Non-Stationary Predictions May Be More Informative: Exploring Pseudo-Labels with a Two-Phase Pattern of Training Dynamics

Hongbin Pei [1] **Jingxin Hai** [1] **Yu Li** [1] **Huiqi Deng** [1]
Denghao Ma [2] **Jie Ma** [1] **Pinghui Wang** [1] **Jing Tao** [1] **Xiaohong Guan** [1]

Abstract

Pseudo-labeling is a widely used strategy in semi-supervised learning. Existing methods typically select predicted labels with high confidence scores and high training stationarity, as pseudo-labels to augment training sets. In contrast, this paper explores the pseudo-labeling potential of predicted labels that **do not** exhibit these characteristics. We discover a new type of predicted labels suitable for pseudo-labeling, termed *two-phase labels*, which exhibit a two-phase pattern during training: *they are initially predicted as one category in early training stages and switch to another category in subsequent epochs.* Case studies show the two-phase labels are informative for decision boundaries. To effectively identify the two-phase labels, we design a 2-*phasic* metric that mathematically characterizes their spatial and temporal patterns. Furthermore, we propose a loss function tailored for two-phase pseudo-labeling learning, allowing models not only to learn correct correlations but also to eliminate false ones. Extensive experiments on eight datasets show that **our proposed 2-*phasic* metric acts as a powerful booster** for existing pseudo-labeling methods by additionally incorporating the two-phase labels, achieving an average classification accuracy gain of 1.73% on image datasets and 1.92% on graph datasets.

1. Introduction

Pseudo-labeling (Lee et al., 2013) is a widely employed strategy in many semi-supervised learning methods, *e.g.*, self-training (Amini et al., 2025), co-training (Blum & Mitchell, 1998), consistency regularization (Sohn et al., 2020), and partial label learning (Tian et al., 2024), to address the practical challenge of labeled data scarcity. This strategy uses predicted labels of unlabeled samples as pseudo-labels to augment training set, facilitating model training via reducing model uncertainty (also called *epistemic* uncertainty) (Hüllermeier & Waegeman, 2021) and promoting a robust decision boundary in low-density regions (Chapelle & Zien, 2005). Pseudo-labeling has achieved success in various tasks, *e.g.*, computer vision (Rizve et al., 2021), text mining (Yang et al., 2023), and graph learning (Sun et al., 2020).

High-quality pseudo-labels are essential to this strategy, yet they are challenging to obtain. The *confidence score* is the most commonly used and intuitive metric for selecting pseudo-labels; however, it suffers from a poor calibration issue, especially on out-of-distribution data (Guo et al., 2017; Kage et al., 2024). Recently, researchers discovered training dynamics—*i.e.,* the trajectory of model predictions during training—contains rich information about prediction uncertainty (Swayamdipta et al., 2020; Jia et al., 2023). They show that if a model makes consistent predictions for an unlabeled sample throughout training, that predicted label is highly likely to be correct. Consequently, *training stationarity* metric is proposed to identify these predicted labels to act as pseudo-labels (Song et al., 2019; Zhou et al., 2020; Chen et al., 2021; Pleiss et al., 2020; Pei et al., 2024b).

In this paper, in contrast to existing works, we explore pseudo-labeling potential of a new type of predicted labels that ***do not*** exhibit high confidence score and high training stationary, as illustrated by type II labels in Fig.1(A)[1]. These predicted labels are rarely exploited for pseudo-labeling, as they are associated with atypical patterns and tend to have a high risk of misclassification. Incorrect pseudo-labels can significantly degrade model performance due to introducing noise and misleading patterns (Wang et al., 2023). However, we find that despite their low correctness, correctly predicted labels of this type could offer greater information gain[2] for pseudo-labeling, than the commonly used type

[1]MOE KLINNS Lab, Xi'an Jiaotong University, China; [2]Beijing Information Science and Technology University, China. Correspondence to: Jing Tao <jtao@mail.xjtu.edu.cn>, Pinghui Wang <phwang@mail.xjtu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]In Fig.1(A), we use an inverse indicator, non-stationary, which can be easily calculated, as detailed in the Appendix A.

[2]We quantify the information gain by measuring the gradient change from pseudo labels, detailed in Appendix E.

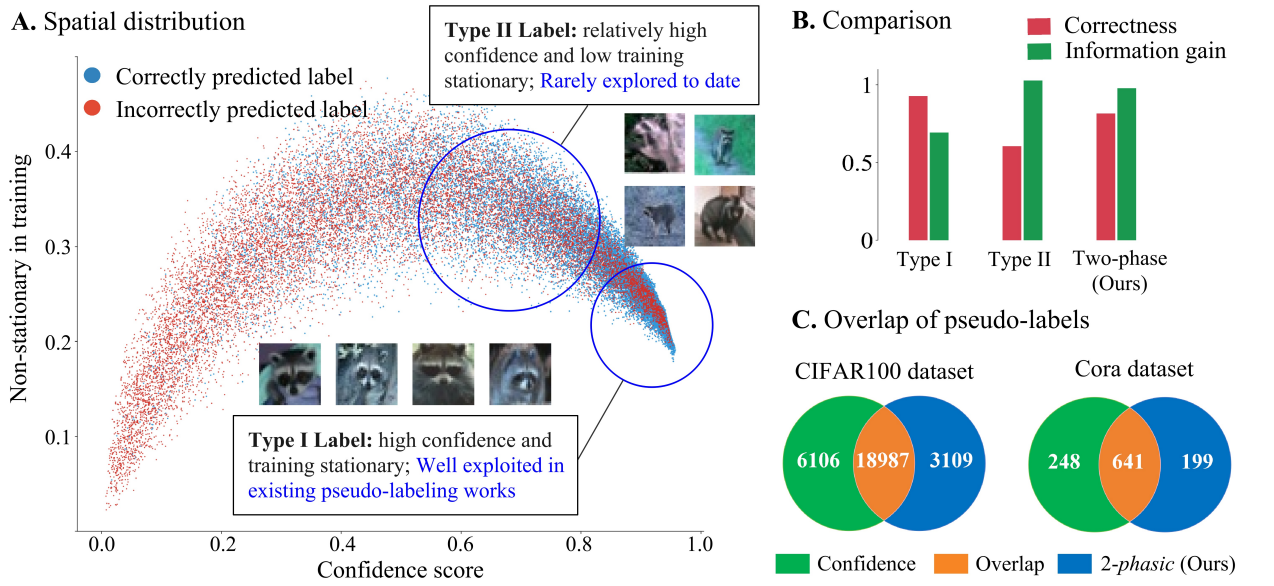

Figure 1. **(A)** Distribution of predicted labels for CIFAR100 dataset, based on a pre-trained Vision Transformers (ViT) classifier. The x-axis represents confidence score, and the y-axis denotes non-stationary during training. Colors indicate correctness of predicted labels. **Type I labels** have high confidence and high training stationarity, corresponding to samples with typical patterns; these labels have been well exploited in existing pseudo-labeling works. **Type II labels** exhibit relatively high confidence and low training stationarity, corresponding to samples with complicated patterns; these labels are rarely explored for pseudo-labeling. **(B)** Type I labels show high correctness, while Type II labels provide high information gain. Our proposed two-phase labels combine both high correctness and relatively high information gain. **(C)** Overlap of effective pseudo-labels (75% accuracy) selected by confidence and our 2-*phasic* metric. The pseudo-labels only identified by 2-*phasic* metric (blue) indicate that *many two-phase labels do not have high confidence scores.*

I labels, as shown in Fig.1(B). **This finding raises a new research problem:** *How can we leverage the significant information gain from type II labels for pseudo-labeling while minimizing the risk of misclassification?*

Toward exploring this problem, this paper uncovers a subset within type II labels, termed *two-phase labels*. The two-phase labels refer to predicted labels whose training dynamics exhibit a two-phase pattern: they are initially predicted as one category in early training stage (the first phase), and switch to another category in subsequent epochs (the second phase), as shown in Fig.2(A). They show a strong potential for pseudo-labeling, exhibiting high information gain and relatively high correctness, as shown in Fig.1(B). *Notably, two-phase labels cannot be easily identified by confidence scores,* as evidenced by non-overlaps in Fig.1(C).

We propose the 2-*phasic* metric to effectively identify the two-phase labels by mathematically characterizing their spatial and temporal patterns. We further analyze the rationale behind using two-phase labels for pseudo-labeling: (1) Two-phase labels provide valuable information about decision boundaries, as two-phase samples are positioned much closer to the boundaries than samples with high confidence; (2) They can enable models to not only learn correct correlations, but also eliminate false correlations. To fully release the two potentials, we design a specialized pseudo-labeling loss function tailored for two-phase labels.

We validate the proposed 2-*phasic* metric-based pseudo-labeling method on eight benchmark datasets, including four image datasets and four graph datasets. Experimental results show: (1) Our proposed 2-*phasic* metric can widely enhance existing pseudo-labeling methods as a booster by additionally incorporating the two-phase labels, achieving an average classification accuracy gain of 1.73% on image data and 1.92% on graph data. (2) Tow-phase labels are often overlooked by commonly used confidence metrics, and they exhibit high quality in terms of information gain and correctness. We analyze the proposed loss function by an ablation study and examine method limitations. The experiment code is released in our Github repository[3].

In summary, our contributions in this paper are three-fold:

- We uncover two-phase labels, which have strong potential for pseudo-labeling and can be a valuable complement to pseudo-labels provided by existing methods. We analyze the rationale of two-phase labels for pseudo-labeling from different perspectives.
- We propose the 2-*phasic* metric to effectively and efficiently identify two-phase labels by mathematically characterizing their spatial and temporal patterns. Furthermore, we propose a loss function specifically designed for two-phase pseudo-labeling learning.

[3]URL: https://github.com/XJTU-Graph-Intelligence-Lab/two-phasic-for-pseudo-labeling

- We conduct extensive experiments to validate the 2-*phasic* metric-based pseudo-labeling method on eight image and graph datasets. Experimental results show that the proposed 2-*phasic* metric acts as a powerful booster for existing pseudo-labeling methods.

2. Related Works

2.1. Pseudo-labeling Approach

The pseudo-labeling approach is dominant in semi-supervised learning, and it involves a class of methods that assign predicted labels to unlabeled samples that are then acted as labeled samples for training (Kage et al., 2024).

In **self-training**, pseudo-labels are generated by a model trained on labeled data and subsequently used to further train the same model (Lee et al., 2013; McClosky et al., 2006). This process can be conducted iteratively to achieve curriculum learning (Cascante-Bonilla et al., 2021; Xie et al., 2020). In **co-training**, pseudo-labels are generated by the teacher model and then used to augment the student model's training set, allowing for complementary learning (Wang & Zhou, 2013), knowledge distilling (Hinton et al., 2015), and iteratively optimizing pseudo-label assignment policy through the teacher model (Blum & Mitchell, 1998; Pham et al., 2021). Pseudo-labeling is also employed in **consistency regularization**, where the objective is to enforce consistency among the pseudo-labels (soft or hard) of perturbed samples—generated by various data augmentation methods—and the original sample (Berthelot et al., 2020; Sohn et al., 2020; Hu et al., 2021). In **partial label learning**, pseudo-labels are leveraged to disambiguate among candidate labels, thereby identifying the true label of each sample (Jin & Ghahramani, 2002; Tian et al., 2024). This paper uncovers a new type of pseudo-labels, the two-phase labels, *which are under-explored in existing research and thus may broadly benefit the aforementioned methods.*

2.2. Metrics for Pseudo-label Selection

Pseudo-label selection aims to identify which samples should be assigned pseudo-labels. It is essential to pseudo-labeling methods. In literature, the most commonly used metric for pseudo-label selection is the *confidence score* derived from softmax distribution (Sun et al., 2020; Cascante-Bonilla et al., 2021; He et al., 2023). However, this metric suffers from poor calibration, which means high confidence scores are often assigned to incorrectly predicted labels, leading to incorrect pseudo-labels (Guo et al., 2017). Moreover, it has been argued that confidence scores should not be trusted for out-of-distribution data (Gal & Ghahramani, 2016). Confidence scores are also susceptible to manipulation by adversarial examples (Nguyen et al., 2015).

Prediction uncertainty is also a critical metric for selecting pseudo-labels (Gawlikowski et al., 2023; Zhao et al., 2020). A popular method for measuring the uncertainty is *Monte Carlo dropout* (Gal & Ghahramani, 2016; Rizve et al., 2021), which measures the variance of predicted labels during multiple dropout operations in testing. Another important uncertainty metric used for pseudo-label selection is *training stationarity* (also called *time-consistency*), which is based on the training dynamics (Song et al., 2019; Zhou et al., 2020; Chen et al., 2021; Pleiss et al., 2020; Pei et al., 2024b). The metric can be viewed as a self-ensemble method, where models at different epochs act as ensemble members (Liu et al., 2022b). It relies on the observation that if a model consistently predicts the same label for an unlabeled sample throughout training, that label is likely to be correct. Unlike existing works, this work is, to the best of our knowledge, **the first to explore predicted labels with non-stationary training dynamics for pseudo-labeling.**

2.3. Training Dynamics-based Analysis

Training dynamics has been applied in various analysis tasks, as it contains rich information about both models and data. In addition to pseudo-labeling, training dynamics is used to identify samples that are frequently forgotten to address catastrophic forgetting issues (Toneva et al., 2019; Pan et al., 2020). It is employed in active learning to find ambiguous samples for expert querying, helping to refine the training set and support out-of-distribution generalization (Kye et al., 2023; Wang et al., 2022a). It is also used to measure learning difficulty in curriculum learning (Baldock et al., 2021; Zhang et al., 2025), identify important samples to reduce the training set (Paul et al., 2021), and detect mislabeled samples for data cleaning (Jia et al., 2023; Swayamdipta et al., 2020). Notably, in computer vision, many studies utilize training dynamics to enhance the embedding space; such approaches are commonly referred to as memory bank methods (Liu et al., 2022c).

3. Problem Definition

Let $D_L = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N_L}$ be a labeled dataset with N_L samples, where the vector $\boldsymbol{x}^{(i)}$ represents input features of sample i. Each sample belongs to one of C categories and is labeled with the one-hot vector $\boldsymbol{y}^{(i)} \in \{0,1\}^C$. Let $D_U = \{(\boldsymbol{x}^{(i)})\}_{i=1}^{N_U}$ be an unlabeled dataset with N_U samples, which does not include sample labels. The pseudo-labeling approach trains a parameterized model f_θ on both the labeled samples in D_L and pseudo-labeled samples in $D_P = \{(\boldsymbol{x}^{(i)}, \tilde{\boldsymbol{y}}^{(i)})\}_{i=1}^{N_P}$, where $\tilde{\boldsymbol{y}}^{(i)}$ denotes pseudo-label. These pseudo-labels are generated from the unlabeled dataset D_U. An essential issue of this approach is the selection of pseudo-labels, which aims to identify unlabeled samples whose predicted labels can be confidently assigned as pseudo-labels, thereby enhancing the training of the model f_θ. These predicted labels may be given by the model f_θ from a previous iteration in self-training or by a teacher

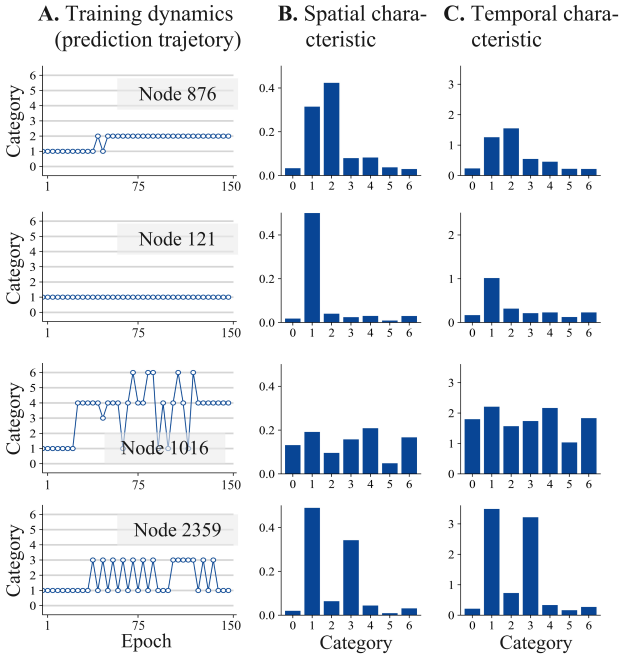

Figure 2. (**A**) Training dynamics with a two-phase pattern (Node 876), stationarity (Node 121), and frequent oscillations (Nodes 1016 and 2359), generated by a Graph Convolutional Networks (GCN) on Cora dataset. (**B**) and (**C**) summarize spatial and temporal characteristics of corresponding training dynamics, respectively

model in co-training. The pseudo-label selection problem can be modeled as designing a metric that evaluates the suitability of predicted labels to act as pseudo-labels.

4. Method

In this section, we first design the 2-*phasic* metric to effectively and efficiently identify two-phase labels, then analyze the rationale behind using two-phase labels for pseudo-labeling, and finally propose a pseudo-labeling loss that can fully extract the information within two-phase labels.

4.1. 2-*phasic* Metric

The key to designing a metric to identify two-phase labels lies in mathematically characterizing the two-phase pattern: the label is initially predicted as one class by the model in the first training phase and then shifts to a different class in the second phase, as shown in Fig. 2A. Our proposed 2-*phasic* metric characterizes this pattern through both spatial and temporal measures, defined as follows:

$$2-phasic^{(i)} := (\mu^{(i)}_{spatial})^{\eta_1}(\mu^{(i)}_{temporal})^{\eta_2}, \qquad (1)$$

where $\mu^{(i)}_{spatial}$ and $\mu^{(i)}_{temporal}$ represent the measures of spatial and temporal pattern, respectively, of the training dynamics for sample i. The parameters η_1 and η_2 are exponential weights. The product operator implies that both the spatial and temporal patterns must be satisfied simultaneously.

4.1.1. Spatial Measure

We summarize the spatial characteristics of training dynamics by averaging multiple predictions made by the model throughout model training, as follows:

$$P^{(i)} = 1/|\mathcal{T}| \sum_{t\in\mathcal{T}} \boldsymbol{\sigma}^{(i)}_t.$$

Here, the set $\mathcal{T}$ denotes a memory bank, in which each element $t \in \mathcal{T}$ specifies a training epoch used for averaging, and $\mathcal{T}$ is expected to cover the training process unbiasedly. $\boldsymbol{\sigma}^{(i)}_t = \mathrm{softmax}(\boldsymbol{z}^{(i)}_t)$ represents the softmax distribution given by the model at epoch t, with $\boldsymbol{z}^{(i)}_t$ being the C-dimensional logit vector of sample i.

We find that, as two-phase samples (*i.e.,* samples with two-phase labels) undergo a change in predicted labels during training, **their distribution $P^{(i)}$ exhibits a clear bimodal pattern**, with probabilities of two categories significantly higher than those of others, as shown in Fig.2(B). This spatial pattern effectively distinguishes two-phase samples from many other types. For example, the distribution $P^{(i)}$ of node 121 is unimodal, which is preferred by training stationarity-based methods (Pei et al., 2024b), while the distribution of node 1016 resembles a uniform distribution.

We specifically design a spatial measure to characterize and quantify the bimodal pattern of $P^{(i)}$, as

$$\mu^{(i)}_{spatial} := \mathbb{H}_{\mathrm{LMO}}[P^{(i)}] = -\sum_{c\in\,\mathcal{C}_{\mathrm{LMO}}} P^{(i)}_c \log P^{(i)}_c.$$

Here, we propose a new entropy measure, termed **L**eave-**M**aximum-**O**ut entropy (LMO entropy for short), denoted by $\mathbb{H}_{\mathrm{LMO}}[P^{(i)}]$ to measure bimodal distributions. Unlike traditional entropy, the uniqueness of LMO entropy lies in that it excludes the category with the highest probability from the entropy calculation. Specifically, the set $\mathcal{C}_{\mathrm{LMO}}$ includes every category c, except the one with the highest probability in the distribution $P^{(i)}$. The smaller the spatial measure $\mu^{(i)}_{spatial}$, the clearer the bimodal pattern.

As empirically validated in Appendix B, the proposed LMO entropy effectively captures bimodal patterns. Its rationale is as follows: *by removing the category with the highest probability, a bimodal distribution transforms into a sharply unimodal distribution, resulting in high entropy.*

4.1.2. Temporal Measure

We further design the temporal measure to complementally capture the two-phase pattern. We summarize temporal characteristics of training dynamics by accumulating changes of successive predictions during training,

$$Q^{(i)} = \sum_{t',t''\in\mathcal{T}} |\boldsymbol{\sigma}^{(i)}_{t'} - \boldsymbol{\sigma}^{(i)}_{t''}|,$$

where t' and t'' are successive epochs in set $\mathcal{T}$, with t' being the immediate predecessor of t''. The C-dimensional vector $Q^{(i)}$ captures the magnitude of prediction changes throughout training, reflecting the evolution of predictions.

We observe two distinguishing temporal patterns in two-phase samples. **Observation 1**: The magnitude of changes in prediction probabilities is relatively small and occurs within the two bimodal categories of $P^{(i)}$ rather than other categories. **Observation 2**: The changes in prediction probabilities are directional, shifting from the predicted category in the first phase to the one in the second phase.

Modelling Observation 1. We first define two deltas,

$$\Delta_{bi}^{(i)} = \frac{1}{2}(Q_{c_1}^{(i)} + Q_{c_2}^{(i)}), \quad \Delta_{\neg bi}^{(i)} = \frac{1}{C-2}\sum_{c \neq c_1, c_2} Q_c^{(i)},$$

where $\Delta_{bi}^{(i)}$ captures the average change of the two bimodal categories c_1 and c_2, which correspond to the predicted label in the first phase and the second phase, respectively. $\Delta_{\neg bi}^{(i)}$ represents the average change of other categories. We further apply a sigmoid function to the deltas to model tolerance to the changes in Observation 1, as follows

$$O_{bi}^{(i)} = \frac{1}{1 + e^{-(\Delta_{bi}^{(i)} - \epsilon_{bi})}} + \frac{1}{1 + e^{-(\Delta_{\neg bi}^{(i)} - \epsilon_{\neg bi})}}.$$

Here, ϵ_{bi} and $\epsilon_{\neg bi}$ are thresholds that control the sensitivity of the sigmoid functions. We set $\epsilon_{bi} > \epsilon_{\neg bi} > 0$ to allow for greater tolerance to changes within the two bimodal categories compared to the other categories. The measure $O_{bi}^{(i)}$ can be used to filter out samples whose predicted labels frequently switch, such as node 2359 in Fig.2, even if they exhibit a bimodal distribution $P^{(i)}$ of spatial pattern.

Modelling Observation 2. To measure the directional change of prediction probabilities, we first calculate the difference in probabilities from the bimodal category c_2 to c_1,

$$g_t^{(i)} = \boldsymbol{\sigma}_t^{(i)}(c_2) - \boldsymbol{\sigma}_t^{(i)}(c_1).$$

Notably, the difference $g_t^{(i)}$ is signed and defined at each epoch t. We then measure the directional change by

$$O_{dr}^{(i)} = g_T^{(i)} - g_{\min}^{(i)}, \text{ and } g_{\min}^{(i)} = \min\{g_t^{(i)} | \, t \in \mathcal{T}\},$$

where T denotes the latest epoch in set $\mathcal{T}$. A large $O_{dr}^{(i)}$ indicates a directional shift in prediction probabilities from c_1 to c_2, which can filter out samples whose predicted labels remain consistently ambiguous between c_1 and c_2 throughout training. These samples cannot be detected by either the measure $O_{bi}^{(i)}$ or the spatial measure.

Finally, we design the temporal measure of the two-phase pattern as a combination of measures $O_{di}^{(i)}$ and $O_{dr}^{(i)}$,

$$\mu_{temporal}^{(i)} := (O_{bi}^{(i)})^{\varphi_1}(1/O_{dr}^{(i)})^{\varphi_2},$$

where parameters φ_1 and φ_2 are exponential weights.

4.2. Rationale Analysis to Two-phase Labels

From two distinct perspectives, we analyze why our proposed two-phase labels are well-suited for pseudo-labeling.

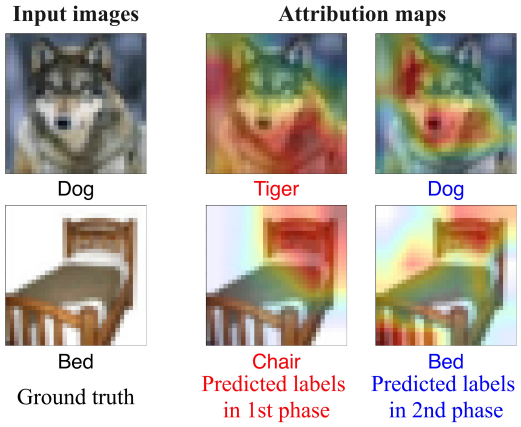

Figure 3. Attribution maps of two-phase samples in CIFAR100. (**Left**) Input images. (**Middle**) and (**Right**) are attribution maps based on ViT model in the first and second phases, respectively.

Pattern Learning Perspective. The two-phase pattern in training dynamics implies a significant shift in the patterns learned by the model, transitioning from patterns associated with category c_1 (the predicted label in the first phase) to those corresponding to category c_2 (the predicted label in the second phase). Existing studies have shown that models initially learn simple and typical patterns, gradually progress to learning more complex patterns (Arpit et al., 2017; Siddiqui et al., 2023). Given that patterns learned in later training stages are more stable and reliable (Liang et al., 2020), we conclude that two-phase samples contain two types of patterns: *simple and non-class-exclusive patterns associated with* ***category*** $\mathbf{c}_1$*; complex and class-exclusive patterns associated with* ***category*** $\mathbf{c}_2$.

To visualize these shifts of learned patterns, we present two representative images whose training dynamics exhibit a clear two-phase pattern, as shown in Fig.3. We use Grad-CAM (Selvaraju et al., 2017), a widely used attribution interpretability technique, to highlight the key regions in two-phase samples that the model relies on for classification. Details of the attribution technique are provided in Appendix C. In the first phase, the model primarily focuses on simple and non-class-exclusive patterns. For example, in the "Bed" image, the model initially emphasizes the headboard, which visually resembles the backrest of a chair, resulting in a misclassification as "Chair". In the second phase, however, the model shifts its attention to more complex, detailed, and class-exclusive patterns, such as the footboard in the "Bed" image and the nose and ears in the "Dog" image. This shift enables the model to finally make correct classifications.

This shift suggests two-phase labels are well-suited for pseudo-labeling in the following two ways: (i) Two-phase labels can help **eliminate false correlations** between labels and simple, non-exclusive patterns, *i.e.,* the c_1 category and the patterns learned in the first phase, thereby reducing the risk of misclassification. (ii) Two-phase labels can provide

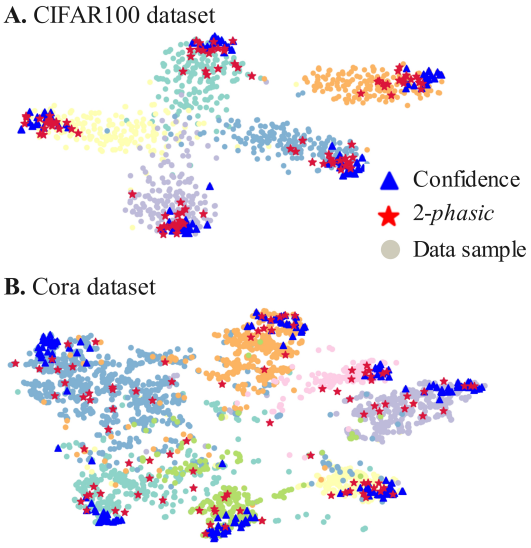

Figure 4. Two-phase samples are closer to decision boundaries in 2D representation space. Blue triangles and red stars denote samples with the highest confidence and *2-phasic* metric, respectively.

additional information to **strengthen correct correlations** between labels and complex, class-exclusive patterns, *i.e.,* the c_2 category and the patterns learned in the second phase. To fully release the two potentials of two-phase labels, we design a specialized pseudo-labeling loss function tailored for two-phase labels, as presented in the next subsection.

Decision Boundary Perspective. To observe how two-phase samples are positioned in representation space, we conduct a visualization study. We extract representations from the last layer of a ViT model for images in CIFAR100 dataset and from the last layer of a GCN (Kipf & Welling, 2017) for nodes in the Cora dataset. These high-dimensional representations are projected into a 2D space using t-SNE (Van der Maaten & Hinton, 2008) for visualization.

As shown in Fig.4, samples with high confidence scores (blue triangles) are distant from decision boundaries, especially on Cora dataset, indicating they are safe choices as pseudo-labeling but may not contribute much additional information to model. In contrast, two-phase samples selected by 2-*phasic* metric (red stars) are positioned much closer to the decision boundaries, suggesting they can offer valuable information about boundaries. Existing studies have shown that samples near decision boundaries generally have a greater influence on model performance than those positioned farther away (Cortes, 1995; Wei et al., 2021).

4.3. Two-phase Pseudo-labeling Loss

As analyzed above, two-phase labels can both enhance correct correlations and eliminate false correlations. To fully leverage this valuable information to train model, we design a specialized pseudo-labeling loss function. For a general multi-label classification task, we present a modified binary cross-entropy loss for a two-phase label:

$$\mathcal{L}(\tilde{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)}) = -(\tilde{y}^{(i)}_{c_2} \log(\hat{y}^{(i)}_{c_2}) + (1-\tilde{y}^{(i)}_{c_1}) \log(1-\hat{y}^{(i)}_{c_1})),$$

where $\tilde{\boldsymbol{y}}^{(i)}$ denotes a two-phase pseudo-label, specifically with the bimodal categories $\tilde{y}^{(i)}_{c_2} = 1$ and $\tilde{y}^{(i)}_{c_1} = 0$, and $\hat{\boldsymbol{y}}^{(i)} = f_\theta(\boldsymbol{x}^{(i)})$ represents the prediction probability given by the model. In this loss function, the first c_2 term is designed to learn correct correlations between complex, class-exclusive patterns and category c_2, and the second c_1 term enforces the elimatation to false correlations between simple, non-exclusive patterns and category c_1. The two terms are further analyzed in the following ablation study. A pseudo-labeling algorithm using 2-*phasic* metric is given in Appendix D.

5. Experiment

We empirically validate our proposed 2-*phasic* metric and pseudo-labeling algorithm on both image and graph datasets, by answering four key questions. **Q1 (Booster Test):** Can the 2-*phasic* metric enhance existing pseudo-labeling methods as a booster by additionally incorporating two-phase labels? **Q2 (Complementary Analysis):** Are two-phase labels high-quality pseudo-labels that are missed by existing pseudo-labeling metrics? **Q3 (Ablation Study):** How do the two terms in our specialized loss function contribute to performance? **Q4 (Parameter Sensitivity):** How do the parameters in the 2-*phasic* metric impact performance?

Baselines. We evaluate our proposed method based on five state-of-the-art baselines, including both confidence-based and uncertainty-based pseudo-labeling methods. On graph datasets, we use four baselines: Confidence score, AUM (Sosea & Caragea, 2022), MoDis (Pei et al., 2024b), and DR-GST (Liu et al., 2022a); on image datasets, we use four baselines: Confidence score, MoDis, UPS (Rizve et al., 2021), and Softmatch (Chen et al., 2023).

Datasets. Experiments are conducted on eight benchmark datasets, including four image datasets, CIFAR-100 (Krizhevsky, 2012), EuroSAT (Helber et al., 2019), Semi-Aves (Su & Maji, 2021), and STL-10 (Coates et al., 2011), and four graph datasets, Cora (McCallum et al., 2000), Citeseer (Sen et al., 2008), Pubmed (Namata et al., 2012), and AmazonComputers (McAuley et al., 2015). Details of baselines, datasets, and metrics are provided in Appendix E.

5.1. Experiment 1: Booster Test

Experimental protocol. To evaluate the pseudo-labeling potential of two-phase labels, we test whether the 2-*phasic* metric can **enhance baselines as a booster by additionally incorporating two-phase labels**. We design a controlled experiment with a baseline trial and a 2-*phasic* trial following a three-stage protocol, as illustrated in Fig.5. In stage 1, a base model is established by either fine-tuning a pre-trained model or training a model on labeled data. In

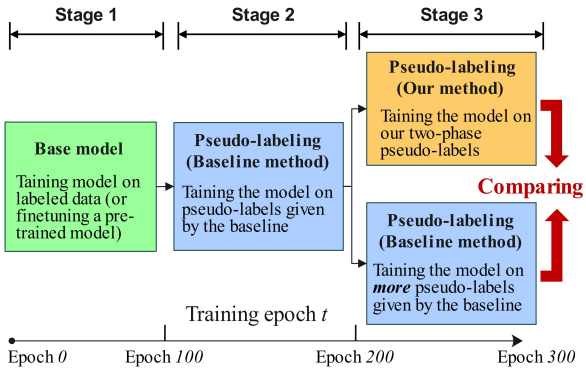

Figure 5. An illustration of the protocol of experiment 1.

stage 2, the model is trained using pseudo-labels generated by a baseline. In stage 3, for the baseline trial, the baseline continues to generate pseudo-labels for model training; for the 2-*phasic* trial, the model is trained on two-phase labels given by our 2-*phasic* metric. Finally, we compare the performance of models trained in the two trials to evaluate whether the 2-*phasic* metric improves the baseline.

We employ the widely used GCN as the base model for graph datasets, which is initially trained using only 3, 5, and 10 labeled nodes per category (L/C), following the existing setup (Liu et al., 2022a) to create challenging scenarios with minimal label information. We employ a pre-trained ViT model for image datasets and fine-tune it using a few labeled images. The total number of the labeled images (# Labels) is set according to (Li et al., 2023; Wang et al., 2022b), as shown in Table 1. Additionally, we introduce a more challenging scenario with only 100 labeled images on CIFAR-100. More details can be found in Appendix E.

Table 1 summarizes the classification accuracies after stage 3 of the experimental procedure on eight image and graph datasets. We repeat the experiments 10 times and report average results on the graphs to minimize the impact of randomness. We observe that incorporating two-phase labels through our 2-*phasic* metric generally improves the performance of these baselines. Specifically,

- The 2-*phase* trials consistently outperform the baseline trials across baselines, datasets, and base models. The average accuracy gain reaches 1.73% for image datasets and 1.92% for graph datasets, indicating additional information is provided by the two-phase labels in stage 3, compared to the baselines.
- The improvements from 2-*phasic* become more significant when the number of labeled samples is lower. This suggests that two-phase labels play a more important role in scenarios with extremely limited labels.
- The SoftMatch achieves strong performance because it well combines both a pseudo-labeling module and an unsupervised contrastive learning module.

Table 1. Classification accuracy of pseudo-labeling learning (%). L/C denotes the number of labeled nodes per category; # Labels denotes the total number of labeled images used.

Graph data	**Cora**			**Citeseer**		
L/C	3	5	10	3	5	10
Confidence	66.21	71.38	73.73	60.91	65.23	67.16
+2-*phase*	70.41	74.07	77.17	64.63	68.09	69.59
AUM	63.50	69.86	75.88	61.10	68.27	67.87
+2-*phase*	66.95	71.84	77.71	64.99	69.72	69.64
MoDis	68.61	71.85	76.32	65.74	69.17	71.88
+2-*phase*	70.10	72.83	78.25	68.63	71.92	72.01
DR-GST	71.01	77.07	81.12	61.05	69.62	73.55
+2-*phase*	73.88	78.53	81.56	66.84	71.02	74.50
Graph data	**PubMed**			**AmazonCS**		
L/C	3	5	10	3	5	10
Confidence	64.02	69.94	72.78	73.55	75.28	80.75
+2-*phase*	65.99	72.32	74.01	75.80	77.00	82.49
AUM	65.19	70.22	72.14	73.65	75.28	80.84
+2-*phase*	66.38	72.24	73.22	75.49	77.02	81.77
MoDis	65.11	71.56	74.14	75.31	76.14	81.34
+2-*phase*	65.69	72.66	74.43	76.64	78.23	82.04
DR-GST	68.88	72.82	77.20	79.76	80.00	81.21
+2-*phase*	73.06	74.56	77.76	80.04	81.30	82.45
Image data	**CIFAR100**			**Semi-Aves**		
# Labels	100	200	400	3959		
Confidence	53.70	66.82	75.37	50.88		
+2-*phase*	55.40	67.24	77.33	52.03		
UPS	52.62	66.84	76.02	50.25		
+2-*phase*	54.40	68.43	77.65	52.40		
Modis	54.06	66.99	75.24	50.92		
+2-*phase*	58.40	68.28	76.78	51.08		
SoftMatch	63.54	75.68	81.24	51.88		
+2-*phase*	65.31	77.11	82.45	54.09		
Image data	**Euro-SAT**			**STL-10**		
# Labels	20		40	40		100
Confidence	75.88		85.59	75.63		87.77
+2-*phase*	77.31		88.55	76.73		89.29
UPS	75.93		86.55	76.99		88.30
+2-*phase*	76.29		87.90	78.45		88.50
Modis	76.09		85.24	74.42		87.01
+2-*phase*	76.76		88.74	75.31		89.96
SoftMatch	91.30		92.48	85.26		89.06
+2-*phase*	95.20		95.65	87.65		90.19

5.2. Experiment 2: Complementary Analysis

We analyze the complementarity of two-phase labels with pseudo-labels generated by existing methods, *i.e.,* whether two-phase labels are high-quality pseudo-labels that are currently overlooked. To this end, we compare pseudo-labels generated by 2-*phasic* metric and confidence score in terms of correctness, information gain, and their overlap. The correctness refers to the proportion of correct pseudo-labels among all generated pseudo-labels. We employ the information gain measure proposed in (Pei et al., 2024b), which quantifies the magnitude of changes in model gradients after adding pseudo-labels to training. We adopt Intersection over Union (IoU), also known as Jaccard similarity, to measure

Table 2. Complementary analysis of two-phase labels. We selected the top-k pseudo-labels with the highest confidence and 2-*phasic* metrics to compare their accuracy, information gain, and overlap.

Graph data	**Cora**			**Citeseer**		
L/C	3	5	10	3	5	10
Correctness of pseudo-labels (%)						
Confidence	89.55	91.75	96.40	84.15	82.85	86.10
2-*phasic*	90.50	91.00	96.95	84.65	83.15	87.00
Information gain of pseudo-labels						
Confidence	2.81	3.60	1.78	2.18	2.48	2.76
2-*phasic*	3.78	4.88	2.67	3.10	3.03	3.66
Overlap of the two pseudo-label sets						
IoU	0.61	0.48	0.60	0.60	0.68	0.64

Image data	**CIFAR100**			**Euro-SAT**	
# Labels	100	200	400	20	40
Correctness of pseudo-labels (%)					
Confidence	56.09	69.10	83.47	77.29	88.60
2-*phasic*	61.83	75.69	86.91	80.05	91.31
Information gain of pseudo-labels					
Confidence	252	344	344	356	528
2-*phasic*	614	554	453	426	551
Overlap of the two pseudo-label sets					
IoU	0.16	0.17	0.21	0.41	0.42

Table 3. Classification accuracy of pseudo-labeling learning (%) in ablation study. "c_1 & c_2 terms" denote the complete loss function; "c_2 term" refers to the one with only c_2 term.

Dataset	**Cora**			**CIFAR100**		
L/C | # Labels	3	5	10	100	200	400
Confidence	66.21	71.38	73.73	53.70	66.82	75.37
+ c_2 term	68.96	72.91	77.06	54.33	67.07	76.31
+ c_1 & c_2 terms	70.16	74.07	77.17	55.40	67.24	76.52
c_1 improvement	1.20	1.16	0.11	1.07	0.17	0.21

the overlap of pseudo-labels generated by different methods. We use the model at the last epoch of stage 2 to generate pseudo-labels that were not previously used, and then we select the top-k pseudo-labels with the highest value of the corresponding metric for analysis. The number k is set to 200 for graph datasets and 10,000 for image datasets.

The comparison results are summarized in Table 2, which shows that: (1) The two metrics are at the same level in terms of pseudo-label correctness; (2) Two-phase labels generated by the proposed 2-*phasic* metric are more informative for the model at the last epoch of stage 2; (3) The overlap of the two pseudo-label sets is not high, indicating that two-phase labels are often overlooked by the confidence score.

5.3. Experiment 3: Ablation Study

Here, we evaluate the contribution of the two terms in the specialized loss function introduced in Section 4.3. Following the protocol of experiment 1, we modified the training process in stage 3 by using the loss function with only the c_2 term and with both the c_1 and c_2 terms, respectively. We adopt Cora and CIFAR100 as two representative datasets and use the confidence score as the baseline.

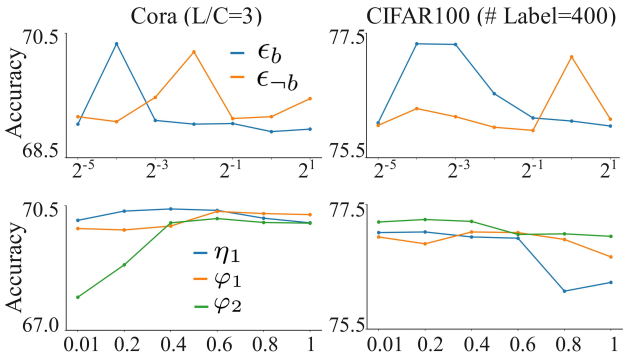

Figure 6. Parameter sensitivity analysis. The y-axis shows the classification accuracy on Cora (**Left**) and CIFAR100 (**Right**); the x-axis denotes the values of parameters in 2-*phasic* metric.

Ablation study results are summarized in Table 3, which shows that: (1) Both the c_1 and c_2 terms contribute to the overall performance, with the c_2 term being the dominant factor; (2) The improvement by the c_1 term is more significant when the number of labeled samples is lower. This indicates that, as analyzed earlier, *eliminating false correlations between simple, non-exclusive patterns and labels is especially important when labeled data is scarce.*

5.4. Experiment 4: Parameter Sensitivity

We analyze the sensitivity of the parameters in the proposed 2-*phasic* metric following the protocol of experiment 1. The analysis involves three weight parameters (η_1, φ_1, and φ_2) and two bias parameters (ϵ_b and $\epsilon_{\neg b}$). Notably, the weight η_2 is dependent on φ_1 and φ_2 in practice and is therefore excluded from our analysis. We adopt Cora ((L/C=3) and CIFAR100 (# Labels = 400) as experimental scenarios and use the confidence score as the baseline. This is a controlled experiment where we adjust only the target parameter and observe the resulting changes in model performance, while keeping all other hyperparameters as fixed values.

As shown in Fig.6, all parameters in 2-*phasic* metric influence model performance, indicating that each component of 2-*phasic* metric is effective and essential. In the figure, none of the curves oscillate up and down, indicating that the parameters can be linearly searched to quickly obtain optimized values. The optimal value of bais ϵ_b precedes that of $\epsilon_{\neg b}$, which aligns with our modeling for Observation 1, as discussed in subsection 4.1.2.

6. Discussion

Calculating the proposed 2-*phasic* metric introduces additional space complexity due to the use of a memory bank to store training dynamics. The additional space complexity is $\mathcal{O}(N|\mathcal{T}|)$, where N denotes the number of unlabeled samples and $|\mathcal{T}|$ is the number of epochs sampled to capture

training dynamics. Our experiments show that a small $|\mathcal{T}|$ is sufficient to effectively identify two-phase samples. For instance, in the experiment of booster test, $|\mathcal{T}|$ is set to 50, meaning the additional space complexity is $\mathcal{O}(N)$

As shown in Fig.6, the optimal values of the parameters vary between the two datasets. Manually tuning these parameters can be time-consuming in practice. In future work, we plan to automatically learn the parameters by adopting the strategy proposed in (Kendall et al., 2018). Future work also includes applying the proposed method to cutting-edge models such as MTGCN (Pei et al., 2024a) on graph data, and extending it to broader learning scenarios, such as active learning and contrastive learning.

7. Conclusion

This work uncovers two-phase labels, a new type of pseudo-labels, which are highly complementary to existing pseudo-labeling methods. These labels are high-quality pseudo-labels that are often overlooked by current methods. Behind this discovery lies an important novel problem — *How can significant information from predicted labels with non-stationary training dynamics be leveraged for pseudo-labeling?* The problem may inspire further research in this underexplored area. We designed the 2-*phasic* metric to identify two-phase labels and proposed a specialized loss function in which the two-phase labels help models both learn correct correlations and eliminate false correlations. Extensive experiments demonstrated that incorporating two-phase labels significantly enhances existing pseudo-labeling methods due to their high information gain and correctness.

Acknowledgements

The authors would like to thank all the anonymous reviewers and chairs for their constructive comments. This work was supported by the National Natural Science Foundation of China under grant 62202369, 62372362, and 62306229.

Impact Statement

This paper may have a broad impact on pseudo-label-based semi-supervised learning methods, as it expands the scope of pseudo-labels to include not only high-confidence, training-stationary predictions but also those exhibiting non-stationary training dynamics. This expansion offers valuable additional information for training, which is particularly significant given the widespread use of pseudo-labeling methods across various algorithms and applications. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, É., and Maximov, Y. Self-training: A survey. *Neurocomputing*, 616:128904, 2025.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242, 2017.

Baldock, R., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*, pp. 10876–10889, 2021.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Eleventh Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.

Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordonez, V. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, pp. 6912–6920, 2021.

Chapelle, O. and Zien, A. Semi-supervised classification by low density separation. In *International Workshop on Artificial Intelligence and Statistics*, pp. 57–64, 2005.

Chen, C., Dong, S., Tian, Y., Cao, K., Liu, L., and Guo, Y. Temporal self-ensembling teacher for semi-supervised object detection. *IEEE Transactions on Multimedia*, pp. 3679–3692, 2021.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

Cortes, C. Support-vector networks. *Machine Learning*, 1995.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pp. 1–77, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

He, H., Aminian, G., Bu, Y., Rodrigues, M., and Tan, V. Y. How does pseudo-labeling affect the generalization error of the semi-supervised gibbs algorithm? In *International Conference on Artificial Intelligence and Statistics*, pp. 8494–8520, 2023.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Hu, Z., Yang, Z., Hu, X., and Nevatia, R. Simple: Similar pseudo label exploitation for semi-supervised classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15099–15108, 2021.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Jia, Q., Li, X., Yu, L., Bian, J., Zhao, P., Li, S., Xiong, H., and Dou, D. Learning from training dynamics: Identifying mislabeled data beyond manually designed features. In *AAAI Conference on Artificial Intelligence*, pp. 8041–8049, 2023.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, 2002.

Kage, P., Rothenberger, J. C., Andreadis, P., and Diochnos, D. I. A review of pseudo-labeling for computer vision. *ArXiv preprint arXiv:2408.07221*, 2024.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. *University of Toronto*, 2012.

Kye, S. M., Choi, K., Byun, H., and Chang, B. Tidal: Learning training dynamics for active learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 22335–22345, 2023.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML 2013 Workshop on Challenges in Representation Learning*, pp. 896, 2013.

Li, M., Wu, R., Liu, H., Yu, J., Yang, X., Han, B., and Liu, T. Instant: semi-supervised learning with instance-dependent thresholds. In *International Conference on Neural Information Processing Systems*, 2023.

Liang, R., Li, T., Li, L., Wang, J., and Zhang, Q. Knowledge consistency between neural networks and beyond. In *International Conference on Learning Representations*, 2020.

Liu, H., Hu, B., Wang, X., Shi, C., Zhang, Z., and Zhou, J. Confidence may cheat: Self-training on graph neural networks under distribution shift. In *ACM Web Conference*, pp. 1248–1258, 2022a.

Liu, J., Qi, Z., Wang, B., Tian, Y., and Shi, Y. Self-llp: Self-supervised learning from label proportions with self-ensemble. *Pattern Recognition*, 129:108767, 2022b.

Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., and Li, W. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19366–19375, 2022c.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. Image-based recommendations on styles and substitutes. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163, 2000.

McClosky, D., Charniak, E., and Johnson, M. Effective self-training for parsing. In *Human Language Technology Conference of the NAACL*, pp. 152–159, 2006.

Namata, G. M., London, B., Getoor, L., and Huang, B. Query-driven active surveying for collective classification. In *Workshop on Mining and Learning with Graphs*, 2012.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

Pan, P., Swaroop, S., Immer, A., Eschenhagen, R., Turner, R., and Khan, M. E. E. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems*, pp. 4453–4464, 2020.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, pp. 20596–20607, 2021.

Pei, H., Li, Y., Deng, H., Hai, J., Wang, P., Ma, J., Tao, J., Xiong, Y., and Guan, X. Multi-track message passing: Tackling oversmoothing and oversquashing in graph learning via preventing heterophily mixing. In *Forty-first International Conference on Machine Learning*, 2024a.

Pei, H., Xiong, Y., Wang, P., Tao, J., Liu, J., Deng, H., Ma, J., and Guan, X. Memory disagreement: A pseudo-labeling measure from training dynamics for semi-supervised graph learning. In *ACM Web Conference*, pp. 434–445, 2024b.

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. Meta pseudo labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021.

Pleiss, G., Zhang, T., Elenberg, E., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, pp. 17044–17056, 2020.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Siddiqui, S. A., Rajkumar, N., Maharaj, T., Krueger, D., and Hooker, S. Metadata archaeology: Unearthing data subsets by leveraging training dynamics. In *The Eleventh International Conference on Learning Representations*, 2023.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pp. 596–608, 2020.

Song, H., Kim, M., and Lee, J.-G. SELFIE: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pp. 5907–5915, 2019.

Sosea, T. and Caragea, C. Leveraging training dynamics and self-training for text classification. In *Findings of the Association for Computational Linguistics*, pp. 4750–4762, 2022.

Su, J.-C. and Maji, S. The semi-supervised inaturalist-aves challenge at fgvc7 workshop. *ArXiv*, abs/2103.06937, 2021.

Sun, K., Lin, Z., and Zhu, Z. Multi-stage self-supervised learning for graph convolutional networks on graphs with few labeled nodes. In *AAAI Conference on Artificial Intelligence*, pp. 5892–5899, 2020.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293, 2020.

Tian, S., Wei, H., Wang, Y., and Feng, L. Crosel: Cross selection of confident pseudo labels for partial-label learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19479–19488, 2024.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11): 2579–2605, 2008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Wang, B., Li, J., Liu, Y., Cheng, J., Rong, Y., Wang, W., and Tsung, F. Deep insights into noisy pseudo labeling on graph data. In *Conference on Neural Information Processing Systems*, 2023.

Wang, H., Huang, W., Wu, Z., Tong, H., Margenot, A. J., and He, J. Deep active learning by leveraging training dynamics. In *Advances in Neural Information Processing Systems*, pp. 25171–25184, 2022a.

Wang, W. and Zhou, Z.-H. Co-training with insufficient views. In *Asian Conference on Machine Learning*, pp. 467–482, 2013.

Wang, Y., Peng, J., and Zhang, Z. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 9092–9101, 2021.

Wang, Y., Chen, H., Fan, Y., Sun, W., Tao, R., Hou, W., Wang, R., Yang, L., Zhou, Z., Guo, L., Qi, H., Wu, Z., Li, Y., Nakamura, S., Ye, W., Savvides, M., Raj, B., Shinozaki, T., Schiele, B., Wang, J., Xie, X., and Zhang, Y. USB: A unified semi-supervised learning benchmark for classification. In *Advances in Neural Information Processing Systems*, 2022b.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Yang, W., Zhang, R., Chen, J., Wang, L., and Kim, J. Prototype-guided pseudo labeling for semi-supervised text classification. In *Annual Meeting of the Association for Computational Linguistics*, pp. 16369–16382, 2023.

Zhang, S., Zhang, H., Lou, S., Wang, Z., Zeng, Z., Wang, Y., and Luo, M. Ptcl: Pseudo-label temporal curriculum learning for label-limited dynamic graph. *arXiv preprint arXiv:2504.17641*, 2025.

Zhao, X., Chen, F., Hu, S., and Cho, J.-H. Uncertainty aware semi-supervised learning on graph data. In *Advances in Neural Information Processing Systems*, pp. 12827–12836, 2020.

Zhou, T., Wang, S., and Bilmes, J. Time-consistent self-supervision for semi-supervised learning. In *International Conference on Machine Learning*, pp. 11523–11533, 2020.

Appendix

This appendix is organized into following six sections, presented in the order as they are referenced in the main paper. Notably, the Code to replicate the experimental results is available in our GitHub repository.

- In Appendix A, we present details for plotting the Fig.1(A) in the main paper.
- In Appendix B, we validate the effectiveness of our proposed LMO-entropy to detect bimodal distribution.
- In Appendix C, we provide details of the attribution method used in Section 4.2 in the main paper.
- In Appendix D, we propose a pseudo-labeling algorithm based on the proposed 2-*phasic* metric.
- In Appendix E, we provide details of the experiments in the main paper, including experimental protocols, baselines, datasets, and evaluation metrics.
- In Appendix F, we present supplementary experiments to further analyze the proposed 2-*phasic* metric.

A. Details for Plotting the Figure 1(A)

We provide details on visualizing the distribution of predicted labels, as shown in Fig.1(A) in the main paper. Specifically, we first fine-tune a pre-trained Visual Transformer (ViT) model (Wang et al., 2022b) on the CIFAR100 dataset using 400 labeled samples, achieving an accuracy of 71.30%, while recording the training dynamics of the ViT.

In the figure, red points represent labels that are incorrectly predicted in the final model output, while blue points indicate correctly predicted labels. Using the recorded training dynamics, we calculate two properties of the predicted labels for each unlabeled sample: the confidence score and non-stationarity, which corresponds to the x-axis and y-axis in the figure, respectively. Here, the confidence score is defined as the average predicted probability of the predicted label's category across epochs by

$$\mu^{(i)} = \frac{1}{T}\sum_{t=1}^{T} P_{\boldsymbol{\theta}_t}\left(y_F^{(i)} \mid \boldsymbol{x}^{(i)}\right),$$

where t specifies a epoch, $\boldsymbol{\theta}_t$ denotes the model parameters at epoch t, and $y_F^{(i)}$ denotes the predicted label for the model after training. The *non-stationary* is defined as the standard deviation of the prediction probability of the predicted labels's category throughout training,

$$\sigma^{(i)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(P_{\boldsymbol{\theta}_t}\left(y_F^{(i)} \mid \boldsymbol{x}^{(i)}\right) - \mu^{(i)}\right)^2}.$$

We use the non-stationary as an inverse indicator of the training stationary, which is easy to calculate.

B. Validation of LMO Entropy

We design this experiment to validate the effectiveness of the proposed Leave-Maximum-Out entropy (LMO entropy for short), a key component for calculating the spatial measure in 2-*phasic* metric. Specifically, we test the ability of LMO entropy to distinguish bimodal distributions (corresponding to two-phase labels) from other types of distribution, such as unimodal and random distribution.

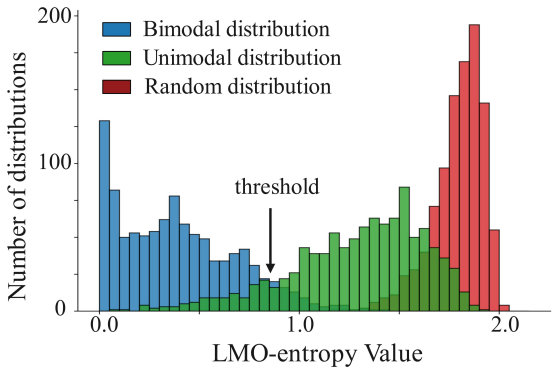

Figure 7. A histogram of LMO entropy values of the generated distribution samples, with distribution types distinguished by colors. The horizontal axis represents the LMO-entropy values, and the vertical axis indicates the number of distribution samples.

To this end, we synthesize three types of distributions: random, unimodal, and bimodal distributions. Each of the distribution is discrete and with ten categories. We generate unimodal distributions using a normal distribution with a random mean and a random variance between 1 and 2. We generate bimodal distributions by applying the softmax function to the combination of two normal distributions, each with random means and random variances between 3 and 5. We generate random distributions by randomly perturbing uniform distributions. For each distribution type, we generate 1,000 distribution samples, and then calculate and analyze the LMO entropy of the the distribution samples.

Fig.7 presents the histogram of LMO entropy values for the generated distribution samples. The results clearly show the LMO entropy values of bimodal distributions are significantly lower than those of other types of distributions. This indicates that a simple threshold of LMO entropy can effectively distinguish bimodal distributions (corresponding to two-phase labels) from other types of distribution.

C. Details of Attribution Method

In Section 4.2 in the main paper, we employed Grad-CAM (Selvaraju et al., 2017) to plot the attribution map Fig.3, *i.e.*, to highlight the key regions in two-phase samples that the first-phase and second-phase models rely on for classification. Grad-CAM is a widely used attribution technique for explaining the attention mechanisms of DNNs.

Grad-CAM begins by computing the gradients of the classification confidence score for a specific category *w.r.t.* all feature activations of each channel in the final convolutional

layer. By applying a global average pooling operation on these gradients, it derives the importance weights of each channel (each feature map) for the given category. These weights for different channels are then used to perform a weighted combination of feature map activations, followed by applying a ReLU activation function, to generate the Grad-CAM attribution map. The resulting attribution map highlights the regions of the image that contribute most significantly to the prediction of the target category.

In experiments, we adopt a ViT model as the base model, which is trained on the CIFAR100 dataset with 400 labeled samples provided. We record the models trained after 250 steps (achieving an accuracy of 66.02%) as the first-phase model, and models trained after 6250 steps (achieving an accuracy of 74.55%) as the second-phase model. We extract two-phase image samples along with their corresponding predicted categories. Then, the Grad-CAM method is employed to visualize the critical image regions for classification of the first-phase and second-phase, respectively.

D. 2-*phasic* based Pseudo-labeling Algorithm

The procedure of the entire pseudo-labeling algorithm is as follows: (1) Base model training: train a base model f_θ on the labeled dataset D_L, and record the training dynamics of the model. (2) Two-phase label generation: Select two-phase samples using the recorded training dynamics and the proposed 2-*phasic* metric. Predict the two-phase labels based on the trained base model. (3) Pseudo-labeling learning: The two-phase labels are used to augment the labeled dataset for further training of the model, utilizing the proposed two-phase pseudo-labeling loss (described in Section 4.3 of the main paper). In the following, we first present the 2-*phasic* based pseudo-label selection algorithm, detailed in Algorithm 1. We then incorporate this selection algorithm into a pseudo-labeling learning framework, as described in Algorithm 2.

In Algorithm 1, the base model f_θ is trained on the labeled data D_L. Throughout the training, we record the softmax distributions $\boldsymbol{\sigma}_t^{(i)}$ of predictions at specific epochs (or batch ID in batch training) in the memory bank, $t \in \mathcal{T}$ (Line 4-8). Subsequently, the recorded training dynamics are used to calculate 2-*phasic* metric for every unlabeled sample using Eq. (1) from the main paper (Line 10-11). Finally, unlabeled samples with a 2-*phasic* value below the threshold τ are selected as pseudo-labeles (Line 13-15). The output pseudo-labels and the corresponding bimodal categories D_P are then used to augment the limited labeled data.

In Algorithm 2, we present 2-*phasic* based pseudo-labeling learning method by incorporating Algorithm 1 into a multi-stage pseudo-labeling learning framework. This algorithm comprises K stages, with $K = 1$ for experiments on graph datasets and $K > 1$ for experiments on image datasets. In each stage, pseudo-labels are generated by Algorithm 1 by leveraging currently available labels in D_L (Line 2-3). Then, the labeled data D_L is augmented by these pseudo-labels, and the augmented labeled data serves as the foundation for the next stage of pseudo-labels generation (Line 4-5). After iterating through the K stages, the final model is trained using the ultimately augmented label set D_L (Line 7). The final model is then used to predict labels for all unlabeled samples (Line 8-10). Notably, we add an equal number of pseudo-labels for each category in every stage so as to avoid the issue of label imbalance.

Algorithm 1 2-*phasic* based Pseudo-label Selection

Input: Labeled samples D_L, unlabeled samples D_U, and initial model f_θ
Output: Pseudo-label samples D_P
Parameter: Threshold τ, epoch indexes $\mathcal{T}$

1: Initialize model parameters θ, and $D_P \leftarrow \varnothing$
2: **for** training epoch $t = 1$ to max_epoch **do**
3: Update θ using the gradient calculated on D_L
4: **if** epoch $t \in \mathcal{T}$ **then**
5: **for** each sample $\boldsymbol{x}^{(i)}$ in D_U **do**
6: Compute $\boldsymbol{\sigma}_t^{(i)}$ via f_θ and record
7: **end for**
8: **end if**
9: **end for**
10: **for** each sample $\boldsymbol{x}^{(i)}$ in D_U **do**
11: Calculate its 2-*phasic* metric by using Eq. (1) in the main paper
12: Record its bimodal categories $\boldsymbol{bc}^{(i)} = [c_1, c_2]$
13: **if** 2-*phasic*$^{(i)} < \tau$ **then**
14: $\tilde{\boldsymbol{y}}^{(i)} = f_\theta(\boldsymbol{x}^{(i)})$; $D_P \leftarrow D_P \cup \{\boldsymbol{x}^{(i)}, \tilde{\boldsymbol{y}}^{(i)}, \boldsymbol{bc}^{(i)}\}$
15: **end if**
16: **end for**
17: **Return** D_P

Algorithm 2 2-*phasic* based Pseudo-labeling Learning

Input: Labeled samples D_L, unlabeled samples D_U, and initial model f_θ
Output: Predicted labels $D^* = \{(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}}^{(i)}) | \boldsymbol{x}^{(i)} \in D_U\}$

1: $D^* \leftarrow \varnothing$
2: **for** each stage k = 1 to K **do**
3: D_P = Algorithm 1(D_L,D_U,f_θ)
4: $D_L \leftarrow D_L \cup D_P$; $D^* \leftarrow D^* \cup D_P$
5: Update D_U according to D_L
6: **end for**
7: Train model f_θ on augmented D_L
8: **for** each sample $\boldsymbol{x}^{(i)} \in D_U$ **do**
9: $\tilde{\boldsymbol{y}}^{(i)} = f_\theta(\boldsymbol{x}^{(i)})$; $D^* \leftarrow D^* \cup \tilde{\boldsymbol{y}}^{(i)}$
10: **end for**
11: **Return** D^*

E. Details of Experiments

E.1. Details of Baselines

We choose the confidence score and MoDis as baselines in experiments on both graph and image datasets.

- Confidence score(Lee et al., 2013): The likelihood or probability assigned to the predicted category, which is a widely used metric for selecting pseudo-labels.
- Memory Disagreement (MoDis) (Pei et al., 2024b): By leveraging training dynamics to quantify the prediction uncertainty, which favors unlabeled samples with higher prediction consistency during training.

In experiments on image datasets, we additionally choose the following two baselines:

- Uncertainty-aware Pseudo-label Selection (UPS) (Wang et al., 2021): UPS utilizes prediction uncertainty to reduce noise from poorly calibrated models, there mitigating overconfidence in pseudo-labeling.
- SoftMatch (Chen et al., 2023): SoftMatch proposes a uniform alignment approach that weights samples based on their confidence levels, thereby enhancing the utilization of weak learners.

In experiments on graph datasets, we additionally choose the two following two baselines:

- Area Under the Margin (AUM) (Sosea & Caragea, 2022): AUM captures the divergence between the logit of annotated labels and predicted labels in training.
- Distribution Recovered Graph Self-Training (DR-GST) (Liu et al., 2022a): DR-GST introduces a distributional correction mechanism into confidence based self-training on graphs.

E.2. Details of Datasets

Our experiments are conducted on eight benchmark datasets, including four graph datasets, specifically Cora (McCallum et al., 2000), Citeseer (Sen et al., 2008), PubMed (Namata et al., 2012), and AmazonComputer (McAuley et al., 2015), and four image datasets, specifically CIFAR100 (Krizhevsky, 2012), EuroSAT (Helber et al., 2019), STL-10 (Coates et al., 2011), and Semi-Aves(Su & Maji, 2021). Statistics of image and graph datasets are summarized in Table 4 and Tables 5, respectively.

In experiments on graph datasets, we follow the setup in (Liu et al., 2022a; Pei et al., 2024b), use only 3, 5, and 10 labeled nodes per class (L/C), which presents challenging scenarios with minimal label information. In experiments on image datasets, we use different configurations, as shown in Table 5, according to existing works (Wang et al., 2022b).

In experiments, we employ Optuna (Akiba et al., 2019) to searche parameters in 2-*phasic* metric, with the parameter search ranges detailed in the Table 6.

E.3. The Calculation of Information Gain

In Table 2 of the main paper, we employ information gain to evaluate the quality of pseudo-labels. Specifically, we measure the information gain from pseudo-labels by summarizing the positive contribution of correct pseudo-labels and the adversarial impact of incorrect pseudo-labels,

$$\rho = \rho^{(+)} - \rho^{(-)},$$

where $\rho^{(+)}$ and $\rho^{(-)}$ denote the positive contribution and the adversarial impact, respectively. Taking positive contribution $\rho^{(+)}$ as an example, the information gain from the inclusion of correct pseudo-labels is calculated based on model perturbations. Specifically, we quantify the gradient changes induced by the inclusion of correct pseudo-labels:

$$\rho^{(+)} = \left\| \nabla \ell \left(D_L \cup D_P^{(+)}; f_\theta \right) \right\|_F - \left\| \nabla \ell \left(D_L; f_\theta \right) \right\|_F ,$$

where $\nabla \ell(\cdot)$ denotes the model's gradient, D_L represents the labeled samples, and $D_P^{(+)}$ is the set of correct pseudo-labels. Here $\|\cdot\|_F$ signifies the Frobenius norm. To calculate adversarial impact $\rho^{(-)}$, we replace the set $D_P^{(+)}$ with $D_P^{(-)}$, which denotes the set of incorrect pseudo-labels.

In Fig.1(B), we compare the information gain of Type 1 and Type 2. The information gain is caculated based on samples within each type and divided by the mean information gain of all samples. In the complementary analysis in 5.2, we compute the average information gain per sample by dividing the information gain of the pseudo-label set by the number of samples in the set.

E.4. Details of Booster Test

The booster experiment (Section 5.1 in the main paper) is structured into three stages, as illustrated in Fig.5 in the main paper. We conducted the experiment on both graph and image datasets. For image classification, we selected CIFAR100, EuroSAT, STL-10, and Semi-Aves datasets, with detailed descriptions in Appendix E.1. We utilized the pre-trained ViT model provided in the (Wang et al., 2022b) as the backbone. In the first phase, we directly loaded the pre-trained checkpoint of the ViT model. During the second phase, in accordance with the settings from the USB, we set the batch size to 8 and employed a baseline algorithm to select pseudo-labels for training while recording the training dynamics within batches. A total of 4 epochs were conducted in the stage 2. In stage 3, the experiment was divided into two trials. We use pseudo-labels selected by the 2-*phasic* metric and the baseline to train the model, respectively, and report the final experimental results of both trials. In the baseline trial, the model continued to use the baseline algorithm to select pseudo-labels for model training. In the two-phase trial, we leveraged the training dynamics obtained from stage 2 to select two-phase pseudo-labels using 2-*phasic* according to Algorithm 1.

Table 4. Statistics of graph datasets.

Graph dataset	# Nodes	# Edges	# Categories	# Features	# Labels used for training pre category
Cora	2,708	5,429	7	1,433	3 / 5 / 10
Citeseer	3,327	4,732	6	3,703	3 / 5 / 10
PubMed	19,717	44,338	3	500	3 / 5 / 10
AmazonComputer	13,752	245,778	10	767	3 / 5 / 10

Table 5. Statistics of image datasets.

Image dataset	# Images in training set	# Images in test set	# Categories	# Labels used for training pre category
CIFAR-100	50,000	10,000	100	1 / 2 / 4
STL-10	5,000	8,000	10	4 / 10
EuroSat	16,200	5,400	10	2 / 4
Semi-Aves	3,959	4,000	200	15

Table 6. Parameter search setup. We use Optuna (Akiba et al., 2019) to search for parameters within the following ranges.

Parameters	Range for graphs	Range for images
ϵ_b	(0, 0.1)	(0, 0.1)
$\epsilon_{\neg b}$	(0.1, 2)	(0.1, 2)
η_1	(0, 1)	(0, 1)
η_2	depends on φ_1 and φ_2	
φ_1	(0, 1)	(0, 1)
φ_2	(0, 1)	(0, 1)
lr	5e-3 for Amazon (1e-5, 1e-3) for others	(1e-4, 1e-3)
τ	(0.8, 1.8)	Mean of *2-phasic* of all the unlabeled

Table 7. Classification accuracy of pseudo-labeling learning (%).

Image Data	**CIFAR100**	
# Labels	200	400
Confidence	66.84	75.38
+2-*phase*	68.24	77.98

According to the experimental results in (Wang et al., 2022b), the baseline models are convergent around the tenth epoch. To speed up our comparison experiments, we did not train for 200 epochs as the seeting in (Wang et al., 2022b); Instead, we run 6 epochs in stage 3, during which models in both trials successfully converged. To prove the convergence, we train model for 200 epochs using one baseline. The results are reported in Table 7, where the perfoermance is very closed to the results in Table 1 in the main paper.

E.5. Details of Ablation Study

We conduct an ablation study to evaluate the contribution of each component in the loss functions in Section 4.3 of the main paper, *i.e.,* c_1 and c_2 terms. The c_2 loss is defined as:

$$\mathcal{L}\left(\tilde{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)}\right) = -\tilde{y}_{c_2}^{(i)} \log\left(\hat{y}_{c_2}^{(i)}\right),$$

where $\tilde{\boldsymbol{y}}^{(i)}$ denotes the two-phase pseudo-Label, $\hat{\boldsymbol{y}}^{(i)} = f_\theta(\boldsymbol{x}^{(i)})$ is the model's prediction probability. The loss with both c_1 and c_2 terms are the same as the loss functions in the main paper. For graph data, we select the GCN model trained on the Cora dataset, as the base model. Our experiments are performed on three partitions of Cora, specifically, L/C = 3, 5, and 10. For image data, the base model is a ViT trained on the CIFAR100 dataset. Experiments are conducted with 100, 200, and 400 pseudo-labels.

Table 8. The node classification accuracy of pseudo-labeling algorithms on graphs (%) with the backbone GAT. L/C represents the number of labeled nodes per category.

Graph Data	**Cora**			**Citeseer**		
L/C	3	5	10	3	5	10
Confidence	67.04	75.12	75.37	61.16	65.33	67.86
+2-*phase*	70.72	77.00	78.53	62.56	67.48	68.63
AUM	68.99	73.14	76.00	59.38	63.45	66.83
+2-*phase*	69.70	77.21	79.44	59.70	66.60	68.65
MoDis	72.67	74.21	75.13	63.09	66.98	67.27
+2-*phase*	74.42	76.06	77.22	64.10	68.37	68.66
DR-GST	73.58	79.42	82.35	52.47	68.13	72.91
+2-*phase*	74.28	80.13	83.01	56.86	70.84	73.39
Graph Data	**Pubmed**			**AmazonCS**		
L/C	3	5	10	3	5	10
Confidence	63.06	64.46	68.60	75.66	77.09	80.81
+2-*phase*	63.43	70.08	72.16	75.75	78.63	81.56
AUM	60.48	65.27	68.01	74.52	77.34	80.29
+2-*phase*	63.92	69.55	72.16	75.20	78.19	80.73
MoDis	65.37	69.22	71.80	74.70	77.18	81.43
+2-*phase*	66.00	69.67	73.03	77.06	78.20	82.38
DR-GST	66.51	73.93	76.48	78.96	80.56	82.66
+2-*phase*	67.89	74.26	76.84	81.54	81.78	83.02

F. Supplementary Experiments

F.1. Additional Booster Test

We also use the widely used graph learning model, Graph Attention Networks (GAT) (Veličković et al., 2018) as an additional base model for booster test. All experimental configurations are the same as the experiment of booster test, described in the main paper and Appendix E.4.

As shown in Table 8, the two-phase labels generated by the 2-*phasic* metric consistently improve classification accuracy across four graph datasets and four baselines, with an average enhancement of 1.76%. These results further demonstrate that adding two-phase labels boosts the performance of existing pseudo-labeling methods, regardless of the base model used, highlighting the broad applicability of our proposed 2-*phasic* metric.

Table 9. Classification accuracy of pseudo-labeling learning (%) in component ablation study. "Temporal only" denotes that only $\mu_{temporal}$ is used in the metric and "Spatial only" denotes that only $\mu_{spatial}$ is used. "2-*phasic*" represents the complete metric.

Dataset	**Cora**		
L/C	3	5	10
Temporal only	69.76	73.40	75.99
Spatial only	68.24	72.70	73.87
2-*phasic*	70.16	74.07	77.17

F.2. Ablation Study on 2-*phasic* Metric

To further analyze the contributions of the spatial measure and the temporal measure in the 2-*phasic* metric, we conducted an ablation study. Specifically, we use GCN as the base model on Cora dataset with L/C=3, 5, and 10. In the booster test experiments, we employ confidence as pseudo-labeling metric in stage 2, and use either $\mu_{temporal}$ or $\mu_{spatial}$ alone as pseudo-labeling metric in stage 3. The experimental results are summarized in Table 9.

The results demonstrate that both temporal and spatial components contribute positively to the overall performance, with the temporal features having a more significant impact.