Supporting Information

Integration of Microstructural Image Data into Machine Learning Models for Advancing High-Performance Perovskite Solar Cell Design

Haotian Liu[a#], Antai Yang[a#], Chengquan Zhong[a#], Xu Zhu[a], Hao Meng[a], Zhuo Feng[a], Jixin Tang[a], Chen Yang[c], Jingzi Zhang[a,b*], Jiakai Liu[c,d*], Kailong Hu[a*], and Xi Lin[a*]

[a]*School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, China*

[b]*School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China*

[c]*Xinjiang Key Laboratory of Separation Material and Technology, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China*

[d]*Center of Materials Science and Opto-electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China*

Corresponding authors

*E-mail: linxi@hit.edu.cn

*E-mail: hukailong@hit.edu.cn

*E-mail: liujk@ms.xjb.ac.cn

*E-mail: zjzhang@hit.edu.cn

[#]These authors contributed equally to this work.

1. Perovskite thin film microstructure feature extraction network

ResNet is employed to extract microstructure feature vectors. Specifically, the feature extraction network for SEM images is refined based on the ResNet18 architecture. The original network has five stages, each containing two residual blocks. As ResNet18 is designed for the large-scale, multi-class ImageNet dataset, the model's depth and width are set relatively large. However, smaller datasets can result in slow training speeds and overfitting. Therefore, we adjusted the network structure to suit the task of material performance prediction. We removed the original fourth stage and set the channels of the first four stages to 16, 32, and 64. Fusion of features from each channel was performed using 1×1 convolutional layers, resulting in 32 high-level semantic feature maps reflecting material microstructure information. Finally, an SEM image feature vector was obtained using the GAP method.

ResNet18 is composed of multiple convolutional layers with different numbers of convolutional kernels. Convolution operations at different depths can extract features at different levels, while multiple convolutions with varying widths can extract multiple types of features at the same level.[1-3] In the early layers of the network, the feature maps retain more spatial information, which helps to understand the local structure of the image. As the network deepens, the spatial resolution of the feature maps decreases, but the abstraction of the features increases.[4-6] In this study, fine spatial information from the lower layers and abstract features from the higher layers are combined through skip connections to improve model performance. Assuming the size of the acquired Feature Map is $h \times w \times c$, as expressed in Eqs. (1) where h and w are the height and width of a single feature map, and c is the number of channels. Let v_l^k represent the k-th element interval of SEM images feature vector (v1), and let feature map kj denote the value at row i and column j of the k-th feature map, the interval can be expressed by the following formula:

$$v_1^k = \frac{1}{h \times w} \sum_{i=1}^{n} \sum_{j=1}^{w} FeatureMap_{ij}^k \quad (1)$$

2. Composition and processing feature extraction network

The material composition and processing feature extraction network for PSC

materials is designed to analyze and interpret the battery's material composition information and the nuances of the fabrication process. This network is proficient at deciphering the complex interplay between the components that make up the PSC layers. The network begins by assimilating the battery material composition data, which is then systematically directed through a series of fully connected layers. Specifically, the data passes through layers with 16, 20, and 32 neurons, respectively.[18] This step aims to optimize the extraction of features related to the material composition. Incorporating the ReLU activation function after the initial layers enhances the network's ability to extract meaningful insights from SEM images. In the subsequent stage, the network shifts its focus to the fabrication process of PSCs. Here, the traditional one-hot encoding is eschewed in favor of an Embedding module, which adeptly translates the experimental parameters into a dense vector. This pivotal transformation facilitates a more nuanced distance measurement between different process parameters. The data then proceeds through additional fully connected layers, each designed to refine further and capture the intricacies of the fabrication process. The network's final output is a feature vector (v3), which encapsulates the essence of the fabrication process characteristics. This vector and the material composition feature vector (v2) form the cornerstone of the predictive model, providing a robust foundation for predicting PSCs performance regression. The program is developed using the PyTorch deep learning framework and runs on a Windows 10 operating system. The hardware setup comprises an Intel Core i9-10900K processor and an NVIDIA GTX 3090 graphics card, which provide the computational capabilities needed for training the model effectively.

3. The procedure for extracting and calculating SEM feature factors

The SEM feature factors were manually collected following a specific procedure. First, the image labeling and selection were set using ImageJ software (version 1.53) based on the scale bar in the SEM images through the “Set Scale” function. Next, grains were segmented using the “Color Threshold” function. The “Analyze Particles” function was then used to automatically extract the perimeter and area of each grain,

while the lengths of the major and minor axes were manually measured using the straight-line tool on the segmented grains.

The grain size indicators include the equivalent circular diameter and major axis length. The equivalent circular diameter is a core indicator of grain size, used to describe the average size of particles. It simplifies complex, irregularly shaped grains by assuming each particle is a circle with the same area, thus converting them into circular grains with equal area. The formula for calculating the equivalent circular diameter is:

$$D_{eq} = 2\sqrt{\frac{A}{\pi}} \tag{2}$$

A represents the actual area of the grain (μm²). Larger grain sizes typically reduce the number of grain boundaries, thereby decreasing the recombination probability of charge carriers at the interfaces and improving the optoelectronic conversion efficiency.[19] Additionally, consistency in grain size helps enhance the film's compactness, reduce internal defects, and improve the stability of battery performance. The grain area can be directly measured by analyzing the SEM images of perovskite films, and the equivalent circular diameter can be further calculated. Studies have found that in high-efficiency PSCs, the average equivalent circular diameter of the grains is usually larger and more uniformly distributed.

The major axis length refers to the length of the long axis of the ellipse that best fits the grain and can characterize the morphological features of the grains. The major axis length not only reflects the degree of elongation of the grains but also reveals the orientation characteristics of the grains within the film. A longer major axis may indicate that the crystal has preferential growth in a certain direction, which is closely related to the deposition process and material properties. The distribution of the major axis length can influence the packing density and light scattering properties of the film, thus indirectly affecting its optoelectronic performance. Through SEM image analysis, the major axis length can be automatically extracted, and the uniformity of the film can be accessed via statistical distribution. For high-performance perovskite films, the

distribution of the major axis length should be as uniform as possible to avoid the presence of abnormal grains.

GBLD is used to measure the total length of grain boundaries per unit area of the film. Its calculation formula is:

$$GBLD = \frac{\sum P}{A_{film}} \tag{3}$$

$\sum P$ represents the total perimeter of all grains in the analyzed region, and A_{film} represents the total area of the SEM image. A higher GBLD typically indicates a greater number of grain boundaries, increasing the likelihood of charge carrier recombination. Conversely, lower GBLD improves charge transport properties and reduces recombination losses. Additionally, films with more regular grain shapes can reduce internal stress distribution, enhancing mechanical stability. By extracting the grain perimeters and total image area from SEM images, the distribution of grain boundary length density can be calculated. High-efficiency perovskite films often aim to reduce GBLD to optimize optoelectronic performance.

The aspect ratio is the ratio of the grain's major axis to its minor axis and is used to measure the elongation of the grain. The calculation formula is:

$$Aspect\ ratio = \frac{Major\ axis}{Minor\ axis} \tag{4}$$

The aspect ratio reflects whether the grains tend to directional growth, which may influence the optoelectronic properties of the film. An appropriate aspect ratio can also improve the surface coverage of the film, reduce defects, and enhance the overall performance of the solar cell. By analyzing the statistical distribution of the aspect ratio from SEM images, the processing conditions for film preparation can be further optimized to improve grain morphology and performance.

The standard deviation (SD) the equivalent circular diameter describes the uniformity of the grain size distribution. The calculation formula is:

$$\sigma_D = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \mu_D)^2} \tag{5}$$

D_i represents the equivalent circular diameter of a single grain (μm); μ_D represents the average value (μm); N represents the total number of grains. Here, represents the equivalent circular diameter of a single grain, represents the average value, and represents the total number of grains. A small standard deviation indicates that the grain size distribution is more uniform, which helps to improve the continuity and carrier transport properties of the film. At the same time, a low standard deviation in grain distribution usually reflects the stability of the film preparation process. By image analysis, the standard deviation of the equivalent circular diameter of the grains can be statistically obtained and used to assess the quality of the film. In high-performance perovskite films, the standard deviation of grain size is usually small, and this uniformity ensures the compactness and performance stability of the film.

The SD of aspect ratio is used to evaluate the uniformity of grain shape, and its calculation formula is:

$$\sigma_{\text{Aspcect ratio}} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{AR_i} - \mu_{\mathrm{AR}})^2} \tag{6}$$

$\mathrm{AR_i}$ represents the aspect ratio of a single grain, and μ_{AR} represents the average value. Grains with a uniform aspect ratio distribution can reduce defects in the film, increase the packing density, and thereby improve the optoelectronic performance. A low SD of aspect ratio indicates a high consistency in grain shape, which is particularly important for high-efficiency perovskite films. By statistical analysis of the aspect ratio from SEM images, the morphological characteristics of the film can be revealed, guiding the optimization of the film preparation process. These SEM image features play a key role in the analysis of perovskite film SEM images. They not only quantitatively describe the morphological characteristics of the grains but also provide important references for optimizing optoelectronic performance. These feature factors provide high-quality input data for machine learning models, significantly enhancing the interpretability and predictive capability of the models.

4. Convolutional Neural Network

Convolutional Neural Network (CNN) is a deep learning model that is particularly suitable for processing image data. It consists of multiple convolutional layers, activation layers, pooling layers, and fully connected layers[2]. The convolutional layer slides over the input image using small filters that capture local features such as edges and texture. The parameters of these filters are automatically learned during the training process, enabling automatic feature extraction. The activation layer typically employs the ReLU function, which introduces nonlinearities that allow the network to learn more complex patterns[3]. The pooling layer, on the other hand, reduces the computational effort by decreasing the dimensionality of the feature map and increasing the invariance to displacement. In the deeper layers of the network, features are gradually abstracted from low to high level and eventually mapped to the output through a fully connected layer.

5. Attention mechanisms

Attention mechanisms are an important technique in deep learning that allows a model to focus on the most important parts of the input data when processing information. The frequency domain channel attention mechanism is an attention mechanism that operates in the frequency domain, and it is commonly used in image processing and audio analysis. This mechanism enhances important information and suppresses unimportant information by transforming the input data into the frequency domain (e.g., using a Fourier transform) and then weighting channels of different frequencies in the frequency domain. This approach captures signal properties at different frequencies and improves the sensitivity of the model to specific frequencies. Self-attention mechanism, also known as the internal attention mechanism, is a method that allows the model to focus on information at different locations within the sequence while processing the sequence data. It generates an attention weight matrix by calculating the relationship of each element in the sequence to the other elements and then uses this weight matrix to weight and sum to get the output of each element. The self-attention mechanism is particularly effective when dealing with long sequential data because it captures the long distance dependencies within the sequence. This

mechanism is widely used in the Transformer model and is one of the key techniques for realizing long-range dependency capture[6].

6. Regression model and classification model

The training parameters are set with a batch size of 32, and a learning rate of 0.001, and the ADAM optimizer is utilized. The training is conducted over 30 epochs using the mean squared error (MSE) loss function. We used a fully connected layer of MLP containing 32 neurons as a regression predictor, whose main function is to output PCEs such as power conversion efficiency of materials based on semantic vector, combined with microstructure, material composition, and processing technology. As expressed in Eqs. (3) suppose x_1, x_2, x_3 represent SEM image, material composition, and processing process, respectively; g_1, g_2, g_3 represent microstructure feature extraction network, material composition feature extraction network, and processing process feature extraction network, respectively; σ represents the module adaptive Fusion; f represents the regression predictor; and P represents the battery performance, i.e., PCE. Then, the process of battery performance prediction can be expressed using the following equation.

$$P = f(\sigma(g_1(x_1), g_2(x_2), g_3(x_3))) \tag{7}$$

The classification algorithm is based on a convolutional neural networks (CNN) model and employs the Visual Geometry Group (VGG) architecture. Within the VGG framework, each convolutional layer employs small-sized kernels combined with ReLU activation functions to extract detailed features from the images. Successive convolutional and max-pooling operations progressively extract and integrate these features, which are then flattened via fully connected layers. The Softmax layer is used to predict the probability distribution of grain sizes.

The model's input solely consists of annotated perovskite thin film scanning electron microscope (SEM) images. To validate the effectiveness of CNN in feature extraction, 868 SEM images with varying grain sizes were annotated, and the equivalent circle diameter of each image was measured. The images were classified into three categories based on their equivalent circle diameter: small size (less than 0.4 μm, 289 images), medium size (0.4–0.8 μm, 305 images), and large size (greater than

0.8 μm, 274 images). The primary objective of the classification model is to validate the effectiveness of the image feature extraction model within the multimodal fusion framework presented in this work.

7. Adaptive feature fusion

The PCE of PSCs is influenced by various factors such as material composition, processing techniques, and microscopic structure of the materials. The more information used in establishing regression prediction models, the more comprehensive the mapping relationship becomes. This paper comprehensively investigates these three crucial factors, aiming to enhance the completeness of the model's mapping relationship and improve predictive accuracy. By integrating these three types of information, SEM images feature vector, material composition feature vector, and processing feature vector are obtained through the three respective branches of the network. These feature vectors represent information in three modes, each affecting material properties differently. Therefore, an adaptive feature vector fusion module based on an attention mechanism is designed to merge these diverse feature vectors effectively.

In this module, integrated features a_i are computed through three independent fully connected layers based on v_i. Subsequently, the weight coefficients w_i are obtained through the Sigmoid function. These coefficients determine the influence of microstructure information, material composition information, and processing technology information on the final predictive results of the model. The weight coefficients can be adaptively learned and adjusted during the model training process. Suppose the size of the feature vector v_i for certain mode information of the PSCs is 1×32. The Linear layer and weight acquisition function used to obtain different mode semantic vectors can be expressed as:

$$a_i = v_i W_i + b_i \tag{8}$$

As expressed in Eqs. (8) where w_i and b_i are learnable parameters, and the weight coefficients of different modal information are mainly calculated based on these parameters. We fused the different modal information after mapping them to a dimensionally fixed common semantic space.

The effects of the microstructure of the perovskite film, the material composition

of the PSCs, and processing on the performance of PSCs vary. The most intuitive way to directly add feature vectors of different modes in the PCE implies that we assign the same influence weights to the three different types of information, which is unreasonable. However, it is also not known what the influence weights of the different modal information are on the PCE. Adaptive Fusion can learn on its own. The method automatically learns the importance of different modal information through three independent fully connected layers using the Sigmoid function to achieve weighted fusion and get higher quality fused feature vectors, thus improving the prediction performance of the model.

8. Materials

ITO glass substrates, Formamidinium Iodide (FAI), Methylammonium chloride (MACl), and PbI_2 were purchased from Advanced Electron Technology Co., Ltd. Methylammonium bromide (MABr), Methylammonium acetate (MAAc), Formamidine acetate (FAAc), and Spiro-OMeTAD were provided by Xi'an Polymer Light Technology Co. BMIMAc (1-butyl-3-methylimidazolium acetate) was sourced from Shanghai Aladdin Biochemical Technology Co., Ltd. Lithium bis(trifluoromethane)sulfonimide (Li-TFSI, 99%), 4-tert-butylpyridine (tBP, 96%), and CsI (99.99%) were obtained from Sigma-Aldrich. Tin(IV) oxide (SnO2) was purchased from Alfa Aesar. Solvents DMF, DMSO, IPA, and acetonitrile were acquired from Aladdin. All chemicals were used as received without further purification.

9. Device Fabrication

ITO substrates were cleaned by sonication for 30 minutes in detergent, deionized water, acetone, and isopropanol, respectively. After cleaning, the substrates were treated with ultraviolet ozone for 30 minutes. A SnO_2 electron transport layer was deposited by spin-coating a SnO_2 colloidal solution in water at 4000 rpm for 30 seconds, followed by annealing at 150°C for 30 minutes on a hotplate. A two-step sequential method was used to fabricate the perovskite layer in a glovebox. In the first step, a 1.5 M PbI_2 solution with BMIMAc was prepared by dissolving in 950 μL DMF, with 50 μL CsI solution (390 mg CsI in 1 mL DMSO) added. The solution was spin-coated onto the SnO2/ITO substrate at 1500 rpm for 30 seconds and dried at 70°C for 1 minute. In

the second step, a solution containing FAI, MABr, and MACl (90 mg, 9 mg, and 9 mg, respectively, in 1 mL IPA) was dropped onto the PbI_2 film and spin-coated at 2000 rpm for 30 seconds. The perovskite film was then annealed at 150°C for 15 minutes under 30-40% relative humidity. The spiro-OMeTAD solution was prepared by dissolving 72.3 mg spiro-OMeTAD, 30 μL TBP, and 35 μL Li-TFSI solution (260 mg in 1 mL acetonitrile) in 1 mL chlorobenzene. The solution was spin-coated onto the perovskite film at 4000 rpm for 30 seconds. Finally, a 100 nm gold electrode was thermally evaporated onto the device.

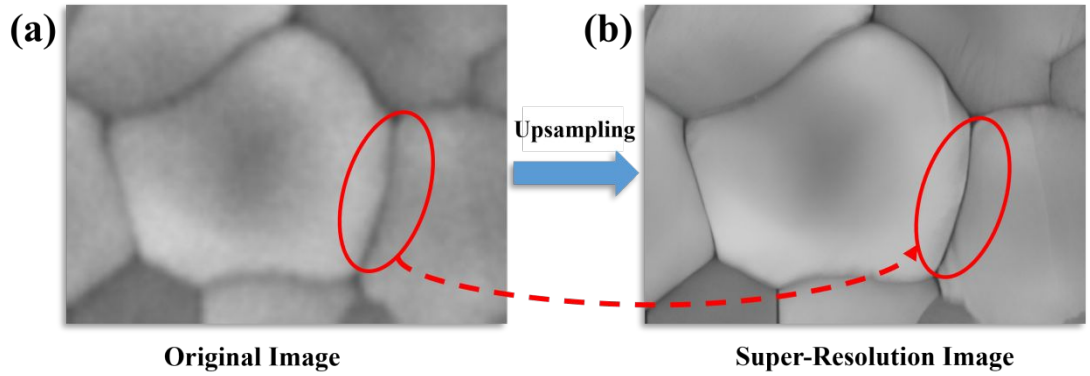

Figure S1 Comparison of original and super-resolution images.

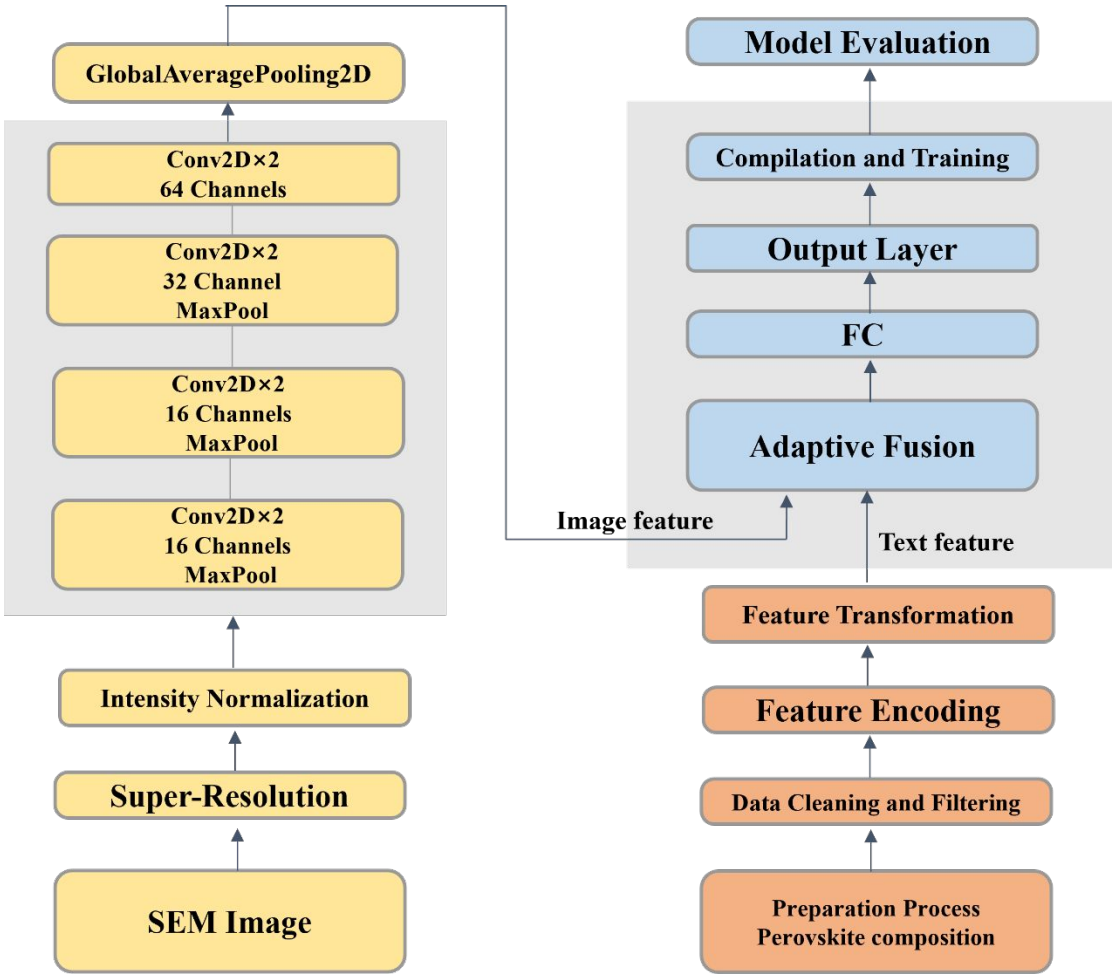

Figure S2. Structure of the multimodal fusion material property prediction model.

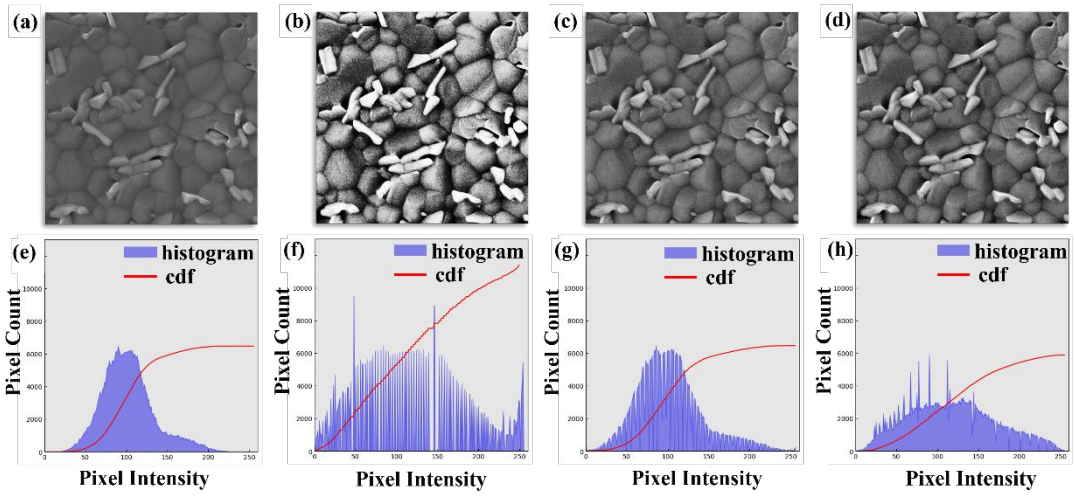

Figure S3. Grayscale images and histograms: (a), (e) the original version; b), (f) the version after a min-max normalization; (c), (g) the version after a Z-score normalization; (d), (h) the version after an adaptive histogram normalization.

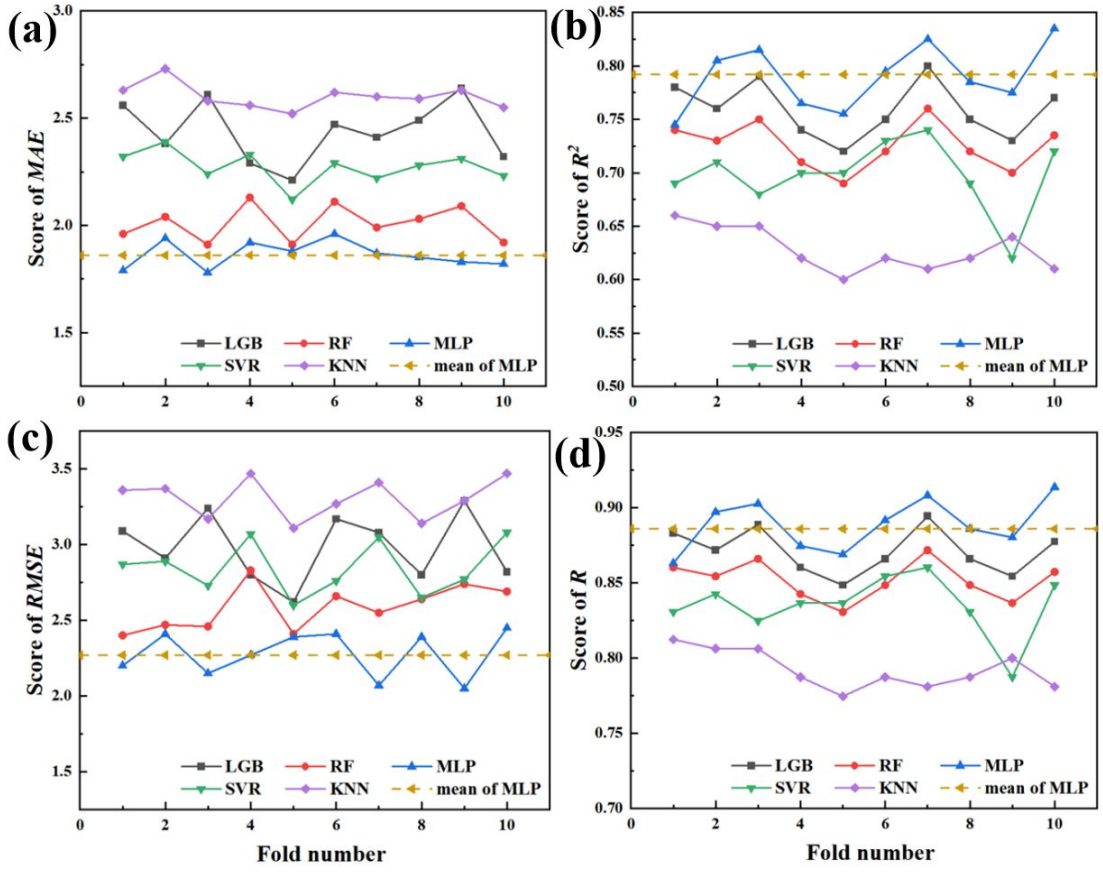

Figure S4. The MAE, R^2, RMSE, and R scores of the five models during the ten-fold cross-validation process: (a) MAE, (b) R^2, (c) RMSE, (d) R.

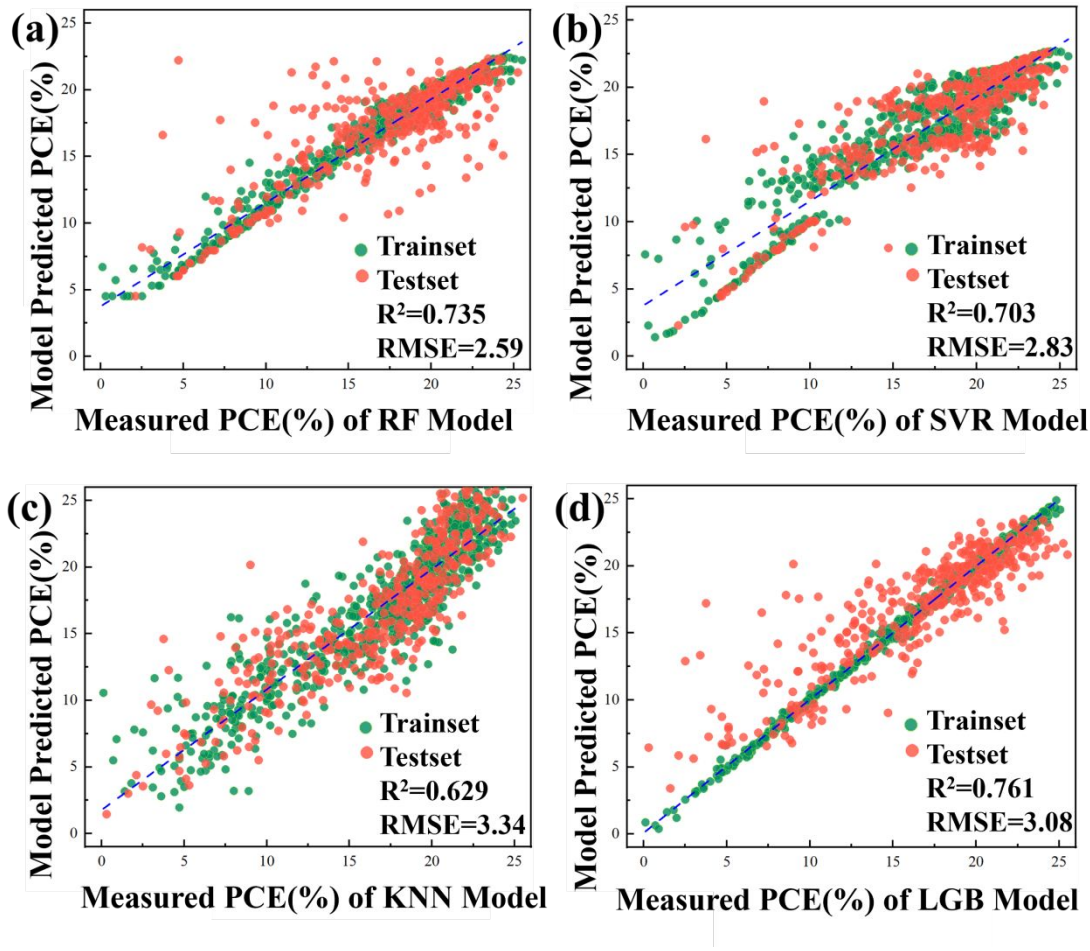

Figure S5. (a) Fitted PCE results of the random forest model. (b) Fitted PCE results of the SVR model. (c) Fitted PCE results of the KNN model. (d) Fitted PCE results of the LGB model. All figures use green to represent the training set and red to indicate the test set.

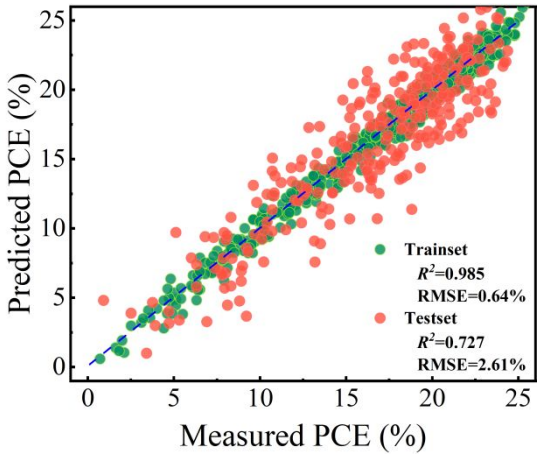

Figure S6. Fitted PCE results of the LGB model using feature factors, material composition features, and processing features as inputs.

Table S1 Dataset of PSCs material compositions and manufacturing processes.

Feature type	Compositions and manufacturing parameters
Perovskite type	$MAPbI_3$, $MAPbBr_3$, $MAPbBr_{3-x}Cl_x$, $MAPbI_{3-x}Br_x$, $MAPbI_{3-x}Cl_x$, $MASn_{1-x}Pb_xI_3$, FA based, Cs based, Sn based, mixed cation, mixed halide perovskite (MA=CH_3NH_3, FA=$CH(NH_2)_2$)
Perovskite deposition process	One-step/two-step
Perovskite deposition method	spin, spin 2–3, spin-dip, spin 2–3-dip, dip coating, VASP, CVD, evaporation-spin, spin-spray, spray, spin-dripping
Precursor solution	DMF, DMA, DMF+DMSO, DMF+CHP, DMF+DIO, DMF+GBL, DMF+H_3P, DMF+HI, DMSO, DMSO+GBL, DMSO+SnF_2, GBL, 2-methoxyethanol+CHP
Antisolvent	Chlorobenzene, toluene, diethyl ether, trifluorotoluene, ethyl acetate, ethanol, no antisolvent treatment
Electron transport layer	C_{60}, SnO_2, TiO_2, TiO_2-doped, ZnO, ZnO-doped, Fe_2O_3, graphene, PCBM, doped PCBM, PCBM+C_{60}, no electron transport layer
Hole transport layer	Spiro-OMeTAD, CuSCN, P3HT, PEDOT, PTAA, SWNT, no hole transport layer
HTL additive	LiTFSI+TBP, LiTFSI+TBP+FK102, LiTFSI+TBP+FK209, LiTFSI, LiTFSI+2,6-lutidine, no additive

Table S2. The PSNR values for each algorithm in the super-resolution reconstruction of perovskite solar cell SEM images.

Scale	Bicubic Interpolation (BICUBIC)	Super-Resolution Convolutional Neural Network (SRCNN)	Residual Channel Attention Network (RCAN)	Stable Attention Network (SAN)	Swin Transformer for Image Restoration (SWINIR)
2	36.84	37.48	38.14	38.31	**39.16**
3	33.91	34.33	35.20	36.27	**36.43**
4	33.07	33. 27	34.13	34.33	**35.64**

Table S3. The SSIM values for each algorithm in the super-resolution reconstruction of perovskite solar cell SEM images.

Scale	BICUBIC	SRCNN	RCAN	SAN	SWINIR
2	0.8684	0.8804	0.8928	0.8860	**0.9308**
3	0.8233	0.8112	0.8476	0.8579	**0.8607**
4	0.7875	0.8082	0.8174	0.8265	**0.8489**

Table S4. Details of SEM images feature extraction network.

Name	Operation	Output size (W×H×C)
Stage 1	Conv 3×3/16 Max pool 2×2, stride 2 CBAM	112×112×16
Stage 2	Conv 3×3/16 Conv 3×3/16 Max pool 2×2, stride 2	56×56×16
Stage 3	Conv 3×3/32 Conv 3×3/32 Max pool 2×2, stride 2	28×28×32
Stage 4	Conv 3×3/64 Conv 3×3/64 Max pool 2×2, stride 2	14×14×64
Compression	Conv 1×1/32	14×14×32
Output	GAP	1×1×32

Table S5. Convolutional Neural Network Formulas and Concepts.

Formula Type	Mathematical Expression	Description
Convolution	$f \times g \times x = \int f(t)g(x-t)dt$	The convolution operation is implemented through integration to achieve a weighted sum.
Convolution Output	$O_{ij} = (I \times K)_{ij} = \sum_m \sum_n I_{m+i,n+j} K_{mn}$	The output of a convolution operation between an input image I and a kernel K .
Activation Function	ReLU(x)=max(0,x)	The Rectified Linear Unit activation function.
Fully Connected Layer	Z=W×A+b	The linear transformation of

		inputs A with weights W and bias b in a fully connected layer.
Loss Function	$L = -\sum_{1}^{C} y_{oc} \log(p_{oc})$	The Cross-Entropy Loss function used for classification tasks in neural networks.
Weight Update	$W = W - \eta \frac{\partial L}{\partial W}$	The update rule for adjusting the weights W in the direction that minimizes the loss L.

Table S6. Detailed parameters of ML models.

ML model name	Parameters
MLP	Solver: 'adam' Alpha: 0.0001 Learning_rate: 'adaptive' Max_iter: 200 Random_state: 2024
SVR	Degree: 3 C: 300 Gamma: scale
LGB	N_estimators: 1000 Num_leaves: 25 Max_depth: -1 Subsample: 0.6 Random_state: 2024
RF	N_estimators: 500 Max_depth: 12 Random_state: 2024
KNN	N_neighbors: 5 Weights: uniform Leaf_size: 40

Table S7. Detailed parameters of the CNN model.

ML model name	Parameters
CNN	Convolutional Channels: [16, 16, 32, 64] Pooling Layers: [2×2 max pooling with stride 2] Fully Connected Layers: [512, 512, 1000] Activation Function: [ReLU] Initial Learning Rate: [0.005] Momentum: [0.9] Weight Decay: [5×10^{-4}] Dropout Probability: [0.4] Batch Size: [256]

References

[1] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 33 (2021) 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

[2] O'shea K, Nash R. An introduction to convolutional neural networks, arXiv:1511.08458 [cs], 2015. https://arxiv.org/abs/1511.08458.

[3] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377. https://doi.org/10.1016/j.patcog.2017.08.012.

[4] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62. https://doi.org/10.1016/j.neucom.2020.07.110.

[5] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey, Comput. Vis. Media 8 (2022) 331–368. https://doi.org/10.1007/s41095-022-0271-y.

[6] Yang X. An overview of the attention mechanisms in computer vision, J. Phys. Conf. Ser. 1693 (2020) 012173. https://doi.org/10.1088/1742-6596/1693/1/012173.