PAPER • OPEN ACCESS

Adapted Swin Transformer-based real-time plasma shape detection and control in HL-3

To cite this article: Qianyun Dong *et al* 2025 *Nucl. Fusion* **65** 026031

OPEN ACCESS
IOP Publishing | International Atomic Energy Agency
Nuclear Fusion
Nucl. Fusion **65** (2025) 026031 (15pp)
https://doi.org/10.1088/1741-4326/ada2fe

Adapted Swin Transformer-based real-time plasma shape detection and control in HL-3

Qianyun Dong[1], Zhengwei Chen[2], Rongpeng Li[1,*], Zongyu Yang[2], Feng Gao[3], Yihang Chen[2], Fan Xia[2], Wulyu Zhong[2,*] and Zhifeng Zhao[3]

[1] Zhejiang University, Hangzhou 310058, China
[2] Southwestern Institute of Physics, Chengdu 610043, China
[3] Zhejiang Lab, Hangzhou 311500, China

E-mail: lirongpeng@zju.edu.cn and zhongwl@swip.ac.cn

Received 30 June 2024, revised 17 December 2024
Accepted for publication 24 December 2024
Published 7 January 2025

Abstract

In the field of magnetic confinement plasma control, the accurate feedback of plasma position and shape primarily relies on calculations derived from magnetic measurements through equilibrium reconstruction or matrix mapping method. However, under harsh conditions like high-energy neutron radiation and elevated temperatures, the installation of magnetic probes within the device becomes challenging. Relying solely on external magnetic probes can compromise the precision of EFIT in determining the plasma shape. To tackle this issue, we introduce a real-time, non-magnetic measurement method on the HL-3 tokamak, which diagnoses the plasma position and shape via imaging. Particularly, we put forward an adapted Swin Transformer model, the Poolformer Swin Transformer (PST), to accurately and fastly interpret the plasma shape from the Charge-Coupled Device Camera images. By adopting multi-task learning and knowledge distillation techniques, the model is capable of robustly detecting six shape parameters under visual interference conditions such as bright light from the divertor and gas injection, thereby avoiding cumbersome manual labeling. Specifically, the well-trained PST model capably infers R and Z within the mean average error below 1.1 cm and 1.8 cm, respectively, while requiring less than 2 ms for end-to-end feedback, an 80% improvement over the smallest Swin Transformer model, laying the foundation for real-time control. Finally, we deploy the PST model in the Plasma Control System using TensorRT, and achieve 500 ms stable PID feedback control based on the PST-computed horizontal displacement information. In conclusion, this research opens up new avenues for the practical application of image-computing plasma shape diagnostic methods in the realm of real-time feedback control.

* Authors to whom any correspondence should be addressed.

Keywords: shape detection and control, real-time, Swin Transformer, HL-3 tokamak

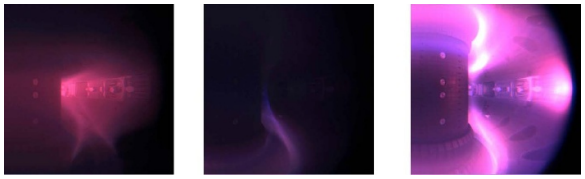
(Some figures may appear in colour only in the online journal)

1. Introduction

For magnetic confinement fusion devices like tokamak [1], accurate measurement of plasma shape is one of the most preliminary steps toward subsequent diagnostics and control. Currently, most measurement techniques like magnetic equilibrium reconstruction code EFIT are largely contingent on magnetic field sensors like probes, the accuracy of which is largely affected by the applicability of magnetic sensing [2]. Long plasma discharges can suffer from drift of the magnetic signals due to the integral nature of these measurements [3]. Similarly, plasmas with a low plasma current and large distance to the magnetic pick-up coils (for example during the ramp-up and ramp-down phase of the discharge or ITER first plasma) may result in weak magnetic signals [4]. Meanwhile, under severe conditions like high-energy neutron radiation and high temperatures, the performance significantly deteriorates and could even completely malfunction due to heat shock and magnetic forces. Therefore, to satisfy the requirements in future extremely high-temperature environments like ITER, developing an alternative method for real-time, non-magnetic measurement is highly desired [5].

Correspondingly, the optical plasma boundary reconstruction for plasma position control sounds promising. For example, Hommen *et al* proposed to reconstruct the plasma boundary from dual camera-based, poloidal-view wavelength images, and attained high qualitative and quantitative agreement with EFIT in the MAST device [6]. Subsequently, they implemented real-time plasma vertical displacement control in the TCV device [4]. However, the work [4, 6] is restricted to the analysis and reconstruction of plasma discharges of a single configuration, the 'TCV standard shot' [4], since it relies on an appreciable distance between the plasma boundary and the first wall results in well-defined boundary features in the camera images, with little polluting light from reflections or plasma-wall interaction in the regions of interest (ROI). Ravensbergen *et al* extracted the optical plasma boundary and radiation front for detached divertor plasmas from multi-spectral imaging, and showed the possibilities to reliably detect the divertor leg and radiation front by lightweight image processing tools [7]. But the magnetic field distribution on the divertor leg can affect the system's sensitivity [7]. Luo *et al* utilized a Least Square-based method to delineate ROI in Charge-Coupled Device (CCD) images manually and demonstrated the successful reconstruction of plasma boundaries in EAST tokamak [8]. Nevertheless, the manual set of ROI makes these works [4, 7, 8] less competent in automatically accomplishing the plasma shape inference task. In addition, the algorithmic design may be constrained by the availability of optical sources. For instance, in the HL-3 tokamak, the proximity to plasma heating regions makes the tangential CCD cameras susceptible to lens coating due to material deposition [9], which significantly degrades image quality and leads to inaccurate observations of the plasma shape. Consequently, compared to the more information-abundant tangential views, using an extreme wide-angle lens poses considerable extra challenges in achieving a comprehensive capture of the plasma configuration.

On the other hand, attributed to the powerful representation ability and efficient parallel computing, the application of machine learning (ML) has yielded significant progress in equilibrium reconstruction [10–12] and solver [13, 14], plasma control [15, 16] and boundary detection [17, 18]. In particular, by learning from labeled thermal patches on the initial plasma wall of the W7-X device, Szucs *et al* employed a vision-centric, YOLOv5-based method to infer possible thermal patchs while achieving near-real-time inference times [17]. Based on manually labeled plasma boundaries, Yan *et al* utilized the U-Net neural network to identify the boundary on EAST [18]. Boyer *et al* [19] proposed a relatively simpler Convolutional Neural Network (CNN) model to identify the detachment front and strike positions (X point) from camera images on the DIII-D device. Besides suffering from manual labeling errors, the complexity of tokamak plasma poses another considerable challenge for accurate shape reconstruction. As shown in figure 1, factors like shifting plasma position, temperature, brightness, and current density distribution can impact the overall brightness within the device, thereby increasing the difficulty of predicting positional parameters. Additionally, the light spots caused by inner wall windows and gas injection interfere visually with plasma shape features, thus further adding to the reconstruction difficulty. Fortunately, experiments have confirmed that the existence of scaling laws in deep learning, which indicate that larger, more powerful deep neural networks (DNN) generally demonstrate improved generalization capabilities and robustness [20, 21], given their enhanced capability to capture intricate patterns within the data and handle out-of-distribution examples. Meanwhile, attention-based Transformer [22] backbones promise remarkable superiority in visual tasks than its convolutional counterparts [23]. Hence, it is natural to apply more powerful Transformer [22] for inference from CCD images of the complex environment. Notably, though DNN-based surrogate models promise faster inference speed with satisfactory accuracy [24, 25], it assumes the adoption of Multi-Layer Perceptron (MLP) and LSTM during the implementation. However, as the width and depth escalate, the computational complexity of the Transformer neural network increases triply [22]. Therefore, achieving real-time computation on cost-effective hardware platforms (e.g. Nvidia GeForce RTX 2080 Ti) becomes more challenging. In other words, for the real-time detection and control of plasma shape, it becomes imperative to strike the balance between accuracy and inference speed.



(a) (b) (c)

Figure 1. Images of plasma shape captured with a camera on the HL-3 device. (*a*) Displays the inner limiter shape in shot #06696, with window interference present on the wall. (*b*) Is the divertor shape related to shot #06696, wherein the image exhibits low brightness levels, leading to a loss of edge recognition. (*c*) Is a gas puffing interference image in shot #06255, with bright air delivery interference spots visible on the wall.

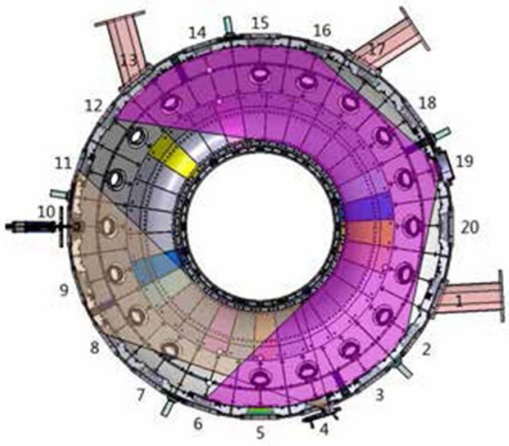


Figure 2. Toroidal cross section of the HL-3 tokamak. © [2024] IEEE. Reprinted, with permission, from [9]. Including the tangential view diagnostic (orange color, from #4 equatorial port), the extreme wide-angle view diagnostic (magenta color, from #19 equatorial port), and the downward-looking view diagnostic installed in the #5 sector, which is indicated by a green arrow. The location of the gas puffing entry at #10 equatorial port is indicated by a red arrow. In this study, we use the extreme wide-angle view diagnostic.

In this work, on top of Swin Transformer [26], we develop a series of Poolformer [27, 28] Swin Transformer (PST) models, which demonstrate sufficient inference speed and relatively higher plasma reconstruction accuracy after training from large-scale data from the HL-3. In particular, we adopt a multi-task learning [29] framework by taking extreme wide-angle CCD images as the input while simultaneously utilizing the output of EFIT including radial position R and vertical position Z of plasma geometric center, minor radius a, elongation κ, upper triangularity δ_u, and lower triangularity δ_l as labels. Such a design also avoids the cumbersomeness of manually labeling ROIs and contributes to learning consistency. To combat the unreliability of EFIT calculations during stages where the absolute value of plasma current is small but the rate of change is large, we incorporate a dynamic weight strategy as well [30]. Furthermore, a knowledge distillation procedure is utilized for further compressing the model. The adapted Swin Transformer model manifests the comprehensive adaptability to the visual complexity of the HL-3 device plasma, including overly blurred boundaries, Neutral Beam Injection (NBI) and gas puffing interference, sudden bright spots, and wall hole interference. This specific model can seamlessly integrate with a plasma control system (PCS) and effectively support immediate magnetic field control.

2. Methods

2.1. Data collection and pre-processing

HL-3 is a medium size tokamak with an aspect ratio of 2.8: plasma current $I_p = 2.5$–3 MA, toroidal field $B = 2.2$–3 T, major radius $R = 1.78$ m, minor radius $a = 0.65$ m, and elongation $\kappa \leqslant 1.8$; triangularity $\delta \leqslant 0.5$ [31]. HL-3 was designed to have a flexible configuration in order to explore multiple divertor configurations. Three heating and current drive (HCD) systems are able to provide a total power of 27 MW, including 15 MW of NBI, 8 MW of electron cyclotron resonance frequency (ECRF), and 4 MW of lower hybrid current drive (LHCD). In this work, given the clear evidence of the positive relationship between the CCD image and EFIT of a plasma shape [8, 18], we aim to directly learn the characteristics of a plasma CCD image and reconstruct the plasma shape according to the output of EFIT. In other words, the input is the CCD image taken from HL-3, while the output is the corresponding parameters (i.e. R, Z, a, κ, δ_u, and δ_l) for the plasma shape output by EFIT. Table 1 summarizes the details of the input and output.

Specifically, the HL-3 device is equipped with an advanced visible light diagnostic system, providing three disparate views—tangential, extreme wide-angle, and downward-looking—of the dynamics of each discharge. As shown in figure 2, the tangential view theoretically provides clearer information on the upper and lower divertor geometry, enhancing the capture of lateral plasma behavior. Yet, the tangential lens' vicinity to the heating area on HL-3 often leads to lens coating, resulting in over-blurred lens imagery, limited effective data, and complications in plasma observation. Consequently, we primarily employ CCD images from the extreme wide-angle view due to their robustness. Albeit its sub-optimality, our results demonstrate that it still yields fairly accurate reconstruction results.

The CCD camera for the extreme wide-angle view outputs images in the Bayer GB8/Bayer GB10 format, with a high frame rate of up to 2000 fps and a resolution of 1920×1080 pixels, each measuring 10 μm $\times$ 10 μm. This high frame rate ensures the capture of subtle dynamic changes in plasma, vividly reflecting the morphology and dynamic changes of the plasma, and providing abundant visual information for diagnosis. Additionally, the Bayer algorithm can convert these images into an RGB format without loss of quality. Notably, the exposure time is taken into account as well, given the inherent brightness differences caused by adjustments of exposure time (i.e. 1$\sim$2 ms) from the camera API.

During the experiment, we collect a dataset of CCD images with exposure time information from Shot #05554 to #06226. In particular, we select 271 effective shots, characterized by the plateau phase lasting a minimum of 500 ms

Table 1. Input and Output of CCD-based Plasma Shape Reconstruction Model.

Input/Output	Variable	Unit	Dimension	Description
Input	Image	N/A	$t \times 3 \times 120 \times 120$	Input image for the CCD perception model.
Input	Exposure Time	ms	1	Interval time of image acquisition by the camera.
Output	R	cm	1	Radial position of plasma geometric center
Output	Z	cm	1	Vertical position of plasma geometric center
Output	a	cm	1	Minor radius
Output	κ	dimensionless	1	Elongation
Output	δ_u	dimensionless	1	Upper triangularity
Output	δ_l	dimensionless	1	Lower triangularity

and a consistent plasma current of 100 kA during the phase. This selected dataset corresponds to a combination of 324 911 images. For each image, it undergoes initial cropping followed by linear interpolation to achieve a size of 120×120 pixels. Afterward, the compressed images undergo a group normalization (GN) operation. While the primary purpose of GN is to ensure stable training by normalizing feature values across groups of channels, it also helps maintain consistency in feature scaling among different images, which indirectly supports the robustness to adapt changes in brightness and contrast. Meanwhile, no denoising operation is conducted, as this helps to effectively retain the nuanced contrasts within the images. Among the selected shots, 236 are designated for training, with the remaining 35 allocated for testing. Given the typically high similarity between adjacent shots, we carefully divide the dataset to guarantee the existence of noticeable discrepancies in reference templates and plasma shape for shots under both the training and testing datasets. This process involves to review the discharge logs to select testing shots with discharge parameters and configuration waveforms significantly different from those in the training set, ensuring that the model is tested across a wider variety of conditions. This approach provides a more accurate assessment of its generalization capabilities.

On the other hand, the EFIT solution employed on the HL-3 device has undergone extensive validation for accuracy and reliability, especially in high-current plasma shots, such as those achieving 1 MA with an elongation ratio of 1.6 [33]. Additionally, comparisons of EFIT outputs with other diagnostic tools, including the microwave reflectometer and static electric probes, have confirmed its accuracy in determining key plasma parameters, such as plasma position and strike points in the divertor. These validations underscore the robustness of EFIT in guiding real-time plasma control. Nevertheless, EFIT might encounter convergence issues for certain cases, especially in correcting errors in rapid shape changes incurred by cases like vertical displacement [33], which could cause significant interference to the training of the CCD model. Furthermore, manually locating and rectifying all poorly converged EFIT data points can prove to be arduous, particularly when dealing with tens of thousands of data points. On the contrary, as demonstrated in evaluations on the HL-3 device [32], EFITNN [10, 13] is able to automatically learn and correct inconsistencies within the data through mapping relationships between magnetic diagnostic data and plasma shape information, thus providing a qualified basis for model training and testing. The EFITNN method demonstrates high accuracy compared to the standard EFITs, as reported in [32]. For instance, table V in [32] (table 2 in this paper) shows that the correlation coefficient (ρ) for plasma parameters exceeds 0.97, and the coefficient of determination ($\mathcal{R}^2$) approaches 0.96 for the validation and test sets. The mean absolute value of normalized residuals $|\Delta|$ is also consistently low, indicating the reliability of EFITNN as training labels. Therefore, during model training and offline testing, we use EFITNN, a surrogate model, as the benchmark; while for on-device deployment & evaluation, the more authoritative EFIT solution is used instead to ensure reliability. In particular, during training and testing, we use six primary parameters output by EFITNN (R, Z, a, κ, δ_u, and δ_l) as labels in our model. As shown in figure 3, the shape parameters distribution is plotted using the kernel density estimation method. For each parameter, the actual maximum and minimum are extracted from effective shots' statistics, and a min-max normalization operation, which brings any parameter $x \in \{R, Z, a, \kappa, \delta_u, \delta_l\}$ to an interval of $(0, 1)$ as $x \leftarrow \frac{x-\min}{\max-\min}$, is applied accordingly. We find that such an operation can prevent potential biases in the learning process.

Table 2. Correlation coefficient ρ, coefficient of determination $\mathcal{R}^2$, and the mean absolute value of the normalized residuals $|\Delta|$ between the predicted and the EFIT values of plasma parameters for the validation and test sets (shots #3309–#4186). The above data represents the average of calculations performed separately for each shot [32].

Name	ρ	$\mathcal{R}^2$	$\lvert\Delta\rvert$
R	0.974	0.941	0.0047
Z	0.991	0.977	0.0287
a	0.976	0.939	0.0101
κ	0.992	0.979	0.0087
δ_u	0.984	0.929	0.0409
δ_l	0.974	0.923	0.0561

2.2. ML model

We initially consider following the DNN structure of Swin Transformer [26] to learn the relationship between CCD images and plasma shape parameters (R, Z, a, κ, δ_u, and δ_l). Specifically, to connect an image to the output, Swin

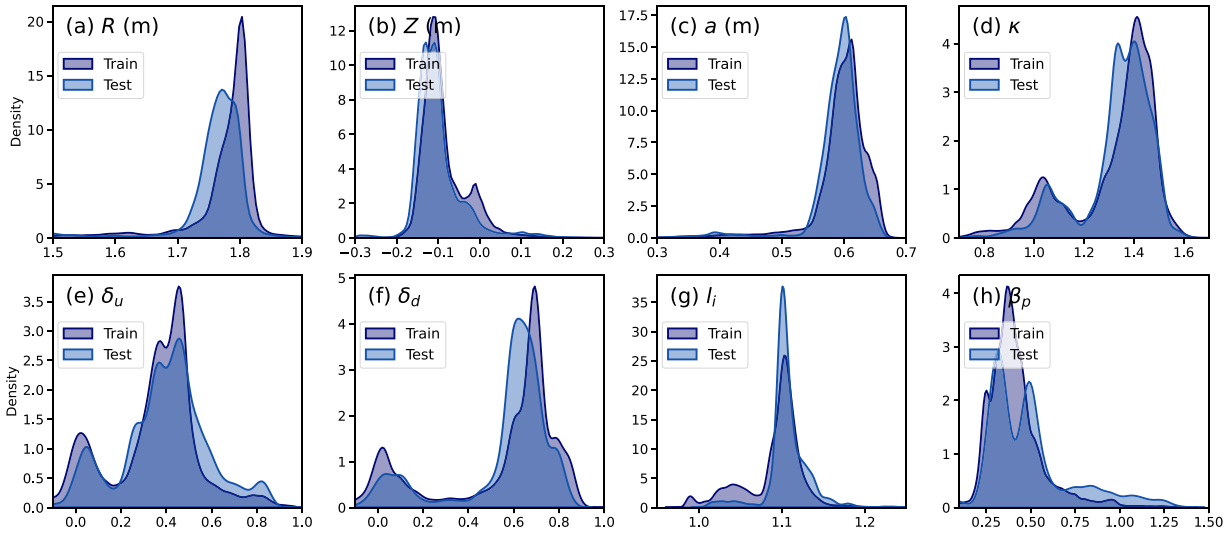


Figure 3. Distribution of shape parameters in the training and testing datasets, based on the kernel density estimation method. This visualized distribution illustrates the range and variations of each corresponding parameter.

Transformer [26] adopts a parallelizable multi-head attention mechanism in Transformer [22, 34], which completely dispenses recurrence and convolutions. Swin Transformer then employs a CNN-alike methodology to extract the hierarchical feature maps of a CCD image. Generally, Swin Transformer consists of 4 stages, each beginning with a patch merging and layer normalization operation to gradually downsample the image. Intuitively, as for an RGB image of $3 \times 120 \times 120$ (indicated as $3 \times H \times W$), Swin Transformer changes it to a size of $C \times \frac{H}{4} \times \frac{W}{4}$, $2C \times \frac{H}{8} \times \frac{W}{8}$, $4C \times \frac{H}{16} \times \frac{W}{16}$, $8C \times \frac{H}{32} \times \frac{W}{32}$ after each stage. Notably, to maintain an acceptable computational cost, the Swin Transformer does not directly calculate the global attention from the entire image. Instead, it partitions the image into fixed-size windows (in our study, we adopt a window size of 5×5), and employs a Windows-Multi-head Self-Attention (W-MSA) mechanism [26] for each window. However, the lack of inter-window information exchange confines the model to capture the long-range relationship. Therefore, a Shifted Window-MSA (SW-MSA) block, which shifts each window half of the window size downward and rightward respectively, enhances the performance. In practice, Swin Transformer alternately uses W-MSA and SW-MSA block. Detailed implementation procedures are provided in the appendix.

Swin Transformer offers multiple versions, including Swin-tiny, Swin-small, Swin-base and Swin-Large, each increasing in model size and complexity. We utilize the simplest and lightest one, Swin-tiny, with a depth (i.e. no. of blocks) of 2, 2, 6, and 2 at each stage. As discussed lately, Swin Transformer can accurately interpret local details and global brightness distribution of the plasma, producing precise parameter predictions even under a wide and variable range of plasma conditions. Nevertheless, even for the smallest Swin-tiny model, it takes at least 10 ms to perform the inference on Nvidia GeForce RTX 2080 Ti, which hinders its applications in high-frequency real-time magnetic control.

To address this issue, we take inspiration from [27, 28] and develop a PST-tiny DNN. Particularly, PST-tiny only keeps a portion of blocks (i.e. the first 1, 1, 2, and 2 blocks) for four stages, and reduces the dimension of the hidden layer from 64 in Swin-tiny to 32. More specifically, the Swin block in the first stage is replaced with a more computation-efficient Poolformer block [27, 28]. In particular, as its name implies, Poolformer divides the input tokens (i.e. partitioned images) into multiple groups and then performs a pooling operation within each group. Theoretically, for the number of input tokens n, MSA has a computational complexity of $O(n^2)$, while Poolformer only needs $O(n)$ computations [27, 28].

To mitigate overfitting and enhance model generalization, we apply regularization techniques such as dropout [35] and stochastic depth [36], which randomly deactivate neurons and skip layers, respectively, to boost model robustness. In addition, we leverage convolution and batch normalization operations rather than layer normalization in Swin Transformer, given their typically faster speed and easier parallelizable implementation on modern hardware. Notably, our experimental results show that the DNN structure re-design can contribute to a reduction in inference time of over 50% compared to the Swin Transformer. As a comparison, we name the Swin-tiny with an added Poolformer as PST-base. Finally, the DNN hyperparameters configured for PST-tiny model are summarized in table 3.

2.3. Model training

In this part, we first outline the design of the loss functions. Specifically, we leverage a multi-task learning loss with a dynamic weight strategy to efficiently learn the six plasma shape parameters, and adopt a knowledge distillation loss to further compress the model. Following this, we detail the practical training procedures.

Table 3. DNN hyperparameter configuration of the PST-tiny model.

Hyperparameter	Default Value
Depths	$(1,1,2,2)$
Patch Size	4
Input Channels	3
Embedding Dimension	96
Number of Heads	$(3,6,12,24)$
Window Size	5
MLP Ratio	4.0
Dropout Rate	0.1
Attention Dropout Rate	0.1
Drop Path Rate	0.2

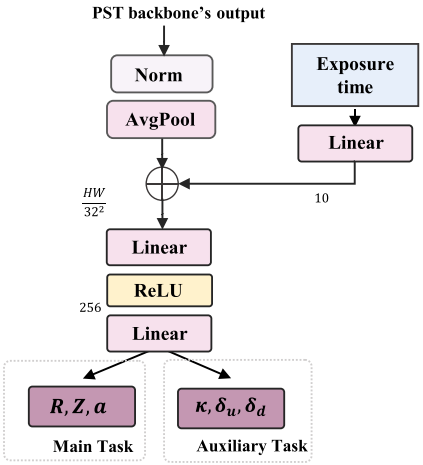


Figure 4. Final DNN part of PST for multi-task learning framework.

2.3.1. Loss function design

2.3.1.1. Individual task loss. In response to the observed difficulty of the EFIT solution to achieve convergence during the ramp-up and ramp-down phases of plasma discharge, we opt to implement a particular loss function. In this context, given the potential outliers for R, Z, and a calculated during the ramp-up and ramp-down phases, we choose the Huber loss function, which combines Mean Squared Error (MSE) and Mean Absolute Error (MAE), to effectively mitigate the negative impact. Mathematically, for $x \in \{R,Z,a\}$,

$$L_{\text{Huber}}(x,x') = \begin{cases} \frac{1}{2}(x-x')^2, & |x-x'| \leqslant \delta; \\ \delta|x-x'| - \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \quad (1)$$

where x' is the output by the PST model corresponding to a CCD image while x denotes the corresponding EFITNN-output label. Besides, the hyperparameter δ is set as 1. Meanwhile, for the parameters κ, δ_u and δ_d, which have a certain tolerance for prediction errors, we use the MAE loss function $L_{\text{MAE}}(x,x') = |x-x'|$ exclusively to avoid over-penalization. The corresponding formula for a batch can be summarized as

$$R(x) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} L_{\text{Huber}}(x_i,x_i'), & x = R,Z,a; \\ \frac{1}{N}\sum_{i=1}^{N} L_{\text{MAE}}(x_i,x_i'), & x = \kappa,\delta_u,\delta_d \end{cases} \quad (2)$$

where N represents the batch size.

2.3.1.2. Multi-task learning loss. Since R, Z and a in a CCD image exhibit more intuitively distinctions than κ, δ_u, and δ_d, and the PCS in HL-3 is contingent on R, Z, and plasma current I_p for control, we perform a Multi-Task Learning (MTL) by treating the learning of R, Z, and a as main tasks while regarding the perception of κ, δ_u, and δ_d as auxiliary tasks. The structure of this MTL design is depicted in figure 4. Such an MTL design contributes to learning the similarities across tasks, while the incorporation of auxiliary tasks allows for more effective optimization of the DNN parameters, thereby alleviating overfitting. Although the previous study [30] has shown that jointly trained DNNs outperform those trained separately for each task, determining appropriate MTL weights is costly, especially when multiple plasma shape parameters are required. Therefore, to effectively balance the individual contribution of the six shape parameters, we employ a multi-task loss strategy [30].

Accordingly, the loss function for a batch in MTL can be written as

$$L_{\text{MTL}}(x) = \sum\nolimits_{x\in\{R,Z,a\}} \left(\frac{1}{2c_x^2}R(x) + \ln\left(1+c_x^2\right)\right) + \sum\nolimits_{x\in\{\kappa,\delta_u,\delta_d\}} \alpha\left(\frac{1}{2c_x^2}R(x) + \ln\left(1+c_x^2\right)\right), \quad (3)$$

where $\ln(1+c_x^2)$ is applied as a regularization term, and the learnable network parameter $c_x, \forall x \in \{R,Z,a,\kappa,\delta_u,\delta_l\}$ contributes to automatically accounting for the different variances and biases among single-task losses. This approach effectively balances the loss weights of the six parameters, thereby achieving superior performance. During practical training, the three parameters in the auxiliary tasks, which receive less attention, are initialized as smaller values (0.5, 0.5, 0.5), whereas the initial values for the parameters in the main tasks are set as 1. The hyperparameter α is used for the auxiliary tasks to prevent them from becoming overly important, and it is set at 0.2 in our study.

2.3.1.3. Knowledge distillation loss for model compression. Despite the enhanced time efficiency of PST-tiny, it exhibits a significant accuracy gap compared to PST-base model. In order to compensate for the performance loss, we employ the method of knowledge distillation, by using PST-tiny as a more streamlined student DNN model and PST-base as the teacher model. This modification allows PST-tiny to deliver more reliable results while enhancing its efficiency.

Concretely, the plasma shape reconstruction task falls into the scope of regression learning, where conventional label-oriented distillation techniques struggle to generate consistent continuous output. Consequently, we turn to feature-based knowledge distillation techniques and adopt Masked Generative Distillation (MGD) [37], which uses features

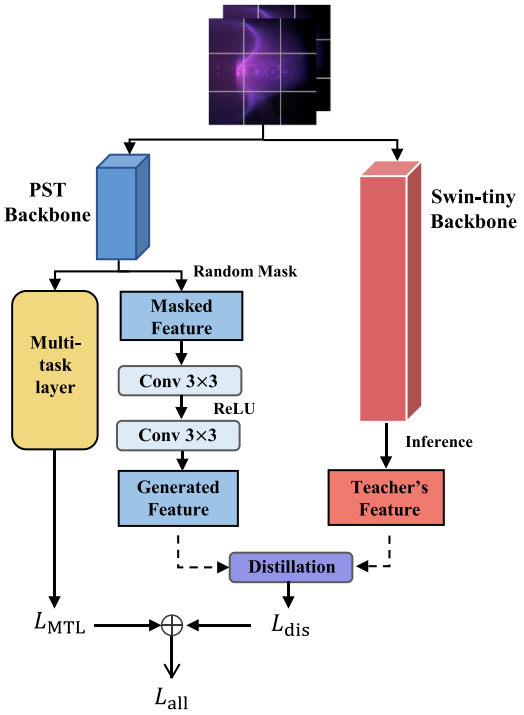


Figure 5. Illustration of the Masked Generative Distillation (MGD) structure. The student's feature is randomly masked and then the projector layer is utilized to induce the student to generate the teacher's feature using the masked one.

(i.e. the hidden layer output) demonstrated by a well-trained teacher DNN to guide the learning of the student DNN. The structure of this distillation framework is illustrated in figure 5, showing the interaction between the teacher and student networks and the role of the adaptation layer. Furthermore, MGD applies random masking to pixels of the student's feature and forces it to generate the teacher's full feature through a simple adaptation layer. In MGD, random pixels are used in each iteration, ensuring all pixels are eventually utilized throughout the training process. Meanwhile, the distillation occurs at the final layer of the Swin Transformer blocks in both networks with MSE computed for pixel values at corresponding positions on the feature maps. Mathematically, for a feature map with size $C \times H \times W$,

$$L_{\mathrm{dis}} = \sum_{k,i,j=1}^{C,H,W} \left(F^{\mathbf{T}}_{k,i,j} - G\left(f_{\mathrm{align}}\left(F^{\mathbf{S}}_{k,i,j}\right) \times K_{k,i,j}\right)\right)^2, \qquad (4)$$

where the superscripts S and T correspond to the student and teacher DNNs, respectively. f_{align} represents the adaptive layer that aligns student features with teacher features, while the student's feature map is randomly masked according to the masking matrix K. $G = W_{l2}(\mathrm{ReLU}(W_{l1}(F)))$, which incorporates two convolution layers (W_{l1}, W_{l2}) and one activation layer (ReLU), represents the projector layer and leads to consistent dimensions of feature maps in student and teacher models.

Since the teacher DNN only serves as a guide for student DNNs to restore features rather than requiring direct imitation, it effectively enhances the capability of the student DNN to generate and understand complex structural data.

2.3.2. Training procedures. For each image with a particular exposure time, six corresponding plasma shape parameters are used as labels. Considering the unreliability of EFIT calculations during periods where the absolute value of plasma current I_p is small, but the rate of change is large, owing to factors like eddy currents, we choose to dynamically adjust the confidence level of the labels. For each set of training data, we regard the first 150 ms as the ramp-up period and the last 150 ms as the ramp-down period. Accordingly, given our limited confidence in EFIT-output labels in these two periods, we assign smaller weights when calculating the loss function. On the contrary, due to the high confidence in the label accuracy during the plateau stage, we give it a higher weight. Such a dynamic weight strategy guides the model to focus more on improving performance during the plateau stage, which indirectly enhances its efficacy in the other phases. Accordingly, the MTL loss function is updated as

$$L_{\mathrm{MTL}} = \gamma \cdot L_{\mathrm{MTL}}(x) \qquad (5)$$

where the parameter γ, which is associated with the specific moment t within a shot, can be written as

$$\gamma = \begin{cases} 0.8, & \text{if } t < 150 \text{or } t > t_{\max} - 150, \\ 1, & \text{otherwise.} \end{cases} \qquad (6)$$

Here $t_{\max}$ refers to the moment at which the discharge ends.

On the other hand, by jointly involving the MTL loss and knowledge distillation loss, the combined loss function can be formulated as

$$L_{\mathrm{comb}} = (1-\zeta)L_{\mathrm{MTL}} + \zeta L_{\mathrm{dis}}, \qquad (7)$$

where ζ is a hyperparameter used for balance loss. Once the magnitudes of the original loss and feature loss are unified, we set $\zeta = 0.1$ for our experiments. This approach successfully ensures the performance of the PST-tiny model while significantly improving its efficiency.

2.4. Model deployment

Given a well-trained PST-tiny model, we utilize ONNX to convert it from the PyTorch model saving format (.pth) to the more universal ONNX format. This open and widespread model ONNX format facilitates convenient model sharing across multiple deep learning frameworks. Additionally, we apply the ONNX Simplifier tool, which capably infers the entire computation graph and replaces redundant operators with constant outputs, to simplify the model and enhance inference speed.

Ultimately, our model has been successfully deployed on a Windows operating system, with model inference running on Nvidia GeForce RTX 2080 Ti. Notably, we apply the powerful deep learning inference optimizer and runtime environment, Nvidia TensorRT [38], for subsequent optimization of models in the ONNX format. Typically, the model optimization, which

Table 4. Comparison of inference speed and deployment time for different models tested on Nvidia GeForce RTX 2080 Ti.

Model	Depth	Inference Time (ms)	Deployment Time (ms)
PST-base	(2,2,6,2)	3.4	4.3
PST-tiny	(1,1,2,2)	1.0	1.8

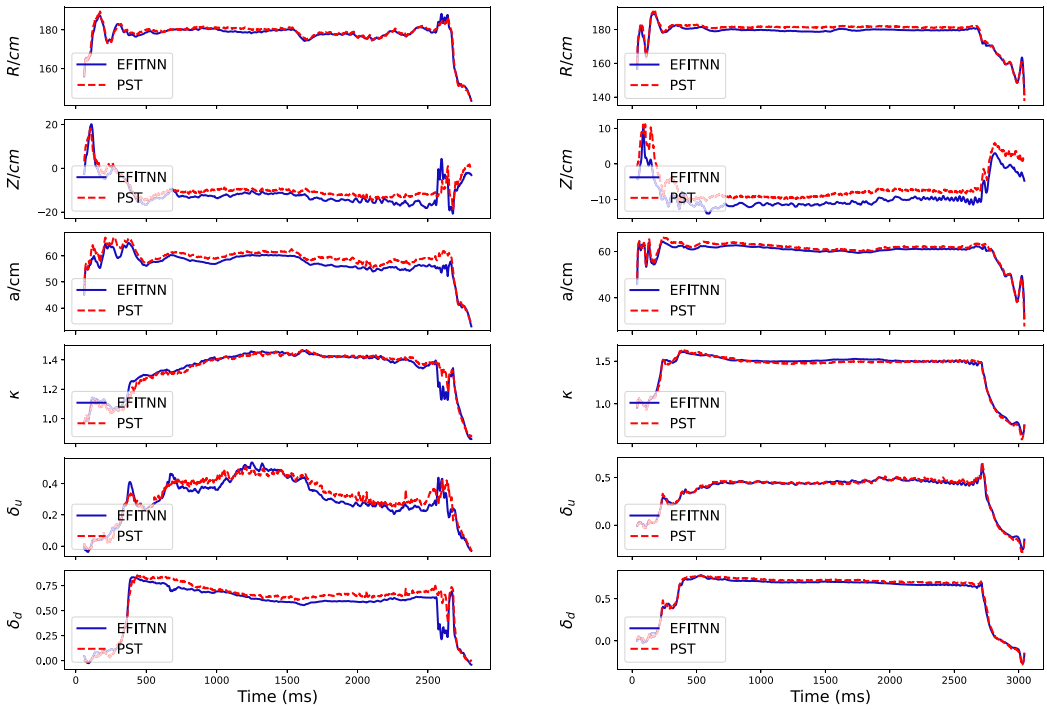


Figure 6. The comparison between the PST-based plasma shape detection results and EFITNN output on (*a*) #06227 shot and (*b*) #06236 shot.

involves layer fusion, operator fusion, model pruning, and precision calibration, further increases the model run efficiency and boosts the model performance. Our offline results suggest that the optimization contributes to further reducing half of the inference time. Table 4 summarizes the inference and deployment time. Meanwhile, as discussed lately, the model compression trivially affects the inference performance with less than 0.01 cm compared to the PST-base model.

3. Results

3.1. Offline results

In this part, we investigate the offline performance of DNN-based plasma shape detection results and evaluate their accuracy in comparison with EFITNN. As mentioned in section 2.1, the testing dataset encompasses 35 complete discharge shots. To ensure the evaluation completeness, the dataset is inclusive of the ramp-up, plateau, and ramp-down phases, and comprises an even distribution of plateau phases reaching 300 kA and 500 kA.

Figure 6 illustrates the inferred six shape parameters by the PST model throughout the entire discharge process for shots #06227 and #06236, each reaching 300 kA and 500 kA during the flat-top period, respectively. Generally, the model can qualitatively capture subtle shape changes in key parameters, and the prediction accuracy of R and a is particularly impressive, with an MAE margin within 1.5 cm. However, the accuracy for Z is slightly lower with approximately 1.8 cm on average. Correspondingly, κ, δ_u, and δ_l, all highly correlated with vertical displacement Z, exhibit discrepancies in some cases. Notably, such inferiority aligns with the adoption of the extreme wide-angle view, as opposed to the tangential view, which might miss some important changes in the upper and lower divertor. On the other hand, it can be observed that all shape parameters exhibit some degree of systematic offset, primarily due to variations in overall image brightness caused by heating and high-density conditions. Particularly, the largest deviation occurs in the vertical displacement Z. This phenomenon is likely attributable to the lower temperature and higher brightness observed in the divertor region, which are significantly affected by density variations and impurity radiation. Such effects adversely impact the accuracy of predicting the Z parameter as well as other shape parameters. To compensate for these local brightness variations and enhance model

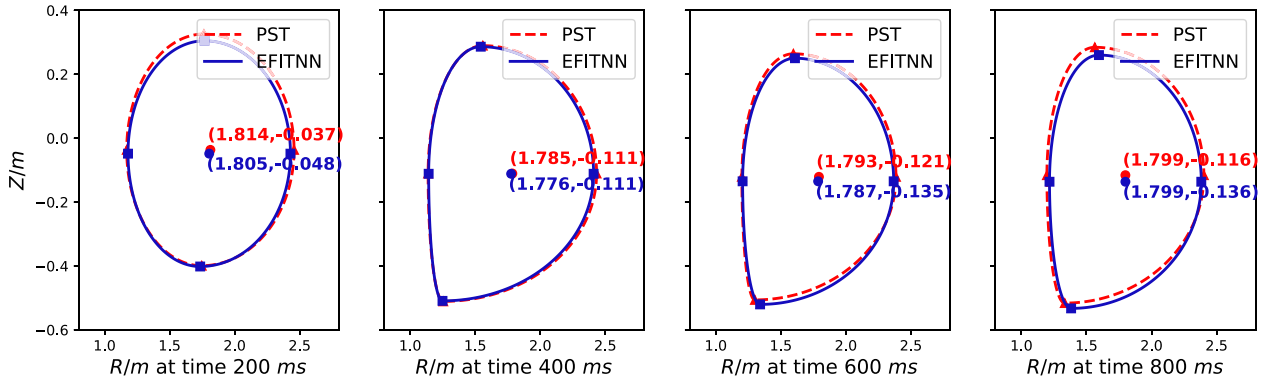


Figure 7. Comparison of plasma boundary coordinates for shot #06227 reconstructed using six parameters from the PST model and those using EFITNN.

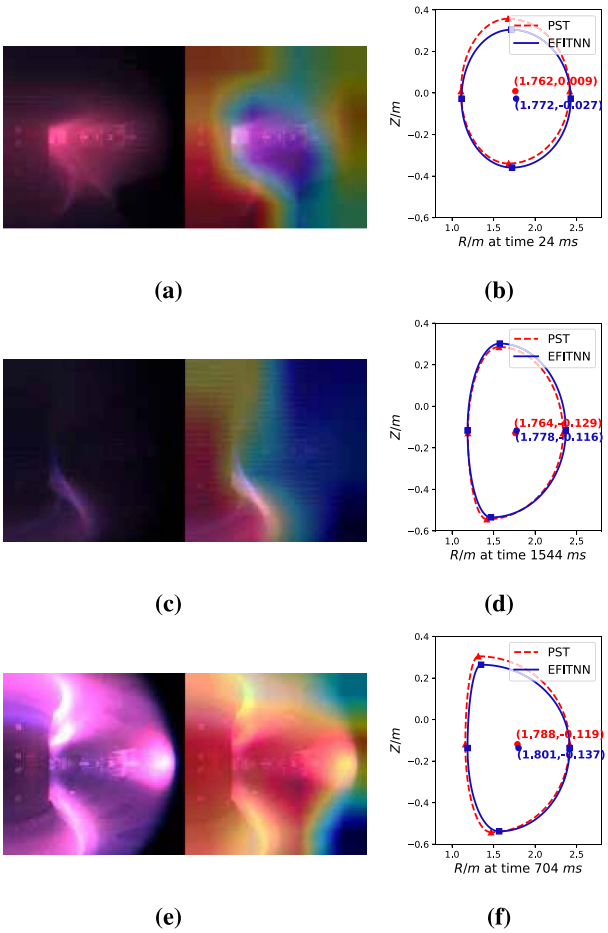


Figure 8. Visualization of feature maps and boundary detection results. Subfigures (*a*), (*c*), and (*e*) respectively represent the feature maps of the plasma under the states of inner limiter configuration, biased limiter configuration, and gas puffing process in figure 1, while subfigures (*b*), (*d*), and (*f*) provide the inferred boundary. Notably, the transition from blue to red in feature maps indicates increased attention or DNN weights to the related pixels during the computation.

robustness, potential research directions include adopting advanced image preprocessing techniques or increasing the diversity of training data. Furthermore, we try to explore the potential of spatio-temporal fusion techniques to mitigate image noise errors, by stacking T consecutive images. Although this approach promises enhanced accuracy, its lengthy inference and deployment time hinders practical application.

To maintain a clear understanding of the plasma shape, we compute the extreme points of the four sections from detected shape parameters and then approximate the plasma boundary as the linkage of four quarter-ellipses. Figure 7 presents the corresponding result for shot #06227, and demonstrates the proficient mimicking of the plasma's progression towards the divertor shape. On the other hand, for cases in figure 1, the left part of figure 8 illustrates the feature maps, which

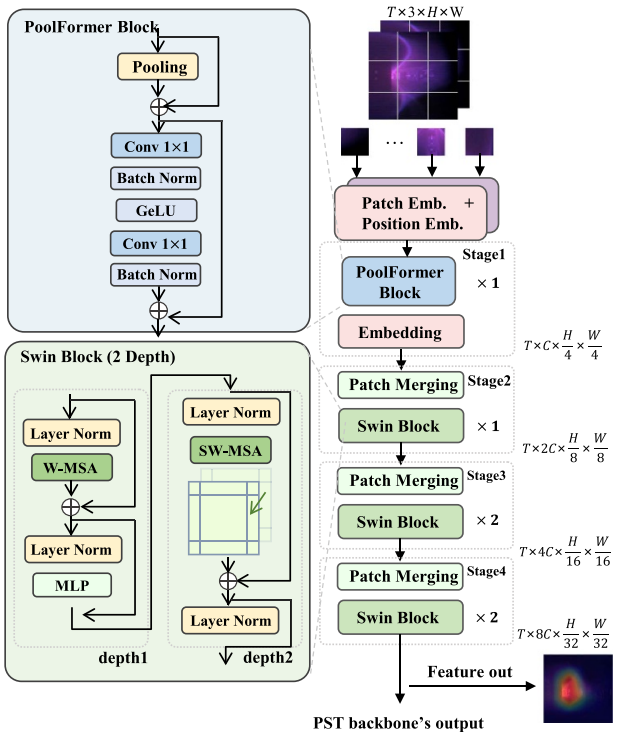


Figure 9. Illustration of the shared base DNN of PST.

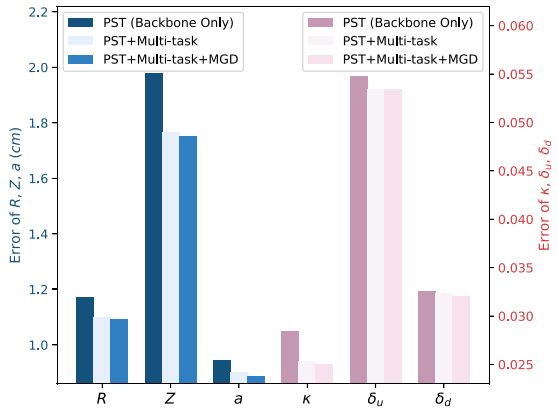


Figure 10. Bar chart of the PST model's gain after adding different modules.

are attained by extracting the intermediate outputs from the shared base DNN of the model in figure 9. Overall, the color in the feature map reflects the importance of individual pixels for subsequent computations, and the transition from blue to red indicates increased attention (i.e. larger DNN weights) to related pixels. Despite the consistently high attention on RoIs, subtle differences can still be observed in local feature maps. For internal limiter plasmas with lower plasma current I_p in figure 8(a), the feature map distinctly outlines plasma shape boundary and likely contributes to a more defined and easily identifiable boundary. For the divertor-shaped plasmas shown in figure 8(c), the feature map predominantly focuses on the pixel points at the divertor target plate, thereby enhancing the spatial positioning of the divertor configuration. Meanwhile, for gas puffing in figure 8(e), where the gas injection point overlaps with the outer boundary of the plasma, the presence of extra bright spots severely interferes with the boundary recognition. Consequently, the feature map puts more emphasis on the changes in brightness inside the plasma. Correspondingly, the recovered plasma boundary shown in the right part of figure 8 demonstrates high consistency with EFITNN in these challenging cases. Meanwhile, we investigate the gains of modular design in PST and provide the results in figure 10. It can be observed that additional enhancements in model performance can be anticipated after the integration of MTL and MGD techniques, especially in terms of the vertical displacement Z.

Finally, table 5 summarizes the comparison between PST-tiny and other models in terms of MAE. Bold values in the table indicate the best performance among all models for each parameter. It's worth noting that the PST model, upon the incorporation of the Swin Transformer, demonstrates comparable accuracy but significantly faster inference speed over Swin-Tiny. On the contrary, RestNet18 and UNet exhibit reduced accuracy and possibly further deteriorate under conditions of enhanced brightness and high contrastness. Taking the example of shot #06696 in figure 11, while all models produce comparable results under normal conditions, RestNet18 and UNet turn to perform poorly when the `convertScaleAbs` function from the OpenCV library is applied with a contrast factor α of 2 and a brightness offset β of 10 to adjust the image's contrast and brightness, respectively. Instead, the PST model maintains superior stability. Notably, operations such as gas puffing and enhancing the ionization rate often lead to increased density, resulting in over-bright CCD images. Such observation is consistent with the literature that Transformers exhibit strong robustness against common corruptions, perturbations, distribution shifts, and natural adversarial examples [39]. On the contrary, CNN and U-Net models, which rely heavily on convolutional operations, suffer from the sensitiveness to local feature dependencies, leading to degraded regression results due to brightness and contrast variations. Furthermore, figure 12 evaluates the accuracy of PST-tiny with goodness-of-fit. Again, the accuracy remains strong across the full range of plasma currents. However, a slight decrease in performance is observed at lower currents during the flat-top phase, which may be due to the lower brightness of the images under these conditions. Based on this, it might be possible that future observations of higher-parameter plasmas will require cameras operating in different spectral bands to maintain accuracy.

Table 5. Comparison of MAE for plasma shape parameters detected by different image perception models. (The CNN model adopts the architecture proposed in DIII-D-related studies [19]).

Model	Depth	R (cm)	Z (cm)	a (cm)	κ	δ_u	δ_l	Inference Time (ms)
PST-tiny	$(1,1,2,2)$	1.091	**1.752**	0.887	0.025	0.053	0.032	1.00
PST-base	$(2,2,6,2)$	**1.013**	1.895	**0.823**	**0.022**	**0.047**	**0.030**	3.40
CNN	—	1.515	2.579	1.148	0.034	0.066	0.045	0.26
ResNet18	—	1.302	2.125	1.134	0.023	0.056	0.038	0.79
UNet	—	1.938	2.985	1.195	0.028	0.0623	0.038	0.89

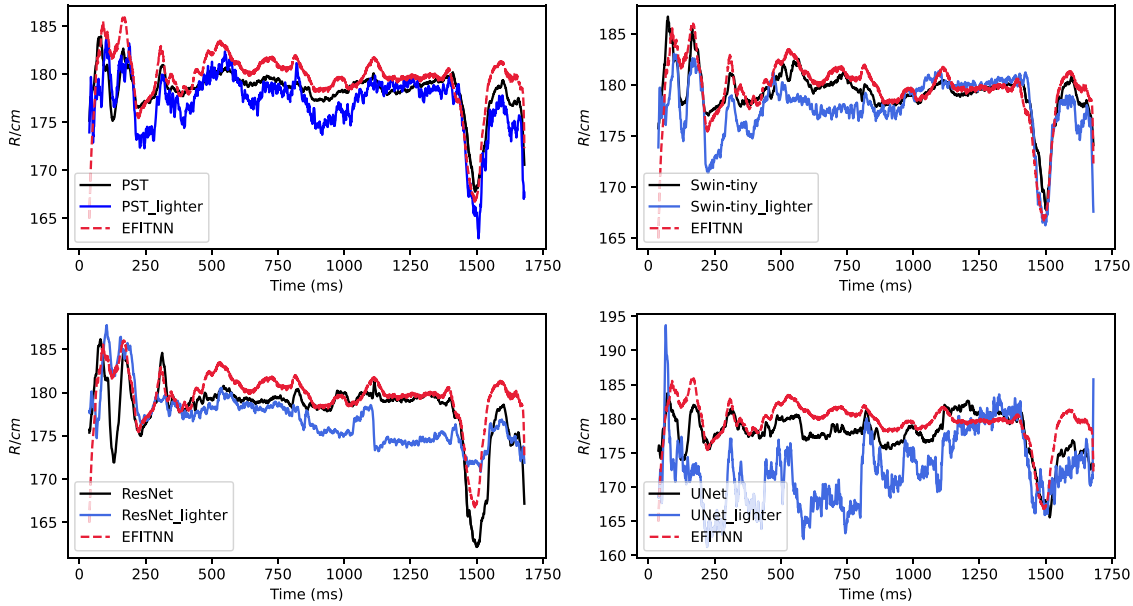


Figure 11. Comparison of the horizontal displacement R outputs from different models and the EFITNN output after enhancing the brightness and contrast of the overall shot #06696. The black lines represent the inference results of each model on the original image, the red lines represent the EFITNN calculation results, and the colored lines are the results after increasing brightness and contrast.

3.2. On-device results

Transitioning from offline simulations to real-time applications, we build a framework as in figure 13 to seamlessly process captured CCD images and integrate the PST-based outputs into the PCS of HL-3. Notably, the PST model infers from CCD images with camera exposure time and delivers six plasma shape parameters into the PCS through reflective memory. Afterward, the PID controller, a classic feedback control mechanism, reads data from the corresponding memory location every 0.1 ms and adjusts the control variables according to the evolution of R as well as other plasma shape parameters from magnetic diagnosis. Notably, as implied in table 4, the plasma shape configurations are updated in memory every 1.8 ms on average, which allows real-time control of the plasma shape based on the most recent available data.

As mentioned in section 2.1, during the on-device deployment & evaluation phase, we use a more authoritative EFIT solution as the benchmark to ensure reliability. Figure 14(a) presents the real-time control result with shot #07059, where, in the duration of 2000–2500 ms, the horizontal displacement R outputs from the PST model is activated as an input into the PID control system, replacing the magnetic diagnostic R. Within the 500 ms, the PID control maintains a stable control of the plasma. Figure 14(b) provides a detailed comparison during this period, incorporating the EFIT calculation method as well, revealing that the PST model's output aligns more closely with EFIT. On the contrary, while the original PID control method is based on magnetic measurements and the M matrix, a systematic deviation (an upward shift of 1–2 cm) is observed. This validates that PST model can effectively utilize the inference results from CCD images as a diagnostic source to control the plasma shape, promising real-time capability with improved accuracy.

4. Conclusion and future research

In this work, we have developed a lightweight DNN model—PST, for accurately and quickly perceiving CCD plasma images without any manual labeling on HL-3. This model predicts the plasma shape parameters, taking the CCD image as input and outputting six parameters including the radial

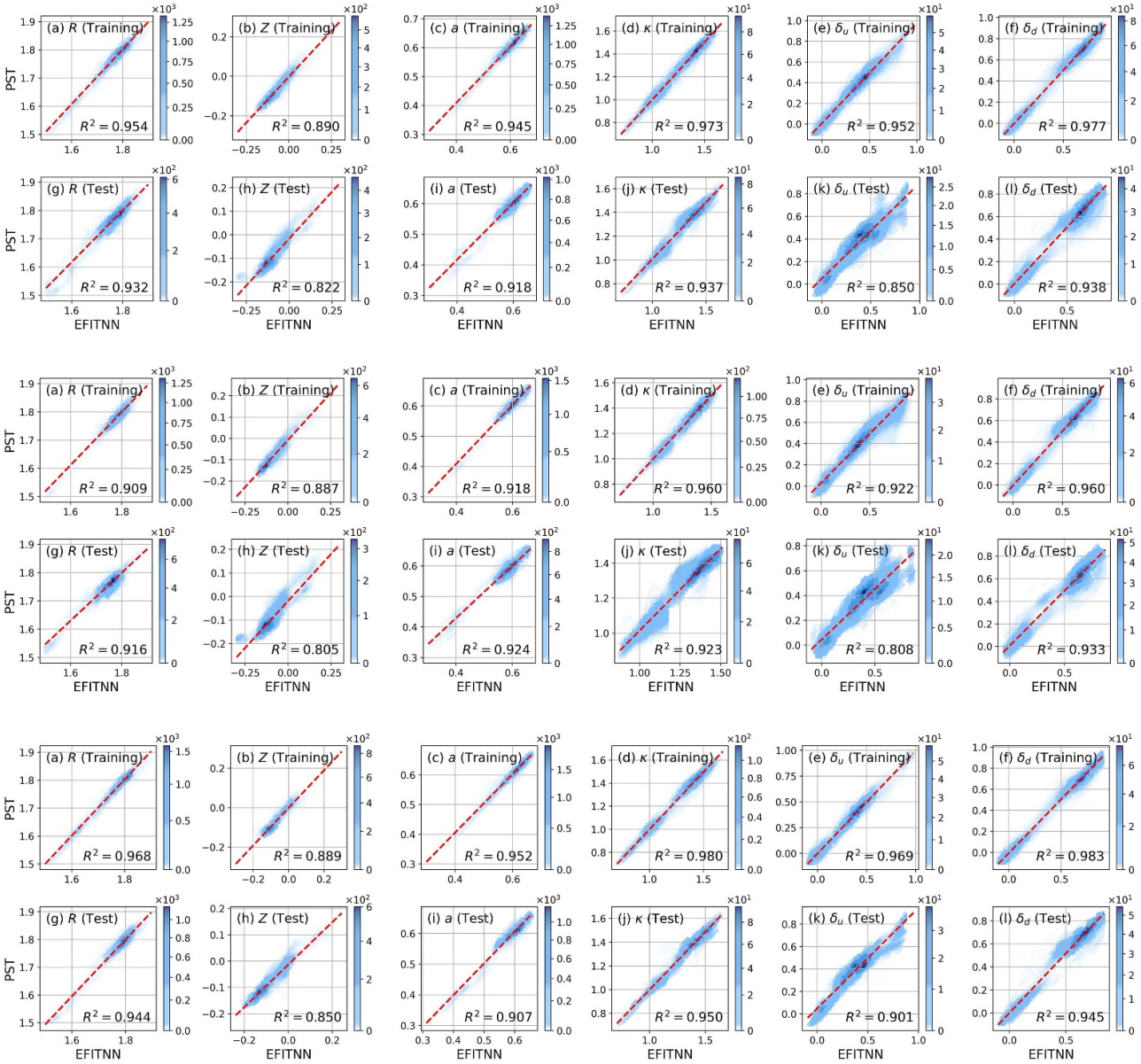


Figure 12. Goodness-of-fit of plasma parameters in terms of the coefficient of determination R^2 in the training and testing sets, based on kernel density estimation. (*a*) The results for all shots in the dataset; (*b*) the results for shots with $I_p \leqslant 300$ kA; and (*c*) the results for shots with $I_p > 300$ kA, where the parameters R, Z, and a are in meters (m).

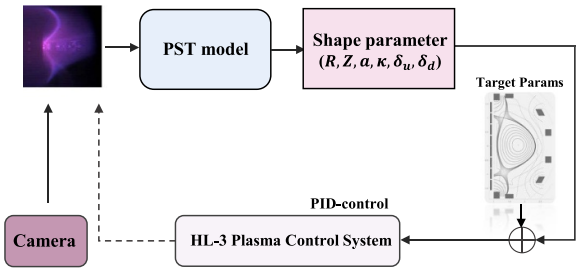
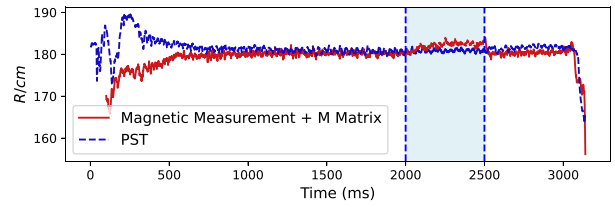
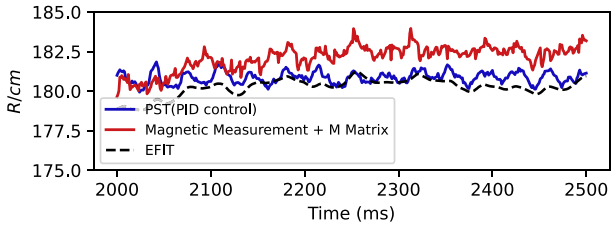


Figure 13. The illustration of the PST model serving as a diagnostic means for real-time PID control.



(a)



(b)

Figure 14. Results of deploying the PST model online and implementing PID feedback control on shot #07059. The blue region represents the phase in which R from the PST is used as input for feedback control. The original input for PID control is obtained using magnetic measurement and the M Matrix method of magnetic measurements. (a) Is the real-time detection results of the entire shot, and (b) represents a detailed comparison between the control segment of the PST model and other computational methods.

position R and vertical position Z of the plasma geometric center, minor radius a, elongation κ, upper triangularity δ_u, and lower triangularity δ_l. Specifically, the PST model adapts Poolformer and Swin Transformer towards a lighter-weight design. Meanwhile, we incorporate masked generative distillation, and adopt a multi-task learning framework with the dynamic weight strategy, obtaining high inference accuracy on the PST model. The PST model can predict six parameters within the entire process of 1.8 ms, with the average MAE for R and Z reaching 1.1 cm and 1.8 cm respectively. Furthermore, the PST model manifests the comprehensive adaptability to the visual complexity of the HL-3 device plasma, including overly blurred boundaries, NBI and gas puffing interference, sudden bright spots, and wall hole interference. This specific model can seamlessly integrate with the PCS and effectively support immediate magnetic field control. We have completed 500 ms PID stable control according to the PST-yielding horizontal displacement R parameter. This preliminary verifies the stability of the PST model in real-time control.

In summary, the results of this research further broaden the potential applications of AI in the field of tokamak plasma control and lay a solid foundation for the development of related technologies in the future. Notably, many works remain to be done. For example, we can explore more sophisticated methods for determining the trust parameter γ, including the use of EFIT chi-squared values and plasma current dynamics, such as $\frac{dI_p}{dt}$, particularly during ramp-up and ramp-down phases. Additionally, given its potential to balance speed and precision in real-time plasma control, implementing a dual-loop system similar to RT-EFIT, where a fast loop ensures quick real-time response, while a slower loop can improve accuracy by processing stacked images or using more refined calculations, is worth considering. Furthermore, In this study, the system achieves real-time feedback at the millisecond scale, which is well-suited for controlling future large-scale fusion devices such as ITER. However, for smaller devices like HL-3, the faster evolution of dynamics and the corresponding finer resolution (i.e. 50 μs for Z control) exceeds the capabilities of our current system's inference speed. Addressing these challenges by enhancing data acquisition and computation efficiency will also be the focus of future work.

Acknowledgments

This work was in part supported by the National Natural Science Foundation of China under Grant No. U21A20440, the Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant No. BX202222, and the Natural Science Foundation of Sichuan Province under Grant No. 2024NSFSC1335.

Appendix. Details in swin transformer

A.1. Patch merging

Swin Transformer adopts Patch Merging, a downsampling technique to the resolution of the input data while increasing the depth of the information. Within this context, a 'Patch' is defined as the smallest unit in the feature map. To put it differently, a feature map with 14×14 pixels comprises $196 (= 14 \times 14)$ patches. As illustrated in figure 15, Patch Merging aggregates each cluster of adjacent $n \times n$ patches by depth-concatenation, reducing the dimensions by a factor of 2. Consequently, the input dimensionality is transformed from $C \times H \times W$ to $4C \times \frac{H}{2} \times \frac{W}{2}$, where H, W, and C representing height, width, and channel depth respectively. Subsequently, the feature map undergoes a fully connected layer, which adjusts the channel dimension to half of its original size. Through the hierarchical design, it is conducive to capturing more complex features with effectively reduced computational load [26].

A.2. Self-attention block

Traditional Transformers perform attention calculations on a global scale, which results in substantial computational complexity. In contrast, the Swin Transformer confines these calculations within individual windows, thereby achieving a significant reduction in computational load.

As depicted in figure 16, compared to the conventional MSA module, the W-MSA divides the feature map into individual windows of size $M \times M$ (with $M = 2$ in this example),

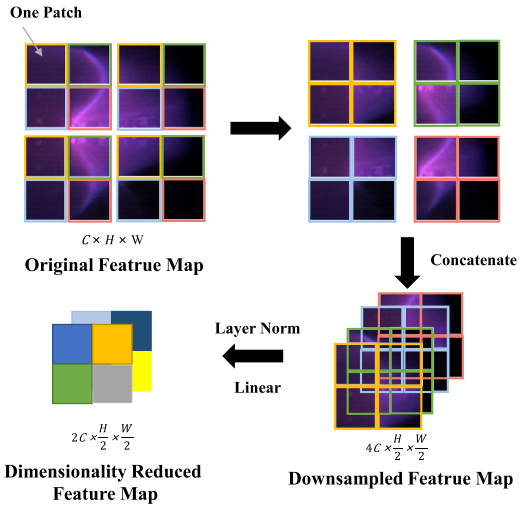


Figure 15. Schematic illustration of the Patch Merging process in Swin Transformer.

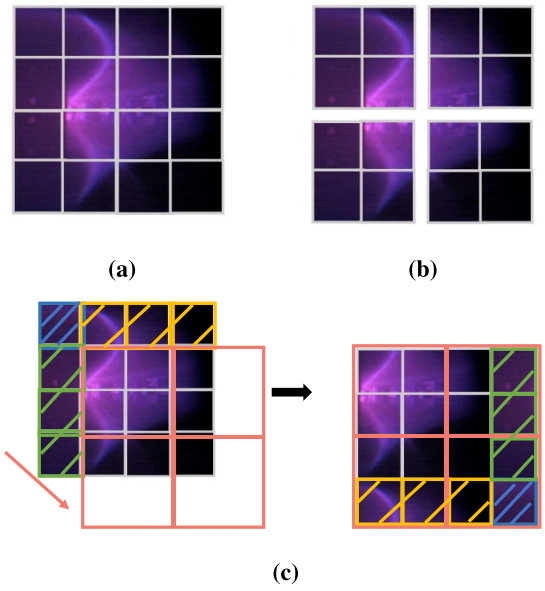


Figure 16. Illustration of Self-Attention Blocks: (*a*) Multi-head Self-Attention (MSA), (*b*) Windows-Multi-head Self-Attention (W-MSA), and (*c*) Shifted Window-MSA (SW-MSA).

and subsequently performs self-attention computations independently within each window. However, the window-limited computations in the W-MSA module inhibit the underlying usefulness of the information transfer between different windows. To circumvent this shortcoming, the Swin Transformer introduces the SW-MSA module, which shifts the window by a factor of $\frac{M}{2}$ both downward and rightward and moves patches into vacant slots to form a complete window for computing efficiency. This innovative strategy effectively resolves the issue of limited information exchange and improves the performance of Swin Transformer [26].

ORCID iDs

Qianyun Dong https://orcid.org/0009-0007-6020-6122
Rongpeng Li https://orcid.org/0000-0003-4297-5060
Zongyu Yang https://orcid.org/0009-0000-4083-1552
Wulyu Zhong https://orcid.org/0000-0001-8217-9400
Zhifeng Zhao https://orcid.org/0000-0002-5479-7890

References

[1] Ferron J.R., Walker M.L., Lao L.L., St John H.E., Humphreys D.A. and Leuer J.A. 1998 Real time equilibrium reconstruction for tokamak discharge control *Nucl. Fusion* **38** 1055
[2] Hutchinson I.H. 2002 *Principles of Plasma Diagnostics* 2nd ed (Cambridge University Press)
[3] Donné A.J.H., Costley A.E. and Morris A.W. 2012 Diagnostics for plasma control on demo: challenges of implementation *Nucl. Fusion* **52** 074015
[4] Hommen G., de Baar M., Duval B.P., Andrebe Y., Le H.B., Klop M.A., Doelman N.J., Witvoet G. and Steinbuch M. (the TCV Team) 2014 Real-time optical plasma boundary reconstruction for plasma position control at the TCV Tokamak *Nucl. Fusion* **54** 073018
[5] Santos J., Guimarãis L., Zilker M., Treutterer W. and Manso M. (the ASDEX Upgrade Team) 2012 Reflectometry-based plasma position feedback control demonstration at ASDEX Upgrade *Nucl. Fusion* **52** 032003
[6] Hommen G., De Baar M., Nuij P., McArdle G., Akers R. and Steinbuch M. 2010 Optical boundary reconstruction of tokamak plasmas for feedback control of plasma position and shape *Rev. Sci. Instrum.* **81** 113504
[7] Ravensbergen T. *et al* 2020 Development of a real-time algorithm for detection of the divertor detachment radiation front using multi-spectral imaging *Nucl. Fusion* **60** 066017
[8] Luo H., Luo Z., Xu C. and Jiang W. 2018 Optical plasma boundary reconstruction based on least squares for EAST tokamak *Front. Inf. Technol. Electron. Eng.* **19** 1124–34
[9] Liu L. *et al* 2024 Multidirectional observations of the high-performance plasmas by visible imaging diagnostics on the HL-2M tokamak *IEEE Trans. Plasma Sci.* **52** 1328–36
[10] Lao L.L. *et al* 2022 Application of machine learning and artificial intelligence to extend EFIT equilibrium reconstruction *Plasma Phys. Control. Fusion* **64** 074001
[11] Dong G., Wei X., Bao J., Brochard G., Lin Z. and Tang W. 2021 Deep learning based surrogate models for first-principles global simulations of fusion plasmas *Nucl. Fusion* **61** 126061
[12] Wai J.T., Boyer M.D. and Kolemen E. 2022 Neural net modeling of equilibria in NSTX-U *Nucl. Fusion* **62** 086042
[13] Joung S., Kim J., Kwak S., Bak J.G., Lee S.G., Han H.S., Kim H.S., Lee G., Kwon D. and Ghim Y.-C. 2019 Deep neural network Grad–Shafranov solver constrained with measured magnetic signals *Nucl. Fusion* **60** 016034
[14] Cheng L., Illarramendi E.A., Bogopolsky G., Bauerheim M. and Cuenot B. 2021 Using neural networks to solve the 2D poisson equation for electric field computation in plasma fluid simulations (arXiv:2109.13076)
[15] Degrave J. *et al* 2022 Magnetic control of tokamak plasmas through deep reinforcement learning *Nature* **602** 414–9
[16] Seo J., Kim S., Jalalvand A., Conlin R., Rothstein A., Abbate J., Erickson K., Wai J., Shousha R. and Kolemen E. 2024 Avoiding fusion plasma tearing instability with deep reinforcement learning *Nature* **626** 746–51

[17] Szűcs M., Szepesi T., Biedermann C., Cseh G., Jakubowski M., Kocsis G., König R., Krause M. and Sitjes A.P. (W7-X Team) 2022 A deep learning-based method to detect hot-spots in the visible video diagnostics of Wendelstein 7-X *J. Nucl. Eng.* **3** 473–9
[18] Yan H. *et al* 2023 Optical plasma boundary detection and its reconstruction on EAST tokamak *Plasma Phys. Control. Fusion* **65** 055010
[19] Boyer M.D., Scotti F. and Gajaraj V. 2024 Neural networks for estimation of divertor conditions in DIII-D using C III imaging *Nucl. Fusion* **64** 106056
[20] Hestness J., Narang S., Ardalani N., Diamos G., Jun H., Kianinejad H., M.M.A Patwary, Yang Y. and Zhou Y. 2017 Deep learning scaling is predictable, empirically (arXiv:1712.00409)
[21] Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Jeffrey W. and Amodei D. 2020 Scaling laws for neural language models (arXiv:2001.08361)
[22] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I. 2017 Attention is all you need *The 31st Annual Conf. on Neural Information Processing Systems* (*Long Beach, CA, USA*, *17 December 2017*) (NIPS vol 30) (available at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
[23] Zhou H.-Y., Chixiang L., Yang S. and Yizhou Y. 2021 Convnets vs. transformers: whose visual representations are more transferable? *Proc. IEEE/CVF Int. Conf. on Computer Vision* (*Montreal, BC, Canada*, *11–17 October 2021*) pp 2230–8
[24] Kates-Harbeck J., Svyatkovskiy A. and Tang W. 2019 Predicting disruptive instabilities in controlled fusion plasmas through deep learning *Nature* **568** 526–31
[25] Kim S.K. *et al* 2024 Highest fusion performance without harmful edge energy bursts in tokamak *Nat. Commun.* **15** 3990
[26] Liu Z., Lin Y., Cao Y., Han H., Wei Y., Zhang Z., Lin S. and Guo B. 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proc. IEEE/CVF Int. Conf. on Computer Vision* (*Montreal, QC, Canada*, *10–17 October 2021*) pp 10012–22
[27] Yu W., Luo M., Zhou P., Chenyang S., Zhou Y., Wang X., Feng J. and Yan S. 2022 Metaformer is actually what you need for vision *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*New Orleans, LA, USA*, *18–24 June 2022*) pp 10819–29
[28] Li Y., Yuan G., Wen Y., Hu J., Evangelidis G., Tulyakov S., Wang Y. and Ren J. 2022 Efficientformer: vision transformers at mobilenet speed *Advances in Neural Information Processing Systems* vol 35 pp 12934–49
[29] Liebel L. and Körner M. 2018 Auxiliary tasks in multi-task learning (arXiv:1805.06334)
[30] Kendall A., Gal Y. and Cipolla R. 2018 Multi-task learning using uncertainty to weigh losses for scene geometry and semantics *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*Salt Lake City, UT, USA*, *18–23 June 2018*) pp 7482–91
[31] Duan X.R. *et al* 2022 Progress of HL-2A experiments and HL-2M program *Nucl. Fusion* **62** 042020
[32] Zheng G.H., Yang Z.Y., Liu S.F., Ma R., Gong X.W., Wang A., Wang S. and Zhong W.L. 2024 Real-time equilibrium reconstruction by multi-task learning neural network based on HL-3 tokamak *Nucl. Fusion* **64** 126041
[33] Duan X.R. *et al* 2024 Recent advance progress of HL-3 experiments *Nucl. Fusion* **64** 112021
[34] Niu Z., Zhong G. and Hui Y. 2021 A review on the attention mechanism of deep learning *Neurocomputing* **452** 48–62
[35] Srivastava N., Hinton G., Krizhevsky A., Sutskever I. and Salakhutdinov R. 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[36] Huang G., Sun Y., Liu Z., Sedra D. and Weinberger K.Q. 2016 Deep networks with stochastic depth *Computer Vision–ECCV 2016* (*Amsterdam, The Netherlands*, *11–14 October*) vol 9908 (Springer) pp 646–61
[37] Yang Z., Li Z., Shao M., Shi D., Yuan Z. and Yuan C. 2022 Masked generative distillation *Computer Vision—ECCV 2022* (*Tel Aviv, Israel*, *23–27 October 2022*) vol 13671 (Springer) pp 53–69
[38] Xia X., Li J., Wu J., Wang X., Xiao X., Zheng M. and Wang R. 2022 TRT-ViT: tensorRT-oriented vision transformer (arXiv:2205.09579)
[39] Paul S. and Chen P.-Y. 2022 Vision transformers are robust learners *Proc. AAAI Conf. on Artificial Intelligence* vol 36 (*Virtual Conference*, *28 June 2022*) pp 2071–81