 3 shows that the grouping depends more on the distribution of ψ than on δ because the results were clustered by considering the four variables as independent. As a change in ϕ is twice as sensitive as a change in θ due to its periodicity, the distribution of ψ significantly impacts clustering. The two variables (δ,ψ) denote a point on the surface of the Poincaré sphere, and their transformation into (θ,ϕ) does not eliminate their inherent dependence. As the transformed variables $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi)$ are also constrained by the geometry of the sphere, they must satisfy a special dependency relationship to remain on the Poincaré sphere. Therefore, treating them as independent variables in clustering analysis would result in loss of the OP consistency and misrepresentation of the underlying structure.

Statistical methods employed to address dependent relationships typically use principal component analysis (PCA) that eliminates correlations between variables and the Mahalanobis distance that condenses correlations into a single distance function using the variance-covariance matrix[53]. The latter approach is appropriate to perform clustering while preserving dependent relationships on the Poincaré sphere because it considers the geometric structure of data and normalizes correlations, thereby enabling a more accurate representation of similarity in a curved space. However, as the Mahalanobis distance reduces four variables into a single distance measure, local structural details in the dataset may be lost. To address this issue, we introduce a long short-term memory (LSTM) network that captures variations while preserving the continuity of the T-series dataset (3λ, 1θ, 1ϕ, $3,362T$, 42,280 pixels)[54, 55]. Consequently, the underlying dynamic structure of the reduced data can be recovered, leading to more accurate clustering. To establish OP resolution-aware analysis, 3DCAE is introduced for convolutional processing with different SRF sizes and extracting both local OP variations and global OP patterns. Unlike conventional PCA, which is applied to independent observations, T-series PCA (TsPCA) is specifically designed for T-series data; it eliminates correlations while preserving the sequential structure[56]. As the LSTM output exhibits autocorrelation and potential unit root effects due to the T-series nature of δ and ψ, TsPCA is applied for stabilizing data and reorganizing the T-series structure before feeding the dataset into the 3DCAE. 3DCAE analysis with different SRF sizes reveals the effect of inter-pixel correlations that have not been considered in conventional PLM images. However, high inter-pixel correlations can lead to increased multicollinearity. Unlike linear regression or PCA-based clustering, decision tree-based clustering is robust to multicollinearity because it does not rely on inverse variance-covariance matrices or eigenvalue decompositions, which can be unstable when predictor variables are highly correlated[57–59]. Decision trees recursively partition the feature space based on entropy or variance reduction; thus, they are well suited for capturing nonlinear dependencies between variables.

3. Data analysis

A data-driven analysis was performed herein using the birefringence imaging datasets at 231 MPa[27] on a computer equipped with an Intel(R) Core(TM) i9-13900 CPU (up to 5.60 GHz) and 128 GB of memory. The NVIDIA RTX A4000 GPU was used to efficiently accelerate the computations. A combination of advanced statistical analysis, machine learning, and deep learning techniques were used within the Python (v. 3.10.12) environment. The Mahalanobis distance for dependent variables in the birefringence imaging dataset was calculated using the “scipy” library (v. 1.10.1) independently for each λ. To ensure a robust statistical analysis, the Mahalanobis distance was computed based on the mean vector and the variance-covariance matrix for each λ-specific dataset. The inverse variance-covariance matrix was used to consider scale dependencies in distance calculations and accurately adjust for correlations between variables. “TensorFlow Keras” API (v. 2.14.0) was used to construct and train the LSTM model, which comprised three stacked LSTM layers followed by five parallel dense layers, and their outputs are linearly combined before the final output layer. Both the training and prediction processes were accelerated using CUDA (v. 11.8) on an NVIDIA RTX A4000 GPU to improve computational efficiency. The training dataset was preprocessed using the “MinMaxScaler” function from the “Scikit-learn” library (v. 1.0.2), and optimization was performed using the “Adam” optimizer. A time series PCA approach was employed for TsPCA using the “Scikit-learn” library, and the “StandardScaler” function was applied for feature normalization. The 3DCAE was implemented using the “PyTorch” library (v. 2.5.1+cu118), with both training and prediction accelerated by a GPU using the CUDA support. This model was designed to extract hierarchical OP features while preserving spatial correlations, thus facilitating the reconstruction of OP structures in birefringence imaging data. T-series clustering was performed using the K-shape algorithm implemented in the “tslearn” library (v. 0.6.3). This method was specifically designed for clustering T-series data based on shape similarity. Before clustering, the “StandardScaler” function from the “Scikit-learn” library was used to normalize each feature independently to ensure comparability across different λ. For the T-series classification of the 3DCAE dataset, a time-series forest model implemented via the “RandomForestClassifier” function of the “Scikit-learn” library was employed. This ensemble-based method allows classification of temperature series data without requiring explicit feature extraction. The

"ProcessPoolExecutor" function was used for parallel execution and achieve efficient processing of large T-series datasets with improved computational efficiency.

4. Results and discussion

The birefringence imaging data were obtained by continuously decreasing T while measuring birefringence at three different λ[27]. The values of θ vary at different λ, which affected the measurement accuracy; the accuracy decreased as θ approached 0° or 180°. Collecting data from multiple λ is possible to compensate for each other and improve the reliability of the analysis[37]. This advantage has been effectively exploited in K-means multivariate clustering and K-shape multivariate clustering methods[29, 30]. Due to the spectral characteristics of the white LED light source used in the experiment, the highest intensity was achieved at 575 nm among the three λ that resulted in the best signal-to-noise ratio. Therefore, we focused primarily on the analysis results obtained at 575 nm for clarity. The results obtained at 523 and 543 nm have been discussed in Supplementary Materials.

This study focused on demonstrating the robustness and consistency of clustering results that are independent of variations in the SRF size rather than quantitatively evaluating the improvements in "polarization resolution." To this end, sequential computational approaches such as statistical analysis, machine learning, and deep learning were employed to clarify latent OP states across adjacent pixels that could not be adequately distinguished via conventional pixel-based or single-scale analyses. Figure 4 shows a flowchart of the analysis process, including the dataset dimensionality at each step. At some analysis steps, K-shape multivariate clustering was performed to confirm the consistency and overall trends of computational results; however, this approach is not explicitly shown in the flowchart. This sequential approach handled overlapping OP states robustly and consistently, thereby enabling more stable and reliable clustering than that achieved using conventional PLM. Each analysis step and the evaluation and interpretation of results are described in subsequent sections.

4.1. Mahalanobis distance

The four variables $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi)$ exhibit inherent dependencies due to the geometric structure of the Poincaré sphere. The Mahalanobis distance (Mh) represents dependent relationships between variables. Although Mh is computationally expensive, scale differences between variables are adjusted using a variance-covariance matrix that considers correlations. This allows the OP states to be treated as a unified measure. PCA is commonly employed to deal with dependent relationships because it assumes that the basis functions are linearly dependent. It applies a linear transformation to maximize the variance along the principal component (PC) axes. However, PCA assumes that the transformation axes are globally optimal in terms of variance, which may not fit well with the local geometric structure of the OP states on the Poincaré sphere. Thus, Mh is a more appropriate choice for preserving the relationships between OP components. To ensure the treatment of λ-dependent OP, Mh is computed separately at each λ. As Mh is inherently standardized during calculation, no additional preprocessing is required.

As a result, the dataset was simplified from (3λ, 2θ, 2ϕ, $3{,}362T$, 42,280 pixels) to (3λ, $1Mh$, $3{,}362T$, 42,280 pixels). Considering the computational cost, we integrated the dataset based on Mh into T intervals of 0.5 K. Binned regression was employed to stabilize the T-series data by reducing local fluctuations while preserving important phase transition trends. The T dependence of θ and ϕ associated with ferroelectric and structural phase transitions was relatively gradual[30]. The clustering result shown in Figure 3 was obtained using the dataset with 0.34 K intervals. Therefore, an interval of 0.5 K was sufficient to capture relevant variations. Assuming a linear relationship $(Mh = a_0 + a_1 T)$ within the 0.5 K bin width, the least-squares method was used to estimate the coefficients (a_0, a_1) for each bin. As the birefringence measurements were performed by decreasing T to 14.1 K, the input data were ordered in the decreasing order of T. The respective bin ranges were defined as the half-open interval $[(160 - 0.5 \times n) - 0.25\,\mathrm{K}, (160 - 0.5 \times n) + 0.25\,\mathrm{K})$, where $(n = 0, 1, \cdots, 292)$. The estimated Mh for each bin was computed using the representative T, which is expressed as $(160 - 0.5 \times n)$ K. This binned regression reduced the dataset size and created a T-series dataset with uniform T intervals: (3λ, $1Mh$, $293T$, 42,280 pixels). Although no experimental data were obtained below 14.1 K, padding was not applied to estimate the values at 14.0 ± 0.25 K. Instead, all available data at and below 14.25 K were used for binned regression.

As our analysis focuses on the variations in T between 130.0 and 14.0 K, Figure 5 shows the computed results at two representative pixel positions, b1 and b2 shown in Figure 2(a). These positions were selected from the full dataset of 42,280 pixels, each containing data points for (3λ, $1Mh$, $233T$). The specific choice of b1 and b2 is not essential to the workflow in Figure 4; they are included merely as illustrative examples. Using the bin regression method with four to six T points set in each bin reduced white noise by more than half, resulting in smoother T-series curves. At 575 nm in Figure 5, anomalies reflecting the structural phase transition appeared at 110–120 K, whereas peaks associated with the ferroelectric phase transition were observed at around 20–30 K. The results at 543 nm showed the same trend as those at 575 nm; however, the shape of the $Mh(T)$ curve at 523 nm was clearly different because T range for which $\delta > \lambda$ was different for each λ. Using this difference in the shape of the $Mh(T)$ curves for different λ, K-shape multivariate clustering was performed using the dataset (3λ, $1Mh$, $233T$) $\times 42,280$ pixels. Figure S1 in the Supplementary Materials shows the results of the elbow method, Silhouette score, and gap statistic for determining the optimal number of clusters (k)[60, 61, 62]. The elbow method evaluates the sum of squared errors (SSE) that measured the variance within each cluster. As the number of k increases, the SSE decreases; however, at a certain point, the decrease becomes marginal and forms an “elbow” in the plot. This elbow point represents the optimal trade-off between compact clusters and avoiding excessive fragmentation. The Silhouette score measures the clustering quality by considering both intra-cluster cohesion and inter-cluster separation. The score ranges from -1 to 1, with higher values indicating better defined clusters. The gap statistic compares the clustering dispersion in the actual dataset with that of randomly generated reference dataset, and a significant gap indicates that the clustering structure is meaningful. The optimal number of clusters is identified as the number that maximizes the gap statistic while maintaining statistical significance. However, this approach is based on random reference data; if the analysis dataset has large variability, statistical significance may decrease and the interpretation will be less straightforward. As shown in Figure S1, the elbow method exhibits a decreasing trend in SSE, with an elbow for four to six clusters. The Silhouette score peaks at two clusters but remains relatively high for three to five clusters, indicating that clustering solutions within this range may also be reasonable. The gap statistic exhibits fluctuations, making interpretation less straightforward. As expected, these methods yield inconsistent results, but $k = 2–5$ seem to be the reasonable choices. Figure 6 shows the spatial distribution of the clustering results for $k = 4$ and 5. Figure S2 in the Supplementary Materials shows the results for $k = 2$ and 3. From these results, the stripe structures observed in Figure 3 are not clearly visible regardless of k. This loss of structure is likely due to the reduction from the original variables of (3λ, 2θ, 2ϕ, $233T$) × 42,280 pixels to only (3λ, $1Mh$, $233T$) × 42,280 pixels.

4.2. Long short-term memory (LSTM)

The K-shape multivariate clustering method classifies time-series data based on the similarity of the T-dependent curvature[50, 51, 52]. Herein, the spatial distributions of the OP state were observed continuously and T decreased gradually (-0.25 K/min); thus, the measurement errors were expected to be different for each T because it changed over time even for the same pixel position[27, 28, 30]. Although the small measurement errors were smoothed by the binned regression, the correlation between the bins along T axis could not be considered. The LSTM algorithm was used in the analysis step, which is a representative case of a recurrent neural network that can efficiently learn by focusing on the long-term time variation or overall trend in time-series data[54, 55]. This allows it to learn gradual changes associated with structural and ferroelectric phase transitions. As the 42,280-pixel dataset was analyzed individually for each pixel, the input dataset (3λ, $1Mh$, $293T$) for the LSTM model was considered as a single three-λ T-series data. Normally, a T-series dataset of 42,280 pixels would be divided into training and test sets to evaluate generalization performance. In materials science, acquiring additional datasets for standard cross-validation is often impractical due to experimental constraints such as limited sample availability and measurement conditions[63 − 66]. As experimentalists, we are responsible for analyzing the entire dataset because each measurement contains valuable, often irreproducible, physical information. In our study, we applied the 0.5 K bin regression to suppress white noise and ensure that the remaining signal accurately reflects the system’s intrinsic physical state. Consequently, our analysis strictly confines itself to the measured data domain and does not involve extrapolation beyond it. Under these conditions, concerns about overfitting are not critical because our objective is to extract the underlying latent structure within the measured data, not to generalize to unseen conditions[67 − 69]. As a preprocessing step, the “MinMaxScaler” function was employed to independently normalize the T-dependent Mh for each λ over all T values in the dataset; each feature was scaled to the range $[0,1]$. This normalization ensured that the Mahalanobis distances at 523, 543, and 575 nm retained their individual dynamic range while maintaining consistency over the entire T-series dataset. As the LSTM was trained while decreasing T, $60T$ points (30.0 K) on the higher T side were chosen as the “look_back” width while estimating Mh at a given T. This width was determined by considering the extent to which the precursors of the structural and ferroelectric phase transitions appeared on the high T side. Specifically, 30.0 K was chosen to account for the T range in which T_c and T_F were determined using the Bayesian method[30]. Moreover, since $SrTiO_3$ undergoes three-dimensional phase transitions, the high-T data at 2 T_F and 1.2 T_c has a negligible impact on the physical changes at low Ts and can be considered statistically independent[40, 70, 71]. Figure 5 shows that the 30.0 K range may adequately capture the long-term trend of the $Mh(T)$ curve. However, medium- and short-term fluctuations will be considered during the analysis of the following steps.

A three-layer LSTM network was constructed (Figure 7) for training and prediction. The T-series data were extracted in the format of (3λ, $1Mh$, $60T$) from the high T side for use as the input data, and the Mh value at the 61st T point was predicted. This process was repeated by moving one T point at a time to the low T side and the $Mh(T)$ curves were trained and predicted for the three λ of 42,280 pixels in the T range [14.0 K, 130.0 K]. Autocorrelation and potential unit root effects must be considered, but the LSTM model can train such effects[72, 73]. The three-layer LSTM structure was designed to progressively reduce the number of units in each layer. The first LSTM layer (100 units) extracted broad trends in the T-series variations, capturing dominant phase transition features. The second LSTM layer (70 units) refined these representations by suppressing smaller fluctuations, and the third LSTM layer (50 units) improved local consistency before passing the data to the fully connected layers. This network architecture was determined via an experimental evaluation in which several configurations were tested. The chosen structure exhibited good results in capturing the T-dependent features of the dataset. The fully connected layer comprises five parallel dense layers, each of which is expected to capture different aspects of the extracted features. In the final output layer, these dense layers are integrated, and the dataset (3λ, $1Mh'$, $233T$) is output for each pixel. The five dense layers are used to balance the computational cost and the diversity of the extracted features. By enabling multiple linear combinations to be performed in parallel, features with different T changes (such as gradual or sharp changes) can be easily extracted. However, interpreting the specific phenomena processed by each layer is challenging as this step is performed by the automatic learning algorithm. This approach is consistent with the role of fully connected layers, which integrate information from the third LSTM layer while refining meaningful feature representations.

To achieve fast, stable, and adaptive learning by automatically adjusting the learning rate for each parameter during training, the mean-squared error (MSE) is used as the loss function for training, and "Adam" optimization is used. The learning rate was set to 0.0001, and the computations were performed on the GPU with a batch size of 32. Approximately one hour per epoch was required for training, and the loss function decreased sufficiently after 15 epochs (Figure 8(a)). Using this trained model, the Mahalanobis distance was predicted at 233 T points for the three λ by inputting the entire dataset in the same way as that during model training. As the structure of existing data was analyzed herein, typical overfitting countermeasures such as "dropout," "early stopping," or "validation data partitioning" were not applied. Instead, the model was designed to fully capture T-dependent features in the dataset. Five parallel dense layers were used to recover the inherent diversity of the data structure, particularly subtle variations that might otherwise be lost. No predictions were made below 14.0 K because experimental data were not gathered at this stage; this ensured that model extrapolation beyond the measured range could be avoided. Figures 8(b–c) show the $Mh(T)$ curves at 575 nm derived from the experiments along with the LSTM predictions, i.e., $Mh'(T)$ curves, at b1 and b2. Figure S3 in the Supplementary Materials shows the results obtained at 523 and 543 nm. These $Mh'(T)$ predictions agreed well with the $Mh(T)$ experimental data, demonstrating that the LSTM model could effectively train the T-dependent structure of the dataset. As the LSTM model reconstructed the input data rather than generating entirely new predictions, information loss was minimized. The learning process ensured that the extracted features emphasized T-dependent variations, making the subsequent clustering process more effective. By training on the sequential structure of the dataset, LSTM improved the representation of phase transitionrelated features that are critical for identifying the OP states.

We focused on five T-series data at five parallel dense layers. As these data naturally contained all relevant information, they served as a powerful basis for data structure analysis. Figure 9 shows the T dependence of the five variables at b1 and b2. Anomalies reflecting the structural phase transition appeared at 110–120 K, whereas kinks associated with the ferroelectric phase transition were observed around 20–30 K. The five layers extracted different features and underwent a final linear transformation to predict the $Mh'(T)$ curve. They shared a linear relationship. Furthermore, the coefficients of the linear transformation of five variables were determined during training so that they had the same value for all pixels. As a result, the corresponding five layers should have the same roles for similar T-series data but different roles for T-series data with different trends. Thus, these five layers created a feature map in a five-dimensional space. Although we focused on a specific dense layer, extracting useful information will be difficult because the specific features to be focused on in each layer differ for each pixel with a different trend. As PCA is also a linear transformation, it naturally adapts to the basic five-dimensional spatial structure; thus, it is an appropriate method for analyzing variation patterns without introducing additional nonlinear distortions. Nonlinear activation functions such as “ReLU,” “sigmoid,” or “tanh” were not applied between the fully connected layer and the final output in this model. This ensured that the transformation remained strictly linear, thereby reinforcing the validity of applying PCA to these outputs. In contrast, the Mahalanobis distance was used in the previous step to integrate multiple dependent variables into a single metric while considering the variance-covariance matrix. Thus, Mh was particularly effective in classification problems such as linear discriminant analysis (LDA), where maximizing class separability is critical. However, as Mh transformed the feature space based on the variance-covariance structure, it may obscure the individual contributions of correlated variables; this makes it less suitable for analyzing the internal dependencies among the five T-series outputs. Therefore, these five outputs represented distinct but correlated features and PCA was employed to preserve their variance while projecting them onto an uncorrelated basis of the five PCs. PCA was applied due to its strong multicollinearity among the five T-series outputs and their spatial correlations in the dataset. When clustering is performed directly on a high-dimensional dataset with redundant information, the underlying structure can be obscured and results in unstable or misleading cluster formation. Highly correlated variables can introduce artificial clustering patterns that do not reflect actual physical differences in the dataset. Moreover, strong spatial multicollinearity between adjacent pixels can amplify redundant patterns, making it difficult to distinguish between structurally meaningful variations. PCA mitigates these problems by projecting the data onto the PC axes that maximize variance, thereby ensuring that clustering is driven by fundamental T-dependent variations rather than random fluctuations. Although this multibranch design may seem unorthodox compared with the typical single-output architectures, it helps to mitigate overfitting in the LSTM stage: by extracting multiple distinct feature representations in parallel, we avoid collapsing all the information into a single path. As a result, finer-grained signals are preserved and model overfitting to noise or outliers is prevented.

As PCA is sensitive to scale differences, the “StandardScaler” function was used herein to standardize each of the five T-series variables before running TsPCA. Unlike the “MinMaxScaler” function, which rescales the data to a fixed range, the “StandardScaler” function preserves the variance structure by transforming each variable to have zero mean and unit variance. This ensures that each variable contributes proportionally to the PCs, thereby preventing features with larger absolute values from dominating the analysis. As the fully connected layer of the LSTM inherently adjusts feature scales during training, the original scale is not preserved for analysis. A sliding window method was subsequently used to extract local features from the T-series data[56]. The window width of 19 T points (9.5 K) was applied, and TsPCA was performed on the five variables within each window for transforming them into five PCs. 19 T points were chosen as a compromise between capturing local T-dependent variations and maintaining consistency with the LSTM's broader training range of 60 T points (30.0 K). No padding was applied to the LSTM model, and the analysis was performed using only the available data. If the window size is considerably small, it may fail to capture meaningful correlations in T-dependent variations along the T axis. Autocorrelation and unit root effects can be mitigated by reducing the window size; these effects might otherwise obscure the underlying structural patterns in the dataset. By continuously applying TsPCA while shifting the window one T point at a time to the lower T side, the five T-dependent PCs can be tracked in detail. TsPCA was performed herein using the eigenvalue decomposition of the variance-covariance matrix, and both the explained and cumulative explained variance ratios of each PC were computed.

Figure 10 shows the T dependence of the five PCs for 575 nm at b1 and b2. Unlike Figure 9, the T dependencies in this figure are difficult to understand intuitively and exhibit many sudden variations. Figure 11 shows the T dependence of the cumulative explained variance ratio. As the cumulative variance ratio exceeds 90%, we concluded that $PC1-PC3$ retain the majority of the variance in the dataset. If the dataset was completely random, each PC would contribute 20% of the total variance. As $PC1$ and $PC2$ exceed this threshold and $PC3$ compensates for the decreasing variances of PC1 and PC2 at around 60 K, these three components can sufficient capture the overall trend. Thus, two or three clusters are deemed reasonable for classification. However, the exact physical meaning of each PC remains uncertain because PCA only optimizes the variance representation. Nevertheless, $PC1-PC3$ may capture the dominant T-dependent fluctuations in the OP states, whereas $PC4$ and $PC5$ may represent finer-scale fluctuations or noise components. Herein, all PCs were included in the analysis to ensure that no relevant information was discarded. Figure S4 in the Supplementary Materials shows a movie that visualizes the T-dependent changes in the spatial distributions of $3\lambda \times 5PC$s, revealing horizontal stripe structures that differ from the data originally entered into the LSTM model.

To further validate this analysis, the K-shape multivariate clustering method was applied to the dataset (3λ, $5PC$, $233T$) × 42,280 pixels. Figure S5 in the Supplementary Materials shows the results of the elbow method and Silhouette score for determining the optimal number of clusters (k). The elbow method shows that the intra-cluster variance decreases considerably from $k = 2$ to 3, but the reduction becomes marginal beyond $k = 3$; this indicates that additional clusters do not provide substantial new insight. The Silhouette score is the highest for $k = 2$ and remains relatively high for $k = 3$ but decreases significantly for $k > 5$. This suggests that increasing k beyond 5 leads to less well-defined clustering, making the classification less interpretable. Based on these results, clusters with $k \geq 8$ were not analyzed considering they would likely produce unstable classifications while imposing unnecessary computational cost. Figure 12 shows the clustering results for $k = 2$ and 3, where the expected stripe structure is clearly visible. Figure S6 in the Supplementary Materials shows the results for $k = 4$ and 5. These results confirm that the T-series dataset with five PCs contains substantial structural information of the datasets. The diversity of the dataset was effectively restored during the training of the LSTM model and that the TsPCA-based reconstruction is successful. However, pixel-to-pixel correlations are not explicitly considered in this analysis step. To further refine clustering and incorporate spatial relationships, the 3DCAE was applied to the T-series data represented by the five PCs.

4.3. 3D convolutional autoencoder (3DCAE)

As shown in Figure 1, conventional PLM causes OP components between adjacent pixels to overlap and obscures spatial features, necessitating appropriate correction. As super-resolution focuses primarily on pixel-wise interpolation, it struggles to separate OP components while preserving the overall trends in an image. We investigate the impact of the convolution of OP states from adjacent pixels on clustering results using the 3DCAE. Here, “3D” refers to the data structure that includes a T axis and two-dimensional spatial dimensions. Figure 13 shows a flowchart of the 3DCAE network used herein. The input dataset is structured as (PC, T, W, H) = (5, 233, 302, 140) for each λ, where W and H denote the width and height of the image, respectively. As a preprocessing step, all pixels and all T steps were treated as a single dataset, and each PC was normalized to the range [0, 1] using the “MinMaxScaler” function. The five normalized variables were transformed into 64 feature maps, and a convolution process was applied with an SRF size of 3×3 pixel. However, no convolution was performed along the T axis because adjacent T points lacked direct physical interactions, making convolution less meaningful. Instead, a sliding window of 5 T points (2.5 K) was used, with each step shifting one T point toward lower T [21,74–76]. No padding was applied during TsPCA, and it was performed using only the available data. As a result, the model could learn spatial structures while maintaining T continuity compared with conventional 2D CNNs. The 3DCAE analysis focused on rapid variations occurring within a narrow T range of 2.5 K because the LSTM was trained over a broader T range of 30.0 K, and TsPCA aggregated data over 19.5 K. During spatial convolution, the “BatchNorm3D” function was applied to normalize each batch and the “ReLU” activation function was used to mitigate the vanishing gradient problem. The batch size was set to four due to computational constraints because the model processed a large dataset containing T-dependent information. Although small batch sizes pose a higher risk of overfitting, this was not a concern herein because the same dataset was used for both training and prediction. Moreover, our primary goal was to analyze intrinsic data structures rather than to generalize unseen data, so any potential overfitting does not compromise the validity of analysis. To extend the analysis beyond the SRF size of 3×3, additional SRF sizes of 5×5 and 10×10 pixels were evaluated. The pixel sizes were generated by downsampling using the “max pooling” function. To avoid periodicity artifacts at image boundaries, zero padding was applied during spatial convolution. The SRF size plays an important role during model training. A considerably small SRF size may fail to capture global regularities, whereas considerably large SRF size may lead to overgeneralization and loss of finer details. An SRF size of 3×3 captures directly adjacent pixels, whereas that of 5×5 extends coverage to the next-nearest neighbors. Both sizes can effectively model local overlapping OP components while preserving spatial details. As the stripe widths observed in the clustering results (Figure 3) ranged from 5 to 10 pixels, the SRF size of 10×10 was tested to evaluate whether the 3DCAE could learn broader periodic structures such as the global periodicity of stripe patterns across the substrate.

“SmoothL1Loss()” was used as the hybrid loss function, which is a combination of MSE ($L2$ loss) and mean absolute error (MAE, $L1$ loss). The MSE is highly sensitive to large errors, so it tends to overreact to outliers. To improve robustness, the function switched to the MAE when the error exceeded a threshold. Conversely, using only the MAE would result in a constant gradient that will slow down the optimization. Therefore, the MSE was used for smaller errors to ensure smoother convergence. This adaptive approach ensured stable training even with high-variance OP data. The default setting of the “SmoothL1Loss()” function was used herein, where the transition threshold was automatically set to 1.0. In other words, errors below this value followed the MSE behavior, whereas larger errors were handled by the MAE. The “Adam” optimizer was used with a learning rate of 0.0001. To improve convergence, the “ReduceLROnPlateau()” function was used because it reduced the learning rate by a factor of 0.5 if the loss did not improve for a given number of iterations. For efficient computation, 16-bit floating-point arithmetic was introduced to reduce memory consumption while maintaining high accuracy. The “autocast()” and “GradScaler()” functions dynamically adjusted computations between 16-bit and 32-bit to ensure accuracy in critical calculations. A batch size of four was chosen to optimize GPU data transfer while maintaining computational stability. Contrary to standard practice, the setting “pin_memory = False” was used to improve stability when processing large datasets, although it could reduce data transfer speed to the GPU. Training was performed for 1,000 epochs. Figure 14(a) shows the representative results with a training time of approximately 30 min. Figures S7(a) and S8(a) in the Supplementary Materials show the loss at each epoch. These loss functions converged after sufficient iterations, confirming stable learning. For prediction, the same data structure, (PC, T, W, H) = (5, 233, 302, 140) for each λ, was used as the input under identical conditions and produced output in the same format.

Figures 14(b–c) show the prediction results for the five PCs at b1 and b2 for the 3×3 SRF size. As the input data were normalized to [0,1] before spatial convolution, the scale of the output data was different from that shown in Figure 10; thus, these were rewritten as ($PC1'$, $PC2'$, $PC3'$, $PC4'$, $PC5'$). The results for the 5×5 and 10×10 cases are shown in Figures S7(b–c) and S8(b–c) of the Supplementary Materials. Although the same data were entered, the prediction results differed considerably depending on the SRF sizes, i.e., the range of change with varying T was small for the 3×3 and 5×5 cases but large for the 10×10 case. The sliding window width along the T axis was set to 2.5 K; therefore, it was much flatter for the 3×3 and 5×5 cases. This indicated a difference in learning trends between the 5×5 and 10×10 cases. In addition, some prediction results exceeded the range [0,1]. As the final output layer does not apply an explicit activation function, the network minimized the reconstruction error in an unconstrained manner. This resulted in small deviations outside the normalization range. Although these deviations were possibly caused by computational artifacts, they may also reflect meaningful variations in the data structure. To assess whether these deviations correspond to meaningful patterns, we analyzed the variations in Mh'' obtained from ($PC1'$, $PC2'$, $PC3'$, $PC4'$, $PC5'$) as a function of the SRF size. Figure 15 shows the results of converting the dataset (3λ, $5PC'$s, $233T$) to (3λ, $1Mh''$, $233T$) at b1 and b2 and comparing them to the LSTM results shown in Figures 8(b–c). The fine spike structures varied with the SRF size, the overall trend remained consistent. The $Mh''(T)$ curves show peaks around T_F and T_c, suggesting that the model successfully captured the essential T-dependent patterns. Possible anomalies were also observed in the 60–100 K range. In this region, the tetragonal quantum paraelectric state was realized. In future, further analysis will be performed to determine whether these anomalies reflect physical phenomena or arise from data variability.

After converting the five PC's into a single Mh'', the K-shape multivariate clustering was applied to the dataset (3λ, $1Mh''$, $233T$) × 42,280 pixels for each SRF size. Figures S9–S11 in the Supplementary Materials show the results of the elbow method, Silhouette score, and gap statistic analysis used to determine the optimal number of clusters (k) for SRF sizes of 3×3, 5×5, and 10×10. As identifying an optimal value of k remains challenging, the candidate values between $k=2$ and 5 are considered. Figures S12–S15 in the Supplementary Materials show the detailed clustering results for each SRF size at $k=2$–5 along with summary tables. The segmentation patterns remain largely consistent across different SRF sizes. This robustness suggests that the essential structural patterns have been consistently extracted, reinforcing the validity of the clustering approach.

To determine whether the 3DCAE effectively learned meaningful representations, the effective receptive field (ERF) and layer-wise relevance propagation (LRP) distributions across different SRF sizes must be analyzed[77, 78]. If the model has not learned properly, the ERF distributions are expected to change randomly across different SRF sizes, without showing any clear pattern. In contrast, if the learned representations are physically meaningful, the ERF and LRP will consistently highlight relevant structural features such as stress-concentration regions and/or periodic stripe structures as a function of SRF size. Therefore, the comparison of ERF and LRP distributions serves as a critical validation step to ensure that the 3DCAE extracts meaningful spatial features and does not merely reflect noise or poor adaptation. Figures 16 and 17 show ERF and LRP images acquired at 575 nm, respectively, and Figures S16–S19 in the Supplementary Materials show the results at 523 and 543 nm. To evaluate the effect of the convolutional layers (“Conv3D”), these results show the raw convolutional output before applying normalization (“BatchNorm3D”) or nonlinear transformations (“ReLU”). During ERF analysis, the regions of higher intensity indicate that the network assigns greater importance to learned features within the SRF size. For an SRF size of 5×5, the ERF intensities of three λ are higher in the stress-concentration regions. This pattern is reasonable because capturing trends in spatially inhomogeneous regions requires integrating information from the adjacent pixels. In contrast, the ERF for an SRF size of 3×3 does not emphasize stress-concentration regions as much as that of 5×5, and an opposite trend is observed at 523 nm. This difference is possibly observed due to the number of convolutional layers: the 5×5 and 10×10 cases use two layers, whereas the 3×3 case has only one (Figure 13). This structural difference indicates that the hierarchical arrangement of convolutional layers may influence feature extraction. However, whether the observed differences are due to the increased network depth or the larger SRF size remain unclear. Meanwhile, the ERF for the 10×10 case follows a completely different trend. Compared to that shown in Figures 2(b) and 16(c), the spatial distribution of the ERF is similar to that of ψ. As ψ represents the slip plane structure formed by dislocations, the ERF patterns in the 10×10 case suggest that the network may capture broader spatial correlations across the substrate. This dramatic variation in the ERF distribution is likely influenced by the stripe width of 5–10 pixels, which is comparable to the SRF size and affects the scale and direction of the extracted features. This suggests that larger SRF sizes integrate more global information, resulting in learned representations that differ from those of smaller SRF sizes that focus primarily on local variations. However, performing direct numerical comparisons is difficult due to the combined effects of spatial correlations and T-dependent trends in each ERF. In contrast, the LRP highlights the most-relevant regions that contribute to the model’s prediction. The LRP results in Figure 17 show the same trend as the ERF results in Figure 16. The results shown in the Supplementary Materials also show similar trends. The agreement between ERF and LRP suggests that the model effectively uses learned features in its predictions which increases the stability and interpretability of the model. Thus, the network is not overfitting to irrelevant patterns but rather captures meaningful OP structures.

Spatial frequency characteristics were analyzed by applying a 2D fast-Fourier transform (FFT) to the ERF data. If the power spectrum was dominated by low-frequency components, it indicated that the convolution emphasized large-scale spatial structures; this may correspond to broad stripe patterns or other global variations across the substrate. In contrast, a dominance of high-frequency components in the power spectrum indicates that the model prioritizes finer spatial details, capturing sharp transitions such as edges in the OP state. To ensure comparability between different SRF sizes, each ERF dataset was standardized using the "StandardScaler" function before FFT analysis. A Hann window was applied to reduce edge effects, and zero padding was used to increase the image size from 302×140 to $(302+100)\times(140+100)=402\times240$ by adding 100 pixels in each dimension[79]. The 100-pixel padding size was chosen to improve frequency resolution by increasing the number of resolvable frequency bins within the Nyquist limit; this enabled a finer sampling of spatial frequency components while maintaining computational efficiency. 2D FFT was computed using the "fft2" function without applying the "fft.fftshift()" function. This was because preserving the default frequency order allowed for the direct interpretation of low- and high-frequency components in their original arrangement without introducing artificial recentering effects.

Bandpass filters were not applied during this analysis to ensure that the 2D FFT results retained all frequency components of the original data. As a result, significant noise was observed in the spectrum. To identify the frequency components ($\mathrm{F}FT(x,y)$) used by the 3DCAE, statistically significant modes were extracted. The absolute values of the frequency components are given by

$$F(u,v)=\left|\mathrm{F}FT(x,y)\right|, \tag{5}$$

where $F(u,v)$ is the amplitude spectrum. The Nyquist frequency range was set to $[-0.5,0.5]$, and the frequency scale was adjusted using the "np.fft.fftfreq()" function. To estimate the background signal intensity, the mean intensity over the entire FFT spectrum ($\mu_{\mathrm{B}ack}$) and its variance ($\sigma^2_{\mathrm{B}ack}$) can be computed as follows:

$$\mu_{\mathrm{B}ack}=\mathrm{mean}(F(u,v)), \tag{6}$$
$$\sigma^2_{\mathrm{B}ack}=\mathrm{var}(F(u,v)). \tag{7}$$

To test whether each frequency mode $F(u,v)$ followed the same statistical distribution as the background, we set the null hypothesis (H_0) to "Each frequency component originates from the same statistical process as the background noise." Based on H_0, the corresponding χ^2-test statistic was expressed as

$$\chi^2(df=2)=\frac{(F(u,v)-\mu_{\mathrm{B}ack})^2}{\sigma^2_{\mathrm{B}ack}}, \tag{8}$$

where $\chi^2(df=2)$ follows a χ^2 distribution with two degrees of freedom[80]. The choice of setting $df=2$ is justified by the statistical properties of the Fourier transform.

Assuming a stationary process, the real and imaginary components of the Fourier transform at each frequency (u,v)—denoted as $\mathrm{R}e[F(u,v)]$ and $\mathrm{I}m[F(u,v)]$—are independently normally distributed with zero mean and equal variance. The power spectrum $S(u,v)$ is given by

$$S(u,v) = | F(u,v) |^2 = \mathrm{R}e[F(u,v)]^2 + \mathrm{I}m[F(u,v)]^2. \tag{9}$$

As $S(u,v)$ is the sum of the squares of two independent normal variables, it follows a χ^2 distribution with $df = 2$:

$$S(u,v) \sim \chi^2(df = 2). \tag{10}$$

This statistical test determines whether a given frequency mode is considerably different from the background noise distribution. To account for multiple hypothesis testing, the false discovery rate (FDR) correction was applied using the Benjamini-Hochberg method; this enables controlling the expected proportion of false positives[80]. If the corrected p-value (q-value) satisfies $q < 0.05$, H_0 is rejected and the corresponding frequency mode is classified as statistically significant. Figure 18 shows the amplitude spectrum of the 2D FFT of ERF for each SRF size at 575 nm, and Figures S20 and S21 in the Supplementary Materials show the results at 523 and 543 nm. The modes highlighted in red represent frequency components identified as significant at the 95 confidence level. The same statistical analysis was applied to the LRP results, which showed similar trends to the ERF results. Detailed FFT LRP results are shown in Figures S22–S24 of the Supplementary Materials. To visualize the dominant frequency structures learned by the 3DCAE, Figure 19 shows an inverse Fourier transform (“ifft2()”) reconstruction of a 302×140 pixel spatial image using only the statistically significant frequency modes at 575 nm. Figures S25 and S26 in the Supplementary Materials show results at 523 and 543 nm, and Figures S27–S29 show the inverse Fourier-transform LRP analysis.

For the SRF sizes of 3×3 and 5×5, the 2D FFT results in Figures 18(a–b) show distinct modes in the top-right diagonal direction; this indicates the presence of a periodic stripe structure oriented in the top-left diagonal direction. The reconstructed images in Figures 19(a–b) further support this interpretation, demonstrating that the 3DCAE effectively learns the stripe structure. The absence of significant modes near the DC component also suggests that the network does not prioritize learning large-scale structural patterns across the substrate. However, as the 3DCAE applies nonlinear transformations, global-scale features may still be captured in a nontrivial manner. In contrast, the 2D FFT spectrum for the 10×10 SRF size in Figure 18(c) shows a different trend, wherein significant modes are observed near the DC component and additional modes are symmetrically distributed around it, resembling satellite reflections. These satellite reflections typically arise from periodic modulations that generate additional harmonics in the spatial frequency domain. This suggests that the 3DCAE for the SRF size of 10×10 captures both the primary periodicity and secondary variations in the spatial structure. In the reconstructed image for the 10×10 case in Figure 19(c), the stripe width appears larger and the periodicity is more pronounced compared with the results for the 3×3 and 5×5 cases. This suggests that the 3DCAE for the 10×10 case, due to its larger SRF size, integrates long-range spatial correlations that potentially capture the overall distortion patterns of the substrate, rather than focusing solely on the localized features as observed in the 3×3 and 5×5 cases.

Occlusion sensitivity analysis (OSA) is used to quantitatively evaluate the spatial regions that are considered important by the 3DCAE[81]. By measuring the effect of occlusions, OSA provides insights into whether the learned features correspond to meaningful physical properties such as stress-concentration and/or periodic stripe modulations or are merely artifacts of the training process. Figure 20 shows the OSA results at 575 nm, and Figures S30 and S31 in the Supplementary Materials show the results at 523 and 543 nm, respectively. In this analysis, specific input regions are masked with a 5×5 pixel window and their effect on the model output is measured using the MSE as the loss function. A significant increase in the MSE indicates that the masked region contains features that are important for the model's decision making, thereby highlighting its importance in the learned representation. During OSA, the sliding window width along the T axis is set to 5 T points (2.5 K) and the sensitivity is measured by shifting the window by one T point. These T conditions are the same as those used for the 3DCAE analysis. The final sensitivity values were obtained by averaging the MSE variations over all the T shifts for each masked region. For the SRF sizes of 3×3 and 5×5, the 3DCAE successfully distinguished between stress-concentration and uniform regions. However, this tendency varied with λ. Specifically, at 523 and 543 nm, the stress-concentration regions exhibited higher sensitivity compared to the uniform regions; however, this trend reversed at 575 nm. Moreover, the stripe structures appeared even in the uniform regions, indicating that the network learned periodic modulations in the OP state. This tendency was consistent with the ERF and LRP results, both of which emphasized the spatial periodicity in the learned and predicted features. In contrast, the OSA results for the SRF size of 10×10 showed a more spatially uniform sensitivity distribution, which was considerably different to the localized variations observed in the 3×3 and 5×5 cases. This indicated that as the SRF size increased, the network integrated information over broader spatial scales across the substrate. The alignment of the square lattice pattern in Figure 20(c) with the significant modes observed in the 2D FFT analysis (Figure 18), particularly the satellite reflections, indicated that the SRF size of 10×10 emphasized large-scale periodicity rather than local stress variations. The considerably larger OSA values observed for the 10×10 SRF, compared to the 3×3 and 5×5 cases, implied that variable output was obtained due to the presence of multiple distinct patterns within a single receptive field. These findings highlighted the sensitivity of the model to spatial scale and its adaptation to the complexity of the input as a function of SRF size.

Based on the OSA results shown in Figure 20, the learned representations were further investigated by examining the feature maps and weight distributions to gain deeper insights into the impact of different SRF sizes on feature extraction and model interpretation. Figure 21 shows the feature maps and feature weight histograms at 575 nm[82]. Figures S32 and S33 in the Supplementary Materials show the results obtained at 523 and 543 nm. Similar to the OSA results, these feature maps showed the raw output of the "Conv3D" function before the "BatchNorm3D" and "ReLU" functions were applied. Some feature map activations have negative values, but regions with large absolute values can be considered highly active. For the SRF size of 3×3, the feature map shows stronger activations in the stress-concentration regions at 523 and 543 nm, whereas the activations are more pronounced in the uniform regions at 575 nm, indicating a shift in the learned feature emphasis. This λ dependence is consistent with the OSA patterns, suggesting that the learned representations of the network vary across λ. The weight histogram in Figure 21(f) shows a symmetric distribution centered around zero, confirming that the model captures meaningful representations without overfitting[83]. A narrow and symmetric distribution suggests that the learned features are accurately regularized and evenly distributed, whereas a skewed distribution indicates that certain feature representations dominate and potentially lead to biased clustering results. Although the feature map and weight distribution for the 10×10 case appear similar to those for the 5×5 case, the OSA results in Figure 20 and the 2D FFT analysis results in Figure 18 reveal distinct differences in the learning of spatial correlations and periodic patterns. As the 3DCAE captures the spatial distribution and T-dependent variations, the spatial feature maps alone may not fully reveal the differences in the learned representations. In the weight histogram shown in Figures 21(g–h), both the first and second layers are symmetrically spread around zero; however the latter layer are narrower. This strongly suggests that the first layer captures the overall large features, whereas the second layer learns more detailed, local features[83]. As shown in Figure 13, the "BatchNorm3D" function stabilizes training by normalizing the activations in each layer and reduces internal covariate shifts. The "Dropout" function, set to 30%, prevents overfitting by reducing co-adaptation between neurons. These mechanisms encourage the network to distinguish between local and global structures; thus, the two-layer histogram trend shown in Figure 21 (g–h) is a natural result of the network's effective learning process. In future, T-dependent trends must be further analyzed to distinguish whether the observed patterns arise from the spatial structure or the T-driven effects. As this study focuses on analyzing the degree of overlap in the OP states, an SRF size of 10×10 is less effective for resolving fine local variations due to its broader spatial integration. In contrast, the SRF sizes of 3×3 and 5×5 strike a balance between local and global feature extraction, making them more suitable for this study. These results confirm that the 3DCAE effectively captures overlapping OP effects at different spatial scales and successfully maintains structural integrity while learning physically meaningful representations.

4.4. Temperature series forest (Tsf)

As the 3DCAE output ($5\,PC'$s, $233T$, 42,280 pixels) $\times 3\lambda$ and its transformed dataset (3λ, $1Mh''$, $233T$, 42,280 pixels) incorporated spatial correlations via the SRF, the resulting T-series data are expected to have more spatial coherence compared with the original TsPCA-based results (Figure 10). Therefore, clustering method must be reconsidered. Although the 3DCAE does not perform direct convolution along the T axis, it uses the sliding window to account for local variations in the OP state with decreasing T. By focusing more effectively on extracting T-dependent trends, the classification primarily reflects T variations rather than spatial artifacts with correlation. The K-shape method is designed to cluster T-series data based on the overall shape similarity rather than absolute values[50–52]. This approach effectively reduces noise sensitivity; however, it struggles with a dataset containing frequent anomalies such as jumps and baseline drifts (Figure 14), which can distort clustering. To address these challenges, temperature series forest (Tsf) was introduced; it is an adaptation of the time-series forest that comprises a decision tree-based ensemble learning approach known for its robustness to noise and outliers[84–86]. Unlike distance-based clustering methods such as the K-shape method, Tsf employs a rule-based classification mechanism that naturally mitigates multicollinearity and improves interpretability. For both training and prediction, the T-series dataset is transformed into multiple representations by varying the T window width and T range. A narrower T window width captures local fluctuations, whereas a wider T window width accounts for broader T-dependent trends. This approach prevents overfitting and mitigates the autocorrelation and unit root effects, thereby improving adaptation to periodicity and random fluctuations in the T-series dataset. As Tsf is a supervised learning method, we first used the K-shape multivariate clustering results for each SRF size shown in Figures S12–S15 of the Supplementary Materials as the initial annotations. Based on the results shown in Figures S9–

S11, the candidate for the number of clusters was set to $k = 2–5$, regardless of the SRF sizes. Using decision trees inherently optimized classification based on the given annotations. However, a potential problem was that the model may simply remember initial annotations rather than generalizing meaningful patterns from the dataset. To mitigate this issue and exploit ensemble learning, a two-step learning method with bootstrap sampling was employed. In this method, pixel labels were randomly sampled from the 42,280-pixel T-series dataset with replacement to generate multiple training subsets. By training each decision tree on different but overlapping data subsets, the overall clustering process became more stable and less sensitive to noise and dataset-specific variations.

First, bagging-based learning was applied using random sampling with replacement, i.e., 20% of the data from each cluster, as defined by the initial annotations, was randomly sampled for each bootstrap sampling. To account for class imbalance, the sample weights were adjusted inversely proportional to the number of samples in each cluster (class_weight = "balanced"). This ensured that underrepresented clusters received adequate training. In total, 1,000 bootstrap datasets were generated and 1,000 classifiers were trained accordingly to increase the ensemble effect and improve the robustness of classification. To introduce diversity among the classifiers, two different T range conditions were considered: (1) a dataset covering the entire T range and (2) a dataset restricted to 50 K or below, focusing on the ferroelectric phase transition. For each T range, successive T-series subsamples were generated using a random or fixed T window width. The starting T position for subsampling was chosen randomly within the defined T range to ensure variability in the training data. As a result, four different classifier types were generated (Table 1). Six feature values derived from statistical descriptors computed from the five PC' s, namely mean, standard deviation, minimum, maximum, linear slope, and coefficient of variation (standard deviation $\div$ mean), were used for this analysis. As a result, the total number of features was $3\lambda \times 5\,PC'$ s $\times 6$ features $= 90$ variables. For the settings of the random forest model[57], the number of decision trees ("n_estimators") was set to 100, the Gini impurity ("gini") was used as the splitting criterion, and the maximum tree depth was left unrestricted. At each split, approximately the square root of the total number of features (i.e., 9 for 90 features) was randomly selected (max_features = "sqrt"). This process ensured randomness in feature selection while keeping the number fixed per split. This configuration was chosen to increase the feature diversity among individual trees and improve the robustness of the ensemble model.

In conventional bagging-based learning, majority voting is typically used to assign a single cluster label to each pixel. However, a more stringent criterion was introduced herein to identify "consistent clusters." This classification was determined based on the predictions of 1,000 classifiers, and the cluster that received the highest number of votes was assigned as the winning cluster label. To enhance the robustness and minimize dependence on the initial annotations, additional conditions were imposed. If the cluster label receiving the most votes matches the initial annotation one, it must receive at least 600 out of 1,000 votes to be considered the winning cluster label. This threshold was introduced to reduce over-reliance of the model on initial annotations. The value of 600 was determined based on bootstrap sampling properties. As each classifier was trained on a random subset containing 20% of the total data, the probability of a given pixel that appeared in at least one classifier was approximately 18.1% ($= 1 - e^{-0.2}$). Consequently, about 200 classifiers (181 to be exact) were strongly influenced by these initial annotations. To ensure that the final decision reflects the ensemble learning effect rather than overfitting to the initial annotations, at least 400 of the remaining 800 classifiers must support the winning cluster label, resulting in the 600-vote threshold. If the number of votes is less than 600, the winning cluster label is marked as "not applicable." Additionally, pixels without a valid label in this step proceed to the second stage of cluster confidence-based adaptive learning (CCAL). In contrast, if the cluster label that receives the most votes differs from the initial annotations, it is directly adopted as the winning cluster label without applying the 600-vote threshold. As a result, the cluster labels change freely if the majority of classifiers support a different label even if some residual influence from the initial annotations remains. After the above steps are completed, if the same winning cluster labels are obtained across all four analysis conditions, then the pixel is classified as a "consistent cluster." This approach ensures that only "consistent clusters" are identified, regardless of variations in the initial cluster counts or the T range conditions used during training. Figures S34–S37 in the Supplementary Materials show the clustering results based on "consistent clusters" only, as well as a transition matrix that tracks changes from the initial annotations. These results indicate that the clusters in the stress-concentration and uniform regions remain stable, whereas other clusters show greater variability and often merge with those in the uniform regions. Across all SRF sizes and initial cluster counts, the training step takes approximately 30 min and the prediction step takes several hours per dataset, possibly because time-consuming file writing is involved while saving the clustering results.

The number of pixels classified as "consistent clusters" via bagging-based learning accounted for 65%–85% of the total pixels, leaving 15%–35% of the pixels unclassified. To deal with the remaining unclassified pixels and further strengthen the learning of "consistent clusters," the second step of CCAL was introduced. First, two types of annotations were prepared: (1) the initial annotations obtained from the K-shape method and (2) the "consistent cluster" labels determined via bagging-based learning. Based on these annotations, 1,000 bootstrap samples with replacement were generated to create training datasets. Specifically, 10% of pixels from each cluster were randomly sampled based on initial annotations, whereas 50% of pixels from each cluster were randomly sampled by the "consistent clusters" labels. To account for class imbalance, the sample weights were inversely proportional to the number of samples in each cluster; this ensured that underrepresented clusters contributed proportionally during training (class_weight = "balanced"). As bootstrap sampling comprised both the initial annotations and "consistent cluster" labels, the same pixel may receive different cluster assignments within a single training subset; this can potentially lead to labeling conflicts. However, the decision trees in the ensemble model recursively partition the feature space based on dominant patterns, and the majority voting mechanism naturally mitigates these inconsistencies. To enhance model generalization and training diversity, four different training conditions were introduced by varying both the T window width and starting T position during training and prediction (Table 2). This approach improved the ability of the model to learn over different T ranges. By incorporating multiple variations in T window width and starting T position, the model effectively captured a wider range of T-dependent trends and performed more robust classification of previously unclassified pixels. Across all SRF sizes and initial cluster counts, the training step took approximately one hour and the prediction step took approximately six hours per dataset.

In the CCAL, the winning cluster labels for each learner were determined based on the majority of votes without applying the 600-vote threshold used in the bagging-based learning. As the CCAL aims to refine the classification of previously unclassified pixels rather than ensuring label stability, strict vote thresholds are not required at this stage. In case of a tie, the pixel is temporarily marked as "not applicable." However, these pixels are not left unclassified but proceed to the next voting step, where the final cluster label is determined based on the aggregated votes of all 1,000 classifiers across four configurations (4,000 votes in total). If at least three of the four learners assign the same winning cluster label, this label is adopted as the final label. This criterion—which corresponds to an agreement threshold of three out of four (75%)—ensures consistency across different training conditions and allows for minor variations. If this match criterion is not met, the pixel is temporarily marked as “not applicable.” For such pixels, the final cluster label is assigned based on the most frequently selected cluster label among the 4,000 votes. This approach ensures that the classification reflects the most dominant trend observed in the ensemble. Moreover, no cases were observed wherein the most frequently assigned cluster label results in a tie. Figures S38–S41 in the Supplementary Materials show the final clustering results using the initial annotations with $k = 2$–5 and a transition matrix showing the evolution of cluster labels through the CCAL. These results show that the “consistent clusters” remain stable after the CCAL, with only minor label adjustments. In contrast, most of the previously unclassified pixels are integrated into the uniform stress regions, suggesting that these regions lack sufficient distinctiveness to form separate clusters.

Figures S38–S41 in the Supplementary Materials show the final clustering results of the CCAL that exhibit a similar trend across different the SRF sizes. To evaluate the clustering performance, the Silhouette score was computed that ranged from -1 to $+1$ using two types of T-series datasets[60, 61]. The first dataset contained the input of the LSTM model between 130.0 and 14.0 K, i.e., (3λ, $1Mh$, $233T$, 42,280 pixels), as shown in Figure 5. The second dataset contained the output of the 3DCAE (3λ, $5PC'$s, $233T$, 42,280 pixels), as shown in Figures 14(b–c). Figure 22 shows the variation in the Silhouette scores with the number of clusters (k) as the initial annotation when the sliding T window width is set to 5 T points (2.5 K). Figure S42 in the Supplementary Materials shows the results for a wider sliding T window of 19 T points (9.5 K). Similar to the aforementioned analysis method, the sliding T window shifted by one T from 130.0 K to 14.0 K and the Silhouette score was computed repeatedly. Specifically, if the data within the sliding T window were outside the range of 130.0–14.0 K, the calculation was performed using only the existing data without padding. The final Silhouette score was the average of the 233 calculated results obtained by shifting the sliding T window, and the 95% confidence interval was expressed as an error bar. Figures 22 and S42 show that the highest Silhouette score is achieved at $k = 2$ for the SRF sizes of 3×3 and 10×10, respectively. For the SRF size of 5×5, the highest Silhouette score is achieved at $k = 3$, but the Tsf leads to the integration of three clusters into two. Thus, regardless of the SRF size, the final clustering results are consistently separated into two groups that correspond to stress-concentration and uniform regions. As shown in Figures S1, S5, and S9–S11 of the Supplementary Materials, the Silhouette score is consistently the highest for $k = 2$, regardless of the dataset. The cumulative explained variance ratio in Figure 11 indicates that the TsPCA dataset can be divided into two or three clusters, indicating that splitting the data into two groups is justified. When the optimal number of clusters k was evaluated using the gap statistic, no significant differences were observed in the results even when the k value and SRF size were varied. In general, gap statistic evaluates the stability of the global cluster structure, whereas the Silhouette score measures the local compactness and separation between clusters[60]. If the overall clustering pattern is already well defined, gap statistic will remain stable even when k changes. This indicates that further subdivision will not considerably alter the overall structure. However, the Silhouette score may still vary because it reflects how well individual data points fit into their assigned clusters. In other words, even if the fundamental clustering framework is robust (as indicated by the gap statistic), finer adjustments in partitioning (reflected in the Silhouette score) may improve intra-cluster cohesion. This distinction explains the consistency across both metrics despite their different behaviors. Thus, the stability of the gap statistic suggests that the underlying cluster structure is robust, whereas the variation in the Silhouette score highlights the existence of optimal partitioning subdivisions within that structure.

Figure 23 shows the final clustering results, indicating the highest Silhouette score for each SRF size. At the SRF size of 3×3, the stripe length is long compared with the previous clustering result shown in Figure 3. In contrast, for the SRF sizes of 5×5 and 10×10, the stripe length is shorter and the stress-concentration regions seem to extract only the larger δ regions compared with that of for the SRF size of 3×3. This clearly indicates that no single "correct" clustering solution is available. Based on the report on statistical causal inference using the clustering result (Figure 3), when considering the treatment effect on δ due to stress, the stress-concentration regions E1 and E2 (10,571 pix) are considered the treatment group. Contrarily, the uniform regions E3 and E4 (31,709 pix) are classified as the untreated group[26]. As shown in Figure 23, the number of pixels assigned to the treated and untreated groups is $15,432:26,848$ for the 3×3 SRF size at $k=2$, $7,948:34,332$ for the 5×5 case at $k=3$, and $6,970:35,310$ for the 10×10 case at $k=2$. This trend indicates that increasing the SRF size allows for a more selective extraction of stress-concentration regions, potentially leading to a more accurate evaluation of the treatment effect. As an alternative analytical approach, the reliability of the cluster labels can be exploited by using only the "consistent clusters" for causal inference. Based on these conditions, the 10×10 ($k=2$) case is too broad and primarily captures the overall substrate regularity, whereas the 3×3 ($k=2$) case may not sufficiently isolate the stress-concentration regions. We therefore conclude that the 5×5 ($k=3$) clustering result shown in Figure 23(b) provides an optimal balance between preserving local sensitivity and ensuring spatial coherence, thereby making it the most appropriate choice for stress distribution analysis. Based on the 5×5 ($k=3$) clustering result, Figures S43–48 in the Supplementary Materials show the distributions of δ and ψ at different T divided into the two groups for 3×3 ($k=2$), 5×5 ($k=3$), and 10×10 ($k=2$) cases. These results confirm that stress-concentration and uniform regions are successfully separated at all T. However, while the distribution of ψ remains consistent regardless of T, δ exhibits a T dependence. Even within the cluster H2 for 5×5 ($k=3$), the intensity ratio of δ in regions H2-1, H2-2, and H2-3 (Figures S45 and S46) varies with T. As these three regions mainly contain "consistent clusters" identified in Figure S35 of the Supplementary Materials, their classification as H2 is not random. Future analysis must entail further understanding of these trends along the T axis for these three regions.

Notably, unlike conventional super-resolution techniques that primarily focus on pixel-wise interpolation and do not inherently separate overlapping OP states, the proposed method uses the 3DCAE to spatially convolve OP states in combination with temperature variations. The learned patterns revealed by FFT analysis of the ERF vary with SRF sizes; however, the clustering results remain remarkably consistent and distinguish between stress-concentration and uniform regions, irrespective of SRF size. This consistency highlights the robustness of the proposed clustering approach and emphasizes that it reliably extracts latent OP features without requiring a quantitative evaluation of "polarization resolution." By effectively capturing spatial OP patterns and preserving their T-dependent properties, this framework provides a more accurate and reliable characterization of OP states. In future, the potential of inverse convolutional learning can be explored using only the decoder component ("ConvTranspose3D"), as shown in Figure 13; this will directly improve "polarization resolution" by reconstructing fine spatial structures based on the learned hierarchical features without introducing artificial smoothing effects, thereby complementing the OP coherence achieved herein.

5. Conclusions

In this study, we proposed a novel deep learning-based framework for analyzing T-dependent birefringence images of stress-induced ferroelectric $SrTiO_3$. The proposed approach addressed the fundamental challenge of overlapping OP components, which is a limitation that cannot be resolved by conventional super-resolution techniques. Rather than quantifying the improvement in “polarization resolution,” we robustly identified and separated intrinsic OP states by ensuring consistency in clustering results, independent of variations in the SRF sizes.

Statistical analysis, machine learning, and deep learning methods were integrated into a sequential analytical pipeline. First, the Mahalanobis distance was used to represent dependent variables and preserve their interrelationships. Then, an LSTM network was used to extract T-dependent OP features via multiple dense layers, and TsPCA was applied to address multicollinearity. The 3DCAE was then introduced to convolve spatially overlapping OP states while accounting for T variations. Despite significant variations in learned feature patterns with varying SRF sizes, Tsf analysis yielded robust and stable clustering results that consistently distinguished between stress-concentrated and uniform regions. The stress-induced ferroelectric states were effectively divided into two distinct OP states, thereby allowing a more accurate assessment of stress-concentration and uniform regions. As a result, the treated and untreated groups could be appropriately discriminated, which is essential for statistical causal inference. Using only the “consistent clusters” identified via ensemble learning, causal relationships could be clarified with higher accuracy.

Unlike conventional super-resolution techniques that rely on pixel-wise interpolation, the proposed method successfully reconstructed OP features while preserving local structure. The deep learning-based methods provided a powerful tool for analyzing OP states in complex materials. In the future, this framework can be extended by integrating transformers for improved sequential feature extraction along the T axis and graph neural networks for capturing complex spatial correlations. These advances will help to further refine feature extraction and broaden the applicability of OP analysis to diverse material systems.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially supported by a Grant-in-Aid for Scientific Research (KAKENHI) (Grant Numbers JP21K04897 and JP23K03283) from the Japan Society for the Promotion of Science.

References

[1] Abbe E. Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. Arch

Mikrosk Anat. 1873;9(1):413–468.

[2] Born M, Wolf E. Principles of optics: electromagnetic theory of propagation, interference and diffraction of light. Cambridge University Press; 1999.

[3] Betzig E, Trautman JK. Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. Science. 1997;257(5067):189–195.

[4] Hell SW, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett. 1994;19(11):780–782.

[5] Betzig E, Patterson GH, Sougrat R, et al. Imaging intracellular fluorescent proteins at nanometer resolution. Science. 2006;313(5793):1642–1645.

[6] Rust MJ, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods. 2006;3(10):793–795.

[7] Gustafsson MG. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. J Microsc. 2000;198(2):82–87.

[8] Moerner WE, Orrit M. Illuminating single molecules in condensed matter. Science. 1999;283(5408):1670–1676.

[9] Huang B, Wang W, Bates M, et al. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science. 2008;319(5864):810–813.

[10] Rivenson Y, Göröcs Z, Günaydin H, et al. Deep learning microscopy. Optica. 2017;4(11):1437–1443.

[11] Ouyang W, Aristov A, Lelek M, et al. Deep learning massively accelerates super-resolution localization microscopy. Nat Biotechnol. 2018;36(5):460–468.

[12] Weigert M, Schmidt U, Boothe T, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. Nat Methods. 2018;15(12):1090–1097.

[13] Wang H, Rivenson Y, Jin Y, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. Nat Methods. 2019;16(1):103–110.

[14] Zhang H, Fang C, Xie X, et al. High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. Biomed Opt Express. 2019;10(3):1044–1063.

[15] Jin L, Kondoh E, Iizuka Y, et al. Lateral ellipsometry resolution for imaging ellipsometry measurement. Jpn J Appl Phys. 2021;60(5):058003.

[16] Azzam RM, Bashara NM. Ellipsometry and polarized light. Amsterdam: North Holland;

1977.

[17] Lee W, Yeh Y. Polarization diversity system for mobile radio. IEEE Trans Commun. 1972;20(5):912–923.

[18] Dinc T, Chakrabarti A, Krishnaswamy H. A 60 GHz CMOS full-duplex transceiver and link with polarization-based antenna and RF cancellation. IEEE J Solid-State Circuits. 2016;51(5):1125–1140.

[19] Tan M, Xu X, Wu J, et al. Orthogonally polarized RF optical single sideband generation with integrated ring resonators. J Semicond. 2021;42(4):041305.

[20] Wang Y, Xie Z, Xu K, et al. An efficient and effective convolutional auto-encoder extreme learning machine network for 3d feature learning. Neurocomputing. 2016;174:988–998.

[21] Mei S, Ji J, Geng Y, et al. Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification. IEEE Trans Geosci Remote Sens. 2019;57(9):6808–6820.

[22] Zhao J, Hu L, Dong Y, et al. A combination method of stacked autoencoder and 3D deep residual network for hyperspectral image classification. Int J Appl Earth Obs Geoinf. 2021;102:102459.

[23] Pintelas E, Pintelas P. A 3D-CAE-CNN model for deep representation learning of 3D images. Eng Appl Artif Intell. 2022;113:104978.

[24] Özdemir OB, Koz A. 3D-CNN and autoencoder-based gas detection in hyperspectral images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:1474–1482.

[25] Li C, Cai R, Yu J. An attention-based 3D convolutional autoencoder for few-shot hyperspectral unmixing and classification. Remote Sens. 2023;15(2).

[26] Seike K, Manaka H, Miura Y. Causal inference in statistics insights into stress-induced ferroelectric states in $SrTiO_3$: Disentangling piezoelectric and flexoelectric effects from birefringence images. Sci Technol Adv Mater Meth. 2025;5:2503698.

[27] Manaka H, Uetsubara K, Korogi S, et al. Microscopic observation of ferroelectric and structural phase transitions in $SrTiO_3$ under uniaxial stress using birefringence imaging techniques. J Phys Soc Jpn. 2022;91(8):084704.

[28] Manaka H, Uetsubara K, Miura Y. Stress-induced ferroelectricity in quantum paraelectric $SrTiO_3$ observed by birefringence imaging. J Phys Conf Ser. 2023;38:011112.

[29] Toyoda K, Manaka H, Miura Y. Improvements of birefringence imaging techniques to observe stress-induced ferroelectricity in $SrTiO_3$ based on K-means clustering with circular statistics. Sci Technol Adv Mater Meth. 2023;3:2278322.

[30] Manaka H, Toyoda K, Miura Y. Multivariate temperature-series analysis of stress-induced ferroelectricity in $SrTiO_3$: a machine learning approach with *K*-shape clustering and hierarchical bayesian estimation. Sci Technol Adv Mater Meth. 2024;4:2342234.

[31] Manaka H, Yagi G, Miura Y. Development of birefringence imaging analysis method for observing cubic crystals in various phase transitions. Rev Sci Instrum. 2016;87(7):073704.

[32] Manaka H, Sasaki Y, Miura Y. Re-examination of successive structural phase transitions in $(C_3H_7NH_3)_2$ $CuCl_4$ using birefringence imaging and electron paramagnetic resonance spectroscopy. J Phys Soc Jpn. 2017;86(11):114710.

[33] Miura Y, Okumura K, Fukuda T, et al. Observation of ferroelastic domains in layered magnetic compounds using birefringence imaging. J Phys Conf Ser. 2018;969:012153.

[34] Manaka H, Okumura K, Tokunaga K, et al. Observations of successive local-structure and ferroelectric phase transitions in $(C_2H_5NH_3)_2$ $CuCl_4$ using birefringence imaging and electron paramagnetic resonance spectroscopy. J Phys Soc Jpn. 2022;91(11):114701.

[35] Masetti E, de Silva MP. Development of a novel ellipsometer based on a four-detector photopolarimeter. Thin Solid Films. 1994;246(1-2):47–52.

[36] Sato T, Araki T, Sasaki Y, et al. Compact ellipsometer employing a static polarimeter module with arrayed polarizer and wave-plate elements. Appl Opt. 2007;46(22):4936–4937.

[37] Manaka H, Fukuda T, Miura Y. Birefringence imaging measurements on various structural phase transitions in $(C_nH_{2n+1}NH_3)_2$ $MnCl_4$ with n=1,2 and 3 using multiple wavelengths. J Phys Soc Jpn. 2016;85(12):124701.

[38] Cowley RA. Lattice dynamics and phase transitions of strontium titanate. Phys Rev. 1964;134:A981.

[39] Courtens E. Birefringence of $SrTiO_3$ produced by the 105° K structural phase transition. Phys Rev Lett. 1972;29:1380.

[40] Cowley RA. The phase transition of strontium titanate. Phil Trans R Soc A. 1996;354(1720):2799–2814.

[41] Sawaguchi E, Kikuchi A, Kodera Y. Dielectric constant of strontium titanate at low temperatures. J Phys Soc Jpn. 1962;17(10):1666–1667.

[42] Hegenbarth E. Die feldstärkeabhängigkeit der dielektrizitätskonstanten von $SrTiO_3$-einkristallen im temperaturbereich von 15 bis 80° K. Phys Status Solidi. 1964;6:333.

[43] Fleury PA, Worlock JM. Electric-field-induced raman scattering in $SrTiO_3$ and $KTaO_3$. Phys Rev. 1968;174:613.

[44] Hemberger J, Lunkenheimer P, Viana R, et al. Electric-field-dependent dielectric constant and nonlinear susceptibility in $SrTiO_3$. Phys Rev B. 1968;174:613.

[45] Hemberger J, Nicklas R M Viana, Lunkenheimer P, et al. Quantum paraelectric and induced ferroelectric states in $SrTiO_3$. J Phys Condens Matter. 1996;8:4673.

[46] Manaka H, Nozaki H, Miura Y. Microscopic observation of ferroelectric domains in $SrTiO_3$ using birefringence imaging techniques under high electric fields. J Phys Soc Jpn. 2017;86(11):114702.

[47] Manaka H, Nozaki H, Miura Y. Development of birefringence imaging techniques under high electric fields. J Phys Conf Ser. 2018;969:012119.

[48] Manaka H, Tateishi K, Miura Y. Real-space imaging by magnetic birefringence for $KNiF_3$ under inhomogeneous stress. J Phys Soc Jpn. 2019;88(12):124702.

[49] Fisher NI. Statistical analysis of circular data. Cambridge University Press; 1993.

[50] Senin P. Dynamic time warping algorithm review. Inf Comput Sci Dep Univ Hawaii Manoa Honolulu, USA. 2008;855(1–23):40.

[51] Box GEP, Jenkins GM, Reinsel GC, et al. Time series analysis: forecasting and control. John Wiley & Sons; 2015.

[52] Paparrizos J, Gravano L. *K*-shape: Efficient and accurate clustering of time series. In: SIGMOD Rec.; Vol. 45; 2016. p. 69–76.

[53] Bishop CM. Pattern recognition and machine learning. Springer; 2006.

[54] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.

[55] Gers FA, Schmidhuber J. Recurrent nets that time and count. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium; Vol. 3; 2000. p. 189–194.

[56] Wang X, Kruger U, Irwin GW. Process monitoring approach using fast moving window PCA. Ind Eng Chem Res. 2005;44(15):0888–5885.

[57] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[58] Verbeke A. Beyond addressing multicollinearity: robust quantitative analysis and machine learning in international business research. J Int Bus Stud. 2022;53:1307–1314.

[59] Yildirim H. The multicollinearity effect on the performance of machine learning algorithms: case examples in healthcare modelling. Acad Platform J Eng Smart Syst. 2024;12:68–80.

[60] Řezanková H. Different approaches to the Silhouette coefficient calculation in cluster evaluation. In: Applications of Mathematics and Statistics in Economics 2018, Conference Proceedings; 2018. p. 259–268.

[61] Dinh D, Fujinami T, Huynh V. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. Commun Comput Inf Sci. 2019;:1–17.

[62] Sagala NTM, Gunawan AAS. Discovering the optimal number of crime cluster using elbow, Silhouette, gap statistics, and NbClust methods. ComTech: Comput Math Eng Appl. 2022;13(1):1–10.

[63] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004;20(3):374–380.

[64] Klement S, Madany Mamlouk A, Martinetz T. Reliability of cross-validation for SVMs in high-dimensional, low sample size scenarios. In: Advances in Data Analysis, Data Handling and Business Intelligence. Springer; 2008. p. 41–50.

[65] Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage. 2018;180:68–77.

[66] Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size. Plos One. 2019;14(11):e0224365.

[67] Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018;4(1):25.

[68] D'souza RN, Huang PY, Yeh FC. Structural analysis and optimization of convolutional neural networks with a small sample size. Sci Rep. 2020;10(1):8348.

[69] Xu P, Ji X, Li M, et al. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):85.

[70] Itoh M, Wang R, Inaguma Y, et al. Ferroelectricity induced by oxygen isotope exchange in strontium titanate perovskite. Phys Rev Lett. 1999;82:3540–3543.

[71] Maity A, Habicht K, Merz M, et al. Soft phonon and the central peak at the cubic-to-tetragonal phase transition in $SrTiO_3$. Phys Rev B. 2025;111:134108.

[72] Siami-Namini S, Siami Namin A. Forecasting economics and financial time series: ARIMA vs. LSTM. arXiv preprint. 2018;.

[73] Siami-Namini S, Tavakoli N, Siami Namin A. A comparison of ARIMA and LSTM in forecasting time series. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018. p. 1394–1401.

[74] Mei S, Ji J, Geng Y, et al. Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification. IEEE Trans Geosci Remote Sens. 2019;57(9):6808–6820.

[75] Wang D, Zhuang L, Gao L, et al. Sliding dual-window-inspired reconstruction network for hyperspectral anomaly detection. IEEE Trans Geosci Remote Sens. 2024;62:1–15.

[76] Yang B, Mao Y, Liu L, et al. Change representation and extraction in stripes: Rethinking unsupervised hyperspectral image change detection with an untrained network. IEEE Trans Image Process. 2024;33:5098–5113.

[77] Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks. arXiv preprint. 2017;.

[78] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7):e0130140.

[79] Harris F. On the use of windows for harmonic analysis with the discrete fourier transform. Proc IEEE. 1978;66(1):51–83.

[80] Dalgaard P. Introductory statistics with R. Springer; 2008.

[81] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. arXiv preprint. 2013;.

[82] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization. arXiv preprint. 2015;.

[83] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization. arXiv preprint. 2017;.

[84] Bagnall A, Davis L, Hills J, et al. Transformation based ensembles for time series classification. Proceedings of the 2012 SIAM International Conference on Data Mining (SDM). 2012;:307–318.

[85] Lines J, Bagnall A. Time series classification with ensembles of elastic distance measures. Data Min Knowl Disc. 2015;29:565–592.

[86] Goehry B, Yan H, Goude Y, et al. Random forests for time series. REVSTAT-Stat J. 2023;21(2):283–302.

Table 1. Extraction conditions for T-series data in bagging-based learning. A successive T-series dataset, with either a fixed or randomly selected T window width, is extracted within the specified T range. The starting T position for subsampling within the specified T range is chosen randomly.

Temperature range (K)	Data points	Window width for learning	Window width for prediction
130.0 → 14.0	233	60 (Fixed)	60 (Fixed)
130.0 → 14.0	233	19–60 (Random)	19–60 (Random)
50.0 → 14.0	73	19 (Fixed)	19 (Fixed)
50.0 → 14.0	73	5–19 (Random)	5–19 (Random)

Table 2. Extraction conditions for the T-series data in cluster confidence-based adaptive analysis. A successive T-series dataset, with either a fixed or randomly selected T window width, is extracted within the specified T range. The starting T position for subsampling within the defined T range is chosen randomly.

Temperature range (K)	Data points	Window width for learning	Window width for prediction
130.0 → 14.0	233	19–60 (Random)	60 (Fixed)
130.0 → 14.0	233	19–60 (Random)	19–60 (Random)
50.0 → 14.0	73	5–19 (Random)	19 (Fixed)
50.0 → 14.0	73	5–19 (Random)	5–19 (Random)

Figure 1: Schematic of resolution concepts in polarized light microscopy. (a) Overlapping distributions ("optical resolution") that cannot be separated merely using only pixel size refinement. (b) Well-separated distributions ("polarization resolution"), where different polarization components are resolved.

Figure 2: Spatial distributions of (a) retardance δ and (b) fast-axis direction ψ at 14.1 K for $\lambda = 575$ nm. Crosses in (a) indicate the positions of b1 and b2 used in the subsequent analysis. This dataset was previously reported in Ref. [28].

Figure 3: Results of K-shape multivariate clustering using 12 variables derived from four independent components, namely $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi)$, evaluated at temperature intervals of -0.34 K. This analysis is based on the dataset reported in Ref. [30].

Figure 4: Overview of the data processing pipeline—from experimental data acquisition to robust clustering. The workflow is categorized into three stages: statistical analysis, machine learning, and deep learning.

Figure 5: Temperature dependence of the Mahalanobis distance (Mh) at three wavelengths at positions (a) b1 and (b) b2 shown in Figure 2(a).

Figure 6: Results of K-shape multivariate clustering with three variables applied to a dataset of (3 λ, 1 Mh, 233 T, 42,280 pixels) for (a) $k=4$ and (b) $k=5$.

Figure 7: Flowchart of the long short-term memory (LSTM) model. The input dataset (3 λ, 1 Mh, 233 T, 42,280 pixels) is processed through five dense layers and the predicted Mahalanobis distance (Mh') is the output obtained via a final linear transformation.

Figure 8: (a) Evolution of the loss function during LSTM training using the dataset (3 λ, 1 Mh, 293 T) across all 42,280 pixels. Temperature dependence of the Mahalanobis distance (Mh') for 575 nm predicted using the LSTM at (b) b1 and (c) b2. Circles denote the experimental Mahalanobis distance data (Mh), and solid lines indicate LSTM predictions (Mh').

Figure 9: Temperature dependence of the extracted features at 575 nm from the five dense layers of the LSTM model at (a) b1 and (b) b2.

Figure 10: Results of temperature-series principal component analysis (TsPCA) at (a) b1 and (b) b2.

Figure 11: Cumulative explained variance ratio as a function of temperature obtained via TsPCA.

Figure 12: Results of K-shape multivariate clustering with 15 variables applied to a dataset (3 λ, 5 PCs, 233 T, 42,280 pixels) shown for (a) $k=2$ and (b) $k=3$.

Figure 13: Flowchart of the 3D convolutional autoencoder (3DCAE). The input datasets reconstructed from the TsPCA results have dimensions of (5 PCs, 233 T, 302 pixels, 140 pixels) for three wavelengths (λ). These datasets are processed via convolutions with the spatial receptive field (SRF) sizes of 3×3, 5×5, and 10×10. The output datasets have dimensions of (5 PC's, 233 T, 302 pixels, 140 pixels) for three λ.

Figure 14: (a) Evolution of the loss function during the training of 3DCAE using the dataset (5 PCs, 233 T, 302 pixels, 140 pixels) at 575 nm for the SRF size of 3×3. Temperature dependence of the five principal components (PC's) predicted by the 3DCAE at (b) b1 and (c) b2.

Figure 15: Temperature dependence of the Mahalanobis distance (Mh'') at 575 nm for different SRF sizes, derived from the 3DCAE predictions (PC's), at positions (a) b1 and (b) b2. Triangles represent the Mahalanobis distances (Mh'') obtained from the 3DCAE. As shown in Figures 8(b–c), red circles denote the experimental Mahalanobis distance data (Mh) and solid lines indicate the LSTM predictions (Mh').

Figure 16: Comparison of effective receptive field (ERF) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

Figure 17: Comparison of layer-wise relevance propagation (LRP) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

Figure 18: Comparison of the amplitude spectra of 2D fast Fourier-transform (FFT) ERF images at 575 nm for different spatial receptive field sizes: (a) 3×3, (b) 5×5, and (c) 10×10. Red circles indicate frequency components identified as significant at the 95% confidence level.

Figure 19: Comparison of inverse Fourier-transform ERF images at 575 nm for different spatial receptive field sizes: (a) 3×3, (b) 5×5, and (c) 10×10, using only the statistically significant frequency modes identified in Figure 18.

Figure 20: Comparison of occlusion sensitivity analysis (OSA) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

Figure 21: Comparison of feature maps at 575 nm for different SRF size sizes: (a) 3×3, (b–c) 5×5, and (d–e) 10×10. The 3×3 SRF size includes a single-layer structure, whereas the 5×5 and 10×10 SRF sizes include two layers. The histograms of feature weights are shown for (f) 3×3, (g) 5×5, and (h) 10×10.

Figure 22: Silhouette scores for different initial numbers of clusters (k) computed using temperature-series forests (Tsf) for the SRF sizes of (a) 3×3, (b) 5×5, and (c) 10×10. The sliding temperature window width is fixed at 5 points (2.5 K).

Figure 23: Final clustering results obtained via cluster confidence-based adaptive learning for different SRF sizes and initial cluster numbers (k): (a) 3×3 with $k=2$, (b) 5×5 with $k=3$, and (c) 10×10 with $k=2$. These results highlight the effect of SRF sizes on the classification of polarization states.

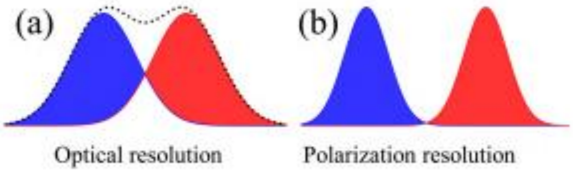

Figure 1. Schematic of resolution concepts in polarized light microscopy. (a) Overlapping distributions ("optical resolution") that cannot be separated merely using only pixel size refinement. (b) Well-separated distributions ("polarization resolution"), where different polarization components are resolved.

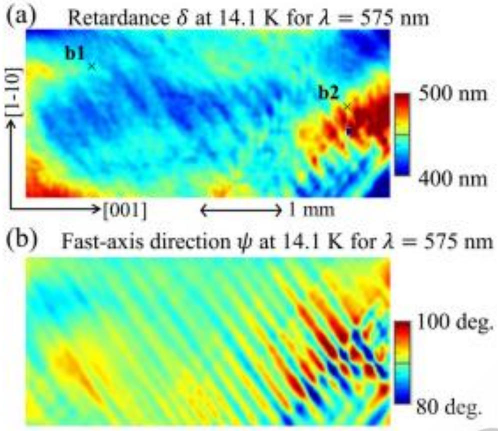

Figure 2. Spatial distributions of (a) retardance δ and (b) fast-axis direction ψ at 14.1 K for $\lambda = 575$ nm. Crosses in (a) indicate the positions of b1 and b2 used in the subsequent analysis. This dataset was previously reported in Ref. [28].

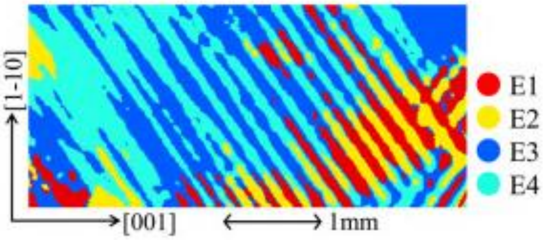

Figure 3. Results of K-shape multivariate clustering using 12 variables derived from four independent components, namely $(\cos\theta/\overline{R}_\theta, \sin\theta/\overline{R}_\theta, \cos 2\phi/\overline{R}_\phi, \sin 2\phi/\overline{R}_\phi)$, evaluated at temperature intervals of -0.34 K. This analysis is based on the dataset reported in Ref. [30].

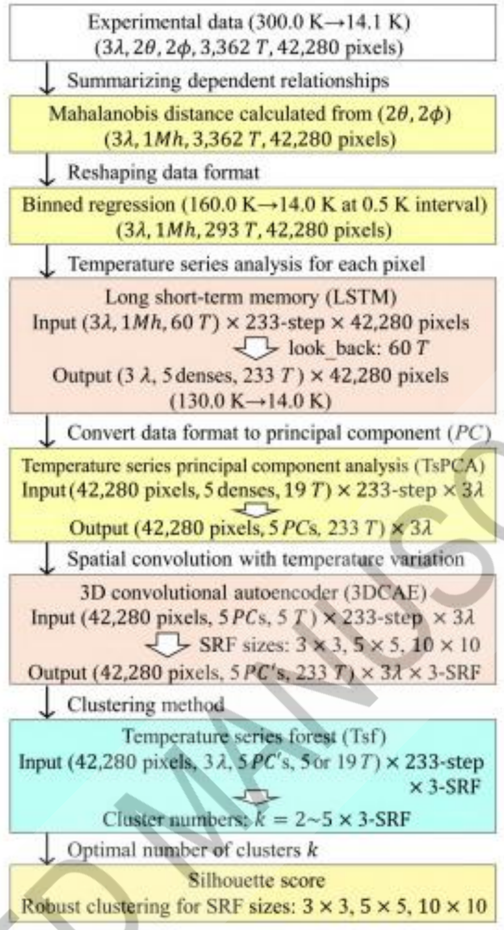
Experimental data (300.0 K→14.1 K)
(3λ, 2θ, 2ϕ, 3,362 T, 42,280 pixels)
Summarizing dependent relationships
Mahalanobis distance calculated from (2θ, 2ϕ)
(3λ, 1Mh, 3,362 T, 42,280 pixels)
Reshaping data format
Binned regression (160.0 K→14.0 K at 0.5 K interval)
(3λ, 1Mh, 293 T, 42,280 pixels)
Temperature series analysis for each pixel
Long short-term memory (LSTM)
Input (3λ, 1Mh, 60 T) × 233-step × 42,280 pixels
look_back: 60 T
Output (3 λ, 5 denses, 233 T) × 42,280 pixels
(130.0 K→14.0 K)
Convert data format to principal component (PC)
Temperature series principal component analysis (TsPCA)
Input (42,280 pixels, 5 denses, 19 T) × 233-step × 3λ
Output (42,280 pixels, 5 PCs, 233 T) × 3λ
Spatial convolution with temperature variation
3D convolutional autoencoder (3DCAE)
Input (42,280 pixels, 5 PCs, 5 T) × 233-step × 3λ
SRF sizes: 3 × 3, 5 × 5, 10 × 10
Output (42,280 pixels, 5 PC′s, 233 T) × 3λ × 3-SRF
Clustering method
Temperature series forest (Tsf)
Input (42,280 pixels, 3λ, 5 PC′s, 5 or 19 T) × 233-step × 3-SRF
Cluster numbers: k = 2~5 × 3-SRF
Optimal number of clusters k
Silhouette score
Robust clustering for SRF sizes: 3 × 3, 5 × 5, 10 × 10

Figure 4. Overview of the data processing pipeline—from experimental data acquisition to robust clustering. The workflow is categorized into three stages: statistical analysis, machine learning, and deep learning.

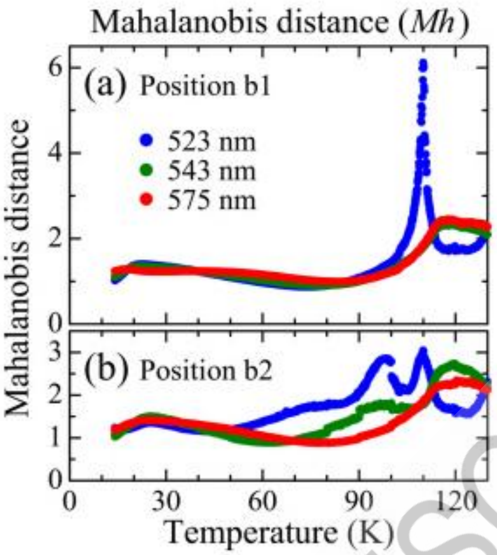

Figure 5. Temperature dependence of the Mahalanobis distance (Mh) at three wavelengths at positions (a) b1 and (b) b2 shown in Figure 2(a).

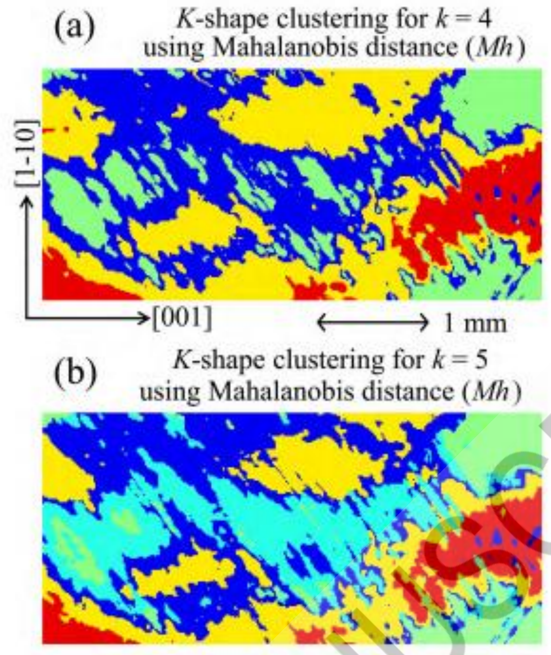

Figure 6. Results of K-shape multivariate clustering with three variables applied to a dataset of (3 λ, 1 Mh, 233 T, 42,280 pixels) for (a) $k = 4$ and (b) $k = 5$.

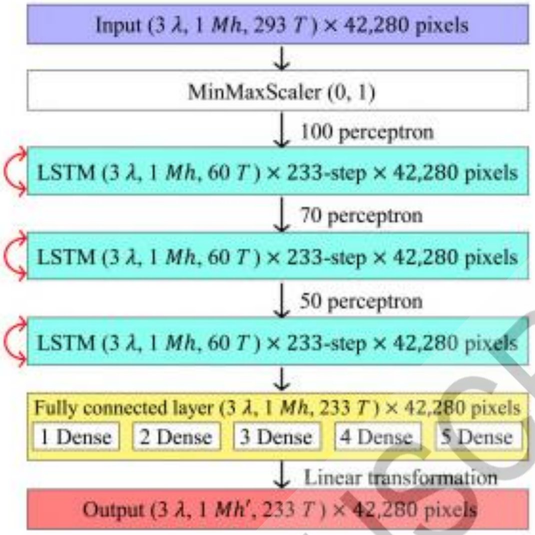
Input (3 λ, 1 Mh, 293 T) × 42,280 pixels
MinMaxScaler (0, 1)
100 perceptron
LSTM (3 λ, 1 Mh, 60 T) × 233-step × 42,280 pixels
70 perceptron
LSTM (3 λ, 1 Mh, 60 T) × 233-step × 42,280 pixels
50 perceptron
LSTM (3 λ, 1 Mh, 60 T) × 233-step × 42,280 pixels
Fully connected layer (3 λ, 1 Mh, 233 T) × 42,280 pixels
1 Dense
2 Dense
3 Dense
4 Dense
5 Dense
Linear transformation
Output (3 λ, 1 Mh′, 233 T) × 42,280 pixels

Figure 7. Flowchart of the long short-term memory (LSTM) model. The input dataset (3 λ, 1 Mh, 233 T, 42,280 pixels) is processed through five dense layers and the predicted Mahalanobis distance (Mh') is the output obtained via a final linear transformation.

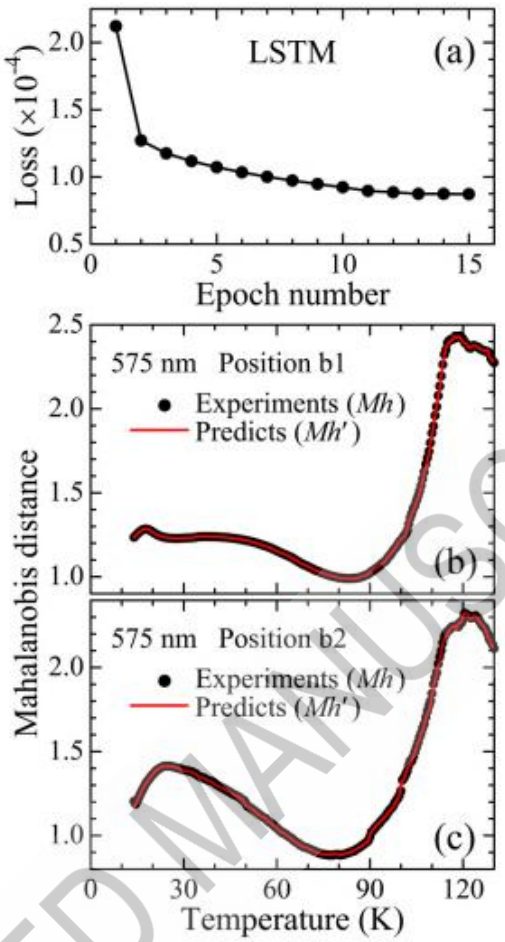

Figure 8. (a) Evolution of the loss function during LSTM training using the dataset (3 λ, 1 Mh, 293 T) across all 42,280 pixels. Temperature dependence of the Mahalanobis distance (Mh') for 575 nm predicted using the LSTM at (b) b1 and (c) b2. Circles denote the experimental Mahalanobis distance data (Mh), and solid lines indicate LSTM predictions (Mh').

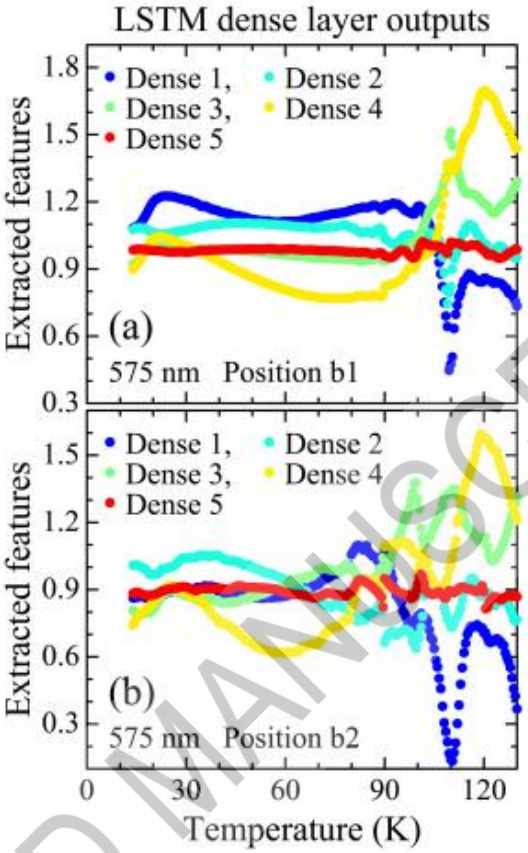

Figure 9. Temperature dependence of the extracted features at 575 nm from the five dense layers of the LSTM model at (a) b1 and (b) b2.

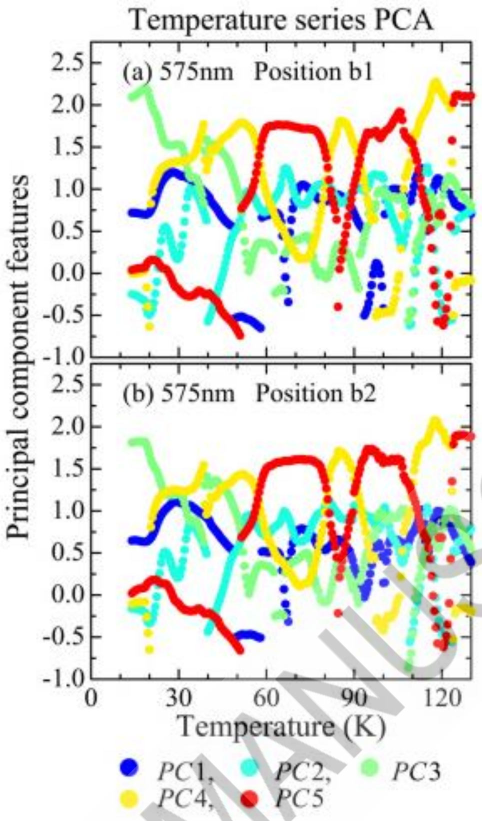

Figure 10. Results of temperature-series principal component analysis (TsPCA) at (a) b1 and (b) b2.

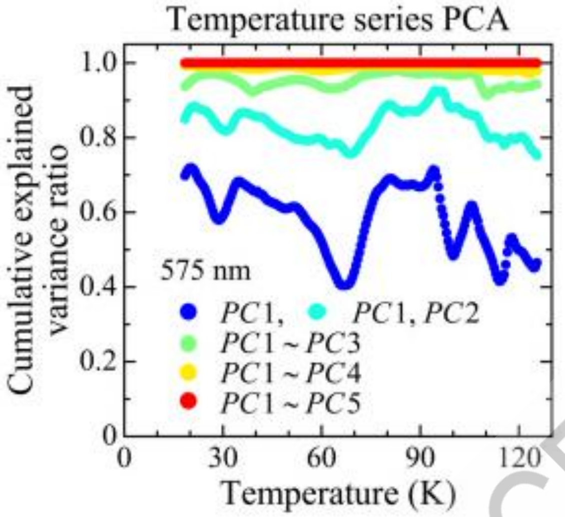

Figure 11. Cumulative explained variance ratio as a function of temperature obtained via TsPCA.

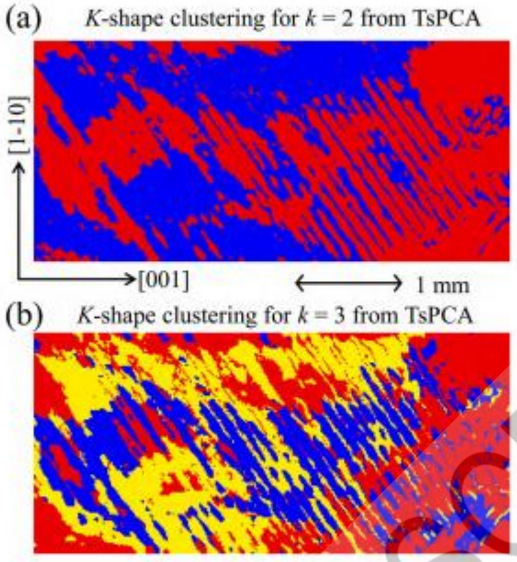

Figure 12. Results of K-shape multivariate clustering with 15 variables applied to a dataset (3 λ, 5 PCs, 233 T, 42,280 pixels) shown for (a) $k = 2$ and (b) $k = 3$.

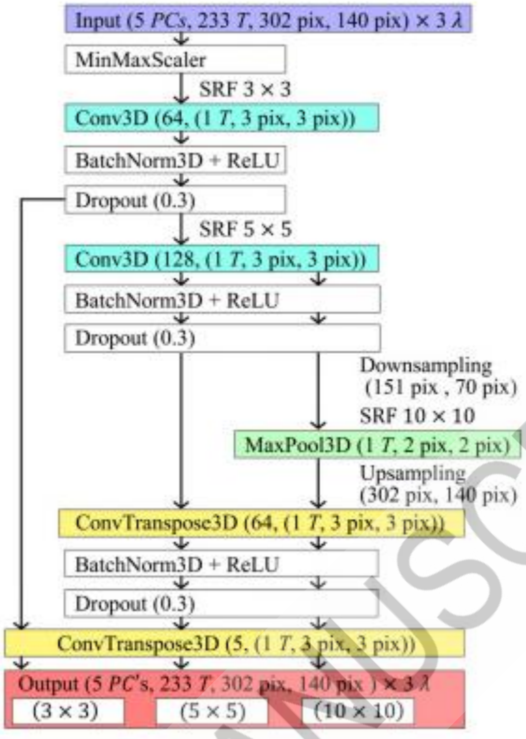
Input (5 PCs, 233 T, 302 pix, 140 pix) × 3 λ
MinMaxScaler
SRF 3 × 3
Conv3D (64, (1 T, 3 pix, 3 pix))
BatchNorm3D + ReLU
Dropout (0.3)
SRF 5 × 5
Conv3D (128, (1 T, 3 pix, 3 pix))
BatchNorm3D + ReLU
Dropout (0.3)
Downsampling
(151 pix , 70 pix)
SRF 10 × 10
MaxPool3D (1 T, 2 pix, 2 pix)
Upsampling
(302 pix, 140 pix)
ConvTranspose3D (64, (1 T, 3 pix, 3 pix))
BatchNorm3D + ReLU
Dropout (0.3)
ConvTranspose3D (5, (1 T, 3 pix, 3 pix))
Output (5 PC's, 233 T, 302 pix, 140 pix) × 3 λ
(3 × 3)
(5 × 5)
(10 × 10)

Figure 13. Flowchart of the 3D convolutional autoencoder (3DCAE). The input datasets reconstructed from the TsPCA results have dimensions of (5 PCs, 233 T, 302 pixels, 140 pixels) for three wavelengths (λ). These datasets are processed via convolutions with the spatial receptive field (SRF) sizes of 3×3, 5×5, and 10×10. The output datasets have dimensions of (5 PC's, 233 T, 302 pixels, 140 pixels) for three λ.

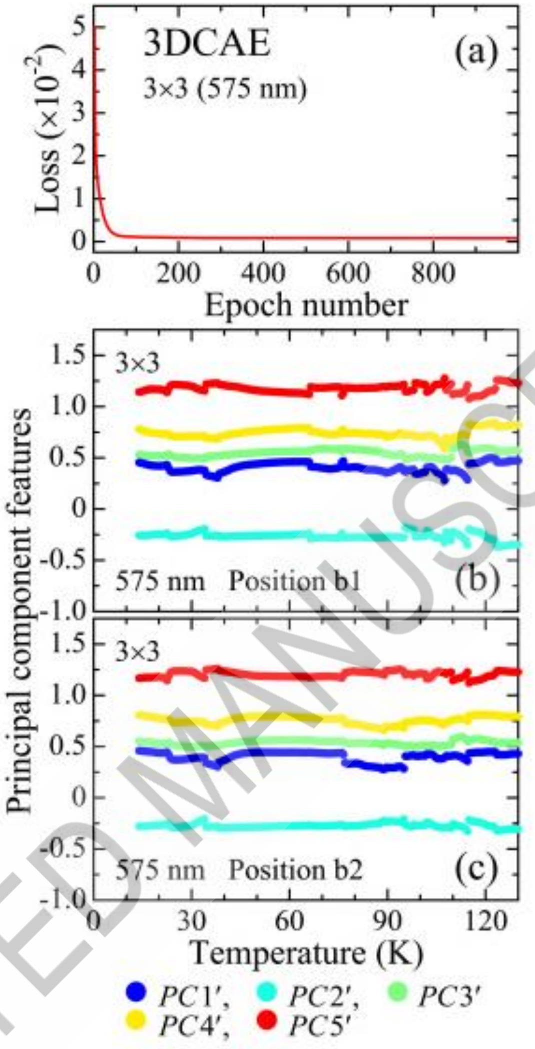

Figure 14. (a) Evolution of the loss function during the training of 3DCAE using the dataset (5 PCs, 233 T, 302 pixels, 140 pixels) at 575 nm for the SRF size of 3×3. Temperature dependence of the five principal components (PC's) predicted by the 3DCAE at (b) b1 and (c) b2.

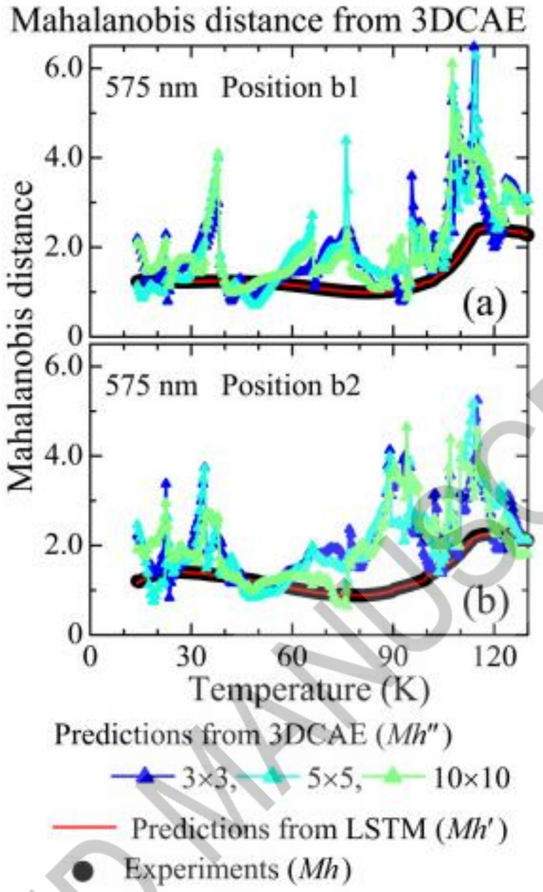

Figure 15. Temperature dependence of the Mahalanobis distance (Mh'') at 575 nm for different SRF sizes, derived from the 3DCAE predictions (PC's), at positions (a) b1 and (b) b2. Triangles represent the Mahalanobis distances (Mh'') obtained from the 3DCAE. As shown in Figures 8(b–c), red circles denote the experimental Mahalanobis distance data (Mh) and solid lines indicate the LSTM predictions (Mh').

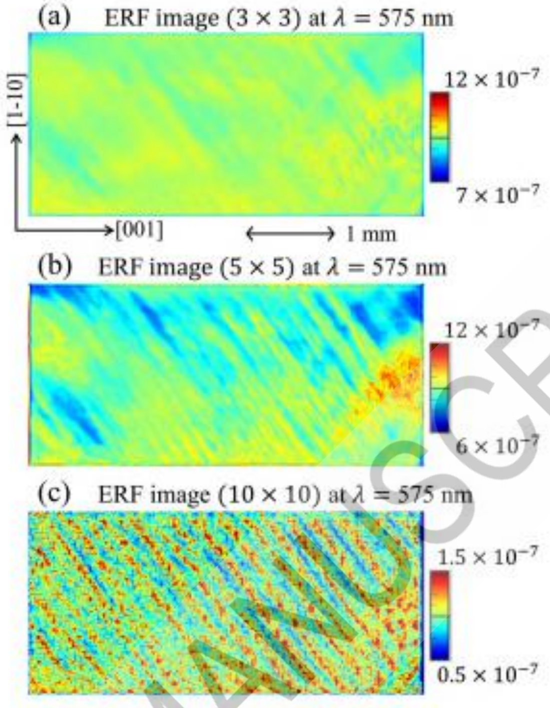

Figure 16. Comparison of effective receptive field (ERF) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

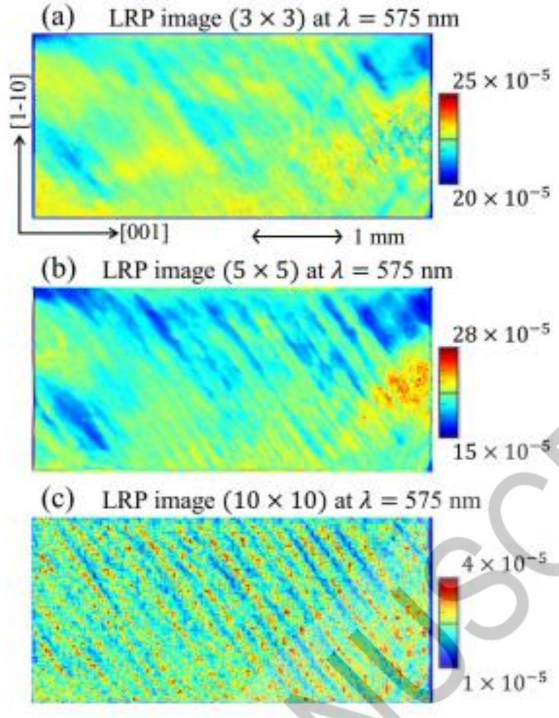

Figure 17. Comparison of layer-wise relevance propagation (LRP) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

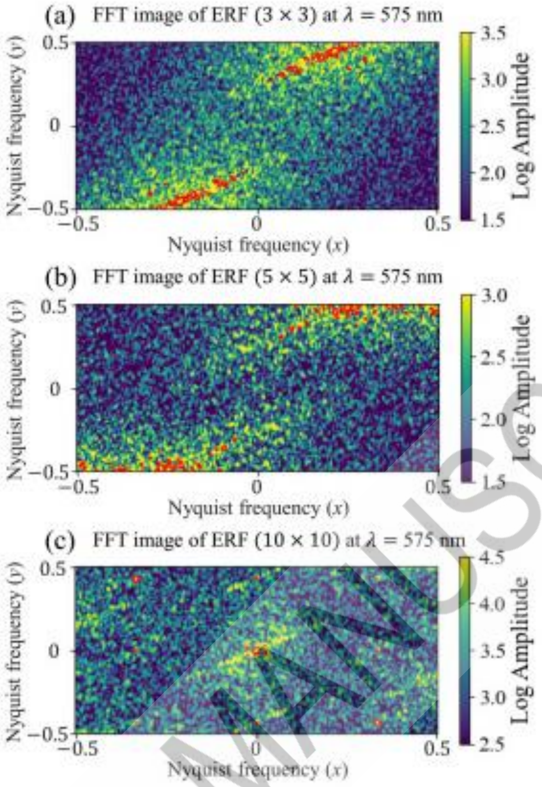

Figure 18. Comparison of the amplitude spectra of 2D fast Fourier-transform (FFT) ERF images at 575 nm for different spatial receptive field sizes: (a) 3×3, (b) 5×5, and (c) 10×10. Red circles indicate frequency components identified as significant at the 95% confidence level.

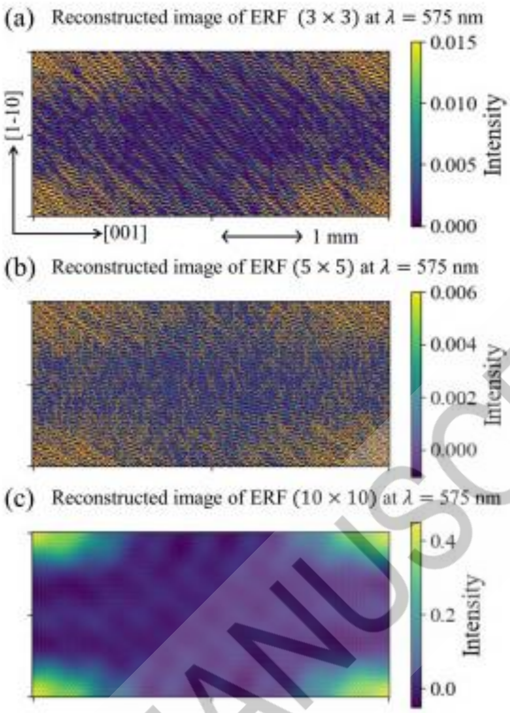

Figure 19. Comparison of inverse Fourier-transform ERF images at 575 nm for different spatial receptive field sizes: (a) 3×3, (b) 5×5, and (c) 10×10, using only the statistically significant frequency modes identified in Figure 18.

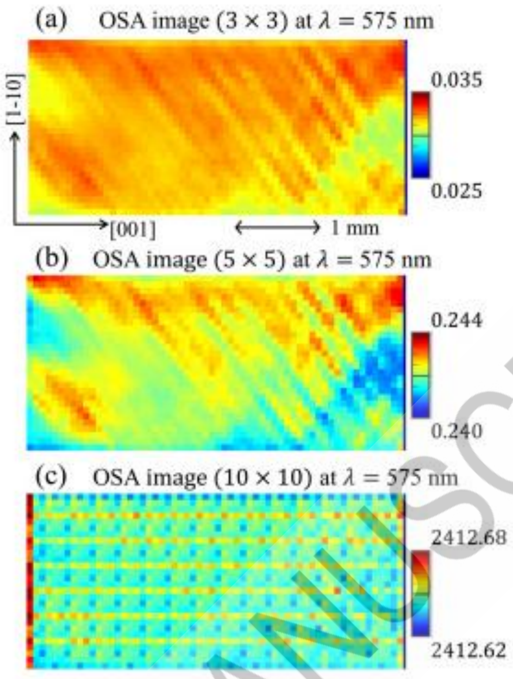

Figure 20. Comparison of occlusion sensitivity analysis (OSA) images for different SRF sizes at 575 nm: (a) 3×3, (b) 5×5, and (c) 10×10.

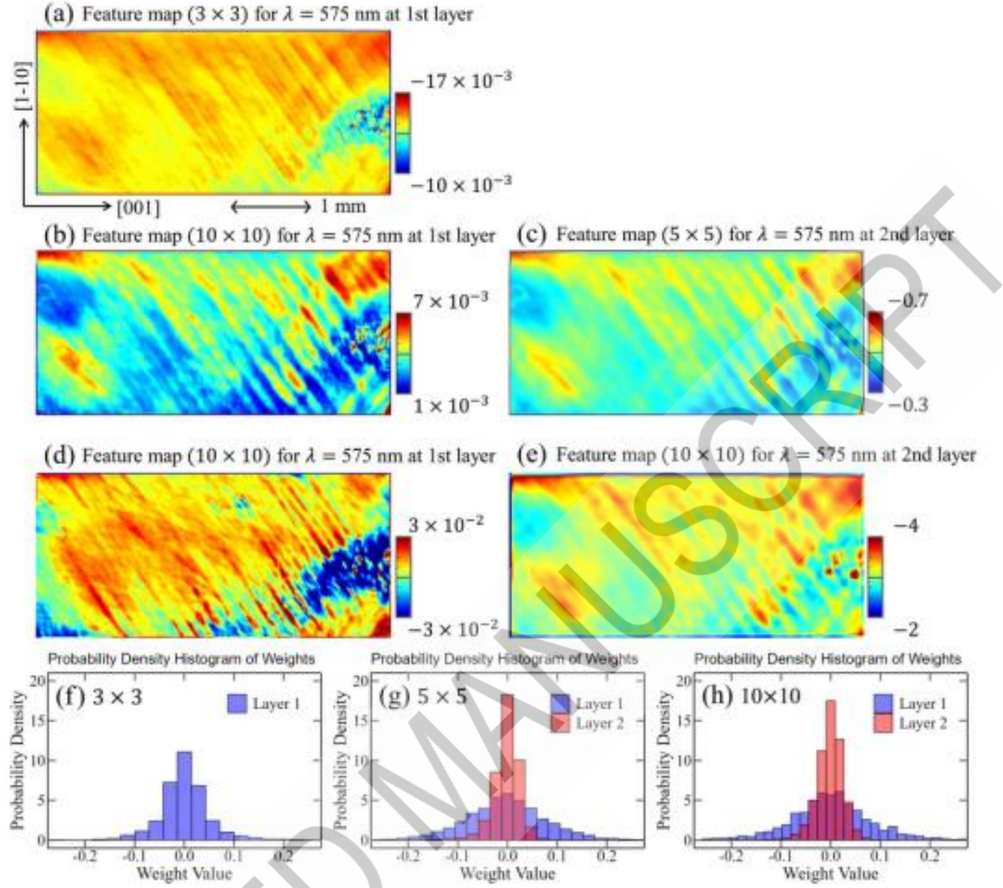

Figure 21. Comparison of feature maps at 575 nm for different SRF size sizes: (a) 3×3, (b–c) 5×5, and (d–e) 10×10. The 3×3 SRF size includes a single-layer structure, whereas the 5×5 and 10×10 SRF sizes include two layers. The histograms of feature weights are shown for (f) 3×3, (g) 5×5, and (h) 10×10.

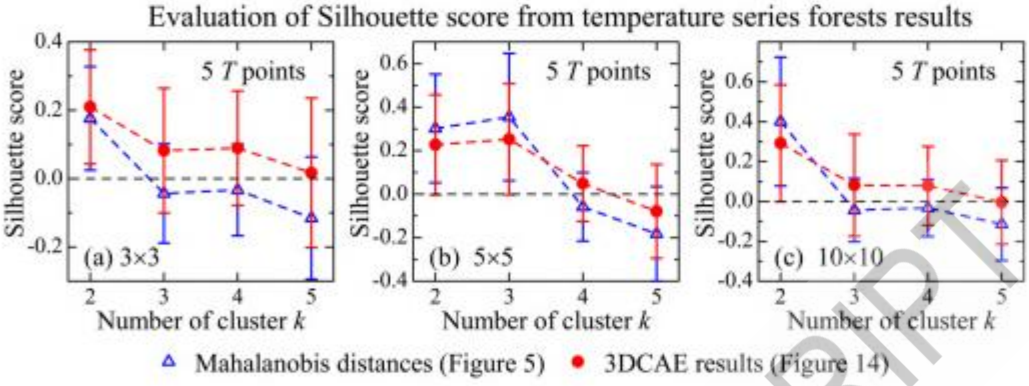

Figure 22. Silhouette scores for different initial numbers of clusters (k) computed using temperature-series forests (Tsf) for the SRF sizes of (a) 3×3, (b) 5×5, and (c) 10×10. The sliding temperature window width is fixed at 5 points (2.5 K).

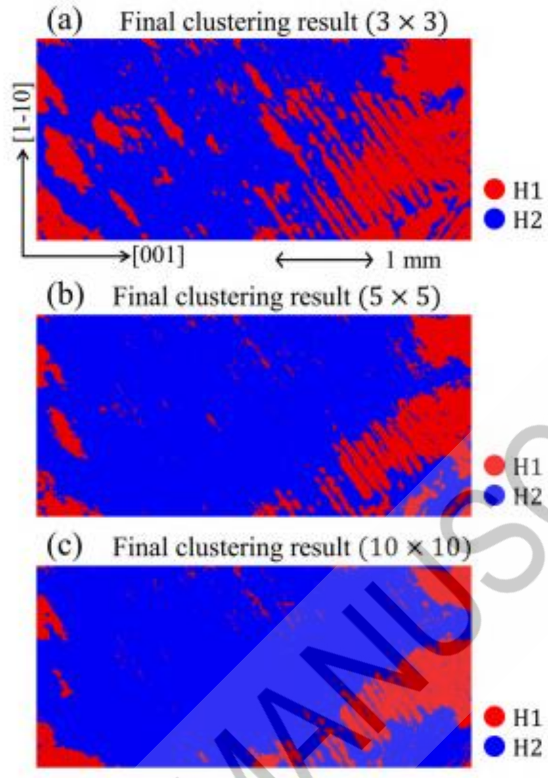

Figure 23. Final clustering results obtained via cluster confidence-based adaptive learning for different SRF sizes and initial cluster numbers (k): (a) 3×3 with $k = 2$, (b) 5×5 with $k = 3$, and (c) 10×10 with $k = 2$. These results highlight the effect of SRF sizes on the classification of polarization states.

Novelty

A deep learning framework is developed to enable clustering that accounts for overlapping polarization components in polarized light microscopy. This approach reconstructs intrinsic birefringence features in the stress-induced ferroelectric $SrTiO_3$.