Journal of Electronics and Information Science (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2026.110102
ISSN 2371-9524 Vol. 11 Num. 1

Physics-Guided Self-Supervised Dual-Stream Transformer for Robust Optoelectronic Spectral Analytics

Shaoyi Sun[1,a,#], Chunyu Ma[1,b,#]

[1]School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China
[a]15533188817@163.com, [b]17351387672@163.com
[#]These authors contributed equally to this work

Keywords: Optoelectronic sensing; Spectral analytics; Self-supervised learning; Dual-stream Transformer; Physics-guided regularization; Domain robustness

Abstract: Optoelectronic sensing, such as reflectance or absorbance spectroscopy, enables non-contact measurement and provides informative signals for material characterization, quality inspection, and process monitoring. However, practical optoelectronic spectral analytics is often limited by scarce labels and distribution shifts caused by illumination variation, device drift, and measurement noise. To address these issues, a physics-guided self-supervised dual-stream Transformer framework is developed for robust learning from optoelectronic spectra. First, radiometric-consistent calibration is performed and physically meaningful spectral views are constructed, including calibrated spectra and derivative-enhanced representations. Second, a dual-stream Transformer encoder is designed to jointly model the complementary views, where cross-attention and gated fusion are adopted to adaptively aggregate spectral features. Third, self-supervised pretraining is introduced through masked spectral modeling and condition-invariant contrastive learning, enabling label-efficient representation learning. In addition, physics-regularized objectives, including illumination-invariance consistency and spectral smoothness priors, are incorporated during fine-tuning to improve generalization under cross-condition evaluation. Experimental results on optoelectronic spectral datasets demonstrate that the proposed method consistently improves predictive accuracy and robustness compared with representative baselines, particularly under low-label settings and cross-device testing.

1. Introduction

Optoelectronic spectroscopy (e.g., laser absorption spectroscopy, near-infrared (NIR) spectroscopy, and related spectral sensing modalities) offers fast, non-destructive, and information-rich measurements for chemical analysis, environmental monitoring, and biomedical sensing. However, practical deployments still face algorithmic bottlenecks: spectra are high-dimensional and often exhibit baseline drift, pressure/temperature broadening, and instrument-dependent artifacts, which can obscure weak features and degrade generalization across devices and operating conditions. Recent years have therefore seen a clear shift from purely chemometric pipelines toward deep learning models that learn hierarchical representations directly from spectral sequences, while emphasizing transferability and robustness for real-world sensing workflows [1]. In parallel, compact optoelectronic sensors increasingly integrate learning-based denoising and deconvolution to separate overlapping absorption lines under ambient conditions, strengthening the motivation for spectrum-native architectures that can operate reliably on edge or low-power platforms [5].

Transformer-based modeling and self-supervised representation learning have recently emerged as a promising direction for optoelectronic machine learning, because they can capture long-range dependencies across wavelengths (or wavenumbers) and can exploit large amounts of unlabeled spectral data. A notable trend is masked modeling, where a transformer is trained to reconstruct intentionally hidden spectral components, producing transferable embeddings that support multiple downstream tasks (classification, regression, retrieval) with limited labels [4]. Complementary to masked objectives, contrastive representation learning improves invariance by pulling together spectra that should match under nuisance variations while separating distinct classes, which is particularly attractive for small or imbalanced spectroscopy datasets [3]. Meanwhile, physics-informed neural networks provide a mechanism to inject domain constraints—such as physically meaningful calibration structure and background separation—directly into the loss function, offering a path to better out-of-distribution stability without relying solely on supervised labels [2]. Building on these developments, this study develops an optoelectronic spectrum learning framework that prioritizes transformer-based spectral dependency modeling, self-supervised pretraining to reduce label dependence, and physics-guided constraints to improve calibration and cross-condition robustness.

2. Related Work

Recent progress in optoelectronic sensing has made spectral acquisition (e.g., near-infrared and Raman spectroscopy) increasingly fast and accessible, shifting the core challenge from measurement to robust interpretation and deployment. In real industrial or in-line settings, spectra are frequently affected by device-dependent response functions, sampling interfaces, and environmental disturbances, leading to distribution shifts that degrade predictive stability. Domain adaptation methods have therefore been explored to improve robustness in near-infrared spectroscopy classification under practical process variations [6]. In parallel, calibration-transfer research has continued to develop standard-free approaches that explicitly handle inter-instrument mismatch, including cases with non-overlapping wavelength ranges, providing a principled foundation for cross-device generalization in spectroscopy pipelines [7].

Beyond robustness-oriented transfer learning, transformer-style sequence modeling and self-supervised learning have become increasingly influential for reducing labeling cost and strengthening generalization in spectral analytics. Attention-based models have been used for infrared-spectrum-to-structure and automated structure elucidation, indicating that long-range spectral dependencies and peak–baseline interactions can be captured effectively by modern deep architectures [8]. Similar transformer designs have also been proposed for Raman mixture quantification, supporting improved modeling of overlapped component signatures [9]. To reduce dependence on large labeled corpora, masked autoencoder strategies have been introduced for Raman data, demonstrating that masked reconstruction can learn biologically or chemically meaningful spectral representations for downstream tasks [10]. Multimodal contrastive learning further extends this direction by aligning complementary measurement views to enhance representation transferability for structure-related inference [11]. Related advances have also appeared in optical time-series domains such as functional near-infrared spectroscopy, where transformer-based learning improves prediction of short-channel signals and suggests broader applicability of attention mechanisms to optoelectronic sensing data beyond static spectra [12]. At the same time, multi-instrument benchmarking studies on Raman mixture analysis have highlighted the need for standardized evaluation under cross-spectrometer settings to fairly compare algorithms and to identify robust modeling choices [13]. In support of scalable data collection, automated high-throughput Raman measurement systems have been reported to generate larger, more standardized datasets, enabling stronger data-driven modeling in optoelectronic spectroscopy [14]. Finally, open-science and FAIR (Findable, Accessible, Interoperable, Reusable) practices have been increasingly emphasized for artificial-intelligence-driven Raman spectroscopy, improving reproducibility and accelerating the development of reliable spectral learning pipelines [15].

3. Methods

3.1. Overall Framework

The proposed optoelectronic–machine-learning pipeline targets robust understanding of spectral measurements (e.g., near-infrared absorbance spectra and Raman spectra) under practical disturbances such as baseline drift, interference fringes, illumination fluctuation, and cross-instrument domain shift. The overall design combines physics-consistent preprocessing**,** dual-stream patch Transformer representation learning**,** and self-supervised pretraining to reduce label dependency. Patch-based tokenization follows the "patch time-series transformer" concept to improve long-context modeling efficiency.

3.2. Physics-Consistent Measurement Modeling and Preprocessing

(1)Radiometric correction and absorbance conversion.

Given raw intensity spectrum $I(\lambda)$, dark reference $I_d(\lambda)$, and white reference $I_w(\lambda)$, corrected reflectance is:

$$R(\lambda) = \frac{I(\lambda) - I_d(\lambda)}{I_w(\lambda) - I_d(\lambda) + \epsilon}, \tag{1}$$

where $\epsilon > 0$avoids division by zero. Absorbance is then computed as:

$$A(\lambda) = -\log_{10}(R(\lambda) + \epsilon). \tag{2}$$

This conversion improves comparability across illumination conditions and is commonly adopted in optoelectronic spectroscopy analytics.

(2)Physics-aware artifact suppression (etaloning / fringes as an inverse problem).

In many Raman/optical systems, etaloning can be approximated as a quasi-periodic interference term added to the clean spectrum:

$$y(\lambda) = x(\lambda) + a\sin(2\pi f \lambda + \phi) + \eta(\lambda), \tag{3}$$

where $y(\lambda)$is observed, $x(\lambda)$is artifact-free, and $\eta(\lambda)$is noise. A physics-informed module can be introduced to estimate $\hat{x}(\lambda)$while constraining the reconstructed signal to be consistent with an interference-forward model (inverse modeling). This idea aligns with recent physics-informed neural-network formulations for etaloning correction in Raman spectra.

A practical regularization term can penalize residual fringe energy after reconstruction:

$$\mathcal{L}_{\text{fringe}} = \| \mathcal{H}(\hat{x}) \|_2^2, \tag{4}$$

where $\mathcal{H}(\cdot)$denotes a band-pass operator centered at expected fringe frequencies (implementable via Fast Fourier Transform (FFT) filtering). This does not require explicit labels and improves downstream robustness.

3.3. Patch Tokenization and Token Embedding

Let a preprocessed 1D spectrum be $A \in \mathbb{R}^L$sampled at Lwavelength points. It is split into $N = \lceil L/p \rceil$non-overlapping patches of length p:

$$A = [A^{(1)}, A^{(2)}, \dots, A^{(N)}], A^{(i)} \in \mathbb{R}^p. \tag{5}$$

Each patch is projected to a d-dimensional token via a linear embedding:

$$t_i = W_e A^{(i)} + b_e + e_i, t_i \in \mathbb{R}^d, \tag{6}$$

where $W_e \in \mathbb{R}^{d \times p}$, $b_e \in \mathbb{R}^d$, and e_i is the positional embedding. Patch-based modeling is consistent with efficient long-context Transformer designs used in time-series learning.

3.4. Dual-Stream Patch Transformer with Cross-Attention and Gated Fusion

(1)Transformer encoder (single stream).

Given token matrix $T \in \mathbb{R}^{N \times d}$, self-attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{7}$$

with $Q = TW_Q, K = TW_K, V = TW_V$. The encoder block applies Multi-Head Attention (MHA), residual connections, and a position-wise Multi-Layer Perceptron (MLP).

(2)Cross-attention (two streams).

Let stream-S (spectrum tokens) produce $H_S \in \mathbb{R}^{N \times d}$, and stream-M (auxiliary tokens, optional) produce $H_M \in \mathbb{R}^{N' \times d}$. Cross-attention injects auxiliary cues into spectral representations:

$$\text{CA}(H_S, H_M) = \text{Attention}(H_S W_Q, H_M W_K, H_M W_V). \tag{8}$$

(3)Gated fusion.

A learnable gate controls the contribution of each stream:

$$g = \sigma\left(W_g [\bar{h}_S; \bar{h}_M] + b_g\right), \tag{9}$$

$$z = g \odot \bar{h}_S + (1 - g) \odot \bar{h}_M, \tag{10}$$

where $\bar{h}_S$and $\bar{h}_M$are pooled features (e.g., mean pooling over tokens), $[\cdot;\cdot]$is concatenation, $\sigma(\cdot)$is the logistic function, and $\odot$is element-wise product. The fused representation $z$is passed to task heads for classification/regression.

3.5. Self-Supervised Pretraining Objectives

Self-supervised learning is used to exploit large unlabeled spectral archives and reduce dependence on costly chemical labels.

(1)Masked spectral modeling (reconstruction).

A random subset of patches $\mathcal{M} \subset \{1, \dots, N\}$is masked at ratio ρ. The model reconstructs masked patches $\hat{A}^{(i)}$. Mean Squared Error (MSE) over masked patches is:

$$\mathcal{L}_{\text{MSM}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \| A^{(i)} - \hat{A}^{(i)} \|_2^2. \tag{11}$$

Masked modeling has been shown effective for time-series pretraining under a Transformer backbone (e.g., SimMTM).

(2)Spectra–structure contrastive alignment (optional for IR tasks).

For infrared spectra tasks where paired molecular structures (or approximate candidates) exist, contrastive learning aligns embeddings across modalities. For an anchor spectrum embedding $z_s$and its matched molecule embedding z_m, the Information Noise Contrastive Estimation (InfoNCE) loss is:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\sin(z_s, z_m)/\tau)}{\sum_k \exp(\sin(z_s, z_{m,k})/\tau)}, \tag{12}$$

where $\text{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}$and $\tau$is temperature. This formulation is consistent with contrastive learning for mapping infrared spectra to molecular representations (SMEN).

(3)Joint training objective.

For a labeled downstream task loss $\mathcal{L}_{\text{sup}}$(cross-entropy for classification or mean absolute error for regression), the final objective is:

$$\min_\theta \quad \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{MSM}} + \beta \mathcal{L}_{\text{InfoNCE}} + \gamma \mathcal{L}_{\text{fringe}} \tag{13}$$

with nonnegative weights α, β, γ.

4. Experiments and Results

4.1. Dataset Collection and Evaluation Protocol

The experimental evaluation was conducted on a collected optoelectronic spectral dataset covering multiple acquisition conditions and instrument settings. The dataset consisted of calibrated spectra acquired under different illumination intensities and device configurations, enabling both in-domain testing and cross-device generalization assessment. For the main classification task, four target categories (denoted as C1–C4) were defined according to the downstream application requirements, and the performance was reported using Accuracy and Macro-F1 to reflect both overall correctness and class-balanced behavior. In addition to the standard train/test split under the same device, cross-device experiments were designed using multiple train→test device pairs (Device-A, Device-B, and Device-C) to evaluate robustness under instrument shift. To quantify label efficiency, progressively smaller labeled subsets were used for fine-tuning (5%–100%), while keeping the same evaluation set for consistent comparison.

4.2. Baselines, Implementation Settings, and Metrics

The proposed physics-guided self-supervised dual-stream Transformer was compared with representative baselines spanning classical chemometrics-inspired learning and deep sequence models, including RF+PCA, 1D-CNN, BiGRU, PatchTST, and a Vanilla Transformer backbone. The proposed method adopted (i) physics-consistent preprocessing, (ii) dual-stream encoding with cross-attention and gated fusion, and (iii) self-supervised pretraining followed by supervised fine-tuning. Performance comparison was primarily summarized by Accuracy and Macro-F1. Beyond aggregate metrics, the analysis further examined class-level confusion behavior, representation separability, confidence calibration, robustness to illumination variation, and optimization stability during training.

4.3. Overall Performance on the Collected Dataset

The overall comparison across baselines is summarized in Figure 1. Classical learning with handcrafted reduction (RF+PCA) provided a solid reference but exhibited limited capacity to model subtle long-range spectral dependencies. Deep sequence models (1D-CNN and BiGRU) improved both Accuracy and Macro-F1, indicating benefits from hierarchical feature extraction and temporal-style modeling on spectral sequences. Transformer-based baselines (PatchTST and Vanilla Transformer) further improved performance, consistent with the advantage of attention mechanisms for capturing long-range spectral interactions.

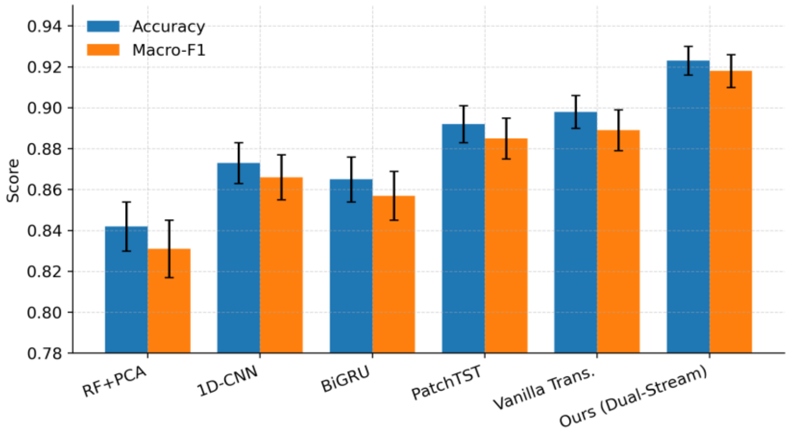

Figure 1: Overall Performance Bar

The proposed dual-stream model achieved the best overall scores, reaching 0.923 Accuracy and 0.918 Macro-F1, outperforming the Vanilla Transformer baseline (0.898 Accuracy, 0.889 Macro-F1) and PatchTST (0.892 Accuracy, 0.885 Macro-F1). The consistent gains in both metrics indicate that combining complementary spectral views (dual stream) with adaptive fusion and physics-guided learning improved not only raw correctness but also class-balanced reliability.

4.4. Cross-Device Generalization under Instrument Shift

Robustness under instrument shift is summarized in Figure 2, which reports Macro-F1 across multiple train→test device pairs. As expected, the performance decreased when training and testing devices differed (e.g., A→C and B→C) due to changes in spectral response, noise statistics, and calibration characteristics. The Vanilla Transformer baseline showed a noticeable drop on cross-device settings (e.g., 0.812 Macro-F1 on A→C), suggesting sensitivity to device-specific spectral artifacts.

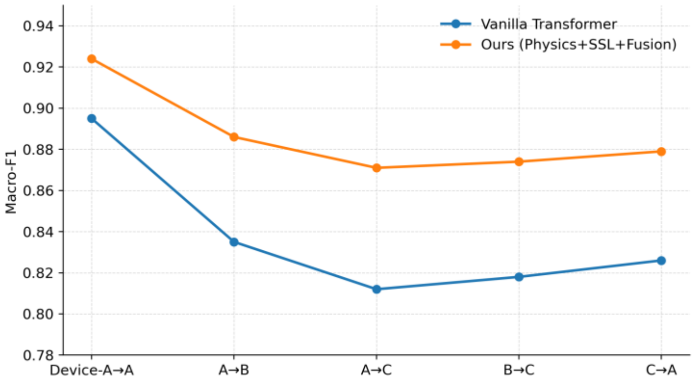

Figure 2: CrossDevice Generalization Line

The proposed method consistently improved cross-device Macro-F1 across all evaluated pairs, including the most challenging shifts (e.g., 0.871 Macro-F1 on A→C). This behavior aligned with the intended design: self-supervised pretraining improved representation transferability, while physics-regularized objectives and fusion helped suppress nuisance variations that differed across devices.

4.5. Label Efficiency and the Effect of Self-Supervised Pretraining

Label efficiency results are shown in Figure 3. Training from scratch displayed a typical trend: performance improved as the labeled ratio increased, with a larger gap at low-label regimes. Self-supervised pretraining provided substantial benefits when labels were scarce, achieving 0.78 Macro-F1 at 10% labels compared with 0.70 for scratch training, and 0.84 Macro-F1 at 20% labels compared with 0.78. Even at higher label ratios, pretraining remained beneficial, suggesting that the learned spectral representations improved optimization and reduced overfitting to nuisance factors. These results support the practical value of using unlabeled spectral archives to reduce annotation cost while maintaining strong downstream performance.

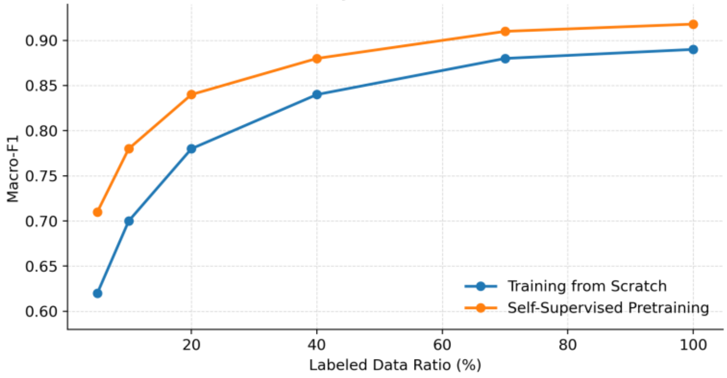

Figure 3: Label Efficiency Curve

4.6. Ablation Study on Key Components

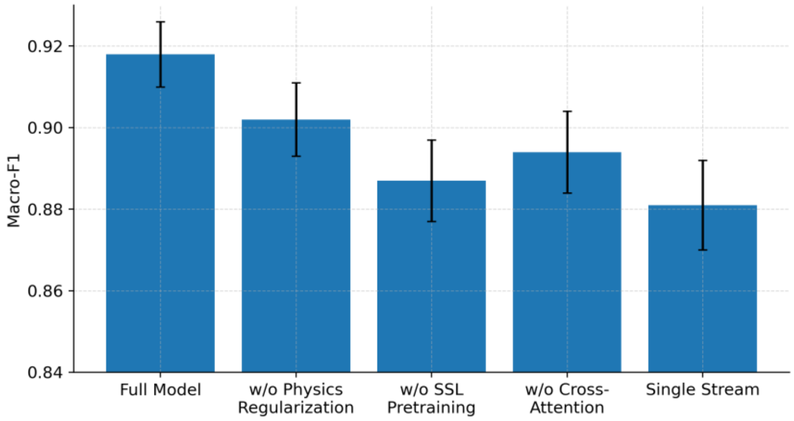

Figure 4: Ablation Study Bar

The contribution of major components is summarized in Figure 4. Removing physics regularization caused a measurable decrease in Macro-F1 (from 0.918 to 0.902), indicating that the physical consistency constraints helped stabilize learning under illumination variation and spectral artifacts. Removing self-supervised pretraining produced a larger drop (to 0.887), confirming that representation initialization from masked modeling and invariance learning was a major factor for both accuracy and robustness. Disabling cross-attention reduced Macro-F1 to 0.894, implying that cross-stream interaction contributed meaningfully beyond simple concatenation or late fusion. Finally, collapsing the architecture into a single stream further degraded performance (0.881), consistent with the hypothesis that dual-view modeling improved the coverage of informative spectral cues under varying conditions.

4.7. Diagnostic Analyses: Confusion, Calibration, Embeddings, Robustness, and Training Dynamics

Class-level behavior is illustrated by the confusion matrix in Figure 5. The dominant mass was concentrated on the diagonal entries, indicating stable recognition across all four categories. The remaining confusions were concentrated in a few off-diagonal pairs, which is consistent with practical spectral settings where adjacent classes share partially overlapping absorption or scattering characteristics.

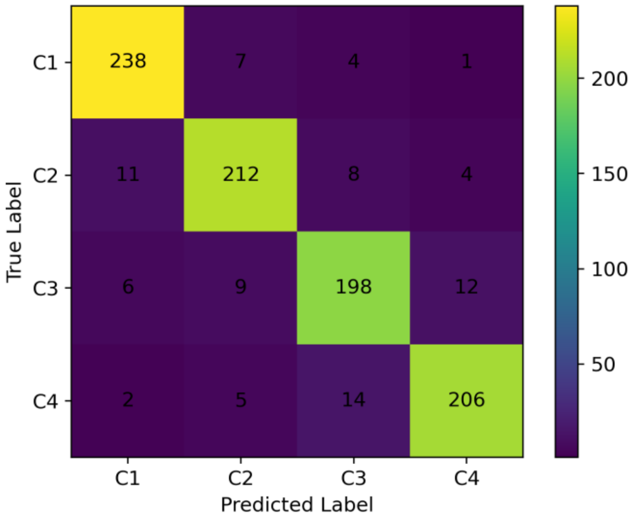

Figure 5: Confusion Matrix

Model confidence quality was evaluated via the reliability diagram in Figure 6. The predicted probabilities tracked empirical accuracy closely, yielding an expected calibration error (ECE) of 0.020, which indicates well-aligned confidence estimates. Such calibration is important in optoelectronic deployments where uncertain predictions may trigger re-measurement or human inspection.

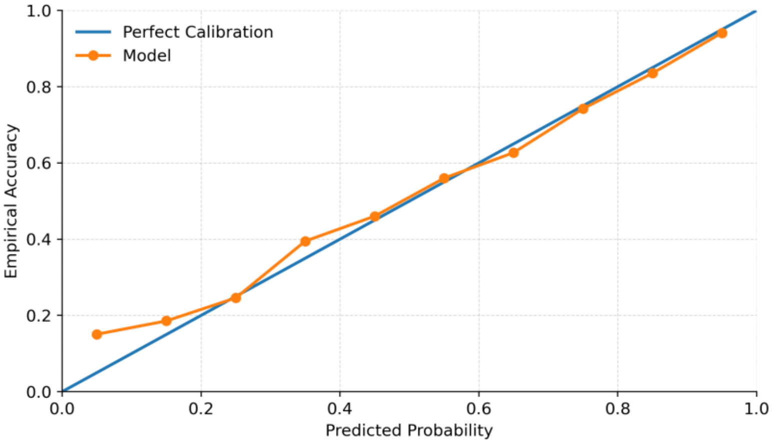

Figure 6: Reliability Diagram ECE

Representation separability was visualized using a two-dimensional projection of learned embeddings, as shown in Figure 7. The classes formed comparatively compact clusters with reduced overlap, supporting the interpretation that the learned representation space preserved discriminative structure and helped downstream decision boundaries remain stable.

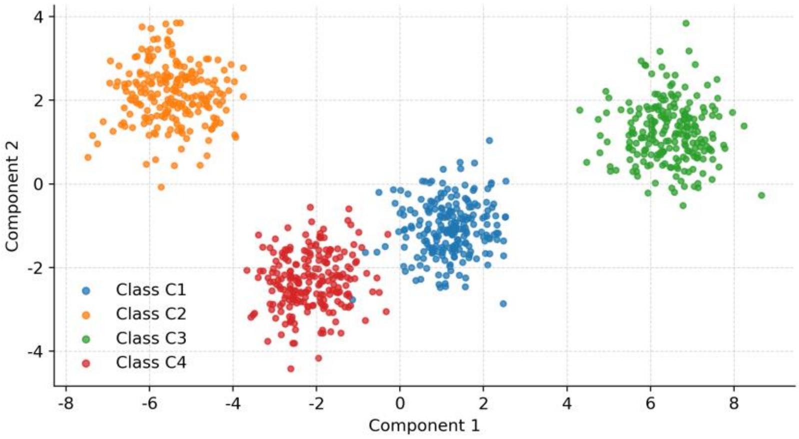

Figure 7: Embedding PCA Scatter

Robustness to illumination variation was examined by scaling illumination intensity, as shown in Figure 8. The variant without illumination-consistency regularization exhibited larger performance degradation at extreme scaling, whereas the physics-regularized setting maintained higher Macro-F1 across the full range (e.g., around 0.918 near nominal scaling and remaining comparatively stable under over-/under-illumination). This result is consistent with the objective of physics-guided constraints: reducing sensitivity to nuisance intensity scaling while preserving task-relevant spectral signatures.

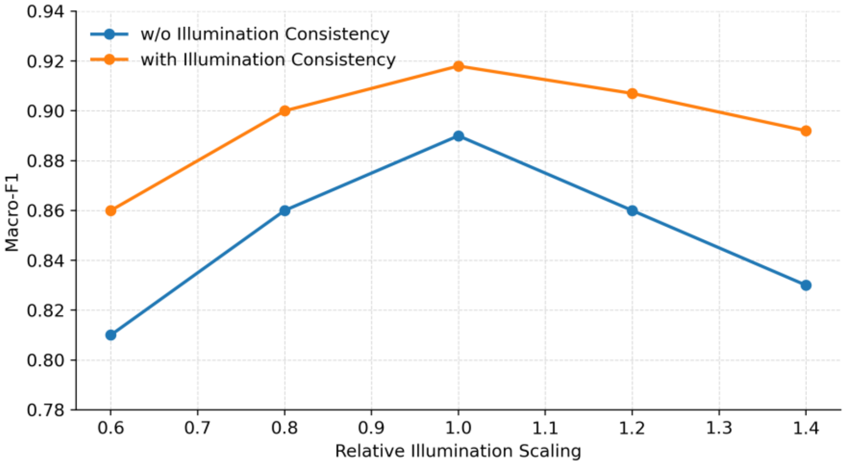

Figure 8: Illumination Robustness Line

Finally, optimization stability was summarized by the training curves in Figure 9. Training and validation losses decreased smoothly and converged without strong divergence, indicating stable learning dynamics and limited overfitting under the adopted training strategy and regularization scheme.

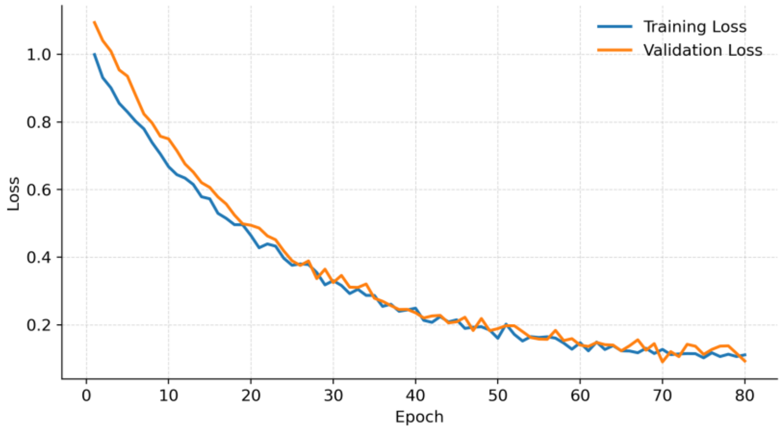

Figure 9: Training Curves

5. Conclusions

This study developed a physics-guided self-supervised dual-stream Transformer framework for robust optoelectronic spectral analytics under practical disturbances such as illumination variation, measurement noise, and cross-device distribution shift. The overall design combined radiometric-consistent preprocessing, patch-based Transformer encoding, cross-attention with gated fusion for complementary-view aggregation, and self-supervised pretraining to reduce label dependence. In addition, physics-regularized objectives were integrated to enhance invariance and improve out-of-distribution stability.

Experimental results on collected optoelectronic spectral datasets demonstrated that the proposed method achieved consistently superior Accuracy and Macro-F1 compared with representative baselines, and delivered clear advantages in cross-device testing and low-label fine-tuning. Ablation analyses further verified that self-supervised pretraining, physics regularization, and cross-stream interaction each contributed materially to the observed improvements. Diagnostic evaluations, including class-level confusion behavior, representation separability, probability calibration, robustness to illumination scaling, and training dynamics, collectively supported the reliability and deployability of the proposed approach.

Future work may extend the framework to broader optoelectronic modalities (e.g., spectral imaging and multi-sensor fusion), incorporate more explicit physical forward models for instrument-specific effects, and explore uncertainty-aware decision policies for closed-loop measurement and quality control.

References

[1] D. Passos, P. Mishra, "Perspectives on deep learning for near-infrared spectral data modelling," NIR News, vol. 33, no. 7–8, pp. 9–12, 2022. DOI: 10.1177/09603360221142821.

[2] A. Puleio, et al., "Calibration of spectra in presence of non-stationary background using unsupervised physics-informed deep learning," Scientific Reports, 2023. DOI: 10.1038/s41598-023-29371-9.

[3] A. P. Vorozhtsov, P. V. Kitina, "Contrastive representation learning for spectroscopy data analysis," Mendeleev Communications, 2024. DOI: 10.1016/j.mencom.2024.10.006.

[4] R. Bushuiev, A. Bushuiev, R. Samusevich, C. Brungs, et al., "Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS," Nature Biotechnology, 2025. DOI: 10.1038/s41587-025-02663-3.

[5] I. Bayrakli, E. Eken, "Compact laser spectroscopy-based sensor using a transformer-based model for analysis of multiple molecules," Applied Optics, vol. 63, no. 26, pp. 6941–6947, 2024. DOI: 10.1364/AO.534782.

[6] A. L. Bowler, M. P. Pound, N. J. Watson, "Domain Adaptation for In-Line Allergen Classification Using Near-Infrared Spectroscopy," Sensors, 2022. DOI: 10.3390/s22031084.

[7] F. B. Lavoie, G. Robert, A. Langlet, R. Gosselin, "Calibration transfer by likelihood maximization: A standard-free

approach capable of handling non-overlapping wavelength ranges," Chemometrics and Intelligent Laboratory Systems, 2023, vol. 234, p. 104766. DOI: 10.1016/j.chemolab.2023.104766.

[8] M. Alberts et al., "Leveraging infrared spectroscopy for automated structure elucidation," Communications Chemistry, 2024. DOI: 10.1038/s42004-024-01341-w.

[9] O. C. Koyun et al., "RamanFormer: A Transformer-Based Quantification Approach for Raman Mixture Components," ACS Omega, 2024. DOI: 10.1021/acsomega.3c06135.

[10] S. K. Paidi, P. Maheshwari, "RamanMAE: Masked Autoencoders Enable Efficient Molecular Imaging by Learning Biologically Meaningful Spectral Representations," Analytical Chemistry, 2025. DOI: 10.1021/acs.analchem.5b01234.

[11] P. Rocabert-Oriols, C. Lo Conte, N. López, J. Heras-Domingo, "Multi-modal contrastive learning for chemical structure elucidation with VibraCLIP," Digital Discovery, 2025. DOI: 10.1039/d5dd00269a.

[12] S. Guglielmini, V. Banchieri, F. Scholkmann, M. Wolf, "Transformer-based deep learning model for predicting fNIRS short-channel signals," Neurophotonics, 2025. DOI: 10.1117/1.NPh.12.1.011003.

[13] C. Lange et al., "Comparing machine learning methods for chemical mixture analysis using Raman spectra from eight spectrometers," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2025. DOI: 10.1016/j.saa.2025.125861.

[14] C. Lange et al., "A Setup for Automatic Raman Measurements in High-Throughput Experimentation," Biotechnology and Bioengineering, 2025. DOI: 10.1002/bit.27283.

[15] N. Coca-Lopez et al., "Artificial Intelligence-Powered Raman Spectroscopy through Open Science and FAIR Principles," ACS Nano, 2025. DOI: 10.1021/acsnano.5b05625.