Bridging Class Imbalance and Partial Labeling via Spectral-Balanced Energy Propagation for Skeleton-based Action Recognition

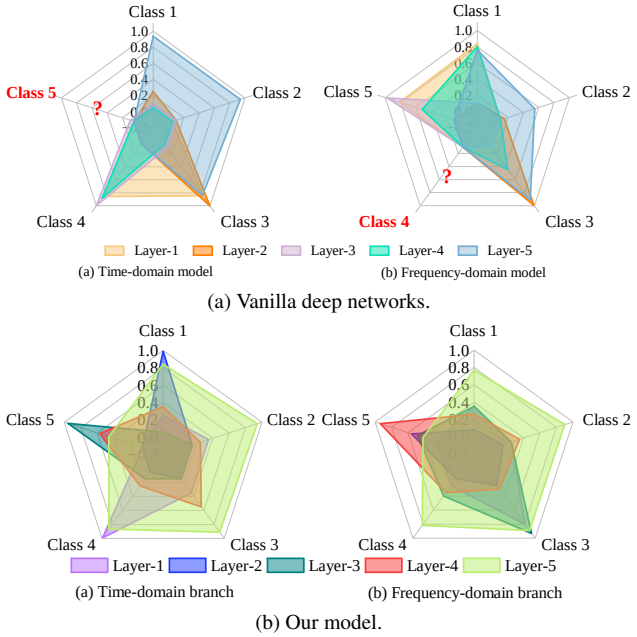
Supplementary Material

In this Supplementary Material, we first provide the layer-wise analysis of deep neural networks in Section 1. Proofs of the theorems presented in the main paper are included in Section 2. Furthermore, we detail the experimental setup in Section 3 and present further analysis of our experimental results in Section 4.

1. Layer-wise analysis of deep networks

In our research, to further evaluate the effectiveness of both time and frequency domain representations, we conducted experiments on traditional time series classification tasks using a 5-layer Fully Convolutional Network (FCN) [9] as the backbone. These analyses on standard time series benchmarks provide more generalizable insights into representation learning. Our findings, illustrated in Fig. A1 (a), reveal that the time-domain model struggles to capture features of Class 5, while the frequency-domain model overlooks Class 4, underscoring the need for cross-domain integration to enhance representation capabilities across various time series analysis tasks. More specifically, the layer-wise analysis indicates that deep layers of both models perform poorly on minority classes 4 and 5, highlighting the challenge posed by class imbalance in cross-domain approaches that primarily focus on deep representations [10]. In contrast, shallow and middle layers effectively identify minority classes, suggesting that high-level feature biases toward majority classes overshadow distinctive minority class features during data propagation. Therefore, leveraging information from these intermediate layers is crucial for developing robust representations that effectively address class imbalance.

Fig. A1 (b) shows the layer-wise performance of our model with the proposed shallow and middle feature fusion modules. Compared to Fig. A1 (a), our attention modules imposed on the shallow and middle layers of the backbone effectively preserve critical information from minority samples in both time and frequency domains, especially in minority classes 4 and 5. Additionally, as discussed in the main text. Specifically, the shallow-layer attention module is engineered to preserve critical information from minority classes, while the mid-layer attention module further refines the model's discriminative capability for their unique features. Furthermore, the time-frequency cross-domain feature fusion improves the model's comprehensive understanding of the input as well, thereby further enhancing its overall representation capabilities.



(a) Vanilla deep networks.

(b) Our model.

Figure A1. Layer-wise performance of the time and frequency domain models on the SleepEDF [5] dataset with 10% labeled to total samples. The supervision for unlabeled samples were provided by vanilla label propagation [2].

2. Mathematical Proofs

2.1. The proof of Theorem 3.1

Since Label Propagation (LP) algorithm can be formulated as a linear system [11], in this subsection, we analyze this linear system formulation [7] to investigate how class imbalance affects LP's performance. Specifically, we focus on scenarios where the class imbalance ratio does not compromise the connectivity of the propagation graph, ensuring that information can still propagate effectively between nodes.

Definition 1 *(Steady State). For a Label Propagation (LP) algorithm , a state* $\boldsymbol{M}^* \in \mathbb{R}^{n\times C}$ *is called a steady state if it satisfies:*

(i) $\boldsymbol{M}^*=\alpha\boldsymbol{S}\boldsymbol{M}^*+(1-\alpha)\boldsymbol{Y}$*, where* $\alpha\in(0,1)$.

(ii) $\|\boldsymbol{M}(t+1)-\boldsymbol{M}(t)\| \leq \epsilon$ *for some small* $\epsilon > 0$*, where* $\boldsymbol{S}$ *is the normalized similarity matrix,* $\boldsymbol{Y}$ *is the initial label matrix.*

The initial error $\epsilon(0)=\boldsymbol{M}(0)-\boldsymbol{M}^*$ reflects the levels of class imbalance and label scarcity, which can be expressed

as a linear combination of the eigenvectors and constant related to the initial condition: $\sum_{i=1}^{B}\eta_i\boldsymbol{v}_i$. According to the stability theory of linear dynamic systems [4], we have:

$$\boldsymbol{M}(t)=\tilde{\boldsymbol{M}}+\epsilon(t). \tag{A1}$$

Ordering the eigenvalues in non-increasing order by absolute value, we have $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_B$, with $|\lambda_i|\leq 1$ for all i. Let $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_B\}$ be the corresponding orthogonal eigenvectors, satisfying $\boldsymbol{S}\boldsymbol{v}_i=\lambda_i\boldsymbol{v}_i$ for each i. Then, we have $(\alpha\boldsymbol{S})^t\boldsymbol{v}_i=(\alpha\lambda_i)^t\boldsymbol{v}_i$. Substituting it into the LP updating rule $\boldsymbol{M}(t+1)=\alpha\boldsymbol{S}\cdot\boldsymbol{M}(t)+(1-\alpha)\boldsymbol{Y}$, we can transform the error into:

$$\epsilon(t+1)=\sum_{i=1}^{B}\eta_i(\alpha\lambda_i)^{t+1}\boldsymbol{v}_i. \tag{A2}$$

Note that the steady-state solution $\boldsymbol{M}^*$ corresponds to the principal eigenvalue $\lambda=1$, while $\epsilon(t)$ is associated with the remaining eigenvalues $\lambda_i<1$. Applying the triangle inequality yields:

$$\|\epsilon(t)\|\leq B\cdot\alpha^t\cdot\left(\max_{\lambda_k\in\sigma(\boldsymbol{S}),|\lambda_k|<1}|\lambda_k|\right)^t\cdot\sum_{j=k}^{B}|\eta_j|. \tag{A3}$$

where $\sigma(\boldsymbol{S})$ denotes the eigenvalues of $\boldsymbol{S}$.

2.2. The proof of Theorem 3.2

For a linear system designed to perform label propagation under class-balanced conditions, we have:

$$\left(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}}\right)\tilde{\boldsymbol{M}}=(1-\alpha)\,\tilde{\boldsymbol{Y}}, \tag{A4}$$

where $\boldsymbol{I}$ is the identity matrix, $\alpha\in(0,1)$ is the propagation parameter, $\tilde{\boldsymbol{S}}$ denotes the similarity matrix under class-balanced conditions, $\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}}$ denotes the system matrix, $\tilde{\boldsymbol{M}}$ and $(1-\alpha)\tilde{\boldsymbol{Y}}$ represent the steady-state solution and the constant term under class-balanced conditions, respectively. The perturbed linear system is expressed as:

$$(\boldsymbol{I}-\alpha\boldsymbol{S})(\tilde{\boldsymbol{M}}+\Delta\boldsymbol{M})=(1-\alpha)\boldsymbol{Y}, \tag{A5}$$

where $\boldsymbol{S}=\tilde{\boldsymbol{S}}+\Delta\boldsymbol{S}$ and $\boldsymbol{Y}=\tilde{\boldsymbol{Y}}+\Delta\boldsymbol{Y}$ denote the perturbed matrices, and $\Delta\boldsymbol{M}$ denotes the perturbations caused by class imbalance. The expression is simplified as:

$$\Delta\boldsymbol{M}=\left(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}}\right)^{-1}\left((1-\alpha)\,\Delta\boldsymbol{Y}+\alpha\Delta\boldsymbol{S}\boldsymbol{M}\right). \tag{A6}$$

As per our pre-assumption, the perturbation in the LP similarity matrix, $\Delta\boldsymbol{S}$, is a result of the induced class imbalance $\Delta\boldsymbol{Y}$, which can be expressed as:

$$\boldsymbol{S}=\tilde{\boldsymbol{S}}+d\left(g(\Delta\boldsymbol{Y})\right). \tag{A7}$$

Here, g is a function that maps changes in the label distribution $\Delta\boldsymbol{Y}$ to corresponding shifts in the feature space. d is a distance function that quantifies the difference between the class-balanced similarity matrix $\tilde{\boldsymbol{S}}$ and the perturbed one $\boldsymbol{S}$, measuring the impact of the class imbalance on the similarity structure. Therefore, $\Delta\boldsymbol{S}$ can be expressed as a higher-order term in $\Delta\boldsymbol{Y}$. Moreover, the inherent normalization properties of $\boldsymbol{S}$ further weaken the impact of $\Delta\boldsymbol{S}$.

This indicates that $\Delta\boldsymbol{S}$ is a smaller perturbation compared to $\Delta\boldsymbol{Y}$, and its impact on the overall system can be considered negligible for small class imbalance. Thus, this leads to the simplified expression for $\Delta\boldsymbol{M}$:

$$\Delta\boldsymbol{M}\approx\left(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}}\right)^{-1}\left((1-\alpha)\,\Delta\boldsymbol{Y}\right). \tag{A8}$$

Here, $(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}})^{-1}$ is an invertible matrix. Taking the L2-norm of $\Delta\boldsymbol{M}$, we get:

$$\|\Delta\boldsymbol{M}\|\leq\|(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}})^{-1}\|\,(1-\alpha)\,\|\Delta\boldsymbol{Y}\|. \tag{A9}$$

Since

$$\begin{aligned}&\|(1-\alpha)\boldsymbol{Y}\|=\|(\boldsymbol{I}-\alpha\boldsymbol{S})\boldsymbol{M}\|\leq\|\boldsymbol{I}-\alpha\boldsymbol{S}\|\cdot\|\boldsymbol{M}\|\\&\Rightarrow\|\boldsymbol{M}\|\geq\frac{(1-\alpha)\cdot\|\boldsymbol{Y}\|}{\|\boldsymbol{I}-\alpha\boldsymbol{S}\|}.\end{aligned} \tag{A10}$$

Therefore, the relative error in the solution satisfies:

$$\frac{\|\Delta\boldsymbol{M}\|}{\|\boldsymbol{M}\|}\leq\|(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}})^{-1}\|\cdot\|\Delta\boldsymbol{Y}\|\cdot\frac{\|\boldsymbol{I}-\alpha\boldsymbol{S}\|}{\|\boldsymbol{Y}\|} \tag{A11}$$

Note that while both structural changes ($\Delta\boldsymbol{S}$) and label perturbation ($\Delta\boldsymbol{Y}$) contribute to the error bound, the main impact of class imbalance is predominantly captured by label perturbation $\Delta\boldsymbol{Y}$. Consequently, we can obtain a more concise error bound without losing key insights into the system's behavior. Thus, for small disturbance, we have:

$$\frac{\|\Delta\boldsymbol{M}\|}{\|\boldsymbol{M}\|}\leq\frac{1-\alpha\lambda_{\min}}{1-\alpha}\cdot\frac{\|\Delta\boldsymbol{Y}\|}{\|\boldsymbol{Y}\|}. \tag{A12}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue of $\boldsymbol{S}$.

2.3. Derivation of Equation (3)

Let $\boldsymbol{M}=\tilde{\boldsymbol{M}}+\Delta\boldsymbol{M}$ denotes the perturbed propagation matrix, for the perturbed label propagation linear system in Eq. (A5), we have:

$$\left(\boldsymbol{I}-\alpha\left(\tilde{\boldsymbol{S}}+\Delta\boldsymbol{S}\right)\right)\boldsymbol{M}=(1-\alpha)\left(\tilde{\boldsymbol{Y}}+\Delta\boldsymbol{Y}\right). \tag{A13}$$

Thereby,

$$\Delta\boldsymbol{S}=\frac{1}{\alpha}\left[\left(\boldsymbol{I}-\alpha\tilde{\boldsymbol{S}}\right)\Delta\boldsymbol{M}-(1-\alpha)\,\Delta\boldsymbol{Y}\right]\boldsymbol{M}^{-1}. \tag{A14}$$

To qualitatively explore the effect of class imbalance and label perturbation on the eigenvalue of $\boldsymbol{S}$, we analyze how they influence the $\Delta\boldsymbol{S}$. Specifically, we can derive the L2-norm of $\Delta\boldsymbol{S}$ as:

$$\|\Delta\boldsymbol{S}\| \leq \frac{1-\alpha}{\alpha}\|\Delta\boldsymbol{Y}\| \cdot \|\boldsymbol{M}^{-1}\| \cdot C, \tag{A15}$$

where C is a constant related to the state of system. This implies that the upper bound of the norm of $\Delta\boldsymbol{S}$ increases with $\|\Delta\boldsymbol{Y}\|$, i.e., the degree of class imbalance. Since $\tilde{\boldsymbol{S}}$ is symmetric, by the spectral theorem, it is diagonalizable and there exists an orthogonal matrix $\boldsymbol{V}$ such that $\tilde{\boldsymbol{S}}=\boldsymbol{V}\Lambda\boldsymbol{V}^\top$. Therefore, the Bauer-Fike theorem [1] can be applied to qualitatively investigate the relationship between class imbalance and eigenvalue deviation in matrix $\boldsymbol{S}$, demonstrating how class imbalance can induce spectral shifts in the system. According to the Bauer-Fike theorem, any eigenvalue λ of $\boldsymbol{S}$ satisfies:

$$\min_{\tilde{\lambda}\in\sigma(\tilde{\boldsymbol{S}})} |\tilde{\lambda}-\lambda| \leq \kappa(\boldsymbol{V}) \cdot \|\Delta\boldsymbol{S}\| = \|\Delta\boldsymbol{S}\|. \tag{A16}$$

where $\sigma(\tilde{\boldsymbol{S}})$ denotes the eigenvalue set of the perturbed matrix $\tilde{\boldsymbol{S}}$. Since $\boldsymbol{V}$ is an orthogonal matrix, we have the condition number $\kappa(\boldsymbol{V}) = \|\boldsymbol{V}\| \cdot \|\boldsymbol{V}^{-1}\| = \|\boldsymbol{V}\| \cdot \|\boldsymbol{V}^\top\| = 1$.

Therefore, the upper bound of the eigenvalue deviation simplifies to:

$$\min_{\tilde{\lambda}\in\sigma(\tilde{\boldsymbol{S}})} |\tilde{\lambda}-\lambda| \leq \frac{1-\alpha}{\alpha}\|\Delta\boldsymbol{Y}\| \cdot \|\boldsymbol{M}^{-1}\|. \tag{A17}$$

2.4. Derivation of Equation (6)

Let $[M_{ic}, M_{i1}, \ldots, M_{i,C-1}]$ denote the propagation scores of sample $\boldsymbol{x}$, with $M_{ic} \geq M_{i1} \geq \ldots \geq M_{i,C-1}$, where M_{ic} represents the propagation score for the target class c. The weighted propagation energy of $\boldsymbol{x}_i$ is then calculated as:

$$E_p(\boldsymbol{x}_i) = -\log\left(\sum_{j=1}^{Q} \exp\frac{M_{ij}}{p_j}\right). \tag{A18}$$

Further, let the score margin between the target class c and a non-target class $k\in\{1,\ldots,C-1\}$ be defined as $\delta=M_{ic}-M_{ik}$. We then analyze the behavior of $E(\boldsymbol{x})$ under two distinct scenarios based on the magnitude of δ:

Case 1: For all non-target classes $k \neq c$, when the propagation scores satisfy $M_{ij} \ll M_{ic}$, $E(\boldsymbol{x}_i)$ can be approximated as:

$$E_p(\boldsymbol{x}_i) \approx -\log\left(\exp\frac{M_{ij}}{p_j}\right) = -\frac{M_{ij}}{p_j}. \tag{A19}$$

Therefore, when the score difference δ is sufficiently large, $E(\boldsymbol{x}_i)$ is primarily determined by the propagation score of the target class. This is desirable for a label propagation metric, as it effectively reflects the target class scores.

Case 2: Assuming there exists a non-target class k such that the score difference $\delta = M_{ic}-M_{ik}\approx 0$, and for $j\notin\{c(\boldsymbol{x}), k\}$, assume $M_{ij} \ll M_{ic}$. Thus, $E(\boldsymbol{x}_i)$ can be expressed as follows:

$$\sum_{j=1}^{C} \exp\frac{M_{ij}}{p_j} = \exp\frac{M_{ic}}{p_c} + \exp\frac{M_{ik}}{p_k} + \sum_{j=1, j\notin\{c,k\}}^{C} \exp\frac{M_{ij}}{p_j}, \tag{A20}$$

Since $M_{ik} = M_{ic} - \delta$, we have $\exp\frac{M_{ik}}{p_k} = \exp\frac{M_{ic}-\delta}{p_k} = \exp\left(\frac{M_{ic}}{p_k} - \frac{\delta}{p_k}\right)$. When $\delta\approx 0$, we can use Taylor expansion: $\exp\left(-\frac{\delta}{p}\right) \approx 1 - \frac{\delta}{p}$, thus, $1 + \exp\left(-\frac{\delta}{p}\right) \approx 2 - \frac{\delta}{p}$. For small values of $\frac{\delta}{p}$, using logarithmic approximation: $\log\left(2 - \frac{\delta}{p}\right) \approx \log(2) - \frac{1}{2}\cdot\frac{\delta}{p}$.

Substituting this expression into Eq.(A18), we have:

$$\begin{aligned} E_p(\boldsymbol{x}) &\approx -\frac{M_{ic}}{p_c} - \log\left(1 + \exp\left(-\frac{\delta}{p_k}\right)\right) \\ &\approx -\frac{M_{ic}}{p_c} - \frac{\delta}{2p_k} - \log 2. \end{aligned} \tag{A21}$$

This means that even in ambiguous scenarios where a non-target class score overshadows the target class score, the propagation energy metric can still reliably reflect the high-activation for the target class c. This allows the model to treat these marginal samples as reliable and effectively leverage them for training.

The above two cases effectively cover practical scenarios in label propagation: (1) Case 1 represents high-confidence predictions where the target class score significantly dominates others ($M_{ic} \gg M_{ij}$), and (2) Case 2 captures ambiguous situations where at least one competing class score approaches the target class score ($M_{ic} \approx M_{ik}$).

2.5. Derivation of Equation (13)

To derive the derivative of the eigenvalue λ_r with respect to the weight p_m, we begin with the normalized similarity matrix defined as $\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}}$. $\boldsymbol{D}$ is the degree matrix with entries $D_{ii} = \sum_j \hat{w}_{ij}$. The eigenvalue problem associated with $\boldsymbol{S}$ is given by $\boldsymbol{S}\boldsymbol{v}_i = \lambda_r\boldsymbol{v}_i$, where $\boldsymbol{v}_i$ is the normalized eigenvector corresponding to the i-th eigenvalue $\lambda_i \in [-1, 1]$.

For $\lambda = 1$, we have $\boldsymbol{S}\boldsymbol{v}_r = \boldsymbol{v}_r$. This leads to $\boldsymbol{S}\boldsymbol{D}^{\frac{1}{2}}\mathbf{1} = \boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{D}^{\frac{1}{2}}\mathbf{1}$. Note that $\boldsymbol{D}$ is the degree matrix of $\hat{\boldsymbol{W}}$, namely $\hat{\boldsymbol{W}}\mathbf{1} = \boldsymbol{D}\mathbf{1}$. Thus, we have $\boldsymbol{S}\boldsymbol{D}^{\frac{1}{2}}\mathbf{1} = \boldsymbol{D}^{\frac{1}{2}}\mathbf{1}$. Therefore, $\boldsymbol{D}^{\frac{1}{2}}\mathbf{1}$ is the eigenvector of $\boldsymbol{S}$ corresponding to $\lambda = 1$, indicating the steady-state distribution of $\boldsymbol{S}$. When the weight matrix is adjusted, the degree matrix $\boldsymbol{D}$ is updated accordingly. Since $\boldsymbol{S}$ is normalized by $\boldsymbol{D}^{-\frac{1}{2}}$, any changes in the weight matrix are offset in $\boldsymbol{S}$, leaving the eigenvectors with eigenvalue 1 unchanged.

For $\lambda_r \neq 1$, considering that

$$\frac{\partial \hat{w}_{ij}}{\partial p_m} = -\frac{\hat{w}_{ij}(2p_m + \Delta_i + \Delta_j)}{2(p_m + \Delta_i)(p_m + \Delta_j)}, \tag{A22}$$

for the diagonal elements of the degree matrix D_{ii}, we have

$$\frac{\partial D_{ii}}{\partial p_m} = -\sum_j \frac{\hat{w}_{ij}(2p_m + \Delta_i + \Delta_j)}{2(p_m + \Delta_i)(p_m + \Delta_j)}. \tag{A23}$$

Thus, we get:

$$\frac{\partial D_{ii}^{-\frac{1}{2}}}{\partial p_m} = \frac{1}{4} D_{ii}^{-\frac{3}{2}} \sum_j \frac{\hat{w}_{ij}(2p_m + \Delta_i + \Delta_j)}{(p_m + \Delta_i)(p_m + \Delta_j)}. \tag{A24}$$

Next, use the chain rule to differentiate $\boldsymbol{S}$:

$$\begin{aligned} \frac{\partial \boldsymbol{S}}{\partial p_m} &= \frac{\partial}{\partial p_m}\left(\boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}}\right) \\ &= \left(\frac{\partial \boldsymbol{D}^{-\frac{1}{2}}}{\partial p_m}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}} + \boldsymbol{D}^{-\frac{1}{2}}\frac{\partial \hat{\boldsymbol{W}}}{\partial p_m}\boldsymbol{D}^{-\frac{1}{2}} + \boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{W}}\frac{\partial \boldsymbol{D}^{-\frac{1}{2}}}{\partial p_m}\right) \end{aligned} \tag{A25}$$

The derivative of the eigenvalue λ_r with respect to p_m can be decomposed into two distinct terms as follows:

$$\begin{aligned} \frac{\partial \lambda_r}{\partial p_m} &= \boldsymbol{v}_r^\top \frac{\partial \boldsymbol{S}}{\partial p_m}\boldsymbol{v}_r \\ &= \underbrace{2\boldsymbol{v}_r^\top\left(\frac{\partial \boldsymbol{D}^{-\frac{1}{2}}}{\partial p_m}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}}\right)\boldsymbol{v}_r}_{T_1} + \underbrace{\boldsymbol{v}_r^\top \boldsymbol{D}^{-\frac{1}{2}}\frac{\partial \hat{\boldsymbol{W}}}{\partial p_m}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{v}_r}_{T_2}. \end{aligned} \tag{A26}$$

Since $\boldsymbol{D}^{-\frac{1}{2}}\hat{\boldsymbol{W}}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{v}_r = \lambda_r \boldsymbol{v}_r$, we can express this component-wise as $\sum_j \hat{w}_{ij} D_{jj}^{-\frac{1}{2}} v_{rj} = D_{ii}^{\frac{1}{2}}\lambda_r v_{ri}$. Consequently, the terms T_1 and T_2 in Equation (A26) can be simplified as follows:

$$\begin{aligned} T_1 &= 2\sum_i v_{ri}\left(\frac{\partial D_{ii}^{-\frac{1}{2}}}{\partial p_m}\right)\sum_j \hat{w}_{ij} D_{jj}^{-\frac{1}{2}} v_{rj} \\ &= -\frac{1}{2}\sum_i v_{ri} D_{ii}^{-\frac{3}{2}} \sum_j \frac{\hat{w}_{ij}(2p_m + \Delta_i + \Delta_j)}{(p_m + \Delta_i)(p_m + \Delta_j)} D_{ii}^{\frac{1}{2}}\lambda_r v_{ri} \\ &= -\frac{1}{2}\lambda_r \sum_{i,j} \frac{(2p_m + \Delta_i + \Delta_j)}{(p_m + \Delta_i)(p_m + \Delta_j)}, \end{aligned} \tag{A27}$$

$$\begin{aligned} T_2 &= \boldsymbol{v_r}^\top \boldsymbol{D}^{-\frac{1}{2}}\frac{\partial \hat{\boldsymbol{W}}}{\partial p_m}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{v_r} \\ &= -\frac{1}{2}\lambda_r \sum_{i,j} \frac{(2p_m + \Delta_i + \Delta_j)}{(p_m + \Delta_i)(p_m + \Delta_j)}, \end{aligned} \tag{A28}$$

Therefore, Eq.(A26) can be rewritten as:

$$\frac{\partial \lambda_r}{\partial p_m} = -\lambda_r \sum_{i,j} \frac{(2p_m + \Delta_i + \Delta_j)}{(p_m + \Delta_i)(p_m + \Delta_j)}, \tag{A29}$$

where $\lambda_r \in [-1, 1)$.

3. More Experimental Details

3.1. Datasets

NTU RGB+D 60 Dataset (NTU 60) [8] consists of 56,880 3D skeleton samples from 40 subjects performing 60 actions, captured by Kinect V2. Each skeleton has 25 joints with 3D coordinates. For cross-subject evaluation, 20 subjects are used for training, and the rest for testing. For cross-view evaluation, data from Cameras 2 and 3 are for training, while Camera 1 is for testing.

NTU RGB+D 120 Dataset (NTU 120) is an extension to NTU 60, consisting of 114,480 videos with 120 categories. Two recommended protocals are presented: 1) Cross-Subject (X-sub): the training data are collected from 53 subjects, while the other 53 subjects are for testing. 2) Cross-Setup (X-setup): the training data use even setup IDs, while testing data use odd ones.

Kinetics-Skeleton [3] is derived from the Kinetics 400 video dataset through automated pose estimation. This dataset comprises 240,436 training samples and 19,796 evaluation skeleton sequences distributed across 400 action categories.

HAR includes multi-channel sensor signals from 30 subjects (aged 19–48), performing six activities. Signals capture 3-axial acceleration and angular velocity at 50Hz.

SleepEDF provides single-channel EEG signals (100Hz), covering five sleep stages: Wake (W), Non-REM (N1, N2, N3), and REM.

Epilepsy contains EEG signals from 500 subjects. Following preprocessing in [6], the data is divided into two classes for classification tasks.

3.2. Experimental Setup

In our study, SpeLER first undergoes T_{sup} epochs of supervised training, followed by T_{semi} epochs of semi-supervised learning. In our experiments, T_{sup} is set to 5 for all classical time series classification datasets and UCI HAR dataset. T_{semi} is set to 75 for the SleepEEG dataset and 115 for the others. The Adam optimizer is used with a learning rate of $1e{-}3$, the batch-size is set to 128. For the NTU 60, NTU 120 and Kinetics-Skeleton datasets, the model is trained for 300 epochs, with the first 10 focused on supervised learning, and the learning rate is set to $5e{-}4$. We implemented an early stopping mechanism with a patience value of 20 epochs. In all experiments, the contrastive loss weight μ is fixed at 0.3, and the base threshold for the propagation energy $\hat{\boldsymbol{\tau}}$ is set to -9.5. For each mini-batch, sam-

ples from the previous 3 mini-batches are used for pseudo-label generation. For fair comparison, we adapt the supervised skeleton-based action recognition backbones, CTR-GCN, FreqMix, and class-imbalanced model BRL, to our semi-supervised settings.

To create partially labeled, class-imbalanced versions of the dataset, we randomly discard training samples while maintaining a pre-defined imbalance ratio. Given an imbalance ratio of π and maximum number n_1, the number of labeled samples for each class c is calculated as $N_c{=}n_1{\times}\pi^{-\frac{c-1}{C-1}}$. We assume that both labeled and unlabeled data follow the same imbalanced distribution, reflecting a common real-world scenario.

In the main text, SpeLER with different layer configurations is compared against baselines. To determine the optimal network depth, we conducted extensive experiments varying the number of layers in our architecture, and the results are shown in Table A1. We found that the 5-layer configurations achieved the best balance between model expressivity and computational efficiency across most datasets. For the 3-layer SpeLER, a single intra-domain attention module is employed between the first and third layers. For the 7-layer SpeLER, the shallow feature fusion module is deployed between the first and the fourth layers, while the middle feature fusion module is applied between the fifth and final layers. The 5-layer configuration consistently outperformed both shallower and deeper alternatives, demonstrating that this architecture effectively captures the hierarchical information necessary for robust time series representation while avoiding the overfitting issues that can emerge with excessive depth.

Table A1. Top-1 accuracy of SpeLER with different layer configurations on NTU-60 X-Sub.

Layer number	$\pi = 60$	$\pi = 30$
3	63.91%	65.55%
5	65.02%	67.00%
7	64.85%	66.83%

4. More Experimental Analysis

4.1. Analysis on propagation energy-based tightened reliability assessment

As depicted in Fig. A2, propagation energy scoring effectively identifies the high-activation borderline samples (marked by red circles), which are potentially reliable while deemed unreliable by softmax-based confidence due to the smoothing effect. The visualization demonstrates that propagation energy allows our method to better leverage borderline samples during model training, which is critical for improving classification performance in class-imbalanced and label-scarce scenarios.

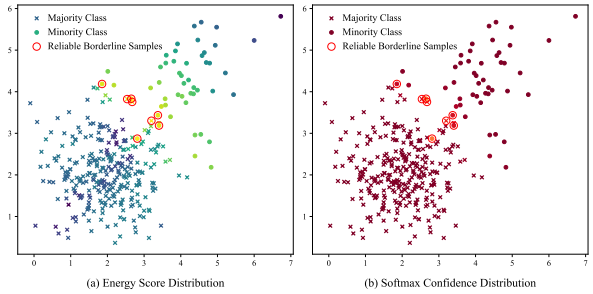


Figure A2. Comparison of pseudo-label reliability assessments for minority class samples using propagation energy and softmax confidence on NTU-60 (X-sub) ($\pi = 60$). Lighter colors indicate lower confidence. The base threshold for energy scoring is -9.5, while that for softmax-based confidence is 0.95.

4.2. Parameter sensitivity analysis

To examine the impact of predefined parameters on the SpeLER model, we performed parameter sensitivity experiments, keeping other parameters at their optimal values. As shown in Fig. A3 (a), increasing the μ value causes the model to prioritize contrastive learning, which can negatively impact the classification performance. Fig. A3 (b) shows that setting a lower $\hat{\tau}$ value restricts training to only the most reliable unlabeled samples. Therefore, setting a reasonable initial threshold or using adaptive threshold adjustment is crucial for effectively preserving borderline samples during training. Besides, using data from three mini-batches for label propagation yields the best performance, balancing effective propagation with minimal complexity from the data volume. Note that the propagation matrix has dimensions $s \times$ batch_size. Our empirical analysis reveals that the condition number remains below 5×10^3 when $s < 5$ with feature dimension 128, consistent with spectral convergence properties in random matrix theory. Furthermore, while increasing the number of neighbors k enhances pseudo-labeling precision and classification boundary definition, there exists an optimal threshold beyond which additional neighbors introduce noise and diminish performance by incorporating contextually irrelevant data points.

4.3. Model efficacy validation

To assess SpeLER, we compared its inference time with the TS-TFC model on a wearable sensor-based human action recognition dataset UCI HAR, two classical TSC datasets, which also consists of time- and frequency-domain branches. Additionally, we evaluated an ablation variant of SpeLER implementing a ResNet backbone architecture, which employs residual connections to preserve and propagate shallow representations to deeper network layers, anal-

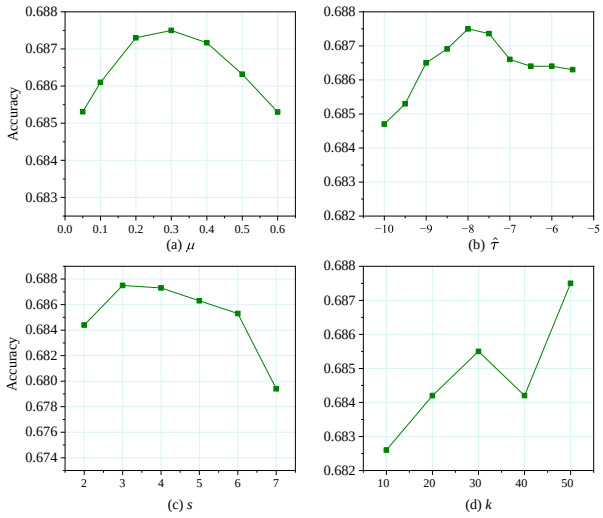


Figure A3. Parameter sensitivity experimental results on NTU-60 (X-sub). μ controls the weight of the contrastive loss in the total loss function, $\hat{\tau}$ sets the base threshold for the propagation energy, s indicates the number of mini-batch used during label propagation, and k defines the number of nearest neighbors considered during label propagation.

ogous to the function of our proposed shallow-middle attention mechanism.

Table A2. Model inference time. TS-TFC model adopts only one branch in the test phase. For fairness, we report the test time after combining results from both branches of TS-TFC.

Models	HAR	SleepEDF	Epilepsy
TS-TFC	**0.17** s	**2.85** s	0.22 s
SpeLER-Res	0.33 s	4.56 s	0.21 s
SpeLER-FCN	0.19 s	3.94 s	**0.20** s

As shown in Table A2, our SpeLER-FCN model demonstrates comparable inference efficiency to the TS-TFC model on the SleepEEG, HAR, and Epilepsy datasets, despite TS-TFC using a 3-layer CNN in each branch. This efficiency highlights the effectiveness of our intra-domain and cross-domain feature fusion modules, which enhance classification performance by efficiently integrating features across domains without significantly adding to the computational burden. The slightly longer inference time of SpeLER model on the SleepEDF dataset can be attributed to the increased processing demands of longer sequences. Furthermore, SpeLER-Res with two residual blocks exhibits the longest inference time across all datasets.

References

[1] F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960. 3

[2] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. *Semi-Supervised Learning*, pages 193–216, 2006. 1

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[4] Chi-Tsong Chen. *Linear system theory and design*. Saunders college publishing, 1984. 2

[5] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021. 1

[6] Zhen Liu, Qianli Ma, Peitian Ma, and Linghao Wang. Temporal-frequency co-training for time series semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, pages 8923–8931, 2023. 4

[7] Daniel Pfrommer, Max Simchowitz, Tyler Westenbroek, Nikolai Matni, and Stephen Tu. The power of learned locally linear models for nonlinear policy optimization. In *International Conference on Machine Learning*, pages 27737–27821, 2023. 1

[8] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 4

[9] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *International Joint Conference on Neural Networks*, pages 1578–1585, 2017. 1

[10] Chixuan Wei, Zhihai Wang, Jidong Yuan, Chuanming Li, and Shengbo Chen. Time-frequency based multi-task learning for semi-supervised time series classification. *Information Sciences*, 619:762–780, 2023. 1

[11] Zhenfeng Zhu, Jian Cheng, Yao Zhao, and Jieping Ye. Lsslp – local structure sensitive label propagation. *Information Sciences*, 332:19–32, 2016. 1