scientific reports

Check for updates

OPEN

Uncertainty-weighted semi-supervised learning with dynamic entropy masking and Bhattacharyya-regularized loss

Mohammed Talal Ghazal[1,2], Jafar Tanha[1]✉, Nasrin Shahi[1] & SeyedEhsan Roshan[1]

Semi-supervised learning (SSL) leverages labeled and unlabeled data for modern classification tasks. However, existing SSL approaches often underutilize moderately uncertain samples and may propagate errors from highly uncertain pseudo-labels, leading to suboptimal performance, in noisy and class-imbalanced datasets. We introduce an SSL framework with an uncertainty-weighted training mechanism that prioritizes moderately uncertain samples while deferring extremely uncertain samples via a dynamic entropy mask. Training on unlabeled data combines masked cross-entropy with a Bhattacharyya-regularized alignment term between weak and strong predictions, improving view consistency and distribution alignment. A dynamic entropy threshold (ϵ_t) that adapts over training, filtering only extremely uncertain pseudo-labels and thereby limiting error propagation while retaining informative unlabeled data. The proposed framework is evaluated on several benchmark datasets, including CIFAR-10, SVHN and STL-10 under label-scarce and class-imbalanced protocols, achieving up to 3–5% absolute accuracy gains over strong SSL baselines (e.g., FixMatch, ReMixMatch, FreeMatch). Our results show that the proposed approach improves model generalization and robustness, particularly in scenarios involving label noise, class imbalance, and limited labeled data, while remaining comparable on clean, class-balanced settings.

Keywords Semi-supervised learning, Uncertainty-weighted training, Bhattacharyya-Regularized divergence, Dynamic masking, Pseudo-labels

Semisupervised learning (SSL) has become an important research focus in modern machine learning, and provides a sound paradigm for situations where a large amount of unlabeled data is available while few labeled data exist[1]. SSL techniques are especially beneficial in a variety of domains including computer vision, natural language processing and medical imaging, where data annotation is expensive, time-consuming or even impossible[2]. The basic principle of SSL is to use labeled and unlabeled instances together to increase a model's predictive power and generalization ability[3]. Two dominant families in SSL are consistency regularization and pseudo-labeling[4]. Consistency regularization assumes that model predictions under different input perturbations should remain consistent, while pseudo-labeling uses the model's predictions on unlabeled samples as targets for training. Pseudo-labeling remains widely used due to its simplicity but is sensitive to confirmation bias[5].

Existing pseudo-labeling approaches face fundamental challenges, in particular with respect to the processing of uncertain samples[6]. In the current research work, we define two kinds of uncertain samples, such as the moderate uncertain samples which are informative but not confidently classified, and the highly uncertain samples which have high entropy and unreliable pseudo-labels. Most pseudo-labeling methods employ a fixed confidence threshold to filter out uncertain instances. While this stabilizes training, it may discard potentially valuable data from the moderately uncertain regime[7]. The exclusion of these instances leads to suboptimal generalization and can induce confirmation bias toward majority and easy classes[8]. Some leading semi-supervised learning techniques, such as FixMatch, MixMatch, and ReMixMatch, are based on probability thresholds for unlabeled data pseudo-label selection[9,10]. Nonetheless, there remain challenges in determining reliable pseudo-labels whilst retaining informative data, and such approaches may be prone to confirmation bias towards some classes, if confidence estimation is not accurate[11].

[1]Department of Electrical & Computer Engineering, University of Tabriz, Tabriz, Iran. [2]Department of Medical Instrumentation Technology, Technical Engineering College, Northern Technical University, Mosul, Iraq. ✉email: tanha@tabrizu.ac.ir

We present an SSL framework based on uncertainty-weighted training and dynamic entropy masking which focuses attention on moderately uncertain samples and defers highly uncertain samples. Instead of rejecting the uncertain samples directly, our method corrects their contribution by a confidence score estimated from the entropy of the distribution of the normalized activation values, making use of the easy and moderately uncertain samples. This dynamic adaptation mechanism enables the model to learn from both trusted and non-trusted information, ensuring that valuable information is not missed. Furthermore, by improving the robustness of the model to noise and label uncertainty, our method also improves generalization on a variety of datasets - even when there is only a small amount of labeled data.

A key component of our approach is a hybrid unlabeled objective that combines masked cross-entropy with a Bhattacharyya-regularized divergence between weak and strong predictions for the same unlabeled sample. This symmetric divergence stabilizes view alignment and mitigates overconfident errors without requiring a fixed alignment to the labeled distribution. Such alignment is especially helpful when labeled and unlabeled distributions are mismatched (e.g., long-tailed class distributions). In addition, we introduce a dynamic entropy masking strategy to address unreliable pseudo-labels. Rather than a fixed cutoff, the entropy threshold (ϵ_t) evolves during training, so that only extremely uncertain pseudo-labels are filtered while informative samples are retained. This dynamic gating limits error propagation and improves stability.

The contributions of this work are threefold:

(1) An uncertainty-weighted training mechanism that links entropy-based uncertainty to loss weighting, emphasizing moderately uncertain samples while down-weighting easy ones and deferring extremely uncertain samples.
(2) A hybrid unlabeled loss that combines cross-entropy with Bhattacharyya-regularized divergence, enhancing distribution alignment and improving robustness in noisy and imbalanced settings.
(3) A dynamic entropy masking strategy (ϵ_t) that evolves over training to exclude only extremely uncertain pseudo-labels, reducing confirmation bias and stabilizing learning.

To demonstrate the main difference between our approach and the traditional ways of implementing the method of SSL, Fig. 1 compares the fixed-threshold pseudo-labeling with our uncertainty-aware pipeline. Whereas the previous approaches can ignore or severely down-weight uncertain examples (e.g., FixMatch[9] and Mean Teacher[10] our framework leaves only highly uncertain examples and rewards more moderately uncertain ones through weighting and masking transition. This preserves useful information with no increase of noise which we demonstrate translates to gains in difficult protocols.

The rest of this paper is organized in the following way. Section 2 will be a review of the literature on the topic of SSL, with particular attention to the issues of dealing with hard-to-label data and the approaches that have been suggested to resolve these problems. Section 3 explains the methodology of our proposed framework. Section 4 gives the experimental setup and results and a discussion of the findings is given in Sect. 5. Lastly, we conclude the paper in Sect. 6 and provide possible future research directions.

Related works

Semi-supervised learning SSL has also developed, and numerous approaches use large sets of unlabeled data to improve generalization in the case of limited labeled data[5,7]. Previous SSL tended to assume that the latent distributions of labeled and unlabeled samples were identical, but in reality, this fails in the presence of noise, class imbalance, and ambiguous labeling, which worsens performance and reliability[9,11]. In this landscape, two pillars, pseudo-labeling and consistency regularization, describe much of the progress of SSL and its continuing limitations when uncertain data are not treated with care[4,6].

Foundations and evolving challenges

In the standard SSL pipeline, a model is first trained on labeled data and then used to assign pseudo-labels to unlabeled samples; these predicted targets are treated as additional supervision during subsequent training rounds[12,13]. Although the mechanism is simple and commonly used, it is also delicate: as soon as false pseudo-labels enter, they are likely to be strengthened with each iteration, a phenomenon, often referred to as the confirmation bias, is said to be prone to[14]. Confidence thresholds were introduced to curb this effect, and FixMatch operationalized the idea by coupling high-confidence selection with consistency between weakly and strongly augmented views of the same input[9,10]. Self-training is a natural extension of this paradigm, adding high-confidence pseudo-labeled data and training larger and larger sets of data to ensure coverage[15], and Noisy Student makes use of a teacher-student ensemble to stabilize targets and reduce overfitting[16]. Despite these refinements, fixed acceptance rules can simultaneously discard moderately uncertain (yet informative) samples and admit extremely uncertain (unreliable) ones, effects that are amplified under label noise and class imbalance[5,11].

Consistency regularization provides a complementary principle: model predictions should remain stable under input perturbations[17]. Representative approaches include Mean Teacher, nudging a student toward an exponential-moving-average teacher, and Virtual Adversarial Training (VAT), which regularizes against worst-case local perturbations[10,18]. MixMatch later combined weak and strong augmentations and target interpolation to produce smoother decision boundaries and stronger invariance[19]. However, purely consistency-driven training does not by itself decide which uncertain samples to use or how strongly to weight them. When many unlabeled points are hard to classify, training can still privilege easy and majority cases while sidelining data that would most benefit minority classes.

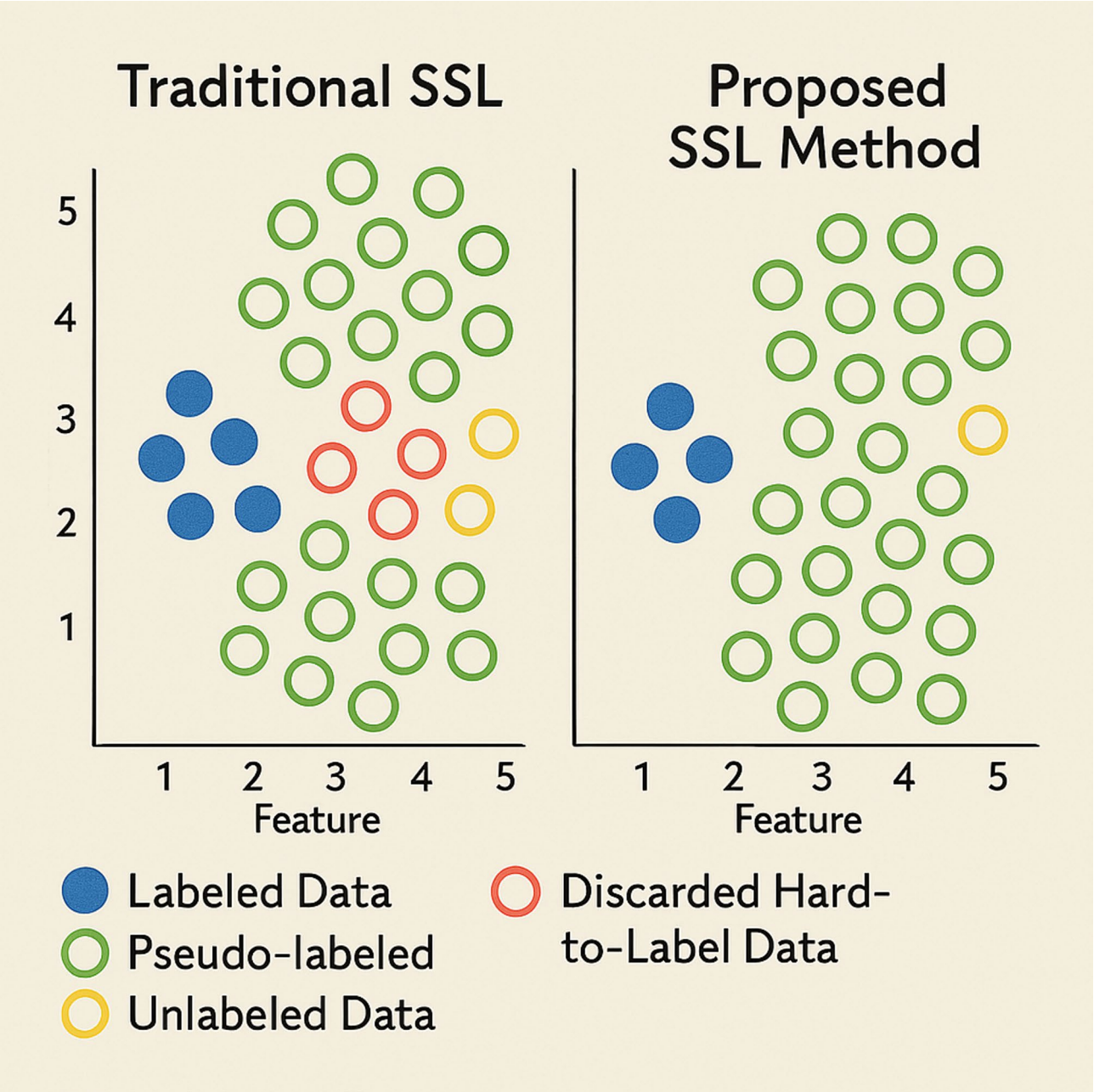

Fig. 1. Comparison of uncertain data (Hard-to-Label) utilization in SSL methods.

Uncertainty, Imbalance, and alignment under mismatch

To address uncertain-sample selection, a substantial line of work estimates uncertainty explicitly and uses it to guide data usage[20,21]. Uncertainty-guided cross-teaching and curriculum methods (e.g., UTCS) gradually expose the learner to more difficult examples as confidence improves[22], while ensemble self-training (e.g., UGE-ST) aggregates multiple predictors to stabilize pseudo-labels and reduce variance[23]. These strategies often increase robustness, but many rely on static heuristics that may be miscalibrated as the model evolves, over-filtering or over-admitting uncertain data at different stages. This observation motivated adaptive schemes in which thresholds or weights vary with training, concentrating learning where uncertainty is informative and deferring samples whose entropy indicates unreliability. In parallel, practical baselines broaden the pseudo-labeling toolbox: FreeMatch introduced self-adaptive thresholding that synchronizes confidence gates with the model's learning status[24] and AdaMatch unified SSL with domain adaptation by aligning distributions while retaining confidence-based selection[25]. Related ensemble low-label lines in few-shot classification which combining various training and adaptation algorithms for ensemble few-shot classification[26] and prototype-neighbor network with task-specific enhanced Meta-learning for few-shot classification[27], underscore the value of calibrated targets and neighborhood-aware consistency when supervision is scarce, a principle we transfer to SSL despite the setting differences.

Distribution alignment forms a third pillar of modern SSL. ReMixMatch and Noisy Student demonstrated that stronger augmentation, consistency, and pseudo-label interpolation can shrink gaps between labeled and unlabeled distributions[16,28], but alignment alone remains brittle under long-tailed class imbalance. Accordingly, imbalance-aware SSL explicitly counteracts majority dominance: DARP refines pseudo-labels via distribution-aware optimization to correct class bias, CReST rebalances self-training schedules to amplify trustworthy minority signals, and DASO blends semantic and linear pseudo-labels while introducing a semantics-oriented alignment loss[29–31]. Complementary to these, a semi-supervised resampling method for class-imbalanced learning[32] directly rebalances exposure, offering another route to counter prior skew during SSL. These approaches improve minority recall by correcting class priors or reshaping exposure, yet most do not simultaneously couple (i) an asymmetric agreement objective across unlabeled weak and strong views with (ii) a dynamic selection rule, an interaction that becomes crucial when selection and alignment influence one another under mismatch[5,7,9]. Unified Consistency Regularization (UCR) moves toward an integrated treatment by adjusting learning dynamics in response to distributional cues[33], and multi-model designs such as DUMM exploit both sample- and pixel-level uncertainties to focus attention where it matters most[34].

Beyond selection and exposure, the choice of divergence for enforcing agreement between weak and strong predictions also shapes optimization stability and calibration under overlap and imbalance. KL divergence is asymmetric and can produce sharp gradients when the teacher is overconfident; Jensen-Shannon is symmetric but can still exhibit peaky gradients in practice; approximate Wasserstein adds geometric fidelity but increases computational and critic-tuning burden. By contrast, symmetric and overlap-aware criteria (such as Bhattacharyya-based regularization) yield smoother, bounded gradients where class distributions mingle, providing a more stable signal in noisy or long-tailed regimes. Empirically, divergence-swap ablations that replace the symmetric term with KL/JS/Wasserstein help quantify accuracy and calibration (Expected Calibration Error, ECE; Negative Log-Likelihood, NLL) trade-offs and clarify when symmetry substantively improves training stability. Taken together, prior work suggests that the most reliable trajectories recognize uncertainty as a graded signal and adapt thresholds and weights over time, while employing symmetric alignment to stabilize learning when labeled an unlabeled data differ markedly in class frequency or noise level[5,7,9–11,13–16,18–31]. Within this view, dynamic entropy masking paired with uncertainty-weighted training and symmetric weak and strong alignment offers a unified response to selection and alignment under distribution mismatch.

Methodology

The current paper proposes a robust semi-supervised learning (SSL) framework that utilizes both labeled and unlabeled data simultaneously, allowing the use of data that are difficult to label, which is often ignored by standard SSL algorithms.

This method combines an uncertainty-weighted training method which dynamically modulates the bias of the model on the uncertain data points so that both the easy and moderately uncertain data are used in the learning process of the model. The proposed framework consists of the following steps: model design, hybrid loss, dynamic entropy masking and entropy-based weighting. During training at the first stage, the model is trained using a rather small amount of labeled data; the results of the trained model on the unlabeled data serve as pseudo-labels. These pseudo-labels are then further refined with an iterative process where easy and hard-to-label instances are included in the process of learning in the model. In comparison to the traditional SSL, where low-confidence pseudo-labels are dropped[7], the suggested framework makes sure that uncertainty is quantified and used gradually (via τ for confidence and a dynamic entropy threshold ϵ_t), without the loss of valuable information.

The suggested framework utilizes a Vision Transformer (ViT)[35], which is unique due to the ability to detect long-range relationships in image data. ViT can break input images into patches and process them sequentially which yield better outcomes on tasks with strong structural and spatial variation[36]. This perfectly fits the case of the SSL in Fig. 2. Each patch is processed by the self-attention mechanism to produce representations that encode global interactions among patches. These representations illustrate long-range relations, which makes the model useful in cases of the application of SSL.

A hybrid loss function, which is a combination of supervised cross-entropy and the Bhattacharyya-regularized divergence on unlabeled data leads to learning in the model. The Bhattacharyya divergence[37] promotes conformity to weakly and strongly augmented views and thus makes class separability and resistance stronger, particularly in the presence of noise and imbalance. Bhattacharyya, in comparison with KL, Jensen-Shannon and Wasserstein substitutes, is symmetric and has an overlap sensitivity, which generates smoother, bounded gradients when predictions are partially inconsistent, which makes it especially convenient in the context of noisy or small-sample data and in the context of training with these requirements. The proposed method also uses a dual-augmentation model in which the trained model makes ensemble predictions of unlabeled inputs with weak and strong data augmentations. Weak augmentation yields less hostile pseudo-labels whereas strong augmentation is used to enforce consistency across augmented views. Collectively, the operations will decrease the chances of misclassification, enhance decision constraints, and augment pseudo-label dependability via the uncertainty-sensitive masking and weighting strategy.

Overview of the proposed framework

The key idea in the proposed architecture is an adaptive mechanism that guides the network to exploit uncertain, hard-to-label instances. The model calculates the entropy of its predictions, and through an uncertainty-weighted training mechanism prioritizes moderately uncertain samples rather than discarding them. This ensures efficient learning, particularly in noisy or ambiguous data. The framework applies weak and strong augmentations and enforces consistency between their predictions. The predictions of the network are then improved through entropy-guided weighting and dynamic masking, where uncertainty (entropy) and confidence levels are used to

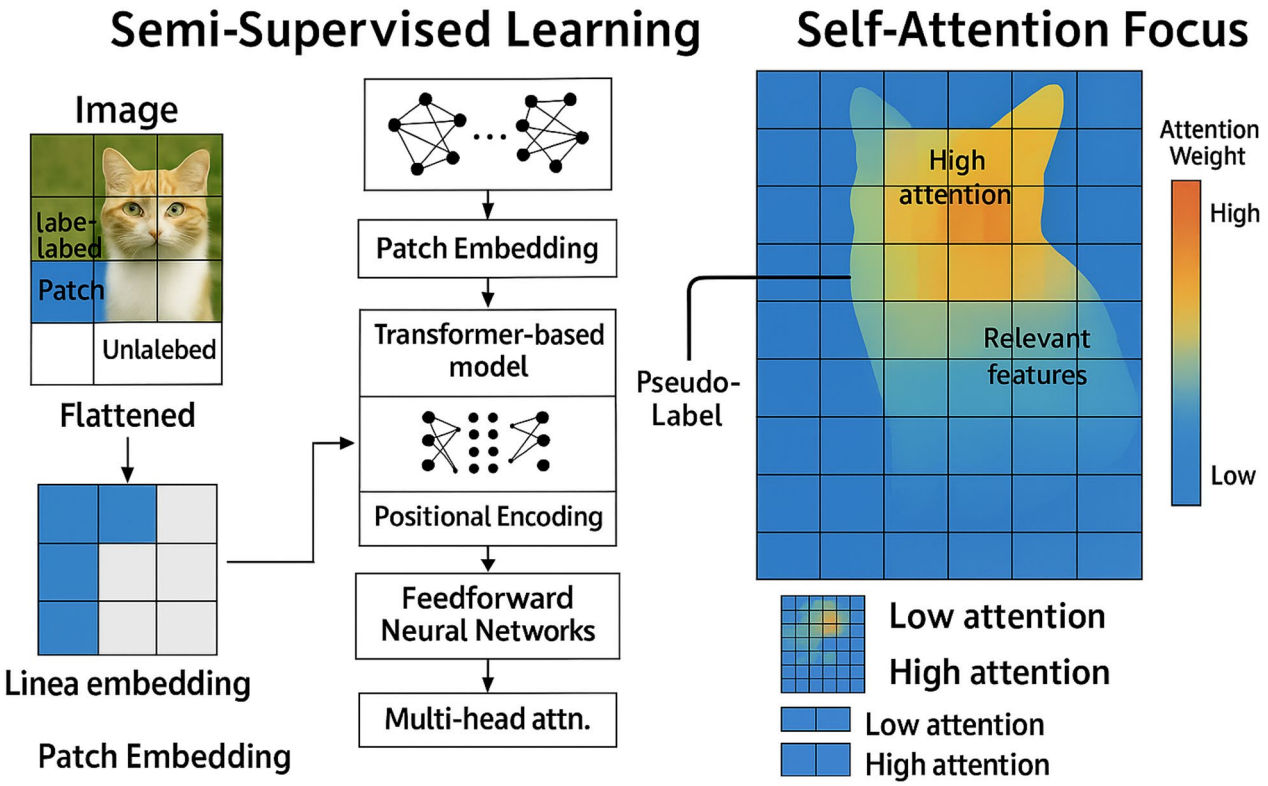

Fig. 2. ViT architecture and self-attention mechanism in SSL.

adjust the weight of each sample. The hybrid loss function makes model outputs coincide with the correct data distribution, whereas the dynamic masking strategy ensure only trustworthy pseudo-labels are used to train the model (see Eqs. (2–5), while extremely uncertain samples are deferred via an entropy threshold ϵ introduced in (Eq. 7).

Model components

The proposed framework adopts the Vision Transformer (ViT) architecture, a method that efficiently processes large-scale image data by detecting both local and global relationships through patch-based processing. ViT is especially applicable in semi-supervised learning cases, in which labeled and unlabeled data are present. For the labeled data, the model is trained on them and the supervised loss is calculated, while for the unlabeled data, the unsupervised loss is processed through pseudo-labeling and consistency regularization.

The supervised loss function L_s is focused on utilizing labeled data for accurate classification. It is defined as the cross-entropy loss between the true labels and predicted labels for the labeled data, which guides the model to learn correct class assignments. The loss function is defined as:

$$L_s = \frac{1}{n} \sum {}_{i=1}^{n} H\left(P_b . P_m\left(Y \mid A_w\left(x_i^l\right); \lambda_s\right)\right) \tag{1}$$

Where P_b is the true (one-hot) label distribution, $P_m\left(Y \mid A_w\left(x_i^l\right); \lambda_s\right)$ is the model's predicated distribution for weakly augmented labeled data x_i^l. While $H(p, q)$ denotes the cross-entropy between distributions p and q, and λ_s is the supervised weight.

The unsupervised loss is calculated using the unlabeled data, which are augmented through weak and strong transformations. This loss is based on two components: Cross-entropy loss and the Bhattacharyya divergence. For unlabeled data, the cross-entropy loss L_{ce} between the predicted probability distribution and the pseudo-label is computed as follows:

$$L_{ce} = \frac{1}{\mu m} \sum {}_{i=1}^{\mu m} \mathbb{1}\left(\max\left(q_b\right) \geq \tau\right) H\left(\widehat{q}_b, P_m\left(Y \mid A_s\left(x_i^u\right)\right)\right) \tag{2}$$

Here, q_b is the weak-view prediction (pseudo-label) generated from $A_w\left(x_i^u\right)$, $\widehat{q}_b$ is its temperature-sharpened form, τ is a confidence threshold, $A_s\left(x_i^u\right)$ is the strong augmentation, μ is the unlabeled-to-labeled ratio per iteration, and m is the labeled batch size. Also, we include τ as confidence threshold and $P_m\left(Y \mid A_s\left(x_i^u\right)\right)$

is for model's prediction for strongly augmented unlabeled data. The second component of the unsupervised loss measures the divergence between the weakly and strongly augmented versions of the same data. The Bhattacharyya divergence D_{bh} ensures that predictions from different augmentation are consistent. The Bhattacharyya divergence loss L_d is computed as:

$$L_d = \frac{1}{\mu\, m} \sum_{i=1}^{\mu\, m} \ll (\max(q_b) > \tau)\ D_{bh}\left(P_s^w \,|\, P_m\left(Y \mid A_s\left(x_i^u\right)\right)\right) \tag{3}$$

While the Bhattacharyya divergence is calculated as:

$$D_{bh} = -\ln\left(\sum_{i \in I} \sqrt{P_s^w * P_m\ (Y \mid A_s\left(x_i^u\right)\ )}\right) \tag{4}$$

And the sharpened prediction for weakly augmented data P_s^w is given by:

$$P_s^w\left(Y \mid A_w\left(x_i^u\right)\right) = \exp\left(\frac{g_b}{T}\right) / \sum_k \exp\left(\frac{g_k}{T}\right) \tag{5}$$

with g_b the class logit corresponding to the predicted class and T a temperature parameter that mitigates overconfidence. Finally, the total loss function L_{total} is the combination of the supervised L_s and unsupervised loss $(L_{ce} + L_d)$ components. The final loss function is given by:

$$L_{total} = \lambda_s L_s + \lambda_u (L_{ce} + L_d) \tag{6}$$

Where λ_s and λ_u are the supervised and unsupervised weights.

Uncertainty weighting and dynamic masking in pseudo-labeling

A central feature of the proposed framework is the combination of uncertainty weighting and dynamic entropy masking, which together ensure that only reliable pseudo-labels contribute to training while moderately uncertain samples are still exploited. In the setting of SSL, low-confidence pseudo-labels may introduce significant noise and cause error propagation. To mitigate this risk, the model estimates the uncertainty of each prediction using entropy. Instead of discarding all uncertain pseudo-labels, the framework distinguishes between moderately uncertain samples, which may still be informative, and extremely uncertain ones, which are likely unreliable.

The dynamic masking strategy formally determines which samples should be excluded. For each unlabeled sample, a binary mask m_i is applied:

$$m_i = \begin{cases} 1. & if\ H(p_i) < \epsilon \\ 0. & if\ H(p_i) \geq \epsilon \end{cases} \tag{7}$$

Where $H(p_i)$ is the entropy of the class probability distribution for the $i-th$ samples, and ϵ is the threshold controlling the exclusion of highly uncertain pseudo-labels. By filtering out only these unreliable samples, the strategy stabilizes training and reduces the risk of propagating noisy pseudo-labels. Unlike conventional methods with a fixed ϵ, our framework employs a dynamic entropy threshold that evolves during training, computed as $\epsilon_t = \mu_t + k_t \sigma_t$, where μ_t and σ_t denote the exponentially-weighted moving average(EMA) mean and standard deviation of batch entropies at epoch t. k_t is a scheduled factor that gradually relaxes the gate; in early epochs $\epsilon_t \in [\epsilon_{min}, \epsilon_{max}]$ for stability and compute entropies from weak view $p_i = softmax\left(logits_i^w\right)$. Complementing masking, the uncertainty weighting mechanism adjusts the contribution of each unlabeled sample. The entropy of each prediction is normalized as:

$$e_i = H(p_i) / \log C \tag{8}$$

Where C is the number of classes, and converted into a bell-shaped weight:

$$w_i = 4 e_i (1 - e_i) \tag{9}$$

This weight $w_i \in [0,\ 1]$ peaks at $e_i = 0.5$, down-weights near certain and ambiguous samples, and emphasizes the moderately uncertain regime that tends to be most informative, Empirically, this reduces confirmation bias toward majority and easy classes while avoiding over trusting highly ambiguous pseudo-labels. The overall unsupervised loss thus becomes a weighted and masked combination of cross-entropy and Bhattacharyya divergence:

$$L_{unsup} = \frac{\sum_i w_i m_i (L_{ce}(x_i) + D_{Bh}(x_i))}{\sum_i w_i m_i} \tag{10}$$

Where both weighting and masking operate jointly to emphasize useful samples while suppressing extreme noise. In optimization, we use $L_{total} = \lambda_s L_s + \lambda_u L_{unsup}$; when $w_i \equiv 1$ and $m_i \equiv 1$, Eq. collapses to $L_{ce} + D_{bh}$ as in Eq. (6), explicitly connecting Eq. (10). For class-imbalance scenarios, an optional distribution-alignment correction can be applied to weak-view probabilities before masking to reduce prior skew without changing Eqs. (7)–(10). The model's uncertainty estimates are continuously updated during training, allowing the weighting and masking mechanisms to adapt dynamically to the evolving decision boundaries. Figure 3

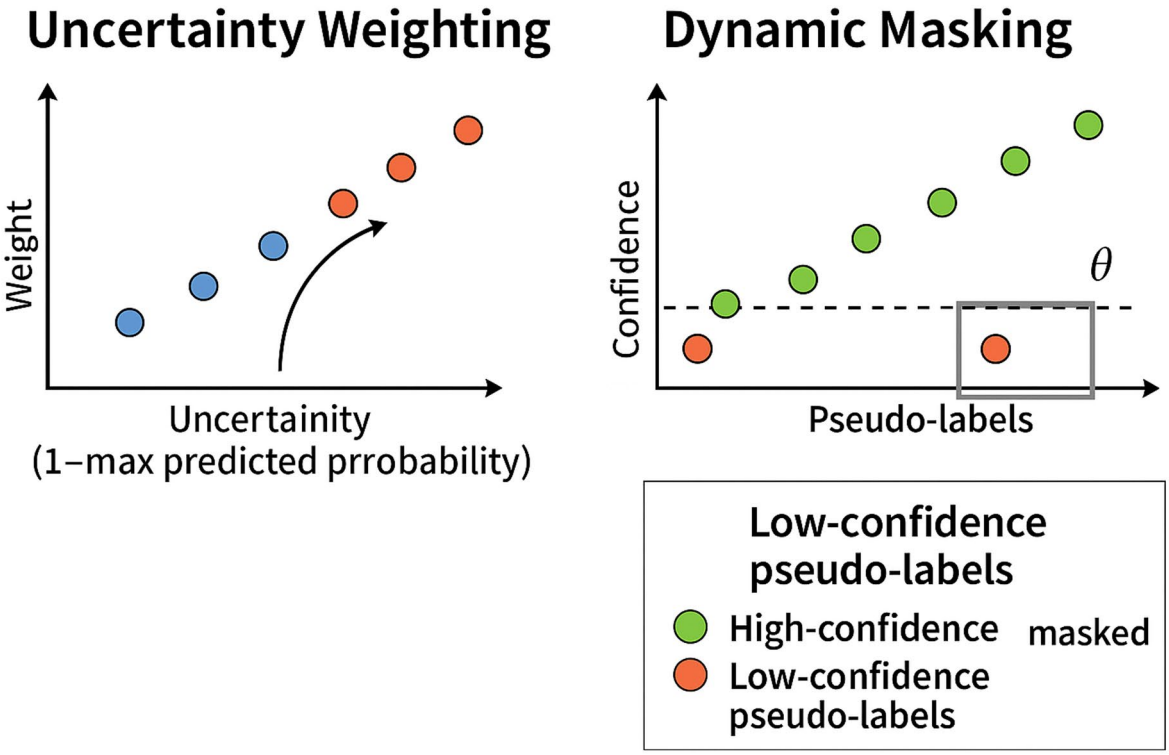

Fig. 3. Uncertainty (1-max predicted probability) and Dynamic Masking in Pseudo-Labeling.

demonstrates how uncertainty weighting is used to guide the training process, with uncertain data points given higher importance. Dynamic masking deliberately filters pseudo-labels with high entropy while preserving moderately uncertain ones, thereby limiting error propagation yet maintaining access to informative unlabeled signal.

Algorithm pseudocode

The following pseudocode concisely summarizes the proposed training loop and its components. Each step is annotated with the corresponding equations (Eqs. 1–10), making explicit where the supervised loss, masked unlabeled cross-entropy, Bhattacharyya alignment, and uncertainty mechanisms. An optional distribution-alignment correction (DisAlign) is indicated before masking to mitigate class-prior skew under imbalance.

Inputs:

- $Labeled\ batch\ X = \{(x_b, y_b)\}\ for\ b = 1..B$
- $Unlabeled\ batch\ U = \{u_b\}\ for\ b = 1..(\mu B)$
- $Confidence\ threshold\ \tau, unlabeled\ ratio\ \mu$
- $Loss\ weights\ \lambda_s, \lambda_u, temperature\ T$
- $Number\ of\ classes\ C$
- $Dynamic\ entropy\ threshold\ \varepsilon_t = \mu_t + k_t \cdot \sigma_t\ (EMA\ mean\ \mu t\ and\ std\ \sigma t\ over\ batch\ entropies)$
- $Optional: DistAlign\ for\ prior\ correction\ before\ masking$

$Supervised\ loss\ on\ labeled\ data:$
$L_s = (1/B) \cdot \sum_{b=1}^{B} H(y_b, p_m(y \mid A_w(x_b)))$ // $cross-entropy\ on\ weakly\ augmented\ labeled\ data$

For $b = 1\ to\ \mu B\ do$ // $loop\ over\ unlabeled\ samples$

$logits_w \leftarrow f_\theta(A_w(u_b))$ // $weak\ view$

$logits_s \leftarrow f_\theta(A_s(u_b))$ // $strong\ view$

$p_w \leftarrow softmax(logits_w)$

if $DistAlign\ enabled:$ // $optional\ distribution\ alignment\ pre-masking$

$p_w \leftarrow DistAlign(p_w)$

$H_b \leftarrow H(pw)$ // $entropy\ of\ weak\ prediction$

$E_b \leftarrow H_b / log(C)$ // $normalize\ entropy\ to\ [0,1]$

$m_b \leftarrow 1\ if\ H_b < \varepsilon_t\ else\ 0$ // $dynamic\ masking\ of\ extremely\ uncertain\ samples$

$w_b \leftarrow 4 \cdot e_b \cdot (1 - e_b)$ // $bell-shaped\ weight\ peaks\ at\ mid-uncertainty$

$q_b \leftarrow softmax(log(p_w) / T)$ // $sharpened\ pseudo-label\ (or\ one\_hot(argmax\ p_w))$

$L_{ce\_b} \leftarrow 1[max(p_w) \geq \tau] \cdot H(qw, softmax(logits_s))$ // $CE\ for\ high-confidence, masked\ later$

$\hat{p}_s \leftarrow softmax(logits_s / T)$

$D_{Bh\_b} \leftarrow -ln \sum_c \sqrt{(q_b[c] \cdot \hat{p}_s[c])}$ // $Bhattacharyya\ divergence\ for\ view\ alignment$

$Accumulate\ per-sample\ terms:$

$num\ +\!= w_b \cdot m_b \cdot (L_{w_b} + D_{Bh_b})$

$denom\ +\!= w_b \cdot m_b$

End For

$Unlabeled\ objective\ (uncertainty-normalized)$

$L_{unsup} = num / max(denom, \varepsilon)$ // $\varepsilon\ is\ a\ small\ constant\ to\ avoid\ divide\ by\ zero$

$Total\ objective$

$L_{total} = \lambda_s \cdot L_s + \lambda_u \cdot L_{unsup}$

$Return\ L_{total}, and\ optionally:$

$util\_ratio = (1/(\mu B)) \cdot \sum_b m_b$ // $fraction\ of\ unlabeled\ samples\ used\ this\ step$

Algorithm.

Experimantal setup

In this section, we describe the experimental setup used to evaluate the performance of the proposed semi supervised learning SSL framework. We detail the datasets, evaluation metrics, comparison with other SSL methods such as, Pseudo-Label, Π-Model, MixMatch, ReMixMatch, VAT, FreeMatch, AdaMatch, and FixMatch[5], as well as imbalance- and noise-aware SSL baselines. We also implementation details and hyper-parameter settings. Our experiments are designed to assess the robustness of the proposed method in various settings,

particularly in the presence of noisy or imbalanced data, which are common challenges in real world-machine learning tasks. All results are averaged over three seeds (mean ± std).

Datasets

We evaluate the proposed model on several benchmark datasets commonly used for SSL tasks. These datasets consist of a mixture of labeled and unlabeled data, allowing us to assess how effectively the model leverages unlabeled data for training.

- *CIFAR-10*: A standard dataset for image classification, containing 60,000 32 × 32 color images in 10 classes, with (50,000 train/10,000 test)[38]. We follow common SSL protocols using low-label splits (e.g., 40-label and 250-label settings with class-balanced selection), and treat the remaining images as unlabeled.
- *SVHN (Street View House Numbers)*: A real-world dataset consisting of digit images 32 × 32 extracted from Google Street View. the dataset contains 73,257 label training images and about 26,000 labeled testing images[39]. We evaluate at low-label (e.g., 250 labels) and medium-label (e.g., 1000 labels) settings; the remainder forms the unlabeled pool.
- *STL-10*: Inspired by CIFAR-10 dataset, with 5,000 labeled training images, 100,000 unlabeled images, and 8,000 testing images[40]. We use the standard SSL protocol with a 1000-label subset when explicitly noted.
- *Imbalanced CIFAR-10 (long-tailed)*: To evaluate robustness under class imbalance, we create long-tailed version of CIFAR-10 following a standard imbalance protocol[9]. Class counts decay exponentially according to imbalance ratio ρ. We consider two settings: ρ = 50 (moderate imbalance) and ρ = 100(severe imbalance). Only the labeled subset is imbalanced while the unlabeled pool remains class-balanced, isolating the effect of imbalanced supervision.
- *Noisy Settings*: To further test robustness, we corrupt the labeled subset of CIFAR-10 and SVHN with symmetric label noise at 20% and 40%. The unlabeled pool remains clean to emulate realistic annotation noise while preserving unsupervised signal.

Evaluation metrics

We evaluate the performance of the proposed model using Top-1 Accuracy and analyze the confusion matrix to understand per-class behavior.

- Top-1 Accuracy. The percentage of times the model's top prediction matches the actual label.

$$\text{Top1 Acc} = \frac{\text{No. correct predictions}}{\text{total number of predictions}} \times 100 \tag{10}$$

- Confusion Matrix: A table used to evaluate the performance of a classification model. It shows the actual versus predicted classifications, giving insight into how well the model is performing across all classes.
- Macro-F1. Reported in imbalance settings to account for class-frequency skew and to complement accuracy.
- Calibration metrics. For divergence ablations (Bhattacharyya vs. KL/JS/Wasserstein), we also report Expected Calibration Error ECE and Negative Log-Likelihood NLL to quantify stability and calibration under noisy and long-tailed regimes.

Comparison with baseline methods

To evaluate the effectiveness of the proposed method, we compare it against several state-of-the-art semi-supervised learning algorithms. These baseline methods have been widely used in previous research and include:

- FixMatch: A strong SSL algorithm based on consistency regularization. It utilizes weak and strong augmentations and pseudo-labeling for unlabeled data.
- MixMatch: A method that combines the concepts of consistency regularization and pseudo-labeling, using both weak and strong augmentations and mixing labeled and unlabeled data to enhance learning.
- Pseudo-labeling: One of the most basic approaches in SSL, where the models' predictions on unlabeled data are used as pseudo-labels and added to the training data.
- VAT (Virtual Adversarial Training): An SSL method that introduces adversarial perturbation to the input data to regularize the model and improve its generalization.
- ReMixMatch, FreeMatch, AdaMatch: Are domain and threshold adaptation and alignment.

We compare the proposed model with these baselines in term of classification top-1 accuracy. Our goal is to show that the proposed model outperforms or at least matches the performance of existing methods, particularly in challenging settings such as noisy and imbalanced data.

Implementation details

The model is implemented using Semilearn, a popular deep learning framework from Microsoft. The training is performed using the Stochastic Gradient Descent SGD optimizer with a momentum of 0.9; the initial learning rate is 0.03; the total number of training iterations is 2^{20}; the labeled batch size is 64; the ratio of unlabeled data μ is set to 7; the confidence threshold τ is set to 0.95; and the sharpening temperature T is set to 0.5. The weak augmentation for the unlabeled data includes standard random cropping and flipping transformations, while the

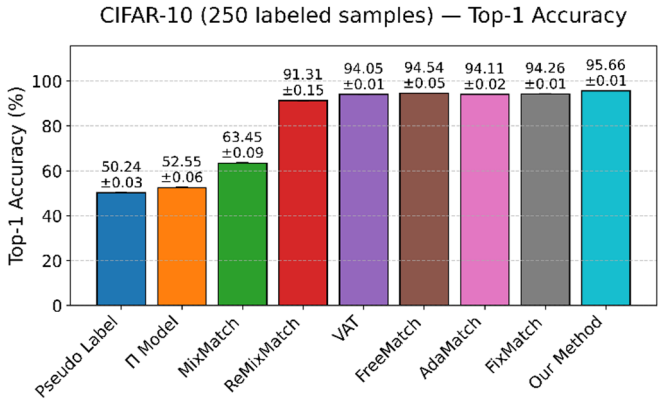

Fig. 4. Top-1 Accuracy on CIFAR-10 for Various SSL Methods using 250 labeled samples.

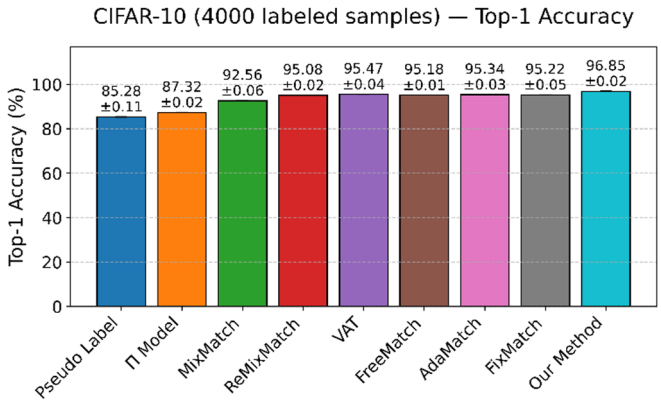

Fig. 5. Top-1 Accuracy on CIFAR-10 for Various SSL Methods using 4000 labeled samples.

RandAugmen has been used in the strong augmentation. For class-imbalance experiments, we optionally apply distribution alignment (DisAlign) to weak-view probabilities before masking.

The training process starts by initializing the ViT backbone with pre-trained weights, applying weak augmentations to labeled data for supervised learning, and generating pseudo-labels from weak views for unlabeled samples. The supervised loss cross-entropy; the unlabeled objective combines masked cross-entropy with Bhattacharyya divergence to align weak and strong predictions of the same sample. Dynamic entropy masking with $\epsilon_t = \mu_t + k_t\sigma_t$ (EMA statistics) filters only extremely uncertain pseudo-labels, while the entropy-based weight $w_i = 4e_i(1 - e_i)$ emphasizes moderately uncertain samples. We conduct sensitivity studies by varying $\tau \in \{0.80, 0.90, 0.95\}$, $T \in \{0.5, 0.7, 1.0\}$, and scheduling k_t to demonstrate the effect of the mask and weighting. Unless specified, results are averaged over three random seeds; validation sets are used for model selection, and the final model is evaluated on the held-out test set. A supervised-only baseline (no unlabeled loss) is trained under the same schedule for comparison.

Results and discussion

In this section, we present the results of our experiments, evaluating the performance of the proposed semi-supervised learning SSL framework on multiple benchmark datasets, including CIFAR-10, SVHN and STL-10. We compare the performance of our model with several state-of-the-art SSL methods, such as Pseudo Label, Π Model, MixMatch, ReMixMatch, VAT, FreeMatch, AdaMatch and FixMatch. We focus on Top-1 accuracy as the evaluation metric and report macro-averaged F1 (Macro-F1) for class-imbalanced protocols and calibration metrics (ECE/NLL) for the divergence ablations. Our results also demonstrate the importance of Bhattacharyya regularization, which consistently improves distributional alignment between weak and strong augmentations and enhances robustness under noise and imbalance.

Results on CIFAR-10

CIFAR-10 is a well-known dataset for image classification with 10 classes of 32×32 color images. We evaluate two labeled-data regimes: 250 labels and 4000 labels (Figs. 4 and 5). Figure 4 compares Top-1 accuracy for the 250-label case; our method achieves 95.66% with FixMatch as the second-best performance. While the fully supervised cifar-10 achieves 95.4%. Moreover, the results on 4000 labels (Fig. 5), our method reaches 96.85% and

Method	250 labels ECE	250 labels NLL	4000 labels ECE	4000 labels NLL
FixMatch	0.081 ± 0.02	0.42 ± 0.04	0.052 ± 0.01	0.29 ± 0.06
Ours (KL in unlabeled align.)	0.071 ± 0.05	0.39 ± 0.02	0.048 ± 0.05	0.27 ± 0.03
Ours (Bhattacharyya)	**0.057** ± 0.06	**0.33** ± 0.03	**0.041** ± 0.01	**0.23** ± 0.02

Table 1. CIFAR-10 (class-balanced) calibration: ECE/NLL under 250 and 4000 labels.

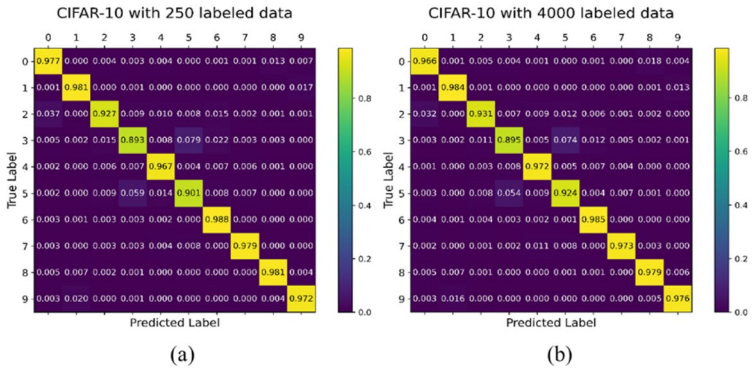

Fig. 6. Confusion matrix of proposed method on CIFAR10. (a) using 250 labeled samples. (b) using 4000 labeled samples.

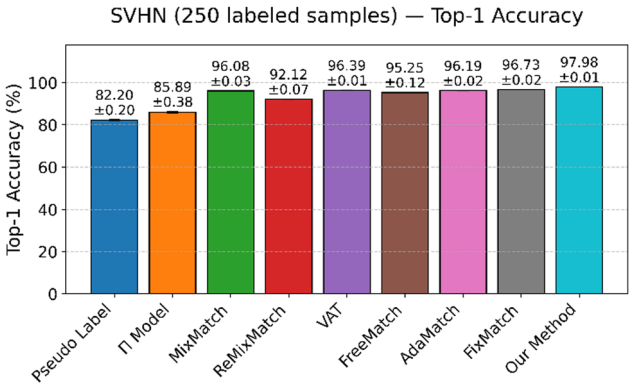

Fig. 7. Top-1 Accuracy on SVHN for various SSL Methods with 250 labeled samples.

remains on par or better than recent SSL baselines. Across both regimes, ECE decreases compared to FixMatch (relative reduction $\sim\sim 20-30\%$), indicating more calibrated pseudo-labels (Table 1).

Additionally, we include the confusion matrices for 250-label and 4000-label settings (Fig. 6) to visualize per-class behavior. Compare to baselines, minority-like categories show reduced confusion with visually similar classes, consistent with the entropy-weighted emphasis on moderately uncertain samples.

Results on SVHN

SVHN comprises real-world house-number images. We evaluate 250-label and 1000-label settings (Figs. 7 and 8). The proposed method achieves 97.98% accuracy with (250 labels) and 98.02% with (1000 labels), outperforming alternatives in both regimes. Gains are more pronounced at 250 labels, suggesting that uncertainty weighted training is particularly beneficial when labeled supervision is scarce. While the fully supervised method on SVHN achieves 97.1%. We also observe lower ECE and NLL than KL-based and JS-based variants in our divergence ablations, indicating more stable training under SVHN's complex background noise (Table 2).

Results on STL-10

STL-10 contains higher resolution 96×96 images. With 1000 labels (Fig. 9), our method achieves 94.39%, outperforming FreeMatch and FixMatch. While the fully supervised method on STL-10 dataset achieves 94.1%. The symmetric overlap-aware Bhattacharyya term contributes noticeably in the higher-resolution dataset by aligning distributions across augmentations, resulting in better calibration of pseudo-labels.

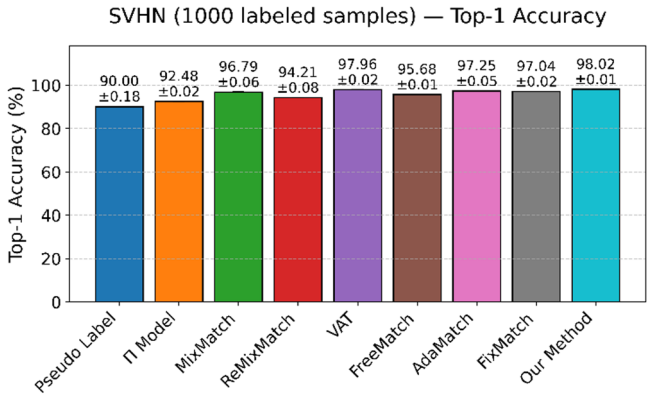

Fig. 8. Top-1 Accuracy on SVHN for various SSL Methods with 1000 labeled samples.

Divergence in unlabeled align	Top-1 Acc (%)	ECE	NLL
Ours w/KL	97.10 ± 0.01	0.034 ± 0.06	0.29 ± 0.02
Ours w/Jensen-Shannon	97.32 ± 0.04	0.030 ± 0.01	0.27 ± 0.01
Ours w/approx. Wasserstein	97.41 ± 0.01	0.028 ± 0.02	0.26 ± 0.03
Ours w/Bhattacharyya	**97.98** ± 0.02	**0.023** ± 0.01	**0.22** ± 0.01

Table 2. SVHN (250 labels) divergence ablation: Top-1 ACC/ECE/NLL.

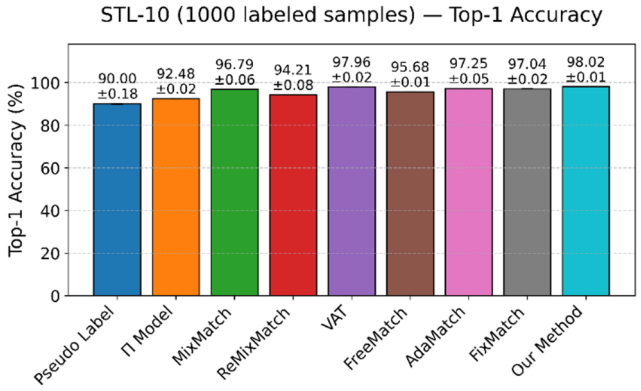

Fig. 9. Top-1 Accuracy on STL 10 for various SSL methods with 1000 labeled samples.

Method	ρ = 50 Acc (%)	ρ = 50 Macro-F1	ρ = 100 Acc (%)	ρ = 100 Macro-F1
FixMatch	87.3 ± 0.3	82.0 ± 0.5	83.9 ± 0.8	79.0 ± 0.6
FreeMatch	88.2 ± 0.2	83.0 ± 0.7	84.7 ± 0.5	80.1 ± 0.4
AdaMatch	89.0 ± 0.5	83.9 ± 0.4	85.5 ± 0.2	81.0 ± 0.4
Ours (no DA)	**92.4 ± 0.3**	**85.9 ± 0.2**	**89.6 ± 0.5**	**84.5 ± 0.5**
Ours (+DA)	**93.1 ± 0.4**	**87.2 ± 0.3**	**90.5 ± 0.2**	**85.9 ± 0.3**

Table 3. Imbalanced CIFAR-10 (labeled long-tailed), Top-1 accuracy (%) and Macro-F1.

In Fig. 9, it is clear that the proposed method performs robustly across all models, achieving the highest Top-1 accuracy on STL 10. This demonstrates the effectiveness of our method in handling different assets and label amounts.

Dataset	Method	20% Noise	40% Noise
CIFAR-10	FixMatch	91.2 ± 0.4	85.6 ± 0.2
	FreeMatch	92.0 ± 0.1	86.3 ± 0.1
	AdaMatch	92.8 ± 0.4	87.1 ± 0.4
	Ours	**94.7 ± 0.9**	**92.1 ± 0.3**
SVHN	FixMatch	95.4 ± 0.5	91.0 ± 0.5
	FreeMatch	95.9 ± 0.3	91.8 ± 0.3
	AdaMatch	96.2 ± 0.6	92.3 ± 0.4
	Ours	**97.6 ± 0.2**	**96.2 ± 0.7**

Table 4. CIFAR-10 and SVHN with symmetric label noise on labeled subset. Top-1 accuracy (%).

Dataset	Method variant	20% Noise ECE	20% Noise NLL	40% Noise ECE	40% Noise NLL
CIFAR-10	Ours w/KL	0.072 ± 0.01	0.58 ± 0.02	0.093 ± 0.01	0.67 ± 0.04
	Ours w/Jensen-Shannon	0.067 ± 0.03	0.55 ± 0.03	0.088 ± 0.01	0.62 ± 0.02
	Ours w/Bhattacharyya	**0.059 ± 0.01**	**0.49 ± 0.01**	**0.068 ± 0.02**	**0.52 ± 0.02**
SVHN	Ours w/KL	0.041 ± 0.04	0.44 ± 0.01	0.062 ± 0.03	0.57 ± 0.01
	Ours w/Jensen-Shannon	0.038 ± 0.01	0.41 ± 0.02	0.059 ± 0.04	0.53 ± 0.01
	Ours w/Bhattacharyya	**0.36 ± 0.01**	**0.36 ± 0.03**	**0.047 ± 0.02**	**0.45 ± 0.02**

Table 5. Calibration under noisy supervision (ECE/NLL).

Results on imbalanced CIFAR-10

we construct long-tailed CIFAR-10 where labeled class counts decay exponentially with imbalance ratio ρ. Two imbalance levels are considered: $\rho = 50$ (moderate) and $\rho = 100$ (severe). Only the labeled subset is imbalanced, while the unlabeled pool remains balanced, ensuring that the challenge arises solely from supervision imbalance. Table 3 shows that our framework consistently outperforms other baselines, with a 5.7% gain in Top-1 accuracy at $\rho = 100$ and +3–5 pp Macro-F1 improvement, indicating better minority-class recall. Optional distribution alignment (applied before masking) further reduces prior skew and improves Macro-F1 by $\approx 1 - 2$pp. The gain is largely due to the Bhattacharyya divergence, which aligns predictions across weak and strong augmentation.

Results on noisy settings

For label-noise robustness, we corrupt the labeled subset of CIFAR-10 and SVHN with symmetric noise at 20% and 40%, leaving the unlabeled data clean. Table 4 shows that our method degrades more gracefully than baselines, with a +6.5-pp top-1 margin over FixMatch at 40% noise on CIFAR-10 and +5.2-pp on SVHN. Calibration also improves (ECE and NLL drop relative to KL/JS variants) as in (Table 5).

Data utilization

In the CIFAR-10 dataset with only 250 labeled examples, our model unequivocally demonstrates its superior capacity to exploit unlabeled data, as evidenced by the model confidence curves over 4000 training iterations. As depicted in Fig. 10, our model maintains a consistently high confidence level remaining above the overconfidence threshold of 0.95—throughout the entire training period. In contrast, the other models exhibit a gradual yet volatile increase in confidence, peaking near 0.9 but never reaching the robust levels observed with the proposed method. This stark difference highlights the strength of our approach; by integrating advanced consistency regularization, a hybrid loss function, dynamic masking, and the uncertainty-weighted training mechanism. This work effectively leverages all unlabeled images to produce stable and reliable pseudo-labels. As a result, our model not only mitigates the detrimental impact of noisy pseudo-labels but also reinforces the learning signal across training iterations. The consistent high confidence indicates that the proposed method builds a more stable decision boundary and achieves a more refined alignment between predictions over different augmentation variants. These outcomes are particularly compelling given the limited amount of labeled data, underscoring that the additional supervisory signal extracted from the entirety of the unlabeled data significantly enhances overall generalization performance.

Ablation studies

Ablation experiments were conducted on CIFAR-10 with 250 labels to quantify the contribution of each component. The results are summarized in Table 6. Removing uncertainty weighting reduces accuracy by 2.1%, as the model fails to leverage moderately uncertain samples. Eliminating dynamic masking decreases accuracy by 1.8%, due to error propagation from low-confidence pseudo-labels. Replacing the Bhattacharyya divergence with KL divergence lowers accuracy by 1.2%, confirming the advantage if a symmetric divergence measure. When all components are present, the full model achieves the highest accuracy and lower ECE.

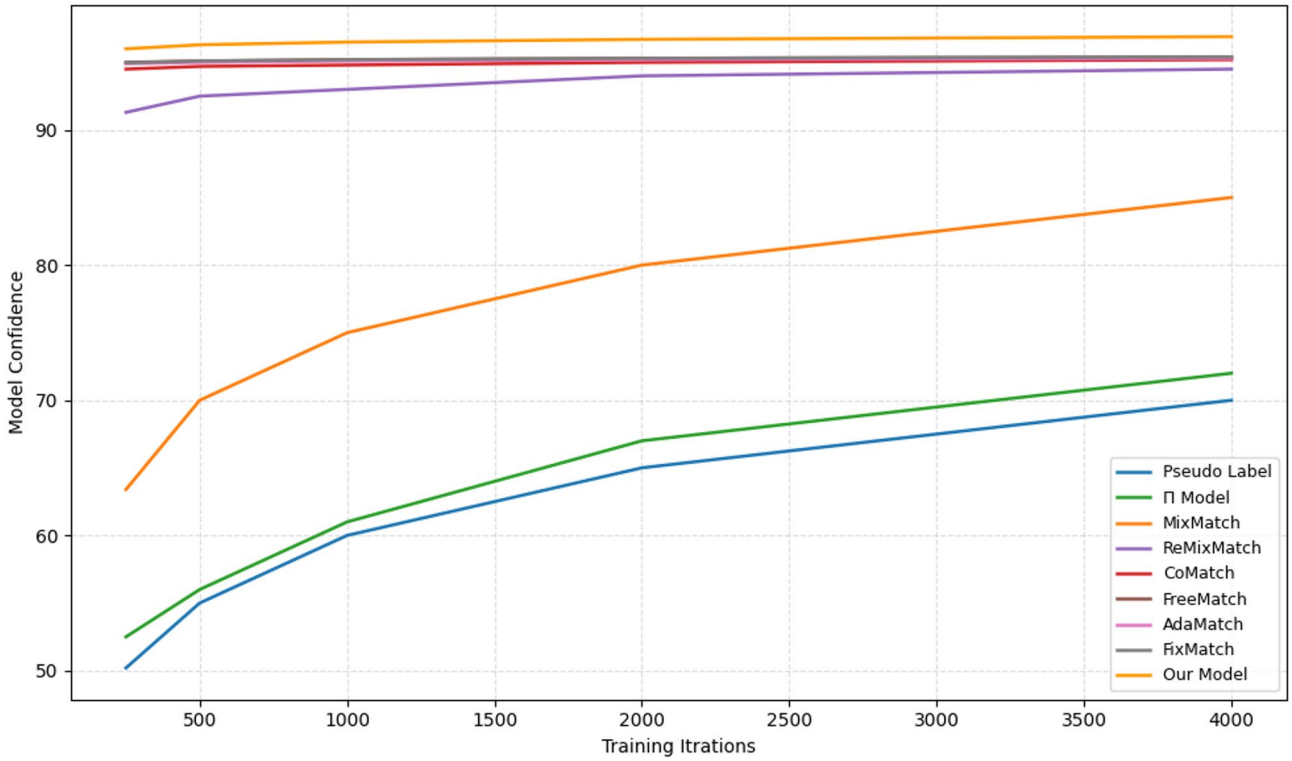

Fig. 10. Comparison of Unlabeled Data Utilization on CIFAR10-250.

Model Variant	Accuracy (%)	ECE	NLL
Full Model **(Ours)**	95.66 ± 0.4	0.057 ± 0.01	0.33 ± 0.03
w/o Uncertainty Weighting	93.56 ± 0.5	0.073 ± 0.03	0.41 ± 0.02
w/o Dynamic Masking	93.82 ± 0.6	0.070 ± 0.04	0.40 ± 0.01
w/o Bhattacharyya Divergence	94.46 ± 0.5	0.063 ± 0.07	0.36 ± 0/07
Bhattacharyya → KL Divergence	94.39 ± 0.5	0.071 ± 0.01	0.39 ± 0.02
Masking with Weighting	95.01 ± 0.3	0.061 ± 0.03	0.35 ± 0.01
Weighting with Bhattacharyya	95.39 ± 0.2	0.059 ± 0.02	0.34 ± 0.03

Table 6. Ablation study on CIFAR-10 with 250 labeled samples (accuracy % ± std).

Discussion

By studying the related works and baselines, the main difference between our method and other approaches of semi-supervised learning is how they deal with unlabeled data under uncertainty, noise, and imbalance. Prior approaches often emphasize high-confidence samples and may discard informative, moderately uncertain ones. Our framework combines dynamic masking, entropy-weighted training and symmetric view alignment via Bhattacharyya regularization. The results from CIFAR-10, SVHN and STL 10 demonstrate the superior performance of the proposed method across multiple data sets and label configurations in term of Top-1 accuracy, and Macro-F1/ECE/NLL for imbalance and noise setting. For CIFAR-10, the model achieves 95.66% accuracy with just 250 labeled samples, outperforming all other models. This shows the method's strength in utilizing unlabeled data effectively, particularly when labeled data is scarce. The given approach demonstrates a notable performance on Street View House Numbers (SVHN) dataset, reaching an accuracy of 97.98% trained with only 250 labeled samples. Such an outcome represents a significant improvement over the results of FixMatch and other competitive methods, which demonstrates the potential of the model to make consistent predictions in the context of scarce labeled data. The model achieves an accuracy of 94.39% on the STL-10 data set when trained under 1000 labeled data, which is higher than FreeMatch and FixMatch. These results may indicate that the model is effective not only in situations when strong training data can be used but is also robust to different datasets. The achieved improvement can be explained by the uncertainty-based training scheme, which focuses on the examples that are naturally challenging to classify in the training process, and the dynamic masking scheme that reduces interference by low-confidence pseudo-labels. Therefore, the proposal teaches the network on the most credible data points.

Limitation and weaknesses

While the proposed method shows clear benefits under label scarcity, class imbalance, and noisy supervision, its improvements on clean, class-balanced settings are modest (typically ≤ 0.3–0.8% in Top-1) because far fewer samples fall into the "extremely uncertain" region; in these cases, our policy behaves similarly to high-performing baselines. The approach is also sensitive to a small set of hyperparameters (τ , T, k_t): overly conservative masking (large k_t) can under-use informative data, whereas overly permissive settings can admit noisy pseudo-labels (see Sensitivity to ϵ_t , τ). In addition, the dual-view training and EMA statistics introduce extra compute and training time. Finally, our experiments focus on image classification; generalization to non-vision domains remains to be validated. These trade-offs motivate future work on lighter training schedules, adaptive threshold and temperature schedules, and broader domain evaluations.

Conclusion

The paper introduces a new semi-supervised learning (SSL) framework that makes the best use of both labeled and unlabeled data by combining uncertainty-weighted training, dynamic masking, and a hybrid loss function that contains Bhattacharyya divergence loss. Concretely, the method pairs an entropy-aware weighting scheme with a dynamic entropy threshold to defer only extremely uncertain pseudo-labels, and adds a symmetric, overlap-aware Bhattacharyya-regularized weak and strong alignment term to stabilize training and improve calibration. Experimental tests demonstrate better results than strong SSL baselines on CIFAR-10, SVHN and STL-10, especially in cases when the annotation is limited. Across challenging regimes, the approach yields consistent gains up to +3–5 pp Top-1, +3–5 pp macro-F1 on long-tailed CIFAR-10, and 20–30% relative reductions in ECE/NLL, while remaining competitive on class-balanced settings. By focusing on hard-to-label instances through uncertainty-aware selection, the method enhances generalization and robustness. The experimental results show that the proposed method in not only efficient but also generalizes well across various datasets, making it promising approach for SSL tasks in real-world scenarios. The limitations of this work represented by the gains are modest on clean class balanced data performance is sensitive to τ , T, k_t (risk of over or under masking); and training incurs extra compute due to dual views and EMA. Future work will reduce this overhead and explore adaptive schedules and broader domains.

Data availability

The data are publicly available at [Github] (https://github.com/mhmdghazal1981/deep-learning-benchmark-datasets).

Received: 22 September 2025; Accepted: 20 November 2025
Published online: 27 November 2025

References

1. Li, Q., Chen, G. & Wang, D. Mineral prospectivity mapping using Semi-supervised machine learning. *Math. Geosci.* **57**, 275–305. https://doi.org/10.1007/S11004-024-10161-6/METRICS (2025).
2. Pani, K. & Chawla, I. Examining the quality of learned representations in self-supervised medical image analysis: a comprehensive review and empirical study. *Multimed Tools Appl.* **84**, 6295–6325. https://doi.org/10.1007/S11042-024-19072-4/METRICS (2025).
3. Han, K. et al. Deep semi-supervised learning for medical image segmentation: A review. *Expert Syst. Appl.* **245**, 123052. https://doi.org/10.1016/J.ESWA.2023.123052 (2024).
4. Sosea, T. & Caragea, C. MarginMatch: Improving semi-supervised learning with pseudo-margins. in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 15773–15782 (2023).
5. Yang, X., Song, Z., King, I. & Xu, Z. A survey on deep Semi-Supervised learning. *IEEE Trans. Knowl. Data Eng.* **35**, 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219 (2023).
6. Jiao, R. et al. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Comput. Biol. Med.* **169**, 107840. https://doi.org/10.1016/J.COMPBIOMED.2023.107840 (2024).
7. Duarte, J. M. & Berton, L. A review of semi-supervised learning for text classification. *Artif. Intell. Rev.* **56**, 9401–9469. https://doi.org/10.1007/S10462-023-10393-8/METRICS (2023).
8. Serrano-Pérez, J. & Sucar, L. E. Semi-supervised hierarchical multi-label classifier based on local information. *Int. J. Approx Reason.* **181**, 109411. https://doi.org/10.1016/J.IJAR.2025.109411 (2025).
9. Sohn, K. et al. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* (2020). https://arxiv.org/pdf/2001.07685 (accessed July 16, 2025).
10. Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017-December**, 1196–1205 (2017). https://arxiv.org/pdf/1703.01780 (accessed July 16, 2025).
11. Abdulrazzaq, M. M. et al. Consequential advancements of self-supervised learning (SSL) in deep learning contexts. *Math.* **12**, 758. https://doi.org/10.3390/MATH12050758 (2024).
12. Li, J., Yang, M. & Feng, M. Confidence-Guided Open-World Semi-supervised learning. *Lect Notes Comput. Sci.* **14428**, 87–99. https://doi.org/10.1007/978-981-99-8462-6_8 (2024).
13. Lee, D. H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. in *Work. Challenges Represent. Learn.* 896 (ICML, 2013).
14. Min, Z., Bai, J. & Li, C. Leveraging local variance for Pseudo-Label selection in Semi-supervised learning. *Proc. AAAI Conf. Artif. Intell.* **38**, 14370–14378. https://doi.org/10.1609/AAAI.V38I13.29350 (2024).
15. Jeong, J., Lee, S. & Kwak, N. Selective Self-Training for semi-supervised Learning (n.d.).
16. Xie, Q., Luong, M. T., Hovy, E. & Le, Q. V. Self-training with Noisy Student improves ImageNet classification. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 10684–10695. https://doi.org/10.1109/CVPR42600.2020.01070 (2019).
17. Yu, K., Ma, H., Lin, T. R. & Li, X. A consistency regularization based semi-supervised learning approach for intelligent fault diagnosis of rolling bearing. *Measurement* **165**, 107987. https://doi.org/10.1016/J.MEASUREMENT.2020.107987 (2020).
18. Miyato, T., Maeda, S. I., Koyama, M. & Ishii, S. Virtual adversarial training: A regularization method for supervised and Semi-Supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1979–1993. https://doi.org/10.1109/TPAMI.2018.2858821 (2017).
19. Berthelot, D. et al. MixMatch: A holistic approach to semi-supervised learning. *Adv. Neural Inf. Process. Syst.* (2019). https://arxiv.org/pdf/1905.02249 (accessed July 16, 2025).

20. Qiao, F. & Peng, X. Uncertainty-guided model generalization to unseen domains. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 6786–6796. https://doi.org/10.1109/CVPR46437.2021.00672 (2021).
21. Ji, K. et al. Uncertainty-guided learning for improving image manipulation detection. *Proc. IEEE Int. Conf. Comput. Vis.* 22399–22408. https://doi.org/10.1109/ICCV51070.2023.02052 (2023).
22. Xu, R. et al. Uncertainty-guided cross teaching semi-supervised framework for histopathology image segmentation with curriculum self-training. *Appl. Soft Comput.* **180**, 113328. https://doi.org/10.1016/J.ASOC.2025.113328 (2025).
23. Zhang, Y., Gong, Z., Zhao, X. & Yao, W. Uncertainty guided ensemble self-training for semi-supervised global field reconstruction. *Complex. Intell. Syst.* **10**, 469–483. https://doi.org/10.1007/S40747-023-01167-4/TABLES/7 (2024).
24. Wang, Y. et al. Freematch:Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246* (2022).
25. Berthelot, D., Roelofs, R., Sohn, K., Carlini, N. & Kurakin, A. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732* (2021).
26. Jiang, Z., Zhao, L., Lu, Y., Zhan, Y. & Mao, Q. A semi-supervised resampling method for class-imbalanced learning. *Expert Syst. Appl.* **221**, 119733 (2023).
27. Jiang, Z., Tang, N., Sun, J. & Zhan, Y. Combining various training and adaptation algorithms for ensemble few-shot classification. *Neural Netw.* **185**, 107211 (2025).
28. Berthelot, D. et al. ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *8th Int. Conf. Learn. Represent. ICLR (2020)* (2019) https://arxiv.org/pdf/1911.09785 (accessed July 16, 2025).
29. Kim, J. et al. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Adv. Neural. Inf. Process. Syst.* **33**, 14567–14579 (2020).
30. Wei, C., Sohn, K., Mellina, C., Yuille, A. & Yang, F. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 10857–10866 (2021).
31. Oh, Y., Kim, D. J. & Kweon, I. S. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 9786–9796 (2022).
32. Zhen, J., Feng, Z. & Niu, B. Prototype-Neighbor networks with Task-Specific enhanced Meta-learning for Few-Shot classification. *Neural Netw.* 107761 (2025).
33. Lee, S., Kim, H. & Chun, D. UCR-SSL: Uncertainty-Based consistency regularization for Semi-Supervised learning. *Int. Conf. Electron. Inform. Commun. ICEIC 2023*. https://doi.org/10.1109/ICEIC57457.2023.10049938 (2023).
34. Qiu, Z., Gan, W., Yang, Z., Zhou, R. & Gan, H. Dual uncertainty-guided multi-model pseudo-label learning for semi-supervised medical image segmentation. *Math. Biosci. Eng.* **21**, 2212–2232. https://doi.org/10.3934/MBE.2024097 (2024).
35. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for image recognition at scale. *ICLR 2021–9th Int. Conf. Learn. Represent.* (2020). https://arxiv.org/pdf/2010.11929 (accessed July 16, 2025).
36. Wang, Z., Li, T., Zheng, J. Q. & Huang, B. When CNN Meet with ViT: Towards Semi-Supervised Learning for Multi-Class Medical Image Semantic Segmentation, *Lect Notes Comput. Sci. 13807 LNCS* 424–441. https://doi.org/10.1007/978-3-031-25082-8_28. (2022).
37. Xuan, G. et al. Feature selection based on the Bhattacharyya distance. *Proc. - Int. Conf. Pattern Recognit.* **3**, 1232–1235. (2006). https://doi.org/10.1109/ICPR.2006.558
38. Krizhevsky, A. G. Hinton, others, learning multiple layers of features from tiny images (2009).
39. Netzer, Y., Wang, T., Coates, A., Bissacco, A. & Wu, B. A.Y. Ng, others, reading digits in natural images with unsupervised feature learning. in *NIPS Work. Deep learn. Unsupervised Featur Learn.* 4 (2011).
40. Coates, A., Ng, A. & Lee, H. *An Analysis of Single-Layer Networks in Unsupervised Feature Learning* 215–223 (2011). https://proceedings.mlr.press/v15/coates11a.html (accessed July 16, 2025).

Author contributions

M.G: Conceptualization, Methodology, Writing, Data testing, Software. J.T: Conceptualization, Supervision, Formal analysis, Reviewing. N.SH: Resources, Writing. S.R: Writing - review & editing.

Funding

This research received no funding.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.