processes

MDPI

Article

Research on the Prediction Method for the Wear State of Aerospace Titanium Alloy Cutting Tools Based on Knowledge Distillation

Bengang Liu [1,2,3], Baode Li [2], Bo Xue [2], Zeguang Dong [2] and Wenjiang Wu [1,3,4,*]

[1] Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110136, China; liubengang18@mails.ucas.ac.cn
[2] Shenyang Aircraft Industry (Group) Co., Ltd., Shenyang 110136, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] Shenyang Zhongke Numerical Control Technology Co., Ltd., Shenyang 110136, China
* Correspondence: wwj@sict.ac.cn

Abstract: To address the challenges of high labeling costs and insufficient cross-condition generalization in training cutting tool wear state prediction models, this paper proposes a model optimization method based on knowledge distillation. By constructing a teacher model using a Bidirectional Gated Recurrent Unit (BiGRU-GRU) and a student model utilizing a Transformer architecture, we jointly employ KL divergence distillation, feature Euclidean distance distillation, and cross-entropy supervised training. A multi-objective joint loss function is designed to facilitate knowledge transfer. Using a self-collected dataset of aerospace TC18 titanium alloy cutting tools and the publicly available PHM2010 dataset, we conduct comparative experiments based on two different data partitioning strategies: tool grouping and mixed grouping. The results indicate that knowledge distillation significantly enhances the performance of the student model. The Transformer-BiGRU model is more robust to knowledge distillation. Positive results were obtained for the Transformer-BiGRU model in both tool grouping and mix grouping experiments. Additionally, comparative experiments show that the proposed method outperforms traditional methods (such as GS-XGBoost and Attention-CNN) in both grouping types, validating the effectiveness of knowledge distillation in transferring knowledge related to tool wear discrimination and addressing the issue of insufficient training data. The research demonstrates that combining knowledge distillation with the Transformer architecture can effectively enhance model generalization capabilities, providing theoretical support for monitoring the state of aerospace titanium alloy machining tools. However, the misclassification issue during the severe wear stage still requires further optimization based on physical principles.

Keywords: knowledge distillation; wear state prediction; aerospace titanium alloy machining; knowledge transferring; Transformer

check for updates

Academic Editor: Massimo Caruso

Received: 11 March 2025
Revised: 15 April 2025
Accepted: 22 April 2025
Published: 24 April 2025

Citation: Liu, B.; Li, B.; Xue, B.; Dong, Z.; Wu, W. Research on the Prediction Method for the Wear State of Aerospace Titanium Alloy Cutting Tools Based on Knowledge Distillation. *Processes* **2025**, *13*, 1300. https://doi.org/10.3390/pr13051300

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep learning models have been widely applied to tool wear prediction. Single-layer or single-type deep learning models can effectively extract spatial or temporal information but struggle to learn high-dimensional feature representations. As a result, many ensemble models have been developed to enhance robustness and generalization capabilities, leading to significant research achievements in this area. For instance, some researchers have proposed a state recognition method that combines continuous wavelet transform with an

improved MobileViT lightweight network to accurately, quickly, and efficiently identify milling tool wear states [1].

To predict tool wear more accurately during machining processes, a hybrid model combining variational mode decomposition (VMD) and backpropagation (BP) neural networks (VMD-BP) has been introduced [2]. Most current research focuses on complex signal processing techniques or advanced deep learning algorithms, prompting exploration into end-to-end data processing paradigms based on deep learning to improve prediction performance [3]. Zhao et al. proposed a convolutional bidirectional long short-term network to extract local features and encode temporal information, thereby improving tool wear monitoring [4]. Li et al. constructed a new physical information element learning framework for monitoring tool wear rates under different conditions [5]. To learn multi-scale features, Qiao et al. introduced multi-scale convolutional long-term memory models for tool wear monitoring [6], combined parallel CNNs with deep residual networks [7], and integrated parallel residual networks with stacked BiLsSTMs [8]. These models were validated under fixed operating conditions; however, in real machining processes, tool wear is influenced by various factors, including material properties, tool design, structure, and cutting parameters. To account for these conditions, some studies have fused machining information with extracted features as inputs to deep learning models [9]. Additionally, Karandikar et al. proposed a physics-guided logical classification method to model the nonlinear reduction of tool life with cutting speed [10]. In reality, variations in cutting signals caused by different working conditions are more directly related to tool wear. Therefore, there is an urgent need to enhance tool condition prediction methods tailored for the complex cutting conditions of aerospace titanium alloys.

Although models based on encoder-decoder structures have been validated on actual TC18 titanium alloy samples, significantly improving performance in predicting tool wear states, these models require large amounts of labeled data for training. Obtaining such data is challenging and costly in practical industrial applications. Furthermore, differences in data collected under various equipment and conditions limit the models' generalization capabilities. To address these issues, this paper proposes a tool wear state prediction model based on knowledge distillation. This model compression and transfer learning method enables knowledge transfer from a pre-trained teacher model to a student model, allowing the student model to learn the teacher model's discriminative capabilities with less data, thereby enhancing its performance in predicting tool wear states. It aims to solve the difficult and costly problem of obtaining large amounts of accurately labeled data in real industrial applications, while improving the generalization ability of the model. Knowledge distillation (KD) is a commonly used method for knowledge transfer, which extracts and transfers the knowledge contained in a pre-trained model (teacher model) to another model (student model) [11]. KD has been widely applied in natural language processing tasks, such as machine translation [12] and relation extraction [13].

In the research on tool wear state prediction, the knowledge distillation framework consists of three components: the teacher model, the student model, and the knowledge distillation loss function. The teacher model is a well-trained model that performs excellently on a specific dataset, while the student model learns knowledge from the teacher model through the distillation process. This paper focuses on constructing a tool wear state prediction model based on knowledge distillation. By defining the knowledge distillation loss function, the study analyzes the impact of the distillation method on different student models. It conducts comparative experiments on a self-constructed dataset for predicting the wear state of titanium alloy cutting tools, evaluates the performance of knowledge transfer models against existing superior methods, and proposes future research directions

and key areas for optimizing the comprehensive performance of tool wear state prediction models in titanium alloy machining processes.

The introductory part of this thesis focuses on the results and limitations achieved in the research on tool wear state prediction, and presents the research focus and contributions of this paper. The second part proposes a distillation-based network architecture for tool wear state prediction, in which BiGRU-GRU is used for the teacher model and the Transformer-GRU, Transformer-BiGRU, and Transformer-FC models are selected for the student model. The joint loss function is then defined. In the third part, the aerospace titanium alloy TC18 cutting tool default allowable wear experiments are carried out, the tool wear data samples are constructed, the data features are extracted and analysis is performed, and the knowledge distillation experimental test is carried out. In the fourth part, comparative experiments are carried out to verify the effectiveness of the proposed method. The fifth part is the conclusion of the paper.

2. Tool Wear State Prediction Encoder–Decoder Structure

In knowledge distillation, the types of knowledge, distillation strategies, and teacher–student architectures play a crucial role in the student's learning process. Traditional knowledge distillation methods use the logits of a large deep model as teacher knowledge [14]. Intermediate layer activation functions, neurons, or features can also serve as knowledge to guide the student model's learning [10]. The relationships between different activation functions, neurons, or data pairs contain rich information learned by the teacher model [15]. Additionally, the parameters of the teacher model (or the connections between layers) also contain another form of knowledge [16]. This paper will adopt response-based and feature-based knowledge distillation methods.

(1) Response-based knowledge distillation usually refers to mimicking the response of the output layer of the teacher model, i.e., directly learning the final prediction of the teacher model, and the method is simple, effective, and widely used. Given the depth model output vector y, its distillation loss is defined as:

$$L_{\mathrm{ResD}}(y_t, y_s) = L_R(y_t, y_s) \tag{1}$$

where L_{ResD} denotes the distillation loss of logits, and y_t and y_s are the logits of the teacher and the student, respectively. the most commonly used response-based knowledge distillation in classification tasks is known as soft targeting. The soft logit distillation loss can be rewritten as:

$$L_{\mathrm{ResD}}(p(y_t), p(y_s)) = L_R(p(y_t), p(y_s)) \tag{2}$$

Obviously, optimizing Equation (1) or Equation (2) matches student and teacher logits. The concepts are concise and easy to understand, and are especially suitable for tacit knowledge transfer. However, this type of distillation usually utilizes only the output layer and fails to provide the intermediate layer of supervisory information. Since soft logits are actually category probability distributions, response-based knowledge distillation is also limited to supervised learning.

(2) Deep neural networks implement representation learning through the abstraction of multi-level features, and thus can utilize the middle layer output (feature mapping) of the network to supervise student models. Feature-based knowledge distillation is an efficient extension of response-based distillation, and its loss function is usually expressed as:

$$L_{\mathrm{FeaD}}(f_t(x), f_s(x)) = L_F(\Phi_t(f_t(x)), \Phi_s(f_s(x))) \tag{3}$$

where $\phi_t(f_t(x))$ and $\phi_s(f_s(x))$ are the middle layer features of the teacher and student models, respectively. When the dimensions of the two are different, the transformation

function and can be utilized to adjust. Specifically, L2 paradigm distance, L1 paradigm distance, cross-entropy loss, and maximum mean difference loss can be used as similarity functions for the characterization of teacher and student model features.

2.1. Tool Wear State Prediction Knowledge Distillation Framework

The knowledge distillation-based tool condition monitoring method consists of three components: the teacher model, the student model, and the knowledge distillation loss function, as shown in Figure 1. Both the teacher and the student model are composed of a feature extraction module, a feature normalization module, an encoder module, a decoder module, and a tool wear state output module. The knowledge distillation loss function includes two functions: KL loss and distance loss.

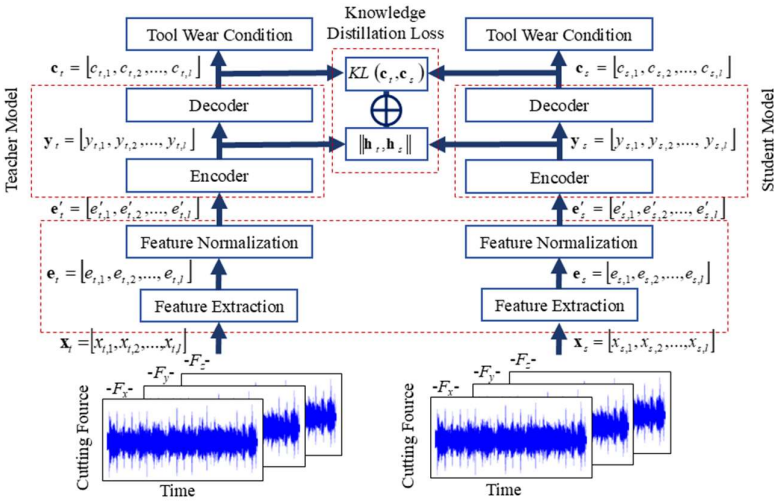


Figure 1. Framework of knowledge distillation method.

2.2. Knowledge Distillation Model Construction

(1) Construction of the Knowledge Distillation Teacher Model

The tool wear state assessment knowledge distillation teacher model selected in this paper is the best performing set of models for the PHM2010 dataset. Due to space limitation, the specific experimental results of the 72 combination models of the encoder and decoder on PHM2010 will not be elaborated on. Table 1 demonstrates the winning statistics of the 72 combinations of encoder and decoder on the PHM2010 dataset. Considering the performance on the PHM2010 dataset and the real dataset, the performance of existing commonly used neural networks, such as CNN-LSTM, is not as good as that of BiGRU-GRU without loss of generality. This paper chooses BiGRU-GRU as the teacher model, which is shown in Figure 2.

Table 1. Encoder and decoder win counts on PHM2010.

Number of Encoder Wins		Number of Decoder Wins	
Bi-GRU	465	GRU	359
GRU	426	Bi-GRU	314
RNN	384	Bi-LSTM	302
Bi-RNN	372	FC	299
Bi-LSTM	328	RNN	290
LSTM	279	LSTM	286
CNN	214	Bi-RNN	281
Transformer	35	CNN	198
		Transformer	174

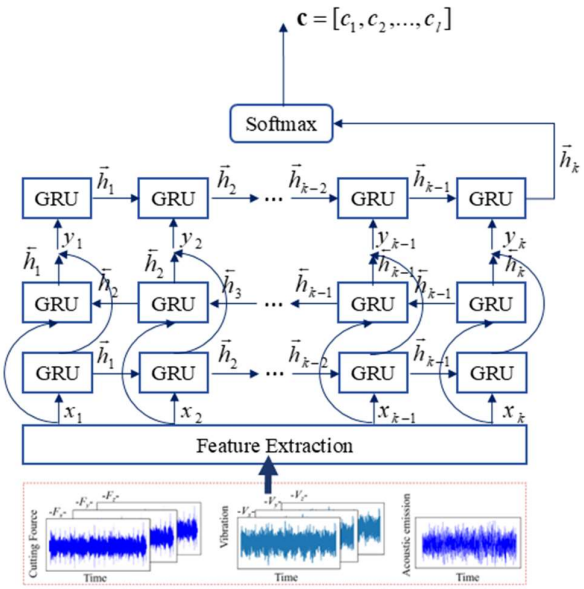


Figure 2. Knowledge distillation teacher model.

Considering the performance on the PHM2010 dataset and the real dataset, the performance of existing commonly used neural networks such as CNN-LSTM is not as good as that of BiGRU-GRU without loss of generality, and the BiGRU-GRU model is chosen as the teacher model in this paper. The teacher model is obtained from training on the PHM2010 dataset. The student model is trained on a self-collected dataset. The output vector of the tool wear state probability distribution from the teacher model $c_t = [c_{t,1}, c_{t,2}, \cdots, c_{t,l}]$, along with the tensor output from the encoder $h_t = [h_{t,1}, h_{t,2}, \cdots, h_{t,l}]$, serves as the supervisory signal to guide the training of the student model, thereby transferring knowledge from the teacher model to the student model.

(2) Construction of the Knowledge Distillation Student Model

The student model is trained using a self-constructed dataset. The output vector of the tool wear state probability distribution $c_t = [c_{t,1}, c_{t,2}, \cdots, c_{t,l}]$ from the teacher model, along with the tensor output $h_t = [h_{t,1}, h_{t,2}, \cdots, h_{t,l}]$ from the encoder, serves as the supervisory signal to guide the training of the student model, thereby transferring knowledge from the teacher model to the student model. For example, as shown in Figure 3, the Transformer-BiGRU architecture is used, where the Transformer structure acts as the encoder for the tool condition detection signals. The encoder's input is a preprocessed tensor of tool condition detection signals, consisting of a data sequence of length l. In the Transformer model, the query matrix (Q), key matrix (K), and value matrix (V) are computed using parameter matrices W_Q, W_K, and W_V, respectively, within the self-attention mechanism. Specifically, the matrix product of the query matrix Q and the key matrix K is calculated and scaled by a factor d_k to obtain the relevance scores for each position in the sequence. These scores are then transformed into a probability distribution using the Softmax function and multiplied by the corresponding value matrix V to produce the weighted output matrix A. The Transformer model employs a multi-head attention mechanism to generate multiple sets of matrix A, allowing for more diverse representations. These matrices are concatenated to form the final output. The output matrix A from the attention module is then fed into the feedforward network module. The output of the feedforward network is the output of the Transformer, which corresponds to the output y of the tool condition detection signal data encoder. Subsequently, the encoder's output y is fed into the decoder. For a tool condition feature encoding a tensor of length l, using a bidirectional gated recurrent unit (BiGRU) network as an example, the hidden layer at time t receives not only the current input but

also the output from the previous hidden layer. Ultimately, the output at the final time step l serves as the decoder's output, which is transformed into a probability distribution vector for the tool wear state via a Softmax layer. For the forward and backward networks at time k, their respective neural units take the forward output vector h_k, the backward output vector h_k, and the encoder's output vector as inputs, producing the forward vector h_k and the backward vector h_k as outputs. The outputs of the recurrent neural units from the BiGRU at the final time step are concatenated, and the resulting vector is passed through the Softmax layer to obtain the probability distribution vector for the tool wear state.

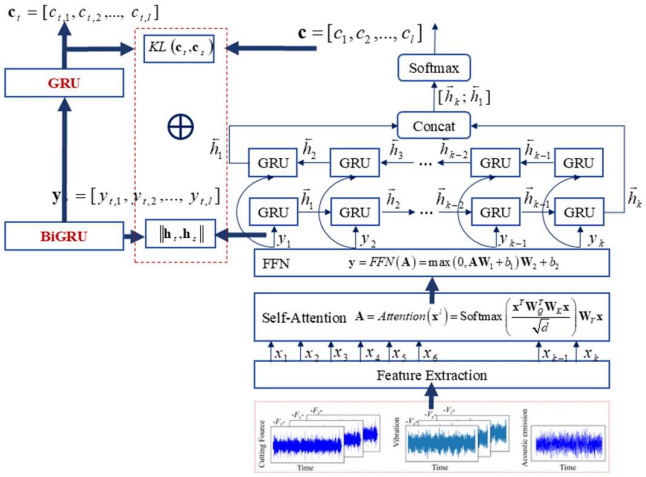


Figure 3. Knowledge distillation student model, taking Transformer-BiGRU as an example.

2.3. Definition of the Knowledge Distillation Loss Function

The teacher model and the student model both employ an encoder–decoder structure. The teacher model processes the input signal sequence $x_t = [x_{t,1}, x_{t,2}, \cdots, x_{t,l}]$, obtaining the feature representation $e'_{t,1} = \left[e'_{t,1}, e'_{t,2}, \cdots, e'_{t,n}\right]$ through feature extraction and normalization, which is then input into the encoder to produce the tensor $h_t = [h_{t,1}, h_{t,2}, \cdots, h_{t,n}]$. In contrast, the student model processes the input signal sequence $x_s = [x_{s,1}, x_{s,2}, \cdots, x_{s,n}]$, obtaining the tensor $h_s = [h_{s,1}, h_{s,2}, \cdots, h_{s,n}]$ after feature extraction, normalization, and encoding. Here, $h_{t,i}$ and $h_{s,i}$ represent the vector representations of the i-th feature for the teacher and student models, respectively. The outputs h_t and h_s are obtained from the decoders of the teacher and student models, yielding the tool wear state distribution vectors $c_t = [c_{t,1}, c_{t,2}, \cdots, c_{t,l}]$ and $c_s = [c_{s,1}, c_{s,2}, \cdots, c_{s,l}]$.

This paper employs both feature distillation and response distillation methods, defining the Euclidean distance loss function and the Kullback–Leibler divergence (KL divergence) loss function, respectively. By minimizing these loss functions, the transfer of the teacher model's encoder feature representation capability is achieved, while minimizing the KL divergence loss function effectively guides the student model to learn from the teacher model's outputs. Additionally, the cross-entropy loss function is used to measure the difference between the student model's output and the true tool wear state labels.

(1) Euclidean Distance Loss Function

Feature-based knowledge distillation [17–22] focuses more deeply on the intermediate layers of the model, using the Euclidean distance loss function as the optimization objective, as shown in Equation (4). Variables h_t and h_s represent the output tensors of the teacher model and student model encoders, respectively.

$$L_F(h_t, h_s) = \|h_t - h_s\| = \frac{1}{n}\sum_{i=1}^{n}(h_{t,i} - h_{s,i})^2 \quad (4)$$

(2) KL Divergence Loss Function

Response-based knowledge distillation allows the student model to directly mimic the output of the teacher model [21–26]. The tool wear state distribution probabilities are computed using the Softmax function, as shown in Equation (5):

$$p(c_i) = \frac{\exp(c_i)}{\sum_i \exp(c_j)} \tag{5}$$

In this context, c_i represents the i-th dimension of the model encoder's output vector corresponding to the i-th class of tool wear state labels, k is the total number of tool wear states, and *exp* denotes the exponential operation. The calculation formula for the response-based distillation loss is shown in Equation (6):

$$L_R(p(c_t), p(c_s)) = \sum_{i=1}^{k} p(c_{t,i}) \log \frac{p(c_{t,i})}{p(c_{s,i})} \tag{6}$$

(3) Cross-Entropy Loss Function

The cross-entropy loss function is used to measure the difference between the tool wear state output of the student model and the true labels, as shown in Equation (7). y_i is the tool wear state label, and $p(c_{s,i})$ is the probability predicted by the student model for the tool wear state $c_{s,i}$.

$$L_{CE} = -\sum_{i=1}^{k} y_i \log(p(c_{s,i})) \tag{7}$$

(4) Combined Loss Function

To balance the weights of the three components of the loss, the loss function is adjusted using weight parameters α_1 α_2, and α_3, with the condition that the sum of the weights satisfies $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

$$L_{distill} = \alpha_1 \cdot L_{CE} + \alpha_2 \cdot L_F + \alpha_3 \cdot L_R \tag{8}$$

3. Experiments

3.1. TC18 Titanium Alloy Machining Tool Wear Experiment

(1) Tool Cutting Experiment and Signal Acquisition

The dataset for tool wear states during the machining of aerospace titanium alloy is based on actual sample data collected from the established machining experiment platform for aerospace titanium alloys. As shown in Figure 4, the tool used is a carbide milling cutter. First, optimal machining parameters for the tools are selected through preliminary experiments; next, the formal cutting experiment plan and signal acquisition methods are determined.

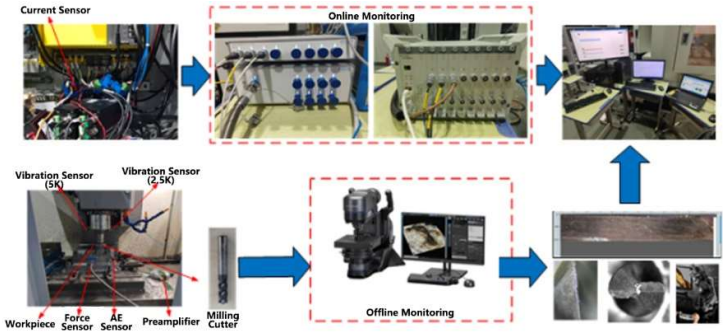


Figure 4. TC18 aerospace titanium alloy cutting experiment and signal acquisition platform.

Tool wear assessment includes direct and indirect measurements. In order to not affect or interfere with normal machining, monitoring of the machining process's indirect signal characteristics is the preferred choice for tool machining process state assessment and dynamic process optimization. According to the analysis of the previous section of the titanium alloy machining tool wear mechanism, due to the intense interaction between the tool and the workpiece, accompanied by heat and vibration and other phenomena, the assessment of these signals and mining analysis can reflect the state of the machining process, and then the degree of tool wear is assessed. Tool breakage due to the instantaneous aggregation of its process signal characteristics is more obvious, usually based on intense vibration or a sudden change in the time domain or frequency domain. Other analyses make it easy to assess and identify in a timely manner. Therefore, the tool wear assessment studied in this paper is mainly for the gradual change in the complex wear state. With the development of sensor technology, combined with the complexity of the working conditions at the processing site, and taking into account the feasibility of engineering applications and signal complementarity and the principle of exhaustive mining, this paper selects cutting force, vibration signal, and acoustic emission signal as the evaluated signal. The actual typical processing state of the tool state's real-time assessment requirements are ensured, along with a reasonable setting of the signal sampling frequency, after comprehensive consideration by the tool, with a rotation frequency of 5–10 times.

This study designed multiple sets of preliminary experimental parameters to obtain the optimal cutting parameters for tools in the machining of aerospace titanium alloys. Considering the interactions among various parameters and the complexity of the experimental conditions, the approach of orthogonal experimental design was referenced, reducing the 27 experimental sets to 9. The preliminary experimental groups are shown in Table 2.

Table 2. TC18 aerospace titanium alloy cutting experimental machining parameter.

Serial Number	Spindle Speed (r/min)	Feed Speed (mm/min)	Cutting Depth (mm)	Cutting Width (mm)
1	1000	240	0.6	4
2	1000	360	0.8	5
3	1000	480	1.0	6
4	1500	240	0.8	5
5	1500	360	0.6	6
6	1500	480	1.0	4
7	2000	240	1.0	5
8	2000	360	0.6	6
9	2000	480	0.8	4

Considering surface quality, tool wear, and machining efficiency, the fifth set of machining parameters was prioritized as the official cutting experiment parameters. Each machining pass was set to 300 mm, and various signals, including force, vibration, and acoustic emission, were collected for each pass. Additionally, the width of the wear land on the cutting edge was measured using a super-depth-of-field 3D microscope system (VHX-2000C). Based on the preliminary experiment results, each tool was set to complete 140 passes. Ultimately, in this paper, a total of 7 tools with full life cycle data were obtained, with 140 sets of data per tool, resulting in 840 valid data samples. The sample size was balanced across the phases (24.9% to 25.2% of the sample) and each tool contributed a similar percentage (T1–T7 each provided about 12% to 17% of the data), indicating that there was no category imbalance in the data.

(2) Data Feature Extraction and Partitioning Strategy

To address the randomness and non-stationarity of the collected signals, a wavelet threshold denoising method was employed. Through wavelet decomposition, threshold selection, and reconstruction, noise signals were effectively removed, enhancing the validity of the signals. The denoising effect was measured using the signal-to-noise ratio, allowing for the selection of suitable wavelet basis types and decomposition levels. Various key features were extracted through time domain, frequency domain, and time–frequency domain analyses, as shown in Table 3. Time domain features included absolute mean, peak value, root mean square, root amplitude, skewness, and kurtosis, as well as dimensionless indicators derived from these dimensional features, such as the waveform factor and pulse factor. These features reflect the characteristics of tool wear over time.

Table 3. Tool state extraction feature table.

Category	Multi-Domain Features				
Time domain	Absolute mean	Peak value	Root mean square	Root amplitude	Skewness
	Kurtosis	Waveform factor	Pulse factor	Kurtosis factor	Peak factor
	Margin factor				
Frequency domain	Centroid frequency	Mean square frequency	Root mean square frequency	Frequency variance	
Time–frequency domain	Number of bases in wavelet packet decomposition: 8				

Frequency domain features were extracted using Fourier transform, including centroid frequency, mean square frequency, root mean square frequency, and frequency variance, revealing the tool wear state from a frequency perspective. Time–frequency domain features were obtained through wavelet packet decomposition, allowing for fine decomposition of the signal across different frequency bands. Energy was used to represent the tool state information. Finally, principal component analysis (PCA) was utilized to perform dimensionality reduction on the features extracted from multiple domains, filtering out effective information closely related to the tool wear state, thereby enhancing the computational efficiency and performance of the model.

Based on this, the measured sample data were grouped according to two strategies: tool-based training and mixed training.

1. Tool-Based Training Strategy: This strategy involves grouping the collected data (from a total of 7 tools) by tool type. Specifically, the training set uses data from tools T1, T2, T3, T4, T5, and T6, while the test set uses data from tool T7. This strategy aligns well with the practical application environment of the tool wear state prediction model, which continuously monitors and predicts the state of a specific tool during its application.

2. Mixed Training Strategy: This strategy involves mixing data from multiple tools. Following the typical approach of training machine learning models, training and test data are randomly extracted in a 6:1 ratio. During this extraction process, care is taken to ensure that the data distribution for each tool's wear state is as uniform as possible. This strategy is a general data partitioning approach, making it suitable for model training, testing, and comparison with other methods.

3.2. Evaluation Method

The evaluation method employs precision (P), recall (R), and the F1 score (F1) using macro averaging. The calculation methods are outlined in Equations (9)–(12), where C represents the number of categories, which is set to 4, corresponding to the four wear states of the tools.

$$F1 = \frac{1}{C}\sum_{C=1}^{C} F1_C, P = \frac{1}{C}\sum_{C=1}^{C} P_C, R = \frac{1}{C}\sum_{C=1}^{C} R_C \quad (9)$$

$$F1_C = \frac{2 \times P_C \times R_C}{P_C + R_C} \quad (10)$$

$$P_C = \frac{TP}{TP + FP} \times 100\% \quad (11)$$

$$R_C = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

3.3. Knowledge Distillation Experiment and Analysis

(1) Model Hyperparameters and Experimental Environment

In the experiment, model training utilized a batch training strategy with a batch size set to 4. The Adam optimizer was used with default parameters, and the learning rate was selected from the set {0.0001, 0.0005, 0.001}. A dropout rate of 0.1 was employed to prevent overfitting. The software and hardware environment for the experiment is detailed in Table 4 below.

Table 4. Experimental environment.

Name	Configuration
CPU	Intel(R) Xeon(R) Gold 6230
GPU	NVIDIA GeForce RTX 3090Ti
Operating system	Ubuntu 18.04
Python	3.8
Platform	Pytorch 1.6
CUDA	10.1

(2) Knowledge Distillation Experiment

As shown in Table 5, a performance comparison between the knowledge distillation model and the model without knowledge distillation was conducted under different configurations in the tool grouping experiment. Knowledge distillation demonstrated performance improvements across all three student models, indicating its effectiveness within these model structures. Notably, the Transformer-GRU(S) model benefited the most from knowledge distillation, with significant increases in precision (P), recall (R), and F1 score. This suggests that knowledge distillation can serve as an effective model optimization method, particularly for the Transformer-GRU(S) model.

Table 5. Experimental results of knowledge distillation (tool grouping). ↑ represents an improvement compared to the no-teacher model. ↓ represents a decrease compared to the no-teacher model.

Teacher	Student	P	R	F1
BiGRU-GRU (T)	Transformer-GRU (S)	55.46% ↓	54.99% ↓	54.77% ↓
-	Transformer-GRU (S)	55.71%	56.01%	55.43%
BiGRU-GRU (T)	Transformer-BiGRU(S)	**57.12%** ↑	**57.18%** ↑	**56.75%** ↑
-	Transformer-BiGRU(S)	54.53%	55.00%	54.18%
BiGRU-GRU (T)	Transformer-FC(S)	52.86% ↓	52.80% ↓	52.19% ↓
-	Transformer-FC(S)	53.87%	53.81%	53.28%

As shown in Table 6, a performance comparison between the knowledge distillation model and the model without knowledge distillation was conducted under different configurations in the mixed grouping experiment. The data indicate that knowledge distillation demonstrated performance improvements across all three student models, highlighting its effectiveness within these model structures. Among them, the Transformer-GRU(S) model benefited the most from knowledge distillation, with significant increases in precision (P), recall (R), and F1 score. This suggests that knowledge distillation can serve as an effective model optimization method, particularly for the Transformer-GRU(S) model.

Table 6. Experimental results of knowledge distillation (mixed grouping). ↑ represents an improvement compared to the no-teacher model.

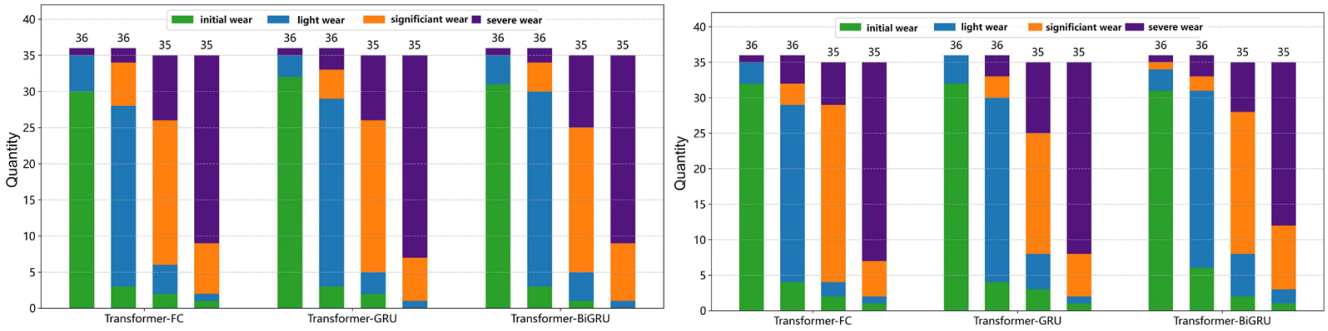
Teacher	Student	P	R	F1
BiGRU-GRU (T)	Transformer-GRU (S)	**72.55%** ↑	**72.67%** ↑	**72.36%** ↑
-	Transformer-GRU (S)	69.24%	69.31%	69.09%
BiGRU-GRU (T)	Transformer-BiGRU(S)	**70.66%** ↑	**70.27%** ↑	**70.12%** ↑
-	Transformer-BiGRU(S)	69.74%	69.82%	69.61%
BiGRU-GRU (T)	Transformer-FC(S)	**69.98%** ↑	**69.66%** ↑	**69.59%** ↑
-	Transformer-FC(S)	68.50%	68.24%	68.10%

Based on the aforementioned tool grouping and mixed grouping experiments, the following conclusions can be drawn:

Applicability of Knowledge Distillation: The effectiveness of knowledge distillation is highly dependent on the experimental grouping and the specific structure of the student models. There are significant differences in the effects of knowledge distillation under various experimental groupings and student model structures. In the mixed grouping experiment, knowledge distillation significantly improved the performance of all student models, particularly the Transformer-GRU(S) model. However, in the tool grouping experiment, knowledge distillation significantly enhanced the Transformer-BiGRU(S) model but had a negative impact on the Transformer-GRU(S) and Transformer-FC(S) models. This indicates that the applicability of knowledge distillation is greatly influenced by the experimental grouping and the specific structure of the student models, and it is evident that knowledge distillation is better suited for mixed grouping.

Robustness of Transformer-BiGRU Model: The Transformer-BiGRU model exhibited consistent positive effects from knowledge distillation in both the tool grouping and the mixed grouping experiments. This suggests that under the experimental conditions set in this study, the Transformer-BiGRU model is more robust in accepting transferred knowledge. This may be related to the decoding robustness of the bidirectional GRU results.

As shown in Figure 5, the classification results of the three models, Transformer-FC, Transformer-GRU, and Transformer-BiGRU, are analyzed in terms of the four wear levels (initial wear, slight wear, significant wear, and severe wear) by the hybrid grouping strategy, which gives the effect of the undistilled model versus the distilled model. The classification results for each model are presented in the form of stacked bar charts, where each bar chart represents one model and different colors in the bar charts represent different levels of wear, from which the following key conclusions can be drawn:



(**a**) Undistilled results (**b**) Distillation model results

Figure 5. Identification results of the four modes for different cutting tool wear states.

In terms of model comparison: Transformer-FC performed best in the classification of initial and severe wear. Transformer-GRU also performed well in the classification of initial and severe wear, but was slightly inferior to Transformer-FC in the classification of minor and significant wear. Transformer-BiGRU performed well in the classification of initial wear and severe wear, but slightly worse than Transformer-BiGRU in the classification of minor and significant wear.

In terms of wear stages, all models performed consistently in the classification of initial wear, indicating that these models have high accuracy in identifying initial wear. Transformer-FC and Transformer-GRU performed better in the classification of minor wear, while Transformer-BiGRU had a slightly lower number. Transformer-FC performed best in the classification of significant wear. All models performed consistently in the classification of severe wear, indicating that these models also have high accuracy in identifying severe wear.

In terms of knowledge distillation, comparing the effects of distilled and undistilled models, knowledge distillation improved the models' recognition at all stages. Especially for Transformer-BiGRU, there was an improvement in the initial wear stage and severe wear stage. In addition, the model had fewer samples misclassified to the initial wear stage in the latter three stages after distillation.

Furthermore, as illustrated in Figure 6, the chart data comprehensively reveal the confusion matrices for the Transformer-FC, Transformer-GRU, and Transformer-BiGRU models under the mixed grouping across four stages (A: initial wear, B: light wear, C: significant wear, D: severe wear). Misclassification between the categories of mild and significant wear was most common, and severe wear was also more frequently misclassified as mild or significant wear. After knowledge distillation, misclassification between the categories of mild and significant wear decreased and misclassification of significant wear improved. It also provides the accuracy for each stage and the overall accuracy, comparing the effects before and after distillation. From this figure, the following conclusions can be drawn:

General Performance Improvement: Knowledge distillation universally enhances model performance. After applying knowledge distillation, all models showed improved overall performance. This indicates that knowledge distillation, as a model training technique, can effectively enhance the learning ability and generalization capability of models. Transformer-FC performed best in classifying initial wear and severe wear. Transformer-GRU also performed well in classifying initial wear and severe wear, though it slightly lagged behind Transformer-FC in classifying light wear and significant wear. Transformer-BiGRU showed similar performance to the first two in classifying initial wear and severe wear, but its performance in classifying light wear and significant wear was slightly inferior.

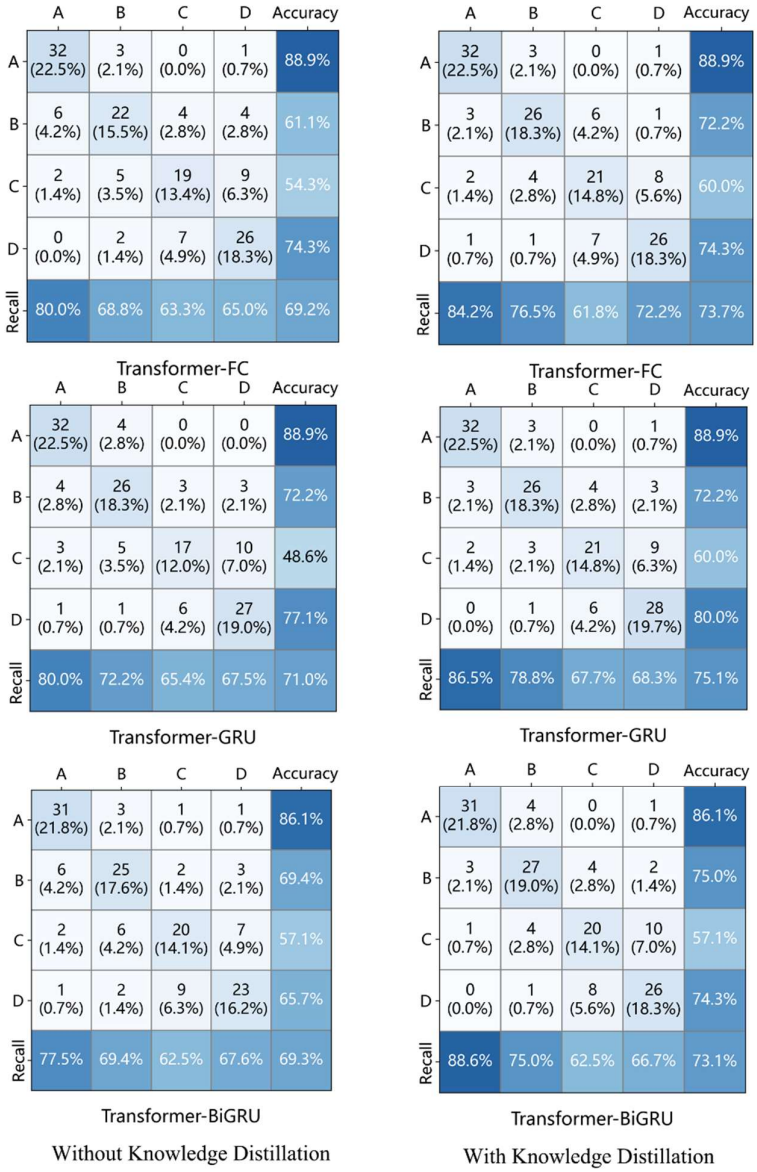


Figure 6. Classification results of the model at different stages of wear.

Significant Performance Improvement in Light, Significant, and Severe Wear Stages: Under the influence of knowledge distillation, the models demonstrated particularly noticeable performance improvement in the light wear, significant wear, and severe wear stages. For initial wear, all models performed consistently in the classification of initial wear.

This indicates that these models have high accuracy in recognizing initial wear. For light wear, both Transformer-FC and Transformer-GRU performed well in classification, while Transformer-BiGRU was slightly lower. In the case of significant wear, Transformer-FC excelled in classification. For severe wear, all models performed consistently, demonstrating high accuracy in identifying severe wear. This may suggest that the data for these categories were challenging to learn during the original training, and knowledge distillation, by imitating a superior model (the teacher model), helped the student models better capture the features of these categories. However, the performance for initial wear did not improve, possibly because the data for this category were already well learned by the model during the original training, leading to limited additional benefits from knowledge distillation. From a confusion perspective, the models showed reduced confusion for earlier classes, which aligns with the irreversible trend of wear stages developing linearly, but the

models did not significantly account for this property; hence, the confusion classification issue remains.

4. Comparative Experiments

To validate the overall performance improvement of the tool wear state prediction model due to knowledge distillation, this paper designed a comparative experimental scheme, as follows:

The effects of the Transformer-GRU(S), Transformer-BiGRU(S), and Transformer-BiGRU(S) knowledge distillation model were validated using actual cutting experiment data sets from aerospace titanium alloys.

GS-XGBoost [27]: A model based on Laplacian eigen-spectrum and ensemble learning methods. This model employs multi-algorithm feature filtering based on random forest (RF) and extreme gradient boosting (XGBoost), combined with the Laplacian Eigenmap (LE) algorithm for feature fusion and dimensionality reduction, and uses grid search (GS) to optimize the parameters of the XGBoost algorithm.

Attention-CNN [28]: A deep learning method based on channel-space attention mechanisms. It transforms cutting force signals into time–frequency images through continuous wavelet transforms, enabling feature extraction of cutting tool wear, and establishes a channel–space attention mechanism to assess the effectiveness of different cutting tool wear signal features.

PCNN-BiLSTM [29]: A deep learning model that combines a 1D convolutional network with a bidirectional long short-term memory network. It processes multi-dimensional features using a multi-path parallel 1D convolutional network, followed by a bidirectional LSTM, and employs a residual mechanism.

The experimental results are shown in Table 7 below, which provides the experimental results of the proposed knowledge distillation model in comparison with other methods. The mixed grouping is more suitable for knowledge distillation, and the overall performance of all models is improved after applying knowledge distillation. This indicates that knowledge distillation, as a model training technique, can effectively improve the learning ability and generalization of models. However, the applicability of knowledge distillation is highly dependent on the experimental grouping and the specific structure of the student models. It is not difficult to find that the knowledge distillation model is significantly better than other methods in both tool grouping and mixed grouping. It shows that the proposed knowledge distillation method can effectively extract the knowledge of tool wear discrimination from the teacher model to the student model, and improve the prediction effect of the student model due to insufficient data and differences in collection conditions.

Table 7. Comparative experimental results of each method.

Data Splitting Method	Model	P	R	F1
Tool grouping	Transformer-FC	52.86%	52.80%	52.19%
	Transformer-GRU	55.46%	54.99%	54.77%
	Transformer-BiGRU	**57.12%**	**57.18%**	**56.75%**
	GS-XGBoost	48.11%	52.67%	50.70%
	Attention-CNN	46.18%	45.64%	45.20%
	PCNN-BiLSTM	52.31%	51.88%	51.74%
Mixed grouping	Transformer-FC	69.98%	69.66%	69.59%
	Transformer-GRU	**72.55%**	**72.67%**	**72.36%**
	Transformer-BiGRU	70.66%	70.27%	70.12%
	GS-XGBoost	58.88%	57.34%	58.01%
	Attention-CNN	58.01%	58.09%	57.65%
	PCNN-BiLSTM	60.19%	60.03%	59.82%

5. Conclusions

In response to the demand for efficient tool wear state prediction in the processing of aerospace titanium alloy structural components, this paper proposes a knowledge distillation-based tool wear state prediction model. This model aims to enhance the feasibility and performance of encoder–decoder structured neural network models in predicting tool wear states, addressing the challenges of obtaining large amounts of accurately labeled data in practical industrial applications, which is often difficult and costly, while simultaneously improving the model's generalization capability.

The results indicate that in the mixed grouping experiment, knowledge distillation significantly improved the performance of all student models, particularly the Transformer-GRU(S) model. In the tool grouping experiment, knowledge distillation had a marked positive impact on the Transformer-BiGRU(S) model but negatively affected the Transformer-GRU(S) and Transformer-FC(S) models. This suggests that the applicability of knowledge distillation is highly dependent on the experimental grouping and the specific structure of the student models.

Furthermore, the results show that the knowledge distillation model outperformed other methods in both the tool grouping and the mixed grouping, effectively transferring tool wear discrimination knowledge from the teacher model to the student model. This alleviates issues arising from insufficient data and variations in collection conditions that lead to poor performance in the student models.

Although the proposed method significantly enhances model performance, the confusion matrix reveals that these distilled models sometimes misclassify severe wear stages as initial or light wear, which contradicts the continuous physical principles governing tool wear states. Therefore, effectively leveraging the physical information behind the data is a promising avenue for improving the predictive capabilities of neural network models. Future work could focus on developing physics-informed neural networks (PINNs) that incorporate the physical laws governing tool wear into the neural network, enabling auxiliary predictions based on physical principles and constraining the neural network outputs with physical information. This approach would enhance both the prediction accuracy and the compliance with physical principles, continuously improving model performance and reliability. Ultimately, this work aims to enhance the significant advantages and potential application value of tool wear state prediction models for aerospace titanium alloy machining, providing a theoretical foundation and valuable reference for subsequent research and the development of tool wear prediction models and systems tailored to actual cutting conditions.

Author Contributions: Conceptualization, B.L. (Bengang Liu) and W.W.; methodology, B.L. (Bengang Liu) and Z.D.; software, B.L. (Bengang Liu); validation, B.L. (Baode Li), B.X. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Science and Technology Key Program (grant number 2024ZD0712801).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors thank the Shenyang Institute of Computing Technology, Chinese Academy of Sciences, for the experimental support provided.

Conflicts of Interest: Authors Bengang Liu, Bo Xue, Bode Li and Zeguang Dong were employed by the Shenyang Aircraft Industry (Group) Co., Ltd.; Author Wenjiang Wu was employed by the Shenyang Zhongke Numerical Control Technology Co., Ltd. The remaining author declares that the

research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, S.; Li, M.; Gao, Y. Deep Learning Tool Wear State Identification Method Based on Cutting Force Signal. *Sensors* **2025**, *25*, 662. [CrossRef] [PubMed]
2. Wang, K.; Wang, A.; Wu, L. Research on Tool Wear Monitoring Technology Based on Variational Mode Decomposition and Back Propagation Neural Network Model. *Sensors* **2024**, *24*, 8107. [CrossRef] [PubMed]
3. Wang, C.; Wang, G.; Wang, T.; Xiong, X.; Ouyang, Z.; Gong, T. Exploring the Processing Paradigm of Input Data for End-to-End Deep Learning in Tool Condition Monitoring. *Sensors* **2024**, *24*, 5300. [CrossRef] [PubMed]
4. Zhao, R.; Yan, R.; Wang, J.; Mao, K. Learning to monitor machine health with convolutional Bi-Directional LSTM networks. *Sensors* **2017**, *17*, 273. [CrossRef] [PubMed]
5. Li, Y.; Wang, J.; Huang, Z.; Gao, R.X. Physics-informed meta learning for machining tool wear prediction. *J. Manuf. Syst.* **2022**, *62*, 17–27. [CrossRef]
6. Qiao, H.; Wang, T.; Wang, P. A tool wear monitoring and prediction system based on multiscale deep learning models and fog computing. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 2367–2384. [CrossRef]
7. Xu, X.; Wang, J.; Zhong, B.; Ming, W.; Chen, M. Deep learning-based tool wear prediction and its application for machining process using multi-scale feature fusion and channel attention mechanism. *Measurement* **2021**, *177*, 109254. [CrossRef]
8. Liu, X.; Liu, S.; Li, X.; Zhang, B.; Yue, C.; Liang, S.Y. Intelligent tool wear monitoring based on parallel residual and stacked bidirectional long short-term memory network. *J. Manuf. Syst.* **2021**, *60*, 608–619. [CrossRef]
9. Cai, W.; Zhang, W.; Hu, X.; Liu, Y. A hybrid information model based on long short-term memory network for tool condition monitoring. *J. Intell. Manuf.* **2020**, *31*, 1497–1510. [CrossRef]
10. Karandikar, J.; Schmitz, T.; Smith, S. Physics-guided logistic classification for tool life modeling and process parameter optimization in machining. *J. Manuf. Syst.* **2021**, *59*, 522–534. [CrossRef]
11. Sun, S.; Cheng, Y.; Gan, Z. Patient Knowledge Distillation for BERT Model Compression. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 4323–4332.
12. Gumma, V.; Dabre, R.; Kumar, P. An Empirical Study of Leveraging Knowledge Distillation for Compressing Multilingual Neural Machine Translation Models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, Tampere, Finland, 12–15 June 2023; European Association for Machine Translation: Tampere, Finland, 2023; pp. 103–114.
13. Zhu, F.; Chen, Y. A Knowledge Distillation Network Combining Adversarial Training and Intermediate Feature Extraction for Lane Line Detection. In Proceedings of the 2024 Australian & New Zealand Control Conference (ANZCC), Gold Coast, Australia, 1–2 February 2024; pp. 92–97.
14. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
15. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.
16. Liu, J.; Wen, D.; Gao, H.; Tao, W.; Chen, T.W.; Osa, K.; Kato, M. Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
17. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
18. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
19. Ahn, S.; Hu, S.X.; Damianou, A.; Lawrence, N.D.; Dai, Z. Variational information distillation for knowledge transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9163–9171.
20. Heo, B.; Lee, M.; Yun, S.; Choi, J.Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 3779–3787. [CrossRef]
21. Yu, L.; Yazici, V.O.; Liu, X.; Weijer, J.V.D.; Cheng, Y.; Ramisa, A. Learning metrics from teachers: Compact networks for image embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2907–2916.
22. Chen, G.; Choi, W.; Yu, X.; Han, T.; Chandraker, M. Learning efficient object detection models with knowledge distillation. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 742–751.
23. Kim, J.; Park, S.U.; Kwak, N. Paraphrasing complex network: Network compression via factor transfer. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 2765–2774.

24. Mirzadeh, S.I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 5191–5198. [CrossRef]
25. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
26. Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; Liang, J. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11953–11962.
27. Xie, Y.; Gao, S.; Zhang, C.; Liu, J. Tool wear state recognition and prediction method based on laplacian eigenmap with ensemble learning model. *Adv. Eng. Inform.* **2024**, *60*, 102382. [CrossRef]
28. Li, R.; Wei, P.; Liu, X.; Li, C.; Ni, J.; Zhao, W.; Zhao, L.; Hou, K. Cutting tool wear state recognition based on a channel-space attention mechanism. *J. Manuf. Syst.* **2023**, *69*, 135–149. [CrossRef]
29. Cheng, M.; Jiao, L.; Yan, P.; Jiang, H.; Wang, R.; Qiu, T.; Wang, X. Intelligent tool wear monitoring and multi-step prediction based on deep learning model. *J. Manuf. Syst.* **2022**, *62*, 286–300. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.