Proceedings of CONF-MSS 2026 Syposium: Mechanical Control and Automation
DOI: 10.54254/2755-2721/2026.BJ32973

Physics-Regularized DINO Representation Learning for Cross-Condition Few-Shot Fault Diagnosis

Runyu Liu

School of Electronics and Control Engineering, Chang'an University, Xi'an, China
liuraiinn@gmail.com

Abstract. Cross-condition fault diagnosis with limited labeled data remains a significant challenge due to distribution shifts across operating conditions. While self-supervised learning methods improve label efficiency by capturing transferable representations, semantic objectives alone may suppress subtle, fault-relevant physical characteristics. Conversely, utilizing physics-informed methods as the primary learning framework can limit representational flexibility. To address this gap, a physics-regularized DINO framework is proposed. The model employs a DINO-style student-teacher vision transformer as the primary representation learner, utilizing wavelet-based time-frequency maps (TFMs) for self-distillation. In parallel, an auxiliary physics head acts as a regularization pathway by predicting the short-time Fourier transform (STFT)-based average energy density from data-driven, fault-sensitive frequency bands. Evaluated under a strict, leakage-aware protocol that enforces source-group isolation, the proposed framework demonstrates robust few-shot performance, achieving over 92% accuracy in 20-shot scenarios for both kNN and linear probing. Ablation studies verify that the interactive design between physics regularization and DINO centering yields consistent gains. This approach successfully balances semantic transferability and physical consistency, providing an effective solution for industrial condition monitoring.

Keywords: Cross-condition fault diagnosis, Few-shot learning, Self-supervised learning, DINO, Physics-regularized learning

1. Introduction

Rotating machinery fault diagnosis is essential for industrial condition monitoring and predictive maintenance, yet its practical application remains challenging in cross-condition settings, where changes in operating conditions often cause distribution shifts [1] between training and testing data. This difficulty becomes more pronounced in few-shot settings because the limited availability of labeled samples restricts the effectiveness of conventional supervised methods [2-4]. Recent self-supervised learning approaches have improved label efficiency by learning transferable representations [5-8], but semantic objectives alone may not fully preserve fault-relevant physical characteristics [9]. Meanwhile, physics-informed methods introduce useful prior knowledge [10], although they may limit representational flexibility when used as the primary learning framework [10]. To address this gap, this study investigates whether self-distillation with no labels (DINO) [11]

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BJ32973

can be enhanced through physics regularization for cross-condition few-shot fault diagnosis. Specifically, a physics-regularized DINO framework is developed, in which a DINO-style student–teacher vision transformer serves as the main representation learner, while an auxiliary physics head predicts STFT-based average energy density as a regularization target [12] during pretraining. To evaluate this design, the learned representations are examined under both the full protocol and the strict protocol. In the strict protocol, train/test construction is made leakage-aware by enforcing source-group isolation, so that samples derived from the same source file or group cannot simultaneously appear in both training and testing splits.

2. Method

2.1. Task setting and overall framework

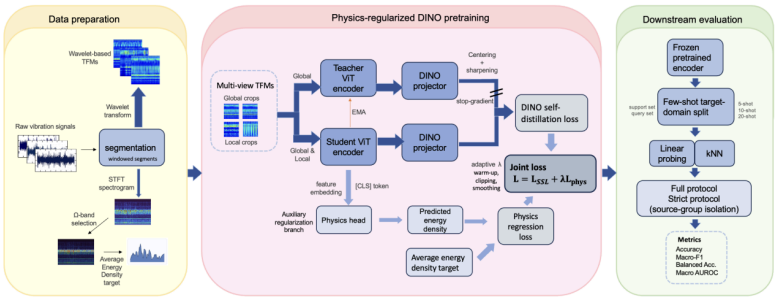


Figure 1. Overall framework of the proposed physics-regularized DINO for cross-condition few-shot fault diagnosis

As illustrated in Figure 1, the proposed framework follows a unified pipeline for cross-condition few-shot fault diagnosis under limited labeled data. In this setting, variations in operating conditions introduce distribution shifts between training and testing data, which significantly degrade the generalization performance of conventional supervised methods. The objective is therefore to learn representations that are both condition-invariant and fault-relevant.

Raw vibration signals are first converted into wavelet-based time-frequency maps (TFMs) [13], which serve as the visual inputs to the DINO backbone. Multiple augmented views, including global and local crops, are then generated from each TFM for self-distillation with no labels [11]. In parallel, STFT-based spectrograms are used to compute the auxiliary average energy density target for physics regularization.

The overall framework consists of a DINO-style student–teacher vision transformer as the main representation backbone [13] and an auxiliary physics head attached during pretraining. The DINO objective learns transferable representations from multi-view TFMs, whereas the auxiliary physics head regularizes these representations by predicting average energy density. Accordingly, the physics component is used as an auxiliary regularization pathway rather than as an independent learning framework.

After pretraining, the encoder is frozen and evaluated for cross-condition few-shot diagnosis under both the full protocol and the strict protocol by kNN and linear probing.

2.2. Physics-regularized DINO pretraining

A DINO-style student–teacher architecture is employed to learn representations from multi-view TFMs. The student network is trained to match the momentum-updated teacher outputs through self-

DOI: 10.54254/2755-2721/2026.BJ32973

distillation with global and local crops [11]. In this process, the teacher parameters are updated by exponential moving average (EMA), and the teacher predictions are stabilized through the centering and temperature-based sharpening mechanism [11] within the DINO framework. This design encourages consistency across views and yields transferable embeddings for few-shot diagnosis. However, a purely semantic objective may suppress subtle but fault-relevant physical characteristics in the vibration signal [9].

To address this limitation, an auxiliary physics head is introduced, as illustrated in Figure 1. Specifically, based on the STFT time–frequency representation, the power spectrum is defined as

$$S(\tau,\omega) = |\ STFT(x)(\tau,\omega)\ |^2 \tag{1}$$

where τ and ω denote the time-frame index and frequency-bin index, respectively. Based on this spectrum, the average energy density over the selected frequency region Ω is computed as [12]

$$E = \frac{1}{T|\Omega|} \sum_{\tau} \sum_{\omega \in \Omega} S(\tau,\omega) \tag{2}$$

where T is the number of time frames and $|\Omega|$ is the number of selected frequency bins. This quantity is used as the auxiliary physics target during pretraining. To enhance fault sensitivity, an automatic Ω -band selection strategy is adopted by ranking frequency components [14] with statistical indicators such as standard deviation and kurtosis. In practice, these statistics are combined into a weighted frequency score, after which the top-ranked frequency bins are selected, merged into contiguous bands, and filtered to remove excessively short isolated regions, thereby forming Ω . The resulting target emphasizes frequency regions that are more likely to contain fault-related information while suppressing less informative background components. In this way, the auxiliary target is not computed from the full spectrum, but from a compact fault-sensitive region selected in a data-driven yet physically interpretable manner. The auxiliary physics head then predicts this average energy density from the learned representations, thereby introducing a physically meaningful regularization signal during pretraining.

Within this design, the DINO backbone provides the main representation learning capability, whereas the auxiliary physics head encourages the embeddings to retain fault-relevant physical information. As a result, the pretrained representations are expected to remain both transferable across operating conditions and discriminative for fault diagnosis.

2.3. Joint objective and training strategy

The pretraining objective combines the DINO self-distillation loss with an auxiliary physics regularization loss in a unified framework. The overall loss is defined as

$$L = L_{ssl} + \lambda L_{phys} \tag{3}$$

where L_{ssl} denotes the DINO loss, L_{phys} denotes the physics regularization loss instantiated as average-energy-density regression, and λ controls the balance between the two objectives [12,15].

DOI: 10.54254/2755-2721/2026.BJ32973

To stabilize joint optimization, an adaptive weighting strategy for λ is adopted based on the relative magnitudes of the DINO and physics losses. Specifically, λ is adjusted according to the loss ratio between the two objectives, so that the auxiliary branch remains comparable in scale without overwhelming the semantic objective. Initially, a warm-up stage prioritizes the DINO objective, preventing the auxiliary physics loss from dominating early representation learning, while clipping and smoothing are introduced to mitigate abrupt fluctuations during training. Consequently, the auxiliary branch acts as a controlled regularizer rather than an equally weighted objective, maintaining stable semantic learning while enforcing fault-relevant physical consistency.

During joint optimization, the semantic DINO loss and the auxiliary energy-prediction loss are minimized simultaneously, with the teacher network updated via an exponential moving average (EMA) of the student parameters.

After pretraining, the encoder is frozen and its representations are evaluated by k-nearest neighbors and linear probing for cross-condition few-shot diagnosis. This setup tests whether the pretrained embeddings remain effective when only limited labeled target samples are available.

3. Experiments

3.1. Experimental setup

The experiments were conducted on the Case Western Reserve University (CWRU) [16] bearing dataset under a cross-condition few-shot fault diagnosis setting [17], where training and testing data came from different operating conditions. In this study, 12 kHz drive-end vibration signals together with normal baseline data were used to construct the dataset. The considered categories covered normal, inner-race fault, outer-race fault, and rolling-element fault conditions, and the cross-condition setting was established by separating samples from different operating conditions between pretraining/evaluation stages. Raw vibration signals were segmented by a sliding window of length 20480 with a hop size of 10240 and were then converted into time-frequency representations. In the final pipeline, wavelet-based time-frequency maps (TFMs) served as the visual inputs to the DINO backbone, whereas STFT-based average energy density was used to construct the auxiliary physics target. For STFT, the FFT length, hop length, and window length were set to 256, 128, and 256, respectively, with the power spectrum computed from the squared magnitude of the STFT coefficients. The resulting maps were normalized and resized to 224×224 before being fed into the encoder. In the auxiliary branch, the energy target was further normalized before regression to improve the stability of joint optimization.

For implementation, a ViT-tiny (patch16-224) backbone [18] initialized from ImageNet was used, and the DINO projector output dimension was set to 1024. During pretraining, multi-view augmentation with global and local crops was adopted following DINO, where two global views and multiple local views were generated for self-distillation. The model was trained with a batch size of 48 using the AdamW optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-4}) and gradient clipping. The objective jointly optimized the DINO and physics regularization losses. The auxiliary physics head—regulated by early warm-up and dynamic adjustment—was restricted to pretraining. Meanwhile, the teacher network was updated via EMA, and the default DINO centering mechanism was retained.

For evaluation, frozen representations were assessed via k-nearest neighbors and linear probing (trained for 40 epochs, learning rate 10^{-2}, weight decay 10^{-4}). Performance across 5-, 10-, and 20-shot settings was measured using Accuracy, Macro-F1, Balanced Accuracy, and Macro AUROC. In these settings, per-class labeled support sets were constructed, with the remaining target samples

DOI: 10.54254/2755-2721/2026.BJ32973

used as queries, and each shot setting was repeated multiple times to reduce sampling variance. To evaluate robustness, performance under the full and strict protocols was compared. The strict protocol enforced source-group isolation, ensuring samples from the same source file were not shared across support and query sets, and support/query construction was performed preferentially in a group-safe manner. Following a class-level feasibility check, strict results were evaluated only on a leakage-aware, group-feasible subset, and thus are not directly comparable to full-class headline results.

3.2. Main results under full and strict protocols

Table 1 summarizes the main few-shot results under the full protocol and the strict protocol. Overall, the proposed framework achieves strong performance across all reported shot settings, indicating that the pretrained representations remain effective for few-shot diagnosis even when only a limited number of labeled target samples are available. Under the full protocol, kNN reaches 87.43%, 87.49%, and 93.24% accuracy in the 5-shot, 10-shot, and 20-shot settings, respectively, while linear probing achieves 86.28%, 85.69%, and 92.74%. These results show that the learned embeddings support both kNN evaluation and linear probing, and that performance improves further as more labeled target samples become available.

Table 1. Main few-shot results under the full and strict protocols

Shot	Full-kNN	Full-Linear	Strict-kNN	Strict-Linear
5	87.43	86.28	86.35	82.49
10	87.49	85.69	86.78	81.80
20	93.24	92.74	92.39	92.27

More importantly, the performance remains robust under the strict protocol. In the lower-shot settings, only moderate degradation is observed relative to the full protocol, while the overall performance remains competitive. In particular, the strict protocol still yields 86.35% and 86.78% accuracy for kNN in the 5-shot and 10-shot settings, respectively. At 20-shot, both classifiers remain above 92% accuracy, with 92.39% for kNN and 92.27% for linear probing. This result suggests that the learned representations preserve strong transferability under the strict protocol, rather than relying primarily on leakage-related bias.

Another observation is that the gap between the full protocol and the strict protocol becomes small at 20-shot. This trend indicates that once a modest number of labeled target samples is available, the pretrained representations can support stable few-shot diagnosis under cross-condition shift. In addition, kNN appears slightly more stable than linear probing in the lower-shot strict setting, suggesting that the pretrained embedding space already exhibits a coherent neighborhood structure before supervised fitting [13]. Taken together, these results support the main claim of the paper: the proposed physics-regularized DINO framework is effective for low-data cross-condition fault diagnosis and remains robust under the strict protocol.

DOI: 10.54254/2755-2721/2026.BJ32973

3.3. Core ablation on collaborative design

Table 2. Ablation results of the collaborative design under the strict protocol

Variant	Probe	Acc.	Macro-F1	Bal-Acc.	AUROC
No-Physics + Center	kNN	77.0	71.7	72.9	88.6
	Linear	80.5	76.6	77.2	96.4
No-Physics + No-center	kNN	83.5	78.0	80.6	92.4
	Linear	83.9	78.8	81.0	98.5
Physics + Center	kNN	89.7	88.1	87.9	95.9
	Linear	85.5	83.8	85.4	96.9
Physics + No-center	kNN	88.8	86.5	86.9	96.7
	Linear	89.7	88.3	87.9	99.2

To examine whether the observed gain truly arose from physics regularization rather than from accidental split advantages, the core ablation was conducted under the strict protocol. Four representative variants were compared: No-Physics + Center, No-Physics + No-center, Physics + Center, and Physics + No-center. Overall, introducing physics regularization led to consistently stronger results than the corresponding no-physics counterparts, indicating that the improvement was stable rather than a fluctuation in a single metric.

The quantitative results further support this trend (Table 2). Under No-Physics + Center, the linear probe achieved 80.5% accuracy, 76.6% macro-F1, 77.2% balanced accuracy, and 96.4% AUROC. Replacing it with No-Physics + No-center improved the linear result to 83.9% / 78.8% / 81.0% / 98.5%. After introducing physics regularization, Physics + Center already raised the kNN performance to 89.7% / 88.1% / 87.9% / 95.9%, while the best overall configuration was obtained by Physics + No-center + Linear, which reached 89.7% accuracy, 88.3% macro-F1, 87.9% balanced accuracy, and 99.2% AUROC. Relative to the representative no-physics settings, physics regularization brought positive gains in both accuracy and macro-F1, including +12.8 pp / +16.4 pp for Center + KNN and +5.9 pp / +9.5 pp for No-center + Linear.

A further observation is that the effect of centering depends on whether physics regularization is present and on which probe is used. Without physics regularization, No-center performed better for both kNN and linear probing. After physics regularization was introduced, however, the preferred configuration became probe-dependent: kNN favored Center, whereas Linear favored No-center. Therefore, the best result did not come from merely adding an auxiliary physics head, but from the interaction between physics regularization and centering within DINO pretraining. Taken together, the ablation results suggest that the observed improvement is associated with this interaction rather than with architectural expansion alone.

3.4. Class-level evidence and error discussion

To complement the averaged metrics, class-level evidence was further examined under the merged 10-class setting. The overall evaluation reached 87.19% accuracy, 82.39% macro-F1, 83.96% balanced accuracy, and 95.41% macro AUROC. Several categories achieved near-perfect recognition, with IR007, IR014, OR021, and NC all reaching 100% recall, while OR007 and RB007 still achieved 95.6% and 93.3%, respectively. At the same time, a few difficult classes remained challenging, especially RB014, IR021+, RB021+, and OR014, whose recalls were 53.3%, 55.0%,

DOI: 10.54254/2755-2721/2026.BJ32973

66.7%, and 76.7%, respectively. These results indicate that the learned representations are broadly discriminative across most fault modes, although several hard classes still limit the overall balance of performance.

The remaining errors also exhibit a clear structure rather than random failure. The dominant confusion routes were concentrated on a few physically similar fault patterns, including IR021+ → OR007, RB014 → IR014, and RB021+ → IR014. This suggests that the current bottleneck lies mainly in local boundary overlap between similar fault modes, rather than in a collapse of the whole representation space. Overall, the class-level results further support the main conclusion of the paper: the pretrained representations are effective for most categories, while the remaining errors are concentrated on a few physically similar hard classes [13].

4. Conclusion

This study aimed to improve cross-condition few-shot fault diagnosis under limited labeled data by enhancing DINO representation learning with auxiliary physics regularization. To this end, a physics-regularized DINO framework was developed, in which a DINO-style student-teacher vision transformer served as the main representation backbone, while an auxiliary physics head predicted STFT-based average energy density during pretraining. Within this design, the physics branch functioned as a controlled regularization pathway rather than as a separate learning objective, so that transferable semantic representation learning remained the dominant component.

The experimental results support the effectiveness of this collaborative design. Under both the full protocol and the strict leakage-aware protocol, the learned representations maintained strong few-shot diagnostic performance with both kNN and linear probing. The core ablation further showed that introducing physics regularization consistently improved performance relative to no-physics counterparts under the strict protocol, indicating that the gain was associated with the regularization effect rather than with split-related bias alone. In addition, the class-level analysis showed that the learned representations were broadly discriminative across most fault categories, while the remaining errors were concentrated on several physically similar hard classes.

Overall, these results indicate that auxiliary physics regularization can strengthen the robustness and transferability of DINO representations for low-data cross-condition fault diagnosis. Nevertheless, several difficult classes still exhibited boundary overlap, and the current study was conducted on the CWRU dataset only. Future work may extend the framework to more diverse datasets and explore stronger strategies for handling physically similar fault categories.

References

[1] M. Russell, P. Wang. (2023) Maximizing model generalization for machine condition monitoring with self-supervised learning and federated learning, Journal of Manufacturing Systems, 71, 274–285.
[2] F. Kibrete, D.E. Woldemichael, H.S. Gebremedhen. (2024) Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review, Measurement, 114658.
[3] S. Tang, J. Ma, Z. Yan, Y. Zhu, B.C. Khoo. (2024) Deep transfer learning strategy in intelligent fault diagnosis of rotating machinery, Engineering Applications of Artificial Intelligence, 134, 108678.
[4] I. Misbah, C.K. Lee, K.L. Keung. (2024) Fault diagnosis in rotating machines based on transfer learning: Literature review, Knowledge-Based Systems, 283, 111158.
[5] X. Zhao, M. Jia, J. Bin, T. Wang, Z. Liu. (2021) Multiple-order graphical deep extreme learning machine for unsupervised fault diagnosis of rolling bearing, IEEE Transactions on Instrumentation and Measurement, 70, 1–12.
[6] S. Liu, J. Chen, S. He, E. Xu, H. Lv, Z. Zhou. (2021) Intelligent fault diagnosis under small sample size conditions via bidirectional infomax GAN with unsupervised representation learning, Knowledge-Based Systems, 232, 107488.

DOI: 10.54254/2755-2721/2026.BJ32973

[7] Y. Ding, J. Zhuang, P. Ding, M. Jia. (2022) Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings, Reliability Engineering & System Safety, 218, 108126.
[8] T. Wang, M. Qiao, M. Zhang, Y. Yang, H. Snoussi. (2020) Data-driven prognostic method based on self-supervised learning approaches for fault detection, Journal of Intelligent Manufacturing, 31, 1611–1619.
[9] Z. Qin, Y. Chen, G. Zhu, E. Zhou, Y. Zhou, Y. Zhou, C. Zhu. (2024) Enhanced pseudo-label generation with self-supervised training for weakly-supervised semantic segmentation, IEEE Transactions on Circuits and Systems for Video Technology, 34(8): 7017–7028.
[10] R. Hassannejad, M.M. Ettefagh, Y.B. Mossayebi. (2025) Adaptive wavelet-based physics-informed CNN for bearing fault diagnosis, International Journal of Prognostics and Health Management, 16(1).
[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin. (2021) Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640.
[12] J. Lee, S. Hong, S. Lee, J. Lee. (2026) Physics-guided self-supervised learning for rotating-machinery fault diagnosis: Role of average energy density equation, Advanced Engineering Informatics, 69, 103876.
[13] Y. Jin, L. Hou, M. Du, Y. Chen. (2022) A wavelet transform and self-supervised learning-based framework for bearing fault diagnosis with limited labeled data, arXiv preprint, arXiv: 2207.10432.
[14] S. Shen, H. Lu, M. Sadoughi, C. Hu, V. Nemani, A. Thelen, M.A. Zuo, A.K. Srivastava, S. Kenny. (2021) A physics-informed deep learning approach for bearing fault detection, Engineering Applications of Artificial Intelligence, 103, 104295.
[15] W. Deng, K.T. Nguyen, K. Medjaher. (2022) Physics informed self-supervised learning for fault diagnostics and prognostics in the context of sparse and noisy data, in: Proceedings of the PHM Society European Conference, 7(1): 574–576.
[16] Case Western Reserve University. (2021) Case Western Reserve University Bearing Data Center, [Online]. Available: http: //csegroups.case.edu/bearingdatacenter/home (accessed August).
[17] G. Wang, D. Liu, J. Xiang, L. Cui. (2024) Attention guided partial domain adaptation for interpretable transfer diagnosis of rotating machinery, Advanced Engineering Informatics, 62, 102708.
[18] A. Dosovitskiy, L. Beyer, et, al. (2021) An image is worth 16×16 words: Transformers for image recognition at scale, arXiv preprint, arXiv: 2010.11929.