Geometry-Aware Visual Representation for Remaining Useful Life Prediction

Trung Hieu Vu[1], Tien Thanh Nguyen[1], and Eyad Elyan[1]

Robert Gordon University, Aberdeen, United Kingdom
{h.vu1, t.nguyen11, e.elyan}@rgu.ac.uk

Abstract. Accurate Remaining Useful Life (RUL) prediction is fundamental to Prognostics and Health Management, yet remains challenging due to stochastic degradation and non-stationary vibration patterns. Existing image-based approaches predominantly rely on time–frequency representations (e.g., spectrograms or wavelet transforms), which capture spectral energy variations but overlook the intrinsic geometric structure of system dynamics. In this work, we reformulate RUL prediction through a phase-space-inspired visual representation. We introduce the Phase Space Density Image (PSDI), a novel representation that encodes the spatial density of time-delay embedded trajectories. Unlike spectral heatmaps, PSDI characterizes degradation as a progressive geometric dispersion of the system attractor, revealing structural transitions invisible to conventional representations. To ensure that these geometric changes reflect true degradation rather than coordinate drift, we further propose a Globally Anchored Reconstruction strategy that enforces consistent phase-space alignment across samples and time. We evaluate the effectiveness of the proposed representation across multiple vision architectures and integrate it with a compact knowledge distillation framework that transfers useful visual priors while reducing prediction jitter through temporal aggregation. Experiments on benchmark datasets demonstrate that PSDI achieves competitive and robust performance compared with classical signal representations and state-of-the-art time-series models, supporting phase-space density maps as an effective visual representation for vibration-based degradation tracking.

Keywords: remaining useful life · forecasting · time series · phase space reconstruction · knowledge distillation

1 Introduction

The prognosis of Remaining Useful Life (RUL) for critical mechanical components is fundamental to the reliability of modern industrial systems. Despite rapid progress in deep learning architectures, ranging from recurrent neural networks to Transformer-based architectures, accurate RUL prediction remains challenging. The difficulty arises from the stochastic nature of degradation and the non-linear, chaotic dynamics that govern mechanical systems [9]. Raw vibration signals often obscure the underlying physical health states, as they represent

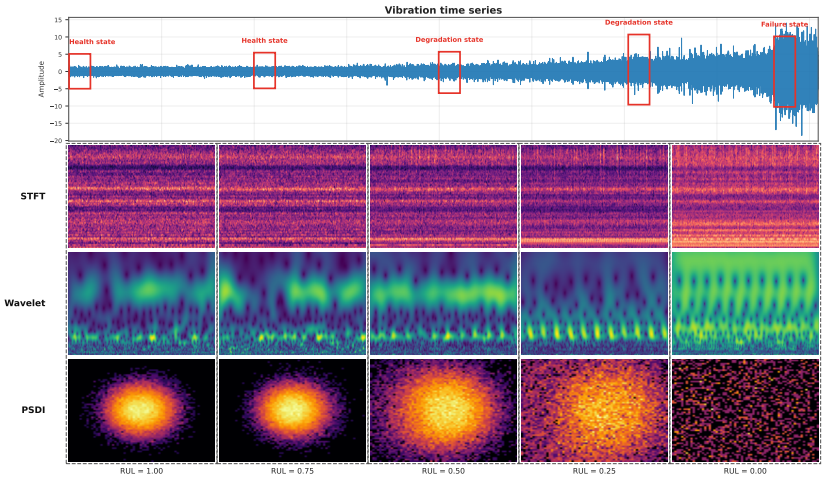

Fig. 1: Signal representations from healthy to failure states. Unlike time-domain and time–frequency views, PSDI reveals a clear geometric dispersion in phase space as degradation progresses.

complex dynamical processes rather than simple linear trends. This suggests that the limitation may not solely lie in architectural capacity, but raises a fundamental question: *To what extent do existing models reflect the underlying progression of the degradation process?*

A closer inspection of vibration signals provides an important insight. As illustrated in Figure 1, the raw waveform shows increasing amplitude and variance as failure approaches, while its local oscillatory pattern remains largely similar across health states. Time–frequency analysis techniques such as the short-time Fourier transform (STFT) and wavelet transform are widely used to analyze non-stationary signals because they reveal how spectral energy evolves over time. These methods decompose measured signals into localized time–frequency components and are commonly used to extract diagnostic features from vibration data [7]. However, since they operate directly on measured signals, they mainly characterize variations in signal intensity and spectral content rather than the underlying structural dynamics or degradation mechanisms governing system health [16]. In contrast, phase-space reconstruction offers a different perspective. Through time-delay embedding, the signal is lifted into a higher-dimensional state space that approximates the system's underlying dynamics [17]. In this space, degradation appears as a structural transition: trajectories progress from a compact and concentrated pattern in healthy conditions to increasingly dispersed and fragmented distributions near failure. This suggests that degradation is better characterised by changes in state-space trajectory structure rather than by amplitude or spectral variations alone, highlighting a limitation of approaches that treat vibration signals purely as temporal sequences.

Motivated by this insight, we reinterpret RUL prediction as learning from the geometric evolution of reconstructed system states. Building upon Takens' embedding theorem [17], we reconstruct phase-space trajectories from time-delayed observations, providing a structured view of the dynamical behaviour underlying measured vibration signals. We then propose PSDI, a visual representation that encodes the spatial density of trajectories in the reconstructed state space. Under a shared healthy reference frame, PSDI produces coordinate-consistent density images that expose degradation-related concentration, dispersion, fragmentation, and drift patterns. From a computer vision perspective, these patterns can be naturally processed by standard visual inductive biases such as convolutional locality and patch-based attention. Our goal is therefore not to design a new vision backbone, but to provide a degradation-aware visual representation that makes vibration-based RUL prediction more compatible with existing vision architectures. To the best of our knowledge, this is the first work to formulate vibration-based RUL prediction using coordinate-consistent phase-space density images. Our contributions can be summarized as follows:

- We introduce Phase-Space Density Image (PSDI), a phase-space-inspired visual representation that encodes the spatial density of time-delay embedded trajectories, enabling vision backbones to learn degradation-related patterns from vibration signals.
- To mitigate coordinate drift in phase-space reconstruction, we propose Globally Anchored Reconstruction (GAR), which uses a shared healthy reference frame to improve the comparability of PSDI representations over degradation time.
- We extensively evaluate PSDI across diverse vision architectures and develop a compact knowledge distillation framework. Benchmark results demonstrate competitive and robust performance compared with state-of-the-art time-series forecasting and RUL prediction methods, validating coordinate-consistent density maps as an effective representation for degradation tracking.

2 Related Work

Transforming time series into 2D images allows researchers to harness high-capacity vision architectures for prediction tasks. The prevailing approach relies on time-frequency heatmaps generated via STFT [6] or Wavelet Transforms [4]. Recent works have successfully integrated these representations with advanced vision models; for example, Zeng et al. [23] employ spectrograms within a multimodal Vision Transformer, and Namura et al. [13] utilize wavelet images for anomaly detection via foundation models. Beyond spectral methods, structural transformations like Recurrence Plots (RP) [5] and Gramian Angular Fields (GAF) [21] attempt to preserve temporal dependencies by mapping state recurrences or polar correlations (e.g., Barra et al. [1] for financial forecasting).

Driven by the success of these 2D representations, deep temporal modeling has increasingly shifted from traditional 1D sequence models toward vision-centric paradigms [22]. A prominent trend treats time series as structured im-

ages to leverage 2D convolutional priors. For instance, structural methods like TimesNet decomposes time series into multi-periodic 2D representations, exposing periodic structure that is difficult to model in the 1D temporal domain [22]. TimeMixer++ extends this direction by performing multi-scale mixing operations across both temporal and feature dimensions, enabling richer cross-scale interactions [19]. A parallel line of work explores model-level adaptation by transferring knowledge from vision models pre-trained on natural images. VisionTS demonstrates the effectiveness of Masked Autoencoders (MAE) applied to grayscale time-series images, showing that self-supervised visual pre-training can generalize to temporal prediction tasks [3]. TimeVLM takes this further by incorporating pre-trained Vision-Language Models (VLMs) into a multimodal forecasting framework, aligning temporal, visual, and textual modalities for improved forecasting performance [24]. Despite their success, these methods primarily focus on periodic or spectral patterns for generic forecasting. In contrast, Remaining Useful Life (RUL) prediction requires capturing non-stationary degradation dynamics. Existing transformations often overlook the underlying geometric structure of reconstructed dynamical trajectories, failing to preserve the evolution of system states. This motivates our exploration of geometry-consistent visual representations specifically tailored for degradation modeling.

3 Methodology

Given the historical observations up to time $t \in \{1, \ldots, T\}$, the objective is to learn a parametric predictor:

$$\hat{y}_t = f_\theta(\mathbf{X}_t, \mathbf{P}_t) \tag{1}$$

where $\mathbf{X}_t \in \mathbb{R}^{C \times H \times W}$ denotes the 2D Phase-Space Density Image (PSDI) with C channels and spatial size $H \times W$, and $\mathbf{P}_t \in \mathbb{R}^{C \times P}$ denotes the corresponding 1D trajectory signal. The predictor f_θ estimates the true normalised RUL $y_t \in [0, 1]$.

3.1 Geometry-Aware Representation

Phase Space Reconstruction. In many real-world monitoring scenarios, including vibration-based condition assessment, the observed signal is a one-dimensional projection of an underlying high-dimensional nonlinear process. Such scalar measurements provide only partial access to the latent system dynamics. To recover this hidden structure, we employ Phase Space Reconstruction (PSR), which lifts the scalar time series into a higher-dimensional embedding space. For a chosen time delay $\tau_d \in \mathbb{N}^+$ and embedding dimension m, let $T = N - (m-1)\tau_d$ denote the valid trajectory length. Each delay vector is constructed as a row vector $r_i = [x_i, x_{i+\tau_d}, \ldots, x_{i+(m-1)\tau_d}] \in \mathbb{R}^{1 \times m}$. Stacking all T delay vectors vertically

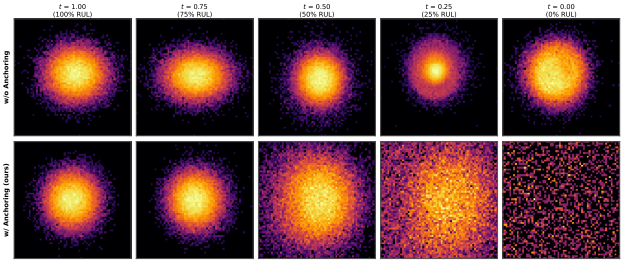

Fig. 2: Comparison of PSDI evolution with and without anchoring.

yields the reconstructed phase space matrix:

$$R = \begin{bmatrix} x_1 & x_{1+\tau_d} & \dots & x_{1+(m-1)\tau_d} \\ x_2 & x_{2+\tau_d} & \dots & x_{2+(m-1)\tau_d} \\ \vdots & \vdots & \ddots & \vdots \\ x_T & x_{T+\tau_d} & \dots & x_N \end{bmatrix} \in \mathbb{R}^{T \times m} \tag{2}$$

This construction yields $N - (m-1)\tau_d$ state vectors in an m-dimensional embedding space, forming a reconstructed trajectory that provides a phase-space view of the signal dynamics. Under standard embedding assumptions, it preserves relevant dynamical structure up to a smooth coordinate transformation, complementary to time- and frequency-domain views.

Globally Anchored Reconstruction. While Takens' theorem [17] preserves the underlying system dynamics, it does not guarantee a unique coordinate system. Consequently, if each sample estimates its own projection basis, it establishes an independent coordinate space. This arbitrary basis choice introduces severe coordinate drift, confounding actual physical changes with random geometric transformations. As illustrated in Figure 2 (top), unanchored embeddings drift unpredictably across stages, rendering healthy and fault states visually indistinguishable. To prevent this coordinate drift, we propose a globally anchored reconstruction that enforces a shared reference frame (outlined in Algorithm 1). This anchoring yields a stable PSDI evolution (Figure 2, bottom), successfully exposing true degradation as a progressive spatial dispersion.

Global Anchor Estimation. We build a reference frame using early-life (healthy) data. Given a healthy subset $\mathcal{D}_{ref} \subset \mathcal{D}_{train}$, we estimate the delay τ (via Mutual Information minimization [8]) and the embedding dimension m (via Cao's method [2]) for each healthy sequence. We then take the median across sequences to obtain global parameters (m^*, τ^*), which improves robustness to noisy trials and outliers. By fixing (m^*, τ^*), all samples are reconstructed in the same phase-space dimension. Next, we delay-embed all healthy signals to form a pooled reference set $\mathcal{R}_{ref} \in \mathbb{R}^{(N \cdot T) \times m^*}$, in which $N = |\mathcal{D}_{\text{ref}}|$. We compute a global

Algorithm 1 Globally Anchored PSR-Based Representation

Require: Training set $\mathcal{D}_{train}$, healthy reference subset $\mathcal{D}_{ref} \subset \mathcal{D}_{train}$, sample $\mathbf{x}_{b,t} \in \mathbb{R}^{C\times N}$

Ensure: 2D phase-space image $\mathbf{X}_{b,t} \in \mathbb{R}^{C\times H_{\text{img}}\times W_{\text{img}}}$

Global Anchor Estimation *(Computed once from healthy reference data)*

Estimate global embedding parameters (τ^*, m^*): (see Appendix for details)
$\tau^* \leftarrow \text{Median}(\textsc{EstimateDelay\_MI}(\mathcal{D}_{ref}))$
$m^* \leftarrow \text{Median}(\textsc{EstimateDim\_Cao}(\mathcal{D}_{ref}))$

Reconstruct phase-space trajectories based on Equation 2 for all healthy signals:
$\mathcal{R}_{ref} \leftarrow \{\textsc{DelayEmbed}(x; m^*, \tau^*) \mid x \in \mathcal{D}_{ref}\}$

Calculate global centroid and fit a 2D PCA basis on centered reference trajectories:
$\boldsymbol{\mu} \leftarrow \textsc{MeanPoints}(\mathcal{R}_{ref})$
$\mathbf{U} \leftarrow \text{PCA}_{2D}(\{R - \mathbf{1}_T\boldsymbol{\mu} \mid R \in \mathcal{R}_{ref}\})$

Project all healthy trajectories into 2D phase space:
$\mathcal{Z}_{ref} \leftarrow \left\{ (R - \mathbf{1}_T\boldsymbol{\mu})\,\mathbf{U}^\top \mid R \in \mathcal{R}_{ref} \right\}$

Compute bin edges from healthy 2D distribution:
$\mathcal{B} \leftarrow \textsc{QuantileBounds}(\mathcal{Z}_{ref})$

Anchored Dual Representation *(Using anchors $\boldsymbol{\mu}$, $\mathbf{U}$, $\mathcal{B}$ from Phase 1)*

Initialize $\mathbf{X}_{b,t} \leftarrow \emptyset$
for $c \leftarrow 1$ to C **do**
  $R^{(c)} \leftarrow \textsc{DelayEmbed}\left(\mathbf{x}_{b,t}^{(c)};\, m^*, \tau^*\right)$
  $Z^{(c)} \leftarrow \left(R^{(c)} - \mathbf{1}_T\boldsymbol{\mu}\right)\mathbf{U}^\top \in \mathbb{R}^{T\times 2}$ ▷ Project onto global healthy PCA basis
  $H^{(c)} \leftarrow \textsc{Hist2D}\left(Z^{(c)}, H_{\text{img}}, W_{\text{img}}, \mathcal{B}\right)$ ▷ 2D density histogram using fixed bin bounds
  $I^{(c)} \leftarrow \textsc{MinMaxNorm}\left(\log(1 + H^{(c)})\right)$ ▷ Log-compress and normalise to $[0, 1]$
  Append $I^{(c)}$ to $\mathbf{X}_{b,t}$
end for
return $\mathbf{X}_{b,t}$

centroid $\boldsymbol{\mu} \in \mathbb{R}^{1\times m^*}$ and learn a 2D PCA basis $\mathbf{U} \in \mathbb{R}^{2\times m^*}$ from the centered reference points. This basis serves as a shared low-dimensional coordinate system anchored to healthy dynamics. Crucially, $(\boldsymbol{\mu}, \mathbf{U})$ is learned once from healthy data and then frozen for all training and inference samples, ensuring that observed geometric changes in the projected trajectories reflect dynamical evolution rather than coordinate re-estimation.

After projection, each healthy trajectory $R \in \mathcal{R}_{ref}$ is mapped to a 2D trajectory $Z = (R - \mathbf{1}_T\boldsymbol{\mu})\mathbf{U}^\top \in \mathbb{R}^{T\times 2}$, where $\mathbf{1}_T \in \mathbb{R}^{T\times 1}$ is a column vector of ones used to broadcast the row vector $\boldsymbol{\mu} \in \mathbb{R}^{1\times m^*}$. We denote the set of all projected healthy trajectories as $\mathcal{Z}_{ref} = \{(R - \mathbf{1}_T\boldsymbol{\mu})\mathbf{U}^\top | R \in \mathcal{R}_{ref}\} \in \mathbb{R}^{(N\cdot T)\times 2}$. We define the histogram boundaries $\mathcal{B}$ from the empirical distribution of $\mathcal{Z}_{ref}$ using quantile-based limits rather than min-max ranges. Specifically, for each projected coordinate, we compute the lower and upper bounds as empirical quantiles (e.g., 1% and 99%) of the healthy projection distribution. This strategy reduces sensitivity to a small number of extreme points and ensures that the spatial support

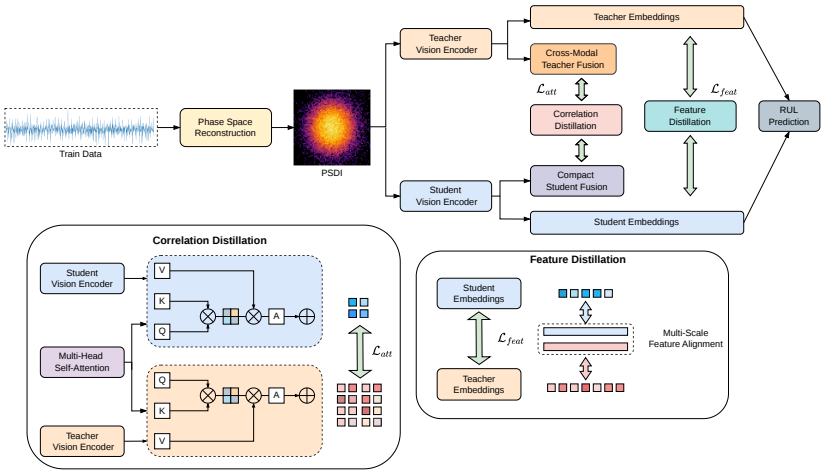

Fig. 3: Overview of the Knowledge Distillation framework

reflects the typical region of healthy dynamics. As a result, the density images are more stable and geometrically consistent across samples.

Anchored Dual Representation. Given a new input sample $\mathbf{x}_{b,t} \in \mathbb{R}^{C\times N}$, the frozen anchors $(\boldsymbol{\mu}, \mathbf{U}, \mathcal{B})$ are applied to each channel independently. The signal is delay-embedded using (m^*, τ^*), centered by subtracting $\mathbf{1}_T\boldsymbol{\mu}$, and projected onto the healthy PCA basis to produce a 2D trajectory. A phase-space density image is then computed by counting how frequently the trajectory visits each spatial bin defined by $\mathcal{B}$. To reduce the large dynamic range caused by highly visited regions, we apply a logarithmic compression $\log(1+H)$ to the histogram counts, which enhances low-density structures while preventing dominant regions from overwhelming the image. Finally, the result is normalized to $[0,1]$ to ensure consistent input scaling across samples. The first principal component (PC1) represents the dominant direction of variation in healthy dynamics. By extracting and resampling this coordinate over time, we obtain a compact 1D descriptor that preserves the temporal evolution of the system. Unlike the density image, which summarizes spatial visitation frequency without explicit temporal ordering, the PC1 trajectory retains time structure. Since both representations are derived from the same projected trajectory, they remain strictly synchronized and reflect consistent underlying dynamics.

3.2 Knowledge Distillation Framework

Figure 3 provides an overview of the proposed knowledge distillation framework. The KD framework is motivated by two factors. First, high-capacity teachers pretrained on natural images provide transferable visual priors but may also carry semantic redundancy irrelevant to PSDI degradation patterns. Second, framewise RUL predictions are sensitive to vibration noise and short-term transients,

causing local jitter. The student is thus not a mere compressed copy of the teacher: it learns PSDI-relevant cues via feature and attention distillation, while sliding-window aggregation suppresses short-term fluctuations and improves stability, explaining why a compact student can match or exceed the teacher.

Teacher Encoder. Given a PSDI input $\mathbf{X}_t \in \mathbb{R}^{C\times H\times W}$, we convert it to 3 channels for compatibility with pretrained vision backbones. The teacher encoder f_{θ_T} is instantiated using a high-capacity pretrained vision backbone. It outputs a global feature $\mathbf{h}_t \in \mathbb{R}^{D_T}$ (e.g., CLS token for ViT-based teachers), and the RUL head predicts $\hat{y}_t$ obtained on input $\mathbf{h}_t$.

The teacher is trained with SmoothL1 regression loss:

$$\mathcal{L}_{\text{teacher}} = \text{SmoothL1}(\hat{y}_t, y_t). \tag{3}$$

In our default setting, the pretrained teacher backbone is frozen and only the regression head is optimized. For ViT-based teachers, patch tokens are additionally retained for downstream attention distillation.

Student Encoder. We adopt a compact vision encoder E_s as the student backbone to encode PSDI images, achieving substantial parameter reduction. Given a PSDI frame $\mathbf{X}_t \in \mathbb{R}^{C\times H\times W}$, the image is tokenized into N_v non-overlapping patches and embedded into a sequence of visual tokens. The encoder produces:

$$[\mathbf{c}_t,\ \mathbf{I}_{\text{vis}}] = E_s(\mathbf{X}_t) \tag{4}$$

where $\mathbf{c}_t \in \mathbb{R}^{d_s}$ denotes the global (CLS) representation and $\mathbf{I}_{\text{vis}} \in \mathbb{R}^{N_v\times d_s}$ denotes the sequence of visual patch embeddings. Concurrently, the 1D trajectory signal $\mathbf{P}_t \in \mathbb{R}^{C\times P}$ is embedded into trajectory tokens $\mathbf{Q}_p \in \mathbb{R}^{N_p\times d_p}$. Compared to the teacher backbone E_t, the student model is designed with substantially smaller capacity (e.g., fewer layers or lower dimensionality). This architecture encourages the model to focus on degradation-relevant geometric cues in PSDI while suppressing semantic redundancy and overfitting noise.

Sliding-Window Temporal Aggregation. Frame-wise RUL prediction is sensitive to stochastic measurement noise and short-term mechanical transients. Directly mapping individual frames causes non-monotonic prediction jitter. To stabilize representations and filter out spurious noise, we perform temporal aggregation over a sliding window of K consecutive PSDI frames. Given the extracted features $\mathbf{c}_\tau$ from the shared student encoder, we compute the temporal average:

$$\bar{\mathbf{c}}_t = \frac{1}{K}\sum_{\tau=t-K+1}^{t} \mathbf{c}_\tau. \tag{5}$$

The aggregated feature $\bar{\mathbf{c}}_t$ is then forwarded to downstream modules for RUL prediction. From a signal processing perspective, this aggregation acts as a low-pass filter. We formally establish its variance reduction property as follows:

Proposition 1 (Variance Reduction). *Assuming locally stationary degradation, independent measurement noise, and a Lipschitz-continuous prediction head, the variance of the predicted RUL, $\hat{y}_t = g(\bar{\mathbf{c}}_t)$, is strictly bounded inversely by the window size K:*

$$\mathrm{Var}(\hat{y}_t) \leq \mathcal{O}\left(\frac{1}{K}\right). \tag{6}$$

The formal assumptions and complete proof are deferred to Appendix. This confirms that temporal aggregation effectively mitigates high-frequency noise.

Multi-Scale Feature Alignment. To bridge the dimensional gap between teacher and student representations, we adopt a pyramid-style aligner that combines multiple scale-dependent projection functions to capture information at different granularities. Given the aggregated student feature $\bar{\mathbf{c}}_t \in \mathbb{R}^{d_s}$, the aligned student feature $\hat{\mathbf{c}}_t \in \mathbb{R}^{d_t}$ is computed as:

$$\hat{\mathbf{c}}_t = \sum_{i=0}^{S} \alpha_i \, \psi_i(\bar{\mathbf{c}}_t), \tag{7}$$

where each projection function $\psi_i : \mathbb{R}^{d_s} \to \mathbb{R}^{d_t}$ operates with a scale-specific hidden width $r_i = \max(d_s, d_t)/2^i$, $\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{\theta})$ are learnable scale weights with $\boldsymbol{\theta} \in \mathbb{R}^{S+1}$, and S denotes the number of scale levels. It is recognised that $\mathbf{I}_{\mathrm{vis}}$ captures the spatial density of reconstructed states but does not preserve temporal ordering, whereas $\mathbf{Q}_p$ encodes trajectory evolution. To combine modalities, we apply cross-attention to inject temporal dynamics into each PSDI patch, producing trajectory-enriched visual tokens that adaptively highlight degradation-relevant regions (see Appendix for details).

Multi-Component Loss Integration. To transfer geometry-aware representations from a high-capacity teacher to a compact student, we formulate distillation as a multi-level structured alignment problem. Instead of merely matching output predictions, our objective enforces consistency at both feature and attention levels, while jointly optimizing the student for RUL prediction. Moreover, unlike conventional knowledge distillation approaches that rely on manually fixed parameters, our framework treats distillation weights and feature alignment weights as learnable weight components. We integrate several supervision signals to regulate the interaction between task optimization and representation transfer.

Feature Distillation. This component aligns the student representations with the teacher's privileged embeddings, transferring representation-level knowledge via multiple similarity metrics:

$$\mathcal{L}_{\mathrm{feat}} = \lambda_{\mathrm{mse}}\mathcal{L}_{\mathrm{mse}} + \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \tag{8}$$

where $\mathbf{h}_s^{(i)} \equiv \hat{\mathbf{c}}_t^{(i)}$ denotes the aligned student global feature for sample i, and $\mathcal{L}_{\mathrm{mse}} = \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{h}_s^{(i)} - \mathbf{h}_t^{(i)}\|_2^2$ encourages direct point-wise alignment between

student and teacher features. Furthermore, $\mathcal{L}_{\cos} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \cos(\mathbf{h}_s^{(i)}, \mathbf{h}_t^{(i)})\right)$ enforces directional consistency in the embedding space. $\mathcal{L}_{\mathrm{KL}}$ matches the softened feature distributions via temperature-scaled divergence:

$$\mathcal{L}_{\mathrm{KL}} = T_{\mathrm{KL}}^2 D_{\mathrm{KL}}\left(\mathrm{softmax}\left(\frac{\mathbf{h}_t}{T_{\mathrm{KL}}}\right)\middle\|\mathrm{softmax}\left(\frac{\mathbf{h}_s}{T_{\mathrm{KL}}}\right)\right), \tag{9}$$

where T_{KL} is the distillation temperature.

Correlation Distillation. To transfer the teacher's attention patterns, we introduce an auxiliary distillation loss. Let $\mathbf{A}_t^{(i)}$ and $\mathbf{A}_s^{(i)}$ denote the teacher and student attention logits for sample i. We use temperature-scaled KL alignment:

$$\mathcal{L}_{\mathrm{att}} = \frac{T_a^2}{B}\sum_{i=1}^{B} D_{\mathrm{KL}}\left(\sigma\left(\frac{\mathbf{A}_t^{(i)}}{T_a}\right)\middle\|\sigma\left(\frac{\mathbf{A}_s^{(i)}}{T_a}\right)\right), \tag{10}$$

where $\sigma(\cdot)$ is the softmax function and T_a controls distribution smoothness.

Student Total Objective. The combined distillation objective is $\mathcal{L}_{\mathrm{distill}} = \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{att}}\mathcal{L}_{\mathrm{att}}$. During training, the student model minimizes:

$$\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{rul}} + \lambda_{\mathrm{distill}}\mathcal{L}_{\mathrm{distill}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{rul}} = \mathcal{L}_{\mathrm{SmoothL1}}(\hat{y}_t, y_t)$ denotes the student's RUL regression loss, matching the predicted RUL $\hat{y}_t$ to the ground truth y_t. The coefficient $\lambda_{\mathrm{distill}}$ balances RUL prediction against cross-modal knowledge distillation.

4 Experiments

The experiments are designed to demonstrate (i) the performance of the framework with different teacher–student backbone combinations, (ii) the effectiveness of PSDI relative to other representations, (iii) the advantage of the proposed distillation framework over alternative architectures, and (iv) the impact of incorporating anchoring compared to a non-anchored setting.

4.1 Experimental Settings

Datasets and Evaluation Metrics. We evaluated the proposed distillation framework for Remaining Useful Life (RUL) prediction on two widely-used bearing benchmarks: XJTU-SY [18] and the PHM bearing dataset [14]. Following prior work [10], we reported performance using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Full dataset statistics, preprocessing, and the precise metric definitions are provided in the Appendix.

Case studies. We evaluated our method across five cases divided into two categories. Category 1 (Cases 1–3) maintains identical bearing types and working parameters for both training and testing. Category 2 (Cases 4–5) utilizes the same bearing type but under different operational conditions. We employed a cross-bearing validation scheme, training on a subset and testing on an unseen instance to ensure unbiased assessment. See the Appendix for details.

Baselines. We compared against representative RUL prediction approaches from prior literature [11, 12] as well as recent state-of-the-art time-series forecasting methods such as TimeMixer++ [19], TimeMixer [20], and TimesNet [22] and PatchTST [15].

4.2 Results and Analysis

Impact of using different Teacher–Student backbone combinations. We conducted the experiments with different teacher-student combinations across CNN and Transformer architectures. A pool of teacher backbones (MAE variants, CLIP, and EfficientNet-B3) and a pool of lightweight student architectures (EfficientNet-B0, MobileNet-V3, and Tiny-ViT) were employed. We followed the same training protocol across datasets and report results under identical evaluation settings (see Appendix for details). Table 1 presents the performance of the proposed framework across various teacher–student backbone combinations. Homogeneous CNN pairings (e.g., EfficientNet-B3 $\rightarrow$ B0) perform well in Case 1, achieving an RMSE of 0.094. In contrast, the CLIP–Tiny-ViT pair attains state-of-the-art results in Case 4 (0.067 RMSE) and Case 5 (0.060 RMSE). Although MAE-Base–EfficientNet-B0 performs exceptionally in Cases 2 and 3, it underperforms in Cases 4 and 5, in contrast to MAE-Large–Tiny-ViT combination. This empirical analysis yields three important observations: (1) *Robust representation outweighs pure model scale*: CLIP's multimodal pre-training provides more stable distillation signals than the higher-capacity MAE-Large, preventing domain-specific overfitting. (2) *Homogeneous distillation*: Knowledge transfer is most effective when the teacher and student share similar inductive biases, as evidenced by the strong performance of ViT-to-ViT and CNN-to-CNN pairings. (3) *Knowledge overload*: Scaling up the teacher does not guarantee better student performance. For instance, distilling from MAE-Large to EfficientNet-B0 results in performance degradation compared to using a moderately sized teacher such as MAE-Base. Consequently, overwhelming a lightweight CNN student with such unnecessarily complex knowledge impairs its predictions. We selected CLIP-Tiny-ViT as the teacher-student backbone for the comparisons in the next sections.

Comparing to SOTAs. We evaluated the RUL prediction capabilities of our model across case studies and compared it against a wide range of state-of-the-art baselines. As shown in Table 2, the proposed framework achieves the best or highly competitive results across most settings. Among the external baselines, it attains the lowest RMSE in Cases 1, 3, 4, and 5 and the lowest MAPE in Cases 1, 3, and 4, indicating strong predictive accuracy under both identical and shifted operating conditions. The main exceptions are Case 2, where TimeMixer++ achieves lower RMSE and MAPE, and the Case 5 MAPE, where TimeMixer++ remains slightly better. Compared with the Only-Teacher ablation, the full framework lowers prediction error in most cases (e.g., Cases 1, 2, 4, and 5), indicating that the compact student together with sliding-window temporal aggregation provides benefits beyond directly applying the high-capacity

Table 1: Performance comparison of different Teacher-Student model combinations on RUL prediction tasks. Red: best, blue: second best.

Dataset		Case 1		Case 2		Case 3		Case 4		Case 5	
Teacher Model	Student Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
EfficientNet-B3	Tiny-ViT	0.124	0.178	0.119	0.080	0.067	0.037	0.069	0.048	0.064	**0.027**
	EfficientNet-B0	**0.094**	**0.150**	0.037	**0.034**	0.056	0.040	0.094	0.086	0.175	0.117
	MobileNet	0.154	0.225	0.061	0.043	0.098	0.079	0.076	0.064	0.128	0.083
CLIP	Tiny-ViT	0.099	0.156	0.107	0.070	0.071	0.039	**0.067**	0.055	0.060	0.040
	EfficientNet-B0	0.098	0.156	0.040	0.038	**0.052**	**0.035**	0.086	0.070	0.158	0.105
	MobileNet	0.171	0.218	0.056	0.041	0.092	0.069	0.075	0.055	0.149	0.103
MAE-Base	Tiny-ViT	0.112	0.176	0.126	0.088	0.064	0.041	**0.067**	0.058	0.061	0.030
	EfficientNet-B0	0.108	0.182	**0.036**	0.035	**0.052**	**0.035**	0.079	0.066	0.131	0.081
	MobileNet	0.182	0.251	0.067	0.039	0.087	0.063	0.077	0.061	0.151	0.087
MAE-Large	Tiny-ViT	0.125	0.168	0.114	0.082	0.066	0.038	**0.067**	**0.047**	**0.059**	**0.027**
	EfficientNet-B0	0.141	0.219	0.039	0.038	0.059	0.046	0.081	0.073	0.116	0.071
	MobileNet	0.175	0.223	0.053	0.041	0.095	0.070	0.073	0.061	0.117	0.082

Table 2: Results of evaluation metrics for different approaches in RUL prediction. Lower values indicate better performance.

Dataset	Case 1		Case 2		Case 3		Case 4		Case 5	
Methods	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
AE-TCA-LSSVM	0.128	0.243	0.232	0.317	0.096	0.358	0.159	0.281	0.214	1.243
2D-CLSTM	0.324	1.320	0.215	0.302	0.075	0.251	0.081	0.177	0.191	0.266
CNN-ORC	0.164	0.306	0.426	0.391	0.192	1.789	0.225	0.316	0.219	0.290
Bi-LSTMF	0.218	0.291	0.328	0.370	0.129	0.215	0.169	0.241	0.216	0.342
CLSTMF	0.146	0.409	0.190	0.285	0.234	0.540	0.144	0.271	0.091	0.184
PatchTST	0.151	0.234	0.095	0.040	0.120	0.053	0.099	0.070	0.087	0.040
TimesNet	0.195	0.283	0.183	0.195	0.119	0.048	0.117	0.101	0.104	0.053
TimeMixer	0.164	0.257	0.113	0.086	0.120	0.045	0.122	0.087	0.103	0.046
TimeMixer++	0.166	0.270	**0.079**	**0.026**	0.118	0.046	0.094	0.092	0.086	**0.035**
Only Teacher	0.170	0.267	0.125	0.117	**0.063**	0.051	0.096	0.107	0.090	0.078
Ours	**0.099**	**0.156**	0.107	0.070	0.071	**0.039**	**0.067**	**0.055**	**0.060**	0.040

teacher, while the teacher alone remains competitive on Case 3. These results position PSDI as a competitive and robust representation rather than a uniformly dominant one.

Performance across Representations and Architectures. Across Tables 3, 4, 5, we first compared PSDI with different image-based representations, including Recurrence Plots (RP) [5], Gramian Angular Fields (GAF) [21], STFT [6], Wavelet [4] under two standard vision backbones (VGG, ResNet18) and the proposed knowledge distillation (KD) framework. In both VGG and ResNet18 architectures, PSDI achieves the best results in Case 1 (e.g., 0.154 RMSE with VGG and 0.153 RMSE with ResNet18) and Case 4, and remains competitive in the remaining cases. While other representations, such as Wavelet and STFT, occasionally achieve the best results in specific scenarios (e.g., Wavelet in Cases

Table 3: Performance comparison of VGG model combinations on different image-based representations.

Model	Transformation	Case 1		Case 2		Case 3		Case 4		Case 5	
	Metric	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
VGG	RP	0.310	0.434	**0.052**	0.032	0.124	0.054	0.143	0.122	0.063	0.038
	GAF	0.298	0.317	0.079	0.055	0.116	0.062	0.551	0.284	0.158	0.102
	STFT	0.354	0.489	0.084	0.033	**0.078**	0.045	0.097	0.085	0.082	0.067
	Wavelet	0.159	0.230	0.092	**0.029**	0.095	**0.036**	0.065	0.054	**0.062**	**0.031**
	PSDI	**0.154**	**0.225**	0.087	0.049	0.081	0.044	**0.064**	**0.052**	0.134	0.118

Table 4: Performance comparison of ResNet-18 model combinations on different image-based representations.

Model	Transformation	Case 1		Case 2		Case 3		Case 4		Case 5	
	Metric	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ResNet18	RP	0.313	0.453	0.085	0.069	0.117	0.080	0.194	0.190	0.125	0.086
	GAF	0.541	0.470	0.075	0.049	0.114	**0.070**	0.220	0.218	0.172	0.117
	STFT	0.177	0.276	**0.074**	0.037	0.097	0.075	0.070	0.066	0.112	**0.065**
	Wavelet	0.154	0.239	0.088	**0.033**	0.105	0.076	0.083	0.090	**0.106**	0.078
	PSDI	**0.153**	**0.213**	0.079	**0.033**	**0.091**	0.072	**0.067**	**0.065**	0.137	0.121

2 and 5 for VGG, and STFT in Case 2 for ResNet18), PSDI maintains stable and competitive performance across all cases. When applied within the proposed KD framework, PSDI demonstrates its advantage by achieving the best performance in Cases 1, 3, 4, and 5, particularly under RMSE-oriented evaluation. PSDI improves prediction performance even with standard CNN backbones (VGG and ResNet18), indicating that the representation itself provides meaningful information independent of the learning architecture.

We also analysed the impact of using PSDI across different architectures. The results show that the proposed KD framework improves performance in most cases compared with the standalone VGG and ResNet18 models. For example, using PSDI, the KD framework achieves significantly lower RMSE in Case 1 (0.099) compared with VGG (0.154) and ResNet18 (0.153). Similar improvements are observed in Cases 3 and 5, where the KD framework obtains the best RMSE values of 0.071 and 0.060, respectively. These results indicate that combining PSDI with the proposed distillation strategy effectively enhances feature learning and leads to more accurate RUL estimation compared to conventional CNN architectures in most settings. Detailed runtime analysis and model parameters are provided in the Appendix.

Ablation Study. We validate the necessity of the proposed anchoring mechanism by comparing our globally anchored PSDI against a baseline using standard local PSR. As depicted in Figure 4, global anchoring improves lower RMSE and MAPE across diverse test cases, effectively mitigating coordinate drift—a phenomenon where attractor geometry shifts independently of the actual degradation state (see Appendix for extended visualizations). This quantitative gain is further ev-

Table 5: Performance comparison of our proposed knowledge distillation framework combinations on different image-based representations.

Model	Transformation	Case 1		Case 2		Case 3		Case 4		Case 5	
	Metric	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
KD (Ours)	RP	0.125	0.206	**0.084**	0.041	0.097	0.046	0.079	0.060	0.066	0.046
	GAF	0.123	0.203	0.085	0.044	0.097	0.045	0.069	0.056	0.070	0.037
	STFT	0.128	0.212	0.127	0.089	0.094	0.047	0.072	0.059	0.064	**0.027**
	Wavelet	0.119	0.192	0.093	**0.028**	0.096	0.045	0.071	0.062	0.062	0.029
	PSDI	**0.099**	**0.156**	0.107	0.070	**0.071**	**0.039**	**0.067**	**0.055**	**0.060**	0.040

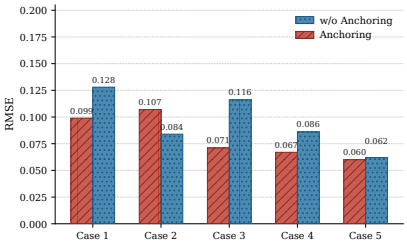

Fig. 4: Impact of global anchoring on prediction error.

Table 6: Analysis of the feature space clustering quality.

Method	Silhouette ↑	Inter ↑	Intra ↓
w/o Anchoring	0.101	0.175	0.308
Anchoring	**0.516**	**0.741**	**0.233**

idenced in Table 6, where anchoring dramatically improves the Silhouette score from 0.101 to 0.516. The significantly enhanced inter-cluster separation confirms that anchoring provides a stable geometric reference, suggesting that the learned feature space better separates degradation stages under the anchored representation.

5 Conclusion

This paper presented PSDI, a phase-space-inspired visual representation for vibration-based RUL prediction. By combining time-delay embedding with a globally anchored healthy reference frame, PSDI produces coordinate-consistent density maps that capture degradation-related changes in reconstructed trajectory distributions. The proposed GAR strategy reduces coordinate drift, while the compact knowledge distillation framework transfers useful visual priors and improves prediction stability through temporal aggregation. Experiments on benchmark bearing datasets show that PSDI is a competitive and robust representation compared with time-frequency, structural image-based, and recent time-series baselines. A limitation is that PSDI is constructed from local signal windows and does not explicitly model full run-to-failure history or cumulative damage accumulation. Future work will explore history-aware PSDI sequences, cumulative degradation indicators, adaptive reference frames, and broader validation beyond bearing vibration data.

Acknowledgements

The authors thank Innovate UK and ADC Energy Ltd. for financial support through the Knowledge Transfer Partnership (KTP) project No. 13518.

References

1. Barra, S., Carta, S.M., Corriga, A., Podda, A.S., Recupero, D.R.: Deep learning and time series-to-image encoding for financial forecasting. IEEE/CAA Journal of Automatica Sinica **7**, 683–692 (2020)
2. Cao, L.: Practical method for determining the minimum embedding dimension of a scalar time series. Physica D: Nonlinear Phenomena **110**, 43–50 (1997)
3. Chen, M., Shen, L., Li, Z., Wang, X.J., Sun, J., Liu, C.: Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In: International Conference on Machine Learning (ICML) (2025)
4. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory **36**(5), 961–1005 (1990)
5. Eckmann, J.P., Kamphorst, S.O., Ruelle, D.: Recurrence plots of dynamical systems. In: Turbulence, Strange Attractors and Chaos, pp. 441–445. World Scientific (1995)
6. Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing **32**, 236–243 (1984)
7. Jardine, A.K., Lin, D., Banjevic, D.: A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing **20**, 1483–1510 (2006)
8. Jiang, A.H., Huang, X.C., Zhang, Z.H., Li, J., Zhang, Z.Y., Hua, H.X.: Mutual information algorithms. Mechanical Systems and Signal Processing **24**, 2947–2960 (2010)
9. Khan, S., Yairi, T.: A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing **107**, 241–265 (2018)
10. Li, Y., Wang, H., Li, J., Tan, J.: A 2-d long short-term memory fusion networks for bearing remaining useful life prediction. IEEE Sensors Journal **22**, 21806–21815 (2022)
11. Ma, M., Mao, Z.: Deep-convolution-based lstm network for remaining useful life prediction. IEEE Transactions on Industrial Informatics **17**, 1658–1667 (2020)
12. Mao, W., He, J., Zuo, M.J.: Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. IEEE Transactions on Instrumentation and Measurement **69**, 1594–1608 (2019)
13. Namura, N., Ichikawa, Y.: Training-free time-series anomaly detection: Leveraging image foundation models. arXiv preprint arXiv:2408.14756 (2024)
14. Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., Varnier, C.: Pronostia: An experimental platform for bearings accelerated degradation tests. In: IEEE International Conference on Prognostics and Health Management. pp. 1–8 (2012)
15. Nie, Y., H. Nguyen, N., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR) (2023)
16. Randall, R.B., Antoni, J.: Rolling element bearing diagnostics—a tutorial. Mechanical Systems and Signal Processing **25**, 485–520 (2011)

17. Takens, F.: Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence, Warwick 1980. pp. 366–381. Springer Berlin Heidelberg (1981)
18. Wang, B., Lei, Y., Li, N., Li, N.: A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. IEEE Transactions on Reliability **69**, 401–412 (2018)
19. Wang, S., Li, J., Shi, X., Ye, Z., Mo, B., Lin, W., Ju, S., Chu, Z., Jin, M.: Timemixer++: A general time series pattern machine for universal predictive analysis. In: International Conference on Learning Representations (ICLR) (2025)
20. Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J.Y., Zhou, J.: Timemixer: Decomposable multiscale mixing for time series forecasting. In: International Conference on Learning Representations (ICLR) (2024)
21. Wang, Z., Oates, T.: Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence (2015)
22. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2d-variation modeling for general time series analysis. In: International Conference on Learning Representations (ICLR) (2023)
23. Zeng, Z., Kaur, R., Siddagangappa, S., Balch, T., Veloso, M.: From pixels to predictions: Spectrogram and vision transformer for better time series forecasting. In: Proceedings of the Fourth ACM International Conference on AI in Finance. pp. 82–90 (2023)
24. Zhong, S., Ruan, W., Jin, M., Li, H., Wen, Q., Liang, Y.: Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In: International Conference on Machine Learning (ICML) (2025)